Scientific Research Publishing

Computational Chemistry, 2025, 13(3), 45-68
https://www.scirp.org/journal/cc
ISSN Online: 2332-5984
ISSN Print: 2332-5968

QSAR Models: Exploring Limits in Three Cases

El Hadji Sawaliho Bamba

Constitution and Reaction of Matter Laboratory, Training and Research Unit in Structural, Material and Technological Sciences, Felix Houphouet-Boigny University, Abidjan, Côte d'Ivoire
Email: bambaelhadjisawaliho@yahoo.ca

How to cite this paper: Bamba, E.H.S. (2025) QSAR Models: Exploring Limits in Three Cases. *Computational Chemistry*, **13**, 45-68.
https://doi.org/10.4236/cc.2025.133003

Received: May 9, 2025
Accepted: June 22, 2025
Published: June 25, 2025

Abstract

This article critically assessed the validity of five multiple linear regression models across three separate studies. The first examined the cytotoxic properties of N-tosyl-1,2,3,4-tetrahydroisoquinoline compounds. The second evaluated the antiproliferative effects of 1,3,5-arylidene rhodanines. The last explored the antitumour potential of thiazoline or thiazine derivatives. Despite limited sample sizes, the model validation showed robust performance and predictive capabilities. However, their forecasts lacked accuracy. The authors validated their models by assessing the fit training data and generalization ability. The gaps weren't clearly defined, and outliers were only partially considered. The cytotoxicity study of N-Tosyl-1,2,3,4-Tetrahydroisoquinoline used a ±2 standardized residual. Non-random sampling can introduce selection bias. Ignoring dispersion and employing fixed molecules can reduce model accuracy. Adhering to MLR premises aids in validation. Analysis of secondary data from three articles showed that all five MLR models were invalid, emphasizing the need to verify MLR assumptions before utilizing the QSAR approach.

Keywords

Multiple Linear Regression, QSAR, Tetrahydroisoquinoline, Arylidene Rhodanine, Thiazoline, Thiazine

1. Introduction

Research in chemistry, pharmaceuticals, and therapeutics renewed interest in Quantitative Structure-Activity Relationship (QSAR) modelling, which predicts biological activity of molecules based on their structure. Multiple Linear Regression (MLR) is used to link these two parts in this approach. The Dependent Variable (DV) is the amount needed for a response, while Independent Variables (IV) are the compounds' properties. QSAR's assumption was that a similar configura-

tion has analogous activities, making it a powerful tool across scientific fields. Models were employed in drug discovery to project the effectiveness and safety of new molecules. Their performances were assessed using the coefficient of determination (R^2), which indicated how well they fit the training data. Prediction accuracy was measured by Root Mean Square Error of the Training (RMSTr), set and by Root Means Square Error of Testing set. The correlation coefficient of Cross-Validation Q^2_{CV} evaluated the predictive power of Model. This work selected three articles to assess the relevance of the standards used in several QSAR-based research studies. Furthermore, these papers proposed models aimed at projecting the activities of molecular families against serious diseases such as cancer. They offered the opportunity to judge the conformity of their models to the premises underlying a valid MLR. One key factor behind this decision was the small sample size, which may have resulted in the breach of certain MLR assumptions [1]. The standards affecting the corroboration of the three studies' models were investigated.

Case 1 analyzed the cytotoxicity of 12 N-Tosyl-1,2,3,4-Tetrahydroisoquinoline molecules [2]. The models explained 79.19% of the variance for DV MOLT3 (Model 1) and 98.49% for DV HepG2 (Model 2). Model 1 demonstrated a predictive power of 56.9%, with forecasts' accuracy of 22.3% in training and 32.5% in testing. Model 2 predicted with 91.5% power. Its precision was 2.8% in training and 7.3% in testing. Model 1 exhibited high performance with low prediction accuracy and reliability. Model 2 showed excellent performance and prediction power, but its precision remains low. These statistics suggest that a model's performance and projection capabilities (precision and power) can be contradictory. They noted that using a high R^2 to validate a model has limitations. The small sample sizes (N = 12 for DV MOLT3, N = 7 for DV HepG2) may reduce models' predictive power [1] [3] [4]. Model 1's low predictive capacity is evident. Model 2's high value is questionable due to its smaller sample size (N = 7 vs. N = 12).

Case 2 evaluated the antiproliferative activity of 13 1,3,5-Arylidene Rhodanine compounds [5]. The performance was 92.7% for Model 3 (ActMDA) and 88.2% for Model 4 (ActNCI). The predictive power of Model 3 was 95.4%, while that of Model 4 was 92.6%, demonstrating their overall effectiveness. However, their sample sizes (N = 13) hardly explained this level of the two models' external validity. The absence of precise forecast data hindered the evaluation of this key component in both models.

Case 3 aimed to predict pulmonary effects of Thiazine or Thiazoline on A-549 cells [6]. Model 5's performance reached 90.5%, with a prediction accuracy of 10.6% for the test sample. Its predictive power was also 90.5%. It was consistent with the performance. Nevertheless, the low reliability of its predictions didn't support it. Furthermore, the small sample size (N = 14) is difficult to reconcile with predictive power or high generalizability of the results [1] [3] [4].

The analyzed cases showed that their models' performances were elevated, effectively accounting for the observed variances. Their predictive powers were high

(except for Model 1). Conversely, they weren't precise. Reference [2] [5] [6] employed the criterion $R^2 - Q^2_{CV} < 0.3$ to validate their models. This criterion included two parameters: R^2 evaluating their fit to the training data and Q^2_{CV} assessing their ability to generalize to new molecules. The interpretation of the gap was unclear. Another weakness was the partial treatment of outliers [1].

Reference [7] noted that Pearson's R^2 estimate was affected by outliers, resulting in overestimation, particularly in small samples. The author [4] recommended removing them before conducting MLR analysis. In case 1, reference [2] used a standardized residual of ±2. Outliers were retained within this interval, impacting the model's validation parameters. The reviewed articles employed MLR training and testing on a single data distribution without considering sampling dispersion. According to [8], Leave-One-Out Cross-Validation (LOO-CV) may introduce a bias selection and reduce model variance in small, unrepresentative samples, using nearly all data in each iteration. These samples can negatively affect its accuracy and performance. Methodological issues can produce unreliable models. Validating their adherence to MLR premises can statistically support them [1] [4] [9] [10]. This article analyzed five models from three studies utilizing their secondary data to address the following question:

How valid were models generated using the QSAR approach?

The research proposed that these models may not meet the premises of MLR due to the small sample sizes, which could affect their validity. This paper aimed to examine the violations or compliance with these prerequisites. It also targeted the consideration of the pros and cons of adopting them. The article analyzed assumptions and methods, assessing model conformity. Internal validity was evaluated by R^2 performance and RMSTr accuracy. External one shows predictive power in forecasting new molecular activities. A model is statistically valid if it significantly enhances these two indicators. This paper comprises five sections: premises of an MLR, materials and methods, results and discussion, and conclusion.

2. Premises of Multiple Linear Regression

The models are required to meet the MLR guidelines, which encompass the relationships between DV and IVs. The quality of the database is analyzed as referenced by [1] [4] [7]. Cook's Distance was employed to identify outliers. Its value has to be above 0.5 [1] [4]. Given the small sample sizes (N) of the five models, a threshold of 0.5 is considered more suitable than 4/N, as the latter results in higher thresholds. Both thresholds are conventional. It's suggested to assess data homogeneity by using the Coefficient of Variation (CV). This latter is the ratio of the standard deviation to the mean. As noted by [1] and [7], data are classified as uniform if the CV is less than 15%. According to [1], it's necessary to establish the continuity of DV and IV, ensuring that they're quantitative and not subject to any constraints. The independence of the DV is demonstrated by considering that data came from different molecules. The normality of its distribution is verified by em-

ploying the Kolmogorov-Smirnov (KS) test, as documented by [1] [4] [9]. IV variances are checked to guarantee they aren't zero. The overall linearity must be examined.

Reference [10] suggests using the Loess line and scatter plots of residuals RES1 and RES2 to verify linear relationships between a DV and its IVs. RES1 represents model error term, while RES2 is derived from another, treating the predictor as a DV. The Loess curve is included in the graph. Linearity is confirmed when it aligns closely with the regression line. The absence of multicollinearity between the IVs must be demonstrated. The Variance Inflation Factor (VIF) makes it possible [9]. According to [1] [9] [11], it must be less than 10. The sample size, N, is compared to its theoretical value using inequalities 1 and 2 as described by [3] and [4] respectively.

$$\mathbf{N} \geq 50 + 8\ \mathbf{vi} \tag{1}$$

$$\mathbf{N} \geq 104 + \mathbf{vi} \tag{2}$$

vi represents the number of IVs.

Three key assumptions about residuals need to be demonstrated.

- Homoscedasticity is checked using the Breusch-Pagan test, which is required to be significant [9] [12].
- Errors are normally distributed around a zero mean, verified by employing a histogram superimposed on a Gaussian curve and a normal P-P plot [1] [4] [9].
- The independence of residuals is confirmed with the Durbin-Watson statistic. A value close to 2 indicates no autocorrelation [1] [9]. Furthermore, the materials and methods were also described.

3. Materials and Method

The research examined three articles that forecasted molecule activity in combating serious diseases, treating each as an individual case.

3.1. Data Sampling

The procedures for analyzing the data were detailed comprehensively. Reference [2] aimed to predict outcomes using Model 1 with 12 molecules for MOLT3 cell lines and seven for HepG2 ones.

Case 1

The activity (IC50) values of N-Tosyl-1,2,3,4-Tetrahydroisoquinoline derivatives were measured. They utilized LOO-CV to create training and test sets, identifying outliers with standardized residual cut-offs of ±2. The modelling was performed with Weka [12]. Descriptors were computed using Gaussian 03 DFT/B3LYP/6-31(d) and Dragon, with redundant indicators removed by the Unsupervised Forward Selection algorithm. Key predictors were pinpointed employing stepwise SPSS analysis, and the main findings were summarized. Two models were involved in Case 1. The expression for Model 1 was

$$MOLT3 = 2.01312*Mor32u + 64.533*Gu - 12.2097 \quad (3)$$

Model 2's wording was

$$HepG\ 2 = -892.215*PJI3 + 1.1322*Mor32u - 1.048\ 3*Mor31v + 6.6454 \quad (4)$$

Gu referred to the symmetric index, Mor32u, Mor31v, and Mor32u were 3D-MoRSE indexes, and PJI3 was the Petitjean 3D index. Case 2 aimed to predict the antiproliferative activity of 13 5-Arylidene Rhodamines and compare descriptors.

Case 2

Nine molecules were modelled for inhibiting human lung tumours (NCI-H727) and ductal carcinoma (MDA-MB-231), with four used for testing. MLR was performed with Excel and Gaussian 09. It generated IVs at the DFT/B3LYP/6-31 G(d) level. Model 3 for the MDA-MB-231 cell line had the expression of:

$$ActMDA\ (PCI50) = -459.091\ 76 + 1.892\ 18*ELUMO + 0.08176*\nu_{C=O} + 39.433\ 17*d\_C\text{-}N \quad (5)$$

Model 4 for the NCI-H727 cell line was documented as follows:

$$ActNCI\ (PCI50) = -612.154\ 55 + 2.145\ 18*ELUMO + 0.11987*\nu_{C=O} + 30.62007*d\_C\text{-}N \quad (6)$$

ELUMO represented the lowest unoccupied molecular orbital energy, $\nu_{C=O}$ indicated the CO frequency of the five-membered ring, and d_C-N referred to its CN distance. The third study analyzed 14 compounds related to Thiazoline and Thiazine for their antitumour properties.

Case 3

The case 3 predicted their pulmonary effects on A-549 cells. Model 5 used 10 molecules for training and four for testing with Gaussian 09 at DFT/B3LYP/6-31+G(d, p) levels. MLR in Excel calculated IVs from their expressions. Model 5 from this case was described as follows:

$$PIC50 = 2.26432 - 0.77981*mu + 0.44572*LogP \quad (7)$$

LogP measured lipophilicity, and mu represented the molecular dipole moment.

3.2. Data Analysis

The research undertaken by [2] [5] [6] supplied secondary data concerning both independent and dependent variables. These are organized in the Appendix across **Table A1**, **Table A2**, **Table A3** and **Table A4**. They were employed to verify that Models 1 to 5 adhered to MLR premises. Their compliance with this latter was conducted using SPSS Statistics version 27.

Cook's distance identified outliers utilizing Analysis > Regression > Linear > Save, then selecting "Cooks" in the "Distance" box. Calculate CV with Descriptive Analyze > Descriptives to find the mean and standard deviation. The KS test assessed normality of DV via Analyze > Nonparametric > Univariate tests > KS test, with the null hypothesis assuming its normal distribution. Multicollinearity was checked by employing VIF: Analyze > Regression > Linear > Statistics > Collinearity Diagnostics. For linearity, a scatter plot of predicted values (x-axis) and residuals (y-axis) was generated: Analysis > Regression > Linear > Graphs, plotting

ZRESID vs. ZPRED. In graph editor, select Elements, Fit Curve to Total, Loess; a curve oscillating around Y = 0 indicates linearity. The sample size was verified by comparing the molecules' quantity exploited to the estimates in Equations (1) [3] and (2) [4]. Homoscedasticity of error term was examined as per [12]. The process included standard regression analysis, saving non-standardized predicted values and residuals, and performing the test. Unstandardized residuals were squared and utilized in a subsequent regression model. Achieve another one with RES_squared as the DV employing the same IVs. Realize the Breusch-Pagan test via: Analyze > Regression > Linear > Save, ensuring that Unstandardized for Predicted Values and Residuals boxes are checked, which will generate PRE1 and RES1 columns in Data View. The residuals were squared via the Transform menu: Compute Variable, naming RES_squared and entering RES1*RES1 as the formula. The p-value from the Breusch-Pagan test in the ANOVA table indicates heteroscedasticity if below 0.05. The normal distribution of residuals is assessed with a P-P graph using linear regression. The author [9] suggests configuring statistics by selecting: "Estimation," "Hypothesis test," and "Residual predictions." After choosing "Histograms" and "Normal P-P Chart" in the chart menu, the PP plot is generated. Points near the diagonal suggest a normal distribution of residuals. The Durbin-Watson D. identifies error term independence [9]. References [1] and [4] describe the standard regression procedure: select the "Statistics" option and check the "Durbin-Watson" box under "Residuals." Durbin-Watson D. close to 2 indicates independent residuals, next to 0 shows positive autocorrelation, and around 4 proposed a negative one [9]. MLR premise conformity analyses were performed for all models. The findings are presented and discussed.

4. Results and Discussion

Models 1 and 2 related to N-Tosyl-1,2,3,4-Tetrahydroisoquinoline were evaluated. According to [1] [4] [7], DVs and IVs data must be free of outliers.

4.1. Compliance with Data on N-Tosyl-Tetrahydroisoquinoline Compounds

In Model 1, molecule 8 (4h) had a Cook's distance of 0.507 [1]. In Model 2, compound 4 (4h) had a statistic of 4.372, above 0.5, while molecule 6 (4k) had 0.812. These data were outliers [1]. They were removed before linear regression analysis [4]. Those for models 1 and 2 were heterogeneous, with CVs over 15% [7]. For Model 1, DV MOLT3 was 39%, IV Mor32u was 41%. In Model 2, DV HepG2 was 16%, IV Mor32u was 59%, and IV Mor31v was 19%.

The DVs and IVs in models 1 and 2 were continuous; they were quantitative with no constraints [1] [4]. DV data were normally distributed (KS test: p = 0.730). MOLT3 and HepG2 activities originated from different molecules, ensuring their independence from each other. The non-zero variance premises were violated. Furthermore, Model 1 had VIFs of 1.028. Those of Model 2 were worth 1.899 for IV Mor32u, 1.872 for Mor31v and 1.030 for PIJ3. All values were below 10, indi-

cating an absence of multicollinearity between the IVs of the two models [1] [11]. Figure 1 indicates that Model 1 has deviated from the y=0 line, thereby compromising its linearity [9] [13]. Similarly, Figure 2, associated with Model 2, depicts a comparable scenario.

The models used small samples. Model 1 employed 11 molecules, less than the 66 and 106 obtained with Equations (1) [3] and (2) [4]. Likewise, Model 2 exploited compounds instead of the needed 74 and 107. The Breusch-Pagan test revealed heteroscedasticity in Model 1 ($F = 0.117$, $p = 0.891$) and Model 2 ($F = 0.545$, $p = 0.665$) [9] [12].

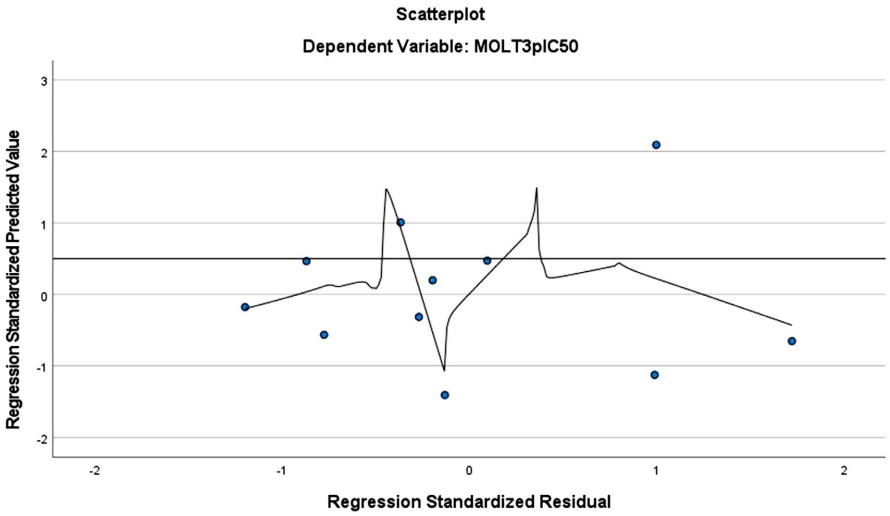


Figure 1. Model 1's linear relationship analysis.

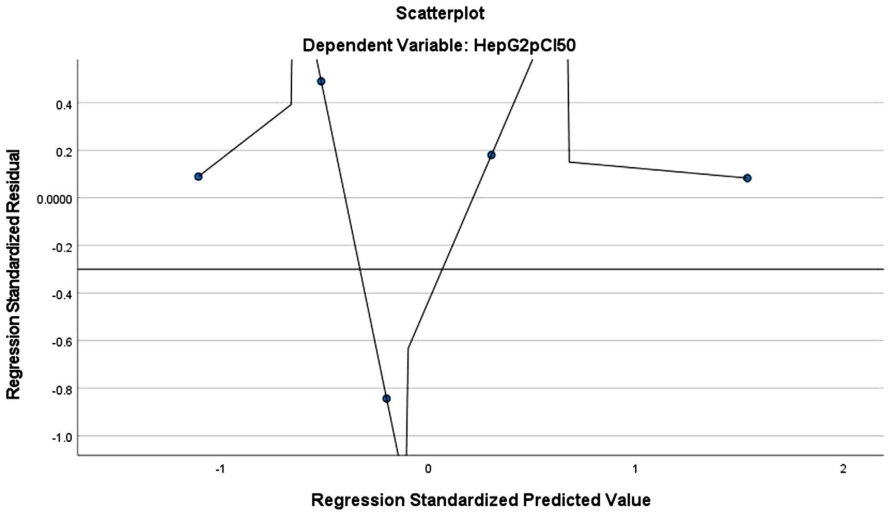


Figure 2. Model 2's linear relationship analysis.

The points in the normal P-P plot of Model 1 didn't align along the diagonal, as shown in Figure 3. Similarly, those in Model 2, illustrated in Figure 4, didn't align either. In both cases, the residuals weren't normally distributed around zero

[9]. Model 1 produced a Durbin-Watson statistic of 2.668, while Model 2 generated 2.736. These values suggest that errors in both models are independent [8]. A summary of the results can be found in Table 1. The models respected the same premises.

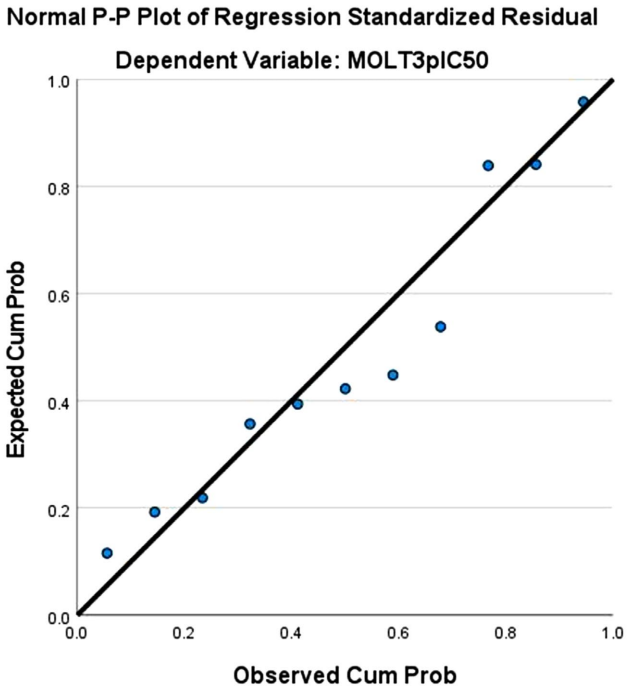


Figure 3. Assessment of Model 1 residuals for normal distribution.

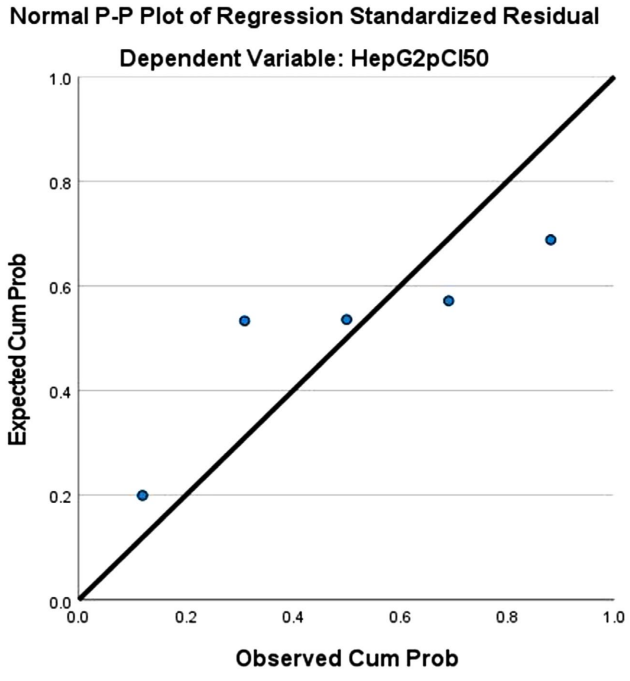


Figure 4. Assessment of Model 2 residuals for normal distribution.

Table 1. Analysis results: compliance summary for Models 1 and 2.

	Model 1		Model 2	
	Violated Premises	Respected Premises	Violated Premises	Respected Premises
Data homogeneity	Data homogeneity		Data homogeneity	
DV		DV continuity DV normality DV independence		DV continuity DV normality DV independence
IV	Non-zero variance	IV continuity Absence of multicollinearity	Non-zero variance	IV continuity Absence of multicollinearity
Relationship between DV and its IVs	Overall linearity		Overall linearity	
Sample	Sample size		Sample size	
Residues' characteristics	Residues' homoscedasticity Normality of residual distribution	Residues' Independence	Residues' homoscedasticity Normality of residual distribution	Residues' independence

The two models followed the assumptions of continuity, normal distribution, and independence concerning the DVs. They adhered to the continuity of the IVs and avoided multicollinearity. They violated basic hypotheses related to non-zero variances of IVs, overall linearity of the model, and adequate sample size. They also transgressed premises regarding homoscedasticity and independence of the error term. The effectiveness of Models 1 and 2 is determined by their capacity to accurately predict the cytotoxic activity of N-Tosyl-Tetrahydroisoquinoline molecules.

Respect for the premises underlines this possibility. Without multicollinearity, the relationship between the DV and its IVs can be pinpointed. This precisely identifies the β_X coefficients and accounts for variations in the IVs. In concert with [1] [4], the DVs MOLT3 and HepG2 can be exactly predicted due to their continuity, normal distribution, and independence. Independent residuals ensure accurate confidence intervals [9] [14] and parameter estimates for both models [9]. According to [9] [14], this improves the reliability of their outcomes and forecasts. Evaluating confidence intervals correctly avoids biases in estimations and predictions. Deviation from premises reduces the projection precision of the two models.

The nonlinearity impacts β_X estimates, making it difficult to understand the relationship between HepG2 and its predictors PIJ3, Mor32u, and Mor31v, leading to unreliable projections [9] [15]. A small sample size may result in incorrect R^2 and β_X coefficients [1] [4] [15] [16]. It raises the variance of estimators, complicating the detection of IV effects [1] [3]. According to [9], heteroscedasticity affects hypothesis testing and confidence intervals by misestimating variances, resulting in incorrect β_X coefficients. Non-normally distributed residuals increase

the risk of type I errors, leading to inaccuracies in model predictions [1] [9].

Violating global linearity, homoscedasticity, and normal distribution of residuals reduces the precision of projections from Models 1 and 2. Respecting multicollinearity doesn't sufficiently compensate for these issues. The quality of each model is also evaluated based on their ability to generalize results.

The small sample size increases sensitivity to fluctuations and lower external validity, making it harder to generalize predictions to additional N-Tosyl-Tetrahydroisoquinoline molecules [1] [3] [4]. The residuals 'heteroscedasticity can lead to unstable MOLT3 and HepG2 predictions, causing incorrect conclusions about IVs variations [9]. However, their independence ensures that each model is robust and applicable to new derivatives from N-Tosyl-Tetrahydroisoquinoline compounds [17]. Generalizing the outcomes of both models is difficult despite the contribution of their quality. Models 1 and 2 don't meet the premises, allowing the activity of the molecules studied to be accurately predicted and generalized to others in the same family. Consequently, their validity remains uncertain. Additionally, Models 3 and 4 related to 5-Arylidene Rhodanine compounds were analyzed thoroughly.

4.2. Compliance with Data on 5-Arylidene Rhodanine Compounds

In Model 3, the DV ActMDA score for molecule 4 was an outlier with a Cook's distance of 0.662, exceeding 0.5. It was removed [4]. Cook distances in Model 4 ranged from 0.007 to 0.236, all below 0.5. No outlier was found in DV ActNCI [1]. Furthermore, IVs data were homogeneous for each model [7]. The CV of IV ELUMO was 8%, and those of IV d_CN and νC=O were identical at 0.2%. The DVs data were heterogeneous [7], with CVs of 57.6% for DV ActNCI and 58.8% for DV ActMDA.

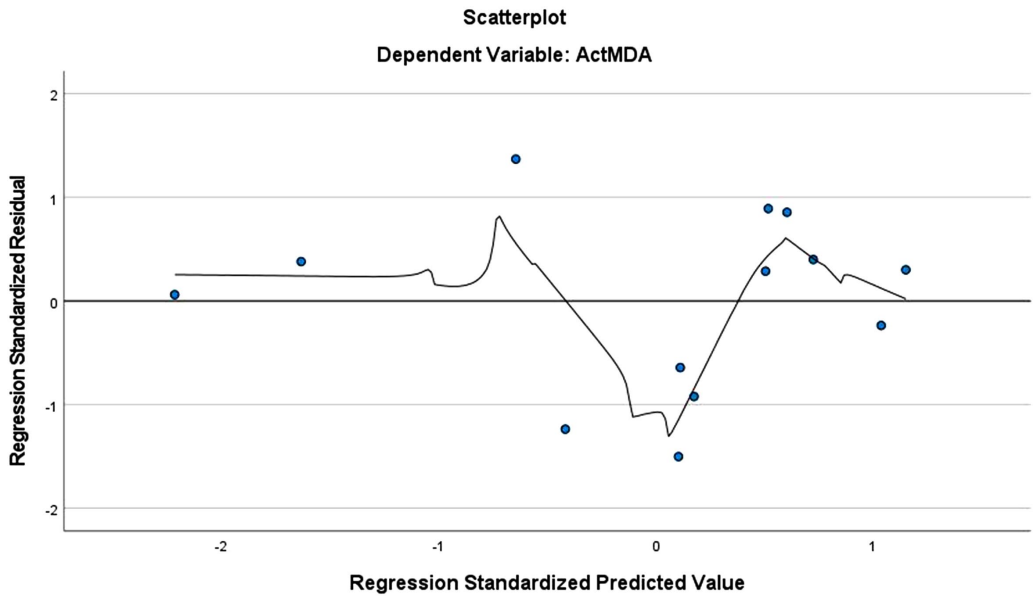


Figure 5. Model 3's linear relationship analysis.

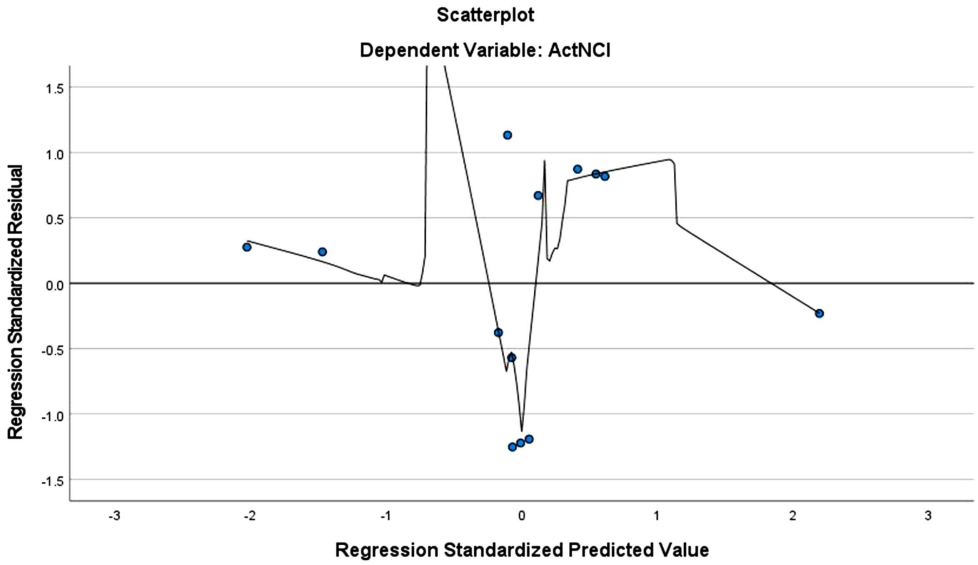


Figure 6. Model 4's linear relationship analysis.

The DVs data of both models exhibited a normal distribution (p = 0.200) [1] [4]. They were continuous as they were quantitative and didn't face any limitations, such as the IVs ELUMO, $\nu_{C=O}$ and d_CN [4]. The DVs data originated from different 5-Arylidene Rhodanine compounds that were independent. In Models 3 and 4, ELUMO and $\nu_{C=O}$ displayed non-zero variances, while d_CN presented zero variance. Model 3 indicated VIFs of 1.134, and Model 4 had VIFs of 1.077, 1.659, and 1.63, all below 10, suggesting no multicollinearity [1] [11]. Figure 5 shows Model 3's Loess line deviating significantly from zero. Figure 6 depicts Model 4's deviation. The condition for overall linearity wasn't met [13]. The sample sizes for the two models were insufficient, with 13 compounds instead of the required 74 or 107 as outlined by Equations (1) [3] and (2) [4].

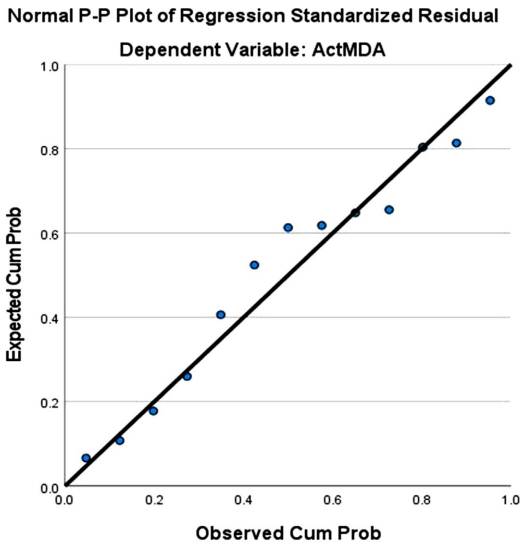


Figure 7. Assessment of Model 3 residuals for normal distribution.

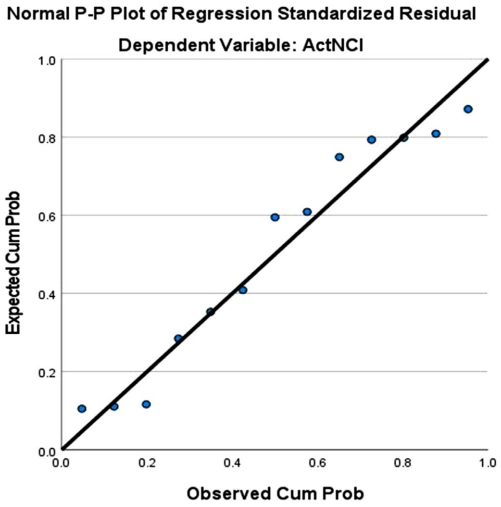


Figure 8. Assessment of Model 4 residuals for normal distribution.

The Breusch-Pagan distribution tests showed no significant results, with a Fisher coefficient of F = 1.049 (p = 0.418) for Model 3 and F = 0.284 (p = 0.836) for Model 4. Neither model met the assumption of residuals' homoscedasticity [4] [12]. Figure 7, related to Model 3, demonstrates the non-alignment of the points on the P-P graph with the diagonal. Similarly, Figure 8, associated with Model 4, shows the same pattern. Consequently, under these conditions, the errors for DV ActMDA and ActNCI weren't normally distributed [4] [9]. According to [1] [9], the Durban-Watson D. suggested positive autocorrelation (ActMDA: 1.184; ActNCI: 0.997). The residues for models 3 and 4 weren't independent. A summary of the results can be found in Table 2.

Table 2. Analysis results: compliance summary for Models 3 and 4.

	Model 3		Model 4	
	Violated Premises	Respected Premises	Violated Premises	Respected Premises
Data homogeneity		Data Homogeneity		Data Homogeneity
DV		DV continuity DV normality DV independence		DV ActMDA continuity DV ActMDA normality DV ActMDA independence
IV	Non-zero variance	IV continuity Absence of Multicollinearity	Non-zero variance	IV continuity Absence of Multicollinearity
Linear Relationship	Overall linearity		Overall linearity	
Sample	Sample size		Sample size	
Residues' characteristics	Residues' homoscedasticity Normality of Residuals Distribution Residues' independence		Residues' homoscedasticity Normality of Residuals Distribution Residues independence	

Models 3 and 4 met data homogeneity criteria, ensuring continuity and avoiding multicollinearity of the IVs. However, they fall to respect non-zero variances, linearity, adequate sample size, homoscedasticity, normal distribution, or independence of residual requirements. Their effectiveness is linked to their ability to accurately predict the antiproliferative activity of 5-Arylidene Rhodanine molecules. Compliance with certain assumptions underlying MLR predisposes them.

As noted by references [1] and [4], the normal distribution and independence of DV data in both models enable precise estimation of molecular activities, enhancing consistency between dependent and independent variables. Continuity in IVs ELUMO, $\nu_{C=O}$, and d_CN achieves similar outcomes. The absence of multicollinearity simplifies models 3 and 4 [18]. However, predicting the antiproliferative effects of 5-Arylidene Rhodanines remains challenging due to certain premise violations.

The non-zero variance of IVs (ELUMO, freq. CO, and d_CN) limits models 3 and 4 in assessing the accuracy of relationships between variables. Moreover, the absence of linearity can cause unclear predictions for DVs ActMDA and ActNCI [9]. As highlighted in references [9] [15], it introduces bias into β_X coefficients. This lack of precision doesn't significantly improve the accuracy of forecasts. Consequently, parameter estimates become incorrect, resulting in unreliable values for IV ELUMO, $\nu_{C=O}$, and d_CN, which increases the risk of type II errors [1]. Thus, the determination of R^2 and β_X remains imprecise [15]. Heteroscedasticity reduces the variance of ActMDA and ActNCI, impacting hypothesis tests and confidence intervals similarly to β_X [9]. It may lead to false conclusions regarding the effects of IVs ELUMO, $\nu_{C=O}$, and d_CN [9]. According to [9] [14], error dependence results in incorrect parameter estimates and affects significance tests and confidence intervals, indicating unincorporated structures in both models. Furthermore, non-normally distributed residuals elevate the likelihood of type I errors, resulting in model prediction inaccuracies [1] [9].

Models 3 and 4's prediction accuracy is reduced due to violations of linearity, homoscedasticity, normal distribution, and independence of residuals. Additionally, their effectiveness also depends on the generalizability of the results.

The small sample size prevents extrapolating the results of the two models to another 5-Arylidene Rhodanine molecule [3]. For [9], Heteroscedasticity may introduce instability in the effects of the IVs ELUMO, $\nu_{C=O}$, and d_CN. Models 3 and 4 can't measure variations in VI. Moreover, the error term dependence can decrease performance and restrict the generalizability of results to new molecules from 5-Arylidene Rhodanine [9] [17]. Additionally, these two models don't accurately predict their antiproliferative activity. Hence, it's advised that they be re-evaluated. The study further examined Model 5 in relation to Thiazine and Thiazoline compounds.

4.3. Compliance with Data Premises on Thiazoline and Thiazine Compounds

The cook's distance for outliers ranged from 0.003 to 0.346, all below 0.5. Model

5 had none [1]. IV mu data were uniform (CV 3%), while those of DV PIC50 and IV logP weren't (CVs 23% and 30%) [7]. DV PIC50 and its IVs were continuous [4]; their data were quantitative and unrestricted. The KS test confirmed the DVPIC50's normal distribution (p = 0.328) [1] [4]. Data from distinct Thiazoline and Thiazine compounds were independent. Variances were 0.013 for IV mu and 0.450 for IV LogP, meeting non-nullity assumption. VIFs were 1.072, indicating no multicollinearity [1] [11]. According to [9] [13], the Loess curve approached zero. **Figure 9** illustrates that the model exhibited overall linearity [9] [13]. With only 14 compounds studied, the sample size was insufficient as Equations (1) [3] and (2) [4] required 66 or 106 molecules. The Breush-Pagan test wasn't significant (F = 1.546; p = 0.256), confirming the heteroscedasticity of this model [12]. The normal P-P plot in **Figure 10** shows the Model 5 error distribution, but the points didn't align with the diagonal.

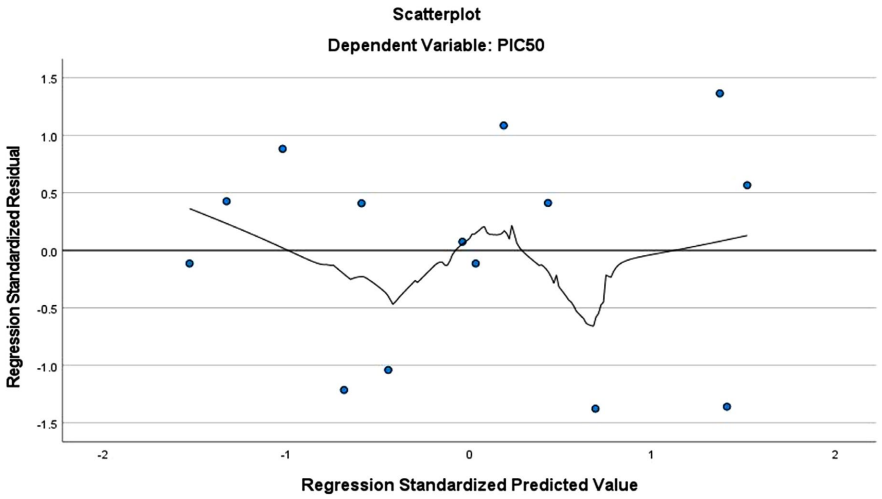


Figure 9. Model 5's linear relationship analysis.

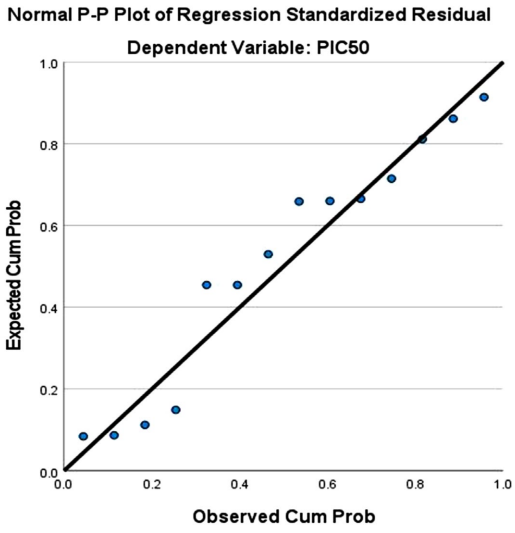


Figure 10. Assessment of Model 5 residuals for normal distribution.

Model 5 residues weren't normally distributed around zero [4] [9], although they were independent [1] [9]; Durbin D. value was 1.916, indicating no autocorrelation [9]. The primary findings of this analysis are presented in **Table 3**.

Model 5 met the DV PCI50 criteria for continuity, independence, and normality. It adheres to IVs continuity without multicollinearity issues. Although linearity and normally distributed error terms were followed, data homogeneity, sample size, and residuals' homoscedasticity were inadequate. Model 5's prediction accuracy for molecules' antitumour activity depends on specific premises.

The continuity of the DV PIC50 and its IVs mu and LogP, along with non-zero variances and the absence of multicollinearity, facilitate a clearer understanding of the structural relationships between them [1] [4] [18]. Non-zero variances and no multicollinearity ensure small confidence intervals and reliable significance tests [1] [18]. According to [10], Model 5's linearity allows for appropriate correlations between variables and valid confidence intervals, which enhances forecast accuracy. Independent residuals result in precise parameter estimates [9] and confidence intervals [14]. This improves the reliability of its outcomes and forecasts. Correctly evaluated confidence intervals avoid biases in parameter estimation and predictions [1]. Nonetheless, violating several premises can restrict the model's precision.

Table 3. Analysis results: compliance summary for Model 5.

	Violated Premises	Respected Premises
Data homogeneity	Data homogeneity	
DV characteristics		DV PIC50 continuity DV PIC50 normality DV PIC50 independence
IV characteristics		IV mu and LogP continuity Absence of multicollinearity Non-zero variance
Linear Relationship		Overall linearity
Sample	Sample size	
Residues' characteristics	Residues' homoscedasticity	Normality of residual distribution Residues' independence

The limited sample size compromises the reliability of parameter estimators, rendering them inadequate for accurately representing Thiazoline or Thiazine activity [19]. According to [9], error term's heteroscedasticity affects the correct understanding of the effects of IVs LogP and mu. It can misestimate the variance of DV PIC50. This results in less precise βx coefficients.

The non-normality of the residual distribution impacts reliability and invalidates confidence intervals and hypothesis tests [1] [4] [9]. Model 5's validation evaluates outcome generalizability, but the insignificant sample size may influence reliability. Changes in DV PIC50 and IVs mu and LogP increase sensitivity, af-

fecting adaptability and predictive accuracy [3]. Results vary by Thiazoline and Thiazine studied. On the other hand, the independence of residuals makes Model 5 robust and applicable to different datasets [17]. The small sample size remains a significant concern, which hinders the generalization of the findings of Model 5 to additional molecules from Thiazine or Thiazolines [3] [4]. This limitation invalidates it. Model 5 excels in forecasting the pulmonary effects of Thiazine or Thiazoline on A-549 cells. However, it struggles to apply this prediction to similar molecules. Increasing the sample size could corroborate it. **Table 4** outlines how the five models conform to MLR premises. It offers a perspective to conclude the article.

Table 4. Summary of compliance analysis for all models.

	MLR Premises	Model 1	Model 2	Model 3	Model 4	Model 5
Data set	The data are homogeneous	Violated	Violated	Respected	Respected	Violated
DV	DV is continuous	Respected	Respected	Respected	Respected	Respected
	The DV follows a normal distribution.	Respected	Respected	Respected	Respected	Respected
	DV data are independent	Respected	Respected	Respected	Respected	Respected
IV	IVs are continuous	Respected	Respected	Respected	Respected	Respected
	The IVs exhibit non-zero variances	Violated	Violated	Violated	Violated	Respected
	No multicollinearity between IVs	Respected	Respected	Respected	Respected	Respected
Relationship between the DV and its IVs	The DV-IV relationship is usually linear.	Violated	Violated	Violated	Violated	Respected
Sample	The sample size satisfies Green's or Pallant's conditions.	Violated	Violated	Violated	Violated	Violated
Residues' characteristics	Residues' homoscedasticity	Violated	Violated	Violated	Violated	Violated
	Residuals are normally distributed.	Violated	Violated	Violated	Violated	Respected
	The residues are independent	Respected	Respected	Violated	Violated	Respected

5. Conclusions

This research assessed the compliance of five QSAR models with MLR premises, uncovering limitations due to validation criteria and sample size. Assumptions were summarized, and conformity was analyzed by comparing data characteristics to MLR premises. The accuracy of its predictions and the possibility of generalizing its results were emphasized. Models 1, 2, and 3 excluded outliers to meet MLR requirements. Homogeneity was proven in models 3 and 4. While all adhered to the basic hypotheses for DVs, they partially respected those of IVs. Except Model 5, others violated non-zero variance assumptions for IVs. All models complied with the premise that IVs shouldn't exhibit multicollinearity. They failed to satisfy sufficient sample size criteria. They breached the principles relating to homoscedasticity of residuals, but models 1, 2 and 5 conform to those associated with their independence. Model 5 demonstrated global linearity with an error

term that was normally distributed.

Adherence to MLR assumptions enhances forecast accuracy and generalizability. Conversely, violations of these assumptions reduce their effectiveness. Factors such as sample size, residuals' heteroscedasticity, and interdependence limit generalizability. Models 1 to 4 face issues with nonlinearity and non-normal residuals, while Models 3 and 4 don't address the error term dependence. Despite challenges in validation due to a small sample size, Model 5 accurately predicted the pulmonary effects of Thiazine or Thiazoline on A-549 cells. Verify that the data matches MLR premises when using these projected models.

Conflicts of Interest

The author declares no conflicts of interest concerning this paper.

References

[1] Field, A. (2009) Discovering Statistics Using SPSS. Sage Publication Ltd.

[2] Pingaew, R., Worachartcheewan, A., Nantasenamat, C., Prachayasittikul, S., Ruchirawat, S. and Prachayasittikul, V. (2013) Synthesis, Cytotoxicity and QSAR Study of N-tosyl-1,2,3,4-tetrahydroisoquinoline Derivatives. *Archives of Pharmacal Research*, **36**, 1066-1077. https://doi.org/10.1007/s12272-013-0111-9

[3] Pallant, J. (2023) SPSS Survival Manual: A Step-by-Step Guide to Data Analysis Using IBM SPSS. 7th Edition, Open University Press.

[4] Green, S.B. (1991) How Many Subjects Does It Take to Do a Regression Analysis. *Multivariate Behavioral Research*, **26**, 499-510. https://doi.org/10.1207/s15327906mbr2603_7

[5] Coulibaly, W.K., Affi, S.T., James, T., Koné, M.G.-R., Yao, A.E.B., Dago, C.D., *et al.* (2022) Anti-Proliferative Activity Study on 5-Arylidene Rhodanine Derivatives Using Density Functional Theory (DFT) and Quantitative Structure Activity Relationship (QSAR). *International Journal of Computational and Theoretical Chemistry*, **10**, 1-8.

[6] Dembelé, G.S., Tuo, N.T., Konaté, F., Soro, D., Konaté, B. and Ziao, N. (2022) Quantitative Structure Activity Relationship (QSAR) Study of a Series of Molecules Derived from Thiazoline and Thiazine Multithioether Having Activity against Antitumor Activity (A-549). *International Journal of Chemical and Life Sciences*, **11**, 2426-2435. https://www.researchgate.net/publication/364965605_Quantative_Structure_Activity_Relationship_QSAR_Study_of_a_Series_of_Molecules_Derived_from_Thiazoline_and_Thiazine_Multithioether_Having_Activity_against_Antitumor_Activity_A-549

[7] Baillargeon, G. (2010) Méthodologies et techniques statistiques, Trois-Rivières: Bibliothèque nationale du Québec, SMG.

[8] Lv, L., Song, X. and Sun, W. (2020) Modify Leave-One-Out Cross Validation by Moving Validation Samples around Random Normal Distributions: Move-One-Away Cross Validation. *Applied Sciences*, **10**, Article No. 2448. https://doi.org/10.3390/app10072448

[9] Flatt, C. and Jacobs, R.L. (2019) Principle Assumptions of Regression Analysis: Testing, Techniques, and Statistical Reporting of Imperfect Data Sets. *Advances in Developing Human Resources*, **21**, 484-502. https://doi.org/10.1177/1523422319869915

[10] Schmidt, A.F. and Finan, C. (2018) Linear Regression and the Normality Assumption.

Journal of Clinical Epidemiology, **98**, 146-151.
https://doi.org/10.1016/j.jclinepi.2017.12.006

[11] Yassine, T. (2020) Contribution des technologies de l'information et de la communication au succès de la collaboration client-fournisseur en développement de produits nouveaux. Université de Grenoble Alpes.
https://theses.hal.science/tel-02931916/

[12] Johnson, R. and Wichern, D. (2018) Applied Multivariate Statistical Analysis. 6th Edition, Pearson.

[13] Nachhilfe, S. (2022) Comment vérifier la condition de linéarité pour le modèle de régression linéaire dans R et SPSS?
https://statistiknachhilfe.ch/fr/2022/12/09/voraussetzung-lineare-regression-linearitat/

[14] Dodge, Y. and Rousson, V. (2004) Analyse de régression appliquée. Dunod.

[15] Moore, D.S., McCabe, G.P. and Craig, B. (2021) Introduction to the Practice of Statistics. 10th Edition, W.H. Freeman.

[16] Sing, V. (2025) Multicollinéarité dans la régression: Un guide pour les scientifiques des données.
https://www.datacamp.com/fr/tutorial/multicollinearity?form=MG0AV3

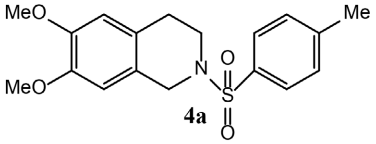
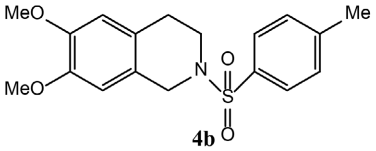
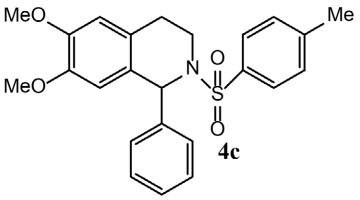
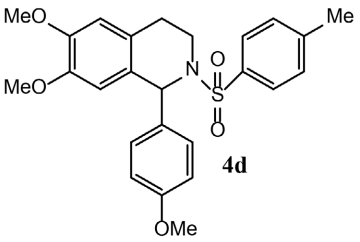
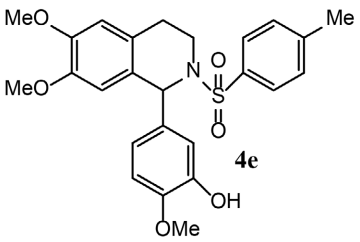
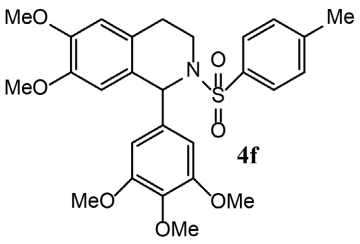
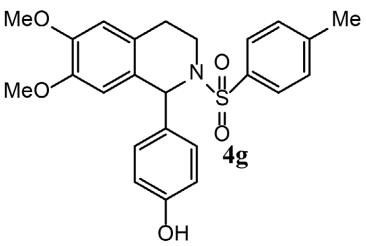
[17] Della-Vedova, C. (s.d.) Régression linéaire simple: Quand les hypothèses ne sont pas satisfaites.
https://delladata.fr/regression-lineaire-simple-quand-les-hypotheses-ne-sont-pas-satisfaites/

[18] Lind, A.D., Marchal, W., Mason, D.R., Satya, G.D., Santosh, K. and Singh, J. (2007) Méthodes statistiques pour les sciences de la gestion. Les éditions de la Chenelière Inc.

[19] Draper, N.R. and Smith, H. (1998) Applied Regression Analysis. 3rd Edition, Wiley.
https://doi.org/10.1002/9781118625590

Appendix. Pooled Training and Test Data

Table A1. N-Tosyl-1, 2, 3, 4-Tetrahydroisoquinoline Derivatives: Descriptors and Experimental Cytotoxic Activities against MOLT3 Cell Lines [1].

Order Number	Molecular Structure/Code	Gu	Mor32u	Experimental activity (pIC50)
1	 4a	0.171	−0.211	−1.817
2	 4b	0.177	−0.305	−1.747
3	 4c	0.169	−0.171	−1.190
4	 4d	0.167	−0.272	−2.042
5	 4e	0.176	−0.305	−1.564
6	 4f	0.176	−0.025	−0.966
7	 4g	0.181	−0.343	−1.302

Continued

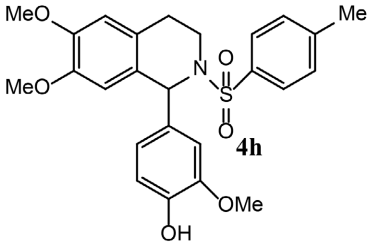
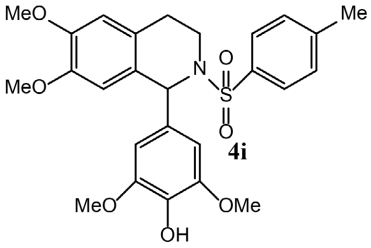
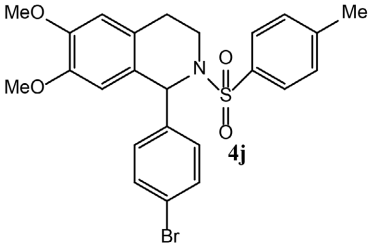
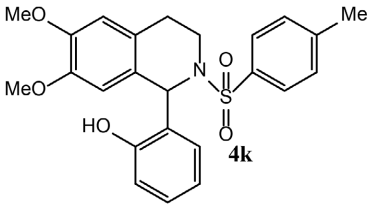
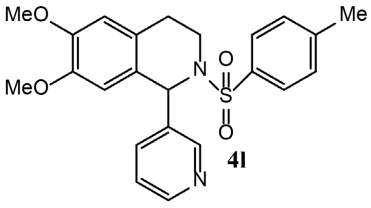
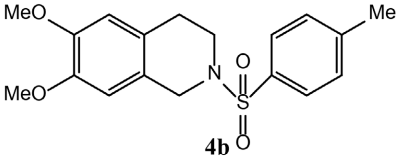
Order Number	Molecular Structure/Code	μ (eV).	LogP	Experimental activity PIC50
8		0.176	−0.439	−1.615
9		0.175	−0.447	−1.609
10		0.183	−0.345	−1.356
11		0.192	−0.265	−0.089
12		0.179	−0.226	−1.094

Table A2. N-Tosyl-1, 2, 3, 4-Tetrahydroisoquinoline Derivatives: Descriptors and Experimental Cytotoxic Activities against HepG2 Cell Lines [1].

Order Number	Molecular Structure/Code	PJI3	Mor32u	Mor31v	Experimental activity (pIC50)
1		0.900	−0.305	0.352	−2.094

Continued

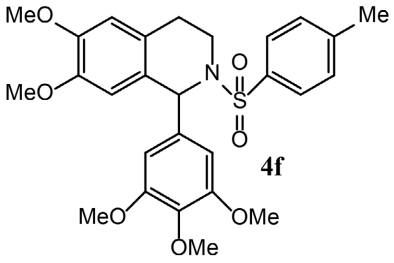
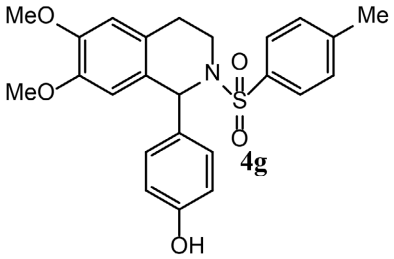
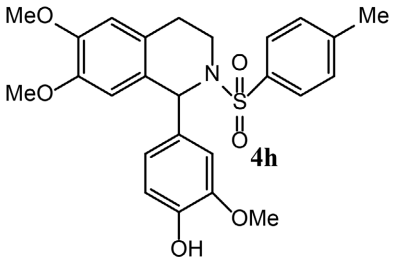
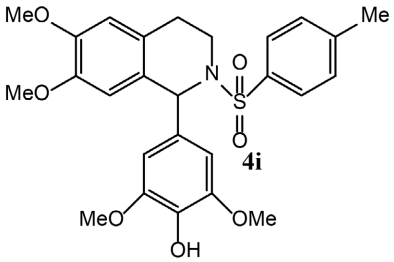
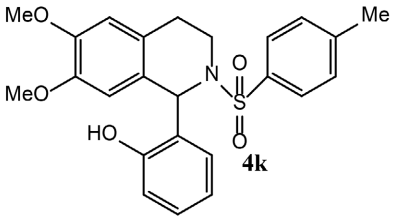
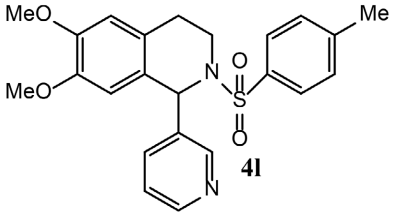
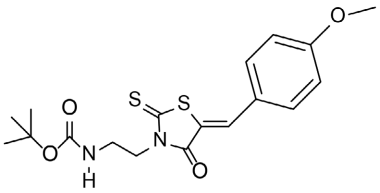
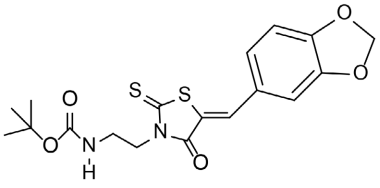
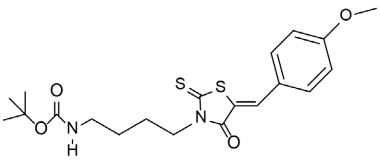
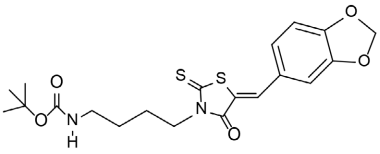
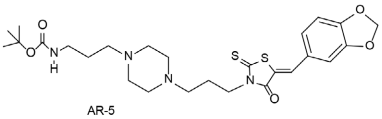
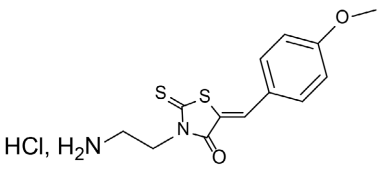
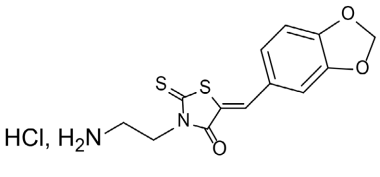
2	 **4f**	0.869	−0.025	0.223	−1.356
3	 **4g**	0.869	−0.343	0.344	−1.901
4	 **4h**	0.873	0.439	−0.258	−1.930
5	 **4i**	0.867	−0.447	0.344	−1.903
6	 **4k**	0.863	−0.265	0.28	−1.650
7	 **4l**	0.859	−0.226	0.396	−1.694

Table A3. 1,3,5-Arylidene Rhodanine Derivatives: Descriptors and Experimental Antitumour Activities against NCI-H727 (lung carcinoma) and MDA-MB 231 (breast carcinoma) Cell Lines [2].

Order Number	Molecular Structure/Code	ELUMO (eV)	$\nu_{C=O}$ (cm^{-1})	d_CN ($\dot{A}$)	Experimental Activity IC_{50} (μM)	
					NCI-H727	MDA-MB 231
1	 AR-1	−5.923	1781.900	1.373	114	110
2	 AR-2	−5.901	1782.730	1.372	97	86
3	 AR-3	−5.897	1783.920	1.371	109	110
4	 AR-4	−5.205	1787.490	1.365	109	110
5	 AR-5	−5.266	1775.920	1.371	60	33
6	 AR-6	−5.887	1783.020	1.373	13	94
7	 AR-7	−5.951	1789.880	1.365	83	74

Continued

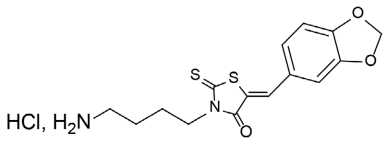
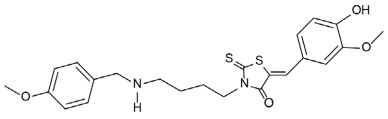
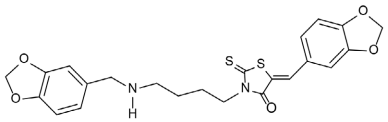
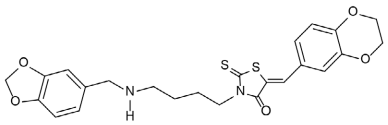
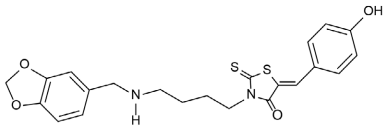
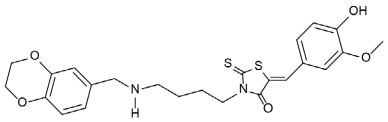
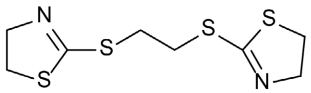
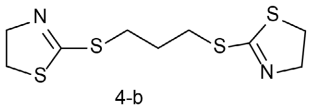
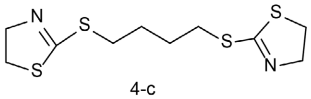
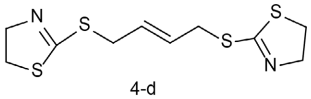
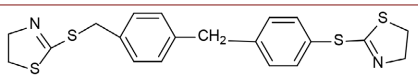
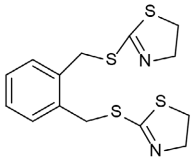
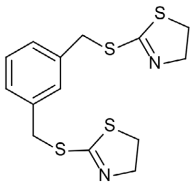
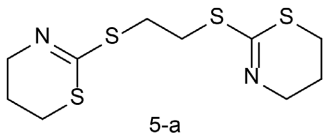
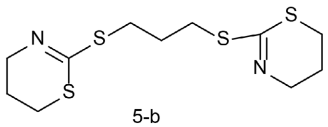
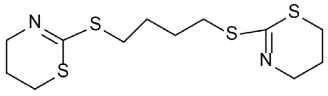
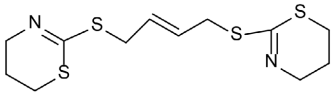
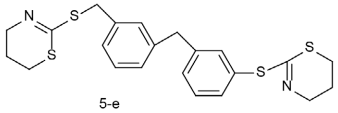
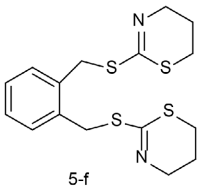
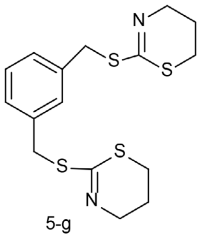
8	 AR-8	−5.834	1786.670	1.371	109	97
9	 AR-9	−5.671	1781.750	1.370	15	11
10	 AR-10	−5.373	1785.940	1.370	49	43
11	 AR-11	−5.558	1784.710	1.370	17	9
12	 AR-12	−5.432	1785.830	1.370	42	33
13	 AR-13	−4.427	1784.570	1.370	55	55

Table A4. Thiazoline and Thiazine Derivatives: Descriptors and Antitumour Activities against A-549 Cell Lines [3].

Order Number	Molecular Structure/Code	μ (eV).	LogP	Experimental activity PIC50
1	 4-a	−3.5804	1.1400	0.9787
2	 4-b	−3.5214	1.3500	1.1287
3	 4-c	3.4475	1.6500	1.2952
4	 4-d	−3.6778	1.6900	1.3441

Continued

5	 4-e	−3.4584	2.3800	1.4417
6	 4-f	−3.7123	2.5500	1.4252
7	 4-g	−3.7014	2.3800	1.6463
8	 5-a	−3.4267	1.9100	1.0455
9	 5-b	−3.3830	2.1300	1.1460
10	 5-c	−3.3702	2.4300	1.4517
11	 5-d	−3.4827	2.4600	1.6868
12	 5-e	−3.6212	3.1600	1.6409
13	 5-f	−3.5495	3.3200	1.9944
14	 5-g	−3.5973	3.1600	2.0830