RESEARCH ARTICLE

Editorial Process: Submission:07/05/2025 Acceptance:19/10/2026 Published:24/10/2026

Quantitative Study of the Structure–Activity Relationships of 4-(Thio)chromenone 6-O-Sulfamate Derivatives Using QSAR Analysis

Souad Madani[1], Fatima- Zahra Abou-Elaaz[2], Khadija El-Moustaqim[3], Riad Maroua[1], Najat Qisse[1], Ahmed Al-Shami[1], Youssef El Mourabit[4], Ibrahim Alsayer[5], Jamal Mabrouki[1]*

Abstract

Objective: Breast cancer is the most common cancer in women; its mortality ranking varies by region, despite recent improvements in early detection and treatment. To account for these factors and modify them during the design of novel drugs, our research aimed to identify potential descriptors influencing Steroid Sulfatase (STS) inhibitory activity by analyzing a series of molecules, specifically 4-(thio)chromenones, to establish a quantitative relationship between the inhibitory activity of Steroid Sulfatase (STS) and the structure of these molecules. **Methods:** Using 24 derivatives of 2-substituted 4-(thio)chromenone 6-O-sulfamate compounds, obtained from PubChem and whose biological activity was assessed experimentally, we created a QSAR model. Multiple linear regression (MLR) was used as the statistical technique to create the 2D QSAR model after the 2D descriptors were computed using PaDEL-Descriptor. We developed the QSAR model and used Lipinski's Rule of Five and Veber's rule to validate the pharmacokinetic profile of the 24 compounds. Consequently, we examined combinations of two and three descriptors. Multiple linear regression (MLR), which is helpful for producing highly predictive and comprehensible QSAR models, was utilized to create the QSAR model equation using the selected combinations. **Results:** At the molecular level, Steroid Sulfatase (STS) is expressed in 87% of breast cancers. Inhibiting STS could decrease estrogen levels. A 2D QSAR analysis using the multiple linear regression (MLR) method was conducted on a dataset of 24 derivatives of 2-substituted 4-(thio)chromenone 6-O-sulfamate acting as Steroid Sulfatase inhibitors to establish a quantitative relationship between biological activity and molecular descriptors and to identify potential descriptors affecting STS inhibition. A robust and predictive model was developed with a good R^2 value of 0.89 and a leave-one-out cross-validation coefficient (LOO Q^2) of 0.83, indicating that the model adequately explains the quantitative relationship. In addition to the QSAR model, the pharmacokinetic profiles of the 24 compounds were validated according to Lipinski's Rule of Five and Veber's rule to assess their potential as drug candidates. **Conclusion:** Provided that untested compounds fall within the model's applicability domain, the validation results indicate that this model may be used to predict their biological activity. Furthermore, based on a comparison of the computational results with the criteria of Lipinski's Rule of Five and Veber's rule, the 24 compounds can be considered potential drug candidates, as they satisfy the requirements of both rules.

Keywords: Steroidal sulfatase, 4(thio)chromenone 6-Osulfamate analogs, QSAR, molecular docking, SAR

Asian Pac J Cancer Prev, **27 (10)**, 3543-3553

Introduction

Breast cancer is the most commonly diagnosed cancer in women. Despite recent advancements in early detection and treatment, breast cancer remains the second leading cause of cancer-related mortality in women worldwide. At the time of diagnosis, approximately 70% to 75% of patients have estrogen receptor (ER) positive breast cancer [1]. Among the two types of estrogen receptors, α and β, ERα is the predominant subtype expressed in breast tumor tissues [2]. This receptor is a major mediator of estrogen's tumorigenicity in cells. By binding to its receptors,

*[1]Laboratory of Spectroscopy, Molecular Modeling, Materials, Nanomaterials, Water and Environment, (CERNE2D), Mohammed V University in Rabat, Faculty of Science, AV Ibn Battouta, BP1014, Agdal, Rabat, Morocco. [2]Mohammed V University in Rabat, Scientific Institute of Rabat, AV Ibn Battouta, BP1014, Agdal, Rabat, Morocco. [3]Improvement and Valuation of Plant Resources, Faculty of Sciences, Ibn Tofaïl University KENITRA-University Campus, Kenitra 14000, Morocco. [4]Science and Technical Research Lab, ESEF, University of Ibn Zohr, Morocco. [5]Department of Chemical and Materials Engineering, College of Engineering, Northern Border University, Arar, Saudi Arabia. *For Correspondence: jamal.mabrouki@um5r.ac.ma*

estrogen activates signalling cascades that regulate the expression of numerous genes controlling metabolism, cell cycle progression, or apoptosis, ultimately promoting tumor growth [3]. Dependence on estrogen has been successfully exploited in ER-positive cancers to develop anti-hormonal therapies. These therapies are primarily based on depriving tumor cells of estrogen by ovarian suppression (OS) or aromatase inhibitors, or by blocking estrogen receptors [2].

Many studies have been conducted to identify a new target for inhibiting estrogen synthesis in the mammary gland, revealing the existence of two enzymes: aromatase and steroid sulfatase (STS), both of which increase estrogen concentration in circulation [4]. However, STS activity in cancerous tissues is a hundred times greater than that of the aromatase pathway [5], making it a more promising target for discovering new candidates for breast cancer treatment [6].

Several studies have shown the importance of substituted sulfamates in inhibiting STS. Our study aims to identify potential descriptors affecting STS inhibitory activity in order to consider and modify them when designing new drugs [7, 8]. For this purpose, we developed a QSAR model using 24 derivatives of 4-(thio) chromenone 6-O-sulfamate 2-substituted compounds retrieved from PubChem, whose biological activity is determined experimentally. The 2D descriptors were calculated using PaDEL-Descriptor, and multiple linear regression (MLR) was adopted as the statistical method to develop the 2D QSAR model. In addition to developing the QSAR model, we verified the pharmacokinetic profile of the 24 molecules based on Lipinski and Veber's rules [9].

This study aims to analyze a series of molecules, specifically 4-thiochromenone 6-O-sulfamates consisting of 24 derivatives, to establish a quantitative relationship between the inhibitory activity of Steroid Sulfatase (STS) and the structure of these molecules. Additionally, we will examine the pharmacokinetic properties of these inhibitors based on Lipinski's rule of five and Veber's rule.

Materials and Methods

QSAR model construction

The principle of a QSAR approach, as its name suggests, relies on establishing a mathematical relationship linking the experimentally determined biological activity of a set of compounds under the same experimental conditions to molecular descriptors calculated from the structure of the molecules based on statistical methods. QSAR aims to analyze structural data to detect factors influencing biological activity (Figure 1) [10-12].

Generally, to predict biological activity using a QSAR method, it is first necessary to construct a mathematical model, which requires following a certain number of steps.

Database Selection and Preparation

The choice of the database is a crucial step in developing a QSAR model because it heavily relies on reference experimental data. These data need to be as reliable as possible and derived from the same experimental protocol. Several databases are available for collecting data, including ChEMBL, Ligand Database, PubChem, etc [13].

The dataset used in this study consists of 24 compounds selected from publicly available databases such as PubChem and ChEMBL. These compounds were chosen based on the availability of reliable experimental biological activity data (pIC_{50}) against steroid sulfatase (STS).

Although the dataset size is relatively limited, it is comparable to datasets used in similar QSAR studies, where the emphasis is placed on data quality, consistency, and structural diversity rather than quantity.

The collected data must be divided into two types: a training dataset, which will be used for constructing the QSAR model, and a prediction or test dataset, which will be used for validating the former model. Several approaches can be used for this selection. The most commonly used approach is random division, which involves splitting the database through simple random selection [14].

The chemical structures were selected from ChEMBL along with their inhibition values in nm. After selecting

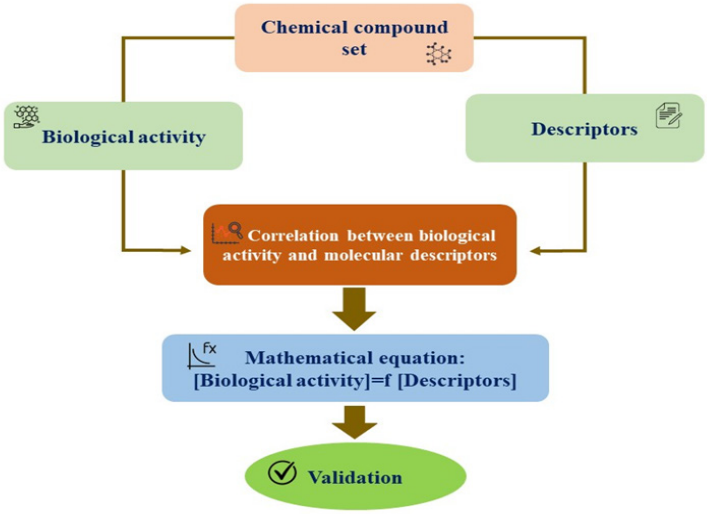

Figure 1. Principle of QSAR Method

the molecules for this study, structural optimization with energy minimization was performed using the Avogadro software version 1.2.0. The optimized structures were then used as inputs for calculating descriptors [15].

The inhibition of STS in nm was converted to a logarithmic scale by taking the negative logarithm (pIC50 = -log10IC_{50}) so that it could be used as the dependent variable for QSAR model development [16].

The dataset was divided into two sets: a 70% training set, which was used for model development, and a 30% test set, which was used for external model validation. This selection was made using "Random selection" by the Drug Theoretics and Cheminformatics (DTC) Laboratory software version 1.0.6.

Model Development

To develop a model, a data analysis method is needed to quantitatively relate the studied activity with the set of calculated descriptors. Indeed, there are several statistical methods for building a model and analyzing its statistical data. Some methods are linear, such as multiple linear regression (MLR) and partial least squares regression (PLS), while others are non-linear, such as artificial neural networks (ANN) [17, 18].

QSAR Model Equation

Using the DTC QSAR tool software, the "best-set selection" option allowed us to choose the combination of descriptors that would provide good prediction quality. Therefore, we analyzed combinations of two descriptors and three descriptors. The chosen combinations were used to develop the QSAR model equation through multiple linear regression (MLR), which is useful for generating highly predictive and interpretable QSAR models [19].

Statistical model validation

After developing the model, all statistical parameters must be analyzed and interpreted, and its robustness and quality must be verified through the validation process. Therefore, this validation will be divided into two parts: to assess the robustness of the model, internal validation is employed, and to evaluate predictive capacity, external validation is necessary. The applicability domain of a QSAR model is the physicochemical, structural, or biological space, the knowledge or information on which the model's training set was developed, and for which it is applicable to make predictions for new compounds [20]. Indeed, the absence of this domain means that each QSAR model can formally predict the activity of a chemical even if its structure is completely different from those included in the subset.

The robustness of the QSAR model is validated to increase confidence in predictions and to verify the applicability domain of the models. In this study, all required measures of internal and external validation are calculated using the respective training and test sets [21].

The Lipinski and Veber rules are applied to identify "drug-like" compounds among the 29 inhibitors. To calculate the following descriptors: NHA, NHD, MW, logP, and PAS-RBC, we used the software Marvin Sketch.

Software

The QSAR (2D) studies were conducted using the QSARINS version 2.2.3, (University of Insubria and Varese, Italy) [http://www.qsar.it] [22]. QSARINS employs a hybrid approach combining genetic algorithms (GA) with multiple linear regression (MLR), which is useful in predicting highly predictive and interpretable QSAR models.

Docking studies were performed using the molecular modelling simulation software AutoDock Tools 1.5.7, developed by The Scripps Research Institute, San Diego, California [http://www.autodock.scripps.edu/] [23]. The molecular structures were sketched in ChemSketch version 2.5 [http://www.acdlabs.com] [24]. And they were cleaned using the clean structure option available in ChemSketch tools, followed by energy minimization using the Avogadro software version 1.2.0 [https://Avogadro.cc], employing the steepest descent algorithm in the MMFF94s force field [25, 26]. PLIP (Protein-Ligand Interaction Profiler) analyses were conducted using the free online software available at https://projects.biotec.tu-dresden.de/plip-web/plip/.

Results

Structure-activity relationship studies

An ensemble of data of derivatives of 4-thiochromenone is collected and analyzed to determine the relationship between the chemical structures represented by the descriptors and the biological activity. The chemical structures of 24 representative compounds and their experimental pIC50 values are illustrated in the Figure 2:

Among the 24 compounds, eight (103350841-103372905-103372957-103373388-103373057-103373504-103373510-103373550) were randomly chosen for testing the model (test set) (30%), and the remaining 16 compounds were kept for the training set (70%).

Data preprocessing resulted in the elimination of 426 descriptors with constant values equal to 0, and 254 intercorrelated descriptors. In the step of selecting the best combination of descriptors, an analysis of 219,453 combinations of 2 descriptors was performed, and all generated combinations yielded either moderate or poor prediction quality. However, the analysis of 2,147,483,647 combinations of 3 descriptors resulted in 15 combinations with good prediction quality. Multiple linear regression was used to develop the QSAR models from these 15 combinations, and the table below shows the results obtained (Table 1).

According to the obtained models, we noticed that model number 9 has the best R^2 value (0.89), with the lowest standard error of the estimate (SEE) value (0.42). However, among these 15 models, model number 9 is chosen.

$$pIC_{50} = 26.3519\,(\pm 1.7493) - 3.7003(\pm 0.5982)\times GATS8p - 10.7364(\pm 1.2478)\times GATS2p + 23.4981(\pm 2.6116)\times MATS1s$$

$$N=24 \quad R^2 = 0.89 \quad R^2_{adj} = 0.87 \quad S = 0.42 \quad F=46.14$$

The R^2 value represents the coefficient of determination,

Table 1. QSAR Models Obtained from 15 Combinations

Numbers	Equation MLR	Internal Validation metrics							External Validation metrics						
1	pIC50 = -26.1865 -1.2716 x maxssO -0.1934 AATSC6v +8.2311 MWC2	R^2	R^2a	SEE	Q^2_{LOO}	$SPED_{LOO}$	$Avg.rm^2_{(LOO)}$	$\Delta rm^2_{(LOO)}$	MAE	$Q^2(F1)_{(test)}$	$Q^2(F2)_{(test)}$	$Avg.\ rm^2_{(test)}$	$\Delta rm^2_{(test)}$	$CCC_{(test)}$	$MAE,_{(test)}$
2	pIC50 = -7.0318 -0.1997 x AATSC6v +0.1283 TWC -0.2533 SssO	0.82	0.79	0.54	0.73	0.6	0.65	0.077	0.48	0.76	0.74	0.71	0.11	0.88	0.35
3	pIC50 = 6.9325 -0.1933 AATSC6v -1.2817 max-ssO +0.8526 SP-3	0.79	0.75	0.59	0.71	0.62	0.61	0.14	0.5	0.65	0.63	0.63	0.04	0.83	0.53
4	pIC50 = 22.2899 +0.0765 ATSC2i -3,5523 GATS8p -15.9647 GATS1s	0.8	0.77	0.57	0.73	0.6	0.64	0.12	0.45	0.77	0.76	0.71	0.15	0.88	0.34
5	pIC50 = -23.9697 -0.0023 ATSC6v +4.9383 MWC4 -0.2173 SssO	0.77	0.73	0.62	0.7	0.63	0.59	0.15	0.43	0.63	0.61	0.53	0.14	0.79	0.53
6	pIC50 = 44.584 - 0.257 AATS3i +8.1492 GATS6v +27.0155 MATS1s	0.84	0.81	0.52	0.73	0.6	0.64	0.1	0.41	0.69	0.67	0.61	0.14	0.83	0.43
7	pIC50 = 3.023 - 0.2413 SssO -0.1952AATSC6v +0.7208 SP-3	0.78	0.74	0.6	0.71	0.62	0.61	0.15	0.46	0.68	0.67	0.64	0.13	0.84	0.45
8	pIC50 = -0.796 -3.3957 AATSC3i +1.3062 MAXDP +1.3734 SdssC	0.76	0.72	0.63	0.63	0.7	0.51	0.2	0.51	0.67	0.66	0.57	0.21	0.82	0.49
9	pIC50 = 26.3519 -3.7003 GATS8p -10.7364 GATS2p +23.4981 MATS1s	0.89	0.87	0.42	0.83	0.48	0.77	0.009	0.32	0.82	0.81	0.83	0.008	0.92	0.37
10	pIC50 = 25.2715 -9.042GATS2i -3.7251 GATS8p +27.4367MATS1s	0.87	0.85	0.45	0.79	0.52	0.73	0.006	0.35	0.83	0.82	0.82	0.04	0.92	0.36
11	pIC50 = 8.7327 + 1.4332MAXDP -6.8511SIC5 -6.6343GATS2c	0.74	0.7	0.65	0.64	0.69	0.52	0.19	0.51	0.69	0.67	0.67	0.17	0.84	0.47
12	pIC50 = 8.0914 + 4.3825 GATS3v -8.0743GATS1s +1.3179maxdssC	0.73	0.68	0.67	0.57	0.76	0.44	0.16	0.54	0.62	0.6	0.45	0.28	0.77	0.51
13	pIC50 = -4.7795 + 3.911 GATS3i + 1,3695MAXDP +1,3013 SdssC	0.76	0.71	0.63	0.63	0.7	0.5	0.2	0.52	0.7	0.68	0.63	0.18	0.84	0.45
14	pIC50 = 12,1843 +1.0117 maxdssC -8,7358GATS1s +0.9308 VC-3	0.7	0.65	0.7	0.56	0.76	0.43	0.18	0.54	0.64	0.62	0.51	0.24	0.79	0.5
15	pIC50 = 2.4715 +0.6777 VP-2 +0.9467 ndssC -0.0071SpAbs_Dzp	0.67	0.61	0.74	0.54	0.78	0.39	0.24	0.53	0.66	0.64	0.5	0.25	0.79	0.42

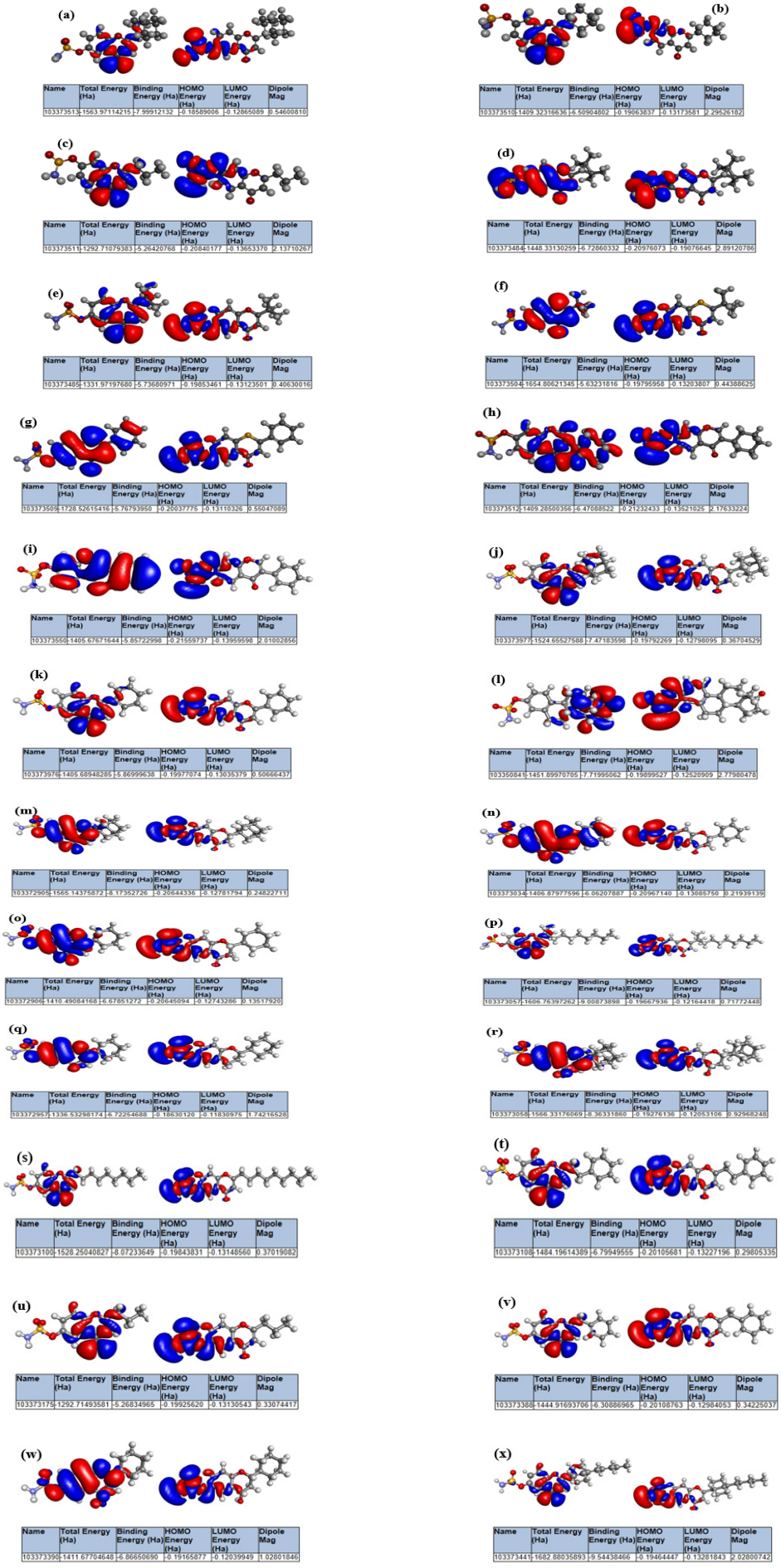

Figure 2. Chemical Structure and Experimental Biological Activity of Derivatives of 4-(thio)chromenone 6-O-sulfamate 2-substituted.

indicating the proportion of the variance in y (pIC50) explained by the independent variables (descriptors) in the model. A value closer to 1 indicates a more adequate model, with a higher R^2 suggesting a better fit of the regression line to the data [27].

In our case, R^2 = 0.89, strongly indicating the correlation between the descriptors used and the studied biological activity (STS inhibition). However, R^2 has a limitation: its value increases with the addition of independent variables to the model, which can be misleading as some added variables may be insignificant. The adjusted R^2 (R^2_{adj}) overcomes this issue by decreasing if the added variable is insignificantand increasing if it is significant. In our equation, the R2adj value is 0.87, indicating a very good agreement between the correlation and the data variation.

The F value = 46.14 strongly suggests that the regression relationship is statistically significant and not by chance. The correlation matrix for the biological activity pIC50 and the selected descriptors used to build the QSAR model is presented in Table 2, demonstrating the absence of correlation between the selected descriptors.

Table 3 shows the Variance Inflation Factors (VIF) of the descriptors in the QSAR model. The VIF values of the three descriptors are all below 5, suggesting the absence of collinearity among the selected descriptors and indicating that the resulting model has good stability.

Validation of the Chosen Model

Internal Validation

The predictive power of the equation is validated by the method of cross-validation. This is a practical and reliable method for testing the significance of a model. Parameters such as Q^2_{LOO}, SEE, $SPED_{(LOO)}$, Avg. $rm^2_{(LOO)}$, $\Delta rm^2_{(LOO)}$, and MAE are calculated using this validation method (Table 4).

Q^2_{LOO} is the coefficient of cross-validation (leave-one-out) and reflects the stability and predictive ability of the model. In this study, Q^2_{LOO} = 0.83, indicating good model robustness.

SEE (Standard Error of Estimation)

Measures the dispersion of predicted values around the regression line [28]. A lower SEE value indicates better accuracy. In our case, SEE = 0.42, confirming that the predictions are close to the regression line.

Table 2. Correlation Matrix Among Selected Descriptors

	MATS1s	GATS2p	GATS8p
MATS1s	1	0.298	-0.013
GATS2p	0.298	1	-0.189
GATS8p	-0.013	-0.189	1

Table 3. Variance Inflation Factors (VIF)

Term	Coef	VIF
Constant	26.35	
MATS1s	23.5	1.1
GATS2p	-10.74	1.14
GATS8p	-3.69	1.04

Table 4. Internal Validation Results

Q^2_{LOO}	SEE	$SPED_{(LOO)}$	Avg. $rm^2_{(LOO)}$	$\Delta rm^2_{(LOO)}$	MAE
0.83	0.42	0.48	0.77	0.009	0.32

MAE (Mean Absolute Error)

Represents the average absolute difference between predicted and experimental values [29]. A lower value indicates better predictive performance. Here, MAE = 0.32, demonstrating good model accuracy.

$SPED_{(LOO)}$ (Standard Deviation of Prediction Errors)

Represents the standard deviation of prediction errors. Lower values indicate better predictive capability. In this study, SPED showed the lowest value among the tested models, confirming its reliability [30].

Both Avg. $rm^2_{(LOO)}$ and $\Delta rm^2_{(LOO)}$ are also important for evaluating the quality of QSAR model prediction. Their values should be > 0.5 and < 0.2, respectively. For our case, Avg. $rm^2_{(LOO)}$ = 0.77 and $\Delta rm^2_{(LOO)}$ = 0.009, proving the predictive capacity of the model.

Overall, all parameters calculated for internal validations have demonstrated the robustness of the model. However, this validation alone is not sufficient; external validation is essential to evaluate the predictive capacity of new compounds that were not used to generate the model (the test set).

External validation

External validation, as mentioned earlier, is performed by applying the model obtained from the training set to the test set dataset, and measuring performance using various criteria such as $Q^2(F1)_{Test}$, $Q^2(F2)_{Test}$, Avg. $Rm^2_{(test)}$, $\Delta rm^2_{(test)}$, $CCC_{(test)}$, et $MAE_{(Test)}$. The results obtained are illustrated in Table 5.

From the initial overview, we observe that the obtained values (Avg. $Rm^2_{(test)}$ = 0.83, $\Delta rm^2_{(test)}$ = 0.008, $MAE_{(Test)}$ = 0.37) In external validation, the obtained values are very close to those of internal validation. For the $Q^2(F1)_{Test}$ External Test for the test set, it is determined by the following equation [31]:

$$Q^2(F1) = 1 - \frac{PRESS}{SS'ext)\ y(train)}$$

Where $\bar{y}TR$ represents the mean of the responses of the training set, PRESS is the sum of predictive squares, and SS(ext)($\bar{y}TR$) is the total sum of squares of the external set calculated by the mean of the training set.

The value of the $Q^2(F2)$ Test is calculated by the following equation:

$$Q^2(F1) = 1 - \frac{PRESS}{SS'ext)\ y(test)}$$

The $Q^2(F2)$ Test differs from the $Q^2(F1)$ Test by using the mean of the test set instead of the training set. The

Table 5. External Validation Results

$Q^2(F1)_{Test}$	$Q^2(F2)_{Test}$	Avg. $Rm^2_{(test)}$	Δrm^2 (test)	$CCC_{(test)}$	$MAE_{(Test)}$
0.82	0.81	0.83	0.008	0.92	0.37

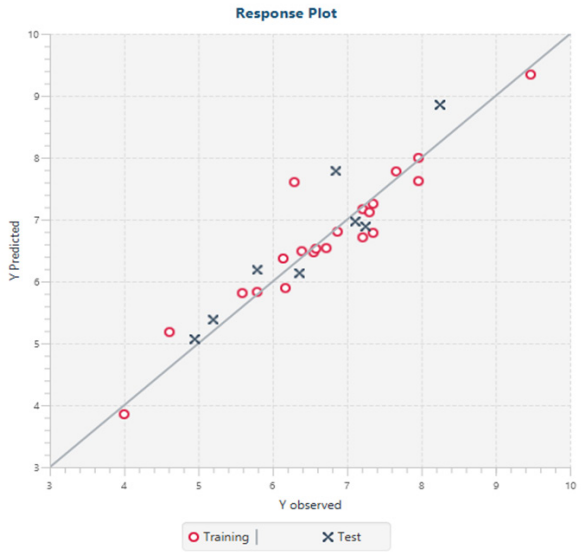

Figure 3. Linear Relationship between Experimental pIC_{50} and Predicted pIC_{50}

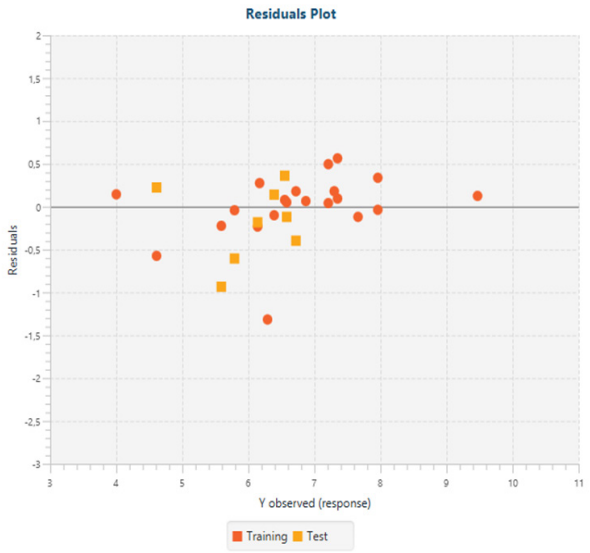

Figure 4. Graphical Representations of Residual Values against Predicted Values.

mean of the training set is used as a reference value. So, we compare the $Q^2(F2)$ Test with the $Q^2(F1)$ Test, and if their values are close and high, it demonstrates the predictive ability of the model [32].

CCC is the Concordance Correlation Coefficient. It is used to measure the agreement between experimental and predicted data. The CCC measures both precision (how much observations deviate from the regression line) and accuracy (how much the regression line deviates from the concordance line), and therefore any deviation of the regression line from the concordance line results in a CCC value less than 1.

From the results of external validation, we notice that the values (Avg. $Rm^2_{(test)} = 0.83$, $\Delta rm^2_{(test)} = 0.008$, $MAE_{(Test)} = 0.37$) in external validation are very close to those of internal validation, and the value of $Q^2(F1)$ Test (0.82) ≈ $Q^2(F2)$ Test (0.83) and CCC = 0.92 approaches 1, indicating that the regression line does not deviate much from the concordance line. This demonstrates the predictive capability of the model on the test set (Figures 3 and 4).

The applicability domain

The AD, or Application Domain, is the theoretical region of chemical space defined by model descriptors and modelled responses, and thus by the chemical properties of the training set, represented in each model by specific molecular descriptors. Since even a robust, significant, and validated QSAR cannot reliably predict modelled properties for the entire universe of chemicals, its application domain must be defined, and predictions are considered reliable only for chemicals within this domain [33, 34].

From the analysis of Table 6, we observe that compound (103373511) in the training set is considered an outlier, indicating that it is dissimilar to the majority of compounds in the training set. Table 7 indicates that compound (103373504) is outside the applicability domain of the model.

Descriptor Contribution

Figure 5 illustrates that the three descriptors (Table 8) are essential for constructing our model. Each of the three contributes either positively or negatively to the pIC_{50}.

Interpretation of selected descriptors

GATS2p is an autocorrelation descriptor. Autocorrelation descriptors of chemical compounds can be represented at the atomic level or the molecular surface level, reflecting the structural characteristics of molecular topology and atomic properties. The negative sign of GATS2p indicates that an increase in this descriptor leads to a decrease in inhibitory activity [35].

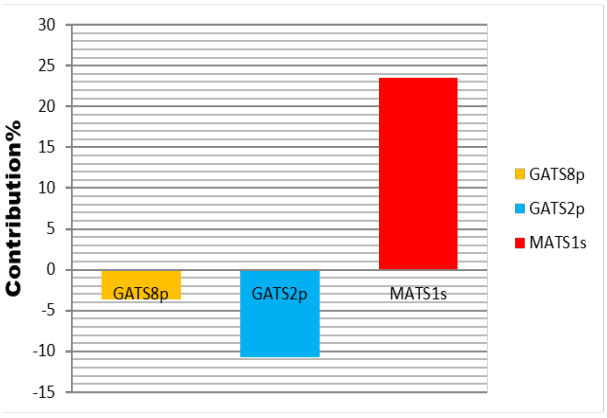

Figure 5. Descriptor Contribution in the Model

Table 6. Status of the Applicability Domain of the Training Set

Name	GATS8p	GATS2p	MATS1s	pIC_{50}	Predicted Response	Status(AD)
103372906	0.94	1.09	-0.2	6.55	6.471242101	In
103373034	0.94	1.02	-0.26	5.59	5.812903954	In
103373058	0.85	1.05	-0.23	6.58	6.528785844	In
103373100	0.79	1.14	-0.2	6.39	6.489473793	In
103373108	0.72	1.05	-0.25	6.72	6.539868481	In
103373134	0.75	1.04	-0.28	5.79	5.831279268	In
103373175	0.76	1.03	-0.26	6.14	6.3716017	In
103373390	0.82	1.12	-0.26	4.61	5.183305649	In
103373441	0.8	1.04	-0.18	7.96	7.996071713	In
103373446	0.84	1.08	-0.21	7.21	6.713659358	In
103373465	0.9	0.88	-0.18	9.47	9.343860532	In
103373480	0.75	1.17	-0.18	7.35	6.785357589	In
103373483	0.81	1.06	-0.22	6.87	6.804416567	In
103373484	1.02	0.88	-0.25	7.35	7.254952698	In
103373485	0.63	0.9	-0.28	7.66	7.778415598	In
103373509	0.98	0.88	-0.26	7.21	7.167985449	In
103373511	1.44	1.03	-0.26	4	3.855368652	Outlier
103373512	0.91	1.04	-0.2	7.3	7.119072098	In
103373513	0.94	1.05	-0.17	6.29	7.605640748	In
103373976	0.82	1.01	-0.28	6.17	5.894347107	In
103373977	0.93	1.03	-0.18	7.96	7.622391103	In

* GATS8p indicates that the presence of polarizable atoms at a topological distance equal to 8 influences inhibitory activity. It is important to note that this contribution is always negative because the mathematical relationship only implies negative values [36, 37].

*MATS1s is also a descriptor belonging to the 2D autocorrelation block [38].

The descriptors GATS2p and GATS8p (weighted by atomic polarizabilities) represent the polarizability distribution of atoms at various topological distances for each atom in a molecule. The negative coefficients of the QSAR equation suggest that if there is too much polarizability distribution across a molecule's atomic elements, this may be detrimental to how well a molecule interacts with the STS active site since this could alter both the electronic balance and steric fit between the two species, which is necessary for effective binding. GATS2p provides a measure of polarizability interactions that occur over two bonds, while GATS8p reflects the way electrons are distributed throughout the entire molecule. The combination of these two descriptors indicates that both local and global electronic arrangements will greatly impact the inhibition of an agonist to its receptor. MATS1s is a Moran autocorrelation descriptor which is weighted based on the atoms' intrinsic state (or I-state). This I-state indicates both how many electrons are in the atom's electron configuration as well as how readily those electrons can move within the electron environment around an atom. A positive coefficient indicates a favorable electronic distribution and an organized state of atoms potentially providing increased strength of STS inhibition as they will increase intermolecular interactions with residues of the enzyme binding pocket.

Veber's rule is separate from Lipinski's Rule of 5

Lipinski's Rule of 5 is a method used to identify compounds with high absorption or permeability compared to those with low absorption. Therefore, for a compound to be considered a drug, its molecular weight should not exceed 500 Da. Otherwise, its diffusion through

Table 7. Status of the Applicability Domain of the Test Set

Name	GATS8p	GATS2p	MATS1s	PIC50	Predicted Response	Status(AD)
103350841	1.21	0.98	-0.19	7.25	6.8881339	Inside-AD
103372905	0.95	1.03	-0.17	6.85	7.783365207	Inside-AD
103372957	0.88	1.11	-0.26	4.95	5.068649029	Inside-AD
103373057	0.8	1.07	-0.21	7.11	6.96903701	Inside-AD
103373388	0.81	1.03	-0.26	5.79	6.186584564	Inside-AD
103373504	0.78	0.77	-0.27	8.25	8.854076409	Outside-AD
103373510	1.2	1.08	-0.21	5.2	5.381535979	Inside-AD
103373550	0.9	0.96	-0.28	6.36	6.135139405	Inside-AD

Table 8. Significance of the Chosen Descriptors

Descriptor	Significance
GATS8p	GATS8p is lag8/weighted by atomic polarizabilities
GATS2p	GATS2p is lag2/weighted by polarizabilities
MATS1s	Moran autocorrelation of lag 1 weighted by I-state

biological membranes will be difficult, and it will be poorly absorbed [39]. Additionally, there should be 5 or fewer hydrogen bonds because too many hydrogen bonds make the molecule too polar, hindering its diffusion into lipophilic areas. Also, the hydrogen bond acceptors should be 10 or fewer, and the log P ≤5.

In addition to Lipinski's rule, Veber introduced two other criteria to identify compounds that do not have problems with oral bioavailability, namely, the number of rotatable bonds ≤10 rotatable bonds and PSA ≤140 Å².

Figure 6 indicates that the values of all molecules are in line with Lipinski's rule, with the number of hydrogen bond acceptors being less than 10, the number of hydrogen bond donors being less than 5, and the logarithm of P is less than 5.

Figure 7 shows that the molecular weight of all compounds is less than 500 Da. In Figure 8, we observe that the number of rotatable bonds is less than 10 for all compounds, and the polar surface area is less than 140 Å².

Based on all these results, none of the compounds violated Lipinski's Rule of 5, indicating that these inhibitors have a pharmaceutical profile and can be considered as potential drug candidates for breast cancer. The same applies to Veber's rule, as none of the compounds in our series violated the rule, demonstrating good compliance with passive molecular transport across membranes.

Discussion

Limitations and future perspectives

Acknowledge more explicitly (29 compounds is small for MLR with 3 descriptors). However, the compounds were carefully selected based on the availability of reliable and consistent experimental biological activity data (pIC50) against steroid sulfatase (STS).

Furthermore, the dataset covers a representative chemical space of 4-(thio)-chromenone derivatives, ensuring the robustness and relevance of the developed

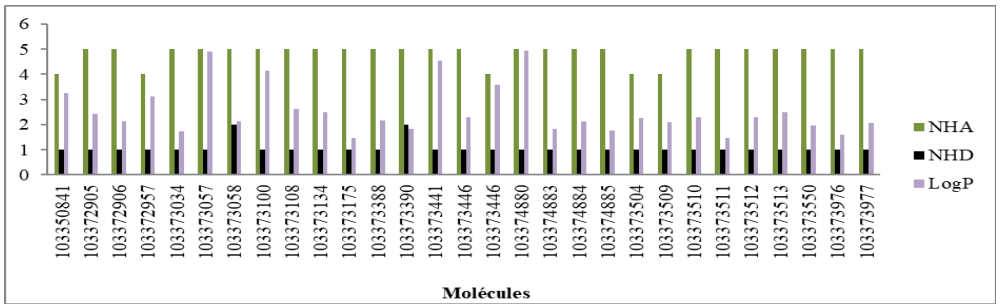

Figure 6. Lipinski Parameter Values (NHA = H-bond acceptors, NHD = H-bond donors, LogP) of the 29 compounds.

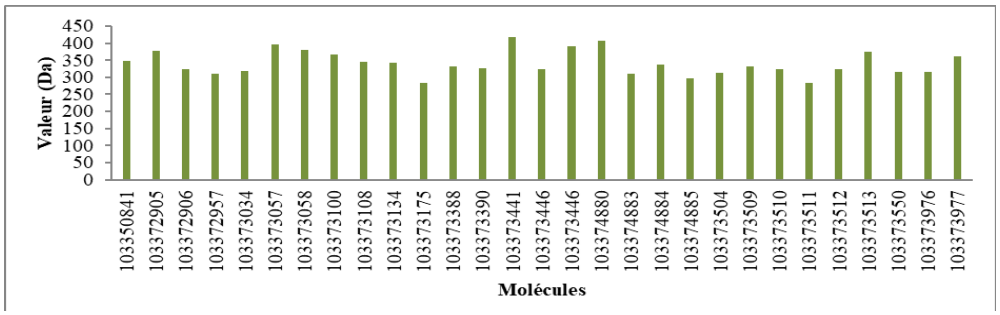

Figure 7. Molecular Weight Values of the 29 Compounds

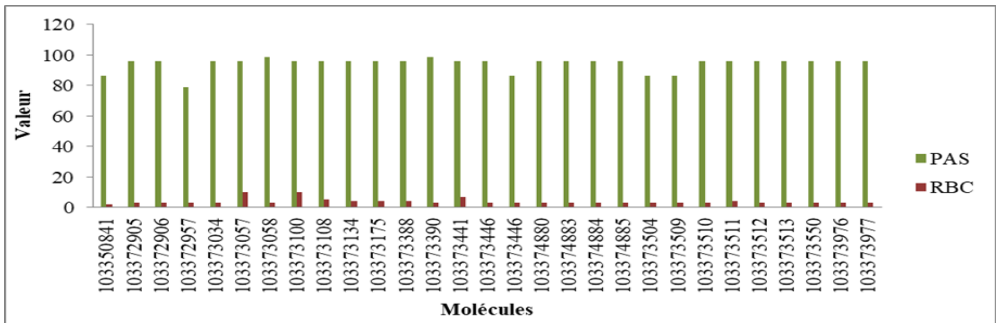

Figure 8. Veber Parameters of the 29 Compounds

QSAR model within this specific class of compounds. This is critical – the study has no experimental validation at all to further enhance the predictive performance of the model.

Although multiple linear regression (MLR) is a well-established method in QSAR modeling, its application remains relevant due to its simplicity, interpretability, and robustness. This integrative approach enhances the understanding of structure–activity relationships and supports the identification of key descriptors influencing biological activity.

The molecular descriptors that were selected for use in this analysis depended on statistical relevance, physical characteristics, as well as providing an explanation of how to perform QSAR assessments. This has allowed for a clearer picture of the relationship between the molecular structure of compounds and STS inhibition activity, therefore improving the predictive accuracy of the model generated. The use of well-established descriptors allows for a clearer understanding of the relationship between molecular structure and biological activity. Moreover, these descriptors have been widely validated in the literature, supporting the reliability and robustness of the developed QSAR model. Therefore, no experimental (in vitro or in vivo) validation has been performed. Further experimental studies are required to confirm the predicted biological activities and to validate the reliability of the proposed model.

In conclusion, during this work, we conducted a QSAR study on a series of 29 molecules of 4-(thio) chromenone 6-O-sulfamate 2-substituted using the MLR statistical analysis method. This study aimed to establish a quantitative structure-activity relationship. This study showed that the inhibitory activity of STS is influenced by 3 2D autocorrelation descriptors (GATS8p, GATS2p, and MATS1s). These 3 descriptors contribute either positively or negatively to the biological activity, thus providing insight into the design of new compounds through structural modification. The robustness of the obtained model was evaluated using the cross-validation method (Leave-One-Out), showing a strong correlation ($Q^2 = 0.83$) between experimental values and predicted values of biological activity A predictive and reliable QSAR model with excellent ability to fit internally (to the sample data) was achieved by the QSAR modeling with an R^2 of 0.89 and an R^2 from leave-one-out cross-validation (Q^2_{LOO}). This demonstrates the ability of the QSAR model to predict effectively, giving it considerable confidence. Thus, the predictive power was confirmed by external validation using a test data set and showed better predictive capability. The obtained validation results allow us to conclude that this model can be used to predict the biological activity of untested molecules, provided they are within the model's applicability domain. The comparison of the calculation results with the criteria of the Lipinski and Veber rules suggests that the 29 molecules exhibited favorable drug-likeness properties according to Lipinski's and Veber's criteria, suggesting their potential suitability for further pharmacological investigation since they comply with Lipinski's rule of 5 and also Veber's rule.

Author Contribution Statement

Conceptualization, M.R., F.A and K.ELM.; methodology, Z.S., A.ALS., Y.ELM; validation, J.M. and I.A.; investigation, N. Q. and A.ALS.; writing-original draft preparation, Y. El MM.R. and K.ELM.; writing-review and editing, J.M. and Y.ELM. All authors have read and agreed to the published version of the manuscript

Acknowledgements

I would like to express my deepest gratitude to my professor, Ms. Ilham Kandoussi, for her invaluable support, wise advice, and availability throughout the completion of this project. Her expertise and guidance were essential to its successful completion. Please state that the article is based on the work from this thesis, along with other improvements. Here is Zineb Soubai's email address: zineb.soubai@uit.ac.ma

Data Availability Statement

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to their use in current research works.

Conflict of interest

The authors declare no conflict of interest.

References

1. Hammond J, Lambert N, Voutsadakis IA. Clinical and Pathological Features and Survival Outcomes of Breast Cancers with Intermediate ER Expression. Cancers. 2025;17(13):2252. https://doi.org/10.3390/cancers17132252
2. Volkova YA, Kozlov AS, Kolokolova MK, Uvarov DY, Gorbatov SA, Andreeva OE, et al. Steroidal N-Sulfonylimidates: Synthesis and biological evaluation in breast cancer cells. Eur J Med Chem. 2019;179:694-706. https://doi.org/10.1016/j.ejmech.2019.06.048
3. Purohit A, Foster PA. Steroid sulfatase inhibitors for estrogen- and androgen-dependent cancers. J Endocrinol. 2012;212(2):99-110. https://doi.org/10.1530/JOE-11-0266
4. Srimathi R, Kathiravan MK. Lead optimization of 4-(thio)-chromenone 6-O-sulfamate analogs using QSAR, molecular docking and DFT - a combined approach as steroidal sulfatase inhibitors. J Recept Signal Transduction. 2021;41(2):123-137. https://doi.org/10.1080/10799893.2020.1794004
5. Nakata T, Takashima S, Shiotsu Y, Murakata C, Ishida H, Akinaga S, et al. Role of steroid sulfatase in local formation of estrogen in post-menopausal breast cancer patients. J Steroid Biochem Mol Biol. 2003;86(3-5):455-460. https://doi.org/10.1016/S0960-0760(03)00357-1
6. El Fouhi M, Benider A, Kagambega Zoewendbem AG, Mesfioui A. Profil épidémiologique et anatomopathologique du cancer de sein au CHU Ibn Rochd, Casablanca. Pan Afr Med J. 2020;37. https://doi.org/10.11604/pamj.2020.37.41.21336
7. World Health Organization. Cancer [Internet]. [cited 2026 Apr 6]. Available from: https://www.who.int/fr/news-room/fact-sheets/detail/cancer
8. El-Moustaqim K, Mabrouki J, Azrour M. Monitoring of Water Toxicity Through the Internet of Things to Protect

the Health of the Population. IoT, Machine Learning and Data Analytics for Smart Healthcare. 2023;85-92. https://doi.org/10.1201/9781003430735-7
9. Rossi L, Mazzara C, Pagani O. Diagnosis and Treatment of Breast Cancer in Young Women. Curr Treat Options Oncol. 2019;20(12):86. https://doi.org/10.1007/s11864-019-0685-7
10. Hassiotou F, Geddes D. Anatomy of the human mammary gland: Current status of knowledge. Clin Anat. 2013;26(1):29-48. https://doi.org/10.1002/ca.22165
11. Sbihi K, Elhamji S, Lghoul S, Aziz K, El Maallem A, Mabrouki J, et al. Biosorption of Hexavalent Chromium by Freshwater Microalgae Craticula subminuscula from Aqueous Solutions. Sustainability. 2024;16(2):918. https://doi.org/10.3390/su16020918
12. Azrour M, Mabrouki J, Farhaoui Y, Guezzaz A. Experimental Evaluation of Proposed Algorithm for Identifying Abnormal Messages in SIP Network. Intelligent Systems in Big Data, Semantic Web and Machine Learning. 2021;1344:1-10. https://doi.org/10.1007/978-3-030-72588-4_1
13. Ghamali M, Chtita S, Bouachrine M, Lakhlifi T. Méthodologie générale d'une étude RQSA/RQSP. (General methodology for an RQSA/RQSP study). Revue Interdisciplinaire. 2016;1(1).
14. Roy K, Kar S, Das RN. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. Academic Press; 2015. Available from: https://shop.elsevier.com/books/understanding-the-basics-of-qsar-for-applications-in-pharmaceutical-sciences-and-risk-assessment/roy/978-0-12-801505-6
15. Dépistage du cancer du sein. Anatomie du sein [Internet]. [cited 2026 Apr 6]. Available from: https://www.depistagesein.ca/anatomie-du-sein/
16. Macias H, Hinck L. Mammary gland development. Wiley Interdiscip Rev Dev Biol. 2012;1(4):533-57. https://doi.org/10.1002/wdev.35.
17. El Ouardi M, Drioiche A, El Makhoukhi F, Mabrouki J, Hakmi M, Al Kamaly O, et al. Chemical composition, antimicrobial, and antioxidant properties of essential oils from Artemisia herba-alba asso. and Artemisia huguetii caball. from Morocco: in vitro and in silico evaluation. Front Chem. 2024;12:1456684. https://doi.org/10.3389/fchem.2024.1456684
18. Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA. 2019;321(3):288. https://doi.org/10.1001/jama.2018.19323
19. Soubai Z. Exploration of STS inhibitors using a QSAR approach for the treatment of breast cancer [thesis]. Rabat, Morocco: Faculty of Medicine and Pharmacy in Rabat; 2022.
20. Fleurot E. Mise en évidence d'un dialogue entre la signalisation oestrogénique et le syndécane-1 dans les cellules de carcinome mammaire humain MCF7 [Evidence of crosstalk between estrogen signaling and syndecan-1 in human MCF7 breast carcinoma cells] [doctoral dissertation]. Normandie Université; 2017.
21. Hall JM, Couse JF, Korach KS. The Multifaceted Mechanisms of Estradiol and Estrogen Receptor Signaling. J Biolo Chem. 2001;276(40):36869-36872. https://doi.org/10.1074/jbc.R100029200
22. Delozier T. Hormonothérapie du cancer du sein. J Gynecol Obstet Biol Reprod. 2010;39(8):F71-F78. https://doi.org/10.1016/j.jgyn.2010.10.004
23. Desta Z, Nguyen A, Flockhart D, Skaar T, Fletcher R, Weinshilboum R, et al. Antiestrogen pathway (aromatase inhibitor). Pharmacogenetics Genom. 2009;19(7):554-555. https://doi.org/10.1097/FPC.0b013e32832e0ec1
24. Mujumdar P, Poulsen S. Natural Product Primary Sulfonamides and Primary Sulfamates. J Nat Prod. 2015;78(6):1470-1477. https://doi.org/10.1021/np501015m
25. Bencheikh I, Azoulay K, Ben Baaziz M, Mabrouki J. Ecological Risks Related to the Influence of Different Environmental Parameters on the Microplastics Behavior. Adv Energy Sci Technol. 2023;117-128. https://doi.org/10.1007/978-3-031-25662-2_10
26. Winum J, Scozzafava A, Montero J, Supuran CT. Sulfamates and their therapeutic potential. Med Res Rev. 2005;25(2):186-228. https://doi.org/10.1002/med.20021
27. Tiano JP, Mauvais-Jarvis F. Importance of oestrogen receptors to preserve functional β-cell mass in diabetes. Nat Rev Endocrinol. 2012;8(6):342-351. https://doi.org/10.1038/nrendo.2011.242
28. Ferlay J EM, Lam F, Laversanne M, Colombet M, Mery L et al. Global Cancer Observatory: Cancer Today [site Web]. International Agency for Research on Cancer; 2026.
29. Fouad A, Yousra A, Kaoutar Z, Omar EM, Afaf A, Sanae B. Classification moléculaire du cancer du sein au Maroc. Pan Afr Med J. 2012;13(91). https://doi.org/10.11604/pamj.2012.13.91.1392
30. Bettaieb A, Paul C, Plenchette S, Shan J, Chouchane L, Ghiringhelli F. Precision medicine in breast cancer: reality or utopia?. J Transl Med. 2017;15(1):139. https://doi.org/10.1186/s12967-017-1239-z
31. Kołaczek A, Fusiarz I, Ławecka J, Branowska D. Biological activity and synthesis of sulfonamide derivatives: a brief review. Chemik. 2014;68(7):620-8.
32. Foenix-Riou B. Médicaments.gouv.fr: la base de données publique des médicaments [Médicaments.gouv.fr: the public database of medicinal products]. Bases. 2013;(308):7.
33. Mabrouki J, Benchrifa M, Azoulay K, Bencheikh I, Azrour M, Raji H. Deep Learning-Based Tracking System for Detecting Toxic Elements in Water to Protect Public Health. Sustainable and Green Technologies for Water and Environmental Management. 2024;267-277. https://doi.org/10.1007/978-3-031-52419-6_19
34. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inform. 2010;29(6-7):476-488. https://doi.org/10.1002/minf.201000061
35. 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
36. Zhang L, Zhu H, Oprea TI, Golbraikh A, Tropsha A. QSAR Modeling of the Blood–Brain Barrier Permeability for Diverse Organic Compounds. Pharm Res. 2008;25(8):1902-1914. https://doi.org/10.1007/s11095-008-9609-0
37. El-Moustaqim K, Tallou A, Mabrouki J, Slaoui M, Hmouni D. Phytoremediation processes for the removal of heavy metals. Integrated Solutions for Smart and Sustainable Environmental Conservation. 2024:33-46.https://doi.org/10.1007/978-3-031-55787-3_3
38. Sreedharan R, Prajapati J, Engineer P, Prajapati D. Leave-one-out cross-validation in machine learning. InEthical issues in AI for bioinformatics and chemoinformatics 2023 Nov 14 (pp. 56-71). CRC Press.
39. Kar S, Roy K, Leszczynski J. Applicability Domain: A Step Toward Confident Predictions and Decidability for QSAR Modeling. Comput Toxicol. 2018;1800:141-169. https://doi.org/10.1007/978-1-4939-7899-1_6

This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.