Molecular Property Filters Describing Pharmacokinetics and Drug Binding

A.T. García-Sosa[1], U. Maran[1] and C. Hetényi*,[2]

[1]*Institute of Chemistry, University of Tartu, 14A Ravila, 50411 Tartu, Estonia*

[2]*Departments of Biochemistry and Genetics, Eötvös University, Pázmány sétány 1/C, 1117 Budapest, Hungary*

Abstract: Drug-target binding affinity and pharmacokinetics are equally important factors of drug design. Simple molecular properties such as molecular size have been used as pharmacokinetic and/or drug-likeness filters during chemical library design and also correlated with binding affinity. In the present study, current property filters are reviewed, a collection of their optimal values is provided, and a statistical framework is introduced allowing calibration of their selectivity and sensitivity for drugs. The role of ligand efficiency indices in drug design is also described. It is concluded that the usefulness of property filters of molecular size and lipophilicity is limited as predictors of general drug-likeness. However, they demonstrate increased performance in specific cases, e.g. in central nervous system diseases, emphasizing their future importance in specific, disease-focused library design instead of general drug-likeness filtering.

Keywords: Binding site, entropy, free energy, molecule, pocket, protein, structure, target, logP, Wiener index.

1. INTRODUCTION

The effects of drug molecules are produced by their interactions with one or more macromolecular targets [1, 2], constituents of the human body. Therefore, small molecule drug design strategies involve multiple screening steps [3, 4] using the structure [5] of drug candidates (ligands) in complex with targets and also the corresponding thermodynamic measures of equilibrium binding affinities [6], the free energy changes (ΔG). In general, an appropriate ΔG is a necessary but not a sufficient property of a successful candidate as pharmacokinetic, toxicological, etc. characteristics also influence drug-likeness [7].

Molecular properties of small compounds have been extensively used as descriptors in structure-activity relationships [8, 9]. For example, molecular weight (MW) is atom-type sensitive and related to the molecular size; logP is a measure for partitioning of compounds between lipophylic and aqueous phase; number of heavy atoms (NHA) is the simplest molecular property providing a crude estimate of the size of a molecule; Wiener index, a topological descriptor characterizes the compactness of a molecule and is proportional to the molecular surface area [10, 11]. Such molecular properties were also adopted for the prediction of complex physiological properties and pharmacokinetics: absorption [12], or blood brain barrier penetration [13, 14], and their use culminated in the definition of general drug-likeness ranges. These empirical ranges of the properties were proved to be useful as property filters in the design of compound libraries of drug screening [15-20]. Notably, the selection of high quality (drug-like) compound libraries [3, 21-23] is a primary and key step of the screening process.

Besides their connection to pharmacokinetics, it has been shown in numerous studies that the above size-dependent filters (MW, NHA) are also coupled to ΔG as they correlate with the (maximal) binding affinity achievable by a ligand. To decouple ΔG from ligand size, efficiency indices (EI, also called ligand efficiencies or binding efficiencies) have been defined dividing ΔG by NHA or MW [24].

The present review sketches how complex phenomena of pharmacokinetics and equilibrium binding are coupled with the above molecular properties. An overview of their use is provided, and a summary of available correlations of ligand-based properties with ΔG is assembled. The role of EIs is discussed, and limitations of the general drug-likeness concept are analyzed. Selectivity and sensitivity of the property filters are defined, and a statistical decoupling of ΔG from the properties is suggested for pharmacokinetics-focused analyses. Besides general drug-likeness, disease- and target-specificity is discussed and future perspectives are outlined.

2. MOLECULAR PROPERTY FILTERS DESCRIBING DRUG-LIKENESS

Filtering of large compound sets generated by combinatorial or other techniques [25, 26] is a central issue of library design. As Martin and Critchlow showed [27], merely random selection of compounds for high throughput screening (HTS) is poor both in structural diversity and in distribution of physicochemical properties. Random libraries are systematically biased toward heavy, flexible compounds that have very high or very low lipophilicity and possess inappropriate bioavailability. Thus, the need for effective filtering to produce 'drug-like' libraries was early recognized and several groups have developed filters based on the analysis of molecular property distribution in available drug databases. The present paper is focused on the analysis of simple molecular properties such as MW or logP coupled to both pharmacokinetics and ΔG (Introduction). Other filters including information on e.g. functional groups [28] are beyond the scope of this study.

2.1. Definition of Drug-Likeness

The first drug-likeness studies dealt with pharmacokinetic properties of drug candidates. Lipinski *et al*. [29] found that poor absorption or permeation is more likely if ligand properties such as MW or logP fulfill the 'rule of 5' (Ro5, Table **1**) criterion. Fecik *et al*. [30] also analyzed the relationship between MW and oral bioavailability. Clark and Pickett [31] describe the term general drug-likeness filtering. According to their definition, such filters incorporate substructure searches for toxic or reactive groups and/or include limits on molecular properties which may be generally useful in drug design, i.e. non-specific for disease types. Other early reviews [28] also use the phrase drug-likeness for "*molecules which contain functional groups and/or have physical properties consistent with the majority of known drugs*". Muegge [19] remarked that "*Drug-likeness is mostly a statistical descriptor derived from databases of other compounds. It should, therefore, be used to evaluate the drug-likeness of other compound selections such as screening libraries, combinatorial libraries, or virtual libraries rather than that of a single compound*." Taking into account the general opinion formulated by the above studies the drug-likeness paradigm in the present review can be classified as (i) general drug-likeness (all diseases and mostly oral drug administration); and (ii) specific drug-likeness (classified by disease, administration, target, etc.).

*Address correspondence to this author at the Departments of Biochemistry and Genetics, Eötvös University, Pázmány sétány 1/C, 1117 Budapest, Hungary; Tel: +36-13812173; Fax: +36-13722641; E-mail: csabahete@yahoo.com

1875-533X/12 $58.00+.00

Table 1. General Drug-Likeness Values of Property Filters

Source			Statistics	Property									Database	
Year	**Author**	**Ref**		**NCC**	**HBD**	**HBA**	**logP**	**MW**	**NHA**	**NR**	**NRB**	**PSA**	**Description**	**N**
1997	Lipinski *et al.* (Ro5)	[12]	~90P		5	10	5(c) 4.15(m)	500					World Drug Index (WDI) filtered by USAN, INN names, etc.	2245
1999	Ghose	[33]	MEAN				2.3(a)	357					CMC	6304
			SD				2.6(a)	174						
2001	Sakaeda *et al.*	[34]	MEAN				1.73(c) 1.94(m)	332					Oral drugs	222
			SD				2.21(c) 2.03(m)	140						
2003	Feher and Schmidt	[39]	MEAN	2.3	1.9	5.7	2.2(s)	340	23.5	2.6	5.6		Chapman and Hall Dictionary of Drugs, The Merck Index	10968
			MED	1	1	5	2.3(s)	312	22	2	5			
2003	Wenlock *et al.*	[40]	MEAN		2.1	4.9	2.5(c)	337			5.9		The Physicians' Desk Reference 1999	594
			SD		2.4	3.6	2.5(c)	157			4.5			
			90P		4	8	5.5(c)	473			11			
2004	Leeson and Davis	[42]	MEAN		1.81	5.14	2.27(c)	331		2.56	4.97	21.1%	Oral drugs pre-1982	864
			MED		1	4	2.31(c)	310		3	4	18.5%		
2004	Leeson and Davis	[42]	MEAN		1.77	6.33	2.50(c)	377		2.88	6.42	21.0%	Oral drugs 1983-2002	329
			MED		1	6	2.36(c)	357		3	6	19.4%		
2004	Vieth *et al.*	[43]	MEAN		1.8	5.5	2.3(c)	343.7		2.6	5.4	78	FDA Orange Book	1193
			90P		3	9	5.2(c)	475		4	10	134		
2004	Vieth *et al.*	[43]	MEAN		3	4.5	2.5(c)	300					Lipinski *et al.* recomputed based on the 2001 edition of the WDI	1791
			90P		3	8	5.3(c)	427.5						
2005	Proudfoot	[44]	MEAN		1.5	5.1	2.5	333					Oral drugs 1937-1997	1791
			90P		3	9	4.8	469						
2006	Vieth and Sutherland	[48]	MEAN		1.8	5.5	2.3(c)	345					Vieth *et al.* 2004 updated with FDA release after 2003	1210
			90P		4	9	5.3(c)	478.4						
2009	Tyrchan *et al.*	[49]	MEAN		1.5	3.9	2.74(c)	335.5			5.6	64.7	GVKBIO, IBEX	976
			SD		1.5	2	2.22(c)	109.2			3.6	39.7		
			MED		1	4	2.83(c)	318.5			5	59.1		

Abbreviations. 90P: 90th percentile; HBA: number of H-bond acceptors (O+N); HBD: number of H-bond donors (OH+NH); logP: logarithm of octanol/water partition coefficient (small letters in brackets denote different methods of logP calculation); MED: median; MW: molecular weight; N: number of drugs in database; NCC: number of chiral centers; NHA: number of heavy atoms; NR: number of rings; NRB: number of rotatable bonds; PSA: polar surface area; SD: standard deviation.

2.2. General Drug-Likeness

Ajay *et al.* [32] investigated the possibility of distinction between general drug-likeness and non-drug-likeness by one- or two-dimensional descriptors within neural network-based models. They used the Comprehensive Medicinal Chemistry (CMC) and the MACCS-II Drug Data Report (MDDR) as drug-like data sets and the Available Chemicals Directory (ACD) as a surrogate for non-drugs. It was correctly remarked that using the above databases as drug/non-drug collections is an assumption as "the characteristics of drug molecules today may change in the future". Therefore, the conclusions of dataset-based drug-likeness studies may always reflect the actual state of the common knowledge on drug-likeness and *a priori* include errors. This study can be regarded as a key analysis, which included not only drugs, but also quasi non-drugs in a truly comparative manner.

However, most of the studies reporting drug-likeness thresholds (Table **1**) deal only with drug (lead or bioactive) databases. Ghose *et al.* [33] based their analysis on the CMC database and provided drug-likeness thresholds for MW and logP. They also concluded the priority of some fragments (e.g. benzene ring) occurring in drug structures. The analysis of 222 commercially available oral drugs by Sakaeda *et al.* [34] supported the Ro5. However, the authors also remark that compounds with a sugar moiety, high atomic weight, and/or large cyclic structure were exceptions to the MW=500 upper

threshold. Veber *et al*. [35] also found that molecular weight cutoff at 500 does not itself significantly separate compounds with poor oral bioavailability from those with acceptable values. They analyzed the oral bioavailability of a large data set in rats containing more than 1000 compounds. It was also concluded that compounds of possibly good oral bioavailability possess 10 or fewer rotatable bonds (NRB) and polar surface area (PSA) equal to or less than 140 $Å^2$ or 12 or fewer H-bond donors (HBD) and acceptors (HBA). Their analysis on artificial membrane permeation rates showed that reduced PSA correlated better with increased permeation rate than did ClogP, and an increased NRB had a negative effect on the permeation rate. Lu *et al.* [36] also investigated the predictive power of NRB and PSA on 434 Pharmacia compounds and found that their correlations with bioavailability depended on the therapeutic class.

Hann *et al*. [37] studied the differences in the properties of drug leads and optimized compounds. The data indicates that, on average, drug leads have lower MW, lower ClogP, fewer aromatic rings (NR), fewer HBA than the corresponding drugs. On the contrary, Proudfoot [38] found that most drugs are within 25% of the lead values with regard to MW, and nearly all are within one calculated MLogP unit.

In another interesting comparative analysis of drugs, natural products and combinatorial libraries Feher and Schmidt [39] also emphasized the importance of properties beyond the often used MW and logP. For example, it was shown that the 'number of chiral centers' in a molecule has a great impact on its drug-likeness. They found that while chiral centers are normally present in drug and natural product molecules, they tend to diminish in combinatorial compounds, which is most probably a consequence of the oversimplified synthetic/construction steps in the generation of combinatorial libraries.

Wenlock *et al.* [40] compared distributions of physico-chemical properties such as MW and logP of marketed oral drugs and of compounds in development. In their analysis, the mean MW of orally administered drugs in development decreased on passing through each of the different clinical phases and gradually converged towards the mean molecular weight of marketed oral drugs. In addition, the most lipophilic compounds diminished during development. They compared upper property thresholds below which 90 % of oral drugs in their data set with the results of the Ro5, and good agreement was found (Table **1**).

Besides the thresholds values, the historical trends of, e.g. MW of drug candidates, may be also useful as collected by Lipinski [41] for the period 1960–2004. It was demonstrated that advanced clinical candidates produced by a "rational drug design" approach of Merck had a time-dependent higher MW, higher H-bonding properties, unchanged logP, and poorer permeability. Early candidates from a HTS-based approach of Pfizer (Groton, CT) had higher molecular weight, unchanged H-bonding properties, and higher logP, i.e. poorer aqueous solubility. In another retrospective study, Leeson and Davis [42] showed that mean values of lipophilicity, percent of PSA and HBD had not changed in the period of 1983-2002. In contrast, mean values of MW and the numbers of O + N atoms, HBA, NRB, and number of rings have increased by 13-29%. Similarly, Vieth *et al.* [43] demonstrated that the mean property values for oral drugs do not vary substantially with respect to launch date. The limited change in the most important oral drug-like property values lead the authors to suggest that the range of acceptable oral properties is independent of the synthetic complexity or targeted receptor. Proudfoot [44] analyzed the very long period of 1937-1997. During this period a steady increase was observable in mean and median MW. Only seven marketed drugs with MW>500 were designed in the 15 year period 1937–1951, and thirty two in the comparable period 1983–1997. Mean and median logP was unchanged in the 60 year period examined. Fewer than 5% of oral marketed drugs had more than 4 H-bond donors and just 2% had MW>500 and >3 H-bond donors. An analysis by Leeson and Springthorpe [45] suggested that clogP is the most important molecular property, as it is changing less over decades in launched oral drugs than other properties. As ClogP plays a dominant role in promoting binding to unwanted drug targets, a high logP therefore carries increased risks of developmental attrition. They conclude that a 5% improvement in attrition would double the output of new medicines and that this might be achieved simply by lowering logP. Comparing sets of drugs and their originating leads, Perola [46] also found that on average, the two sets have similar logP, suggesting that the ability to maintain low levels of logP while increasing MW is one of the keys to a successful drug discovery program.

Schneider *et al.* [47] investigated the combined use of drug-likeness property filters in gradual filtering by decision trees. With rapidly computable properties such as MW, XlogP, molar refractivity, and several drug-likeness indices, up to 76% of all non-drugs could be sorted out in the first filtering step. With the aid of sophisticated (quantum chemical) properties in the succeeding steps up to 92% of the initial non-drugs were filtered out, while less than 19% of the actual drugs were lost. In addition to the above examples, Table **1** also lists threshold values given by Vieth and Sutherland [48] and Tyrchan *et al.* [49].

2.3. Limitations of the General Drug-Likeness Concept

Although physicochemical properties are widely used as general drug-likeness filters (Section 2.2), there are several articles pointing to their limitations. As Walters *et al.* [28] envisioned, instead of dealing with the complex problem of drug-likeness, a viable alternative is the prediction of the various pharmacokinetic properties (logP, half-life, plasma protein binding, etc.) that contribute to a drug's success. Remarkably, even the calculation and modeling of these properties themselves is rather complex [50] and extremely difficult in many cases.

The lack of validated sets of drugs and decoy sets of non-drugs [51] also limits the usefulness of any drug-likeness filters as there are compounds, e.g., that can easily fall into either category. Moreover, the filters can only recognize those compounds that resemble existing drugs as drug-like – compounds from completely new classes could be misclassified [31]. Remarkably, the original publication of Lipinski [29], root of many others in this field, addressed the prediction of *only* pharmacokinetic properties (absorption and permeation) and *not* general drug-likeness.

However, collecting sets of good and bad pharmacokinetic properties remains a challenge for property filters due to the above-mentioned complexity of the properties themselves. In addition, the final decision on drug-likeness is just further postponed if a filter can provide information only on one drug-likeness property. In fact, there are several properties to be predicted which can easily give controversial results in ranking of a compound or a library and it is still unclear which property should be prioritized for the final decision, etc. For example, Kubinyi [52] finds that "*inappropriate ADME (Absorption, Distribution, Metabolism, Excretion) characteristics have clearly made far less of a contribution to clinical failures than is widely supposed!*". At the same time, he also accepts that the application of the Ro5 aimed at prediction of "A" of ADME significantly aided improving early combinatorial libraries which had included "*many large and greasy, biologically inactive molecules*". This example of the controversial judgment of the fairly well-studied ADME properties illustrates that it would be indeed very difficult to set the above-mentioned priority order of properties in a decision tree. The questions on the appropriate use of a property, i.e., "*where and to which extent*" seem to remain unanswered in general.

Similarly, an important study by Feher and Schmidt analyzing properties of natural products [39] concluded that: "Drug-like filters, such as the Lipinski rules, are very helpful in isolating likely problem molecules. However, overly strict adherence to it can have the adverse effect of restricting diversity … and hence also reducing similarity to natural products. … A large proportion of natural products is biologically active and has favorable ADME/T properties, despite the fact that they often do not satisfy 'drug-likeness' criteria." Furthermore, Ganesan [53] analyzed a total of 24 unique natural products that led to an approved drug in the period 1970–2006. They found an identical success rate of 50% both for the classes conforming or violating the Ro5. It was also found that natural products are successful in maintaining favorable logP and intermolecular H-bond donating potential even with high MW and large numbers of rotatable bonds.

Lajiness *et al* [54] raise additional concerns regarding drug-likeness studies. They claimed that there are very few studies accompanied by the data sets used for analysis, and therefore, reproducibility of the results is questionable. During collection of data in Table **1**, we also found that in many cases authors refer to, e.g. in-house, company-owned data sets or other resources with no or reduced public availability or a non-defined sub-set of an available database. However there is no guarantee that proprietary collections are adequate for the analysis of general drug-likeness. For example, Lajiness *et al*. [54] mentioned that proprietary collections may be biased due to historical lead optimization efforts focused at particular chemical classes, such as steroids or benzodiazepines. They also concluded that comparing drug-likeness of groups instead of individual compounds was appropriate to achieve significant results.

There are also methodological problems with the properties 'traditionally' used as filters. For example, Bhal *et al*. [55] suggest the cautious use of logP in drug design due to its inability to account for the ionization of compounds under physiological conditions. They conclude that the pH-dependent logD is a more realistic descriptor of lipophilicity under physiological pH's and, therefore, logD should be used preferentially over logP as the descriptor for lipophilicity, especially when working with ionizable compounds. Vistoli *et al*. [51] also mention the problems of pH-dependent properties.

In their seminal paper, Lipinski *et al*. [29] already claimed that antibiotics, antifungals, vitamins, and cardiac glycosides fell outside their Ro5, possibly due to transporter effects. The results of the study of Good and Hermsmeier [56] suggest further discontinuities in drug-like space, beyond those claimed by Lipinski *et al*. [29], in the context of classification. Giménez *et al*. [57] also concluded that Ro5 is very useful to select better compounds in chemical libraries, but it must be used carefully to avoid a possible exclusion of promising compounds. They evaluated the top pharmaceutical products in 2007. Among 60 drugs, 7 (atorvastatin, montelukast, docetaxel, telmisartan, tacrolimus, leuprolide and olmesartan) did not fit the Ro5, and 5 failed one of the threshold values.

Zhang and Wilkinson [58] summarized their criticism of the overemphasis of Ro5 of drug-likeness from two points of view. Firstly, they claim that only 51% of all FDA-approved small molecule drugs are both used orally and comply with the Ro5. This does not even include the increasing number of biologicals of which several have reached 'blockbuster' status. Secondly, the Ro5 does not cover natural product and semisynthetic natural product drugs, which constitute over one-third of all marketed small-molecule drugs (see also Feher and Schmidt [39]).

A further doubt arises from the finding (Dobson and Kell [7]) that general drug-likeness properties such as MW or logP, adequate for passive diffusion, have decreased ability for prediction of carrier-mediated and active uptake of drugs that are more common forms of transport than is usually assumed. For drugs transported by carriers, general property filters are not normally effective in individual cases, and specific data on interactions of drugs and transporters would therefore accelerate research in this field. Similarly to drugs, naturally occurring intermediary metabolites may also require solute carriers to enter cells. Thus, an evaluation of metabolite-likeness (Dobson *et al*.) [59] would be essential to understand the true physiological processes. However, estimation of metabolite-likeness is missing from most of the present drug-likeness studies.

2.4. Specific Drug-Likeness

Considering the diversity of drug profiles, specific approaches of drug-likeness may become an alternative to the limited general concept reviewed in the previous Sections. Drugs achieve their effects through different mechanisms in the body, targeting different proteins, organs or even organisms, as in the case of anti-infective agents. Moreover, dermatological agents used topically may require completely different pharmacokinetic properties than drugs which are inhaled, injected or administered orally. In addition, drugs that affect the central nervous system have to pass yet another obstacle, the blood-brain barrier (BBB).

Besides their general analysis, (Section 2.2) Ghose *et al*. [33] also investigated the property profile (MW, logP, etc.) of seven different classes of drug molecules in the CMC such as central nervous system (CNS), cardiovascular, cancer, inflammation, and infectious diseases (Table **2**). They provided drug-likeness ranges for the different classes and found considerable outliers from the general drug-likeness trend. For example, the antibacterial compounds formed a special class of biologically active compounds very different from regular drugs. The logP of anticancer drugs showed a high standard deviation possibly due to the complexity of cancer, which affects different parts of the body and tissues. On the other hand, the standard deviation of logP of CNS drugs was relatively small due to the requirement that they should cross the BBB. They concluded that for different drug classes the ranges may be considerably tighter than the general drug-likeness ranges. Leeson and Davis [42] also found that significant differences exist between the property distributions of different therapeutic areas of oral drugs of the 1983-2002 period. The distributions of MW and logP among antiinfectives show different trends from the other drug classes probably related to the need for their activity in a non-human organism, and cell wall penetration in the case of antibiotic drugs.

Vieth *et al*. [43] analyzed the differences between routes of administration (Table **2**). It was observed that 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 for oral vs. injectable drugs. However, they concluded that due to the substantial overlap in the range of properties found between the different drug classes, a particular drug cannot be adequately classified as either oral or injectable on the basis of simple physical property calculations. Tronde *et al*. [60] have studied the physicochemical properties and absorption qualities of inhaled drugs, finding that the pulmonary epithelium allows for higher PSA (up to 479 $Å^2$) in compounds, as compared to the intestinal mucosa and BBB. They propose the lung route as an alternative to drugs poorly absorbed through the oral route. Ritchie *et al*. [61] also studied respiratory drugs administered through intranasal/inhaled routes, and found their calculated physicochemical properties to have lower lipophilicity, higher molecular weight, and higher PSA, when compared to drugs administered orally.

Table 2. Specific Drug-Likeness Values of Property Filters

Source			Disease/administration /target family	Statistics	Property							Database	
Year	**Author**	**Ref**			**HBD**	**HBA**	**logP**	**MW**	**NR**	**NRB**	**PSA**	**Description**	**N**
1999	Ghose *et al*.	[33]	cancer	MEAN			1.59(a)	332				CMC	349
				SD			2.5(a)	129					
			cardiovascular/ antiphypertensive	MEAN			1.97(a)	361					269
				SD			2.1(a)	123					
			CNS/antidepressant	MEAN			3.05(a)	291					208
				SD			1.5(a)	69					
			CNS/antipsychotic	MEAN			4.10(a)	380					105
				SD			1.5(a)	83					
			CNS/hypnotic	MEAN			2.20(a)	277					74
				SD			1.5(a)	99					
			infection	MEAN			2.38(a)	339					39
				SD			2.7(a)	139					
			inflammation	MEAN			3.09(a)	335					290
				SD			1.5(a)	122					
2004	Leeson and Davis	[42]	Cancer	MEAN	1.00	4.5	3.02(c)	313	2.36	5.00	20.8 %	Oral drugs 1983-2002	14
				MED	1	4.5	3.01(c)	299	2	3.5	18.3 %		
			cardiovascular	MEAN	1.46	6.73	3.05(c)	389	2.84	8.23	19.8 %		79
				MED	1	7	3.00(c)	396	3	8	18.6 %		
			gastrointestinal and metabolism	MEAN	2.71	6.84	1.90(c)	378	2.32	7.63	26.7 %		38
				MED	2	6	2.28(c)	357	2.5	7	20.7 %		
			infection	MEAN	2.41	8.78	1.56(c)	456	3.45	6.83	24.6 %		64
				MED	2	7	0.94(c)	389	3	5	21.5 %		
			nervous system	MEAN	1.50	4.32	2.50(c)	310	2.85	4.70	16.3 %		74
				MED	1	4	2.55(c)	307	3	4.5	14.3 %		
			respiratory and inflammation	MEAN	1.37	4.24	3.34(c)	396	3.02	5.52	20.5 %		46
				MED	1	4	2.90(c)	353	3	4.5	19.3 %		
2004	Vieth *et al*.	[43]	absorbent	MEAN	3	6.5	1.6(c)	392.3	2.5	7.9	100.5	FDA Orange Book. Drugdex	116
				10-90P	0-7	2-14	-2.3 to 4.8(c)	172-666	0-4	2-16	20-219		
			injectable	MEAN	4.7	11.3	0.6(c)	558.2	3.2	12.7	143.6		308
				10-90P	0-11	3-23	-3.3 to 4.9(c)	196-1085	1-6	2-27	28-311		
			topical	MEAN	1.9	5	2.9(c)	368.5	2.9	5.3	75.4		112
				10-90P	0-3	2-8	-0.6 to 6.0(c)	188-495	1-5	1-9	21-114		
2006	Vieth and Sutherland	[48]	CYP450	MEAN	0.7	2.9	3.4	300.5				Vieth *et al*. 2004 updated with FDA release after 2003	12
				90 %	2	5	8.8	399.4					
			GPCR-bio	MEAN	1.3	4.2	2.8	326.8					216
				90 %	3	7	5.1	435.4					
			GPCR-lipid	MEAN	1.8	5.0	5.5	414.9					8
				90 %	3	9	8.5	586.2					
			GPCR-pep	MEAN	1.6	8.5	5.0	484.8					11
				90 %	2	12	7.5	600.2					
			ion channel	MEAN	1.3	4.9	2.5	305.5					115
				90 %	2	9	5.0	443.2					
			kinase	MEAN	2	7.0	4.6	439.4					5
				90 %	3	8	5.6	493.6					
			NHR	MEAN	1.4	3.8	4.1	381.8					58
				90 %	3	6	7.2	445.8					
			PDE	MEAN	0.9	6.9	1.7	331.9					15

(Table 2). Contd…..

Source			Disease/administration /target family	Statistics	Property							Database	
Year	**Author**	**Ref**			**HBD**	**HBA**	**logP**	**MW**	**NR**	**NRB**	**PSA**	**Description**	**N**
				90 %	2	10	4.2	480.2					
			protease	MEAN	4.5	7.2	2.3	430.6					35
				90 %	5	11	5.9	636.6					
			transporter	MEAN	1.3	4.2	3.0	304.7					37
				90 %	3	7	5.5	423.5					
1999	Ajay *et al.*	[64]	CNS	MEAN			2.8	354				CMC and MDDR	1050 + 16785
				MED			2.9	351					
				90 %			0.0-5.2	200-540					
1999	Kelder *et al.*	[65]	CNS	MAX							120 $Å^2$	Passively transported oral drugs	776
				~MEAN							60-70$Å^2$		
2009	Chico *et al.*	[69]	CNS	~MAX			4	400			80 $Å^2$	Brain-penetrant small molecules	448
2001	Sakaeda *et al.*	[34]	CNS	MEAN			2.67(c) 2.80(m)	285					44
				SD			2.03(c) 1.98(m)	91					
			inflammation	MEAN			2.63(c) 2.66(m)	279					17
				SD			1.37(c) 1.47(m)	107					
			microbial	MEAN			-0.18(c) -0.13(m)	371					48
				SD			1.88(c) 1.59(m)	161					

Abbreviations. 90P: 90^{th} percentile; HBA: number of H-bond acceptors (O+N); HBD: number of H-bond donors (OH+NH); logP: logarithm of octanol/water partition coefficient (small letters in brackets denote different methods of logP calculation); MED: median; MW: molecular weight; N: number of drugs in database; NHA: number of heavy atoms; NR: number of rings; NRB: number of rotatable bonds; PSA: polar surface area; SD: standard deviation.

Another study of Vieth and Sutherland [48] investigated the distribution of drug-likeness property filters by targeted proteomic families. For proteases, nuclear hormone receptors, lipid and peptide G-protein-coupled receptors (GPCRs), the corresponding drugs significantly exceed Ro5 limits, while others targeting cytochrome P450s, biogenic amine GPCRs, and transporters had significantly lower values for certain properties. It is also an interesting question whether ligands targeting different proteomic families have statistical difference in their property ranges. According to the results of Morphy [62], the ligands of peptide GPCRs and integrin receptors, possess significantly higher median property values than those for aminergic targets, such as monoamine transporters and GPCRs. Agonists for monoamine GPCRs, opioid receptors and ion channels had smaller MW and clogP than the antagonists, but there was no difference between the agonists and the antagonists for peptide GPCRs and nuclear receptors. Paolini *et al.* [63] also found distinct differences in the distribution of molecular properties between sets of compounds active against different families. For example, they also found that the mean MW of ligands binding to aminergic GPCRs is 378(±93), whereas the mean MW of peptide GPCR ligands is greater at 514(±202).

The design of libraries of CNS-active compounds has been a goal of many research groups since the early applications of drug-likeness property filters (Ajay *et al.* [64]) such as MW or logP. Kelder *et al.* [65] found a significant difference in the polar surface area distribution of 776 CNS and 1590 non-CNS drugs. It was concluded that orally active drugs with passive transcellular transport should not exceed a PSA of 120 $Å^2$, and a 60-70 $Å^2$ for appropriate BBB permeability. MW was identified as a good descriptor of BBB penetration [66, 67], and applied in fact as a key property filter together with logP in testing a 3042 compound screening library [68] for CNS-compatibility. Chico *et al.* [69] claimed that kinase inhibitor drugs for CNS indications required a modification of the property limits set by the Ro5. They found that most of the brain-penetrating small molecules had a MW<400, logP<4 and PSA<80$Å^2$. In addition to above examples Table **2** provides also threshold values fro three disease families by Sakaeda *et al.* [34].

3. MOLECULAR PROPERTY FILTERS DESCRIBING BINDING AFFINITY

Besides the use of molecular properties (MW, NHA, logP, etc.) as filters (see previous Sections), several studies investigated the correlation between these properties and the binding affinity of a ligand to its macromolecular target (Eq. 1). Remarkably, during the formation of the [Ligand:Target] complex some water molecules (k in Eq. 1) may leave the binding interface, whilst others may join it [70, 71]. The binding affinity (also called '*in vitro* potency') can be described in terms of thermodynamic equilibrium constants of association, dissociation or inhibition (K_a, K_d. K_i) which can be related to ΔG (Eq. 2). In some cases, the logarithm of inhibitor

concentration at 50 % inhibition (pIC_{50}) is also applied as a measure of binding affinity, but pIC_{50} cannot be directly related to ΔG by Eq. 2.

$$Ligand(H_2O)_n + Target(H_2O)_m \rightleftharpoons [Ligand:Target](H_2O)_{n+m-k} + k\,H_2O \quad \text{Eq. 1}$$

$$\Delta G = -RTlnK_a = RTlnK_{d/i} \quad \text{Eq. 2}$$

(R is the gas constant, T is the thermodynamic temperature)

In a seminal article Kuntz *et al*. [72] plotted experimental ΔG values of a large set of complexes of macromolecular targets and their strongest-binding ligands against the NHA of the ligand molecules. They found that ΔG increases with NHA with an initial slope of ca. -1.5 kcal/mol (1cal = 4.18 J) per atom. Beyond 15 NHAs the increase dropped dramatically suggesting a logarithmic relationship between ΔG and NHA for large molecules. Reynolds *et al*. [73, 74] found a similar, non-linear relationship when plotting the most potent ligands of the BindingDB database. The 'maximal affinities' as measured by pIC_{50} increased rapidly up to 20 heavy atoms, but a plateau existed beyond 25. A recent study of Ferenczy and Keserű [75] also presented a non-linear plot of pK_d-NHA with a plateau starting from 40 heavy atoms.

Ferrara *et al*. [76] calculated the Pearson R value between the experimental ΔG and the logarithm of the MW for different data sets (Table **3**) and found significant correlations in many cases. The logarithmic function was chosen according to the above detailed logarithmic dependence of ΔG on NHA shown by the study of Kuntz *et al*. [72]. Velec *et al*. [77] also calculated a Spearman's rank order correlation coefficient of 0.56 between experimental ΔGs of 100 complexes and the MWs of participant ligands. Affinity predictions purely based on the ligand's MW gave in fact better results for the 100 complexes than many scoring functions involving other terms on, e.g. interaction with the target. Wells and McClendon [78] collected the ΔG of highest-affinity fragments and small molecules that target seven different protein–protein interfaces, and found an R=0.77 (R^2=0.59) correlation between ΔG and NHA. Kim and Skolnick [79] published correlations between pK and logMW values of various data sets (Table **3**).

Olsson *et al*. [80] measured a considerable correlation of ΔG with apolar surface area burial (including both ligand and protein surface) upon complex formation (R^2=0.65) and the change in ligand apolar solvent accessible surface area (ASA, R^2=0.44) using a diverse set of 254 complexes of the SCORPIO database. Notably, binding pocket ASA was shown [81] to correlate with ligand MW at an R^2=0.77 too. For peptide ligands, estimation of ΔH was considered using a linear combination of ΔASA values [82].

The background of the correlations of ΔG (logK) with ligand-based, size-dependent properties (MW and NHA, Table **3**) has not been elucidated yet. According to Eq. 3, for the analysis of correlation of the properties with ΔG it may be a plausible idea to analyze their correlations with the binding enthalpy (ΔH) and entropy (ΔS) changes, respectively. Using the data set published by Reynolds and Holloway [83], no correlation can be observed between NHA and ΔH or TΔS, respectively, but with ΔG, a slight R^2=0.28 can be calculated. This finding hints that such a dissection of ΔG into ΔH and TΔS may not help in finding the reasons of the correlations of Table **3**.

$$\Delta G = \Delta H - T\Delta S \quad \text{Eq. 3}$$

Table 3. Correlations Between Binding Affinity and Molecular Properties

Source			Correlated quantities		R^2	Database	
Year	**Author**	**Ref**	**Binding affinity**	**Property**		**Description-target protein**	**N**
2004	Ferrara *et al*.	[76]	pK_i	logMW	0.36	LPDB-all	189
					0.23	LPDB-oxidoreductase	37
					0.81	LPDB-serine protease	25
					0.58	LPDB-metalloprotease	13
					0.50	LPDB-immunoglobulin	10
					0.18	LPDB-lyase	8
					0.16	LPDB-L-arabinose binding protein	9
2005	Velec *et al*.	[77]	pK_d	MW	0.31[a]	Wang *et al*.	100
2007	Wells and McClendon	[78]	ΔG	NHA	0.59	Ligands of seven different targets	13
2008	Kim and Skolnick	[79]	pK_i/pK_d	logMW	0.38	CDSa(CDS1-7)	146
					0.24	CDS3-HIV-1 protease	28
					0.53	CDS5-Ribonuclease a	13
					0.76	CDS6-Thermolysin	10
					0.50	CDS7-Beta trypsin	47
					0.59	Protein Ligand Database v1.3 CDS8-Beta trypsin	7
					0.88	CDS9-Carbonic anhydrase II	15
					0.40	CDS11-HIV-1 protease	6
					0.49	CDS12-Thrombolysin	9
2011	Reynolds and Holloway	[83]	ΔG	NHA	0.28[b]	BindingDB	102
2012	Present study		ΔG	logMW	0.14	Non-drugs	320
				logW	0.15		
				logP	0.19		

Abbreviations. logP: logarithm of octanol/water partition coefficient; MW: molecular weight; N: number of data; NHA: number of heavy atoms; W: Wiener index.
[a]Spearman's R^2; [b]Calculated using the data in the reference.

$$\Delta G \approx \Delta H_{inter} + \Delta H_{intra} - T\Delta S_{config} + \Delta G_{sol} \quad \text{Eq. 4}$$

$$\Delta H_{inter} \approx \Delta E_{Coulomb} + \Delta E_{LJ} + \ldots \quad \text{Eq. 5}$$

ΔG can be approximated (Brooijmans and Kuntz) [84] further by separating the terms of Eq. 3 into enthalpy changes (Eq. 4) coming from changes of intra (ΔH_{intra})- and intermolecular (ΔH_{inter}) interactions, configurational entropy change (ΔS_{conf}), and a free energy change coupled to (de)solvation processes (ΔG_{sol}), such as release of interface waters (Eq. 1) during binding. ΔG_s includes both enthalpic and entropic contributions of changes of solute-solvent interactions during complex formation. (Notably, there is an unclosed debate in the literature on the separability of the entropic terms for individual (molecular) contributions which may affect the above separation of ΔG_s from other terms of ΔG [85, 86]. In many ΔG calculators [84], ΔH_{inter} is estimated involving pair-additive potential terms such as the Coulomb ($E_{Coulomb}$) or the Lennard-Jones (E_{LJ}) formulas (Eq. 5) for electrostatic and van der Waals-interactions, respectively. Jacobson and Karlén [87] found that ΔG calculators built mostly on such enthalpic terms of ligand-target interactions (Eq. 5) produced high correlations with NHA hinting that ΔH_{inter} accounting for protein-ligand interactions is partly described by NHA. One possible explanation is that NHA can be related to surface area, and hence, to van der Waals interactions and, therefore, a high NHA can translate into a high ΔH_{inter}.

Besides ΔH_{inter}, some parts of the configurational entropy (S_{config}) can be also related to MW (Eq. 6)

$$S_{config} = S_{trans} + S_{rot} + S_{vib} \quad \text{Eq. 6}$$

$$S_{trans} + S_{rot} = R\ln(aMW) \quad \text{Eq. 7}$$

where trans, rot, and vib denote respectively, the translational, rotational, and vibrational ΔS contributions to the configurational entropy change, and 'a' is a constant. Several studies [88-94] calculate S_{config} using classical formulas relating S_{trans} and S_{rot} to the logarithms of MW and the principal moments of inertia, respectively. As known, the principal moments of inertia are also dependent on molecular size (and shape). Simplified formulas [95, 96] were also introduced (Eq. 7) showing the dependence of part of S_{config} on MW. However, this dependence was suggested to be very weak or zero for the change of S_{config}, i.e. for ΔS_{config} of the binding process [97, 98].

In summary, several studies have published relationships (Table **3**) at various correlation levels between experimental binding affinity and molecular property filters such as MW, NHA, etc. Since the article of Gilson *et al*. [97], which had also dealt with the ΔG-MW correlation, experimental collections have been published presenting new data. A collection of recent correlations was provided in Table **3** and the thermodynamic background was sketched to illustrate the problems of explaining these correlations. While the above considerations suggest that individual components of ΔG such as ΔH_{inter} are related to molecular size, and some of them, such as ΔS_{conf}, are probably not correlated with MW, the final explanation on the moderate, but significant correlations of ΔG with ligand size is still awaiting. Notably, these relationships are probably not linear as quantities obtained by simple normalization of ΔG with, e.g. MW, are still dependent on MW (see next Section for details).

4. THE CONCEPT OF LIGAND EFFICIENCY (EFFICIENCY INDEX, EI)

The dependence of binding affinity on ligand size (MW, NHA) discussed in the previous section raises the question whether it is possible to define a measure, the binding efficiency for comparison of 'intrinsic' binding affinities of ligands of any sizes *via* 'decoupling' ΔG from molecular size. In an early work, Andrews *et al*. [99] hinted at the possibility of definition of such intrinsic ΔGs for a limited number of functional groups of a molecule by using average values calculated from experimental ΔGs. Later, DeWitte and Shaknovich [100] calculated the intrinsic binding affinity per heavy atom and correlated these values with experimental K_i-s. Kuntz *et al*. [72] also used this intrinsic measure and showed that ΔG/NHA rapidly decreases up to ca. 15 NHA (see also previous Section).

Based on the above results, Hopkins *et al*. [101] recommended the introduction of ligand efficiency in the following explicit form (Eq. 8). The work of Wells and McClendon [78] provides information on the actual values of 'efficient' molecules. They collected several potent small molecules inhibiting protein–protein interactions and obtained $|EI_{NHA}|$ values of 0.2…0.4 for their data set. An alternative, idealized value of 0.5 has been recommended by others [63, 101, 102].

$$EI_{NHA} = \frac{\Delta G}{NHA} \quad \text{Eq. 8}$$

To note, throughout this review we use the name 'efficiency index (EI)' instead of 'ligand efficiency' to emphasize that this measure of intrinsic ΔG is a rational definition of the efficiency of a ligand, however, it is not the only possible definition.

Definition of other EIs was provided by Abad-Zapatero and Metz [24] using MW (EI_{MW}) and PSA (EI_{PSA}) in the denominator of Eq. 8 instead of NHA. A series of other EIs were introduced based on various size-dependent properties for normalization among which the Wiener-index (W) was found particularly useful in the form of EI_W [103]. Leeson and Springthorpe [45] proposed a ligand-lipophilicity-based efficiency index (EI_{lipo}, Eq. 9) to be used in "maximizing the minimally acceptable lipophilicity" per unit of binding affinity during drug design. They suggest that an average drug has an EI_{lipo} of 5-7 or greater.

$$EI_{lipo} = pIC_{50} \text{ (or } pK_i) - \text{clogP (or logD)} \quad \text{Eq. 9}$$

Although the definition of EIs involves normalization by ligand size (Eq. 8), Reynolds *et al*. [74] found that EI_{NHA} is still dependent on ligand size, as a very dramatic decline was observed in EI_{NHA} as size increases. Notably, Orita *et al*. [104], and Keserü and Makara [105] described a similar trend of EI_{NHA} vs. NHA. The drop in EI_{NHA} was large between ca. NHA=10…20, and flattened toward very large sizes (NHA>40). They found an interesting similarity between the maximal EI_{NHA} vs. NHA and the ASA vs. NHA curves suggesting that the primary driving forces behind the systematic decline in maximal EI_{NHA} with increasing molecular size is the reduced effective surface area for the larger compounds. In other words, large molecules possess relatively large buried surface area unavailable for binding. In a recent study, Reynolds and Holloway [83] concluded that the strong size dependence of EI_{NHA} (average or optimal) is mostly a consequence of the dependence of the enthalpic, and not the entropic part of EI. To eliminate the above size-dependency of EI_{NHA}, Reynolds *et al*. [74] introduced a new functional form called 'fit quality', and Nissink [106] derived a size-independent ligand efficiency measure of the form of binding affinity/$NHA^{0.3}$.

The concept of ligand efficiency is a simple way to merge binding and pharmacokinetic characteristics of a ligand into a single measure. EI has already been applied in many studies and it is suggested to become a useful tool of fragment-based drug discovery [102, 104, 107-109], lead optimization [46], and drug chemical (molecular) space localization for some diseases or organs [110]

5. SENSITIVITY AND SELECTIVITY OF PROPERTY FILTERS

Tables **1** and **2** list general or specific drug-likeness values of molecular properties. Most of these values are descriptive statistics (mean, median, percentile, etc.) of data sets including only drugs. That is, counter-examples of a set of non-drugs are generally not

considered. Notably, the strict definition of such sets is not obvious (Section 2.2) due to possible change/evolution of the drug/non-drug status of any compounds. However, if sets of drugs were collected, then it is fairly plausible to expect a non-drug set for comparison. Introduction of a new statistical term on the selectivity of the property filter is also necessary showing the ability of the property to distinguish drugs from non-drugs. Since the investigated molecular properties (MW, NHA) are coupled to both pharmacokinetic drug-likeness (Section 2) and ΔG (Section 3), it would be also advantageous to 'switch off' the ΔG-coupling in an analysis to investigate the properties' selectivity only for drug-likeness. In the forthcoming Sections, selectivity and sensitivity measures of drug-likeness filters are introduced using a 631-compound database as an example.

5.1. Data Sets

Details of the collection of the data sets are provided in the Appendix and the sets are listed in the Supplementary Material. To decouple the ΔG-dependence (Section 3) of the property filters, the two sets (320 non-drugs and 311 drugs) were designed to have the same range of maximal experimental ΔG. To evaluate data sets and assess the similarity/dissimilarity of the distributions, a standard protocol of statistical analysis was followed (Appendix). The distribution of the data was checked, and it was found that the ΔG values in the sets and also in the entire database followed non-normal distributions ($p<0.001$). To check the distribution of an even larger sample of available experimental ΔG data, the same tests were performed for a set of more than 4,000 binding affinity values from the BindingDB [111] database and it showed a non-normal distribution as well. As the normality tests failed for the ΔG data sets, two non-parametric tests were applied and showed equal medians and distributions of ΔG between the drug and non-drug populations ($p>0.1$, $p>0.05$). In addition to the statistical tests, a high degree of overlap between the distributions of the two ΔG populations can be seen from the plot of their histograms (Fig. **1a**), and from the fitted mixed normal probability density functions (PDF, Fig. **1b**). The comparison of descriptive statistics also emphasizes the equality of drug and non-drug ΔG populations. The medians of the samples are in good agreement ($\Delta \approx 0.5$ kcal/mol) and the median difference between percentiles of the two samples (Appendix) is a marginal 3 % (Fig. **1c**). Details of the statistics are included as Supplementary Material.

In conclusion, a database of drug and non-drug compounds was collected wherein the two sets have ΔG distributions of significantly high similarity. Importantly, such criterion was not applied for the distribution of molecular properties and EIs of the two sets. Thus, it could be tested if the properties can describe general drug-likeness 'decoupling' effects common with ΔG. The outcome of this test is summarized in the next Section.

5.2. General Drug-Likeness Filters

Similarly to the previous section, the results of normality tests indicate that most of the investigated drug-likeness property filters (MW, NHA, W, logP) and the corresponding EIs (Section 4 and Appendix, EI_{MW}, EI_{NHA}, EI_W) are not normally distributed ($p<0.001$). In contrast with the previous section, the non-parametric tests of equivalence resulted in a highly significant difference ($p<0.001$) between the property/EI distribution of the drug and non-drug sets. There is a considerable increase in the medians of MW, NHA, and W with $\Delta \approx 150$, 10, and 2000 units respectively, for non-drugs compared with drugs. Similarly, the corresponding median percentile differences are in the range of 15-230%, which is significantly larger than that of ΔG (Fig. **1c**). The histograms and Probability Density Functions (PDF's) (Fig. **2a**, **b**, **e**, & Supplementary Material) show a change in the shape of the distributions. Whereas the ΔG distributions (Fig. **1a**, **b**) are rather rounded, well-defined peaks appear in the case of MW, NHA and logW, reflected also by a change in the kurtosis value from negative to positive. For drugs, a sharp peak and a high kurtosis value appear, while the non-drug histogram is flat with a long tail.

A similar separation of the two sets can be observed using EIs (Fig. **2c**, **d** & Supplementary Material). The EI_{MW} histograms (Fig. **2c**) do not resemble the non-separable ΔG distributions of drugs and non-drugs (Fig. **1a**). The partial separation of EI values seems to be a plausible consequence of the differentiating power of the parent MW. By definition, in EIs the populations of ΔG and MW or NHA are connected and the distributions of EI_{MW} and EI_{NHA} reflect the shape of the one-peaked MW (Fig. **2a**) or NHA distributions, which are more suitable candidates for statistical evaluations than the flat ΔG distributions with dual maxima (Fig. **1a**, **b**).

Whereas significant separation power of the filters can be concluded from the above analysis, a considerable overlap of the drug and non-drug histograms can also be observed especially in the cases of W and EIw where the distributions have an exponential shape (Supplementary Material). Notably, taking the logarithm of W (logW) resulted in separate peaks (Fig. **2e**). For logP, (Fig. **2f**) the drug population is centered in a well-defined peak in the hydrophobic region (logP>0), as can be expected for drugs [12, 42], whereas the distribution of non-drugs is similar to the case of ΔG. The considerable overlap of drug and non-drug populations in the hydrophobic region, along with a separate non-drug sub-population

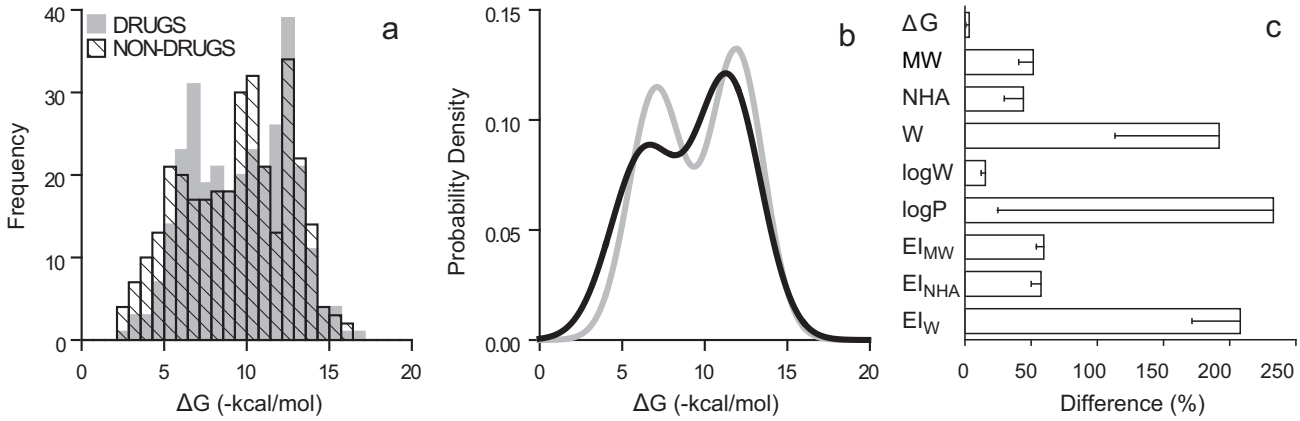

Fig. (1). Comparison of binding affinity distributions of sets of drugs and non-drugs. The two compound sets with N=311 and 320 members respectively, were designed to be non-separable by ΔG. Overlapping histograms of ΔG values in part (**a**) and two-component normal mixture probability density functions fitted to the histograms in part (**b**) reflect the similarity of the two datasets. In part (**c**), the median differences between the series of percentiles of the two sets are shown. Whereas the difference is marginal in the case of ΔG, it is significant for the filters. Error bars represent median absolute deviations.

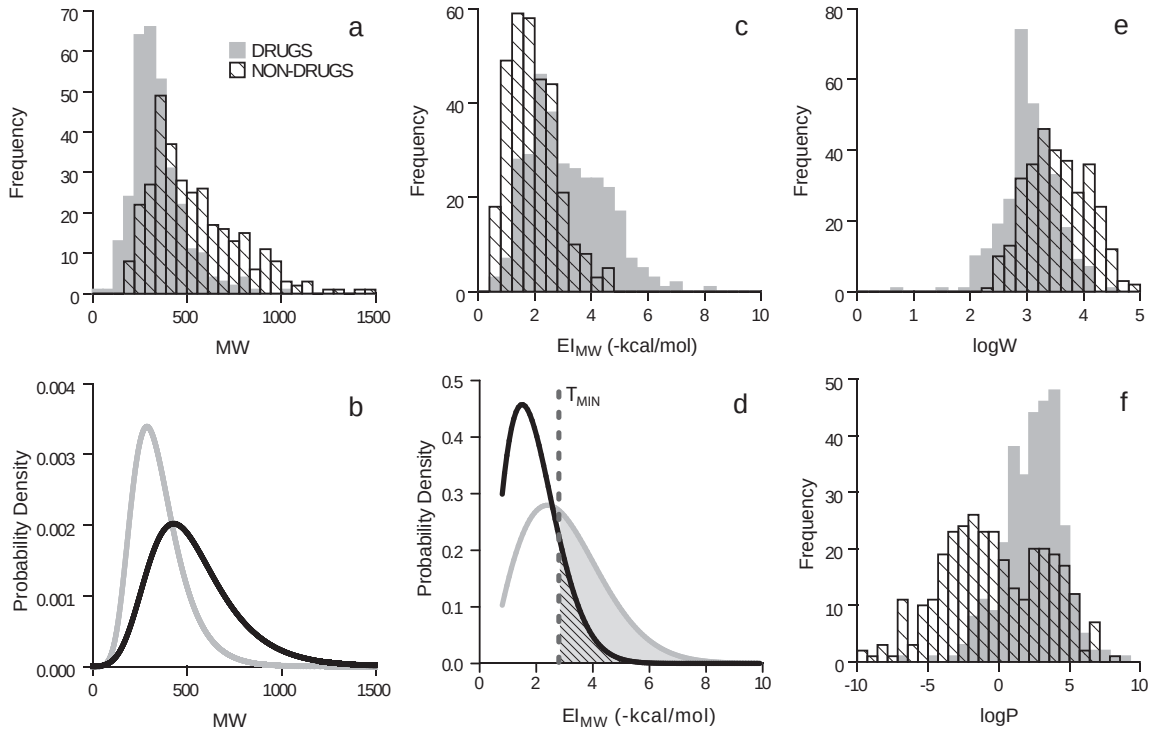

Fig. (2). Separation of sets of drugs and non-drugs by drug-likeness filters. Histograms in parts (**a**), (**c**), (**e**), (**f**), and fitted probability density functions in parts (**b**) and (**d**) reflect separation of the two compound sets by various filters. Part (**d**) also features key terms of this study with an example of EI_{MW}. The shaded area below the drugs group curve represents the sensitivity (σ) of EI_{MW} above the threshold T_{MIN}=2.8 kcal/mol. The ratio of this shaded area and the striped area below the non-drug curve shows that drugs can be found with three fold higher probability than non-drugs above T_{MIN} and by definition this equals the drug-likeness ratio (DR=3).

in the hydrophilic region (logP<0) explains the high spread of median differences at logP (Fig. **1c**).

5.3. Definition of Selectivity and Sensitivity of Drug-Likeness Filters

As it was shown in the previous section, sharp borders cannot be drawn between the partly overlapping drug and non-drug populations for the properties investigated. To achieve a coherent formulation of selectivity and sensitivity, fits of Probability Density Functions (PDF) of continuous distributions (Weibull, Gumbel, Exponential) were performed for histograms of the properties (Figs. **2b**, **d**). Using these explicit forms of PDFs, an analytical comparison of the distributions of drugs and non-drugs has become possible for the EI's and MW (Fig. **3** and Appendix). In the following discussion we will use the example of EI_{MW} for the introduction of PDF-based sensitivity and selectivity of the filters.

The probability that a drug adopts an EI_{MW} larger than a minimum threshold (T_{MIN}) is expressed as a percentage (Eqs. A4 and A6) and named sensitivity (σ) as it reveals whether a large enough section of the entire drug population is included in the region under question. A σ=51% is represented by an shaded area in Fig. (**2d**). In this case, 51% of the total drug population is located in the region above T_{MIN}. The larger the sensitivity of a filter, the fewer drugs are excluded erroneously above a minimum threshold T_{MIN}. Detailed definitions of probabilities are shown in the Appendix. Decidedly, σ is a necessary, but not a sufficient parameter of a property filter.

Further inspection of the fitted PDFs of EI_{MW} (Fig. **2d**) reveals that in the region starting from T_{MIN}, the probability that a drug adopts an EI_{MW} is three times higher than this probability for non-drugs. Thus, the ratio of the shaded area below the PDF curve of drugs and the striped area (Fig. **2d**) corresponding to non-drugs is three. Generalizing the previous observations, we introduce another measure of selectivity (Eqs. A5 and A7), the Drug-likeness Ratio (DR), relating the population of drugs with that of non-drugs by the ratio of their probabilities. In terms of the above-mentioned example, DR equals 3 as there is a three-fold higher chance for a compound to be a drug than a non-drug above T_{MIN}.

After fitting the PDFs, thresholds can be fine-tuned for a drug-likeness filter using the DR and σ functions as calibration curves (Fig. **4a**), i.e. the T_{MIN} value can be read from the curve plot at a required level of DR or σ. According to the relative position of DR and σ functions, drug-likeness filters can be categorized into three types (see also Appendix for details): those with limits of T_{MIN} (Fig. **4a**), both T_{MIN} and a maximum threshold (T_{MAX}, Fig. **4b**), or only T_{MAX} (Fig. **4c**). EI_{MW} can be categorized under the first type (Fig. **4a**). In the above-mentioned example (Fig. **2d**), a T_{MIN} of 2.8 kcal/mol is a realistic lower EI_{MW} threshold at levels of DR=3, and σ=51%. As σ decreases with increasing DR (Fig. **4a**), thresholds with DR>10 may have no practical importance.

In Table **4**, general thresholds calculated for all filters at DRs from 2 to 3, and σ>50% are listed. Compared with values from the literature (Table **1**), it can be concluded that the calibrated range of 129-369 for MW (Fig. **4b**) correspond to drugs. Calibrated thresholds of NHA, EI_{MW} and EI_{NHA} at similar DR and σ values are also located at the drug/lead border (Table **1**). For logP there are various data published and our estimated range of 0.7-4.3 between T_{MIN} and T_{MAX} agrees well with the values from the literature (Table **1**). The above results allow experimenting with calibration, and fine-tuning of thresholds at different DR and σ levels depending on the nature of desired applications, i.e. if hits, leads or drugs are investigated requiring small/large selectivity and sensitivity criteria, etc. Example thresholds at various DR and σ levels and details of the calculations can be found in the Supplementary Material. Importantly, while the above description

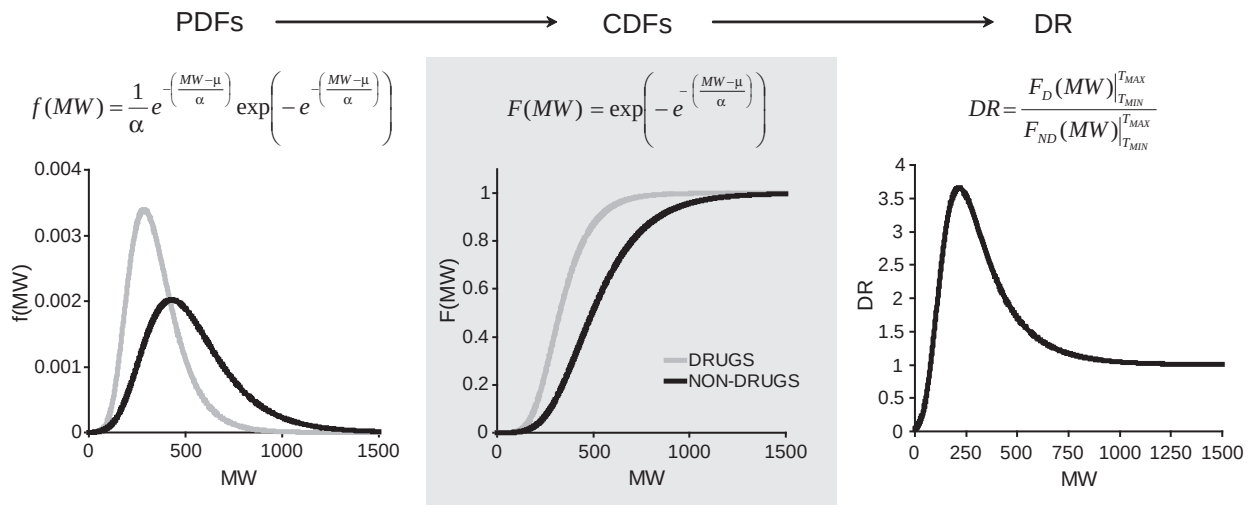

Fig. (3). An example of the use of fitted probability density function (PDF, f) and the corresponding cumulated density function (CDF, F) for the analytical calculation of selectivity and sensitivity measures DR and σ of MW. As F(MW)≈0 for small MWs, T_{MIN} was omitted from function DR. Gumbel distributions were fitted for both drugs (D) and non-drugs (ND) sets (see also Fig. **1**). Notably, the general functional formulae are provided in this figure and different scale (α) and location (μ) parameters were obtained for the two sets (see Supplementary Material for numerical values of the parameters and details of fit). The σ can be directly calculated (Eq. A6) from the CDFs according to $\sigma=100[F_D(T_{MAX})-F_D(T_{MIN})]$. Since the DR function has a maximum on the MW≤1500 domain investigated, T_{MIN} and T_{MAX} thresholds can be calculated (Fig. **4**) for DR values up to ca. DR=3.75. Plausibly, a DR≥ 1 is of interest.

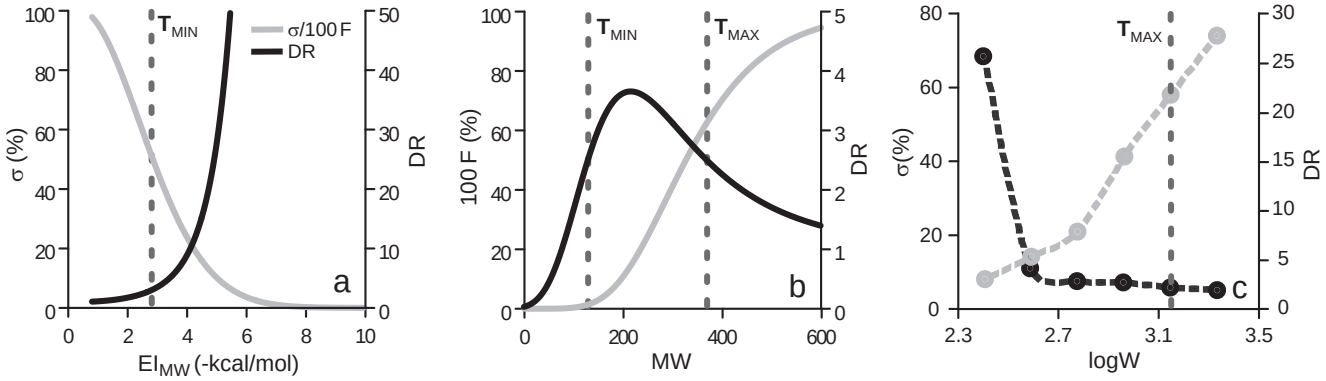

Fig. (4). Calibration curves of drug-likeness thresholds. The shape and relative location of sensitivity (σ) and Drug-likeness Ratio (DR) functions facilitate calibration of the three types of drug-likeness thresholds (T) of filters. (**a**) In the case of EI_{MW}, the DR function increases and σ decreases on the domain investigated. Thus, a minimum threshold (T_{MIN}=2.8 kcal/mol) can be calibrated (following the previous example of Fig. **2d**) with a DR=3, which is large enough that EI_{MW} can separate drugs from non-drugs. At the same time, the sensitivity of EI_{MW} is also acceptable (σ=51 %) for recognition of drugs above this threshold. (**b**) In case of MW, the DR function has a maximum, and therefore there are two thresholds (T_{MIN} and T_{MAX}) with the same DR value specifying a favorable MW interval with sufficiently high DR values. Here, $\sigma=100[F_D(T_{MAX})-F_D(T_{MIN})]$, where F_D is the cumulative distribution function of drugs. At higher DR values, i.e. narrower (T_{MIN}, T_{MAX}) intervals σ becomes smaller. (**c**) For a decreasing DR, the maximum of logW can be set (T_{MAX}), below which the separation of drugs from non-drugs is possible by logW. To note, the logW-related curves are not continuous functions, the points are derived from raw histogram data.

of filters with DR and σ were used for drug/non-drug (drug-likeness) separations, our present approach can easily be easily adopted to describe the filters in drug/lead or lead/hit relations (lead-likeness).

5.4. Disease-Specific Drug-Likeness

It is informative to characterize the discriminating power of the filters between drugs and non-drugs beyond general terms according to disease categories (Section 2.4). For this characterization, the set of drugs was divided into sub-sets by disease categories according to the classification of DrugBank [112]. Similarly to the case of general drug-likeness (Section 5.2), a non-drug companion with the closest ΔG was selected for each drug in each disease category. This method resulted in selected disease category sub-sets of non-drugs that are inseparable from the corresponding drugs by ΔG (Fig. **5**). In all cases, statistical comparisons of sub-sets of drugs and non-drugs were performed for ΔG and for all 8 filters. The overall results on separation of inter-quartile ranges are shown as a matrix (Fig. **5**), other details can be found in the Supplementary Material. (Notably, due to the relatively small number of drug/non-drug members of the sub-sets σ and DR were not calculated in this analysis by disease types. In forthcoming studies we plan to extend the selectivity and sensitivity calculation of the filters on large disease-specific data sets.)

In 70% of the cases, separation of the sub-sets at different levels can be observed (Fig. **5**), and in the remaining cases, the inter-quartile ranges of drugs and non-drugs are completely overlapping.

Table 4. Calibrated Thresholds, Selectivity, and Sensitivity of Property Filters

Property filter	General thresholds				Disease-specific thresholds
	T_{MIN}	T_{MAX}	DR	σ	T_{MIN}-T_{MAX}[c]
MW	129	369	2.5	61	206-322[d], 262-342[e], 258-342[f]
NHA[a]	9	27	2.0	67	
W[a]	-	2037	2.0	73	578-1180[f]
logW[a]	-	3.1	2.1	58	2.76-3.07[f]
logP[a]	0.7	4.3	2.6	67	1.02-3.27[g], 1.74-3.59[d]
EI_{MW}[b]	2.8	-	3.0	51	
EI_{NHA}[b]	4.2	-	3.0	52	
EI_W[b]	7.5	-	3.0	56	6.51-15.87[e], 6.01-17.78[h], 7.35-21.93[f]

Abbreviations. DR: drug-likeness ratio (selectivity); EI: efficiency index; logP: logarithm of octanol/water partition coefficient (small letters in brackets denote different logP definitions); MW: molecular weight; NHA: number of heavy atoms; σ: sensitivity; T: threshold; W: Wiener-index.
[a]Calibrated values of this filter were estimated from histograms and not from fitted distributions. [b]The dimension of EIs is –kcal/mol. [c]T_{MIN}-T_{MAX} is a (modified) inter-quartile range of the drugs set (no DR and σ values are given). [d]Musculo-skeletal system. [e]Nervous system. [f]Various. [g]Dermatologicals. [h]Respiratory system.

There is a minority (10%) of cases in which a high separation (>90%) was found. The "worst performance" occurred in the categories of Antineoplastic agents and Anti-infectives, with cancer drugs presented in the former category. This finding implies that the failure of new drug discovery in these areas [113, 114] may be partly due to the inefficacy of drug-likeness filters investigated in this study. There are also numerous cases where only a partial separation was achieved as, e.g. the cardiovascular system compounds at MW (Fig. **5b**).

Based on these observations, drug-likeness thresholds (Table **4**) were estimated for disease groups using the inter-quartile ranges of properties that provide the highest level of separation (>90%), which show agreement with Table **2**. These disease-specific thresholds provide in some cases (MW, logP) narrower drug-likeness ranges than the general thresholds found within reference values available in the literature.

Whereas the evaluation of the above mentioned negative or partly successful cases is not an easy task, certain positive results can be readily explained. For example, logP, well-known to describe skin permeability [115-117], performed well for the category of dermatological drugs which require absorption through the skin (Fig. **5c**). Similarly, MW, a good filter of nervous system drugs in this study (Fig. **5d**), describes blood-brain barrier penetration [66, 67], an important issue of drug design for CNS diseases. The MW-threshold calculated for nervous system diseases (Table **4**) is in good agreement with the MW<400 value recommended by other studies [34, 69].

Interestingly, the performance of EIs does not always correspond to their parent ligand-based properties (MW, NHA, W), emphasizing their different information contents. Besides the well-known drug-likeness filters such as MW and logP, the recently introduced EI_W [103] was one of the best separators according to the present analysis, emphasizing the benefits of using EIs.

6. SUMMARY AND FUTURE OUTLOOK

Molecular properties of drug candidates have been extensively used as drug-likeness filters of compound libraries. In the present

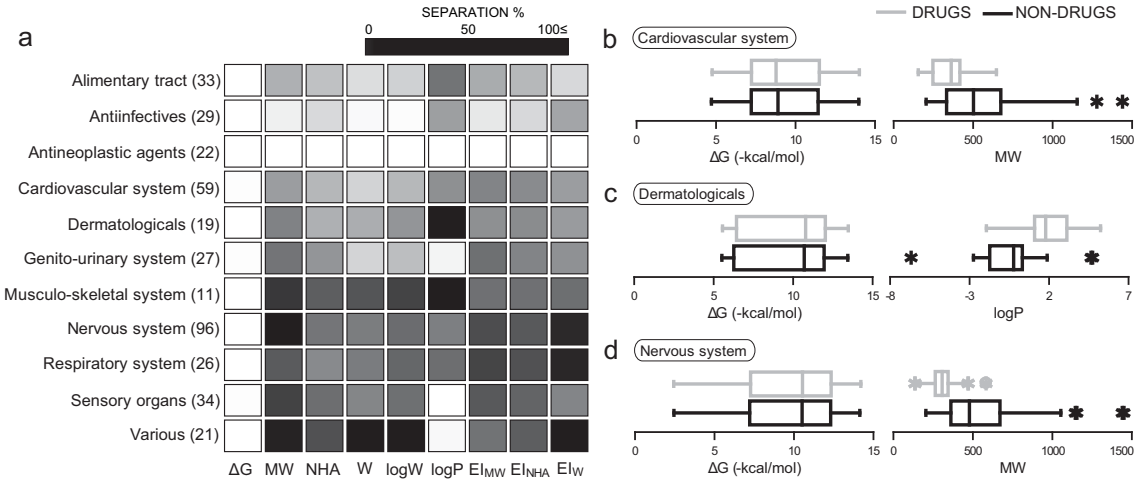

Fig. (5). Disease-specificity of drug-likeness filters. The set of drugs (N=311) used in this study was split into sub-sets according to various disease categories. The number (N) of members of these sub-sets is marked in brackets on the left side of the shaded matrix (**a**). Sub-sets of non-drug compounds possessing the same ΔG distribution as the sub-sets of drugs were formed. Each cell of the shaded matrix shows the level of separation of the inter-quartile range of a sub-set of drugs from that of non-drugs according to ΔG or a filter. The separation is 0 % if the two ranges are completely overlapping. This is the situation for all disease categories in the ΔG column due to the aforementioned selection of sub-sets of non-drugs. The separation is between 0 and 100 % if there is a partial overlap between the ranges as shown in the box plot for cardiovascular system drugs according to MW (**b**). If there is no overlap in the ranges, then the separation is 100 %. The latter situation is featured in examples of box plots of dermatologicals according to logP (**c**), and nervous system drugs according to MW (**d**).

review, a distinction was made between general and specific drug-likeness. While the investigated properties significantly differentiated between the sets of oral drugs and non-drugs in general, a considerable overlap remained between the two sets. Certain disease types or drug administration routes may require specific filter values instead of the broader, general ones. It was also discussed to which extent the molecular properties are coupled to ΔG. Statistical comparison of drug sets with non-drugs of similar binding affinity and use of selectivity and sensitivity measures were introduced as an improved description of the overlapping distributions of filter values. With the new measures filtering thresholds gain statistical meaning: namely, their selectivity against non-drugs (DR) and sensitivity for drugs (σ). In addition to the positive results of the general drug-likeness concept, relevant criticism and limits of its applicability were also surveyed.

Filtering thresholds can help in the future design of standardized, compound libraries assembled for binding assays, HTS, or other *in vivo* tests. However, precise statistical calibration of filtering thresholds − as shown in this work − may be required beyond simple descriptive statistics (mean, median) to assess full reliability of the thresholds. Disease-, target-, or administration-specific drug-likeness filters may help the design of focused libraries which may become a competitive alternative to general compound sets. Molecular property and EI-based filters have an increasing impact also in fragment-based design [101, 118], and in the chemical optimization of physico-chemical properties of natural products [113, 119] or other lead compounds.

CONFLICT OF INTEREST

None declared.

ACKNOWLEDGEMENT

This study was financed by the European Social Fund (grant agreement no. TAMOP-4.2.1/B-09/1/KMR-2010-0003), an Estonian Science Foundation grant JD80, an Estonian Ministry for Education and Research grant SF0140031Bs09, and an Innove Foundation grant 1.0101-0310. C.H.'s work was financed by a Bolyai Scholarship of the Hungarian Academy of Sciences. We are grateful to Professor David van der Spoel for critically reading of the manuscript and Professor David Meredith (Department of Mathematics, San Francisco State University) for providing the program Xplore.

APPENDIX

Collection and Verification of Compound Sets of Drugs and Non-Drugs

The structure of 311 drugs and 320 non-drugs and their experimental binding affinities (mostly as inhibition equilibrium constants, K_i) were collected from the following sources: PDBbind v2005 [120], KiBank [121, 122], SCORPIO [123], and from a previous study [124]. The BindingDB [110] database was also used for normality test comparisons. Where it was necessary, ΔG (precisely the standard Gibbs free energy change, ΔG^o − the standard sign is omitted in this study for simplicity) values were obtained from K_i by $\Delta G = RT\ln K_i$, using T=25 °C (298.15 K). The complete sets of drugs and non-drugs, as well as their raw and converted ΔG values are available as Appendices of the Supplementary Material. Similarly to other studies [72, 81], maximal ΔG values, i.e. ligand binding affinities corresponding to the complex with the relevant, strongest binding protein partner were collected. In the case of drugs, ΔG values with the pharmacologically relevant targets were considered. Whereas ΔG correspond to a multi-molecular interaction between the ligand compound, target, and solvent shell, it has been shown that ΔG is also related to molecular properties (MW, NHA, logP) [72, 78, 81] of the ligand only, as these properties hold information on both enthalpic (ΔH) and entropic (ΔS) constituents [103] of ΔG (through $\Delta G = \Delta H - T\Delta S$). Consequently, there is a ligand-based part of ΔG explained by the above properties (see also Section 3), which is constant regardless of the actual target. Since a compound can bind as a ligand to various targets, it can adopt different ΔG values due to target-specific interactions (ΔH) and, therefore, the maximal experimental ΔG, i.e. the maximum ΔG value of a compound with its relevant target(s), were collected for both sets in the present study. Using these maximum ΔG values helps decreasing target-specificity of the interaction (ΔH) part as they correspond to the ideal binding affinity of a compound.

The selection procedures of the two sets are following. (i) The list of all small-molecule approved drugs was downloaded from the DrugBank database, which also contain disease-specific data on drugs approved by the FDA (U.S. Food and Drug Administration agency). A standard, programmed procedure was applied to ensure purity the two sets. (ii) The ligand names were extracted from the PDB files, and queried in the DrugBank [112] database to identify those ligands that are FDA-approved drugs, and to avoid contamination of drug molecules in the non-drug collection. (iii) The set of non-drugs was designed to have overlapping binding affinity distribution with the set of drugs (Figs **1a**, **b**). While non-drugs were selected with a similar ΔG as drugs, but such criterion was not applied for the filtering properties of the compounds. Thus, there were no circumstances in the sampling which affected the composition of non-drugs set so as to determine/guarantee its similar/different property (MW, NHA, etc.) distribution compared with the drugs set (Fig. **1c**).

Filters

There are various properties applied in drug design as size, structural, or property filters. Whereas size filters such as molecular weight (MW) or number of heavy atoms (NHA) require solely the knowledge of a compound's atomic composition, structural descriptors also involve intra-molecular connectivity. The Wiener index (W) is a typical structural descriptor reflecting the branching and complexity of the molecule. The W is a robust measure as it does not depend on the molecular conformation. To be able to calculate the W of a compound, knowledge of its Lewis-structure is sufficient (Eq. A1). In this study, its logarithm (logW) is also used.

$$W = \frac{1}{2}\sum_{i,j}^{NHA} d_{ij} \qquad \text{Eq. A1}$$

where d_{ij} is the number of bonds in the shortest path connecting the pair of atoms i and j in the molecule. There are also other property filters, e.g. the logarithm of octanol/water partition coefficient (logP) which is generally applied as a measure of hydrophobicity for a non-ionized compound. The binding affinity and the aforementioned size or structural properties have been combined into hybrid filters called the efficiency indices (EI). The EIs are ΔGs normalized by these filters (Eq. A2). Exponents of ten were used as multipliers in the formulae to obtain human-readable EI values.

$$EI_{MW} = 100\frac{\Delta G}{MW}$$
$$EI_{NHA} = 10\frac{\Delta G}{NHA} \qquad \text{Eq. A2}$$
$$EI_{W} = 1000\frac{\Delta G}{W}$$

The program XLOGP v2.060 [125] was used to calculate the logarithm of octanol/water partition coefficient (logP) by an atom-additive method including correction factors. The calculations of molecular formula and number of heavy atoms (NHA), molecular

weight (MW), and Wiener index were performed with Marvin Beans v4.1.861 [126]. The experimental ΔG's, calculated physicochemical properties and EI's are available as an Appendix in the Supplementary Material.

Descriptive Statistics

To check the similarity/dissimilarity of the distributions, a standard protocol of statistical analysis was followed for all ΔG and filter sets. A complete descriptive statistics including histogram (Figs. **2a**, **c**, **e**, **f**), minimum, maximum, range, median, median absolute deviation, arithmetic mean, standard error of arithmetic mean, 95.0% confidence interval, trimmed mean (10%, two sided), standard deviation, variance, coefficient of variation, skewness, kurtosis and a data vector of percentiles (1, 5, 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, 95, 99%) was calculated for all data sets with program package Systat 12_[127]. The median of differences (%, Fig. **1c**) between vectors ($\vec{p}$) of tabulated percentiles of two sets (drugs and non-drugs) was calculated according to Eq. A3.

$$\text{Difference}(\%) = \text{median}\,(\vec{p}_{DIFF})\ ;\ \{p_{DIFF}\}_i = \frac{100\ |\{p_{DRUGS}\}_i - \{p_{NON\text{-}DRUGS}\}_i\ |}{\min(|\{p_{DRUGS}\}_i\ |;|\{p_{NON\text{-}DRUGS}\}_i\ |)} \qquad \text{Eq. A3}$$

Where $\vec{p}_{DIFF}$ is the difference vector and $\{p_{...}\}_i$ denotes the element of a vector. The spread of $\vec{p}_{DIFF}$ was given as median absolute deviation. Results of descriptive statistics are tabulated in the Supplementary Material**.**

Statistical Tests

The Shapiro-Wilk [128], Kolmogorov-Smirnov [129], and Anderson-Darling [130] tests were applied to check if the data sets came from a normally distributed population (α=0.05). The null hypothesis was that the population is normally distributed. If the p-value was smaller than significance level α, then the null hypothesis was rejected (the data are not from a normally distributed population). If the p-value was larger than α, then the null hypothesis that the data came from a normally distributed population was accepted. The statistics and p-values are tabulated in the Supplementary Material. As the data populations are not normally distributed, non-parametric tests are valuable, since they do not require assumptions on the distribution of the population and therefore are sometimes called distribution-free [131]. Thus, in the present study the non-parametric two-sided Kruskal-Wallis test (also called Wilcoxon rank sum test or Mann-Whitney [132] U test, α=0.1) and the two-sided Kolmogorov-Smirnov two sample test (α=0.05) were used to decide if two data sets came from the same population. The null hypothesis was that the two samples came from the same population and have the same distribution. If the p-value was less than the α level, then the null hypothesis was rejected (the data are not from the same distribution). If the p-value was greater than α, then the null hypothesis that the data came from the same population was accepted. The statistics and p-values are tabulated in the Supplementary Material. All tests were performed with Systat 12, many cases were counterchecked and p-values were calculated in parallel with the program R [133].

Fitting Distributions

In all cases where the data allowed, PDF's of the following 21 distributions were fitted to histograms of each data sets (and their parameters estimated by the respective methods) using Systat 12. Beta, Chi-square, Erlang, Gamma, Gumbel, Logistic, Loglogistic, Smallest extreme value (method of moments); Normal, Lognormal, Logit normal, Exponential, Double exponential (Laplace), Gompertz, Inverse Gaussian (Wald), Pareto, Rayleigh, Weibull, Uniform, (maximum likelihood method); Cauchy (method of quantiles or order statistics); Triangular (modified maximum likelihood and moments).

In all cases, 12 bin histograms were prepared for the fits. In the case of mixed normal distribution (Fig. **1b**), and for refinement of some fits (especially for calculation of location parameters of Weibull distributions), the software Dataplot [134] along with the probability plot correlation coefficient plot (PPCC) method was used. Quality of fits was confirmed by Kolmogorov-Smirnov and Anderson-Darling tests with Systat 12. Only highly significant PDF's (α=0.1) were selected for further use. The analytical form of PDFs (Fig. **2b**, **d**) facilitated the mathematically accurate calculation of calibration of thresholds (Fig. **4a**, **b**). Statistics of tests of fit, formulae of selected distributions and values of their location, shape and scale parameters of the PDF are listed in the Supplementary Material.

Calibration of Thresholds

The probability (P_D) that a filter χ adopts a value between thresholds T_{MIN} and T_{MAX} for drugs is expressed as a percentage and named sensitivity (σ) in this study (Eq. A4), as it reveals whether a large enough section of the entire drug population is included in the region under question. The random variable ξ_D^{χ} corresponds to the statistical event when a filter χ adopts a value for drugs (D).

$$\sigma = 100 P_D(T_{MIN} \le \xi_D^{\chi} \le T_{MAX}) \qquad \text{Eq. A4}$$

The drug-likeness ratio (DR) is expressed (Eq. A5) as the ratio of P_D and the corresponding probability for non-drugs (P_{ND}).

$$DR = \frac{P_D(T_{MIN} \le \xi_D^{\chi} \le T_{MAX})}{P_{ND}(T_{MIN} \le \xi_{ND}^{\chi} \le T_{MAX})} \qquad \text{Eq. A5}$$

In the cases where fitted continuous PDFs are available for drugs (f_D) and non-drugs (f_{ND}), the σ and DR of a filter χ can be expressed as Eqs. A6, and A7 respectively.

$$\sigma = 100 \int_{T_{MIN}}^{T_{MAX}} f_D(\chi) d\chi \qquad \text{Eq. A6}$$

$$DR = \frac{\int_{T_{MIN}}^{T_{MAX}} f_D(\chi) d\chi}{\int_{T_{MIN}}^{T_{MAX}} f_{ND}(\chi) d\chi} \qquad \text{Eq. A7}$$

Eqs. A6 and A7 and the cumulative distribution functions could be used in cases of χ = EI_{MW}, EI_{NHA}, EI_W, and MW.

Depending on the types of the DR and σ functions, i.e. the relative location of the f functions, there are three cases to consider (Fig. **4**). (I) If DR is increasing on the investigated domain of filter χ, then T_{MAX}=+∞ and a T_{MIN} can be calculated. This situation was experienced at the EI's. (II) If DR has a maximum on the domain then both T_{MIN} and T_{MAX} can be calculated as in the case of MW. (III) Finally, if DR is decreasing then T_{MIN}=-∞ and T_{MAX} can be calculated as for logW.

The T_{MIN} and/or T_{MAX} thresholds were calculated by solution of Eqs. A6 and A7, for a set of different σ and DR values using the integral forms, i.e. the cumulative distribution functions of the respective PDF's at χ = MW, EI_{MW}, EI_{NHA}, EI_W. The equations were solved with the aid of Xplore, a program by Prof. David Meredith (Department of Mathematics, San Francisco State University). In the cases where continuous PDF's (χ = NHA, logP, W, logW) could not be fitted, the histograms were used to estimate

thresholds applying the definitions of Eqs. A4, and A5**.** The details of the calculations can be found in the Supplementary Material.

Disease Specificity

A set of 309 drugs of this study (excluding the very small molecules ethanol and piperazine) was divided into sub-sets according to the 14 disease categories of the DrugBank database. These 14 disease categories were: Alimentary tract and metabolism, Blood and blood forming organs, Cardiovascular system, Dermatologicals, Genito-urinary system and sex hormones, Systemic Hormonal preparations (excluding sex hormones and insulins), Antiinfectives for systemic use, Antineoplastic and immunomodulating agents, Musculo-skeletal system, Nervous system, Antiparasitic products, including insecticides and repellents, Respiratory system, Sensory organs, and others which do not fit in the above categories (Various). Non-drug molecules having the closest ΔG were selected for each member of each drug sub-sets using an in-house program. The difference in ΔG was set not to exceed 1 kcal/mol for the drug-non-drug pairs, and indeed it was much less for all cases. One non-drug was used only once for each sub-set. Thus, two sub-sets of compounds (drugs and non-drugs) with overlapping ΔG distributions were available for a disease-specific analysis. Descriptive statistics (median, median absolute deviation, mean, standard deviation, 1st and 3rd quartiles, minimum and maximum) were calculated for all disease categories, and filters by the same program. Boxplots generated by the program R were used for visual comparison of distributions. Only sub-sets having N>10 members were used for final discussion (Fig. **5**). Details of the disease-specific analysis can be found in the Supplementary Material.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] Hopkins, A.L. Network pharmacology. *Nat. Biotechnol.*, **2007**, *25*(10), 1110-1111.

[2] Yildirim, M.A.; Goh, K.-I.; Cusick, M.E.; Barabási, A.L.; Vidal, M. Drug-target network. *Nat. Biotechnol*., **2007**, *25*(10), 1119-1126.

[3] Shoichet, B.K. Virtual screening of chemical libraries. *Nature*, **2004**, *432*, 862-865.

[4] Hopkins, A.L.; Witty, M.J.; Nwaka, S. Mission possible. *Nature*, **2007**, *449*, 166-169.

[5] Villoutreix, B.O.; Eudes, R.; Miteva, M.A.; Structure-based ligand screening: Recent success stories. *Comb. Chem. High T. Scr*., **2009**, *12*, 1000-1016.

[6] Rajamani, R.; Good, A. C. Ranking poses in structure-based lead discovery and optimization: Current trends in scoring function development. *Curr. Opin. Drug Disc. Devel.*, **2007**, *110*, 308-315.

[7] Dobson, P. D.; Kell, D. B. Opinion - Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nat. Rev. Drug Disc*., **2008**, *7*, 205-220.

[8] Karelson, M. *Molecular descriptors in QSAR/QSPR*. Wiley- Interscience publication: New York, 2000.

[9] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*, *2nd ed.*; WILEY-VCH: Weinheim, Germany, 2009

[10] Amidon, G.L.; Anik, S.T. Comparison of several molecular topological indexes with molecular surface area in aqueous solubility estimation. *J. Pharm. Sci*. **1976**, *65*, 801-806.

[11] Gutman, I.; Körtvélyesi, T. Wiener indices and molecular surfaces. *Z. Naturforsch*. **1995**, *50a*, 669–671.

[12] Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature*, **2004**, *432*, 855-861.

[13] Katritzky, A.R.; Kuanar, M.; Slavov, S.; Dobchev, D.A.; Fara, D.C.; Karelson, M.; Acree, W.E. Correlation of blood-brain penetration using structural descriptors. *Bioorgan. Med. Chem.*, **2006**, *14*, 4888-4917.

[14] Hitchcock, S.A. Blood-brain barrier permeability considerations for CNS-targeted compound library design. *Curr. Opin. Chem. Biol*., **2008**, *12*, 318-323.

[15] Charifson, P.S.; Walters, W.P. Filtering databases and chemical libraries. *Mol. Divers*., **2000**, *5*, 185-197.

[16] Charifson, P.S.; Walters, W.P. Filtering databases and chemical libraries. *J. Comput. Aid. Mol. Des.*, **2002**, *16*, 311-323.

[17] Tounge, B.A.; Pfahler, L.B.; Reynolds, C.H. Chemical information based scaling of molecular descriptors: A universal chemical scale for library design and analysis. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 879-884.

[18] Walters, W.P.; Murcko, M.A. Prediction of 'drug-likeness'. *Adv. Drug Deliver. Rev*., **2002**, *54*, 255-271.

[19] Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev*., **2003**, *23*(3), 302-321.

[20] Zheng, S.; Luo, X.; Chen, G.; Zhu, W.; Shen, J.; Chen, K.; Jiang, H. A new rapid and effective chemistry space filter in recognizing a druglike database. *J. Chem. Inf. Model*., **2005**, *45*, 856-862.

[21] Jorgensen, W.L. The many roles of computation in drug discovery. *Science*, **2004**, *303*, 1813-1818.

[22] Böcker, A.; Schneider, G.; Teckentrup, A. Status of HTS data mining approaches. *QSAR Comb. Sci.*, **2004**, *23*, 207-213.

[23] Blomberg, N.; Cosgrove, D.A.; Kenny, P.W.; Kolmodin, K. Design of compound libraries for fragment screening. *J. Comput. Aid. Mol. Des.*, **2009**, *23*, 513-525.

[24] Abad-Zapatero, C.; Metz, J.T. Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today*, **2005**, *10*(7), 464-469.

[25] Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem. Int. Ed.*, **2005**, *44*, 1504-1508.

[26] Blum, L.C.; Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc*., **2009**, *131*, 8732-8733.

[27] Martin, E.J.; Critchlow, R.E. Tailoring combinatorial libraries for drug discovery. *J. Comb. Chem.*, **1999**, *1*, 32-45.

[28] Walters, W.P.; Ajay; Murcko, M.A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol*., **1999**, *3*, 384-387.

[29] 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 Deliv. Rev.*, **1997**, *23*, 3-25.

[30] Fecik, R.A.; Frank, K.E.; Gentry, E.J.; Menon, S.R.; Mitscher, L.A.; Telikepalli, H. The search for orally active medications through combinatorial chemistry. *Med. Res. Rev.*, **1998**, *18*(3), 149-185.

[31] Clark, D.E.; Pickett, S.D. Computational methods for the prediction of 'drug-likeness'. *Drug Discov. Today*, **2000**, *5*(2), 49-58.

[32] Ajay; Walters, W.P.; Murcko, M.A. Can we learn to distinguish between "drug-like" and "Nondrug-like" molecules? *J. Med. Chem*., **1998**, *41*, 3314-3324.

[33] Ghose, A.K.; Viswanadhan, V.N.; Wendolowski, J.J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem*., **1999**, *1*, 55-68.

[34] Sakaeda, T.; Okamura, N.; Nagata, S.; Yagami, T.; Horinouchi, M.; Okumura, K.; Yamashita, F.; Hashida, M. Molecular and pharmacokinetic properties of 222 commercially available oral drugs in humans. *Biol. Pharm. Bull*., **2001**, *24*(8), 835-940.

[35] 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**, *45*, 2615-2623.

[36] Lu, J.J.; Crimin, K.; Goodwin, J.T.; Crivori, P.; Orrenius, C.; Xing, L.; Tandler, P.J.; Vidmar, T.J.; Amore, B.M.; Wilson, A.G.E.; Stouten, P.F.W.; Burton, P.S. Influence of molecular flexibility and polar surface area metrics on oral bioavailability in the rat. *J. Med. Chem.*, **2004**, *47*, 6104-6107.

[37] Hann, M.M.; Leach, A.R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci*., **2001**, *41*, 856-864.

[38] Proudfoot, J.R. Drugs, leads, and drug-likeness: An analysis of some recently launched drugs. *Bioorgan. Med. Chem. Lett.*, **2002**, *12*, 1647-1650.

[39] Feher, M.; Schmidt, J.M. Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 218-227.

[40] Wenlock, M.C.; Austin, R.P.; Barton, P.; Davis, A.M.; Leeson, P.D. A comparison of physicochemical profiles of development and marketed drugs. *J. Med. Chem.*, **2003**, *46*, 1250-1256.

[41] Lipinski, C.A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol*., **2000**, *44*, 235-249.

[42] Leeson, P.D.; Davis, A.M. Time-related differences in the physical property profiles of oral drugs. *J. Med. Chem.*, **2004**, *47*, 6338-6348.

[43] Vieth, M.; Siegel, M.G.; Higgs, R.E.; Watson, I.A.; Robertson, D.H.; Savin, K.A.; Durst, G.L.; Hipskind, P.A. Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem*., **2004**, *47*, 224-232.

[44] Proudfoot, J.R. The evolution of synthetic oral drug properties. *Bioorgan. Med. Chem. Lett*., **2005**, *15*, 1087-1090.

[45] Leeson, P.D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.*, **2007**, *6*, 881-890.

[46] Perola E. An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. *J. Med. Chem.*, **2010**, *53*, 2986–2997.

[47] Schneider, N.; Jäckels, C.; Andres, C.; Hutter, M.C. Gradual in silico filtering for druglike substances. *J. Chem. Inf. Model*., **2008**, *48*, 613-628.

[48] Vieth, M.; Sutherland, J.J. Dependence of molecular properties on proteomic family for marketed oral drugs. *J. Med. Chem*., **2006**, *49*, 3451-3453.

[49] Tyrchan, C.; Blomberg, N.; Engkvist, O.; Kogej, T.; Muresan, S. Physicochemical property profiles of marketed drugs, clinical candidates and bioactive compounds. *Bioorgan. Med. Chem. Lett*., **2009**, *19*, 6943-6947.

[50] Krämer, S.D.; Lombardi, D.; Primorac, A.; Thomae, A.V.; Wunderli-Allenspach, H. Lipid-bilayer permeation of drug-like compounds. *Chem. Biodivers.*, **2009**, *6*, 1900-1916.

[51] Vistoli, G.; Pedretti, A.; Testa, B. Assessing drug-likeness - what are we missing? *Drug Discov. Today*, **2008**, *13*(7/8), 285-294.

[52] Kubinyi, H. Drug research: myths, hype and reality. *Nat. Rev. Drug Discov.*, **2003**, *2*, 665-668.

[53] Ganesan, A. The impact of natural products upon modern drug discovery. *Curr. Opin. Chem. Biol*., **2008**, *12*, 306-317.

[54] Lajiness, M.S.; Vieth, M.; Erickson, J. Molecular properties that influence oral drug-like behaviour. *Curr. Opin. Drug Disc.*, **2004**, *7*(4), 470-477.

[55] Bhal, S.K.; Kassam, K.; Peirson, I.G.; Pearl, G.M. The rule of five revisited: Applying logD in place of logP in drug-likeness filters. *Mol. Pharm.*, **2007**, *4*(4), 556-560.

[56] Good, A.C.; Hermsmeier, M.A. Measuring CAMD technique performance. 2. How "druglike" are drugs? Implications of random test set selection exemplified using druglikeness classification models. *J. Chem. Inf. Model*., **2007**, *47*, 110-114.

[57] Gimenez, B.G.; Santos, M.S.; Ferrarini, M.; Fernandes, J.P.S. Evaluation of blockbuster drugs under the rule-of-five. *Pharmazie*, **2010**, *65*(2), 148-152.

[58] Zhang, M.-Q.; Wilkinson, B. Drug discovery beyond the 'rule-of-five'. *Curr. Opin. Biotech*., **2007**, *18*, 478-488.

[59] Dobson, P.D.; Patel, Y.; Kell, D.B. 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discov. Today*, **2009**, *14*(1/2), 31-40.

[60] Tronde, A.; Nordén, B.; Marchner, H.; Wendel, A.-K.; Lennernäs, H.; Hultkvist Bengtsson, U.; Pulmonary absorption rate and bioavailability of drugs *in vivo* in rats: Structure-absorption relationships and physicochemical profiling of inhaled drugs. *J. Pharm. Sci.*, **2003**, *92*(6), 1216-1233.

[61] Ritchie, T.J.; Luscombe, C.N.; Macdonald, S.J.F. Analysis of the calculated physicochemical properties of respiratory drugs: Can we design for inhaled drugs yet? *J. Chem. Inf. Model.*, **2009**, *49*, 1025-1032.

[62] Morphy, R. The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds. *J. Med. Chem*., **2006**, *49*, 2969-2978.

[63] Paolini, G.V.; Shapland, R.H.B.; van Hoorn, W.P.; Mason, J.S.; Hopkins, A.L. Global mapping of pharmacological space. *Nat. Biotechnol.*, **2006**, *24*(7), 805-815.

[64] Ajay; Bemis, G.W.; Murcko, M.A. Designing libraries with CNS activity. *J. Med. Chem.*, **1999**, *42*, 4942-4951.

[65] Kelder, J.; Grootenhuis, P.D.J.; Bayada, D.M.; Delbressine, L.P.C.; Ploemen, J.-P. *Pharm. Res*., **1999**, *16*(10), 1514-1519.

[66] Levin, V.A. Relationship of octanol-water partition coefficient and molecular-weight to rat-brain capillary-permeability. *J. Med. Chem*., **1980**, *23*, 682-684.

[67] Kaliszan, R.; Markuszewski, M. Brain/blood distribution described by a combination of partition coefficient and molecular mass. *Int. J. Pharm*., **1996**, *145*, 9-16.

[68] Barn, D.; Caulfield, W.; Cowley, P.; Dickins, R.; Bakker, W.I.; McGuire, R.; Morphy, J.R.; Rankovic, Z.; Thorn, M. Design and synthesis of a maximally diverse and druglike screening library using REM resin methodology. *J. Comb. Chem*., **2001**, *3*, 534-541.

[69] Chico, L.K.; Van Eldik, L.J.; Watterson, D.M. Targeting protein kinases in central nervous system disorders. *Nat. Rev. Drug Discov*., **2009**, *8*, 892-909.

[70] Garcia-Sosa, A.T.; Mancera, R.L. Free energy calculations of mutations involving a tightly bound water molecule and ligand substitutions in a ligand-protein complex. *Mol. Inf.*, **2010**, *29*, 589-600.

[71] Lie, M.A.; Thomsen, R.; Pedersen, C.N.S; Schiøtt, B.; Christensen, M.H. Molecular docking with ligand attached water molecules. *J. Chem. Inf. Model*., **2011**, *51*(4), 909-917.

[72] Kuntz, I.D.; Chen, K.; Sharp, K.A.; Kollman, P.A. The maximal affinity of ligands. *Proc. Natl. Acad. USA*, **1999**, *96*, 9997-10002.

[73] Reynolds, C.H.; Bembenek, S.D.; Tounge, B.A. The role of molecular size in ligand efficiency. *Bioorgan. Med. Chem. Lett*., **2007**, *17*, 4258-4261.

[74] Reynolds, C.H.; Bembenek, S.D.; Tounge, B.A. Ligand binding efficiency: trends, physical basis, and implications. *J. Med. Chem.*, **2008**, *51*, 2432–2438.

[75] Ferenczy, G.G.; Keserü, G.M. Enthalpic efficiency of ligand binding, *J. Chem. Inf. Model*., **2010**, *50*, 1536-1541.

[76] Ferrara, P.; Gohlke, H.; Price, D.J.; Klebe, G.; Brooks, C.L. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem*., **2004**, *47*, 3032-3047.

[77] Velec, H.F.G.; Gohlke, H.; Klebe, G.; DrugScore^{CSD}-Knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem*., **2005**, *48*, 6296-6303.

[78] Wells, J.A.; McClendon, C.L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, **2007**, *450*, 1001-1009.

[79] Kim, R.; Skolnick, J. Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem*., **2008**, *29*, 1316–1331.

[80] Olsson, T.S.G.; Williams, M.A.; Pitt, W.R.; Ladbury, J.E. The Thermodynamics Of Protein-Ligand Interaction And Solvation: Insights For Ligand Design. *J. Mol. Biol.*, **2008**, *384*, 1002-1017.

[81] Cheng, A.C.; Coleman, R.G.; Smyth, K.T.; Cao, Q.; Soulard, P.; Caffrey, D.R.; Salzberg, A.C.; Huang, E.S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol*., **2007**, *25*(1), 71-75.

[82] Perozzo, R.; Folkers, G.; Scapozzo, L. Thermodynamics of Protein-Ligand Interactions: History, Present, and Future Aspects. *J. Recept. Sig. Transd.*, **2004**, *24*, 1–52.

[83] Reynolds, C.H.; Holloway, M.K. Thermodynamics of ligand binding and efficiency. *ACS Med. Chem. Lett*., **2011**, *2*(6), 433–437.

[84] Brooijmans, N.; Kuntz, I.D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct*., **2003**, *32*, 335–373.

[85] Brady, G.P.; Sharp, K.A. Entropy in protein folding and in protein-protein interactions. *Curr. Opin. Struct. Biol*., **1997**, *7*, 215-221.

[86] Zhou, H.-X.; Gilson, M.K. Theory of free energy and entropy in noncovalent binding. *Chem. Rev*., **2009**, *109*, 4092-4107.

[87] Jacobsson, M.; Karlén, A. Ligand bias of scoring functions in structure-based virtual screening. *J. Chem. Inf. Model*., **2006**, *46*, 1334-1343.

[88] Steinberg, I.Z.; Scheraga, H.A. Entropy changes accompanying association reactions of proteins. *J. Biol. Chem.*, **1963**, *238*(1), 172-181.

[89] Karplus, M.; Kushick, J.N. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, **1981**, *14*, 325-332.

[90] Tidor, B.; Karplus, M. The contribution of vibrational entropy to molecular association. The dimerization of insulin. *J. Mol. Biol*., **1994**, 238, 405-414.

[91] Yu, Y.B.; Privalov, P.L.; Hodges, R.S. Contribution of translational and rotational motions to molecular association in aqueous solution. *Biophys. J*., **2001**, *81*, 1632-1642.

[92] Schwarzl, S.M.; Tschopp, T.B.; Smith, J.C.; Fischer, S. Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas correction? *J. Comput. Chem.*, **2002**, *23*, 1143-1149.

[93] Carlsson, J.; Åqvist, J. Absolute and relative entropies from computer simulation with applications to ligand binding. *J. Phys. Chem. B*, **2005**, *109*, 6448-6456.

[94] Irudayam, S.J.; Henchman, R.H. Entropic cost of protein-ligand binding and its dependence on the entropy in solution. *J. Phys. Chem. B*, **2009**, *113*, 5871–5884.

[95] Searle, M.S.; Williams, D.H. The cost of conformational order: Entropy changes in molecular associations. *J. Am. Chem. Soc*., **1992**, *114*, 10690-10697.

[96] Murray, C.W.; Verdonk, M.L. The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput. Aid. Mol. Des*., **2002**, *16*, 741-753.

[97] Gilson, M.K.; Given, J.A.; Bush, B.L.; McCammon, J.A. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J*.; **1997**, *72*, 1047-1069.

[98] Finkelstein, A.V.; Janin, J. The price of lost freedom: entropy of biomolecular complex formation. *Prot. Engin*., **1989**, *3*, 1-3.

[99] Andrews, P.R.; Craik, D.J.; Martin, J.L. Functional group contributions to drug-receptor interactions. *J. Med. Chem*., **1984**, *27*, 1648-1657.

[100] DeWitte, R.S.; Shakhnovich, E.I.; SMoG: de Novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.*, **1996,** *118*, 11733-11744.

[101] Hopkins, A.L.; Groom, C.R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today*, **2004**, *9*(10), 430-431.

[102] Abad-Zapatero, C. Ligand efficiency indices for effective drug discovery. *Expert Opin. Drug Discov.*; **2007**, *2*(4), 469-488.

[103] Hetényi, C.; Maran, U.; García-Sosa, A.T.; Karelson, M. Structure-based calculation of drug efficiency indices. *Bioinformatics*, **2007**, *23*(20), 2678-2685.

[104] Orita, M.; Ohno, K.; Niimi, T. Two 'golden ratio' indices in fragment-based drug discovery. *Drug Discov. Today*, **2009**, *14*(5/6), 321-328.

[105] Keserü, G.M.; Makara, G.M. The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov*., **2009**, *8*, 203-212.

[106] Nissink, J.W.M. Simple size-independent measure of ligand efficiency. *J. Chem. Inf. Model*., **2009,** *49*, 1617–1622.

[107] Bembenek, S.D.; Tounge, B.A.; Reynolds, C.H. Ligand efficiency and fragment-based drug discovery. *Drug Discov. Today*, **2009**, *14*(5/6), 278-283.

[108] Reitz, A.B.; Smith, G.R.; Tounge, B.A.; Reynolds, C.H. Hit triage using efficiency indices after screening of compound libraries in drug discovery. *Curr. Top. Med. Chem.*, **2009**, *9*, 1718-1724.

[109] Murray, C.W.; Rees, D.C. The rise of fragment-based drug discovery. *Nat. Chem*., **2009**, *1*, 187-192.

[110] Garcia-Sosa, A.T.; Oja, M; Hetényi, C.; Maran, U. Disease-specific differentiation between drugs and non-drugs using principal component analysis of their molecular descriptor space. *Molecular Informatics*, **2012**, *In the press*..

[111] Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R. N.; Gilson, M.K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*., **2007**, *35*, D198-D201.

[112] Wishart, D.S.; Knox, C.; Guo, A.C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico

drug discovery and exploration. *Nucleic Acids Res*., **2006**, *34*, D668-D672 Sp. Iss. SI.
[113] Neidle, S.; Thurston, D. E. Chemical approaches to the discovery and development of cancer therapies. *Nat. Rev. Cancer* **2005**, *5*, 285-296.
[114] von Nussbaum, F.; Brands, M.; Hinzen, B.; Weigand, S.; Habich, D. Antibacterial natural products in medicinal chemistry - Exodus or revival? *Angew. Chem.-Int. Edit.* **2006**, *45*, 5072-5129.
[115] Payne, D. J.; Gwynn, M. N.; Holmes, D. J.; Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* **2007**, *6*, 29-40.
[116] Barratt, M.D. Quantitative structure-activity relationships for skin permeability. *Toxicol. Vitro*, **1995**, *9*, 27-37.
[117] Cross, S.E.; Magnusson, B.M.; Winckle, G.; Anissimov, Y.; Roberts, M.S. Determination of the effect of lipophilicity on the *in vitro* permeability and tissue reservoir characteristics of topically applied solutes in human skin layers. *J. Invest. Dermatol*., **2003**, *120*, 759-764.
[118] Rees, D.C.; Congreve, M.; Murray, C.W.; Carr, R. Fragment-based lead discovery. *Nat. Rev. Drug Discov*., **2004**, *3*, 660-672.
[119] Koehn, F.E.; Carter, G.T. The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discov*., **2005**, *4*, 206-220.
[120] Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem*., **2004**, *47*, 2977-2980.
[121] Aizawa, M.; Onodera, K.; Zhang, J.-W.; Amari, S.; Iwasawa, Y.; Nakano, T.; Nakata. K. KiBank: A database for computer-aided drug design based on protein-chemical interaction analysis. *Yakugaku Zasshi*, **2004**, *124*, 613-619.
[122] Zhang, J.–W; Aizawa, M.; Amari, S.; Iwasawa, Y.; Nakano, T.; Nakata, K. Development of KiBank, a database supporting structure-based drug design. *Comput. Biol. Chem*., **2004**, *28*, 401-407.

Received: October 11, 2011 **Revised: January 26, 2012** **Accepted: January 27, 2012**

[123] Ababou, A.; Ladbury, J.E. Survey of the year 2005: literature on applications of isothermal titration calorimetry. *J. Mol. Recognit*., **2007**, *20*, 4-14.
[124] Hetényi, C.; Paragi, G.; Maran, U.; Timár, Z.; Karelson, M.; Penke, B. Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins. *J. Am. Chem. Soc*., **2006**, *128*, 1233-1239.
[125] Wang, R.; Gao, Y.; Lai, L. Calculating partition coefficient by atom-additive method. *Perspectives in Drug Discovery and Design*, **2000**, *19*, 47-66.
[126] *Marvin 4.8.1*, ChemAxon **2008**, www.chemaxon.com.
[127] *SYSTAT 12*, SYSTAT Software, Inc., 1735 Technology Dr., Ste. 430, San Jose, CA 95110.
[128] Shapiro, S.; Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591–611.
[129] Kolmogorov, A. *Foundations of the Theory of Probability*, 2nd ed.; New York: Chelsea, 1956.
[130] Anderson, T.W.; Darling, D.A. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, **1952**, *23*, 193–212.
[131] Otto, M. *Chemometrics*, Wiley-VCH: Weinheim, Germany, 1999.
[132] Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **1947**, *18*, 50–60.
[133] *Program package R*. http://www.r-project.org (accessed 24 February **2009**).
[134] Heckert, N.A.; Filliben, J.J. *NIST Handbook 148: DATAPLOT Reference Manual, Volume I: Commands*, National Institute of Standards and Technology Handbook Series, 2003.