

Surrogate data – a secure way to share corporate data

Igor V. Tetko^{a,d,*}, Ruben Abagyan^b & Tudor I. Oprea^c

^a*Institute of Bioorganic and Petroleum Chemistry, Ukrainian Academy of Sciences, Kyiv, Ukraine;*

^b*Molecular Biology, The Scripps Research Institute, La Jolla, CA, 92037, USA;* ^c*Division of Biocomputing, University of New Mexico School of Medicine, University of New Mexico, 87131-0001, Albuquerque, NM, USA;* ^d*GSF – Forschungszentrum fuer Umwelt und Gesundheit, GmbH, Institute for Bioinformatics, D-85764, Neuherberg, Germany*

Received 17 May 2005; accepted 6 August 2005

© Springer 2005

Key words: drug design, structure–property prediction, information content of a molecule, representation of molecules, surrogate data, lipophilicity prediction

Summary

The privacy of chemical structure is of paramount importance for the industrial sector, in particular for the pharmaceutical industry. At the same time, companies handle large amounts of physico-chemical and biological data that could be shared in order to improve our molecular understanding of pharmacokinetic and toxicological properties, which could lead to improved predictivity and shorten the development time for drugs, in particular in the early phases of drug discovery. The current study provides some theoretical limits on the information required to produce reverse engineering of molecules from generated descriptors and demonstrates that the information content of molecules can be as low as less than one bit per atom. Thus theoretically just one descriptor can be used to completely disclose the molecular structure. Instead of sharing descriptors, we propose to share surrogate data. The sharing of surrogate data is nothing else but sharing of reliably predicted molecules. The use of surrogate data can provide the same information as the original set. We consider the practical application of this idea to predict lipophilicity of chemical compounds and we demonstrate that surrogate and real (original) data provides similar prediction ability. Thus, our proposed strategy makes it possible not only to share descriptors, but also complete collections of surrogate molecules without the danger of disclosing the underlying molecular structures.

In spite of an increased investment in preclinical pharmaceutical research, the overall number of new drugs registered by regulatory agencies remained approximately unchanged over the last decade. It is estimated that the total pre-approval cost of production of a new drug is in the range from US \$800 millions [1] to more than \$1.7 billion [2]. The cost of drug development is currently 55% higher than the average cost from 1995 to 2000 and it is rising largely as a result of an

increasing failure rate for prospective drugs in clinical trials [2]. One of the main reasons of high cost of drug production is the failure of candidate drugs in the later phases of clinical testing due to poor pharmacokinetic and toxicological (ADME/T) profiles of these compounds. Thus, many pharmaceutical companies are changing their strategy for drug development by emphasizing ADME/T properties in the early phases of drug discovery. Both biological and physicochemical properties of compounds are considered in the early stages. This strategy, however, requires the evaluation of a large number of compounds, which

*To whom correspondence should be addressed. Fax: +49-89-3187-3585; E-mail: itetko@vcclab.org

is both time- and cost-intensive. The development of reliable computational models can dramatically speed up this phase of drug development and decrease the costs.

Despite the growing number of commercial software products available for ADME/T property calculations, the quality of predictions remains poor even for a simple and well-characterized property such as the 1-*n*-octanol/water partition coefficient (logP) [3, 4]. The general explanation of this phenomenon is the inadequate availability of data that can be used to develop and test new programs. Data sources are usually limited to what is disclosed in the public domain; at the same time, the private sector has produced relatively large amounts of in-house data, which could be used to enhance the publicly available datasets and thus lead to the development of more reliable programs [5]. Extended to properties such as cardiac toxicity (e.g., Q-T prolongation syndrome) and other difficult-to-evaluate properties that are relevant in the clinic, an effort to increase the amount and quality of data available in the public domain could be mutually beneficial to the public at large (patients), and to the pharmaceutical companies that participate in such data exchange, as the quality of ADME/T predictions is likely to improve dramatically.

Unfortunately, the pharmaceutical houses are unlikely to freely release such data to the public domain. The reason stems in the competitive nature of this business, and the need of the private sector to keep secure the structures of those chemicals that have not been disclosed in patents or other publications. Such information could benefit competitors and may hurt the interests of the company disclosing data. Therefore, the interest in developing methods for the safe exchange of chemical information has increased.

In the current article we analyze some theoretical limits on the number of descriptors that can be used to decode molecules from its 2D-descriptors. The limit appears to be prohibitive and thus chemical descriptors may at least in theory lead to disclosure (via reverse-engineering) of the molecular structure from their descriptors. Therefore, we propose a new strategy based on encoding molecules as surrogate data. Furthermore, we apply this procedure to prediction of lipophilicity of chemical compounds.

Data

The data used in this study were collected from two sources. The first collection (Pub_NZ) included molecules publicly available from the LMC Chemical Information Services available at the National Cancer Institute (<http://cactus.nci.nih.gov/PubDBs>) and a number of different vendors collected at ZINC database [6]. The SMILES from databases with at least 10,000 were selected from the LMC site and all databases were downloaded from the ZINC. The second collection includes molecules available at iResearch™ library, Chem-Navigator (<http://www.chemnavigator.com>). The disconnected molecules (such as salts), molecules that could not be processed with ALOGPS program [7] (freely available at <http://www.vcc-lab.org>) and molecules without carbon atoms were excluded from the analysis. For each collection we extracted all the unique, non-stereoisomeric SMILES. Statistic for both datasets is shown in Table 1.

Method

The Associative Neural Network (ASNN) [8, 9] represents an ensemble of N neural networks

Table 1. The statistics of analyzed databases.

Database	SMILES	atoms ¹	size, bytes	atoms ^a /molecule	method	compressed size, bytes	bits/atom
Pub_NZ	2,454 537	64,431 783	117,477 654	26.3	gzip 1.3.3	28,877 816	3.58
					WinRar 3.42	16,686 921	2.07
					bzip2 1.0.2	20,068 848	2.49
IResearch library	13,333 629	405,609 998	712,299 408	30.4	gzip 1.3.3	80,857 933	1.59
					WinRar 3.42	47,648 330	0.94
					bzip2 1.0.2	53,946 272	1.06

^aNumber of non-hydrogen atoms.

trained on various subsets of the initial training set. Naturally, for the current analyses the molecules were represented by their descriptors. Each row of indices (data case) corresponded to one molecule. The learning dataset of each network in the ensemble is selected by chance as 50% of the initial training set. The remaining part of the initial dataset, the validation set, is not used directly to update neural network weights but is applied to terminate the neural network training when performance of the network on the validation set is maximum. This training process is known as early stopping over ensembles and it is well illustrated by Figure 2 from Ref. 10. Following the training of the ensemble, one can calculate an ensemble average

$$\bar{y}_i = \frac{1}{N} \sum_{j=1, \dots, N} y_i^j \quad (1)$$

where y_i^j is calculated value of neural network j for data case i and \bar{y}_i is ensemble average for the same case. Indeed, since we have an ensemble of $j=1, \dots, N$ networks (models), each network, j , will calculate for each data case, i , one value, y_i^j . The formula (1) corresponds to a simple way one can deal with ensemble of predictions, i.e. just to average all values. The use of ensemble averages decreases the variance of neural networks. The ASNN introduces a similarity of data cases i, t in space of residuals, $y_i^j - \bar{y}_i$, of neural network (NN) as Pearson

$$\rho(i, t) = \frac{\sum_{j=1, \dots, N} (y_i^j - \bar{y}_i)(y_t^j - \bar{y}_t)}{\sqrt{\sum_{j=1, \dots, N} (y_i^j - \bar{y}_i)^2 \sum_{j=1, \dots, N} (y_t^j - \bar{y}_t)^2}} \quad (2)$$

or Spearman rank correlation coefficients (in the later case the values y_i^j are substituted by their ranks).

The prediction of the ASNN is calculated as

$$\bar{y}_i' = \bar{y}_i + \frac{1}{K} \sum_{j=1, \dots, K} (f_j - \bar{y}_j) \quad (3)$$

where f_j is experimental logP value of compound j and the summation is performed over K – nearest neighbors (this number is optimized in the ASNN method) determined with Equation 2 of the analyzed case i .

The ALOGPS 2.1 program [7–9] used in this study was developed with the ASNN (64 networks in the ensemble), based on 75 E-state indices [11–13] and 12,908 compounds from the PHYS-

PROP database [14]. A molecule calculated with the ALOGPS program is represented by a set of 64 ranks, i.e., a set of 64 integer values. The ranks 1 and 64 corresponded to the largest and the smallest values, respectively, predicted by the neural networks in the ensemble. Examples of different representations of *Morphinan-3-ol*, *17-methyl-* and *Levallorphan* are shown in the supplementary materials (Table A1). The ALOGPS ranks were calculated for all analyzed molecules and were used for all studies reported in this article.

Real/Surrogate data selection: For each analyzed molecule from the PHYSPROP database correlation coefficients to each molecule from the reference database (Pub_NZ or iResearch library) were calculated using Spearman rank correlation. Thus for each PHYSPROP molecule we calculated 2,454,537 and 13,333,629 correlation coefficients to molecules from PUB_NZ and iResearch libraries, respectively. A threshold of $r^2=0.3$ was used to filter out small and negative ($r<0$) correlations. Thus each PHYSPROP molecule had two lists of molecules correlated to it above the threshold value, i.e. one from Pub_NZ and one from the iResearch library. Then, for each PHYSPROP molecule we selected a molecule from the corresponding list that had the D th largest correlation coefficient, where D is the specified dilution rate ($D=100$ or $D=1000$). The selected molecules formed the *surrogate datasets*, i.e. one set per library. If there were less than D molecules in a list, no surrogate molecule for this molecule was selected for the corresponding surrogate dataset. The molecules from PHYSPROP for which the surrogate molecules were found formed so-called *real dataset*. A pseudo-code for this procedure is shown in attachment (Schema A1). By derivation, the real and the surrogate data sets had identical sizes. A comparison of the models developed using real and surrogate subsets provided a good estimation of the quality of surrogate data set and its ability to substitute the original molecular structures.

ASNN models: The neural network programs developed with surrogate and real dataset were trained using the same neural network parameters and descriptors used to develop the original ALOGPS program [7].

3D models: The molecules from surrogate/real dataset were converted from SMILES (unique,

non-stereoisomeric representation) using Corina [15]. The Dragon software [16] was used to calculate 763 3D descriptors for each molecule. The calculated 3D descriptors were de-correlated (for each pair of descriptors with $r^2 > 0.8$ we eliminated each second descriptor in the data file) and analyzed using the Multiple Linear Regression program. At the end of the training, the developed models were used to predict all molecules from the PHYSPROP database.

Results

Why can we not share molecules at the level of descriptors?

The possibility to disclose molecules via descriptors depends on the availability of software to generate such descriptors (indices) and on molecular libraries or/and software to generate molecules that step-by-step converge towards the descriptor values of the target molecule(s). Problem-solving methods such as genetic algorithms have been applied to reach target descriptor values, e.g., the GROK method [17].

If the software used to generate the descriptors is unknown and/or not generally available, the task of identifying the molecule becomes impossible. Such approach is indeed very secure and is used in common projects between companies [18]. However, this case cannot be considered as a sharing of information since such data could not be used to predict new molecules outside of the company that initially generated the data.

Two possible scenarios can occur when disclosing molecules with associated descriptors. We will assume that a “hacker” is interested in guessing the molecular structures. In one scenario, the “hacker” has access to a virtual library that already contains the molecule in question. If each molecule has a unique representation in the descriptor space, then the hacker simply needs to differentiate between the descriptors of the molecule in question and the descriptors associated to all other molecules in the set. The problem arises when several molecules have exactly the same descriptors referred to as “confused descriptors” [19]. Thus, using “confused descriptors” makes it rather difficult to unambiguously assign one structure to an associated descriptor space. If the

number of molecules with identical descriptors is relatively small, i.e. less than 100, the “hacker” can use all such molecules for his analysis and the approach cannot be considered as secure.

A more complex scenario occurs when the molecule in question is absent in the virtual library. In this case the “hacker” can analyze all molecules in the available virtual library, select molecules that best match the descriptors of the molecule in question, e.g. using some similarity metric between the target molecule and the virtual compounds, and use these molecules as starting points for an exhaustive search with, e.g. EA-inventor [20]. Eventually after some steps, the “hacker” can identify a molecular structure that will provide a perfect match to the target descriptors. It is possible that the currently existing algorithms cannot always find the perfect match, i.e., they can get trapped in local minima and unable to converge to the ideal solution, the target molecule. However, one cannot rely on this assumption. Therefore, let us assume that the target molecule is available in the database, i.e. we consider only the first scenario.

Let us analyze how many descriptors may contain sufficient information to unambiguously decode a molecule. SMILES codes represent a simple and efficient coding of molecular structures using a string of symbols. All further information about the molecule can be easily derived from the analysis of SMILES string alone. The information content of a string is given by the Minimum Description Length (MDL) or a minimum number of bits required to store the string and decode it without loss of precision [21]. The MDL is a theoretical number since, in general, the algorithm that provides such optimal encoding is unknown. However, an approximate estimation of this number can be achieved using data compression algorithms.

In order to estimate the information content of molecules we used popular data compression algorithms, gunzip and bzip2 (Linux) and WinRar (Windows). The original and compressed sizes of SMILES for a data file containing SMILES from Pub_NZ and iResearch datasets are shown in Table 1.

The bit rate per molecule (per atom) is defined as a number of bits required on average to store and decode the molecule (atom) without lost of information using corresponding data compression algorithm.

The minimal bit rate calculated by WinRAR is approximately equal to 2 bits/atom for the Pub_NZ data set and less than 1 bit per atom for the iResearch™ dataset. Similar rates were reported elsewhere [22]. This result indicates that Pub_NZ data set contains on average more diverse structures than iResearch. The information content of a molecule with 35 atoms from the later dataset is on average just 32 bits. The actual MDL number, however, can be even lower since it is possible that the SMILES based representation of molecules and/or the used (text compression) algorithms are not the optimal ones. For example, one could probably achieve better compression rates by using one-symbol encoding of aromatic rings which are abundant elements in the both datasets.

The calculated number puts a severe limitation on a number of descriptors that can be derived from a molecule without disclosing its structure. For example, float value index uses 32 bits. Thus, at the theoretical limit, as little as one float index could be sufficient to completely recover a molecule with 35 atoms from the iResearch library! This number is not related to the nature of the descriptors, e.g. topological, quantum-mechanical or 3D, but to the information content of the descriptors. The limit also does not assume the existence of an algorithm that is able to reverse-engineer the structure of a molecule from its descriptors. Some descriptors could be easier to decode, while others may require the development of more complex algorithms. However, the absence of an algorithm (or the absence of knowledge that such algorithm does exist) cannot be considered as insurance that reverse engineering is impossible. Thus, unfortunately, we need to conclude that molecular descriptors can be, at least in theory, used to reverse engineer the underlining molecular structure.

Surrogate data

So far we considered the case of releasing descriptors generated from proprietary data to the public domain. But is this the only way to disclose the data?

One of us, Ruben Abagyan, proposed to simplify the exchange and distribute molecular structures along with the data with one caveat, the structures will be different and only closely resem-

ble the originals. This method does not require any predefined indices and special software to generate, read, and interpret them [23]. The permutation may include exchange of rings and groups, which would make it impossible to recognize the original molecule. This approach, however, will not work unless *we know* which changes should be allowed (i.e., they should not influence the structure–property relationships (SPR) with the target property) and which changes should be excluded (since they dramatically alter the SPR). This is only possible *by modeling* the given property. If we have a model, we can predict which changes should be allowed and which are not. However, within the model we can also predict how each alteration of the molecular structure *will influence* the property of the molecule. Thus, assuming one can reliably predict new compounds and we are confident about our accuracy, one could simply release the predicted molecules, i.e. surrogate data, instead of the original compounds used to develop the model. The surrogate molecules should be accompanied by surrogate activities, i.e. activities that are predicted by the model for these molecules.

Thus, instead of releasing descriptors of real molecules and be exposed to reverse engineering attempts, the companies may release surrogate data associated with complete (but fake) molecular structures. Such data could still be valuable to the public domain, since it could be used to develop new methods that could result in improved predictions for chemical compounds. Assuming the surrogate data set is prepared with some minimal caution, such disclosures would make reverse engineering efforts impossible, and the original structures could not be recovered.

Below we consider a practical example to develop a program to predict lipophilicity of chemical compounds using surrogate data.

Surrogate logP models

In our previous studies we introduced the distance between molecules in space of models as a correlation of residuals of the ensemble of neural network models [8, 9]. The same distance can be used with any ensemble of models, not necessarily based on neural networks. It was also shown that the distance in space of models can be used to evaluate a quality of predictions in ALOGPS 2.1 program [24]. Molecules with a low maximum

correlation coefficient to any other molecule in the training set are likely to have large prediction error. For example, if the maximum correlation of a new molecule to any molecule in the training set is less than 0.3, the expected error is > 0.7 log. The expected error is however 0.3 log unit if maximum correlation to any molecule in the training set is $r^2 > 0.6$ [24].

The correlation coefficients calculated between all molecules in Pub_NZ and iResearch datasets and PHYSPROP database are shown in Figure 1. We also include the Spearman correlation coefficients calculated by random shuffling of the ranks order. This later distribution evaluates the noise. In first, it is clear that correlations $r^2 > 0.35$ are non-random. Furthermore, Pub_NZ contains more molecules that are more similar to the PHYSPROP training set molecules, compared to the iResearch library. This result is not surprising since Pub_NZ incorporates a large number of small molecules from NCI and PubChem that are not always “drug-like”. On the contrary, iResearch contains commercial compounds catered to the pharmaceutical industry, which are expected to be more “drug-like”. Since most PHYSPROP compounds are not “drug-like”, the PHYSPROP and Pub_NZ data sets are more similar.

Even though Pub_NZ is 5 times smaller than iResearch, one can expect to find on average, for any molecule in PHYSPROP, $N_{\text{sel}} = 96$ and $N_{\text{sel}} = 142$ molecules in Pub_NZ and iResearch datasets with the correlation coefficient of $r^2 \geq 0.3$, respectively. Of course, because of the chance

effect, one can select some completely unrelated molecules. The noise curve predicts that for the same databases, one can find $N_{\text{noise}} = 12$ and $N_{\text{noise}} = 58$ correlations with $r^2 \geq 0.3$ completely by chance. Thus, if we use this threshold and select surrogate molecules, the chance to select non-related molecules with $r^2 \geq 0.3$ is 8:1 for Pub_NZ and 2.5:1 for iResearch. The standard way to measure the signal to noise ratio is $20 \cdot \log(N_{\text{sel}}/N_{\text{chance}})$; the corresponding curve for both databases is shown at Figure 2.

The molecules in the PHYSPROP database range from very simple, e.g. ethane (MW = 30), to very large ones, e.g. Pengitoxin (CAS RN 7242-04-8, MW = 991). Depending on their structure they can have different number of molecules correlated to them above some threshold. Figure 3 demonstrates the distribution of the number of molecules in PHYSPROP having more than N molecules with $r^2 \geq 0.3$ in iResearch and Pub_NZ. There are 11 molecules in PHYSPROP each having more than 8192 molecules correlated above the threshold in iResearch library.

The selection of the most correlated molecules in the iResearch or PubNZ datasets would provide the best surrogate data to develop the model. However, there is the possibility that these molecules share the same chemotype. It is therefore safer to order all molecules correlated with the proprietary one, and to select as the surrogate molecule not the nearest but, e.g., the $D = 100$ th or even $D = 1000$ th molecule in the list. We will call the number D the dilution rate. Thus, the

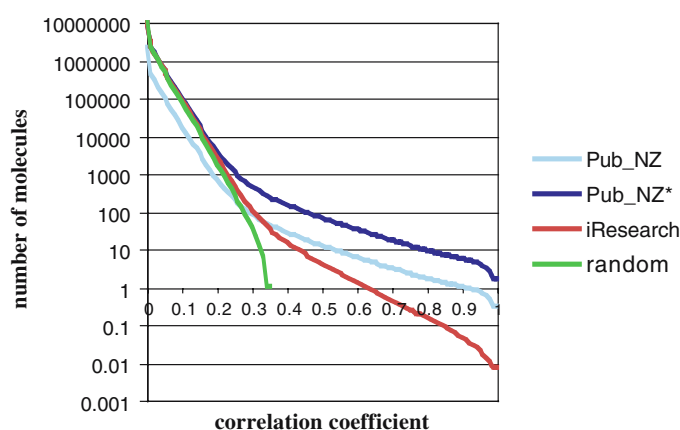


Figure 1. The average number of molecules in respective databases with higher or equal correlation coefficient to an average molecule in the PHYSPROP database. The Pub_NZ results (gray) were multiplied by 5.4 (Pub_NZ*), to match the number of compounds in iResearch library.

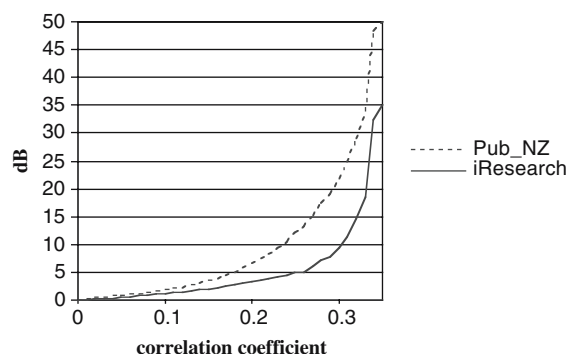


Figure 2. Signal/noise ratio to have a PHYSPROP molecule correlated to molecules in each databases above the threshold.

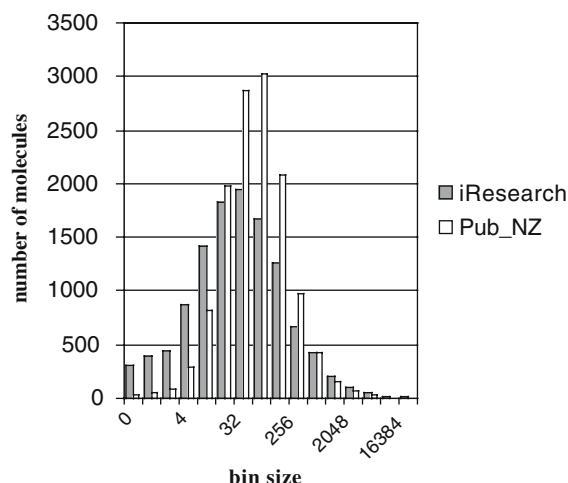


Figure 3. Histogram of molecules in the PHYSPROP having molecules with $r^2 > 0.3$ in iResearch and Pub_NZ datasets. Each bin of the histogram counts the total number of molecules in PHYSPROP having the number of neighbors within the half bin to the bin range. For example bin 32 counts number of molecules in PHYSPROP each having from 16 to 32 molecules with $r^2 > 0.3$ in iResearch or Pub_NZ. There are 1940 and 2873 such molecules in PHYSPROP for iResearch and Pub_NZ datasets, respectively.

proprietary and surrogate molecules will be separated at least $D=100$ or $D=1000$ other molecules. Even if the “hacker” would be able to use precisely the same measure of similarity and the same databases to perform his search, he/she will face with a problem of having 100 or 1000 molecules as the result of this search. In addition, the “hacker” will never be sure that the target molecule is in his collection. It is also possible to use much larger databases of chemicals or to select compounds from databases calculated by complete enumeration of all chemical structures. In addition

to the correlation in space of models, one could also use Tanimoto or other similarity coefficients to select non-related molecules.

Some molecules in PHYSPROP and their diluted analogs at $D=100$ and $D=1000$ in iResearch library are shown in Figure 4. As it is clear, dilution at $D=100$ makes it impossible even to guess the original molecules in the PHYSPROP database. Of course, all molecules have some similarity, i.e. they carry chemical groups important for their lipophilicity.

Using the dilution rate of $D=100$ we selected molecules that had at least this number of neighbors with $r^2 \geq 0.3$ in both databases, respectively. This provided us with $N=1949$ and $N=1671$ surrogate molecules for Pub_NZ and iResearch datasets, respectively. The activities of the surrogate data were predicted using the ALOGPS program. Thus, we could find sufficiently diluted molecules in the databases only for 13–15% of molecules. The probability of a molecule to be used as prototype for the surrogate data was highest for molecules with the number of non-hydrogen atoms in the range from 20 to 40 (Figure 5). Small molecules as well as very large molecules were less correlated (more dissimilar) from the molecules in both analyzed databases.

The surrogate datasets as well as their corresponding real datasets (i.e., molecules which were used as prototypes of the surrogate data) from the PHYSPROP were used to develop “new” logP prediction programs which were tested to predict all molecules from the PHYSPROP database. The calculated results indicate that the surrogate and real datasets gave practically the same prediction ability for the PHYSPROP database (Table 2).

The prediction performance calculated using the Pub_NZ as a source for surrogate data was considerably higher than that of the surrogate data set created using the iResearch library. This result is clear if we take into account the signal-to-noise ratio at the used threshold correlation coefficient (Figure 2). The surrogate data produced using Pub_NZ set were cleaner, the dataset was larger and thus the overall performance of the program developed using these data is higher.

So far we used the same system of descriptors (E-state indices [11–13]) and the same method (Associative Neural Networks) to select surrogate data and then build new models. However, in practice, one would like to use different sets of

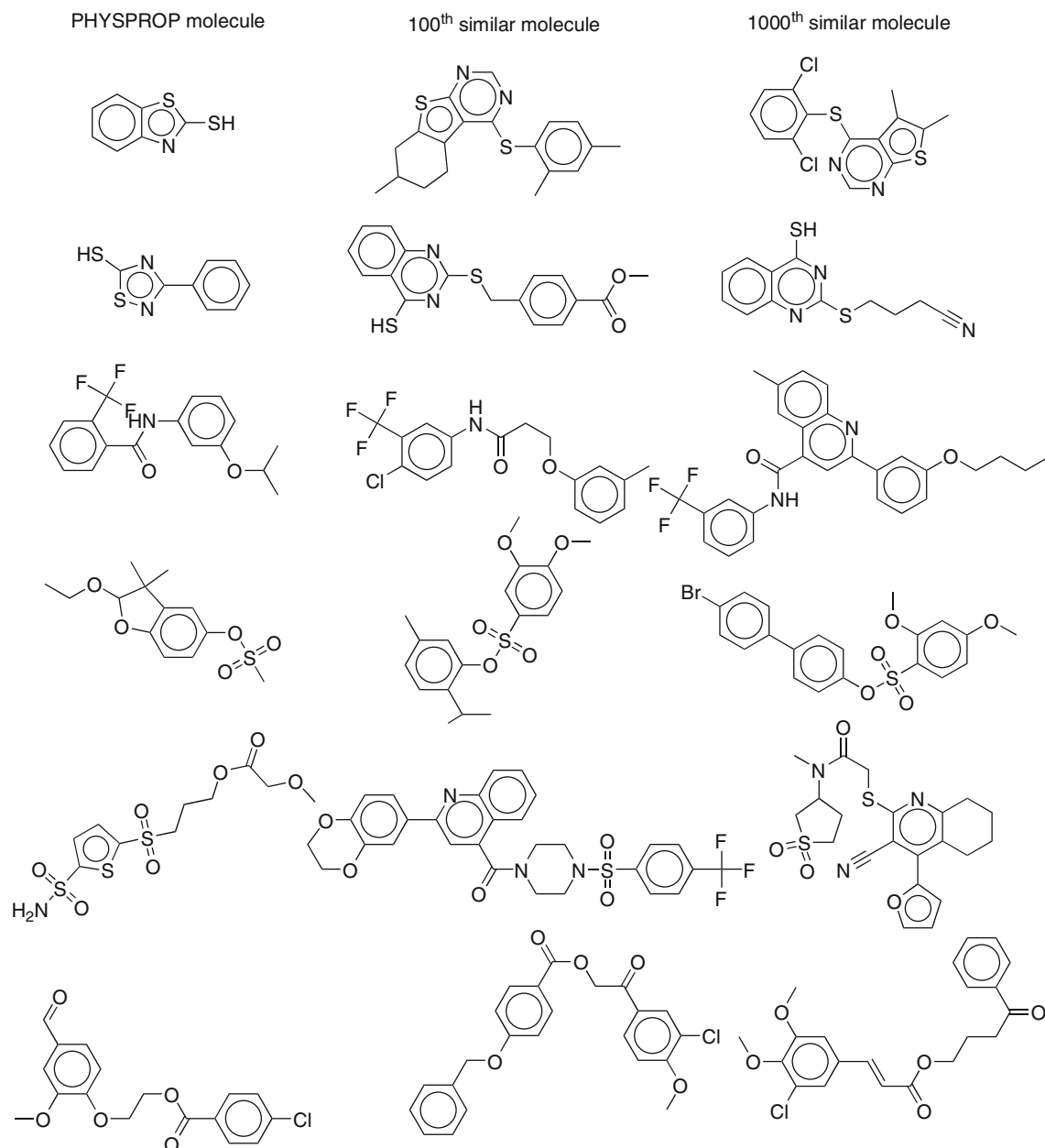


Figure 4. Examples of real molecules (molecules with experimental logP values from the PHYSPROP dataset) and surrogate molecules at $D=100$ and $D=1000$ dilution rates selected from Pub_NZ dataset. A surrogate molecule at rate D corresponds to a molecule that has D th largest correlation coefficient in the Pub_NZ database to the corresponding original molecule. The surrogate molecules together with their *predicted* logP values created surrogate datasets.

descriptors and methods to develop predictive models. It is possible that the performance of different sets of descriptors may depend on their similarity to the original descriptors used to select the surrogate molecules. However, if the predictions for surrogate molecules are correct, one

should be able to develop new reliable models using completely different descriptor sets and statistical models.

In order to test one of the most difficult cases, we have developed a model using 3D descriptors and multiple linear regression method, as

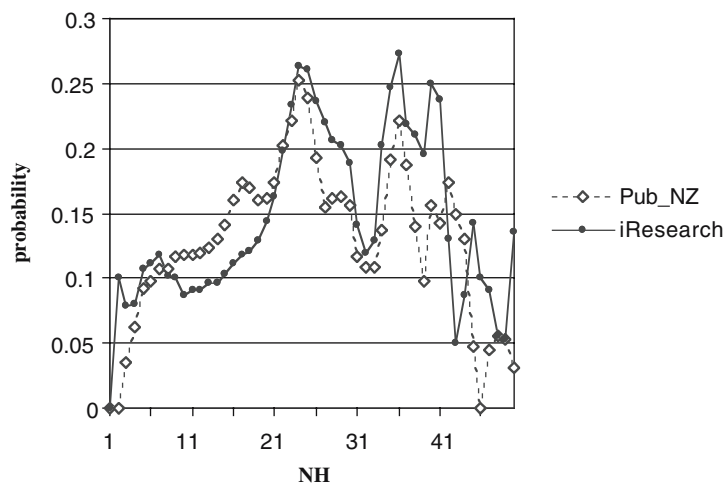


Figure 5. A probability of a molecule to be used as a prototype for the surrogate data set as a function of the number of non-hydrogen atoms (NH). The very simple molecules composed of just few non-hydrogen atoms ($NH < 10$) or very large molecules ($NH > 40$) have lower probabilities.

described in the method section. The 3D descriptors represent a very different method to describe the property of molecules, compared to topological indices; the overlap in information content between 2D and 3D descriptors has been estimated as 40% [25]. Surprisingly, the performance of the surrogate model for the prediction of the PHYSPROP set was even better than the one based on the original molecules. A closer look at the predicted values revealed that a number of

molecules, particular simple ones, had an extremely poor prediction with both real and surrogate data models. For example, CS_2 , $\log P = 1.94$, and CO_2 , $\log P = 0.83$, had predicted $\log P$ values of 13.4 and 6.86, respectively, using the original molecules, and 4.03 and 4.15 respectively when using surrogate data. This result has a simple explanation: The surrogate and original molecules contained a small number of simple molecules (Figure 5), therefore the prediction ability of

Table 2. Prediction accuracy of models developed with real and surrogate models for the 12903 molecules from the PHYSPROP database [14].

Indices/Method	Training set size	prediction performance for the PHYSPROP set		
		r^2	$RMSE$	MAE
<i>real datasets</i>				
E-state/NN ^a	1949	0.87	0.69	0.47
E-state/NN ^b	1671	0.88	0.65	0.46
3D/MLRA ^a	1949	0.45 (0.84) ^c	1.75 (0.72)	0.74 (0.56)
<i>surrogate datasets</i>				
E-state/NN ^a	1949	0.85	0.72	0.50
E-state/NN ^b	1671	0.67	1.1	0.64
3D/MLRA ^a	1949	0.69 (0.82) ^c	1.04 (0.73)	0.71 (0.57)

^aReal and surrogate sets were selected by mapping of PHYSPROP molecules to the Pub_NZ database.

^bReal and surrogate sets were selected by mapping of PHYSPROP molecules to the iResearch library.

^cStatistical parameters after filtering of heavy outliers with absolute error > 2 log units (647 and 559 out of 12903 PHYSPROP molecules for real and surrogate models, respectively) are shown in parentheses. E-state/NN models were developed using 75 E-state indices and ASNN. 3D/MLRA models were developed using 3D descriptors and multiple linear regression analysis. r^2 is square of Pearson correlation coefficient between predicted and calculated values. $RMSE$ is root mean squared error. MAE is mean absolute error.

models developed from both the original and surrogate molecule subsets are particular poor for these classes of compounds. The same problem is, of course pertinent for the E-state indices. However, the E-state indices were weighted with a function of molecular weight [7] and are apparently less sensitive to this problem.

After excluding all prediction outliers, identified as molecules that had prediction error higher than 2 log units, the prediction ability of models developed using original and surrogate data was quite similar but lower to that of the models developed using E-state indices. Of course, a more accurate analysis of 3D data, e.g. an optimization of molecular structures taking into consideration stereoisomerism and using quantum-chemical calculations, may produce better 3D models. This was not, however, the primary goal of the current study and we did not pursue it further. We simply wanted to demonstrate that surrogate molecules can be used to calculate new sets of indices and build new models, and that models calculated using the same number of molecules with real and surrogate data have similar prediction performance.

Discussion

The theoretical analysis of the information content of molecules provided in this article suggests that it is possible to reverse-engineer molecular structures from descriptors. Thus the sharing of molecular indices in the usual way, i.e. to generate indices from the molecular structure and to provide them along with property data, is not safe and may expose the underlying chemical structures.

A traditional way to share such information is a confidentiality agreement between the company that releases the molecules and the academy/industry providing the data analysis. However, given the global reach of the Internet, there is a danger that such information can be intercepted by unauthorized third parties, i.e. by unauthorized access to computer. Encoding molecules using public surrogate data is a reliable alternative that will make impossible decoding the underlying structure, even if the data are publicly released.

Regarding bit rates and the statement that as little as 1 descriptor can be used to decode the molecule: While these numbers sound very drastic,

a simple example may help to understand them better. A comprehensive discussion of information theory is definitely beyond the scope of this paper. The internal representation of a molecule with 35 atoms in the WinRar zipped file is less than 32 bits per molecule as given by the size of this file. It is possible to take the bits corresponding to such molecules and convert them to a long integer (or float value) number representing a molecule from the database with just one descriptor. Of course, in this example the "WinRar" descriptor can be any value from the whole possible range of long integer values, i.e. from -2147483648 to 2147483647 and numbers should have all digits as significant ones. Such descriptor perhaps looks too artificial. However, if two indices instead of one are used, the range of these descriptors, -32768 to 32767, may appear more "reasonable". Four indices will have range from -255 to 256 only, i.e. below the accuracy of usually used float descriptors. Definitely, the considered example is an oversimplification but it demonstrates some theoretical limit on the number of bits. In practice, nobody uses WinRar to generate molecular indices. However, the attempts to develop non-redundant system of indices covering the molecules, e.g. large systems like fragment-based descriptors [26, 27], or E-state indices [12], or even small systems like the newly proposed Shannon-based descriptors [28] may contain enough information to completely decode the molecules. Thus, there is always a danger that a given system of descriptors, or combination of them, may contain sufficient information to reverse engineer the original molecular structures. In general, each system of indices should be mathematically proven to be a secure one, but this can be as difficult as to solve Fermat's last theorem [29]. However, to break the system, it is enough to design an algorithm to reverse-engineer structures from indices. Thus, it will be difficult to get a consensus on methods to publicly release data with indices for which the reverse engineering issue is unknown. From another perspective, an attempt to design "unbreakable" system of indices may result in decreased quality of the generated indices below the level that they can be used in the QSAR studies.

The proposed approach is an amalgamation of ideas discussed in relationship to "confused descriptors" [19], "permuted molecules" [23] and sharing data using shuffled ranks of molecules

[24]. Two former approaches were already mentioned. The later approach proposed to share molecules as shuffled ranks of models. Unfortunately, it can be used only with the model from which the ranks were generated. The surrogate molecule can be considered of as having shuffled ranks compared to the real molecule. Thus the surrogate molecule is the “materialization” in a chemical entity of “a ghost” represented by the shuffled ranks. While the concept of surrogate data proposed in this study is essentially the same as that of shuffled ranks, the consequences of having molecular structures associated with data for modeling, instead of the abstract ranks, provides a dramatic improvement in the development of methods for safe exchange of chemical information.

The surrogate data also is an extension of the secure linear regression method [30]. Rather than to transmit the (sufficient) model parameters, we transmit surrogate molecules that themselves carry sufficient statistics to build new models. A simple illustration of this idea can be done using support vectors in the SVM method [31]. The support vector for the classification task is the training point that determines the boundary (that is optimal according to some criteria) for the classification. Thus support vectors contain the sufficient statistics about the model. The surrogate vectors can be identified as some, e.g. linear, combinations of original support vectors in such a way that they contain the same information as the original support vectors. The surrogate vectors can then be back-projected to the space of molecules to define the surrogate data.

In this article we clearly demonstrated that the use of surrogate molecules provides the possibility of developing models that are as predictive as those developed from the original data. Moreover, this method allows one to identify and calculate new sets of descriptors, and to build new predictive models. Indeed, we used two completely different systems of indices, E-state and *3D*; both of them produced comparable results starting from the real and surrogate data sets. This result was possible because ALOGPS reliably predicted the logP of surrogate molecules. Therefore we could develop new models using predicted rather than the experimental values.

Instead of sharing the original data, companies may share surrogate data, i.e. data for which their internal or external model (e.g., ALOGPS) makes

a confident and precise property prediction. Sharing surrogate data is nothing more than sharing fake molecules associated with reliable and precise property predictions. As long as the shared compounds contain sufficient statistics about the underlying chemical or biological property, the models developed using shared data will provide the same prediction ability as the original data. The proposed methodology can work very well for development of ADME/T properties but may be not applicable to models of biological activity, i.e., binding constants to receptor. Moreover, in some cases even the type of the biological activity could not be named since it may disclose the field in which the company is working on. Of course, if some companies are not interested to share some data there is no way to do it.

However, even development of a robust model to predict only the lipophilicity of chemicals can be very important. Despite the traditional opinion that this property is easy to calculate, the evaluations of several popular programs by Pfizer [3], AstraZeneca [4] and Merck [32] calculated poor performance of considered algorithms (the absolute mean average errors were quite often above 1 log units) for datasets of tens thousands compounds. There is, however, no reason why the current method will not work for aqueous solubility, which is even a more difficult property to predict. The surrogate data should be separately generated for each property and data for logP may not work for aqueous solubility or CaCO₂ permeability of molecules. Definitely, the molecular features important for these types of activity may not be conserved in the surrogate data generated using the lipophilicity model.

One question that remains to be addressed by the practice of safe exchange of chemical information is the following: Given that reliable predictive models are needed to initiate the process, will the final outcome be positive? In other words, will the effort of collecting large datasets of surrogate molecules and their associated predicted properties, lead to an overall improvement in our predictive abilities? Will we, for example, be able to get the ultimate logP prediction models? Or will “local” vs. “global” modeling effects continue to be present?

The proposed solution to the safe exchange of chemical information can only be used in

connection with global models. If the developed model is very local and explains just a particular series of (confidential) compounds, only molecules with this particular chemotype can be used as surrogate data. It will further depend on the descriptors used to generate the model. Some descriptors will generalize to predict compounds across series, and some are likely to prove useless.

The solution proposed in the current article, to select molecules with reliable predictions, was based on the analysis of the distance in space of models and on searching molecules in publicly available databases. The use of very large databases will make it possible to create even more appropriate surrogate data. It would be very interesting to investigate databases derived from the complete enumeration of molecular structures. Such efforts have already been considered for alkanes with MW up to 300 [33]. Such a database would provide an exhaustive coverage of molecules, and would allow the release of a significantly larger amount of surrogate data, compared to the 13–15% achieved in this study.

For some compounds, it will be impossible to find a good quality surrogate. These are, for example very simple molecules in PHYSPROP, like ethane or ethylene. They have unique combinations of indices such that no structures could be highly correlated to them. In one scenario, such molecules could be excluded and no surrogate structures will be provided. Another possibility is to release not one, but several molecules that are correlated with the “difficult” structure, but selected below the threshold correlation level. Such data might contain sufficient information about the difficult molecule to derive a useful model. Using the noise/signal framework will make it possible to provide theoretical estimations how many surrogate molecules should be released for each noise to signal ratio.

There is a simple and elegant explanation of an apparent superior performance of models with surrogate data using 3D structures ($MAE=0.74$ vs. $MAE=0.71$ for prediction of the PHYSPROP set using models with real and surrogate data, respectively). One can speculate that the surrogate data contained more information about outlining molecules than the real molecules. Indeed, the activity prediction of surrogate data was done with the ALOGPS model that was calculated using all PHYSPROP

data. Contrary to that, the real molecules contained just experimental values for themselves. Thus the models calculated using the real data had lower extrapolation properties compared to models calculated using surrogate data set of similar size.

Notice that in this study we completely ignored the chirality of molecules before using them in 2D \rightarrow 3D conversion with Corina program. This was done intentionally, since the quality of 3D structures was different for real and surrogate data and could dramatically bias the prediction ability of the models. The use of chirality-free information allowed us to compare both sets under equal conditions. The use of correct 3D structures may significantly increase the prediction ability of models for both real and surrogate data and can be a topic of further studies; however, it is well known that lipophilicity does not change with chirality. A further improvement could be to use another more appropriate 3D descriptors for the study. This may considerably improve the prediction rate of both models. There is, however, no reason to suppose that the comparative performance of both models will change.

The measure of similarity as correlation ranks of models is very specific. For example, only 5876 out of 2,4 M compounds (0.23%) in Pub_NZ have coinciding ranks, i.e. “confusing descriptors”. The number of such molecules in iResearch library is 36,304 or just 0.27%. These numbers are by an order of magnitude smaller than the corresponding “confusion rates” for the iResearch library calculated using MDL-320 (14.5%), Daylight 512-FP (15.2%) or even Daylight 2048-FP keys (5.3%) [19]. The molecules with the same rank are usually from the same chemical series and are just different in positions of several substituents. Thus the ALOGPS rank coding makes an excellent effort in separating the molecules. This is intentional, otherwise the program would not differentiate molecules with different logP values and would not provide reliable predictions for new compounds. Thus we anticipate that attempts to develop highly “confusing” descriptors may run the risk of having reduced prediction abilities.

The proposed surrogate data approach will not confuse all the data, just those structural features that are not related to the underlying property.

The produced surrogate data can be highly specific for the property they are generated to represent. They may also be used, perhaps, to develop a model for closely related properties. For example we recently demonstrated that the ALOGPS program, developed to predict logP of neutral compounds can reliably predict the logD of charged species using simple correction of its results in ASNN method [3, 4]. Thus, the correlation measures in both logP and logD spaces are rather similar, yet obviously different.

Conclusions

The 229th ACS symposium challenge asked if secure data sharing is possible [34]. An answer was given for the common idea of data exchange using indices and a new method using surrogate molecules was proposed. The validity of the method for prediction of lipophilicity of compounds was demonstrated and a sound estimation of the

complexity to make “reverse-engineering” of the proposed approach was provided. Thus, in this study we introduced a new approach to share complete collections of surrogate molecules without the danger of disclosing the underlying molecular structures.

Acknowledgement

The authors thank Scott Hutton for providing compounds from iResearch library (ChemNavigator) used in the current study, Cristian Bologa (University of New Mexico Division of Biocomputing) and Philip Wong (Institute for Bioinformatics) for their technical help. The authors thank Robert S. Pearlman for his constructive comments. Part of this work was supported by INTAS “Virtual Computational Chemistry Laboratory” <http://www.vcclab.org> grant (IVT) and by New Mexico Tobacco Settlement Funds for Biocomputing (TIO).

Table A1. An example of different representations of molecules.

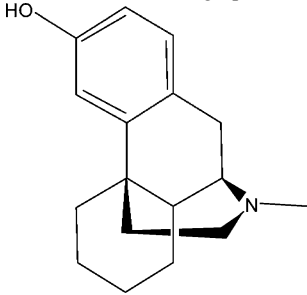
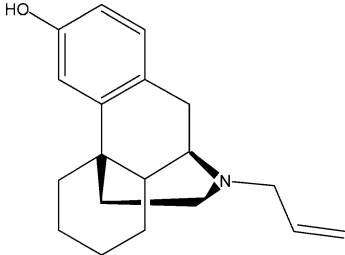
<i>Morphinan-3-ol, 17-methyl-</i>	<i>Lavallorphan</i>
Representation of molecules as SMILES <chem>C12=C(C=C(O)C=C1)C34C(C(C2)N(C)CC3)CCCC4</chem>	<chem>C12=C(C=C(O)C=C1)C34C(C(C2)N(CC=C)CC3)C-CCC4</chem>
Representation of molecules as 2D graphs 	
Representation of molecules as E-state indices SaaCH 6.1180 SaasC 3.4341 SsCH3 2.3002 SsOH 9.9140 SssCH2 9.1443 SsssCH 1.5420 SsssN 2.5873	SaaCH 6.1120 SaasC 3.3943 SdCH2 3.9441 SdsCH 2.0577 SsOH 9.9849 SssCH2 9.9861 SsssCH 1.4475

Table A1. Continued.

SssC 0.3767	SssN 2.6498																		
SsOH(phen) 9.9140	SsssC 0.3402																		
SsssN(al) 2.5873	SsOH(phen) 9.9849																		
Se1C1N3s 2.5833	SsssN(al) 2.6498																		
Se1C2C2ss 5.6089	Se1C2C2ds 1.6881																		
Se1C2C3sa 1.4264	Se1C2C2ss 5.5361																		
Se1C2C3ss 2.1553	Se1C2C3sa 1.4099																		
Se1C2C4ss 1.8407	Se1C2C3ss 2.0911																		
Se1C2N3ss 1.9941	Se1C2C4ss 1.7867																		
Se1C3C3ss 0.7539	Se1C2N3ss 3.8970																		
Se1C3C4as 1.0486	Se1C3C3ss 0.7090																		
Se1C3C4ss 0.5697	Se1C3C4as 1.0293																		
Se1C3N3ss 1.8091	Se1C3C4ss 0.5334																		
Se1C3O1a 6.6004	Se1C3N3ss 1.8102																		
SeaC2C2aa 2.2600	Se1C3O1a 6.6434																		
SeaC2C3aa 7.2003	Se2C1C2s 3.2680																		
SeaC3C3aa 1.6493	SeaC2C2aa 2.2642																		
SeaC2C3aa 3.2361	SeaC2C3aa 7.1936																		
	SeaC3C3aa 1.6400																		
Representation of molecules as 64 predicted neural networks values (rounded to 3 significant digits) ¹																			
3.63	3.51	3.69	3.70	3.81	3.76	3.82		4.25	4.07	4.42	4.22	4.43	4.30	4.29					
3.63	3.62	3.76	3.47	3.71	3.81	3.72		4.31	4.11	4.44	4.08	4.24	4.24	4.38					
3.79	3.54	3.53	3.76	3.66	3.66	3.81		4.34	4.28	4.12	4.24	4.21	4.29	4.34					
3.61	3.58	3.47	3.85	3.81	3.65	3.63		4.18	3.91	4.04	4.50	4.44	4.24	4.14					
3.63	3.70	3.81	3.70	3.65	3.38	3.57		4.30	4.24	4.34	4.13	4.21	4.07	4.33					
3.62	3.46	3.79	3.52	3.67	3.73	3.63		4.34	4.15	4.24	3.95	4.26	4.21	4.21					
3.41	3.63	3.39	3.51	3.65	3.49	3.72		4.02	4.28	3.83	4.06	4.41	4.01	4.31					
3.67	3.67	3.56	3.58	3.69	3.65	3.80		4.44	4.35	3.92	4.17	4.30	4.30	4.41					
3.52	3.53	3.73	3.78	3.53	3.65	3.70		4.05	4.21	4.25	4.43	4.38	4.22	4.25					
3.84								4.51											
Representation of molecules as ranks of the neural networks ²																			
37	56	25	24	05	14	03	38	44	13	32	55	08	39	07	24	25	20	52	03
60	20	07	19	11	50	51	15	31	30	53	37	36	12	16	27	51	34	41	26
06	45	47	59	02	04	35	41	40	21	14	46	63	58	02	05	35	49	23	38
08	22	33	64	48	43	61	10	54	27	15	50	42	54	18	17	48	33	61	29
17	39	62	42	63	57	36	58	18	28	43	45	59	28	64	56	10	60	19	04
29	49	46	26	34	09	55	52	16	12	13	62	47	22	21	09	57	44	31	06
53	32	23	01							11	40	30	01						

¹Notice, that predictions of all 64 networks in the ALOGPS ensemble are similar but not exactly the same. Thus they can be ordered and ranked.

²The Spearman correlation of these two molecules using the ranks is $r^2 = 0.47$, i.e. molecules are highly and significantly correlated.

```

BEGIN:

FOR each molecule i from PHYSPROP
    Calculate ranks using ALOGPS
    Store ranks
ENDFOR

FOR each molecule i from iResearch Library
    Calculate ranks using ALOGPS
    Store ranks
ENDFOR

FOR each molecule i from PHYSPROP
    FOR each molecule j from iResearch Library
        Calculate Spearman rank correlation coefficient,  $r_{ij}$ , using stored ranks
        IF  $r_{ij} > 0$  and  $r_{ij}^2 > 0.3$ 
            THAN
                ADD molecule j to a list L
            ENDIF
        ENDFOR

        IF size of the candidate list L  $\geq$  D
            THAN
                ADD molecule i with its experimental value to the REAL DATASET
                SELECT molecule k from the list L, which has  $D^h$  largest
                correlation coefficient to the molecule i
                ADD molecule k to the SURROGATE DATASET
            ENDIF
        ENDFOR

    FOR each molecule from the SURROGATE DATASET
        Predict its logP value
        Store molecule with its predicted value in the SURROGATE DATASET
    ENDFOR

END:

```

At the end of the procedure the **REAL DATASET** will contain molecules from PHYSPROP database with experimental logP values and the **SURROGATE DATASET** will contain the same number of molecules selected from the iResearch Library with calculated logP values. Thus the **SURRODATE DATASET** will not contain any one experimental value.

The same procedure was used to select **REAL/SURROGATE DATASET** using Pub_NZ dataset.

Schema A1. Pseudo-code to create real/surrogate dataset using iResearch Library.

References

1. DiMasi, J.A., Hansen, R.W. and Grabowski, H.G., J. Health. Econ., 22 (2003) 151.
2. Landers, P., The Wall Street Journal, 12/8/2003, 2003.
3. Tetko, I.V. and Poda, G.I., J. Med. Chem., 47 (2004) 5601.
4. Tetko, I.V. and Bruneau, P., J. Pharm. Sci., 93 (2004) 3103.
5. Tetko, I.V., Drug Discov. Today, in press (2005).
6. Irwin, J.J. and Shoichet, B.K., J. Chem. Inf. Model., 45 (2005) 177.
7. Tetko, I.V., Tanchuk, V.Y. and Villa, A.E., J. Chem. Inf. Comput. Sci., 41 (2001) 1407.
8. Tetko, I.V., Neur. Proc. Lett., 16 (2002) 187.
9. Tetko, I.V., J. Chem. Inf. Comput. Sci., 42 (2002) 717.
10. Tetko, I.V., Villa, A.E.P., Aksenova, T.I., Zielinski, W.L., Brower, J., Collantes, E.R. and Welsh, W.J., J. Chem. Inf. Comput. Sci., 38 (1998) 660.
11. Hall, L.H. and Kier, L.B., J. Chem. Inf. Comput. Sci., 35 (1995) 1039.
12. Kier, L.B. and Hall, L.H. Molecular Structure Description: The Electropotological State. Academic Press, London, 1999.

13. Kier, L.B. and Hall, L.H., *Pharm. Res.*, 7 (1990) 801.
14. PHYSPROP database is available from Syracuse, Inc. <http://www.syrres.com>, 31/07/2005.
15. Sadowski, J., Gasteiger, J. and Klebe, G., *J. Chem. Inf. Comput. Sci.*, 34 (1994) 1000.
16. Todeschini, R. and Consonni, V. *Handbook of Molecular Descriptors*. WILEY-VCH, Weinheim, 2000.
17. Weininger, D., Blaney, J.M. and Dixon, S., 1993 USA.
18. Clement, O.O. and Guner, O.F. 229th American Chemical Society National Meeting & Exposition, ACS, San Diego, CA, March 13–17, 2005.
19. Bologa, C., Olah, M. and Oprea, T.I. 229th American Chemical Society National Meeting & Exposition, ACS, San Diego, CA, March 13–17, 2005.
20. Shen, L., Smith, K.M., Masek, B.B. and Pearlman, R.S. 229th American Chemical Society National Meeting & Exposition, ACS, San Diego, CA, March 13–17, 2005.
21. Li, M. and Vitanyi, P. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, Heidelberg, 1997.
22. Filimonov, D. and Poroikov, V.V. 229th American Chemical Society National Meeting & Exposition, ACS, San Diego, CA, March 13–17, 2005.
23. Abagyan, R. 229th American Chemical Society National Meeting & Exposition, ACS, San Diego, CA, March 13–17, 2005.
24. Tetko, I.V. 229th American Chemical Society National Meeting & Exposition, ACS, San Diego, CA, March 13–17, 2005.
25. Oprea, T.I., *J. Braz. Chem. Soc.*, 13 (2002) 811.
26. Solov'ev, V.P., Varnek, A. and Wipff, G., *J. Chem. Inf. Comput. Sci.*, 40 (2000) 847.
27. Trepalin, S.V., Gerasimenko, V.A., Kozyukov, A.V., Savchuk, N.P. and Ivaschenko, A.A., *J. Chem. Inf. Comput. Sci.*, 42 (2002) 249.
28. Mestres, J. and Gregori-Puigjané, E. 229th American Chemical Society National Meeting & Exposition, ACS, San Diego, CA, March 13–17, 2005.
29. http://www-groups.dcs.st-and.ac.uk/~history/HistTopics/Fermat's_last_theorem.html, 31/07/2005.
30. Young, S.S., Karr, A. and Sanil, A.P. 229th American Chemical Society National Meeting & Exposition, ACS, San Diego, CA, March 13–17, 2005.
31. Vapnik, V.N. *Statistical Learning Theory*. Wiley, New York, 1998.
32. Walker, M.J., *QSAR Comb. Sci.*, 23 (2004) 515.
33. Kappler, M.A., Allu, T.K. and Oprea, T.I. *J. Chem. Inf. Model.*, (2005) in preparation.
34. Wilson, E.K., *Chem. Eng. News*, 83 (2005) 24.