Cent. Eur. J. Chem. • 9(1) • 2011 • 165-174
DOI: 10.2478/s11532-010-0135-7

VERSITA

Central European Journal of **Chemistry**

Analysis of the co-evolutions of correlations as a tool for QSAR-modeling of carcinogenicity: an unexpected good prediction based on a model that seems untrustworthy

Research Article

Alla P. Toropova[1], Andrey A. Toropov[1,*], Rodolfo Gonella Diaza[1],
Emilio Benfenati[1], Guesippina Gini[2]

[1]*Institute of Pharmacologic Researches by Mario Negri,
20156 Milan, Italy*

[2]*Departmen of Electronics and Information,
Polytechnic Institute of Milan, 20133 Milan, Italy*

Received 11 August 2010; Accepted 30 October 2010

Abstract: To validate QSAR models an external test set is increasingly used. However the definition of the compounds for the test set is still debated. We studied, co-evolutions of correlations between optimal descriptors and carcinogenicity (pTD50) for the sub-training, calibration, and test set. Weak correlations for the sub-training set are sometimes accompanied by quite good correlations for the external test set. This can be explained in terms of the probability theory and can help define a suitable test set. The simplified molecular input line entry system (SMILES) was used to represent the molecular structure. Correlation weights for calculating the optimal descriptors are related to fragments of the SMILES. The statistical quality of the model is: n=170, r^2=0.6638, q^2=0.6554, s=0.828, F=331 (sub-training set); n=170, r^2=0.6609, r^2_{pred}=0.6520, s=0.825, F= 331 (calibration set); and n=61, r^2=0.7796, r^2_{pred}=0.7658, $R_m{}^2$=0.7448, s=0.563, F=221 (test set). The calculations were done with CORAL software (http://www.insilico.eu/coral/).

Keywords: *QSAR • Carcinogenicity • Optimal descriptor • Balance of correlations • Co-evolutions of correlations*

1. Introduction

There is growing debates about quantitative structure - activity relationships (QSAR) models [1-5]. The use of QSAR for regulatory purposes is widely discussed at European Chemical Agency (http://echa.europa.eu/) as well as at Organization for Economic Co-operation and Development (http://www.oecd.org/dataoecd/33/37/37849783.pdf). An important aspect of this problem is the preliminary estimation of ecologic effect of various pesticides [6]. This would require an evaluation of the models performance, which includes verification of the models predictive power using a series of statistical tools [7,8]. The risk of a model being over-trained is well known: this is a situation where the satisfactory prediction for the training set is accompanied by the unsatisfactory prediction for the external test set. One solution accepted for validation is the use of an external set of compounds that have never been used to build up the model.

Some authors suggest that the training set should be smaller than 75-85% of the original data set. We have experimentally assessed how this size affects results, considering data on carcinogenicity, which is a complex biochemical phenomenon. For this aim, five distributions into subtraining, calibration, and test sets, *i.e.*, 134-134-133 where training and calibration sets are 66%, 170-170-61 (85%), 255-85-61 (85%), 85-255-61 (85%), and 185-185-31 (92%), respectively, were examined. Substances characterized by extremely poor prediction of the carcinogenic potential were extracted during preliminary Monte Carlo experiments. These were inserted in subtraining sets. Using 401 chemicals, we evaluated the probabilistic principles of external validation of QSAR models for carcinogenicity (pTD50) calculated

* *E-mail: andrey.toropov@marionegri.it*

Springer

with optimal descriptors based on the representation of the molecular structure by the simplified molecular input line entry system (SMILES) [7,8].

Thus, the study of measure of influence on the quality of prediction of the number of substances involved in the modeling process (by means of the optimal descriptors) is the aim of this work.

2. Calculating Details

2.1. Data

Experimental values for the carcinogenicity of 401 organic compounds were taken from the Internet (at http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html, Accessed 05.03.2010). We examined substances with numerical data on their carcinogenicity, and a positive carcinogenic dose. Thus, carcinogenicity is expressed as the dose that induces cancer in male and female rats (TD50, in mg/kg body weight). These values were converted into mmol/kg body weight and we examined the pTD50 (*i.e.*, the negative decimal logarithm of TD50) as the endpoint. We divided the compounds into sub-training, calibration, and test sets. The calibration set is a preliminary test set: these substances are used for optimizing the balance of correlations in order to avoid overtraining.

We studied five distributions of chemicals: 134-134-133, 170-170-61, 255-85-61, 85-255-61, and 185-185-31. Each split was randomly composed. However substances with 'atypical behavior' (*i.e.*, substances for which predicted and experimental values are strongly different for a series of probes of the Monte Carlo optimization) were inserted in the sub-training set. For each distribution three experiments (*i.e.*, the Monte Carlo optimization) were done, repeating the random splitting three times.

Taking into account the complex nature of carcinogenicity phenomenon, the placement of the 'atypical' substances in the training set can be considered as realistic conditions for the computational experiments.

2.2. Optimal descriptor

Optimal descriptors are calculated as the following

$$DCW(T)=\sum_{k=1}^{E}W(^1S_k)+\sum_{k=1}^{E-1}W(^2S_k)+\sum_{k=1}^{E-2}W(^3S_k) \qquad (1)$$

where $^1S_{k'}$ $^2S_{k'}$ and ^3S_k are one-, two-, and three-elements SMILES attributes; E is the total number of SMILES elements for a given molecular structure; $W(^1S_k)$, $W(^2S_k)$, and $W(^3S_k)$ are the correlation weights of the attributes. The SMILES element comprises one or two symbols which should be examined as one (*e.g.* 'Cl', 'Br', *etc.*). The threshold is a value used to classify attributes as either rare or active. For instance, if the threshold is 5, then attributes found in four (or fewer) SMILES structures of the training set should be classified as rare. The correlation weights of rare attributes are blocked with their values fixed at zero. E is the number of $^1S_{k'}$. If a SMILES is a sequence of element 'ABCDE', then the construction $^1S_{k'}$, $^2S_{k'}$ and ^3S_k can be represented as:

'ABCDE' → 'A', 'B','C', 'D', 'E' ($^1S_{k'}$)
'ABCDE' → 'AB', 'BC','CD', 'DE' ($^2S_{k'}$)
'ABCDE' → 'ABC', 'BCD','CDE' ($^3S_{k'}$)

Correlation weights (for calculation with Eq. 1) were defined by the Monte Carlo method optimization procedure using three functional sets of compounds: the sub-training, calibration, and test sets. Optimization is based on correlation coefficients between the DCW(T) and pTD50 for the sub-training and calibration sets. The target function is the following:

$$\mathbf{TF=R+R'-abs(R\text{-}R')*dR_{weight}-} \\ \mathbf{-abs(C0+C0'+C1\text{-}C1')*dC_{weight}} \qquad (2)$$

where R and R' are correlation coefficients between endpoint and optimal descriptor for the sub-training and calibration sets; C0 and C0' are intercepts for the sub-training and calibration sets; C1 and C1' are slopes for the sub-training and calibration sets. $\mathbf{dR_{weight}}$ and $\mathbf{dC_{weight}}$ are empirical parameters.

Thus, the sub-training set is used to construct the model; the calibration set is a preliminary test of the model (in order to avoid overtraining), and the test set is used for final assessment of the model.

2.3. Co-evolution of correlations

The number of epochs (N_{epoch}) of the training (optimization) clearly influences the statistical quality of the model. We examined the correlation coefficients for the sub-training, calibration, and test sets for a range of the N_{epoch} from 1 to 50. The calculations were done with CORAL freeware (at http://www.insilico.eu/coral/, Accessed 05.05.2010).

3. Results and Discussion

Fig. 1 shows the co-evolutions of correlations for three random 134-134-133 splits into sub-training, calibration and test sets with range from 1 to 50 the number of

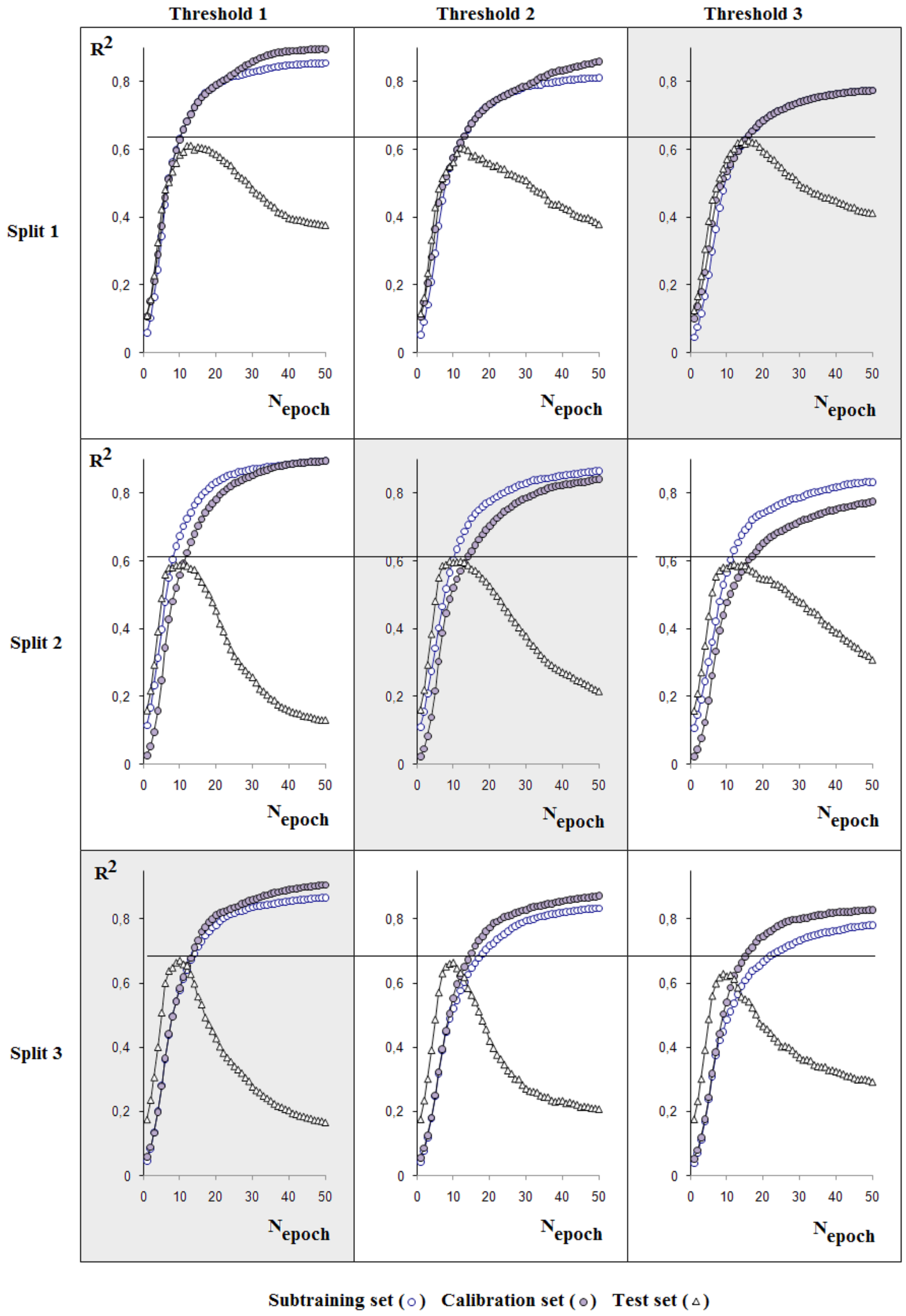

Figure 1. The 134-134-133 models: best predictions (maximum of the r^2 for test set) are indicated by a grey background.

epochs of the Monte Carlo optimization. Figs. 2-5 show the results for the 170-170-61, 255-85-61, 85-255-61, and 185-185-31, splits.

There is a reproducibility of results of the Monte Carlo optimization (each computational experiment is repeated three times).

The statistical quality of the model on the external test set was used as the criterion for defining the best N_{epoch}. According to [8], the preferable N_{epoch} (to obtain the maximum correlation coefficient for the test set) is 10. In Figs. 1-5, the preferable value of the N_{epoch} is near 10, but not exactly 10.

An example of the co-evolutions of correlations for the first 170-170-61 split (this was recently examined in [8]), where the preferable N_{epoch} is 13. Fig. 6 shows the correlation coefficient between the experimental pTD50 and pTD50 calculated for the 170-170-61 split as a mathematical function of the threshold and N_{epoch}. The statistical quality of the model for 170-170-61 (split 1) is best when the threshold is 2 and there are 13 epochs of optimization (Fig. 6).

For the first 170-170-61 split the model for pTD50, obtained with the number of epochs of the training (optimization) N_{epoch}=13 and a threshold of 2, is the following:

$$\mathbf{pTD50 = -0.1602(\pm 0.0082) + +0.0945(\pm 0.0004) * DCW(2)} \qquad (3)$$

n=170, r^2=0.6638, q^2=0.6554, s=0.828, F=331 (sub-training set)

n=170, r^2=0.6609, r^2_{pred}=0.6520, s=0.825, F= 331(calibration set)

n=61, r^2=0.7796, r^2_{pred}=0.7658, R_m^2=0.7448, s=0.563, F=221 (test set)

where n is the number of compounds in the set; r is the correlation coefficient; q^2 is the determination coefficient of the LOO-cross-validation for the sub-training set; r^2_{pred} is the determination coefficient of the LOO-cross-validation for the calibration and test sets; s is the standard error of estimation; and the F is the Fisher F-ratio. R_m^2 is a measure of the predictability of the model. According to [9], a model is satisfactory if the R_m^2 is larger than 0.5.

Fig. 7 shows graphically the model calculated with Eq. 3.

The statistical characteristics of the model calculated with Eq. 3 for the test set are better than those of the models described in [7] (n=61, $r^2_{(test)}$=0.723, s=0.676, F=164) and the model described in [8] (n=61, $r^2_{(test)}$=0.7541, s=0.682, F=181). The model calculated with Eq. 3 is also simpler than the above-mentioned SMILES-based models, because it is built up without the dC parameter [7,8]. This improvement reflects the rational selection of the N_{epoch}.

The range of the threshold 1, 2, and 3 is selected from the following reasons. Threshold 0 is nonsense, because in this case, attributes which are absent in the sub-training set can be involved in the modeling process. In the case of the threshold 4, the statistical quality of carcinogenicity models becomes poorer. Having denoted this circumstance, we have described results obtained with threshold 1, 2, and 3. It is possible to carry out the described calculations with 5 or 7 splits for each distributions (*i.e.,* 134-134-133, 170-170-61, *etc*.). However, in this case, statistical quality of the results will be approximately the same, in spite of increase of the number of splits.

The property/activity can be examined as a mathematical function of molecular structure represented by SMILES elements and/or attributes (combinations of SMILES elements). Some substances show ‘average’ behavior: their SMILES attributes provide the necessary information for adequate calculating their pTD50 (with a model similar to Eq. 3). Some substances show ‘atypical’ behavior as regards the pTD50: their SMILES attributes do not take into account some important features of the real molecule (in real conditions) which influence on pTD50. It was noted above that all substances with ‘atypical behavior’ were defined in preliminary experiments and were inserted in the sub-training set.

We suggest that overtraining in the case of CORAL modeling is a ‘compensation’ of the impossibility of involving these important features of molecules for the process of modeling. Under these circumstances, available weights contribute to improving of the statistical quality for the sub-training set (and maybe also for the calibration set), but they give misleading information for the external test set.

There are two phases of the Monte Carlo optimization. The first phase involves an increase of the correlation coefficient between DCW(T) and pTD50 for the sub-training, calibration, and test sets. The second phase involves an increase in this correlation coefficient for the sub-training and calibration sets and a decrease of the correlation coefficient for the external test set (Figs. 1-5). The critical point between these two phases (*i.e*., the specific number of iterations, N_{epoch}, for the Monte Carlo optimization) can be an indicator of the preferable model.

The majority of substances show ‘average’ behavior (Fig. 7), and they are the basis for building up the pTD50 model. However, there are substances with ‘atypical’ behavior in the sub-training set. During the first phase of the Monte Carlo optimization the main contribution for building up of the model comes from substances

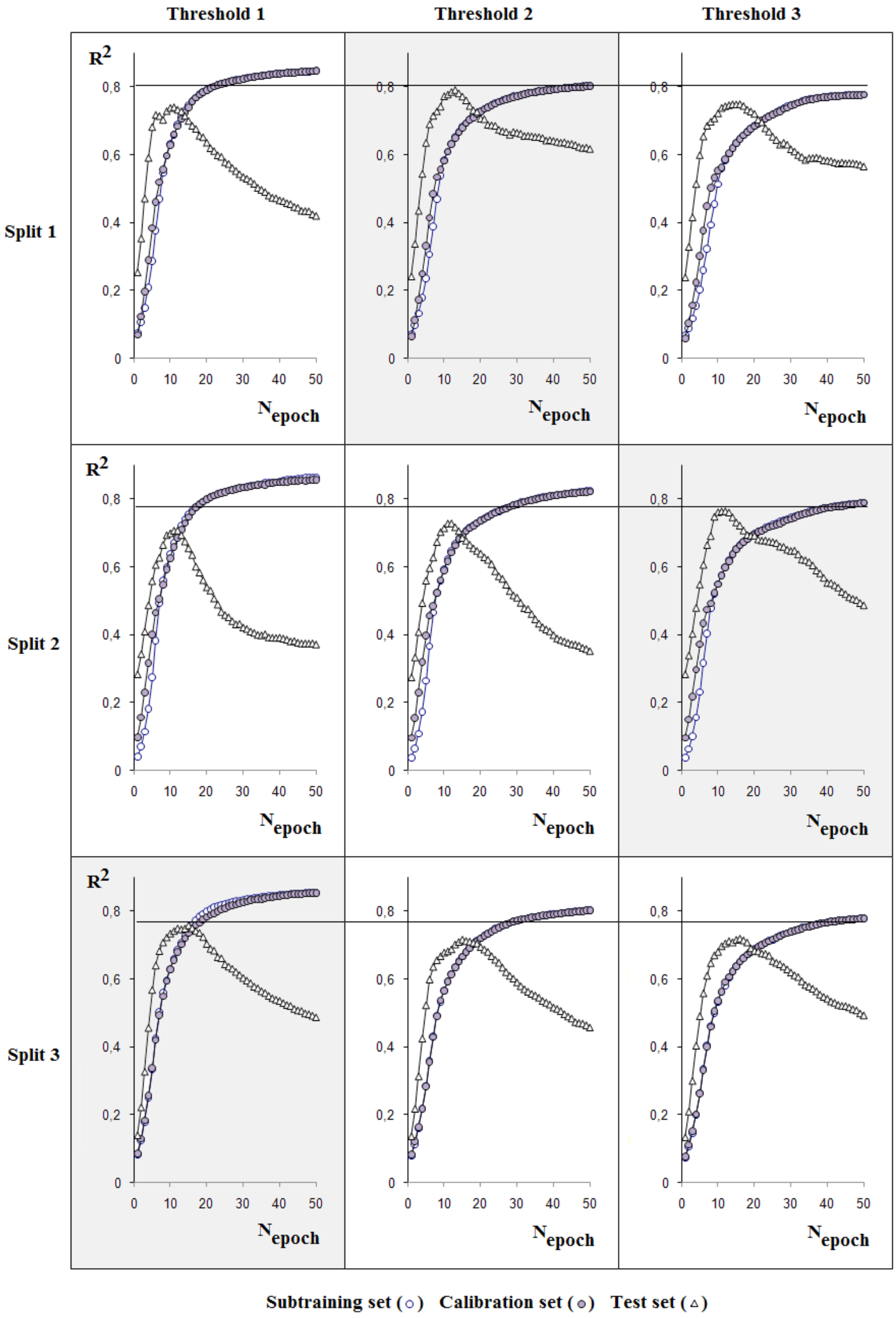

Figure 2. The 170-170-61 models: best predictions (maximum of the r^2 for test set) are indicated by a grey background.

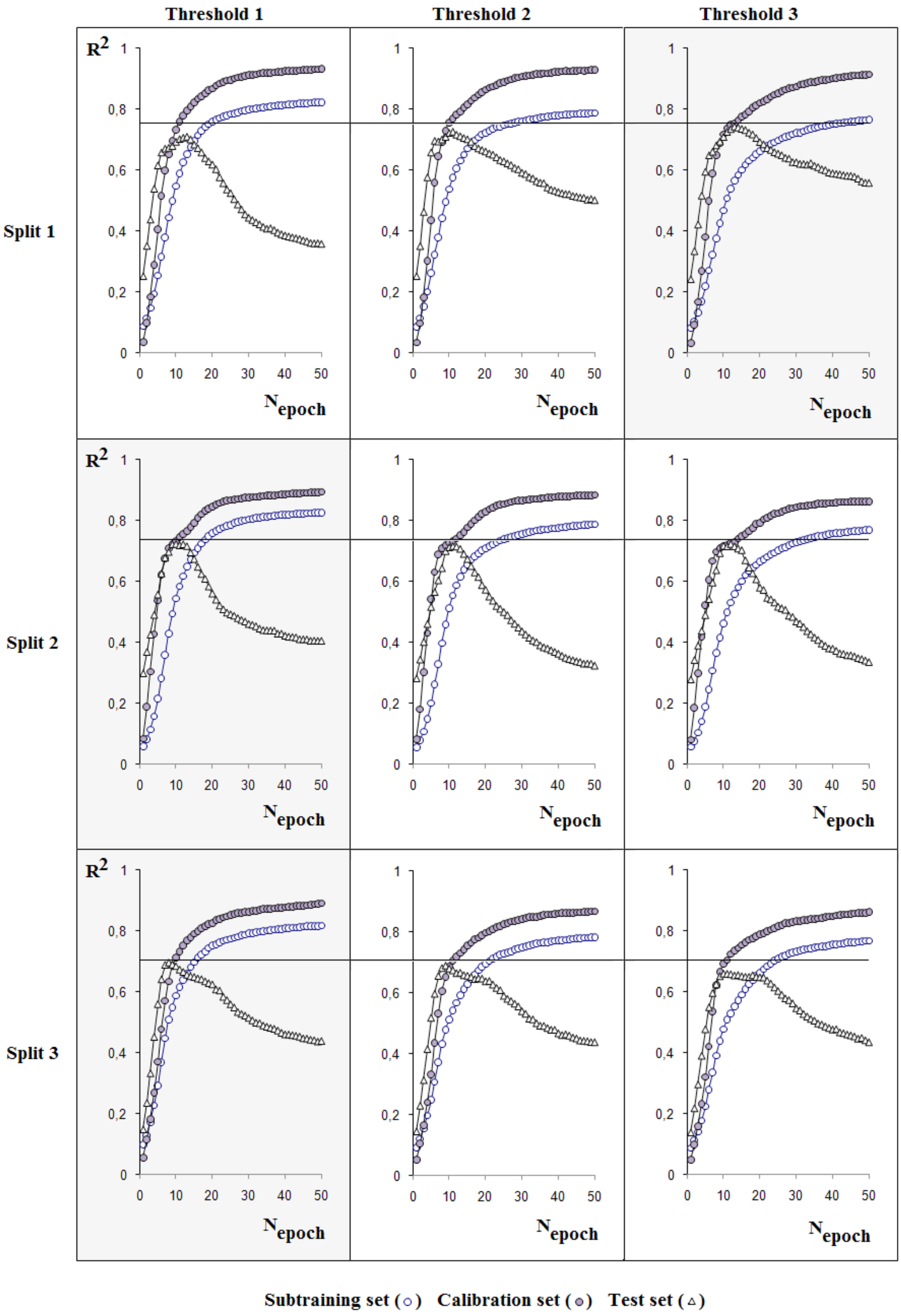

Figure 3. The 255-85-61 models: best predictions (maximum of the r^2 for test set) are indicated by a grey background.

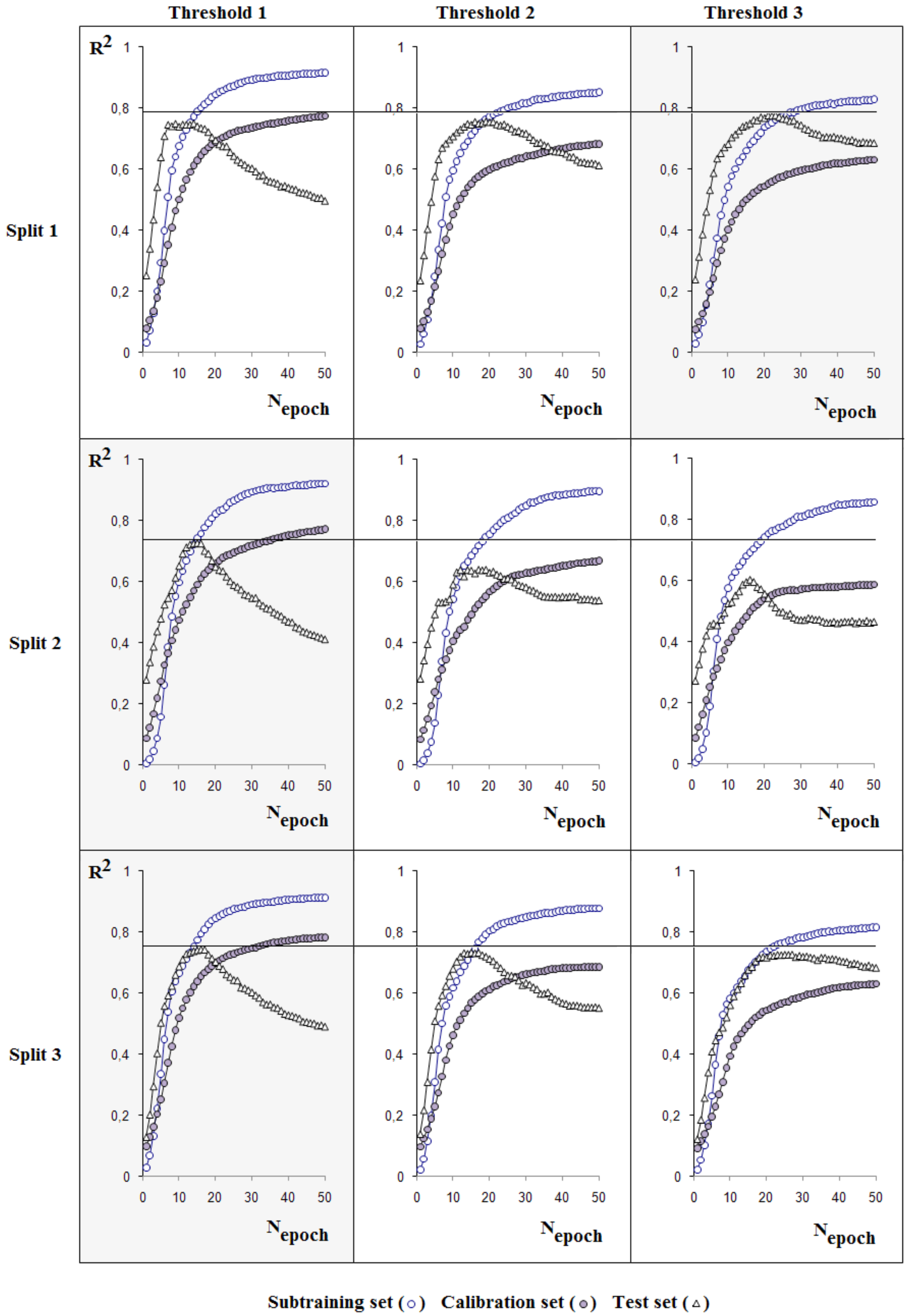

Figure 4. The 85-255-61 models: best predictions (maximum of the r^2 for test set) are indicated by a grey background.

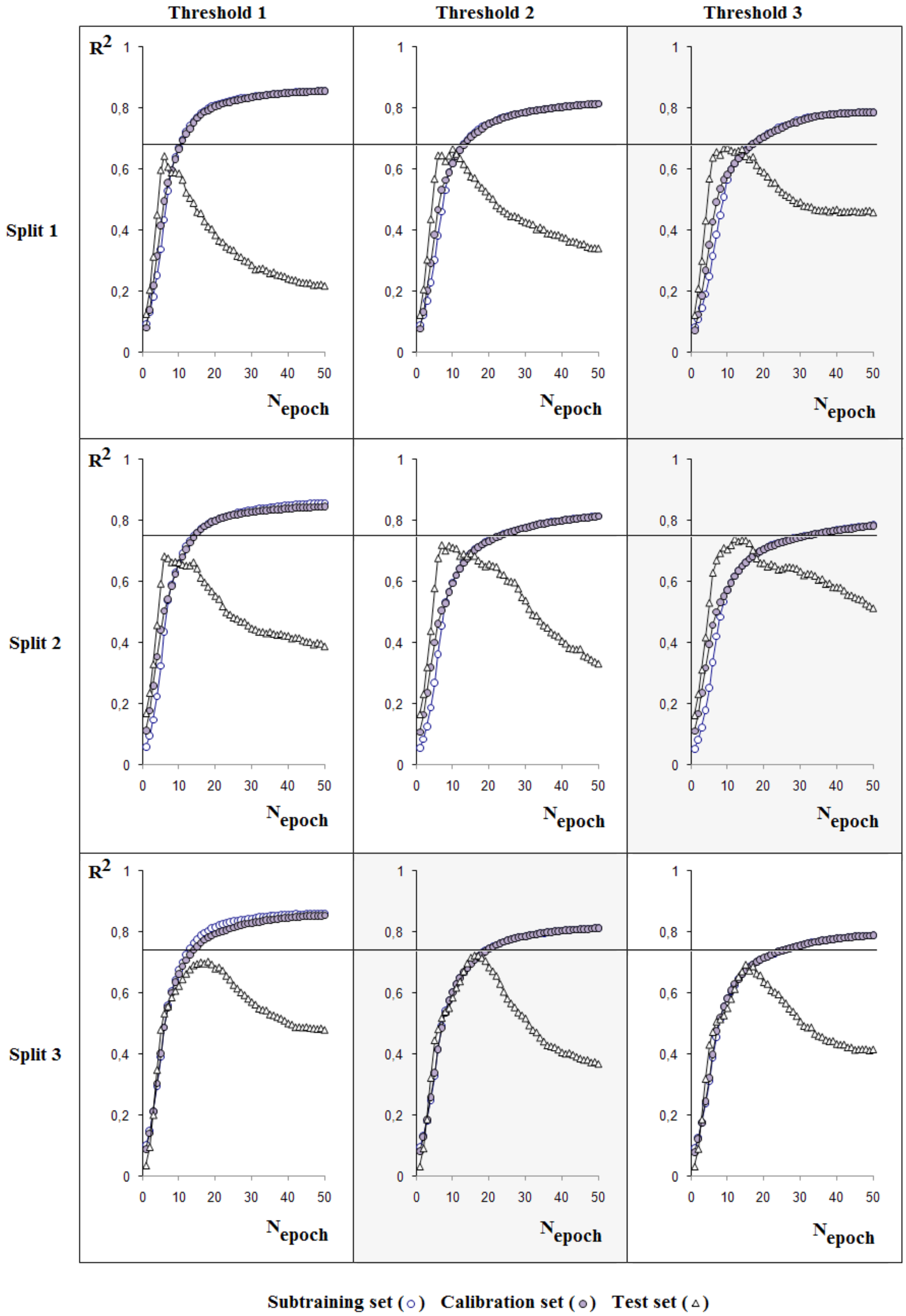

Figure 5. The 185-185-31 models: best predictions (maximum of the r^2 for test set) are indicated by a grey background.

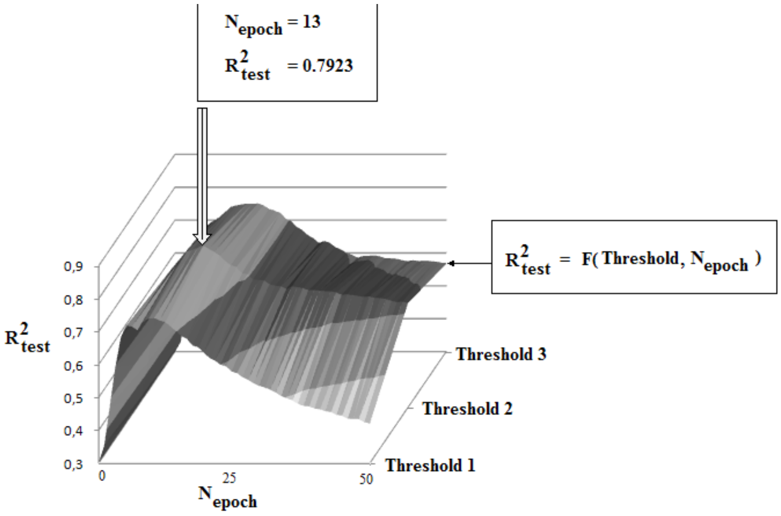

Figure 6. Correlation coefficient between the pTD50 experimental and pTD50 calculated with Eq. 3, as a mathematical function of the threshold and N_{epoch}.

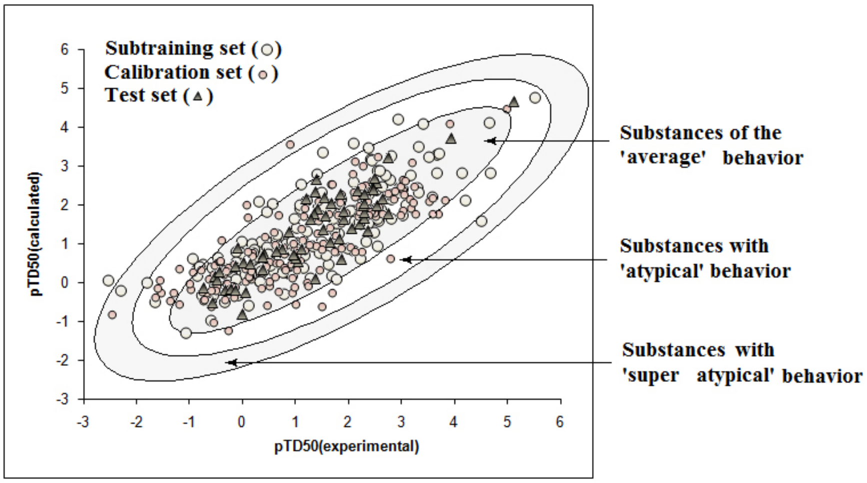

Figure 7. Experimental pTD50 values and those calculated with Eq. 3.

Table 1. Statistical quality of QSAR for carcinogenicity obtained with taking into account co-evolution of correlations for five distributions.

Distribution*	Split	T**	Aver±disp***	N_{epoch}	Aver±disp	$r^2_{(test)}$	Aver±disp	$s_{(test)}$	Aver±disp
134-134-133	1	3	2±0.82	14	11.0±2.16	0.6231	0.63±0.03	0.829	0.828±0.006
	2	2		9		0.5981		0.820	
	3	1		10		0.6691		0.836	
235-85-61	1	3	1.7±0.94	15	11.3±2.62	0.7415	0.72±0.02	0.610	0.665±0.039
	2	1		10		0.7222		0.700	
	3	1		9		0.6900		0.686	
85-235-61	1	3	1.7±0.94	23	18.0±3.55	0.7755	0.75±0.02	0.604	0.640±0.035
	2	1		15		0.7259		0.688	
	3	1		16		0.7416		0.629	
170-170-61	1	2	2.0±0.00	13	13.7±0.94	0.7796	0.75±0.02	0.563	0.617±0.039
	2	2		13		0.7609		0.635	
	3	2		15		0.7207		0.652	
185-185-61	1	2	2.0±0.00	10	12.7±2.49	0.6677	0.71±0.03	0.578	0.597±0.021
	2	2		12		0.7382		0.627	
	3	2		16		0.7220		0.585	

*) *Distribution = the number of chemicals in sub training set – the number of chemicals in calibration set – the number of chemicals in test set;*
**) *T = threshold;*
***) *Aver = average, disp=dispersion.*

with ‘average’ behavior. When the informative reserve of the substances of ‘average’ behavior is expired, the overtraining starts. The essence of overtraining involves modification of the correlation weights of available attributes to improve the model for the sub-training set. As noted before, unfortunately this reduces the predictive potential of the model for the external test set.

However, the preferable N_{epoch} value can be obtained from computational experiments (Figs. 1-5). These results (statistical quality of the models) are reproduced in a series of the probes of the Monte Carlo method optimization. Thus, the statistical quality of the models is reproducible. It is best for the external test set for the 170-170-61 model. Thus, the size of the test set is about 15% of the total compounds. However, for all splits, curves in coordinates of N_{epoch} against the correlation coefficient (r^2, for the external test set) show a maximum that is an indicator of the preferable N_{epoch}. In agreement with a previous report [8] the best N_{epoch} is about 10, but it varies for different splits into sub-training, calibration, and test sets. *Supplementary materials* section contains representation of Figs. 1-5 and technical details for model calculated with Eq. 3.

We deem the generalized reasonable empirical rule for the definition of N_{epoch} is the following: the Monte Carlo optimization should be stopped when $0.5*[r^2_{(sub\text{-}training)} + r^2_{(calibration)}] \approx 0.7$, where r is correlation coefficient. However, this rule is formulated for models of the examined carcinogenicity.

4. Conclusions

Fifteen splits into sub-training, calibration, and test sets have been studied in QSAR analysis of carcinogenicity of 401 substance by means of optimal SMILES-based descriptors.

Analysis of co-evolution of correlations (*i.e.*, correlation coefficients for sub-training, calibration, and test sets for each epoch of the Monte Carlo optimization) can be used for definition of the preferable number of epochs.

In the case of models for carcinogenicity the preferable number of epochs can be defined by empirical rule: the Monte Carlo optimization should be stopped when $0.5*[r^2_{(sub\text{-}training)} + r^2_{(calibration)}] \approx 0.7$.

Acknowledgement

The authors express their gratitude to OSIRIS for financial support and to Dr. L. Cappellini (Istituto di Ricerche Farmacologiche Mario Negri, Milano) for technical assistance.

References

[1] P.P. Roy, J.T. Leonard, K. Roy, Chemomet. Intell. Lab. 90, 31 (2008)

[2] W. Tong, Q. Xie, H. Hong, L. Shi, H. Fang, R. Perkins, Environ. Health. Persp. 112, 1249 (2004)

[3] A.A.Toropov, A.P.Toropova, D.V.Mukhamedzhanova, I. Gutman, Indian J. Chem. 4A,1545 (2005)

[4] G. Melagraki, A. Afantitis, H. Sarimveis, P.A. Koutentis, G. Kollias, O. Igglessi-Markopoulou, Mol. Divers. 13, 301 (2009)

[5] E. Vicente, P.R. Duchowicz, E.A. Castro, A. Monge, J. Mol. Graph. Model. 28, 28 (2009)

[6] E. Benfenati (Ed.), Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes (Elsevier Science, Amsterdam, 2007)

[7] A.A. Toropov, A.P. Toropova, E. Benfenati, A. Manganaro, Mol. Divers. 13, 367 (2009)

[8] A.A. Toropov, A.P. Toropova, E. Benfenati, Int. J. Mol. Sci. 10, 3106 (2009)

[9] P.P. Roy, K. Roy, QSAR Comb. Sci. 27, 302 (2008)