Discover Chemistry

METHODOLOGY Open Access

Python-based QSAR modeling protocol for antioxidant activity: a case-study using a library of di(hetero)cyclic amines or amides

Cristiano Mateus[1] and Rui M. V. Abreu[1*]

*Correspondence:
Rui M. V. Abreu
ruiabreu@ipb.pt
[1]CIMO, LA SusTEC, Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300- 253 Bragança, Portugal

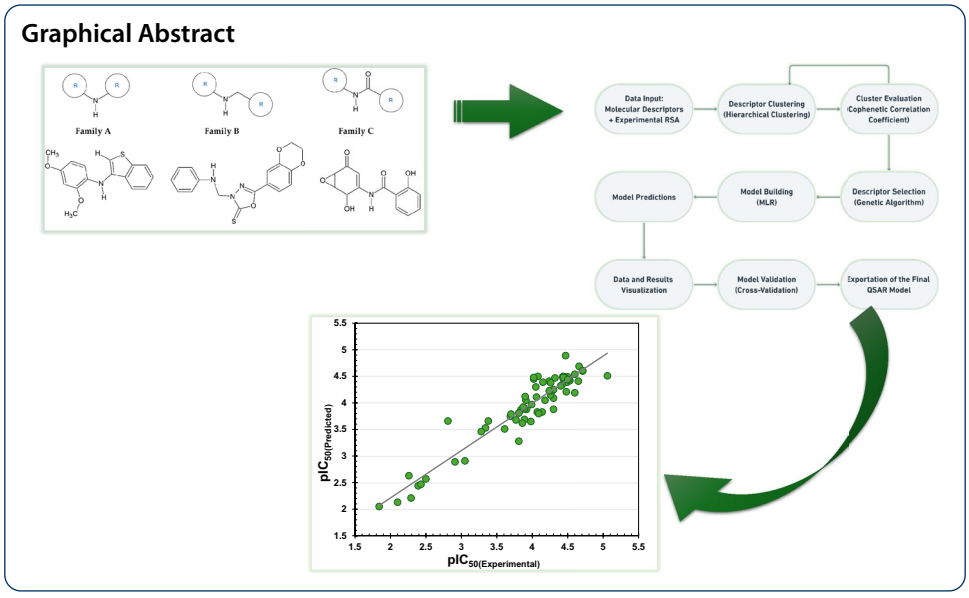
Abstract

Constructing a QSAR model involves several critical steps, including chemical structure preparation, molecular descriptor calculation and selection, and model development and validation. This study presents a comprehensive methodology for preparing QSAR models using freely open-source software tools. A detailed, step-by-step protocol outlines the entire process, from compound library preparation to statistical validation. As a case study, we developed a QSAR model to predict the antioxidant activity, specifically radical scavenging activity, of 70 di(hetero)aryl amine and amide compounds. Molecular descriptors (12,072 total) were calculated using the OCHEM platform, and PyQSAR built-in tools were used for descriptor selection and model construction. Four key descriptors (B06[C-O], Eig04_AEA(dm), JGI2, and J_Dz(p)) were selected to develop a MLR model with strong statistical performance (Q^2_{CV} = 0.8676, RSR_{CV} = 0.3518). Internal validation showed strong predictive stability, while external validation demonstrated the model's generalizability with a $Q^2_{EXT} > 0.5$. This study not only demonstrates the application of a freely open-source QSAR approach but also contributes to ongoing efforts in identifying and designing potent antioxidant agents with potential therapeutic applications. All relevant files and the detailed protocol are provided, allowing other researchers to replicate the antioxidant QSAR model and apply the methodology to develop QSAR models for other compound libraries and biological activities.

Scientific contribution

- New methodology to prepare QSAR models using free and open-source tools.
- Antioxidant QSAR model of di(hetero)aryl amines and amides with robust statistical parameters.
- Complete step-by-step protocol to replicate the antioxidant QSAR model or prepare new models.

Keywords QSAR modeling, Molecular descriptors, Methodology, Antioxidant activity, Di(hetero)aryl amines and amides



1 Introduction

Quantitative structure–activity relationships (QSAR) modeling is an in silico methodology to predict small molecules' physical or biological properties. A QSAR analysis generally attempts to find correlations between an experimental biological activity and molecular descriptors, either calculated from the compound's chemical structure or experimentally obtained [1, 2]. QSAR modeling has diversified and evolved from its application to small series of similar compounds, using relatively simple regression methods, to the analysis of much larger datasets spanning thousands of molecules, using various statistical techniques [3]. Continuous improvements have allowed QSAR modeling to be widely employed in the chemical, medical, and pharmaceutical industries and by government institutions to predict properties such as bioactivity, toxicity, solubility, and permeability [4].

A series of steps are typically needed to perform a QSAR analysis: chemical data preparation (literature search for compounds with an IC_{50} value defined experimentally by comparable biological assays), dataset splitting, molecular descriptor calculation, descriptor selection, mathematical model building and model validation (Fig. 1).

Currently, many software tools are available, and typically, each software performs one or a small number of steps needed to build a QSAR model [3, 5] or is commercial software that is not readily available for many researchers. PyQSAR is a free, open-source software that performs several steps of QSAR model development [6]. At the same time, OCHEM is another free-for-academic-use web-based platform that calculates a large number of molecular descriptors [7]. The calculated descriptors range from 1D descriptors, based on the compounds' physicochemical properties or molecular formula, to 2D descriptors, calculated from a 2D representation, and 3D descriptors obtained from a 3D representation of the compounds [8]. Recently, a workflow was presented by Mansouri et al. (2024) based on KNIME software that automated the initial steps of QSAR modeling, specifically chemical library structure preparation [9]. Additionally, a framework for QSAR model building was published by Kausar et al. (2018); however, the number of

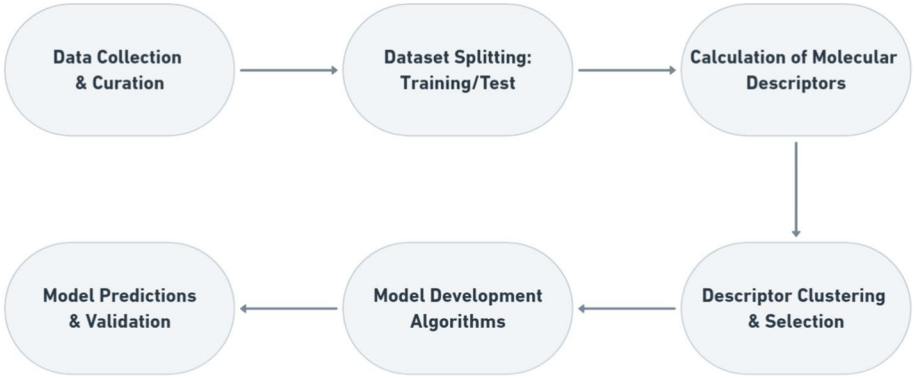


Fig. 1 General workflow for building a QSAR model

available molecular descriptors is limited, and extensive knowledge of the KNIME platform is needed [10].

Reactive oxygen species (ROS), which are commonly generated as byproducts of chemical reactions within the human body, inflict harmful effects on living cells. Oxidative stress occurs when the body's natural antioxidant defense system cannot effectively neutralize the free radicals produced within the body [11–13]. The antioxidant activity is a biological property of small molecules that has been widely studied and is related to the treatment and prevention of several diseases, including inflammatory diseases (e.g., rheumatoid arthritis), diabetes, and neurodegenerative diseases (e.g., Alzheimer's disease) [14, 15]. Several antioxidants have been developed and used in the cosmetic, pharmaceutical, and food industries [16]. Because of their diverse structures and established biological significance, di(hetero)aryl amines and amides are often considered a typical chemical class of great antioxidant capacity due in significant part to their ability to stabilize free radicals through resonance and electron-donating actions. For instance, a study made by Rajíc et al. (2010) consisted of the synthesis of novel ketoprofen amides and evaluation of their antioxidant capability, where findings indicated that these amides exhibited superior reducing activity compared to their counterparts, providing a direct experimental demonstration of their role as effective antioxidants [17]. Moreover, in a study carried out by Brizzi et al. (2022), the biological characterization of lipoic/capsaicin-related amides highlighted their efficacy as TRPV1 agonists with significant antioxidant properties, thus establishing a connection between the structural features of amides and their antioxidant capacity [18]. Additionally, a pertinent study by Afzal and Ahsan (2023) elaborates on the synthesis of 1-(1,3-Dioxoisoindolin-2-yl)-3-aryl urea analogs, demonstrating significant antioxidant properties, especially with specific substituents on the aryl ring, thus illustrating that structural modifications can modulate antioxidant activity [19]. In addition to the in-depth study of the antioxidant capacity of molecules that include di(hetero)aryl amine and amide scaffolds, these compounds are also excellent candidates for QSAR optimization since they have been widely investigated for use in various industries, such as materials science, pharmaceuticals and cosmetic research fields [20–23].

Due to the importance of finding new compounds with better antioxidant capacities, several antioxidant QSAR studies of different compound classes have been reported, consubstantiating the applicability of this type of study [24–27]. For instance, a study by Goya Jorge et al. (2016) built an antioxidant activity QSAR model using a library of

1373 compounds. As the experimental variable, the radical scavenging activity (RSA) was obtained using the DDPH (2,2-diphenyl-l-picrylhydrazyl) method. To get the QSAR model, Goya Jorge et al. (2016) used a neural network method called multilayer perceptron (MLP) while using DRAGON® software to calculate the molecular descriptors [24]. In a previous study, a library of 26 di(hetero)arylamine and amine derivatives was also used to prepare an antioxidant QSAR model. To build the QSAR model, Abreu et al. (2009) used the partial least squares projection of latent structures (PLS) method and DRAGON® software to calculate the molecular descriptors [25]. The antioxidant activity of each compound was experimentally determined using the DPPH method, which assesses each compound's free RSA, and the results are presented as IC_{50} values. The statistical performance was outstanding, with an R^2 value of 0.881 and a $Q_{ext}{}^2$ value of 0.843. The di(heteroaryl) amines and amides are thus promising scaffolds for developing new compounds with potent antioxidant activities [25]. Another study by Zhan et al. (2017) prepared a 3D-QSAR model to analyze the relationship between the antioxidant activity of 15 amine derivative compounds as additives of trimethylolpropane trioleate lubricating oil [27].

Despite the growing number of QSAR studies targeting antioxidant activity, few standardized, easily available procedures include free, open-source technologies from chemical library preparation to model validation. Furthermore, the implementation of replicable, step-by-step techniques that non-experts may follow is not well-documented. This fragmentation presents a barrier for researchers who wish to develop predictive models without costly software or extensive computational expertise, particularly in academic and resource-limited environments. Therefore, there is a need for well-documented methodologies that demonstrate the complete QSAR workflow using freely available resources. This article presents a complete methodology for preparing QSAR models using free and open-source software tools, from chemical library preparation to calculating and selecting molecular descriptors to QSAR model building and validation. As an example of an application of the methodology, an antioxidant QSAR model was prepared using a library of 70 di(hetero)aryl amines or amides. The main tools used were the OCHEM platform to calculate the molecular descriptors while PyQSAR was used to build the QSAR model. The developed antioxidant QSAR model presents excellent standard and cross-validated statistical parameters and can be used to predict and guide the synthesis of new di(hetero)aryl derivatives with improved antioxidant activities. We describe in detail the methodology used and present an easy-to-follow step-by-step protocol for developing similar QSAR models for different biological activities that may interest other researchers (Additional file 1).

2 Materials and methods

2.1 Library preparation and molecular descriptor calculations

A dataset of 70 di(hetero)aryl amines and amides was selected through an extensive literature search. Only compounds presenting the following characteristics were selected: (1) two aryl or heteroaryl scaffolds connected by a linker containing an amine or amide group; (2) a linker with a maximum length of three covalent bonds between rings; and (3) the availability of an experimental IC_{50} using the DPPH (2,2-diphenyl-1-picrylhydrazyl) method. The rationale for this selection was based on the study by Abreu et al. (2009) [25], in which a QSAR model was prepared using compounds with amine linkers

and benzothiophene rings. The types of linkers were expanded to linkers containing amine and amine groups, and the types of rings were extended to all those containing at least one aryl group. The IC_{50} value is the concentration required to inhibit DPPH radical formation by 50%. In total, 70 compounds presented the intended structure and were divided into 3 families (A, B and C) according to linker type (Fig. 2). Family A, with 26 compounds, contains an amine group (NH); family B, with 24 compounds, contains a methanamine group; and family C, with 20 compounds, contains an amide linker (NHCO). Figure 2 presents the scaffold structure for each family and an example of a compound belonging to each family. All compound structures are represented in Additional file 2.

As the antioxidant efficiencies of the 70 compounds used in this study differed by several orders of magnitude, for each compound, the IC_{50} values were transformed into pIC_{50} values (Table 1) according to the following equation:

$$pIC50_{50} = -\log(IC50_{50}) \quad (1)$$

The steps performed, from library selection to molecular descriptor, are presented in Fig. 3. The structures of the 70 compounds were drawn using Chemsketch software and then transformed into SMILES format using a conversion tool present in the same software [28]. The list and SMILES representation of the complete library is included as Additional file 3 in a ready-to-use format so that the complete QSAR modeling process performed in this study can be easily replicated. The structures were then uploaded to the OCHEM platform. OCHEM calculates a large number of molecular descriptors using different descriptor software, including Dragon, alvaDesc, MERA, MERSY, and MORDRED [7]. A total of 12,072 molecular descriptors were calculated (Fig. 3).

2.2 QSAR modeling using PyQSAR

The PyQSAR platform was used to prepare the antioxidant QSAR model. A primary objective of QSAR modeling is predicting the biological activity, physicochemical properties, or toxicity of novel chemicals within the developed model's applicability domain (AD). The importance of AD analysis is highlighted by the fact that the accuracy of such predictions is inextricably linked to how well the novel compounds fit within this

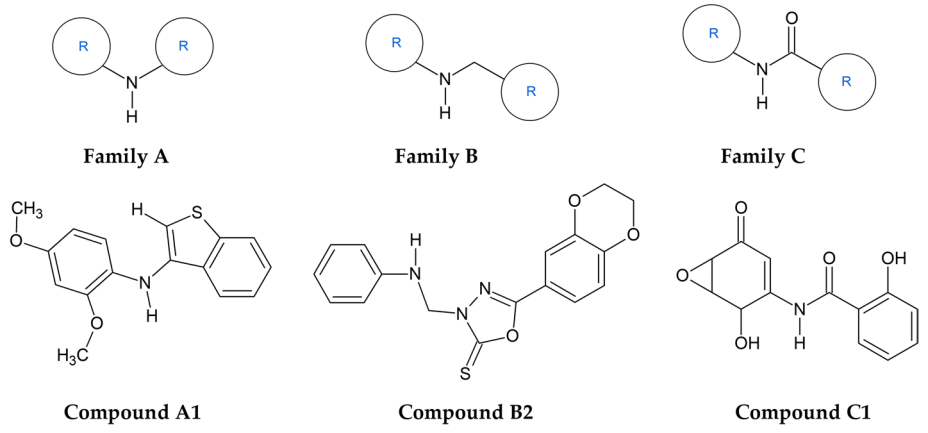


Fig. 2 Scaffold representation of the library of 70 di(hetero)aryl amine or amide compounds. Each family contains 2 heterocyclic groups (R) connected by a different linker. Above is the 2D scaffold representation for each family of compounds (**A**, **B** and **C**)

Table 1 Experimental and predicted pIC50 values and respective residuals for each compound

Compound	$pIC_{50\ (Experimental)}$	$pIC_{50\ (Predicted)}$	$\lvert\Delta pIC_{50}\rvert$	Compound	$pIC_{50\ (Experimental)}$	$pIC_{50\ (Predicted)}$	$\lvert\Delta pIC_{50}\rvert$
A1 [25]	4.06	4.11	0.05	B10 [36]	4.50	4.49	0.01
A2 [25]	3.61	3.51	0.10	B11 [36]	4.46	4.48	0.02
A3 [25]	3.77	3.68	0.09	B12 [36]	4.44	4.50	0.06
A4 [25]	3.84	3.87	0.03	B13 [36]	4.44	4.45	0.01
A5 [25]	3.91	4.05	0.14	B14 [36]	4.15	4.39	0.24
A6 [25]	3.81	3.28	0.53	B15 [36]	4.44	4.48	0.04
A7 [25]	4.07	3.83	0.24	B16 [36]	4.32	4.47	0.15
A8 [25]	4.30	4.09	0.21	B17 [36]	4.42	4.34	0.08
A9 [25]	3.89	3.69	0.20	B18 [36]	5.06	4.51	0.55
A10 [25]	4.14	3.83	0.31	B19 [36]	4.08	4.50	0.42
A11 [25]	3.38	3.66	0.28	B20 [36]	4.60	4.54	0.06
A12 [25]	2.39	2.44	0.05	B21 [36]	4.50	4.44	0.06
A13 [25]	2.29	2.21	0.08	B22 [36]	4.65	4.41	0.24
A14 [25]	3.81	3.81	0.00	B23 [36]	4.02	4.45	0.43
A15 [25]	3.98	3.65	0.33	B24 [37]	4.02	4.48	0.46
A16 [25]	3.69	3.75	0.06	C1 [38]	4.24	4.41	0.17
A17 [25]	3.05	2.91	0.14	C2 [39]	4.28	4.23	0.05
A18 [25]	2.43	2.47	0.04	C3 [39]	4.47	4.89	0.42
A19 [25]	2.91	2.89	0.02	C4 [39]	4.18	4.05	0.13
A20 [25]	2.10	2.13	0.03	C5 [39]	4.26	4.38	0.12
A21 [25]	4.26	4.15	0.11	C6 [39]	4.48	4.21	0.27
A22 [25]	2.81	3.66	0.85	C7 [40]	3.34	3.53	0.19
A23 [25]	4.30	3.88	0.42	C8 [40]	3.28	3.46	0.18
A24 [25]	1.84	2.05	0.21	C9 [40]	3.92	3.88	0.04
A25 [25]	2.50	2.57	0.07	C10 [40]	3.86	3.62	0.24
A26 [25]	2.26	2.63	0.37	C11 [41]	3.90	4.12	0.22
B1 [36]	4.71	4.60	0.11	C12 [41]	3.70	3.79	0.09
B2 [36]	4.49	4.39	0.10	C13 [41]	3.89	3.87	0.02
B3 [36]	4.43	4.49	0.06	C14 [42]	4.60	4.19	0.41
B4 [36]	4.53	4.42	0.11	C15 [42]	4.40	4.32	0.08
B5 [36]	4.66	4.69	0.03	C16 [43]	4.09	3.80	0.29
B6 [36]	4.17	4.61	0.44	C17 [43]	3.87	3.92	0.05

Table 1 (continued)

Compound	$pIC_{50\ (Experimental)}$	$pIC_{50\ (Predicted)}$	$\lvert\Delta pIC_{50}\rvert$	Compound	$pIC_{50\ (Experimental)}$	$pIC_{50\ (Predicted)}$	$\lvert\Delta pIC_{50}\rvert$
B7 [36]	4.45	4.49	0.04	C18 [43]	3.99	3.97	0.02
B8 [36]	4.05	4.30	0.25	C19 [44]	4.30	4.25	0.05
B9 [36]	4.52	4.45	0.07	C20 [44]	4.24	4.23	0.01

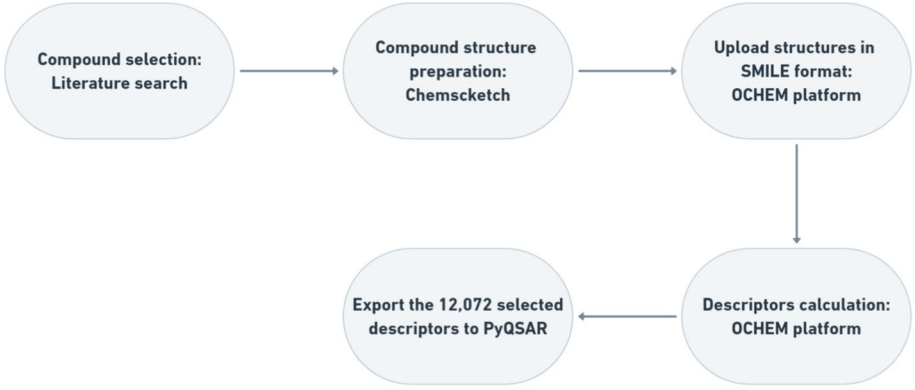


Fig. 3 The scheme represents the protocol for obtaining the molecular descriptors through the OCHEM platform (detailed steps are provided in Additional file 1)

specified domain [29]. Based on the structural and physicochemical similarity between the query compound and those in the model's training set, the AD offers a framework for calculating prediction uncertainty [30, 31]. PyQSAR provides built-in tools for separating training and test sets and defining the AD, which estimates the boundaries of the chemical space in which the QSAR model can reliably make predictions. Specifically, this study AD was determined based on the distance-based method in the descriptor space of the training set.

The 12,072 calculated descriptors and the experimental pIC_{50} values for each compound were uploaded to the PyQSAR platform and used as input (Fig. 4). The RSA of the 70 compounds (pIC_{50}), the value to be predicted with the QSAR model, was used as an independent variable. A QSAR model typically consists of fewer than 10 molecular descriptors [6]. For instance, if a model utilizes only four descriptors out of a pool of 5,000 descriptors, the potential number of combinations amounts to ${}_{5000}C_4$ (permutation of 5,000 by 4), equal to 6.24×10^{14} [6]. For this work, 12,072 molecular descriptors were calculated, which, using the same permutation formula, ${}_{12072}C_4$ (permutation of 12,072 by 4), gives a total of 2.12×10^{16} possible descriptor combinations. The vast search space renders an exhaustive approach impractical in such scenarios [32].

PyQSAR implements hierarchical clustering and genetic algorithms (GA) for descriptor selection to address this issue. The clustering process involves forming groups of descriptors with similar calculation values, starting from the first input descriptor (Fig. 4). After evaluating the obtained clusters, PyQSAR selected the molecular descriptors to be included in the QSAR model. First, the descriptors were grouped into clusters using a hierarchical clustering method. After forming the clusters, PyQSAR used the cophenetic correlation coefficient to evaluate the quality of the hierarchical clustering

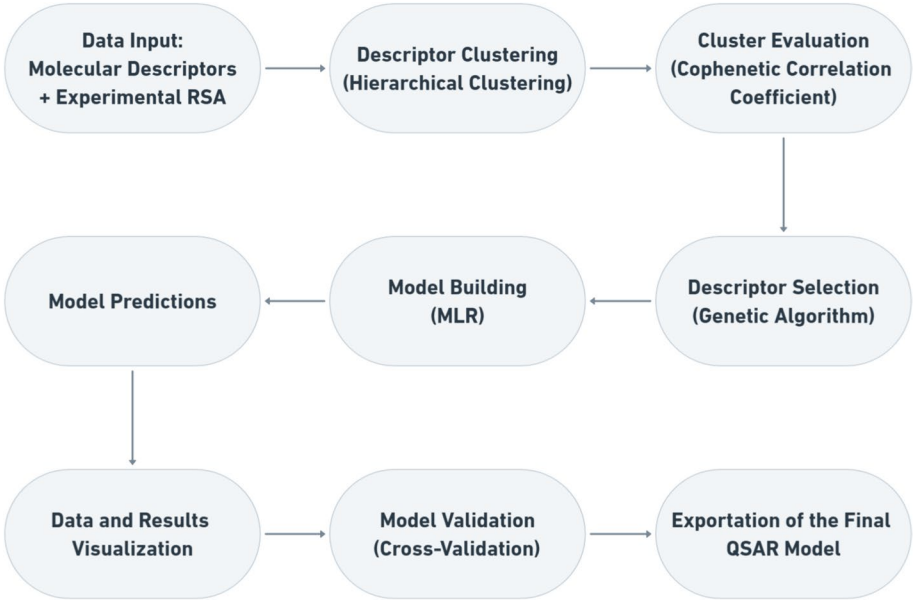


Fig. 4 Scheme demonstrating the QSAR export model creation process through PyQSAR software [6]

solution. This coefficient is the correlation between the pairwise distances among the objects in the original dataset and the pairwise distances among the objects in the dendrogram produced by the hierarchical clustering algorithm [33]. Following the formation and evaluation of the clusters, a GA was used to select a set of descriptors for the different clusters. This method helps to prevent the selection of descriptors with similar properties (belonging to the same cluster). The selections based on the GA were repeated until the ideal set of descriptors was obtained. Each time a new set of descriptors was generated, multiple linear regression (MLR) was performed to obtain the coefficients for a given set of descriptors. Then, the sets of descriptors were organized according to their score, and only a predefined number of descriptors advanced to the next step in the iterative selection process. During the description selection step, a Student's *t*-test assessed each molecular descriptor's statistical significance concerning the compounds' biological activity. Specifically, PyQSAR compares the distribution of descriptor values between two groups (based on a median activity threshold) using a univariate *t*-test for each descriptor. Low *p*-value descriptors (p-value < 0.05) are considered statistically significant and kept for subsequent modeling stages. This method guarantees that only descriptors that significantly contribute to activity are included in the QSAR model and help to minimize dimensionality. A step-by-step protocol is available to detail the complete descriptor selection process (Additional file 1).

2.3 QSAR model statistical validation

PyQSAR calculated several statistical parameters to evaluate the quality of the final QSAR model, including square correlation coefficient (R^2), root mean squared error (RMSE) and relative standard deviation of the residuals (RSR). The R^2 value represents the variation in the dependent variable that can be explained by the independent variables in a regression model. The RMSE is calculated as the average squared difference between the predicted and experimental values. The RMSE quantitatively assesses the difference between the predicted and experimental values in a QSAR model [34]. The

RSR value was calculated as the ratio between the RMSE and the observed standard deviation ($STDEV_{obs}$) according to the equation $RSR = RMSE/STDEV_{obs}$ [35]. The RSR provides an estimate of the spread of the residuals, which is the difference between the observed and predicted biological activity values.

The predictive stability and robustness of the QSAR model were evaluated through internal cross-validation using a repeated random sub-sampling approach. Specifically, cross-validation was performed 100 times by randomly removing five compounds from the training set. The model was retrained using the remaining compounds, and predictions were generated for the excluded compounds. This procedure allowed for the calculation of several key statistical parameters, namely:

- Q^2_{CV} (Cross-validated R^2): This metric represents the average predictive squared correlation coefficient obtained across the 100 iterations. It reflects the model's ability to predict unseen data and serves as an indicator of internal predictivity. A higher Q^2_{CV} value suggests that the model has good generalizability.
- $RMSE_{CV}$ (Root Mean Squared Error of Cross-Validation): This measures the average magnitude of the prediction error across all cross-validation runs. Lower $RMSE_{CV}$ values indicate that the predicted values are, on average, closer to the actual values, suggesting higher predictive accuracy.
- RSR_{CV} (Relative Standard Deviation of Residuals in Cross-Validation): This parameter captures the residual variability, normalized by the mean of observed values. A low RSR_{CV} value indicates that the model predictions are accurate and consistent across different cross-validation runs.

Collectively, these parameters provide a comprehensive assessment of the model's predictive performance, highlighting its accuracy and robustness under internal validation.

3 Results

3.1 QSAR model for determining the antioxidant activity of di(hetero)aryl amines or amides

At the end of the PyQSAR modeling process, an antioxidant QSAR model was obtained for the library of 70 di(hetero)aryl amines and amides. The QSAR model developed used four descriptors:

B06[C-O]: (alvaDesc) – presence/absence of C-O at topological distance 6.
Eig04_AEA(dm): (alvaDesc) – eigenvalue n. 4 from the augmented edge adjacency mat, weighted by edge degree.
JGI2: (alvaDesc) – mean topological charge index of order 2.
J_Dz(p): (alvaDesc) – Balaban-like index from the Barysz matrix weighted by polarizability.

The QSAR model equation and the statistical parameters were as follows:

$$pIC50_{50(Predicted)} = 4.3384 + 1.1236 * B06\,[C-O] + (-1.7150) * Eig04_{AEA(dm)} + 1.5673 * JGI2 + (-2.0874) * J_{Dz(p)} \quad (2)$$

$N = 70; STDEV_{obs} = 0.7012 R^2 = 0.8905; Q^2_{CV} = 0.8676 RMSE = 0.2328;$
$RMSE_{CV} = 0.2467 RSR = 0.3320; RSR_{CV} = 0.3518$

where N is the number of compounds used, R^2 is the square correlation coefficient found using the linear regression method, RMSE is the root mean square error, and RSR is the relative standard deviation of the residuals. $RMSE_{CV}$ and RSR_{CV} correspond to the RMSE and RSR, respectively, after performing the cross-validation method. Finally, Q^2_{CV} is the average value of the MLR method score after cross-validation was performed one hundred times.

For each compound, the experimental and predicted pIC_{50} values, calculated by Eq. (2), are shown in Table 1. To graphically assess the performance of the QSAR model, a plot of the predicted versus the experimental pIC_{50} values is displayed in Fig. 5. The residual value of each compound ($|\Delta pIC_{50}|$) is also given in Table 1.

3.2 QSAR model validation

The model was validated for its robustness and predictive power by calculating different statistical parameters, including cross-validation parameters, as demonstrated by the R^2, Q^2_{CV}, RMSE and RSR values. It is generally accepted that, as a rule of thumb, a QSAR model has good predictive power when the R^2 value is greater than 0.6 and the Q^2_{CV} is greater than 0.5 [45]. For both correlation coefficients, the values obtained were significantly above the considered threshold values, with values of 0.8905 and 0.8909 for R^2 and Q^2_{CV}, respectively. Q^2_{CV} is a parameter calculated by applying the average score of the MLR method after the cross-validation process is performed one hundred times. This value reflects the ability of the model to predict values close to the experimental values and, in essence, the precision of the model [46].

The RMSE value is also widely used as a validation indicator, as the lower the RMSE is, the better the model performance. Nevertheless, no definite RMSE threshold value is generally accepted. Singh et al. (2005) proposed another statistical parameter, RSR,

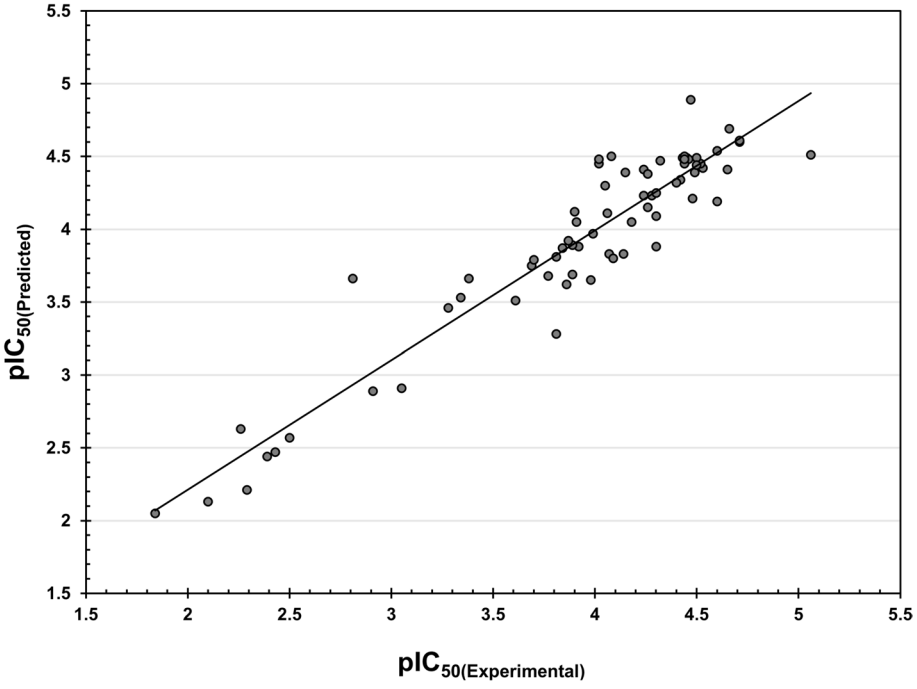


Fig. 5 Predicted versus experimental pIC50 values of the 70 di(hetero)aryl amine and amide libraries used in the QSAR model

that normalizes the RMSE values by including $STDEV_{obs}$ as a scaling factor [47]. An RSR value below 0.5 is considered suitable for statistically validating a QSAR model. For the presented model, the RSR and RSR_{CV} were 0.3320 and 0.3518, respectively, which are significantly below the threshold of 0.5, thus thoroughly validating the model [35].

For external validation, the dataset of 70 compounds was randomly divided into a training set (80%) and a test set (20%). A QSAR model was independently developed using only the training set, including fresh descriptor selection via a Genetic Algorithm. This model was then applied to predict the test set compounds' antioxidant activity (pIC_{50}). The predictive performance of this unseen data was evaluated using the external coefficient of determination (Q^2_{EXT}), using the formula bellow:

$$Q^2_{EXT} = 1 - \frac{\sum (Y_{EXP} - Y_{PRED})^2}{\sum (Y_{EXP} - \overline{Y}_{TRAIN})^2} \quad (3)$$

where Y_{EXP} corresponds to the experimental pIC_{50} values of the test set, Y_{PRED} are the pIC_{50} values predicted by the model for the test set and $\overline{Y}_{TRAIN}$ corresponds to the mean of the experimental pIC_{50} values of the training test used to build the model. After calculation, we obtained a value for Q^2_{EXT} equals to 0.6589. This value is higher than the commonly accepted minimum limit of 0.5 for predictive QSAR models, confirming the model's generalizability and robustness [45]. Following this assessment, the final QSAR model was constructed using the complete dataset (all 70 compounds) to maximize predictive accuracy and descriptor significance.

The correlation matrix between the $pIC_{50\ (experimental)}$ values of antioxidant activity and the selected molecular descriptors is shown in Table 2. Obtaining a correlation between each pair of the selected molecular descriptors is crucial in developing a QSAR model to assess the importance of each molecular descriptor and verify that the model has been built with molecular descriptors that showed low collinearity between them.

As shown in Table 2, the correlations ranged from 0.0734 for the B06[C-O]/JGI2 pair to 0.7417 for the Eig04_AEA(dm)/J_Dz(p) pair. These values show a lack of a strong correlation between the selected molecular descriptors, indicating that they all contribute to the variation in the antioxidant activity of the di(hetero)aryl amides and amine library and probably describe different characteristics of the compounds.

4 Discussion

4.1 QSAR model: di(hetero)aryl amine and amide library selection and statistical validation

An antioxidant QSAR model was prepared using a library of 70 di(hetero)aryl amines and amide derivatives. When selecting the compounds, great care was taken to choose a large enough number of compounds. Although not standardized, it is generally accepted that at least 5 compounds should be used per molecular descriptor. In this study, 4

Table 2 Correlation matrix between the pIC50 (experimental) values and the molecular descriptors

	B06[C-O]	Eig04_AEA(dm)	JGI2	J_Dz(p)	$pIC_{50\ (Experimental)}$
B06[C-O]	1.0000	0.5456	0.0734	0.5617	0.8007
Eig04_AEA(dm)	–	1.0000	0.2927	0.7417	0.5565
JGI2	–	–	1.0000	0.1421	0.1706
J_Dz(p)	–	–	–	1.0000	0.7511
$pIC_{50\ (Experimental)}$	–	–	–	–	1.0000

descriptors were used, meaning that the minimum number of compounds used would be 20 compounds. The library of compounds used consisted of 70 compounds, far exceeding the minimum accepted value. Although the general heteroaryl-linker-heteroaryl structure was considered, there was substantial variety in the heteroaryl rings and linker types (Fig. 2 and Additional file 2). In general, compounds containing amine and amide groups present radical scavenging activity. This ability is usually related to the presence of nitrogen atoms with lone pairs of electrons (as in the case of amine and amide functional groups), which are capable of interacting with free radicals through the hydrogen atom transfer (HAT) pathway [26]. Additionally, heteroaryl rings are considered to have radical scavenging activity through the radical adduct formation (RAF) reaction, where the double bonds in the aryl ring neutralize HOO• radicals. The outstanding statistical validation parameters in such a large and diverse library confirm that the QSAR model will perform well when predicting the antioxidant activity of other compounds with congeneric structures.

The software tools used in this study are all freely accessible for academic use and are readily available for use by others. For this reason, a detailed protocol of the different steps performed was prepared and is provided in Additional file 1. The complete library used as input in the OCHEM platform is also available in XLS file format as Additional file 3. The protocol can easily replicate this antioxidant QSAR model to prepare other antioxidant QSAR models using different compound libraries of interest or to model other biological activities of interest.

4.2 QSAR model interpretation

The primary objective of a QSAR model is to predict the modeled bioactivity of new compounds not used in the model implementation. Nevertheless, an analysis of the molecular descriptors in the equation may lead to a better understanding of the chemical features that promote the modeled bioactivity and elucidate possible mechanisms of action. However, many descriptors are calculated using complex equations, and often, the equations are not available, making understanding each descriptor contribution difficult.

Four descriptors were selected for the antioxidant QSAR model presented, and an initial analysis was performed where the correlation between each pair of descriptors was determined (Table 2). All correlation pairs were less than 0.75, indicating that each descriptor describes different chemical features contributing to the compound's antioxidant activity. Of the four descriptors, two contributed positively to the predicted pIC_{50} (B06[C-O] and JGI2), while the other two contributed negatively (Eig04_AEA(dm) and J_Dz(p)). This contribution can be seen by the positive or negative coefficient in Eq. (2).

The first molecular descriptor selected, B06[C-O], belongs to the family of 2D atom pair descriptors and is of straightforward significance, indicating the presence or absence of C-O pairs at a topological distance of 6 covalent bonds, meaning that pairs of C-O atoms separated by 6 covalent bonds are a favorable feature for antioxidant activity. The presence of C-O motifs, particularly in hydroxyl, carbonyl, or ether groups, is well known to influence antioxidant activity due to their roles in HAT or single electron transfer (SET) mechanisms [48]. The spatial distribution of such functional groups may also impact molecular reactivity with radical species. Figure 6 presents a compound with

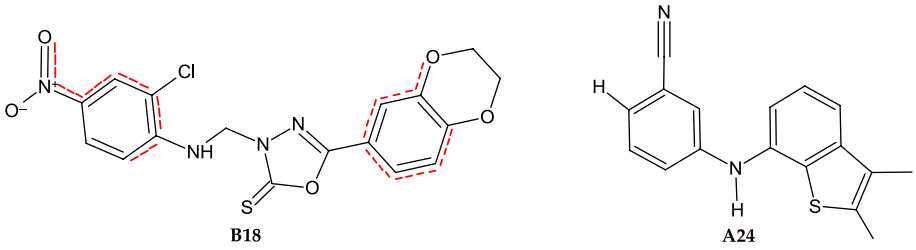


Fig. 6 Compounds with the presence (B18) or absence (A24) of a C-O pair at the topological distance of the bonds. Examples of the 6 covalent bonds between pairs of C-O are represented by traced red lines

C-O pairs at a topological distance of 6 bonds (B18), a B06[C-O] value of 1, and another compound without C-O pairs (A24), with a B06[C-O] value of 0.

The other descriptors are more challenging to analyze in terms of significance. JGI2 and J_Dz(p) are 2D-type descriptors representing the chemical structure as a topological graph or matrix. JGI2 is calculated considering the charge transfer between pairs of atoms at a topological distance of 2 and reflects the electron distribution over the molecular graph. The descriptor J_Dz(p) combines molecular topology with atomic polarizability, highlighting the role of electron cloud flexibility and molecular branching in stabilizing radical species [49]. Finally, Eig04_AEA(dm) is obtained from a binary zero–one matrix (also called the adjacency matrix) that describes the nature of molecules based on their dipole moment [49]. This descriptor accounts for molecular topology and electronegativity differences, which influence electron transfer and interaction with reactive species, which are key factors in antioxidant mechanisms.

The JGI2 descriptors suggest a positive dependence between the antioxidant activity of the compounds, and it's extremely important as molecules capable of delocalizing charge across conjugated systems are often more effective in radical scavenging due to enhanced stabilization of radical intermediates [50]. On the other hand, J_Dz(p) and Eig04_AEA(dm) negatively contribute to the antioxidant activity, suggesting that higher values of the dipolar moment and polarizability of the compounds have a negative effect on the antioxidant activity. This negative correlation between dipole moment and antioxidant activity was also verified in a study using diphenylpropionamide derivatives as a library of compounds, further confirming the mechanisms observed in a library of di(hetero)aryl amines and amides [51].

Internal and external validation procedures were rigorously performed to comprehensively evaluate the robustness and predictive power of the developed QSAR model. Internal validation was carried out through repeated random sub-sampling cross-validation (100 iterations), which yielded a high cross-validated coefficient of determination ($Q^2_{CV} = 0.8909$), indicating excellent predictive stability within the training set. The associated $RMSE_{CV}$ and RSR_{CV} values were low, with $RSR_{CV} = 0.3518$, confirming the consistency and accuracy of the model across multiple validation runs. Furthermore, the model's fit to the training data was demonstrated by a high R^2 value of 0.8905, reinforcing the strength of the linear correlation between the selected molecular descriptors and antioxidant activity.

For external validation, the dataset was split into a training set (80%) and a test set (20%). A separate model was constructed exclusively from the training data and then applied to predict the antioxidant activity of the test set compounds. The model achieved a $Q^2_{EXT} = 0.6589$, surpassing the commonly accepted threshold of 0.5 for predictive

QSAR models. This confirms the model's ability to generalize beyond the training set and make reliable predictions on unseen compounds. Combining these internal and external metrics strongly supports the statistical validity, robustness, and real-world applicability of the developed QSAR model.

5 Conclusion

QSAR modeling has become an essential in silico approach for predicting the biological activities of small molecules, particularly in the context of antioxidant activity, which is vital for addressing various health issues linked to oxidative stress. This study presents a comprehensive methodology for developing QSAR models using only free and open-source software tools. A detailed protocol is included with all the steps required for QSAR model preparation, from chemical library preparation to molecular descriptor calculation and QSAR model computation and validation.

An antioxidant QSAR model using a library of 70 di(hetero)aryl amines and amides was successfully modeled using the presented methodology. The OCHEM platform was used to calculate 12,072 molecular descriptors, and PyQSAR was used to construct the antioxidant QSAR model. A GA was used for the clusterization as well as for descriptor selection methods, and in total, four key descriptors were selected and used to construct the QSAR model: B06[C-O], Eig04_AEA(dm), JGI2, and J_Dz(p). These descriptors summarize the structural and chemical properties of the compounds and were carefully chosen to capture distinct chemical features that can have a key role in antioxidant activity. The positive contribution of descriptors like B06[C-O] and JGI2 indicates the importance of functional groups and electron delocalization in radical scavenging. In contrast, negative contributions from J_Dz(p) and Eig04_AEA(dm) highlight the influence of dipole moment and polarizability on antioxidant efficiency.

The antioxidant QSAR model was then implemented using the MLR method. Rigorous validation procedures, including standard and cross-validation, demonstrated excellent predictive stability with a high Q^2_{CV} value (0.8909), while external validation confirmed the model's ability to generalize to unseen data, achieving a Q^2_{EXT} value of 0.6589. These results underscore the statistical robustness and reliability of the QSAR model.

Furthermore, the availability of the software tools, protocols, and compound library ensures that this model can be easily replicated for future studies, offering a versatile platform for developing other antioxidant QSAR models or models targeting different biological activities.

Abbreviations

QSAR	Quantitative structure–activity relationships
ROS	Reactive oxygen species
RSA	Radical scavenging activity
MLP	Multilayer perceptron
PLS	Partial least squares
GA	Genetic algorithm
MLR	Multiple linear regression
HAT	Hydrogen atom transfer
RAF	Radical adduct formation
SET	Single electron transfer
AD	Applicability domain

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s44371-025-00290-0.

Additional file 1. Protocol for implementing a QSAR model using the antioxidant QSAR model prepared in this study as an example

Additional file 2. Detailed chemical structures of the library of 70 di(hetero)aryl amines and amides

Additional file 3. Complete library of di(hetero)aryl amines and amides in XLS format with the SMILES format and experimental pIC_{50} values for all compounds

Acknowledgements
This work was supported by national funds through FCT/MCTES (PIDDAC): CIMO, UIDB/00690/2020 (https://doi.org/10.54499/UIDB/00690/2020) and UIDP/00690/2020 (https://doi.org/10.54499/UIDP/00690/2020); and SusTEC, LA/P/0007/2020 (https://doi.org/10.54499/LA/P/0007/2020).

Author contributions
C.M. and R.A. conceptualized the study; C.M. wrote the main manuscript and prepared the figures; R.A. corrected and prepared the final version. All authors reviewed the article.

Funding
This research received no external funding.

Data availability
Data generated or analysed during this study are included in this published article and its supplementary information files.

Declarations

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 13 January 2025 / Accepted: 5 September 2025
Published online: 16 September 2025

References

1. Patel HM, Noolvi MN, Sharma P, Jaiswal V, Bansal S, Lohan S, et al. Quantitative structure–activity relationship (QSAR) studies as strategic approach in drug discovery. Med Chem Res. 2014;23:4991–5007. https://doi.org/10.1007/s00044-014-1072-3.
2. Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010;29:476–88. https://doi.org/10.1002/minf.201000061.
3. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. Qsar modeling: where have you been? Where are you going to? J Med Chem. 2014;57(12):4977–5010. https://doi.org/10.1021/jm4004285.
4. Dimitrov SD, Diderich R, Sobanski T, Pavlov TS, Chankov GV, Chapkanov AS, et al. QSAR toolbox–workflow and major functionalities. SAR QSAR Environ Res. 2016;27(3):203–19. https://doi.org/10.1080/1062936X.2015.1136680.
5. De P, Kar S, Ambure P, Roy K. Prediction reliability of QSAR models: an overview of various validation tools. Arch Toxicol. 2022;96(5):1279–95. https://doi.org/10.1007/s00204-022-03252-y.
6. Kim S, Cho KH. PyQSAR: a fast QSAR modeling platform using machine learning and jupyter notebook. Bull Korean Chem Soc. 2019;40(1):39–44. https://doi.org/10.1002/bkcs.11638.
7. Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, et al. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J Comput Aided Mol Des. 2011;25(6):533–54. https://doi.org/10.1007/s10822-011-9440-2.
8. Gramatica P, Papa E. Qsar modeling of bioconcentration factor by theoretical molecular descriptors. QSAR Comb Sci. 2003;22(3):374–85. https://doi.org/10.1002/qsar.200390027.
9. Mansouri K, Moreira-Filho JT, Lowe CN, Charest N, Martin T, Tkachenko V, et al. Free and open-source QSAR-ready workflow for automated standardization of chemical structures in support of QSAR modeling. J Cheminform. 2024;16(1):1–18. https://doi.org/10.1186/s13321-024-00814-3.
10. Kausar S, Falcao AO. An automated framework for QSAR model building. J Cheminform. 2018;10:1–23. https://doi.org/10.1186/s13321-017-0256-5.
11. Ruba SL, Nisar B, Raza AR, Saadia M, Tahir MN, Sajjad N, et al. Synthesis and antioxidant screening of novel indole amines. J Iran Chem Soc. 2022;19:2693–704. https://doi.org/10.1007/s13738-021-02482-8.
12. Yen GC, Hsieh CL. Antioxidant effects of dopamine and related compounds. Biosci Biotechnol Biochem. 1997;61(10):1646–9. https://doi.org/10.1271/bbb.61.1646.
13. Ahmed D, Mir H. Study of antioxidant, anti-protease and anti-urease potential of Schiff bases of acetophenone with different amines. J Chem Soc Pak. 2014;36(5):915.
14. Geronikaki AA, Gavalas AM. Antioxidants and inflammatory disease: synthetic and natural antioxidants with anti-inflammatory activity. Comb Chem High Throughput Screen. 2006;9(6):425–42. https://doi.org/10.2174/138620706777698481.

15. Kontogiorgis AC, Pontiki AE, Hadjipavlou-Litina D. A review on quantitative structure-activity relationships (QSARs) of natural and synthetic antioxidants compounds. Mini-Rev Med Chem. 2005;5(6):563–74. https://doi.org/10.2174/1389557054023233.
16. Rani K. Role of antioxidants in prevention of diseases. J Appl Biotechnol Bioeng. 2017;4(1):00091. https://doi.org/10.15406/jabb.2017.04.00091.
17. Rajić Z, Hadjipavlou-Litina D, Pontiki E, Kralj M, Šuman L, Zorc B. The novel ketoprofen amides–synthesis and biological evaluation as antioxidants, lipoxygenase inhibitors and cytostatic agents. Chem Biol Drug Des. 2010;75(6):641–52. https://doi.org/10.1111/j.1747-0285.2010.00963.x.
18. Brizzi A, Maramai S, Aiello F, Baratto MC, Corelli F, Mugnaini C. Lipoic/capsaicin-related amides: synthesis and biological characterization of new TRPV1 agonists endowed with protective properties against oxidative stress. Int J Mol Sci. 2022;23(21):13580. https://doi.org/10.3390/ijms232113580.
19. Afzal O, Ahsan MJ. An efficient synthesis of 1-(1, 3-dioxoisoindolin-2-yl)-3-aryl urea analogs as anticancer and antioxidant agents: an insight into experimental and in silico studies. Molecules. 2023;29(1):67. https://doi.org/10.3390/molecules290100 67.
20. Lee A, Mercader AG, Duchowicz PR, Castro EA, Pomilio AB. Qsar study of the DPPH radical scavenging activity of di (hetero) arylamines derivatives of benzo [b] thiophenes, halophenols and caffeic acid analogues. Chemom Intell Lab Syst. 2012;116:33–40. https://doi.org/10.1016/j.chemolab.2012.03.016.
21. Etxebeste-Mitxeltorena M, Plano D, Astrain-Redin N, Morán-Serradilla C, Aydillo C, Encío I. New amides and phosphoramidates containing selenium: studies on their cytotoxicity and antioxidant activities in breast cancer. Antioxidants. 2021;10(4):590. https://doi.org/10.3390/antiox10040590.
22. Feng J, He Y, Zeng X, Xiang B, Chen S, Shi P, et al. Synthesis and application of cardanol-heterocyclic amines as photostabilizers and antioxidants in polylactic acid. ACS Sustain Chem Eng. 2023;11(37):13527–34. https://doi.org/10.1021/acssuschemeng.3c00777.
23. Petchey MR, Grogan G. Enzyme-catalysed synthesis of secondary and tertiary amides. Adv Synth Catal. 2019;361(17):3895–914. https://doi.org/10.1002/adsc.201900694.
24. Goya Jorge E, Rayar AM, Barigye SJ, Jorge Rodríguez ME, Sylla-Iyarreta VM. Development of an in silico model of DPPH• free radical scavenging capacity: prediction of antioxidant activity of coumarin type compounds. Int J Mol Sci. 2016;17(6):881. https://doi.org/10.3390/ijms17060881.
25. Abreu RM, Ferreira IC, Queiroz MJR. Qsar model for predicting radical scavenging activity of di (hetero) arylamines derivatives of benzo [b] thiophenes. Eur J Med Chem. 2009;44(5):1952–8. https://doi.org/10.1016/j.ejmech.2008.11.011.
26. Ngo TC, Nguyen TH, Dao DQ. Radical scavenging activity of natural-based cassaine diterpenoid amides and amines. J Chem Inf Model. 2019;59(2):766–76. https://doi.org/10.1021/acs.jcim.8b00847.
27. Zhan S, Duan H, Hua M, Xu H, Shang H, Jin Y, et al. Studies of antioxidant performance of amine additives in lubricating oil using 3D-QSAR. Sci China Technol Sci. 2017;60:299–305. https://doi.org/10.1007/s11431-016-0560-9.
28. ACD/ChemSketch (Freeware) 2021.2.1, Advanced Chemistry Development, Inc. (ACD/Labs), Toronto, ON, Canada, www.acdlabs.com.
29. Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. J Mol Graph Model. 2008;26(8):1315–26. https://doi.org/10.1016/j.jmgm.2008.01.002.
30. Zhong S, Lambeth DR, Igou TK, Chen Y. Enlarging applicability domain of quantitative structure–activity relationship models through uncertainty-based active learning. ACS EST ENGG. 2022;2(7):1211–20. https://doi.org/10.1021/acsestengg.1c00434.
31. Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemom Intell Lab Syst. 2015;145:22–9. https://doi.org/10.1016/j.chemolab.2015.04.013.
32. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
33. Saraçli S, Doğan N, Doğan İ. Comparison of hierarchical cluster analysis methods by cophenetic correlation. J Inequal Appl. 2013;1:1–8. https://doi.org/10.1186/1029-242X-2013-203.
34. Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK. Validation of QSAR models-strategies and importance. Int J Drug Des Discov. 2011;3:511–9. https://doi.org/10.37285/ijddd.2.3.1.
35. Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL. Model evaluation guidelines for sys-tematic quantification of accuracy in watershed simulations. Trans ASABE. 2007;50(3):885–900. https://doi.org/10.13031/2013.23153.
36. Ma L, Xiao Y, Li C, Xie ZL, Li DD, Wang YT, et al. Synthesis and antioxidant activity of novel mannich base of 1, 3,4-oxadiazole derivatives possessing 1,4-benzodioxan. Bioorg Med Chem. 2013;21(21):6763–70. https://doi.org/10.1016/j.bmc.2013.08.002.
37. Amigó M, Terencio MC, Mitova M, Iodice C, Payá M, De Rosa S. Potential antipsoriatic avarol derivatives as antioxidants and inhibitors of PGE2 generation and proliferation in the HaCaT cell line. J Nat Prod. 2004;67(9):1459–63. https://doi.org/10.1021/np049873n.
38. Saitoh T, Suzuki E, Takasugi A, Obata R, Ishikawa Y, Umezawa K, et al. Efficient synthesis of (±)-parasitenone, a novel inhibitor of NF-κB. Bioorg Med Chem. 2009;19(18):5383–6. https://doi.org/10.1016/j.bmcl.2009.07.120.
39. Koufaki M, Calogeropoulou T, Detsi A, Roditis A, Kourounakis AP, Papazafiri P, et al. Novel potent inhibitors of lipid peroxidation with protective effects against reperfusion arrhythmias. J Med Chem. 2001;44(24):4300–3. https://doi.org/10.1021/jm010962w.
40. Jabbari A, Davoodnejad M, Alimardani M, Assadieskandar A, Sadeghian A, Safdari H, et al. Synthesis and SAR studies of 3-allyl-4-prenyloxyaniline amides as potent 15-lipoxygenase inhibitors. Bioorg Med Chem. 2012;20(18):5518–26. https://doi.org/10.1016/j.bmc.2012.07.025.
41. Jung JC, Lim E, Lee Y, Kang JM, Kim H, Jang S, et al. Synthesis of novel trans-stilbene derivatives and evaluation of their potent antioxidant and neuroprotective effects. Eur J Med Chem. 2009;44(8):3166–74. https://doi.org/10.1016/j.ejmech.2009.03.011.
42. Lee H, Lee K, Jung JK, Cho J, Theodorakis EA. Synthesis and evaluation of 6-hydroxy-7-methoxy-4-chromanone-and chroman-2-carboxamides as antioxidants. Bioorg Med Chem. 2013;15(11):2745–8. https://doi.org/10.1016/j.bmcl.2005.03.118.
43. Yogeeswari P, Sharma M, Samala G, Gangadhar M, Karthick S, Mallipeddi S, et al. Discovery of novel tetrahydro-pyrazolo [4,3-c] pyridines for the treatment of neuropathic pain: synthesis and neuropharmacology. Eur J Med Chem. 2013;66:211–20. https://doi.org/10.1016/j.ejmech.2013.05.022.

44. Alexiou P, Demopoulos VJ. A diverse series of substituted benzenesulfonamides as aldose reductase inhibitors with antioxidant activity: design, synthesis, and in vitro activity. J Med Chem. 2012;53(21):7756–66. https://doi.org/10.1021/jm101008m.
45. Golbraikh A, Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Mol Divers. 2000;5:231–43. https://doi.org/10.1023/A:1021372108686.
46. Consonni V, Ballabio D, Todeschini R. Comments on the definition of the Q 2 parameter for QSAR validation. J Chem Inf Model. 2009;49(7):1669–78. https://doi.org/10.1021/ci900115y.
47. Singh J, Knapp HV, Arnold JG, Demissie M. Hydrological modeling of the Iroquois River watershed using HSPF and SWAT 1. J Am Water Resour Assoc. 2005;41(2):343–60. https://doi.org/10.1111/j.1752-1688.2005.tb03740.x.
48. Shahidi F, Ambigaipalan P. Phenolics and polyphenolics in foods, beverages and spices: antioxidant activity and health effects–a review. J Funct Foods. 2015;18:820–97. https://doi.org/10.1016/j.jff.2015.06.018.
49. Khan K, Kumar V, Colombo E, Lombardo A, Benfenati E, Roy K. Intelligent consensus predictions of bioconcentration factor of pharmaceuticals using 2D and fragment-based descriptors. Environ Int. 2022;170:107625. https://doi.org/10.1016/j.envint.2022.107625.
50. Charlton NC, Mastyugin M, Török B, Török M. Structural features of small molecule antioxidants and strategic modifications to improve potential bioactivity. Molecules. 2023;28(3):1057. https://doi.org/10.3390/molecules28031057.
51. Urbani P, Ramunno A, Filosa R, Pinto A, Popolo A, Bianchino E, et al. Antioxidant activity of diphenylpropionamide derivatives: synthesis, biological evaluation and computational analysis. Molecules. 2008;13(4):749–61. https://doi.org/10.3390/molecules13040749.

Publisher's Note