In Silico Estimation of DMSO Solubility of Organic Compounds for Bioscreening

Article *in* Journal of Biomolecular Screening · March 2004

DOI: 10.1177/1087057103260006 · Source: PubMed

Konstantin Balakin
Moscow Institute of Physics and Technology
94 PUBLICATIONS **1,762** CITATIONS
SEE PROFILE

Yan A Ivanenkov
ChemDiv, Inc.
80 PUBLICATIONS **923** CITATIONS
SEE PROFILE

Yuri V. Nikolsky
Prosapia Genetics
230 PUBLICATIONS **14,881** CITATIONS
SEE PROFILE

In Silico Estimation of DMSO Solubility of Organic Compounds for Bioscreening

Konstantin V. Balakin, Yan A. Ivanenkov, Andrey V. Skorenko, Yuri V. Nikolsky, Nikolay P. Savchuk and Andrey A. Ivashchenko

J Biomol Screen 2004; 9; 22
DOI: 10.1177/1087057103260006

In Silico Estimation of DMSO Solubility of Organic Compounds for Bioscreening

KONSTANTIN V. BALAKIN, YAN A. IVANENKOV, ANDREY V. SKORENKO,
YURI V. NIKOLSKY, NIKOLAY P. SAVCHUK, and ANDREY A. IVASHCHENKO

Solubility of organic compounds in DMSO is an important issue for commercial and academic organizations handling large compound collections or performing biological screening. In particular, solubility data are critical for the optimization of storage conditions and for the selection of compounds for bioscreening compatible with the assay protocol. Solubility is largely determined by the solvation energy and the crystal disruption energy, and these molecular phenomena should be assessed in structure-solubility correlation studies. The authors summarize our long-term experimental observations and theoretical studies of physicochemical determinants of DMSO solubility of organic substances. They compiled a comprehensive reference database of proprietary data on compound solubility (55,277 compounds with good DMSO solubility and 10,223 compounds with poor DMSO solubility), calculated specific molecular descriptors (topological, electromagnetic, charge, and lipophilicity parameters), and applied an advanced machine-learning approach for training neural networks to address the solubility. Both supervised (feed-forward, back-propagated neural networks) and unsupervised (Kohonen neural networks) learning methods were used. The resulting neural network models were validated by successfully predicting DMSO solubility of compounds in independent test selections. (*Journal of Biomolecular Screening* 2004:22-31)

Key words: neural networks, Kohonen self-organizing maps, DMSO, solubility, quantitative structure-property relationship

INTRODUCTION

DMSO: Properties and role in the pharmaceutical industry

DMSO is recognized as the most powerful of any readily available organic solvent. It dissolves the great variety of organic substances to the highest loading level, including carbohydrates, polymers, peptides, and many inorganic salts and gases. Loading levels of 50-60 wt% are often observed with DMSO (compared with 10-20 wt% with typical solvents). The solvating ability of DMSO is primarily related to its high dielectric constant, exceeding that of most other common dipolar aprotic solvents, such as dimethylformamide, N,N-dimethylacetamide, and N-methylpyrrolidone.[1] Another, more subtle, phenomenon affecting solvating ability is stereochemistry (Fig. 1). The structure of the DMSO molecule is not flat but is a trigonal pyramid in shape. There is a highly directional lone pair of electrons at the apex of the pyramid, which further helps complex and otherwise solvate typical solute molecules.

DMSO has low toxicity by every route of administration (oral, inhalation, and dermal) and has low environmental toxicity.[2-5]

DMSO is miscible and does not form azeotropes with water. At low concentrations, it usually does not have any serious biological effect; therefore, DMSO-water solutions of organic compounds can be used in various bioassays. In case of necessity, trace DMSO impurities in the product may be removed by an aqueous, alcohol, or ethyl acetate wash. Freeze-drying is an option for thermally sensitive organic substances, and this technique is widely used in the pharmaceutical industry for the preparation of dry samples in the form of thin films.

DMSO has widespread pharmacological applications. It was used at high concentrations for modifying the radiation effect on the eye.[6] It is also applied as a drug carrier to cells.[7] DMSO is one of the best cryoprotectors.[8] Due to its physicochemical properties, high solvent power, low chemical reactivity, and relatively low toxicity, DMSO is a solvent of choice for sample storage and handling in the pharmaceutical industry, particularly in stages of primary high-throughput bioscreening.

A potential problem is associated with the ability of DMSO to slowly react with certain classes of chemical compounds, such as Schiff bases, or slowly decompose in the presence of water, organic or inorganic acids, and strong oxidizing agents. The decom-

Chemical Diversity Labs, Inc., San Diego, CA.

Received Jun 3, 2003, and in revised form Sep 5, 2003. Accepted for publication Sep 12, 2003.

Journal of Biomolecular Screening 9(1); 2004
DOI: 10.1177/1087057103260006
Published by Sage Publications in association with The Society for Biomolecular Screening

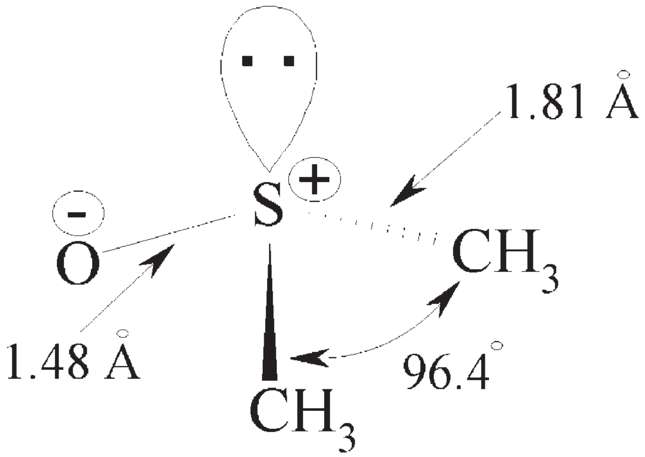

FIG. 1. Spatial parameters of DMSO molecule.

position products, such as formaldehyde, methyl mercaptan, dimethyl sulfide, and dimethyl disulfide can further react with the dissolved organic compounds. As a result, a prolonged storage of organic substances in DMSO may result in their partial chemical degradation. For example, relative stability of trifluoroacetic acid (TFA) salts or adducts in comparison with non-TFA analogs in DMSO was recently studied.[9] It was shown that compounds containing acid-labile groups should not be routinely stored as TFA salts or adducts in DMSO solution in compound collections used for high-throughput screening (HTS).

Nowadays, distribution of samples as DMSO solutions is replacing distribution of powders/solids throughout the drug discovery industry worldwide. Not surprisingly, compound storage in DMSO has been a subject of intense research and discussions in recent years. Thus, the problem of chemical degradation of compounds dissolved in DMSO has been addressed in a series of recent works.[10-12] For instance, the effect of temperature cycling (the impact of repeated freeze-thaw cycles) on dissolved samples from the large chemical repositories was studied.[10] Water absorption by DMSO, as another factor pertaining to compound instability, was also investigated.[11,12] The materials used to produce the storage vials and microplate seals have changed significantly over the past 2 to 3 years for improving compound handling and data quality. To optimize compound storage, different microplate seals were evaluated for seal integrity with the physical stress tests and the detection of materials leached into DMSO.[13,14]

It is believed that compound solubility in DMSO represents a problem at least as serious as compound stability in combinatorial libraries.[15] An apprehension of the factors affecting a compound's DMSO solubility could help in predicting the storage conditions and appropriateness of compounds for primary bioscreening programs. An in silico procedure for estimating DMSO solubility would be a very useful tool for the discovery chemistry vendors and for the companies involved in bioscreening at early stages of drug discovery.

There are three main problems with regard to computerized prediction of DMSO solubility of organic compounds. First, the experimental data on the DMSO solubility of drugs and drug-like compounds are insufficient. Only 109 pharmaceutical agents with measured DMSO solubility are mentioned in a special database from a leading industrial DMSO provider.[16] Such a small training set is not suitable for the generation of a quality quantitative structure-property relationship (QSPR) model. Second, the crucial problem is the right selection of molecular descriptors capable of capturing the molecular features essential for solubility. Finally, the QSPR algorithms should be highly effective for real-time handling of very large virtual and real discovery compound databases built over the years.

In this work, we present the results of our long-term studies of DMSO solubility of organic substances from the Chemical Diversity Labs (CDL) corporate collection of drug-like compounds for bioscreening. We assembled a large database of diverse organic compounds that contain proprietary data on compound solubility and then performed an extensive statistical study using specific calculated molecular descriptors and advanced data-mining methods. Then, we developed a neural network computational approach to perform effective in silico assessment of DMSO solubility of organic compounds.

METHODS

Determination of DMSO solubility

The DMSO solubility was estimated as follows. DMSO (1 mL) was added to 10 µmol of a compound. The mixture was shaken at room temperature for 30 min in a tightly sealed vial and then visually inspected. The completely solubilized compounds were classified as a DMSO well-soluble compound, DMSO(+). Incompletely solubilized compounds were also shaken at 40 ° C for 3 h. The compounds were classified as a DMSO poorly soluble compound, DMSO(–), if their insoluble parts could be visually detected after this prolonged shaking.

The above-mentioned potential chemical instability of organic compounds in DMSO can impair their solubility estimation. Thus, when a well-soluble compound (e.g., a Schiff base) reacts slowly with DMSO, the insoluble reaction products may cause solution turbidity. In this work, we assume that no considerable chemical or thermal decomposition occurs under the experimental conditions. This assumption is supported by regular random reexamination of samples after the described dissolution procedure, as well as by the reported data on the stability of dissolved samples in DMSO.[10-12]

It should be noted that the described estimation method does not provide the exact value of DMSO solubility, as opposed to the conventional solubility measurements. The latter techniques typically require the samples of known purity as well as precise methods of determining the precipitation point, such as light-scattering measurements. Solubility also depends on the experimental conditions (temperature, humidity, crystal form, etc.). The mentioned

Table 1. Diversity Statistics for the Reference Databases

Parameters	*DMSO(+)*	*DMSO(–)*
Total number of compounds	55,277	10,223
Number of screens[a]	13,865	5949
Diversity coefficient	0.780	0.785
Number of unique heterocycles	962	422
Number of combinatorial libraries	905	490

a. Screens are simple structural fragments, centroids, with the topological distance equal to 1 bond length between the central atom and the atoms maximally remote from it.

Table 2. Physicochemical Properties of Compounds from the Reference Databases

	DMSO(+)			*DMSO(–)*		
Property	*Mean*	*Minimum*	*Maximum*	*Mean*	*Minimum*	*Maximum*
Molecular weight	389.7	167.1	790.8	434.3	230.2	796.8
logP	4.05	–3.71	9.97	5.01	–1.82	11.65
pK_a	7.00	–1.00	19.20	6.45	–3.30	17.35

logP and pK were calculated with ChemoSoft™ software.

problems with obtaining high-quality measured solubility data under standard conditions make it difficult to develop high-throughput methods for measuring DMSO solubility. We believe that in practice, the offered experimental "yes or no" categorization is an adequate and reasonable starting point for developing an accurate predictive calculation method.

Databases

A total of 65,500 compounds from the CDL corporate collection of drug-like compounds were studied in this work. A total of 55,277 compounds with DMSO solubility at room temperature higher than 0.01 mol/L were classified as DMSO well-soluble compounds, and 10,223 compounds with DMSO solubility at 40 ° C less than 0.01 mol/L were classified as DMSO poorly soluble compounds. These categories were used in all further statistical and modeling experiments.

The diversity parameters for the positive, DMSO(+), and the negative, DMSO(–), reference compound sets are shown in Table 1. As evident from the number of screens, the number of unique heterocyclic fragments, and the diversity coefficients (all these parameters are calculated using the Diversity module[16] of the ChemoSoft™ software tool), both sets are highly diverse. Compounds included in DMSO(+) and DMSO(–) data sets are included, correspondingly, in 905 and 490 combinatorial libraries produced by solution-phase parallel synthesis. The average molecular weight and the values of pK_a and logP (calculated using the SLIPPER module[17] of the ChemoSoft™ software tool) presented in Table 2 are typical of the databases of organic compounds for bioscreening and are consistent with the widely accepted rules of drug likeness.[18]

Table 3. Eight Descriptors Selected by Principal Component Analysis and Used in This Work

Descriptor	*Definition*
logP	Log of 1-octanol/water partition coefficient
SASA	Total molecular solvent-accessible surface area
DipM	Dipole moment
Zagreb	Zagreb index
PNSA-3	Atomic charge-weighted negative surface area
B_rot	Number of rotatable bonds
R_Gyr	Radius of gyration
HBD	Number of H-bond donors

Molecular descriptors

The molecular descriptors were calculated for the entire 65,500-compound data set using Cerius^2 (Accelrys, Inc.) and ChemoSoft™ (Chemical Diversity Labs, Inc.) software tools. The number of descriptors was reduced to 25 by the omission of low-variable and highly correlated ($R > 0.9$) descriptors. To reduce the dimensionality of the descriptor space and select the appropriate input variables for the QSPR modeling experiments, a principal component analysis for 25 descriptors was performed using ChemoSoft™ software.

To measure the difference between the mean values of the descriptors for the 2 categories of compounds studied, t' statistics have been used. For large, normally distributed compound selections studied in this work, t' values higher than 4 to 5 indicate statistically significant differences in mean values.

Neural network modeling

Three independent randomizations, corresponding to the complementary training/cross-validation/testing sets, were considered. For the generation of neural network models, the total combined set of 65,500 molecules was randomized and subdivided into 3 categories: (1) training set of 32,750 compounds (50% of the total number of compounds), (2) cross-validation set of 16,375 compounds (25%), and (3) test set of 16,375 compounds (25%). The cross-validation set was selected to avoid overtraining while generating neural network models with the supervised learning method. The test sets were used for independent validation of all the models. The ChemoSoft™ suite was used for all neural network operations. Eight descriptors shown in Table 3 were used as input variables.

For supervised learning, feed-forward nets were constructed and trained with the molecular descriptors as input values and the scores as output values. For unsupervised learning, we used a 15 × 15 Kohonen net with a 2D organization of the network nodes (neurons). The training parameters were as follows: the number of interactions for the training runs was 2000, the starting adjustment radius for the training runs was 0.01, and the decay factor was

0.001. It should be noted that in supervised learning, the multivariate objects should be split into 3 sets (the training set, the cross-validation or control set, and the test set). On the contrary, in unsupervised learning, the control set is not required because the learning continues until network stabilization.

RESULTS AND DISCUSSION

Molecular descriptors

The choice of the minimal set of relevant molecular descriptors is key for the generation of an effective predictive QSPR model. These descriptors should be easily computable and, at the same time, should adequately encode the specific molecular parameters influencing the property of interest.

Compound solubility is determined by a complex combination of intramolecular and intermolecular forces, including solvation, electrostatics, dipole interactions, van der Waals forces, and the hydrophobic effect. At the initial stage of our study, we determined the molecular properties related to the mentioned molecular phenomena and selected the ones with the maximum discriminative ability between the 2 compound categories.

Sixty molecular descriptors for the structural and physicochemical molecular properties, such as lipophilicity, electromagnetic and quantum parameters, charge distribution, topological features, and steric and surface parameters, were explored in this study. Some low-variable and highly correlated descriptors were taken out, and the principal component analysis (PCA) for the DMSO(+) and DMSO(–) data sets was performed with the remaining 25 descriptors. Based on the results of PCA, 8 molecular descriptors shown in Table 3 were selected and used in all neural network experiments performed in this work.

Differences between DMSO well-soluble and poorly soluble compounds

Crystal disruption energy and solvation energy are the main factors determining the solubility of a molecule. The interactions in a solvating environment are usually studied in the frameworks of continuum[19,20] or supermolecular [21,22] solvent models. In continuum models, the free energy of solvation is calculated by considering first the free energy of formation of a cavity of the correct size and shape to contain the solute molecule and then adding the free energy of interaction of the solute with the solvent, including the contribution from the free energy of changes in the solvent. A major contribution to the interaction energy arises from the polarization of the medium by the electric field of the solute. However, the solvent consists of discrete molecules rather than a continuum fluid with a dielectric constant, and this is why the variety of continuum solvent models has been only moderately successful. Supermolecular methods are capable of more adequate calculation of the solvation energy, but they require extensive quantum-chemical calculations.

Thus, crystal disruption energy and interactions in a solvating environment are determined by a complicated interplay between physical and chemical forces, including solvation, electrostatics, van der Waals forces, and the hydrophobic effect. Estimation of these forces is key in predicting DMSO solubility. However, real QSPR modeling requires high-throughput operations with large compound databases, which makes it problematic to assess exactly all these above-mentioned interactions, as such calculations are computationally very demanding. Therefore, for practical reasons, we should use more easily computable molecular properties.

Histogram plots of the selected descriptors reveal statistically significant differences between compounds in the studied databases (Fig. 2).

On the basis of these data, we compared the differences between DMSO(+) and DMSO(–) sets (Table 4). Poorly soluble compounds have a higher partial negative surface area (ΔPNSA-3 $\approx$ 16.9). This is related to DMSO's ability to accept hydrogen bonds and other atoms bearing partial positive charges important in the solvation-induced breaking of the molecular lattice. Obviously, the lattices with the higher number of negatively charged atoms are less prone to breakage by DMSO. This also explains why well-soluble compounds have a higher number of H-bond donors (ΔHBD $\approx$ 0.48). The poorly soluble compounds have a substantially larger solvent-accessible surface area (ΔSASA $\approx$ –45.6). Surface area is an important variable for solvation energy as it is related to the reversible work required to create a cavity of the molecular shape of the solute.[23] The poorly soluble compounds also have, in general, a higher Zagreb index (ΔZagreb $\approx$ –18.3). A possible explanation is that this descriptor value positively correlates with molecular branching. The higher value of branching is associated with larger molecular surface and poorer solvation.[24,25] The well-soluble compounds tend to be more flexible (ΔB_rot $\approx$ 0.54), which can be related to the destabilization of the molecular lattice by flexible bonds. An effect similar to this was observed in the studies of melting temperatures of organic compounds. The well-soluble compounds also have a lesser radius of gyration (ΔR_Gyr $\approx$ –0.43), which can be explained by the positive correlation of this descriptor with molecular volume and surface. The higher lipophilicity of poorly soluble DMSO compounds (ΔlogP $\approx$ –0.94) is in line with the general observation that lipophilic molecules are usually poorly solvated by polar solvents.

Interestingly, such an important property as the dipole moment, which influences both solvation ability and molecular lattice energy, is almost identical between two data sets (ΔDipM $\approx$ –0.05, t' = 0.84). This can be explained by the fact that the increase of this parameter results in 2 oppositely directed effects: stabilization of molecular lattice and enhancement of solvation ability by DMSO. Probably, these factors are in balance, and no clear influence of the dipole moment on the composite effect, DMSO solubility, is observed.

The described differences between 2 sets are statistically significant and, as our results show, provide some important molecular features that account for the compounds' DMSO solubility charac-

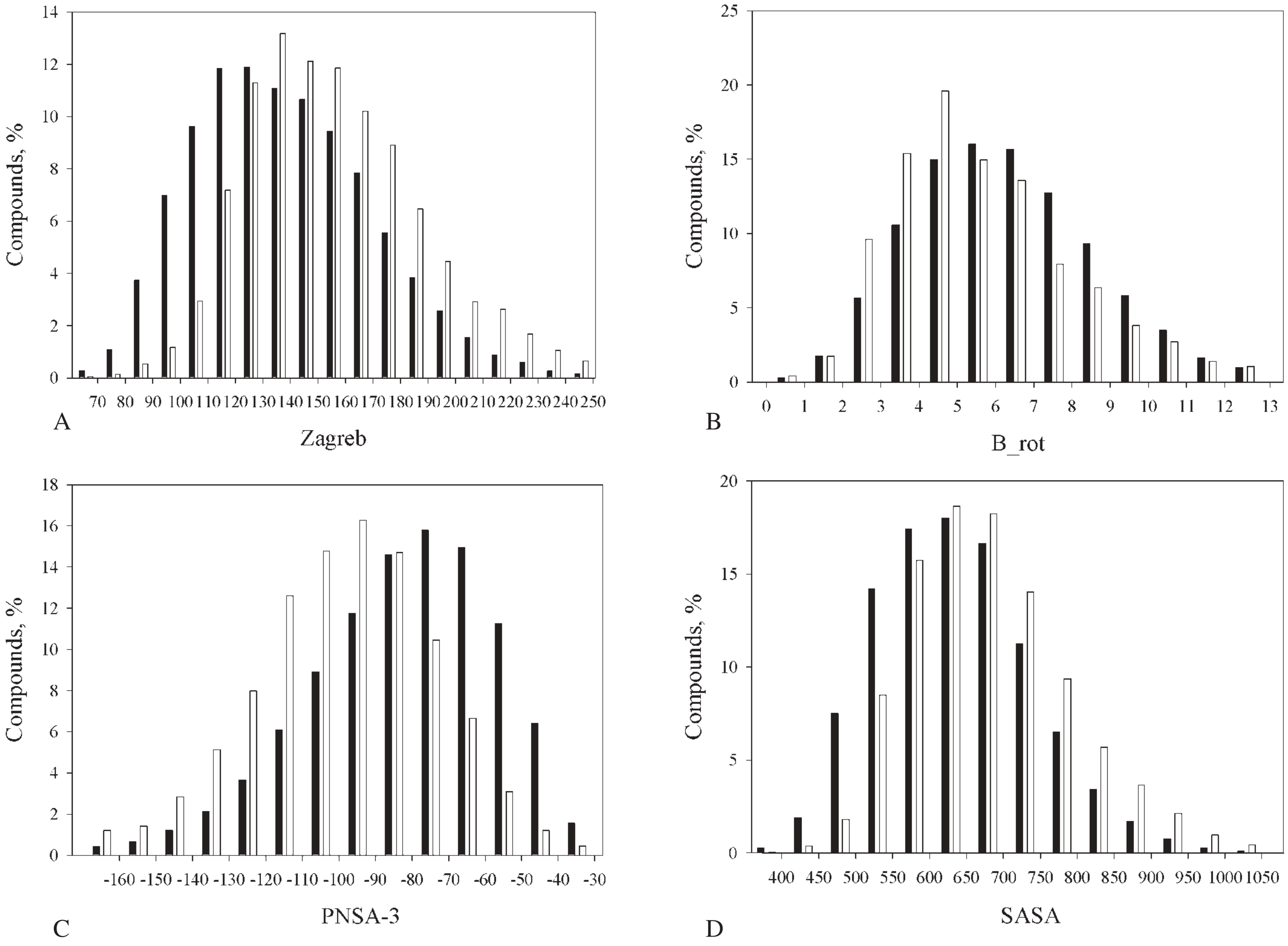

FIG. 2. Property distribution profiles for data sets consisting of 55,277 DMSO well-soluble compounds (black) and 10,223 DMSO poorly soluble compounds (white).

(continued)

teristics. These descriptors, when considered in the aggregate and analyzed with neural network algorithms, serve as a reasonable basis for building the quality classification QSPR models.

Predictive modeling using a supervised learning method

Over the past few years, the classification methods based on artificial neural networks (ANNs) have become popular in computer-assisted drug design. Neural networks were used successfully for segregation of pharmaceutical compounds into categories, such as drug likeness and nondrug likeness.[26,27] Recently, we applied ANN classification methodology for property-based design of GPCR[28] and serine protease-targeted[29] libraries.

Three independent randomizations of the 65,500-compound database were used for the training-testing experiments with the 8-descriptor set. Figure 3 shows the distributions of the calculated scores for the test sets. The gray bars are assigned to DMSO well-soluble compounds and the white bars to DMSO poorly soluble compounds. The classification quality is shown in Table 5: up to 75% of DMSO(+) and 78% of DMSO(–) were correctly predicted in the corresponding test sets (the separation threshold is set to the score value of 0.5). As one can conclude from the distribution histograms, the discrimination efficiency between these compound categories is moderate. Thus, there is a number of false negatives in the score range 0.1 to 0.5 and many false positives in the area 0.5 to 0.9. Despite this fact, the calculated scoring permits us to identify the compound subsets with a significantly increased or decreased percentage of DMSO poorly soluble molecules. The observed level of discrimination for the independent test sets demonstrates a high utility of the developed scoring procedure.

These results suggest that the neural network models based on 8 selected descriptors and the supervised learning method represent a useful tool for estimating DMSO solubility of small-molecule or-

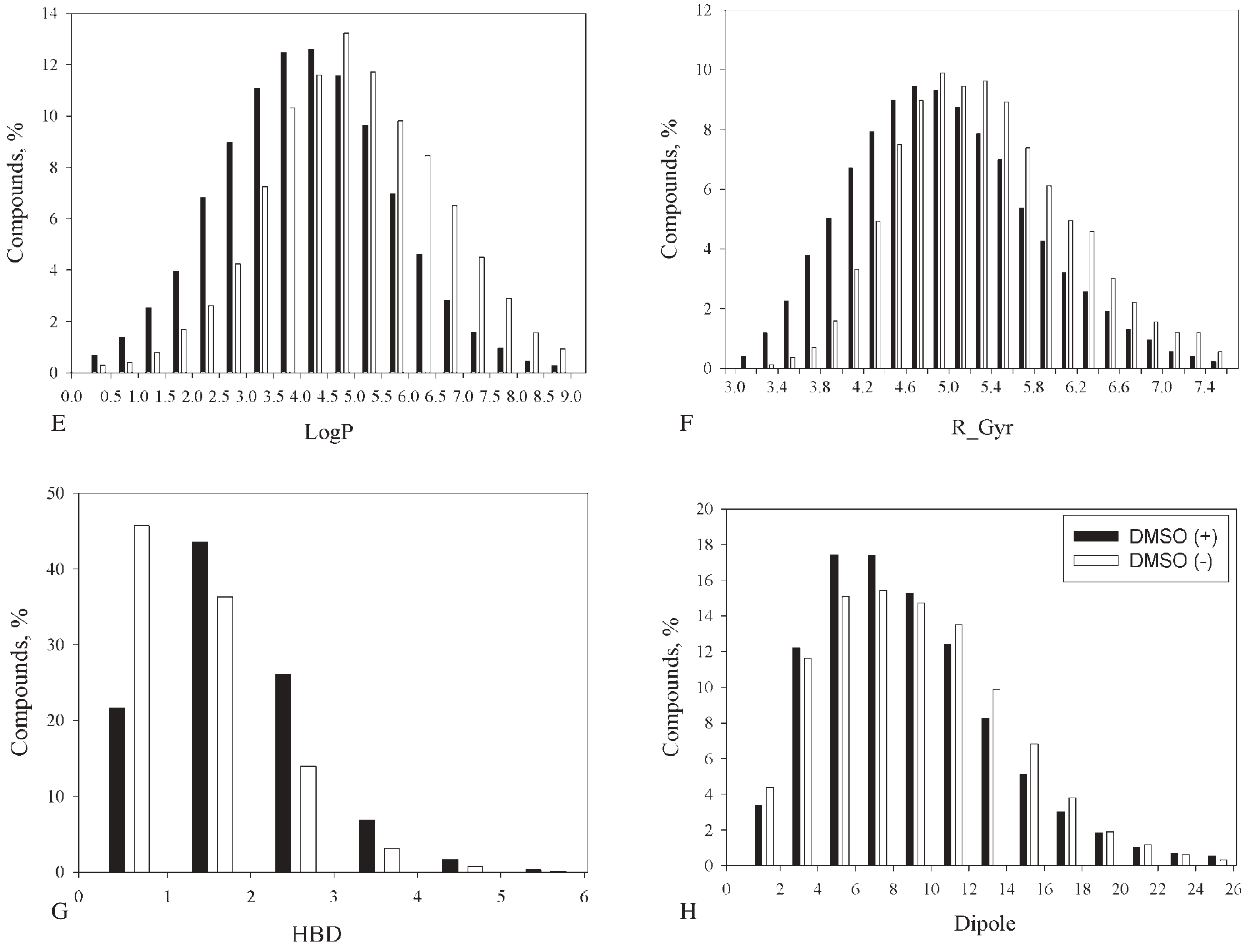

Figure 2 (Continued)

ganic compounds. However, the quality of discrimination is moderate, and we decided to apply an alternative method of classification based on unsupervised learning and generation of Kohonen maps.

Unsupervised Kohonen learning approach

In most studies on the application of neural networks in drug discovery, a supervised learning strategy has been used. The alternative unsupervised learning method becomes popular for comparative analysis and visualization of large ligand data sets.[30] For instance, Kohonen self-organizing maps were used for distinguishing between drugs and nondrugs with a set of descriptors derived from semiempirical molecular orbital calculations.[31] Recently, we reported the application of Kohonen self-organizing

Table 4. Results of *t′* Statistics Calculation for DMSO(+) and DMSO(–) Sets

Descriptors	*Mean DMSO(+)*	*Mean DMSO(–)*	Δ[a]	t′
SASA	630.7	676.3	–45.6	38.3
Zagreb	138.1	156.4	–18.3	52.0
PNSA-3	–73.1	–90.0	16.9	59.8
B_rot	5.77	5.23	0.54	18.7
logP	4.07	5.01	–0.94	51.5
R_Gyr	4.96	5.39	–0.43	44.0
Dipole	9.04	9.09	–0.05	0.84
HBD	1.25	0.77	0.48	49.3

For definitions of descriptors, see Table 3.
a. Difference between mean values.

Table 5. Ratio of Correctly Classified Compounds within the 3 Independent Test Sets (in percentages)

	DMSO(+)				*DMSO(–)*			
Learning Method	*Randomization 1*	*Randomization 2*	*Randomization 3*	*Mean*	*Randomization 1*	*Randomization 2*	*Randomization 3*	*Mean*
Supervised (generalized feed-forward networks)	74.1	72.0	76.1	74.6	77.2	77.1	78.1	77.5
Unsupervised (Kohonen networks)	95.2	92.5	91.5	93.0	93.3	90.1	94.0	92.5

maps for the design of compound libraries for bioscreening with enhanced target-specific informational content.[32] The choice between the supervised and the unsupervised approaches depends on the problem and the available data. For both procedures, objects with known labels are needed. In supervised learning, these labels are directly used to influence the learning system; in unsupervised learning, the labels are needed only to identify and mark the output neurons. In the latter case, the categorization can be conducted more objectively. Therefore, the experiments with unsupervised Kohonen learning can be considered as an additional validation of the suitability of selected descriptors for discriminating between the DMSO(+) and DMSO(–) compounds. Another, more important reason for using the Kohonen networks is as follows. A compound's solubility is affected by a large number of different and sometimes directly opposing factors. As a result, several clusters may appear in the studied property space corresponding to different patterns of DMSO (in)solubility driving forces. Therefore, the ANN algorithm should be able to classify objects into none, 1, or more classes and not only into 1 out of 2 predefined (known in advance) existing classes. For the classification of a large number of objects with likely heterogeneity of properties (presence of different clusters), the unsupervised strategy is more efficient than the supervised one. We have chosen the Kohonen neural network among several different neural networks as the one with the most appropriate architecture and learning strategy. We used a Kohonen net with a 2D organization of the network nodes. A 15 × 15 node architecture was chosen to provide the studied molecules (49,125 compounds from the combined training and cross-validation sets) with sufficient distribution space. The smoothed projection of the combined data set of DMSO(+) and DMSO(–) compounds onto the Kohonen map was conducted using the 8 descriptors selected by principal component analysis. The DMSO(+) compounds are widely distributed throughout the map as irregularly shaped islands (Fig. 4a). The area occupied by the DMSO(+) compounds is relatively large, which reflects the excellent solubilizing ability of DMSO for the large variety of organic compounds. We suggest that the physicochemical properties of a molecule falling into the positive regions of the Kohonen map are consistent with the molecule's ability to be effectively solubilized by DMSO.

For the comparison, we separately show the area of distribution of DMSO(–) compounds within the same Kohonen map (Fig. 4b). This data set occupies 2 compact, relatively small sites on the map that are substantially different from the sites of the well-soluble compounds.

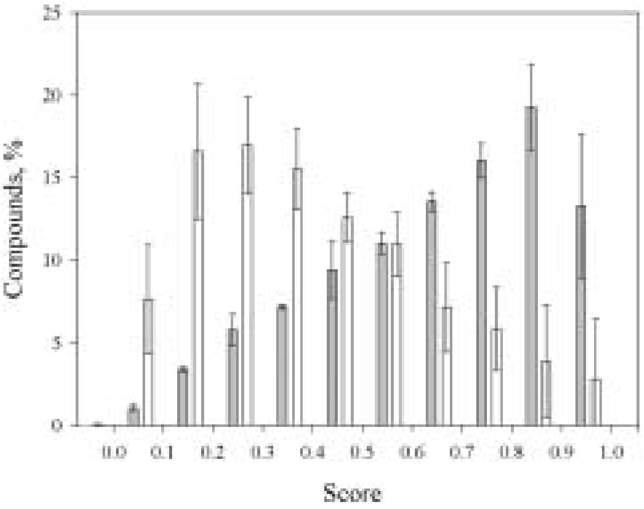

FIG. 3. Compound distributions on the scale of calculated neural network scores (supervised learning approach) for DMSO(+) (gray) and DMSO(–) (white) compounds. Averaged data with the standard deviation interval are shown for 3 independent test sets.

On the basis of these distributions, we built the smoothed contour plots of the occurrences of these 2 compound categories within the Kohonen map (Fig. 5). The area of DMSO(+) compounds is shown in gray, the area of DMSO(–) is in blue, and the low-populated area is in white. The contours correspond to at least 3% of compounds per node belonging to the particular category.

To validate the generated model, we processed the compounds from the independent test sets on the same Kohonen map. Compounds distributed within the corresponding areas shown in Figure 5 were considered as correctly classified. All the other compounds localized within the opposite or the low-populated areas were considered as misclassified. The model correctly classified up to 93% of compounds belonging to each category, as defined by their localization in the corresponding areas of the Kohonen map (Table 5).

Obviously, the unsupervised learning procedure provides much more accurate discrimination between the studied compound categories than the supervised learning, in which the percentage of correctly classified compounds was in the range of 75% to 77% (Table 5). The moderate level of discrimination observed for the back-propagated neural networks with supervised learning can be ex-

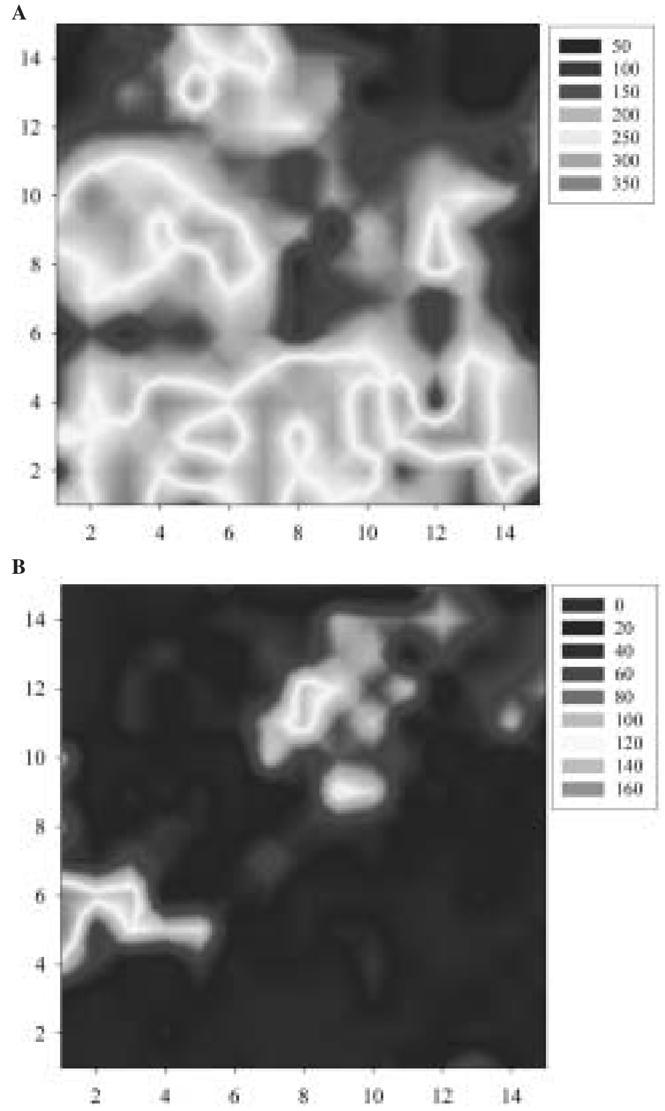

FIG. 4. Distribution of DMSO(+) (**A**) and DMSO(–) (**B**) compounds within the generated 15 × 15 Kohonen map. The corresponding areas are shown separately.

plained by the presence of several distinct clusters in the property space corresponding to separate islands on the Kohonen map. Thus, DMSO(–) compounds exist as 2 distinct types with clearly different physicochemical properties (Fig. 6). As a rule, Kohonen maps outperform the supervised learning methods in discriminating more than 2 clusters. Overall, the unsupervised strategy was more efficient than the supervised method for solving the DMSO solubility modeling problem.

Two types of compounds poorly soluble in DMSO

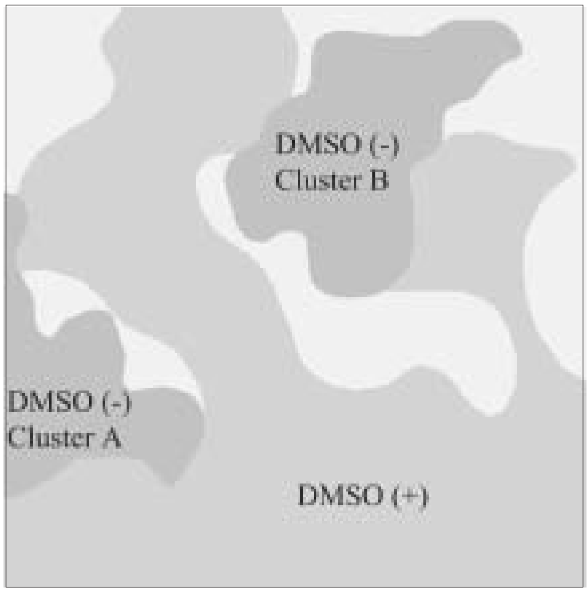

FIG. 5. Smoothed contour plots of the occurrences of DMSO(+) (gray) and DMSO(–) (blue) compounds within the Kohonen map.

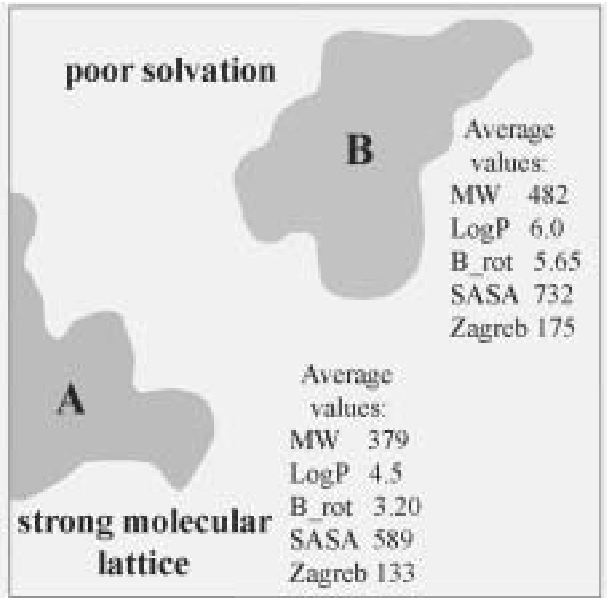

FIG. 6. Localization within the Kohonen map and average properties of 2 types of DMSO poorly soluble compounds. We hypothesize that poor DMSO solubility can be explained by the strong stability of the molecular lattice in the type A subset and low solvation ability in the type B subset.

We have found that DMSO(–) compounds are split into 2 distinct types, A and B, according to their localization within the Kohonen map (Figs. 5, 6). We compared type A and type B compounds using t' statistics calculations (Table 6) and observed a clear difference in some parameters. Type B compounds have significantly higher molecular weight, solvent-accessible surface area, radius of gyration, number of rotatable bonds, logP, and Zagreb index (t' parameter in the range of 40 to 80). They are also characterized by an increased number of H-bond donors and higher atomic charge–weighted negative surface area (t' parameter in the range of 10 to 16). The difference between the dipole moments of these 2 compound groups is negligible. The described

Table 6. *t′* Statistics Calculation for 2 DMSO(–) Clusters on the Kohonen Map

Descriptors	*Mean (Cluster A)*	*Mean (Cluster B)*	Δ[a]	t′
SASA	589.2	732.5	–143.3	83.2
Zagreb	133.4	174.7	–41.3	73.1
R_Gyr	4.75	5.74	–0.99	65.9
B_rot	3.19	5.65	–2.46	58.7
logP	4.48	5.90	–1.42	46.5
PNSA-3	–88.4	–97.7	9.3	16.4
HBD	0.34	0.50	–0.16	10.7
Dipole	9.26	9.33	–0.07	0.63

For definitions of descriptors, see Table 3.

a. Difference between mean values.

differences can be expressed as follows: type B compounds have, on average, higher molecular size and complexity, and they are more flexible and more lipophilic. On the contrary, type A compounds have relatively small, rigid, and more hydrophilic molecules.

On the basis of these differences, we suggest that these 2 types of compounds poorly soluble in DMSO principally differ in their ability to form stable molecular lattices and to be solvated by DMSO. For type A compounds, the poor DMSO solubility can be explained by strong stability of the molecular lattice. The low solvation ability can be suggested as the main factor of poor solubility in the type B subset. However, despite its consistency with the observed experimental results, this hypothesis requires further theoretical and experimental confirmation.

An ability to differentiate between the 2 types of compounds that are poorly soluble in DMSO is practically important. If our hypothesis about the main factors of poor DMSO solubility of type A and B subsets is correct, compounds of type A can be more soluble in DMSO-water mixtures than compounds of type B. Water has enhanced ability to disrupt the molecular lattices (type A) and, as a result, is less prone to solvating the lipophilic compounds (type B). As a result, compounds of type A can be more amenable for the bioassays usually performed in DMSO-water media.

CONCLUSIONS

As part of an ongoing program to eliminate DMSO-insoluble compounds from the CDL corporate collection of drug-like compounds, we developed an automated classification procedure for in silico assessment of DMSO solubility of organic compounds. The study was conducted on a large reference database with proprietary data on compound DMSO solubility, using statistical methods of data mining, including advanced machine-learning approaches. We have shown that DMSO well-soluble and poorly soluble compounds can be effectively differentiated based on a specific combination of their physicochemical properties and generated classification quantitative structure-solubility models using the neural network approach. In our study, in classification ability, Kohonen neural networks outperformed the more frequently used models trained with the supervised learning method. The Kohonen net models correctly classified up to 93% of DMSO(+) and DMSO(–) compounds from independent test sets.

As recently discussed,[15] poor DMSO solubility represents an essential problem for primary bioscreening, particularly in the high-throughput format. The groups responsible for bioscreening and libraries acquisitions would benefit if they applied our models for revealing the DMSO-insoluble compounds as a routine procedure for selecting screening molecules. Such compounds may be incompatible with the assay protocols and impair the statistics analysis as false negatives. An automated version of the described computational algorithm can be applied for "filtering" the large virtual compound collections designed for initial bioscreening programs.

Our findings are also applicable for optimizing the management of large corporate and commercial compound libraries of organic compounds. Our data indicate that up to 20% of the compounds in the commercial libraries are poorly soluble in DMSO. Using our model, compound vendors will be able to reevaluate their collections in terms of DMSO solubility and determine the optimal handling conditions for their compounds.

In this work, we separated 2 types of DMSO poorly soluble compounds corresponding to 2 distinct clusters on the Kohonen map. The presence of different clusters within the studied property space explains the limited classification ability of the multilayer neural networks with a supervised learning procedure, such as the error back-propagation learning algorithm. Although the supervised procedure is predominantly used in chemical engineering, we recommend a choice of unsupervised methods of training in this case. In light of our observations, it can be concluded that any further QSPR modeling studies of DMSO solubility should take into account the wide heterogeneity of properties of DMSO poorly soluble molecules.

ACKNOWLEDGMENTS

We would like to thank Ekaterina Lazareva and Maxim Evstratov for providing DMSO solubility data. We also thank Mary Gotschall for help in preparation of the manuscript.

REFERENCES

1. Budavari S: *The Merck Index: Encyclopedia of Chemicals, Drugs and Biologicals*. Rahway, NJ: Merck and Co., 1989.
2. Brown JH: Double-blind clinical study: DMSO for acute injuries and inflammations compared to accepted standard therapy. *Curr Ther Res* 1971;13:536-540.
3. Small A, Ide RS: Failure to detect nephrotoxicity of chronically administered dimethyl sulfoxide (DMSO) in rats. *Cryobiology* 1976;13:328-333.
4. Raegnier JF, Richard J: Lack of developmental toxicity in rats treated with dimethyl sulfoxide (DMSO). *Toxicologist* 1998;42:256-257.

5. Murdoch LA: Dimethyl sulfoxide (DMSO): overview. *Can J Hosp Pharm* 1982;35:79-85.
6. Hagelman RF, Evans TC, Riley EF: Modification of radiation effect on eye by topical application of dimethyl sulfoxide. *Radiat Res* 1970;44:368-342.
7. Wood DC, Wood J: Pharmacologic and biochemical considerations of dimethyl sulfoxide. *Ann NY Acad Sci* 1975;243:7-19.
8. Lowelock JE, Bishop MWH: Prevention of freezing damage to living cells by dimethyl sulphoxide. *Nature* 1979;183:1394-1395.
9. Hochlowski J, Cheng X, Sauer D, Djuric S: Studies of the relative stability of TFA adducts vs non-TFA analogues for combinatorial chemistry library members in DMSO in a repository compound collection. *J Comb Chem* 2003;5:345-350.
10. Kozikowski BA: The effect of freeze/thaw cycles on the stability of compounds in DMSO. Paper presented at the 7th Annual SBS Conference and Exhibition, Baltimore, September 2001.
11. Turmel M, Spreen R, Nie D: Preliminary investigation of compound stability under various storage conditions. Paper presented at the Drug Discovery Congress, Boston, August 2002.
12. Beckner C, Cheng X, Hepp D: Studies on compound stability in DMSO under different conditions. Paper presented at the Drug Discovery Congress, Boston, August 2002.
13. Yaskanin D: Quality perspectives for compound storage. Paper presented at the Automated Compound Storage & Retrieval Meeting, Somerset, UK, January 2003.
14. Fillers WS: Key factors affecting compound integrity. Paper presented at the LabAutomation Meeting, Palm Springs, CA, February 2003.
15. Lipinski CA: Storage of samples in DMSO: issues and potential solutions. Paper presented at the SBS 8th Annual Conference and Exhibition, Hague, The Netherlands, September 2002.
16. Gaylord Chemical Corp.: Solubility of active pharmaceutical compounds (APCs) in USP grade dimethyl sulphoxide (DMSO) [Online]. Retrieved from http://www.gaylordchemical.com/bulletins/PharmaSolubilities%20booklet%20bulletin.pdf
17. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Delivery Rev* 1997;23:3-25.
18. Raevsky OA, Trepalin SV, Trepalina HP, Gerasimenko VA, Raevskaya OE: SLIPPER-2001: software for predicting molecular properties on the basis of physicochemical descriptors and structural similarity. *J Chem Inf Comput Sci* 2002;42:540-549.
19. Scarsi M, Apostolakis J, Caflisch A: Comparison of a GB solvation model with explicit solvent simulations: potentials of mean force and conformational preferences of alanine dipeptide and 1,2-dichloroethane. *J Phys Chem B* 1998;102:3637-3641.
20. Song X, Chandler D: Dielectric solvation dynamics of molecules of arbitrary shape and charge distribution. *J Chem Phys* 1998;108:2594-2600.
21. Cui Q: Combining implicit solvation models with hybrid quantum mechanical molecular mechanical methods: a critical test with glycine. *J Chem Phys* 2002;117:4720-4728.
22. Jensen L, van Duijnen PT, Snijders JG: A discrete solvent reaction field model within density functional theory. *J Chem Phys* 2003;118:514-521.
23. Pierotti RA: A scaled particle theory of aqueous and nonaqueous solutions. *Chem Rev* 1976;76:717-726.
24. Mezei M, Beveridge DL: Free energy simulations. *Ann Acad Sci NY* 1986;482:1-23.
25. Straatsma TP, McCammon JA: Computational alchemy. *Annu Rev Phys Chem* 1992;43:407-435.
26. Sadowski J, Kubinyi H: A scoring scheme for discriminating between drugs and nondrugs. *J Med Chem* 1998;41:3325-3329.
27. Ajay A, Walters WP, Murcko MA: Can we learn to distinguish between “drug-like” and “nondrug-like” molecules? *J Med Chem* 1998;41:3314-3324.
28. Balakin KV, Tkachenko SE, Lang SA, Okun I, Ivashchenko AI, Savchuk NP: Property-based design of GPCR-targeted library. *J Chem Inf Comput Sci* 2002;42:1332-1342.
29. Lang SA, Kozyukov AV, Balakin KV, Skorenko AV, Ivashchenko AA, Savchuk NP: Classification scheme for the design of serine protease targeted compound libraries. *J Comp-Aid Mol Des* 2002;16:803-807.
30. Anzali S, Gasteiger J, Holzgrabe U, et al: The use of self-organizing neural networks in drug design. In Kubinyi H, Folkers G, Martin YC (eds): *3D QSAR in Drug Design: Volume 2*. Dordrecht, The Netherlands: Kluwer/ESCOM, 1998:273-299.
31. 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.
32. Nikolsky Y, Balakin KV, Ivanenkov YA, Ivashchenko AA, Savchuk NP: Intelligent machine learning technologies in pre-synthetic combinatorial design. *Pharma Chem* 2003;4:68-72.

Address reprint requests to:
Konstantin Balakin
Chemical Diversity Labs, Inc.
11558 Sorrento Valley Road
San Diego, CA 92121

E-mail: kvb@chemdiv.com