DOI https://doi.org/10.1007/978-3-319-56850-8_3
Publisher Name: Springer, Cham
Print ISBN 978-3-319-56849-2
Online ISBN 978-3-319-56850-8

Which performance parameters are best suited to assess the predictive ability of models?

Károly Héberger[a,]* Anita Rácz[a,b], and Dávid Bajusz[b]

[a] *Plasma Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, H-1117 Budapest, Magyar tudósok krt. 2, Hungary;*

[b] *Medicinal Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, H-1117 Budapest, Magyar tudósok krt. 2, Hungary*

* To whom correspondence should be sent: K. Héberger, e-mail: heberger.karoly@ttk.mta.hu

Abstract

We have revisited the vivid discussion in the QSAR-related literature concerning the use of external *vs*. cross-validation in this work, and have presented a thorough statistical comparison of model performance parameters with the recently published SRD (sum of (absolute) ranking differences) method and analysis of variance (ANOVA). Two case studies were investigated, one of which has exclusively used external performance merits. The SRD methodology coupled with ANOVA shows unambiguously for both case studies that the performance merits are significantly different, independently from data preprocessing. While external merits are generally less consistent (farther from the reference) than training and cross-validation based merits, a clear ordering and a grouping pattern of them could be acquired. The results presented here corroborate our earlier, recently published findings (SAR QSAR Environ. Res., 2015, 26, 683–700) that external validation is not necessarily a wise choice, and is frequently comparable to a random evaluation of the models.

Keywords: performance parameters (merits), ranking, cross-validation, external validation, QSAR modeling

Introduction

There is a long lasting and never ending discussion in the literature: How can we estimate the predictive ability of multivariate models (and in particular QSAR models)? Here we cannot recapitulate the entire story, just refer to some basic sources. Generally, there are principally two different ways to evaluate the “goodness” of a QSAR model: to assess the model’s performance with regards to i) description (fitting or recall), *i.e.* evaluating the performance on the existing data; and ii) prediction, *i.e.* evaluating the performance on future data, also called external validation, (how reliable a prediction can the model be made for external data, such as for new molecules).

External validation is usually modelled by a single split (hold-out sample) in the belief that future compounds (objects, samples) will be derived from the same property distribution, which is more or less true for QSAR models within the applicability domain. If the future compounds diverge from the property distribution of the earlier ones (on which the model was built); then the model cannot be applied anymore without updating.

A common choice is to estimate the predictive ability using cross-validation; however, it is debatable how well cross-validation can mimic the prediction performance. Cross-validation is probably the most widely used method for estimating prediction error, but its various implementations inherently call for a compromise in terms of the bias-variance trade-off. As Hastie, Tibshirani and Friedman point out, "[…] five- or ten-fold cross-validation will overestimate the prediction error. Whether this bias is a drawback in practice depends on the objective. On the other hand leave-one-out cross-validation has low bias, but can have high variance. Overall, five- or tenfold cross-validation are recommended as a good compromise" (Hastie et al. 2009).

Some chemists also advocate a separation of an external part for testing (Esbensen and Geladi 2010), while others maintain the opposite: "hold-out sample is far inferior [as compared to leave-one-out cross-validation]" (Hawkins et al. 2003) or "hold-out samples are downward biased. […] small independent hold-out samples are all but worthless" (Hawkins 2004).

As the machine learning community provides a plethora of novel techniques, which can produce 100 % classification or error-free regression on the training set, the assessment of the predictive performance on future samples (*i.e.* validation, test) has gained increasing importance. There is no single best way to determine the predictive performance of a model, though some options such as leave-one-out cross-validation have become a kind of standard. We should emphasize the statistician's view: "If possible, an independent sample should be obtained to test the adequacy of the prediction equation. Alternatively, the data set may be divided into three parts; one part to be used for model selection [model building or variable selection], the second part for the calibration of parameters in the chosen model and the last part for testing the adequacy of predictions" (Miller 1990).

In the machine learning field (artificial neural networks, support vector machines, *etc.*) this is the standard or at least the advocated practice. In many cases the insufficient number of samples leads to the division of the data into two parts. If the calibration of parameters is done using the same part of the data, substantial biases arise.

We should mention two recent sources with opposite conclusions: Esbensen and Geladi insist categorically on external validation with new measurements (Esbensen and Geladi 2010). Meanwhile, Gütlein *et al.* maintain that "contrary to current conception in the community, cross-validation may play a significant role in evaluating the predictivity of (Q)SAR models" (Gütlein et al. 2013). A somewhat intermediate opinion is presented by Gramatica, who agrees that cross-validation will generally give better and less variable results in terms of the prediction error for the available and modeled data, but also argues that only an additional "external evaluation" on totally new chemicals can represent a future working situation of the model (and thus, assess its predictivity) (Gramatica 2014). Her paper, together with an earlier work of her research group (Gramatica et al. 2012), also presents a thorough data splitting approach for external validation.

Recently, we have shown how one can identify the best (most consistent) performance indicators (merits) and demonstrated the capabilities of sum of ranking differences (SRD) in model selection and in the ranking of the performance merits. Based on two case studies from the literature (using a total of four training-test splits for the two case studies), we established that many of the performance parameters – if not all – for external validation are substantially inferior to other merits even if their application can be advantageous in some cases of data fusion (Rácz et al. 2015).

This work complements our earlier study on model performance parameters with two more case studies from the literature: a QSPR study employing a non-conventional technique, multivariate image analysis (MIA) to predict bioactivity-related properties of small peptides against Dengue virus 2 NS3 proteases (Silla et al. 2011), and a recent work by Roy *et al.* suggesting the use of error measures for QSAR model validation (Roy et al. 2016).

Model performance parameters (merits)

Multivariate models can be evaluated with a large number of performance parameters (merits), including correlation-based (*e.g.* R^2, Q^2) and error-like (*e.g.* MAE, RMSE) merits. In the QSAR modeling field – to the best of our knowledge – the QSARINS modeling software from the group of Paola Gramatica provides the largest pool of model performance parameters during QSAR modeling (Gramatica et al. 2013). A comprehensive summary of this set of performance parameters is available in our recent work (Rácz et al. 2015). In Table 1, the performance parameters occurring in at least one of the discussed case studies are included.

Case study 1

The MIA-QSPR application of Silla *et al.* (Silla et al. 2011) involves the comparison of the correlation coefficient R^2 and the root mean square error RMSE, for calibration (cal), leave-one-out cross-validation (loo) and external validation (ext); a total of six performance parameters. While the selection is of moderate size, it provides an illustrative, balanced distribution of performance merits (two for calibration, two for cross-validation and two for external validation) to be compared. (Nonetheless, our recent work has shown that the outcome of SRD calculations is not – or only negligibly – influenced by the apparent “overweighting” of some methods (Bajusz et al. 2015).)

Case study 2

In contrast, the article of Roy *et al.* (Roy et al. 2016) on QSAR model validation deals exclusively with external validation merits. It is also interesting to know, which external merit(s) is (are) acceptable, preferable or which one(s) should be avoided. This work originally reports eight performance parameters for numerous QSAR models, and was complemented with PRESS values from the courtesy of Prof. Kunal Roy, arriving at a total of ten performance parameters. (PRESS values – along with multiple other merits – were calculated for the whole dataset, as well as for 95% of the data points, after omitting 5% high residual data points.)

Table 1. Definition and description of the performance parameters compared

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
$RMSE$	training, internal and, external validation	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$	Root mean square error
MAE	training, internal. and, external. validation..	$MAE = \frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{n}$	Mean absolute error
CCC	training internal. and, external. validation.	$CCC = \frac{2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \hat{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \hat{y})^2 + n(\bar{y} - \hat{y})^2}$	Coefficient of concordance, concordance correlation coefficient (Lin 1989; Lin 1992)
$PRESS$	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_{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
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 ref. (Consonni et al. 2010) for Q^2 of the external test set (Schüürmann et al. 2008), TR: training 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 ref. (Consonni et al. 2010) for Q^2 of the external test set (Shi et al. 2001), EXT: external test set

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

[b] The following notation is 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 ith sample when the ith sample is left out from the training; n: number of samples; i: sample index.

Data preprocessing methods

Performance parameters can be distributed into two groups, which are scaled reversely: similarity (correlation) coefficient-like and error-like measures. To obtain comparable results we reversed the error like measures. Some well-known data preprocessing methods were used for the datasets: normalization (to unit length), rank transformation, range scaling, and standardization. The techniques are discussed in details below:

Normalization (NOR)

Normalization has several types, such as unit vector, area and mean normalization. Normalization based on area is used mostly in chromatography or spectroscopy, because it means that the observations are divided with the sum of all peak area. Mean normalization can be considered a classic choice: here the observations are divided with the row average. Unit vector normalization is also popular, as it is frequently used in the preprocessing phase of pattern recognition methods. In our case, the latter was used. Its basic idea is that all variables are scaled to unit length, which means that the elements of a column are divided with their Euclidian distances of each column:

$$\boldsymbol{x}_{i,j}^{normalized} = \frac{x_{i,j}}{\sqrt{\sum \boldsymbol{x}_j^2}}, \qquad (1)$$

Here *j* means the running index of columns.

Rank transformation (RNK)

Rank transformation is the simplest data transformation technique, because in this case the only task is to order the values of a column (variable) in increasing (or in the reverse case: decreasing) magnitude and give a rank number to each value in the column. Thus the scale of the values will be between zero and the number of samples.

Range scaling (SCL)

With the use of range scaling the variables are transformed into the [0;1] (or other pre-defined) interval in a simple way:

$$x_{i,j}^{range\ scaled} = \frac{\left(x_{i,j} - Min(x_j)\right)}{\left(Max(x_j) - Min(x_j)\right)}, \qquad (2)$$

Where *j* means the running index for columns: *1, 2, ..., m*. In this case there will be at least one zero and one unity in the dataset (or in each column, in case of more variables) by definition. Range scaling is very sensitive to outliers. Both range scaling and standardization increase the measurement errors. Range scaled values can be easily inverted:

$$x_{i,j}^{reversely\ scaled} = 1 - \frac{\left(x_{i,j} - Min(x_j)\right)}{\left(Max(x_j) - Min(x_j)\right)} \qquad (3)$$

Standardization (STD)

In the case of standardization the centered matrix is divided with the column standard deviations. It is absolutely necessary, if the variables in the dataset are expressed in different units. Standardization can transform the variables to the same scale. In this way the variables are scaled to unit standard deviation. Standardization can be used in two different forms: row-wise and column-wise. While row-wise standardization is more important in the field of spectroscopy, in our case the column-wise version was used. The equation of the standardization process is the following:

$$x_{i,j}(standardized) = \frac{x_{i,j} - Average(x_j)}{\text{standard deviation}(x_j)} \qquad (4)$$

Sum of ranking differences (SRD)

Sum of (absolute) ranking of differences is a novel and general ranking (ordering) and pattern recognition method for the comparison of methods, models and other types of features (variables) (Héberger 2010; Kollár-Hunek and Héberger 2013).

In the beginning the dataset should be compiled in the following format: the variables are arranged in the columns and the samples (observations, compounds) are in the rows. A reference column is also needed for the calculation, which can contain exact reference values, but row average, minimum or maximum values are also applicable as consensus approaches. (The choice depends on the dataset, *e.g.* minimums for error rates and maximums for non-error rates are suitable choices.)

In the first step the compounds (samples, observations) are ranked in every column (in the reference column, as well) in increasing magnitude. In the following step, differences are calculated between the ranks of the reference values and the ranks of each variable, for each row (sample). Finally the (absolute) differences are summed in every column: these are the SRD values, based on which the different models and methods can be compared. The smaller the SRD value, the better the method (more consistent with the reference), thus the best features are close to zero.

The validation of SRD calculations is carried out with a randomization test and a bootstrap-like cross-validation. (If the number of cases is smaller than fourteen, leave-one-out cross-validation is used.)

The final result of SRD is an ordering of methods (models, features, *etc*.), visualized on a plot, where both the *x* and the (left) *y* axis show the same SRD values. (Thus, the SRD values are lines instead of points in the plot.) The information is carried by the location of the lines and their proximity to each other and not by the height of the lines. Additionally, a Gauss-like curve corresponding to the distribution of SRD values of the randomization test is plotted, with frequency values on the right *y* axis. Features that overlap with the 95 % of the Gauss-like curve are not significantly better than the use of random numbers in terms of their ranking behavior, as compared to the reference (the 5 % error limit is marked with dotted lines and abbreviation of XX1 in the SRD plots: anything below this limit is significantly different from random ranking at the 5 % error level. Similarly, XX19 means 95 % confidence, Med is the abbreviation for median).

The results of cross-validated SRD are favorably presented in Box and Whisker plots and can constitute the input of factorial ANOVA analysis in those cases, where there is more than one factor (indicator or grouping variable) present. The basic idea of ANOVA is that it tests the significance of differences between the group averages (where samples are grouped according to the indicator variables). ANOVA is a parametric technique and assumes (multi)normal distribution. In the case of factorial ANOVA we can use more than one factor, which means that we can test all the group averages with different group systems one by one and together as well.

Analysis of variance (ANOVA)

ANOVA is a technique used to assess effects of the categorical factors and their interactions (Lindman 1991). The following model was considered:

$$\mathrm{SRD} = b_0 + b_1 * I1 + b_2 * I2 + b_{1\,2} * I1{*}I2 \qquad (1)$$

where SRD stands for the sum of absolute ranking differences, *I1* is the type of preprocessing (four levels NOR, RNK, SCL, STD), *I*2 is the performance parameter: 6 levels in *Case study 1* ($\mathrm{RMSE_{cal}}$, $\mathrm{RMSE_{ext}}$, $\mathrm{RMSE_{loo}}$, r^2_{cal}, r^2_{ext}, r^2_{loo}) and 10 levels in *Case study 2* (CCC, $\mathrm{PRESS_{95}}$, $\mathrm{PRESS_{100}}$, $\mathrm{MAE_{95}}$, $\mathrm{MAE_{100}}$, R^2_{100}, $Q^2_{\mathrm{F1\_95}}$, $Q^2_{\mathrm{F2\_95}}$, $Q^2_{\mathrm{F1\_100}}$, $Q^2_{\mathrm{F2\_100}}$).

Seven repetitions allow us to test the significance of factors and their interactions.

Variance analysis decomposes the effect of the different factors on the SRD values. This unique combination of SRD and ANOVA provides not only the relative importance of factors, but also an overall evaluation, and has proven to be successful in earlier cases, such as comparing evaluation techniques for genotoxicity measurements (Héberger et al. 2014) and comparing similarity measures for molecular fingerprints (Bajusz et al. 2015).

SRD analyses have been carried out with our own scripts, including the recently published SRD-COVAT heatmaps (Andrić et al. 2016), all of which are downloadable from our website:

http://aki.ttk.mta.hu/srd/

ANOVA calculations have been carried out with STATISTICA (version 12.5, StatSoft, Inc., Tulsa, OK 74104, USA, 2014).

Results and discussion

Similarly to our earlier paper (Rácz et al. 2015) we have chosen two examples from the literature as case studies for the comparison of various model performance parameters applied in the QSAR modeling field. While previously we have taken only the raw data from the publications and carried out QSAR modeling ourselves, this time we have selected two papers that have reported a selection of performance parameters for several models that were compared by the authors.

Case study 1

In a 2011 study Silla *et al.* have applied multivariate image analysis of 2D chemical structures to develop QSPR models for the prediction of bioactivity-related properties (substrate cleavage rate constant k_{cat} and Michaelis constant K_M) of small peptides against Dengue virus 2 NS3 proteases (Silla et al. 2011). Since image analysis is an inherently high-dimensional task (each pixel of the image can be considered a dimension), a suitable variable selection technique is of paramount importance in such studies. To that end, the authors compared numerous PLS models, where the variables were selected with one of (or a combination of) three variable selection methods: interval PLS (iPLS), genetic algorithm (GA) and ordered predictors selection (OPS).

Table 2 of the mentioned paper summarizes six performance parameters for 22 models, namely R^2 and RMSE values for calibration, leave-one-out cross-validation and external validation (see Table 1 for definitions). The external test set was compiled randomly and contained 11 compounds (*vs.* the 43 compounds in the training set). The data in the mentioned table are suitable for a detailed statistical analysis, for a fair comparison of performance parameters (merits).

As the merits are measured on different scales, first they have to be placed on the same scale. Four possibilities have been selected for this task: normalization, rank transformation, range scaling and standardization. (Naturally the error-like measures should be reversed to obtain comparable quantities.) Thus, four 6×22 input matrices were formed according to the data preprocessing techniques.

During the SRD analysis, average was used as the benchmark (reference column) with the consideration that all performance parameters express some prediction ability with error. (The maximum likelihood principle would suggest the usage of average as the best estimation.) Figure 1 shows the ordering result of the SRD procedure on the standardized dataset. Here, R^2 values based on calibration and cross-validation are the two performance parameters that are most consistent with the reference, while $\mathrm{RMSE}_{\mathrm{ext}}$ is over the 5 % limit (*i.e.* indistinguishable from random ranking). The process was repeated for all the four data preprocessing methods and all of the four matrices were subjected to a sevenfold cross-validation. In such a way, 192 SRD values were calculated showing characteristic patterns according to the factors: performance merits (Figure 2) and data preprocessing techniques (Figure 3). As an additional validation step, we have made sure that the SRD values resulting from the whole dataset are in conformity with the SRD value distribution acquired from sevenfold cross-validation (data not shown).

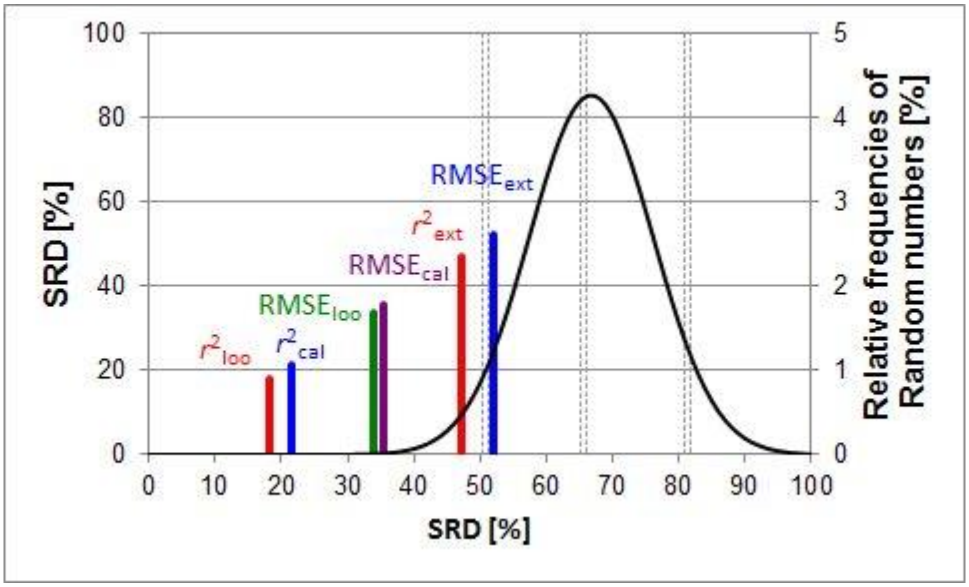


Figure 1. Scaled SRD values (between 0 and 100) compared to random ranking (black Gaussian curve) for the standardized dataset. In this example, $\mathrm{RMSE}_{\mathrm{ext}}$ overlaps with the Gaussian curve and is thus not significantly different from random ranking. r^2_{loo} is the most consistent with the reference (in terms of ranking the models).

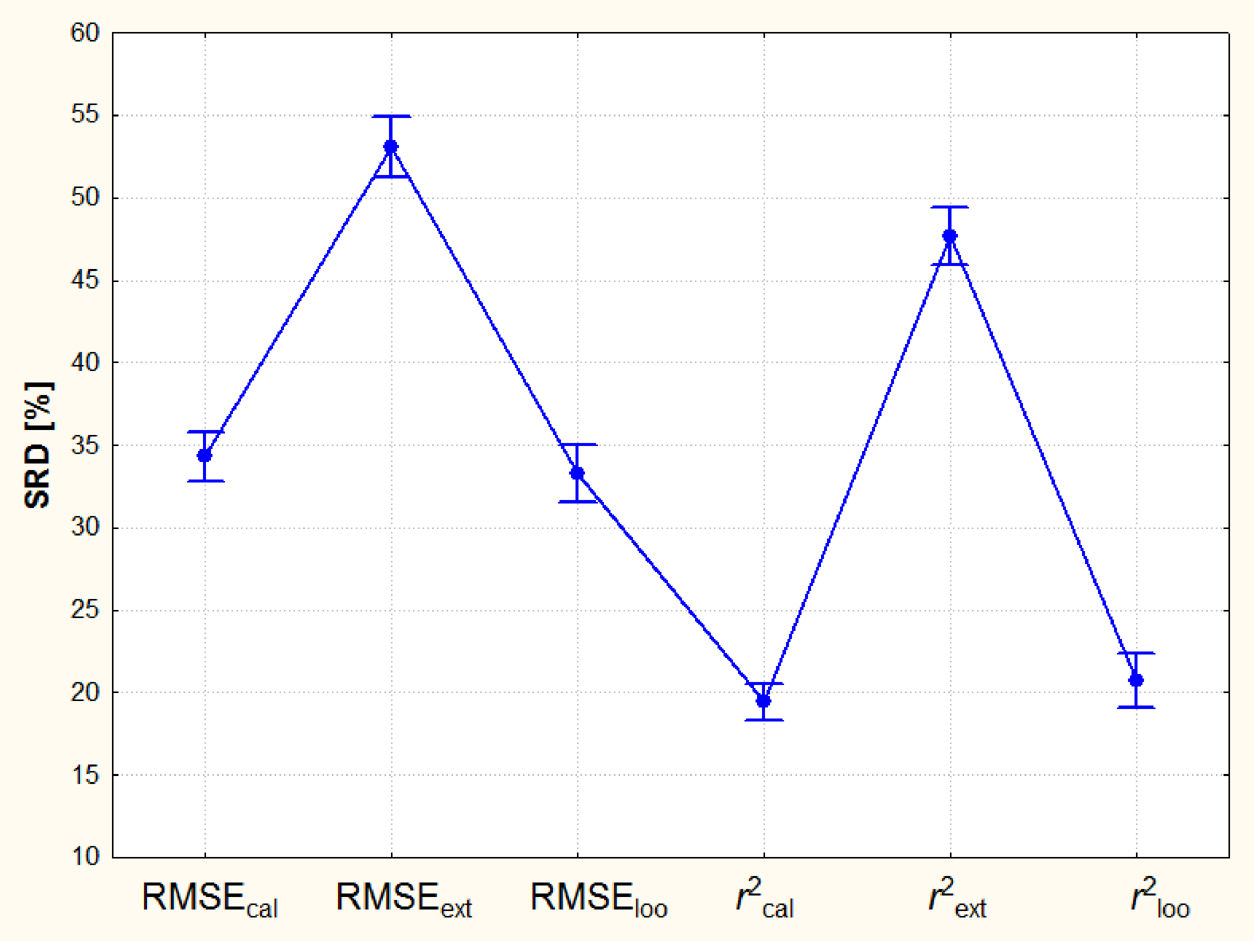


Figure 2. Sevenfold cross-validated SRD results for the comparison of performance parameters. r^2 values based on calibration and leave-one-out cross-validation are the most consistent metrics (as they display the smallest SRD values), RMSE values from the same procedures are intermediate and the r^2 and RMSE values based on external validation are the least consistent with the reference (average).

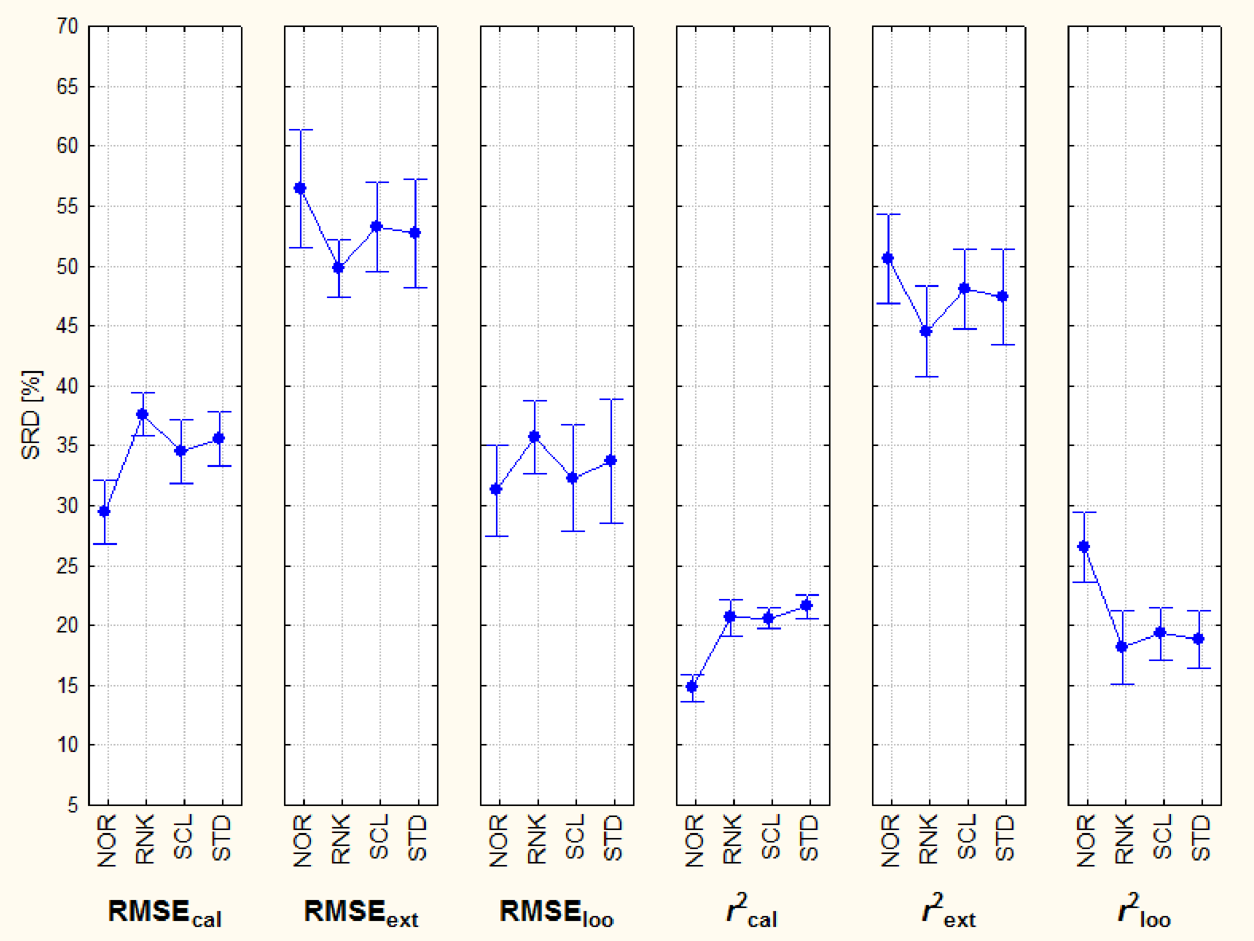


Figure 3. Effects of preprocessing to the SRD values of the performance merits. The preprocessing techniques are generally in good agreement, with the exception of normalization in some cases (such as for r^2_{cal} and r^2_{loo}).

While there is a generally good agreement between ranking, range scaling and standardization, normalization to unit length is peculiar, sometimes the worst (largest), sometimes the best (smallest) one among the preprocessing methods. However, the differences among performance parameters cannot be traced back to the choice of different pretreatment methods, as demonstrated by the ANOVA results in Table 1.

Two-way analysis of variance (ANOVA) has been carried out on the SRD results, with the data preprocessing methods (*I*1) and the performance merits (*I*2) as the two indicator variables. Based on the ANOVA analysis we conclude that the choice of the data preprocessing method does not influence the SRD values significantly (at a 5% error level suggesting that no “artificial” effect was introduced with data preprocessing), while the choice of the performance merit as well as the combination of the two factors does.

Table 1. Results of two-way ANOVA conducted on the cross-validated SRD values, with the data preprocessing methods (*I*1) and the performance parameters (*I*2) as indicator variables, for *Case study 1*. (SS: sum of squares, DOF: degrees of freedom, MS: mean squares)

	SS	DOF	MS	*F*	*p*
Intercept	231607.4	1	231607.4	81233.46	**0.000000**
*I*1	8.6	3	2.9	0.03	0.991579
*I*2	29989.8	5	5998.0	69.61	**0.000000**
*I*1**I*2	1292.5	15	86.2	6.08	**0.000000**
Error	2380.7	168	14.2		

Case study 2

A 2016 study by Roy *et al.* promotes the use of error measures for the evaluation of QSAR models, as a more advantageous alternative to "classic" correlation-based metrics (Roy et al. 2016). The authors argue that while R^2-based performance parameters are easier to comprehend (due to their fixed [0;1] range), they are highly dependent on the range of the response values. However, the study deals exclusively with external validation parameters. In addition, a guideline is proposed to assess the quality of predictions based on the mean absolute error (MAE) and its standard deviation computed from 95 % of the test set predictions (after omitting 5 % high residual data points). Tables 1, 2 and 3 of ref. [11] report various performance parameters based on the external validation of an abundance of QSAR models, and have formed the basis of our analysis. The original tables were complemented with PRESS values from the courtesy of Prof. Kunal Roy. Interestingly, Q^2_{F3} has been left out from the evaluation, though Consonni *et al.* suggested its superiority (Consonni et al. 2010). On another note, Chirico and Gramatica suggested the favorable usage of the coefficient of concordance (Chirico and Gramatica 2011).

The same SRD procedure has similarly been carried out as for *Case study 1*, the average was used as the reference here, as well (see Figure 4). All of the four data preprocessing methods were applied as in *Case study* 1: standardization, normalization, range scaling and rank transformation. Analysis of variance has confirmed the conclusions drawn in the first case study: the choice of the data preprocessing method is not a significant factor (see Table 2) suggesting that no "artificial" effect was introduced with data preprocessing.

Table 2. Results of two-way ANOVA conducted on the cross-validated SRD values, with the data preprocessing methods (*I*1) and the performance parameters (*I*2) as indicator variables, for *Case study 2*. (SS: sum of squares, DOF: degrees of freedom, MS: mean squares)

	SS	DOF	MS	*F*	*p*
Intercept	589413.5	1	589413.5	88787.07	**0.000000**
I1	34.6	3	11.5	1.74	0.159475
I2	45074.0	9	5008.2	754.42	**0.000000**
I1*I2	884.6	27	32.8	4.94	**0.000000**
Error	1858.8	280	6.6		

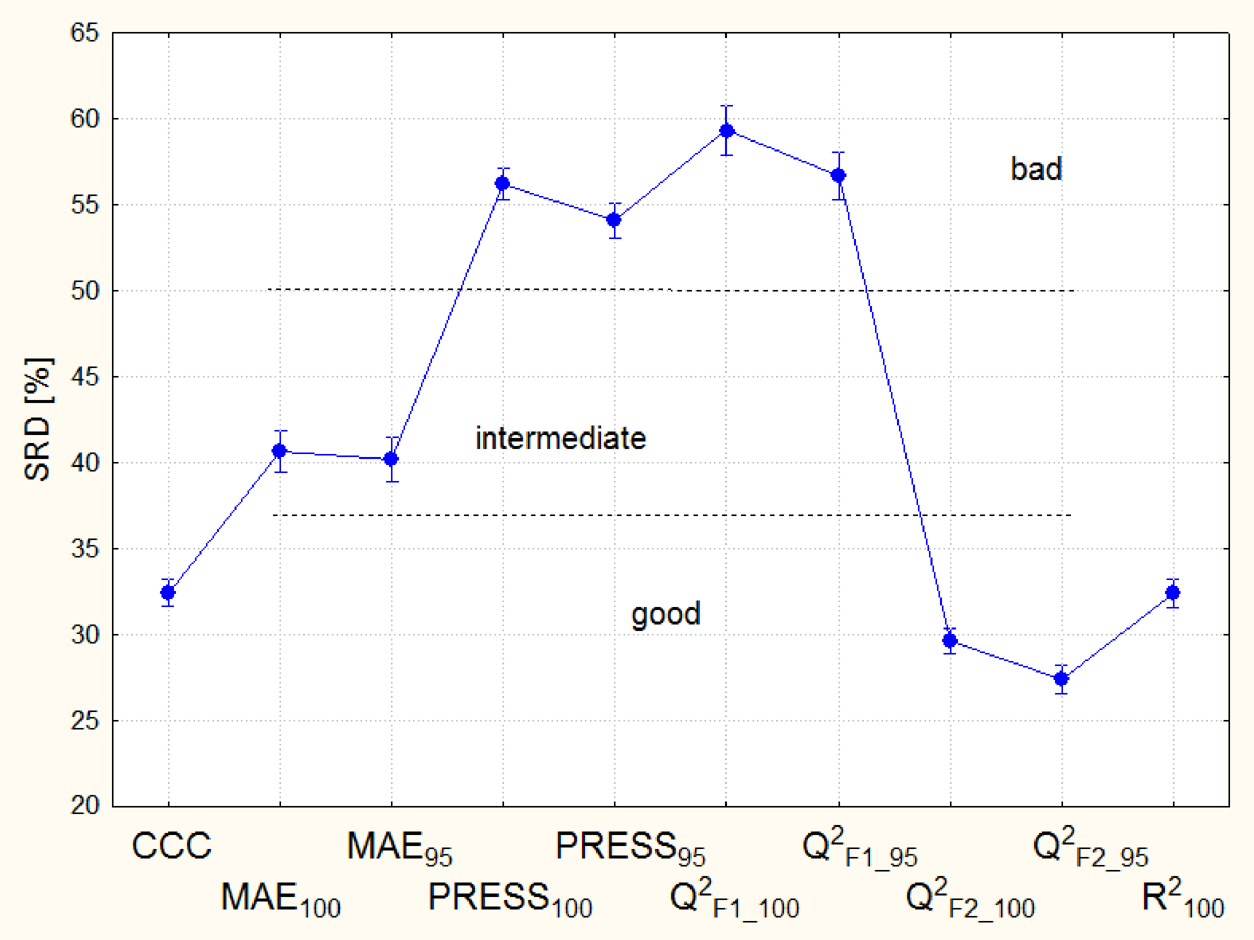
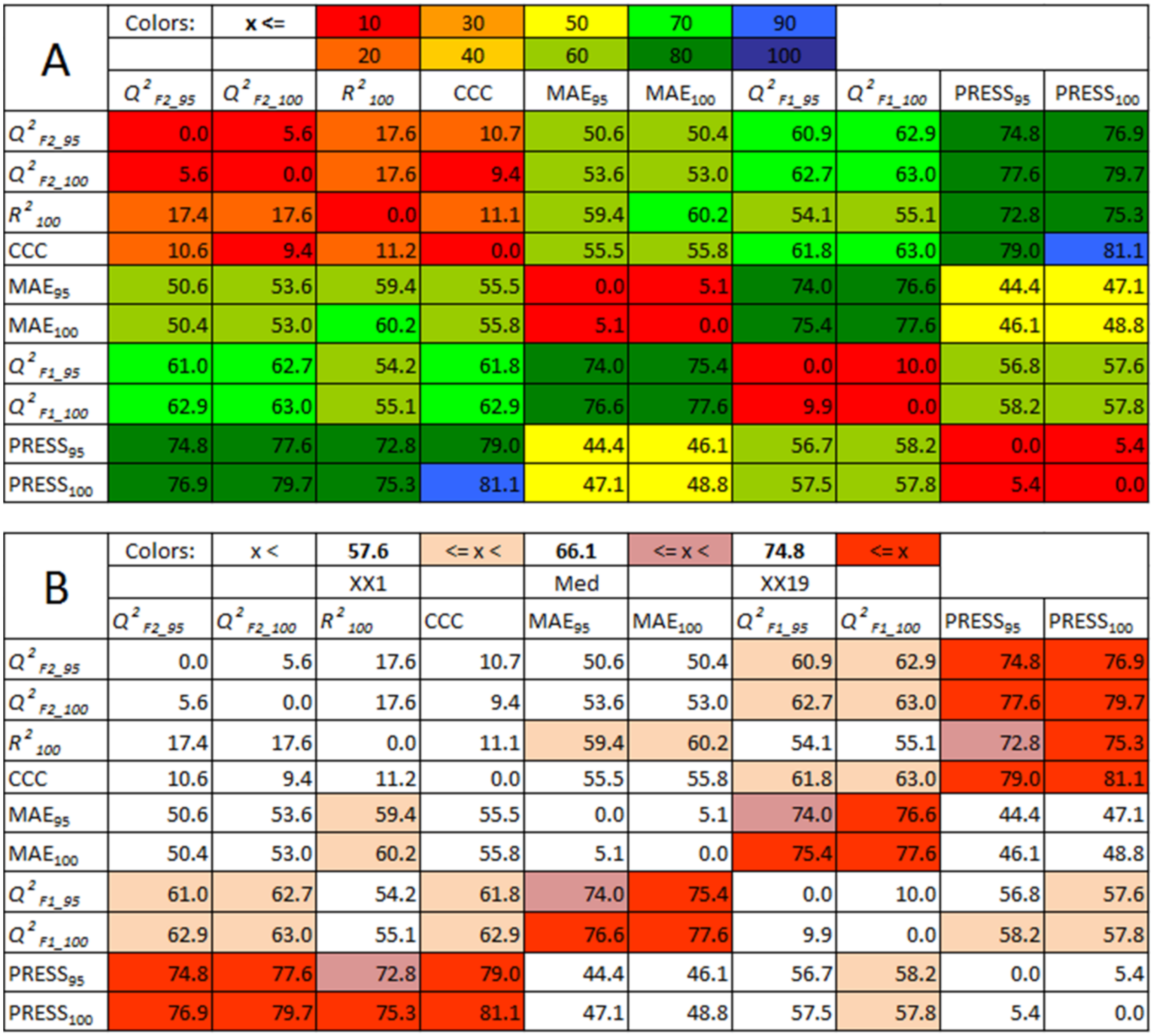


Figure 4. Sevenfold cross-validated SRD results for the comparison of external validation parameters (with average values as the reference vector). It is relatively easy to classify these performance parameters into good (consistent), intermediate and bad (least consistent) ones considering the SRD gaps between groups. The fact that the concordance correlation coefficient is among the good merits strengthens Chirico and Gramatica's recommendation (Chirico and Gramatica 2011) about its usefulness.

One can argue that the SRD results are principally determined by the selection of the reference (benchmark) column, which is true to some extent (but overlooks the maximum likelihood principle and the superiority of the consensus approach over an individual reference variable). Therefore, we have elaborated a technique to examine the underlying data structure to a finer "resolution". In this case, each variable (column) is used as the reference, one at a time and a color-coded matrix (heatmap) is compiled from the results. This approach was termed COVAT – Comparison with One VAriable at a Time – and was introduced in our recent paper on lipophilicity scales (Andrić et al. 2016). The different benchmarks eliminate the problem of golden standard selection: the grouping of the SRD values obtained with the various reference vectors can reveal the underlying connections between the examined variables (here, performance parameters).

For the heatmap calculations, standardization has been selected as the data preprocessing method – keeping in mind that the factor of data preprocessing was proven to be *not* significant. The results are shown in Figure 5.



A	Colors:	x <=	10	30	50	70	90			
			20	40	60	80	100			
	$Q^2_{F2\_95}$	$Q^2_{F2\_100}$	R^2_{100}	CCC	MAE_{95}	MAE_{100}	$Q^2_{F1\_95}$	$Q^2_{F1\_100}$	$PRESS_{95}$	$PRESS_{100}$
$Q^2_{F2\_95}$	0.0	5.6	17.6	10.7	50.6	50.4	60.9	62.9	74.8	76.9
$Q^2_{F2\_100}$	5.6	0.0	17.6	9.4	53.6	53.0	62.7	63.0	77.6	79.7
R^2_{100}	17.4	17.6	0.0	11.1	59.4	60.2	54.1	55.1	72.8	75.3
CCC	10.6	9.4	11.2	0.0	55.5	55.8	61.8	63.0	79.0	81.1
MAE_{95}	50.6	53.6	59.4	55.5	0.0	5.1	74.0	76.6	44.4	47.1
MAE_{100}	50.4	53.0	60.2	55.8	5.1	0.0	75.4	77.6	46.1	48.8
$Q^2_{F1\_95}$	61.0	62.7	54.2	61.8	74.0	75.4	0.0	10.0	56.8	57.6
$Q^2_{F1\_100}$	62.9	63.0	55.1	62.9	76.6	77.6	9.9	0.0	58.2	57.8
$PRESS_{95}$	74.8	77.6	72.8	79.0	44.4	46.1	56.7	58.2	0.0	5.4
$PRESS_{100}$	76.9	79.7	75.3	81.1	47.1	48.8	57.5	57.8	5.4	0.0

B	Colors:	x <	57.6	<= x <	66.1	<= x <	74.8	<= x		
			XX1		Med		XX19			
	$Q^2_{F2\_95}$	$Q^2_{F2\_100}$	R^2_{100}	CCC	MAE_{95}	MAE_{100}	$Q^2_{F1\_95}$	$Q^2_{F1\_100}$	$PRESS_{95}$	$PRESS_{100}$
$Q^2_{F2\_95}$	0.0	5.6	17.6	10.7	50.6	50.4	60.9	62.9	74.8	76.9
$Q^2_{F2\_100}$	5.6	0.0	17.6	9.4	53.6	53.0	62.7	63.0	77.6	79.7
R^2_{100}	17.4	17.6	0.0	11.1	59.4	60.2	54.1	55.1	72.8	75.3
CCC	10.6	9.4	11.2	0.0	55.5	55.8	61.8	63.0	79.0	81.1
MAE_{95}	50.6	53.6	59.4	55.5	0.0	5.1	74.0	76.6	44.4	47.1
MAE_{100}	50.4	53.0	60.2	55.8	5.1	0.0	75.4	77.6	46.1	48.8
$Q^2_{F1\_95}$	61.0	62.7	54.2	61.8	74.0	75.4	0.0	10.0	56.8	57.6
$Q^2_{F1\_100}$	62.9	63.0	55.1	62.9	76.6	77.6	9.9	0.0	58.2	57.8
$PRESS_{95}$	74.8	77.6	72.8	79.0	44.4	46.1	56.7	58.2	0.0	5.4
$PRESS_{100}$	76.9	79.7	75.3	81.1	47.1	48.8	57.5	57.8	5.4	0.0

Figure 5. SRD-COVAT heatmaps of the external validation parameters in *Case study 2* with an equidistant (A) and a "Gaussian" (B) color coding. (Color references are provided on the upper parts of the images.) While panel A highlights four clusters of similar performance parameters, panel B provides information on the significance of SRD values i.e. relative to the distribution of random rankings.

Figure 5a highlights three groups of external validation merits: group 1 (upper left part) contains $Q^2_{F2\_95}$, $Q^2_{F2\_100}$, R^2_{100} and CCC, group 2 (middle) contains MAE_{95} and MAE_{100}, group 3 (lower right) contains $Q^2_{F1\_95}$, and $Q^2_{F1\_100}$, $PRESS_{95}$, and $PRESS_{100}$. While the first two groups confirm the conclusions based on Figure 4 completely, the third "group" can be further divided based on Q^2_{F1} and PRESS values, though they have similar (sometimes overlapping) SRD distributions against the average as reference (see Figure 4). Additionally, the pairs of performance parameters calculated from the whole dataset and 95 % are close to each other, as expected.

Figure 5B offers even more intriguing results, as it shows the SRD values relative to the SRD distribution of random rankings (consult the Gaussian curve on Figure 1 for reference): cells of any other color than white denote that there is no correspondence between the rankings produced by the two external validation parameters indicated in the implied row and column headers. As there are many such cells in the table, we can conclude that the ranking results obtained by most of these (external) performance merits are highly divergent. Ultimately, this can safely be considered as a conclusion supporting the preference of performance parameters based on cross-validation, since there seems to be little consensus among those based on external validation (see Figure 5). From the external validation metrics, we would suggest the use of Q^2_{F2}, R^2_{100} and CCC as the most consensual ones.

Conclusion

We have carried out a comparison of QSAR model performance parameters based on two case studies, with the combination of sum of ranking differences (SRD) and analysis of variance (ANOVA). The first case study has shown cross-validation based performance metrics to be more consistent with the consensus ranking than those based on external validation. In the second case study, we have compared some members of the latter group in more detail and have shown that the rankings produced by them are greatly divergent. The results presented here corroborate our earlier, recently published findings (Rácz et al. 2015) on diverse data sets of independent literature sources.

Showing a model to be predictive for a small external test set does not necessarily mean that it will be predictive for molecules outside of this test set. In other words, in the case of external validation we are delivered to a random test, which might be informative but not necessarily. While we agree that a more meticulous training-test splitting approach (such as the one presented by Gramatica *et al.* (Gramatica et al. 2012)) can significantly improve the reliability of external validation, we would still advise against overemphasizing model performance parameters based on external validation, or preferring them over the ones derived from cross-validation. (In our opinions, a consensus approach might be the best choice here.) In the lack of sufficient test data (which is often the case in QSAR modeling), our results reinforce the conclusions of Hawkins *et al.* (Hawkins et al. 2003), who advise against small holdout samples (to avoid the loss of information in model building) and recommend cross-validation instead.

Acknowledgement

The work was supported by the Hungarian Scientific Research Fund (OTKA, grant number K 119269).

References

Andrić F, Bajusz D, Rácz A, et al (2016) Multivariate assessment of lipophilicity scales—computational and reversed phase thin-layer chromatographic indices. J Pharm Biomed Anal 127:81–93. doi: 10.1016/j.jpba.2016.04.001

Bajusz D, Rácz A, Héberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7:20.

doi: 10.1186/s13321-015-0069-3

Chirico N, Gramatica P (2011) 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 51:2320–35. doi: 10.1021/ci200211n

Consonni V, Ballabio D, Todeschini R (2010) Evaluation of model predictive ability by external validation techniques. J Chemom 24:194–201. doi: 10.1002/cem.1290

Esbensen KH, Geladi P (2010) Principles of Proper Validation: use and abuse of re-sampling for validation. J Chemom 24:168–187. doi: 10.1002/cem.1310

Gramatica P (2014) External Evaluation of QSAR Models, in Addition to Cross-Validation: Verification of Predictive Capability on Totally New Chemicals. Mol Inform 33:311–314. doi: 10.1002/minf.201400030

Gramatica P, Cassani S, Roy PP, et al (2012) QSAR Modeling is not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)triazoles on Algae. Mol Inform 31:817–835. doi: 10.1002/minf.201200075

Gramatica P, Chirico N, Papa E, et al (2013) QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. J Comput Chem 34:2121–2132. doi: 10.1002/jcc.23361

Gütlein M, Helma C, Karwath A, Kramer S (2013) A Large-Scale Empirical Evaluation of Cross-Validation and External Test Set Validation in (Q)SAR. Mol Inform 32:516–528. doi: 10.1002/minf.201200134

Hastie T, Tibshirani R, Friedman J (2001) Overview of supervised learning. In: Elements of Statistical Learning. Data Mining, Inference, Prediction. Springer, New York, NY, USA, pp 9–43

Hastie T, Tibshirani R, Friedman JH (2009) Cross-Validation. In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York, pp 241–249

Hawkins DM (2004) The problem of overfitting. J Chem Inf Comput Sci 44:1–12. doi: 10.1021/ci0342472

Hawkins DM, Basak SC, Mills D (2003) Assessing model fit by cross-validation. J Chem Inf Comput Sci 43:579–86. doi: 10.1021/ci025626i

Héberger K (2010) Sum of ranking differences compares methods or models fairly. TrAC Trends Anal Chem 29:101–109.

Héberger K, Kolarević S, Kračun-Kolarević M, et al (2014) Evaluation of single-cell gel electrophoresis data: combination of variance analysis with sum of ranking differences. Mutat Res Genet Toxicol Environ Mutagen 771:15–22. doi: 10.1016/j.mrgentox.2014.04.028

Kollár-Hunek K, Héberger K (2013) Method and model comparison by sum of ranking differences in cases of repeated observations (ties). Chemom Intell Lab Syst 127:139–146. doi: 10.1016/j.chemolab.2013.06.007

Lin LI-K (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–68.

Lin LI-K (1992) Assay Validation Using the Concordance Correlation Coefficient. Biometrics 48:599. doi: 10.2307/2532314

Lindman HR (1991) Analysis of Variance in Experimental Design. Springer-Verlag, New York

Miller A (1990) Subset selection in regression. Chapman and Hall, London

Rácz A, Bajusz D, Héberger K (2015) Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters. SAR QSAR Environ Res 26:683–700. doi: 10.1080/1062936X.2015.1084647

Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. Chemom Intell Lab Syst 152:18–33. doi: 10.1016/j.chemolab.2016.01.008

Schüürmann G, Ebert R-U, Chen J, et al (2008) External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean. J Chem Inf Model 48:2140–5. doi: 10.1021/ci800253u

Shi LM, Fang H, Tong W, et al (2001) QSAR Models Using a Large Diverse Set of Estrogens. J Chem Inf Model 41:186–195. doi: 10.1021/ci000066d

Silla JM, Nunes CA, Cormanich RA, et al (2011) MIA-QSPR and effect of variable selection on the modeling of kinetic parameters related to activities of modified peptides against dengue type 2. Chemom Intell Lab Syst 108:146–149. doi: 10.1016/j.chemolab.2011.06.009