

Comparative QSAR- and Fragments Distribution Analysis of Drugs, Druglikes, Metabolic Substances, and Antimicrobial Compounds

Emre Karakoc,[†] S. Cenk Sahinalp,[†] and Artem Cherkasov^{*,‡}

School of Computing Science, Simon Fraser University, Burnaby, Canada, and Division of Infectious Diseases, University of British Columbia, British Columbia, Canada

Received April 25, 2006

A number of binary QSAR models have been developed using methods of artificial neural networks, *k*-nearest neighbors, linear discriminative analysis, and multiple linear regression and have been compared for their ability to recognize five types of chemical compounds that include conventional drugs, inactive druglikes, antimicrobial substituents, and bacterial and human metabolites. Thus, 20 binary classifiers have been created using a variety of ‘inductive’ and traditional 2D QSAR descriptors which allowed up to 99% accurate separation of the studied groups of activities. The comparison of the performance by four computational approaches demonstrated that the neural nets result in generally more accurate predictions, followed closely by *k*-nearest neighbors methods. It has also been demonstrated that complementation of ‘inductive’ descriptors with conventional QSAR parameters does not generally improve the quality of resulting solutions, conforming high predictive ability of ‘inductive’ variables. The conducted comparative QSAR analysis based on a novel linear optimization approach has helped to identify the extent of overlapping between the studied groups of compounds, such as cross-recognition of bacterial metabolites and antimicrobial compounds reflecting their immanent resemblance and similar origin. Human metabolites have been characterized as a very distinctive class of substances, separated from all other groups in the descriptors space and exhibiting different QSAR behavior. The analysis of unique structural fragments and substituents revealed inhomogeneous scale-free organization of human metabolites illustrating the fact that certain molecular scaffolds (such as sugars and nucleotides) may be strongly favored by natural evolution. The established scale-free organization of human metabolites has been contemplated as a factor of their unique positioning in the descriptors space and their distinctive QSAR properties. It is anticipated that the study may bring additional insight into QSAR determinants for conventional drugs, inactive chemicals, and metabolic substances and may help in rationalizing design and discovery of novel antimicrobials and human therapeutics with improved, metabolite-like properties.

INTRODUCTION

Previous studies using ‘inductive’ QSAR descriptors and artificial neural networks (ANN) approach model antimicrobial activity of organic compounds¹ and cationic peptides^{1,2} and steroid-like potential³ and construct predictive binary models of ‘antibiotic-likeness’, ‘drug-likeness’, and ‘bacterial-metabolite-likeness’ (BML).⁴ The latter study developed a QSAR model that recognizes substances involved in bacterial metabolism and demonstrated that the BML model exhibits up to a 30-fold preference toward antimicrobial compounds when compared to other substances illustrating its applicability for ‘in silico’ antibiotic discovery.⁴ Numerous other molecular modeling investigations have also been successfully performed with ‘inductive’ QSAR parameters (see ref 1 for a recent review).

At the same time, all previous studies have been based on ANN-modeling with ‘inductive’ descriptors, while no other machine-learning, statistical, or combinatorial ap-

proaches have been examined, and no other types of QSAR parameters have been investigated. In this work, we attempt to complement ‘inductive’ descriptors with various conventional QSAR descriptors and compare the performance of ANN-based QSAR models with the results produced by *k*-nearest neighbor classification (*k*NN), linear discriminative analysis (LDA), and multiple linear regression (MLR). We also anticipate that such comparative analysis conducted with the combined set of antimicrobial compounds, conventional drugs, druglike substances, and bacterial and human metabolites will allow defining common descriptors-based and structural trends among these important chemical types.

MATERIALS AND METHODS

Molecular Data Sets. The data set of antimicrobial compounds has been assembled from several public resources including *ChemIDPlus* service,⁵ the *Journal of Antibiotics* database,⁶ and from the literature.^{7–9} The conventional drug molecules covering a broad range of nonanti-infective therapeutic activities have all been identified from the *Merck Index Database*.¹⁰

Structures of human metabolites have been obtained from the *Metabolomics* database¹¹ developed and maintained by

* Corresponding author phone: (604)875-4588; fax: (604)875-4013; e-mail: artc@interchange.ubc.ca. Corresponding author address: Division of Infectious Diseases, UBC Faculty of Medicine, 2733 Heather Street, Vancouver, British Columbia V5Z 3J5, Canada.

[†] Simon Fraser University.

[‡] University of British Columbia.

Table 1. 30 ‘Inductive’ and 32 Conventional QSAR Descriptors Used in the Study

descriptor	characterization	descriptor	characterization
EO_Equalized	iteratively equalized electronegativity of a molecule	a_acid	number of acidic atoms
Average_EO_Pos	arithmetic mean of electronegativities of atoms with positive partial charge	a_base	number of basic atoms
Average_EO_Neg	arithmetic mean of electronegativities of atoms with negative partial charge	a_count	number of atoms
Sum_Hardness	sum of hardnesses of atoms of a molecule	a_heavy	number of heavy atoms
Sum_Neg_Hardness	sum of hardnesses of atoms with negative partial charge	a_nN	number of nitrogen atoms
Average_Hardness	arithmetic mean of hardnesses of all atoms of a molecule	a_nO	number of oxygen atoms
Average_Pos_Hardness	arithmetic mean of hardnesses of atoms with positive partial charge	a_nS	number of sulfur atoms
Average_Neg_Hardness	arithmetic mean of hardnesses of atoms with negative partial charge	b_count	number of bonds
Largest_Pos_Hardness	largest atomic hardness among values for positively charged atoms	b_double	number of double bonds
Largest_Neg_Hardness	largest atomic hardness among values for negatively charged atoms	b_rotN	number of rotatable bonds
Hardness_of_Most_Pos	atomic hardness of an atom with the most positive charge	b_rotR	fraction of rotatable bonds
Hardness_of_Most_Neg	atomic hardness of an atom with the most negative charge	b_triple	number of triple bonds
Global_Softness	molecular softness – sum of constituent atomic softnesses	chiral	number of chiral centers
Total_Neg_Softness	sum of softnesses of atoms with negative partial charge	density	mass density (AMU/Å ³)
Average_Softness	arithmetic mean of softnesses of all atoms of a molecule	FCharge	sum of formal charges
Largest_Neg_Softness	largest atomic softness among values for positively charged atoms	KierFlex	molecular flexibility
Softness_of_Most_Pos	atomic softness of an atom with the most positive charge	lip_acc	Lipinski acceptor count
Total_Charge_Formal	sum of charges on all atoms of a molecule (formal charge of a molecule)	lip_don	Lipinski donor count
Average_Pos_Charge	arithmetic mean of positive partial charges on atoms of a molecule	logP(o/w)	log octanol/water partition coefficient
Average_Neg_Charge	arithmetic mean of negative partial charges on atoms of a molecule	mr	molar refractivity
Most_Pos_Charge	largest partial charge among values for positively charged atoms	PC+	total positive partial charge
Most_Neg_Charge	largest partial charge among values for negatively charged atoms	PC-	total negative partial charge
Most_Pos_Sigma_mol_i	largest positive group inductive parameter σ^* (molecule→atom) for atoms in a molecule	rings	number of rings
Most_Neg_Sigma_mol_i	largest (by absolute value) negative group inductive parameter σ^* (molecule→atom) for atoms in a molecule	SMR	molar refractivity
Most_Pos_Sigma_i_mol	largest positive atomic inductive parameter σ^* (atom→molecule) for atoms in a molecule	vdw_area	van der Waals surface area (Å ²)
Most_Neg_Sigma_i_mol	largest negative atomic inductive parameter σ^* (atom→molecule) for atoms in a molecule	vsa_acc	VDW acceptor surface area (Å ²)
Sum_Neg_Sigma_mol_i	sum of all negative group inductive parameters σ^* (molecule→atom) within a molecule	vsa_acid	VDW acidic surface area (Å ²)
Largest_Rs_i_mol	largest value of atomic steric influence R_s (atom→molecule) in a molecule	vsa_base	VDW basic surface area (Å ²)
Most_Neg_Rs_mol_i	steric influence R_s (molecule→atom) ON the most negatively charged atom in a molecule	vsa_don	VDW donor surface area (Å ²)
Most_Pos_Rs_i_mol	steric influence R_s (atom→molecule) by the most positively charged atom ONTO the rest of the molecule	vsa_hyd	VDW hydrophobe surface area (Å ²)
apol	sum of atomic polarizabilities	weight	molecular weight

the Wishart group at the University of Alberta; structures of bacterial metabolites have been obtained from the *Analyti-Con-Discovery* company.¹²

Druglike substances used in the study have been randomly selected from the *Assinex Gold* collection¹³ using the expanded druglike criteria: number of H-bond acceptors between 1 and 10; number of H-bond donors between 1 and 5; molecular weight between 200 and 500 Daltons; number of rotating bonds below 12; hydrophobicity in the range 1–7; and the total polar surface area below 140 Å².

The redundancy of the resulting data set containing antimicrobials, drugs, druglike substances, and human and bacterial metabolites has been ensured through the SMILES records. All duplicate entries have been removed; all organometallic structures as well as inorganic components have also been eliminated. All molecules containing basic and/or acidic groups have been converted into un-ionized form.

The resulting final data set included 520 antimicrobial compounds, 959 general drugs, 1202 druglike chemicals, 562 bacterial metabolites, and 1104 human metabolites. The corresponding SMILES records for nonmetabolic substances can be found in the Supporting Information. All molecular structures have been further optimized with the MMFF94

force-field¹⁴ as it is implemented within the MOE modeling package.¹⁵

Descriptors. The optimized structures of 4346 compounds have been used for calculating ‘inductive’ QSAR descriptors used in our previous studies^{1–4} and about 80 conventional QSAR parameters implemented within the MOE-QSAR module.¹⁵ We have chosen only those descriptors that can be rapidly calculated for large molecular data sets (see Table 1). The ‘inductive’ QSAR descriptors have been calculated by the custom SVL scripts that can be freely downloaded through the SVL exchange.¹⁶

To eliminate possible cross-correlation between the independent variables we removed all sets of descriptors that cross-correlated with $R > 0.9$. As a result 30 ‘inductive’ parameters and 32 conventional QSAR descriptors have been selected for modeling. Thus, the final set of 62 descriptors (described in Table 1) has been calculated for 4346 compounds under investigation, and the descriptors values have been normalized within [0÷1] range. The normalized values have then been used to generate QSAR models distinguishing all five types of chemical substances under study.

Principal Component Analysis. Our optimal QSAR models for each biological activity contains up to 60 descriptors, which cannot be visually displayed by 3-D

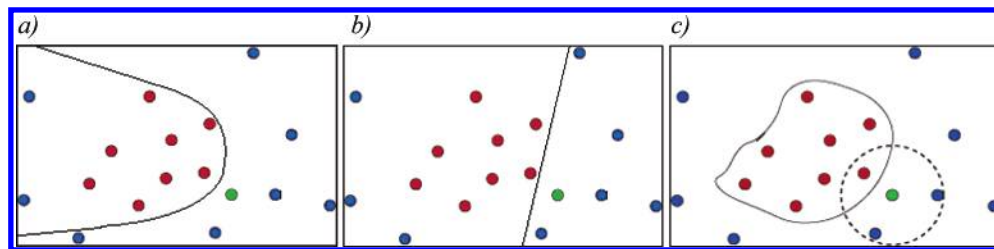


Figure 1. Projection of unknown compound (green point) onto chemical space where active compounds (red points) have been separated from inactive ones (blue) using the ANN approach (a), the LDA method (b), and the *k*NN algorithm (c).

graphics. It is possible to use the top 3 most relevant descriptors for displaying the data in 3-D Euclidean space, but this results in loss of accuracy. To reduce the loss of accuracy, it is possible to use principal component analysis (PCA), which maps the data elements under the QSAR parameters to 3-D Euclidean space with a small distortion as follows. PCA (as implemented by the MOE package¹⁵) first subtracts from each descriptor of every data element, the mean value of that descriptor. On this “mean corrected” data set, PCA calculates the covariance matrix, where the entry (*i,j*) represents the covariance between descriptors *i* and *j*. PCA then computes the eigenvalues and eigenvectors of the covariance matrix. The eigenvectors that correspond to the 3 highest valued eigenvalues are called the principal components of the covariance matrix. PCA completes the mapping of the original descriptor space into the target 3-D Euclidean space by setting the reference coordinates of the target space as the principal components of the covariance matrix.

Artificial Neural Networks. To relate QSAR descriptors of the studied compounds to the Boolean (0|1) indicators of their association with antibiotic, drug, druglike, and metabolite groups we employed the method of artificial neural networks (ANN) as it is implemented by the WEKA open-source package.¹⁷

ANN is known to be one of the most effective classification techniques and has become an essential part of the QSAR research.¹⁸ ANN define the relationship between *n* input variables $input_node_{ij}$ and a dependent parameter's value $output_node_i$ by recursive adjustments of the weights attributes w_{ij} assigned to each network node. In particular, a set of inputs multiplied by each neuron's weights are summed up for each of the *m* hidden neurons:

$$hidden_node_i = \tanh \left[\sum_{j=1}^n (input_node_j * w_{ij} + const_0 * w_0) \right]$$

Then, the transformed sums for the hidden units are multiplied by the output weights

$$output_node = \sum_{i=1}^m (hidden_node_i * w_{ij} + const_0 * w_0)$$

where they are summed a final time, transformed with the learning function $1/(1 + e^{-x})$, and interpreted. The ANN approach is capable of fitting complex training patterns of QSAR parameters in nonlinear fashion and allows predicting ANN outputs (activity) of unknown entries, as long as their input patterns are contained within the range of descriptor values used in the training phase (see Figure 1a).

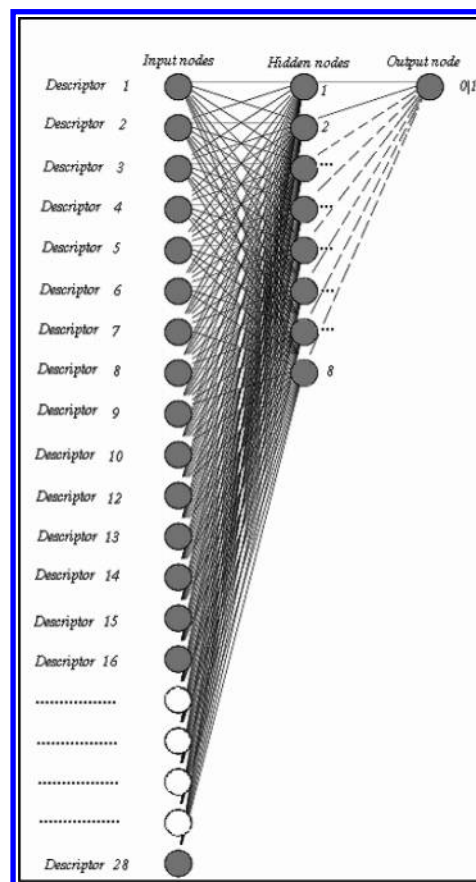


Figure 2. Configuration of artificial neural network used to develop binary QSAR models for drug, druglike compounds, antimicrobials, and bacterial and human metabolites.

The ANN configuration we utilized for the current study consisted of 41 input nodes, 1 hidden layer with 10 nodes, and 1 output node (Figure 2). Input descriptors normalized to the range [0.0÷1.0] have been randomly separated into nonoverlapping training and testing sets (70%/30% split) keeping a constant ratio between active and inactive substances. Separation of the input patterns into the training and testing groups has been done to avoid their overfitting the networks.

All neural networks have been trained with the standard back-propagation algorithm, input shuffling, weight decay and using learning rate and threshold values set to 0.80 and 0.10, respectively. The initial network weights have been randomly assigned in a range of [−1.0÷1.0]. Typically, 20 independent training and testing runs have been conducted for each ANN, and the resulting training/testing statistics have been reported for the averaged network outputs. More details on networks training, validation, and performance will be discussed in the following sections.

Linear Discriminant Analysis. Linear discriminant analysis (LDA) is another technique broadly used in QSAR research for data classification. The objective function of this method aims to maximize the ratio of between-class variance and within-class variance using linear projection of high-dimensional data. The general strategy of the linear discriminant analysis on the set of chemical structures with the assigned Boolean activity measures (0|1) can be described as follows.

First, LDA computes the mean value of each descriptor for both active (μ_A) and inactive compounds (μ_I) and then computes a general mean value $\mu_G = p_A \bullet \mu_A + p_I \bullet \mu_I$ where p_A represents the probability of the active compounds and p_I represents the probability of the inactive compounds observed in the training data. LDA then computes the covariance matrices

$$\text{cov}_A = (x_A - \mu_A)(x_A - \mu_A)^T$$

$$\text{cov}_I = (x_I - \mu_I)(x_I - \mu_I)^T$$

where $x_{A/I}$ represents the descriptor matrix of the active/inactive compounds. Following this, LDA computes the “within-class scatter”, which is the expected covariance of each of the classes which can be defined as

$$S_W = p_A \times (\text{cov}_A) + p_I \times (\text{cov}_I)$$

LDA also computes the “between-class scatter”, which can be defined as

$$S_B = (\mu_A - \mu_G)(\mu_A - \mu_G)^T + (\mu_I - \mu_G)(\mu_I - \mu_G)^T$$

LDA tries to maximize the ratio of between-class scatter and within-class scatter by finding a transformation matrix that maps the original data set into a simpler 2-D or 3-D space where it is possible to separate the active and inactive compounds using, respectively, a line or a plane (Figure 1b). The transformation matrix is calculated through the use of nonzero eigenvectors of the matrix $\text{inv}(S_W) \times S_B$ (for details see ref 24). As a result, given a query compound with unknown activity, it is possible to predict its class by checking to which side of the line/plane its projection (under the above transformation matrix) falls into. Clearly, this can be performed very fast.

Multiple Linear Regression. Multiple linear regression (MLR) is a widely used classification technique that quantifies the activity level of a compound based on a linear function of the descriptors. More specifically, given a data set X of m compounds with n descriptors each, MLR estimates the activity level of a compound X_i as follows

$$\text{activity}(X_i) = c + \sigma_1 \bullet X_i[1] + \sigma_2 \bullet X_i[2] + \dots + \sigma_n \bullet X_i[n]$$

where c is a constant.

If $\text{activity}(X) \geq t$ for a user defined threshold value t , then it is likely that the molecule is active with respect to the bioactivity of interest. Thus the MLR classifier is described by a planar separator in the multidimensional descriptor array space.

There are several different optimization criteria for determining the best fitting MLR solution; among them, partial-least-squares criteria is possibly the most commonly used.

This approach suggests minimizing the sum of squares of the deviations between estimated and actual activity levels; more formally it suggests minimizing

$$S = \sum_{i=1}^m (A(X_i) - c + \sigma_1 \bullet X_i[1] + \sigma_2 \bullet X_i[2] + \dots + \sigma_n \bullet X_i[n])^2$$

Although it is computationally difficult to minimize the value of S , approximate solutions are possible through the use of genetic algorithms, local search heuristics, etc.

k-Nearest Neighbors Approach. The k -nearest neighbors (k NN) classification method requires the definition of distance $D(S, R)$ between any pair of molecules S and R in d -dimensional descriptors space. According to QSAR formalism, such a distance measure should reflect functional association and/or chemical similarity between the molecules. Thus, the k NN approach allows descriptors-based clustering of chemical compounds according to already known biological activity and can be used to classify an untested chemical substance by its proximity to established clusters. The notion of a distance should be a metric for various reasons; i.e., it should satisfy the following three properties. $D(S, S) = 0$, a point has distance 0 to itself; $D(S, R) = D(R, S)$, distance is symmetric; and $D(S, R) \leq D(S, Q) + D(Q, R)$, distance satisfies the triangle inequality. The distance measures satisfying the above conditions include the Hamming distance (i.e. L_1)

$$\sum_{h=1}^d |S_h - R_h|$$

the Euclidean distance (i.e. L_2)

$$\sqrt{\sum_{h=1}^d (S_h - R_h)^2}$$

and the maximum of dimensions (i.e. L_∞): $\max_{1 \leq h \leq d} |S_h - R_h|$ among others.

In the current work we utilized the weighted Hamming distance

$$\sum_{h=1}^d \sigma_h |S_h - R_h| \quad (\forall \sigma_i, \sigma_i \geq 0)$$

that allows differentiating relevance of various QSAR descriptors for a given activity. Weighted Hamming distance representations allow the establishing of optimal σ_i values that maximize separation of active $T^A = \{T^A_1, T^A_2 \dots T^A_m\}$ and inactive $T^I = \{T^I_1, T^I_2 \dots T^I_n\}$ elements in the training set: $T = T^A \gg T^I$.

We utilized the linear programming approach to minimize the function

$$\begin{aligned} f(T) = & \left(\sum_{i=1}^m \sum_{j=1}^m \sum_{h=1}^d \sigma_h \bullet |T^A_i[h] - T^A_j[h]| \right) / m^2 + \\ & \left(\sum_{i=1}^n \sum_{j=1}^n \sum_{h=1}^d \sigma_h \bullet |T^I_i[h] - T^I_j[h]| \right) / n^2 - \\ & \left(\sum_{i=1}^m \sum_{j=1}^n \sum_{h=1}^d \sigma_h \bullet |T^A_i[h] - T^I_j[h]| \right) / m \cdot n \end{aligned}$$

such that the following three conditions are satisfied:

$$(1) \forall T_i^A \in T^A (\sum_{j=1}^m \sum_{d=1}^h \sigma_{h^*} |T_i^A[h] - T_j^A[h]|) / m^2 \leq (\sum_{j=1}^n \sum_{d=1}^h \sigma_{h^*} |T_i^A[h] - T_j^I[h]|) / m \cdot n$$

$$(2) \forall i0 \leq \sigma_i \leq 1$$

$$(3) \sum_{i=1}^d \sigma_i \leq C$$

where C is a used-defined constant.

More details on the adopted k NN procedure can be found in ref 19. It should be outlined that the described mathematical procedure not only maximizes distances between active and inactive elements of the training set but also aims to minimize within-the-class average distances and, therefore, tends to condense activity clusters.

The defined clusters of biologically active and inactive molecules in d -dimensions can then be used to characterize unknown entries (chemicals) by projecting their QSAR parameters into descriptors space (Figure 1c). In particular, a new compound (green point in Figure 1) can be associated with a certain predefined activity cluster by considering affiliations of its k -nearest neighbors. In the current study we considered 3 neighboring points and have been assigning a tested compound to a certain cluster, if ≥ 2 of its neighbors belong to the cluster. Implementation of the above-described k NN method has been done within the open-source linear programming solver CPLEX.²⁰

RESULTS AND DISCUSSION

We tested applicability of the above-described combinatorial, statistical, and machine-learning approaches for creating binary QSAR classifiers that operate by 'inductive' and conventional 2D QSAR variables. The combined molecular data set consisting of 1202 druglike chemicals, 959 drugs, 520 antimicrobials, 562 bacterial metabolites, and 1404 human metabolites (with 62 normalized QSAR descriptors assigned to each entry) has been used to create QSAR models based on ANN, k NN, LDA, and MLR approaches and distinguishing the following activity categories:

(a) '*Antimicrobials vs Others*'. To study this system, 520 antimicrobial compounds have been assigned 1.0 activity value, and the remaining 3826 general drugs, druglikes, and metabolites all have been considered as a negative control and assigned to a null dependent variable.

(b) '*Drugs vs Others*'. In this case, four QSAR models based on ANN, k NN, LDA, and MLR have been trained to separate conventional drugs from the rest of the compounds. We did not include antimicrobials in the general drug category, as our previous QSAR investigations demonstrated distinctive differences between these two classes. Thus, to develop the binary classifiers specifically recognizing 959 general drugs among the studied compounds, we assigned formers to 1.0 activity and considered 3387 other chemicals as inactive, with assigned null dependent variables.

(c) '*Inactive Druglikes vs Others*'. A group of 1202 chemicals from the Assinex Gold collection that justifies the expanded druglike criteria has been considered as presumably inactive chemicals that do not possess any therapeutic

activities and/or metabolite-like properties. Thus, by referring to this group of substances we do not imply their exclusive 'druglike' potential but their presumed inactivity in four other classes of substances. Understandably, the majority of the studied drugs (97%), antimicrobials (84%), and bacterial (89%) and human metabolites (69%) also justify the original Lipinski rule. Thus, when we trained the corresponding ANN, k NN, LDA, and MLR models distinguishing 1202 'druglike' chemicals from 3144 drugs, antimicrobials, and metabolites composing the negative control set, we anticipated that the resulting QSAR binary classifiers captured specific structural features of biologically inactive molecules.

(d) '*Bacterial Metabolites vs Others*'. Similarly to our previous QSAR work on bacterial metabolites,⁴ we considered a set of 562 compounds characterized in bacteria isolates by the *AnalytiCon-Discovery* Company.¹⁰ In the current work we expanded the 'bacterial-metabolite-likeness' (BML) model by including structures of human metabolites in a negative control to ensure that the resulting QSAR approach is not simply biased toward metabolic substances. This is an important development as we have previously established a definite BML character of antimicrobial compounds and speculated that enhancing the 'bacterial-metabolite-like' potential of antibiotic candidates may improve their bio-availability, microbial uptake, etc.

To develop the expanded BML models based on four mathematical approaches under study, we assigned 1.0 dependent variable to 562 *AnalytiCon-Discovery* substances and treated them as actives. All other 3784 molecules have been assigned null activities as four QSAR models attempted separating them from bacterial metabolites.

(e) '*Human Metabolites vs Others*'. The data set of >1000 chemical substances involved in chemical reactions taking place in the human body have recently been catalogued by the group of Professor Wishard at the University of Alberta. These molecules have been incorporated into the larger metabolomics database and have been made available through the Web: <http://www.metabolomics.ca/>. As it has been mentioned above, we decided to include these compounds in a negative control of the BML models to ensure that the developed QSAR classifiers recognize bacterial metabolites specifically, rather than metabolites in general. In addition, we attempted developing 'human-metabolite-likeness' (HML) models hoping that the corresponding QSAR classifiers may become useful tools for assessing potential human therapeutics. We used the methods of ANN, k NN, LDA, and MLR to recognize 1104 human metabolites among 4346 compounds under study.

QSAR Modeling. All five classification systems (a)–(e) have been investigated using 70–30% separation of 4346 substances into training and testing sets that contained proportional fractions of active and inactive entries. Sixty-two normalized descriptors that consist of 30 'inductive' parameters and 32 conventional QSAR descriptors are used for training our models (the details of the resulting solutions can be obtained from the authors upon request). Correlation of the descriptors has no effect on the solutions of the combinatorial and statistical QSAR methods (LDA, MLR, k NN) where it is taken care of during the solution process as described in the Methods section. However artificial neural networks are based on certain heuristics which relies on a stricter elimination of the correlation between the descriptors.

Table 2. Statistics of the Developed QSAR Models for Antimicrobials, Drugs, Druglikes, and Bacterial and Human Metabolites

method	validation	true posit	true negat	false posit	false negat	spec	sens	accur	PPV	NPV
Antibacterials vs (Drugs + Druglikes + Bacteria Metabolites + Human Metabolites)										
kNN	training 70%	269	2610	9	95	0.97	0.74	0.95	0.80	0.96
	testing 30%	117	1119	8	39	0.98	0.75	0.95	0.81	0.97
	LOO	400	3727	9	120	0.97	0.77	0.95	0.80	0.97
LDA	training 70%	364	0	2679	0	0.00	1.00	0.12	0.12	0.00
	testing 30%	156	0	1147	0	0.00	1.00	0.12	0.12	0.00
	LOO	261	3751	75	259	0.98	0.50	0.92	0.78	0.94
MLR	training 70%	194	564	2115	170	0.21	0.53	0.25	0.08	0.77
	testing 30%	61	1129	18	95	0.98	0.39	0.91	0.77	0.92
	LOO	279	3726	100	241	0.97	0.54	0.92	0.74	0.94
ANN	training 70%	294	2651	27	70	0.99	0.81	0.97	0.92	0.97
	testing 30%	129	1132	16	27	0.99	0.83	0.97	0.89	0.98
	LOO	449	3821	5	71	0.99	0.86	0.98	0.99	0.98
Bacteria Metabolites vs (Drugs + Druglikes + Antibacterials + Human Metabolites)										
kNN	training 70%	311	2537	112	83	0.96	0.79	0.94	0.74	0.97
	testing 30%	135	1091	44	33	0.96	0.80	0.94	0.75	0.97
	LOO	455	3637	147	107	0.96	0.81	0.94	0.76	0.97
LDA	training 70%	240	2587	62	154	0.98	0.61	0.93	0.79	0.94
	testing 30%	90	1088	47	78	0.96	0.54	0.90	0.66	0.93
	LOO	336	3665	119	226	0.97	0.60	0.92	0.74	0.94
MLR	training 70%	301	2525	124	93	0.95	0.76	0.93	0.71	0.96
	testing 30%	119	1073	62	49	0.95	0.71	0.91	0.66	0.96
	LOO	406	3603	181	156	0.95	0.72	0.92	0.69	0.96
ANN	training 70%	338	2597	52	55	0.98	0.86	0.96	0.87	0.98
	testing 30%	159	1076	59	10	0.95	0.94	0.95	0.73	0.99
	LOO	534	3780	4	28	0.99	0.95	0.99	0.99	0.99
Drugs vs (Bacteria Metabolites + Druglikes + Antibacterials + Human Metabolites)										
kNN	training 70%	474	2158	214	197	0.91	0.71	0.86	0.69	0.92
	testing 30%	204	928	88	83	0.91	0.71	0.87	0.70	0.92
	LOO	694	3111	277	264	0.92	0.72	0.88	0.71	0.92
LDA	training 70%	0	2372	0	671	1.00	0.00	0.78	0.00	0.78
	testing 30%	0	1014	2	287	0.99	0.00	0.78	0.00	0.78
	LOO	393	3206	182	565	0.95	0.41	0.83	0.68	0.85
MLR	training 70%	279	2234	138	392	0.94	0.42	0.83	0.67	0.85
	testing 30%	109	951	65	178	0.94	0.38	0.81	0.63	0.84
	LOO	329	3253	135	629	0.96	0.34	0.82	0.71	0.84
ANN	training 70%	489	2178	194	182	0.92	0.73	0.88	0.72	0.92
	testing 30%	177	978	39	110	0.96	0.62	0.89	0.82	0.90
	LOO	879	3387	1	79	0.99	0.92	0.98	0.99	0.98
Druglikes vs (Bacteria Metabolites + Drugs + Antibacterials + Human Metabolites)										
kNN	training 70%	674	2043	158	168	0.93	0.80	0.89	0.81	0.92
	testing 30%	281	866	77	79	0.92	0.78	0.88	0.78	0.92
	LOO	957	2935	209	245	0.93	0.80	0.90	0.82	0.92
LDA	training 70%	683	1917	284	159	0.87	0.81	0.85	0.71	0.92
	testing 30%	295	801	142	65	0.85	0.82	0.84	0.68	0.92
	LOO	887	2834	310	315	0.90	0.74	0.86	0.74	0.90
MLR	training 70%	665	1951	250	177	0.89	0.79	0.86	0.73	0.92
	testing 30%	282	812	131	78	0.86	0.78	0.84	0.68	0.91
	LOO	849	2874	270	353	0.91	0.71	0.86	0.76	0.89
ANN	training 70%	734	2086	114	107	0.95	0.87	0.93	0.87	0.95
	testing 30%	334	891	52	27	0.94	0.93	0.94	0.87	0.97
	LOO	1170	3138	5	32	0.99	0.97	0.99	0.99	0.99
Human Metabolites vs (Bacteria Metabolites + Drugs + Antibacterials + Druglikes)										
kNN	training 70%	773	2270	0	0	1.00	1.00	1.00	1.00	1.00
	testing 30%	331	972	0	0	1.00	1.00	1.00	1.00	1.00
	LOO	1104	3242	0	0	1.00	1.00	1.00	1.00	1.00
LDA	training 70%	773	2270	0	0	1.00	1.00	1.00	1.00	1.00
	testing 30%	331	972	0	0	1.00	1.00	1.00	1.00	1.00
	LOO	1104	3242	0	0	1.00	1.00	1.00	1.00	1.00
MLR	training 70%	773	2270	0	0	1.00	1.00	1.00	1.00	1.00
	testing 30%	331	972	0	0	1.00	1.00	1.00	1.00	1.00
	LOO	1104	3242	0	0	1.00	1.00	1.00	1.00	1.00
ANN	training 70%	773	2270	0	0	1.00	1.00	1.00	1.00	1.00
	testing 30%	331	972	0	0	1.00	1.00	1.00	1.00	1.00
	LOO	1104	3242	0	0	1.00	1.00	1.00	1.00	1.00

For training ANN methods, the number of the descriptors is reduced to 41 after further applying more stringent criteria for the cross-correlation among the descriptors using a simple route of the ANN solver.

Five classification systems (a)–(e) have been independently processed by four statistical and machine-learning approaches, and the developed binary solutions have been collected in the Supporting Information. The performance

by the developed 20 QSAR models has been assessed by true/false positive/negative predictions produced during the corresponding training and testing phases. [The parameters of the developed QSAR models can be obtained directly from authors upon request.] The outputs from the ANN-, kNN, LDA, and MLR-based binary QSAR classifiers have been interpreted using 0.5 threshold which allowed computing the corresponding sensitivity, specificity, accuracy, and positive

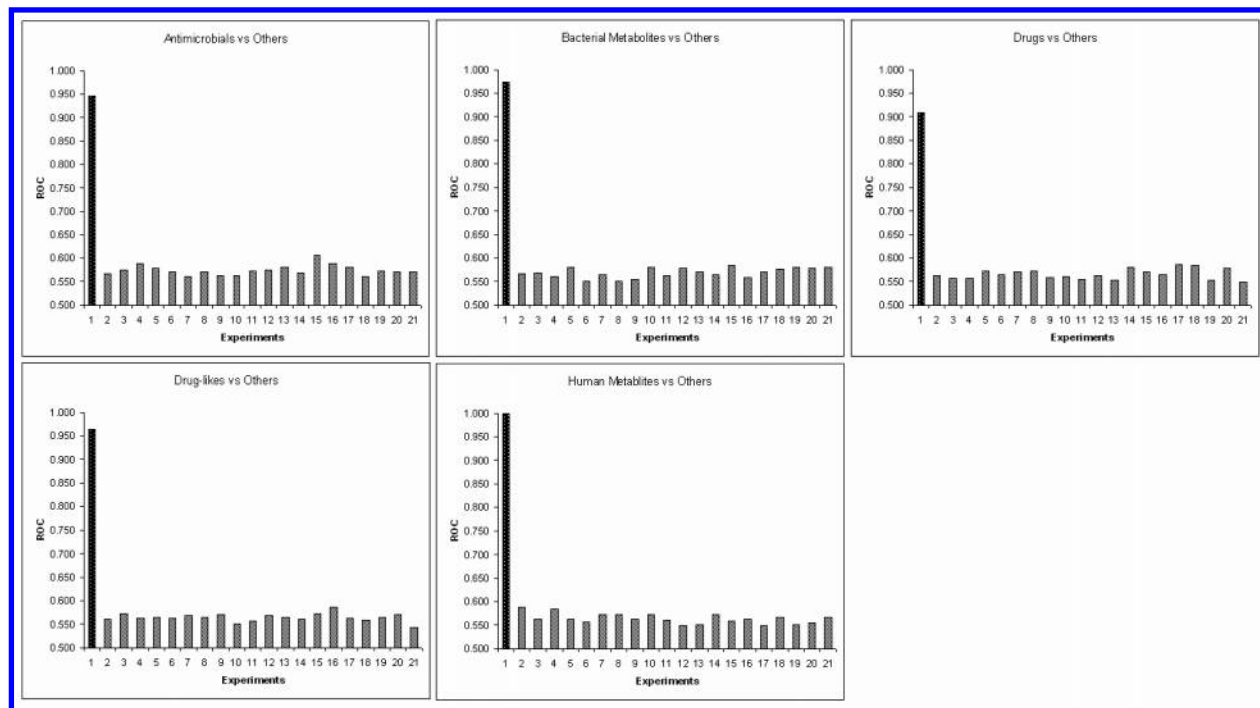


Figure 3. ROC parameters computed for 5 ANN-based QSAR models that have been created on the original molecular sets (left columns on the histograms) and using the ‘scrambled’ data sets where activity values have been assigned to the molecules in a random manner (all other columns corresponding to 20 independent runs).

predictive value and negative predictive value parameters that are all collected in Table 2.

Data in Table 2 illustrate that the method of neural nets allowed generally better separation of actives and inactives in all five systems (a)–(e). The results also demonstrate that complementation of ‘inductive’ parameters with conventional 2D-QSAR descriptors did not significantly improve the accuracy of predictions. Thus, our earlier ANN-based QSAR model utilizing only 28 ‘inductive’ parameters resulted in a 95% and 92% accurate prediction of antimicrobials in the training and testing sets, respectively, while the reported ANN classification of antimicrobials with 41 ‘inductive’ and conventional QSAR descriptors allowed 97% prediction accuracy.

These results confirm high predictive power of ‘inductive’ QSAR descriptors^{1–4} that has been previously attributed to the fact that they cover a broad range of properties of bound atoms and molecules related to their size, polarizability, electronegativity, and electronic and steric interactions and, thus, can adequately capture structural determinants of intra- and intermolecular interactions.

LOO Cross-Validation of the Developed QSAR Models.

To rigorously validate the predictive ability of the developed QSAR approaches and to confirm accuracy of the reported results we have also conducted leave-one-out (LOO) analysis of 20 developed QSAR models. In particular, the above-described five data sets containing active and inactive compounds have been processed with ANN, *k*NN, LDA, and MLR approaches, where each model training cycle has been conducted 4346 times, with 4345 entries in the training set and the remaining molecule used for testing. The resulting 4346 predicted values for all 20 QSAR models have then been transformed into the corresponding LOO confusion matrices using 0.5 output thresholds. The resulting statistic parameters have also been collected into Table 2.

As data in Table 2 illustrate the LOO statistics can generally reproduce the results of QSAR modeling with 70–30% data separation. Thus, the overall superior performance by the ANN approach has been confirmed for all (a)–(e) systems.

The predictions by the *k*NN algorithm also appeared as consistently accurate and comparable with ANN results. The linear discriminative analysis failed to separate antimicrobial substances from the rest of the compounds and did not result in good predictions for conventional drugs. Linear fitting of data with the multiple linear regression also did not adequately perform for those groups.

To illustrate the adequacy and nonrandom character of the developed QSAR models even further, we have applied the ANN approach to all five data sets under study when the corresponding active/inactive Boolean dependent parameters have been assigned to molecules in a random manner (keeping the numbers of ones and zeros the same as in the actual data sets). Thus, we attempted to train ANN-based QSAR models on the noise-type data and assessed the accuracy of the resulting approaches by computing the corresponding ROC values. Figure 3 features ROC parameters for ANN-based solutions (presented as histograms) created in 20 independent runs and using the settings of the original QSAR models for five groups of activities and ‘scrambled’ activity values. As Figure 3 illustrates no meaningful results could be obtained on the ‘scrambled’ data set as the corresponding ROC parameters all fit the ‘random’ range of 0.5–0.6 values, while the native QSAR models result in 0.9–1.0 ROC values (presented as the first bar of the histograms). Thus, the results obtained on the scrambled sets of the studied compounds clearly illustrate that the ANN approach did not just fit the noise signals (that are not uncommon for some 2D descriptors) into the QSAR models.

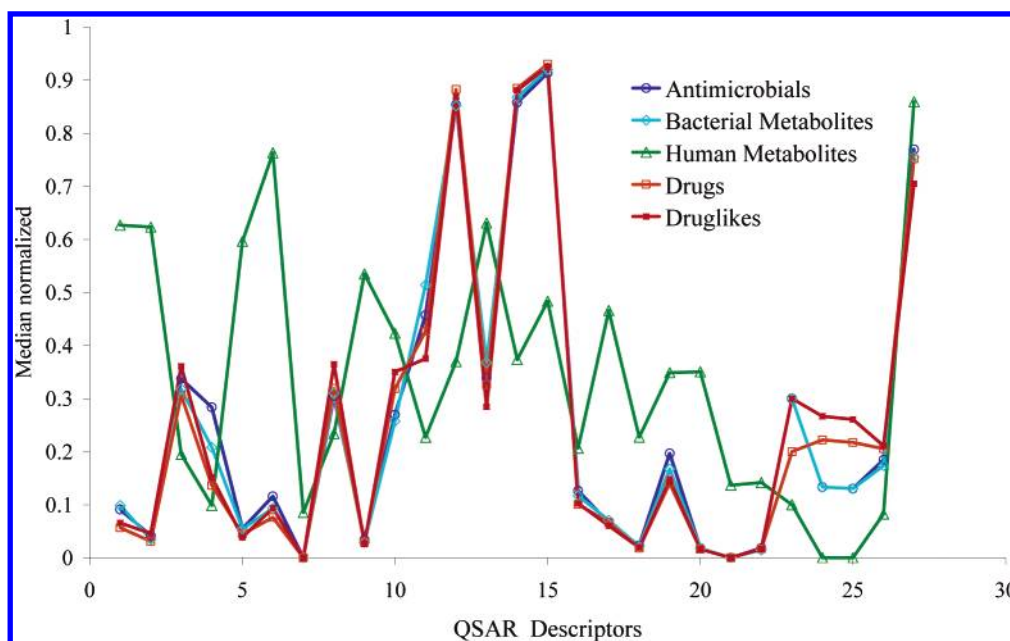


Figure 4. Median values for selected 'inductive' and conventional QSAR descriptors (normalized) calculated independently within studied sets of chemical substances.

Table 3. Origin of the False Positive Predictions by ANN Models

testing models	true positive/no. of active compds (in the testing model)				
	AB	BM	HM	D	DL
AB		0.40	0.00	0.59	0.32
BM	0.56		0.00	0.78	0.27
HM	0.59	0.28			0.51
D	0.05	0.33	0.00		0.87
DL	0.00	0.03	0.00	0.78	

It is also notable, that all four mathematical algorithms under study could recognize human metabolites from the rest of the compounds with remarkable 100% accuracy. As Table 2 indicates, no false positive/negative predictions have been generated by the 'human-metabolite-likeness' models based on ANN, *k*NN, LDA, and MLR approaches. This positions human metabolites as a rather unique group of entries in the descriptors space and raises questions regarding the nature of structural determinants of the chemicals involved in human metabolism.

Cross-Recognition between Drugs, Nondrugs, Antimicrobials, and Metabolites. With the exception of those models that have been built for classification of human metabolites, all other developed QSAR approaches produced a nondismissible number of false positive predictions (see Table 2) determined by overlaps between the studied groups of compounds. Thus, the ANN-based separation of antimicrobials from the rest of the compounds produced a total of 43 false positive predictions (27 during the training phase and 16 during the testing phase), the majority of which correspond to bacterial metabolites that tend to be cross-recognized by the 'antimicrobial-likeness' models.

To investigate the cross-recognition between the developed models even further, we retrained ANN models (a)–(e) leaving one of the activity groups out of consideration and then applied the developed models to the excluded set. The resulting numbers of positive predictions have been collected into Table 3 and transformed into the corresponding fractions of antimicrobials, drugs, druglikes, and bacterial and human

metabolites that have been recognized by the 'nonself' QSAR models. These numbers also reflect a profound similarity between drugs and druglike substances as well as bacterial metabolites and antimicrobials. Thus, the binary classifier trained to recognize inactive druglike substances (in the absence of drugs) could recognize 87% of conventional human therapeutics, while the drugs-trained model treated 78% of druglike substances as potentially active. The cross-recognition between antimicrobials and bacterial metabolites has been established as high as 40–56%, which confirmed previously suggested immanent similarity between these groups of chemicals.⁴

The results of cross-recognition analysis also confirmed uncharacteristic QSAR behavior of human metabolites: the ANN model trained to recognize them in the mixed set of compounds did not produce any false positive predictions. The latter may reflect the fact that QSAR descriptors computed for human metabolites follow different trends when compared to drugs, inactive chemicals, antimicrobials, and bacterial metabolites. To illustrate this point we plotted median and mean values of 'inductive' and 2D-QSAR descriptors that have been computed independently for the studied groups of chemicals (Figures 4 and 5). The charts clearly demonstrate that descriptors computed for human metabolites appear differently.

Separation of Drugs, Nondrugs, Antimicrobials, and Metabolites in Descriptor Space. To gain a better understanding of the distinctive behavior of human metabolites and their positioning in the descriptors space we considered a data set of >2 million druglike chemical structures downloaded from the ZINC database.²² For every substance in that data set we calculated 62 'inductive' and 2D-QSAR descriptors selected for modeling and assumed that such large data set should sufficiently cover all feasible values of QSAR parameters. We conducted the principle component analysis on 62 descriptors calculated for 2 066 905 ZINC entries and 4346 studied antimicrobials, drugs, druglikes, and metabolites and defined three major principal components. The values

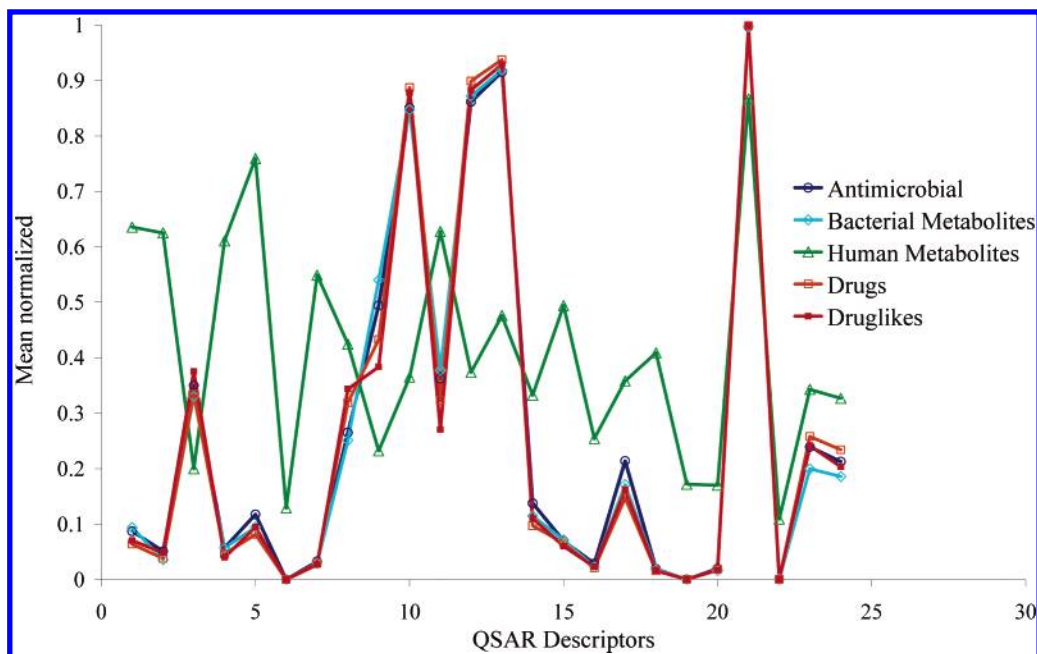


Figure 5. Averaged values of selected 'inductive' and conventional QSAR descriptors (normalized) calculated independently within studied sets of chemical substances.

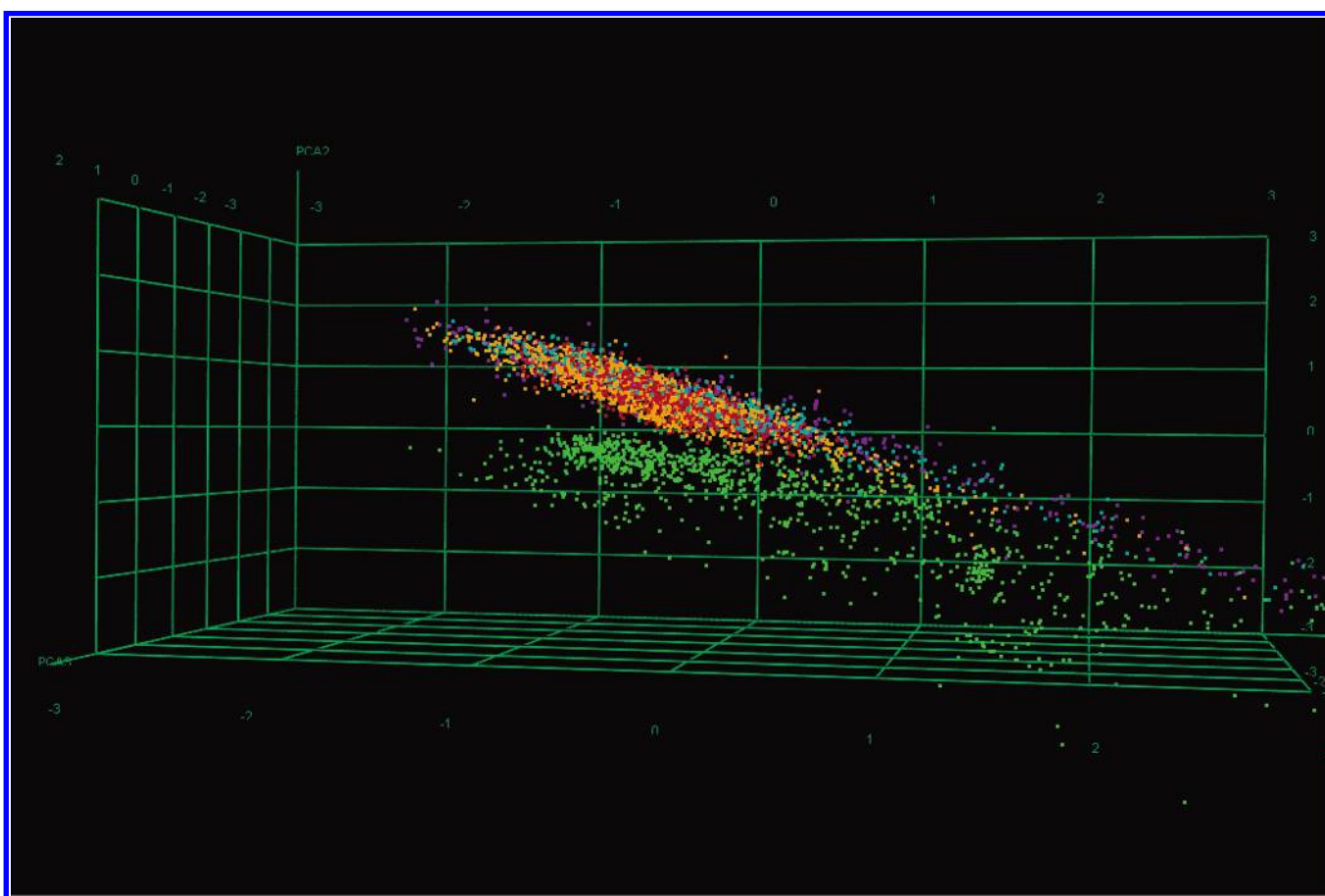


Figure 6. Separation of five groups of the studied compounds in three-dimensional space formed by three principal components derived from 62 QSAR descriptors use in the study. The color coding of points corresponds to the following scheme: red, druglikes; orange, drugs; purple, antimicrobial; cyan, bacterial metabolites; and green, human metabolites.

of the established principal components then were plotted on the three-dimensional chart (Figure 6). Thus, Figure 5 features positioning of the studied compounds (color-coded by their activity type) in descriptors space encompassing the

broadest range of chemical substances. The spatial positioning of human metabolites (green dots) in the descriptors space can be characterized by their distinctive clustering relative to antimicrobial compounds (marked in purple), conventional

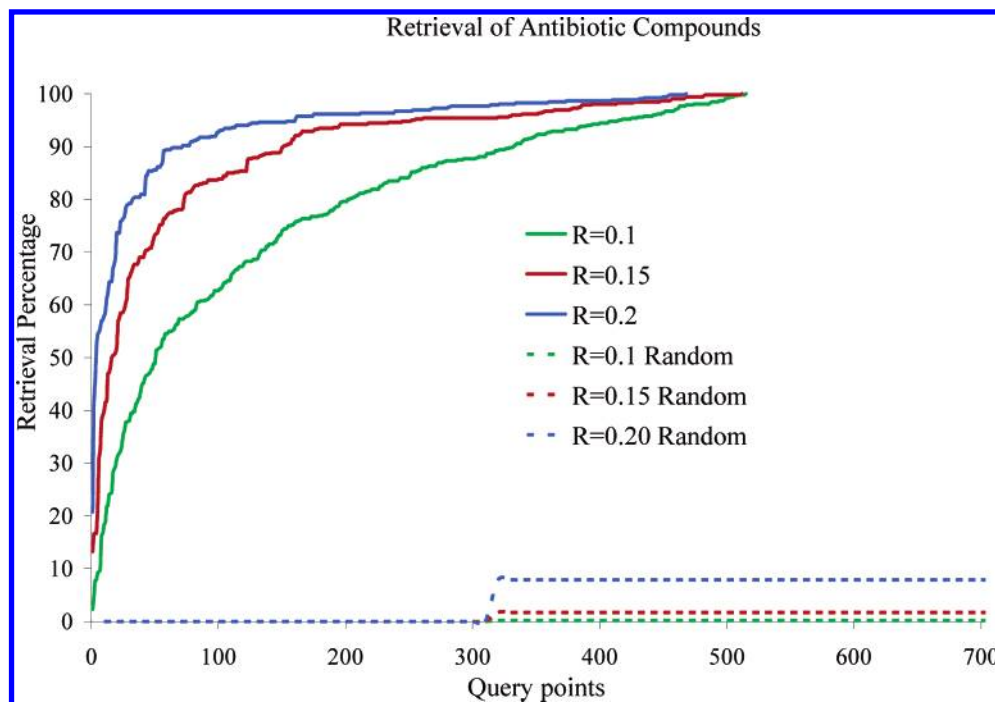


Figure 7. Retrieval of antimicrobial compounds from the general molecular database (>2 M entries) using the k NN model with varying distance constraints (solid lines). The dashed lines correspond to random identification of antimicrobial substances.

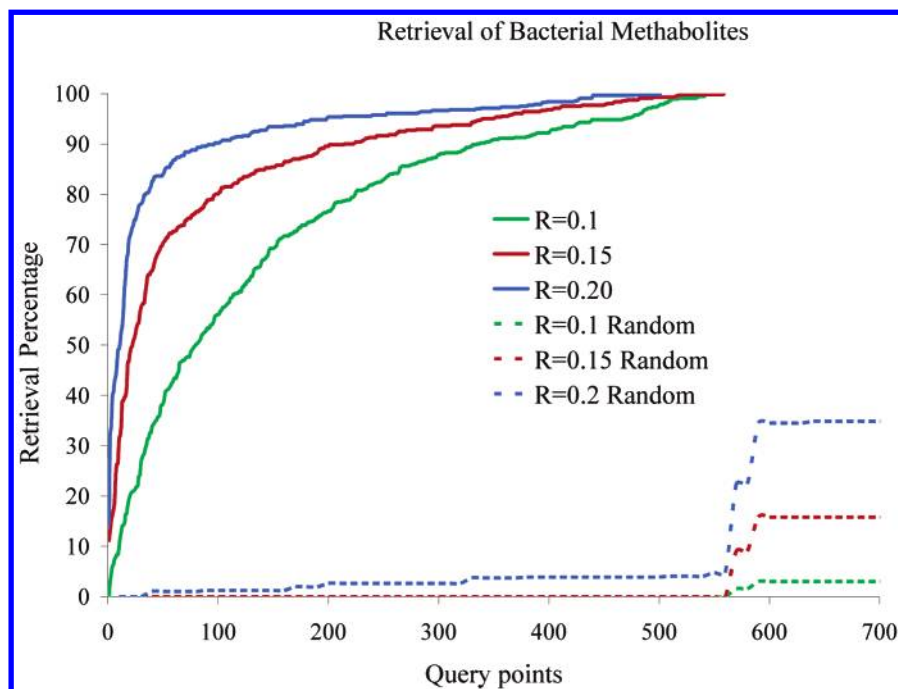


Figure 8. Retrieval of bacterial metabolite substances from the general molecular database (>2M entries) using the k NN model with varying distance constraints (solid lines). The dashed lines correspond to random identification of bacterial metabolites.

drugs (marked in orange), druglikes (marked in red), and bacterial metabolite substances (marked in cyan).

To assess separation between the studied groups in the descriptors space and to sample their compactness and overlaps, we utilized the developed k NN solutions. Thus, for each studied group of chemical substances we considered every constituent molecule as a probe that has then been placed into chemical space consisting of 4346 studied compounds mixed with 2 066 095 ZINC structures. For each probe we applied the developed k NN model (distance function for the corresponding activity class) to identify all

active entries located within a certain radii R . For each studied group of compounds, we have continued such probing until all group's active elements could be identified.

Understandably, the established number of the required probe-based queries strongly depended on the probe radius. Figures 7–11 feature probe-based recovery of antimicrobials, bacterial metabolites, drugs, druglikes, and human metabolites from the pool of 2 071 251 entries using k NN based solution developed for these classes of activity and utilizing search-radius values of 0.10, 0.15, and 0.20. The blue 'recovery curves' in Figures 7–11, corresponding to probing

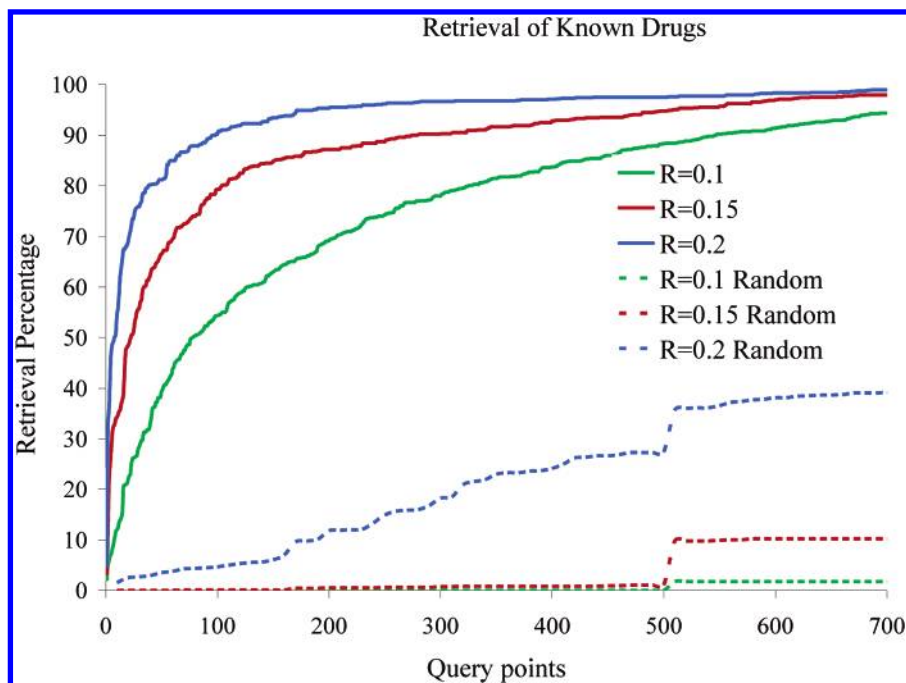


Figure 9. Retrieval of drugs from the general molecular database (>2 M entries) using the k NN model with varying distance constraints (solid lines). The dashed lines correspond to random identification of drug entries.

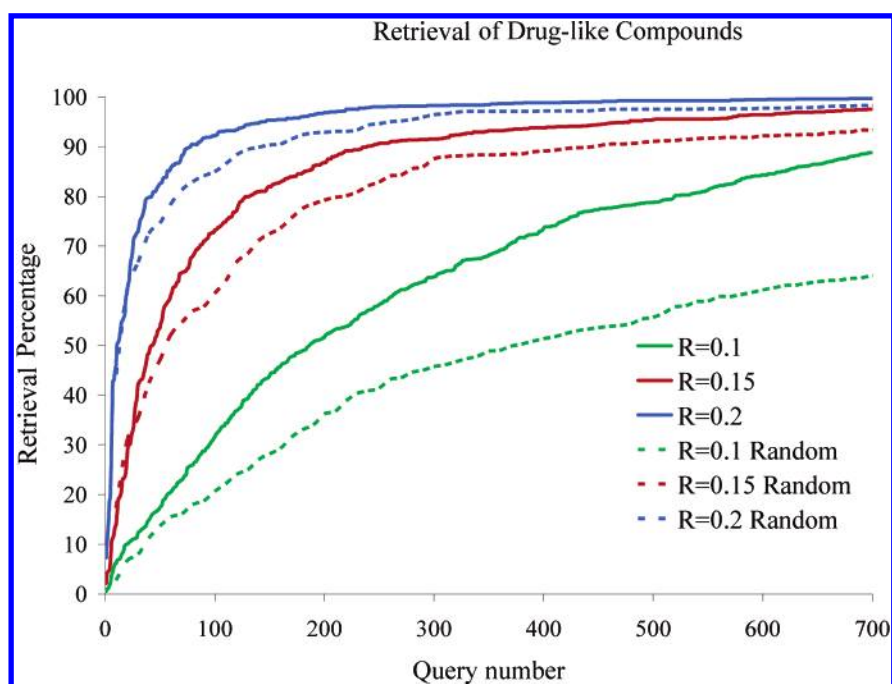


Figure 10. Retrieval of druglike substances from the general molecular database (>2 M entries) using the k NN model with varying distance constraints (solid lines). The dashed lines correspond to random identification of druglike entries.

with $R = 0.20$, illustrate that k NN recovery of the majority of antimicrobials, drugs, nondrugs, and metabolites can be accomplished in less than 100 iterations.

When the database has been queried with $R = 0.15$ and, particularly $R = 0.10$ probes (red and green curves, respectively), the complete recovery may require as many as ~500–700 steps. Figures 7–11 also feature random recovery of 520 antimicrobials, 959 drugs, 1202 druglikes, and 562 bacterial and 1104 human metabolites from the total of 2 071 251 chemicals structures (the corresponding curves are marked in dashed lines). As random recovery curves illustrate, only members of the ‘inactive druglike compounds’

group could be found somewhat efficiently by random placing of probes into chemical space. On another hand, active probe-based recovery of druglike substances was not very efficient either (Figure 10). These observations may justify that druglike entries are spread throughout the descriptors space without distinctive clustering.

In contrast, other types of substances, particularly human metabolites, could be recovered very rapidly by the k NN search, which characterizes them as compact collections of entries.

One useful criterion for assessing clustering of active entries in a large database is the number of k NN probes of

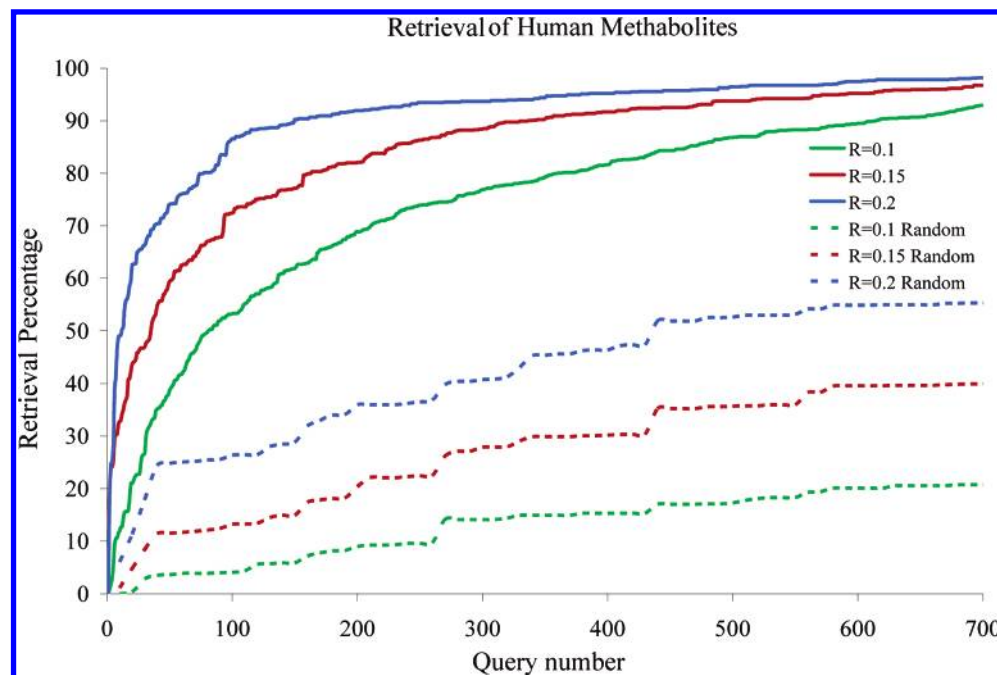


Figure 11. Retrieval of human metabolite substances from the general molecular database (>2 M entries) using the k NN model with varying distance constraints (solid lines). The dashed lines correspond to random identification of human metabolites.

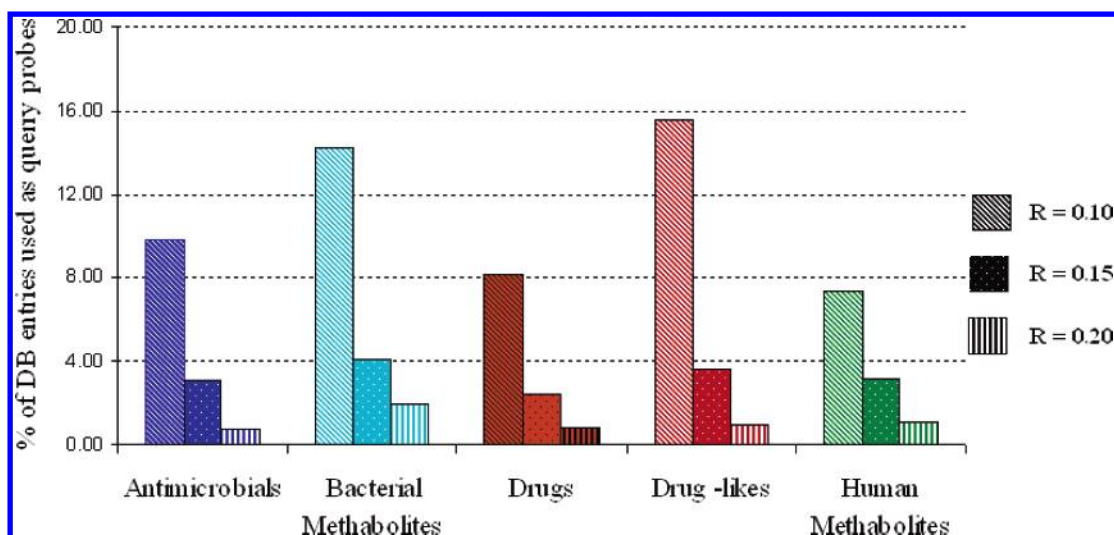


Figure 12. Histograms of $P_{1/2}$ values -fractions of cluster entries required to retrieve 50% members of the corresponding cluster from a large molecular database using the k NN approach. The values have been identified for the searches with varying R parameters.

a certain radius that are required for identification of 50% of active entries. Thus, we have computed the corresponding parameters $P_{1/2}$ for five k NN models with search radius values of 0.10, 0.15, and 0.20. The established numbers of probes required to identify 50% of each group are featured in Figure 12, where they are normalized by the size of the corresponding activity group. Thus, it required only 81 probes (or ~7% of the total number of entries) with $R = 0.10$ radius to identify 552 human metabolites (50% of the total number) from the mixed pool of more than 2 million chemical structures. This illustrates that human metabolite substances are clustered very tightly in multidimensional descriptors space. The grouping becomes less profound for conventional drugs, followed by antimicrobials, bacterial metabolites, and, finally, by the group of druglikes which required more than 15% of actives to be used as probes to locate 50% of the group (Figure 12).

To summarize the results of the above-described experiments it is possible to conclude that groups of antimicrobial compounds, conventional drugs, druglike chemicals, and bacterial and human metabolites form distinctive and relatively compact clusters in chemical space where the dimensions are defined by 'inductive' and conventional QSAR descriptors. Such clustering allows rather accurate recognition of these types of biological activity using various statistical and machine-learning techniques that include methods of artificial neural networks, k -nearest neighbors, linear discriminative analysis, and multiple linear regression.

The QSAR separation of antimicrobials, drugs, nondrugs, and metabolites with these approaches demonstrates a certain degree of similarity between the members of these activity classes resulting in their cross-recognition by the corresponding QSAR models. On another hand, the group of human metabolites demonstrated rather distinctive behavior com-

Table 4. Most Common Distinct Molecular Scaffolds Classified for the Studied Groups of Chemical Substances




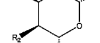
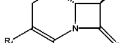

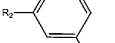


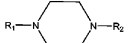

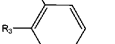

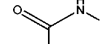
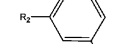

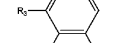

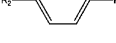
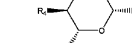

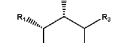

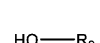
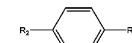

























Class	Average Number of Molecular Features	Rank 1 fragment	Rank 2 fragment	Rank 3 fragment	Rank 4 fragment	Rank 5 fragment
Antibiotics	2.56	 22%	 22%	 6%	 4%	 3%
Drugs	1.70	 25%	 25%	 7%	 4%	 3%
non-Drugs	2.50	 15%	 15%	 6%	 3%	 2%
Bacterial Metabolites	1.57	 32%	 32%	 8%	 3%	 2%
Human Metabolites	0.88	 22%	 22%	 8%	 5%	 3%

Table 5. Most Common Distinct Substituents Classified for the Studied Groups of Chemical Substances

Class	Average Number of Molecular Features	Rank 1 fragment	Rank 2 fragment	Rank 3 fragment	Rank 4 fragment	Rank 5 fragment
Antibiotics	5.08					
Drugs	2.84					
non-Drugs	2.72					
Bacterial Metabolites	4.72					
Human Metabolites	2.64					

pared to all other studied types of chemicals, with the corresponding cluster of entries being the most compact and completely separated from other groups in the descriptors space. Thus, the results of the comparative QSAR analysis allow categorizing human metabolites as a distinctive class of chemical structures and raises questions about structural determinants of their unusual QSAR properties.

Substitutions Analysis of Drugs, Nondrugs, Antimicrobials, and Metabolites. The QSAR descriptor-based solutions utilized for defining commonalities and differences between the studied groups of compounds do not allow interpreting their predictions in conventional terms of chemical structures. Thus, to complement the developed QSAR solutions for antimicrobials, drugs, nondrugs, and metabolites we investigated the occurrence of structural scaffolds and substituents within these classes.

First, we have identified all unique molecular scaffolds and substituents present in 4346 studied compounds and counted their frequencies of occurrence within antimicrobials, drugs, nondrugs, and metabolites. We have used the custom MOE script 'Fragmentation.svl'¹⁶ for this purpose—the module that implements the druglike index approach and separates molecular structures into the constituent 'building blocks'.²¹ Thus, we have determined the most abundant elements of chemicals structures of antimicrobials, conventional drugs, inactive chemicals, and bacterial and human metabolites. The established most abundant scaffolds and substituents are featured in Tables 4 and 5, respectively. The results demonstrate that structures of drugs and druglike substances are dominated with ortho-, para-, and 1,2,4-substituted aromatic rings, while antibiotics and bacterial metabolites are similarly enriched with hydroxyl substituents.

tures and exhibit high abundance of aromatic and sugar fragments. Understandably, the antimicrobials appeared to be over-represented with penicillin derivatives as the corresponding scaffold has been ranked fifth in the frequenters list. Interestingly, the conventional therapeutics appeared to be the only group with a high abundance of *N,N*-disubstituted piperazine scaffold, which, perhaps, may be attributed to its favorable pharmacokinetic potentials.

The group of human metabolites once again demonstrated rather distinctive properties compared to other studied classes, with the three most abundant molecular scaffolds being identified as five- and six-member sugars and a 1,6-disubstituted purine (in contrast to benzene scaffolds common for other studied types of chemicals). Sugars have also been found among abundant substructures for antimicrobials and bacterial metabolites.

The most common substituents featured in Table 5 demonstrate that hydroxyl, methyl, amino, and methoxy groups are abundantly present among antimicrobials, drugs, nondrugs, and metabolites. High abundance of phenyl and halogen substituents among drugs and druglike substances are indicative of their synthetic nature, while the common presence of carboxyl fragments in metabolites and antimicrobials can, perhaps, be attributed to their natural origin.

Tables 4 and 5 also contain values of average numbers of scaffolds and substituents per structure, determined for the studied groups of compounds. The values illustrate that metabolic substances tend to contain a single scaffold, while drugs and synthetic chemicals usually contain 2–3 distinct molecular fragments. The average number of substituents distinguishes bacterial metabolites and antimicrobial which tend to have around 5 substituents, while this number ranges from 2 to 3 for drugs, druglikes, and human metabolites. The tendency of multiple substitutions in antimicrobials and bacterial metabolites is also reflected by the prevalence of polysubstituted scaffolds in these classes (Table 5).

The analysis of the most common structural scaffolds present in Table 5 for five groups of chemical substances may also bring certain insight into previously discussed ~40–50% cross-recognition among such classes of compounds as bacterial metabolites and antimicrobials. Thus, these two subsets share three out of five most common scaffolds, including the HO–R which has been ranked as the ‘number one’ scaffold within both data sets. Similarly, the established 70–80% cross-recognition between drugs and presumably inactive druglike substances (Table 3) can be related to the observation that these classes of chemicals share three of their most abundant scaffolds (and both have para-substituted benzene ranked first). It is feasible to anticipate that enrichment of two groups of compounds with similar molecular fragments should maximize the overlap in the descriptors space and therefore lead to cross-recognition by the corresponding QSAR models. Similarly, the uniqueness of the most frequent human metabolite scaffolds can contribute into their distant positioning relative to other studied groups.

Thus, the analysis of distribution of distinctive molecular fragments among the studied groups of chemical structures may be useful for interpreting some of the results of the binary QSAR modeling with 3D ‘inductive’ and traditional 2D QSAR descriptors.

Analysis of Scale-Free Organization of Molecular Structures of Drugs, Nondrugs, Antimicrobials, and Metabolites. The occurrence of unique molecular fragments in antimicrobials, drugs, druglikes, and metabolites has also been analyzed in terms of their frequencies. Thus, for each group of compounds we have computed the percentiles $f(x)$ of unique scaffolds and substituents that occur more than x times. For example, in the case of conventional drugs there are 549 unique scaffolds that occur ≥ 1 times, with 153 of them occurring ≥ 2 times, 82 scaffolds occurring ≥ 3 times, 59 scaffolds with ≥ 4 occurrence, etc., with the count continued until establishing that para-substituted benzene occurred 219 times among the drugs. Similar inhomogeneous distributions have been observed for other data sets. Such uneven numbers illustrate that representation of unique molecular fragments is highly unequal, with one scaffold type occurring in up to 35% of structures, while hundreds of other unique molecular fragments appear in the data set only once or twice. Typically, such inhomogeneous frequency distributions point toward the possible scale-free organization of a system.

Scale-free distributions have been established for many naturally evolved systems that include Internet, economic, professional, sexual, and social networks, where numbers of connections between nodes are very unevenly distributed, in language structures where occurrence of certain words strongly prevail, in uneven economic systems that can be characterized by the infamous the ‘rich gets richer’ principle.²³ In such unequally populated systems the distribution of components is typically characterized by the power law

$$f(x) = \alpha x^{-\gamma} \quad (1)$$

where α is a constant, and $f(x)$ is a fraction of features with the population occurrence x , with the value of exponent γ typically varying in the range of [1.5–3.5]. Formula 1 can be translated into a simple notion that scale-free heterogeneous systems (such as a molecular universe) are populated with a variety of distinct entries (say unique scaffolds), but the majority of unique entries are present in few instances, while a handful of dominant members are present in great numbers of copies determining the main system’s content.

To establish whether the frequencies of distribution of structural features within the studied classes of compounds obey power law (1), we have applied double-logarithm transformation to the corresponding $f(x) \sim x$ dependences. Figures 13 and 14 feature the resulting linear trends that have been established for the cumulative distributions of unique scaffolds and substituents among drugs, nondrugs, antimicrobials, and bacterial and human metabolites.

The charts illustrate that double-logarithm transformation resulted in high quality linear trends for the frequency-distribution dependences. On another hand, the established slope values corresponding to exponent γ in (1) demonstrate that the occurrence of substituents among the studied groups of compounds (Figure 13) do not generally obey the power law, as the corresponding γ parameters vary in the range of 0.75–0.86. On another hand, the distribution of molecular scaffolds (Figure 13) can be characterized as scale-free, in particular for drugs and human metabolites (the corresponding γ values are 1.26 and 1.35, respectively). This demonstrates that structures of conventional human therapeutics and

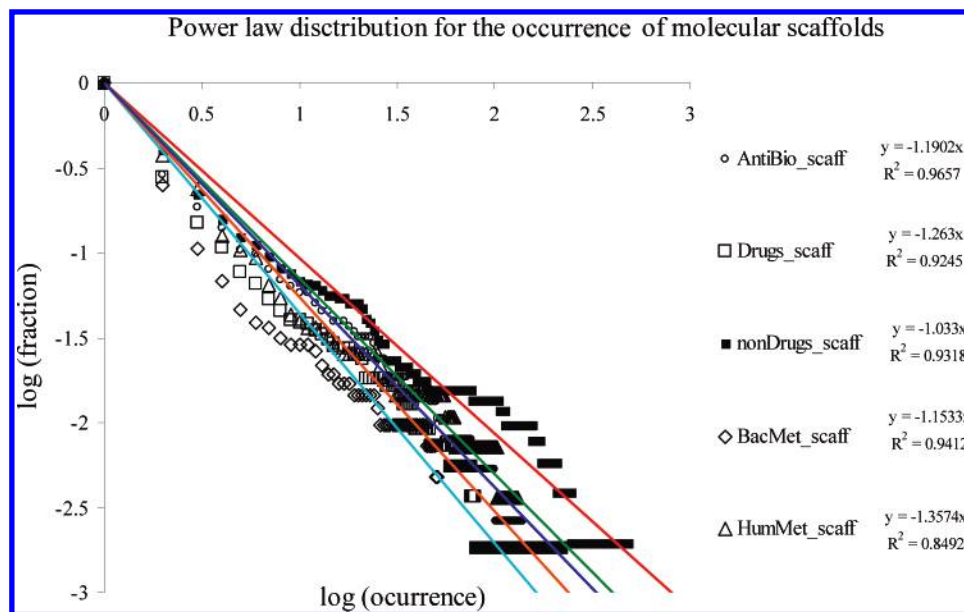


Figure 13. Power-law distribution of frequencies of occurrence of distinct molecular scaffolds within five studied groups of compounds.

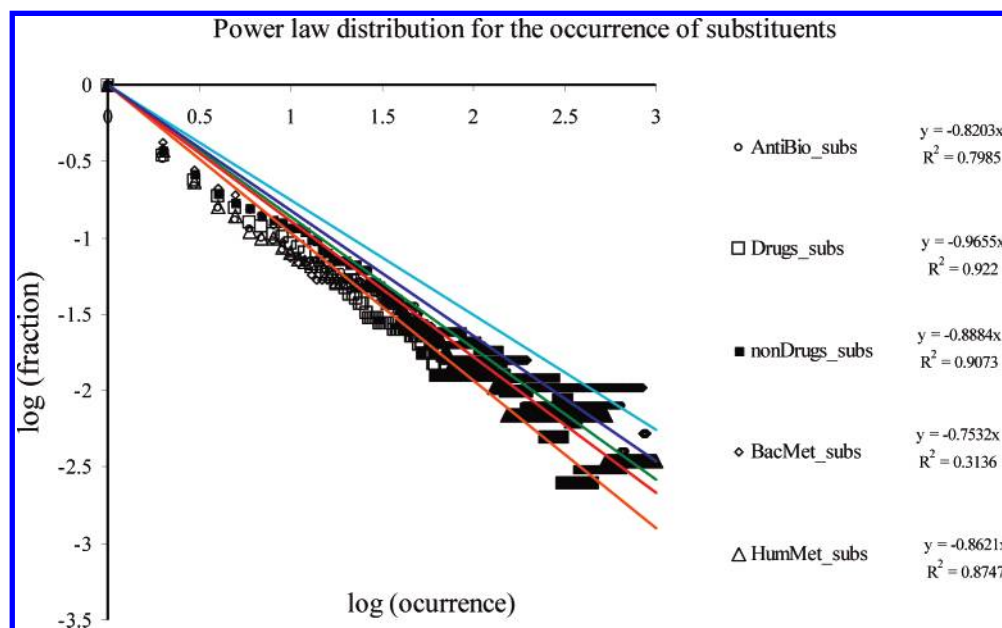


Figure 14. Power-law distribution of frequencies of occurrence of distinct substituents within five studied groups of compounds.

human metabolites are dominated with few scaffolds which may be an indication of certain types of scaffolds being reutilized in the process of drug development (such as para- and meta-substituted benzenes) while others, such as sugar and nucleotide systems, are favored by natural evolution.

The scale-free organization of compounds within the studied groups may also be a factor in their compact clustering in chemicals space, as overrepresentation of a certain scaffold(s) will result in closer positioning of the corresponding compounds. As it has been mentioned before, the degree of the scale-free organization of a system is characterized by the value of power exponent γ . In the context of dependences (1) describing the occurrence of distinct molecular fragments, the power exponents should reflect the extent of overpopulation of a molecular data set with its most frequent scaffold(s). Therefore, the established values γ for antimicrobials, drugs, nondrugs, and bacterial and human metabolites (in Figure 6 they correspond to the

slopes of presented linear dependences) should adequately reflect the compactness of the corresponding clusters.

On another hand, the previously estimated normalized values of the numbers of positives probes required for 50% recovery of actives ($P_{1/2}$ parameters featured in Figure 11) should also reflect density of the corresponding molecular clusters in descriptors space, as more condensed groups of points can be recovered more efficiently by the k NN algorithm. Hence, we decided to examine whether γ values derived for the power-law dependences (1) for antimicrobial, drug, nondrug, and metabolic scaffolds can form any kind of trend with $P_{1/2}$ parameters identified by the k NN probing of these groups of substances. As Figure 15 illustrates such dependence between seemingly unrelated parameters $P_{1/2}$ and γ values could be observed.

We did not consider this trend to be particularly meaningful or expected, but nonetheless it is not unreasonable to suggest that there might a relationship between the degree

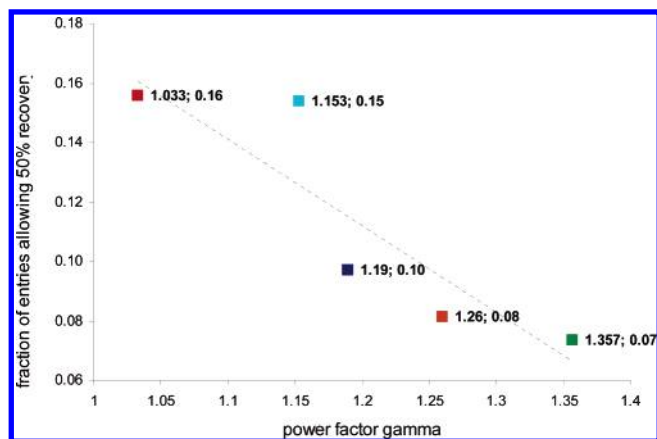


Figure 15. A trend relating power-law (1) exponent values γ established from the distributions of unique structural scaffolds among studied groups of compounds and normalized numbers of database probes with $R = 0.10$ required for retrieval of 50% of actives.

of scale-free organization of a class of chemicals (in other words—the measure of over-representation of molecular scaffolds of a certain type) and density of the corresponding activity cluster in the descriptor space (as similar molecules should produce similar values of QSAR descriptors and will tend to group together). Therefore it is possible to speculate that likely scale-free organization of groups of actives compounds may have an impact on their relative positioning and overlapping in the descriptors space.

CONCLUSIONS

Thus, to summarize the spectra of results presented in the above sections, it is possible to conclude that antimicrobials, conventional therapeutics, and bacterial and human metabolites are organized into rather compact as distinguished clusters in QSAR descriptors space which makes it possible to distinguish these types of chemicals with binary SA models. When we utilized the k -nearest neighbors, multiple linear regression, and linear discriminative analysis algorithms for that purpose, they all allowed generally accurate separation of the activities; however, it is the method of artificial neural networks that provided the most accurate predictions.

The binary QSAR analysis also demonstrated that the studied groups of compounds can be characterized by rather significant overlapping, particularly profound for bacterial metabolites and antimicrobial drugs, which can be attributed to their similar origin. The groups of human metabolites demonstrated distinctive QSAR behavior compared to the other four types of substances; the cluster of human metabolites is the most condensed in the descriptors space and shows no intersections with other groups.

The analysis of scaffolds and substituents distribution among the studied groups of chemicals revealed a likely scale-free nature of distribution of molecular scaffolds for drugs and human metabolites. These observations imply that structures of conventional human therapeutics and human metabolites are dominated with few scaffolds that are correspondingly favored by the drug developers and by natural evolution. It has also been demonstrated that the established scale-free organization of chemical structures may be an important factor of the spatial positioning and overlapping of the studied groups of chemicals in the descriptors space.

The overall results of the conducted comparative QSAR and fragments distribution analysis bring more insight into the nature and structural dominants of the studied classes of chemicals substances and, if necessary, can help rationalizing the design and discovery of novel antimicrobials and human therapeutics with metabolite-like chemical profiles.

ACKNOWLEDGMENT

The authors thank Dr. David Wishart (University of Alberta) for providing us the Database of Human Metabolites.

Supporting Information Available: SMILES notation of the studied compounds (except some bacterial metabolites) and values of 62 descriptors described in the study. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Cherkasov, A. 'Inductive' Descriptors. 10 Successful Years in QSAR. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 21–42.
- (2) Cherkasov, A.; Jankovic, B. Application of 'Inductive' QSAR Descriptors for Quantification of Antibacterial Activity of Cationic Polypeptides. *Molecules* **2004**, *9*, 1034–1052.
- (3) Cherkasov, A.; Shi, Z.; Fallahi, M.; Hammond, G. L. Successful in Silico Discovery of Novel Non-Steroidal Ligands for Human Sex Hormone Binding Globulin. *J. Med. Chem.* **2005**, *48*, 3203–3213.
- (4) Cherkasov, A. Can 'Bacterial-Metabolite-Likeness' Model Improve Odds of 'In silico' Antibiotic Discovery? *J. Chem. Inf. Model.* **2006**, *46*, 1214–1222.
- (5) ChemIDPlus database: <http://chem.sis.nlm.nih.gov/chemidplus/>, May 2006.
- (6) J. Antibiot. database: <http://www.nih.gov.jp/~jun/NADB/byname.html>, May 2006.
- (7) Tomas-Vert, F.; Perez-Gimenez, F.; Salabert-Salvador, M. T.; Garcia-March, F. J.; Jaen-Oltra, J. Artificial Neural Networks Applied to the Discrimination of Antibacterial Activity by Topological Methods. *J. Mol. Struct. (THEOCHEM)* **2000**, *504*, 249–259.
- (8) Cronin, M. T. D.; Aptula, A. O.; Dearden, J. C.; Duffy, J. C.; Netzeva, T. I.; Patel, H.; Rowe, P. H.; Schultz, T. W.; Worth, A. P.; Voutzoulidis, K.; Schuurmann, G. Structure-based classification of antibacterial activity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 869–878.
- (9) Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F. J.; Salabert-Salvador, M. T.; Diaz-Villanueva, W.; Castro-Bleda, M. J.; Villanueva-Pareja, A. Artificial neural networks and linear discriminant analysis: a valuable combination in the selection of new antibacterial compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1031–1041.
- (10) *The Merck Index 13.4 CD-ROM Edition*; CambridgeSoft, Cambridge, MA, 2004.
- (11) *Human Metabolite Database*: http://redpoll.pharmacy.ualberta.ca/~aguo/www_hmdb_ca/HMDB/, May 2006.
- (12) *Analyticon Discovery Company*: www.ac-discovery.com May 2006.
- (13) *Assinex Gold Collection*; Assinex Ltd.: Moscow, 2004.
- (14) Halgren, T. A. Merck molecular force field. 1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.
- (15) *Molecular Operational Environment*; Chemical Computing Group Inc.: Montreal, Canada, 2005.
- (16) *MOE SVL exchange community*: <http://svl.chemcomp.com/index.php>, May 2006.
- (17) WEKA 3: Data Mining Software in Java <http://www.cs.waikato.ac.nz/~ml/weka/>, May 2006.
- (18) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley: New York, 1999; 380p.
- (19) Karakoc, E.; Cherkasov, A.; Sahinalp, S. C. *Distance Based Algorithms for Small Biomolecule Classification and Structural Similarity Search*; ISMB'06 Intelligent Systems for Molecular Biology, 2006.
- (20) CPLEX: High-performance software for mathematical programming <http://www.ilog.com/products/cplex/>, May 2006.
- (21) Xu, J.; Stevenson, J. Drug-like Index: A New Approach To Measure Drug-like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177–1187.
- (22) Irwin, J. J.; Shoichet, B. K. ZINC- a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
- (23) Barabasi, A.-L. *Linked: The New Science of Networks*; Perseus Publ.: Cambridge, MA, 2002; 256p.
- (24) Livingston, D. J. *Data analysis for chemists. Applications to QSAR and chemical product design*; Oxford University Press: 1995; 324p.