author version:
SAR and QSAR in Environmental Research Volume 26, Issue 7-9, 2015
DOI: 10.1080/1062936X.2015.1084647

Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters†

Anita Rácz[a,b] , Dávid Bajusz[c] and K. Héberger[a],*

[a] Plasma Chemistry Research Group, Hungarian Academy of Sciences, Budapest, Hungary;
[b] Department of Applied Chemistry, Corvinus University of Budapest, Budapest, Hungary;
[c] Medicinal Chemistry Research Group, Hungarian Academy of Sciences, Budapest, Hungary

Recent implementations of QSAR modeling software provide the user with numerous models and a wealth of information. In this work, we provide some guidance on how one should interpret the results of QSAR modeling, compare and assess the resulting models and select the best and most consistent ones. Two QSAR datasets are applied as case studies for the comparison of model performance parameters and model selection methods. We demonstrate the capabilities of sum of ranking differences (SRD) in model selection and ranking and identify the best performance indicators and models. While the exchange of the original training and (external) test sets does not affect the ranking of performance parameters, it provides improved models in certain cases (despite the lower number of molecules in the training set). Performance parameters for external validation are substantially separated from the other merits in SRD analyses, highlighting their value in data fusion.

Keywords: model selection, performance parameters, ranking, cross-validation, sum of ranking differences,

*Corresponding author. Email: heberger.karoly@ttk.mta.hu

†Presented at the 8th International Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources, CMTPI-2015, 21–25 June 2015, Chios, Greece.

1. **Introduction**

Model comparison and selection of the best one is an evergreen among scientific investigations. The process is contradictory: bias-variance trade-off, local minima, searching for robust models, the principle of parsimony, *etc*.; all ideas consider various models inherently. One model is better from one point of view, the other should be better from another point of view. Even if one fixes the aim (and algorithm) according to various criteria: R^2, Q^2, Mallows *Cp*, Akaike Information criterion, Bayesian information criterion, *etc.*, their application on the training, validation and test sets will necessarily provide different models for description of existing data and for prediction of future samples. The case is even more complicated with the fact that we deal with random effects: *i.e*. it is relatively easy to find conditions where one of the models is clearly superior compared to other models. Many authors select instinctively or deliberately such datasets, splits, *etc.* for which their own descriptor selection or model building algorithm performs better than the rival approaches.

Kalivas *et al*. suggested selecting harmonious models taking into account the bias–variance trade-off: it is difficult and not unambiguous to find the 'best' model. A biased model provides less variance and *vice versa*. However, harmonious models are not necessarily parsimonious [1]. The scope of the methodology has recently been extended with the idea of sum of ranking differences (SRD) for partial least squares and ridge regression models [2].

Principal-component analysis (PCA) has been applied by Geladi [3,4] and Todeschini *et al*. [5] to find the best and worst regression and classification models, respectively. PCAs were completed on a matrix of regression vectors and dominant patterns (grouping, outliers) could be detected among the models. The interpretation of PCA results is easy: principal component 1 marks the direction of the best and worst regression models. Principal component 2 reflects various behaviors of the regression models on various datasets. The models lying in the middle of the plot (scores near 0) show a similar behavior for all datasets, while models far away from the center have a dissimilar behavior for different datasets.

While the generalization of the pairwise correlation method (GPCM) [6,7] provides the best models for recognition (for description of the existing data), its performance for predictive purposes might be weaker. It is presumed to be the reason why GPCM could not attain general usage.

A scientific investigation should be reproducible in any laboratory: hence a kind of standardization (algorithms, performance parameters, *etc.*) would be expected. Even in this sense no model selection approach was validated properly: different degrees of freedom, different numbers of variables, and different algorithms should and do provide different models as the best ones found and no hints are given as to which one should be accepted and why.

Therefore our aim was to rank and group the various modeling approaches and performance parameters. The results were compared to the model selection algorithm based on multi-criteria optimization as incorporated in the QSARINS approach of Gramatica and coworkers [8].

2. Methods

2.1 Dataset preparation

Two published QSAR datasets were used for our study: a toxicology study of benzene derivatives by Bertinetto and coworkers (from here on *Case study 1*) [9] and an SAR study of N-substituted maleimides by Matuszak and coworkers (from here on *Case study 2*) [10], for which docking and QSAR modeling has been carried out by Wu and

coworkers [11]. For *Case study 1*, toxicity values were expressed as acute toxicities (negative base 10 logarithm of 96-h LC_{50}, or pLC_{50}) for fathead minnow (*Pimephales promelas*), while for *Case study 2*, negative 10-base logarithms of the half-maximal inhibitory concentrations (pIC_{50}) were reported for two enzimes, hMGL and fatty acid amide hydrolase (FAAH). For *Case study 2*, QSAR modeling was carried out just for the activity data on human monoglyceride lipase, hMGL. For a better comparison, the training and (external) test sets reported in [9] and [11] were used without modification: for *Case study 1*, 51 and 18 molecules constituted the training and external test sets (compounds **1-51** and **52-69** in [9]), while for *Case study 2*, 48 and 14 molecules, constituted the training and external test sets [11]. The selection of the two case studies was purposeful: while reliable models exist for prediction of the toxicities of benzene derivatives (*Case study 1*), the prediction of inhibitory concentrations in *Case study 2* is not straightforward, at least not with this training-test set split.

Molecular structures and activity data were manually entered using ChemAxon's Instant Jchem [12], then two sets of molecular descriptors were generated for each dataset: the complete descriptor set (51 descriptors) of QikProp [13] using Schrödinger's Maestro [14], and the complete descriptor set (117 descriptors) of RDKit [15], using KNIME [16], resulting in a total of 168 descriptors. (The two descriptor sets were used simultaneously during QSAR modeling.) Detailed descriptions of the descriptors are available in Table 1.1 of the QikProp user manual [17] and in the RDKit documentation [18], respectively.

2.2 QSAR modeling

For QSAR model building, Gramatica and coworkers' QSARINS 2.2 software was used [8,19]. QSARINS implements a rich toolbox of statistical methods for the generation, validation and ranking of QSAR models. Models are calculated by MLR (Multiple Linear Regression) with Ordinary Least Squares (OLS) and a Genetic Algorithm (GA) [20] procedure is used to explore a large number of descriptor combinations for QSAR modeling. (Enumeration of all possible combinations becomes unfeasible in the case of a large number of descriptors.) The GA used Q^2_{LOO} as the fitness function. As output, QSARINS provides a rich selection of QSAR models, as well as model performance parameters (see Table 1 for the full list of performance parameters calculated by QSARINS). The software also provides a way for model ranking: multi-criteria decision making (MCDM), based on the work of Keller and coworkers [21], is applied to evaluate the models with regards to their performance in fitting and internal and external validation.

Table

Table 1. Description of the performance parameters in QSARINS

Performance parameter	Calculated during[a]	Formula[b]	Description
R^2, R^2_{ext}	training, external validation	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$	Explained variance; coefficient of determination, square of the multiple correlation coefficient
$R^2_{adj.}$	training	$R^2_{adj.} = R^2 - (1 - R^2) \times \frac{p}{n - p - 1}$	R^2 corrected with the degree of freedom
$R^2 - R^2_{adj.}$	training	see above	Difference of the two
LOF	training	$LOF = \frac{RSS}{n\left(1 - \frac{M + d(M-1)/2}{n}\right)^2}$	Friedman lack of fit criteria [40]. M: total number of linearly independent bases in the model, d: degrees-of-freedom cost for each nonlinear basis function
K_x	training	Based on PCA, see [41] for details	Inter-correlation among descriptors
ΔK	training	Based on PCA, see [41] for details	Difference of correlation among descriptors (K_x) and the

			descriptors plus responses (K_{xy})
RMSE	training, int. val., ext. val.	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$	Root mean square error
MAE	training, int. val., ext. val.	$MAE = \frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{n}$	Mean absolute error
RSS	training	$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	Residual sum of squares
CCC	training, int. val., ext. val.	$CCC = \frac{2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \hat{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \hat{y})^2 + n(\bar{y} - \hat{y})^2}$	Coefficient of concordance, concordance correlation coefficient [42,43]
s	training	$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$	Standard error of the estimate
F	training	$F = \left(\frac{\sum_{i=1}^{N}(\bar{y} - \hat{y}_i)^2}{p-1}\right) / \left(\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{n-p}\right)$	Fisher value

Q^2_{LOO}	internal validation	$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i/i}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{PRESS}{TSS}$	Leave-one-out cross-validated square of the (multiple) correlation coefficient
$R^2 - Q^2_{LOO}$	internal validation	see above	Difference of the two
$PRESS$	internal, external validation	$PRESS = \sum_{i=1}^{n}(y_i - \hat{y}_{i/i})^2$	Predicted residual sum of squares (either cross-validated or calculated on the external set)
Q^2_{LMO}	internal validation	$Q^2_{LMO} = 1 - \frac{\sum_{j=1}^{m}\sum_{i=1}^{n}\left(y_i - \hat{y}_{i/j}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	Leave-many-out cross-validated square of the (multiple) correlation coefficient
$R^2_{Y\text{-}SCRAMBLE}$	internal validation	see above	R^2 of the training set with Y-scrambling [44]
$RMSE_{Avg,\ Y\text{-}SCRAMBLE}$	internal validation	see above	Average RMSE with Y-scrambling [44]
$Q^2_{Y\text{-}SCRAMBLE}$	internal validation	see above	Q^2_{LOO} of the training set with Y-scrambling [44]

$R^2_{RND\text{-}DESCR}$	internal validation	see above	R^2 of the training set with randomized descriptors [44]
$Q^2_{RND\text{-}DESCR}$	internal validation	see above	Q^2_{LOO} of the training set with randomized descriptors [44]
$R^2_{RND\text{-}RESP}$	internal validation	see above	R^2 of the training set with randomized responses [44]
$Q^2_{RND\text{-}RESP}$	internal validation	see above	Q^2_{LOO} of the training set with randomized responses [44]
Q^2_{F1}	external validation	$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{EXT}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}}(y_i - \bar{y}_{TR})^2}$	Definition 1 in [45] for Q^2 of the external test set [46], TR: training set, EXT: external test set
Q^2_{F2}	external validation	$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{EXT}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}}(y_i - \bar{y}_{EXT})^2}$	Definition 2 in [45] for Q^2 of the external test set [47], EXT: external test set
Q^2_{F3}	external validation	$Q^2_{F3} = 1 - \frac{\sum_{i=1}^{n_{EXT}}(y_i - \hat{y}_i)^2 / n_{EXT}}{\sum_{i=1}^{n_{TR}}(y_i - \bar{y}_{TR})^2 / n_{TR}}$	Definition 3 in [45] for Q^2 of the external test set [48], TR: training set, EXT: external test set

$\overline{r^2_m}$	external validation	$\overline{r^2_m} = \frac{r^2_m + r'^2_m}{2}$	Here, $r^2_m = R^2 \times (1 - \sqrt{R^2 - R_0^2})$, where R_0^2 is the squared correlation coefficient without intercept. r'^2_m is the same as r^2_m, with the *x* and *y* axes exchanged. [49,50]
Δr^2_m	external validation	$\Delta r^2_m = r^2_m - r'^2_m$	See above.

[a] Parameters that are calculated for more than one subsets are indexed in the main text: *tr* for training, *cv* for cross-validation, *ext* for external validation.

[b] The following notations are used: y_i: single experimental value, $\bar{y}$: mean of experimental values, $\hat{y}_i$: single predicted value, $\hat{y}$: mean of predicted values, $\hat{y}_{i/i}$: predicted value for the *i*th sample when the *i*th sample is left out from the training, $\hat{y}_{i/j}$: predicted value for the *i*th sample when the *j*th part of the dataset is left out from the training (the whole dataset is split into *m* parts), *n*: number of samples, *i*: sample index, *p*: number of variables in the model

2.3 Multi-criteria decision making

Multi-criteria decision making (MCDM) [21] is a multi-parameter optimization technique that utilizes desirability functions [22] to evaluate the performances of several criteria simultaneously, usually as a single number (score) between 0 (worst) and 1 (best). The overall score is the geometric average of the values obtained from the desirability functions of the various criteria. In QSARINS, MCDM values are calculated to assess the QSAR models' performance with regards to fitting (*i.e.* how well does the model reproduce the data from which it was calculated), cross-validation (*i.e.* how well can the model predict smaller segments of the training set) and external validation (*i.e.* how reliable a prediction can the model make for external data, such as new molecules). MCDM of fitting is calculated by maximizing R^2, $R^2_{adj.}$ and CCC_{TR}, while minimizing $R^2 - R^2_{adj,}$, while MCDM of cross validation by maximizing Q^2_{LOO}, Q^2_{LMO} and CCC_{cv} and minimizing $R^2_{Y\text{-}SCRAMBLE}$, and MCDM of external validation by maximizing Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and CCC_{EXT}. From these three, $MCDM_{all}$ is calculated as a consensus. For visualization, $MCDM_{fit}$ can be plotted against $MCDM_{ext}$.

A useful and proven approach for multi-criteria optimization is the use of desirability functions (also applied during the calculation of MCDM values) as defined by Harrington and later by Derringer and Suich [22,23]. However, they inherently involve some subjectivity. In her PhD thesis Manuela Pavan compares total ranking methods and states: 'All the methods are based on a first level of subjectivity, concerning the criteria selected as representative of the system under investigation. Another level of subjectivity is added when the criteria are weighted, as this requires the identification of the more important criteria and the results are strictly influenced by the weight setting' [24]. By contrast, SRD, see below does not introduce such opportunities for subjectivity. Among other investigations, we have assessed the similarities and differences between the results of the two methods.

2.4 Sum of ranking differences (SRD)

SRD is a novel, simple and entirely general procedure for the quick and reliable comparison of models/methods/techniques *etc.* [25–27]. For an input matrix with n columns (methods or models, in this work typically QSAR models or performance parameters) and m rows (samples, in this work typically molecules or QSAR models), SRD is calculated according to the following steps: i) add a reference column to the input matrix (this can be a 'golden standard': a set of known reference values, such as experimental data; or a consensus of the n methods/models can be calculated with a suitable data fusion rule: typically average, minimum or maximum depending on the application), ii) rank the m samples in order of magnitude according to each of the n methods and the reference method; iii) calculate the absolute difference of ranks for each sample between each method and the reference; iv) sum up the ranking differences for each method to calculate the SRD values. (An animation to illustrate how SRD is calculated is published as a supplement to our recent article on similarity metrics [28].)

SRD values are identical to the Manhattan distances between the given method vector (in the space spanned by the samples) and the reference vector: a smaller SRD value means proximity to the reference: the smaller the better. To enable the comparison of different SRD calculations, SRD values are usually normalized:

$$SRD_{nor} = 100SRD / SRD_{max}, \quad (1)$$

where SRD_{max} is the maximum attainable SRD value.

SRD is validated in two ways: Comparison of ranks with random numbers (CRRN) is a randomization test that results in a distribution of SRD values when using randomized ranks for the same SRD calculation (see the Gaussian curves on Figures 2 and 4-10). A model/method/technique is more reliable than random ranking as long as it does not overlap with this Gaussian curve. It also tests, whether the SRD values of the different methods are distinguishable from each other (significantly different): to that end, a bootstrap-like cross-validation is carried out (leave-one-out cross-validation for 13 or fewer samples and sevenfold cross-validation for 14 or more samples).

3. Results and discussion

In this work two QSAR datasets – a toxicology study of benzene derivatives (*Case study 1*) [9] and an SAR dataset of N-substituted maleimides (*Case study 2*) [10,11] – were used to build multilinear regression models (with the application of pLC_{50} and pIC_{50} values as the dependent variable, *y*) and the created models and performance parameters were ranked and grouped with the SRD method.

The aim of the research was to answer the following questions about the models: 1) Can we complement the MCDM method with SRD, which gives consistent results in an easy way in the case of model comparison and selection? 2) Is there any difference in the selection of the best model(s) if we use the predicted *y* values instead of the performance parameters of the created models? 3) Which model performance parameter is the most predictive? 4) How does the consideration of an alternative training-test set split affect the outcome of the previous question?

QSARINS 2.2 of Gramatica *et al*. [7] was applied for MLR-based QSAR modeling. Variable selection was based on filtering out the constant variables (based on the standard deviation) and those that correlated with another variable with a correlation coefficient of 1.000000. Thus, the final numbers of variables for MLR analysis were 69 in *Case study 1* and 62 in *Case study 2*, respectively. Furthermore in the modeling section GA was used as another variable selection method for the creation of better regression models. In *Case study 1* the maximum number of descriptors (for the GA) was six and in *Case study 2* it was seven. The final dataset contained the best ten models for every possible descriptor number (1 to 6 and 1 to 7), thus the number of models was 60 in *Case study 1* and 70 in *Case study 2*. In the following part the calculated performance parameters and the predicted *y* values for each sample will be used for the analysis to answer the four main questions that were posed.

3.1 Comparison of MCDM and SRD methods

How does SRD compare with MCDM in model selection? In the first part of our work MCDM (which is included in QSARINS 2.2) and SRD were used for the selection of the best models. We wanted to compare the usefulness of these methods in model selection. In *Case study 1* twelve performance parameters were used for MCDM analysis, and the best models can be seen in **Figure 1**, where the MCDM values of external prediction are plotted against the MCDM fitting values. Figure 1 shows that there are six models, which have good score values (close to one) for both MCDM parameters.

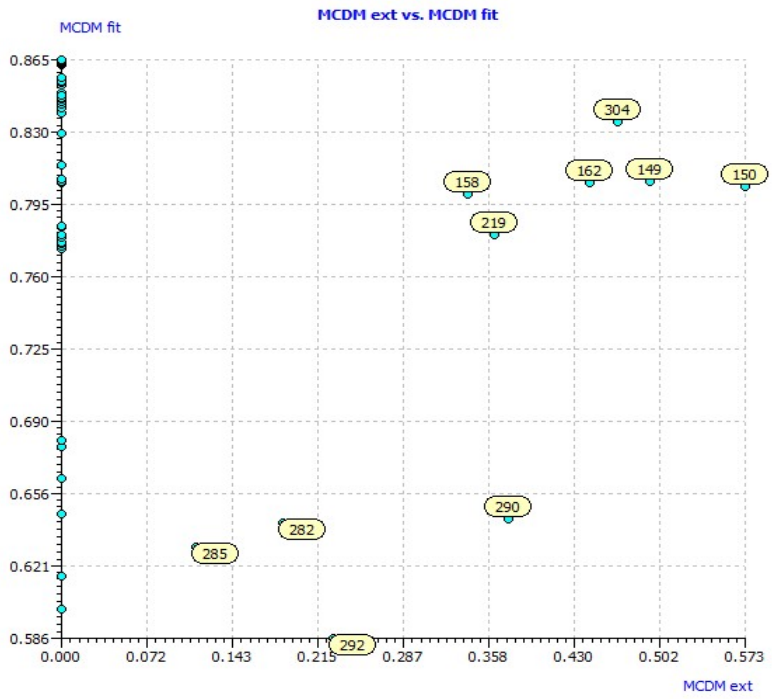


Figure 1 Multi-criteria optimization for model selection in Case study 1. MCDM values of external prediction are plotted against the MCDM values of fitting.

For SRD analysis 35 performance parameters were used as ‘samples’ and the sixty models as ‘variables’ in the input matrix. The reference value was the maximum or minimum value depending on which is the best for each of the performance parameters (to remain comparable with MCDM). **Figure 2** shows that in *Case study 1*, the SRD method gave very similar results to the MCDM method, thus we can conclude that SRD is another good choice for model selection. As we can see, there are a lot of similar models among the examined ones, but all of them are better than the randomly generated numbers (they are located before the XX1 line, which corresponds to the 5 % error limit).

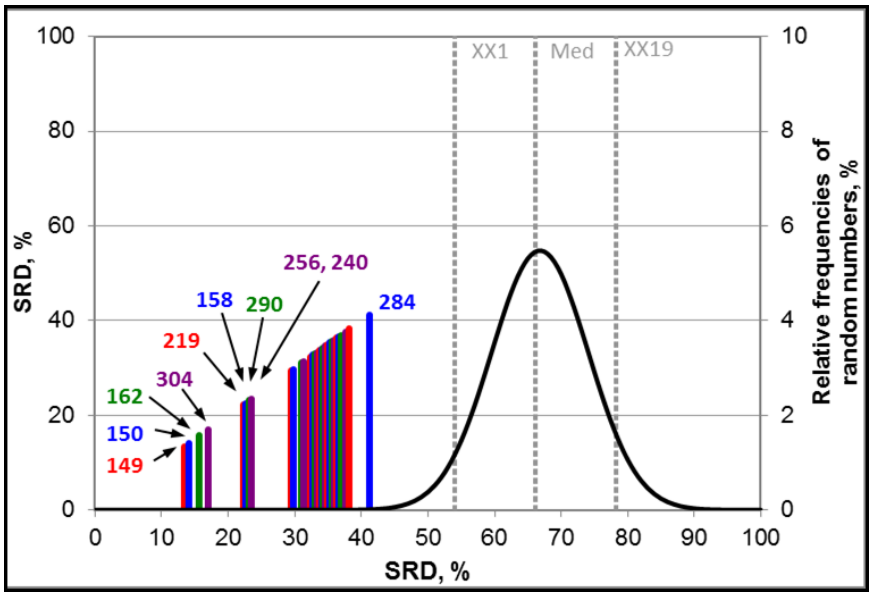


Figure 2 SRD gave the same best model selection as MCDM in Case study 1. Scaled SRD values (between 0 and 100) are plotted on the x axis and left y axis. The right y axis shows the relative frequencies for the black (fitted) Gauss curve (XX1 = 5 % limit, med = median, XX19 = 95 % limit).

In *Case study 2* the same numbers of performance parameters were used for MCDM and SRD analysis as in the previous case. Here the models were still acceptable, but not as good as before. **Figure 3** shows that in the MCDM analysis most of the models have a good MCDM value for the model fitting but worse for the external prediction, or *vice versa*. The 'best' models are located in the lower part of the plot and their MCDM fitting values are not higher than 0.436.

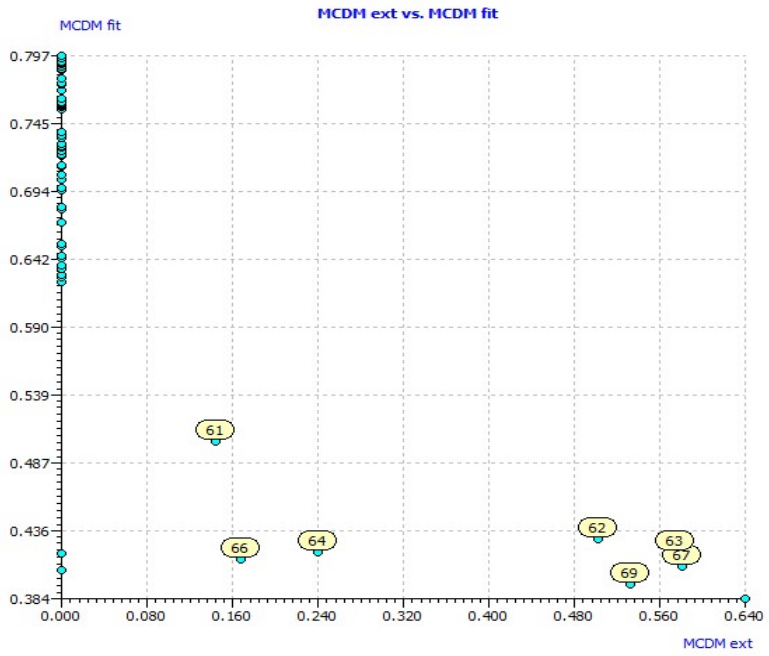


Figure 3: Multi-criteria optimization for model selection in Case study 2. MCDM values of external prediction are plotted against the MCDM values of fitting.

Although the final results are not as attractive as in *Case study 1*, **Figure 4** shows that the SRD method found the same models except for model 70, which is also an acceptable one. All of the models are shifted a little in the direction of bigger SRD values (in comparison with *Case study 1*), but none of them overlaps with the Gaussian curve of random numbers.

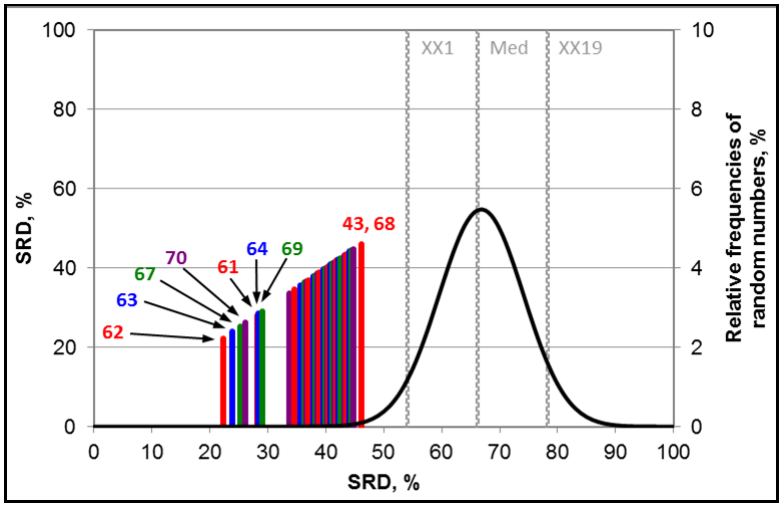


Figure 4: SRD gave similar model selection as MCDM in *Case study 2* – except for model 70. Scaled SRD values (between 0 and 100) are plotted on the *x* axis and left *y* axis. The right *y* axis shows the relative frequencies for the black (fitted) Gauss curve (XX1 = 5 % limit, med = median, XX19 = 95 % limit).

In the first part of the research, the results confirmed that SRD can be a good choice for QSAR model comparison and selection, because it is an easy and fast technique; moreover, the goodness of the results was proven by a comparison with the MCDM method (see Figures 3 and 4). In addition, SRD found one 'new' good model in *Case study 2*, which was not in the group of the best models in the MCDM analysis.

3.2 Usage of predicted y values

Is there any difference in the selection of the best model(s) if we use the predicted y values instead of the performance parameters of the created models?

In the second part of our research, the models were compared by the predicted pLC_{50} and pIC_{50} values of the compounds. SRD was used for the analysis here, where the compounds were included in the rows of the input matrix and the predicted values of each model were placed in the columns (as variables). The comparison was carried out in two ways: first the average was used as the reference (or 'golden standard'), and second the experimental pLC_{50} (or pIC_{50}) values were used as the reference vector.

Average can be a good choice, because it shows us, which models are better or worse than the experimental values. Though the models have systematic and random errors, they eliminate a large portion of experimental error, *i.e.* the error of the modeled values can be less, than the experimental ones. Using average as the reference can be thought of as a consensus in accordance with the 'maximum likelihood principle, which yields a choice of the estimator as the value for the parameter that makes the observed data most probable' (the average) [29].

In the second case, when the experimental values were used as the reference, we wanted to know which model gives the most 'similar' results to the experimental values, and if there is any difference between the usage of the predicted *y* values and the

usage of performance parameters. It is well known that the selection of the reference vector greatly influences the ranking results [30].

In *Case study 1* the dataset contained 51 samples, 60 predicted value columns (for each of the created QSAR models) and the experimental values as the 61st one if average was used as reference. In this latter case the results can be seen in **Figure 5(a)** and (**b)**, where the experimental 'model' is far away from the most consistent ones (closest to the average), which means that the average prediction of the models is quite far from the experimental values. (Note that in the context of this article, and particularly for SRD results, *consistency* is defined as the closeness/similarity of the given model's ranking to the reference ranking.)

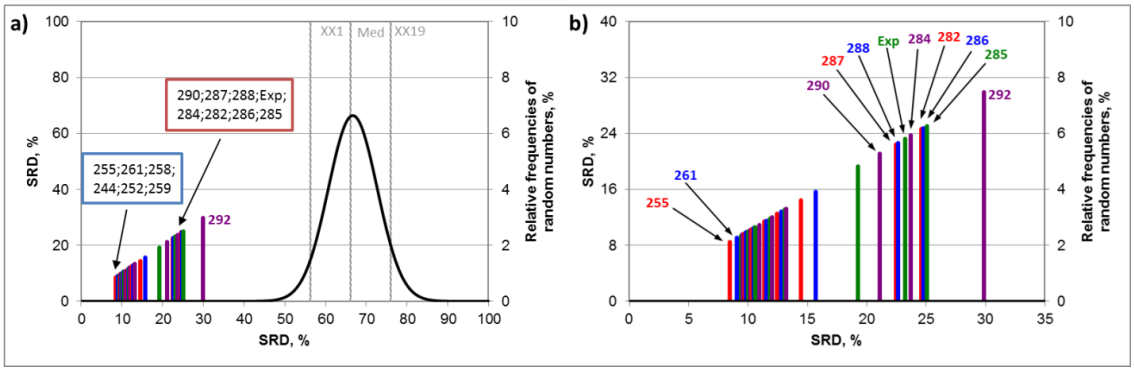


Figure 5(a) and (b): SRD ranking of the model selection with the use of predicted biological activity values. Average was used as the reference. Figure 5b is the magnified version of 5a. Figure 5 can be interpreted in the same way as Figure 4.

As we can see models 255 and 261 were the most consistent ones, their SRD values were under 10 %. An interesting observation is that some of the models, which were mediocre based on the MCDM analysis, were located closer to the Experimental variable, at SRD values of 20-25 %.

If we use the experimental values as reference, the results are somewhat different. **Figures 6(a)** and (**b)** show that here the best models were 346, 347 and 348. Although the SRD values of these models are higher than 20 %, they have the closest proximity to the experimental *y* values. In this case the mediocre models (based on the MCDM analysis) were located at the end of the line.

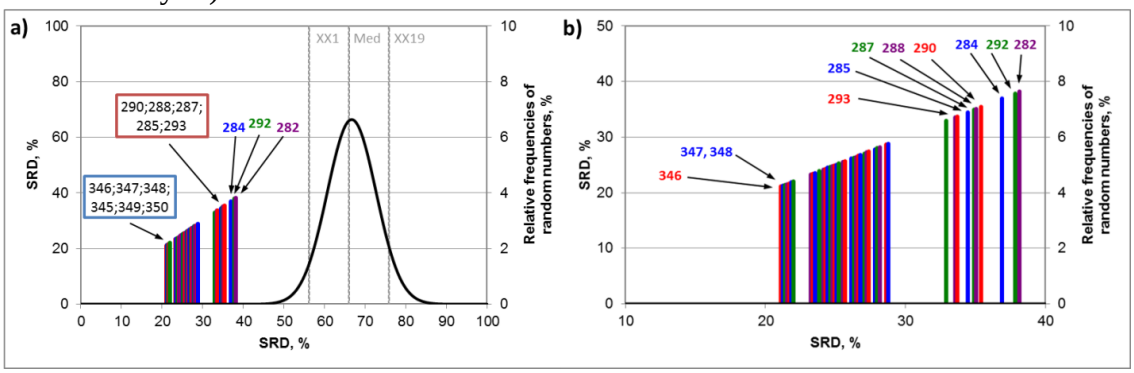


Figure 6(a) and (b): Comparison of the models with the experimental values as reference in *Case study 1*. Figure 6(b) is the magnified version of 6(a). Figure 6 can be interpreted in the same way as Figure 4.

In *Case study 2* we carried out the same analysis, but here the dataset contained 48 rows (samples) and 70 columns (models). When average was the reference, the experimental *y* variable was added as the 71st column. In the latter case the results can

be seen in **Figure 7**, where the experimental 'model' was also in the end of the line. All of the models were better than random rankings, but here models 18, 30 and 8 were the most consistent ones.

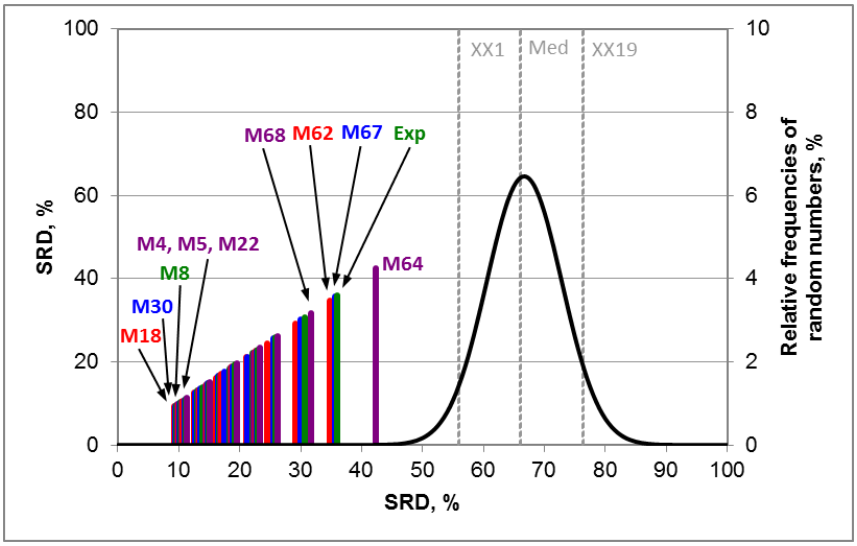


Figure 7: SRD ranking in *Case study 2*, where average was the reference. Figure 7 can be interpreted in the same way as Figure 4.

If we used experimental values as reference, the results were also very interesting, because the model selection differed from the MCDM analysis. **Figures 8a** and **b** show that in this case models 69 and 70 were the best, so they could best approximate the experimental values. Model 69 was also a good one based on MCDM analysis and SRD analysis of the performance parameters, but the model 70 is a new one, which was already identified once by SRD, using performance parameters. Most of the models that have been selected by MCDM analysis were not verified by SRD, except for model 69 in *Case study 2*. We have to admit that the model parameters and all the original values of *Case study 2* are far from being optimal for a straightforward prediction.

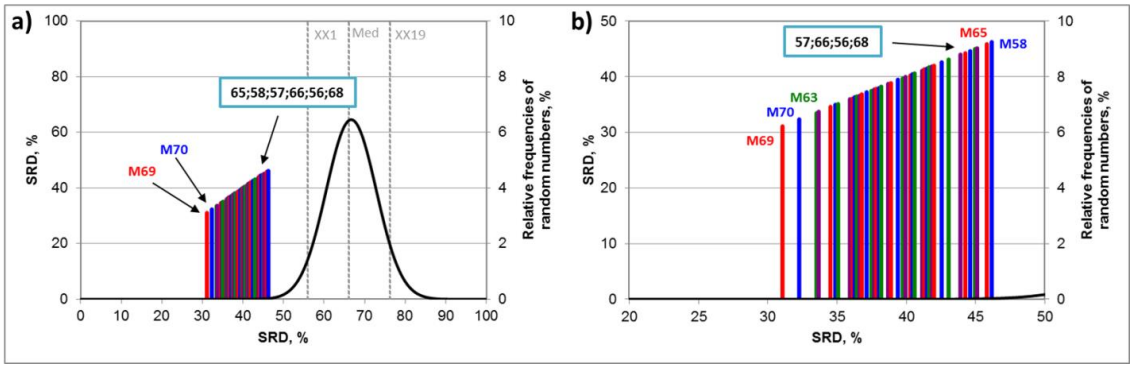


Figure 8(a) and (b): Comparison of the models with the experimental values as reference in *Case study 2*. Figure 8(b) is the magnified version of 8(a). Figure 8 can be interpreted in the same way as Figure 4.

Thus, in the second part, we can conclude that the use of predicted values for SRD opens a new way to model selection, because it can reveal good models other than those identified by MCDM. This alternative is also a valuable approach, as it better accounts for the predictive capability of the regression models, while the calculation of performance parameters unavoidably leads to some information loss. Primarily fitted (modeled) and experimental values have the full information content of the data.

3.3 Choice of merit of performance

Which is the most predictive model performance parameter? In this section, our goal was to choose the most appropriate performance parameter(s) for our datasets. We have selected the following, more commonly used parameters for the comparison: R^2, R^2_{ext}, R^2_{adj}, $\overline{r^2_m}$, CCC_{ext}, CCC_{cv}, CCC_{tr}, MAE_{ext}, MAE_{cv}, MAE_{tr}, $RMSE_{ext}$, $RMSE_{cv}$, $RMSE_{tr}$, Q^2_{LOO}, s, F, Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and Q^2_{LMO}. The dataset in *Case study 1* contained these 20 parameters in the columns and the 60 models in the rows. Row-average was used as reference. Sevenfold cross-validation was used for the verification of the analysis. **Figure 9** (a box and whisker plot) shows that there are a few performance parameters (Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and $RMSE_{ext}$) that overlap with random ranking, but most of the parameters are located between zero and the 5 % limit for random ranking. The first twelve parameters are indistinguishable according to the Wilcoxon matched pair test. To conclude whether the position of CCC_{ext} is a consequence of the peculiar character of the coefficient of concordance, or merely a random effect, we need to make further investigations. Above the horizontal dotted line the performance parameters are indistinguishable from random ranking (at 5 % error level).

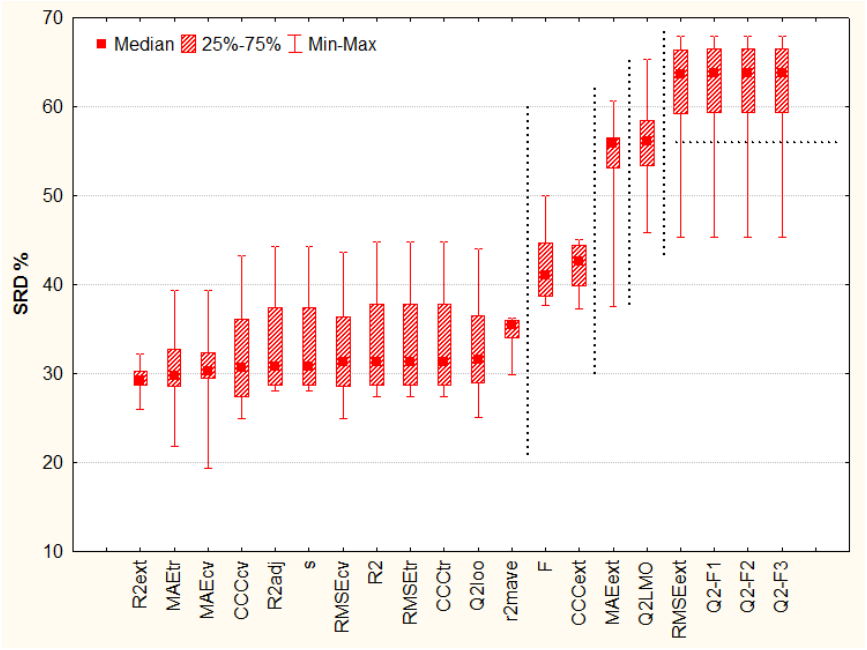


Figure 9: Comparison of performance parameters using sevenfold cross-validation of scaled SRD values. On the box and whisker plot horizontal dotted line shows the 5 % error limit for random ranking. Vertical dotted line shows the 5 % error limit for Wilcoxon matched pair test.

The same examination was carried out for *Case study 2*, where the number of the columns was 20 with the same performance parameters, but the number of the rows was 70 (since here the number of created models is 70).

According to **Figure 10**, quite the same performance parameters have the lowest SRD values for *Case study 2* as well, in a somewhat (not significantly) different order. Here, the most consistent one was $RMSE_{cv}$ and the following ones in order were Q^2_{LOO}, MAE_{tr}, CCC_{cv}, and MAE_{cv}, (but their ordering lacks significance by Wilcoxon matched pair test and at the 5 % level). In this case external validation metrics were the farthest merits from the average, overlapping with the distribution of random rankings. These

differ significantly from the others at the 5 % level (marked with a dotted line). This does not mean that these measures are not useful; on the contrary, as they provide an ordering that is dissimilar from the reference, they present valuable information that can be utilized for *e.g.* for data fusion. Other investigations also support the view that external validation provides comparable results to a single split in many cases.

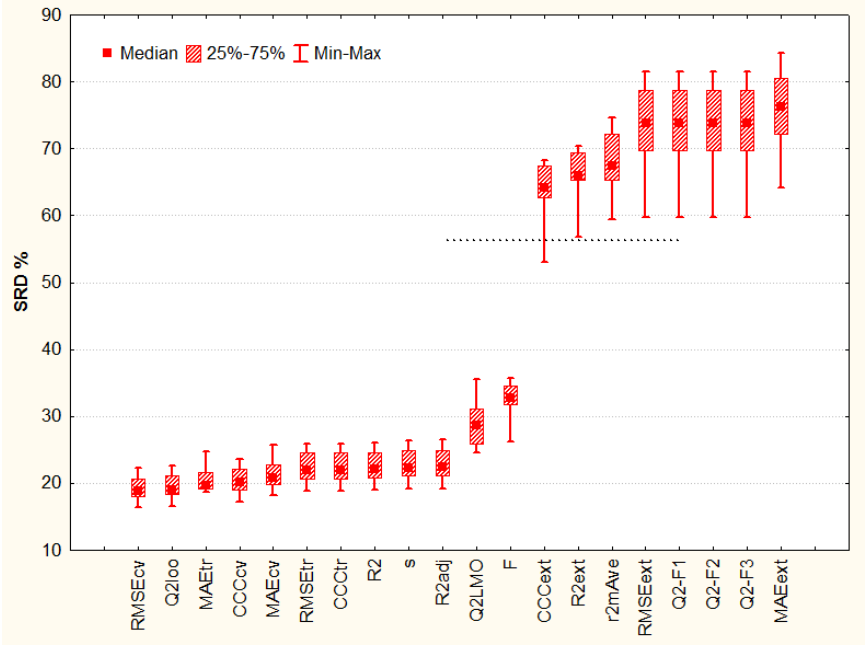


Figure 10: Comparison of performance parameters using sevenfold cross-validation of scaled SRD values. On the box and whisker plot horizontal dotted line shows the 5 % error limit for random ranking.

In Case study 2 the SRD values were considerably smaller than in Case study 1, but in both cases $RMSE_{cv}$, CCC_{cv} and Q^2_{LOO} were among the first group of ranked parameters. The coincidence is striking as two different datasets; two qualitatively different model performances were compared. In Case study 1 the first twelve parameters are indistinguishable by the Wilcoxon matched pair test, while in Case study 2, the first ten parameters are indistinguishable. Two distinct groups of performance parameters can also be seen in **Figure 10**, as well.

A comparison of several performance parameters has already been done applying extensive simulations in refs. [51] and [52]. In ref. [51] 'CCC [coefficient of concordance] was broadly in agreement … with other validation measures in accepting models as predictive, and … it was the most precautionary.' Therefore it was proposed as an external validation parameter for use in QSAR studies. Our findings are in agreement with the cited papers, as we have also identified CCC_{cv} as one of the most consistent performance parameters.

3.4 Use of alternative training-test set splits

Both modeling studies have used the training-test set split from the original publications. To assess the effect of the dataset splitting on the outcome, we repeated the calculations using the external sets for training and the training sets for external validation. Thus, the validation became 'crossed': each element of the left-out part of the data was used in the modeling (training) phase. It is interesting to know, whether such splits are representative for the whole distribution or carry different information.

Many authors favor cross-validation as it 'gives a reliable picture with no apparent systematic over- or underestimation' [31]; 'overfitting is avoided by the

repeated double cross-validation approach' [32]; 'LOO gives too small a perturbation to the data, so that Q^2 approaches the properties of R^2 asymptotically' [33]; 'The cross-validation estimate of prediction error is nearly unbiased but can be highly variable.' [34]. Others rigidly adhere to external validation as 'External figures of merit are the gold standard…' [35]; an external validation set is necessary for obtaining honest validation results after parameter estimation [36]; '… only models that have been validated externally, after their internal validation, can be considered reliable and applicable for both external prediction and regulatory purposes' [37] and 'there is only one valid paradigm, formulated as the test set validation imperative' [38].

SRD is a particularly suitable technique to decide whether cross-validation is appropriate to evaluate the models' predictive performance. SRD is based on the assumption that errors cancel each other and a consensus reference is suitable for ranking and grouping the performance parameters.

Case study 1: Previously, the 12 candidates to substitute all performance parameters were located ahead and now a new modeling of the 20 compounds, variable selection (best subset and GA), and range scaling of 30 models produced 13 performance parameters with the smallest SRD values. (Only a small number of changes was observed.)

Similarly, the least consistent merits were $RMSE_{ext}$, Q^2_{F1}, Q^2_{F2}, and Q^2_{F3} earlier (each of them are comparable with the random ranking) and now all merits based on external validation have proven to be the least consistent ones. The agreement is fairly and surprisingly good. (We note that exactly the same ranking cannot be expected as for model building, and consequently, the calculation of performance parameters are subjects to biases and random errors. Moreover, the model building set was considerably smaller, and different models were built and considered.) Unexpectedly, even better models were built with these 18 compounds, suggesting that these compounds are more characteristic of the substituted benzene derivatives than the larger training set (51).

Case study No 2: A very similar pattern can be observed for the performance parameters with the smaller SRD values. Similarly, seven of the most dissimilar merits were kept for the modeling with the external set: $RMSE_{ext}$, Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}, MAE_{ext}, CCC_{ext} and R^2_{ext}. The agreement is astonishingly good, given that the training set was considerably smaller, and different models were built. (The box and whisker plots of the alternative training-test splits can be seen in the supplementary material as **Figure S1** for *Case study 1* and **Figure S2** for *Case study 2*.)

Some general conclusions can be drawn: leave-one-out and cross-validated merits are among the most representative group. Coefficient of concordance can be favorably used instead of (or beside) the (multiple) correlation coefficient. The most dissimilar results were obtained with the external validation indicators. This section suggests that the property distributions of the training and test sets were similar enough so that mostly the same results were acquired upon their exchange (*e.g.* for the comparison of performance parameters); and yet for *Case study 1*, even better QSAR models could be developed with the use of the (original) test sets. In accordance with the recent work of Roy and coworkers [39], these results highlight the importance and usefulness of considering more than one training-test set splits for QSAR modeling.

4. Conclusion

The procedure based on sum of ranking differences (SRD) agrees well with multi-criteria decision making as it provides very similar rankings for models if the performance merits are used for SRD-based comparison (particularly when considering

good models, as in *Case study 1*).

However, if the primary experimental and predicted data are used, the ranking and clustering of the models are different from the case when performance merits are used for ranking and data fusion. The use of performance parameters leads to a kind of information loss, thus we suggest selecting *consistent* models using primary data and SRD.

Coefficient of concordance can be favorably used instead of (or beside) the (multiple) correlation coefficient. Performance parameters based on external validation were the most dissimilar from the consensus: this can mean that they provide complementary information and their use can be beneficial for *e.g.* for data fusion.

SRD can also be applied to check the consistency of training and (external) test splits. The distributions of the training and test sets in both cases were similar enough, but we suggest using more than one training-test set split for QSAR modeling, as it can provide even better QSAR models in some cases.

Abbreviations

FAAH, fatty acid amide hydrolase; GA, genetic algorithm; MCDM, multi-criteria decision making; hMGL, human monoglyceride lipase; MLR, multiple linear regression; OLS, ordinary least squares; PCA, principal component analysis; SRD, sum of ranking differences.

Acknowledgement

The authors wish to acknowledge Paola Gramatica from the University of Insubria for providing access to QSARINS 2.2.

Disclosure statement

15 No potential conflict of interest was reported by the authors.

ORCID

A. Rácz http://orcid.org/0000-0001-8271-9841
D. Bajusz http://orcid.org/0000-0003-4277-9481
K. Héberger http://orcid.org/0000-0003-0965-939X

References

[1] F. Stout, M.R. Baines and J.H. Kalivas, *Impartial graphical comparison of multivariate calibration methods and the harmony/parsimony tradeoff*, J. Chemom. 20 (2006), pp. 464–475.

[2] J.H. Kalivas, K. Héberger and E. Andries, *Sum of ranking differences (SRD) to ensemble multivariate calibration model merits for tuning parameter selection and comparing calibration methods.*, Anal. Chim. Acta 869 (2015), pp. 21–33.

[3] P. Geladi, J. Swerts and F. Lindgren, *Multiwavelength microscopic image analysis of a piece of painted chinaware: Classification and regression*, Chemom. Intell. Lab. Syst. 24 (1994), pp. 145–167.

[4] P. Geladi, *The regression model comparison plot (REMOCOP)*, in *Frontiers in Analytical Spectroscopy*, D. Andrews and A. Davies, eds., The Royal Society of Chemistry, Cambridge, UK, 1995, pp. 225–236.

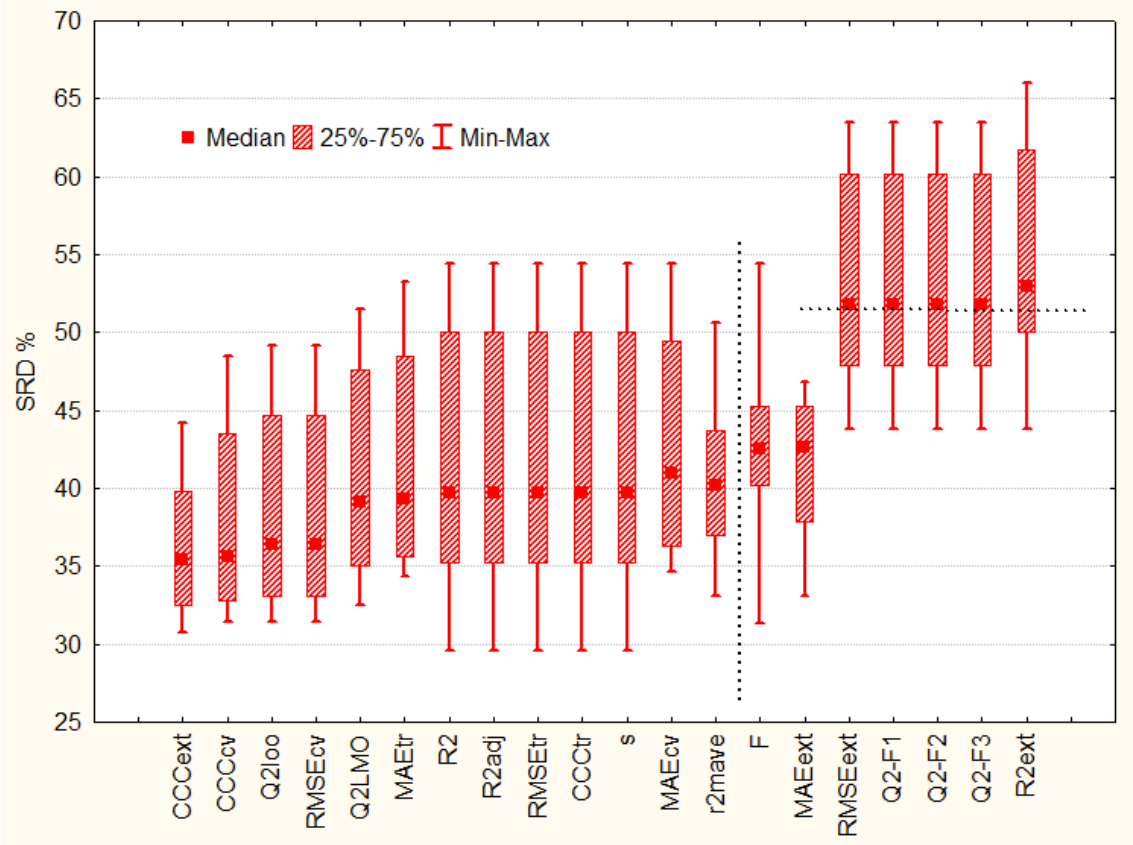
[5] R. Todeschini, D. Ballabio, V. Consonni, a. Mauri and M. Pavan, *CAIMAN (Classification And Influence Matrix Analysis): A new approach to the*

classification based on leverage-scaled functions, Chemom. Intell. Lab. Syst. 87 (2007), pp. 3–17.
[6] K. Héberger and R. Rajkó, *Generalization of pair correlation method (PCM) for non-parametric variable selection*, J. Chemom. 16 (2002), pp. 436–443.
[7] K. Héberger and R. Rajkó, *Variable selection using pair-correlation method. Environmental applications.*, SAR QSAR Environ. Res. 13 (2002), pp. 541–54.
[8] P. Gramatica, N. Chirico, E. Papa, S. Cassani and S. Kovarich, *QSARINS: A new software for the development, analysis, and validation of QSAR MLR models*, J. Comput. Chem. 34 (2013), pp. 2121–2132.
[9] C. Bertinetto, C. Duce, R. Solaro and K. Héberger, *Modeling of the Acute Toxicity of Benzene Derivatives by Complementary QSAR Methods*, MATCH-COMMUNICATIONS Math. Comput. Chem. 70 (2013), pp. 1005–1021.
[10] N. Matuszak, G.G. Muccioli, G. Labar and D.M. Lambert, *Synthesis and in vitro evaluation of N-substituted maleimide derivatives as selective monoglyceride lipase inhibitors.*, J. Med. Chem. 52 (2009), pp. 7410–20.
[11] J. Wu, Y. Wang and Y. Shen, *Molecular docking and QSAR analysis on maleimide derivatives selective inhibition against human monoglyceride lipase based on various modeling methods and conformations*, Chemom. Intell. Lab. Syst. 131 (2014), pp. 22–30.
[12] *Instant JChem*. ChemAxon LLC, Budapest, Hungary, 2014.
[13] *QikProp, version 4.2*. Schrödinger, LLC, New York, NY, USA, 2014.
[14] *Small-Molecule Drug Discovery Suite 2014-4*. Schrödinger, LLC, New York, NY, USA, 2014.
[15] *RDKit: Cheminformatics and Machine Learning Software*. 2014.
[16] *KNIME | Konstanz Information Miner*. University of Konstanz, Konstanz, Germany, 2015.
[17] *Schrödinger Documentation*; available at http://www.schrodinger.com/supportdocs/18/.
[18] *RDKit Descriptor List*; available at http://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors.
[19] P. Gramatica, S. Cassani and N. Chirico, *QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS.*, J. Comput. Chem. 35 (2014), pp. 1036–44.
[20] R.L. Haupt and S.E. Haupt, eds., *Practical Genetic Algorithms*, 2ed., Wiley, 2004.
[21] H.R. Keller, D.L. Massart and J.P. Brans, *Multicriteria decision making: A case study*, Chemom. Intell. Lab. Syst. 11 (1991), pp. 175–189.
[22] E.C. Harrington, *The desirability function*, Ind. Qual. Control 21 (1965), pp. 494–498.
[23] G. Derringer and R. Suich, *Simultaneous optimization of several response variables*, J. Qual. Technol. 12 (1980), pp. 214–219.
[24] M. Pavan, *Total and Partial Ranking Methods in Chemical Sciences*, University of Milano - Bicocca, 2003.
[25] K. Héberger, *Sum of ranking differences compares methods or models fairly*, TrAC Trends Anal. Chem. 29 (2010), pp. 101–109.
[26] K. Héberger and K. Kollár-Hunek, *Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers*, J. Chemom. 25 (2011), pp. 151–158.

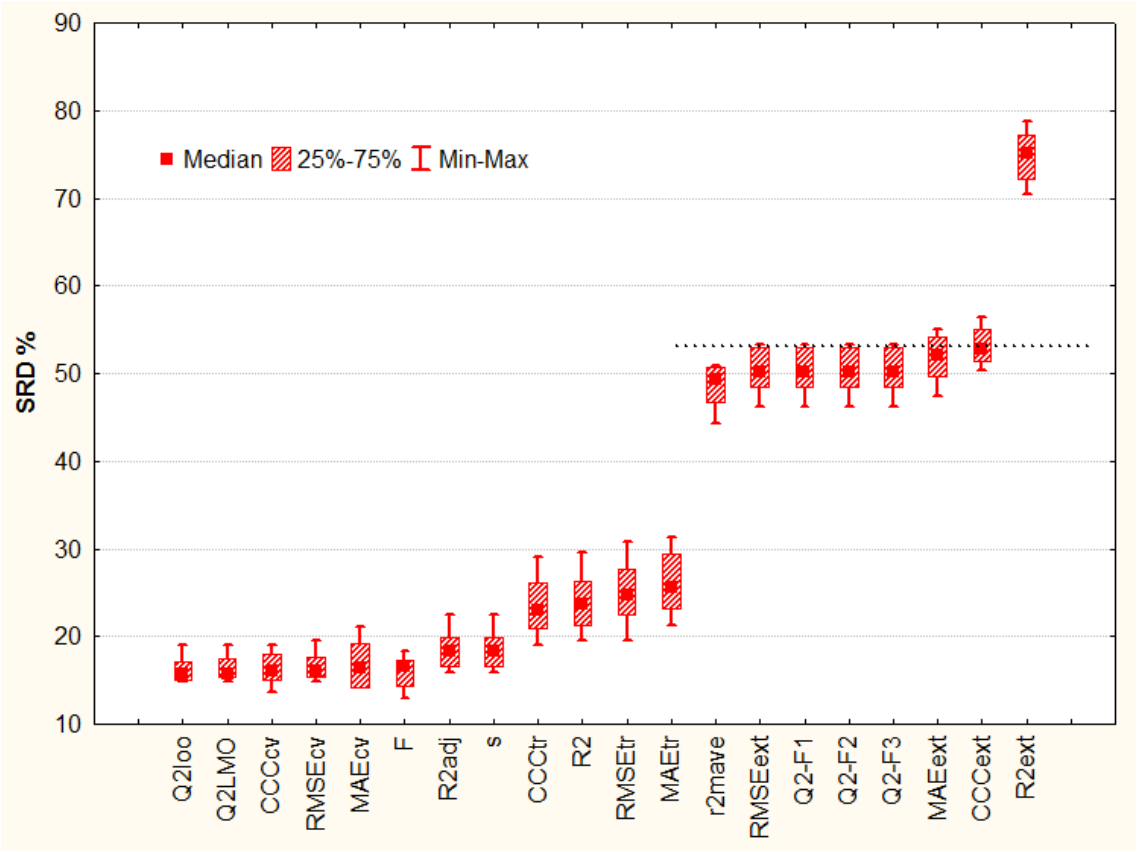
[27] K. Kollár-Hunek and K. Héberger, *Method and model comparison by sum of ranking differences in cases of repeated observations (ties)*, Chemom. Intell. Lab. Syst. 127 (2013), pp. 139–146.
[28] D. Bajusz, A. Rácz and K. Héberger, *Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?*, J. Cheminform. 7 (2015), pp. 20.
[29] T. Hastie, R. Tibshirani and J. Friedman, *Linear Discriminant Analysis*, in *Elements of Statistical Learning. Data Mining, Inference, Prediction*, Springer, New York, NY, USA, 2001, pp. 106–119.
[30] K. Héberger and B. Skrbić, *Ranking and similarity for quantitative structure-retention relationship models in predicting Lee retention indices of polycyclic aromatic hydrocarbons.*, Anal. Chim. Acta 716 (2012), pp. 92–100.
[31] D.M. Hawkins, S.C. Basak and D. Mills, *Assessing model fit by cross-validation.*, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 579–86.
[32] P. Filzmoser, B. Liebmann and K. Varmuza, *Repeated double cross validation*, J. Chemom. 23 (2009), pp. 160–171.
[33] J. Shao, *Linear Model Selection by Cross-validation*, J. Am. Stat. Assoc. 88 (1993), pp. 486–494.
[34] B. Efron and R. Tibshirani, *Improvements on Cross-Validation: The 632+ Bootstrap Method*, J. Am. Stat. Assoc. 92 (1997), pp. 548–560.
[35] K. Baumann, *Chance Correlation in Variable Subset Regression: Influence of the Objective Function, the Selection Mechanism, and Ensemble Averaging*, QSAR Comb. Sci. 24 (2005), pp. 1033–1046.
[36] A. Tropsha, P. Gramatica and V. Gombar, *The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models*, QSAR Comb. Sci. 22 (2003), pp. 69–77.
[37] P. Gramatica, *Principles of QSAR models validation: internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701.
[38] K.H. Esbensen and P. Geladi, *Principles of Proper Validation: use and abuse of re-sampling for validation*, J. Chemom. 24 (2010), pp. 168–187.
[39] K. Roy, I. Mitra, S. Kar, P.K. Ojha, R.N. Das and H. Kabir, *Comparative studies on some metrics for external validation of QSPR models.*, J. Chem. Inf. Model. 52 (2012), pp. 396–408.
[40] J.H. Friedman, *Multivariate Adaptive Regression Splines*, Ann. Stat. 19 (1991), pp. 1–67.
[41] R. Todeschini, V. Consonni and A. Maiocchi, *The K correlation index: theory development and its application in chemometrics*, Chemom. Intell. Lab. Syst. 46 (1999), pp. 13–29.
[42] L.I.-K. Lin, *A concordance correlation coefficient to evaluate reproducibility*, Biometrics 45 (1989), pp. 255–68.
[43] L.I.-K. Lin, *Assay Validation Using the Concordance Correlation Coefficient*, Biometrics 48 (1992), pp. 599.
[44] C. Rücker, G. Rücker and M. Meringer, *y-Randomization and its variants in QSPR/QSAR.*, J. Chem. Inf. Model. 47 (2007), pp. 2345–57.
[45] V. Consonni, D. Ballabio and R. Todeschini, *Evaluation of model predictive ability by external validation techniques*, J. Chemom. 24 (2010), pp. 194–201.
[46] G. Schüürmann, R.-U. Ebert, J. Chen, B. Wang and R. Kühne, *External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean.*, J. Chem. Inf. Model. 48 (2008), pp. 2140–5.

[47] L.M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R.M. Blair, W.S. Branham, S.L. Dial, C.L. Moland and D.M. Sheehan, *QSAR Models Using a Large Diverse Set of Estrogens*, J. Chem. Inf. Model. 41 (2001), pp. 186–195.
[48] V. Consonni, D. Ballabio and R. Todeschini, *Comments on the definition of the Q2 parameter for QSAR validation.*, J. Chem. Inf. Model. 49 (2009), pp. 1669–78.
[49] P.P. Roy and K. Roy, *On Some Aspects of Variable Selection for Partial Least Squares Regression Models*, QSAR Comb. Sci. 27 (2008), pp. 302–313.
[50] P.K. Ojha, I. Mitra, R.N. Das and K. Roy, *Further exploring rm2 metrics for validation of QSPR models*, Chemom. Intell. Lab. Syst. 107 (2011), pp. 194–205.
[51] N. Chirico and P. Gramatica, *Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient*, J. Chem. Inf. Model. 51 (2011), pp. 2320-2335.
[52] N. Chirico and P. Gramatica, *Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection*, J. Chem. Inf. Model. 52 (2012), pp. 2044-2058.

Supplementary figures



Supplementary Figure S1: Box and whisker plots of the alternative training-test splits for *Case study 1* (c.f. figure 9)



Supplementary Figure S2: Box and whisker plots of the alternative training-test splits for *Case study 2* (c.f. figure 10)