Estimation of Random Accuracy and its Use in Validation of Predictive Quality of Classification Models within Predictive Challenges

Lučić, Bono; Batista, Jadranko; Bojović, Viktor; Lovrić, Mario; Sović Kržić, Ana; Bešlo, Drago; Nadramija, Damir; Vikić-Topić, Dražen

Source / Izvornik: **Croatica Chemica Acta, 2019, 92, 379 - 391**

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

https://doi.org/10.5562/cca3551

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:151:237629

ORIGINAL SCIENTIFIC PAPER

CROATICA CHEMICA ACTA

Croat. Chem. Acta **2019**, *92*(3), 379–391
Published online: October 21, 2019
DOI: 10.5562/cca3551

HKD 1926

Estimation of Random Accuracy and its Use in Validation of Predictive Quality of Classification Models within Predictive Challenges

Bono Lučić,[1,*] Jadranko Batista,[2] Viktor Bojović,[1] Mario Lovrić,[1,3,4] Ana Sović Kržić,[5] Drago Bešlo,[6] Damir Nadramija,[1,7] Dražen Vikić-Topić[1,8]

[1] NMR Centre, Ruđer Bošković Institute, P.O. Box 180, HR-10002 Zagreb, Croatia
[2] Faculty of Science and Education, University of Mostar, Matice hrvatske b.b., BA-88000 Mostar, Bosnia and Herzegovina
[3] Srebrnjak Children's Hospital, Srebrnjak 100, HR-10000 Zagreb, Croatia
[4] Know-Center, Inffeldgasse 13, AT-8010 Graz, Austria
[5] Department of Electronic Systems and Informational Processing, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia
[6] Faculty of Agrobiotechnical Sciences Osijek, Josip Juraj Strossmayer University of Osijek, Vladimira Preloga 1, HR-31000 Osijek, Croatia
[7] PharmaS, Radnička cesta 47, Zagreb, Croatia
[8] Department of Natural and Health Sciences, Juraj Dobrila University of Pula, Zagrebačka 30, HR-52100 Pula, Croatia
* Corresponding author's e-mail address: lucic@irb.hr

RECEIVED: July 12, 2019 ✶ REVISED: September 20, 2019 ✶ ACCEPTED: September 20, 2019

Abstract: Shortcomings of the correlation coefficient (Pearson's) as a measure for estimating and calculating the accuracy of predictive model properties are analysed. Here we discuss two such cases that can often occur in the application of the model in predicting properties of a new external set of compounds. The first problem in using the correlation coefficient is its insensitivity to the systemic error that must be expected in predicting properties of a novel external set of compounds, which is not a random sample selected from the training set. The second problem is that an external set can be arbitrarily large or small and have an arbitrary and uneven distribution of the measured value of the target variable, whose values are not known in advance. In these conditions, the correlation coefficient can be an overoptimistic measure of agreement of predicted values with the corresponding experimental values and can lead to a highly optimistic conclusion about the predictive ability of the model. Due to these shortcomings of the correlation coefficient, the use of standard error (root-mean-square-error) of prediction is suggested as a better quality measure of predictive capabilities of a model. In the case of classification models, the use of the difference between the real accuracy and the most probable random accuracy of the model shows very good characteristics in ranking different models according to predictive quality, having at the same time an obvious interpretation.

Keywords: model validation, QSPR, QSAR, two-class variable, classification model, contingency table, estimation, prediction, test set, correlation coefficient, predictive error, classification accuracy, model ranking, random accuracy.

INTRODUCTION

COMPUTER modelling has intensively been used in many areas, including chemistry and life science. As a rule, the main goal of the modelling is to extract useful information or regularities in a form of functional relationship(s) between subsets of data, between one variable and a subset of variables, or just between pairs of variables. In chemistry and life science, the modelling of relationships between a property y_i, (i = 1, …, N) of N compounds represented by a set of M structure-based descriptors $x_{i,j}$, is very often used in the analysis of different problems. That kind of models is known under the acronym QSPR/QSAR (quantitative structure-property/activity relationship).[1]

For validation of model quality, from the beginnings of modern QSPR models correlation coefficient R and root-mean-square error S have been usually calculated from experimental values y_i and values obtained in fitting $\hat{y}_i$.[1] Also, the quality of correlation between topological (graph-theoretical) descriptors and the most

important physico-chemical properties of a molecule[2] was analysed in this journal, and the ranges of correlation coefficients of QSPR models on a set of octane isomers were defined.[3]

The quality of the model is first validated by internal validation procedures on the training set in the fit, or in the Leave-One-Out (LOO) or Leave-*k*-out (L*k*O) cross-validation (CV) procedures.[4] Within CV procedure, *k* (L*k*O) compounds are omitted in each step from the total number of *M* compounds in the training set, and properties of omitted compounds are estimated by the developed model. Such a procedure is repeated as long as the property value of each compound from the training set is estimated by the LOO procedure just once. Later, it has been recommended to define the eligibility of QSPR models. The main recommendation was based on the square of correlation coefficient achieved in the LOO CV process, which must be higher than 0.5 for the high-quality (predictive) model.[5] Although it was introduced in Ref. [5] after the analysis of only one type of model obtained by the k-nearest neighbour algorithm, and although that recommendation was not derived from the strict mathematical analysis but rather from simulations on selected data sets, it is often used and quoted in the scientific literature. However, the importance of testing model's quality on external data set has been pointed out by Tropsha et al.[6] Probably, the main reason why such a practice has not been introduced earlier was the lack of large enough sets of data at that time. Even earlier, other authors have been noticed that CV is not always reliable in estimating the model's quality in predicting properties for new chemical compounds (i.e. for new, never seen examples - an external data set) that were not used in model training and optimisation (e.g. Refs. [7–10]). Thus, validations on external sets have been used in a comparative study between multivariate and neural network structure-property models,[7] in development of models for modelling viscosities of 361 organic compounds (240 in the training and 121 in the test set),[8] and in modelling secondary structure contents (alpha, beta and irregular) in a set of 475 soluble proteins (317 in the training and 158 in the test set).[9]

To estimate or measure the quality of a model, in addition to the statistical parameters calculated in the fit and CV procedure, i.e. the correlation coefficient (R and R_{cv}) or the standard error or estimate (S and S_{cv}) calculated between the estimated and experimental property values, corresponding parameters are introduced and calculated between experimental and predicted values on the test set (R_{pred} and S_{pred}). Also, many other parameters have been calculated and used for model validation in the field of QSAR/QSPR modelling. These parameters are just the basic set calculated and used for estimation of the quality of (almost) all QSPR models.

The afore mentioned QSPR / QSAR methodology and model validation parameters and procedures were accepted for the regulatory purposes for the evaluation and prediction of molecular properties by OECD countries.[10] Later, these recommendations were included in REACH, the EU document regulating the Registration, Evaluation, Authorisation and Restriction of chemical compounds in the European Union.[11,12]

Today, the field of modelling in chemistry and biology is evolving due to the increasing amount of data for many compounds/structures and their measured properties, activities and interactions with other small molecules or macromolecules. These studies are very interesting and important to wider community because they are linked to drug design and environmental research - two issues of global importance. Due to the propulsion of the area and the availability of large databases, researchers from different areas such as scientific computing, machine learning or deep learning, etc. are also intensively involved in modelling. In order to accelerate the exchange of ideas and to estimate/summarise the predictive potential of available (constantly evolving) computer algorithms and procedures in modelling of different chemical and biological problems, predictive modelling challenges have been organised in predicting molecular properties and activities for a new (external) set of compounds (cases, instances). One such consortium (DREAM[13]) has a long experience in organization of different challenges since 2006 and is an interdisciplinary team composed of researchers from biotechnological, pharmaceutical and technological companies.[13,14] During challenges related to chemistry or biology (drug design), data sets containing both structural descriptors (attributes) and experimental activities for training and model development are first given to all participant groups. Also, an additional test set is given to competitors without experimental activities, which should be predicted by the developed model. Evaluation of model quality is estimated by an independent team of scientists, according to pre-defined statistical parameters. Pearson's correlation coefficient and standard error of estimate/prediction were usually used in the evaluation of prediction quality for prediction of continuous properties or activities.[15–17] In a case of classification problems with two classes *A* and *B*, F_1score was used as the main statistical parameter for ranking the quality of models.

Also, we will analyse the suitability of the use of correlation coefficient in estimations of quality of models on an external (test) set by analysis of predictive potential of structure-solubility models on the test set containing 258 organic compounds. All these models are developed on the training data set having 1039 organic compounds.[18] Furthermore, the non-sensitivity of the correlation coefficient to the constant shift of predictions will be analysed on

the training set in cross-validation and on the test set in prediction. Such a characteristic disqualifies the correlation coefficient for its use in ranking modelling methods according to their predictive accuracy on an external test set. In prediction of classification properties (e.g. whether test set compounds are active or inactive), we will apply the recently introduced parameter ΔQ_2 for estimation the difference between real and most probable random accuracy,[19] and compare its properties with other parameters regularly used in evaluation of accuracy of classification models in predictive challenges.[20–22] It comes out that the parameter ΔQ_2 has very good properties in estimating the quality of predictions done by different models. Namely, ΔQ_2 ranks as better models those having higher value of correct predictions of both classes (designated as 1 and 0) and, at the same time, more balanced total values of errors (under- and over-prediction, which are two types of errors defined in analyses of accuracy of classification models meaning the total number of cases when class 1 is predicted as 0, and the total number of cases when class 0 is predicted as class 1, respectively[19]).

METHODS

Mathematical equations for calculation of statistical parameters, whose characteristics and suitability for estimating predictive accuracy of models on external (test) are analysed in this study, are given here. Additionally, data sets used in simulations and in the comparative analysis are also described.

Estimating the Accuracy of Prediction of Continuous Properties

For estimating the quality of models for prediction of continuous properties in predictive challenges,[15–17] a well-known Pearson's correlation coefficient R was used as the main parameter. It is calculated by Eq. (1):

$$R = \frac{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)\left(\hat{y}_i - \overline{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2 \sum_{i=1}^{N}\left(\hat{y}_i - \overline{\hat{y}}\right)^2}} \quad (1)$$

Here, y_i and $\hat{y}_i$ are experimental property values and values predicted by the model, respectively, $\overline{y}$ is the mean of experimental property values and, finally, $\overline{\hat{y}}$ is the mean of property values predicted by the model in prediction on an external test set.

Another parameter that has been regularly used for estimating the predictive accuracy of models is the standard error (root-mean-square-error) of prediction, calculated by Eq. (2):

$$S = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N}} \quad (2)$$

where y_i and $\hat{y}_i$ are described below Eq. (1), and N is the total number of molecules (cases) in the test set. These parameters can be also calculated between experimental values and those estimated by the model on the training set in fitting and CV procedure.

Estimation of Accuracy of a two-state Classification Models

The F_1score (Eq. (3)) has been used in predictive challenges[20–22] for estimation of predictive model accuracy and for ranking classification models according to its values. A higher value of F_1*score* means that the model is more accurate in prediction.

$$F_1\text{score} = \frac{2 \cdot \text{prec} \cdot \text{rec}}{\left(\text{prec} + \text{rec}\right)} \quad (3)$$

This parameter is defined as a harmonic mean of precision (Eq. (4)) and recall (Eq. (5)), and is primarily used for estimating the quality of models developed on (highly) dis-imbalanced data sets.

In Eq. (3), precision is defined by Eq. (4) and recall by Eq. (5):

$$\text{prec} = \text{precision} = \frac{p}{\left(p + u\right)} \quad (4)$$

$$\text{rec} = \text{recall} = \frac{p}{\left(p + o\right)} \quad (5)$$

where p = TP (true positive) is the total number of positive correct predictions of class A (observed class A is correctly predicted by the model to be class A), u = FN (false negative) is under-prediction of class A (experimental class A predicted to be class B) and o = FP (false positive) is over-prediction (class A predicted to be class B).

By putting Eq. (4) and (5) into Eq. (3) and after some simplifications, F_1score can be simply expressed by (Eq. (6)) using only p, u and o:

$$F_1 score = \frac{2p}{2p + u + o} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FN} + \text{FP}} \quad (6)$$

It is useful to compare this equation for estimation of the F_1score with a well-known and often used parameter

named accuracy, classification accuracy,[19] or the percentage of all correct prediction[23] (Eq. (7)):

$$Q_2 = \frac{p+n}{p+n+u+o} = \frac{p+n}{N} \tag{7}$$

where n = TN (true negative) is the total number of negative class correctly predicted as negative, and p, u and o have the same meaning as in Eqs. (4–6). These numbers (p, n, u, o) are elements of the contingency table[19] (called also confusion matrix) that is defined for each classification problem. Parameter Q_2 has been used for balanced classification problems, i.e. those having similar number of positive and negative cases. The F_1*score* has been considered as more appropriate for imbalanced classification problems, in which one class is largely predominant (major *vs* minor class).[15–17] It is interesting to note that one can obtain Eq. (6) from Eq. (7) just by putting $n = p$ in Eq. (7). However, because F_1score is primarily used on imbalanced data sets (where $p << n$, but not $p = n$) such an analogy made by this 'arbitrary' substitution indicates that the comparison of these two parameters is very complicated (or even impossible). Though the interpretation of Q_2 is very simple, the interpretation of F_1score (Eq. (6)) is neither straightforward nor comparable with the interpretation of Q_2.

According to Eq. (6), it seems like the F_1score is not dependent on n. For the fixed set of p, u, and o values, F_1score will have the same value for any n. This is a weak characteristic of the F_1*score* indicating its insensitivity to the size of data set. If we re-write the denominator in Eq. (6) taking into account that $N = p + n + o + u$, and $p + o + u = N - n$ (N is the total number of cases), then we get the following:

$$F_1\text{score} = \frac{2p}{N + p - n} \tag{8}$$

Thus, Eqs. (7) and (8) show that both Q_2 and F_1score can be calculated from the same (three) numbers: p, n, and N.

In case of binary (two-class) classification models, the correlation coefficient between observed predicted variable (Eq. (1)) can be expressed by the elements of the contingency table as:

$$R = \text{Mcc} = \frac{pn - uo}{\sqrt{(p+u)(p+o)(n+u)(n+o)}} \tag{9}$$

This well-known form of a correlation coefficient for estimating the correlation between the observed and the estimated (predicted) two-class target variable is named Matthew's correlation coefficient (Mcc).[24] Although it has also some limitations (like that it is not possible to calculate its values if only one class is predicted or estimated),[25] Mcc is very often used for estimating the accuracy of the model, and as the quality parameter for ranking different models developed on the same data set. Equation (Eq. (2)) for calculation of standard error of estimation/prediction of a binary classification model can also be expressed by the values of under- and over-estimation/prediction o and u (Eq. (10)):

$$S = \sqrt{\frac{u+o}{N}} \tag{10}$$

Estimating the Random Correlation of a two-state Classification Model

It is shown in Ref. [19] that if a binary classification model predicts (p + o) cases to be in class A for (p + u) experimentally determined cases of class B, and (n + u) = N – (p + o) cases to be in class B for (n + o) experimentally determined cases in class B, then the most probable random accuracy $Q_{2,\text{rnd}}$ can be estimated as:

$$Q_{2,\text{rnd}} = \frac{(p+u)(p+o)+(n+o)(n+u)}{N^2}. \tag{11}$$

This value of $Q_{2,\text{rnd}}$ is always between values corresponding to minimal accuracy, i.e. maximal disagreement, and maximal accuracy, i.e. maximal agreement. The maximal range of $Q_{2,\text{rnd}}$ values is between 0 and 1. Both Q_2 and $Q_{2,\text{rnd}}$ can be expressed in percentages. Additionally, the difference (in %) between the real model accuracy Q_2 obtained by a model and the corresponding most probable random accuracy $Q_{2,\text{rnd}}$ (Eq. (11)) can be simply calculated (Eq. (12)):

$$\Delta Q_2 = 100\left(Q_2 - Q_{2,\text{rnd}}\right)(\%). \tag{12}$$

This value can have a maximum of $(\Delta Q_2)_{\text{max}}$ = 50 % in these two cases:

(1) totally equal numbers of elements in both classes (50 : 50 %) in data set, and

(2) perfect model estimation or prediction (u = o = 0).

Thus, ΔQ_2 can be considered as a measure of the contribution of a model to real accuracy of estimation or prediction over the most probable random accuracy level. In analysis of mutual quality of different classification models, ΔQ_2 parameter can serve for models' ranking. The

higher value of parameter ΔQ_2 means that the model contributes a larger amount of useful information over the maximal level of random accuracy, which is a clear interpretation.

Normally, an appropriately optimised model (named *balanced model* in Ref. [19]) estimates (or predicts) the same numbers of states/classes as in the experimental structure (i.e. $p + u = p + o$ and $n + o = n + u$), then $Q_{2,\mathrm{rnd}}$ from Eq. (11) becomes:

$$Q_{2,\mathrm{rnd}} = Q_{2,\mathrm{rnd\text{-}bal}} = \frac{(p+u)^2 + (n+o)^2}{N^2}. \qquad (13)$$

Equation (13) enables the estimation of the most probable random accuracy for balanced model and, in that case, $Q_{2,\mathrm{rnd}}$ can be calculated only using the experimental number of cases in the first class ($p + u$), because, for the second class, we have $n + o = N - (p + u)$. Whenever the value of $Q_{2,\mathrm{rnd}}$ calculated by Eq. (11) is different (i.e. higher) from the one calculated by Eq. (13), it is an indication of the lack of model training process.

Data Sets

Analysis of modelling and prediction of properties having continuous values was performed on aqueous solubility data of 1297 organic compounds.[26] The solubility data set is composed of two aqueous solubility databases AQUASOL[27] and PHYSPROP,[28] and it was partitioned as it was done by Liu and So.[18] Namely, the training set contains 1039 compounds, and the test set used for estimation of external prediction has (remaining) 258 compounds (Table S1). The set of 123 descriptors used in the last stage for selection of the best models are also given in Table S1, and the analysis of statistical parameters in Table 1. More details on the developed models are given in Tables S2 and S3. The weakness of correlation coefficient connected to the distribution of data has been illustrated on simulated data set having three pairs of variables among which (in each pair) the first represents experimental and the second one predicted variable (Table 2).

For analysis related to modelling of classification variables, we used three data sets dealing with a two-class problem. The first one is the data set used in the final phase of the Tumour prediction challenge[22] organised to develop algorithms and models for detection of somatic mutations from cancer genome sequences in order to understand the genetic basis of disease progression. This data set is taken from Cooper *et al*. (Additional file 9, Table S8 in Ref. [20]). It contains prediction results for 70 models in the final phase of predictive challenge (IS3). Among them, prediction performances of 15 top scoring models is given in Table 3, and details of the remaining 55 models are included in Table S4. The data set contains 24687 prediction cases among which 7903 (32 %) is of positive, and the remaining 16784 (68 %) of negative class. Because this data set is imbalanced, F_1score was used as the main scoring (methods' ranking) criteria.

The second classification data set contains six special cases of contingency tables of extremely imbalanced class distribution (5 : 9995) from the critical overview of evaluation metrics applicable for analysis of classification problems.[29]

Additional data sets are composed of:

(1) Four examples of contingency table values from Tables 4 and 5 (special cases) in Ref. [29] having a highly imbalanced class distribution (5 : 95);
(2) Two imbalanced examples (95 : 5 and 94 : 6) of contingency table values given and analysed in Ref. [30] and also in Ref. [31] in order to illustrate the drawbacks of F_1score comparing to Mcc; and
(3) Eight examples constructed in this study in close analogy with examples from literature[29–31] mentioned above in (1) and (2).

To calculate and compare different validation parameters one has to have only 2 x 2 contingency tables for each model estimate/prediction containing *p*, *n*, *o* and *u* values. Some of the data sets are artificial ones containing *p*, *n*, *u* and *o* values selected in a specific way in order to check the values of the corresponding validation parameter in that special cases. The largest set of contingency tables is from modelling done within final phase of the Tumour prediction challenge,[20,22] where many groups developed different classification models; however, descriptions of the models are not given in sufficient details. Comparative analysis of the usefulness and informativeness of the validation parameters presented in this paper on classification problems is completely independent on the algorithms/methods used for model development. For more details on computational methods used in modelling classification problems, interested readers can consult recent literature.[32]

RESULTS AND DISCUSSION

By using four data sets we will illustrate some important problems which can arise from the application of commonly used statistical parameters like Pearson's correlation coefficient (R) in application to validation of prediction accuracy of continuous and two-class problems. Moreover, the limitation of two-class accuracy measures like Q_2, Mcc or F_1score (given by Eqs. (7–9), respectively), and the advance of the use of novel parameter ΔQ_2 given by Eq. (12) in ranking models will be analysed.

We want to point out here the distinction between internal (fit and LOO CV) and external validation

Table 1. Basic statistical accuracy parameters of structure-solubility QSPR models based on Multivariate Linear Regression (MLR) having 1 – 5 most significant descriptors.[a]

N	*I*	*S'*	*S*	*S* – *S'*	*R*	*c* (mean error/difference)
				fitting		
1039	1	1.2265371255	1.2265371264	$9{\cdot}10^{-10}$	0.798	$5{\cdot}10^{-05}$
1039	2	0.9629603589	0.9629603593	$4{\cdot}10^{-10}$	0.881	$-3{\cdot}10^{-05}$
1039	3	0.8812569638	0.8812569639	$3{\cdot}10^{-11}$	0.901	$8{\cdot}10^{-06}$
1039	4	0.8087179434	0.8087179436	$2{\cdot}10^{-10}$	0.918	$-2{\cdot}10^{-05}$
1039	5	0.7628136584	0.7628136585	$6{\cdot}10^{-11}$	0.927	$-1{\cdot}10^{-05}$
				leave-one-out cross-validation		
1039	1	1.2300027	1.2300031	$3{\cdot}10^{-7}$	0.797	$-9{\cdot}10^{-04}$
1039	2	0.9667658	0.9667659	$1{\cdot}10^{-7}$	0.880	$-5{\cdot}10^{-04}$
1039	3	0.8852920	0.8852923	$3{\cdot}10^{-7}$	0.900	$-7{\cdot}10^{-04}$
1039	4	0.8136367	0.8136370	$3{\cdot}10^{-7}$	0.917	$-7{\cdot}10^{-04}$
1039	5	0.7681544	0.7681547	$2{\cdot}10^{-7}$	0.926	$-6{\cdot}10^{-04}$
				prediction on external test set		
258	1	1.2083	1.2123	$4{\cdot}10^{-3}$	0.800	$-1{\cdot}10^{-01}$
258	2	0.9514	0.9536	$2{\cdot}10^{-3}$	0.881	$-6{\cdot}10^{-02}$
258	3	0.8792	0.8804	$1{\cdot}10^{-3}$	0.900	$-5{\cdot}10^{-02}$
258	4	0.8337	0.8344	$8{\cdot}10^{-4}$	0.911	$-4{\cdot}10^{-02}$
258	5	0.8257	0.8261	$3{\cdot}10^{-4}$	0.912	$-2{\cdot}10^{-02}$

[a] *N* is the number of compounds in data set; *I* is the number of descriptors in the model; *S'* is the standard error of estimate calculated (for each model) as the mean deviation of each error from the mean error value (Eq. (14)); *S* is the (normal) standard error of estimate calculated by Eq. (2); *R* is the correlation coefficient (Eq. (1)); *c* is the constant shift defined by Eq. (15). To be able to notice the variation of corresponding statistical parameters, their values are given (as a rule) to the last two digits that differ.

procedures. Consequently, we want to delineate these two procedures by using different more precise terms, in order to avoid possible confusion in tracking results, as well as to help in understanding the main matter and points of the presented research. Thus, the term 'estimate' in this study corresponds to internal validation procedure, i.e. to the calculation of property/activity values by a model on the training data set, which is used for model development and optimisation. However, the term 'prediction' is used in this study in case of pure prediction, i.e. when the model is used for calculation of property/activity values on an external data set, which is not used for model development and optimisation.

Problems in the Evaluation of Prediction Accuracy of Continuous Properties Using the Correlation Coefficient

Using the algorithm for selection of the best subset of descriptors into the Multivariate Linear Regression (MLR) models,[33] we selected the best QSPR for modelling water solubility of organic compounds. Developed models given in Table 1 are based on molecular descriptors selected from 123 descriptors (which are pre-selected from the initial pool containing more than one thousand descriptors) calculated by the Dragon 5.4 program.[34] For a selected set of descriptors, model parameters are optimised using the MLR methodology by the least square fitting procedure, which ensures that developed model will have the lowest standard error of estimate in fitting among all other possible linear models (which could be obtained by the application of other fitting procedures).

To save space, we give in the supplementary Table S1 of the manuscript details on the models from Table 1, like the model equation or details on molecular descriptors involved, because it is not the main subject of this study. The main statistical parameters (correlation coefficients and standard errors of estimate) for the best selected QSPR models are given in Table 1. These models are developed in fitting and internally validated by LOO CV procedure on 1039 compounds from the training, and they are also externally validated on 258 compounds from the test set.

Correlation coefficient given by Eq. (1) can be considered as a measure of linear agreement between two sets of paired data (two variables), and is not sensitive to the constant shift of values of variables considered. In modelling, values of experimental variable y_i are fixed, and only estimated/predicted values $\hat{y}_i$ can have a constant shift. Regularly, experimental and the corresponding values

estimated by the fit and CV procedure on the training set can be in a stronger linear relationship because model parameters were optimised on these experimental data. However, experimental data whose numerical values are unknown to modeller(s) and the corresponding values from the test set predicted by the model do not have to be in a significant linear relationship. If so, Pearson's correlation coefficient given by Eq. (1) is not a good or acceptable measure of agreement between experimental and predicted values.

The standard error of estimate or prediction in Eq. (2) is calculated using the difference (error, deviation) of each estimated/predicted value $\hat{y}_i$ from the corresponding experimental value y_i. However, if predicted values have some constant shift (c) in one or another direction, we can calculate a modified version of the standard error of estimate (S') by Eq. (14):

$$S' = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i - c\right)^2}{N}} \tag{14}$$

The constant shift c for each data set (in fitting and LOO CV on the training set, and in prediction on the test set) is simply obtained as the mean of all differences (errors) by Eq. (15):

$$c = \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)}{N} \tag{15}$$

Parameter S' ($\leq S$) is calculated to see what would be the value of the standard error in case of an ideally optimised model, which introduce in prediction neither constant over-estimation ($c > 0$), nor under-estimation ($c < 0$). Moreover, by considering the differences $S - S'$ for fit, and LOO CV one can see that they are for a factor of 10^3 larger going from the fit to LOO CV, and for a factor 10^4 larger going from LOO CV to prediction (Table 1).

Linear models having one to five descriptors from Table 1 were developed on a large training set of 1039 compounds and have (only) two to six optimised parameters, i.e., 1037 to 1033 degrees of freedom. Thus, these models are far from the over-fitting regime. And even with such models, we notice an increase in constant shifts (though they are very small) going from the fit to LOO CV on the training set, and to prediction on the test set of 258 molecules. By comparing the constant shift values of each model with the same number of descriptors starting from the fit procedure to LOO CV the increase is for a factor of 10, and the increase from LOO CV to prediction is for a factor of 100. Since this can happen with such simple models with a very small number of optimised parameters, we can assume that a constant shift will be also present (being much larger) in the case of more complex and nonlinear models. Evidently, we cannot eliminate it, and hence the largest constant shift is in prediction, being more than 1000 times larger than in the fitting estimate. It is worth mentioning here that the LOO CV procedure done for the models from Table 1 is just the stability test performed to see the difference between the fit statistical parameters and the corresponding ones obtained by the LOO CV procedure. Namely, the model parameters are not selected either optimised by the use of LOO CV. This is not the case with robust methods based on machine learning, which are prevalent methods used in prediction challenges, and which are comprehensively optimised in several cycles of LOO or L*k*O CV procedures.

Taking this into account, the validation of models in prediction on external data set should be primarily evaluated and ranked by the parameter which is not sensitive to constant shift such as the standard error of prediction, or maybe by another variant of correlation coefficient named concordance correlation coefficient.[35]

Besides Pearson's correlation coefficient is not sensitive to the constant shift of predicted towards experimental values, it is also highly sensitive to the distribution of data in the test set. The test set can be small, and its distribution can be (generally) skewed. In such a case, good prediction of only one or two cases located at the far edge of the distribution can cause a relatively large increase of correlation coefficient. This is illustrated on data given in Table 2 with three sets, among which the second and the third set are larger for just one case. Correlation coefficients between experimental and predicted values from these three sets are –0.09, 0.56 and 0.77 for Prediction 1, 2 and 3, respectively. Correlation coefficients between experimental and predicted values for Prediction 2 and 3 are high because correlation coefficient given by Eq. (1) is highly sensitive to the distribution of data values, and insensitive to the constant shift. Obviously, the change of S values between these three predictions is much smaller than the change of R, indicating a greater stability of standard error when applied to the calculation of prediction accuracy on an external set.

It is known that the application of the least square optimisation is not optimal in such a case, and does not give optimal result.[36] Namely, another method, based on the minimisation of absolute deviation (L_1-norm) introduced in 1757 by R. Bošković,[37] seems to be a more convenient solution for data sets with outliers, what can appear quite often in prediction on new external data sets.[38]

There are many problems in chemistry or life sciences having skewed distribution, i.e. in which there are many compounds in the data set which are inactive or only weakly active, and only few of them with high or very high activity. One example of such a distribution is related to

Table 2. Three examples of predictions to illustrate the greater stability of standard error (*S*) comparing to correlation coefficient (*R*) in estimating the quality of prediction on external set.[a]

No.	Exp. 1	Prediction 1	Exp. 2	Prediction 2	Exp. 3	Prediction 3
1	2	4	2	4	2	4
2	3	8	3	8	3	8
3	4	6	4	6	4	6
4	5	7	5	7	5	7
5	6	11	6	11	6	11
6	7	3	7	3	7	3
7	8	9	8	9	8	9
8	9	10	9	10	9	10
9	10	2	10	2	10	2
10	11	5	11	5	11	5
11			15	25	25	25
R	*N* = 10	–0.09	*N* = 11	0.56	*N* = 11	0.77
S		4.24		5.04		4.04
c (shift)[b]		0.0		-1.0		0.0

[a] Prediction 1 and Prediction 2 differ only in one (the last) case. The acronym 'Exp.' is for variable containing 'experimental' values.

[b] Average constant shift calculated by Eq. (13).

activity of 100 polyphenols measured by two assays.[39,40] All antioxidant activity values of polyphenolic compounds determined by the first assay are in the range 0 – 11.6. Among them 51 least active compounds have activity values < 1.0, and 30 values are < 0.1. The reason for this is in the fact that only polyphenols having one or more catecholic OH groups can have high antioxidant activity.

In the fitting procedure, it is possible to optimise the final model to be $\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right) = 0$. However, it will not be the case in the cross-validation procedure or in prediction on an external set of data having an arbitrary distribution. This problem will be much larger in the case of nonlinear models, which are much more dependent on the distribution of the training data set due to the stronger minimisation of the total model error by the introduction of additional nonlinear terms and, consequently, additional optimised coefficients in the model. In the CV procedure, only a slight perturbation of model is randomly introduced in each step by omitting some small portion of data, or only one case (compound) in each step in LOO CV. The rest of the data samples are then used for calculation of model coefficients which differ from the coefficients of model obtained on the complete training data set in fitting procedure.

Prediction on Data Sets Containing two-class Classification Properties

In the analysis of predictive quality of two-class problems we used the data from the paper describing results of the Tumour prediction challenge.[22] Originally, all submitted models in the stage 3 were ranked according to higher value of F_1score. The values of F_1score, Q_2, Mcc and ΔQ_2 are calculated from the contingency table values of 70 submitted models. The top 15 models are given in Table 3, and details of the remaining 55 models are included in Table S4. The code-names of models as indicated in the Tumour prediction challenge[22] are given in the second column of Table 3.

The first three parameters in Table 3 are just the values of pure accuracy parameters, and they can be related just to their maximal (or minimal) value. However, the first three parameters do not consider the accuracy that can be obtained just by the random guessing. Random (guessing) accuracy is higher if data set is more monotonous, i.e. more imbalanced. However, parameter ΔQ_2 takes into account the level of random accuracy, and estimate the real model contribution to accuracy above that level. In the last four columns the ranks of models according to each of four parameters are given. These results in Table 3 are sorted by the values of ΔQ_2 parameter.

Ten best-ranked models according to each of four parameters are within 15 top-ranked models according to ΔQ_2. Additionally, 20 best-ranked models according to each of four parameters is within 21 top-ranked models. The mean absolute differences of ranks according to F_1score, Q_2, Mcc with the ΔQ_2 rank on the top 15 models are (respectively) 4.3, 5.9 and 6.3, and for the complete list of 70 models they are 2.9, 4.1 and 4.0 (Table S4).

Table 3. Ranking the models from the final phase of the Tumour prediction challenge[22] according to F_1score from Ref. [20], Q_2, Mcc, and ΔQ_2 based on their predictions on the test set (IS3).[a]

No.	ID	F_1score	Q_2	Mcc	ΔQ_2	rank_F_1score	rank_Q_2[b]	rank_Mcc	rank_ΔQ_2
1	X2463247	94.55	96.57	0.921	39.680	16	19	19	**1**
2	X2478107	94.70	96.69	0.923	39.674	10	14	14	**2**
3	X2453885	94.79	96.75	0.925	39.666	**5**	9	10	**3**
4	X2478109	94.79	96.76	0.925	39.621	**4**	8	9	**4**
5	X2473029	94.83	96.80	0.926	39.583	**1**	**2**	**2**	**5**
6	X2472860	94.81	96.78	0.926	39.579	**3**	***4***	7	6
7	X2463211	94.70	96.71	0.924	39.529	11	12	12	7
8	X2456287	94.79	96.78	0.926	39.494	6	***4***	6	8
9	X2456202	94.76	96.78	0.926	39.422	7	*6.5*	**5**	9
10	X2476556	94.82	96.82	0.927	39.418	**2**	**1**	**1**	10
11	X2468117	94.70	96.73	0.925	39.396	13	11	11	11
12	X2470044	94.76	96.78	0.926	39.395	8	***4***	**4**	12
13	X2460633	94.60	96.67	0.923	39.366	15	15	15	13
14	X2476415	94.70	96.74	0.925	39.331	12	10	8	14
15	X2476341	94.74	96.78	0.926	39.326	9	*6.5*	**3**	15

[a] Q_2 (%), F_1score, Mcc and ΔQ_2 (%) are calculated by Eqs. (7–9) and (12), respectively.

[b] Rank 4 appears three times because models between ranks 3–5 have the identical value of Q_2, and the identical rank. Analogously, rank 6.5 appears twice, because models 6 and 7 have the identical rank.

Table 4. Values of contingency tables of the best 3–5 models from Table 3 and their ranks according to each of four ranking parameters.[a]

No.	ID	p = TP	n = TN	u = FN	o = FP	r_F_1score	r_Q_2	r_Mcc	r_ΔQ_2	n/p[b]	o/u[b]
A)	top three models according to ΔQ_2 values										
1	X2463247	**7335**	**16506**	**278**	**568**	16	19	19	**1**	2.3	2.0
2	X2478107	7308	16561	223	595	10	14	14	**2**	2.3	2.7
3	X2453885	7291	16594	190	612	5	9	10	**3**	2.3	3.2
B)	top three models according to F_1score values										
1	X2473029	**7253**	**16643**	**141**	**650**	**1**	2	2	5	2.3	4.6
2	X2476556	7191	16711	73	712	**2**	1	1	10	2.3	9.8
3	X2472860	7255	16637	147	648	**3**	*4*	7	6	2.3	4.4
C)	top five models according to Q_2 values										
1	X2476556	**7191**	**16711**	**73**	**712**	2	**1**	1	10	2.3	9.8
2	X2473029	7253	16643	141	650	1	**2**	2	5	2.3	4.6
3	X2472860	7255	16637	147	648	3	***4***	7	6	2.3	4.4
4	X2456287	7226	16666	118	677	6	***4***	6	8	2.3	5.7
5	X2470044	7192	16700	84	711	8	***4***	**4**	12	2.3	8.5
D)	top three models according to Mcc values										
1	X2476556	**7191**	**16711**	**73**	**712**	2	1	**1**	10	2.3	9.8
2	X2473029	7253	16643	141	650	1	2	**2**	5	2.3	4.6
3	X2476341	7169	16722	62	734	9	*6.5*	**3**	15	2.3	11.8

[a] See footnote of Table 3 for details and explanations. In the names of ranking parameters r_F_1score, r_Q_2, r_Mcc and r_ΔQ_2, the part 'r_' means 'rank' (according to the given parameter).

[b] The ratio of true negative and true positive (n/p) and false positive and false negative (o/u) values from the left part of this table calculated for each model.

Table 5. Comparison of accuracy parameters calculated from values of contingency tables as suggested in Ref. [29] (N = 10000).[(a)]

No.	p = TP	u = FN	n = TN	o = FP	true	Q_2	F_1score	Mcc	ΔQ_2
a_1	0	5	9995	0	9995	99.95	0.00	#DIV/0!	0.00
a_2	1	4	9994	1	9995	99.95	0.29	0.32	0.02
a_3	2	3	9993	2	9995	99.95	0.44	0.45	0.04
a_4	3	2	9992	3	9995	99.95	0.55	0.55	0.06
a_5	4	1	9991	4	9995	99.95	0.62	0.63	0.08
a_6	5	0	9990	5	9995	99.95	0.67	0.71	0.10

[(a)] Examples a_1–a_6 are from Ref. [29], Values of Q_2 and ΔQ_2 are given in (%). See footnote of Table 3 for the definition of F_1score, Q_2, Mcc, and ΔQ_2.

The top three models according to ΔQ_2 (X2463247, X2478107, and X2453885) with ranks 1, 2 and 3 have, respectively, ranks 5, 10 and 6 by F_1score, 10, 5 and 6 by Q_2, and 10, 5 and 15 by the Mcc values. To obtain a deeper evidence in the differences between models ranked as the best ones according to these four statistical parameters, we give in Table 4 the values of elements of contingency matrices for the best 3–5 models. Also, the ratio of negative and positive correct predictions (n/p) and the ratio of the over-prediction and under-prediction (o/u) for each model are given in the last two columns of Table 4.

One can see that the ratios of (n/p) are similar for all top models, but the lowest values or ratios (o/u) are for the models ranked as the best ones according to ΔQ_2, being in the range 2.0 – 3.2. The corresponding range for the top three models ranked by F_1score is 4.4 – 9.8 (Table 4, the last column of part B), and in the similar range are the values of o/u for the top models ranked by Q_2 (part C) and Mcc (part D). The analysis of the ratios of n/p and o/u reveals that ranking by ΔQ_2 favour the models with closer values of o and u, having at the same time the ratio o/u closer to the ratio of n/p. It seems reasonable to proclaim as the better one the model having closer values of ratios of n/p and o/u.

Six examples of extremely imbalanced data sets in Table 5 are constructed and suggested for the testing the suitability of statistical parameters in estimation of quality of classification models.[29] The authors suggested that the correct order (ranking) of quality of models should be $a_6 \rightarrow a_5 \rightarrow a_4 \rightarrow a_3 \rightarrow a_2 \rightarrow a_1$. Three parameters from Table 5 give such an order, and Q_2 is not sensitive to variation of values of elements of contingency tables (p, n, u, and o) corresponding to these models, giving the accuracy of 99.95% for all models.

However, the difference in p, n, u, and o between neighbour models is just 1 or -1, and going from the first to the sixth model p, u, and o values are gradually changed in the same direction (from 0 to 5, or vice versa). Accordingly, we can say that two neighbouring models in the sequence given in Table 5 are the closest neighbours according to p, n, u, and o values. If so, a better quality parameter should have equidistant values (although this is not an imperative property) for models in the sequence $a_6 \rightarrow a_5 \rightarrow a_4 \rightarrow a_3 \rightarrow a_2 \rightarrow a_1$. Only the ΔQ_2 parameter gave such values, the difference being 0.02 between each two neighbouring models.

Both F_1score, suggested as the most convenient for estimating the quality of models on imbalanced sets in Refs. [20–22], and Mcc, suggested as the better one in Refs. [30,31], do not give equidistant values for neighbouring models in Table 5. Moreover, Mcc is not defined for model a_1, what could be an important failure of a parameter.

Hereafter, in Table 6 more examples of values of contingency table corresponding to imbalanced models are given. We will analyse these examples to test the adequacy of these parameters in estimating the model quality, as well as in ranking the models based on the values of these quality parameters. It is evident that Q_2 shows a larger redundancy, because it counts only the sum of correct (positive and negative) predictions giving the same values if $p + n$ is constant. Again, F_1score and Mcc are not defined in some specific cases of contingency table values, but Q_2 and ΔQ_2 are defined in all analysed cases of models in Table 6. The values of F_1score for models C_1 – C_6 show relatively large deviations of this parameter, because it is largely sensitive to relatively small changes of p (i.e. of the class having a smaller number of elements, minority class).

Models C_7 and C_8 illustrate that F_1score is not an adequate measure of more populated class (majority class), and two separate variants of this parameter should be calculated for two classes. The comparison of models C_{11} and C_{12} show drastic differences of F_1score and Mcc just because of small differences of p (from 1 to 0) and o (from 0 to 1) values, and similar conclusions can be drawn from the analysis of models C_{13} and C_{14}. These results indicate a relatively large sensitivity of F_1score and Mcc on the distribution of data. All examples of models given in Table 5 and 6 by their contingency table values are developed on imbalanced data having non-symmetric (skewed) distribution. A similar conclusion was obtained for correlation coefficient and illustrated by artificial

Table 6. Comparison of accuracy parameters calculated from values of contingency tables as suggested in Refs. [29–31] (C_1 – C_4, C_7, C_8) and from additional examples (C_5, C_6, C_9 – C_{14}) introduced in this study (N = 100).[a]

No.	p = TP	u = FN	n = TN	o = FP	true	Q_2	F_1score	Mcc	ΔQ_2
C_1	0	0	95	5	95	95	0.000	#DIV/0!	0.00
C_2	5	5	90	0	95	95	0.667	0.69	9.00
C_3	0	5	95	0	95	95	0.000	#DIV/0!	0.00
C_4	5	0	89	6	94	94	0.625	0.65	8.90
C_5	4	2	90	4	94	94	0.571	0.55	7.04
C_6	1	5	94	0	95	95	0.286	0.40	1.88
C_7	95	5	0	0	95	95	0.974	#DIV/0!	0.00
C_8	90	4	1	5	91	91	0.952	0.14	1.40
C_9	0	5	95	0	95	95	0.000	#DIV/0!	0.00
C_{10}	1	4	90	5	91	91	0.182	0.14	1.40
C_{11}	1	0	99	0	100	100	1.000	1.00	1.98
C_{12}	0	0	99	1	99	99	0.000	#DIV/0!	0.00
C_{13}	0	0	100	0	100	100	#DIV/0!	#DIV/0!	0.00
C_{14}	0	1	99	0	99	99	0.000	#DIV/0!	0.00

[a] Values of Q_2 and ΔQ_2 are given in (%). See footnote of Table 3 for definitions of acronyms.

continuous data given in Table 2. Such a result was to be expected because Mcc is just the correlation coefficient written in the form appropriate for calculation from the contingency table values.

We showed here several examples in which the use of the correlation coefficient is not justified. However, R can be a good and reliable measure in analyses of y-variables (i.e. properties or activities) which have symmetric distribution of data values and, in the case of classification variables, approximately equal number of elements of both classes. Correlation coefficient is a standard and useful validation parameter in analysis of accuracy of models in fitting and in cross-validation. Also, R is a useful accuracy measure in cases in which there is no constant shift between values of experimental and estimated/predicted values of y-variables, and when one wants just to predict by a model the correct order of values of y-variable, e.g. to rank correctly a set of molecules according to predicted properties or activities.

The values of parameter ΔQ_2 are very small in the case of models from Tables 5 and 6 developed on the imbalanced data sets. Knowing that ΔQ_2 is the difference between the real classification accuracy Q_2 and the corresponding random accuracy $Q_{2,\mathrm{rnd}}$ given by Eq. (12), it is normal to obtain even very high value of $Q_{2,\mathrm{rnd}}$ for highly imbalanced data sets. Because the maximal value of Q_2 is 100% and if for the largely imbalanced set the $Q_{2,\mathrm{rnd}}$ value is higher than 95 or 99 %, then ΔQ_2 will be very low. According to presented results we strongly suggest the use of parameter ΔQ_2 in analysis of quality of models, together with other useful and appropriate statistical parameters.

All results related to classification models presented here correspond to external validation, i.e. they are related to predictions done on external test sets. External validation can be problematic and can give over-optimistic or under-optimistic results and performance parameters, especially if the test set is small.[41,42] In such a case, the model evaluation parameters can be highly sensitive to the partition of data into the training and test set, and to the distribution of values of variables or descriptors. However, results of analysis related to the usefulness of different validation parameters presented in this study are based only on the elements of contingency tables and, consequently, are not dependent on the size of data sets and the significance of calculated parameters. The analysis of significance of parameters that are compared and analysed in this study is not a primary issue of this paper, although it is known according to the basics of statistics that each statistical parameter will be more significant if it is calculated on/from a larger set of cases. Anyway, we used here several artificial data sets for which elements of contingency tables (p, n, u and o) were either defined by us or taken from literature (Tables 5 and 6), and their size is selected arbitrary. Moreover, the continuous real data set (Tables S1) is large enough having 1039 and 258 cases (compounds) in the training and test set, respectively. Additionally, the real classification data set is very large having 24687 cases in the test set (Tables 3 and S4), and the

corresponding models submitted within the challenge were developed on the training set containing an even larger number of cases.[13,20,22]

CONCLUSION

The validation of models submitted within predictive challenges is not an easy and straightforward task. In several challenges with subjects connected to the prediction of biological properties of small molecules and macromolecules, evaluation criteria (in cases when continuous predictions are expected) is based on correlation coefficient *R* applied on the test set. However, although the correlation coefficient is very useful in comparative analysis of quality of models, in prediction of activity/property values on the test set (because of the possible constant shift to which *R* is not sensitive) its use for ranking models in prediction could be misleading. Moreover, the test set can have an arbitrary and even much skewed distribution of experimental activity/property data that are intended to be predicted. Consequently, prediction of such data by the models could also have largely skewed distribution. We have shown that the correlation coefficient is very dependent on the distribution of data and it can have a very high value just because of the presence of only one activity/property value which is far from the mean of the rest of data values.

In cases of binary classification problems,[20–22] the Matthew's correlation coefficient Mcc and F_1score were used for ranking models. In our study, a novel accuracy parameter ΔQ_2, estimating the real contribution of a model over the most probable random accuracy, was tested in the evaluation of model predictive quality. We presented here results of the analysis and ranking of classification models/methods based on four most often used accuracy parameters. These results (obtained/predicted ranks) are compared with the ones obtained by the parameter ΔQ_2. Also, the corresponding values of the contingency table parameters were analysed. It turns out that the best models ranked according to parameters ΔQ_2 have more true positives (p = TP) of minor class, and more balanced numbers of false positives (o = FP) and false negatives (u = FN) than the corresponding top models, ranked according to values of Q_2, Mcc or F_1score. Thus, ΔQ_2 favours models having more balanced (symmetric) prediction errors over the imbalanced classification models. The symmetry of false positive and negative prediction errors (o and u) is also suggested as a good characteristic of validation parameter in the literature (e.g. by Baldi et al. in Ref. [43]). Additionally, ΔQ_2 is defined for any set of values of the contingency table, and it has (linearly) proportional values with respect to the changes of values of the contingency table. Additionally, ΔQ_2 is defined for each set of elements of contingency tables.

Presented analyses support the use of the standard error of prediction (or the mean absolute error of prediction) for ranking models developed on continuous data and for evaluation of their quality. Also, presented results related to estimation of quality of classification models strongly support the involvement of parameter ΔQ_2 in the standard set of validation parameters for ranking models within predictive challenges. Namely, the parameter ΔQ_2 estimates just what should be the main aim of improvement of models – i.e. the increase of model accuracy (as much as possible) over the most probable random accuracy.

Acknowledgment. The authors were supported by the Croatian Ministry of Science and Education through basic grants given to their institutions, and by the Croatian Government and the European Union through the European Regional Development Fund – the Competitiveness and Cohesion Operational Programme (KK.01.1.1.01) The Scientific Centre of Excellence for Marine Bioprospecting – BioProCro. The work of doctoral student Viktor Bojović has been fully supported by the "Young researchers' career development project – training of doctoral students" of the Croatian Science Foundation financed by the European Union from the European Social Fund.

Supplementary Information. Supporting information to the paper is attached to the electronic version of the article at: https://doi.org/10.5562/cca3551.

REFERENCES

[1] C. Hansch, T. J. Fujita, *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626. https://doi.org/10.1021/ja01062a035

[2] M. Randić, Z. Mihalić, S. Nikolić, N. Trinajstić, *Croat. Chem. Acta*, **1993**, *66*, 411–434.

[3] M. Randić, *Croat. Chem. Acta*, **1993**, *66*, 289–312.

[4] R. D. Cramer, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967. https://doi.org/10.1021/ja00226a005

[5] A. Golbraikh, A. Tropsha, *J. Mol. Graphics Model.* **2002**, *20*, 269–276. https://doi.org/10.1016/S1093-3263(01)00123-1

[6] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR & Comb. Sci.* **2003**, *22*, 69–77. https://doi.org/10.1002/qsar.200390007

[7] B. Lučić, D. Amić, N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 403–413. https://doi.org/10.1021/ci990061k

[8] B. Lučić, I. Bašic, D. Nadramija, et al. *Arkivoc* **2002**, *4*, 45–59. https://doi.org/10.3998/ark.5550190.0003.406

[9] T. Piližota, B. Lučić, N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 113–121. https://doi.org/10.1021/ci034037p

[10] *Report from the expert group on (Q)SARs on principles for the validation of (Q)SARs*, http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2004)24&doclanguage=en, **2004**, (accessed on June 23, 2019)
[11] Regulation (of the European Parliament and of the Council) no 1907/2006 (REACH), https://osha.europa.eu/en/legislation/directives/regulation-ec-no-1907-2006-of-the-european-parliament-and-of-the-council, **2006**, (accessed on June 23, 2019)
[12] *Guidance document on the validation of (Q)SAR models* (of the Europe Union) http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en, **2007**, (accessed on June 23, 2019)
[13] The DREAM Consortium, **2019**, https://dreamchallenges.org, (accessed on June 23, 2019)
[14] NCI DREAM Community, J. C. Costello, L. M. Heiser, et al. *Nat. Biotechnol.* **2014**, *32*, 1202–1212. https://doi.org/10.1038/nbt.2877
[15] A. A. Margolin, E. Bilal, E. Huang, et al. *Sci. Transl. Med*. **2013**, *5*, 181re1. https://doi.org/10.1126/scitranslmed.3006112
[16] Members of the Rheumatoid Arthritis Challenge Consortium, S. K. Sieberts, F. Zhu, et al. *Nat. Commun.* **2016**, *7*, art. no. 12460.
[17] R. Küffner, N. Zach, R. Norel, et al. *Nat. Biotechnol.* **2015**, *33*, 51–57. https://doi.org/10.1038/nbt.3051
[18] R. Liu, S.-S So, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1633–1639. https://doi.org/10.1021/ci010289j
[19] J. Batista, D. Vikić-Topić, B. Lučić, *Croat. Chem. Acta*, **2016**, *89*, 527–534. https://doi.org/10.5562/cca3117
[20] C. I. Cooper, D. Yao, D. H. Sendorek, et al. *BMC Bioinformatics* **2018**, *19*, art. no. 339. https://doi.org/10.1186/s12859-018-2391-z
[21] N. Aghaeepour, G. Finak, The FlowCAP Consortium, The DREAM Consortium, et al. *Nat. Methods* **2013**, *10*, 228–238. https://doi.org/10.1038/nmeth.2365
[22] A. D. Ewing, K. E. Houlahan, Y. Hu, et al. *Nat. Methods* **2015**, *12*, 623–630. https://doi.org/10.1038/nmeth.3407
[23] D. M. W. Powers, Int. *J. Mach. Learn. Technol*. **2011**, *2*, 37–63.
[24] B. W. Matthews, *Biochim. Biophys. Acta* **1975**, *405*, 442–451. https://doi.org/10.1016/0005-2795(75)90109-9
[25] D. Juretić, B. Lučić, N. Trinajstić, *J. Mol. Struct. (THEOCHEM)* **1995**, *338*, 43–50. https://doi.org/10.1016/0166-1280(94)04047-V
[26] J. Huuskonen, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777. https://doi.org/10.1021/ci9901338
[27] S. H. Yalkowsky, R. M. Dannelfelser, *The Arizona Database of Aqueous Solubility* (AQUASOL), College of Pharmacy, University of Arizona, Tucson, AZ, **1990**.
[28] Syracuse Research Corporation. Physical/Chemical Property Database (PHYSOPROP); SRC Environmental Science Center: Syracuse, NY, **1994**.
[29] M. B. Hossin, M. N. Sulaiman, *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 1–11.
[30] D. Chicco, *BioData Mining* **2017**, *10*. Art. no. 35. https://doi.org/10.1186/s13040-017-0155-3
[31] Wikipedia contributors, Matthew's correlation coefficient, *Wikipedia*, *The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Matthews_correlation_coefficient&oldid=887971433 (accessed June 23, 2019).
[32] M. Mathea, W. Klingspohn, K. Baumann, *Mol. Inf.* **2016**, *35*, 160–180. https://doi.org/10.1002/minf.201501019
[33] B. Lučić, N. Trinajstić, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 121–132. https://doi.org/10.1021/ci980090f
[34] Dragon Pro 5.4 program, Milano, Talete srl, Italy (www.talete.mi.it), **2006**.
[35] L. I.-K. Lin, *Biometrics*, **1989**, *45*, 255–268. https://doi.org/10.2307/2532051
[36] A. Sović Kržić, D. Seršić, *Signal Process.* **2018**, *151*, 119–129. https://doi.org/10.1016/j.sigpro.2018.05.002
[37] J. R. Bošković, "De literaria expeditione per pontificiam ditionem et synopsis amplioris operis," *Bononiensi scientiarum et artum instituto atque academia commentarii*, **1757**.
[38] A. Sović, D. Seršić, *Engin. Rev.* **2012**, *32*, 70–77. https://doi.org/10.3103/S1068798X12010285
[39] B. Lučić, D. Amić, N. Trinajstić, *Antioxidant QSAR Mo-deling as Exemplified on Polyphenols* (chapter 16), in *Methods in Molecular Biology,* Vol. 477: *Advanced Protocols in Oxidative Stress I* (Ed.: D. Armstron), Humana Press, New York, NY, USA, **2010**, pp. 207–218. https://doi.org/10.1007/978-1-60327-517-0_16
[40] Y.-Z. Cai, M. Sun, J. Xing, Q. Luo, H. Corke, *Life Sci.* **2006**, *78*, 2872–2888. https://doi.org/10.1016/j.lfs.2005.11.004
[41] S. Majumdar, S. C. Basak, *Curr. Comput. Aided Drug Des.* **2018**, *14*, 284–291. https://doi.org/10.2174/1573409914666180426144304
[42] S. C Basak, S. Majumdar, *Curr. Comput. Aided Drug Des.* **2015**, *11*, 2–4. https://doi.org/10.2174/1573409911011507221421444
[43] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, H. Nielsen, *Bioinformatics* **2000**, *16*, 412–424. https://doi.org/10.1093/bioinformatics/16.5.412