Novel Approach to Evolutionary Neural Network Based Descriptor Selection and QSAR Model Development

Debeljak, Željko; Marohnić, Viktor; Srečnik, Goran; Medić-Šarić, Marica

Source / Izvornik: **International Journal of Computer Aided Engineering and Technology, 2006, 19, 835 - 855**

Journal article, Published version
Rad u časopisu, Objavljena verzija rada (izdavačev PDF)

https://doi.org/10.1007/s10822-005-9022-2

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:239:165922

Journal of Computer-Aided Molecular Design (2006) 19: 835–855
DOI 10.1007/s10822-005-9022-2

Novel approach to evolutionary neural network based descriptor selection and QSAR model development

Željko Debeljak[a,*], Viktor Marohnić[b], Goran Srečnik[c] & Marica Medić-Šarić[d]
[a]*Medicinal Biochemistry Department, Osijek Clinical Hospital, J. Huttlera 4, 31000 Osijek, Croatia;* [b]*AVL-AST, Avenija Dubrovnik 10, 10000 Zagreb, Croatia;* [c]*Analytical Development Department, PLIVA d.d., Prilaz Baruna Filipovića 25, 10000 Zagreb, Croatia;* [d]*Department of Medicinal Chemistry, Faculty of Pharmacy and Biochemistry, University of Zagreb, A. Kovačića 1, 10000 Zagreb, Croatia*

Received 16 June 2005; accepted 12 October 2005

Key words: benzodiazepines, descriptor selection, evolutionary neural networks, QSAR, wrappers

Summary

Capability of evolutionary neural network (ENN) based QSAR approach to direct the descriptor selection process towards stable descriptor subset (DS) composition characterized by acceptable generalization, as well as the influence of description stability on QSAR model interpretation have been examined. In order to analyze the DS stability and QSAR model generalization properties multiple random dataset partitions into training and test set were made. Acceptability criteria proposed by Golbraikh et al. [J. Comput.-Aided Mol. Des., 17 (2003) 241] have been chosen for selection of highly predictive QSAR models from a set of all models produced by ENN for each dataset splitting. All QSAR models that pass Golbraikh's filter generated by ENN for each dataset partition were collected. Two final DS forming principles were compared. Standard principle is based on selection of descriptors characterized by highest frequencies among all descriptors that appear in the pool [J. Chem. Inf. Comput. Sci., 43 (2003) 949]. Search across the model pool for DS that are stable against multiple dataset subsampling i.e. universal DS solutions is the basis of novel approach. Based on described principles benzodiazepine QSAR has been proposed and evaluated against results reported by others in terms of final DS composition and model predictive performance.

Nomenclature: QSAR – quantitative structure activity relationship; descriptor – attribute; molecule – object; input variable – independent variable; output variable – dependent variable; m–n–p – fully connected NN topology with biases where input layer contains m input neurons hidden layer contains n hidden neurons and output layer contains p neurons; ENN – evolutionary neural networks; GNN – genetic neural networks; NN – neural networks; DS – descriptor subsets; LOO – leave one out; EA – evolutionary algorithm; CPU – central processing unit; PC – personal computer; FF – fitness function, also called objective or merit function; MSE – mean squared error; SSE – sum of squared errors; RMSE – root mean squared errors; CV – coefficient of variation (relative standard deviation); LMO – leave many out; ROC – receiver operating characteristic; SCG – scaled conjugated gradient; SNNS – Stuttgart neural network simulator; GABA – γ-aminobutiric acid; IC – inhibitory concentration; $\langle X \rangle$ – average *X value*; *ntot* – total number of molecules; *npart* – number of complete dataset partitions into training and external validation set; *ntest* – number of external validation set molecules; *nsubd* – number of internal validation sets; T – critical *npart* fraction corresponding to number of complete dataset partitions for which specific DS fails to pass at least one of the predictive performance filters; $p(\alpha)$ – probability of type I statistical error; $p(\beta)$ – probability of type II statistical error

*To whom correspondence should be addressed. Phone: +385-91-5755111; E-mail: debeljak.zeljko@kbo.hr

Introduction

ENN and genetic neural networks (GNN) are frequently used evolutionary algorithm (EA) based descriptor selection methods that belong to a group of feature selection/prediction hybrids called wrappers [1, 2]. It has been shown on a number of different QSAR problems that these wrappers could efficiently lower external test set and/or leave-one-out (LOO) prediction error [3–12]. Subject that has not been thoroughly documented before is final ENN/GNN QSAR model interpretation. Most of the authors use only a few best performing QSAR models in terms of preselected fitness from the last ENN/GNN generation into consideration [3–6, 12]. This is not the best selection due to a number of reasons. First of all, such approach results in different models for different random generator seed values used during ENN/GNN training [4]. This raises a question about selection of final QSAR model among different final ENN/GNN solutions corresponding to different random seeds. Moreover, different dataset partitions i.e. different random subsamplings [13] will also cause different final QSAR models i.e. description instability [11, 14]. This behavior, also known as 'Rashomon effect' [15] is well-documented property of both, feature selection methods and neural networks (NN) [16], as well as other prediction tools like random forests [17]. Many authors used model ensembles for improvement of external validation and proved adequacy of this approach [11, 16–19]. But such improvement of external validation results leads to very complex model interpretation [15, 16].

According to the resampling statistics literature three levels of statistical inference are possible: population statistics determination implemented by bootstrap procedures, determination of causal inference that is implemented by rerandomization and analysis of description stability that is implemented by cross-validation [14]. The first generalization level could hardly be achieved in QSAR since it is not possible to make random sampling across complete population of all molecules that are characterized by certain ranges of property values. All members of such population are not known. Without detailed population description and possibility to randomly select population members for population statistics determinations this type of statistical inference is not possible. However, applications of the other two levels of statistical inference are frequently used in QSAR studies. Rerandomization doesn't pose a request for random sampling across complete population of some type of molecules. Instead, it could be done on a non-random sample that is the case in all QSAR studies. The basis of the procedure is random permutation of output variable values and analysis of external validation results of such constructs. This method is known among QSAR community as y-scrambling [20, 21]. Unfortunately, according to literature it has been rarely done in appropriate way [21]. Instead of analysis of y-scrambling influence on the final QSAR model analysis of influence on complete wrapper training should be done. This is the main reason why this type of statistical inference is not easy to achieve when one uses very CPU time demanding wrappers like ENN or GNN.

The last level of statistical inference that could be easily misidentified as generalization is stability of description i.e. model stability. Generalization and, sometimes description stability analysis have been implemented in QSAR by different cross-validation methods. The most frequently used cross-validation method in ENN/GNN based QSAR is LOO. During the last few years the most important cross-validation method became predictive performance analysis based on test set also known as hold out set obtained by dataset partition (Figure 1). According to statistical literature recommended percentage of test objects needed for appropriate external cross-validation is approximately 50% of the complete dataset [14]. This percentage is rarely possible to use in QSAR external validation since the number of all molecules involved in the study is frequently very small. Computational community proposed somewhat lower percentage of test objects needed for correct analysis of model generalization and stability. According to Yao and Liu [19] and Kohavi [22] 25–33% of complete dataset objects should be placed in the test set if one wants to achieve appropriate generalization and/or description stability estimation. External validation result comparison makes important difference between generalization and stability analysis. It is easy to find single dataset splitting for which

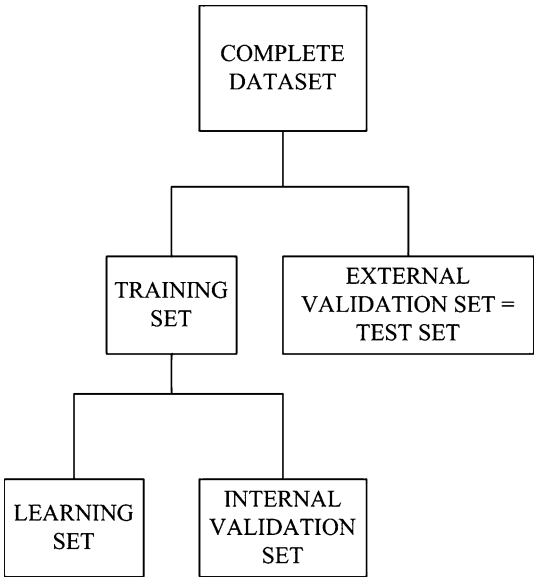

Figure 1. Dataset partition protocol.

one gets acceptable generalization, especially if the number of test set elements (*ntest*) is small. But only multiple external validation results could provide realistic picture [13, 16, 23, 24]. Random subsampling based on multiple partitions of the dataset into training and test set represents extensive application of cross-validation that has been introduced to QSAR by Mattioni et al. [11] and to related fields by Leardi and Gonzales [20]. Stability of solutions against random subsampling could be referred to consistency of QSAR generalization results. In case of feature selection wrappers it could also be referred to consistency of models' DS composition. Only DS that exist in solution pools corresponding to most of or all examined splittings could be treated as non-random and for these DS we introduce term the universal solutions. If such solutions are also characterized by acceptable and consistent generalization it makes search for such solutions an interesting task.

It has been shown that LOO tends to be too optimistic predictive performance estimator [25, 26]. Besides, LOO used as wrapper FF provides only internal model validation. Therefore internal and external validation of QSAR model based on single or multiple hold out test set prediction results became a standard tool for ENN/GNN training and performance evaluation in recent years [11, 12, 25–27]. We do agree with reviewers comment that multiple leave-many-out (LMO) external validation could lead to underfitting due to removal of significant proportion of molecules from already small dataset [28]. However, multiple LMO was selected for both, nested internal and external predictive performance measure. The reason for that is its ability to detect model overfitting even when analyzed set contains a lot of very similar molecules. Namely, if a lot of clones exist in complete dataset multiple random removal of significant number of molecules for internal and external testing leads to inclusion of all or most of the clones into test sets at least in some instances. In case of overfitting such test sets cause significant predictive performance deterioration. Moreover, predictive performance variation is expected to be significantly increased. The more partitions the higher probability of such events. In the following research approximately 44% of complete dataset is set aside for internal and external tests and the dataset splitting is repeated 10–60 times.

If one wants to do evolutionary QSAR model selection based on separate set predictive performance followed by correct external validation at least three subsets are needed. Namely, training set, which is subdivided into learning and internal validation subsets and external test set (Figure 1). Internal validation subset is used as FF during wrapper training. Test set predictive performance i.e. external validation enables selection of the best model or models among a set of different model candidates only after the wrapper training phase has been done. This scheme represents a kind of nested form of model validation.

Among different measures of test set predictive performance i.e. model acceptability multicriterial approach proposed by Golbraikh et al. [27] is used (Equations 1–7):

$$q^2 > 0.5, \tag{1}$$

$$q^2 = 1 - \frac{\sum_{i=1}^{\text{ntot}} (y_i - \hat{y}_i)}{\sum_{i=1}^{\text{ntot}} (y_i - \overline{y}_i)} \tag{2}$$

in all equations y represents output variable (pharmacological activity in this particular case), while $\hat{y}$ represents pharmacological activity prediction. Signed variables represent corresponding averages (within the text averages are represented

by '⟨ ⟩'); *ntot* is the total number of objects, in our case molecules in complete dataset;

$$R^2 > 0.6, \tag{3}$$

$$R^2 = \frac{\left(\sum_{i=1}^{ntest}(y_i - \overline{y}) * (\hat{y}_i - \overline{\hat{y}})\right)^2}{\sum_{i=1}^{ntest}(y_i - \overline{y})^2 * \sum_{i=1}^{ntest}(\hat{y}_i - \overline{\hat{y}})^2}, \tag{4}$$

$$\frac{R^2 - R_0^2}{R^2} < 0.1. \tag{5}$$

R_0^2 represents coefficient of determination corresponding to linear regression equation $y = f(\hat{y}) = K\hat{y}$;

$$0.85 \leq K \leq 1.15, \tag{6}$$

$$\left|R_0^2 - R_0'^2\right| < 0.3. \tag{7}$$

$R_0'^2$ represents coefficient of determination corresponding to equation $\hat{y} = f(y) = K'y$.

It could be seen that Golbraikh's criteria involve both, LOO q^2 and hold out test set R^2 based validation. Most authors use only LOO q^2 or external test set R^2 coefficients [3, 6] but Todeschini et al. [29] and Golbraikh et al. [27] described deficiencies of such approaches and proposed two multicriterial approaches.

In order to reach highest possible external validation results two ingredients are crucial: efficient learning algorithm and internal FF that perfectly correlates with external validation results. Unfortunately, such internal FF generally does not exist [19, 24]. Among different internal FF some FF provide better correspondence with external validation results than others. It has been shown on some QSAR problems that certain wrappers that use validation subset based FF produce better external validation results than their counterparts that use FF based on simple root mean squared error (RMSE) calculated for complete training set [23, 24]. Among different wrapper FF that use internal validation subset prediction quality for QSAR model selection multiple LMO RMSE based FF [21, 23, 24] approach has been selected. Selected FF is based on a large number of random subdivisions (*nsubd*) of training set into learning and internal validation subset. The most of QSAR datasets are quite small and molecules taken into account make scarce coverage of the descriptor and/or output variable spaces. In other words some property value ranges are overrepresented and some property value ranges are underrepresented in the dataset. This problem becomes even more pronounced when dataset partition is needed. Large *nsubd* could minimize the influence of uneven distribution of molecules between learning and validation set that could be caused by small number of molecules. Moreover, Baumann reported that chance correlation is unlikely if one uses this type of FF [24]. Selected LMO FF is represented by Equation 8:

$$\mathrm{FF} = \sqrt{\frac{\sum_{i=1}^{nsubd}\sum_{j=1}^{nval}(y_{i,j} - \hat{y}_{i,j})^2}{nsubd * nval}}, \tag{8}$$

where *nval* stands for number of internal validation set molecules.

During the preparation for this study we also examined some of existing LMO based FF alternatives [9, 12, 30]. Although less computationally demanding and characterized by similar average generalization, application of analyzed alternatives led to a smaller number of eligible models probably due to a single preselection of internal validation set elements. Therefore multiple LMO type of FF was selected.

Random subsampling, large number of different ENN/GNN solutions and application of acceptability criteria led us in a position to propose novel QSAR model building principle and final DS selection. Our primary goal was to produce universal if not unique interpretation of QSAR. This means that small number of DS or even only one final DS was searched for. One way to do that is to apply acceptability criteria on all wrapper results. Application of acceptability criteria ensures model pool shrinkage as well as selection of solutions characterized by acceptable predictive performance. This way obtained set of eligible solutions could be reduced further by the application of stability filter i.e. selection of those DS that exist in all eligible solution subsets corresponding to different complete dataset partitions. Same as frequency of descriptor appearance

in eligible models [11, 31], stability of acceptable DS against random subsampling represents basis for small QSAR model pool selection that consequently simplifies model interpretation. The prerequisite for the success of such double filter based process is existence of large number of QSAR models characterized by acceptable predictive performance that are produced by applied wrappers. ENN/GNN, wrapper that could produce multitude of QSAR models, equipped with suitable internal validation like multiple LMO FF that resembles external DS stability analysis itself seems to be appropriate wrapper candidate. Figure 2 describes model selection protocol.

When stable and/or interpretable models have been found one more QSAR goal remains to be accomplished. That is accurate prediction of output variable value needed for new, most frequently not yet synthesized molecules. If one uses random subsampling even in the case of universal DS existence more than one model should be taken into account. It should be kept in mind that besides DS trained NN makes chromosome i.e. complete QSAR model. Since there could be more than one trained NN counterpart of universal DS it seems appealing to use ensemble of such solutions for final QSAR predictions [21]. Proposed method resembles ensemble based approaches developed by Yao & Liu [19] and well-known bagging and random forests developed by Breiman [16, 17]. But simple model interpretation is the major theoretical advantage of proposed approach, which is also the main goal of this study.

Only when both DS and NN components of the model are obtained simultaneously, and that is the case with universal DS solutions one can use such DS for both, molecular behavior interpretation and output variable value prediction. If universal DS does not exist and one constructs final DS based on descriptor frequencies taken from eligible model pool prediction task is even more complex. There is a high chance of overfit-

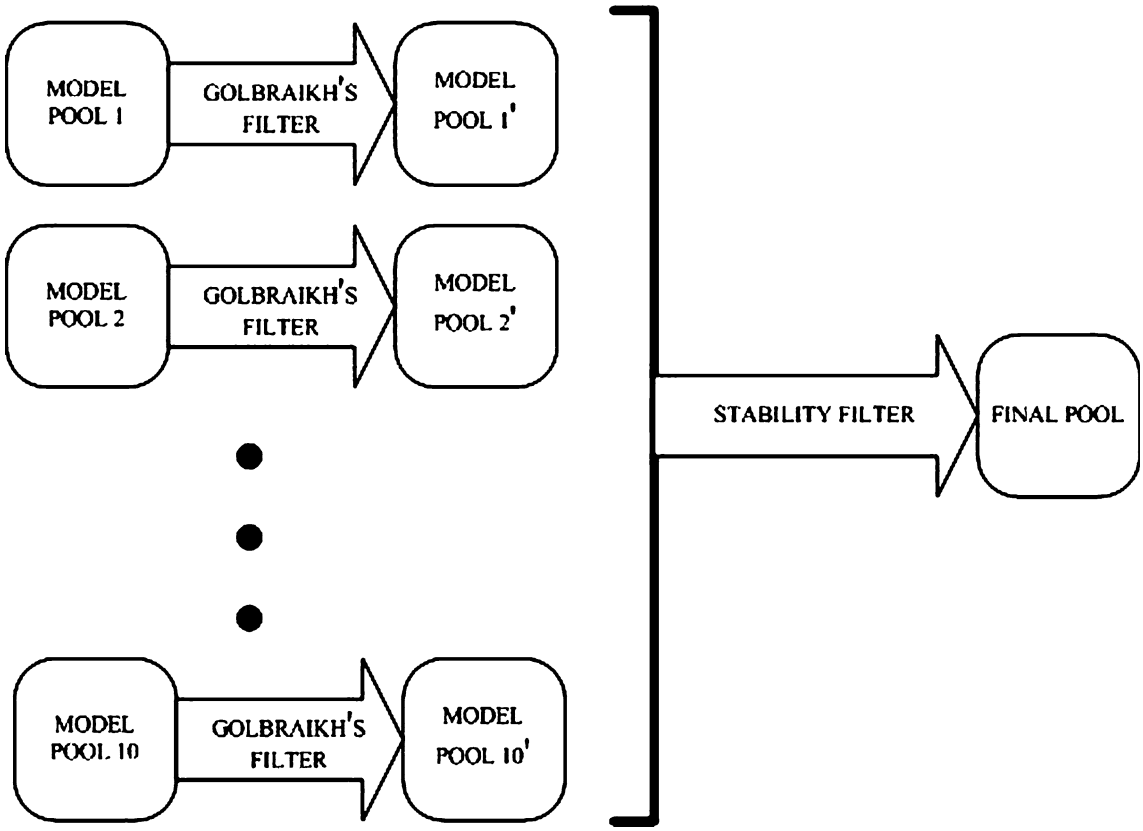

Figure 2. Model selection. Each model pool produced during 50 generations of ENN/GNN based model training and selection contains 10,000 chromosomes. Golbraikh's filter is based on application of acceptability rules while the stability filter is based on existence of one or more DS which exists at least in one copy in each of 10 " ′ " pools.

ting when such, preselected DS construct is reused for final model building. If one constructs final DS based on descriptor frequencies such model doesn't contain NN part. In order to remove this deficiency one can use such DS construct for NN training based on dataset that has already been used for DS construction! This leads to overfitting [22]. Therefore constructed DS based on descriptor frequencies could only be used for interpretation of analyzed QSAR. In order to make correct prediction in such case one can use all models as a model ensemble. Mattioni and coworkers already have applied this approach in QSAR analysis [11]. Novel approach is compared against this, already accepted method that will be referred to as standard method.

According to results published by So and Karplus on GNN based models of benzodiazepine interactions with $GABA_A$ receptor acceptable LOO prediction results have been achieved [4]. Results based on different nonlinear approaches obtained on the same or closely related datasets confirmed high LOO prediction quality of models reported by So and Karplus [32, 33]. Since different groups have previously analyzed this dataset it represents a good candidate for direct final QSAR model comparisons. Moreover, exact descriptor values for each analyzed molecule are available for this set. This is critical for ENN/GNN evaluation reproducibility. Since there are only 42 descriptors there is no need for preselection of descriptors while 57 benzodiazepines is not too large number and it represents usual QSAR experimental settings. Due to the given facts we selected benzodiazepine dataset for novel ENN based DS design evaluation.

Methods

EA considerations

Elitism concept is applied [20]. All high quality chromosomes are selected for breeding no matter if they were offspring or parents in previous generation. Chromosomes featuring appropriate FF values have better chance for multiple breeding this way. ENN training is composed of 50 generations. Two hundred individuals exist in each generation and each parent produces only one child per generation. Reproduction is based on application of DS mutation operator that has been applied for production of all children i.e. by 100% frequency. Crossover operator with defined frequency, that is either kept constant or monotonically changes its value during evolution has been implemented but no improvements were noticed. Although some QSAR results have been reached by the application of such operators [12] we were not able to find any detailed analysis of influence of these functionalities on external validation quality. Moreover, some authors described theoretical disadvantages of crossover operator application in GNN training [19, 34]. Strictly speaking only ENN with fixed relative frequency of DS point mutation equal to 100% has been used in following study.

There is no guarantee that offspring DS has the best performance in combination with initial weight NN values inherited from parent chromosome. Some connectionist packages already implemented random variation of initial NN weights as an NN training option [35]. Since it has been implemented on NN training level it seemed natural to use this functionality as NN genetic operator on evolutionary level. Initial weight mutation operator is based on random generation of variations from actual parent initial NN weight values. DS mutations have better chance this way for adaptation and survival. Values of this operator were set constant throughout all generations and they were fixed at $\pm 50\%$ of parent NN weights.

All applied operators do depend on some random number generator. In order to make adequate comparison of results all random variables were generated by corresponding mutually independent and uniformly distributed random number generators that are controlled by random seed selection.

Random subsampling

Random subsampling has been applied in two ways: for multiple partition of complete dataset into training and test set needed for external validation and for multiple training set partition into learning and validation set needed for internal validation. These settings represent a form of nested cross-validation. Typical experiment is described in Figure 3.

If it is not stated otherwise, for external validation examinations 10 random, complete dataset partitions (*npart*) into training set, composed

of 43 molecules and test set composed of 14 molecules (25% of all molecules) were made.

Random subsampling is also implemented in internal validation i.e. LMO FF value calculation. Multiple LMO FF is dependant on two user-defined variables: *nsubd* and *nval*. Since Baumann reported that selection of *nsubd* value is not of crucial importance we did not analyze the influence of *nsubd* value on experimental outcome. In all experiments *nsubd* has been set to 100. The same 100 partitions were used in all experiments. On the contrary, selection of *nval* value is reported

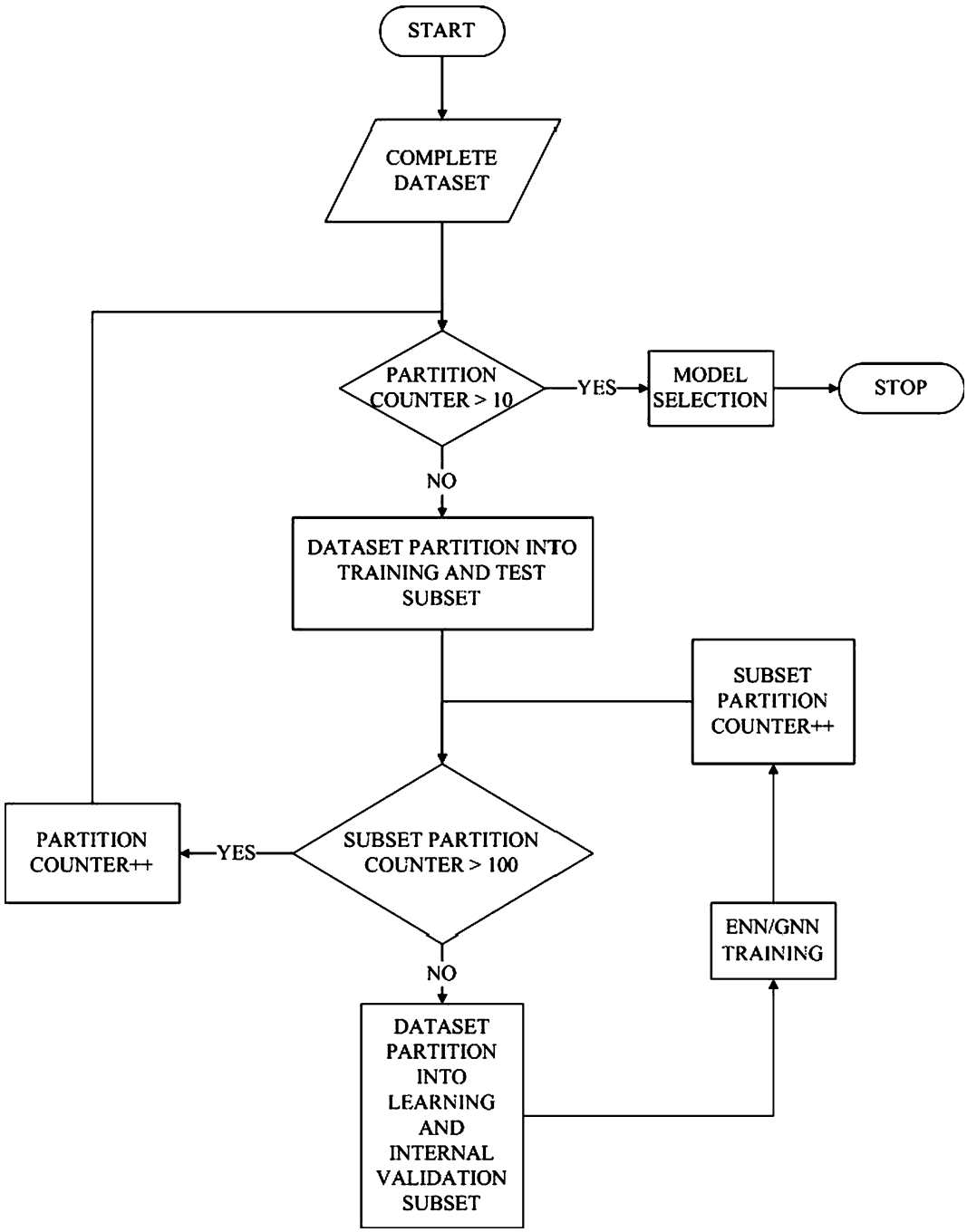

Figure 3. Application of random subsampling.

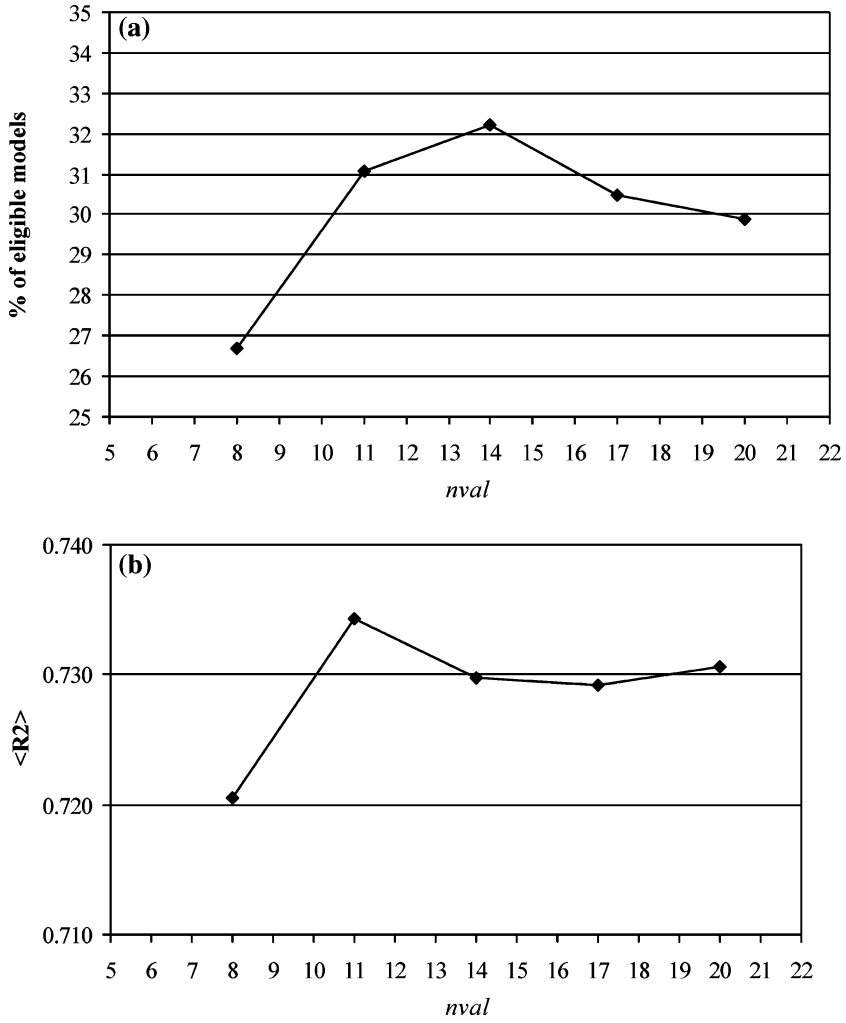

Figure 4. (a) Dependence of eligible models percentage on *nval*. (b) Dependence of eligible models external validation $\langle R^2 \rangle$ on *nval*.

to be very important. Unfortunately, selection of FF *nval* value does not posses firm theoretical ground [23]. Therefore this value should be determined experimentally. Application of external validation is restricted during the wrapper training but it is allowed to use external validation results to select the best model, and consequently the best set of user-defined variables after the wrapper training has been done [20, 23, 24]. We utilized this fact for experimental determination of the most suitable *nval*. Influence of analyzed *nval* values on average external validation results and number of eligible DS obtained by ENN training based on 10 random dataset splittings was analyzed. NN topology with 4 input layer, 2 hidden layer and 1 output layer neurons (4-2-1 NN topology) has been used for *nval* selection while LMO *nval* values tested were 8, 11, 14, 17 and 20. This way from 19% to 47% of the training set is sequestered for internal validation. Training set is obviously very small. It consists of only 43 molecules. Therefore analysis of ENN performance based on higher *nval* values did not seem convenient. Although 40–60% interval has been proposed by Baumann as a starting point for *nval* selection, lower optimal percentages have been already described by the same author. Results of *nval* influence on performance characteristics are given in Figures 4a, b.

According to these results 11 was recognized as near optimal LMO *nval* value. However,

Table 1. Selected user-defined ENN/GNN parameters.

NN evolutionary training parameters	
Initial weight mutation interval boundaries (%)	−50, 50
Percentage of weight mutation interval boundaries change per generation (%)	0
Frequency of selection of candidate NN for link pruning (%)	0
Link pruning selection frequency change per generation (%)	0
General ENN/GNN parameters	
Number of offspring produced per parent	1
Number of parents	200
Number of generations	50
Descriptor selection parameters	
DS point mutation frequency (%)	100
DS point mutation frequency change per generation (%)	0
Frequency of DS cross-mating (%)	0
DS cross-mating frequency per generation (%)	0

Skipped parameters are given in previous text.

nval = 14 is characterized by similar performance characteristics as *nval* = 11. On the other hand, only *nval* = 8 seems to significantly reduce values of both figures of merit. It could be seen that selected *nval* value represents quite small percentage (~26%) of training set and it corresponds to percentage of molecules taken from complete dataset for external validation.

Baumann reported that optimal *nval* value is primarily influenced by dataset structure and/or mathematical relationships between input and output variables and the number of molecules [23, 24]. If one determines optimal *nval* value in one experiment it could be used in similar experiments performed on the same dataset. All following experiments were based on application of ENN on the same dataset. Only differences were the number of input neurons, and consequently the number of variable NN parameters. Accordingly, it seems convenient to set *nval* value to 11 in all following experiments.

NN implementation

Basic facts about NN, in our case three-layered perceptron could be found elsewhere [36, 37]. Still, there are some points specific for this study which are related to NN that should be described here. First of all, batch version of scaled conjugate gradient (SCG) NN learning algorithm is selected [4] since the use of Hessian matrix diagonal elements lessens the chance for getting stuck in local minima during NN training [38]. Number of NN training steps is set to 20 epochs since application of fixed and rather small number of training epochs lowers the probability of chance correlation and ensures faster computations [18].

Two rules of thumb for NN topology construction were considered. Leardi and Gonzales [20], as well as Baumann [21] suggested that the number of molecules should be at least five times the number of DS elements. In order to implement this rule we used fixed number of input neurons [4, 7]. According to another rule of thumb used by Chiu and So [39] and Patankar and Jurs [10] the number of training objects in this case molecules should be approximately twice the number of NN adjustable parameters. Similar or even more demanding rules could be found in NN literature [36]. According to these rules and because only 57 molecules are present in benzodiazepine dataset the choice of NN topology is very restricted. These facts limited our study to 6 input neurons, 1 hidden layer and 2 hidden neurons at maximum (6-2-1 NN topology). Such small NN restricted severely usage of random topology generation and magnitude based link pruning operator that we have implemented [8, 34, 35]. According to the second rule of thumb only 4-2-1, 3-2-1 and 3-3-1 NN topologies with biases included are acceptable. In order to avoid chance correlations [20, 24] instead of NN topology evolutionary adaptation fixed 4-2-1 fully connected NN topology with bias terms included was used as a starting topology. NN with higher and smaller value of number of training elements per number of adjustable NN parameters were also examined.

All descriptors from a set composed of 42 elements were scaled to 0.1–0.9 interval same as output variable, namely benzodiazepine binding affinity for $GABA_A$ receptor represented by log IC50. Initial weights are generated randomly from [−1, 1] interval and logistic function was selected for NN transfer function between input and hidden layer. Hidden and output layer were connected with linear transfer function. SCG

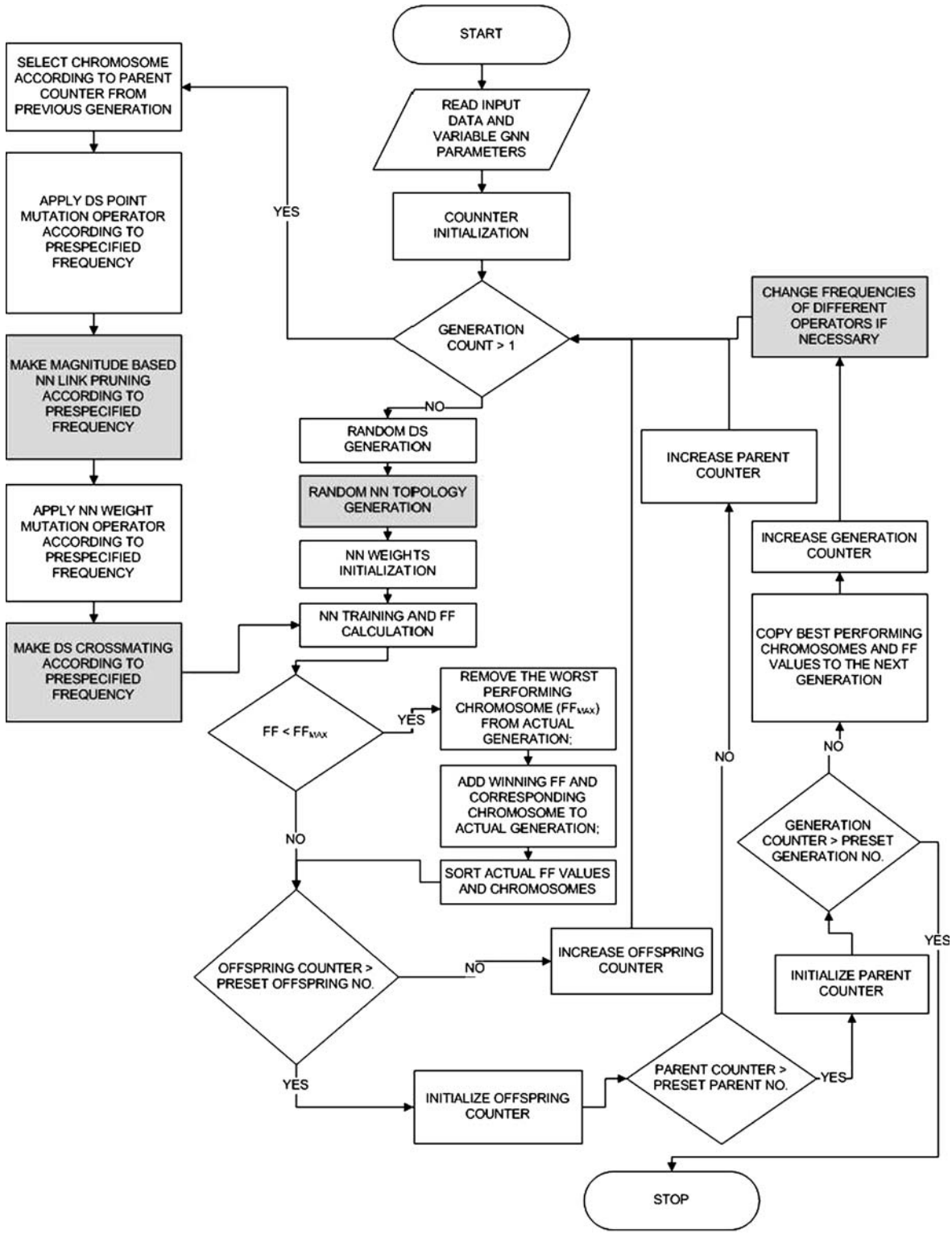

Figure 5. ENN/GNN training. Gray boxes are implemented but they have not been used in following experiments. Random application of specific operators is made in accordance with corresponding operator frequencies.

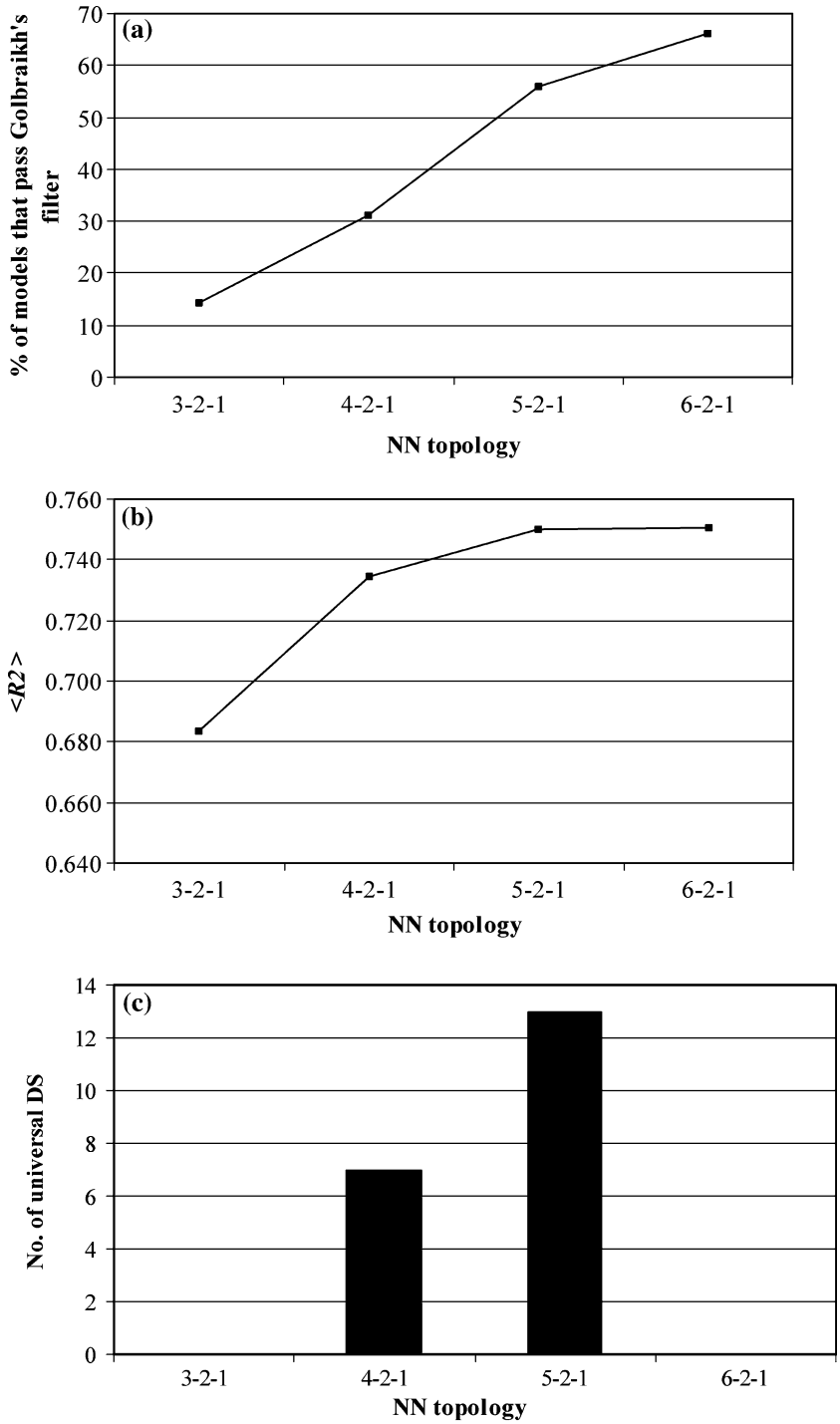

Figure 6. (a) Dependence of percentage of models that pass Golbraikh's filter on the number of DS elements. (b) Dependence of external validation $\langle R^2 \rangle$ corresponding to all eligible models on the number of DS elements. (c) Dependence of number of universal DS on the number of DS elements.

user-defined parameters are set to values proposed by So and Karplus [4] and Møller [38]. Near optimal ENN/GNN training user defined variables used in this study are given in Table 1.

Some of the implemented options were not used. Therefore corresponding user defined variable values are equal to 0%. 200 NN were generated per each of 50 generations. Each, among 10 dataset splittings was used for ENN training and evaluation. This means that 100,000 QSAR models were trained and tested in a typical experiment. Figure 5 represents ENN flow chart.

Computational aspects

Instead of *ab initio* programming approach publicly available software solutions were considered. Few non-commercial software solutions for NN training are available. Stuttgart Neuronal Network Simulator 4.2 (SNNS) has been chosen [35]. SNNS has been employed for different regression and classification problems. Moreover, it has been recognized by QSAR community [40]. The number of NN design and training options that it offers is quite large. Most of GNN routines were written in SNNS interpreter called batchman. Still, there were some limitations that were solved by writing short C/C++ scripts that were precompiled and by some Python scripts. Corresponding executables were called from main batchman program. Obtained ENN results were analyzed by the application of Statistica 6.0 (StatSoft, Inc.) and routines written in Mathematica 5.0 (Wolfram Research, Inc.). In order to make further ENN/GNN development and results comparison possible all routines used in this study are available on request.[1]

Results and discussion

Influence of input layer size on DS stability and predictive performance

The number of DS elements, as well as total number of descriptors is expected to have significant impact on existence of universal DS. The number of DS elements and total number of descriptors directly determine the number of all possible DS combinations. Number of all possible DS combinations inversely correlates with probability of finding single universal DS or any other specific DS in general. In case of 4-2-1 topology number of all possible DS combinations is equal to 111930. This means that performance of at most ~9% of all DS combinations is analyzed during single ENN training under settings given in previous text. Due to a large number of possible DS combinations, determined by number of input neurons search for universal solutions characterized by acceptable predictive performance appears to be very challenging task even in case of 4-2-1 NN topology. Therefore our research started with analysis of the influence of DS elements number i.e. number of input layer neurons on the percentage of eligible models, corresponding predictive performance and existence of universal DS. Results are given in Figure 6a–c.

The percentage of solutions that pass Golbraikh's filter increases almost in linear fashion as the number of DS elements increases. It is expected that inclusion of higher number of descriptors in QSAR model enables capturing more QSAR details in comparison to small QSAR models i.e. small models are prone to underfitting. However, $\langle R^2 \rangle$ reaches the saturation limit around value of 5 (Figure 6b). This behavior indicates arise of inefficient search over the DS combinations space and/or inappropriate evolutionary NN training (underfitting or overfitting). In case of 6 input neurons 5245786 DS combinations are possible. This number represents ~47 times the value that corresponds to 4 input neuron case. Therefore, DS combination space burst, and consequently inefficient search over the DS combination space represents very probable cause of $\langle R^2 \rangle$ curve saturation.

DS combination space search is very efficient for small models while small number of input neurons restricts acquisition of important underlying QSAR features. Very efficient DS combinations space search accompanied by a small number of input neurons inevitably leads to speciation or even learning based on noise i.e. overfitting. It is expected that models obtained this way are characterized by poor external test based predictive performance. Golbraikh's filter removes large number of such models from the model pool and makes existence of universal DS unlikely when small NN are used (Figure 6c).

Problem with large models is the burst of possible DS combinations caused by input layer neuron number increase, which results in multitude of very heterogeneous solutions. Large

number of DS combinations makes detection of universal DS less likely despite the fact that large number of input neurons does not diminish predictive performance (Figure 6b). Increase of the number of DS combinations accompanied by unchanged ENN training settings could lead to negative interaction between applied DS point mutation and NN weight mutation operators. Existence of few copies of the same or closely related DS in a single ENN generation provides better chances for evolutionary adaptation of corresponding NN chromosome parts of such QSAR model classes. On the contrary, when single copies of many heterogeneous models exist in one ENN generation they make NN chromosome part evolutionary improvement provided by weight mutation operator less efficient. Inefficient NN training leads to loss of acceptable chromosome candidates during the evolution. It is expected that among dumped chromosomes exist potentially universal DS. Moreover, large number of different DS accompanied by inefficient NN learning opens possibility of chance correlations. Models characterized by chance correlation may survive Golbraikh's filter based selection. By removing all unstable i.e. non-universal models from the pool stability filter directly restrains QSAR model selection in cases when DS combination burst causes inefficient NN part evolution.[2]

Predictive performance saturation depicted in Figure 6b is only an indication of negative evolutionary operator interaction. Detailed study of this ENN training aspect is beyond the scope of this article. But if the presumption about negative interaction between evolutionary operators is true then chance correlation problem arises and application of double filter based approach could be useful. However, application of double filter stops the NN topology selection process when $\langle R^2 \rangle$ increase corresponding to QSAR models that pass Golbraikh's filter reaches saturation.

Selection approach based on consecutive Golbraikh's and stability filter application yields results given in Figure 6c. It could bee seen that input neuron number gradual increase accompanied by fixed ENN training settings results in maximum on $\langle R^2 \rangle$ curve. This maximum corresponds to 5-2-1 NN topology. This NN topology is in agreement with the first rule of the thumb given in experimental part. On the contrary, number of training objects per number of NN variable parameters is lower than 2. However, experimental evaluation of 5-2-1 topology shows that predictive performance is acceptable. Therefore 5-2-1 NN topology has been used in following experiments.

Influence of npart on definition of DS stability

The second factor that has an important impact on existence of universal DS is *npart*. One can easily create such complete dataset partition that is characterized by highly different training and test

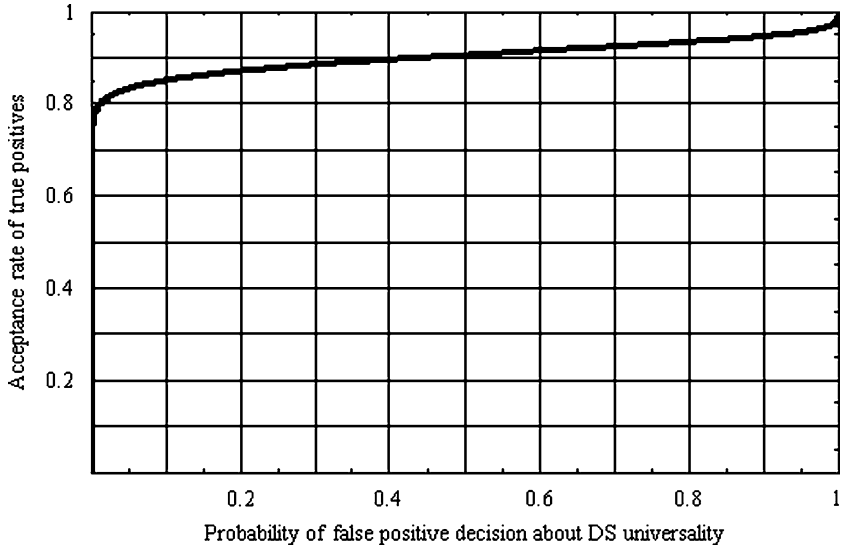

Figure 7. Theoretical ROC curve that corresponds to selection of universal DS based on *npart* = 60 and $T = 6$.

set distributions of one or more descriptors or even output variable. It is obvious that such situation could happen also in case of random generation of dataset partitions. The probability of such event is proportional to *npart*. On the other hand, absolutely efficient model selection tool does not exist and therefore some potentially universal DS could be missed due to a chance during training. Therefore application of all-or-nothing rule in determination whether specific DS is universal or not needs to be redefined.

Possibility of false negative and false positive conclusions i.e. chance correlation that enables some inappropriate model to pass selection raises a question about selection of optimal cut off value (T). If specific DS exists as acceptable solution in more than $npart - T$ complete dataset subsampling cases such DS could be described as universal. On the contrary, DS characterized by T or higher number of failures is considered to be non-universal. Appropriate T should be characterized by smallest possible probabilities of false negative (α) and false positive (β) errors and it expected to be close to 0.

We can assume that failures of specific model to pass filters during *npart* repetitions of described experiment based on random and independent dataset splitting distributes according to binomial distribution [41]. This assumption enables construction of appropriate decision-making strategy. If one finds 10% probability limits corresponding to α and β errors ($p(\alpha)$ and $p(\beta)$) acceptable and also accepts $[0.05\times npart, 0.15\times npart]$ as appropriate location of T then optimal *npart* and corresponding T could be determined. Let assume that true probability of universal behavior corresponding to specific DS equals 0.95. Under given conditions DS should be accepted in at least 90% of cases ($p(\beta) \leq 10\%$). When true probability of universal behavior equals 0.85, 90% of cases would be rejected ($p(\alpha) \leq 10\%$). True probabilities here define 'appropriate location of T'. In this

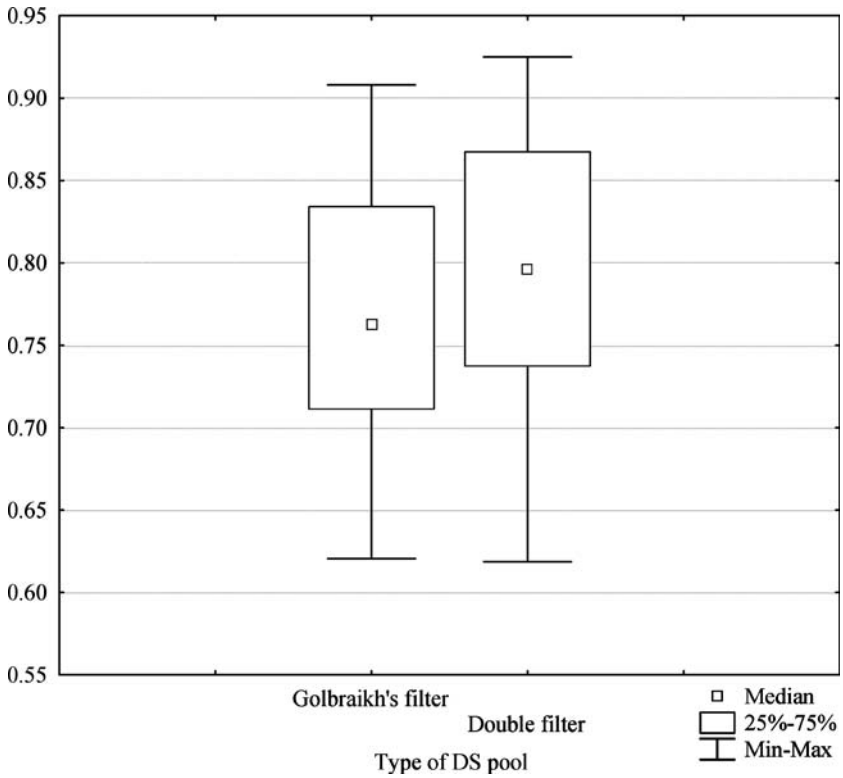

Figure 8. Comparison of $\langle R^2 \rangle$ distribution characteristics corresponding to external predictive performance of all DS that pass Golbraikh's filter against $\langle R^2 \rangle$ distribution characteristics corresponding to external predictive performance of DS that pass Golbraikh's and stability filter (double filter). *Y*-axis represents $\langle R^2 \rangle$. R^2 averaging is made over all members of specific DS set (universal, non-universal) for each complete dataset partition. $npart = 60$, $T = 6$.

way theoretical approach to receiver operating characteristic (ROC) has been defined and based on it optimal *npart* and T have been calculated. Under described assumptions corresponding optimal values for *npart* and T are 60 and 6, respectively. This means that if one makes 60 random partitions of complete dataset corresponding T value for rejection of hypothesis about universality of specific DS equals 6. In other words, if specific model fails selection in 6 or more cases of complete dataset partition and total number of partitions is 60 then such DS is not universal. Corresponding theoretical ROC curve is given in Figure 7.

Due to symmetric behavior of cumulative binomial distribution optimal T value falls near the middle of selected interval. Actual $p(\alpha)$ and $p(\beta)$ under given settings are 7.87% and 9.68%, respectively. We could calculate $p(\alpha)$ and $p(\beta)$ corresponding to 10 and 0 pair of values selected at the beginning of this research for *npart* and T. $p(\beta)$ equals 19.69% while $p(\alpha)$ is quite high and it equals 40.12%. In other words, probability of rejection of hypothesis about specific DS universality when true positive rate equals 0.95 is very large. This shows that such values could only be used for initial DS universality screening. Lower limit values for $p(\alpha)$ and $p(\beta)$ could be required and lower T interval limits seem attractive but such requests dramatically increase computing demands. Tight limitations simply lead to huge increase of optimal *npart*. Of course, tight limitations make decision-making protocol more convincing. For example, application of 5% value $p(\alpha)$ and $p(\beta)$ limits and $[0.01\times npart,\ 0.05\times npart]$ T interval leads to $npart = 181$ and $T = 5$. Therefore, selection of *npart* and T values directly determines DS universality definition and statistical properties of DS stability based selection at the same time.

At the end, it should be noticed that this analysis is grounded on assumption about binomial

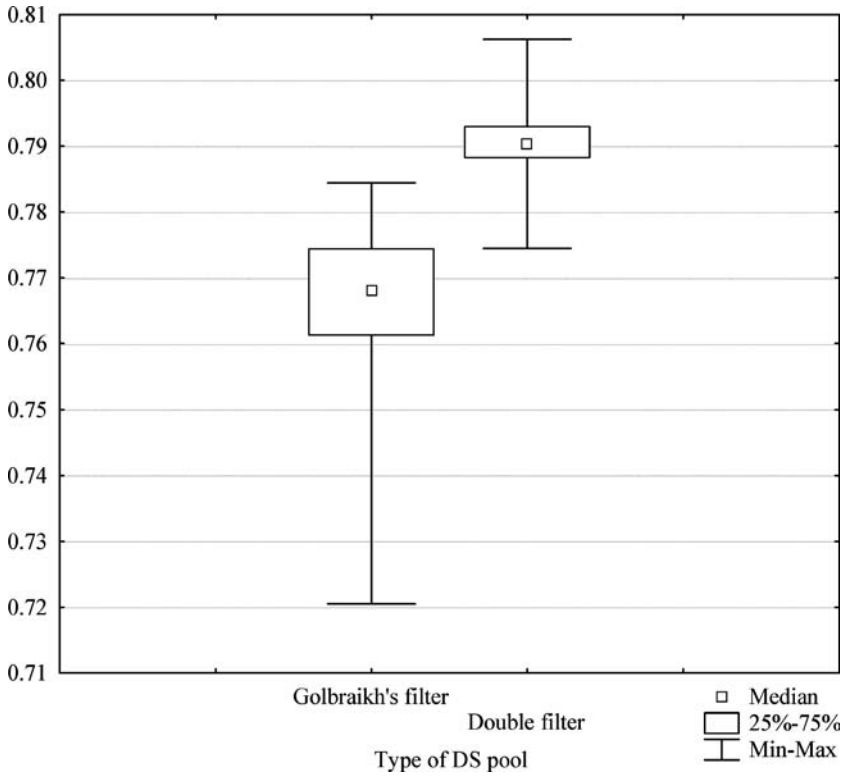

Figure 9. Comparison of $\langle q^2 \rangle$ distribution characteristics corresponding to internal validation performance of all DS that pass Golbraikh's filter against $\langle q^2 \rangle$ distribution characteristics corresponding to internal validation performance of DS that pass Golbraikh's and stability filter (double filter). *Y*-axis represents $\langle q^2 \rangle$. q^2 averaging is made over all members of specific DS sets (universal, non-universal) for each complete dataset partition. $npart = 60$, $T = 6$.

distribution of failures of specific DS to pass filters in multiple external validations. Simulation experiment could be used for generation of more appropriate ROC curves, but such experiment is beyond the scope of this article.

DS stability influence on predictive performance

DS stability is expected to have significant impact on QSAR model interpretation. Still, it is interesting to see whether DS stability affects external and internal predictive performance or not. In order to analyze this 60 partitions of complete dataset were made ($npart = 60$). For each of these partitions ENN training followed by application of double filter has been done. Cut off T value was set to 6. About 17 DS and corresponding QSAR models survived double filter based selection. $\langle R^2 \rangle$ results corresponding to all QSAR models that survive Golbraikh's filter are compared with results corresponding to all universal DS (Figure 8).

It is visible that certain improvement of $\langle R^2 \rangle$ could be achieved by implementation of stability filter in addition to Golbraikh's filter. Both, t-test for dependent samples and Wilcoxon matched pair test confirm significant differences between $\langle R^2 \rangle$ values corresponding to analyzed approaches to QSAR ensemble generation. These statistical hypothesis tests have been selected since the comparison is based on results obtained by application of both QSAR model ensemble-forming principles on each complete dataset partition. According to the statistics at least marginal differences between standard and novel approach exist i.e. application of novel approach could in some cases produce better $\langle R^2 \rangle$ results. More appropriate comparison could be done if one more nesting level was available. In other words, more independent external sets are required for such analysis. These sets could be used as a benchmark for independent comparison between standard and novel QSAR model ensemble predictive performance. Since benzodiazepine dataset is not very large further investigations in this direction are a part of our ongoing research. Briefly, we use double filter approach similar to one described in previous text for selection of acceptable trees and stable descriptors in random forest based QSAR modeling [42] of some

Table 2. Universal DS elements.

No. of successes	% of successes	DS element no.				
		1	2	3	4	5
57	95.00	$\pi 7$	F7	MR1	$\sigma_m 2'$	$\pi 8$
56	93.33	$\pi 7$	F7	$\mu 1$	MR1	$\mu 2'$
56	93.33	$\pi 7$	F7	MR1	R1	$\mu 2'$
56	93.33	$\pi 7$	F7	MR1	$\sigma_p 1$	$\mu 2'$
56	93.33	$\pi 7$	F7	MR1	$\mu 2'$	$\sigma_m 2'$
56	93.33	$\pi 7$	F7	MR1	$\mu 2'$	$\pi 8$
56	93.33	$\pi 7$	F7	MR1	$\mu 2'$	MR8
55	91.66	$\pi 7$	F7	MR1	R1	$\sigma_m 2'$
55	91.66	$\pi 7$	F7	MR1	$\sigma_p 1$	$\sigma_m 2'$
55	91.66	$\pi 7$	F7	MR1	$\mu 2'$	F2′
55	91.66	$\pi 7$	F7	MR1	$\mu 2'$	$\mu 6'$
55	91.66	$\pi 7$	F7	MR1	$\mu 2'$	MR6′
55	91.66	$\pi 7$	F7	MR1	$\mu 2'$	R8
55	91.66	$\pi 7$	F7	MR1	$\sigma_m 2'$	$\pi 6'$
55	91.66	$\pi 7$	F7	MR1	$\sigma_m 2'$	MR6′
55	91.66	$\pi 7$	F7	MR1	$\sigma_m 2'$	$\sigma_p 6'$
55	91.66	$\pi 7$	F7	MR1	$\sigma_m 2'$	F8

First part of descriptor symbol corresponds to physicochemical property of the specific substituent on the benzodiazepine structure represented by second part of symbol [4]. Benzodiazepine structure is given in Figure 10. Substituent physicochemical properties taken into account are: lipophilicity (π), polar constant (F), molar refractivity (MR), dipole moment (μ), resonance constant (R), Hammet meta (σ_m) and para constants (σ_p). The first column represents number of complete dataset partitions for which specific DS survived selection.

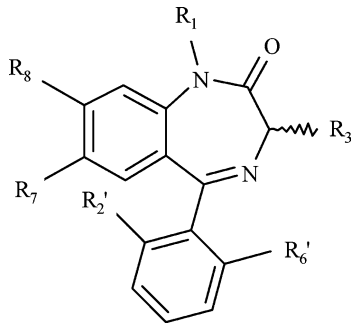

Figure 10. Benzodiazepine structure.

coumarin derivatives. Preliminary results confirm existence of universal solutions, in this case universal descriptors and even the reduction of predictive performance variation. However, either increase or decrease of predictive performance calculated on completely isolated dataset partitions based on application of double filter in comparison to non-filtered results has not been observed. Still, usage of acceptability filter alone leads to some predictive performance improvement. These findings leave the question about possible improvement of predictive performance due to application of different filters still open.

As shown in Figures 8 and 9, median $\langle R^2 \rangle$ values and median $\langle q^2 \rangle$ are almost the same in case of novel approach as well as in the case of standard approach. This is in agreement with analysis of relation between R^2 and q^2 values made by Golbraikh et al. [25]. Variation of $\langle q^2 \rangle$ values that correspond to acceptable QSAR models is very small while at the same time the variation of $\langle R^2 \rangle$ values is somewhat higher.

Figure 8 also shows existence of $\langle R^2 \rangle$ higher than 0.9. According to the same figure consideration of these results that correspond to one or two dataset partitions instead of consideration of the complete pool of results corresponding to all analyzed dataset partitions leads to incorrect predictive performance description. This finding suggests that predictive performance reports which are brought based on results corresponding to small number of complete dataset partitions i.e. hold out sets are in best case insufficient. It is interesting to notice that this finding is in accordance with recommendations made by authors who argue against usage of (single) hold out tests but in favor of LOO validation [28].

Table 3. Selected descriptor ranges that minimize IC50.

Descriptors	Descriptor range that minimizes IC50
MR1	[1.03, 1.03]
$\pi 7$	[−0.28, 0.71]
F7	[0.41, 0.67]
$\mu 2'$	[−1.59, −1.43]
$\sigma_m 2'$	[0.34, 0.37]

Interval limits represent descriptor minima and maxima corresponding to five most potent benzodiazepine derivatives used in this study (Tables 1 and 2 in Ref. [4]).

Benzodiazepine QSAR and description stability significance

As mentioned earlier under given conditions 17 universal DS have been found. These universal DS are given in Table 2 (Figure 10).

According to Table 2 all universal DS contain descriptors $\pi 7$, F7 and MR1. Two more descriptors exist in 10 and 8 out of 17 DS. These descriptors are $\mu 2'$ and $\sigma_m 2'$, respectively. Only one descriptor of five descriptors that constitute DS makes significant difference between most of the pairs of DS. This fact advocates for simplicity of benzodiazepine QSAR model interpretation provided by novel DS forming principle. It is expected that application of $npart = 181$ and $T = 5$, as suggested in previous text leads to even smaller number of eligible DS and smaller difference between DS pairs. Similar effect is expected if 4-2-1 NN topology is used.

In short, molecular refractivity of substituent 1 and lipophilicity as well as polarity of substituent 7 make major contribution to IC50. Substituent 2′ which dipole moment and Hammet meta constant

Table 4. Ten descriptors selected according to their relative frequencies of appearance in DS that pass Golbraikh's filter.

No.	Descriptors	Relative descriptor frequency (%)
1	MR1	99.52
2	$\pi 7$	98.31
3	F7	72.95
4	$\mu 2'$	44.32
5	$\sigma_m 2'$	35.74
6	F2′	19.44
7	$\sigma_m 7$	17.81
8	$\sigma_p 2'$	6.96
9	$\pi 2'$	6.67
10	$\mu 7$	6.41

are very frequent elements of universal DS is the third most important benzodiazepine pharmacological determinant. Based on selected descriptor values that characterize five most potent among all analyzed benzodiazepines Table 3 has been made.

According to study results benzodiazepine candidates should be characterized by descriptor values that lie within descriptor intervals given in Table 3. Instead of interval, only one value is assigned to MR1. This value corresponds to H atom that should be placed at position 1. $\pi 7$ interval indicates that substituent 7 should be characterized by approximately equal affinity for water and *n*-octanol. The same substituent should be as much polar as possible at the same time. Finally, mid-levels of $\mu 2'$ and $\sigma_m 2'$ should be achieved in order to improve benzodiazepine candidate IC50. Described descriptor value optimization is quite conservative since it is based solely on input dataset. Although extensive external model validation lowers the risk of extrapolation errors extrapolation has been avoided. Other IC50 minimization methods based on known functional dependence between IC50 and selected descriptors that involve extrapolation could be found in literature [4].

Table 4 represents selection of 10 most frequent descriptors that exist in DS pool formed by application of Golbraikh's criteria on all DS produced by ENN in each of 60 complete dataset partitions. Qualitative examination confirms correspondence between results obtained by application of standard and novel DS forming principles. Main problem with standard DS forming principle is the cut of value for inclusion of specific descriptor in final DS based on descriptor relative frequency of appearance. This number lacks theoretical framework [11]. The same assumptions about binomial distribution of failures used in previous paragraphs for universal DS selection could be applied this time on descriptors. According to this approach only the first two descriptors, namely MR1 and $\pi 7$ pass selection i.e. only MR1 and $\pi 7$ could be accepted as truly important benzodiazepine features. Since the number of biological determinants is very small this method resolves the problem of QSAR model interpretation in standard DS forming principle. The interpretation is almost the same as in the case of novel DS forming principle application.

Before we start with comparison of presented results with results published by others a few things should be stressed out. In comparison to NN topologies used by So and Karplus [4] and Maddalena and Johnston [32] NN topologies used in our approach are less complex. This is related to input layer size in the first place. So and Karplus used in some experiments topologies with as much as 30 input neurons. However, in most of the experiments they used 10-3-1 NN topology. In analysis that preceded this study influence of 3 hidden layer neurons on predictive performance was analyzed. It was found to be insignificant (results not shown). According to that hidden layer size does not make much impact on benzodiazepine QSAR results comparison. Still, input layer size differences could be important. Another important difference between analyzed benzodiazepine QSAR model development methods is validation method selection. While So and Karplus [4] and Maddalena and Johnston [32] used LOO q^2 we used multiple external validations besides internal LOO q^2 i.e. application of Golbraikh's filter on multiple random dataset partitions. Although Golbraikh's filter contains LOO q^2 based criterion model selection is dominated by external test set R^2 (results not shown). It is not easy to deduce potential influence of validation method difference on following comparison but some of result differences could be attributed to validation method selection.

Correspondence between results published by So and Karplus [4] and other authors that developed their own benzodiazepine QSAR with our results remains to be analyzed. So and Karplus used few best performing DS for descriptor frequency based final model development. MR1, $\pi 7$ and $\sigma_m 2'$ emerged as most important biological determinants. It is interesting to notice that F7 does not belong to the group of most frequent descriptors while $\sigma_m 7$ has similar importance as F7 in their benzodiazepine QSAR. Moreover, one of the most important descriptors that emerged in our analysis, namely $\mu 2'$ does not exist at all on list provided by So and Karplus (Table 4 from [4]). Plausible explanation of these differences is high correlation between F7 and $\sigma_m 7$ as well as $\mu 2'$ and $\sigma_m 2'$. $\sigma_m 2'$ is quite frequent descriptor among descriptors listed by So and Karplus [4]. Although F7 and $\sigma_m 7$ as well as $\mu 2'$ and $\sigma_m 2'$ are characterized by almost the same correlations with IC50

$\sigma_m 7$ and $\sigma_m 2'$ are very rare descriptors among all descriptors from DS pool produced by application of Golbraikh's filter. On the contrary, So and Karplus [4] did not put $\mu 2'$ on their list at all. This result shows that even when there is almost perfect correspondence between input variables accompanied by same correspondence of each input variable with output variable different modeling tools assign different preferences to each input.

Results provided by Maddalena and Johnston [32] are characterized by specific descriptor pattern, but most important descriptors from our point of view are retained in their approach. If one analyzes benzodiazepine QSAR model correspondence between these three groups of results from most important substituent perspective it could be stated that it is satisfactory. All approaches recognized that substituents 7, 1 and 2′ have significant influence on IC50. But model details contained in specific substituent descriptors are different in some instances.

Comprehensive quantitative results comparison is not possible since different validation protocols and somewhat different equations were used. Still, some general aspects could be analyzed. In comparison to results published by So and Karplus [4] less than a perfect internal LOO q^2 characterize most of benzodiazepine QSAR models obtained in this study (Figure 9). This result is a consequence of improvement of external validation R^2 values [25]. More interesting comparison would be comparison of external validation results. Unfortunately, external validation was not a widely accepted practice when analyzed publications were made. However, at the beginning of preliminary study we used LOO q^2 as FF used during wrapper training. Very high LOO q^2 values (>0.95) were achieved then, but external performance was poor. Therefore we abandoned this type of internal validation FF.

Finally, practical aspects of these methods are compared in the following text. In general, novel approach is not much more demanding than approach used by Mattioni et al. [11]. Still, large increase in CPU time needed for computations in comparison to So and Karplus [4] and Maddalena and Johnston [32] is obvious. Applications of multiple external validations and/or multiple LMO FF are main causes of large increase of CPU demands. We have shown that single dataset partition results are insufficient for appropriate predictive performance description. Therefore multiple external validations are unavoidable. One of the goals of any type of QSAR is to find model with acceptable performance. It has been shown by others [21, 23, 24] that internal multiple LMO FF correlates well with external validation results. This fact makes multiple LMO FF suitable training FF that drives evolutionary NN training towards selection of models characterized by acceptable external predictive performance. Unessential performance differences between models that have 3 hidden layer neurons in comparison to models that have 2 hidden layer neurons advocate for application of GA partial least squares hybrid [20] instead of ENN if CPU time is limitation factor. This switch can save significant amount of CPU time. Another option is application of taboo search [21, 23, 24] or some similar descriptor selection tool instead of CPU time consuming GA. In our current research we use random forest method which application is characterized by significant reduction of CPU demands along with acceptable predictive performance.

Conclusions

Standard and novel ENN based QSAR model-forming principles were applied on benzodiazepine dataset. Predictive performance and resultant pharmacological determinants were examined. External $\langle R^2 \rangle$ coefficient of variation (CV) calculated over 60 partitions i.e. 60 ENN ensemble training experiments was $\sim$10% while internal $\langle q^2 \rangle$ CV was less than 2% for both approaches. Although $\langle R^2 \rangle$ CV values are somewhat higher than expected overall predictive performance stability is considered to be acceptable. Predictive performance itself is acceptable in both cases. Reasonable $\langle R^2 \rangle$ and corresponding CV values also show that possible existence of molecular clones does not affect predictive performance reliability.

Besides acceptable and stable predictive performance DS stability has been achieved by application of novel approach. Under given assumptions 17 universal and very similar DS were collected. Any of them fails to pass Golbraikh's filter step in less than 6 occasions i.e. in less than 10% of all experiments. On the other hand, standard model building principle based on

ENN ensemble results obtained on multiple dataset partitions was improved. Theoretical framework based on binomial distribution properties for pharmacological determinants selection from a set of many candidates produced by standard final DS forming principle is given. Direct consequence of presented results is simple and straightforward final QSAR model interpretation.

Benzodiazepine QSAR has been revisited. Minor differences among analyzed model building methods were detected. Both studied approaches designated MR1 and $\pi 7$ as benzodiazepine pharmacological determinants. Besides these two descriptors some of 2′ substituent specific descriptors play an important role in IC50 predictions according to analyzed methods. It is quite surprising that major benzodiazepine substituents recognized in this study almost perfectly correspond to substituents reported by other authors who did not use external validation and multiple dataset partitions. On the other hand, selected descriptors could be affected by chosen validation protocol. It can be concluded that these validation protocols make little impact on model interpretation when molecules are solely represented by substituent specific descriptors. This is even more obvious in cases when descriptors corresponding to a single substituent are correlated. It would be interesting to see how these validation protocols affect descriptor selection in cases when molecules are represented by descriptors that describe molecule as a whole.

Finally, DS stability related issues are not the only important issues related to QSAR model interpretation. As Kohavi and John [43] pointed out relevant attributes need not to correspond to good predictors. This raises important question whether DS stability could establish more reliable link between descriptor predictive performance and its relevance. Together with detailed analysis related to DS stability in different settings (different datasets and different prediction tools) this issue represents an important topic for the future research.

Notes

1 Interested readers should send their requests by e-mail to Željko Debeljak, zdebelja@inet.hr

2 Possible solution of negative operator interaction problem is application of higher number of chromosomes per generation and/or application of evolutionary operator frequency modulation, both needed for improvement of large NN training efficiency.

Acknowledgement

This work was supported by the Ministry of Science and Technology of the Republic of Croatia through Grant 0006541.

References

1. Guyon, I. and Elisseef, A., J. Mach. Learn. Res., 3 (2003) 1157.
2. Molina, L.C., Belanche, L. and Nebot, A., 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, September 09–12, 2002.
3. So, S.S. and Karplus, M., J. Med. Chem., 39 (1996) 1521.
4. So, S.S. and Karplus, M., J. Med. Chem., 39 (1996) 5246.
5. So, S.S. and Karplus, M., J. Med. Chem., 40 (1997) 4347.
6. So, S.S. and Karplus, M., J. Med. Chem., 40 (1997) 4360.
7. So, S.S., van Helden, S.P., van Geerestein, V.J. and Karplus, M., J. Chem. Inf. Comput. Sci., 40 (2000) 762.
8. Kyngäs, J. and Valjakka, J. Quant. Struct.-Act. Relat., 15 (1996) 296.
9. Patankar, S.J. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 40 (2000) 706.
10. Patankar, S.J. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 42 (2002) 1053.
11. Mattioni, B.E., Kauffman, G.W., Jurs, P.C., Custer, L.L., Durham, S.K. and Pearl, G.M., J. Chem. Inf. Comput. Sci., 43 (2003) 949.
12. Hemmateenejad, B., Akhond, M., Miri, R. and Shamsipur, M., J. Chem. Inf. Comput. Sci., 43 (2003) 1328.
13. Kohavi, R., A study of cross-validation and bootstrap for accuracy estimation and model selection, Technical Report. Computer Science Department, Stanford University, Stanford, 1995.
14. Lunneborg, C.E., Data Analysis by Resampling: Concepts and Applications. Duxbury Press, Pacific Grove, 2000.
15. Breiman, L., Statist. Sci., 16 (2001) 199.
16. Breiman, L., Bagging predictors, Technical Report. Department of Statistics, University of California, Berkley, 1994.
17. Breiman, L., Mach. Learn., 45 (2001) 5.
18. Tetko, I.V., Livingstone, D.J. and Luik, A.I., J. Chem. Inf. Comput. Sci., 35 (1995) 826.
19. Yao, X. and Liu, Y., IEEE Trans. Syst. Man. Cybern. B Cybern., 28 (1998) 417.
20. Leardi, R. and Lupianez Gonzalez, A., Chemometr. Intell. Lab. Syst., 41 (1998) 195.
21. Baumann, K., Trends Anal. Chem., 22 (2003) 395.
22. Kohavi, R. and Sommerfeld, D., Feature subset selection using wrapper method: overfitting and dynamic search

space topology, Technical Report. Computer Science Department, Stanford University, Stanford, 1995.
23. Baumann, K., Albert, H. and von Korff, M., J. Chemomet., 16 (2002) 339.
24. Baumann, K., von Korff, M. and Albert, H., J. Chemomet., 16 (2002) 351.
25. Golbraikh, A. and Tropsha, A., J. Mol. Graphics Model., 20 (2002) 269.
26. Tropsha, A., Gramatica, P. and Gombar, V.K., QSAR Comb. Sci., 22 (2003) 69.
27. Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H. and Tropsha, A., J. Comput.-Aided Mol. Des., 17 (2003) 241.
28. Hawkins, D.M., Basak, S.C. and Mills, D., J. Chem. Inf. Comput. Sci., 43 (2003) 579.
29. Todeschini, R., Consonni, V., Mauri, A. and Pavan, M., Anal. Chim. Acta, 515 (2004) 199.
30. Clark, R.D., J. Comput.-Aided Mol. Des., 17 (2003) 265.
31. Shen, Q., Jiang, J.-H., Shen, G.-L. and Yu, R.-Q., Anal. Bioanal. Chem., 375 (2003) 248.
32. Maddalena, D.J. and Johnston, G.A.R., J. Med. Chem., 38 (1995) 715.
33. Aoyama, T., Suzuki, Y. and Ichikawa, H., J. Med. Chem., 33 (1990) 2583.
34. Abraham, A., Optimization of Evolutionary Neural Networks Using Hybrid Learning Algorithms, Technical Report, School of Business Systems, Monash University, 2002.
35. Zell, A. (Ed.), Stuttgart Neural Network Simulator User Manual, Version 4.2. University of Stuttgart and University of Tübingen, 1998.
36. Haykin, S. Neural Networks: A Comprehensive Foundation, 2nd ed., Prentice-Hall Inc, Upper Saddle River, NJ, 1999.
37. Gasteiger, J. and Zupan, J., Angew. Chem. Int. Ed. Engl., 32 (1993) 503.
38. Møller, M.F., Neural Networks, 6 (1993) 525.
39. Chiu, T.-L. and So, S.S., QSAR Comb. Sci., 22 (2003) 519.
40. Wagener, M., Sadowski, J. and Gasteiger, J., J. Am. Chem. Soc., 117 (1995) 7769.
41. Dudewitz, E.J. and Mishra, S.N. Modern Mathematical Statistics. John Wiley and Sons, New York, 1988.
42. Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P., J. Chem. Inf. Comput. Sci., 43 (2003) 1947.
43. Kohavi, R. and John, G.H., Artif. Intell., 97 (1997) 273.