JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

DrugLogit: Logistic Discrimination between Drugs and Nondrugs Including Disease-Specificity by Assigning Probabilities Based on Molecular Properties

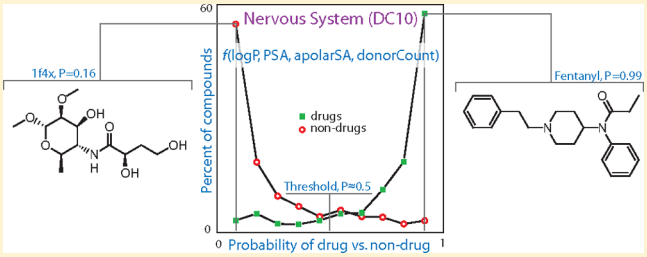
Alfonso T. García-Sosa,[†,*] Mare Oja,[†] Csaba Hetényi,[‡] and Uko Maran[†]

[†]Institute of Chemistry, University of Tartu, Ravila 14a, Tartu 50411, Estonia

[‡]Molecular Biophysics Research Group, Hungarian Academy of Sciences, Budapest, Hungary

Supporting Information

ABSTRACT: The increasing knowledge of both structure and activity of compounds provides a good basis for enhancing the pharmacological characterization of chemical libraries. In addition, pharmacology can be seen as incorporating both advances from molecular biology as well as chemical sciences, with innovative insight provided from studying target-ligand data from a ligand molecular point of view. Predictions and profiling of libraries of drug candidates have previously focused mainly on certain cases of oral bioavailability. Inclusion of other administration routes and disease-specificity would improve the precision of drug profiling. In this work, recent data are extended, and a probability-based approach is introduced for quantitative and gradual classification of compounds into categories of drugs/nondrugs, as well as for disease- or organ-specificity. Using experimental data of over 1067 compounds and multivariate logistic regressions, the classification shows good performance in training and independent test cases. The regressions have high statistical significance in terms of the robustness of coefficients and 95% confidence intervals provided by a 1000-fold bootstrapping resampling. Besides their good predictive power, the classification functions remain chemically interpretable, containing only one to five variables in total, and the physicochemical terms involved can be easily calculated. The present approach is useful for an improved description and filtering of compound libraries. It can also be applied sequentially or in combinations of filters, as well as adapted to particular use cases. The scores and equations may be able to suggest possible routes for compound or library modification. The data is made available for reuse by others, and the equations are freely accessible at http://hermes.chem.ut.ee/~alfx/druglogit.html .



INTRODUCTION

The adequate description of drug molecules' properties and their differences from those of nondrug ones are important in the design and discovery of new therapeutic compounds. In addition, pharmacology recently incorporates both advances from molecular biology as well as chemical sciences, with innovative insight provided from studying target-ligand data from a ligand molecular point of view.[1−4] In principle, the better their characterization in chemical and pharmacological terms, the easier it would be to distinguish drug-like properties of compounds, as well as to best ascertain their specific interactions with different targets and modes of action. "Drug-likeness" may thus be defined as some set of properties or fragments that are present in currently accepted drug compounds, but less so in nondrugs, so that compound libraries can have chemical characteristics similar to those of known drugs. These characteristics also change over time[5] given new and retracted drugs. Complying with "drug-likeness rules" will not mean that a compound is likely to become a drug, rather that it has chemical properties similar to those of drug compounds. Other issues will be critical, such as pharmacodynamics, pharmacokinetics, side-effects, toxicity (the latter being dose-dependent), therapeutical windows, market considerations, competitors, and intellectual property, among others.[6]

Drug compounds have been studied from the point of view of their oral bioavailability,[7] their drug-likeness,[8−12] as well as their lead-likeness,[13] number[14] and topology of rings,[15] molecular frameworks,[16,17] fragments,[6,18−22] using linear discriminant analysis[23] and decision trees,[24] among others.[25−29] Their properties have been classified in analogy to Global Positioning System (GPS) coordinates,[30] as well as in cartography terms[31] and chemical space.[32] Some early studies used Bayesian[8] or feed forward[9] neural networks to distinguish between drugs and nondrugs. Some studies have focused on target-family specific drug properties,[33] such as protease inhibitors and nuclear hormone receptors,[34,35] kinases,[36,37] and GPCRs,[38] using principal components analysis[39] and also using naive Bayesian methods, neural networks, and support vector machines in chemogenomics.[40−50] However, due to various limiting factors, there is no commonly accepted, routine method for a complete and simple comparison of drug properties relative to nondrugs

Received: December 9, 2011
Published: July 25, 2012

(i.e., negative controls) and especially for involving specificity on disease or location of target in the organism.[5] The latter may assist other drug design techniques that are based on ligand similarity and/or target structure principles. Drug repurposing[51−53] (i.e., the discovery or development of additional or alternative therapeutic indications for drug compounds) and polypharmacology or systems pharmacology[54] (i.e., the interactions between a compound and many targets, or many compounds against many targets at different organizational levels), as well as changing techniques[55] applied over time, will also create moving definitions of disease categories. One way of controlling for this is limiting the classification of compounds and using only those that belong to a specific disease category. Another factor to include is the different biomolecular targets of drugs, such as the recent realization that breast cancer consists of at least 10 different types of disease on the basis of their different gene expression.[56] Organ compartmentalization thus can provide the best answers when coupled to other structure- and ligand-based methods.

The present study is based on logistic regressions (LR) and on a large set of structural and experimental activity data. LR has been proven useful in medical and biological research because it can smoothly relate the probability of a deterministic outcome (disease present or not present, for example), from one or more dependent variables that can be continuous or discrete.[57,58] It is a special case of the linear model that is particularly useful for classifying binary outcomes (binomial regression) on the basis of several variables. It has been used in drug design to select water molecules appearing tightly bound in protein X-ray crystal structures,[59] as well as to improve neural network classifications of drugs and nondrugs[60] and to select between druggable and nondruggable binding sites in proteins.[61] The outcome probability is distributed evenly between probability values of 0 and 1 through the logarithm of odds ratio (see Appendix 1 in ref 59), and so it is well suited for comparing the properties of compounds that can have a varying degree of drug-like or pharmaceutically relevant character.

Classification of drug compound properties and in contrast to nondrugs has been previously achieved through statistical comparison of probability density functions,[5] as well as using principal components analysis.[62] These previous studies highlighted the usefulness of considering negative examples, as well as building data sets of experimentally available data for drugs (belonging to multiple administration routes and disease categories) and nondrugs (biologically active compounds that have similar binding affinity ranges). In the present work, multivariate LR is used to study the different chemical and pharmacological properties of drug and nondrug data sets, in order to classify compounds on the basis of simple, readily available, and physicochemically rational properties and produce their probability of classifying as a drug or a nondrug, as well as their probability of belonging to a highest (broadest) anatomical level, i.e., disease category specificity.

One of the advantages of the present method is that a quantitative measure is provided of the drug-like nature of a compound. This is beneficial because there are no strict cutoffs, and different properties can be involved. The use of a strict cutoff such as in Lipinski et al.,[7] can miss important compounds that lie slightly outside the defined ranges for the 90% of compounds.[5] However, it is becoming important to recognize that the molecular properties that are characteristic of drug compounds are a gradual spectrum,[63,64] and therefore, the method presented here is well suited for this task. New molecular properties not used before for characterizing drug-likeness are provided in this work. In addition, the most relevant properties for each drug category are able to be selected, and this may be of use when designing and filtering compounds of libraries for drug design.

METHODS

Data Sets. The training[5] and validation[62] sets of compounds were adapted from previous publications.

Training Set. A training collection of drug compounds and their inhibition or dissociation constants were extracted from several data sets, including the PDBBind version 2005,[65] the SCORPIO data set,[66] and KiBank.[67] The training set of nondrug compounds was also collected from these sources, and their nonexistence as drugs was verified in the DrugBank[68] database. Together, they composed a balanced set with $n = 311$ for drugs and $n = 320$ for nondrugs. Important features of the sets are that their distribution of binding energies and the number of compounds is similar for both drugs and nondrugs and that the drugs include all administration routes, not only oral. This can be seen in their superposed histograms in Figure S1 in the Supporting Information. This introduces a challenge to distinguish drugs from active, nontherapeutic compounds (nondrugs) because the differences between drugs and nondrugs are not judged by their binding energy. All structure files were checked for consistency and errors corrected.

Validation Set. A different set of drug and nondrug compounds (independent validation set) was compiled with newer collections of the PDSP database[69] and the PDBBind database version 2009[70] that were not available at the time of collecting the training set (i.e., compounds already present in the PDBBind database version 2005 were excluded from the validation set). As such, compounds of the validation set are completely independent of the training data set. Drug or nondrug status was verified using the DrugBank. The validation data set contained 106 drugs and 106 nondrugs. Predictions were run on the values of the (whole molecule) variables for each compound in the validation set using the regression equations obtained previously with the training set. The predicted outcome for each compound (the probability of classifying as drug or nondrug) was compared to their true status and percents of miss-prediction for the validation set were calculated. Mathew's correlation coefficients were also computed, and distributions of probabilities for each group were plotted. Receiver−operator characteristic curves and areas under the curves were also analyzed.

Further Tests. Another set of 224 compounds was also used for further tests to evaluate DC classification. Thus, the total number of compounds used in this study was 1067.

Physicochemical Properties. A small set of readily obtainable properties was calculated: log*P* (the logarithm of the octanol/water partition coefficient) was obtained using the atom-additive XLOGP method;[71] Marvin Beans version 5.3.8[72] was used for calculating number of heavy atoms (NHA), exact mass (MW), number of carbons (NoC), atom count, hydrogen count, bond count, ring count, aliphatic ring count, aromatic ring count, aromatic atom count, hydrogen bond donor count, hydrogen bond acceptor count, rotatable bond count, molecular surface area (MSA), polar surface area (PSA), apolar surface area (APSA), molecular polarizability (molpol), Wiener index (Wiener), Balaban index, Harary index, hyper-Wiener index, Platt index, Randic index, Szeged index, and Wiener polarity. The binding energy, ΔG_{bind}, of compounds to their binding partner proteins was calculated as previously reported,[5,62] using the

experimental equilibrium inhibition or dissociation constants. Ligand efficiency indices (LE) were calculated by using LE = ΔG_{bind}/NF, where the normalization factors (NF) were NHA,[73] MW,[74] NoC,[75] PSA,[76] MSA, APSA, Wiener index,[77] and P (obtained as $10^{\log P}$).[78,79]

Disease Categories. The drugs in the data set were characterized with an additional dimension according to correspondence in the 14 disease categories (DC) of the first and highest, anatomical, level of the Anatomical Therapeutic Chemical (ATC) classification, as shown in the DrugBank: DC1 = alimentary tract and metabolism, DC2 = blood and blood forming, DC3 = cardiovascular system, DC4 = dermatological, DC5 = genito-urinary system and sex hormones, DC6 = systemic hormonal drugs excl. sex hormones and insulins, DC7 = anti-infectives, DC8 = anti-neoplastic and immunomodulating agents, DC9 = musculo-skeletal system, DC10 = nervous system, DC11 = antiparasitics, insecticides and repellants, DC12 = respiratory system, DC13 = sensory organs, and DC14 = various drugs. There are instances of drugs belonging to more than one DC at a time, such as with glucocorticoids (for a detailed description, see ref 62).

Multivariate Regression. The module glm in the statistical computing package *R*,[80] was used for correlations, statistical tests, and regression analysis. Direct Pearson correlations were calculated between all of the variables described above. Direct univariate logistic regressions were calculated between each variable and the outcome variable of drug (coded 1) or nondrug (coded 0). Further, all the possible combinations were generated between the variables taking two, three, four, and five elements at a time, and multivariate logistic regressions were computed using these combinations and the outcome variable of drug or nondrug. The predicted response (probability (P) of outcome being drug, called P_{drug}) is calculated using the intercept (β) and coefficients ($\alpha_1,...,\alpha_n$) of the variables ($X_1,...,X_n$) in logistic regression, according to[57,59]

$$P_{drug} = \frac{e^{(\beta+\alpha_1X_1+...\alpha_nX_n)}}{1 + e^{(\beta+\alpha_1X_1+...\alpha_nX_n)}} \quad (1)$$

The probability of classifying as a nondrug is thus, $P_{nondrug} = 1 - P_{drug}$. A univariate regression corresponds to the particular case of $n = 1$ in eq 1. In addition to the drug/nondrug comparisons, all the Pearson noncorrelated combinations of variables were used to calculate multivariate logistic regressions with the outcome variable of belonging to each disease category (coded 1 for membership, 0 for nonmembership).

Statistical Tests. The regressions were accepted if each of their variables were statistically significant at the 95% confidence level or higher ($p < 0.05$), by rejecting the null hypothesis that the association between variables could be due to random variation. In addition, the overall model was also required to be statistically significant at the 95% confidence level or higher, against a null model ($p < 0.05$). The difference between the residual deviance for the model with predictors and the residual deviance for the null model provides the test statistic, which follows a chi-squared (χ^2) distribution with degrees of freedom equal to the number of predictor variables in the model.[81]

Predictions. A percent of miss-prediction was calculated by substituting the variable values into eq 1 and comparing the predicted (probability of being a drug compound, threshold $P = 0.5$) to the true outcome variable (drug (1) or nondrug (0)). In our previous work, different thresholds were selected to differentiate the classes of compounds through their probability density functions on the basis of selectivity and drug to nondrug ratio, allowing for tuning of these parameters according to different function forms.[5]

In logistic regression, there does not exist a correlation coefficient directly comparable to the Pearson's $R_{Pearson}$ of ordinary regression. Instead, the regression coefficients are calculated by maximum likelihood methods common to all generalized linear models, computed numerically by using iteratively reweighted least-squares, and the regression's quality is evaluated by testing the statistical significance of the χ^2 statistic as described above. In addition, further tests of accuracy are provided by the described percent of correct predictions, as well as by Mathew's correlation coefficients (MCC, eq 2, where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives)[82]

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP})}} \quad (2)$$

Recovery of Actives. Receiver−operator characteristics (ROC) curves were computed as true positive rate (i.e., fraction of known drugs) vs false positive rate using the obtained regressions, and their areas under the curve (AUC) were calculated using the Python module CROC[83] and xmgrace.[84] A similar measure, the drug to nondrug ratio, was calculated in our previous work using a different set of function forms.[5] Enrichment was determined through the ROC curves by calculating the percent of true positive rate at 5% of false positive rate.

Bootstrapping. The program *R*[80] was used to bootstrap the regression intercepts and coefficient values and perform 1000 fold substitution and resampling runs from which standard errors, bias, and 95% confidence intervals were established for measuring robustness in the functions.

Variable Terms. The names, drug status, training or validation set, and disease category of all compounds are shown in Table S1 in the Supporting Information. The molecular properties were calculated in a few minutes for the 1067 compounds. A Pearson correlation matrix was computed for the entire cross terms (i.e., pairs of descriptors). For all cross terms that were either strongly correlated ($R_{Pearson} > 0.6$) or strongly anticorrelated ($R_{Pearson} < -0.6$) with each other, the multivariate logistic regressions including them in their variables were discarded, thus avoiding collinearity. The discarded pairs of cross terms are shown in Table S2 in the Supporting Information. From the single direct, linear, Pearson correlation calculations, only one variable (hydrogen bond acceptor count) was strongly anticorrelated ($R_{Pearson} < -0.6$) with the outcome of being a drug or nondrug. For the logistic regressions, the total number of regression expressions using up to five variable terms at a time and one outcome variable was 52,521,875 (35^5). The full molecular properties calculated for all the compounds in the data sets are available in Table S3 in the Supporting Information.

RESULTS AND DISCUSSION

Multivariate Logistic Regression Analysis. Scripts were written to build these cleaned terms (i.e., terms of variables excluding the Pearson correlated pairs of descriptors mentioned above), feed them to regression analysis in the software package *R*, and then to check for statistical significance of each variable and of the regression equation as a whole, as well as their selectivity between DCs for the DC comparisons. Regression equations were accepted if both (i) all variables had statistical significance at the 95% confidence level or higher ($p < 0.05$), thus

Table 1. Logistic Regression Expressions Corresponding to Equations with Statistical Significance at the 95% Confidence Level or Higher ($p < 0.05$) To Predict the Probability (P) of a Compound Being a Drug or Nondrug[a]

	Intercept (Std. Err.) and Coefficients (Std. Err.)*Variables			
expression	training set prediction accuracy (%)	validation set prediction accuracy (%)	median of deviance residuals	sensitivity (%)
3	3(0.3) − 0.38(0.03)*Acceptor count			
	79	70	−0.03	91
4	2.7(0.2) − 0.026(0.002)*PSA			
	79	76	−0.03	91
5	1.7(0.3) + 0.51(0.07)*ring Count − 0.030(0.002)*PSA			
	82	74	−0.02	90
6	2.3(0.3) + 0.8(0.1)*aliphatic ring count + 0.50(0.05)*log *P* − 0.010(0.001)*MW			
	85	78	−0.01	88
7	2.7(0.3) + 0.8(0.1)*aliphatic ring count − 0.031(0.003)*PSA − 0.011(0.005)*bond count			
	82	77	−0.02	88
8	2.1(0.3) + 0.15(0.06)*log *P* − 1.19 × 10^{-4}(4.7 × 10^{-5})*Wiener index − 0.018(0.003)*PSA			
	82	74	−0.01	92
9	3.5(0.7) − 0.8(0.2)*Balaban index − 0.3(0.1)*aromatic ring count − 0.76(0.07)*donor count − 1.9(0.7)*ΔG_{bind}/NHA			
	84	75	−0.01	90
10	0.9(0.4) + 0.8(0.1)*aliphatic ring count + 0.15(0.05)*log *P* − 0.026(0.003)*PSA − 56.2(15.5)*ΔG_{bind}/Wiener			
	83	77	−0.02	91

[a] P threshold set at $P = 0.45$.

discarding that a correlation between the variables could have arisen by random variation and (ii) the chi-squared (χ^2) test statistic for the overall model was lower than 0.05 (i.e., 95% confidence level or higher, $p < 0.05$), thus discarding that a null model would perform better. Thanks to this, a selected number of regression equations were filtered from the large number of calculated regression equations. These regression equations were then compared between themselves using the number of true positives, true negatives, false positives, and false negatives, Mathew's coefficients, and by the percent of miss-prediction using the values of the variables of the training data set as well as those of a validation data set. The regressions that produced the highest number of true positives and true negatives, as well as lowest numbers of false positives and false negatives, were selected among the best. These would have the highest MCC values, indicating a good recall of positives and discrimination of negatives. In addition, the selection procedure for the best regressions included choosing those with a high accuracy of prediction both for training and validation data sets, as well as those regressions with small residual deviations and small standard errors as determined by the bootstrapping procedure. The distribution of predicted values was also analyzed in order to select the best regressions and discard those that did not produce a clear enough separation of the compounds. A final criterion for selection of the best regressions was the smallest possible number of variables to retain the ability to rationalize the equations in terms of their physicochemical meaning. A selection of the best regression expressions is presented in Table 1, including their percent of accurate predictions and miss-predictions for training and validation data sets, mean deviance of residuals, and standard errors.

One or several of the expressions in Table 1 can be used to classify compounds based on their predicted probability of being a drug compound or not. They can be selected on the basis of their prediction accuracy, the relevance or availability of their descriptor variables, or their physicochemical relevance to a data set at hand. They can also be selected on the basis of their mean of deviance residuals or on the standard errors in their intercept and coefficient(s). They can also be used to filter compounds based on one or several of the expressions 3−10, for example, in hierarchical (i.e., successively applied or daisy-chained) filters. It is important to note that due to the small number of variables in the equations, they remain immediately interpretable in chemical and physical terms. They are not hidden, nor do they suffer from overfitting or complex relationships between variables and outcome. As such, they should be easily applied in many settings.

The standard errors for the regression intercepts and coefficient variables are low. The relative standard errors (Relative Std. Err. = (Std. Err./Mean)*100) are mostly around 10% or lower, which indicate the reliability and precision of the statistical estimates. The 1000 fold bootstrapping (1000 resampling runs or replicates) provided values of standard error and bias estimate that show the reliability and small variance (robustness) in these parameters, as well as reasonably tight 95% confidence intervals (Table S4 in the Supporting Information).

Predicting Drugs or Nondrugs. From Table 1, one of the simplest regression equations is that for the univariate hydrogen bond acceptor count (eq 3), and it serves well to illustrate the regressions. The predicted outcome for this variable for the training data set was calculated by substituting eq 3 in eq 1, to produce eq 11, and plotted in Figure 1 together with the corresponding values for the validation set compounds using the regression equation thus developed (eq 11).

$$P_{drug} = \frac{e^{(3-0.38*\mathrm{AcceptorCount})}}{1 + e^{(3-0.38*\mathrm{AcceptorCount})}} \quad (11)$$

Figure 1 shows the smooth transition between the probability of drugs and nondrugs based on their values of hydrogen bond acceptor count. The good agreement between the two data sets indicates that the regression equation is valid and can be used on a wide variety of compounds (those for which the descriptor can be calculated) in order to quantify a degree of drug-likeness.

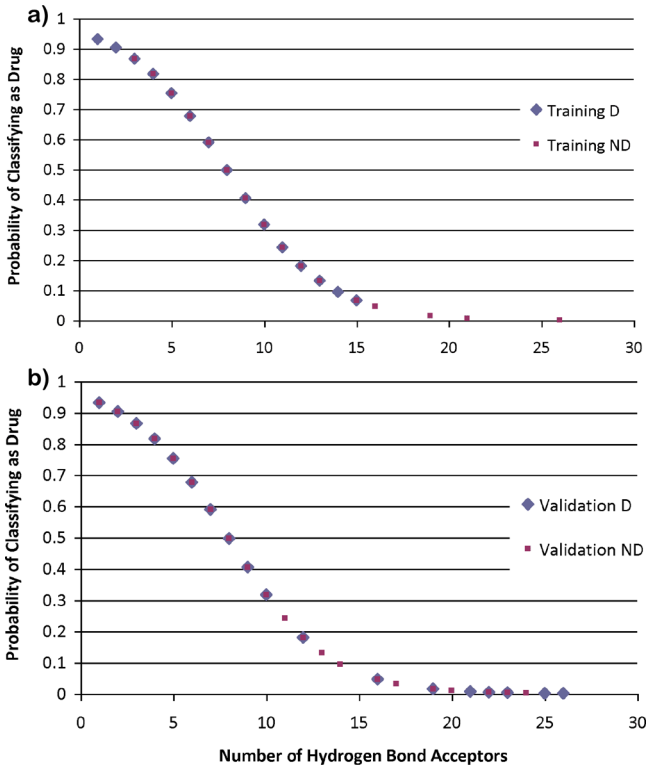


Figure 1. Logistic regression (eqs 3 and 11) between the number of hydrogen bond acceptors of a compound in the (a) training data set or (b) validation data set and its predicted probability of being classified as a drug compound (D, $P_{drug} = 1$), as opposed to a nondrug compound (ND, $P_{drug} = 0$). Training data set, $n^{drugs} = 311$, $n^{nondrugs} = 320$. Validation data set, $n^{drugs} = 106$, $n^{nondrugs} = 106$.

Figure 1a shows how true drugs mostly predominate in the early (top) part of the curve, while nondrugs are mostly in the bottom part. The shape of the plots follows an inverse sigmoidal function (i.e., showing anticorrelation), decreasing gradually from probabilities close to 1 for a small number of hydrogen bond acceptors to a probability close to zero for a large number of hydrogen bond acceptors. The transition between the two reflects the slowly differing nature between compounds due to their number of hydrogen bond acceptors, and there is a threshold around 8 acceptors (inflection point) for both the training set and the validation set. These results are in agreement with the threshold of less than 10 acceptors included in Lipinski's rule-of-five[7] and particularly well with the threshold of 8 acceptors determined for 90% of marketed drugs.[27] The threshold of close to $P = 0.5$ produced a good distinction between drugs and nondrugs, with only a small overlap between both groups of compounds at that P value (see Figure 4 below). The inverse relationship between the independent and outcome variable is also evidenced in the negative coefficient in eqs 3 and 11. The equations also allow a quantitative analysis: in eq 11, for every unit increase in the number of acceptors, the probability of a compound being classified as drug decreases, i.e., the log odds of being a drug decreases by 0.375. This can also be expressed as for every unit increase in number of acceptors, the odds of being classified as a drug compound versus a nondrug decrease by a factor of 1.455. Alternatively, this can also be expressed as every unit increase in number of acceptors decreases the odds of being a drug by 45.5%. The rationale is that the more hydrogen bond acceptors a molecule may have, after reaching a limit, they will become an obstacle for a compound to reach its target, as well as have more groups that may complicate metabolism, as well as increase the desolvation penalty for extracting from bulk solvent and binding to its target.

Figure 2 shows the logistic relationship between PSA (in Å^2) with the probability of a compound's classifying as a drug (P_{drug}) for all compounds studied, according to eq 4.

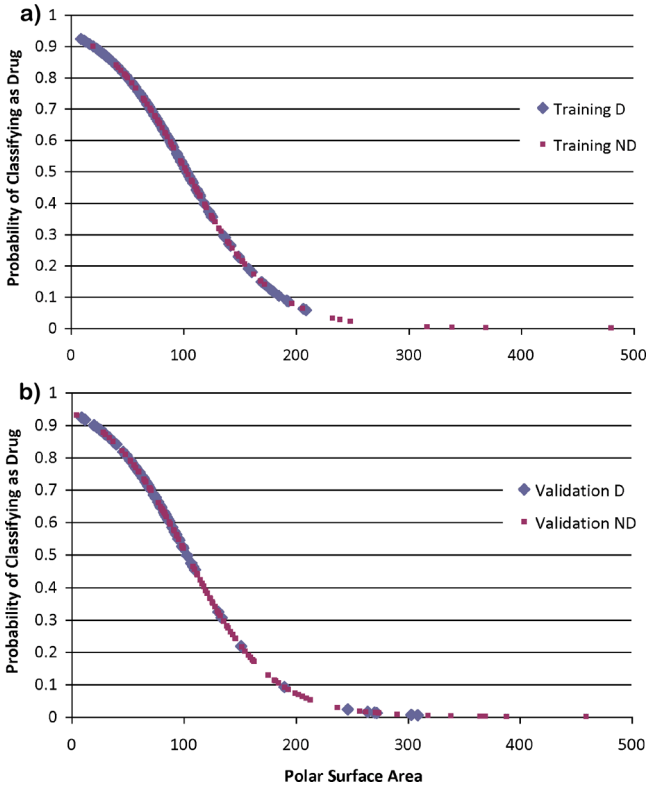


Figure 2. Logistic regression (eq 4) between the polar surface area of a compound in the (a) training or (b) validation data set and its predicted probability of being classified as a drug compound (D, $P_{drug} = 1$), as opposed to a nondrug compound (ND, $P_{drug} = 0$). Training data set, $n^{drugs} = 311$, $n^{nondrugs} = 320$. Validation data set, $n^{drugs} = 106$, $n^{nondrugs} = 106$.

As was the case for expressions 3 and 11, the properties and probabilities of training and validation data sets are in good agreement. There is also a higher proportion of drugs both for the training set, Figure 2a, as well as for the validation set, Figure 2b, in the early or top part of the curve, corresponding to higher P values, as well as a higher proportion of nondrugs for both sets in the bottom part of the curve, indicating less probability of classifying as a drug compound, i.e., less drug-likeness. The degree of miss-prediction in the curve is indicated by the number of drugs in the bottom part and nondrugs in the higher part of the curves. A perfect curve would have all drugs in the top half of the curve and all nondrugs in the bottom half. The curves also allow to see the gradual (here quantitatively described) transition between drugs and nondrugs. Here, the inflection point at $P_{drug} = 0.5$ occurs at PSA = 100 to 105 Å^2. This is also in agreement with previous studies where values under 140 Å^2 were identified as important to identify good oral bioavailability in compounds,[85] as well as with a study where values under 120 Å^2 were reported to be important for orally active drugs that are transported passively through cells.[86]

Some of the expressions in Table 1 also show a balance between terms. For example, in eq 6 the positive coefficient terms

dx.doi.org/10.1021/ci200587h | J. Chem. Inf. Model. 2012, 52, 2165−2180

Table 2. Mathew's Correlation Coefficient, True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for the Training and Validation Data Sets, Area under the Curve (AUC) and Enrichments at First (top) 5% of Ranked Data Set[ab]

equation	Mathew's coefficient	training data set TP, TN, FP, FN	validation data set TP, TN, FP, FN	AUC	enrichment
3	0.58	244, 257, 68, 64	96, 52, 54, 10	0.86	36
4	0.59	257, 244, 55, 77	97, 65, 41, 9	0.86	43
5	0.64	264, 254, 48, 67	95, 62, 44, 11	0.88	44
6	0.69	274, 261, 38, 60	93, 73, 33, 13	0.90	37
7	0.65	259, 262, 53, 59	93, 71, 35, 13	0.89	45
8	0.64	269, 249, 43, 72	97, 60, 46, 9	0.87	43
9	0.67	274, 255, 38, 66	95, 64, 42, 11	0.91	55
10	0.64	263, 257, 49, 64	96, 67, 39, 10	0.90	44
eq 6 in consensus with eq 9	0.77	257, 238, 43, 21	94, 62, 43, 12	–	–

[a] P threshold set at 0.45. [b] Equations 3–10 are shown in Table 1. Training data set, $n^{drugs} = 311$, $n^{nondrugs} = 320$. Validation data set, $n^{drugs} = 106$, $n^{nondrugs} = 106$.

for aliphatic ring count and for log$P$ are counterbalanced by the negative coefficient for MW, effectively balancing a larger number of rings and higher hydrophobicity against a smaller overall molecular size. A similar behavior is seen in eq 7, where a larger aliphatic ring count is offset by negative terms for increasing polar surface area and increasing bond count, i.e., constraining size and number of polar groups while permitting few rings. Equation 8 contains the Wiener index, and eq 10 contains the derived ΔG_{bind}/Wiener. This topologically derived real number increases with the size of a compound as well as with its linearity (e.g., extended chains) and is calculated through the number of bonds and their connectivity in the molecular graph. It can thus be related to the "compactness" of a molecule, i.e., a measure of how branched a molecule is, as well as to its intermolecular interactions.[87,88] Equation 9 contains the Balaban index, another chemical graph derived measure that is less dependent on the size of the molecule and its number of rings, and has elements that are lower in less branched isomers.[88] Thus, for both of these indices, a smaller number indicates a more compact or less branched molecule.

Ligand Efficiency. Ligand efficiency indices are also present among the terms of the expressions in Table 1 (eqs 9 and 10). They were usually found multiplied by negative coefficients that applied on the negative sign from the ΔG_{bind} part produce a positive correlation effect on the predicted outcome probability of being a drug. This makes sense because the larger the value of a LE the more effective the compounds will be due to there being fewer parts of the compound that do not participate in binding. For a large LE value, either or both the binding energy must be high and the NF must be small. Ligand efficiency indices are thus useful in characterizing drug compounds as they had previously been shown for distinguishing probability density functions,[5] as well as improving correlations between experimental LE and calculated LE values,[75] and grouping compounds in PCA analysis.[62]

Lipinski and "Drug-Like" Tests. Tests were also carried out for a combined term that included the four variables described in Lipinski's rule-of-five (R-o-5). Interestingly, the combined terms did not show correlation with the outcome variable for the training set, but coding them as a test (i.e., 1 for passing Lipinski's rule-of-five criterion, 0 for not) did return an acceptable regression expression

$$-1.466 + 1.983*\text{R-o-5} \quad (12)$$

The equation using this equation had a low miss-prediction rate of 32% for the training data set, though this same regression equation produced a high 58% miss-prediction rate for the validation data set. Similar results were produced by another test based on the Lipinski test and in addition using the criteria of PSA < 140 Å^2 and number of rotatable bonds under 10 (Veber, also coded 1 for a pass, 0 otherwise)[85]

$$-1.719 + 2.206*\text{Veber} \quad (13)$$

This may be explained by the fact that Lipinski's rule-of-five was developed using orally administered compounds that might provide an indication of bioavailability, and our data sets contain drug compounds with different administration routes other than just the oral route. Both the regression equation for Lipinski's test and for the "drug-like" test show that the intercept nearly cancels out the variable, which can be only 1 or 0, so the predicted outcomes have only two values, one for each value of the variable. Another explanation is that the rule-of-five could be extended and modified on the basis of new results that include different routes of administration and a new set of rules derived for each administration route.

Near Misses and Prospective Compounds. In addition to Table 1, the complete numbers of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) as well as Mathew's correlation coefficients for all the equations based on expressions in Table 1 are shown in Table 2. In addition, the area under the curve (AUC) for each ROC graph for each equation is also shown in Table 2, as well as their enrichments at the first (top ranked) 5% of the data set, while the ROC plots are shown in Figure 3.

From Table 2, it can be seen that there is good accuracy for the amount of true positives and true negatives, as well as relative symmetry among them, with some equations being better than others. A scenario can also be foreseen where several of the equations can be used as successive filters. For example, using eqs 6 and 9 in consensus for the training set leads to a high Mathew's correlation coefficient of 0.774, with 257 true positives, 238 true negatives, 43 false positives, and 21 false negatives. Thus, using multiple equations as filters may sometimes increase the power of discrimination.

The ROC curves from Figure 3 as well as the high AUC values in Table 2 for these equations show a good and early discrimination between known drugs (true positives) and false positives, as well as between known nondrugs (true negatives) and false negatives, much better than random selection. A random selection would follow the diagonal line and have an AUC of 0.5, whereas all of these equations have around 0.9 units. It also shows that the procedure in the equations is adequate and

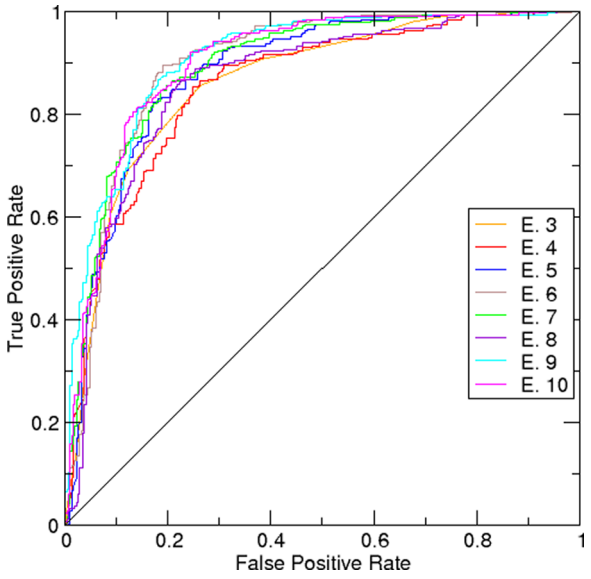


Figure 3. ROC curves for eqs 3−10 and for random selection (diagonal line).

useful for recovering known actives, in this case known drugs. This good selection is also proven by plotting the distribution of predicted probabilities as shown in Figure 4 for eq 9.

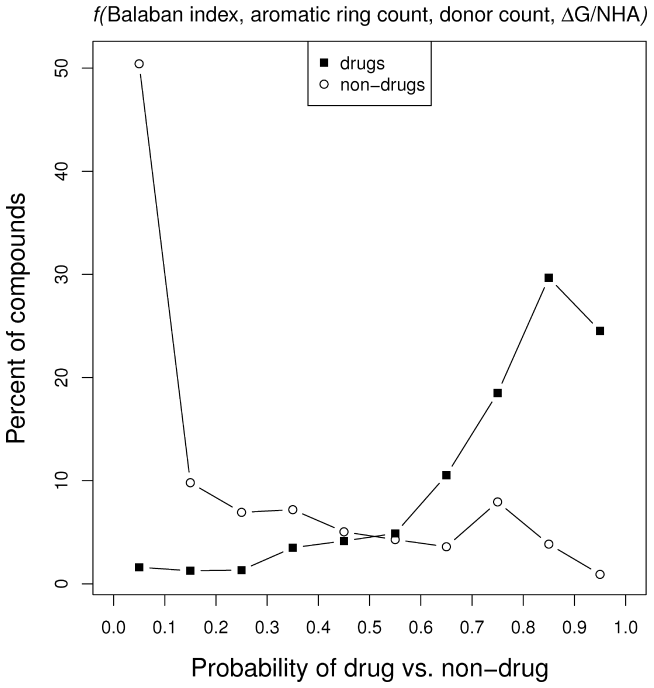


Figure 4. Distribution of predicted probabilities according to eq 9.

Figure 4 shows that the majority of drugs and nondrugs are clearly separated by their predicted probabilities by eq 9, and only a very minor fraction of less than 5% of the compounds are in the region of transition around $P = 0.5$. Regions of probability that are more extreme than 0.5 at either end provide a larger separation between drugs and nondrugs.

Variability near the threshold of $P = 0.5$ is normal for logistic regression. Some of the miss-predicted compounds, or "near misses", are interesting in their features and chemistry. For example, for eq 4 the naturally produced drug adenosine slightly missed the threshold with a P of 0.4. A look at its chemical structure (Figure 5) shows it has a moderate to high number of

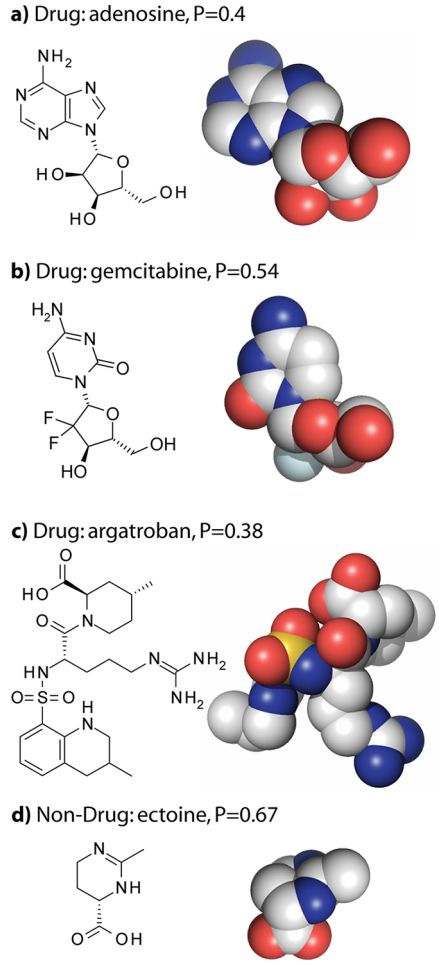


Figure 5. Chemical structures, drug status and probablility, and van der Waals surface areas colored according to polar atoms for a selection of compounds predicted as near-misses (a, c, d) and hits (b) according to eq 5.

rings of three and a moderate to high PSA of 113.5 Å^2. A similar compound, the modified nucleoside gemcitabine, has one ring less and nearly 29 Å^2 less of PSA. Gemcitabine is the result of synthesis and design that most likely resulted in improved physicochemical properties with respect to adenosine, which are taken into account by eq 5 to produce a $P = 0.54$. The drug argatroban (Figure 5) had a P of 0.38. Its values resemble adenosine's because it also has a moderate to high number of rings (three) and a moderate to high PSA (122.8 Å^2). On the other hand, the nondrug ectoine (1,4,5,6-tetrahydro-2-methyl-4-pyrimidinecarboxylic acid, 2vpn, Figure 5), had a $P = 0.67$ due to a low number of rings (one) and a low PSA of 49.7 Å^2. In fact, it is a natural product produced by bacteria to protect against salt and temperature stress and is used in cream products for human use. Thus, the equations produced show that a nondrug compound with appropriate chemical features can be classified as a useful bioactive and may have a potential to be developed into a therapeutic compound (provided many other factors). In addition, the drug character of compounds can be increased by reducing their number of rings and PSA. The chemical structures as well as van der Waals surface areas colored according to their polar atoms for these three compounds are shown in Figure 5.

dx.doi.org/10.1021/ci200587h | J. Chem. Inf. Model. 2012, 52, 2165−2180

Table 3. Predicted Probabilities by Eq 9 for the Available Best Selling Drugs

drug	*P*	drug	*P*	drug	*P*	drug	*P*
amlodipine	0.62	enalapril	0.70	losartan	0.76	ranitidine	0.72
amphetamine	0.91	fenofibrate	0.89	memantine	0.93	risperidone	0.95
anastrazole	0.90	fentanyl	0.94	methylphenidate	0.89	ropinirole	0.89
aripiprazole	0.86	fexofenadine	0.42	nifedipine	0.78	rosiglitazone	0.91
bicalutamide	0.60	fluconazole	0.90	olanzapine	0.85	salmeterol	0.37
bosentan	0.67	gabapentin	0.58	omeprazole	0.87	sertraline	0.89
bupropion	0.82	gemcitabine	0.55	ondansetron	0.95	sildenafil	0.87
candesartan	0.65	imatinib	0.72	paclitaxel	0.22	sumatriptan	0.72
celecoxib	0.78	irbesartan	0.76	paroxetine	0.91	tacrolimus	0.56
ciprofloxacin	0.64	lamivudine	0.75	pioglitazone	0.90	topiramate	0.82
clarithromycin	0.25	lamotrigine	0.71	pramipexole	0.90	valproate	0.67
diclofenac	0.69	lansoprazole	0.91	pravastatin	0.33	vardenafil	0.87
dorzolamide	0.79	latanoprost	0.54	progesterone	0.97	venlafaxine	0.86
doxazosin	0.88	levetiracetam	0.76	quetiapine	0.82	zidovudine	0.56
duloxetine	0.86	lidocaine	0.78	raloxifene	0.66	zolpidem	0.94

Table 4. Logistic Regression Expressions and Equations with Statistical Significance at the 95% Confidence Level or Higher To Predict the Probability of a Compound Being a Drug of a Specific Disease Category Compared to Other Drugs

disease category (expression)	Intercept (Std. Err.) and Coefficients (Std. Err.)*Variables				
	median of deviance residuals	Mathew's coefficient (TP, TN, FP, FN)	accuracy (%) for correct prediction, miss-prediction	AUC	enrichment (at 5%)
DC8 (eq 14)	−4.9(0.6) + 0.4(0.2)***NoC** + 0.03(0.01)***PSA** − 0.23(0.08)***molecular polarizability**				
	−0.21	0.39 (5, 343, 2,16)	95, 5	0.83	50
DC10 (eq 15)	2.4(0.9) + 1.5(0.3)***aromatic ring count** + 0.8(0.2)***aliphatic ring count** − 0.21(0.05)***NoC** − 0.029(0.005)***PSA** + 47.8(18.6) ***ΔG_{bind}/MSA**				
	−0.46	0.35 (45, 229, 26, 66)	74, 26	0.77	12.2

Equation 9 provided good characterization for the best selling drugs in the data,[89] and are shown in Table 3.

As seen in Table 3, there is a good prediction of different types of drugs, according to eq 9. The few miss-predictions include the originally natural products paclitaxel, clarithromycin, and pravastatin, which may have a more complex chemical structure than synthesized compounds. Fexofenadine scores low for eq 9 because its Balaban index is higher than other drugs, given that it is not as small as other drugs, as well as having three aromatic rings, three hydrogen bond donors, and a comparatively high number of NHA as compared to other drugs, all of these property values being disfavored by eq 9. A different equation, such as using eqs 3, 5, 6, or 8 would be more suitable for fexofenadine, giving a $P = 0.75$, 0.73, 0.70, and 0.65, respectively. Equations 4 and 10 give $P = 0.58$ and $P = 0.57$, respectively, while eq 7 gives $P = 0.43$. In addition, an average over all the P values for fexofenadine for eqs 3−10 (all of those in Table 1), gives a $P = 0.64$. Salmeterol, on the other hand, is also slightly atypical for a drug compound in that it has a large, bulky, and lipophilic group, in addition to four hydrogen bond donors that also make this compound score low for eq 9. A different equation would be more suited for this compound, such as using eq 3 ($P = 0.75$), eq 4 ($P = 0.70$), eq 5 ($P = 0.65$), and eq 8 ($P = 0.73$). Equations 6 and 7 give $P = 0.55$ and $P = 0.43$, respectively. In addition, the average for salmeterol over eqs 3−10 gives $P = 0.62$.

Predicting Disease Category Specificity. A similar procedure was carried out for classifying drug compounds according to their disease category. Compounds were removed if they were part of more than one DC in order to reduce the noise that multi-DC compounds may produce. In this case, drugs were differentiated among themselves to produce regression equations that would distinguish a specific disease category from the others (i.e., drugs in a DC against other drugs in different DCs). Separation between the individual DCs is not as clear-cut as in the case of drugs versus nondrugs, and only a few equations with Mathew's coefficients higher than 0.35 were obtained. These are shown in Table 4.

It should be noted that eq 14 provides distinction at a P threshold of 0.5 between drugs belonging to DC8 versus other DCs (and eq 15 for DC10) after inserting their terms, respectively, into eq 1. The same rationalization as performed for univariate logistic regression with eq 3 can be performed for a multivariate case, where holding other variables fixed, the coefficient for a variable shows its change in the log odds of having a predicted outcome of 1 per unit increase in that variable. Table 4 also shows enrichments at the first (top) 5% of data set and AUCs for the respective ROC curves presented in Figure 6.

From Figure 6, it can be seen that these equations still perform better than random and have areas under the curve of around 0.8 units. The fact that no equations were possible for other DCs

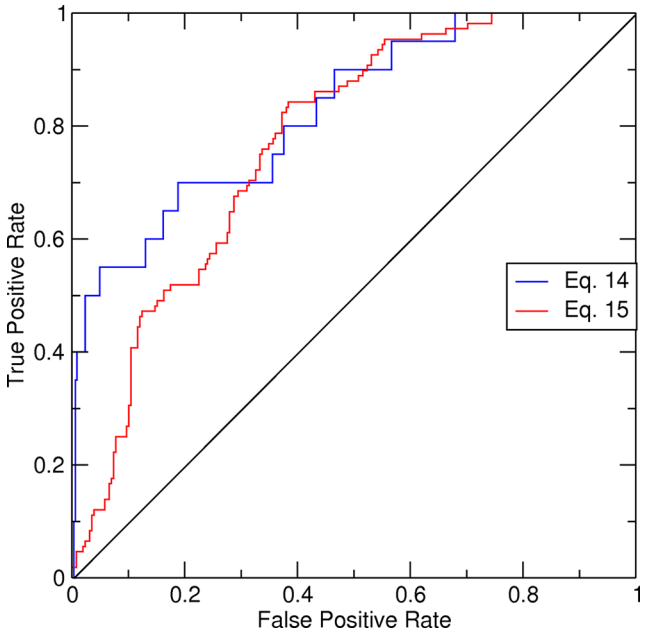


Figure 6. ROC curves for eqs 14 (DC8, anti-neoplastics) and 15 (DC10, nervous system agents) and random selection (diagonal line).

than those in Table 4 may be due to the small number of members of each individual group in the data set and the promiscuity of a large number of compounds. This may also be due to the fact that most drugs are optimized for pharmaceutical action and safety but are not perfectly optimized for disease specificity. This is seen in the many side effects that drugs have, which can sometimes be exploited for a new indication of a drug. This may also speak of some disease categories being harder to predict than others. A collection of typical class compounds for the DCs is shown in Figure 7.

DC8 contains cancer treatment drugs. Some of these are enzyme inhibitors, though many are receptor agonists or antagonists. As such, a majority of these compounds are moderately large and with a moderate amount of functionality (e.g., compound tamoxifen, Figure 7). Equation 14 reflects this with positive coefficients for the number of carbon atoms and for PSA, balanced by a negative coefficient for molecular polarizability. The latter is a measure of the charge distribution within a molecule and can be related to its solvation energy through solvent/solute interactions, as well as being useful in describing the lipophilicity (hydrophobicity) of compounds.[87,88]

Drugs that act on the nervous system (DC10) are generally moderate in size and relatively hydrophobic (e.g., compound diazepam, Figure 7). A large number of them act on the brain and thus have an extra barrier that is relatively impermeable to cross. Many contain the CNS active phenylethylamine substructure. Equation 15 shows these effects in negative coefficients for the number of carbons and PSA combined with positive and small contributions from the number of aromatic or aliphatic rings as well as for ΔG_{bind}/MSA (which translates as favoring hydrophobic compounds).

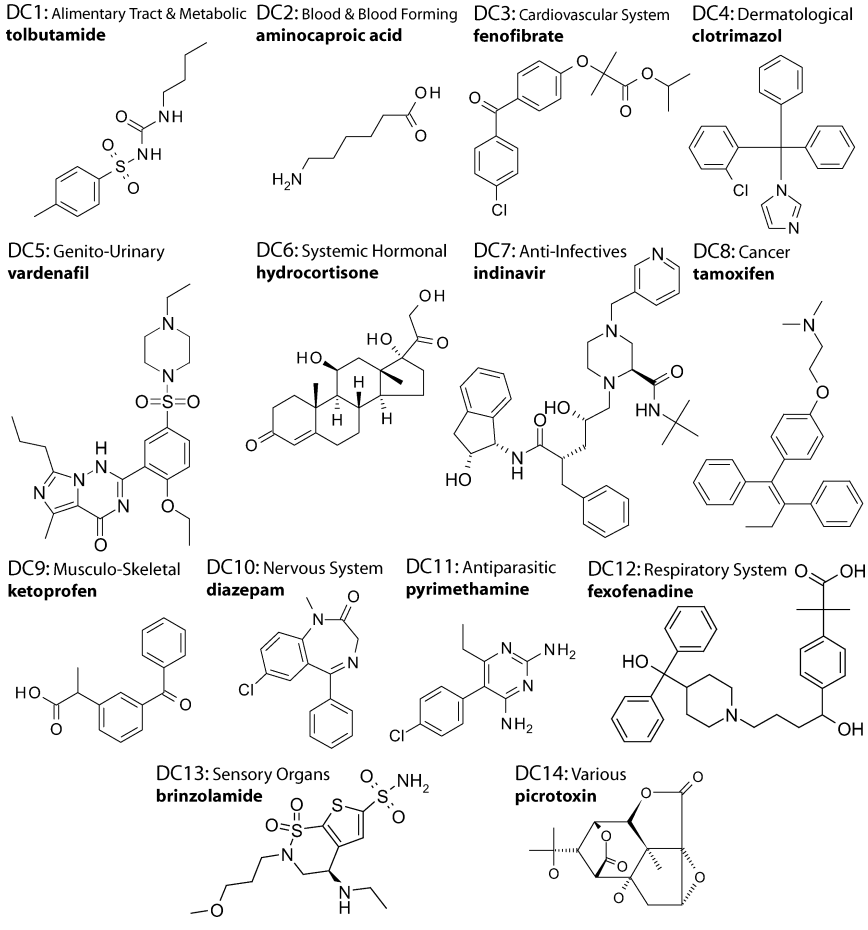


Figure 7. Typical compounds for specific disease categories.

dx.doi.org/10.1021/ci200587h | J. Chem. Inf. Model. 2012, 52, 2165−2180

Table 5. Logistic Regression Expressions and Equations with Statistical Significance at the 95% Confidence Level or Higher To Predict the Probability of a Compound Being a Drug of a Specific Disease Category Compared to Nondrugs[a]

disease category	Intercept (Std. Err.) and Coefficients (Std. Err.)*Variables				
	median of dev. res.	Mathew's coefficient (TP, TN, FP, FN) for training and validation data sets	accuracy (%) correct prediction, misprediction; training and validation sets	AUC	enrichment (at 5%)
DC1 (eq 16)	−3.2(1.5) + 2.0(0.6)*aromatic ring count − 0.02(0.01)*PSA − 156(61)*ΔG_{bind}/APSA				
	0.0	0.72 (28, 27, 5, 4) 0.63 (7, 6, 2, 1)	86, 14; 81, 19	0.91	65
DC2	No selective, statistically significant equation found				
DC3 (eq 17)	5.9(1.6) − 1.3(0.6)*Balaban index + 0.9(0.2)*log *P* − 0.006(0.002)*MSA − 0.17(0.07)*aromatic atom count				
	0.02	0.85 (57, 54, 6, 3) 0.91 (10, 11, 0, 1)	92, 8; 91, 9	0.95	73
DC4	No selective, statistically significant equation found.				
DC5 (eq 18)	−0.5(1.3) + 1.5(0.6)*aromatic ring count − 0.07(0.02)*PSA + 0.1(0.04)*Wiener polarity				
	−0.02	0.81 (24, 34, 3, 3) 0.78 (8, 8, 1, 1)	91, 9; 89, 11	0.95	89
DC6	No selective, statistically significant equation found.				
DC7 (eq 19)	3.4(1.4) + 1.5(0.4)*ring count − 0.4(0.2)*NoC − 0.07(0.02)*PSA + 0.2(0.09)*Wiener polarity				
	−0.00	0.86 (27, 27, 2, 2)	91, 9; −, −	0.95	79
DC8	No selective, statistically significant equation found.				
DC9	No selective, statistically significant equation found.				
DC10 (eq 20)	4.7(1) − 0.6(0.2)*log P − 0.04(0.01)*PSA + 0.01(0.01)*APSA − 1.2(0.2)*donor count				
	0.05	0.86 (92, 88, 9, 5) 0.77 (27, 26, 4, 3)	93, 7; 88, 12	0.97	92
DC11	No selective, statistically significant equation found.				
DC12 (eq 21)	−5.7(2.9) + 4(1.2)*aromatic ring count − 2.15 × 10^{-4}(9.7 × 10^{-5})*hyper − Wiener index − 180.2(91.8) *ΔG_{bind}/APSA				
	0.01	0.81 (23, 24, 2, 3) 0.39 (13, 9, 7, 3)	90, 10; 69, 31	0.97	84
DC13 (eq 22)	−2.5(1.6) + 0.9(0.3)*log *P* − 0.04(0.01)*PSA − 6.5(2)*ΔG_{bind}/NoC				
	0.01	0.91 (33, 32, 2, 1) 0.87 (14, 12, 2, 0)	96, 4; 93,7	0.93	67
DC14 (eq 23)	0.3(1.4) − 0.8(0.2)*donor count − 104.8(51)*ΔG_{bind}/APSA				
	−0.01	0.77 (19, 20, 2, 3) 0.67 (18, 17, 4, 3)	89, 11; 81, 19	0.93	82

[a]Median of dev. res. = Median of deviance residuals. AUC = Area under the curve of ROC plots in Figure 8.

Nondrugs vs DC Drugs. A similar method to that employed in the drugs vs nondrugs was employed but now comparing nondrug compounds to drugs belonging to a particular DC. Compounds belonging to more than one DC were removed. A selection of the best statistically significant equations is shown in Table 5, together with values from ROC plots shown in Figure 8.

The ROC plots in Figure 8 together with the AUC (higher than 0.9 units) and enrichment values from Table 5 show a very good retrieval of known drugs of particular DC as compared to known nondrugs. This is also evident in the plots of the distribution of values (line histograms) corresponding to the predicted values according to the equations based on expressions in Table 5 and shown in Figure 9.

The expressions in Table 5 are specific for each DC, i.e., they are not shared with another DC and as such can help distinguish drugs belonging to different disease types. The numbers of TP, TN, FP, and FN and the plot distributions show good symmetry. DC1 corresponds to alimentary tract and metabolism diseases. From Table 5, eq 16 has positive coefficients for the number of aromatic rings that combined with the negative signs for PSA and ΔG_{bind}/APSA (i.e., favoring small compounds) provides a balanced correlation with the predicted outcome of classifying a drug as belonging to DC1. An example compound in DC1 is tolbutamide (Figure 7). DC2 contained only seven compounds, and no selective, statistically significant regression equation was found.

For DC3, small size and a moderate number of polar groups are important to distinguish specificity for cardiovascular drugs as seen in eq 17. One can observe this effect in the example of the structure of fenofibrate (see Figure 7). The place of action of these drugs is the heart and circulatory system, and as such, they must reach their targets through several membranes. The properties described in eq 17 (negative coefficient for Balaban index, MSA, and number of aromatic atoms, i.e., favorable to small molecules, and positive coefficient for log*P*, i.e., favoring nonpolar compounds) are able to select compounds for this type of therapeutics.

dx.doi.org/10.1021/ci200587h | J. Chem. Inf. Model. 2012, 52, 2165−2180

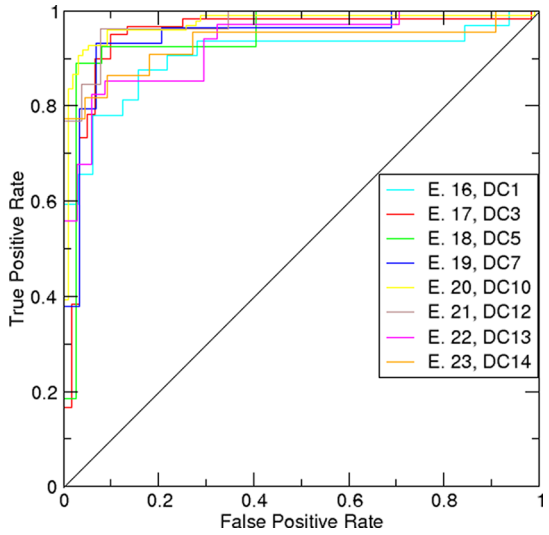


Figure 8. ROC curves for eqs 16−23 and random selection (diagonal line).

A typical compound for DC4, dermatological agents is clotrimazol (Figure 7). The drugs for this DC tend to be hydrophobic because they are applied directly to the skin, and transdermal absorption would be important. However, these properties are not exclusive to this DC, and the equations found for this DC were also applicable to other DCs (hence, none are shown in Table 5). This can be due to the fact that many dermatological drugs also belong to other DCs (such as erithromycin, also a topical antibiotic) and are therefore less specific.

DC5 contains genito-urinary drugs and many act on receptors that are cell membrane bound. Reaching them imposes a balance in the chemical properties of these drugs. The number of aromatic rings, Wiener polarity, and PSA are important as shown by their respectively positive, positive, and negative coefficients in eq 18, as well as the structure of vardenafil (Figure 7). Wiener polarity measures the number of pairs of graph vertices and relates to the flexibility of acyclic structures and their steric effects.[88]

Hydrocortisone (Figure 7) is a typical example compound for DC6, Systemic hormonal drugs. Most compounds in this group have a similar shape (flat) and several rings. These compounds bind to receptors that have similar binding pockets, and thus, they possess relatively similar chemical structure properties. However, they also belong to several DCs, and so no selective, statistically significant equations were found for this DC. They also typically possess a variety of side effects such as is the case with glucocorticoids.

Equation 19 presents an interesting case. DC7 contains anti-infectives, and many of them are relatively large compounds but with some functionality. Accordingly, eq 19 has a positive coefficient for ring count and Wiener polarity and at the same time a negative coefficient for PSA. Thus, compounds in this DC are distinguished by their size and limited functionality, which are useful properties due to their mainly enzyme (transpeptidases, HIV protease, lanosterol 14 α-demethylase, etc.) inhibitor character (e.g., compound indinavir, Figure 7). No drugs in the validation set belonged to this DC.

DC8 contains anticancer compounds. Given the wide variety of their molecular targets and sometimes their lack of organ specificity in their mechanism of action, they are very varied. No selective, statistically significant regression equations were found for this DC vs nondrug comparison (though there was an equation for inter-DC comparison as seen above) nor for musculo-skeletal drugs (DC9) nor antiparasitics (DC11).

Figures 8 and 9 show the good distinction between drugs for DC10 and nondrugs. This is also reflected in the calculated values for the probabilities of the available best selling drugs for eq20 shown in Table 6, where all of the available best selling drugs for this DC scored well.

DC12 (respiratory system) drugs tend to be of moderate size and have at least one aromatic ring (e.g., fexofenadine, Figure 7), which is reflected in eq 21, with a positive coefficient for number of aromatic rings balanced by negative coefficients for hyper-Wiener index and ΔG_{bind}/APSA. The hyper-Wiener index is a measure of the "expandedness" of a chemical graph of a molecule, more sensitive than Wiener index for expanded graphs.[88]

Different shaped compounds are characteristic for sensory organ drugs (DC13, see brinzolamide, Figure 7), given the different target organs they act upon. Equation 22 favors hydrophobic compounds.

DC14 contains various drugs, i.e., drugs that do not belong to other disease categories. Their chemical structures are quite varied, as expected (see compound picrotoxin, Figure 7). However, there were equations found for this group, specifically, eq 23 has negative coefficients for number of donors and for ΔG_{bind}/APSA, which would act to reduce in general the size of the compounds. Although this function can be viewed to apply to all drugs, the equation was specific for this DC. Compounds in DC14 are indeed varied in their structure and mechanism. Equation 23 does provide a score, one that is based on relatively general factors such as donor count, binding energy, and apolar surface area, which in principle characterize general drug-likeness as opposed to specific DC characteristics. It should be noted that eqs 14 and 15 were generated considering drugs of different DCs vs DC8 and DC10, respectively. However, eq 23 for DC 14 was generated using DC14 drugs vs nondrugs. In that respect, eqs 14 and 15 are more discriminating between specific DCs than eq 23.

In our previous study, we found that drugs and their disease categories could be charted and ranges established on the basis of their molecular properties and score plots using PCA analysis.[62] The present work shows that drugs can be distinguished from nondrugs as well as from drugs for different diseases or organs with statistical significance through probability based on multivariate logistic regression analysis.

The most difficult DCs to assign unambiguously are perhaps DC7 (anti-infectives) and DC8 (anti-neoplastic), probably due to their action in many different organs or locations in the body, and therefore, overlap may occur with other DCs. The other DCs appear to be better localized. This also shows that certain targets of drug actions, for example, estrogen receptors or other receptors, may be located in different degrees of expression and of subtypes in different organs or parts of the body. Classifying compounds according to their chemical molecular properties as we have done may allow distinguishing the chemical space available to separate receptors at different target organs or tissues. Alternatively, in some occasions, if the chemical space available for a certain disease overlaps with another DC, it will strongly suggest the possibility of multiple effects or indications for a particular drug. For example, raloxifene is a genito-urinary system and sex hormone disease category (DC5) drug that shares the same estrogen receptor and mechanism as tamoxifen, an anti-neoplastic (endocrine therapy, DC8) drug. Another option of looking for specificity among DCs is taking into account

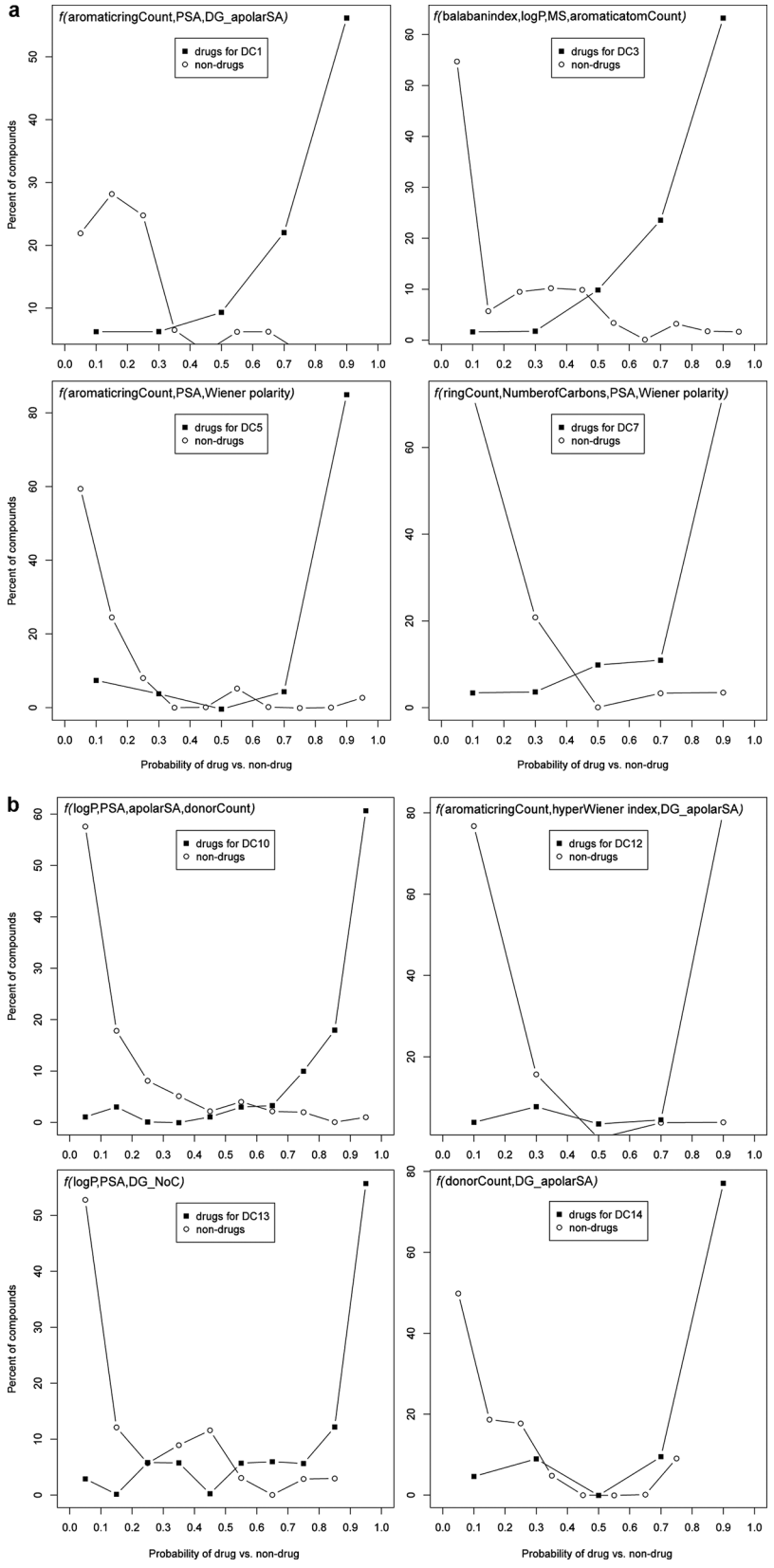


Figure 9. (a,b) Distribution of predicted probabilities for drugs belonging to specific disease categories and nondrugs.

descriptors that are unique to a specific DC, such as molecular polarizability, which is unique to eq 14 for DC8 (anti-neoplastic agents) or hyper-Wiener index, which is unique to eq 21 for DC12 (respiratory system).

Further Tests. After the completion of the regression study and validation tests, new compounds were obtained to further test the equations produced before. Experimentally verified and in-use drug compounds were found, mostly due to them not

dx.doi.org/10.1021/ci200587h | J. Chem. Inf. Model. 2012, 52, 2165−2180

Table 6. Predicted Probabilities by Eq 20 for Best Selling Drugs in Disease Category 10, Nervous System

drug	P_{DC10}	drug	P_{DC10}
amphetamine	0.93	paroxetine	0.96
aripiprazole	0.95	pramipexole	0.87
bupropion	0.95	quetiapine	0.78
duloxetine	0.89	risperidone	0.97
fentanyl	0.99	ropinirole	0.97
gabapentin	0.80	sertraline	0.93
lamotrigine	0.59	sumatriptan	0.77
levetiracetam	0.96	topiramate	0.84
lidocaine	0.97	valproate	0.84
memantine	0.89	venlafaxine	0.94
olanzapine	0.93	zolpidem	0.98

having binding data available at the time when the data sets were collected or having been approved since. These 224 compounds were able to be correctly classified into DCs 3, 5, 7, and 10, according to eqs 17−20, respectively, and with good accuracy (reported as sensitivity, see eq 24, also named "recall" in some sources) of at least 70% or higher. The results are shown in Table 7, with the compound names and associated DC shown in Table S1 in the Supporting Information.

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \quad (24)$$

Table 7. Success Rates for Classifying Additional Compounds into Specific DCs According to the Developed Equations

disease category (equation)	True Positives	False Negatives	total number	accuracy % (as sensitivity)
DC3 (eq 17)	66	4	70	94
DC5 (eq 18)	17	4	21	81
DC7 (eq 19)	42	18	60	70
DC10 (eq 20)	60	13	73	82

Given that predicting drug-likeness is a challenging task as well as a constantly transforming environment due to changes in legislation and drug approval and removal, as well as the changing nature of candidate compounds over time, achieving a degree of accuracy of at least 70% can help in better defining quantitative and gradual measures of drug-likeness. Through these equations and in combination with others, better profiling of drug libraries may be possible. One of the possible desirable goals would be the development of organ-specific chemical libraries. The regression expressions obtained in the present work have been coded and are available for use at http://hermes.chem.ut.ee/~alfx/druglogit.html.

CONCLUSIONS

The method described and validated allows calculating a predicted probability of classification as drug or nondrug for a compound on the basis of simple, readily available (or able to be calculated on-the-fly) properties. A variety of logistic regression equations are presented, from one term to the combination of up to five terms, and their implementation are straightforward to profiling compound libraries and selecting compounds with desired outcome qualities, such as organ-specific chemical libraries. The small amount of terms in the regression expressions and equations allows easy understanding of the drugs' properties based on their physicochemical attributes. In addition, testing for collinearity and removal of correlated terms prevented overfitting of variables to observables in all regressions.

A selection of the equations allows classifying the disease category of a compound. They are rationalized based on the different mechanism of action, administration mode, and target organs of different disease categories. It is envisioned that the method described here can be further extended to particular drug targets in order to achieve even finer specificity.

Sometimes a drug can be found by comparing ligand similarity to known active compounds (for example, using ligand-based design) and/or its interactions with its therapeutically relevant biomolecular target (enzyme, receptor, nucleic acid, etc., using, for example, pharmacophores or structure-based design[90]). However, in some cases the same receptor or enzyme may be located in different tissues, and therefore, an organ-based specificity would be desirable in order to better target the active compound. The regression functions presented here can be used in conjunction to other drug discovery and design techniques in order to further compartmentalize drug action. They can be extended or substituted using different molecular properties and can be applied in a similar manner to different problems in molecular and drug design. Additionally, the filters can be applied successively or in consensus in order to increase desired features, as well as used in conjunction with other filters to combine drug or DC drug features with others such as oral-bioavailability or lead-likeness.

The near-misses of the regression functions allowed comprehending the behavior of compounds based on their chemical properties and suggesting routes for compound modification. The regression functions also allow for a broad view of disease categories on the basis of chemical features.

ASSOCIATED CONTENT

Supporting Information

Figure S1 with histograms of ΔG_{bind} for drugs and nondrugs; Table S1 with names, drug status, training or validation set, and disease category of all compounds; Table S2 with discarded cross-terms due to strong correlation or anticorrelation; Table S3 with full results of molecular properties for all compounds studied, training and validation sets; and Table S4 with bootstraps (nonparametric) on 1000 resamples. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: alfonsog@ut.ee. Tel: +372 737 5270. Fax: +372 737 5264.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Estonian Ministry for Education and Research (Grant SF0140031Bs09), Estonian Science Foundation (Grant 7709), and the University of Tartu High Performance Computing Centre for computational resources. C.H. is thankful to the Hungarian Academy of Sciences for a Bolyai Scholarship.

ABBREVIATIONS LIST:

DC, disease category; LR, logistic regression; NHA, number of heavy atoms; NoC, number of carbons; LE, ligand efficiency indices (also called binding efficiencies); MSA, molecular surface

area; PSA, polar surface area; APSA, apolar surface area; ATC, Anatomical Therapeutic Chemical Classification

REFERENCES

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

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

(3) Hert, J.; Keiser, M. J.; Irwin, J. J.; Oprea, T. I.; Shoichet, B. Quantifying the relationships among drug classes. *J. Chem. Inf. Model.* **2008**, *48*, 755−765.

(4) Keiser, M. J.; Irwin, J. J.; Shoichet, B. K. The chemical basis of pharmacology. *Biochemistry* **2010**, *49*, 10267−10276.

(5) García-Sosa, A. T.; Maran, U.; Hetényi, C. Molecular property filters describing pharmacokinetics and drug binding. *Curr. Med. Chem.* **2012**, *19*, 1646−1662.

(6) Ursu, O.; Rayan, A.; Goldblum, A.; Oprea, T. I. Understanding drug-likeness. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 760−781.

(7) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(8) 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.

(9) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(10) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of 'drug-likeness'. *Drug Discovery Today* **2000**, *5*, 49−58.

(11) Ajay. Predicting drug-likeness: Why and how? *Curr. Top. Med. Chem.* **2002**, *2*, 1273−1286.

(12) Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **2003**, *23*, 302−321.

(13) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.

(14) Ritchie, T. J.; Macdonald, S. J. F. The impact of aromatic ring count on compound developability − Are too many rings a liability in drug design? *Drug Discovery Today* **2009**, *14*, 1011−1020.

(15) Chen, H. M.; Yang, Y. D.; Engkvist, O. Molecular topology analysis of the differences between drugs, clinical candidate compounds, and bioactive molecules. *J. Chem. Inf. Model.* **2010**, *50*, 2141−2150.

(16) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(17) Yang, Y. D.; Chen, H. M.; Nilsson, I.; Muresan, S.; Engkvist, O. Investigation of the relationship between topology and selectivity for druglike molecules. *J. Med. Chem.* **2010**, *53*, 7709−7714.

(18) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.

(19) Gálvez, J.; de Julián-Ortiz, J. V.; García-Domenech, R. General topological patterns of known drugs. *J. Mol. Graphics Modell.* **2001**, *20*, 84−94.

(20) Ajay. Predicting drug-likeness: Why and how? *Curr. Top. Med. Chem.* **2002**, *2*, 1273−1286.

(21) Batista, J.; Bajorath, J. Mining of randomly generated molecular fragment populations uncovers activity-specific fragment hierarchies. *J. Chem. Inf. Model.* **2007**, *47*, 1405−1413.

(22) Hu, Y.; Bajorath, J. Scaffold distributions in bioactive molecules, clinical trials, and drugs. *ChemMedChem* **2010**, *5*, 187−190.

(23) Wang, J. M.; Hou, T. J. Drug and drug candidate building block analysis. *J. Chem. Inf. Model.* **2010**, *50*, 55−67.

(24) Schneider, N.; Jäckels, C.; Andres, C.; Hutter, M. C. Gradual in silico filtering for druglike substances. *J. Chem. Inf. Model.* **2008**, *48*, 613−628.

(25) Brüstle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. Descriptors, physical properties, and drug-likeness. *J. Med. Chem.* **2002**, *45*, 3345−3355.

(26) Lee, M. L.; Schneider, G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: Application in the design of natural product-based combinatorial libraries. *J. Comb. Chem.* **2001**, *3*, 284−289.

(27) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physicochemical properties of development and marketed oral drugs. *J. Med. Chem.* **2003**, *46*, 1250−1256.

(28) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881−890.

(29) Tyrchan, C.; Blomberg, N.; Engkvist, O.; Kogej, T.; Muresan, S. Physicochemical property profiles of marketed drugs, clinical candidates, and bioactive compounds. *Bioorg, Med. Chem. Lett.* **2009**, *19*, 6943−6947.

(30) Oprea, T. I.; Gottfries, J. Chemography: The art of navigating chemical space. *J. Comb. Chem.* **2001**, *3*, 137−166.

(31) Abad-Zapatero, C.; Perisic, O.; Wass, J.; Bento, A. P.; Overington, J.; Al-Lazikani, B.; Johnson, M. E. Ligand efficiency indices for an effective mapping of chemico-biological space: the concept of an atlas-like representation. *Drug Discovery Today* **2010**, *15*, 804−811.

(32) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824−828.

(33) Vieth, M.; Sutherland, J. J. Dependence of molecular properties on proteomic family for marketed oral drugs. *J. Med. Chem.* **2006**, *49*, 3451−3453.

(34) Hopkins, A. L.; Paolini, G. V. Chemogenomics in Drug Discovery − The Druggable Genome and Target Class Properties. In *Comprehensive Medicinal Chemistry II, 4 (Computer-Assisted Drug Design)*; Mason, J. S., Ed.; Elsevier: Amsterdam, The Netherlands, 2007; pp 421−433.

(35) Mestres, J.; Martín-Couce, L.; Gregori-Puigjané, E.; Cases, M.; Boyer, S. Ligand-based approach to in silico pharmacology: Nuclear receptor profiling. *J. Chem. Inf. Model.* **2006**, *46*, 2725−2736.

(36) Sutherland, J. J.; Higgs, R. E.; Watson, I.; Vieth, M. Chemical fragments as foundations for understanding target space and activity prediction. *J. Med. Chem.* **2008**, *51*, 2689−2700.

(37) Sprous, D. G.; Palmer, K.; Swanson, J. T.; Lawless, M. QSAR in the pharmaceutical research setting: QSAR models for broad, large problems. *Curr. Top. Med. Chem.* **2010**, *10*, 619−637.

(38) Jacob, L.; Hoffman, B.; Stoven, V.; Vert, J. P. Virtual screening of GPCRs: An in silico chemogenomics approach. *BMC Bioinformatics* **2008**, *9*, 363.

(39) Strömbergsson, H.; Kleywegt, G. J. A chemogenomics view on protein−ligand spaces. *BMC Bioinformatics* **2009**, *10* (suppl 6), S13.

(40) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882−1889.

(41) Savchuk, N. P.; Balakin, K. V.; Tkachenko, S. E. Exploring the chemogenomic knowledge space with annotated chemical libraries. *Curr. Opin. Chem. Biol.* **2004**, *8*, 412−417.

(42) Kuhn, M.; Campillos, M.; González, P.; Jensen, L. J.; Bork, P. Large-scale prediction of drug-target relationships. *FEBS Lett.* **2008**, *582*, 1283−1290.

(43) Bender, A.; Young, D. W.; Jenkins, J. L.; Serrano, M.; Mikhailov, D.; Clemons, P. A.; Davies, J. W. Chemogenomic data analysis: Prediction of small-molecule targets and the advent of biological fingerprint. *Comb. Chem. High Throughput Screening* **2007**, *10*, 719−731.

(44) Young, D. W.; Bender, A.; Hoyt, J.; McWhinnie, E.; Chirn, G. W.; Tao, C. Y.; Tallarico, J. A.; Labow, M.; Jenkins, J. L.; Mitchison, T. J.; Feng, Y. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* **2008**, *4*, 59−68.

(45) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124−1133.

(46) Azzaoui, K.; Hamon, J.; Faller, B.; Whitebread, S.; Jacoby, E.; Bender, A.; Jenkins, J. L.; Urban, L. Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem* **2007**, *2*, 874−880.
(47) Gregori-Puigjané, E.; Mestres, J. Coverage and bias in chemical library design. *Curr. Opin. Chem. Biol.* **2008**, *12*, 359−365.
(48) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197−206.
(49) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175−181.
(50) Cases, M.; Mestres, J. A chemogenomic approach to drug discovery: Focus on cardiovascular diseases. *Drug Discovery Today* **2009**, *14*, 479−485.
(51) Gregori-Puigjané, E.; Mestres, J. A ligand-based approach to mining the chemogenomic space of drugs. *Comb. Chem. High Throughput Screening* **2008**, *11*, 669−676.
(52) Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 262−266.
(53) Oprea, T. I.; Bauman, J. E.; Bologa, C. G.; Buranda, T.; Chigaev, A.; Edwards, B. S.; Jarvik, J. W.; Gresham, H. D.; Haynes, M. K.; Hjelle, B.; Hromas, R.; Hudson, L.; Mackenzie, D. A.; Muller, C. Y.; Reed, J. C.; Simons, P. C.; Smagley, Y.; Strouse, J.; Surviladze, Z.; Thompson, T.; Ursu, O.; Waller, A.; Wandinger-Ness, A.; Winter, S. S.; Wu, Y.; Young, S. M.; Larson, R. S.; Willman, C.; Sklar, L. A. Drug repurposing from an academic perspective. *Drug Discovery Today: Ther. Strategies* **2011**, *8*, 61−69.
(54) Taboureau, O.; Nielsen, S. K.; Adouze, K.; Weinhold, N.; Edsgärd, D.; Roque, F. S.; Kouskoumvekaki, I.; Bora, A.; Curpan, R.; Jensen, T. S.; Brunak, S.; Oprea, T. I. ChemProt: A disease chemical biology database. *Nucleic Acids Res.* **2011**, *39*, D367−D372.
(55) Walters, W. P.; Green, J.; Weiss, J. R.; Murcko, M. A. What do medicinal chemists actually make? A 50-year retrospective. *J. Med. Chem.* **2011**, *54*, 6405−6416.
(56) Curtis, C.; et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **2012**, DOI: 10.1038/nature10983.
(57) Agresti, A. An introduction to Categorical Data Analysis. In *Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley: New York, 1996.
(58) Kokko, H. In *Modelling for Field Biologists and Other Interesting People*. Cambridge University Press: Cambridge, UK, 2007.
(59) García-Sosa, A. T.; Mancera, R. L.; Dean, P. M. WaterScore: A novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein−ligand complexes. *J. Mol. Model.* **2003**, *9*, 172−182.
(60) Givechi, A.; Schneider, G. Impact of descriptor scaling on the classification of drugs and nondrugs with artificial neural networks. *J. Mol. Model.* **2004**, *10*, 204−211.
(61) Schmidtke, P.; Barril, X. Understanding and predicting druggablity. A high-throughput method for detection of binding sites. *J. Med. Chem.* **2010**, *53*, 5858−5867.
(62) García-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. *Mol. Inf.* **2012**, *31*, 369−383.
(63) Kirkpatrick, P. Medicinal chemistry: Shades of chemical beauty. *Nat. Rev. Drug Discovery* **2012**, *11*, 107−107.
(64) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90−95.
(65) 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.
(66) Ababou, A.; Ladbury, J. E. Survey of the year 2005: Literature on applications of isothermal titration calorimetry. *J. Mol. Recognit.* **2007**, *20*, 4−14.
(67) 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.
(68) 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.
(69) National Institute of Mental Health (NIMH) Psychoactive Drug Screening Program. PDSP Ki Database. http://pdsp.med.unc.edu/pdspImg.php (accessed December 5, 2010).
(70) *Shanghai Institute of Organic Chemistry*. PDBbind database. http://www.pdbbind-cn.org/ (accessed December 5, 2010).
(71) Wang, R.; Gao, Y.; Lai, L. Calculating partition coefficient by atom-additive method. *Perspect. Drug. Discovery* **2000**, *19*, 47−66.
(72) *Marvin Beans*, version 5.3.8; ChemAxon: Budapest, Hungary, 2010. http://www.chemaxon.com
(73) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997−10002.
(74) Hopkins, A. L.; Groom, C. R. Ligand efficiency: A useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430−431.
(75) García-Sosa, A. T.; Hetényi, C.; Maran, U. Drug efficiency indices for improvement of molecular docking scoring functions. *J. Comput. Chem.* **2010**, *31*, 174−184.
(76) García-Sosa, A. T.; Sild, S.; Takkis, K.; Maran, U. Combined approach using ligand efficiency, cross-docking, and antitarget hits for wild-type and drug-resistant Y181C HIV-1 reverse transcriptase. *J. Chem. Inf. Model.* **2011**, *51*, 2595−2611.
(77) Hetényi, C; Maran, U.; García-Sosa, A. T.; Karelson, M. Structure-based calculation of drug efficiency indices. *Bioinformatics* **2007**, *23*, 2678−2685.
(78) García-Sosa, A. T.; Sild, S.; Maran, U. Docking and virtual screening using distributed grid technology. *QSAR Comb. Sci.* **2009**, *28*, 815−821.
(79) García-Sosa, A. T.; Sild, S.; Maran, U. Design of multi-binding-site inhibitors, ligand efficiency, and consensus screening of avian influenza H5N1 wild-type neuraminidase and of the oseltamivir-resistant H274Y variant. *J. Chem. Inf. Model.* **2008**, *48*, 2074−2080.
(80) *R* Development Core Team. *R*: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria. http://www.R-project.org. (accessed January 1, 2011).
(81) UCLA: Academic Technology Services, Statistical Consulting Group. R: Logit Regression. http://www.ats.ucla.edu/stat/r/dae/logit.htm (accessed April 12, 2011).
(82) Mathews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.
(83) Swamidass, S. J.; Azencott, C.-A.; Daily, K.; Baldi, P. A CROC stronger than ROC: Measuring, visualizing, and optimizing early retrieval. *Bioinformatics* **2010**, *26*, 1348−1356.
(84) Grace Development Team. Grace. http://plasma-gate.weizmann.ac.il/Grace/ (accessed January 1, 2011).
(85) 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.
(86) Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J.-P. Polar molecular surface area as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* **1999**, *16* (10), 1514−1519.
(87) Karelson, M. In *Molecular Descriptors in QSAR/QSPR*. 1st ed.; Wiley Interscience: New York, 2000.
(88) Todeschini, R; Consonni, V. In *Molecular Descriptors for Chemoinformatics*; Mannhold, R.; Kubinyi, H.; Folkers, G., Eds.; Series: Methods and Principles in Medicinal Chemistry; Wiley-VCH: Weinheim, Germany, 2009.

(89) UBM Canon. PharmaLive.com. MedAdNews 200 - World's Best-Selling Medicines, MedAdNews, July 2007. http://www.pharmalive.com/magazines/medad/?date=07%2F2007 (accessed January 1, 2011).
(90) García-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.