

Research article

Open Access

Chemometric QSAR Modeling and *In Silico* Design of Antioxidant NO Donor Phenols

Indrani MITRA¹, Achintya SAHA², Kunal Roy *¹

¹ Drug Theoretics and Cheminformatics Lab, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India.

² Department of Chemical Technology, University College of Science and Technology, University of Calcutta, 92, A. P. C. Road, Kolkata 700 009, India.

* Corresponding author. E-mail: kunalroy_in@yahoo.com (K. Roy)

Sci Pharm. 2011; 79: 31–57

doi:10.3797/scipharm.1011-02

Published: December 2nd 2010

Received: November 2nd 2010

Accepted: December 2nd 2010

This article is available from: <http://dx.doi.org/10.3797/scipharm.1011-02>

© Mitra, Saha and Roy; licensee Österreichische Apotheker-Verlagsgesellschaft m. b. H., Vienna, Austria.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

An acceleration of free radical formation within human system exacerbates the incidence of several life-threatening diseases. The systemic antioxidants often fall short for neutralizing the free radicals thereby demanding external antioxidant supplementation. Therein arises the need for development of new antioxidants with improved potency. In order to search for efficient antioxidant molecules, the present work deals with quantitative structure-activity relationship (QSAR) studies of a series of antioxidants belonging to the class of phenolic derivatives bearing NO donor groups. In this study, several QSAR models with appreciable statistical significance have been reported. Models were built using various chemometric tools and validated both internally and externally. These models chiefly infer that presence of substituted aromatic carbons, long chain branched substituents, an oxadiazole-N-oxide ring with an electronegative atom containing group substituted at the 5 position and high degree of methyl substitutions of the parent moiety are conducive to the antioxidant activity profile of these molecules. The novelty of this work is not only that the structural attributes of NO donor phenolic compounds required for potent antioxidant activity have been explored in this study, but new compounds with possible antioxidant activity have also been designed and their antioxidant activity has been predicted *in silico*.

Keywords

Antioxidants • Chemometric tools • Structure-activity relationships • Phenolic derivatives

Introduction

Free radicals (reactive oxygen species) like superoxide and hydroxyl radicals are generated as a result of partial reduction of molecular oxygen [1]. Free radicals are constitutively produced during various metabolic functions of the body. In addition to the lethal actions, they bear several beneficial effects also. The immune system utilizes these free radicals for detection of foreign invaders or damaged tissues that are needed to be eliminated from the human system [2]. However, excessive free radical production resulting from heavy exercise, exposure to environmental pollutants, smoking etc may endanger healthy livelihood through an aggravation of their deleterious effects. Recent research implicates a close association of the free radicals (reactive oxygen species accumulating within the human system) with the etiology and/or progression of a number of diseases as well as aging [3]. Most of the fatal degenerative diseases like Parkinson's disease [4], atherosclerosis involving cardiovascular damage [5] etc have their origin from the deadly effects of these toxic free radicals. The free radicals are also involved in DNA damage [6], induction of lipid peroxidation in cell membranes and inactivation of membrane-bound enzymes.

The free radical attack to the human system can be controlled to a large extent through utilization of antioxidants. Antioxidants [7] are molecules which can safely interact with free radicals and terminate the chain reaction before vital molecules are damaged. To prevent free radical damage, the body has a defense system of antioxidants. But this endogenous antioxidant supply falls short under conditions of excessive oxidative stress. Although there are several enzyme systems within the body that scavenge free radicals, the principle micronutrient (vitamin) antioxidants are vitamin E, beta-carotene and vitamin C [8]. Epidemiologic observations show lower cancer rates in people whose diets are rich in fruits and vegetables suggesting that such diets rich in antioxidants protect the human system against the development of cancer. Antioxidants are also thought to have a role in slowing the aging process and preventing heart disease and strokes. At the molecular level, the antioxidant mechanism of action can be explained based on the electron–proton transfer theories: (a) hydrogen atom transfer (HAT), (b) single-electron transfer–proton transfer (SET-PT) and (c) sequential proton loss electron transfer (SPLET) [9].

The structural features and properties of a molecule determine its biological activity profile. Quantitative structure-activity relationship (QSAR) is a method of studying a series of molecules of different structures with varying observed properties and attempting to find empirical relationships between structure and property or activity [10]. Starting from the period of Hansch [11], QSAR has been widely used for lead optimization and drug discovery process. This technique has also been used by several researchers for designing of newer antioxidant molecules with improved activity. Rastija et al. [12] modeled antioxidant activity of wine polyphenols using QSAR technique with descriptors calculated from 2D and 3D representation of the molecules and inferred that arrangement of free hydroxyl groups on the flavonoid skeleton, or on the phenolic ring together with the shape, size, mass and steric properties of the molecules bear considerable effects on the activity profile of these molecules. Ray et al. [13] performed QSAR modeling using

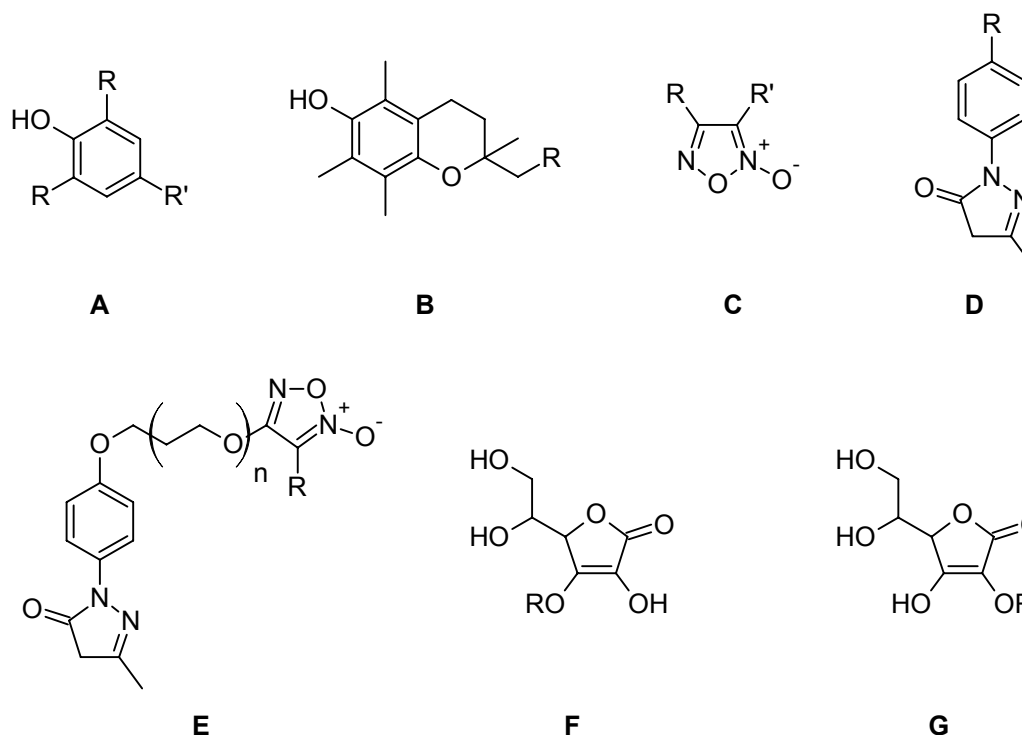
electrotopological state atom (E-state) parameters in order to determine the antiradical properties of flavonoids as studied in a methanolic solution of DPPH (2,2-diphenyl-1-picrylhydrazil) and the antioxidant activity of flavonoids in a beta-carotene-linoleic acid model system and revealed the importance of the substituent effect and structural changes for optimal antioxidant activity of the flavonoids. In order to determine the key chemical features imparting antioxidant activity to this class of molecules, Mitra et al. [14] performed pharmacophore mapping of arylamino-substituted benzo[*b*]thiophenes as free radical scavengers. Various QSAR models of antioxidant molecules have been recently reviewed [15].

Viewing the immense utility of antioxidants for fighting the array of present day diseases, in the present work, an attempt has been made to develop models capable of assessing the structural attributes of a series of molecules required for exhibiting potent antioxidant activity. A series of phenolic compounds with NO donor functions in the molecular structure having significant antioxidant activity was used for QSAR model development in the present work. Besides internal validation, the models developed were validated externally using compounds not included in the model development process. A comparison of the developed QSAR models with a previously reported model for this class of phenolic derivatives has also been performed in the present work. It may be noted that in the previous QSAR report, lower number of compounds were used for model development than those considered in the present work. Based on the QSAR models developed here, a new series of compounds has been designed and their possible antioxidant activity has been predicted *in silico*. The novelty of this QSAR study is not only that the structural requirements for antioxidant activity have been explored in this work, but the developed models have also been used for design of new molecules with possible potent antioxidant activity.

Materials and methods

The dataset

The data used for this analysis has been collected from the reports of Boschi et al. [16], Chegaev et al. [17] and Cena et al. [18]. The dataset comprises of 33 phenolic compounds, most of them bearing the NO donor functions, exhibiting a wide range of antioxidant activity. The antioxidant activities of the compounds were reported to be measured using the TBARS (Thiobarbituric acid reactive substance) assay method. For the present work, the IC₅₀ (50% inhibitory concentration) values of the compounds were expressed in millimolar units and converted to negative logarithmic scale (pIC₅₀). The observed and calculated/predicted activities of the compounds together with their structures are summarized in Tab. 1.

Tab. 1. Molecular structure together with the observed and predicted activity data of the 33 phenolic derivatives.

Compd. No. / Structure	R	R' / n	pIC ₅₀ [log(1/IC ₅₀)] [16–18]	Activity predicted ^a	Activity predicted ^b
1	A	H	0.538	-0.405	0.639
2	A	OCH ₃	1.745	1.760	1.700
3	A	<i>t</i> -Bu	2.770	3.039	2.732
4*	B	H	3.770	3.475	3.451
5*	C	OEt	0.959	1.403	1.367
6	A	H	0.845	1.275	0.932
7	A	OCH ₃	2.229	2.434	2.180
8	A	<i>t</i> -Bu	2.699	2.583	2.817
9	B		3.824	4.064	3.452
10	A	H	0.733	0.780	0.910
11*	A	OCH ₃	2.268	2.031	2.057
12	A	<i>t</i> -Bu	2.585	2.408	2.688
13*	A	H	1.328	1.277	1.132

Tab. 1. (Cont.)

Compd. No. / Structure	R	R' / n	pIC ₅₀ [log(1/IC ₅₀)] [16–18]	Activity predicted ^a	Activity predicted ^b
14 A	OCH ₃		2.469	2.676	2.310
15 A	<i>t</i> -Bu		2.699	2.490	3.039
16 B		–	3.310	3.649	3.448
17 A	<i>t</i> -Bu		2.921	2.519	3.218
18 B		–	3.854	4.063	3.626
19 D	H	–	1.770	2.021	1.626
20* D	OCH ₃	–	1.699	2.006	2.170
21* D		–	2.469	2.254	2.171
22 D		–	2.678	2.964	2.280
23* D		–	2.538	2.532	2.348
24* E	Ph	1	2.886	3.023	2.936
25 E	SO ₂ Ph	1	2.420	2.617	2.341
26 E	CONH ₂	0	2.102	1.810	2.169
27 E	CN	0	2.237	2.004	2.254
28 F		0.343	-0.128	0.829	
29 F		1.097	1.327	0.803	
30 F		1.553	2.115	0.869	
31 F		1.770	2.074	1.361	
32 F		1.097	1.422	0.865	
33 G		0.407	-0.914	1.548	

* Compounds selected as test set based on *k*-means clustering;^{a,b} Activity predicted (LOO predicted for the training set) based on Eqs. 8 and 10, respectively.

Descriptor calculation

The molecular structures of the compounds were drawn using the ACD Lab software [19] and were exported to the Cerius2 software [20] for the calculation of descriptors. Initially, conformational analysis of the molecules was performed using 'optimal search method' within the Cerius2 software. This method allows the automatic selection of the best method to generate the lowest-energy conformers for structures in the study table. This selection is done among the three methods (grid scan, random sampling and Boltzmann jump) available for conformer generation in the Cerius2 software. Grid scan [20] method performs a simple systematic search in which each specified torsion angle is varied over a grid of equally spaced values. Random sampling [20] perturbs the starting conformation of a structure by randomly altering values of all variable torsion angles and the Boltzmann jump [20] method involves random change of the torsion angles of a molecule within a specified angle window. The lowest energy conformers were energy minimized using the smart minimizer under open force field and the subsequent charge calculation of the lowest energy conformer was performed using Gasteiger method [20]. Followed by this, descriptors belonging to different categories were calculated using the Descriptor+ module of the Cerius2 software version 4.1 [20] (listed in Tab. 2). The calculated topological indices include descriptors like Wiener, Zagreb, Balaban J, connectivity indices ($^0\chi$, $^1\chi$, $^2\chi$, $^3\chi_P$, $^3\chi_C$, $^0\chi^v$, $^1\chi^v$, $^2\chi^v$, $^3\chi^v_P$, $^3\chi^v_C$), kappa shape indices ($^1\kappa$, $^2\kappa$, $^3\kappa$, $^1\kappa_{am}$, $^2\kappa_{am}$, $^3\kappa_{am}$) and E-state parameters. Besides these, spatial (Jurs charged partial surface area descriptors and shadow indices), structural, physicochemical and electronic descriptors were also calculated [20]. After excluding those descriptors having variance lower than 0.0001, a total of 86 descriptors were chosen. The values of the significant descriptors for all 33 compounds are given in supplementary materials (Tab. S1). Initially a QSAR model was built based on the entire dataset of 33 compounds. Considering the small size of the dataset, full leave-one-out cross-validation [21] has been performed for the model. This was followed by splitting of the dataset into training and test sets for further validation and determination of the external predictive ability of the derived models.

Tab. 2. List of descriptors used for present work.

Category of descriptors	Descriptor type
Topological indices	Wiener, Zagreb, Balaban, connectivity indices, kappa shape indices, E-state parameters
Structural	Hbond acceptor, Hbond donor, Rotlbonds, Chiral centers
Thermodynamic	LogP, AlogP, AlogP98, Molar refractivity
Spatial	Jurs descriptors, Shadow indices, Radius of Gyration, Molecular surface area, Density, Principal moment of inertia, Molecular volume.
Electronic	Dipole moment, HOMO (Highest occupied molecular orbital energy), LUMO (Lowest unoccupied molecular orbital energy), Superdelocalizability (Sr).

Selection of training set

The training set was utilized for the development of the QSAR model while the test set was used for the external validation purpose. The selection of the training and test sets serves

as a critical step in the QSAR model development process. The selection of the training set should be such that it captures all the features and characteristics of the whole set of molecules. It should also span the activity range of the entire dataset. For the present work, the selection of the training set was done based on the *k*-means clustering technique. Cluster analysis [22] is a method of arranging objects into groups. It divides objects into groups in such a manner that the degree of association between two objects is maximum if they possess same group and otherwise minimum. There are two types of clustering techniques: (a) hierarchical and (b) non-hierarchical. *k*-Means clustering is one of the best known non-hierarchical clustering techniques [22]. In this method, clustering starts randomly and then cluster means are calculated in the descriptor space. Molecules are reassigned to clusters whose means are closer to the position of the molecules. In the present work, clustering was performed with the standardized descriptor matrix using about 25% of the whole dataset compounds as the test set and the remaining as the training set.

Chemometric tools

Stepwise multiple linear regression (MLR) technique was used for the QSAR model development using the entire dataset. Stepwise MLR method is based on forward selection and backward elimination techniques for inclusion and rejection of descriptors. The selection of the significant descriptors for developing the model was done according to the 'stepping criteria' [23] (*F*) with *F* = 4 for inclusion and *F* = 3.9 for exclusion. The *F*-value used for inclusion or exclusion of a variable in the stepwise regression process is a test for partial regression coefficient and it is obtained by dividing the difference between reductions of sum of squares with and without the variable being included or excluded with error mean square of the equation [23]. The *F*-value for inclusion or exclusion of a variable in a MLR equation during stepwise process is square of the *t*-value of the regression coefficient of the variable being included or excluded.

For the development of the QSAR models using the training set data, two different chemometric tools were employed, viz., GFA (genetic function approximation) and G/PLS (genetic partial least squares). A genetic algorithm (GA) is a search technique [24] employed as a computational tool to find out exact or approximate solutions to optimization and search problems. Genetic function approximation was originally conceived from: (i) genetic algorithm originally developed by Fraser and others (later popularized by Holland) and (ii) Friedman's multivariate adaptive regression splines (MARS) algorithm. In this technique, an initial population of equations is built by random selection of descriptors followed by cross over between pairs of those equations. The progeny equations thus built are again subjected to cross over and the fitness of the final equations is assessed based on the lack of fit (LOF) value (given by Eq. 1). The model quality improves as the value of LOF diminishes.

$$\text{Eq. 1. } LOF = \frac{LSE}{\left(1 - \frac{c + d \times p}{m}\right)^2}$$

Here, LSE is the least square error, *c* is the number of basis functions, *d* is the smoothing parameter which was set at the default value of 1, *p* is the number of descriptors and *m* is

the number of observations in the training set. In effect, 'd' is the user's estimate of how much detail in the training data set is worth modeling. Smaller equations are obtained for larger values of 'd'. Since the GFA technique builds a population of equations, the range of variations in this population gives added information on the quality of fit and importance of the descriptors. GFA builds models not only with linear polynomials but also uses higher-order polynomials, splines and other nonlinear functions.

G/PLS technique [25] is the combination of (i) genetic function approximation and (ii) partial least squares methods. These are valuable analytical tools for QSAR modeling where number of descriptors is more than samples. The variables are selected using the GFA technique and the PLS regression method is used to weigh the relative contributions of the selected variables in the final model. Application of G/PLS allows the construction of larger QSAR equations while avoiding overfitting and eliminating most variables. Moreover the PLS technique takes into consideration a large number of noisy and collinear variables. Additionally, PLS provides a description of the available data using minimum number of adjustable parameters and consequently, maximum precision and stability of regression model is achieved using this technique.

Model quality

Various statistical parameters are calculated in order to assess the fitness of the developed model. The correlation coefficient, R, measures how closely the observed data tracks the fitted regression line and thus helps to quantify any variation in the calculated data with respect to the observed data. The F statistic, calculated from R^2 and the number of data points, determines the statistical significance of the regression equation at specified degrees of freedom (df). Other statistical parameters used to test the quality of generated regression equations include the standard error of the estimate (s) and adjusted R^2 (R_a^2) [23]. Although the value of R^2 increases with the addition of descriptors to the developed QSAR model but this may not necessarily indicate that the predictive ability of the model improves. Thus, to check the predictive potential of the developed models, internal and external validation experiments are performed on them.

Model validation

The QSAR models thus developed were validated using both internal and external validation techniques. In case of internal validation, the predictive ability of the models is judged based on the training set compounds. On the contrary, external validation deals with a new set of compounds which are not included in the QSAR model building process. Hence, the latter technique measures the ability of the model to predict the activity of a new series of compounds.

Internal validation

This technique involves calculation of cross-validated squared correlation coefficient (Q^2) (Eq. 2) and predicted residual error of sum of squares (PRESS) [23] based on the observed and predicted activity data of the training set molecules. In the present work, leave-one-out (LOO) cross-validation technique was used for determination of Q^2 . For the calculation of LOO- Q^2 , each of the compounds of the training set is deleted once and models are built with the remaining compounds. The activity of the deleted compound is thus predicted using the model developed. The cycle is repeated until all the compounds

are deleted at least once and the predicted activity data obtained for all the training set compounds are used for the calculation of above mentioned internal validation parameters.

$$\text{Eq. 2. } Q^2 = 1 - \frac{\sum (Y_{obs(train)} - Y_{pred(train)})^2}{\sum (Y_{obs(train)} - \bar{Y}_{training})^2}$$

Here, $Y_{obs(train)}$ is the observed activity, $Y_{pred(train)}$ is the predicted activity and $\bar{Y}_{training}$ is the mean observed activity of the training set compounds. A model is considered to be satisfactory if the value of Q^2 exceeds the stipulated value of 0.5.

External validation

The value of Q^2 signifies the ability of the model to predict the activity of molecules which are very much alike the training set ones. But to determine the predictive potential of the QSAR model for a new set molecules differing in some aspects from the training set ones, external validation is needed to be performed. In this case, the predictive capacity of a model is judged by its application for prediction of activity values of the test set compounds and subsequent calculation of Q^2_{ext} , i.e., predictive R^2 (R^2_{pred}) [26] value as given by Eq. 3:

$$\text{Eq. 3. } Q^2_{ext} = R^2_{pred} = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \bar{Y}_{training})^2}$$

In the above equation, $Y_{obs(test)}$ and $Y_{pred(test)}$ are the observed and predicted activity data of the test set compounds. A value of R^2_{pred} (given by Eq. 3) greater than the stipulated value of 0.5 reflects efficient prediction for the test set molecules by the developed model.

Calculation of r_m^2 metrics

It can be inferred from Eq. 3 that the value of the external predictive parameter (R^2_{pred}) primarily depends on the mean activity value of the training set compounds and its distance from the activity values of the test set compounds. Now, since the value of R^2_{pred} is dependent on the sum of squared differences between the observed activity data of the test set compounds and the training set mean, the value of R^2_{pred} increases as these differences for individual compounds increase. Thus, compounds with a wide range of activity data may exhibit a large value for this parameter, but this may not indicate that the predicted activity values are very close to those observed. In such a case, there remains a considerable difference between these values although they maintain a good overall correlation. Thus, to obviate this error and to better indicate the model predictive ability, the r_m^2 metrics [27] with threshold values of 0.5 (Eq. 4) were calculated.

$$\text{Eq. 4. } r_m^2 = r^2 \times \left(1 - \sqrt{(r^2 - r_0^2)}\right)$$

In Eq. 4, r^2 and r_0^2 are the squared correlation coefficient values between the observed and predicted activity data [LOO predicted activity of training set compounds in case of $r_{m^2(\text{LOO})}$ and the predicted activity of the test set compounds in case of $r_{m^2(\text{test})}$] with and without intercept respectively. As the above equation (Eq. 4) suggests, the value of r_m^2 depends solely on the observed and predicted activity data of the molecules and hence, any large deviation between these values will be reflected through the r_m^2 parameter. Similarly, based on the predicted activity values of both the training and test sets, values of $r_{m^2(\text{overall})}$ [27] were calculated. The parameter r_m^2 has been used by different groups of authors to check external predictability of QSAR models [28].

Randomization tests

Validation of the developed models was also performed using the randomization or Y-scrambling technique. In this technique, the Y column (activity data) is permuted keeping the remaining X matrix (descriptors) unchanged. Thereafter, models are built based on this scrambled matrix and average squared correlation coefficient of the randomized models (R_r^2) was calculated. Two types of randomization were performed in the present work, namely, process and model randomization. In case of process randomization, the entire descriptor matrix was used and scrambling of data was done using the total pool of descriptors at 90% confidence level. This technique ensures the reliability and robustness of the process employed for the development of the QSAR model. In addition to this, model randomization was also performed at 99% confidence level using the descriptors occurring in the respective models in order to verify whether the developed QSAR model was the outcome of a chance correlation or not. Values of R_r^2 lower than those of R^2 for the respective model signify a robust model. However, since no guideline is given as to how much this difference should be, another parameter, R_p^2 (threshold value=0.5) [27, 29] was calculated (Eq. 5). This parameter penalizes the model R^2 for small differences between the values of R^2 and R_r^2 . Thus, models having an acceptable value for this parameter (>0.5) are considered to be robust enough and are not obtained merely by chance.

$$\text{Eq. 5.} \quad R_p^2 = R^2 \times \sqrt{(R^2 - R_r^2)}$$

However, in an ideal case, the average value of R_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, the corrected formula of R_p^2 (${}^cR_p^2$) as proposed by Todeschini [29] is given as (Eq. 6):

$$\text{Eq. 6.} \quad {}^cR_p^2 = R \times \sqrt{(R^2 - R_r^2)}$$

Applicability domain

The domain of applicability constitutes an important concept in QSAR analysis that enables estimation of uncertainty in the prediction of a particular molecule based on its similarity to the compounds used for developing the model [30, 31]. It refers to a chemical space as defined by the molecular descriptors and the modeled response. A QSAR model exhibits reliability in prediction only for molecules lying within this defined chemical space

referred to as its applicability domain. Thus, for a dissimilar compound lying outside the domain of applicability, reliable prediction of activity becomes unlikely. Consequently, a QSAR should only be used for making predictions of molecules within the specified domain by interpolation thereby enabling avoidance of any unjustified extrapolation for activity prediction. The need to characterize the model applicability domain is also reflected in the OECD guidelines for QSAR model validation [32, 33]. In the present work, applicability domain of the best model selected according to the $r_m^2(\text{overall})$ criterion has been assessed. Since the model has been developed based on the G/PLS technique, the DModX method [25] implemented in the SIMCA software [34] has been utilised for detecting the applicability domain of the developed model. In this technique, the residuals of Y and X are used as diagnostic values for ensuring the quality of the model [25]. The standard deviation (SD) of the X-residuals of the corresponding row of the residual matrix E is proportional to the distance between the data point and the model plane in X-space, often called DModX (distance to the model in X-space). Here, X is the matrix of predictor variables, of size $N \times K$ [where, N is number of objects (cases, observations) and k is the index of X-variables ($k=1, 2, \dots, K$)], Y is the matrix of response variables of size $N \times M$ [m is the index of Y-variables ($m=1, 2, \dots, M$)] and E is the $N \times K$ matrix of X-residuals. A DModX value larger than around 2.5 times the overall SD of the X-residuals (corresponding to an F-value of 6.25) indicates that the observation is outside the applicability domain of the model [25].

Results and discussion

Initially, an attempt was made to develop a QSAR model using stepwise regression applied on the whole dataset. This was followed by division of the dataset into training and test sets. Models were developed based on the training set and the developed models were used for prediction of test set activity. Using two different chemometric techniques (GFA and G/PLS), three types of QSAR models were developed based on different combination of descriptors: (a) models developed with topological, structural and thermodynamic descriptors, (b) models developed with spatial, electronic and thermodynamic descriptors and (c) models developed using combined set of descriptors. All the significant models developed in the present work are summarized in Tab. 3. The critical F values at different degrees of freedom at 98% significance level are given at the end of Tab. 3. The results infer that since the F value for each of the QSAR models developed is higher than the corresponding critical value, all the developed models are statistically significant. However, among all the developed models, models developed with the spatial, electronic and thermodynamic descriptors are of poor statistical quality in comparison to the other two types of models. The GFA models were developed using 5000 iterations considering both linear and spline options. The models thus developed are nonlinear, and the spline terms are expressed as truncated power splines and denoted with angular brackets. E. g. $\langle f(x) - a \rangle$ is equal to zero if the value of $f(x) - a$ is negative, else it is equal to $f(x) - a$. The constant 'a' is called the knot of the spline. G/PLS was performed with 1000 iterations, scaled variables and with the option of no fixed length of equation. The maximum number of components or latent variables (LVs) fixed for variable selection was 3. These components are the functions of the original descriptors and they encode data as represented by the descriptors. Following the model development step, new compounds were designed *in silico* based on the information available from all the developed models (*vide infra*). The activities of all the newly designed compounds were predicted using all the developed QSAR models and their consensus activity values were reported (Tab. 4).

Tab. 3. Comparison of the statistical quality of the various QSAR models developed in the present work.

Using topological, structural and thermodynamic descriptors										
Mod.	Stat.	Eq.	Descriptors	LVs	n _{train.}	s	R ²	R ² _a	F*	PRESS
A1	GFA-linear	–	SC-0, S _{aa} CH, S _{dss} C, S _{dO}	–	25	0.369	0.889	0.867	40.07	4.756
A2	GFA-spline	8	<SC-3_P-20>, ³ X _p , <1.79401-S _s CH ₃ >, S _{aas} C	–	25	0.315	0.919	0.903	56.95	3.021
A2a	GFA-spline	9	³ X _p , (³ X _p) ² , <1.79401-S _s CH ₃ >, S _{aas} C	–	25	0.369	0.889	0.867	40.17	4.557
A3	G/PLS-linear	–	S _{aa} CH, S _{aas} C, S _{ds} N, S _s OH, MolRef	2	25	0.401	0.856	0.843	65.48	5.150
A4	G/PLS-spline	10	<6.68154- ¹ χ>, ³ X _p , <1.98556-S _s CH ₃ >, S _{aas} C	2	25	0.323	0.906	0.897	106.44	3.022
Mod.	Stat.	Eq.	Descriptors	Q ²	r _m ² (LOO)	n _{test}	R ² _{pred}	r _m ² (test)	r _m ² (overall)	
A1	GFA-linear	–	SC-0, S _{aa} CH, S _{dss} C, S _{dO}	0.806	0.676	8	0.859	0.839	0.685	
A2	GFA-spline	8	<SC-3_P-20>, ³ X _p , <1.79401-S _s CH ₃ >, S _{aas} C	0.877	0.757	8	0.924	0.899	0.777	
A2a	GFA-spline	9	³ X _p , (³ X _p) ² , <1.79401-S _s CH ₃ >, S _{aas} C	0.814	0.677	8	0.917	0.818	0.711	
A3	G/PLS-linear	–	S _{aa} CH, S _{aas} C, S _{ds} N, S _s OH, MolRef	0.790	0.771	8	0.879	0.887	0.790	
A4	G/PLS-spline	10	<6.68154- ¹ χ>, ³ X _p , <1.98556-S _s CH ₃ >, S _{aas} C	0.877	0.870	8	0.884	0.812	0.872	

Tab. 3. (Cont.)

Using spatial, electronic, and thermodynamic descriptors									
Mod.	Stat.	Descriptors	LVs	$n_{\text{train.}}$	s	R^2	R_a^2	F*	PRESS
B1	GFA-linear	MR, Jurs-TASA	–	25	0.443	0.824	0.808	51.55	5.271
B2	G/PLS-linear	MR, Jurs-SASA, Jurs-PPSA-3, Jurs-TASA	3	25	0.434	0.839	0.816	36.41	5.516
B3	G/PLS-2, spline	<55.0428-MR>, Jurs-PNSA-121.354-Jurs-WNSA-1>, Jurs-WPSA-3, <472.813-Jurs-TASA>	3	25	0.397	0.865	0.846	44.88	4.901
Mod.	Stat.	Descriptors	Q^2	$r_m^2(\text{LOO})$	n_{test}	R_{pred}^2	$r_m^2(\text{test})$	$r_m^2(\text{overall})$	
B1	GFA-linear	MR, Jurs-TASA	0.785	0.645	8	0.754	0.683	0.639	
B2	G/PLS-linear	MR, Jurs-SASA, Jurs-PPSA-3, Jurs-TASA	0.775	0.754	8	0.773	0.661	0.775	
B3	G/PLS-2, spline	<55.0428-MR>, Jurs-PNSA-121.354-Jurs-WNSA-1>, Jurs-WPSA-3, <472.813-Jurs-TASA>	0.800	0.787	8	0.678	0.525	0.761	
Using combined descriptors									
Mod.	Stat.	Descriptors	LVs	$n_{\text{train.}}$	s	R^2	R_a^2	F*	PRESS
C1	GFA-linear	$^3\chi_c^v$, Zagreb, S_aaCH, Jurs-RPSA	–	25	0.284	0.935	0.921	71.35	2.237
C2	GFA-spline	<4.19273- $^1\chi^v$ >, S_aasC, <1.83917-S_sCH ₃ >, RadOfGyration	–	25	0.304	0.925	0.910	61.31	2.756
C3	G/PLS-linear	$^0\chi^v$, S_aaCH, S_aasC, S_sOH, Jurs-TASA, HOMO	2	25	0.362	0.883	0.872	82.95	5.205
C4	G/PLS-spline	<1.78363-S_sCH ₃ >, <S_aasC-1.50199>, <5.22431-RadOfGyration>, <133.005-Jurs-WPSA-2>	1	25	0.294	0.919	0.915	260.7	3.064
Mod.	Stat.	Descriptors	Q^2	$r_m^2(\text{LOO})$	n_{test}	R_{pred}^2	$r_m^2(\text{test})$	$r_m^2(\text{overall})$	
C1	GFA-linear	$^3\chi_c^v$, Zagreb, S_aaCH, Jurs-RPSA	0.909	0.808	8	0.894	0.834	0.822	
C2	GFA-spline	<4.19273- $^1\chi^v$ >, S_aasC, <1.83917-S_sCH ₃ >, RadOfGyration	0.888	0.768	8	0.892	0.826	0.791	
C3	G/PLS-linear	$^0\chi^v$, S_aaCH, S_aasC, S_sOH, Jurs-TASA, HOMO	0.788	0.758	8	0.880	0.800	0.785	
C4	G/PLS-spline	<1.78363-S_sCH ₃ >, <S_aasC-1.50199>, <5.22431-RadOfGyration>, <133.005-Jurs-WPSA-2>	0.875	0.848	8	0.800	0.737	0.829	

* Critical values of F distribution (two-tailed) at 98% significance level: $F_{4,20} = 4.431$, $F_{2,22} = 5.719$, $F_{3,21} = 4.874$, $F_{1,23} = 7.881$

Tab. 4. Activity predicted for the newly designed compounds based on the 11 QSAR models developed in the present work.

Cpd. No.	Model A1	Model A2	Model A3	Model A4	Model B1	Model B2	Model B3	Model C1	Model C2	Model C3	Model C4	APA*
N1	4.085	9.362	4.281	3.915	3.213	2.697	3.282	4.900	7.824	3.851	8.831	5.113
N2	4.245	10.777	4.368	3.945	3.818	3.170	3.055	4.497	9.414	4.117	10.525	5.630
N3	3.414	7.474	3.644	3.493	3.533	3.254	3.652	3.971	6.578	3.493	7.281	4.526
N4	3.567	8.805	3.731	3.560	3.278	3.038	3.481	4.041	7.506	3.508	8.386	4.809
N5	3.517	7.759	3.696	3.558	3.594	3.278	3.911	3.982	6.929	3.557	7.759	4.685
N6	3.674	9.155	3.785	3.600	3.558	3.249	3.739	4.011	8.023	3.635	9.012	5.040
N7	3.678	9.276	3.783	3.633	3.592	3.261	3.736	4.110	8.044	3.646	9.053	5.074
N8	3.834	10.614	3.872	3.665	3.526	3.186	3.567	4.067	9.171	3.711	10.338	5.414
N9	3.772	10.125	4.738	4.387	4.727	4.039	2.878	4.405	9.143	4.500	9.643	5.669
N10	3.931	11.553	4.825	4.416	3.462	2.898	2.950	4.368	9.469	4.211	10.740	5.711
N11	3.787	8.678	4.513	4.362	4.300	3.708	3.063	3.858	7.928	4.153	8.685	5.185
N12	3.948	10.130	4.599	4.402	4.330	3.754	2.886	3.936	9.029	4.248	9.979	5.567
N13	3.844	8.990	3.749	3.566	3.494	3.154	3.676	3.816	7.809	3.591	8.747	4.949
N14	3.624	8.496	4.364	4.291	4.113	3.561	3.233	3.668	7.546	3.974	8.492	5.033
N15	3.784	9.786	4.449	4.323	3.239	2.816	3.107	4.069	8.181	3.798	9.347	5.173

*.average predicted activity.

Model developed with the whole dataset


$$pIC_{50} = -0.373 + 0.302(\pm 0.069) \times S_aasC + 0.009(\pm 0.001) \times JursTASA - 0.017(\pm 0.006) \times MR - 0.360(\pm 0.088) \times \chi_c^v - 2.11(\pm 0.608) \times$$

Eq. 7. $S_ddsN - 1.80(\pm 0.821) \times JursRPCG$


$$n = 33, s = 0.317, R^2 = 0.914, R_a^2 = 0.894, F = 46.10(df 6,26),$$

$$PRESS = 4.091, Q^2 = 0.866, true r_m^2(Loo) = 0.578$$

In the above equation, n is the number of compounds used for developing the QSAR model. Standard errors of the regression coefficients are shown in parentheses. Eq. 7 was developed with the entire dataset of molecules using stepwise regression method [23]. A value of internal predictive variance ($Q^2 = 0.866$) above the stipulated value of 0.5 for the developed model signifies its predictive ability. The positive coefficients for S_aasC and $Jurs TASA$ descriptors indicate that the antioxidant activities of these molecules increase with an increase in the values of these descriptors. The S_aasC descriptor refers to the

summation of E-state values for the  (aromatic carbon) fragments. $Jurs TASA$ (total hydrophobic surface area) is calculated as the sum of solvent-accessible surface areas of atoms with absolute value of partial charges less than 0.2. An increase in the value of S_aasC descriptor indicates increase in substitution on the aromatic nucleus while an increase in the value of $Jurs TASA$ indicates an increase in surface area with reduced partial charge. Again, negative coefficients for MR , χ_c^v , S_ddsN and $Jurs RPCG$ descriptors signify that the antioxidant activity of these molecules is inversely proportional

to the values of these descriptors. *MR* refers to the molar refractivity of the molecule and gives a measure of the size and volume of the molecule. The parameter, ${}^3\chi_c^v$ is a topological descriptor [20] belonging to the category of molecular connectivity indices and defined as the third order cluster index based on valance count. It encodes the number of branch points in a molecule indicating a decrease in branching of the molecule for a decrease in its value. The parameter, *Jurs RPCG* (relative positive charge) is a spatial descriptor obtained by dividing the partial charge of the most positive atom with the total positive charge. *S_ddSN* is a topological descriptor and refers to the summation of E-state

value for the nitrogen atom of type . In this data set, it has referred to the nitrogen of the N-oxide fragment of the oxadiazole ring. It has been observed that the descriptor, *S_ddSN*, assumes a negative value when the 5 position of the oxadiazole ring is substituted with a group containing more electronegative atoms. In such cases, the impact of *S_ddSN* becomes positive on the activity (the coefficient of *S-ddSN* in Eq. (7) is actually negative).



The opposite signs for the coefficients of *Jurs TASA* and *MR* descriptors infer that although an increase in hydrophobic surface area of the molecules favours their activity, the total volume of the molecule should small enough so that the *MR* descriptor attains a lower value. Thus, the increase in hydrophobic surface area should be up to a specific limit. Interestingly, *MR* assumes a positive coefficient in the absence of the *Jurs TASA* descriptor. Again, according to the order of significance, the descriptors occurring in Eq. (7) can be ranked as: (i) *Jurs TASA*, (ii) *S_aasC*, (iii) ${}^3\chi_c^v$, (iv) *MR*, (v) *S_ddSN* and (vi) *Jurs RPCG*. The weightage of these descriptors once again signifies that the *Jurs TASA* descriptor has a greater impact on the activity profile these molecules compared to the *MR* descriptor. Compound numbers (nos.) **4** and **9** bearing conducive values for all these descriptors exhibit maximum activity. Again compound nos. **16**, **17** and **18** despite having large values for the *MR* and ${}^3\chi_c^v$ descriptors exert high range of activity, since these descriptors rank lower in terms of the weightage of the descriptors. Although compound nos. **1**, **10** and **28** satisfy the requirements for most of the descriptors, these compounds exhibit the lowest activity range due to unsatisfactory values of the two most significant descriptors, *Jurs TASA* and *S_aasC*.

It has been argued that in case of a small dataset, considerable amount of information is lost in division of dataset into training and test sets. Alternatively, 'true $r_m^2_{(LOO)}$ ' calculated based on the whole dataset may efficiently reflect the predictive potential of a model. Thus, the value of 'true $r_m^2_{(LOO)}$ ' [21] (threshold value = 0.5) was also calculated for this dataset. For the calculation of this parameter, each molecule in the dataset was deleted once and the variable selection strategy was applied and a new model was built with the remaining molecules. The activity of the deleted molecule was thereafter predicted using the developed model. The cycle was continued till all the molecules in the dataset were deleted at least once. The activity predicted thus for all the molecules was used for the calculation of this 'true $r_m^2_{(LOO)}$ ' parameter. Thus, in the calculation of 'true $r_m^2_{(LOO)}$ ' metric, new variables are selected in each cycle based on the leave-one-out technique. Consequently, this parameter reflects the external predictive ability of the model especially in case of such a small dataset where splitting may result in loss of an appreciable amount of chemical information. In this case, statistically significant result was obtained for the

'true $r_m^2_{(LOO)}$ ' (0.578) parameter indicating ability of the model to predict the activity of new series of molecules of this class.

Models developed from training set data

To indicate the external predictivity of developed models, the dataset was further divided into training and test sets. Subsequent models were built based on the training set and were externally validated based on the test set. Three different types of models were built using the training set compounds such as: (i) models developed with topological, structural and thermodynamic descriptors, (ii) models developed with spatial, electronic and thermodynamic descriptors and (iii) models developed with combined set of descriptors. The predictive ability of the models was judged based on the internal and external validation parameters which are summarized in Tab. 3. All the models bear acceptable values of Q^2 and R^2_{pred} which are considerably greater than the stipulated value of 0.5. In terms of internal predictivity ($Q^2 = 0.909$), model C1 developed with the combined set of descriptors was the better compared to the other ones. But since internal validation alone fails to judge the ability of a model to predict the activity of new series of molecules, values of R^2_{pred} were also taken into consideration. Thus, in terms of both internal ($Q^2 = 0.814$) and external ($R^2_{pred} = 0.917$) predictive parameters, model A2a developed with the topological, structural and thermodynamic descriptors shows maximum statistical significance. However, the value of R^2_{pred} alone fails to judge whether the range of predicted activity data lies within the desired observed activity range. Hence, values of r_m^2 metrics were calculated. The $r_m^2_{(overall)}$ value determines the degree of proximity between the observed and corresponding predicted activity data for both the training and test set molecules. Thus, in terms of all the three parameters ($Q^2 = 0.877$, $R^2_{pred} = 0.884$, $r_m^2_{(overall)} = 0.872$), model A4 developed with topological, structural and thermodynamic descriptors exhibits maximum statistical significance. It can be inferred from Tab. 3 that the models developed with spatial, electronic and thermodynamic descriptors were inferior in terms of their predictive ability compared to the remaining ones and hence are not described below.

The repeated occurrence of the E-state descriptors in the developed QSAR models signifies the importance of the various structural fragments for optimal antioxidant activity of these molecules. The S_{aaCH} and S_{aasC} descriptors refer to the summation of E-state values for unsubstituted () and substituted () aromatic carbon fragments respectively while S_{sCH_3} descriptor refer to the summation of E-state values for the methyl groups (-CH₃) present within the molecular structures. Thus, presence of these descriptors signifies the influence of these structural fragments for the activity profile of these molecules. The parameter S_{aasC} also appeared in Eq. 7 obtained for the whole dataset. Again repeated occurrence of the $Jurs$ descriptors and the various types of connectivity (χ) descriptors indicate that the charged surface area of the molecules as well as their extent of branching influence the antioxidant activity profile of these molecules.

Due to space limitation, only the GFA and G/PLS models (models A2, A2a and A4) obtained using spline option from topological, structural and thermodynamic descriptors are described here.

GFA model

Eq. 8.

$$pIC_{50} = -1.68 - 0.113 (\pm 0.029) \times \langle SC-3_P-20 \rangle + 0.786 (\pm 0.164) \times {}^3\chi_p - 0.637 (\pm 0.099) \times \langle 1.79401 - S_sCH_3 \rangle + 0.329 (\pm 0.049) \times S_aasC$$

$$n_{training} = 25, s = 0.315, R^2 = 0.919, R_a^2 = 0.903, F = 56.95 (df 4, 20), PRESS = 3.022,$$

$$Q^2 = 0.877, r_m^2 (LOO) = 0.757, n_{test} = 8, R^2_{pred} = 0.924, r_m^2 (test) = 0.899, r_m^2 (overall) = 0.777$$

The acceptable values of the internal ($Q^2 = 0.877$) and external ($R^2_{pred} = 0.924$) predictive parameters reflect the predictability of the developed model. Moreover, statistically significant results for all the r_m^2 metrics indicate that the predicted activity values of all the compounds are close to the corresponding observed activity data. Although the model exhibits high predictive ability, there exists significant intercorrelation between two descriptors namely, SC-3_P (number of third-order subgraphs in the molecular graph) and ${}^3\chi_p$ (molecular connectivity index). Intercorrelation matrix (Tab. 5) for all the descriptors appearing in Eq. 8 signifies that there may exist a parabolic relationship between the activity values and these descriptors. Thus, to better express the parabolic behaviour of the developed model, the SC-3_P descriptor was replaced with a second order function of the ${}^3\chi_p$ descriptor.

Tab. 5. Intercorrelation matrix for Eq. 8 (model A2)

Descriptor	SC-3_P	${}^3\chi_p$	S_sCH ₃	S_aasC
SC-3_P	1.000	0.989	-0.133	0.020
${}^3\chi_p$		1.000	-0.062	-0.027
S_sCH ₃			1.000	-0.566
S_aasC				1.000

Eq. 9.


$$pIC_{50} = -0.448 + 0.497 (\pm 0.147) \times {}^3\chi_p - 0.0225 (\pm 0.009) \times ({}^3\chi_p)^2 + 0.300 (\pm 0.110) \times S_aasC - 0.539 (\pm 0.057) \times \langle 1.79401 - S_sCH_3 \rangle$$

$$n_{training} = 25, s = 0.369, R^2 = 0.889, R_a^2 = 0.867, F = 40.17 (df 4, 20),$$

$$PRESS = 4.557, Q^2 = 0.814, r_m^2 (LOO) = 0.677, n_{test} = 8,$$

$$R^2_{pred} = 0.917, r_m^2 (test) = 0.818, r_m^2 (overall) = 0.711$$

Eq. 8 is thus modified to Eq. 9 in order to express the parabolic relationship of the developed QSAR model with respect to ${}^3\chi_p$ descriptor. ${}^3\chi_p$ [20] is the weighted count of four atom (three-bond) fragments and it reflects the degree of branching at each of the four atoms in the fragment. In the above equation, a positive coefficient of the ${}^3\chi_p$ descriptor signifies that the antioxidant activity of these molecules increases with an increase in the value of this descriptor. Thus, it can be inferred that an increase in the degree of branching in these molecules favours their antioxidant activity profile. Maximum antioxidant activity profile of compound nos. **9**, **16** and **18** can be explained by their large ${}^3\chi_p$ descriptor values while the reduced activity of compound nos. **1**, **6**, **10** and **28**, may be attributed to the small values of ${}^3\chi_p$ descriptor. The optimum value of ${}^3\chi_p$ for this series of compounds is 11.044


which approximately matches with the value of this descriptor for the most active compound (compound no. **18**). This in turn explains the parabolic relationship between activity and ${}^3\chi_p$ descriptor. The S_{aasC} descriptor refers to the summation of E-state values for the  (aromatic carbon) fragments. Since the S_{aasC} descriptor bears a positive coefficient, an increase in its value leads to an improvement in the antioxidant activity profile of these compounds as observed in case of compound nos. **9**, **16**, **17** and **18**. Thus, as the number of such fragments increase, an increase in the activity data of these molecules is observed. The S_{sCH_3} descriptor refers to the summation of E-state values for the $-\text{CH}_3$ (methyl group) fragments. In the above equation, a negative coefficient of the spline term with this descriptor indicates that for any value of this descriptor greater than 1.79401, the spline term $\langle 1.79401 - S_{\text{sCH}_3} \rangle$, exerts zero contribution and the compounds show an improvement in activity data as exemplified in compound nos. **9**, **16**, **17** and **18**. On the contrary, compound nos. **6**, **10**, **28** and **33** with zero values of the S_{sCH_3} descriptor exhibit lowest activity range. Thus it can be suggested that presence of methyl substituents favours the antioxidant activity profile of these molecules.

G/PLS model

A statistically significant QSAR model was also obtained using the G/PLS technique together with the spline option.

$$pC = 1.158 - 0.429 \times \langle 6.68154 - {}^1\chi \rangle + 0.108 \times {}^3\chi_p - 0.525 \times \langle 1.98556 - S_{\text{sCH}_3} \rangle + 0.288 \times S_{\text{aasC}}$$

Eq. 10. $n_{\text{training}} = 25, LVs = 2, s = 0.323, R^2 = 0.905, R_a^2 = 0.897, F = 106.44(df 2, 22),$
 $PRESS = 3.023, Q^2 = 0.877, r_m^2(\text{LOO}) = 0.870, n_{\text{test}} = 8, R_{\text{pred}}^2 = 0.884,$
 $r_m^2(\text{test}) = 0.812, r_m^2(\text{overall}) = 0.872$

The predictive power of the developed model is reflected through the statistically significant values of the internal ($Q^2 = 0.877$) and external ($R_{\text{pred}}^2 = 0.884$) validation parameters as well as the r_m^2 metrics. Among all the developed equations, this equation gives a maximum value of the $r_m^2(\text{overall})$ (0.872) parameter indicating that the predicted activity values of all the dataset compounds are in very close proximity to the corresponding observed data. These results signify that the model can be efficiently used for activity prediction of new compounds of this class. The predicted activity values (LOO predicted values for the training set) according to Eq. (10) are given in Tab. 1. According to the order of significance, the descriptors occurring in Eq. (10) can be arranged as: (i) $\langle 1.98556 - S_{\text{sCH}_3} \rangle$, (ii) S_{aasC} , (iii) ${}^3\chi_p$ and (iv) $\langle 6.68154 - {}^1\chi \rangle$. Similar to the previous equation (Eq. 9), ${}^3\chi_p$ and S_{aasC} descriptors bear positive coefficients indicating that the antioxidant activity increases with an increase in the values of these descriptors. Consequently, this observation infers that a high degree of branching (as indicated by high values of ${}^3\chi_p$) and large number of  fragments present within the molecular structure of these NO donor phenolic compounds favour their antioxidant activity profile.

Again, negative coefficients of the spline terms with the S_sCH_3 and ${}^1\chi$ imply that for zero values of these spline terms, the compounds show high activity range. This, in turn, suggests that values of S_sCH_3 descriptor greater than 1.98556 and that of ${}^1\chi$ descriptor greater than 6.68154 account for zero contribution of the spline function and hence explain the high activity range for compounds with such values. This is because a negative value of a spline term indicates zero contribution of the corresponding spline term. ${}^1\chi$ [20] is a topological descriptor referring to simple connectivity index obtained by one bond dissection of the molecule. Since the descriptor encodes the number of non-hydrogen atoms in a molecule, values of this descriptor reflect the size and volume of the molecule along with the degree of branching. Thus the results suggest that an increase in the number of methyl substitution ($-CH_3$ fragment) as well as an increase in the volume and/or branching of the molecule favours their antioxidant activity. This increase in size can be well correlated with the observation of Eq. 7, which infers that increase in total hydrophobic surface area of these molecules up to a definite limit favors their antioxidant activity profile. Compound nos. **9**, **16**, **17** and **18** fulfilling all these necessary structural requirements exert maximum activity range. In case of compound nos. **6**, **10** and **28**, although the $\langle 6.68154-{}^1\chi \rangle$ descriptor attains zero value, the other highly significant descriptors do not meet the necessary criteria and subsequently, results in a lowering of their activity profile. Similarly, compound no. **33** shows reduced activity despite having zero value for the $\langle 6.68154-{}^1\chi \rangle$ descriptor and a large value of the ${}^3\chi_p$ descriptor due to a lower weightage of these descriptors compared to the others.

Major observations from other models

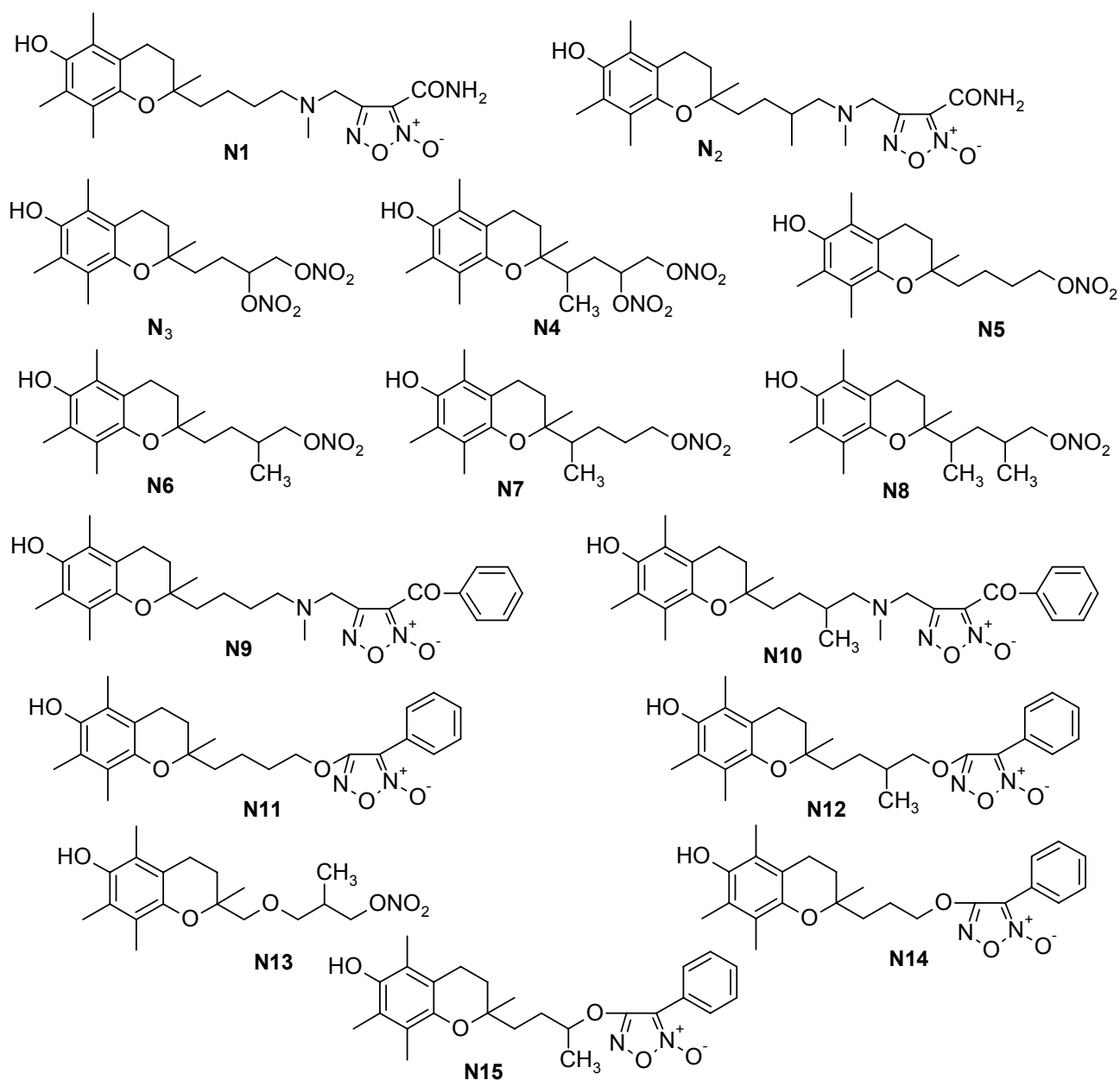
The remaining equations which are not described in detail here also reveal some interesting structure-activity relationships for optimum antioxidant activity of these molecules. The S_dsN descriptor appearing in model A3 refers to the summation of E-state values for the $-N=$ fragment of the pyrazolone ring. The positive coefficient of this term indicates that the presence of this fragment in the molecular structure favours the activity range of these compounds. However, being a descriptor of less relative importance, it does not significantly contribute to the activity profile of these compounds. Again, both models C2 and C4 bear the *Rad of Gyra* descriptor (abbreviation for radius of gyration, which is a size descriptor denoting the distribution of atomic masses in a molecule and measures of molecular compactness and symmetry for long-chain molecules) inferring that long chain unsymmetrical substitution of the parent molecule leads to an increase in their activity profile. Besides these, most of the models signify that an increase in the activity profile of these molecules is achieved with an increase in the degree of methylation and substituted aromatic carbon fragments in their molecular structure.

Tab. 6. Results of validation based on randomization.

Process randomization						
Models with topological, structural and thermodynamic descriptors						
Model No.	Statistical tool	R²	R	R_r	R_r²	^cR_p²
A1	GFA-linear	0.889	0.943	0.531	0.282	0.735
A2	GFA-spline	0.919	0.959	0.533	0.284	0.764
A3	G/PLS-linear	0.856	0.925	0.669	0.448	0.591
A4	G/PLS-spline	0.906	0.952	0.808	0.653	0.479
Models with spatial, electronic and thermodynamic descriptors						
Model No.	Statistical tool	R²	R	R_r	R_r²	^cR_p²
B1	GFA-linear	0.824	0.908	0.467	0.218	0.707
B2	G/PLS-linear	0.839	0.916	0.671	0.450	0.571
B3	G/PLS-spline	0.865	0.930	0.778	0.605	0.474
Models with combined set of descriptors						
Model No.	Statistical tool	R²	R	R_r	R_r²	^cR_p²
C1	GFA-linear	0.935	0.967	0.646	0.417	0.696
C2	GFA-spline	0.925	0.962	0.715	0.511	0.619
C3	G/PLS-linear	0.883	0.940	0.686	0.471	0.603
C4	G/PLS-spline	0.919	0.959	0.795	0.632	0.514
Model randomization						
Model No.		R²	R	R_r	R_r²	^cR_p²
	A2	0.919	0.959	0.379	0.144	0.844
	A3	0.856	0.925	0.058	0.003	0.854
	A4	0.906	0.952	0.12	0.014	0.899
	C1	0.935	0.967	0.368	0.135	0.865
	C2	0.925	0.962	0.377	0.142	0.851
	C4	0.919	0.959	0.113	0.013	0.913

Results of validation based on randomization tests

Further validation of the developed models was done using the randomization technique in order to check the robustness of the genetic QSAR models. Process randomization was performed using the whole descriptor matrix to assess the fitness of the process employed for the development of the QSAR models. Besides this, model randomization was also performed in order to determine whether the model was developed by chance or not. Based on the randomized data, values of R_r^2 were calculated. For all the developed models, the values of R_r^2 were much lower than that of the model R^2 implying the robustness of the developed model. However, since no guideline is given as to how much this difference should be in order to obtain a robust QSAR model, values of ${}^cR_p^2$ were computed [29] (Tab. 6). Models having ${}^cR_p^2$ values greater than 0.5 are considered to be statistically robust. The ${}^cR_p^2$ values of most of the models obtained for the process randomization technique exceed the threshold value of 0.5. The ${}^cR_p^2$ values for all the models described above are well above the stipulated value with model A4 (0.899) and model C4 (0.913) exhibiting maximum values. This indicates that the models developed are not merely the outcome of chance.

Design of new compounds**Fig. 1.** Structures of the designed compounds

The statistically significant QSAR models developed above thus determine the required structural attributes for maximum antioxidant activity. The equations primarily suggest that the presence of substituted aromatic carbon within the molecular structure together with extensive methyl substituent favours the antioxidant activity profile of this series of molecules. Additionally charged surface area and polar surface are of the molecules also play significant role in determining the potency of these molecules indicating that an increase in surface area and volume of the molecules up to a specific limit favours their activity profile. Besides these, charged positive and charged polar surface areas play an important role for activity prediction of these molecules. Moreover, long chain compounds with reduced symmetry and an optimum volume may favour the activity profile of these

compounds. An extensive occurrence of the connectivity descriptors in the above equations signify that proper branching of the molecules essential for potent antioxidant activity of these molecules. Again, the developed QSAR models also suggest that an oxadiazole-N-oxide ring substituted with an electronegative atom containing group at the 5 position may exert a positive impact on the activity of the molecules. Based on these structural attributes, 15 new molecules were designed and their activity was predicted using all the developed models. With the predicted activity obtained for all the developed models, a consensus for predicted activity was computed and it was observed that all the compounds exhibited potent activity profile which was in close proximity to that of the highly active compounds of the present dataset. All the 15 designed compounds and their *in silico* predicted activity data are listed in Fig. 1 and Tab. 4.

Test for applicability domain

According to the r_m^2 (overall) criterion, Eq. (10) is the best model among all the developed models. The applicability domain Eq. 10 (G/PLS model) was checked based on the DModX [25] approach. A bar diagram for the DModX values of the 8 test set compounds as well as the designed compounds for Eq. (10) is shown in Fig. 2. The DModX values thus obtained for all the test compounds as well as the 15 newly designed compounds are below the critical value of 3.225 calculated at the 99% significance level. So, none of the compounds are outside the applicability domain of Eq. (10) and predictions for all the compounds are acceptable with confidence. Moreover, acceptable DModX values for the designed compounds indicate that predictions for antioxidant activity for these compounds according to Eq. (10) are reliable.

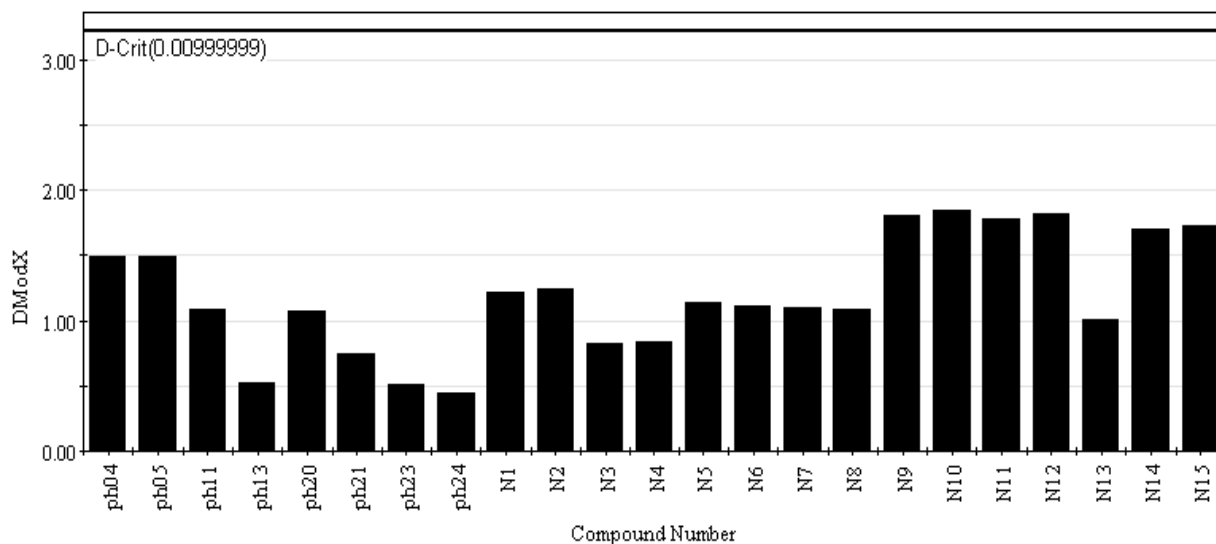


Fig 2. Bar diagram showing the DModX values of the 8 test set compounds and the 15 designed compounds calculated at 99% significance level with the thick horizontal line signifying the critical DModX value (3.225) for Eq. (10).

The best model [Eq. (10)] thus built obeys the 5 guidelines for acceptability of QSAR models laid by the Organisation for Economic Co-operation and Development (OECD) [35]: (i) the model has been built based on an unambiguous algorithm; (ii) a definite

response, viz., antioxidant activity using the TBARS (Thiobarbituric acid reactive substance) assay method has been modeled in the present work; (iii) the molecules predicted using the developed model are rightly located within the model applicability domain; (iv) goodness of fit, robustness and predictivity of the developed models have been appropriately checked using different validation measures and (v) the model provides a suitable mechanistic interpretation for assessing the necessary structural attributes of the molecules for exhibiting optimum response. Hence, the developed model can be satisfactorily used from the regulatory point of view.



Comparison with previously reported work on NO donor phenols

Previously, structure–antioxidant activity relationships for a series of NO-donor phenols have been reported by Tosco et al. [36]. 17 out of the 33 NO donor phenols used in the present work were modeled by Tosco et al. based on their partition coefficient and bond dissociation enthalpy values. Tosco et al. reported structure-activity relationships of these molecules purely based on their internal validation parameters. They developed a bilinear model with 17 compounds and obtained a Q^2 of 0.94. However, as Tosco et al. used limited number of compounds for QSAR model development and did not perform external validation, a direct comparison of these models with those developed by us is not possible. Unlike the present work, they did not report any data regarding external validation and randomization of the dataset.

Overview and conclusion

In the present work, QSAR models were built using a dataset ($n=33$) comprising phenolic derivatives but chiefly constituting compounds with the NO donor functions. For the development of the QSAR models different statistical tools and software were employed. The major chemometric tools used for the present work include the GFA and G/PLS techniques. Initially QSAR model was developed using the entire dataset using stepwise regression and value of “true r_m^2 (LOO)” was calculated in order to determine the predictive ability of the dataset. In order to determine the external predictive ability of models, the dataset was divided into training and test sets using the k -means clustering technique and external validation was done based on the activity prediction of the test set compounds. Eq. 9 (0.917) with maximum value of the R_{pred}^2 parameter indicates significant ability of the developed model to predict the activity of new compounds belonging to this series of phenolic derivatives. Besides these, the r_m^2 metrics were also calculated to determine the distance of the predicted activity data from the corresponding observed ones. A high value of r_m^2 (overall) for Eq. 10 (0.872) implies that the activity data predicted for the test set compounds using the model satisfies the desired range of observed activity data. To check the reproducibility of the developed models, validation was done using both process and model randomization techniques. The results of model randomization test reveal that the $^{\circ}R_p^2$ values [37] for all the models exceed the stipulated value of 0.5. Maximum $^{\circ}R_p^2$ values for model A4 and model C4 infer that the developed models are sufficiently robust and not the outcome of mere chance.

Analysis of the QSAR models developed in the present work reveal the structural requirements of these molecules for exhibiting maximum antioxidant activity. The repeated occurrence of the S_{aasC} and S_{CH_3} descriptors in different models signifies that the

presence of the  fragment and methyl substituents within the molecular structure of these phenolic derivatives is conducive to the antioxidant activity profile of these compounds. The presence of an aromatic carbon without a substitution () hinders the activity profile of these compounds. The presence of an oxadiazole-N-oxide ring with an electronegative atom containing group substituted at the 5 position are conducive for the antioxidant activity of these compounds. Besides these, increase in the positively charged surface area and the volume of the molecules favours the antioxidant activity profile of these compounds. Long chain branched substituents lacking symmetry about the centre of mass of the molecule exhibit improved antioxidant activity. Based on this structural information, 15 new compounds were designed and their activity was predicted using the QSAR models developed in the present work. Since the qualities of the models are good and the observed and predicted activity values of the test set compounds are in good agreement, we can presume that the designed compounds may show potent experimental antioxidant activity as also predicted by the developed models. Thus, the statistically significant QSAR models developed in the present work can be satisfactorily used for activity prediction of new series of molecules of this class. Moreover, the compounds designed in the present work can be utilized further for experimental work.

Acknowledgement

This research work is supported in the form of a major research project to KR and a senior research fellowship to IM by Indian Council of Medical Research (ICMR), New Delhi.

Supporting Information

A table with the values of the important descriptors appearing in the described QSAR models is available in the online version (Format: PDF, Size: < 0.2 MB): <http://dx.doi.org/10.3797/scipharm.1011-02>

Authors' Statement

Competing Interests

The authors declare no conflict of interest.

References

- [1] Genestra M.
Oxyl radicals, redox-sensitive signalling cascades and antioxidants.
Cell Signal. 2007; 19: 1807–1819.
doi:10.1016/j.cellsig.2007.04.009
- [2] Halliwell B, Gutteridge JMC.
Free radicals in biology and medicine.
Oxford: Oxford University Press, 2007.
- [3] Harman D.
Free radical theory of aging: Alzheimer's disease pathogenesis.
Age. 1995; 18: 97–119.
doi:10.1007/BF02436085

- [4] Koutsilieris E, Scheller C, Grünblatt E, Nara K, Li J, Riederer P. Free radicals in Parkinson's disease. *J Neurol.* 2002; 249 (Suppl 2): II/1–II/5. doi:10.1007/s00415-002-1201-7
- [5] Nuttall SL, Kendall MJ, Martin U. Antioxidants therapy for prevention of cardiovascular disease. *Q J Med.* 1999; 92: 239–244. doi:10.1093/qjmed/92.5.239
- [6] Dizdaroglu M, Jaruga P, Birincioglu M, Rodriguez H. Free radical-induced damage to DNA: mechanisms and measurement. *Free Radic Biol Med.* 2002; 32: 1102–1115. doi:10.1016/S0891-5849(02)00826-2
- [7] Gutteridge JMC, Halliwell B. Antioxidants in nutrition, health and disease. Oxford: Oxford University Press, 1994.
- [8] Pokorny J. Natural antioxidants for food use. *Trends Food Sci Technol.* 1991; 2: 223–227. doi:10.1016/0924-2244(91)90695-F
- [9] Wright JS, Johnson ER, DiLabio GA. Predicting the activity of phenolic antioxidants: theoretical method, analysis of substituent effects, and application to major families of antioxidants. *J Am Chem Soc.* 2001; 123: 1173–1183. doi:10.1021/ja002455u
- [10] Helguera AM, Combes RD, Gonzalez MP, Cordeiro MN. Applications of 2D descriptors in drug design: a DRAGON tale. *Curr Top Med Chem.* 2008; 8: 1628–1655. doi:10.2174/156802608786786598
- [11] Hansch C, Maloney PP, Fujita T, Muir RM. The correlation of the biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature.* 1962; 194: 178–180. doi:10.1038/194178b0
- [12] Rastija V, Medic-Saric M. QSAR study of antioxidant activity of wine polyphenols. *Eur J Med Chem.* 2009; 44: 400–408. doi:10.1016/j.ejmech.2008.03.001
- [13] Ray S, Sengupta C, Roy K. QSAR modeling of antiradical and antioxidant activities of flavonoids using electrotopological state (E-State) atom parameters. *Cent Eur J Chem.* 2007; 5: 1094–1113. doi:10.2478/s11532-007-0047-3
- [14] Mitra I, Saha A, Roy K. Pharmacophore mapping of arylaminosubstituted benzo[b]thiophenes as free radical scavengers. *J Mol Model.* 2010; 16: 1585–1596. doi:10.1007/s00894-010-0661-4
- [15] Roy K, Mitra I. Advances in quantitative structure–activity relationship models of antioxidants. *Expert Opin Drug Discov.* 2009; 4: 1157–1175. doi:10.1517/17460440903307409

- [16] Boschi D, Tron GC, Lazzarato L, Chegaev K, Cena C, Stilo AD, Giorgis M, Bertinaria M, Fruttero R, Gasco A.
NO-Donor Phenols: A New Class of Products Endowed with Antioxidant and Vasodilator Properties.
J Med Chem. 2006; 49: 2886–2897.
doi:10.1021/jm0510530
- [17] Chegaev K, Cena C, Giorgis M, Rolando B, Tosco P, Bertinaria M, Fruttero R, Carrupt PA, Gasco A.
Edaravone derivatives containing NO-donor functions.
J Med Chem. 2009; 52: 574–578.
doi:10.1021/jm8007008
- [18] Cena C, Chegaev K, Balbo S, Lazzarato L, Rolando B, Giorgis M, Marini E, Fruttero R, Gasco A.
Novel antioxidant agents deriving from molecular combination of Vitamin C and NO-donor moieties.
Bioorg Med Chem. 2008; 16: 5199–5206.
doi:10.1016/j.bmc.2008.03.014
- [19] ACD/3D Viewer, version 12.00, Advanced Chemistry Development Inc. www.acdlabs.com
- [20] CERIUS2 version 4.1, San Diego, CA, USA: Accelrys Inc.
- [21] Mitra I, Roy PP, Kar S, Ojha PK, Roy K.
On further application of r_m^2 as a metric for validation of QSAR.
J Chemometrics. 2010; 24: 22–33.
doi:10.1002/cem.1268
- [22] Everitt B, Landau S, Leese M.
Cluster Analysis.
London: Arnold Press, 2001.
- [23] Snedecor GW, Cochran WG.
Statistical Methods.
New Delhi: Oxford & IBH, 1967.
- [24] Rogers D, Hopfinger AJ.
Application of genetic function approximation to quantitative structure–activity relationship and
quantitative structure–property relationship.
J Chem Inf Comput Sci. 1994; 34: 854–866.
doi:10.1021/ci00020a020
- [25] Wold S, Sjostrom M, Eriksson L.
PLS-regression: A basic tool of chemometrics.
Chemometrics Intell Lab Syst. 2001; 58: 109–130.
doi:10.1016/S0169-7439(01)00155-1
- [26] Golbraikh A, Tropsha A.
Beware of q^2 !
J Mol Graph Model. 2002; 20: 269–276.
doi:10.1016/S1093-3263(01)00123-1
- [27] Roy PP, Paul S, Mitra I, Roy K.
On two novel parameters for validation of predictive QSAR models.
Molecules. 2009; 14: 1660–1701.
doi:10.3390/molecules14051660
- [28] Toropov AA, Toropova AP, Benfenati E.
QSPR modeling bioconcentration factor (BCF) by balance of correlations.
Eur J Med Chem. 2009; 44: 2544–2551.
doi:10.1016/j.ejmech.2009.01.023
- [29] Todeschini R.
Milano Chemometrics.
Italy (personal communication), 2010.

- [30] Weaver S, Gleeson MP.
The importance of the domain of applicability in QSAR modeling.
J Mol Graph Model. 2008; 26: 1315–1326
doi:10.1016/j.jmgm.2008.01.002
- [31] Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P.
Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs.
Environ Health Perspect. 2003; 111: 1361–1375.
PMid:12896860
- [32] Organisation for Economic Co-operation and Development Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. 2007, OECD Document ENV/JM/MONO(2007)2
- [33] Netzeva TI, Worth AP, Aldenberg T, Benjamin I, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, van de Sandt JJM, Tong W, Veith G, Yang C.
Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships-The Report and Recommendations of ECVAM Workshop 52. ATLA 2005; 33: 155–173.
- [34] UMETRICS SIMCA-P 10.0, info@umetrics.com: www.umetrics.com, Umea, Sweden, 2002
- [35] OECD Principles for the Validation of (Q)SARs, <http://www.oecd.org/dataoecd/33/37/37849783.pdf>
- [36] Tosco P, Marini E, Rolando B, Lazzarato L, Cena C, Bertinaria M, Fruttero R, Reist M, Carrupt PA, Gasco A.
Structure-antioxidant activity relationships in a series of NO-donor phenols.
Chem Med Chem. 2008; 3: 1443–1448.
doi:10.1002/cmdc.200800101
- [37] Mitra I, Saha A, Roy K.
Exploring quantitative structure-activity relationship (QSAR) studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants.
Mol Simul. 2010; 36: 1067–1079.
doi:10.1080/08927022.2010.503326