

Title: New fusion strategies for similarity and diversity indices

Final report for 2018-09-01 - 2020-02-29 time period

Project ID: KH_17 125608

Novel, modern data fusion strategies have been elaborated and tested on various fields and instances. There is no need to recapitulate all results and conclusions as the publications are readily available mostly as open access contributions.

However, at least one characteristic figure has been selected for each issue emphasizing one of the important conclusions.

1) Pattern recognition.

Class modelling methods (sometimes called one class modelling) could be considered as a significant contribution of chemist to the classification tasks. The most frequently applied method is the soft independent modeling of class analogies (SIMCA). Our thorough survey of classification data sets and a rigorous comparison of classification methods clearly show the unambiguous superiority of other techniques over SIMCA in any classification task. The ordering of classification techniques was validated with a randomization test and cross-validation. Whereas SIMCA frequently (but not always) passed the randomization test, cross-validation unambiguously proves its inferiority to other techniques in supervised classification tasks.

The next figure shows a comparison of nine classifiers on 27 highly different data sets (linearly not separated and class in class situations included). Sum of Ranking differences (SRD) is a city block (Manhattan) distance between the individual classifier and the reference ranking (golden standard), here the hypothetical best classifier showing the best performance (maximum of row values) on each data sets.

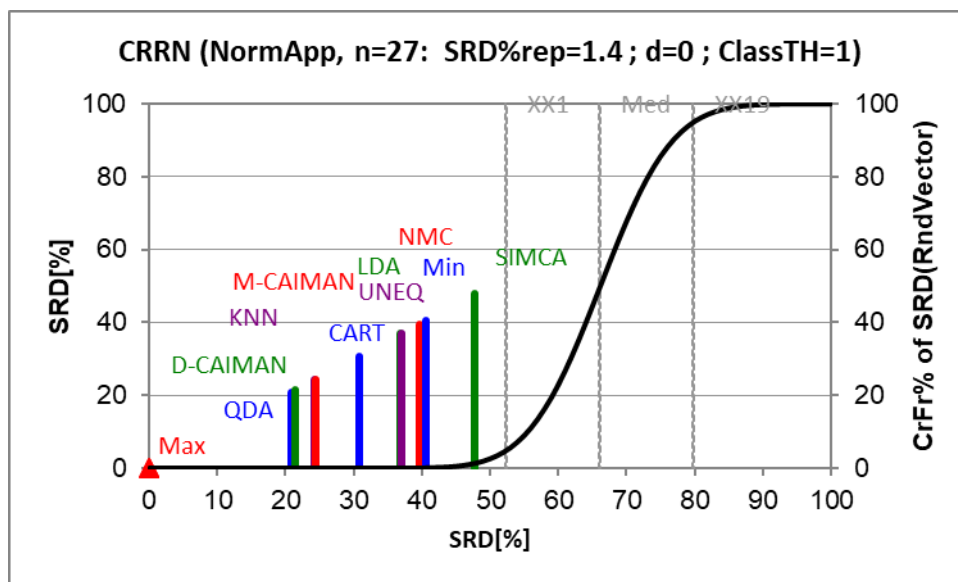


Figure 1. Normalized SRD values (scaled between 0 and 100) compared to random ranking (increasing part of the black cumulative distribution function (CDF) curve, right y axis) for the leave-one-out non-error rates (NER%) of 27 data sets. SRD values are plotted on the x and left y-axes, the right y-axis The 5% error level (XX1) is also given.

Notations: Classification And Influence Matrix Analysis (CAIMAN: D-discriminant, M-modeling), classification and regression trees (CART), *k*-nearest neighbors (KNN), linear

discriminant analysis (LDA), nearest mean classifier (NMC), quadratic discriminant analysis (QDA), soft independent modeling of class analogy (SIMCA) and unequal dispersed classes (UNEQ).

SRD ranks and groups the classifiers. It is easy to perceive that the line for SIMCA is the farthest from the reference: it is worse than the hypothetically worst classifiers (min), *i.e.* SIMCA is the worst from among the studied classifiers. It can happen as the various data sets rank SIMCA reverse order, at least partly.

The publication summarizes six case studies for diverse classification task; all validated with randomization tests and various variants of cross-validation.

[1] Anita Rácz, Attila Gere, Dávid Bajusz, Károly Héberger*, Is soft independent modeling of class analogies a reasonable choice for supervised pattern recognition?

RSC Advances, **8**, pp. 10-21. (2018)

<https://doi.org/10.1039/c7ra08901e>

if(2018)=3.049

2) Comparison of binary similarity coefficients

A comprehensive evaluation of binary similarity measures has been completed for the elucidation of patterns among samples of different botanical origin and various metabolomic profiles. Baroni-Urbani–Buser (BUB) and Hawkins–Dotson (HD) similarity coefficients were selected as the best measures by sum of ranking differences (SRD) and analysis of variance (ANOVA), while Dice (Di1), Yule, Russel-Rao, and Consonni-Todeschini 3 ones ranked the worst.

First binary fingerprints were calculated; then, 44 similarity metrics for nine data sets: coding is 1-metabolic present, 0-absent. The uncertainties were estimated by >50 repeated resampling (bootstrap), *i.e.* a part of the rows were eliminated and SRD procedure was carried out on the remaining objects.

Definitions, labels, and names of similarity metrics are given in the Appendix, Table A1.

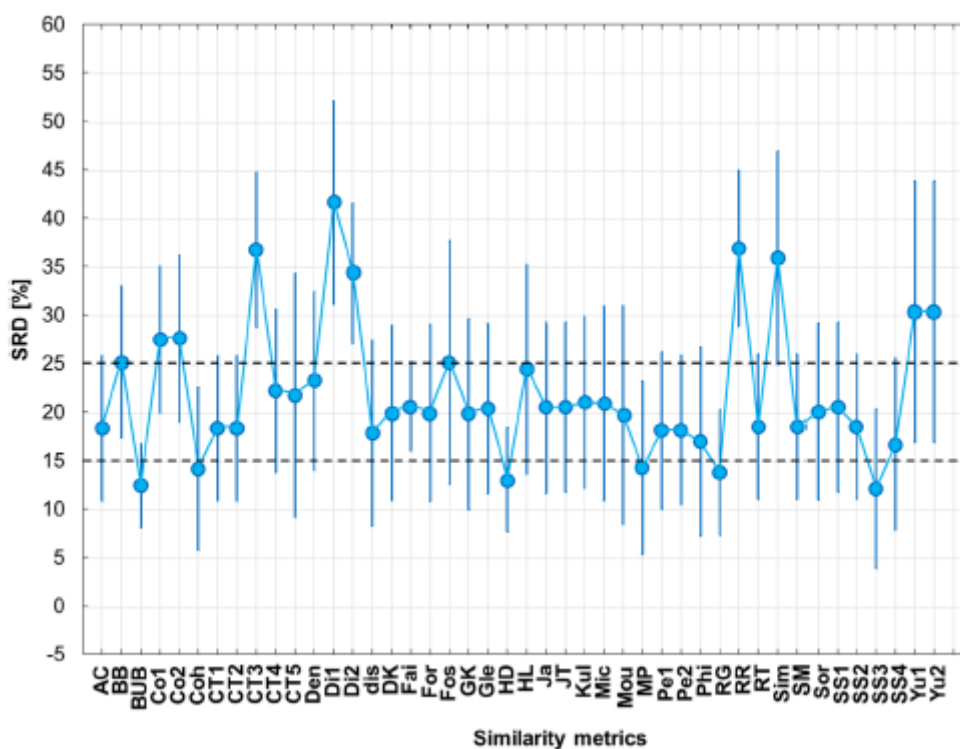


Figure 2. ANOVA decomposition of similarity metrics as a factor. Dashed lines symbolize the limit of the best/consistent (lower part), worst (upper part) and medium groups

of similarity metrics based on SRD values. 95% confidence limits are plotted with vertical bars. The dotted lines are arbitrary.

The similarity metrics can be split to three groups based on this plot: those having smaller SRD values than 15 can be considered the most consistent based on the 9 datasets. Metrics between SRD values of 15 and 25 are in the medium group, while the weakest ones have SRD values greater than 25.

Comparison with the cluster analysis based on quantitative profiles has validated the findings: the best similarity coefficient (BUB) has provided the same pattern (almost exactly) as the quantitative data.

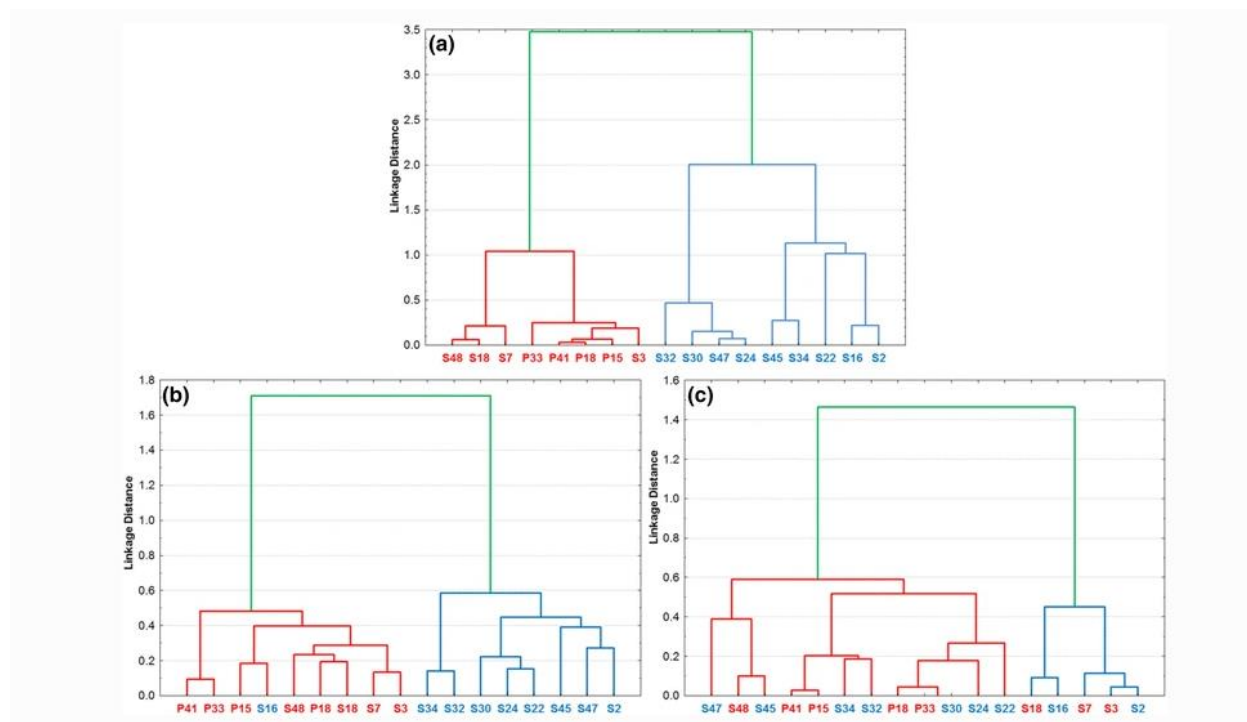


Figure 3. Comparison of cluster analysis trees (linkage rule: Ward's method). (a) Reference (quantitative results). (b) Binary fingerprints with the BUB (best) distance metric. (c) Binary fingerprints with Di1 (worst) distance metric. The two largest clusters (indicated with red and blue) were compared. It is clearly seen that the number of misclassifications (as compared to the reference) is one for the BUB, and 10 for the Di1 measure.

The above figures clearly show that the clustering by binary coding might be similarly useful as clustering by quantitative data, provided optimal binary coefficients is used.

[2] Anita Rácz, Filip Andrić*, Dávid Bajusz, Károly Héberger, Binary similarity measures for fingerprint analysis of qualitative metabolomic profiles, *Metabolomics*, **14**, Article Number: 29. pp. 1-9. (2018)
<https://doi.org/10.1007/s11306-018-1327-y> if(2018)=3.167

3) Model validation

Three types of cross-validation (randomized and stratified fivefold and leave-one-out cross-validation), three types of variable selection algorithms were compared with factorial analysis of variance (ANOVA) tests. We also examined the effect of the applied datasets (calibration, test samples, and both sets) based on the original predicted values. Sum of ranking differences (SRD) values can be used as a promising and useful performance parameter for the ranking and evaluation of numerous regression models. When experimental values were used in data fusion step (for gold standard) the best models could be found based

on their SRD values. We also carried out a systematic comparison of various modeling techniques such as PLS regression, principal component regression (PCR) and support vector machines (SVM). The properly validated support vector machine models proved to be the best for all of the four used datasets.:

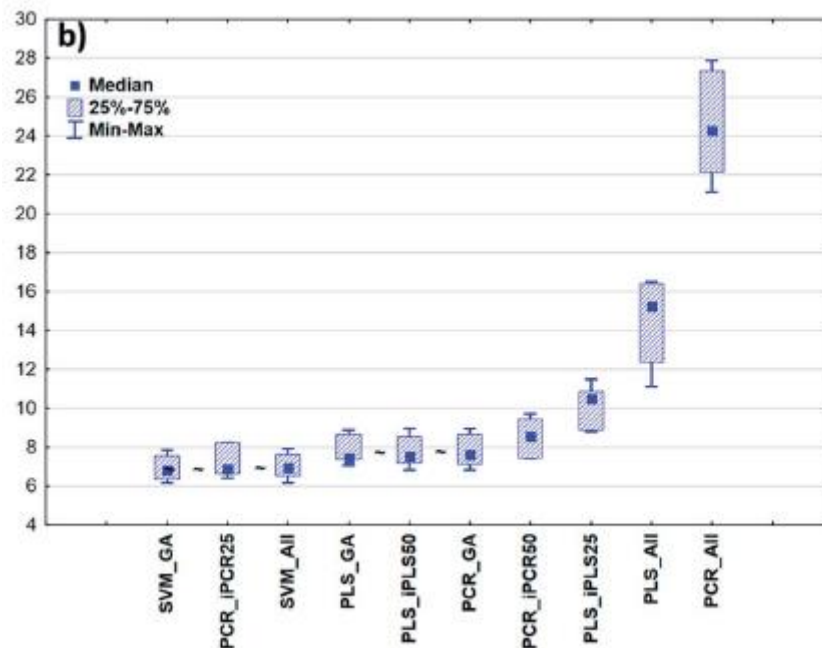


Figure 4. Box and whisker plots of the transmission spectral models based on fivefold cross-validated SRD in the case of dry material content. The “~” mark means that there is no statistically significant difference between the two models according to Wilcoxon's matched pair test at the 5% level.

- [3] Anita Rácz*, Marietta Fodor, Károly Héberger, Development and comparison of regression models for determination of quality parameters in margarine spread samples using NIR spectroscopy, *Analytical Methods*, **10**, Issue 25, pp. 3089-3099. (2018)
<https://doi.org/10.1039/c8ay01055b> if(2018)=2.378

4) Data fusion for cross-validation variants

Prediction performance often depends on the cross- and test validation protocols applied. Several combinations of different cross-validation variants and model-building techniques were examined, applying five-fold cross-validation (with random, contiguous and Venetian blind forms) and leave-one-out cross-validation (CV). External test sets showed the effects and differences between the validation protocols SRD can provide a unique and unambiguous ranking of methods and CV variants. Venetian blind cross-validation proved to be a promising validation realization tool. The variable selection was always advantageous, and the modelling had a larger influence on the performance parameters than any or all of the CV variants.

The use of variable selection, cross-validation (CV) variants and modelling methods were used as factors for ANOVA. SRD analysis was performed with the original experimental values (toxicity values), and with the average as the reference. With the experimental values we can choose those models and combinations of parameters that led to the best predictions of the original experimental values. On the other hand, the use of the average could show us the most consistent models

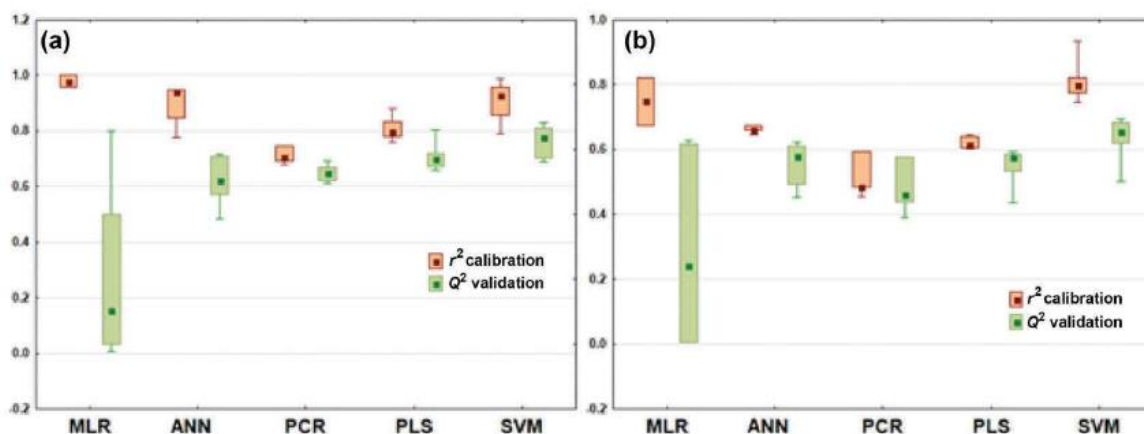


Figure 5. Box and whisker plots of the r^2 and Q^2 values for model building methods: (a) for case study 1; and (b) for case study 2. The median was used as the center points (boxes), the rectangles contain 50% of the data (first and third quartile), and the whiskers are located at the minimum and maximum values. Notations: multiple linear regression (MLR), artificial neural networks (ANN), principal component regression (PCR), Partial least squares regression (PLS), support vector machine (SVM).

The smallest gaps between red and green boxes mean the best validated models; however, on the expenses of the performances on the training data sets. It is somehow odd that the least validated technique, the multiple linear regression (MLR), is the most frequently used one. PCR rarely applied nowadays, though it is the best validated technique.

- [4] Anita Rácz, Dávid Bajusz and Károly Héberger*, Modelling methods and cross-validation variants in QSAR: a multi-level analysis, *SAR and QSAR in Environmental Research*, **29**, Issue 9, pp. 661-674. (2018) <https://doi.org/10.1080/1062936X.2018.1505778> if(2018)=2.287

5) New fusion techniques for interaction fingerprints

As a complementary method to ligand docking, Interaction fingerprints (IFP) can be applied to quantify the similarity of predicted binding poses to a reference binding pose. A large number of similarity metrics (44) were compared, and various parameters of the IFPs themselves have also been customized. In a large-scale comparison, we have assessed the effect of similarity metrics and IFP configurations to a number of virtual screening scenarios with ten different protein targets and thousands of molecules. The performances were compared based on area under curves (AUC) values and on original similarity data. Similarity metrics were evaluated with several statistical tests and the novel, robust sum of ranking differences (SRD) algorithm: we evaluate the consistency (or concordance) of the various similarity metrics to an ideal reference metric, which is provided by data fusion from the existing metrics. We could find better similarity metrics than the Tanimoto coefficient. Better coefficients than the Tanimoto one were identified: Simple Matching (SM), Rogers-Tanimoto, Sokal-Sneath, Consoni-Todeschini 1 and 2, and Austin-Colwell coefficients can be recommended. The recommended indices can be seen in Figure 6 (next Page), these are the metrics, below the first dotted line.

The notations and classification of similarity measures (see Appendix Table A1) were kept the same as in ref. [Todeschini R, Consonni V, Xiang H *et al.* (2012) Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *J Chem Inf Model* 52:2884–2901. <https://doi.org/10.1021/ci300261r>]

A link is given to calculate these coefficients in Python are as follows: <https://github.com/davidbajusz/fpkit>

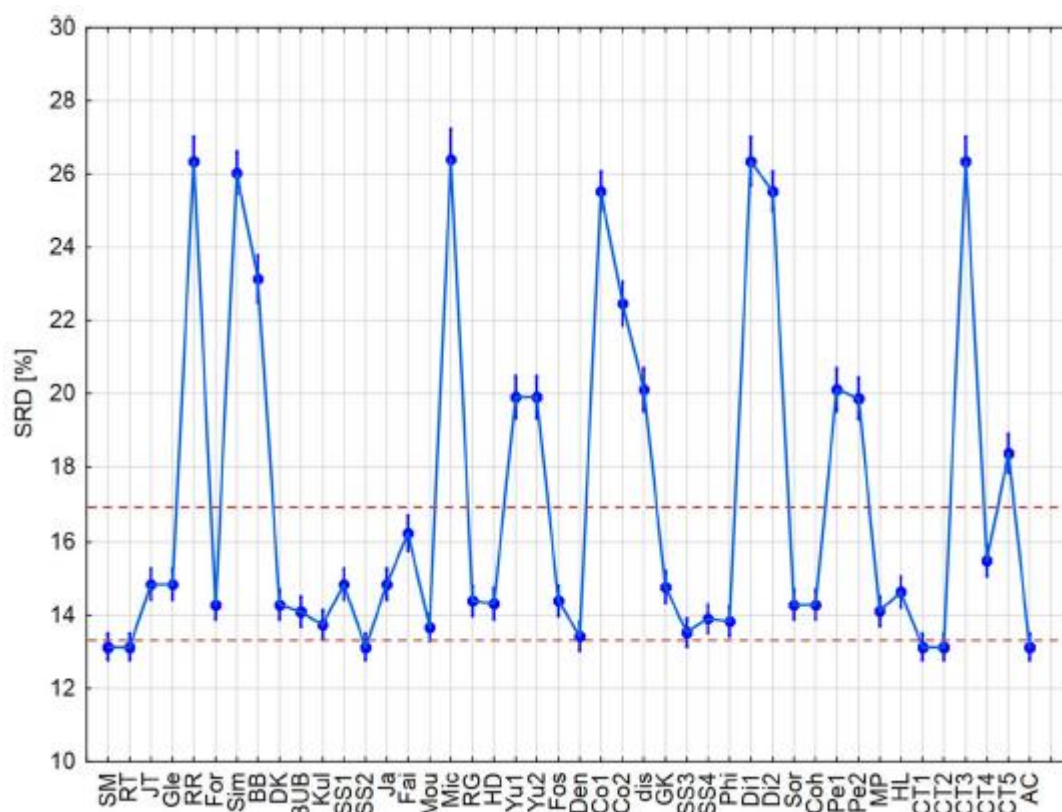


Figure 6. Factorial ANOVA with the similarity measures as one of the factors. Average values are marked with blue dots and the blue lines below and above the dots denote 95% confidence intervals. Normalized SRD values [%] are plotted against the similarity measures. The red dashed lines are arbitrary thresholds defined to select the best few metrics, and to identify the region with the less consistent similarity measures

[5] Anita Rácz, Dávid Bajusz*, Károly Héberger, Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints, *Journal of Cheminformatics* **10**, Article Number: 48 (2018) <https://doi.org/10.1186/s13321-018-0302-y> if(2018)=4.154

6) Row maximum, minimum and mean as data fusion methods

Carefully selected case studies (three) have disclosed similarities and differences in validation variants. The next validation variants have been examined: stratified (contiguous block), repeated Monte Carlo resampling, and how many times the data set is split (5-7-10). The fair method comparison algorithm called sum of ranking differences (SRD) can rank and group the model validation variants. SRD in combination with variance analysis reveals whether the differences among validation variants are significant or merely the play of random errors. In special circumstances any of the influential factors for validation variants can exert significant influence on evaluation as shown by sums of (absolute) ranking differences (SRDs): The optimal validation variant should be determined individually again and again. A random resampling with sevenfold cross-validations seems to be a good compromise to diminish the bias and variance alike.

In this manuscript SRD is considered as a bias term and analysis of variance (ANOVA) decomposes the random fluctuations around the biases according to the factors. Figure 7 was selected as an example of comparison of classifiers. The original publication introduced a new algorithm called random projection (RP).

It is interesting to mention that random projection (applied to the last three classifiers in the figure 7) to optimal dimension is not necessarily a viable option; it can provide low bias and small variance and reversely high bias and high(er) variance. The reason is probably that the optimized solution can hardly be further improved; on the other hand, not perfect technique(s) such as kNN can be significantly improved by random projection.

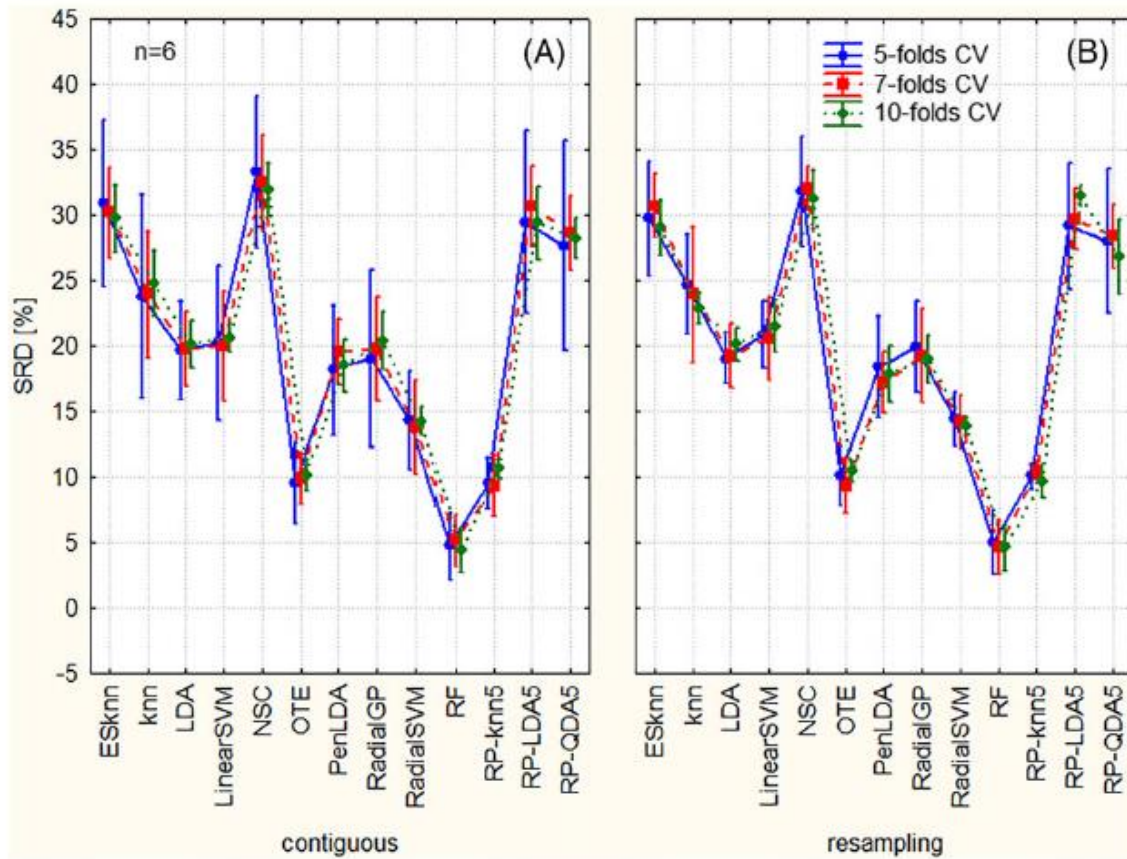


Figure 7. Interaction of the three factors: ways of cross-validation (validation variants): A, blockwise, *i.e.*, contiguous, and B, (Monte Carlo) random resampling. Line plots show the number of folds (5, 7, and 10).

The figure also shows the categorization of classifiers (the best ones have the smallest bias and smallest variance alike). The dotted lines are arbitrary thresholds, visually set. Linear and radial basis function has also been used with GP and SVM, linear and radial, respectively.

Abbreviations for classifiers: ESkNN, ensemble of subset of kNN classifiers; GP, Gaussian process; kNN, k -nearest neighbors; LDA, linear discriminant analysis; NSC, nearest shrunken centroids; OTE, optimal tree ensemble; PenLDA, penalized LDA; QDA, quadratic discriminant analysis; RF, random forest; RP, random projection; SRD, sum of ranking differences; SVM, support vector machine; the number after the abbreviations means “sufficient dimension reduction (SDR=5) assumption. References for non-trivial classifiers can be found in ref. [6] below.

Almost all classifiers are significantly different; however, *post hoc* tests (Bonferroni and Scheffé) amalgamate four of them (LDA, linear SVM, PenLDA, and radial GP, *i.e.* SRD values of 18 and 21). Indeed, they are related techniques; it shows the inner consistency of the SRD analysis.

General trends can be observed: the variance become smaller as the number of folds in cross-validation decreases. Contiguous variant has higher variance, if the original data set

contains structure (systematic arrangement). SRD is sensitive enough to reveal the non-negligible structure in the data. Therefore, a randomization of objects (samples) is recommended before the SRD analysis.

- [6] Károly Héberger* and Klára Kollár-Hunek, Comparison of validation variants by sum of ranking differences and ANOVA, *Journal of Chemometrics*, **33**, pp. 1-14, Article number: e3104 (2019)
<https://doi.org/10.1002/cem.3104> if(2018)=1.847

7) Effects of data reduction in QSAR model building

QSAR/QSPR (quantitative structure-activity/property relationship) modeling has been a prevalent approach in various, overlapping sub-fields of computational, medicinal and environmental chemistry for decades. The generation and selection of molecular descriptors and fingerprints is an essential part of this process. Variable reduction is to be done before any modeling. Variable reduction is an unsupervised technique, generally includes elimination of correlated variables and variables with zero (or minimal variance). In this study, we examined in detail the effect of various possible descriptor intercorrelation limits on the resulting QSAR models.

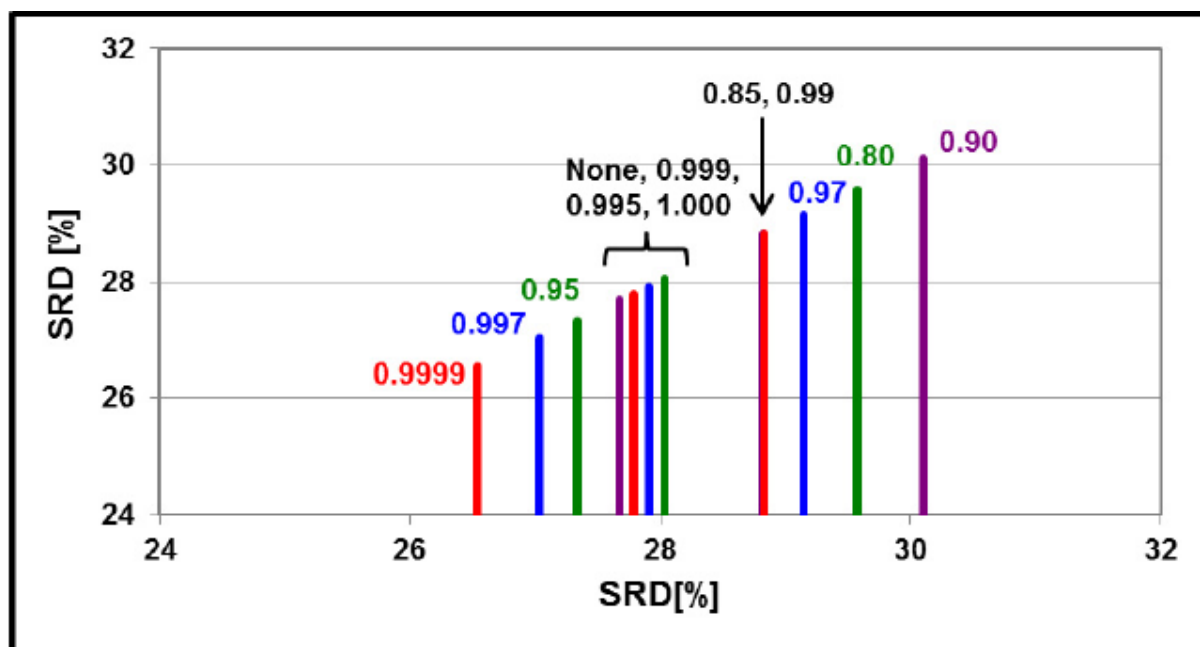


Figure 8. An example of SRD ranking of intercorrelation limits (filtering threshold) The black curve corresponds to the cumulative distribution of SRD values based on random rankings cannot be seen in the blown-up. On the left Y and X axes, normalized SRD [%] values are plotted, while the right Y axis shows the percentages for the distribution of random rankings.

The non-monotonous character of intercorrelation limits can easily be perceived.

- [7] Anita Rácz, Dávid Bajusz, Károly Héberger*, Intercorrelation limits in molecular descriptor preselection for QSAR/QSPR. *Molecular Informatics*, **38**, Article Number: 1800154 (2019)
<https://doi.org/10.1002/minf.201800154> if(2018)=2.375

8) Data fusion for ensemble docking

Ensemble docking is a widely applied concept in structure-based virtual screening—to at least partly account for protein flexibility—usually granting a significant performance gain at a modest cost of speed. In this study, several data fusion methods were tested and compared for ensemble docking. Seven fusion rules were applied and four performance parameters were used for the comparison of the fusion metrics. We have found much better alternatives for the consensus scoring instead of the widely applied minimum fusion rule.

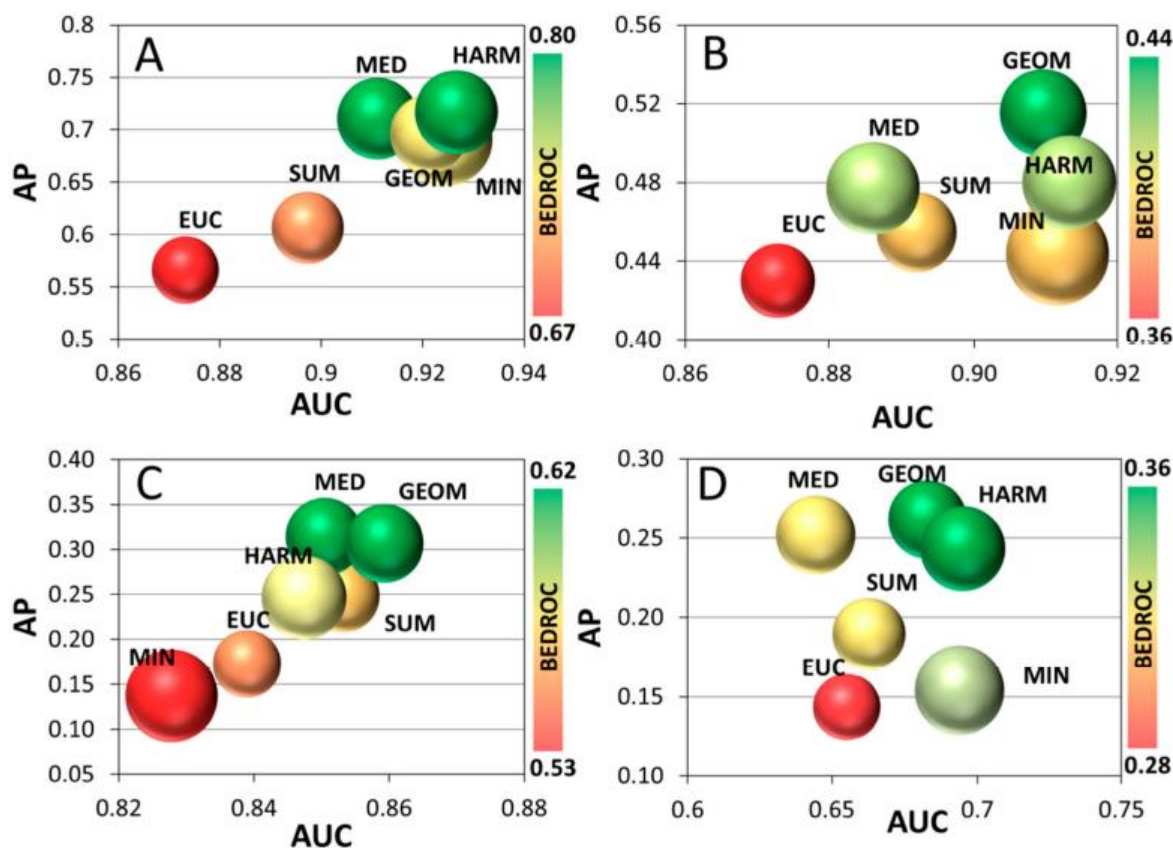


Figure 9. Bubble plots for various (diverse) datasets. Average precision (AP) values are plotted against the area under receiver operating characteristic curves (AUC) values. Bubble sizes correspond to the SRD and the color scale correspond to Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) values, increasing from red to green (see color scale on the right). The abbreviations denote the fusion rules: MED – median, GEOM – geometric mean, HARM – harmonic mean, EUC – Euclidean distance, MIN – minimum, SUM – sum (it equals the mean here).

Bubble plots allow a multidimensional evaluation, in the present case four. The higher the AP, the AUC values the better, green is better than yellow or even the red in case of BEDROC, whereas SRD values the smaller the better. Although there are some conflicts between data sets, several general conclusions can be drawn, *e.g.* The frequently applied minimum fusion rule is never the best, but can be the worst, *etc.*

- [8] Dávid Bajusz, Anita Rácz*, Károly Héberger Comparison of Data Fusion Methods as Consensus Scores for Ensemble Docking, *Molecules*, **24**, Article Number: 2690 (2019) <https://doi.org/10.3390/molecules24152690>

if(2018)=3.060

9) Multi-level comparison of machine learning classifiers

Machine learning classification algorithms are widely used for the prediction and classification of the different properties of molecules such as toxicity or biological activity. Performance metrics give diverse information about the performance of machine learning methods; therefore, compound metrics created with data fusion are highly desirable. The most optimal and robust performance parameters were found in addition to the different classification scenarios, such as balanced or imbalanced groups, 2-class or multi-class problems. The algorithms were compared and their robustness and validation features were revealed.

The workflow applied for the comparisons is summarized in Figure 10

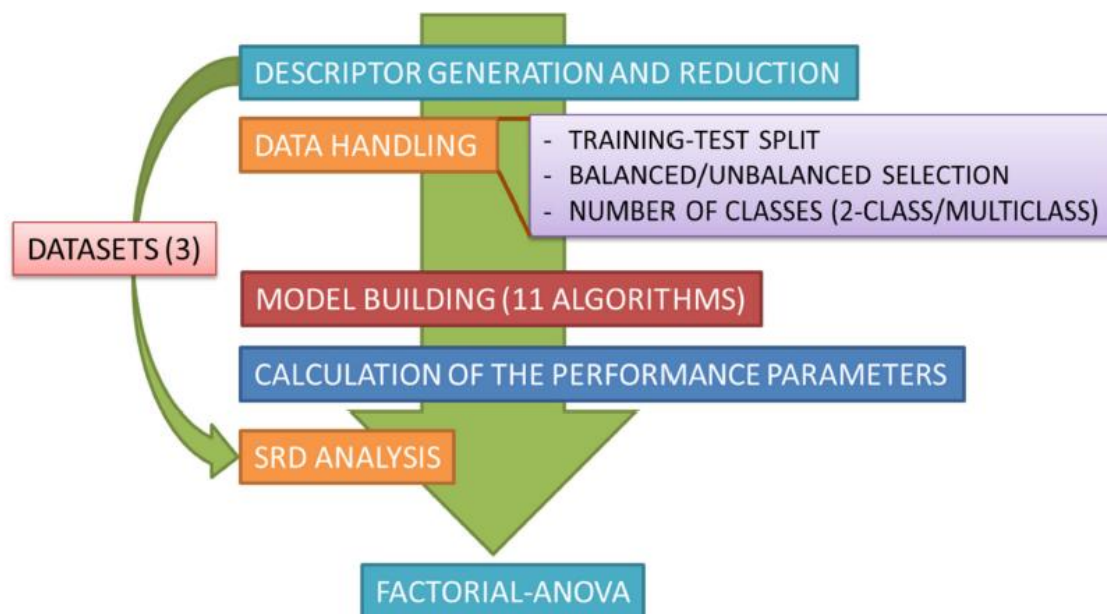


Figure 10. Workflow of the comparative study. Briefly, after descriptor generation and reduction, eleven machine learning methods are applied for model building (for each combination of 2-class/multiclass and balanced/imbalanced cases). After the calculation of the performance parameters, statistical analysis of the results is carried out with sum of ranking differences (SRD) and factorial analysis of variance (ANOVA). The complete process is carried out on three highly diverse datasets.

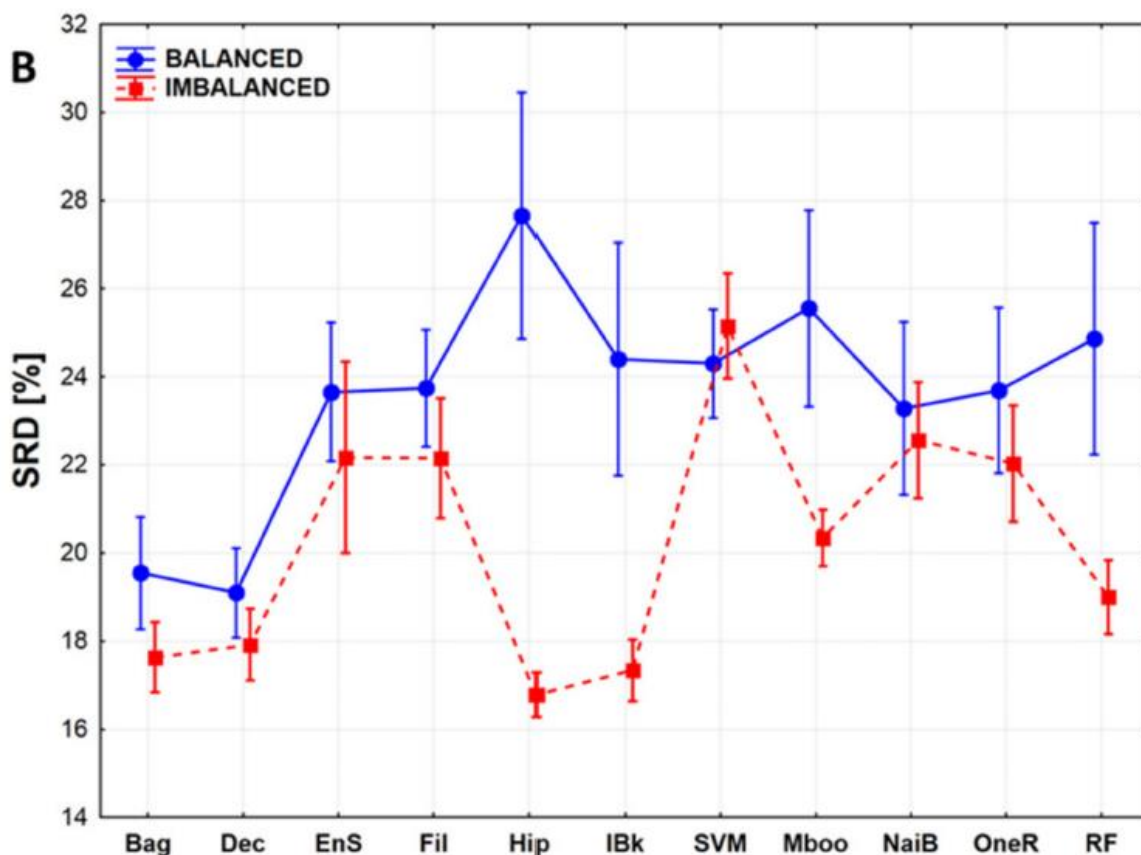
SRD values are, on average, higher for 2-class classification scenarios (farther from the reference), meaning that there is a greater degree of disagreement between the performance metrics in this case, highlighting the importance for their informed selection and application during model evaluation. The difference is even more pronounced if the dataset is imbalanced.

The relatively small distance from the reference (SRD = 0) suggests that the hypothetical best classifier is well approximated with the bagging (Bag), k-nearest-neighbor (1Bk), and Decorate (Dec) methods.

Figure 11 (B, next page) Decomposition of the classifiers according to dataset composition (balanced vs. imbalanced classes). Normalized SRD values for the eleven classifiers. Error bars mean 95% confidence intervals.

It can properly be seen that some classifiers behave considerably differently if the experimental design is balance or imbalanced. Some classifiers, such as Bagging, Decorate, Support vector machine and Naïve Bayes are not sensitive to dataset composition. whereas Hyperpipe, k-nearest neighbor, random forest and to a lesser extent Megaboost are highly

sensitive ones. These findings call for the selection of classifiers according to the experimental design. Only several classifiers are suitable in both cases and the special gain justifies the usage of hyperpipe and k -nearest neighbors for the imbalanced case.



Abbreviations: Naïve Bayes (NaiB), FilteredClassifier (Fil), k -nearest neighbor (IBk), Lazy, HyperPipe (Hip), MultiboostAB (Mboo) library SVM (SVM), oneR, based on 1-rule, (OneR), Bagging (Bag), Ensemble Selection (EnS), Decorate (Dec), Random Forest (RF)

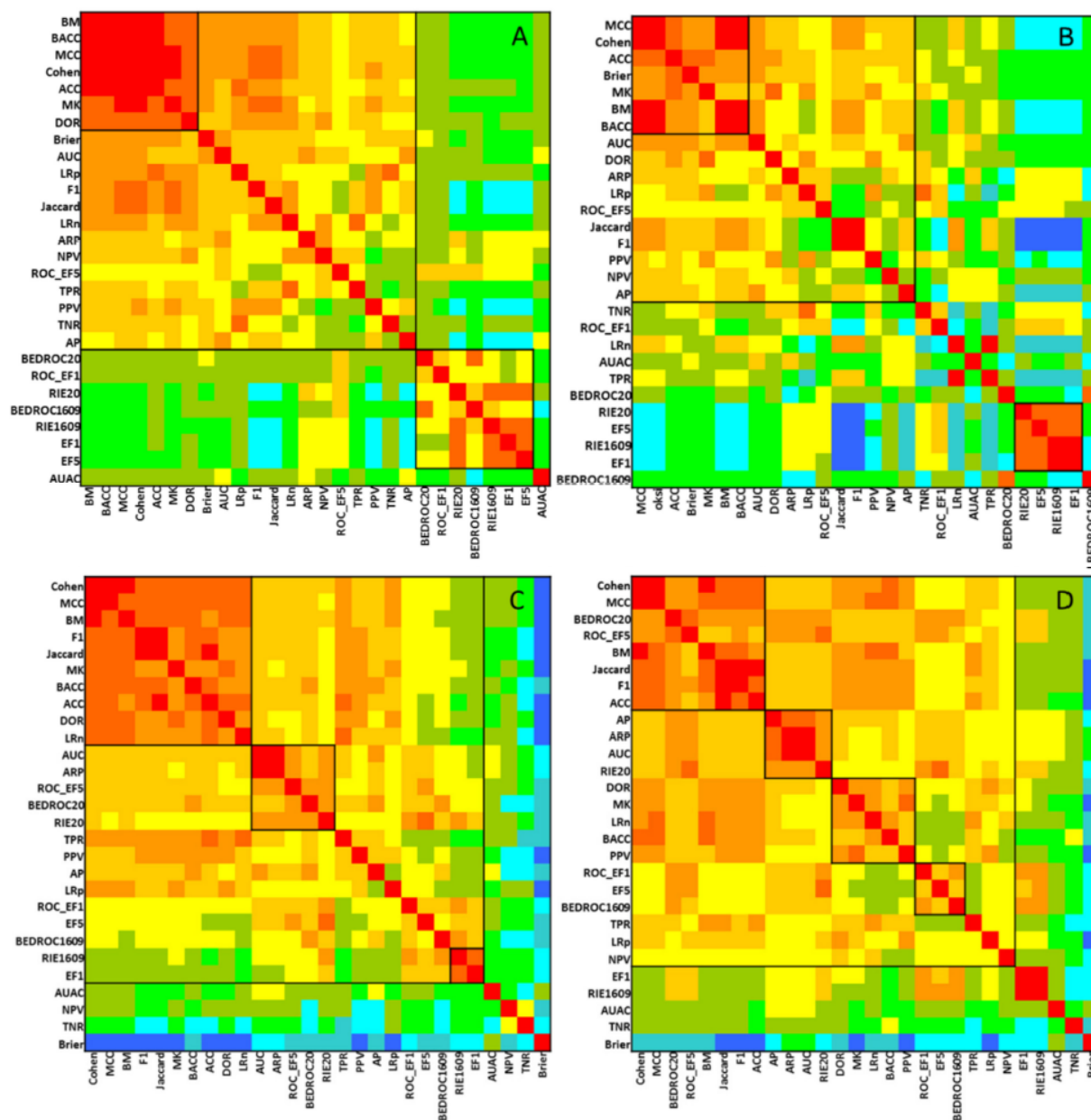
10) Multi-level comparison of performance metrics for binary classification

There is a steady problem with usage of SRD algorithm. In some cases, the definition of reference (gold standard, benchmark) is not obvious and the results are highly dependent on the selected reference. To overcome this difficulty all variables have been selected as reference and ordered them according to the sum of the SRD values. The techniques is called SRD-COVAT (Comparisons with One Variable at a Time. <https://dx.doi.org/10.1016/j.jpba.2016.04.001>)

Figure 12. (next page) Results of the SRD-COVAT method: 2-class classification with balanced (A) and imbalanced (B) classes; and multiclass classification with balanced (C) and imbalanced (D) classes. Clusters of similarly behaving performance parameters are separated with black lines (squares) on the plot based on visual inspection.

Abbreviations for performance parameters: Accuracy (ACC), Area under the accumulation curve (AUAC), Area under the ROC curve (AUC), Average precision (AP), Average rank (position) of actives (positives) (ARP), Balanced accuracy (BACC), Bookmaker informedness (BM), Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC), Brier score loss, (BR), Cohen's kappa (Cohen), Diagnostic odds ratio (DOR) Enrichment factor (EF 1 or 5 %), F1 score (F1), Jaccard score (Jaccard), Matthews correlation coefficient (MCC), Markedness (MK), Negative likelihood ratio (LRn),

Negative predictive value (NPV), Positive likelihood ratio (LRp), Positive predictive value (PPV), ROC enrichment (ROC_EF) Robust initial enhancement (RIE), True positive rate (TPR), True negative rate (TNR). Detailed descriptions and definitions can be found in the Appendix Tables A2-A4 and in ref. [9].



Although three clusters can always be observed, the performance for the individual metrics are different. Balanced accuracy, Matthews correlation coefficient, Cohen kappa, Bookmaker informedness, *etc.* are always among the best representations, whereas the usage of Brier score should be avoided for multiclass classifications.

[9] Anita RÁCz, DÁvid Bajusz*, KÁroly Héberger, Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics, *Molecules*, **24**, Article Number: 2811 (2019) <https://doi.org/10.3390/molecules24152811>

11) Practical applications of data fusion

Data fusion can be used for grouping *trans*-resveratrol and anthocyanin concentrations according to (oak) barrel type and, wine sorts (Kadarka, Kékfraknos, Cabernet France).

- [10] Z Guld, A Rácz*, H Tima, M Kállay, Dn Sárdy, Effects of aging in oak barrels on the *trans*-resveratrol and anthocyanin concentration of red wines from Hungary
Acta Alimentaria, **48**, pp. 349-357. (2019) If(2008-2018)=0.274-0.505
<https://doi.org/10.1556/066.2019.0004>

Similarly, the 100-point OIV sensory test and quantitative descriptive analysis can be amalgamated and as factors decomposed in wine sensory analysis.

- [11] Z. Guld, D. N. Sárdy, A. Gere*, A. Rácz Comparison of sensory evaluation techniques for Hungarian wines
Journal of Chemometrics. **34** e3219. (2020) if(2018)=1.847
<https://doi.org/10.1002/cem.3219>

12) Software development

SRD is developed as an MS Excel macro, and is freely available for download at: <http://aki.ttk.mta.hu/srd>.

FingerPrint Kit - Python-based cheminformatics package for fingerprint-related tasks. <https://github.com/davidbajusz/fpkit/>, are available via: <https://doi.org/10.5281/zenodo.1217969>

We have also developed a python code to calculate numerous classification performance metrics which were compared in our recent paper, ref. [9]. The code is made available upon request and is planned to be released via *github*.

13. Summary

The above summary (Part 1-12) of works on data fusion demonstrate the usefulness and abilities of the novel algorithm called sum of ranking differences (SRD). SRD is not only a simple distance metric, but an algorithm containing three steps [12,13]:

i) a data fusion act: the ideal ranking (benchmark, golden standard, or reference) is fixed: the average assumes a kind of consensus, minimum for errors, misclassification rates, etc. corresponds to the hypothetical best method with the smallest error. Another possibility is the maximum *e.g.* for correlation coefficients, non-error rates. Intuitively, one can easily accept that we are better off, if closer to the golden standard reference (benchmark).

ii) Calculation of the SRD values: it equals with Spearman's footrule (after data fusion) only if no ties present in the input matrix, but we developed a metric, *i.e.* "SRD with ties" [14], then.

iii) Validation options: Exact theoretical distributions were derived for randomization test, if the number of rows (n)<14. The random distribution is approximated reasonably with the Gaussian distribution between $13 < n < 1400$. Cross-validation assigns uncertainties to the SRD values. At present 5-10-fold cross-validation with and without repeated sampling is included in our program [13,15].

In case the 'data structure' is unknown *a priori*; then, a fair method comparison is the best choice by using SRD.

If the variables, (factors, indicators) are conflicting, the only proper solution is an optimization by multicriteria decision making (MDCM) also known as post-Pareto analysis (PPA). Although virtually endless number of tools exist for that, SRD is known to be a consensus of numerous MDCM tools [16]. Recent examinations clearly and unambiguously have shown that SRD realizes a multicriteria optimization. [16,17].

- [12] K. Héberger*, Sum of ranking differences compares methods or models fairly, *TRAC - Trends in Analytical Chemistry* **29**, pp. 101-109. (2010)
<https://doi.org/10.1016/j.trac.2009.09.009>
- [13] K. Héberger* and K. Kollár-Hunek, Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers
Journal of Chemometrics **25**, pp. 151-158. (2011)
<https://doi.org/10.1002/cem.1320>
- [14] K. Kollár-Hunek and K. Héberger*, Method and Model Comparison by Sum of Ranking differences in Cases of Repeated Observations (Ties)
Chemometrics and Intelligent Laboratory Systems **127**, pp. 139-146. (2013)
<http://dx.doi.org/10.1016/j.chemolab.2013.06.007>
- [15] K. Héberger* and K. Kollár-Hunek, Comparison of validation variants by sum of ranking differences and ANOVA
Journal of Chemometrics **33**, pp. 1-14 Article number: e3104 (2019)
<https://doi.org/10.1002/cem.3104>
- [16] J. M. Lourenço and L. Lebensztajn*, Post-Pareto Optimality Analysis with Sum of Ranking Differences,
IEEE Transactions on Magnetics, **54**, Issue: 8 pp. 1-10. Article Sequence Number: 8202810 (2018).
<https://doi.org/10.1109/TMAG.2018.2836327>
- [17] A. Rácz, D. Bajusz and K. Héberger, Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters,
SAR and QSAR in Environmental Research, **26**, pp. 683-700. (2015)
<https://doi.org/10.1080/1062936X.2015.1084647>

14) Conclusion and outlook

Sum of ranking differences (SRD) provides a unique and unambiguous ranking of methods, models, items, *etc.* SRD coupled with analysis of variance (ANOVA) provide a unique and unambiguous way of decomposing the effects and determine the best combination of factors.

SRD is simple, it corresponds to principle of parsimony. SRD is a nonparametric technique, but highly sensitive; it is able to find differences, when other methods cannot.

The validation using randomization test enhances its reliability. Validation by cross-validation assigns uncertainties to the SRD values. Sign test or preferably Wilcoxon's matched pair test is able to provide a statistically correct discrimination at a predefined error limit (say 5%).

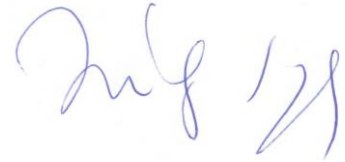
Selection the best item(s), ranking and grouping them belong to the most advantageous features of SRD.

The method can be applied in all laboratories where measurement data are evaluated. It is suitable to determine consistency of training test set splits, the models can be classified into good and bad ones, *etc.* As it has recently become apparent [16,17, see above] that SRD corresponds to the optimum of multi-criteria decision analysis. It has great potential, can be used not only in chemistry but in all other fields, *e.g.* evaluation of performance of athletes, ranking of universities, optimal selection of voting districts, determination of dominant factors for sensory investigations, ranking and grouping the ways of determining the partition coefficient ($\log P$), to list only a few.

Novel techniques are planned to be developed: *e.g.* alternative distance measures, sum of absolute differences (SAD), weighted schemes, sum of percentile distances (PSD), sum of utility distances (SUD), coupled (hybrid) techniques and their applications. Therefore, a new

proposal was submitted to the National Research, Development and Innovation Office of Hungary under Grant Numbers K 134260, the decision has not been made yet.

Budapest, March 27 / 2020

A handwritten signature in blue ink, appearing to read 'Károly Héberger'.

Károly Héberger
Scientific advisor

Binary similarity coefficients and performance parameters are summarized in the Appendix **Tables A1-A3**.

Appendix

Table A1. List of the binary similarity coefficients, their definitions, concordance symmetry and metric properties.

No	Label	Name	Equation	Scaling parameters		Concordance symmetry	Metricity
				α	β		
1	SM	Simple matching, Sokal-Michner	$s_{SM} = \frac{a+d}{p}$	0	1	S	M
2	RT	Rogers-Tanimoto	$s_{RT} = \frac{a+d}{p+b+c}$	0	1	S	M
3	JT	Jaccard-Tanimoto	$s_{JT} = \frac{a}{a+b+c}$	0	1	A	M
4	Gle	Gleason	$s_{Gle} = \frac{2a}{2a+b+c}$	0	1	A	N
5	RR	Russel-Rao	$s_{RR} = \frac{a}{p}$	0	1	A	M
6	For	Forbes	$s_{For} = \frac{pa}{(a+b)(a+c)}$	0	p/a	A	M
7	Sim	Simpson	$s_{Sim} = \frac{a}{\min\{(a+b), (a+c)\}}$	0	1	A	N
8	BB	Braun-Blanquet	$s_{BB} = \frac{a}{\max\{(a+b), (a+c)\}}$	0	1	A	M
9	DK	Driver-Kroeber, Ochiai, cosine	$s_{DK} = \frac{a}{\sqrt{(a+b)(a+c)}}$	0	1	A	N
10	BUB	Baroni-Urbani-Buser	$s_{BUB} = \frac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c}$	0	1	I	M
11	Kul	Kulczynski	$s_{Kul} = \frac{1}{2} \left(\frac{a}{a+b} + \frac{a}{a+c} \right)$	0	1	A	N
12	SS1	Sokal-Sneath (1)	$s_{SS1} = \frac{a}{a+2b+2c}$	0	1	A	M
13	SS2	Sokal-Sneath (2)	$s_{SS2} = \frac{2a+2d}{p+a+d}$	0	1	S	N

No	Label	Name	Equation	Scaling parameters		Concordance symmetry	Metricity
				α	β		
14	Ja	Jaccard	$s_{Ja} = \frac{3a}{3a+b+c}$	0	1	A	N
15	Fai	Faith	$s_{Fai} = \frac{a+0.5d}{p}$	0	1	I	M
16	Mou	Mountford	$s_{Mou} = \frac{2a}{ab+ac+2bc}$	0	2	A	M
17	Mic	Michael	$s_{Mic} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	1	2	Q	N
18	RG	Rogot-Goldberg	$s_{RG} = \frac{a}{2a+b+c} + \frac{d}{2d+b+c}$	0	1	S	M
19	HD	Hawkins-Dotson	$s_{HD} = \frac{1}{2} \left(\frac{a}{a+b+c} + \frac{d}{b+c+d} \right)$	0	1	S	M
20	Yu1	Yule (1)	$s_{Yu1} = \frac{ad-bc}{ad+bc}$	1	2	Q	N
21	Yu2	Yule (2)	$s_{Yu2} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	1	2	Q	M
22	Fos	Fossum	$s_{Fos} = \frac{p(a-0.5)^2}{(a+b)(a+c)}$	1	$(p-0.5^2)/p$	A	M
23	Den	Dennis	$s_{Den} = \frac{ad-bc}{\sqrt{p(a+b)(a+c)}}$	$(p/2)^{1/2}$	$p^{1/2}$	Q	M
24	Co1	Cole (1)	$s_{Co1} = \frac{ad-bc}{(a+c)(c+d)}$	$p-1$	p	Q	N
25	Co2	Cole (2)	$s_{Co2} = \frac{ad-bc}{(a+b)(b+d)}$	$p-1$	p	Q	N
26	dis	Dispersion	$s_{dis} = \frac{ad-bc}{p^2}$	1/4	1/2	Q	N
27	GK	Goodman-Kruskal	$s_{GK} = \frac{2 \min(a,d) - b - c}{2 \min(a,d) + b + c}$	1	2	S	N

No	Label	Name	Equation	Scaling parameters		Concordance symmetry	Metricity
				α	β		
28	SS3	Sokal-Sneath (3)	$s_{SS3} = \frac{1}{4} \left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d} \right)$	0	1	S	M
29	SS4	Sokal-Sneath (4)	$s_{SS4} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	0	1	S	M
30	Phi	Pearson-Heron colligation coefficient	$s_{Phi} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$	1	2	Q	M
31	Di1	Dice (1)	$s_{Di1} = \frac{a}{a+b}$	0	1	A	N
32	Di2	Dice (2)	$s_{Di2} = \frac{a}{a+c}$	0	1	A	N
33	Sor	Sorgenfrei	$s_{Sor} = \frac{a^2}{(a+b)(a+c)}$	0	1	A	N
34	Coh	Cohen	$s_{Coh} = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	1	2	Q	N
35	Pe1	Peirce (1)	$s_{Pe1} = \frac{ad-bc}{(a+b)(c+d)}$	1	2	Q	N
36	Pe2	Peirce (2)	$s_{Pe2} = \frac{ad-bc}{(a+c)(b+d)}$	1	2	Q	N
37	MP	Maxwell-Pilliner	$s_{MP} = \frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)}$	1	2	Q	M
38	HL	Harris-Lahey	$s_{HL} = \frac{a(2d+b+c)}{2(a+b+c)} + \frac{d(2a+b+c)}{2(b+c+d)}$	0	p	S	N
39	CT1	Consoni-Todeschini (1)	$s_{CT1} = \frac{\ln(1+a+d)}{\ln(1+p)}$	0	1	S	M
40	CT2	Consoni-Todeschini (2)	$s_{CT2} = \frac{\ln(1+p) - \ln(1+b+c)}{\ln(1+p)}$	0	1	S	N

No	Label	Name	Equation	Scaling parameters		Concordance symmetry	Metricity
				α	β		
41	CT3	Consoni-Todeschini (3)	$s_{CT3} = \frac{\ln(1+a)}{\ln(1+p)}$	0	1	A	N
42	CT4	Consoni-Todeschini (4)	$s_{CT4} = \frac{\ln(1+a)}{\ln(1+a+b+c)}$	0	1	A	N
43	CT5	Consoni-Todeschini (5)	$s_{CT5} = \frac{\ln(1+ad) - \ln(1+bc)}{\ln(1+p^2/4)}$	0	1	S	M
44	AC	Austin-Colwell	$s_{AC} = \frac{2}{\pi} \arcsin \sqrt{\frac{a+d}{p}}$	0	1	S	M

Performance metrics for binary classifications

Table A2. Local performance metrics for 2-class classification—One-sided.

Name	Alternative Names	Formula	Complementary Metric	Complementary Metric Formula
True positive rate (TPR)	Sensitivity, recall, hit rate	$TPR = \frac{TP}{P} = \frac{TP}{TP+FN} = 1 - FNR$	False negative rate (FNR), miss rate	$FNR = \frac{FN}{P} = \frac{FN}{TP+FN} = 1 - TPR$
True negative rate (TNR)	Specificity, selectivity	$TNR = \frac{TN}{N} = \frac{TN}{TN+FP} = 1 - FPR$	False positive rate (FPR), fall-out	$FPR = \frac{FP}{N} = \frac{FP}{TN+FP} = 1 - TNR$
Positive predictive value (PPV)	precision	$PPV = \frac{TP}{TP+FP} = 1 - FDR$	False discovery rate (FDR)	$FDR = \frac{FP}{TP+FP} = 1 - PPV$
Negative predictive value (NPV)		$NPV = \frac{TN}{TN+FN} = 1 - FOR$	False omission rate (FOR)	$FOR = \frac{FN}{TN+FN} = 1 - NPV$

Table A3. Local performance metrics for 2-class classification—Two-sided. (n : total number of samples, k : total number of classes).

Name	Formula	Description
Accuracy (ACC), or Correct classification rate (CC)	$ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN}$ $ACC = \frac{\text{correctly predicted}}{\text{total}}$	Readily generalized to multiple classes. Complementary metric: misclassification rate (or zero-one loss, or Hamming loss).
Balanced accuracy (BACC)	$BACC = \frac{TPR+TNR}{2}$ $BACC = \frac{\sum_{j=1}^k \frac{n_{j,corr.}}{n_{j,actual}}}{k}$	Alternative of accuracy for imbalanced datasets. Readily generalized to multiple (k) classes. $n_{j,corr.}$: number of samples correctly predicted into class j $n_{j,actual}$: actual number of samples in class j
F1 score (F1), or F measure	$F = 2 \times \frac{PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Harmonic mean of precision and recall
Matthews correlation coefficient (MCC) [24], ϕ coefficient (Pearson) [25]	$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$ $MCC = \frac{n_{correct} \times n - \sum_{j=1}^k n_{j,pred.} \times n_{j,actual}}{\sqrt{(n^2 - \sum_{j=1}^k n_{j,pred.}^2) \times (n^2 - \sum_{j=1}^k n_{j,actual}^2)}}$	Readily generalized to multiple classes. $n_{j,pred.}$: number of samples predicted into class j $n_{j,actual}$: actual number of samples in class j $n_{correct}$: total no. of correctly predicted samples n : total no. of samples
Bookmaker informedness (BM), or Informedness [26]	$BM = TPR + TNR - 1$	
Markedness (MK) [26]	$MK = PPV + NPV - 1$	
Positive likelihood ratio (LR+)	$LR+ = \frac{TPR}{FPR}$	
Negative likelihood ratio (LR-)	$LR- = \frac{FNR}{TNR}$	
Diagnostic odds ratio (DOR)	$DOR = \frac{LR+}{LR-}$	

Enrichment factor (EF)	$EF_{x\%} = \frac{\frac{TP}{PP}}{\frac{P}{P+N}}$	Ratio of true positives in the top x% of the predictions, divided by ratio of positives in the whole dataset.
ROC enrichment (ROC_EF) [27]	$ROC_EF_{x\%} = \frac{TPR}{FPR_{x\%}} = \frac{TPR}{x}$	Ratio of TPR and FPR at a fixed FPR value (x). Independent of dataset composition.
Cohen's kappa [28]	$\kappa = \frac{ACC - baseline}{1 - baseline}$ $baseline = \frac{(PP \times P) + (PN \times N)}{(P+N)^2}$ $baseline = \sum_{j=1}^k \frac{n_{j,pred.} \times n_{j,actual}}{n^2}$	<p>Readily generalized to multiple classes. baseline corresponds to the random agreement probability.</p> <p>$n_{j,pred.}$: number of samples predicted into class j $n_{j,actual}$: actual number of samples in class j n: total no. of samples</p>
Jaccard score (J)	$J = \frac{TP}{TP+FN+FP} = \frac{TP}{P+FP}$	Jaccard-Tanimoto similarity between the sets of predicted and actual (true) labels for the complete set of samples.
Brier score loss (B)	$B = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k (f_{i,j} - o_{i,j})^2$	<p>Readily generalized to multiple classes. $f_{i,j}$ is the predicted probability of sample i belonging to class j, while $o_{i,j}$ is the actual outcome (0 or 1). Requires predicted probability values for each class. The smaller the better.</p>
Robust initial enhancement (RIE) [29]	$RIE = \frac{\sum_{t=1}^P e^{-\alpha r_t/n}}{\sum_{t=1}^P e^{-\alpha r_t/n_r}}$	<p>r_i is the rank of positive sample i in the ordered list of samples and α is a parameter that defines the exponential weight.</p> <p>The denominator corresponds to the average sum of the exponential when P positives are uniformly distributed in the ordered list containing n samples.</p>

All square bracketed references can be found in the original publication: Molecules 2019, 24, 2811; <https://doi.org/10.3390/molecules24152811>

Table A4. Global performance metrics for 2-class classification.

Name	Formula	Description
Area under the ROC curve (AUC) [30]	Area under the TPR-FPR curve	Probability that a randomly selected positive sample will be ranked before a randomly selected negative.
Area under the accumulation curve (AUAC)	Area under the TPR-score (or TPR-rank) curve	If the ranks are normalized, then $0 \leq \text{AUAC} \leq 1$ Probability that a randomly selected positive will be ranked before a randomly selected sample from a uniform distribution.
Average precision (AP)	Area under the precision-recall (PPV-TPR) curve	
Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) [31]	$\text{BEDROC} = \frac{RIE - RIE_{\text{min}}}{RIE_{\text{max}} - RIE_{\text{min}}}$	See the definition of RIE above, α is a parameter that defines the exponential weight. $0 \leq \text{BEDROC} \leq 1$ BEDROC is an analog of AUC that assigns an (exponentially) greater weight to high-ranked samples, thus tackling the "early recognition problem".
Average rank (position) of actives (positives) (r) [32]	$r = \frac{1}{P \times (P+N)} \sum_{i=1}^P r_i$	r_i is the rank of positive sample i in the total ranked list of samples. The smaller the better.

All square bracketed references can be found in the original publication: Molecules 2019, **24**, 2811; <https://doi.org/10.3390/molecules24152811>

