molecular informatics

DOI: 10.1002/minf.201100094

Disease-Specific Differentiation Between Drugs and Non-Drugs Using Principal Component Analysis of Their Molecular Descriptor Space

Alfonso T. García-Sosa,[a] Mare Oja,[a] Csaba Hetényi,[b] and Uko Maran*[a]

Presented at the 18th European Symposium on Quantitative Structure Activity Relationships, EuroQSAR 2010, Rhodes, Greece

Abstract: The physicochemical descriptor space has been extensively mapped and described in the literature for orally administered drugs and lead compounds. However, consideration of negative examples (non-drugs) or disease pathophysiology is not common in many studies. In the present work, a principal component analysis was carried out using drugs and non-drugs taking into account disease- and organ-specific categories, as well as different administration routes in addition to oral. The study involves 1386 relevant small-molecules including natural and synthetic products. Drug-specific as well as disease-category-specific or organ-specific regions and their respective threshold sets (ranges of descriptors) relative to non-drugs were elucidated on the scores plot and validated with external, independent sets of drugs and non-drugs. The respective loadings plot of molecular descriptors was rationalized in terms of physicochemically relevant groups related to the components of solvation free energy. The results of this analysis can contribute to the improved profiling of drug candidates and libraries making use of disease- and organ-specificity coded by physicochemical descriptors and ligand binding efficiency.

Keywords: Drugs · Non-drugs · Principal component analysis

1 Introduction

The structural characteristics of small compounds are major determinants in the early steps of drug design. Various threshold or range sets (such as Lipinski's rule-of-five for orally bioavailable compounds[1] and others[2–19]) are based on molecular descriptors and have been designed and extensively used for this task. However, it is still an open and active area of research how known drugs' and non-drugs' areas of chemical space locate relative to each other. In particular, the relative location of drug areas in the chemical space belonging to distinct disease categories is missing. A better knowledge of these localizations can aid in the improved profiling of chemical libraries, the optimization of compounds, as well as better target compounds to reduce selectivity issues. The property or label of 'drug' for a compound is not inherent, but can change with time, i.e., a compound may receive the label and consequences of a 'drug' by the United States Food and Drug Agency (FDA) or another drug agency, or may even be retracted from market or clinical use,[20] and the label of 'non-drug' can also change. Also important are advances in the last 50 years in molecular biology, synthesis and analysis, high-throughput, and other technologies, that direct and allow medicinal chemists to work with more complex chemistries, and this, as consequence, leads to changes in drug- and non-drug-likeness characteristics.[21] Therefore, one needs to define 'drug-likeness' as a particular set of properties or fragments that are present in currently-defined drug compounds, as opposed to non-drugs, in order to improve compound libraries to have chemical characteristics similar to those of known drugs to perhaps better and expedite drug discovery and design. Thus, 'drug-likeness' characteristics are also subject to change over time due to new and retracted drugs. One should keep also in mind that drug-likeness characteristics are derived from databases of other compounds and are therefore, mostly a statistical description and should not be used in the context of characterizing single compounds, but rather chemical libraries.[10] Evidently, a compound that possesses good 'drug-likeness' will not mean that such compound will become a drug,

[a] *A. T. García-Sosa, M. Oja, U. Maran*
Institute of Chemistry, University of Tartu
Ravila 14A, Tartu 50411, Estonia
phone: (+372) 7375254; fax: (+372) 7375264
**e-mail: uko.maran@ut.ee*

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

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201100094.

just that it has chemical properties similar to those of drug compounds. Other issues are at stake here, such as pharmacodynamics, pharmacokinetics, side-effects, toxicity, therapeutical windows, market considerations, competitors, intellectual property, among others.

Previous work by several groups[19,20,22,23] has shown how to profile compound libraries in terms of their drug- or lead-likeness,[24] using non-drugs as comparative examples, though few studies regularly use non-drugs as negative controls. Certainly, the choice of descriptors and the data available can influence the outcome of any analysis. The mapping of the chemical space location of drug compounds, lead compounds, and non-active compounds has been described in cartography terms such as ChemGPS,[25] and a chemico-biological atlas.[26] Further, compounds designed for a particular target such as protein kinase inhibitors,[27] as well as for some other targets have been profiled by their physicochemical (descriptor) nature (for a review see the literature[28,29]). Also, the use of ligand efficiency values[30–33] as measures to describe binding can provide new profiling prospects. There is important information to be gleaned from considering a large dataset of drug compounds from different disease categories compared to a non-drug "negative" dataset with similar binding strength (ΔG) profile. For instance, multivariate statistical techniques can provide associations between the important compound properties providing discriminative patterns between groups.

In the present work, the purpose was not to pinpoint the disease categories of particular compounds, since they are known to overlap, and more information is accruing on drug repurposing, as well as systems pharmacology (the interaction of drugs or chemical compounds with biological networks at the organism, tissue, cellular, and molecular level, as described, e.g., in Taboreau et al.[34]). Rather, the focus is on describing the chemical (descriptor) space of different disease categories at the highest anatomical level, in order to prove a concept that can be further refined and adapted. Knowledge of these chemical spaces is valuable, however intertwined and co-dependent they may be. In order to achieve these aims, the following goals and tasks were carried out. (a) A previously compiled set of drugs and non-drugs was characterized with molecular descriptors. (b) Multidimensional descriptor space was analysed, distilling the principal components of their variation; (c) Principal components analysis was carried out, as well as the investigation of how structural representation determines the relative position of drug and non-drug subsets. (d) The ranges of molecular descriptors (upper and lower thresholds) for drug disease categories were determined, establishing disease or organ likeness ranges for chemical library design. (e) And finally, the PCA model was validated with several independently compiled validation sets.

2 Methodology

2.1 Experimental Data

The set of data for the analysis (training set) consisted of 631 small molecules (see Table S1) of which 311 are approved drugs and 320 are non-drugs, and was collected for our previous publication.[35] Briefly, structures and experimental inhibition or dissociation constants (K_i or K_d) for drugs were gathered from the KiBank[36,37] and PDBBind version 2005[38] databases, the list of small-molecule approved drugs being obtained from the DrugBank.[39] Non-drug structures and experimental K_i or K_d values were obtained from the PDBBind and SCORPIO[40] databases, as well as two articles,[41,42] and consist of compounds that are not classified as drugs in the DrugBank. Experimental K_i or K_d values were converted to experimental binding affinities (ΔG_{exp}) at 298.15 K. The selection criteria for the small molecules in the training set were: (i) availability of ΔG_{exp} information and (ii) the subsets had to possess a similar range of ΔG_{exp} values. Notably, several studies have shown[43–47] that molecular descriptors such as molecular mass (MW) hold ligand non-specific information on its ΔG. All selected non-drugs are bioactive molecules, but not therapeutics, and only 47% of them pass Lipinski's drug-likeness rules (see Table S1). This provides a challenge to distinguish drugs from active, non-therapeutic non-drugs.

Compounds in the set of drugs had an extra dimension as they were assigned to different disease categories (DC) according to organ or system on which they act, as classified in the DrugBank (first and highest, anatomical, level of ATC classification). In total, 14 different DCs are presented, namely, alimentary tract and metabolism, blood and blood forming, cardiovascular system, dermatological, genito-urinary system, systemic hormonal, anti-infective, antineoplastic and immunomodulating agents, musculo-skeletal system, nervous system, antiparasitic, respiratory system, sensory organs and various drugs (see details in Table S1). The last DC contains drugs, which cannot be included into any other categories. It is important to note that drugs indeed act on several targets and disease categories which may have overlap and, in the current set some drugs belong to more than one category since they may have an effect in several diseases. Increasingly, this is the result of drug repurposing when new uses and indications are proven for already existing drugs.[20,48] In addition, our drug set contains several administration routes in addition to oral.

A separate validation set was also constructed using data different from the training dataset, and as such, can be considered as an independent collection of compounds. The experimental binding information of the validation set was obtained from the PDSP database,[49] as well as recent FDA-approved drugs, and non-drugs from the PDBBind database version 2009.[50] Both sets were checked to exclude previous compounds already included in the training set,

and for being non-redundant. Non-drugs of the validation set were verified to have no drug action reported in the DrugBank. The final number of compounds in the validation set was 118 and 395, for drugs and non-drugs respectively (see Table S2).

The third set of compounds was obtained from the Sigma-Aldrich catalog.[51] It consisted of compounds that did not have 'exotic' elements, i.e., consisted only of the atoms C, N, O, H, P, S, Si, Cl, and F. No pesticides were included and the resulting 350 compounds were not available from any other vendor, i.e., were exclusive to this vendor. Form these compounds, only one tautomer or stereoisomer was selected when multiple were found, and the rest were deleted. In addition, for all compounds their possible activity was checked in the ChEMBL database,[52] and those active (6 compounds) were removed from this set, leaving 242 compounds in order to have a clean, no-activity non-drug dataset for testing (see Table S3). This altogether leads to the 1,386 compounds analysed in the manuscript.

2.2 Conformation Space Analysis and Quantum Chemical Calculations

In order to achieve a consistent representation of the small molecule structures, a search of conformational space for each structure was carried out. This was accomplished with MacroModel[53] as part of the Schrödinger Suite of software packages. The Merck Molecular Force Field (MMFFs) parameterization[54,55] and Monte-Carlo Multiple Minimum (MCMM) search method[56,57] or the Mixed torsional/Low mode sampling method[58] were applied for the conformational space analysis. Depending on the molecule, the maximal number of steps (conformations) scanned was 15,000 and the water environment for the molecules was accounted for using a GB/SA model.[59] For each molecule, conformational search variables were set automatically and the conformer with the lowest energy was selected for the further steps.

The geometry of the lowest energy conformer of each small molecule was further optimized using semi-empirical quantum chemical methods. Namely, the AM1 parameterization[60] was used to characterize molecular structures, and the eigenvector following algorithm[61] was used for geometry optimization. Both methods were used as implemented in the MOPAC 6.0 program.[62] For four molecules the geometry optimization could not converge because of symmetry constraints that the optimizer was not able to solve. These molecules are vasopressin (d303) and oxytocin (d587) for drugs and 1gux (n101) and 1mpa (n170) for non-drugs. These molecules were therefore excluded from further analysis. After the conformational search and geometry optimization, the final dataset consisted of 628 small molecules (310 drugs and 318 non-drugs) in the training set, and 512 small molecules in the validation set (117 drugs and 395 non-drugs), plus 242 compounds in the no-activity non-drugs test set.

2.3 Molecular Descriptors

Molecules were subsequently characterized with molecular descriptors. MOPAC calculations provided the electronic, steric, and energetic parameters needed to calculate the molecular descriptors. Codessa software[63,64] was used to calculate constitutional, topological, geometrical, charge-distribution related, and quantum chemical molecular descriptors. Initially, 627 descriptors were calculated. The ΔG_{exp} was also included in the set of descriptors along with the calculated logarithm of the octanol/water partition coefficient using an atom contribution method (*X*Log*P*),[65] as well as three different efficiency indices for molecular weight (*MW*), Wiener index (*W*) and number of heavy atoms (NHA), in brief: $EIm = |\Delta G_{exp}/MW|$; $EIw = |\Delta G_{exp}/W|$; and $EIh = |\Delta G_{exp}/NHA|$. From the set of 632 descriptors, those with missing values were excluded and the remaining 307 descriptors were subjected to principal component analysis.

2.4 Principal Component Analysis

Principal component analysis (PCA) is a widely used multivariate data exploratory technique for pattern recognition. In PCA, the data matrix (**D**) is expanded as a sum of the principal components defined by scores and loadings, $\mathbf{D} = \mathbf{T} \cdot \mathbf{P} = \sum_{n=1}^{k} \mathbf{t}_n \mathbf{p}_n = \sum_{n=1}^{k} t_{i,n} p_{n,j}$. In the equation, **T** and **P** are the score and loading matrices, respectively; $\mathbf{t}_n$ and $\mathbf{p}_n$ are the score and loading vectors for a given component, which are expanded to their elements $t_{i,n}$ and $p_{n,j}$, respectively. The index *i* corresponds to observations (chemicals) and the index *j* corresponds to variables (descriptors), and *n* is the number of principal components (PC). The number of PCs (scores, loadings) existing in characteristic vector space can be equal to, or less than, the number of variables in the data set. The principal components are uncorrelated, i.e., orthogonal to each other. The first principal component is defined as that giving the largest contribution to the respective PCA of linear relationship exhibited in the data. The second component may be considered as the second best linear combination of variables that accounts for the maximum possible of the residual variance after the effect of the first component is removed from the data. Subsequent components are defined similarly until practically all the variance in the data is exhausted. The graphical plots of the score and loading vectors also reveal relationships between the objects and variables. In our case, the score plots summarized patterns among the drugs and non-drugs (observations) and the loading plots summarized patterns for the molecular descriptors (variables). The loading plot also enables interpretation of the patterns seen in the score plot. Hence, the patterns of these two presenta-

tions aid in the analysis of information encoded by the chemical structure.

A vital issue for the PCA model is the identification of strong and moderate outliers which could skew the model. Strong outliers can be traced in plots of PC scores while moderate outliers can be found by inspecting the model residuals. Generally, the strong outliers tend to significantly shift (rotate) the PCA model towards them. An appropriate statistical method for identifying such outliers is Hotelling's T^2,[66] a generalization of the Student's *t*-statistic. T^2 is graphically presented as an ellipsoid of T^2 range on score plots and indicates deviations far from the defined confidence intervals (95% or 99%). Strong outliers can also be spotted by the distance to the model X (DModX). Observations with a DModX twice over a critical value (D-Crit) are strong outliers to the PCA model.[67]

The set of 307 descriptors was analysed with PCA as implemented in the SimcaP+ software.[68] The descriptor scales were pre-processed to provide all scales with equal weight, with the standard unit variance scaling method, where the data is standardized, centralized and normalized using the sample standard deviation, variance and mean. The absolute value of the variables was used except in the case of descriptors that spanned both positive and negative values in order not to neglect information.

3 Results and Discussion

3.1 PCA Model and Outlier Analysis

The first PCA Model M1 on 307 descriptors resulted in 19 principal components – PCs (Table S4). Analysis of the prediction quality of the model (Q^2) revealed that the first three PCs captured most of the structural variation in the dataset due to unstable prediction quality for more than the first three PCs (Figure S1). Next, outlier analysis was performed for the model including the first three PCs. Analysis by distance to the model (DmodX, Figure S2) revealed seven strong outliers: allopurinol (d5), ethanol (d99), lindane (d151), 1c3x (n49), 1iht (n129), 1l8s (n153) and 1z71 (n261). The Hotelling's T^2 range within a 99% of confidence interval (Figure S3) on score plots reveals nine additional unique strong outliers in the model: piperazine (d224), 1btn (n44), 1hgt (n110), 1joc (n143), 1w1d (n237), 1xd0 (n250), 2hrp (n279), 2msb (n282), and 3er5 (n291). In total, 16 strong outliers were excluded, representing 2.5% of the compounds in the dataset. The skeletal formula of strong outliers (Figure S4) reveals that those chemicals are mostly either very big or very small relative to the remaining set of data. Subsequent analysis of the loadings plot showed several overlapping descriptors. Hence, redundant descriptors were eliminated. Descriptors relating to sites-specific charges and to energy partitioning terms originating from the quantum chemical calculations, and local descriptors were also removed, as well as those with very little contribution to the overall model. A new set of variables consisted of 116, whole molecule descriptors. Further, a new PCA Model M2 was derived that described structural variance with 14 PCs (Table S4).

3.2 Pattern Analysis of Scores and Loadings

Figure 1a shows that the training drugs and non-drugs do occupy different areas on the score plot with a borderline overlapping area. It can be seen that most of the drug molecules (198) are located in the upper right quadrant (tA) on the plot of 1st and 2nd scores, while encompassing the smallest amount of non-drugs (33). The upper left quadrant (tB) and lower right quadrant (tC) also contain a considerable number of drugs (40 and 60 respectively), but a higher

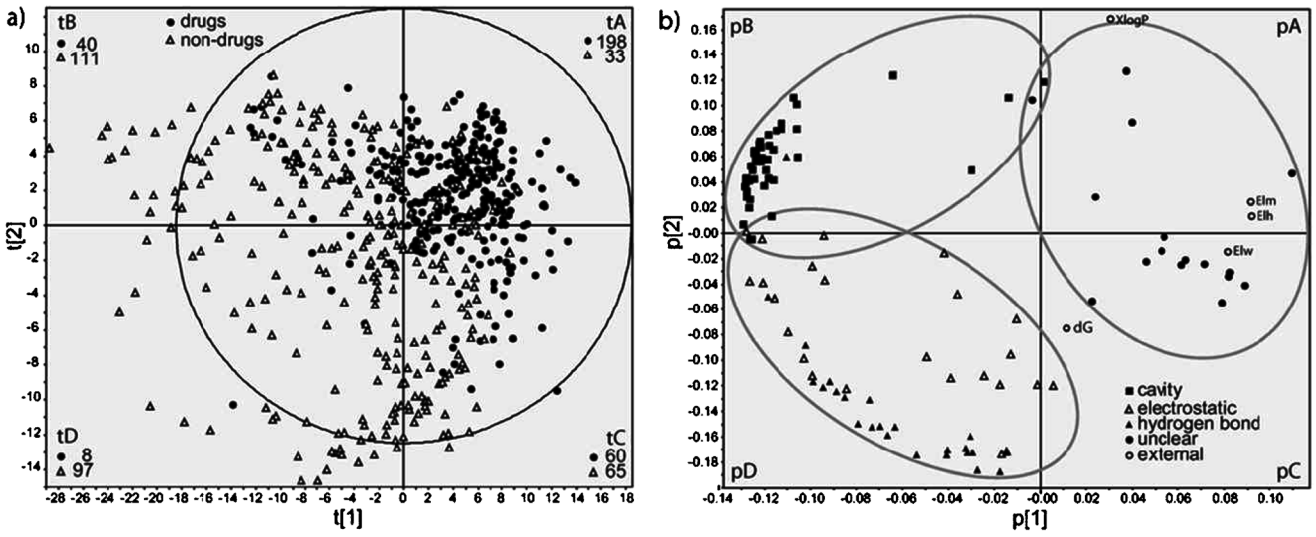

Figure 1. PCA Model M2 plots for 1st and 2nd principal component scores (a) and loadings (b).

proportion of non-drugs. The lower left quadrant (tD), however, contains only eight drugs among a vast majority of non-drugs. Drugs in quadrant D belong to antineoplastic agents (DC8; daunorubicin (d64), doxorubicin (d87), methotrexate (d169), raltitrexed (d246)), alimentary tract & metabolism (DC1; acarbose (d1), aprepitant (d19)), blood and blood forming (DC2; dipyridamole (d80)) and various drugs' group (DC14; ouabain (d205)). All of the drugs in quadrant D are located close to the centre of the model, with the exception of acarbose (d1). All of them, with the exception of aprepitant, dipyridamole, and raltitrexed, are either natural products or modifications of a natural product. The complete list of drug and non-drug compounds, as well as their disease category (in the case of drugs) and quadrant location (tA, tB, tC or tD) is shown in Table S1 in the Supporting Information. 90% of the drug compounds stayed in the following focused range according to the scores in each axis: t[1] −5.8 to 11; t[2] −5.09 to 6.08, t[3] −4.25 to 3.84 (e.g., see the plot of t[1] vs. t[2] in Figure 1a). This presents a direct way of comparing and using these ranges together with other rule based methods based on 90% of drug compound populations.[1,20]

Natural products and their modification have served as starting points for many therapeutics. 53 drug compounds were identified as natural products or semi-synthetic compounds (i.e., those drugs that are produced by using a natural-product intermediate) and located in the map (data not shown). Even though they can be found in all four quadrants, nearly half (22) of them are comprised in quadrant tA, which can be understood by natural systems having fine-tuned chemical compounds by evolution for bioactivity for a long time and number of generations.

The molecular descriptors group in a clear manner on the loadings plot (Figure 1b). The complete list of variables (molecular descriptors) together with their location in quadrants (pA, pB, pC, pD) is shown in Table S5 in the Supporting Information. To explain these groupings, we make use of solvation free energy and its components. The solvation free energy comprises at least four main components[69,70]: $\Delta G_S = \Delta G_{cavity} + \Delta G_{el} + \Delta G_{disp} + \Delta G_{HB}$, where ΔG_{cavity} is the cavity-formation term, ΔG_{el} is the free energy of electrostatic interactions, ΔG_{disp} depicts dispersion interactions, and ΔG_{HB} is the term arising from hydrogen bond formation. As discussed in our earlier work,[71,72] certain molecular descriptors closely reflect the terms of the free energy of solvation. For example, the cavity formation term can be satisfactorily modeled with the use of topological and geometrical descriptors, semi-empirically derived molecular polarizability, and entropy. Electrostatic and quantum chemical descriptors contribute significantly to both non-specific and specific solvation either through atomic charges, charged surface areas, dipole moments, reactivity indices, or other similar structural parameters. Descriptors designed for hydrogen bonding include molecular surface areas that are confined by H-bond donor or acceptor sites, as well as those that merely count such sites derived from atomic charge considerations. Such an approach facilitates the discussion of the main structural characteristics possibly influencing intermolecular interactions of drug and non-drug molecules. In this way we can group descriptors into 4 groups. The biggest single group is formed by the 46 descriptors that are related to the cavity formation term, or interactions that are related to the size of the molecule (see Table S5). They are grouped in the upper left quadrant (pB) of the loadings plot (Figure 1b). The group consists mostly of constitutional (atom, ring counts, etc.) descriptors, topological descriptors, and molecular surface area descriptors. The next two groups are located in the lower left quadrant (pD), forming distinct regions but overlapping with each other. The groups are formed by the descriptors that reflect (i) hydrogen bonding or (ii) electrostatic interaction characteristics of the molecules. Hydrogen bonding is characterized with 25 descriptors that reflect H-bonding surface area or are simple counts of H-bond donor and acceptor sites. Electrostatic interactions are characterized with 24 descriptors that are mostly related to the charge distribution and/or charged partial surface area of molecules. The fourth group occupies upper (pA) and lower (pC) right quadrants and consists exclusively from the descriptors that are relative measures of molecular sizes, i.e. descriptors that are divided by the maximum number of atoms, maximum size/distance, maximum charge, etc. These are descriptors for which their relationship with solvation free energy terms is currently not determined and that most presumably combine different types of interactions.

The model included also a fifth group of descriptors consisting of three efficiency indices (*Elh*, *Elm*, *Elw*), as well as *X*log*P* and ΔG_{exp}. ΔG_{exp} is adjacent to electrostatic descriptors and is located on border-line between quadrants pC and pD. At the same time *X*log*P* is located on the border of quadrant pA that corresponds to the score plot quadrant tA which contains most of the drugs, indicating that drugs are mostly log*P* optimized. The position of efficiency indices in the area of drug molecules indicates that they can be a valuable tool in distinguishing between drugs and non-drugs. They are also relative measures, and as such, may be indicating the simultaneous optimization of several molecular properties that characterize drug compounds over non-drugs.

Deeper analysis of the first three scores shows that PC1 is strongly influenced by the size of the small molecules (Figure 2a), running roughly parallel to the horizontal axis and decreasing from left to right, i.e., decreasing as it approaches to quadrants tA and tC. This is also visible in the loadings plot where size-related descriptors (such as MW and other size-related parameters) are located on the left-hand side (in pB) of this axis. PC2, on the other hand, is strongly influenced by the hydrophobic properties of the molecules (Figure 2b), running also roughly parallel to the vertical axis in the score plot and decreasing as it leaves quadrant tA. PC3 does not correspond to one clear structural effect. The best optimal single molecular characteristic

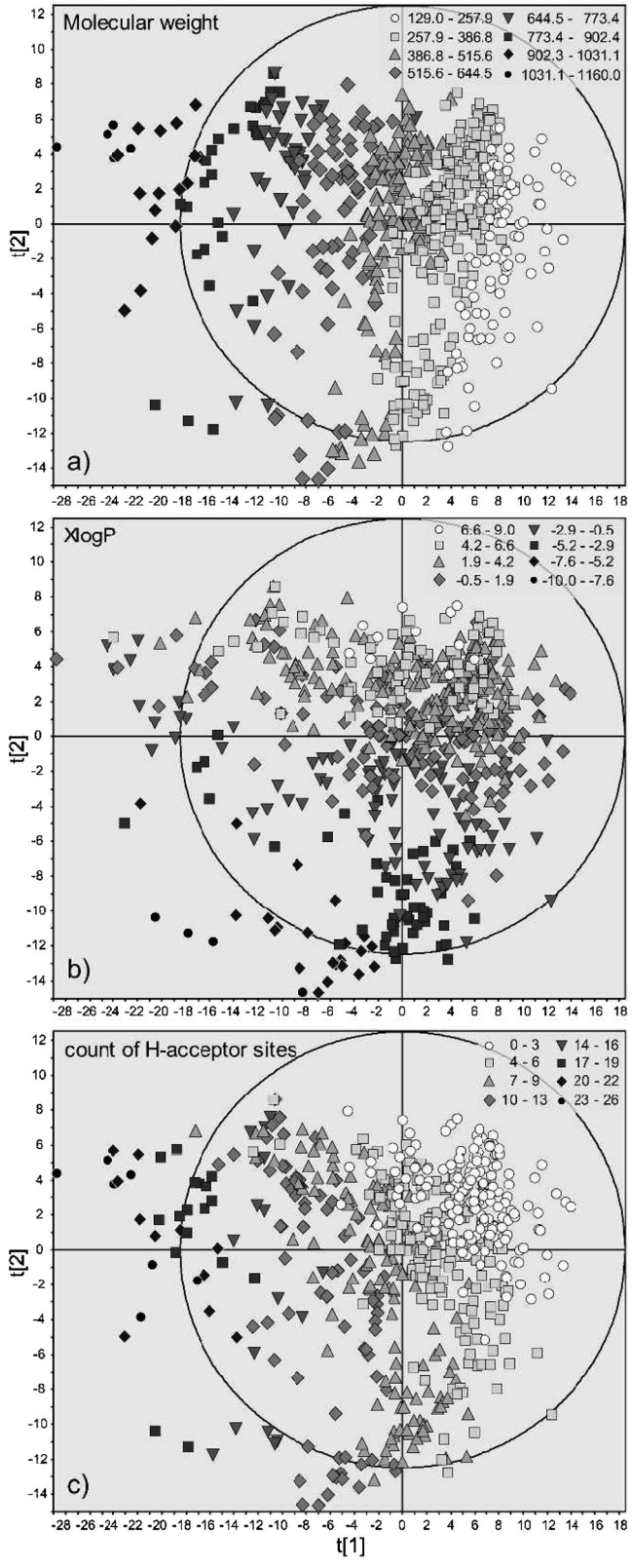

Figure 2. Plot for 1st and 2nd principal components grey scaled according to the gradient in three different descriptors: MW (a), logP (b), and number of hydrogen bond acceptors (c).

to explain PC3 is hydrogen bonding, namely, hydrogen bond acceptor capabilities (Figure 2c) running roughly diagonal between the axes, decreasing as it reaches into quadrant tA. This is also visible on the loadings plot, where hydrogen bonding descriptors are grouped in quadrant pD that corresponds to the area of the score plot (quadrant tD) of compounds with the highest number of hydrogen bonds. As a conclusion, this pattern indicates that most drugs are well-grouped in chemical space and have a relatively small size, a balanced hydrophobicity[73] (as measured by non-extreme values of *X*log*P*), as well as a relatively small amount of hydrogen bond acceptors.

3.3 Pattern Analysis of Specific Disease Categories

The coordinates of the first three components for all 14 DCs, i.e., the coordinates of score ranges, are shown in Table 1. These coordinates define the precise location for each disease category in descriptor space. It can be seen that the spread in coordinates is much wider for some DCs than to others. Based on those observations 14 DCs can be classified into three classes depending on whether: (i) all drugs in the category are clustered close to each other; (ii) similar to the previous class, but only a few drugs are extending from the rest of the group; or (iii) the drugs in the DC are spread considerably over the model. For each class one example is discussed in the main text. PCA plots and chemical structures for all of the drugs broken down per each disease category can be seen on Figures S6–S33 in the Supporting Information.

The well-grouped disease categories (class i) are genito-urinary systems (DC5), systemic hormonal drugs (DC6) and anti-parasitic products (DC11). The prominent example is DC5 (Figures S14–15), drugs for genito-urinary systems, comprised of molecules that are mostly steroids. Many compose a pharmacophore that is responsible for binding to the estrogen receptor or else to the androgen receptor, whereas the remaining are cGMP phosphodiesterase inhibitors, or antifungals. Almost all drugs in this group are placed in the drug-dominated quadrant tA. DC5 drugs that are placed in quadrant tB are more flexible than those in quadrant tA.

In the majority of cases, the disease categories are well grouped with a few compounds extending from the general trend (class ii). For instance, all alimentary tract and metabolism drugs (DC1) are located in almost the same area (Figures S6–7). Only one molecule, acarbose (d1), was set far away. This can be explained by the structure of acarbose which is a long molecule (oligosaccharide) with four rings and an abundance of hydroxyl groups. This molecule is used in the small intestines to inhibit the digestive enzyme alpha glucosidase to reduce levels of sugars, and therefore, does not need to cross any membrane and is distinct from other drugs in this disease group which fall under drugs for gastrointestinal or acid problems, for example, that have a different mechanism (e.g., proton pump inhibitors or his-

Table 1. Thresholds that apply to disease categories for the first three scores [a].

Disease categories	*n*	t[1]		t[2]		t[3]	
		min–max		min–max		min–max	
Class i							
DC5: Genito–urinary system	27	−1.92–9.29		0.81–7.14		−5.53–3.53	
DC6: Systemic hormonal	4	0.84–1.63		−1.58–−0.87		1.77–2.68	
DC11: Anti–parasitic products	4	1.89–7.26		−0.64–3.13		−3.82–−0.12	
Class ii							
DC1: Alimentary tract	32	−1.03–8.08	(−13.83–8.08)	−3.95–5.02	(−10.28–5.02)	−5.53–3.99	
DC2: Blood and blood forming	7	−2.27–8.49	(−4.33–12.07)	−2.51–2.68	(−3.95–2.68)	−2.32–3.85	(−4.00–6.76)
DC3: Cardiovascular system	59	−2.84–11.53	(−4.57–11.54)	−9.38–7.93		−3.26–6.30	(−5.21–6.30)
DC4: Dermatologicals	19	−1.92–8.63	(−12.18–11.79)	−3.86–5.51		−2.84–3.04	(−5.53–4.68)
DC9: Musculo–skeletal system	10	4.51–9.31	(−3.92–13.26)	−1.40–1.91	(−1.40–5.95)	−4.21–−0.09	(−6.11–2.07)
DC10: Nervous system	96	−0.86–13.88	(−4.65–13.88)	−3.64–6.53	(−5.76–6.53)	−6.44–6.90	
DC12: Respiratory system	26	−1.83–9.73	(−1.83–11.24)	−3.95–6.09	(−6.49–7.41)	−2.78–3.21	
DC13: Sensory organs	34	−2.15–9.73	(−10.69–11.96)	−3.95–3.66	(−7.96–5.12)	−5.53–6.31	
DC14: Various	21	1.74–10.00	(−7.45–11.33)	−2.43–3.96	(−2.43–6.56)	−4.26–3.81	(−4.26–7.37)
Class iii							
DC7: Antiinfectives	29	−10.69–8.51	(−12.22–11.79)	−6.99–6.70		−5.53–2.49	(−5.53–5.65)
DC8: Antineoplastic agents	22	−5.86–7.11	(−12.44–12.37)	−7.95–8.59	(−9.44–8.59)	−3.90–3.04	(−6.43–3.04)
All drugs in set	306	−13.83–13.89		−10.28–8.59		−6.44–7.37	

[a] Values in parenthesis are for the cluster with extending molecules; *n*: number of compounds.

tamine antagonists). Epinephrine (d93), another slightly deviating point, is on the contrary a small and very simple molecule.

Two disease groups (DC7 and 8) are well spread over the model plane (class iii). For instance, drugs that act as antineoplastic agents (DC8) are most widely spread in all four quadrants (Figure S20–21). This may be an indication that, structurally and chemically, they include different molecules for a variety of cancers targeting different organs. Also, since the targets of antineoplastic agents are human cells, it may show the lack of specificity as compared to other disease groups. Fluorouracil (d109) is a small, fluorinated molecule while paclitaxel (d208), tacrolimus (d270), vinblastine (d306), and vincristine (d307) are large, natural product molecules located in the same region. Daunorubicin (d64) and doxorubicin (d87) have the same substructure and pharmacophore, and thus they are in almost the same location.

Some drugs are present in several disease categories that illustrate the trend of drug repurposing, and can therefore deviate from central core groups in a given DC, due to their properties. For example, erythromycin (d95) is part of the dermatological agents category (DC4) since it is a drug applied to the skin, but it is also an antibiotic (DC7) since its function is to treat infections such as in acne. Tacrolimus (d270), another deviation, is used to treat infections on the skin (DC4), as well as to suppress the immune system in transplantation therapy (DC8). Molecules near the centre of the plot in Figure 1a may be classified as promiscuous or unspecific, a good example being the glucocorticoid steroids (d30, d68, d126, d230) that even if clustered tightly together, have a plethora of interactions which explain their multiple side-effects (DC: 1, 3, 4, 6, 12, 13) observed in their clinical use.[74] More targeted regions of physicochemical space, and further from the centre of the plots, may be more specific. Alclometasone (d4) and prednisolone (d230) are very similar molecules, alclometasone containing a chlorine atom. Alclometasone is included in only two categories (DC: 4, 13), but prednisolone in six (DC: 1, 3, 4, 6, 12, 13). This shows that a difference of only one atom can dramatically change the therapeutic properties of a molecule. Codeine (d62) and morphine (d180) are almost the same molecule, codeine has an extra methyl group. This methyl decreased the membership in two disease categories (codeine and morphine are part of two and four disease categories, respectively). Drug repurposing will have the effect of extending the number of DCs for such a drug with several indications and uses. In this sense, DC chemical space is also a moving definition. This is of benefit for new treatments and indications. Drug/non-drug comparisons will be evidently better refined than interDC comparisons, since for obvious reasons, targets are often present in more than one DC, given that, e.g., receptors can be present throughout the body, and have different subtypes or concentrations in different organs. However, the focus is on the chemical space of specific diseases or organs, not specific chemical compounds. Therefore, the focus of the present study is the chemical space available for compounds for a specific disease or organ compartmentalization, rather than in the multiple diseases a particular drug can be of use in. The advantages of considering DC chemical space, as opposed to drug-target or drug-many targets interactions, are that patterns can be elucidated irrespective of ligand structure or protein function or sequence. The present method described in this publication can be used to-

gether with polypharmacology or systems pharmacology[34] in a complimentary way.

3.4 Pattern Analysis of Non-Drugs

Quadrant tA, mostly drugs, also includes 33 non-drugs. Compounds in the quadrant share a number of common structural features. For instance, of the non-drugs, three molecules contain chlorine that is a common functional group for drugs: 40 drugs in this quadrant also contain it. A number of non-drug molecules in the quadrant have very flexible structures, whilst most of drugs in the quadrant are rigid. Five non-drugs in this quadrant contained sulfur, similarly to many drug molecules (as sulfonamides, for example). Only one non-drug molecule had a triple bond while almost all drugs with triple bonds are in quadrant tA. These features: chlorines, sulfonamides and triple bonds, being present in non-drugs in this quadrant show that they are features that help classify drugs since they are the overwhelming majority of compounds in this region. The complete list of non-drug molecules in quadrants is provided in Table S1. Here we describe a few examples for each quadrant. A first non-drug example molecule for the quadrant tA (Figure 3 and S5) is the compound taken from 1kdk (n149), which is the naturally produced androgen dihydrotestosterone that has bioactivity and is involved in several pathologies. Its structure is rigid with several fused rings, but not very polar. A second example is the compound from 1if8 (n128), which has a phenylsulfonamide substructure and balanced physicochemical and structural properties, which would indicate a potential drug-like characteristic. A schematic diagram with the location of example non-drug compounds located in their positions on the four quadrants is shown in Figure S5 in the Supporting Information.

Quadrant tB includes the biggest number of non-drugs (Figure 1a). They have a tendency to be large and non-polar, similarly to saccharides and peptides. For example, the ligand from 1hvr (n205) contains six phenyl rings arranged around a central cyclic urea with few polar groups, which is or was an experimental compound for HIV-1 protease inhibition[75] (Figure 3 and S5). Another example is the poly-proline biological compound in 1awi (n13) (Figure 3 and S5), the partner molecule for human platelet profilin, which is a protein that is implicated in cytoskeletal regulation and morphogenesis.

Quadrant tC contains compounds with relatively small size, and those that are near the border region with quadrant tA display interesting, balanced properties similarly to drugs. Notable examples are the experimental compound for human carbonic anhydrase II inhibition, namely the small sulfonamide AL-4623[76] (1bnq (n40), Figure S5, similar in structure to the drugs brinzolamide and dorzolamide), as well as the small benzylsuccinate 1wht (n244) that is part

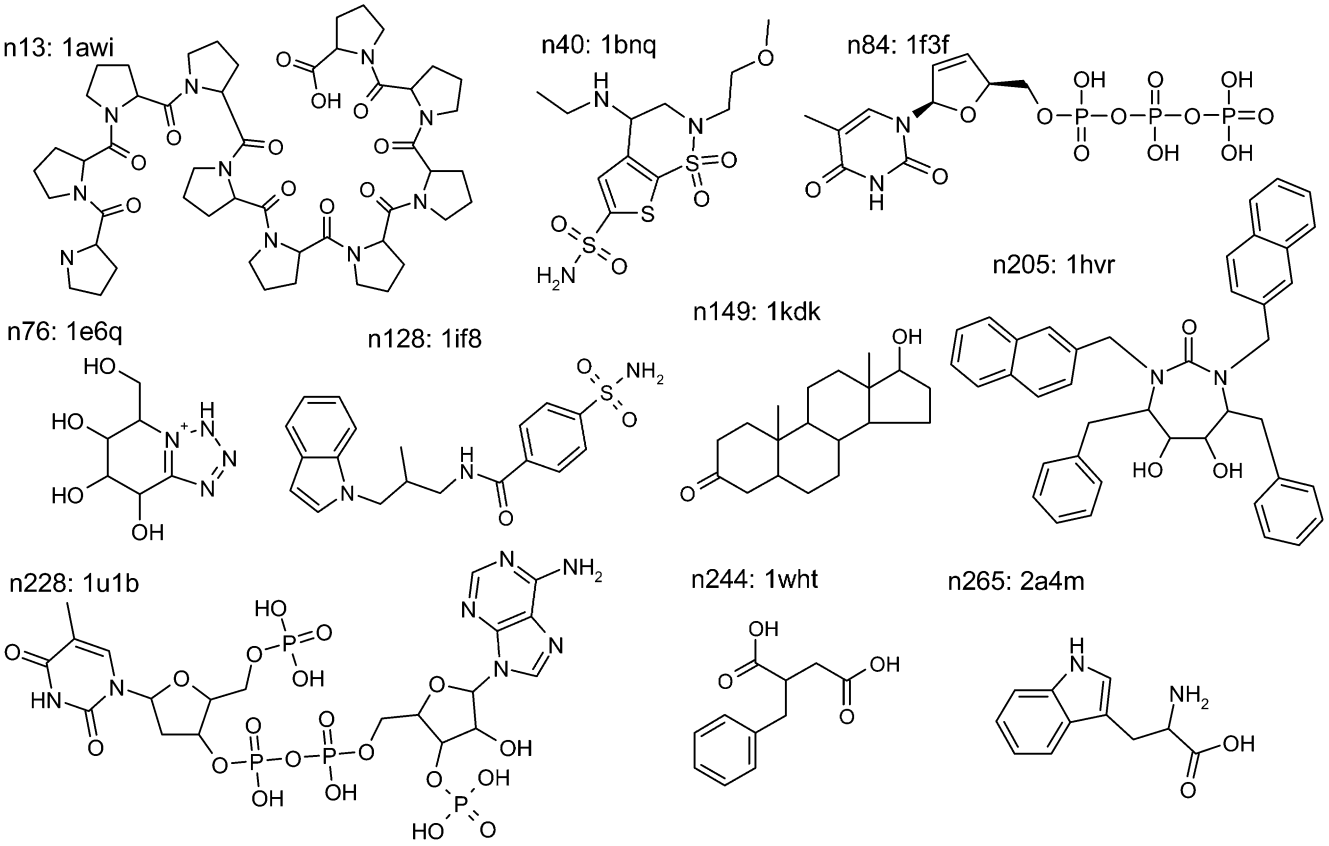

Figure 3. Chemical structures of selected representative non-drugs.

of the benzoate degradation pathway, in addition to 2a4m (n265), which is the natural aminoacid L-tryptophan. The most typical non-drugs in this quadrant, though, are small and very polar compounds, such as 1e6q (n76) (see Figure 3 and S5), as well as small saccharides.

Quadrant tD is almost exclusively composed of non-drugs. They are characterized by being large and polar compounds. Many non-drugs in this region contained one or more phosphate groups (for example, compounds 1u1b (n228) and 1f3f (n84) in Figure 3 and S5). Typical non-drugs also have many carboxyl groups. Some non-drugs are molecules with aldehyde, disulfide or sulfhydryl groups and many of them have a long chain with several cycles attached. Some non-drugs are carbohydrate polymers, similar to acarbose (d1), which are located far away from drugs. Often, non-drugs have more than one of these described structural features and as such, they make these molecules less drug-like.

Some of the non-drugs have activity against the same targets as the drugs, such as methionine aminopeptidase, carbonic anhydrase, or phosphodiesterase. This gives validity to the selection of compounds used as non-drugs, since they have comparable features and binding profiles that the drug compounds have, and provides a non-biased evaluation. They have also been studied extensively, given the availability of protein-ligand crystal structures and biochemical and thermodynamic characterization.

3.5 Disease-Likeness and Respective Ranges of Descriptors

The coordinates of the first three components for all of the disease categories, the coordinates of score ranges, describe the area that they occupy in the PCA model (see Table 1 and discussion in Section 4.3). In addition, two other sets of ranges (upper and lower thresholds) were established using molecular descriptors to define disease-likeness. The first set was defined for each disease category for values of *MW*, *X*log*P* and number of hydrogen acceptor sites, and they were also compared to known drug-likeness and lead-likeness parameters (see Table 2). The second set of ranges were defined for the three efficiency indices *EIh*, *EIm*, and *EIw* for each disease category, as well as for all the drugs in the dataset (see Table 3).

To establish broad reference points (goalposts) for efficiency indices, assuming a compound with sub-micromolar potency and a MW around 500, one could expect an *EIm* of ca. 0.02 kcal mol^{-1} Similarly, assuming a typical drug compound with less than 50 heavy atoms, one could expect a value for *EIh* of 0.2 kcal mol^{-1} NHA^{-1}. These values for a lead like compound would broadly change to around *MW* = 350 and 35 heavy atoms, and thus correspond to *EIm* ca. 0.029 kcal mol^{-1} and *EIh* ca. 0.29 kcal mol^{-1} NHA^{-1}. We have previously determined similar values of *EIw* for smaller sets of drug compounds to be around 0.001 to 0.070 kcal mol^{-1}.[33,35]

Some disease categories have a well-defined threshold set, which give an indication of their specificity. This is especially the case for those DCs in class (i) as well as in class (ii), such as genito-urinary system (DC5) compounds, respiratory system (DC12), as well as nervous system (DC10) compounds. This provides information on how to be more focused in the design of compounds that have a specific human organ in which they will be effective. For example, genito-urinary system drugs have relatively tight thresholds

Table 2. Thresholds that apply to disease categories and known drug-likeness and lead-likeness parameters for *MW*, *X*log*P*, and number of acceptor sites. Values in parenthesis are for the cluster with extending molecules.

Disease categories [a]	*MW* (g mol^{-1})		log*P* (as *X*log*P*)		#H-acceptors	
	min–max		min–max		min–max	
Class i						
DC5	206.3–531.4		0.1–7.3		1–7	
DC6	360.4–392.5		0.5–1.1		5	
DC11	248.7–369.4		2.04–6.3		3–5	
Class ii						
DC1	252.3–534.4	(183.2–645.6)	−0.5–5.5	(−6.8–5.5)	1–7	(1–14)
DC2	183.2–360.5	(131.2–508.6)	0.1–3.2	(−2.8–3.2)	2–8	(2–12)
DC3	153.2–551.6	(153.2–645.3)	0.22–6.2	(−2.7–7.9)	0–8	
DC4	172.2–531.4	(172.2–804.0)	0.22–5.2	(−2.0–5.29)	1–7	
DC9	169.6–381.4	(169.6–609.7)	1.3–3.8		1–5	
DC10	133.2–470.0	(133.2–583.7)	−0.6–5.4	(−2.1–5.4)	1–9	
DC12	165.2–501.7		0.5–5.6	(−2.0–7.1)	1–6	
DC13	147.2–500.6	(147.2–733.9)	−2.1–5.0		1–7	
DC14	166.2–408.5	(166.2–612.6)	−1.2–4.2		1–8	
Class iii						
DC7	137.1–812.0	(172.2–812.0)	−2.0–5.0	(−2.0–6.6)	2–11	
DC8	130.1–853.9	(263.2–853.9)	−2.9–7.1	(−2.9–8.9)	1–11	
All drugs in set	130.1–853.9		−2.9–8.9	(−6.8–8.93)	0–12	(0–14)
Drug-likeness (Lipinski Oral)	> 500		> 5		0–10	
Lead-likeness (Hann and Oprea)[77]	200–460		−4–4.2		0–9	

[a] Full names of the disease categories are in Table 1.

Table 3. Thresholds that apply to disease categories and known drug-likeness and lead-likeness parameters for Efficiency Indices. Values in parenthesis are for the cluster with extending molecules.

Disease categories [b]	*Elm*		*Elh*		*Elw*	
	min–max		min–max		min–max	
Class i						
DC5	0.01–0.06		0.14–0.75		0.002–0.018	
DC6	0.02–0.03		0.35–0.48		0.008–0.008	(0.006–0.008)
DC11	0.03–0.05		0.38–0.71		0.006–0.024	
Class ii						
DC1	0.01–0.05		0.14–0.67		0.001–0.018	(0.001–0.034)
DC2	0.02–0.05		0.29–0.75		0.003–0.034	(0.003–0.059)
DC3	0.01–0.06	(0.01–0.07)	0.15–0.75	(0.15–0.97)	0.001–0.051	(0.001–0.067)
DC4	0.01–0.05		0.22–0.75	(0.12–0.75)	0.001–0.028	(0.001–0.054)
DC9	0.01–0.04		0.20–0.68		0.004–0.028	(0.001–0.052)
DC10	0.01–0.07	(0.01–0.08)	0.15–0.94	(0.15–1.23)	0.001–0.089	
DC12	0.01–0.05		0.29–0.75	(0.17–0.75)	0.001–0.018	(0.001–0.044)
DC13	0.01–0.05	(0.01–0.08)	0.12–0.87	(0.12–1.23)	0.001–0.044	(0.001–0.089)
DC14	0.01–0.06		0.21–0.83		0.002–0.035	(0.002–0.052)
Class iii						
DC7	0.01–0.06		0.11–0.77		0.0004–0.025	(0.0004–0.054)
DC8	0.01–0.06		0.14–0.75	(0.09–0.90)	0.0004–0.025	(0.0004–0.075)
All drugs in set	0.01–0.08		0.09–1.23		0.0004–0.089	

[a] Full names of the disease categories are in Table 1.

of 206 to 531 in *M*W, 0.8 to 5.0 of *X*log*P*, 1 to 7 hydrogen bond acceptors, and *Elm* from 0.010 to 0.056 kcal mol^{-1}, *Elh* from 0.14 to 0.75 kcal mol^{-1} NHA^{-1}, and *Elw* from 0.002 to 0.018 kcal mol^{-1}. Thus, these efficiency indices limits for *Elm* and *Elh* are close to the above described rough estimate guidance values of "drug-like" and "lead-like" values of 0.02 kcal mol^{-1}, and 0.29 kcal mol^{-1} NHA^{-1}, respectively, and, together with their *Elw* values, are similar to those calculated previously for drug compounds.[33,35] This organ-based or disease-based drug design is a further development from considering only oral bioavailability characteristics to profile compounds.

Class (ii) DCs compounds can also be used to establish disease-likeness ranges for descriptors. The values in Table 2 and Table 3 show the limits of the regions that are well clustered in DCs of class (ii). Disease-likeness thresholds for DCs belonging to class (iii) are more broad and difficult to define since they occupy vast regions of chemical space.

Some of the thresholds fit within previously determined "drug-like" or "lead-like" thresholds, especially for those compounds administered orally. However, there are also specific regions which do not fall into the already described filter thresholds, and these may be used for designing drugs with a higher specificity. In particular, it may be the case that known drug-likeness and lead-likeness thresholds are too loose or unspecific relative to diseases and/or organs. The disease/organ/target biomolecule – specific filters determined by improved characterization of physicochemical regions of disease-specific and organ-specific ligands could produce more focussed compound libraries. In fact, it is also possible to combine different ranges of descriptors, so lead-like, genito-urinary (DC5) disease-like compounds would have *MW* between 206 to 460, *X*log*P* between 0 and 4.2, 1 to 7 hydrogen bond acceptors, as well as values of *Elm*, *Elh*, and *Elw* of 0.01 to 0.056 kcal mol^{-1}, 0.14 to 0.75 kcal mol^{-1} NHA^{-1}, and 0.002 to 0.018 kcal mol^{-1}, respectively. On the other hand, it is also observed that drugs that are in more than one DC are generally still within the overall drug region. This is a probable sign that 'once a drug, always a drug', in the sense that it is easier to find new applications for known drugs than for a similar, same binding affinity non-drug. We have previously observed that this is indeed possible for new applications of known drugs in the design of inhibitors for wild-type and drug-resistant H5N1 avian influenza,[78] as well in the design of new compounds for HIV-1 reverse transcriptase.[79]

From the efficiency indices in Table 3, it is observable that some disease groups have a wider range of values, such as drugs used in cancer therapy (DC8) and those used on sensory organs (DC13), as compared to the more tightly distributed systemic hormonal drugs (DC6), for example. We can also compare the high values of *Elh* for all disease category drugs with the rough threshold of 0.24 kcal mol^{-1} NHA^{-1} determined for protein-protein inhibitors.[80] In this respect, it suggests that drug compounds may be characterized by higher *Elh* values than both non-drugs, and protein-protein inhibitors. This may be due to the nature of protein-protein inhibitors, specifically, their requirement of displacing large protein binding partners.

Even if molecular size, hydrophobicity and number of hydrogen bond acceptors have been recognized to contribute to the drug-like character of compounds, we show here that drugs group distinctly according to their target organs or disease groups, providing locations for their positioning

in chemical space and subdivision according to mechanism and mode of action, as well as profiling their efficiency indices to provide specific, localized analysis and description of their structural properties.

3.6 Validation of the Results

The ranges of ΔG values for the validation set of drugs (-3.55 to -14.59 kcal mol^{-1}) and non-drugs (-3.46 to -15.93 kcal mol^{-1}) were the same as for the training set. The validation set was subjected to the same treatment as the training set. One validation set drug, cyclosporine A (d525), appeared to be a strong outlier and was therefore removed from further analysis. The validation set was aligned with the described pattern above. The overlay of training and validation sets can be seen in Figure 4. A general picture emerges where the drug compounds are located in the same region within the same quadrant as the drug compounds in the training set. Comparison with the score ranges defined by 90% of the drug compounds in the training set (Section 3.2) shows that 88%, 93%, and 90% of the validation set drugs overlay within the ranges of t[1], t[2], and t[3], respectively. The non-drugs are also in broad similar regions as before, with the small change that they are closer to the border regions with the drug molecules (71% of them pass Lipinski's rule-of-five, see Table S2). This may indicate progression of drug design strategies over time in that more recent compounds have been designed with a focus on being more drug-like, or on improving pharmacokinetic properties.

Another proof of the utility of the model comes when analyzing the disease-category location of the new set of validation drugs. For most of the cases, they overlapped in the regions defined previously. Examples can be seen in Figures 5, where training and validation drug compounds are superposed for alimentary tract and metabolism (DC1), genito-urinary (DC5), and nervous system (DC10) categories. The overlap can be illustrated numerically by the percent of the compounds from the drugs validation set that fit to the range defined by the training set (Table 4). This indicates that for the majority of the DCs more than 80% of the compounds fit into the ranges. Exceptions are those DCs that have a small number of compounds in the training set, such as DC6 and DC11. Expectedly, the full set of validation drugs nearly completely overlaps with the training set, also indicated by the ranges of *MW*, *X*log*P* and number of hydrogen bond acceptors.

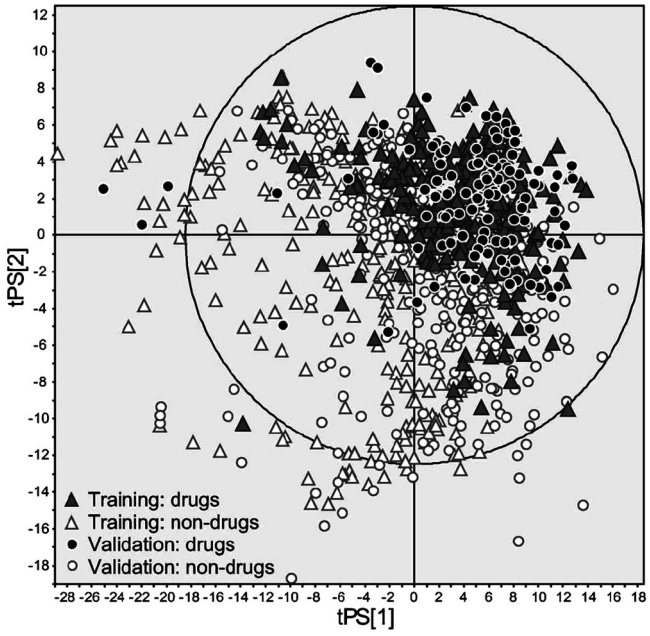

Figure 4. The overlay of both training and validation sets.

3.7 Tests with No-Activity Non-Drugs

In the previous section, the validation of the PCA model was carried out using a set of non-drugs of comparable binding affinity to the set of drugs. This was in order to provide a challenging background for the description of the chemical properties of the molecules. An additional independent test was carried out using a set of compounds with no known activity (none reported to date in the ChEMBL database). Arguably, these can be considered 'very' non-drug compounds, and similar approaches have been used before, using the Available Chemicals Directory (ACD)[20] as well as Sigma Aldrich chemicals.[81] Figure 6 provides this comparison and shows that the no-activity non-drugs form a distinct separate cluster, with some overlapping compounds in quadrants tA and tC, but mostly in a defined group of compounds in quadrant tA, located past the group of drug compounds (i.e., extending further to the edges of the graph than the drug compounds do). Only one of the no-activity non-drugs is present in quadrant tB, and none is present in quadrant tD. Comparison with the score ranges defined by 90% of the training set drugs (Section 3.2) shows that 65%, 17%, and 61% of no-activity non-drugs are located outside of these ranges for t[1], t[2], and t[3], respectively. At the same time, all but one of the no-activity non-drugs pass Lipinski's rule-of-five (see Table S3). This clearly shows that although no-activity non-drugs pass rule-of-five bioavailability-likeness criteria, they are outside of the drug compounds' chemical space. This also serves to illustrate the proposal made in the present work that combining different drug-likeness thresholds or ranges (using more than one in sequential filters, for example) are beneficial to more precisely characterizing chemical libraries for drug-likeness.

An inspection of the chemical nature of these compounds shows that they share some characteristics with drug compounds, such as relatively small size, low number of hydrogen bond acceptors, and some hydrophobicity, but they are more extreme in these parameters than drug compounds are. Hence, a balanced chemical nature (even more refined than that of the rule-of-five) and a moderately well-

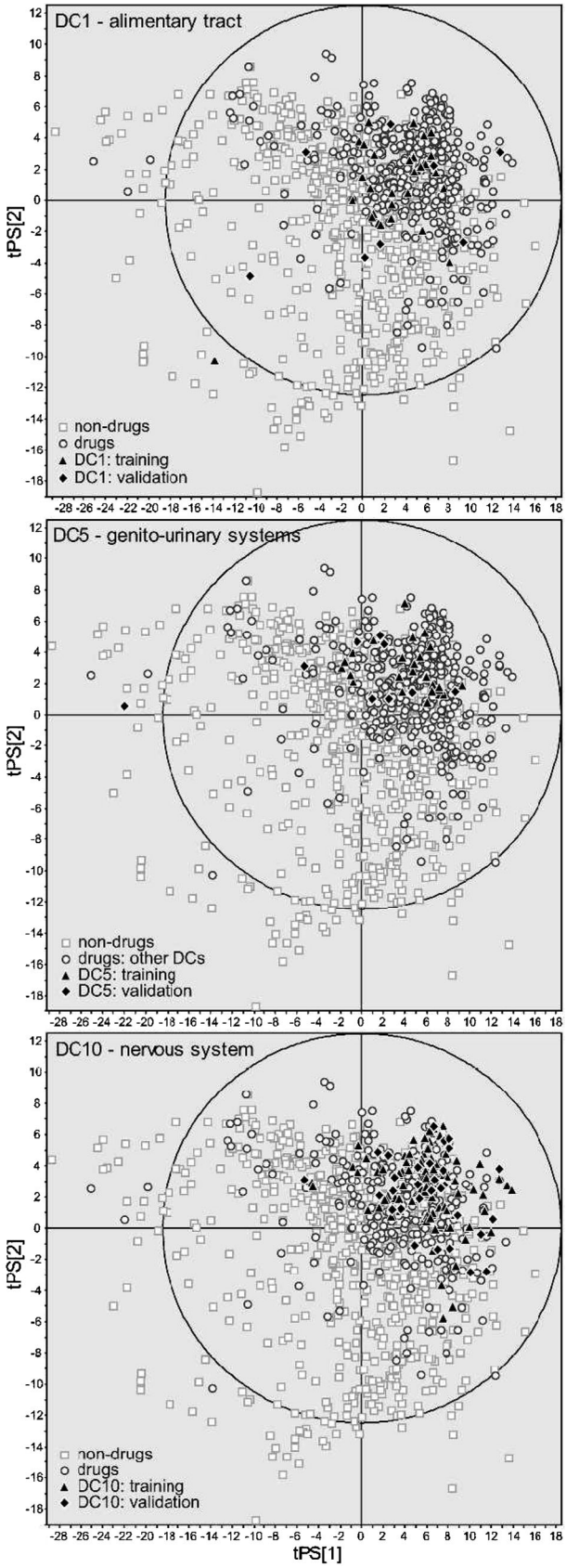

Figure 5. Superposed training and validation drug compounds for DC1, DC5 and DC10.

defined area of chemical space are able to describe some of the properties of drug compounds and of compounds that resemble them.

It is important to note that the present analysis of drug-likeness is not based on binding affinity. In fact, in the frame of the present publication, binding affinity was designed to be of similar magnitude between drugs and non-drugs in order to focus on other parameters, and indeed, it did not contribute to the distribution of the data in the PCA, i.e., it did not describe any difference between the variation of data between drugs and non-drugs. The current work focuses on analysis of molecular descriptors and all the interpretations are made in these terms. Pharmacokinetics and toxicity (themselves dependent on dose) are also accompanying issues, of course, when dealing with drug-likeness. These will determine the outcome of a compound's interactions with the many biomolecules, membranes, and organs of an organism, from phenomena such as absorption, metabolism, excretion, side-effects, and toxicity, which will depend on physical, chemical, and biological interactions of compounds and their metabolites with the many systems in which they will operate.[82]

Additionally, other methods exist for defining drug-likeness (for a review see the literature[20]), such as fragments having different frequency in drugs as in non-drugs.[20,83,84] One cannot overestimate the importance of good data (itself also a time-changing entity since new data are always forthcoming), as well as acknowledge the time-changing nature and definition of DCs, drug-many target and many drugs-one target interactions, and newly available synthetic and commercially available chemical space that, together with drug-like chemical space, is still a very small fraction of possible biologically relevant chemical space.

4 Conclusions

Using principle component analysis we verified if it is possible to differentiate between drugs and non-drugs based on their physicochemical descriptor space, as well as further determine what are the set of ranges of physicochemical descriptors for the specific disease groups and/or target organs. We found that this was indeed possible by separating disease-categories into different classes according to their spread. Based on this localization of physicochemical descriptor and ligand efficiency space, disease-likeness or organ-likeness threshold sets are defined in the vein of 'drug-likeness' and 'lead-likeness' threshold sets. The proposed disease-specific approach provides guidelines to improve profiling and filtering of compound libraries in order to guide a chemical library toward certain molecular characteristics of a certain set, such as those characteristics of drug compounds. This will not make them drugs, of course, since there are many mechanisms involved, not at least target relevance and importance, toxicity (which is a dose-

Table 4. Percent of drugs from the validation set within the range defined by the training set compounds for each criteria used in Tables 2 and 3. *n*: number of compounds; #H-accept.: number of hydrogen bond acceptors.

Disease categories [a]	n	*Elm*		*Elh*		*Elw*		*MW*		*X*log*P*		#H-accept.	
Class i													
DC5	9	100		100		89		78		89		78	
DC6	2	0		0		0		0		50		10	
DC11	3	33		33		100		100		100		100	
Class ii													
DC1	8	75		75		63	63	50	63	63	75	75	100
DC2	1	100		100		0	100	0	100	100	100	0	
DC3	11	91	100	91	100	91	100	91	91	82	100	100	
DC4	4	100		75	100	100	100	75	100	75	75	75	
DC9	12	92		92		92	100	92	100	83		83	
DC10	30	97	100	97	100	100		97	97	97	100	100	
DC12	16	88		69	94	94	94	75		81	88	94	
DC13	14	93	100	100	100	93	100	93	100	93		93	
DC14	20	75		80		60	75	70	80	75		95	
Class iii													
DC7	0												
DC8	4	100		100	100	100	100	75	75	100	100	75	
All drugs in set	116	100		100		99		97		99	99	96	97

[a] Full names of the disease categories are in Table 1.

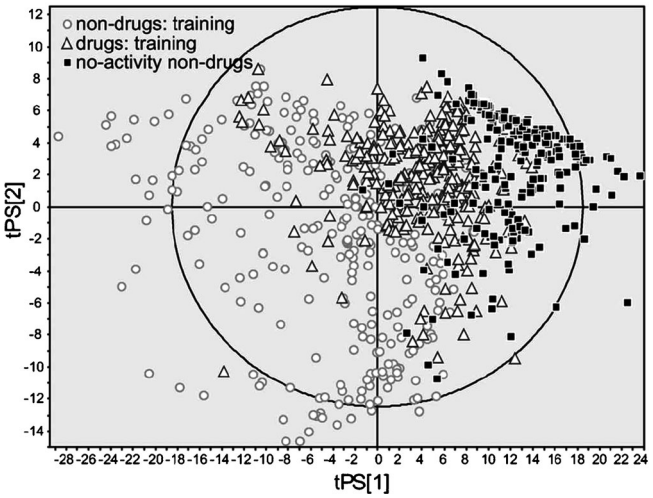

Figure 6. Location of no-activity non-drugs relative to the training set drugs and non-drugs.

dependant characteristic), side-effects, patient response, FDA approval, commercial considerations such as patents and expected revenues, among others, but it is an aid for refining compound libraries. It can also have value in predicting compounds that may exhibit more than one pharmacological activity, since if a compound has chemical properties that fit those for more than one specific disease category, there may be hints towards possible activity in several of them. These threshold sets and mappings include not only orally-available drugs, but also drugs that have a different route of administration. Such analysis is evidently more reasonable in those disease categories that have well-defined chemical space, which is challenging, considering the time-moving nature of 'drug-likeness', as well as pharmaceutical indications. However, the availability of chemical data and target or network interactions can increasingly provide guidance that is both broader than for only orally-administered compounds, as well as more specific towards individual targets and anti-targets of action.

Analysis of the loadings plot was able to produce physicochemically relevant groups of descriptors corresponding to the terms of free energy of solvation. The model was validated by an external, independent set of both drugs and non-drugs, which fell into the same established regions. This concordance was also valid for most disease-group categories, so that some specific DC prediction may be possible based on descriptor and ligand efficiency threshold sets and these ranges are presented.

Natural product drugs are seen to be both on the border next to non-drugs, as well as in the drug-defined region. This may be due to their relatively complex structural features as compared to drugs, the latter having been simplified to facilitate chemical synthesis. Natural products can also provide clues on how to design drugs, since they are naturally evolved to produce a biological effect.

Compounds in a particular DC may have similar structure, and/or pharmacophore, but many do not. Important information is provided by looking at diseases from both the ligand point-of-view, i.e., structural in addition to descriptor characteristics, as well as from their mechanism of action, target organ(s), and disease group(s). Some compounds that have the same activity may share structural characteristics (which may be uncovered by ligand-based design) or interaction partners (uncovered by, e.g., pharmacophore description), but many do not. The present paper shows a way of describing compounds with similar action without relying on their structure or binding functional group distribution. That is, it can help when profiling com-

pounds even when they do not share structural features. Therefore, not all genito-urinary drugs need be steroids, nor will all steroids be necessarily genito-urinary drugs. The results from the analyzed model may aid in the design of compounds and libraries better targeted to specific organs or diseases, and eventually targets, based on information obtained from physicochemical descriptor and ligand efficiency index space.

Supporting Information

Tables include compounds for training set, validation set, no-activity non-drugs, lists of descriptors used in PCA analysis, and statistics for PCA models M1 and M2. Figures include the graphical representation of PCA analysis, and detailed results and chemical structures of all disease categories. The complete matrix with molecular descriptors for all compounds used in the PCA analysis is available upon request.

Acknowledgements

Estonian Science Foundation Grant 7709, *Estonian Ministry for Education and Research* Grant SF0140031Bs09. C. H.'s work was supported by the *EU - European Social Fund* (Grant agreement no. TAMOP-4.2.1/B-09/1/KMR-2010-0003) and a *János Bolyai Research Scholarship* awarded by *the Hungarian Academy of Sciences*. The authors are also appreciative to Prof. Dr. *T. Oprea* for discussions at the Euro-QSAR2010 conference in Rhodes (September 2010). Authors are thankful to the reviewer for the motivation to include an additional test with no-activity non-drugs. The authors declare no financial nor personal conflicts of interest that might bias this work.

References

[1] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Delivery Rev.* **1997**, *23*, 3 – 25.
[2] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887 – 2893.
[3] V. J. Gillet, P. Willett, J. Bradshaw, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165 – 179.
[4] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski, *J. Comb. Chem.* **1999**, *1*, 55 – 68.
[5] D. E. Clark, S. D. Pickett, *Drug Discov. Today*, **2000**, *5*, 49 – 58.
[6] J. Galvez, J. V. de Julián-Ortiz, R. García-Domenech, *J. Mol. Graph. Model.* **2001**, *20*, 84–94.
[7] M. L. Lee, G. Schneider, *J. Comb. Chem.* **2001**, *3*, 284 – 289.
[8] E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1882 – 1889.
[9] M. C. Wenlock, R. P. Austin, P. Barton, A. M. Davis, P. D. Leeson, *J. Med. Chem.* **2003**, *46*, 1250 – 1256.
[10] I. Muegge, *Med. Res. Rev.* **2003**, *23*, 302 – 321.
[11] C. M. Dobson, *Nature*, **2004**, *432*, 824 – 828.
[12] M. Vieth, M. G. Siegel, R. E. Higgs, I. A. Watson, D. H. Robertson, K. A. Savin, G. L. Durst, P. A. Hipskind, *J. Med. Chem.* **2004**, *47*, 224 – 232.
[13] M. Vieth, J. J. Sutherland, *J. Med. Chem.* **2006**, *49*, 3451 – 3453.
[14] P. D. Leeson, B. Springthorpe, *Nat. Rev. Drug Discov.* **2007**, *6*, 881 – 890.
[15] J. J. Sutherland, R. E. Higgs, I. Watson, M. Vieth, *J. Med. Chem.* **2008**, *51*, 2689 – 2700.
[16] C. Tyrchan, N. Blomberg, O. Engkvist, T. Kogej, S. Muresan, *Bioorg, Med. Chem. Lett.* **2009**, *19*, 6943 – 6947.
[17] H. M. Chen, Y. D. Yang, O. Engkvist, *J. Chem. Inf. Model.* **2010**, *50*, 2141 – 2150.
[18] J. M. Wang, T. J. Hou, *J. Chem. Inf. Model.* **2010**, *50*, 55 – 67.
[19] Y. Hu, J. Bajorath, *ChemMedChem* **2010**, *5*, 187 – 190.
[20] O. Ursu, A. Rayan, A. Goldblum, T. I. Oprea, *WIRE Comp. Mol. Sci.* **2011**, *1*, 760 – 781.
[21] W. P. Walters, J. Green, J. R. Weiss, M. A. Murcko, *J. Med. Chem.* **2011**, *54*, 6405 – 6416.
[22] T. I. Oprea, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251 – 264.
[23] M. Brüstle, B. Beck, T. Schindler, W. King, T. Mitchell, T. Clark, *J. Med. Chem.* **2002**, *45*, 3345 – 3355.
[24] T. I. Oprea, A. M. Davis, S. J. Teague, P. D. Leeson, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308 – 1315.
[25] T. I. Oprea, J. Gottfries, *J. Comb. Chem.* **2001**, *3*, 137 – 166.
[26] C. Abad-Zapatero, O. Perišić, J. Wass, A. P. Bento, J. Overington, B. Al-Lazikani, M. E. Johnson, *Drug Discov. Today* **2010**, *15*, 804 – 811.
[27] A. L. Gill, M. Verdonk, R. G. Boyle, R. Taylor, *Curr. Top. Med. Chem.* **2007**, *7*, 1408 – 1422.
[28] D. G. Sprous, R. K. Palmer, J. T. Swanson, M. Lawless, *Curr. Top. Med. Chem.* **2010**, *10*, 619 – 637.
[29] H. Strombergsson, G. J. Kleywegt, *BMC Bioinformatics* **2009**, *10(Suppl 6)*, S13
[30] A. L. Hopkins, C. R. Groom, A. Alex, *Drug Discov. Today* **2004**, *9*, 430 – 431.
[31] I. D. Kuntz, K. Chen, K. A. Sharp, P. A. Kollman, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 9997 – 10002.
[32] C. Abad-Zapatero, J. T. Metz, *Drug Discov. Today* **2005**, *10*, 464 – 469.
[33] C. Hetényi, U. Maran, A. T. García-Sosa, M. Karelson, *Bioinformatics* **2007**, *23*, 2678 – 2685.
[34] O. Taboureau, S. K. Nielsen, K. Adouze, N. Weinhold, D. Edsgärd, F. S. Roque, I. Kouskoumvekaki, A. Bora, R. Curpan, T. S. Jensen, S. Brunak, T. I. Oprea, *Nuceic Acids Res.* **2011**, *39*, D367 – D372.
[35] A. T. García-Sosa, U. Maran, C. Hetényi, *Curr. Med. Chem.* **2012**, *19*, 1646 – 1662.
[36] M. Aizawa, K. Onodera, J.-W. Zhang, S. Amari, Y. Iwasawa, T. Nakano, K. Nakata, *Yakugaku Zasshi* **2004**, *124*, 613 – 619.
[37] J.-W. Zhang, M. Aizawa, S. Amari, Y. Iwasawa, T. Nakano, K. Nakata, *Comput. Biol. Chem.* **2004**, *28*, 401 – 407.
[38] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, S. Wang, *J. Med. Chem.* **2005**, *48*, 4111 – 4119.
[39] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, *Nucleic Acids Res.* **2006**, *34*, D668 – D672.
[40] A. Ababou, J. E. Ladbury, *J. Mol. Recognit.* **2007**, *20*, 4 – 14.
[41] C. Hetényi, G. Paragi, U. Maran, Z. Timár, M. Karelson, M. B. Penke, *J. Am. Chem. Soc.* **2006**, *128*, 1233 – 1239.
[42] W. I. Weis, K. Drickamer, W. A. Hendrickson, *Nature* **1992**, *360*, 127 – 134.
[43] P. D. Dobson, D. B. Kell, *Nat. Rev. Drug Discov.* **2008**, *7*, 205 – 220.

[44] P. D. Leeson, B. Springthorpe, *Nat. Rev. Drug Discov.* **2007**, *6*, 881 – 890.
[45] M.-Q. Zhang, B. Wilkinson, *Curr. Opin. Biotech.* **2007**, *18*, 478 – 488.
[46] L. K. Chico, L. J. Van Eldik, D. M. Watterson, *Nat. Rev. Drug Discov.* **2009**, *8*, 892 – 909.
[47] H. Kubinyi, *Nat. Rev. Drug Discov.* **2003**, *2*, 665 – 668.
[48] C. R. Chong, D. J. Sullivan, *Nature* **2007**, *448*, 645 – 6.
[49] *PDSP K_i Database*, http://pdsp.med.unc.edu/pdspImg.php (last accessed May 31, 2011)
[50] *PDBbind-CN database*, http://www.pdbbind-cn.org/index.asp (last accessed May 31, 2011)
[51] *Sigma Aldrich Catalog of Chemicals*, http://www.sigmaaldrich.-com (last accessed September 18, 2011).
[52] *European Bioinformatics Institute-European Molecular Biology Laboratory Small Molecule Database, ChEMBL*, https://www.e-bi.ac.uk/chembl/ (accessed September 21, 2011).
[53] *MacroModel*, Version 9, Schrödinger, Inc., Portland, **2000**. www.schrodinger.com
[54] T. A. Halgren, *J. Comput. Chem.* **1996**, *17*, 490 – 519, 520 – 552, 553 – 586, 587 – 615, 616 – 641.
[55] T. A. Halgren, *J. Comput. Chem.* **1999**, *20*, 730 – 748.
[56] G. Chang, W. C. Guida, W. C. Still, *J. Am. Chem. Soc.* **1989**, *111*, 4379 – 4386.
[57] M. Saunders, K. N. Houk, Y. D. Wu, W. C. Still, M. Lipton, G. Chang, W. C. Guida, *J. Am. Chem. Soc.* **1990**, *112*, 1419 – 1427.
[58] I. Kolossváry, W. C. Guida, *J. Comput. Chem.* **1999**, *20*, 1671 – 1684.
[59] W. C. Still, A. Tempczyk, R. C. Hawley, T. Hendrickson, *J. Am. Chem. Soc.* **1990**, *112*, 6127 – 6129.
[60] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *J. Am. Chem. Soc.* **1985**, *107*, 3902 – 3909.
[61] J. Baker, *J. Comput. Chem.* **1986**, *7*, 385 – 395.
[62] J. J. P. Stewart, *MOPAC Program Package 6.0*, QCPE, No. 445, **1989**.
[63] A. R. Katritzky, V. S. Lobanov, M. Karelson, *CODESSA: Reference Manual*, University of Florida, Gainesville, **1994**.
[64] A. R. Katritzky, V. S. Lobanov, M. Karelson, *Chem. Soc. Rev.* **1995**, *24*, 279 – 287.
[65] T. Cheng, Y. Zhao, X. Li, F. Lin, Y. Xu, X. Zhang, Y. Li, R. Wang, L. Lai, *J. Chem. Inf. Model.* **2007**, *47*, 2140 – 2148.
[66] R. A. Johnson, D. W. Wichern, *Inferences about a Mean Vector, in Applied Multivariate Statistical Analysis*, Prentice-Hall, Englewood Cliffs, NJ, **1982**, pp. 177 – 225.
[67] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, *PCA, in Multi- and Megavariate Data Analysis*, Umetrics AB, Umeå, Sweden, **2001**, pp. 43 – 69.
[68] *SIMCA-P* Version 12, Umetrics AB, Umeå, Sweden, **2009**,www.umetrics.com.
[69] C. Reichardt, *Solvent and Solvent Effects in Organic Chemistry*, Wiley-VCH, Weinheim, 2003, p. 629.
[70] M. Karelson, *Adv. Quantum Chem.* **1997**, *28*, 141 – 157.
[71] A. R. Katritzky, A. A. Oliferenko, P. V. Oliferenko, P. Petrukhin, D. B. Tatham, U. Maran, A. Lomaka, W. E. Acree Jr., *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794 – 1805.
[72] I. Tulp, D. A. Dobchev, A. R. Katritzky, W. Acree Jr., U. Maran, *J. Chem. Inf. Model.* **2010**, *50*, 1275 – 1283.
[73] P. D. Leeson, A. M. Davis, *J. Med. Chem.* **2004**, *47*, 6338 – 6348.
[74] H. Schäcke, W. D. Döcke, K. Asadullah, *Pharmacol. Therapeut.* **2002**, *96*, 23 – 43.
[75] P. Y. Lam, P. K. Jadhav, C. J. Eyermann, C. N. Hodge, Y. Ru, L. T. Bacheler, J. L. Meek, M. J. Otto, M. M. Rayner, Y. N. Wong, C. H. Chang, P. C. Weber, D. A. Jackson, T. R. Sharpe, S. Ericksonviitanen, *Science* **1994**, *263*, 380 – 384.
[76] P. A. Boriack-Sjodin, S. Zeitlin, H. H. Chen, L. Crenshaw, S. Gross, A. Dantanarayana, P. Delgado, J. A. May, T. Dean, D. W. Christianson, *Protein Sci.* **1998**, *7*, 2483 – 2489.
[77] M. M. Hann, T. I. Oprea, *Curr. Opin. Chem. Biol.* **2004**, *8*, 255 – 263.
[78] A. T. Garcia-Sosa, S. Sild, U. Maran, *J. Chem. Inf. Model.* **2008**, *48*, 2074 – 2080.
[79] A. T. Garcia-Sosa, S. Sild, K. Takkis, U. Maran, *J. Chem. Inf. Model.* **2011**, *51*, 2595 – 2611.
[80] J. A. Wells, C. L. McClendon, *Nature* **2007**, *450*, 1001 – 1009.
[81] V. V. Zernov, K. V. Balakin, A. A. Ivaschenko, N. P. Savchuk, I. V. Pletnev, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048 – 2056.
[82] R. Mannhold, H. Kubinyi, G. Folkers, R. J. Vaz, T. Klabunde, D. Schuster, C. Laggner, T. Langer, in *Antitargets. Prediction and Prevention of Drug Side-Effects* (Eds: R. J. Vaz, T. Klabunde), Wiley-VCH, Weinheim, **2008**, pp. xix–18.
[83] Ajay, *Curr. Top. Med. Chem.* **2002**, *2*, 1273 – 1286.
[84] J. Batista, J. Bajorath, *J. Chem. Inf. Model.* **2007**, *47*, 1405 – 1413.

Received: June 7, 2011
Accepted: January 25, 2012
Published online: May 3, 2012

