Archives of Toxicology (2022) 96:1279–1295
https://doi.org/10.1007/s00204-022-03252-y

REVIEW ARTICLE

Prediction reliability of QSAR models: an overview of various validation tools

Priyanka De[1] · Supratik Kar[2] · Pravin Ambure[3] · Kunal Roy[1]

Received: 28 December 2021 / Accepted: 14 February 2022 / Published online: 10 March 2022

Abstract

The reliability of any quantitative structure–activity relationship (QSAR) model depends on multiple aspects such as the accuracy of the input dataset, selection of significant descriptors, the appropriate splitting process of the dataset, statistical tools used, and most notably on the measures of validation. Validation, the most crucial step in QSAR model development, confirms the reliability of the developed QSAR models and the acceptability of each step in the model development. The present review deals with various validation tools that involve multiple techniques that improve the model quality and robustness. The double cross-validation tool helps in building improved quality models using different combinations of the same training set in an inner cross-validation loop. This exhaustive method is also integrated for small datasets (< 40 compounds) in another tool, namely the small dataset modeler tool. The main aim of QSAR researchers is to improve prediction quality by lowering the prediction errors for the query compounds. 'Intelligent' selection of multiple models and consensus predictions integrated in the intelligent consensus predictor tool were found to be more externally predictive than individual models. Furthermore, another tool called Prediction Reliability Indicator was explained to understand the quality of predictions for a true external set. This tool uses a composite scoring technique to identify query compounds as 'good' or 'moderate' or 'bad' predictions. We have also discussed a quantitative read-across tool which predicts a chemical response based on the similarity with structural analogues. The discussed tools are freely available from https://dtclab.webs.com/software-tools or http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/ and https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home (for read-across).

Keywords QSAR · Validation · Double cross-validation · Small dataset modeling · Intelligent consensus prediction · Read across

Introduction

A growing number of research have been conducted in recent years, wherein computational methods have been used to predict the physicochemical properties and biological activities of chemical compounds. Quantitative structure−activity relationship (QSAR) (Dearden 2016) modeling is a popular in silico technique performed to find out a quantitative correlation between the structural features (known as descriptors) and a known response (activity/property/toxicity) for a set of molecules using various chemometric methodologies. QSAR evolves at the crossroads of chemistry, statistics, biology, and toxicological studies. The main aim is to identify and optimize new leads to shorten the time and reduce expenditure for drug discovery (Hsu et al. 2017). The fundamental assumption regarding QSAR modeling is that a chemical structure possesses unique features (geometric, steric, and electronic properties) responsible for its physical, chemical, and biological properties.

The European Union (EU) envisaged that QSAR models would increasingly be used for hazard and risk assessments of chemicals (Commission of the European Communities 2001). It is also necessary to create and apply QSARs to address animal welfare concerns by replacing, reducing,

✉ Kunal Roy
kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in

[1] Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

[2] Interdisciplinary Center for Nanotoxicity, Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, Jackson, MS 39217, USA

[3] ProtoQSAR S.L., Valencia, Spain

and refining animal testing in toxicological assessments. In November 2004, the European Commission and the OECD (Organisation for Economic Co-operation and Development) member countries adopted principles for validation of QSAR models for use in regulatory assessment of chemical safety (Organisation for Economic Co-operation and Development (OECD 2004). According to the agreed guidelines of OECD, a QSAR model should be developed with

(a) A defined endpoint,
(b) An unambiguous algorithm to guarantee model transparency,
(c) A defined domain of applicability,
(d) Proper measures of validation including internal performance (as determined by goodness-of-fit and robustness) and predictivity (as represented by external validation), and
(e) Possible mechanistic interpretation.

Validation is crucial for the development and application of any QSAR model. It confirms the reliability of the developed model and the acceptability of each step through model development. The debate between internal versus external validation prevails predominantly among QSAR practitioners (Roy 2007). Some QSAR studies reported an inconsistency between internal and external predictivity (Novellino et al. 1995; Norinder 1996). According to researchers, there might be an inconsistency between internal and external predictability, i.e., high internal predictivity may result in low external predictivity and vice versa (Kubinyi 1998). However, external validation is considered the 'gold standard' of checking predictive potential of QSAR models. Some researchers consider cross-validation to be more appropriate for checking the predictive ability of QSAR models to circumvent the loss of information from splitting the dataset into training and test sets (Héberger 2017). Several validation metrics (as discussed later) are used to check the quality of predictions generated by regression-based and classification-based QSAR models (Gramatica and Sangion 2016; Todeschini et al. 2016).

The present review has discussed several prediction reliability tools exploring various strategies to determine model reliability and predictivity. We have discussed the tools that engage in the model-building through a double cross-validation approach on large and small datasets. Furthermore, we have explained the utility of intelligent selection of multiple models and various forms of consensus prediction. We have also mentioned a tool that explains a similarity-based reliability scoring approach to understand the quality of predictions for a new query compound and ensure the developed models' reliability. We have further reported a similarity-based quantitative read-across tool addressing the quality of predictions both quantitatively and qualitatively.

Predictive QSAR model development approaches

Modern QSAR methods use multiple descriptors combined with the application of both linear and non-linear modeling approaches with a strong emphasis on rigorous model validation to afford robust and predictive QSAR models. Several types of research along with our understanding of QSAR model development and validation led us to establish a general outline of QSAR model workflow as described in Fig. 1. This figure illustrates the classical QSAR model development algorithm which includes: (a) collection of pertinent data with a defined endpoint, (b) descriptor calculation and data pre-treatment, (c) model development through analysis of the correlation between input data and descriptors calculated, (d) validation of the model, and (e) design and prediction of the activity of new query molecules. The QSAR modeling scheme is further described briefly in the following section.

(i) **Dataset preparation and data curation:** One of the most challenging parts of QSAR is dataset collection with a "defined endpoint" as explained in OECD principle 1. The intent is to confirm the transparency of the endpoint aimed for prediction models, considering that a given endpoint could be dependent on the experimental protocol and the experimental conditions. Data curation is an essential and time-consuming step in the QSAR model development process. Erroneous data (both in chemical structures and biological data) retrieved from online sources require strict curation to avoid false or non-predictive models (Ambure and Cordeiro 2020).
(ii) **Calculation of molecular descriptors:** The molecular structures applied for QSAR modeling need to be translated into numbers, i.e., molecular descriptors. The molecular descriptor is an encoded representation of the information about a chemical compound in the form of numerical values based on its chemical constitution, allowing the correlation of chemical structure with physical properties, chemical reactions, or biological activity (Consonni and Todeschini 2010). In a QSAR model, descriptors of a molecule, which describe specific aspects of a molecule, are predictors (X) of the dependent variable (Y). A QSAR study uses a variety of descriptors that can be classified into different dimensions or categories, as shown in Table 1.

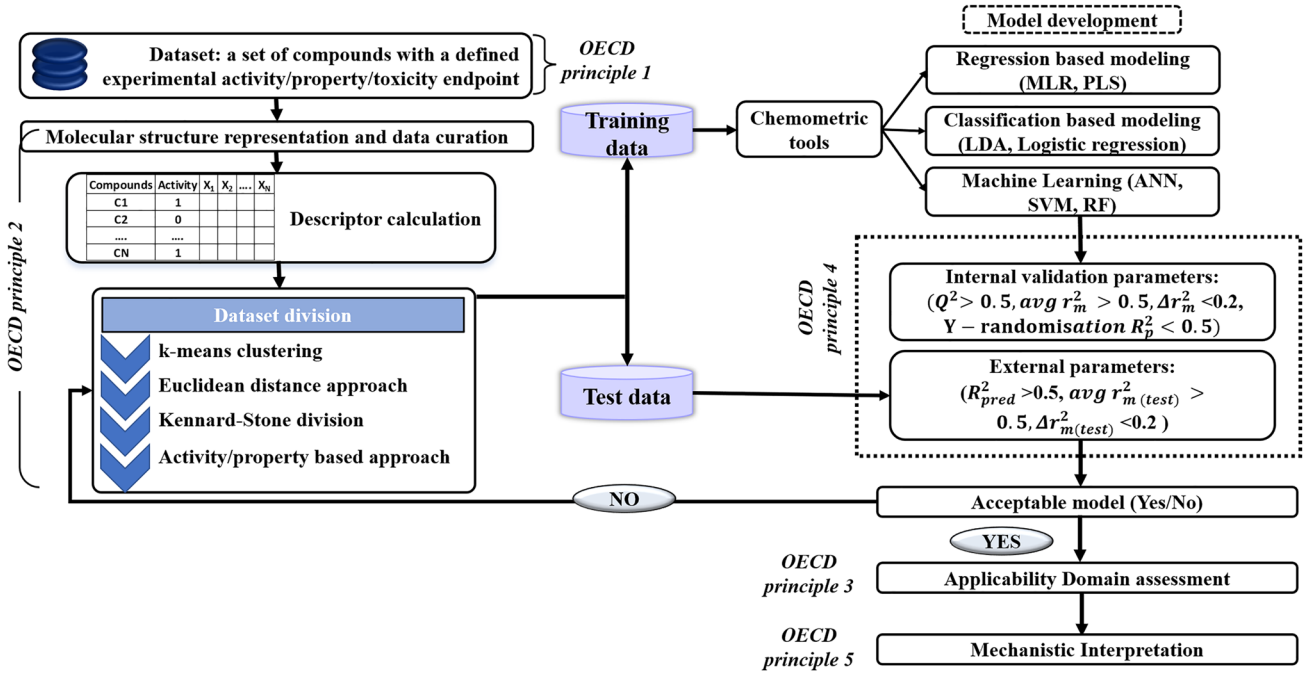

Fig. 1 Schematic representation of QSAR methodology according to OECD guidelines

(iii) **Dataset division:** A predictive model's performance must be determined by dividing the dataset into a training set and a test set. Among all chemicals, only the training set molecules are used for developing QSAR models, and the external predictivity of the models is examined through the use of test set compounds. In developing the QSAR model, it is necessary to select a training set in a way, such that it encompasses a wide chemical domain. The test set compounds must lie within the chemical space of the training set. Dataset division involves different methods including (a) Euclidean distance (diversity-based) (Golmohammadi et al. 2012), (b) Kennard-Stone (Kennard and Stone 1969), (c) k-means clustering (Likas et al. 2003), (d) sorted response (Roy 2018), etc.

(iv) **Feature selection:** A feature selection process is a vital step that involves identifying important predictor variables to develop correlations with the response variable. Feature selection helps decrease the model complexity, decreases the risk of overfitting or overtraining, and helps select the most critical descriptors among a pool of hundreds or thousands. In this way, the dimensionality of input descriptors is minimized without the loss of essential information (Goodarzi et al. 2012). Finally, these selected descriptors are used to build a mathematical model linking to the biological activity of the corresponding compounds. According to the OECD guidelines, several feature selection techniques have been applied using a mechanistic basis including, genetic algorithms, genetic function approximation (GFA), forward selection, backward elimination, stepwise regression, simulated annealing, etc.

(v) **Model development algorithms:** The OECD guideline 2 explains that a QSAR model should be developed using an "unambiguous algorithm" (Directorate 2007). The rule focuses on bringing transparency in model-building, rendering it reproducible to others and making it possible to achieve the endpoint estimates. This embraces the methods implemented during data pretreatment, division of data, feature selection, and model development. Linear modeling techniques involve multiple linear regression (MLR) (Pope and Webster 1972; De and Roy 2018), ordinary least squares (OLS), partial least squares (PLS) (Wold et al. 2001), principal component analysis (PCA) (Abdi and Williams 2010), principal component regression (PCR), etc.

In QSAR, model-building tools can be grouped into two major categories: regression-based approach and classification-based approach. Regression-based approaches are effective when both dependent (response variable) and independent (molecular descriptors) variables are quantitative (Roy et al. 2015a; b). In the case of classification-based modeling, a relationship between the descriptors and the graded values

Table 1 Types of 0D-3D descriptors used in the QSAR study

Dimension of descriptors	Parameters	Examples
0D	Constitutional indices	Number of atoms, number of non-H atoms, number of bonds, number of aromatic bonds, sum of atomic van der Waals volumes (scaled on carbon atom), etc.
	Molecular property Atom and bond counts	Unsaturation count, unsaturation index, hydrophilic factor, unsaturation index
1D	Fragment counts, fingerprints	Atom centered fragments (C-001, H-046, O-056, etc.)
2D	Topological	Wiener index (W), Zagreb group indices, Balaban *J* index, Randic branching index (χ), Molecular connectivity index, subgraph count, Chi indices, etc.
	Structural	Chiral centers, rotatable bonds, HBond donor, HBond acceptor
	Physicochemical parameters (thermodynamic parameters)	Heat of formation (Hf), Log of the partition coefficient using Ghose and Crippen's method (AlogP), Desolvation free energy (Fh2o, Foct)
	Connectivity indices	Average connectivity index, valence connectivity index, solvation connectivity index, modified Randic index, connectivity topochemical index, perturbation connectivity index
	Functional group counts	Number of terminal primary C(sp3), number of total secondary C(sp3), number of ring quaternary C(sp3), number of carboxylic acids, number of hydroxyl groups, etc.
	2D matrix based	Balaban-like index from adjacency matrix, logarithmic spectral positive sum from adjacency matrix, spectral absolute deviation from adjacency matrix, etc.
	2D atom pairs	Presence or absence of any two atoms at a particular topological distance (B01[C–C], B09[C-F], etc.), frequency of two atoms at a particular topological distance (F01[C-F], F05[O-N]), sum of occurrence of two atoms at a particular topological distance (T(N..I), T(O..N))
3D	Electronic	Dipole moment, highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), superdelocalizibility
	Spatial	The radius of gyration, Jurs descriptors, area, density, volume, etc.
	Receptor surface analysis parameters	Hydrophobicity, partial charge, electrostatic (ELE) potential, van der Waals (VDW) potential, and hydrogen bonding propensity
	Molecular shape analysis	Difference volume (DIFFV), Common overlap steric volume (COSV), Common overlap volume ratio (Fo), Noncommon overlap steric volume (NCOSV), Root mean square to shape reference (ShapeRMS)
	Geometric	Molecular eccentricity, spherosity, asphericity, aromaticity index, gravitational index
	Other 3D descriptors	3D matrix based (Wiener like index, Randic like index, Balaban-like index, etc. all from geometric matrix, spectral moment,), 3D autocorrelations (3D Topological distance-based descriptors: unweighted; weighted by mass, polarizability, van der Waals volume, Sanderson electronegativity, ionization potential), 3D Morse descriptors, WHIM descriptors, GETAWAY descriptors, quantum-chemical descriptors

0D, 1D, and 2D descriptors may be collectively grouped under the broad class of 2D descriptors in general

of the response variable(s) is established. Here, the response is offered in a Boolean form like active/inactive and positive/negative or categorical (as observed in linear discriminant analysis, logistic regression, and cluster analysis).

(vi) **Determination of domain of applicability:** One of the most essential checkpoints in QSAR modeling is determining the applicability domain (AD) of a model as explained in OECD principle 3. The applicability domain denotes a physicochemical space (both the response and chemical structure space) within which a QSAR model can predict with a certain degree of reliability (Roy et al. 2015a, b). This space is defined by the features explained by the compounds in the training set and is mandatory to examine whether the prediction of test set molecules is reliable or not. The concept of AD was used to avoid an unjustified extrapolation of property predictions.

(vii) **QSAR model validation:** Before interpreting and predicting biological responses of untested compounds, any QSAR model needs to be validated. Here, the model's predictive power is established, and the ability to reproduce the biological activities of the untested compounds is measured. In consonance with the fourth principle of OECD guidelines, statistical validation of models in terms of goodness-of-fit, robustness, and predictivity is an extremely impor-

tant step during QSAR model development. The validation of QSAR models is crucial if these models are used for virtual screening. Each validation parameter aims to judge the accuracy of prediction, i.e., determining whether the experimental value is close to the model-derived value. The model fitness determined using the coefficient of determination or correlation coefficient from the training set measures the degree of achieved correlation between the experimental (Y_{exp}) and calculated (Y_{calc}) response values. Data fitting does not confirm the predictability of a model but instead demonstrates the model's statistical quality. Different internal and external validation metrics for both regression and classification modeling are utilized to check model prediction quality which is discussed later in the following section.

(viii) **Mechanistic interpretation:** The fifth OECD principle focuses on identifying the features of the variables that may contribute to a more thorough understanding of the response being modeled. Chemicals that act specifically using a specific mechanism can only be designed and developed with absolute certainty using the structural analogues. However, it is evident that furnishing mechanistic information may not always be feasible. The rule suggests that the modeler should report if any such information is available, facilitating future research on that endpoint. A mechanistic interpretation from the literature can be added, and therefore, the fifth OECD principle encourages the reporting of such information to enrich the physicochemical understanding of response being modeled.

Regression and classification validation metrics

The reliability of a developed QSAR model is confirmed through the validation process. The quality of input data, dataset diversity, predictability on an external set, applicability domain determination, and mechanistic interpretability are also confirmed through various validation metrics. QSAR model validation can be classified into two major types: (a) internal validation and (b) external validation. Internal validation in QSAR modeling involves activity prediction of the molecules/compounds used for generating the model. This is followed by estimating metrics for detecting the precision of predictions. Internal validation is useful in the case of cross-validation approaches (Konovalov et al. 2008) where the internal data are partitioned into calibration (training) and validation (test) subsets. The calibration set is used for model-building purposes, and the validation set is utilized for model predictivity assessment. Assessment of prediction capability and applicability of a QSAR model to predict newly designed or untested molecules is done using external validation metrics. In most cases, some compounds from the original datasets are used for validation purpose when true external data points are limited or not available.

Regression-based validation metrics

One of the main quality metrics to check the goodness-of-fit of a regression model is the determination coefficient $\left(R^2\right)$ which measures the variation of observed data with the fitted ones. The maximum possible value for R^2 is 1, which defines a perfect correlation.

Adjusted R^2 (R^2_{adj}) is a modified version of the determination coefficient and is also known as the explained variance. The R^2_{adj} parameter incorporates the information of the number of samples and the independent variables used in the model.

Considering the internal validation for a regression-based QSAR model, the leave-one-out cross-validation (Q^2_{LOO}) metric is calculated. Here, a model is developed by modifying the original training set of n compounds by removing one compound. The activity of the omitted compound is then predicted using the model developed with n-1 compounds. This cycle is repeated until all the training set compounds have been eliminated once and the predicted activity data are obtained for all the training set compounds. The model predictivity is thus measured using the predicted residual sum of squares (PRESS) and cross-validated R^2 (Q^2) (Table 2). The PRESS value is defined as the sum of squared differences between the experimental and leave-one-out predicted data. The standard deviation of error of predictions (SDEP) is calculated from the PRESS value (Table 2). A model is considered satisfactory if the value of Q^2 is higher than the predetermined value of 0.6. However, numerous evidences suggested that leave-one-out prediction should neither be considered as the ultimate standard for judging the predictive power of models nor for model selection (Konovalov et al. 2007; Veerasamy et al. 2011). There is a chance of overfitting and overestimation in LOO due to structural redundancy (Höltje and Sippl 2001). Leave-many-out (LMO) or leave-some-out (LSO) might be a better alternative where a part of the training data is held out ($(1 \leq m < n$, where n is a sample size) and the remaining data are modeled. The model is developed using the remaining compounds in each cycle, and the hold-out compounds are predicted. This cycle continues till all the compounds are predicted, and the predicted values are used for the calculation of Q^2_{LMO}. Therefore, the LMO technique partly reflects external validation in the context of internal validation.

Although, Q^2_{LOO} provides a measure of model robustness, it may not be sufficient to characterize the performance of the model during prediction of new query/test compounds. Furthermore, Q^2_{LOO} can provide an overestimation of model

Table 2 Validation metrics for regression modeling

Parameters	Equation	Description
Determination coefficient $\left(R^2\right)$	$R^2 = 1 - \frac{\sum \left(Y_{\text{obs}} - Y_{\text{pred}}\right)^2}{\sum \left(Y_{\text{obs}} - \overline{Y_{\text{training}}}\right)^2}$	Metric to check the goodness-of-fit of a regression model. It measures the variation of observed data with the predicted ones. The maximum possible value for R^2 is 1, which defines a perfect correlation. Y_{obs} denotes the observed response values for the training set, and Y_{pred} denotes the calculated response values for the training set of compounds. Y_{training} is the mean observed response of the training set compounds
Explained variance or adjusted R^2 (R^2_{adj})	$R^2_{adj} = \frac{\{(n-1)XR^2\}-p}{n-p-1}$	Modified version of the determination coefficient. The R^2_{adj} parameter incorporates the information of the number of samples and the independent variables used in the model. n is the number of training set compounds and p is the number of predictor variables
Leave-one-out cross-validation (Q^2_{LOO})	$Q^2_{LOO} = 1 - \frac{\sum \left(Y_{\text{obs(training)}} - Y_{\text{pred(training)}}\right)^2}{\sum \left(Y_{\text{obs(training)}} - \overline{Y_{\text{training}}}\right)^2}$	Cross-validated $R^2(Q^2)$ is checked for internal validation. $Y_{\text{obs(training)}}$ is the observed response, and $Y_{\text{pred(training)}}$ is the predicted response of the training set molecules based on the leave-one-out (LOO) technique
Predictive R^2 or R^2_{pred} or $Q^2_{\text{ext}(F1)}$	$Q^2_{\text{ext}(F1)} = 1 - \frac{\sum (Y_{\text{obs(test)}} - Y_{\text{pred(test)}})^2}{\sum (Y_{\text{obs(test)}} - \overline{Y_{\text{training}}})^2}$	This metric employed for judging external predictivity. It is a measure of correlation between the observed and predicted data of test set. $Y_{\text{obs(test)}}$ is the observed response, and $Y_{\text{pred(test)}}$ is the predicted response of the test set molecules. $\overline{Y_{\text{training}}}$ denotes the mean observed response of the training set
$Q^2_{\text{ext}(F2)}$	$Q^2_{\text{ext}(F2)} = 1 - \frac{\sum (Y_{\text{obs(test)}} - Y_{\text{pred(test)}})^2}{\sum (Y_{\text{obs(test)}} - \overline{Y_{\text{test}}})^2}$	It helps in the judgment of predictivity of a model using the test set mean $\left(\overline{Y_{\text{test}}}\right)$.
$Q^2_{\text{ext}(F3)}$	$Q^2_{\text{ext}(F3)} = 1 - \frac{\left[\sum (Y_{\text{obs(test)}} - Y_{\text{pred(test)}})^2\right]/n_{\text{test}}}{\left[\sum \left(Y_{\text{obs(train)}} - \overline{Y_{\text{training}}}\right)^2\right]/n_{\text{train}}}$	$Q^2_{\text{ext}(F3)}$ is measured to determine external predictivity employing both training and test set features. $Y_{\text{obs(test)}}$ is the observed response, and $Y_{\text{pred(test)}}$ is the predicted response of the test set molecules. $Y_{\text{obs(training)}}$ is the observed response and $\overline{Y_{\text{training}}}$ denotes the mean observed response of the training set molecules. The threshold for $Q^2_{\text{ext}(F3)}$ is 0.5
Concordance correlation coefficient (CCC)	$CCC = \overline{p_c} = \frac{2\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(y_i-\bar{y})^2+n(\bar{x}-\bar{y})}$	The concordance correlation coefficient (CCC) measures both precision and accuracy detecting the distance of the observations from the fitting line and the degree of deviation of the regression line from that passing through the origin, respectively. 'n' denotes the number of compounds, and x_i and y_i signify the mean of observed and predicted values, respectively
Root mean square error in predictions (RMSE_p)	$RMSE_p = \sqrt{\frac{\sum (Y_{\text{obs(test)}} - Y_{\text{pred(test)}})^2}{n_{\text{test}}}}$	It gives a measure of model external validation. A lower value of this parameter is desirable for good external predictivity
r^2_m metrics	$\overline{r^2_m} = \frac{r^2_m + r'^2_m}{2}$ and $\Delta r^2_m = \left\lvert r^2_m - r'^2_m \right\rvert$ where $r^2_m = r^2 \times (1 - \sqrt{r^2 - r^2_0}))$ $r'^2_m = r^2 \times \left(1 - \sqrt{r^2 - r'^2_0}\right)$	r^2 is the squared correlation coefficient value between observed and predicted response values, and r_0^2 and $r'_0{}^2$ are the respective squared correlation coefficients when the regression line is passed through the origin by interchanging the axes. For the acceptable prediction, the value of all Δr^2_m metrics should preferably be lower than 0.2 provided that the value of $\overline{r^2_m}$ is more than 0.5 (Ojha et al. 2011)
Predicted residual sum of squares (PRESS)	$\text{PRESS} = \sum \left(Y_{\text{obs}} - Y_{\text{pred}}\right)^2$	Sum of squared differences between experimental and predicted data. Y_{obs} and Y_{pred} correspond to the observed and LOO predicted values

Table 2 (continued)

Parameters	Equation	Description
Standard deviation of error of prediction (SDEP)	$\mathrm{SDEP} = \sqrt{\frac{\mathrm{PRESS}}{n}}$	The value of standard deviation of error of prediction (SDEP) is calculated from PRESS. n refers to the number of observations
Mean absolute error (MAE)	$MAE = \frac{1}{n} \times \sum \left\lvert Y_{\mathrm{obs}} - Y_{\mathrm{pred}} \right\rvert$	This is also known as average absolute error (AAE) and is considered a better index of errors in the context of predictive modeling studies

quality as a result of structural redundancy in the training set data. Thus, the performance of a model on an external dataset is considered mandatory for the judgment of predictivity. The metric employed for judging external predictivity is termed as predictive R^2 or R^2_{pred} or $Q^2_{\mathrm{ext}(F1)}$. The $Q^2_{\mathrm{ext}(F1)}$ metric is characterized by a minimum threshold value of 0.6, i.e., models showing a value more than 0.6 are considered to be externally predictive with the ideal value being 1.0. Schüürmann and co-workers (Schüürmann et al. 2008) defined another external validation metric $Q^2_{\mathrm{ext}(F2)}$ for the judgment of the predictivity of a model using the test set. Consonni et al. (2009) introduced another external validation metric $Q^2_{\mathrm{ext}(F3)}$. This metric measures the model predictability and is sensitive to the selection of training dataset and tends to penalize models fitted to a very homogeneous data set even if predictions are close to the truth, with a threshold value being 0.6.

Another metric that checks the model reliability is the concordance correlation coefficient (CCC) metric (Chirico and Gramatica 2011). It measures both precision and accuracy, detecting the distance of the observations from the fitting line and the degree of deviation of the regression line from that passing through the origin, respectively. Any deviation of the regression line from the concordance line (line passing through the origin) gives a value of CCC smaller than 1. The desirable threshold value for CCC is 0.85.

The root-mean-square error in predictions $\left(\mathrm{RMSE}_p\right)$ gives a measure of model external validation. This metric is comparatively simpler and directly depicts the prediction errors for the test set observations concerning the total number of test set samples. A lower value of this metric is desirable for good external predictivity.

The r^2_{m} metrics: the training set mean value and the distance of the mean from the response values of each compound play a decisive role in computing the Q^2 values. The Q^2 value increases with the rise in the value of the denominator in the expression of Q^2. Thus, even for a considerable deviation between the predicted and observed response values, satisfactory Q^2 values may be obtained, if the molecules exhibit a considerably broad range of response data. Using the concept of regression through origin approach, Roy et al. (2012) introduced a new metric r^2_{m} or modified r^2 that penalizes the r^2 value of a model when there is large deviation between r^2 (squared correlation coefficient values between the observed (Y axis) and predicted (X axis) values of the compounds with intercept) and $r_0{}^2$ (squared correlation coefficient values between the observed (Y axis) and predicted (X axis) values of the compounds without intercept) values (Table 1).

MAE-based criteria: in a study, Roy et al. (2016) have shown that the conventional correlation-based external validation metrics ($Q^2_{\mathrm{ext}(F1)}$,$Q^2_{\mathrm{ext}(F2)}$) often provide biased judgment of model predictivity, since such metrics are influenced by factors such as response range and distribution of data. Here, the authors have defined a set of criteria using simple 'mean absolute error' (MAE) and the corresponding standard deviation (σ) measure of the predicted residuals to judge the external predictivity of the models. Note that $\mathrm{MAE} = \frac{1}{n} \times \sum \left\lvert Y_{\mathrm{obs}} - Y_{\mathrm{pred}} \right\rvert$, where Y_{obs} and Y_{pred} are the respective observed and predicted response values of the test set comprising n number of compounds. The response range of training set compounds has been employed here to define the threshold values. Furthermore, the authors have proposed the application of the 'MAE based criteria' on 95% of the test set data by removing 5% data with high predicted residual values precluding the possibility of biased prediction quality due to any outlier prediction. The following criteria for MAE prediction are followed:

i. Good predictions: in easier terms, an error of 10% of the training set range should be acceptable, while an error more than 20% of the training set range should be a very high error. Thus, the criterion for good predictions is as follows:

$$\mathrm{MAE} \leq 0.1 \times \text{training set range and } (\mathrm{MAE} + 3\sigma) \leq 0.2 \times \text{training set range.}$$

Here, σ value indicates the standard deviation of absolute errors for the test data. For a normal distribution pattern, mean ± 3σ covers 99.7% of the data points.

ii. Bad predictions: a value of MAE more than 15% of the training set range is considered high, while an error higher than 25% of the training set range is judged as very high. Thus, prediction is considered bad when

MAE > 0.15 × training set range or (MAE + 3σ) > 0.25 × training set range.

Predictions which do not fall under either of the above two conditions may be considered as of moderate quality. This criterion is applied for judging the quality of test set prediction when there are at least 10 data points signifying statistical reliability and there is no systemic error in model predictions.

Randomisation of response (Y-scrambling)–Randomisation is an assessment to ensure the developed QSAR model is not due to chance, thereby giving an idea of model robustness (Rücker et al. 2007). In this technique, validation metrics are checked by repetitive permutation of the response data (Y) of n compounds in the training set with respect to the X (descriptor) matrix which is kept unchanged. The calculations are repeated with randomized activities, followed by a probabilistic examination of the results. Every run will yield approximations of R^2 and Q^2, which are recorded. For an acceptable QSAR model, the average correlation coefficient (R_r) of randomized models should be less than the correlation coefficient (R) of a non-random model. The difference between mean-squared correlation coefficients of the randomized (R_r^2) and that of the non-random (R^2) models can be obtained through R_p^2 calculation ($R_p^2 = R^2 \times \sqrt{R^2 - R_r^2}$). A robust QSAR model should have R_p^2 value less than 0.5. At the ideal condition, the average value of R^2 for the randomized models should be zero, i.e., R_r^2 should be zero. Consequently, in such a case, the value of R_p^2 should be equal to the value of R^2 for the developed QSAR model. Thus, as proposed by Todeschini, the corrected formula of $R_p^2(c_{R_p}^2)$ is $c_{R_p}^2 = R \times \sqrt{R^2 - R_r^2}$ (Todeschini 2010).

Classification-based QSAR validation metrics

In a binary classification model, several validation metrics are utilized to evaluate the model's performance in terms of accurate qualitative prediction of the dependent variable. Classification models are generally assessed using a statistical method that is based on the Bayesian approach (Ghosh et al. 2020). A binary classification model is typically a two-class model, i.e., positive and negative, or active and inactive. The results obtained can be arranged in a contingency table (also known as confusion matrix) (Table 3). The

Table 3 Contingency table or confusion matrix for classification modeling

		Experimental		Total
		Active	Inactive	
Predicted	Active	True positive (TP)	False positive (FP)	TP+FP
	Inactive	False negative (FN)	True negative (TN)	FN+TN
Total		TP+FN	FP+TN	N = TP+FP+FN+TN

Table 4 Validation metrics for classification modeling

Sl No.	Classification metric	Equation
1	Sensitivity	$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}$
2	Specificity	$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}$
3	Precision	$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$
4	Accuracy	$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+TN+FP}}$
5	F-measure	$\text{F} - \text{measure}(\%) = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Sensitivity}}}$
6	G-means	$\text{G} - \text{means} = \sqrt{\text{SpecificityXSensitivity}}$
7	Cohen's Kappa (κ)	$P_r(a) = \frac{(\text{TP+TN})}{(\text{TP+FP+TN+FN})}$ $P_r(e) = \frac{\{(\text{TP+FP})\times(\text{TP+FN})\}+\{(\text{TN+FP})\times(\text{TN+FN})\}}{(\text{TP+FN+FP+TN})^2}$ $\text{Cohen}'\text{s}K = \frac{P_r(a)-P_r(e)}{1-P_r(e)}$
8	Mathews correlation coefficient (MCC)	$\text{MCC} = \frac{(\text{TP}\times\text{TN})-(\text{FP}\times\text{FN})}{\sqrt{(\text{TP+FP})\times(\text{TP+FN})\times(\text{TN+FP})\times(\text{TN+FN})}}$

$P_r(a)$: relative observed agreement between the predicted classification of the model and the known classification; $P_r(e)$: hypothetical probability of chance agreement

statistical metrics explaining the quality of a classification model are given below and in Table 4.

In classification QSAR modeling, the compounds are classified into four main categories: a) true positives (TP), b) true negative (TN), c) false positive (FP), and d) false negative (FN) (Table 3). Researchers have used a variety of statistical tests to assess the classifier model performance and classification capability. Sensitivity (*Sn*) is the percentage of active compounds correctly predicted and is expressed as the ratio of true-positive results to the total number of positive data. Specificity (*Sp*) is the ratio of true-negative results to the total number of negative data. Accuracy (*Acc*) implies the fraction of correctly predicted compounds. The precision indicates the accuracy of a predicted class (ratio between the true positives and total positives) and *F*-measure refers to the harmonic mean of Recall (or Sensitivity) and Precision. Higher values for recall and precision give higher values for *F*-measure, thereby implying better classification.

G-means is a combination term that includes *Sn* and *Sp* into a single parameter merged via the geometric mean. This allows an easy assessment of the model's ability to distinguish between active or inactive samples.

Cohen's kappa (κ) can be utilized to determine the concordance between classification (predicted) models and known classifications (Cohen 1960). It is a measure of the degree of agreement. It returns value from − 1 (total disagreement) to 0 (random classification) to 1 (total agreement).

Mathews correlation coefficient (MCC) measures the quality of binary classifications and compares different classifiers. In any case, where the number of positive and negative compounds is not equal, the terms sensitivity, specificity, and accuracy are not reliable. MCC uses all four values (TP, TN, FP, and FN) and is directly calculated from the confusion matrix to provide a more-balanced prediction evaluation. Like Cohen's kappa, the value for MCC also ranges from − 1 to 1.

Prediction reliability detection tools

As discussed earlier, the process of QSAR modeling consists of three important steps: model development, model selection, and model interpretation. The model development process involves various feature selection practices including stepwise-multiple linear regression (S-MLR), genetic algorithm, genetic function approximation, etc. Model selection is based on the identification of the finest model (based on validation metric values) from a set of alternative models. When it comes to the reliability of QSAR/QSPR models, validation is essential. After a model has been selected, several internal and external validation metrics are assessed which help in demonstrating the actual predictive performance of the chosen model. Several groups of researchers in QSAR suggested external validation to be the gold standard in demonstrating the predictive ability of a model (Golbraikh and Tropsha 2002; Gramatica and Sangion 2016; Gramatica 2020). Multiple modeling in consensus form has been introduced to achieve a lower degree of predicted residuals for query compounds (Roy et al. 2015b; Khan et al. 2019a; Roy et al. 2019). In the following sections, we will discuss various tools from the DTC Laboratory (https://sites.google.com/site/kunalroyindia/home/qsar-model-development-tools) that help understand the prediction ability of one or more QSAR models.

(i) Double cross-validation (version 2.0) tool

The most common scheme of external validation is by introducing the hold-out method. Here, the original dataset is divided into training and test sets, where the training set is used for model-building purposes followed by model selection based on internal validation metrics, and the test set is used for model validation through external validation metrics. This approach ensures that the test set is never applied during the model-building procedure and it remains unseen by the developed model. However, a single training set does not confirm feature optimization, since a fixed training set composition leads to a bias in feature selection. This issue is more apparent in the case of MLR models than partial least-squares (PLS) or principal component regression (PCR) models which are more robust and generalized methods. Baumann and Baumann (Baumann and Baumann 2014) discussed the concept of double cross-validation (DCV) which Roy and Ambure implemented in a tool (Roy and Ambure 2016) where the training set is further divided into 'n' number of calibration and validation sets. The tool is freely available from http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/. The algorithm comprises two nested cross-validation loops (Bates et al. 2021), namely, the outer loop and the inner loop (Fig. 2). The outer loop consists of data points that are split arbitrarily into disjoint subsets known as training set compounds and test set compounds. The training set is utilized in the inner loop for model development and model selection, and the test set is used exclusively for checking model predictivity. The training set in the inner loop is further split into *k* number of calibration and validation sets in the inner loop by applying the *k*-fold cross-validation technique (Wainer and Cawley 2021). In the *k*-fold cross-validation method, the training data are initially segregated into k subsets followed by preparing k-iterations of calibration and validation sets. At each reiteration, different subset of data is excluded and used as validation set, while the remaining k-1 subsets are used as calibration sets. The data are passed through a stratification process, i.e., data rearrangement which helps

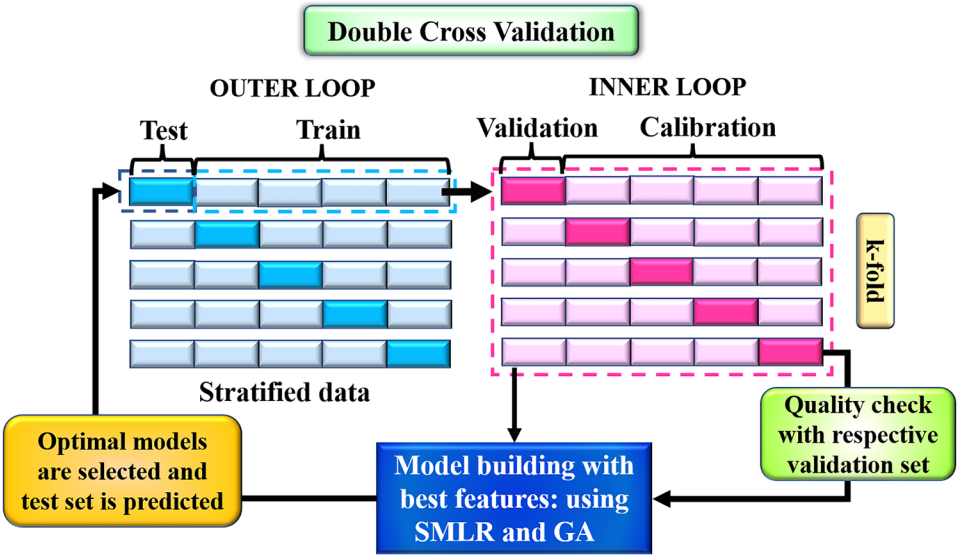

Fig. 2 Schematic diagram of double cross-validation algorithm (colour figure online)

maintain data uniformity (each fold is representative of the whole dataset). Each k-fold calibration set is then used to develop multiple linear regression (MLR) models, while the respective validation sets are applied to find the prediction errors. The tool provides two methods of feature selection: stepwise-multiple linear regression (S-MLR) (Maleki et al. 2014; Ojha and Roy 2018) and genetic algorithm-MLR (GA-MLR) (Leardi 2001). The prediction error is checked using mean absolute error ($MAE_{95\%}$) (Roy et al. 2016). There is also a provision for the generation of PLS models in the tool. Furthermore, the models in the inner loop are selected based on three major criteria as follows:

i) The models with the lowest MAE value (on the validation set) are selected.
ii) Consensus predictions of the top model are selected based on the MAE value of the validation set.
iii) Searching out the best descriptor combination from the top models.

Researchers found the DCV approach to be reliant and useful and thus successfully employed in various applications, for example, quantitative structure–property relationship (QSPR) modeling for sweetness potency of organic chemicals (Ojha and Roy 2018), developing nano-QSAR models for TiO_2-based photocatalysts (Mikolajczyk et al. 2018), inhalational toxicity modeling (Nath et al. 2022), modeling of diagnostic agents (De et al. 2019; De et al. 2020, 2022; De and Roy 2020, 2021), etc.

(ii) Intelligent consensus predictor tool

A well-validated QSAR model engages different classes of descriptors, which accentuate many features of molecular structures. Individual QSAR models may exaggerate a few important features, undervalue other features, and overlook some significant characteristics features. Roy et al. (2018b) proposed an "intelligent" selection of multiple models that would enhance the quality of predictions of query compounds (Roy et al. 2018b). This software helps judge the performance of consensus predictions compared to their quality obtained from the individual MLR models based on the MAE-based criteria (95%). The tool "Intelligent Consensus Prediction" is available from http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/. The tool takes multiple individual models (M1, M2, M3, etc.) as input derived using a different combination of descriptors from the training set. There are four ways of consensus prediction explained in the work:

(i) Consensus model 0 (CM0): it provides a simple average of predictions from all input individual models.
(ii) Consensus model 1 (CM1): it is the average of predictions from all individual qualified models. It is calculated from the arithmetic average of predicted response values attained from the '*n*' qualified models for test compounds rather than from all existing individual models.
(iii) Consensus model 2 (CM2): it is the weighted average prediction (WAP) from all qualified individual models. In CM2, the average is evaluated by giving a proper weightage to the qualified models for a particular test set compound.
(iv) Consensus model 3 (CM3): compound-wise best selection of predictions from qualified individual models. The best model for a particular test compound is selected based on its cross-validated mean absolute error (MAE_{CV}). A model with the lowest value MAE_{CV} is the best for a particular test set compound.

The tool further provides additional selection criteria which include:

(a) Euclidean distance cut-off: this is used to find a fitting model to predict the test set compound, where 10 most similar compounds are selected based on Euclidean Distance score. The user can set a Euclidean cut-off ranging from 0 to 1 to restrict the selection of only those training set compounds with a Euclidean distance score less than or equal to the set cut-off value.
(b) Applicability domain: AD helps to check whether the test/query compound is in the chemical space of the model or not. A simple standardization approach is used for AD determination.
(c) Dixon Q test: this test can be employed to spot and remove a response outlier out of selected similar training set compound.

The complete calculation method is demonstrated in the published article by Roy et al. and the methodology is given in Fig. 3. The ICP method has found good application in the prediction of pharmaceuticals (Khan et al. 2019a), organic chemicals and dyes (Roy et al. 2019; Khan and Roy 2019; Ghosh and Roy 2019; Ojha et al. 2020), determining aquatic toxicity (Hossain and Roy 2018), inhalational toxicity (Nath et al. 2022), polymer properties (Khan et al. 2018), etc.

(iii) Prediction Reliability Indicator tool

A QSAR model is developed based on the physicochemical features of an appropriately designed training set having experimentally derived response data. In contrast, the model is validated using one or more test set(s) for which the experimental response data are available. The suitability of this model for a completely new data set (true external set) for providing a reliable prediction is quite an interesting study.

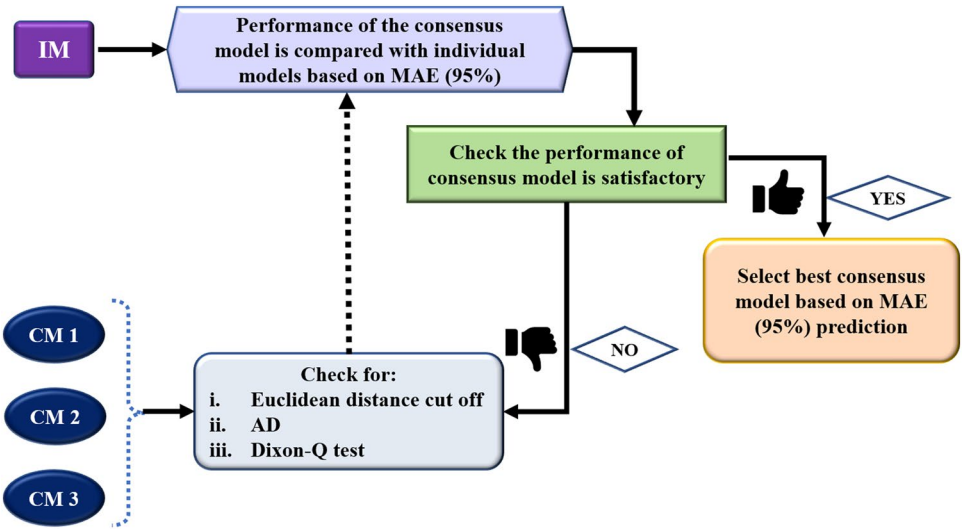

Fig. 3 "Intelligent Consensus Prediction" algorithm

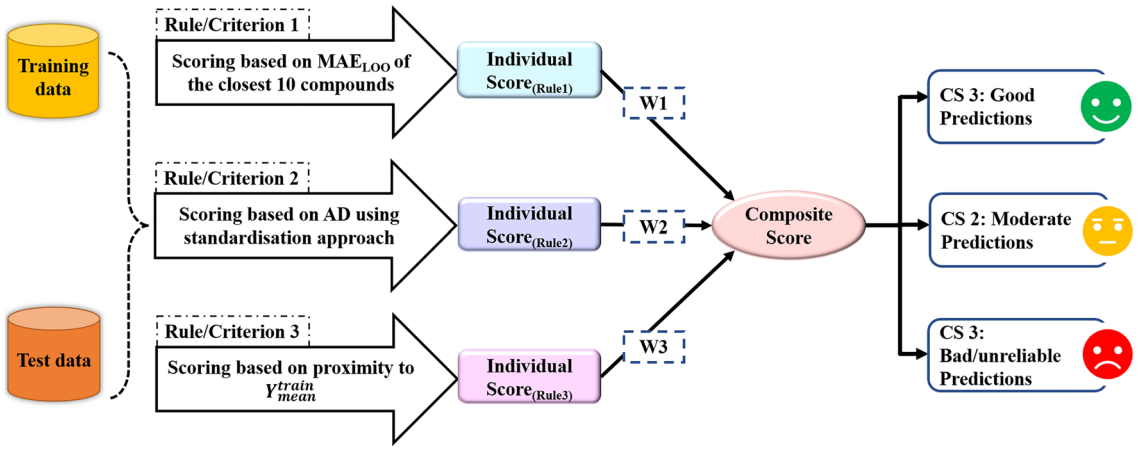

Fig. 4 Methodology applied for scoring test/query compounds in "Prediction Reliability Indicator" tool

Roy et al. (2018a, b) have developed a new scheme (Fig. 4) to define the reliability of predictions from QSAR models for new query compounds and implemented the method in a new tool called "Prediction Reliability Indicator" freely available from http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/. This tool is applicable for predictions from MLR and PLS models. The work aimed at formulating a set of rules/criteria that will ultimately empower the user to estimate the quality of predictions for individual test (external) compounds. Prediction of test/external sets can have varying quality. It might not be good predictions in all cases, while the model can show moderate to bad/unreliable predictions for some of the external set compounds. By keeping the variation of prediction quality, the authors have hypothesized three rules/criteria which might assist in classifying the quality of predictions for individual test/external set compounds into good, moderate, and poor/unreliable ones. We have now discussed the three rules briefly in the following segment:

(a) Rule/criterion 1: the scoring is based on the quality of leave-one-out predictions of the closest 10 training compounds to a test/external compound. Here, 10 most similar compounds are identified for each test/query compound (based on Euclidean distance similarity), followed by which mean of absolute LOO prediction error (MAE_{LOO}) is calculated for the selected closest 10 compounds. For a test/query compound whose MAE_{LOO} is lowest corresponding to its closest training compounds is predicted well and gets the highest prediction score (Prediction Score = 3). Test/query compounds that have medium MAE_{LOO} values with corresponding close training compounds should get a moderate score (Prediction Score = 2), and those test compounds with corresponding close training compounds having high MAE_{LOO} values should get the least score (Prediction Score = 1). The MAE-based criteria (Roy et al. 2016) are applied here for scoring the compounds which involve MAE_{LOO} and standard deviation (σ_{LOO}) of the absolute prediction error values.

(b) Rule/criterion 2: scoring based on the similarity-based AD using standardization method. The applicability domain (AD) of a model plays an important role in identifying uncertainty in the prediction of a specific chemical (test/query) by that model. This is based on how similar is the test/query compound with those in the training set. When a test/query compound is similar to a small fraction or none of the training compounds, the prediction is considered unreliable or fails to fall under the training set AD. Here, a simple AD based on the standardization approach (Roy et al. 2015a, b) is employed.

(c) Rule/Criterion 3: scoring based on the proximity of predictions to the training set observed/experimental response mean. Earlier, the quality of fit or prediction of compounds is better when compounds possess experimental response values (training and test compounds) close to the training set observed response mean. Thus, in rule/criterion 3, the authors have proposed to assess the prediction quality of a test compound based on the closeness of predicted response value to the training set observed/experimental response mean. The predicted response value (Y_{pred}^{test}) of each test compound is first measured using the training set model, and then, this Y_{pred}^{test} value is compared with the training set experimental response mean (Y_{mean}^{train}) and the corresponding standard deviation (σ^{train}). The scoring is based on the following manner:

(i) A test compound with Y_{pred}^{test} value falling within the range inside $Y_{mean}^{train} \pm 2\sigma^{train}$, that is, $(Y_{mean}^{train} + 2\sigma^{train}) \geq Y_{pred}^{test} \geq (Y_{mean}^{train} - 2\sigma^{train})$, can be assumed to be well (good) predicted by the model and thus have a score 3.

(ii) A test compound with Y_{pred}^{test} value falling within the range $(Y_{mean}^{train} + 3\sigma^{train}) \geq Y_{pred}^{test} \geq (Y_{mean}^{train} - 3\sigma^{train})$ and $(Y_{mean}^{train} + 2\sigma^{train}) < Y_{pred}^{test} < (Y_{mean}^{train} - 2\sigma^{train})$ can be presumed to be predicted moderately by the model and thus gets a score 2.

(iii) A test compound with Y_{pred}^{test} value falling within the range $(Y_{mean}^{train} + 3\sigma^{train}) < Y_{pred}^{test} < (Y_{mean}^{train} - 3\sigma^{train})$ can be assumed to be predicted poorly by the model and thus gets a score 1.

Furthermore, after these three criteria are checked, a weighting scheme is employed to compute a composite score for judging the prediction quality of each test compound using all three individual scores. The composite score is defined as follows:

$$\text{Composite score} = W_1 \times \text{score}_{\text{rule1}} + W_2 \times \text{score}_{\text{rule2}} + W_3 \times \text{score}_{\text{rule3}}.$$

Here, $\text{score}_{\text{rule1}}$, $\text{score}_{\text{rule2}}$, and $\text{score}_{\text{rule3}}$ represent the scores obtained after applying respective rules, whereas $W1$,$W2$, $W3$ indicate the weightage (automatic or user-provided) given to each of the three individual scores. The PRI tool offers a unique composite score which can act as a marker of prediction quality of true external test compound. This tool has found application for the prediction of external set/query compounds in many areas, viz., endocrine disruptor chemicals (Khan et al. 2019b), metal oxide nanoparticles (De et al. 2018), organic chemicals (Khan and Roy 2019; Khan et al. 2019c; De et al. 2020; 2022; Nath et al. 2022), etc.

(iv) Small dataset modeler (version 1.0.0) tool

Various specialized datasets involving nanomaterials, properties of catalysts, radiosensitizer molecules, etc. have smaller number of data points where the division of data into training and test sets may not produce robust and predictive models. A small dataset with 25–50 compounds cannot be used for conventional double cross-validation as dividing the data set into training and test sets and further into calibration and validation sets is not possible. Ambure et al. have developed a new tool called the Small Dataset Modeler, version 1.0.0 (http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/) solely for small datasets which includes a double cross-validation approach to develop a model for a small number of data points without training and test sets division of the dataset (Ambure et al. 2019) (Fig. 5). Here, the whole input set (containing n number of compounds) goes into a loop where it is repeatedly split up into calibration and validation sets (same as in the inner loop of DCV). The best possible combinations (k) are tried to obtain using validation sets of r compounds and calibration sets of n–r compounds. The tool asks for the number of compounds (i.e., r) in the validation set from the user based on which all probable combinations of calibration and validation sets are produced. The Multiple Linear Regression (MLR) models are generated using the calibration set compounds employing the Genetic Algorithm-Multiple Linear Regression (GA-MLR) method (Devillers 1996; Venkatasubramanian and Sundaram 2002) of variable selection, while the validation sets are employed to judge the predictive ability of the models. Numerous important internal ($R^2, R^2_{adj}, Q^2_{LMO}, MAE_{LOO}, r^2_m$(LOO) metrics) and external ($Q^2_{F1}, Q^2_{F2}, r^2_m$(test), CCC, MAE_{test}) validation metrics are measured in the exhaustive DCV method for all the chosen models. The tool is designed in such a way that it also develops Partial Least Squares Regression (PLS-R) models based on the descriptors selected in MLR models. The final top model selection can be done in any five of the following recommended ways:

(i) Any model (MLR/PLS) with the smallest MAE (95%) in the validation set is chosen.
(ii) Any model (MLR/PLS) with the smallest MAE (95%) in the modeling set is chosen.
(iii) Any model (MLR/PLS) with the lowest $Q^2_{Leave-Many-Out}$ (modeling set) is chosen.
(iv) Implementing consensus predictions using the best models that are chosen depending on the MAE (95%) in the validation sets. Consensus predictions can be of two types: (a) Using a simple arithmetic average of predictions of the best models. (b) Using a weighted average of predictions (WAP) by assigning proper weights to the top chosen models depending on the mean abso-

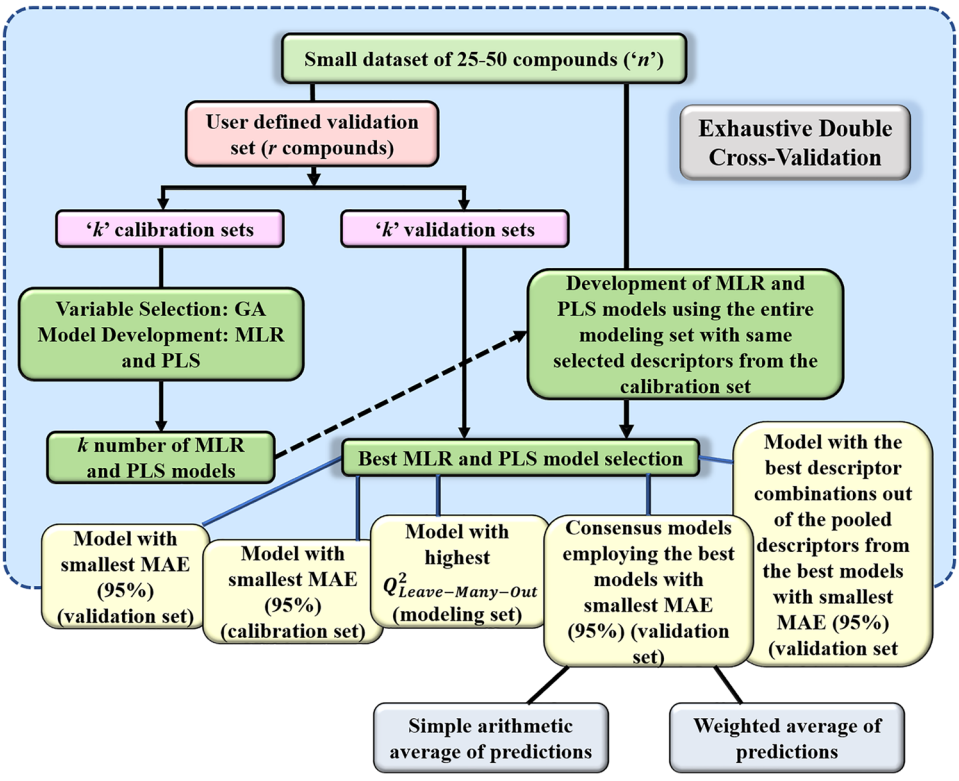

Fig. 5 Methodology behind the "Small Dataset Modeler" (version 1.0.0) tool to perform QSAR modeling for a small set of data points

lute error obtained from leave-one-out cross-validation,MAE_{cv}(95%).
(v) A pool of exclusive descriptors from the best models with the smallest MAE(95%) obtained from the validation set is again employed to build models. In the case of MLR, the best descriptor combinations are put through the *Best Subset Selection* method. However, in the case of a PLS model, descriptors nominated in the top models are pooled together for a PLS run.

The method proposed in the "Small Dataset Modeler" tool confirms internal divisions of small datasets within the DCV technique without taking any test set into account. The approach of "Small Dataset Modeler" tool integrates data curation, exhaustive DCV technique, and ideal modeling techniques entailing consensus predictions to develop models, principally for a small set of data points. The methodology behind the "Small Dataset Modeler" tool is schematically presented in Fig. 5. Small dataset modeling has found use in environmental toxicity modeling including acute toxicity of antifungal agents toward fish species (Nath et al. 2021) and soil ecotoxicity (Lavado et al. 2022), radiosensitization modeling (De and Roy 2020), modeling of Hepatitis C virus inhibitor protein (Ejeh et al. 2021), and modeling anesthetics causing GABA inhibition (Stošić et al. 2020).

(v) Read-Across-v3.1 tool

The read-across methodology has gained immense attention in recent years, because it is a non-testing approach that can be utilized for data-gap filling. The basic aim of the read-across technique is to predict endpoint information for one or more chemicals (i.e., the target chemicals) using data from the same endpoint from another substance (the source chemicals) using the similarity principle. The method is widely used as an alternative tool for hazard assessment to fill data gaps (ECHA 2011). Read-across based predictions seem to be more fitting for small data sets (limited source compounds). Hence, it has provided promising results in nanosafety assessment possessing limited data. Chatterjee and co-workers (2022) developed a new prediction-oriented quantitative read-across approach based on certain similarity principles. The reported work verifies the efficiency of the newly developed read-across algorithm in filling nanosafety data gaps. A tool has been developed to facilitate the implementation of the approach (Fig. 6) for quantitative read-across which is available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. The tool allows the users to optimize different hyperparameters including similarity kernel functions and distance and similarity thresholds to get the best quality of quantitative predictions. Mainly, three types of similarity estimation techniques were introduced involving Euclidean distance, Gaussian kernel function, and Laplacian kernel function. The algorithm developed in this study was optimized using three small nanotoxicity datasets ($n \leq 20$). The algorithm is based on two basic steps: (a) finding the 10 most similar training compounds for each query or test compound; (b) calculating the weighted average prediction of test set compounds from the most similar training set compounds. Different

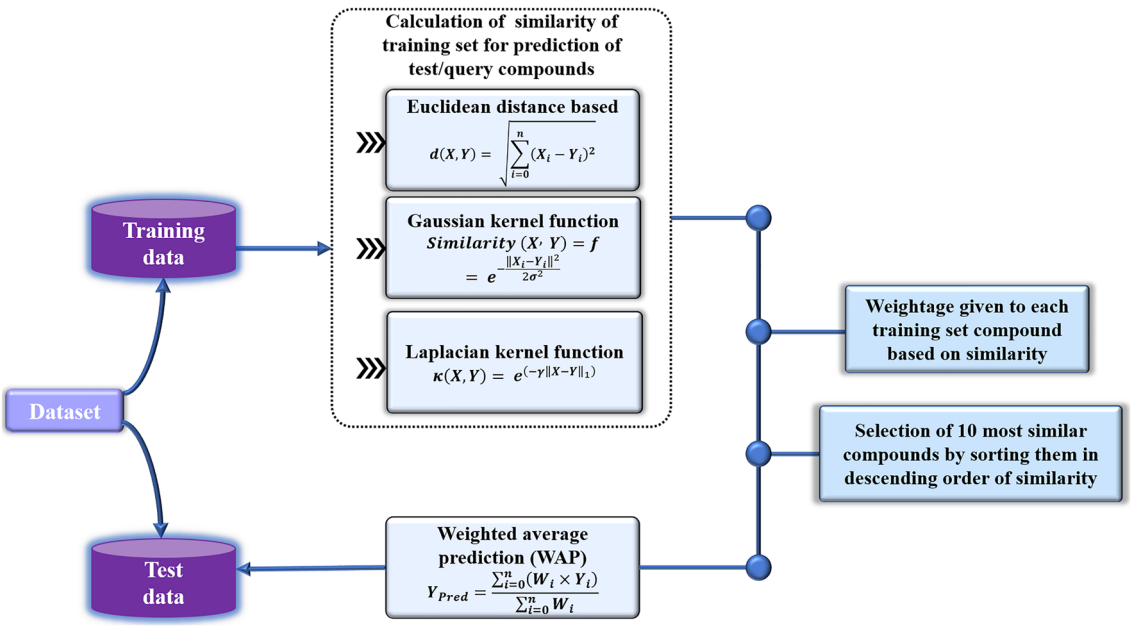

Fig. 6 Quantitative read-across algorithm

hyperparameters like sigma and gamma values in Gaussian and Laplacian kernel functions have been optimized. The effect of the number of close training compounds on the prediction quality has been evaluated; 2–5 close training compounds can efficiently predict the toxicity of query compounds. Another feature incorporation in the tool involves a distance threshold for the Euclidean distance similarity estimation and a similarity threshold for the Gaussian and Laplacian kernel function similarity estimations. This generated better prediction at the distance threshold of 0.4–0.5 and a similarity threshold of 0.00–0.05. This algorithm is easy to use, proficient, and an expert independent alternative method for the nanoparticle toxicity prediction which can further assist in data-gap filling and prioritization. Version 3.1 of this tool also computes classification-based validation metrics and generates receiver operating curve (ROC) for predictions which can be used to estimate the uncertainty of predictions. The tool is also applicable for several endpoints other than nanotoxicity, for example activity/toxicity/property of organic compounds in general.

Future perspectives

Over the past few decades, the QSAR methodology has received both praise and criticism in connection to its reliability, limitations, successes, and failures. The above discussion of the aforementioned tools from the DTC Laboratory provides methods and information relating to QSAR model development and validation, pointing out current trends, unresolved problems, and persistent challenges associated with evolution of QSAR. Furthermore, there are few scopes of further refining the present tools like inclusion of computation of Golbraikh and Tropsha's (Golbraikh and Tropsha 2002) criteria in the Double Cross Validation tool and computation of leave-many-out cross-validation (Q^2_{LMO}) criteria for both the Double Cross Validation tool and Small Dataset Modeler tool (PLS version), etc. Additionally, there is an opportunity to incorporate an uncertainty measure of predictions in the read-across tool which will improve the reliability for quantitative predictions of untested molecules.

Conclusion

The QSAR domain has been expanded substantially in the past few years as databases and their applications have grown. As the field of QSAR evolves through decades, it is necessary to evaluate the effectiveness of the QSAR models in predicting the behavior of new molecules. A QSAR model stands on the pillars of various validation metrics used to assess the quality of a predictive model that portrays the true picture of the prediction errors. The present review explains various internal and external validation metrics necessary for model predictivity assessment. Furthermore, a brief explanation of various innovative QSAR modeling tools developed by Drug Theoretics and Cheminformatics (DTC) laboratory (https://sites.google.com/site/kunalroyindia/home/qsar-model-development-tools) is given for better selection and development of models. These tools are aimed at addressing various features like selection of training set, model development methodology, model selection techniques, the use of multiple models, scoring of query compounds, etc. These improvisations helped in enhancing the quality of predictions of QSAR models. The tools highly assist in the reliability estimation of untested chemicals when experimental data are unavailable. However, most of these tools cannot be used for classification-based/graded data, but are well suited for quantitative models like MLR and PLS regression. Furthermore, the tools have a major role in different fields for predicting chemicals associated with the pharmaceutical industry, cosmeceuticals, polymer chemistry, diagnostic agents, dyes, nano-chemistry, food chemistry, etc.

Acknowledgements PD thanks Indian Council for Medical Research, New Delhi for Senior Research Fellowship.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdiscip Rev 2(4):433–459

Ambure P, Cordeiro MNDS (2020) Importance of data curation in QSAR studies especially while modeling large-size datasets. In: Roy K (ed) Ecotoxicol QSARs. Springer, New York, pp 97–109

Ambure P, Gajewicz-Skretna A, Cordeiro MND, Roy K (2019) New workflow for QSAR model development from small data sets: small dataset curator and small dataset modeler integration of data curation, exhaustive double cross-validation, and a set of optimal model selection techniques. J Chem Inform Model 59(10):4070–4076

Bates S, Hastie T, Tibshirani R (2021) Cross-validation: what does it estimate and how well does it do it? arXiv:210400673

Baumann D, Baumann K (2014) Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. J Cheminform 6(1):1–19

Chatterjee M, Banerjee A, De P, Gajewicz-Skretna A, Roy K (2022) A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. Environ Sci Nano 9(1):189–203

Chirico N, Gramatica P (2011) Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model 51(9):2320–2335

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46
Consonni V, Todeschini R (2010) Molecular descriptors Recent advances in QSAR studies. Springer, New York, pp 29–102
Consonni V, Ballabio D, Todeschini R (2009) Comments on the definition of the Q 2 parameter for QSAR validation. J Chem Inf Model 49(7):1669–2167
De P, Roy K (2018) Greener chemicals for the future: QSAR modelling of the PBT index using ETA descriptors. SAR QSAR Environ Res 29(4):319–337
De P, Roy K (2020) QSAR modeling of PET imaging agents for the diagnosis of Parkinson's disease targeting dopamine receptor. Theor Chem Acc 139:176
De P, Roy K (2021) QSAR and QSAAR modeling of nitroimidazole sulfonamide radiosensitizers: application of small dataset modeling. Struct Chem 32(2):631–642
De P, Kar S, Roy K, Leszczynski J (2018) Second generation periodic table-based descriptors to encode toxicity of metal oxide nanoparticles to multiple species: QSTR modeling for exploration of toxicity mechanisms. Environ Sci Nano 5(11):2742–2760
De P, Bhattacharyya D, Roy K (2019) Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease. Struct Chem 30(6):2429–2445
De P, Bhattacharyya D, Roy K (2020) Exploration of nitroimidazoles as radiosensitizers: application of multilayered feature selection approach in QSAR modeling. Struct Chem 31(3):1043–1055
De P, Bhayye S, Kumar V, Roy K (2022) In silico modeling for quick prediction of inhibitory activity against 3CLpro enzyme in SARS CoV diseases. J Biomol Struct 40(3):1010–1036
Dearden JC (2016) The history and development of quantitative structure-activity relationships (QSARs). Int J Quant Struct-Property Relat 1(1):1–44
Devillers J (1996) Genetic algorithms in molecular modeling. Academic Press, NY
Directorate E (2007) Environment health and safety publications series on testing and assessment No. 69, Guidance document on the validation of (quantitative) structure-activity relationships [(Q) SAR] models. OECD, Paris, France
ECHA (2011) The Use of Alternatives to Testing on Animals for the REACH Regulation. European Chemicals Agency Helsinki, Finland
Ejeh S, Uzairu A, Shallangwa GA, Abechi SE (2021) Computational insight to design new potential hepatitis C virus NS5B polymerase inhibitors with drug-likeness and pharmacokinetic ADMET parameters predictions. Future J Pharm Sci 7(1):1–13
Ghosh S, Ojha PK, Roy K (2019) Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs. Chemosphere 228:545–555
Ghosh K, Bhardwaj B, Amin S, Jha T, Gayen S (2020) Identification of structural fingerprints for ABCG2 inhibition by using Monte Carlo optimization, Bayesian classification, and structural and physicochemical interpretation (SPCI) analysis. SAR QSAR Environ Res 31(6):439–455
Golbraikh A, Tropsha A (2002) Beware of q2! J Mol Graph Model 20(4):269–276
Golmohammadi H, Dashtbozorgi Z, Acree WE Jr (2012) Quantitative structure–activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. Eur J Pharm Sci 47(2):421–429
Goodarzi M, Dejaegher B, Heyden YV (2012) Feature selection methods in QSAR studies. J AOAC Int 95(3):636–651
Gramatica P (2020) Principles of QSAR modeling: comments and suggestions from personal experience. IJQSPR 5(3):61–97
Gramatica P, Sangion A (2016) A historical excursus on the statistical validation parameters for QSAR models: a clarification concerning metrics and terminology. J Chem Inf Model 56(6):1127–1131
Héberger K, Rácz A, Bajusz D (2017) Which performance parameters are best suited to assess the predictive ability of models? Advances in QSAR Modeling. Springer, New York, pp 89–104
Höltje H-D, Sippl W (2001) Rational approaches to drug desing: proceedings of the 13th European symposium on quantitative structure-activity relationships, August 27-Setember, 1, 2000. JR Prous Science
Hossain KA, Roy K (2018) Chemometric modeling of aquatic toxicity of contaminants of emerging concern (CECs) in Dugesia japonica and its interspecies correlation with daphnia and fish: QSTR and QSTTR approaches. Ecotoxicol Environ Saf 166:92–101
Hsu H-H, Hsu Y-C, Chang L-J, Yang J-M (2017) An integrated approach with new strategies for QSAR models and lead optimization. BMC Genom 18(2):1–9
Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11(1):137–148
Khan K, Roy K (2019) Ecotoxicological QSAR modelling of organic chemicals against Pseudokirchneriella subcapitata using consensus predictions approach. SAR QSAR Environ Res 30(9):665–681
Khan PM, Rasulev B, Roy K (2018) QSPR modeling of the refractive index for diverse polymers using 2D descriptors. ACS Omega 3(10):13374–13386
Khan K, Benfenati E, Roy K (2019a) Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: ranking and prioritization of the DrugBank database compounds. Ecotoxicol Environ Saf 168:287–297
Khan K, Roy K, Benfenati E (2019b) Ecotoxicological QSAR modeling of endocrine disruptor chemicals. J Hazard Mater 369:707–718
Khan PM, Roy K, Benfenati E (2019c) Chemometric modeling of Daphnia magna toxicity of agrochemicals. Chemosphere 224:470–479
Konovalov DA, Coomans D, Deconinck E, Vander Heyden Y (2007) Benchmarking of QSAR models for blood-brain barrier permeation. J Chem Inf Model 47(4):1648–1656
Konovalov DA, Llewellyn LE, Vander Heyden Y, Coomans D (2008) Robust cross-validation of linear regression QSAR models. J Chem Inf Model 48(10):2081–2094
Kubinyi H, Hamprecht FA, Mietzner T (1998) Three-dimensional quantitative similarity—activity relationships (3d qsiar) from seal similarity matrices. J Med Chem 41(14):2553–2564
Lavado GJ, Baderna D, Carnesecchi E, Toropova AP, Toropov AA, Dorne JLC, Benfenati E (2022) QSAR models for soil ecotoxicity: development and validation of models to predict reproductive toxicity of organic chemicals in the collembola Folsomia candida. J Hazard Mater 423:127236
Leardi R (2001) Genetic algorithms in chemometrics and chemistry: a review. J Chemom 15(7):559–569
Likas A, Vlassis N, Verbeek JJ (2003) The global k-means clustering algorithm. Pattern Recognit 36(2):451–461
Maleki A, Daraei H, Alaei L, Faraji A (2014) Comparison of QSAR models based on combinations of genetic algorithm, stepwise multiple linear regression, and artificial neural network methods to predict K d of some derivatives of aromatic sulfonamides as carbonic anhydrase II inhibitors. Russ J Bioorganic Chem 40(1):61–75
Mikolajczyk A, Gajewicz A, Mulkiewicz E, Rasulev B, Marchelek M, Diak M, Hirano S, Zaleska-Medynska A, Puzyn T (2018) Nano-QSAR modeling for ecosafe design of heterogeneous TiO 2-based nano-photocatalysts. Environ Sci Nano 5(5):1150–1160
Nath A, De P, Roy K (2021) In silico modelling of acute toxicity of 1, 2, 4-triazole antifungal agents towards zebrafish (Danio rerio)

embryos: application of the small dataset modeller tool. Toxicol in Vitro 75:105205
Nath A, De P, Roy K (2022) QSAR modelling of inhalation toxicity of diverse volatile organic molecules using no observed adverse effect concentration (NOAEC) as the endpoint. Chemosphere 287:131954
Norinder U (1996) Single and domain mode variable selection in 3D QSAR applications. J Chemom 10(2):95–105
Novellino E, Fattorusso C, Greco G (1995) Use of comparative molecular field analysis and cluster analysis in series design. Pharm Acta Helv 70(2):149–154
Ojha PK, Roy K (2018) Development of a robust and validated 2D-QSPR model for sweetness potency of diverse functional organic molecules. Food Chem Toxicol 112:551–562
Ojha PK, Mitra I, Das RN, Roy K (2011) Further exploring rm2 metrics for validation of QSPR models. Chemometr Intell Lab Syst 107(1):194–205
Ojha PK, Kar S, Roy K, Leszczynski J (2020) Chemometric modeling of power conversion efficiency of organic dyes in dye sensitized solar cells for the future renewable energy. Nano Energy 70:104537
Organisation for Economic Co-operation and Development (OECD) (2004) The Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q) SARs] on the Principles for the Validation of (Q) SARs. Series on Testing and Assesment, p 206
Pope P, Webster J (1972) The use of an F-statistic in stepwise regression procedures. Technometrics 14(2):327–340
Roy K (2007) On some aspects of validation of predictive quantitative structure–activity relationship models. Expert Opin Drug Discov 2(12):1567–1577
Roy K (2018) Quantitative structure-activity relationships (QSARs): a few validation methods and software tools developed at the DTC laboratory. J Indian Chem Soc 95(12):1497–1502
Roy K, Ambure P (2016) The "double cross-validation" software tool for MLR QSAR model development. Chemom Intell Lab Syst 159:108–126
Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H (2012) Comparative studies on some metrics for external validation of QSPR models. J Chem Inf Model 52(2):396–408
Roy K, Kar S, Ambure P (2015a) On a simple approach for determining applicability domain of QSAR models. Chemom Intell Lab Syst 145:22–29
Roy K, Kar S, Das RN (2015b) Statistical methods in QSAR/QSPR A primer on QSAR/QSPR modeling. Springer, New York, pp 37–59
Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. Chemometr Intell Lab Syst 152:18–33
Roy K, Ambure P, Kar S (2018a) How precise are our quantitative structure–activity relationship derived predictions for new query chemicals? ACS Omega 3(9):11392–11406
Roy K, Ambure P, Kar S, Ojha PK (2018b) Is it possible to improve the quality of predictions from an "intelligent" use of multiple QSAR/QSPR/QSTR models? J Chemom 32(4):e2992
Roy J, Ghosh S, Ojha PK, Roy K (2019) Predictive quantitative structure–property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs). Environ Sci Nano 6(1):224–247
Rücker C, Rücker G, Meringer M (2007) y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model 47(6):2345–2357
Schüürmann G, Ebert R-U, Chen J, Wang B, Kühne R (2008) External validation and prediction employing the predictive squared correlation coefficient—test set activity mean vs training set activity mean. J Chem Inf Model 48(11):2140–2145
Stošić B, Janković R, Stošić M, Marković D, Stanković D, Sokolović D, Veselinović AM (2020) In silico development of anesthetics based on barbiturate and thiobarbiturate inhibition of GABAA. Comput Biol Chem 88:107318
Todeschini R (2010) Milano Chemometrics. University of MilanoBicocca, Milano, Italy (personal communication)
Todeschini R, Ballabio D, Grisoni F (2016) Beware of unreliable Q 2! A comparative study of regression metrics for predictivity assessment of QSAR models. J Chem Inf Model 56(10):1905–1913
Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK (2011) Validation of QSAR models-strategies and importance. Int J Drug Des Discov 3:511–519
Venkatasubramanian V, Sundaram A (2002) Genetic algorithms: introduction and applications. In: Encyclopedia of computational chemistry 2. Wiley, New Jersey
Wainer J, Cawley G (2021) Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Syst Appl 182:115222
White Paper on a Strategy for a Future Chemicals Policy. Commission of the European Communities. (2001) Brussels, Belgium
Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58(2):109–130