"In-House Likeness": Comparison of Large Compound Collections Using Artificial Neural Networks

Sorel Muresan* and Jens Sadowski

AstraZeneca R&D Mölndal, Pepparedsleden 1, 431 83 Mölndal, Sweden

Received October 5, 2004

Binary classification models able to discriminate between data sets of compounds are useful tools in a range of applications from compound acquisition to library design. In this paper we investigate the ability of artificial neural networks to discriminate between compound collections from various sources aiming at developing an "in-house likeness" scoring scheme (i.e. in-house vs external compounds) for compound acquisition. Our analysis shows atom-type based Ghose-Crippen fingerprints in combination with artificial neural networks to be an efficient way to construct such filters. A simple measure of the chemical overlap between different compound collections can be derived using the output scores from the neural net models.

INTRODUCTION

The compound collection is one of the most important assets of any pharmaceutical company. Having access to a large chemistry base not only provides hits and leads for new targets by using the current HTS technology but also significantly accelerates the lead optimization process. Compound collections larger than 1 million are routinely screened nowadays. These collections are replenished on a regular basis to refill some of the depleted compounds or to complement the in-house chemistry with novel compounds. Commercially and freely available databases are an important source of screening compounds, and many of them are readily accessible on the Internet. ChemNavigator,[1] MDL Screening Compound Directory,[2] Daylight,[3] and ZINC[4] are examples of large databases used in the compound acquisition process. ChemNavigator, a discovery-informatics company, provides the iResearch Library, a compilation of commercially accessible screening compounds from international chemistry suppliers. As of February 2005, the iResearch database contained 21 million structures (14 million unique) and associated data from more than 150 suppliers. Scientists within compound management groups face the challenge to select the 'best' compounds from this huge pool of available chemicals. At AstraZeneca a chemoinformatics platform supported by powerful LINUX clusters has been developed to assist chemists in the compound selection procedure. Criteria used for compound selection include the following: (i) compound availability and price; (ii) purity; (iii) leadlike and druglike filters; (iv) diversity and similarity with in-house compounds; and (v) availability of small clusters of compounds.

The classical approaches to selecting diverse sets of compounds[5] (clustering, dissimilarity-based methods, and cell-based methods) are computationally demanding methods as they are based on a pairwise comparison of the compounds under evaluation. For example, neighborhood analysis is a common way to analyze the overlap between collections of compounds.[6,7] Using various molecular descriptors and similarity measures one can compute the full similarity matrix and estimate the database overlap in terms of number or percentage compounds from set A having close neighbors within set B. Such an analysis has the time complexity $O(N^2)$ for collections of size N and becomes problematic for large databases. Several algorithms have been developed to avoid all pairwise similarity calculations and neighbor list comparisons, see for example ref 8 and references therein.

Neural networks using various sets of descriptors have been successfully applied to several classification problems in computational and medicinal chemistry. An excellent introduction into the field of neural networks for chemists is the book of Zupan and Gasteiger.[9] A few examples of such applications are mentioned in Table 1. In all these applications two classes of compounds at a time were presented to the neural net for classification. These classes were in most of the cases nonredundant, aiming at discriminating, for example, drugs from nondrugs or actives from inactives against a specific target.

In this paper we address the similarity/dissimilarity based selection within the compound acquisition process and propose a new measure for the chemical overlap between two sets of compounds. We investigate the ability of a neural net in connection with Ghose-Crippen[10] atom-type descriptors to discriminate between two unspecified sets of compounds aiming at developing filters for compound acquisition. For this purpose we have also considered redundant data sets, i.e., databases with similar and even identical compounds.

METHODS

Databases. ChemNavigator iResearch database version April 2003 (8 million samples) was used to select various data sets for the experiments described in this study. The ChemNavigator database was initially processed in order to split the salts and remove the small fragments, neutralize charges, generate canonical tautomers, and remove duplicates using in-house tools based on the Daylight[3] toolkit. This step resulted in more than 5 million unique compounds. The

* Corresponding author phone: +46-31-7065283; fax: +46-31-7763792; e-mail: sorel.muresan@astrazeneca.com.

10.1021/ci049702o CCC: $30.25
Published on Web 05/06/2005

Table 1. Examples of Neural Network Models for Classification

example	scope	neural network algorithm	descriptors[a]	ref
1	evaluate drug-likeness	feed-forward neural networks with error back-propagation	120 Ghose-Crippen atom-type descriptors	11
2	evaluate drug-likeness	Bayesian neural network	1D (logP, MW, ND, NA, NR, AR, the kappa index) and 2D (166 ISIS keys)	12
3	evaluate drug-likeness	feed-forward neural networks with error back-propagation	80 CONCORD atom-type descriptors	13
4	evaluate drug-likeness	feed-forward neural networks with error back-propagation	62 topological indices	14
5	classification scheme for serine protease targeted compds	feed-forward neural networks with error back-propagation	60 2D molecular descriptors	15
6	classification scheme for GPCR targeted compds	feed-forward neural networks with error back-propagation	8 1D and 2D molecular descriptors 60 1D and 2D molecular descriptors	16, 17
7	classification scheme for CNS targeted compds	feed-forward neural networks with error back-propagation	120 Ghose-Crippen atom-type descriptors	18
8	classification scheme for bacterial compds	feed-forward neural networks with error back-propagation	62 topological indices	19
9	classification scheme for biological targets	probabilistic neural network	24 atom-type descriptors	20

[a] The number refers to the starting point. Usually, the number of descriptors in the final neural network models is reduced.

Table 2. Molecular Property Distribution (Mean ± Standard Deviation) and Chemotype Composition for Three Different Pairs of 10 000 Compound Data Sets Used To Calibrate the Neural Networks

	MW	CLogP	PSA	selection procedure
set1a	353 ± 75	3.4 ± 1.5	64 ± 25	randomly selected from *filtered* ChemNavigator
set1b	442 ± 107	4.6 ± 2.0	82 ± 33	randomly selected from ChemNavigator
set2a	353 ± 75	3.4 ± 1.5	64 ± 25	randomly selected from *filtered* ChemNavigator
set2b	352 ± 80	3.4 ± 4.6	65 ± 27	selected from ChemNavigator using a genetic algorithm to match set2a MW, ClogP and PSA distributions
set3a	352 ± 76	3.4 ± 1.5	64 ± 25	randomly selected from *filtered* ChemNavigator
set3b	354 ± 75	3.4 ± 1.5	64 ± 25	randomly selected from *filtered* ChemNavigator

cleaned ChemNavigator was further processed using AstraZeneca's current filters for compounds acquisition. These filters include Lipinski-type properties (e.g. MW, ClogP, Polar Surface Area) and chemical filters to flag frequent hitters, reactive groups, toxic compounds, etc. Approximately 3 million compounds that passed all property and chemical filters were designated as the "filtered ChemNavigator" for the purpose of this study.

Molecular Descriptors. An extended set of the Ghose-Crippen atom-types was used as molecular descriptors for neural net modeling. The input for the neural net was a vector counting the occurrence of the 133 atom types in each molecule. If an atom type was found in more than 100 compounds in the entire training set, it was considered frequent enough to be included as a significant descriptor.

Neural Network Modeling. The 'SNNS: Stuttgart Neural Network Simulator' package was used for all neural network calculations.[21] Feedforward networks were constructed that consist of 133 input neurons (Ghose-Crippen atom types), 10 hidden neurons, and 1 output neuron (score). All layers were fully connected. For technical reasons all input and output values were linearly scaled between 0.1 and 0.9. The neural networks were trained using the 'back-propagation with momentum' scheme as implemented in SNNS. The training was performed over 2000 cycles with a learning rate of 0.2 and a momentum term of 0.1. The training data set was shuffled before each cycle, i.e., the training data were in each training cycle presented to the neural network in a new order. Test runs showed that the training process achieved sufficient convergence with these parameters.

Diversity Analysis. Diversity of single and merged databases was evaluated using the centroid-based algorithm proposed by Willett et al.[22] and modified by Trepalin et al.[8] to handle large collections of compounds. This measure provides a numerical description of the similarity relationships that exist within a data set. It also gives a measure of the change in diversity when two databases are merged. The same Ghose-Crippen fingerprints were used as molecular descriptors to assess database diversity. Each molecule was represented by a vector containing 133 bins. In turn, each database is represented by a virtual molecule, a centroid, or a linear combination of the individual molecule vectors. This centroid is the basis of diversity assessment of a database or the change in diversity when databases are merged.

RESULTS AND DISCUSSIONS

Test Case 1. Three pairs of 10 000 compound sets were randomly selected from ChemNavigator database in order to assess the ability of a neural network model to discriminate between sets of compounds with different degrees of similarity. Random selections from the same pool of compounds (e.g. filtered ChemNavigator) will be similar with respect to both physicochemical properties and chemical makeup, while sets of compounds from different pools, in this case filtered vs not-filtered ChemNavigator, are expected to be less similar due to the constraint of the property and chemical filters used for the latter. The selection procedure is summarized in Table 2.

The set1a and set1b data sets have different property distributions and consist of different chemotypes as can be seen from Table 2. These two sets are nonredundant. The set2a and set2b data sets have similar property distributions but different chemotypes. A genetic algorithm was used to

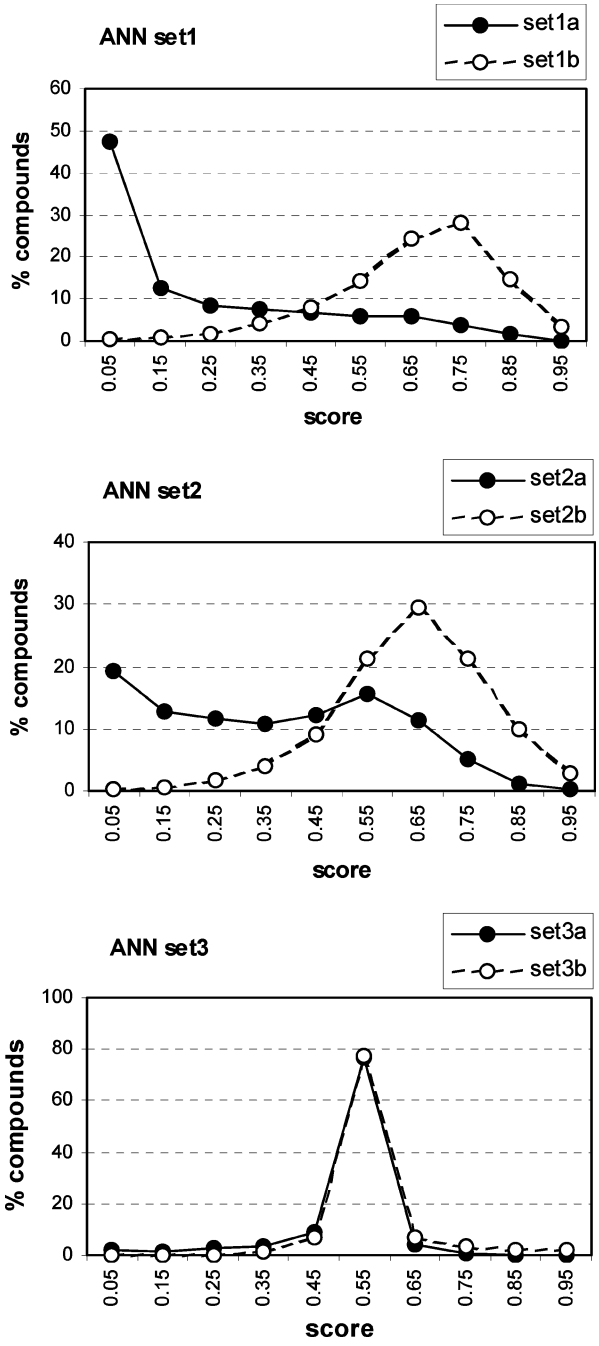

Figure 1. Distribution of neural network predicted scores for three pairs of compounds with different degrees of similarity.

match the property distributions within this pair.[23] The algorithm minimizes the F-test statistics between two data sets in order to obtain a nonsignificant difference between the normal distributions of the three properties. Finally, the set3a and set3b data sets are similar in both property distributions and chemotype composition. These data sets are highly redundant. It has to be mentioned that redundancy here means that the compounds in their Ghose-Crippen fingerprint vectors, bin for bin, are similar. All data sets contained unique compounds, and no exact matches were allowed within a pair.

The optimal network configuration was found by training and evaluating different network architectures. The distribution of the neural network scores for the three pairs of compound sets is presented in Figure 1. There is a clear separation between the filtered and not-filtered compounds. With a scoring value of 0.5 as the natural (symmetric) cutoff for the classification, 83% of the set1a compounds and 85% of the set1b compounds were correctly classified. A somewhat weaker model was obtained for the second pair where 71% of the set2a compounds and 78% of the set2b compounds were correctly classified. For the third pair, having highly similar compounds in the two sets, no relevant model was obtained. As expected, the high redundancy between the two sets introduces noise and drastically reduces the classification power of the neural network. However, the message we get from this analysis is still valuable: the data sets contain similar compounds.

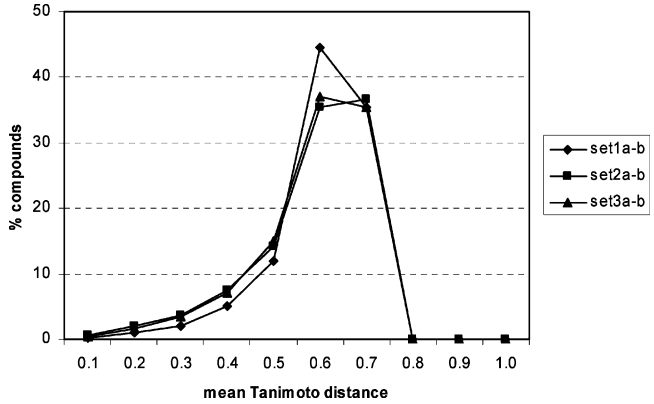

Figure 2. An all-by-all comparison using the mean pairwise Tanimoto distances for the data sets described in Table 2.

A simple measure of the 'chemical overlap' (CO) between two sets of compounds can be derived using the output scores from the neural network models

$$CO = 100 - \frac{1}{2}\sum_{bin=1}^{10} |\%\ cmpdsSetA - \%\ cmpdsSetB|_{bin}$$

where *bin* is a 0.1 step on a score scale between 0 and 1. The CO score is calibrated between 0 and 100, and the higher the score the larger the overlap. The CO values for the three pairs, 32, 49, and 90, respectively, reflect the similarity between the data sets within each pair.

We have compared the same data sets by computing the mean pairwise intermolecular Tanimoto distances: each molecule from set1a, for example, was compared with all molecules from set1b, and the average Tanimoto distance based on Ghose-Crippen descriptors was plotted. The distributions presented in Figure 2 are rather similar, mean ± stdev: set1 (0.55 ± 0.10), set2 (0.53 ± 0.12), and set3 (0.53 ± 0.12). This suggests a similar overlap between the three pairs, in contradiction with the neural network results and the way we have selected the compounds.

Test Case 2. Full compound collections from five vendors were pulled out from the ChemNavigator database based on two criteria: the collection size and their mutual overlap in terms of exact matches. We aimed to select compound sets of different sizes and with various degrees of chemical overlap. Duplicate compounds within each set were removed, and no chemical or property filters were applied to the data sets. The actual supplier is not relevant for this study, and similar sets could have been selected from other available databases. The total number of unique compounds per collection (diagonal) and the number of exact matches for

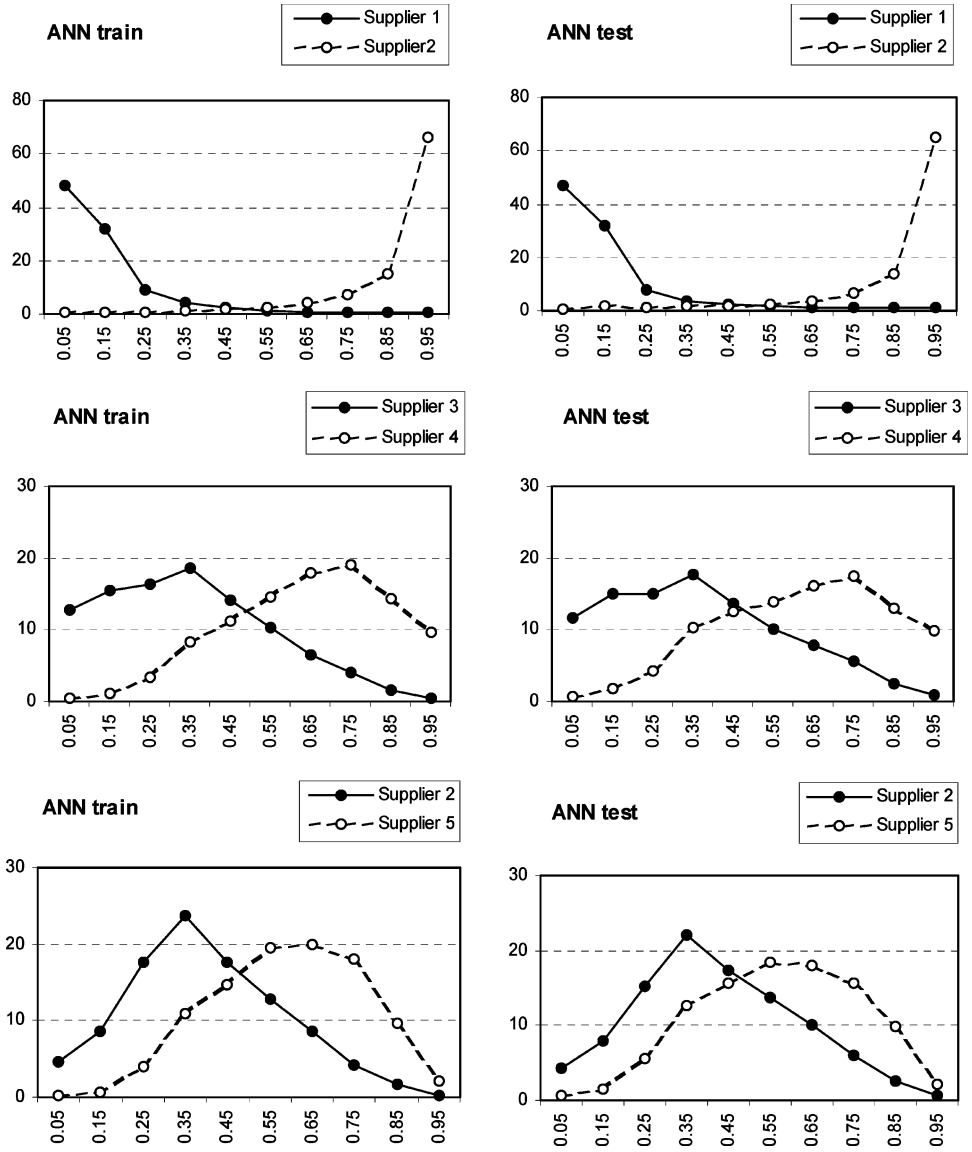

Figure 3. Score distribution (% compounds) for the test and training sets using compounds from different suppliers, see also Table 4.

Table 3. Compound Collections from Different Suppliers Selected from ChemNavigator Database[a]

	supplier 1	supplier 2	supplier 3	supplier 4	supplier 5
supplier 1	**113 370**	226	79	179	2315
supplier 2		**235 012**	38 998	38 567	28 704
supplier 3			**287 353**	16,386	21 111
supplier 4				**83 336**	6883
supplier 5					**162 489**

[a] The diagonal presents the total number of unique compounds while each cell presents the exact matches between pairs of collections.

each pair are presented in the Table 3. The number of exact matches provides the first indication of chemical similarity between the data sets.

Neural networks were trained to distinguish between sets of compounds from different suppliers. In each case, 10 000 compounds per supplier were randomly selected for training, and the rest of the compounds were used as a test set. In this way, the model is not biased toward the collection with more compounds. The neural network models and the chemical overlap derived from the training sets are presented in Table 4. Some examples of the performance of models built using compound collections with different degrees of chemical overlap are presented in Figure 3.

Table 4. Neural Network Models for Compound Collections from Various Suppliers[c]

		% train set correct		% test set correct		
model	NN arch[a]	setA	setB	setA	setB	CO[b]
supplier 1−supplier 2	77−10−1	91.4	96.6	88.9	95.3	11
supplier 1−supplier 3	77−10−1	92.5	96.2	90.4	94.3	11
supplier 1−supplier 4	71−10−1	96.8	92.9	93.5	87.4	10
supplier 1−supplier 5	73−10−1	93.2	94.6	89.3	93.1	12
supplier 2−supplier 3	84−10−1	67.8	81.2	85.4	76.3	51
supplier 2−supplier 4	78−10−1	93.7	33.6	91.0	29.7	59
supplier 2−supplier 5	83−10−1	90.9	37.1	88.7	33.5	63
supplier 3−supplier 4	78−10−1	97.3	35.0	96.1	29.7	51
supplier 3−supplier 5	82−10−1	90.9	53.9	88.8	50.4	49
supplier 4−supplier 5	77−10−1	59.7	88.0	53.7	85.1	48

[a] Neural network architecture: 77−10−1 means 77 input neurons (Ghose-Crippen atom types), 10 hidden neurons, and 1 output neuron (score). All layers were fully connected. [b] Chemical overlap obtained from the neural network output scores for the training data sets. [c] In each case 10 000 compounds per data set were used for training.

Table 5. Diversity of Single and Merged Databases Based on Ghose-Crippen Fingerprints[a]

	supplier 1	supplier 2	supplier 3	supplier 4	supplier 5
supplier 1	**0.225**	0.018	0.030	0.024	0.031
supplier 2		**0.246**	0.009	0.007	0.010
supplier 3			**0.262**	0.003	0.003
supplier 4				**0.272**	−0.001
supplier 5					**0.269**

[a] The diagonal elements contain the diversity of the single database. The *ij*th element of the table contains the change in diversity when the database in the *j*th column is added to the database in the *i*th row.

For comparison, the diversity of the selected databases was estimated using the algorithm developed by Willet et al.[16] on the same training sets and Ghose-Crippen fingerprints. The Willet diversity measure has the advantage that it is fast to compute and therefore can be used for large databases. The analysis is summarized in Table 5 where the diagonal elements contain the diversity of the single databases, while the *ij*th element of the table contains the change in diversity when the database in the *j*th column is added to the database in the *i*th row.

Data in Table 5 show a similar diversity for the databases analyzed. When merged with other databases the diversity of the data set from supplier 1 increases by 8%, while the others stay more or less the same. This supports the idea that increasing the size of a collection does not necessarily cause a corresponding increase in diversity. In some cases, merging two databases is more likely a hole-filling process than adding novel compounds which might explain why the diversity remains practically unchanged. It also points out the current problem of many diversity functions, the ability to correctly handle redundant information. The ideal properties of a molecular diversity measure is discussed by Walman et al.[24] and Harper et al.[25]

The chemical overlap derived by integrating the output scores from the corresponding neural network model provides a better picture of database similarity, Table 4. This parameter offers a simple and rapid estimation of the chemical similarity between two compound collections within a data set. Furthermore, the neural net seems to be able to handle the redundancy given by identical compounds in two data sets. For example, the neural net model corresponding to the supplier 2−supplier 3 pair with ∼15% common structures was able to correctly classify more than 75% of both sets.

When a corporate database is used as one of the data sets for training the neural network, this approach provides an efficient "in-house likeness" filter. Such a filter can be used for compound acquisition and other high-throughput applications to rapidly select similar/dissimilar compounds from a certain data set. In our experience such an in-house likeness filter can reduce significantly the time required for compound selection without overlooking classes of compounds. This is particularly relevant when considering that many suppliers operate on a first come, first served basis, and the selected compounds might not be available by the time a purchasing order is made.

CONCLUSIONS

Neural network models using Ghose-Crippen atom-type descriptors are able to correctly classify compounds from collections of different provenance. As expected, the chemical overlap between two data sets greatly affects the quality of the model. A simple measure of this chemical overlap can be estimated from the output scores of the neural net models.

When a corporate database is used as one of the data sets, this approach provides an efficient "in-house likeness" filter able to discriminate between in-house and external compounds. These filters can be used for compound acquisition as they are able to rapidly screen a large collection of compounds.

ACKNOWLEDGMENT

We thank Niklas Blomberg, Hongming Chen, Dave Cosgrove, and Peter Kenny (AstraZeneca) for providing tools used in this study.

Supporting Information Available: The Ghose-Crippen atom-type descriptors for the three pairs of compound sets described in Table 2. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) ChemNavigator.com, Inc.; http://www.chemnavigator.com
(2) MDL Information Systems, Inc; http://www.mdl.com
(3) Daylight Chemical Information System, Inc.; http://www.daylight.com
(4) 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.
(5) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Kluwer Academic Publishers: Dordrecht, Boston, London, 2003.
(6) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: using the MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(7) Bradley, M. P. An overview of the diversity represented in commercially available databases. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 301−309.
(8) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. P.; Ivaschenko, A. A. New diversity calculations algorithms used for compound selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 249−258.
(9) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists. An Introduction*; VCH: Weinheim, 1993.
(10) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure−Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(11) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(12) Ajay; Walter, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "non drug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(13) Frimurer, T. M.; Bywater R.; Nærum, L.; Lauritsen, L. N.; Brunak, S. Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315−1324.
(14) Murcia-Soler, M.; Péréz-Giménez, F.; García-March, F. J.; Salabert-Salvador, M. T.; Díaz-Villanueva, W.; Castro-Bleda, M. J. Drugs and nondrugs: an effective discrimination with topological methods and artificial neural networks. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1688−1702.
(15) Lang, S. A.; Kozyukov, A. V.; Balakin, K. V.; Skorenko, A. V.; Ivashchenko, A.; Savchuk, N. P. Classification scheme for the design of serine protease targeted compound libraries. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 803−807.
(16) Balakin, K. V.; Tkachenko, S. E.; Lang, S. A.; Okun, I.; Ivashchenko, A. A.; Savchuk, N. P. Property-based design of GPCR-targeted library. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1332−1342.
(17) Balakin, K. V.; Lang, S. A.; Skorenko, A. V.; Tkachenko, S. E.; Ivashchenko, A. A.; Savchuk, N. P. Structure-based versus property-based approaches in the design of G-protein-coupled-receptor-targeted libraries. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1553−1562.

(18) Engkvist, O.; Wrede, P.; Rester, U. Prediction of CNS activity of compound libraries using substructure analysis. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 155−160.
(19) Murcia-Soler, M.; Péréz-Giménez, F.; García-March, F. J.; Salabert-Salvador, M. T.; Díaz-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.
(20) Niwa, T. Prediction of biological targets using probabilistic neural networks and atom-type descriptors. *J. Med. Chem.* **2004**, *47*, 2645−2650.
(21) SNNS: Stuttgart Neural Network Simulator; http://www-ra.informatik.uni-tuebingen.de/SNNS/
(22) Turner D. B.; Tyrrell S. M.; Willett P. Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.
(23) Blomberg, N.; Muresan, S. Work to be published.
(24) Waldman, M.; Li, H.; Hassan, M. Novel algorithms for the optimisation of molecular diversity of combinatorial libraries. *J. Mol. Graphics Modell.* **2000**, *18*, 412−426.
(25) Harper, G.; Pickett, S. D.; Green, D. V. S. Design of a compound screening collection for use in High Throughput Screening. *Comb. Chem. High Throughput Screening* **2004**, *7*, 63−70.

CI049702O