QSAR and Drug-likeness Studies of Thiadiazole Derivatives Against Lung Cancer

Mouad Mouhsin[1*], Mustapha Oubenali[1], Samir Chtita[2], Aimen El orche[1], Mohamed Mbarki[1], Malika Echajia[1], Tarik El ouafy[3] and **Ahmed Gamouh[1]**

[1] Team of ACCNE - Laboratory of Engineering in Chemistry and Physics of Matter, Faculty of Science and Technologies, Beni Mellal, Morocco.
[2] Laboratory of Physical Chemistry of Materials, Faculty of Science Ben M'Sik, Hassan II, University of Casablanca, B.P. 7955, Sidi Othmane, Casablanca, Morocco
[3] Team of ACCNE - Laboratory of Engineering in Chemistry and Physics of Matter, Sultan Moulay Slimane University, Polydisciplinary Faculty, Khouribga, Morocco.
*Address for correspondence:
Team of ACCNE - Laboratory of Engineering in Chemistry and Physics of Matter, Faculty of Science and Technologies, Beni Mellal, Morocco.
Tel: +212649798537; E-mail: mouhsinmouad@gmail.com

ABSTRACT

This study was aimed at building a robust quantitative structure–activity relationship (QSAR) to predict the anti-proliferate activity of 1,3,4-thiadiazole derivatives against the A549 lung cancer cell lines. The semi-empirical PM7 parametrization approach was used to optimize the complete set of 1,3,4-thiadiazole derivatives and various classes of molecular descriptors have been calculated. We built models using Fisher score and the best subset selection for feature selection, and the final model was developed using the multiple linear regression technique, all in accordance with the rigorous Organization for Economic Co-operation and Development (OECD) requirements. Furthermore, various internationally agreed severe validation parameters were used to validate the model. Overall, our established model for quick prediction should be relevant to new, untested, or not yet produced compounds that fall within the applicability domain (AD) of the model. The drug-likeness properties of the 10 compounds with the greatest activity value were also calculated using Lipinski's rule properties.

Keywords: *QSAR, Thiadiazole derivatives, A549, PM7, OECD*

1. INTRODUCTION

Lung cancer is one of the most frequent cancers, with one of the worst fatality rates on the planet (1). It's also the most commonly diagnosed cancer among men in most Northern African countries, with 31.9 cases per 100,000 in Morocco (2). Important progress has been made in anticancer therapy, finding safe and effective anticancer drugs remains one of the most difficult tasks in drug development. Therefore, scientists are investigating a variety of compounds in order to discover novel and effective anticancer agents.

The current research is centred on thiadiazole derivatives, because they are an appealing source of novel and effective anti-cancer therapeutic compounds due to their structural variety and biological behaviour. Those substances have a wide range of pharmacological properties, including as antiviral, antibacterial, antifungal, antiparasitic, anti-inflammatory, and anticancer activities (3, 4).

The research and development of novel drug candidates is a time-consuming and expensive process. Computer-aided drug design techniques have a significant impact in shortening the drug discovery development process (5). The quantitative structure-activity relationship (QSAR) is a powerful method defined as a mathematical equation that correlates the compounds' biological activities to their molecular descriptors, this equation could be used to predict, interpret, and assess novel compounds with desired activities (6, 7). A molecular descriptor is a numerical representation of a molecule's chemical properties. There are five categories of molecular descriptors: geometrical, constitutional, topological, thermodynamic, and electronic (8, 9). Some of these descriptors require minimization and optimization of molecules. Minimization is the process of refining already-built molecules in order to obtain a favorable structure suited for optimization. Geometry optimization entails locating the equilibrium or minimal energy of conformation 10. For optimization compounds, density functional theory (DFT) is an accurate but time-consuming method. The Semi-empirical Hamiltonians method can generate valid molecular parameters for creating QSAR models in a more time-efficient manner, especially when there is a lack of experience with descriptor selection (11-13).

Several studies have recently been developed to determine the structural features that influence the biological activity of these cancer-fighting compounds (14-18). During this last decade for data measured on A549 cancer cell lines, there is no descriptive and predictive quantitative structure–activity relationships (QSARs) of 1,3,4-thiadiazole derivatives' anti-cancer activity.

2. MATERIALS AND METHODS

2.1. Dataset

The data set of 1,3,4-thiadiazole derivatives with anti-proliferative activity in A549 cancer cell lines was assembled from synthesis and experimental determination studies (19). The data set contains a total of 33 molecules (Table 1). The activities of these compounds expressed by IC50 values used as the dependent variables for QSAR modelling analyses.

2.2. Structure optimization and descriptor generation

ChemDraw software (version 16.0) (20) was used to construct the 2D structures of the compounds. The semi-empirical PM7 parametrization approach was used using MOPAC2016 software (21) to optimize the complete set of 1,3,4-thiadiazole derivatives. All the molecules

were merged into a single structural text file (sdf format) using Open Babel (version 2.3.2) (22). The PaDEL (23), Mordred (24), and ChemoPy (25) programs were used to calculate the descriptors based on the 3D geometry representation of molecules. MOPAC output files were also used to derive quantum chemical descriptors such as HUMO and LUMO energies. As a result, we have over 4000 descriptors, which include 1D, 2D, 3D, and quantum molecular descriptors. Intercorrelated descriptors (correlation coefficient cut-off 0.9) and descriptors with constant near-constant value were removed during the initial data pre-treatment (variance cut-off 0.001). After data pre-treatment, 681 descriptors remained for the creation of a 2D-QSAR model.

2.3. Data splitting

The Y-based ranking approach was used to divide data set into training and test sets (2:1 ratio). These compounds were grouped in ascending order of biological activity first, Then, for the test set, every fourth compound was selected. The test set was omitted from the highest and lowest active compounds. As a result, for model construction and validation, a training set of 22 compounds and a test set of 11 compounds were used.

2.4. Descriptor selection and model development

A vital phase in the creation of a QSAR model is the selection of significant and meaningful descriptors from a wide pool of descriptors. therefore f-regression from sklearn (26) has been used as a feature selection method to search for those descriptors with significant explained variance. It is done in two steps: Firstly, the correlation coefficient between each descriptor and biological activity is calculated as presented in equation 1.

$$R = \frac{(x_{ij} - \bar{X}_j)(y_i - \bar{Y})}{S_x S_y} \qquad (1)$$

Where x_{ij} and y_i are the values of each descriptor and activity, respectively for each molecule. $\bar{X}_j$ the mean value for each descriptor j, and $\bar{Y}$ the mean value for activity. Sx and Sy are the standard deviations for each descriptor and activity, respectively.

Then, the value of F score is computed using equation 2.

$$F = \frac{R^2(n-2)}{1-R^2} \qquad (2)$$

Where R^2 is the square of the correlation coefficient and n is the number of molecules in the training set. Finally, the value of F will be converted to p-value.

In this work, 'selectKBest' (from sklearn) has been used which allows the selection of a number of variables based on p-value, i.e., the variables with the lowest p-value. Using these methods, 100 descriptors were chosen to leave only the lucky descriptors explaining biological activity, as well as to reduce the time required to complete the next steps. Using these 100 descriptors, we ran the best subset selection (using MLR BestSubsetSelection 2.1) of DTC lab (27) to generate models. Best subset regression is a linear regression analysis technique for developing exploratory models. This method examines all potential models with a given set of variables and provides the best-fitting models that have one, two, or more variables. We chose one best model from the equations given by the best subset selection, based on the highest Q^2_{LOO} and Q^2_{f1}values. We choose the best model for each combination, resulting in a total of three models.

2.5. Validation of the QSAR models

The basic goal of all QSAR methods is to create a robust model that can appraise a novel molecule empirically, authentically, and meticulously (28). To determine our model's robustness concerning reliability and predictivity, different statistical metrics were used, as the coefficient of determination (R^2), adjusted coefficient of determination (R^2_{adj}), and standard error of estimation (SEE). Additional metrics were used to ensure that the predictions were correct. For internal validation, leave-one-out cross-validation $Q^2_{LOO}$was computed, and for external validation, metrics such as Q^2_{F1} and, Q^2_{F2} were used. Because compounds with a high value of Q^2_{LOO} do not always suggest that the predicted activity data is close to the observed ones, we computed r^2_m metrics for both the training and test sets (29, 30). The parameters suggested by Golbraikh and Tropsha for predicting model acceptability were also calculated (31). Furthermore, the Y-randomization technique is used to assess a QSAR model's robustness. The randomized model's squared mean correlation cR^2_p was calculated, as well as the degree of variance from the non-random model (R^2) (32).

2.6. Applicability domain investigation

Any QSAR model should have a well-defined applicability domain, according to OECD's third principle (33). The broad applicability domain (AD) of a QSAR model is typically what determines its reliability (34). AD denotes a theoretical region in chemical space defined by the respective model descriptors and activities where reliable predictions can be made (35). Thus, only if the compound being predicted falls inside the model's application domain may QSAR be used to predict a modeled response. To determine the AD of QSAR models, several approaches have been offered, the most frequent of which is the Williams plot method (36). Our model's applicability domain was evaluated by the leverage approach expressed as Williams plot. The steps to build this graph is done as follows:

The leverage value for all compounds in the dataset X were calculated from the hat matrix (H) (Equation 3).

$$H = X(X^T X)^{-1} X^T \quad (3)$$

Where X is the two-dimensional $n \times k$ descriptor matrix of the training set compounds that contains n compounds and k descriptors, the transpose of X is represented as XT. The leverage value of the ith compound (hi) was estimated using equation 4.

$$h_i = x_i(X^T X)^{-1} {x_i}^T \quad (i=1\ldots,) \quad (4)$$

The cut-off leverage, h*, which is the limit of normal values for X outliers, was also calculated as presented in equation 4.

$$h^* = \frac{3(k+1)}{n} \quad (5)$$

For the estimation of the standard residuals, equation 6 was employed.

$$Standard\ Residual = \frac{Residual}{RMSE} \quad (6)$$

Where $RMSE$ is the Root Mean Square Error.

The Williams plot was obtained by plotting the standard residuals against the leverage

values.

2.7. Drug-likeness

The failure of drug candidates in clinical trials is primarily due to undesirable pharmacokinetic properties. Since its inception, the idea of drug-likeness has grown in importance in the selection of molecules with optimal bioavailability (37).

The molecular weight (MW), partition coefficient (log P), H-bond acceptors, H-bond donors (HBA and HBD), and topological polar surface area (TPSA) of the 10 molecules with the highest activity value were predicted using the SwissADME online tool (38).

At the preclinical stage of drug development, Lipinski's rule of five (5) is one of the most common and useful rules. Lipinski's rule of five (5) predicts that if a molecule fails more than two of these parameters (MW ≤ 500, HBD ≤ 5, HBA ≤ 10, Log P ≤ 5 and TPSA ≤ 140 Å2), it will be poorly absorbed in the preclinical stage (39).

Log S was also used to determine the water solubility of the chosen molecules (molecule is insoluble or poorly soluble if log S ≤ −6, moderately soluble if −6 < log S ≤ −4, soluble if −4 < log S) (40).

3. RESULTS AND DISCUSSION

3.1. Validation of generated models and analysis of their results

The molecular descriptors obtained for the 1,3,4-Thiadiazole derivatives were pre-treated, as previously stated. Finally, using the ‘best subset selection' technique, 100 descriptors were examined. Using a combination of 4, 5, and 6 descriptors, many models were created using the MLR approach. Based on the two parameters Q^2_{LOO} and Q^2_{f1}, the best model for each combination is selected. These models are represented in Table 2.

The derived models were validated carefully using different validation metrics. Table 3 presents an overview of the three models' internal validation results.

All the three models satisfied the model acceptability standards based on the results of their internal validation, with the third model taking priority because it has the greatest Q^2_{LOO}value. This last one can explain 76.1% of the variance and predicted 72.5% of the variance of biological activity. Meanwhile, the higher values of $\overline{r^2_m}$ = 0.68 and the lower value of $\Delta r^2_{m(LOO)}$ ($\Delta r^2_{m(LOO)} = 0.25$) demonstrated that the model was internal robust.

We also remark that include a descriptor increases the model's quality, but it can lead to obtain overfitting. Thus, an external validation must be done to ensure that the proposed models do not suffer from the overfitting problem.

From the results presented in Table 4, all the models gave encouraging results (R^2> 0.6 and $Q^2_{f1} > 0.6$). Moreover, these generated models passed all the Golbraikh and Tropsha criteria for acceptability, indeed the results of their external validation are within the recommended threshold values as shown in Table 5. It can be observed that model 3 is statistically better than the other models in the external validation parameters.

In addition to assess each descriptor's modeling qualification, we computed its Variable Inflation Factor (VIF) and significance (p-value). The multicollinearity problem lay between the modeling descriptor would be severe if its VIF > 10, slightly with VIF > 5, and inexistence if VIF < 5 (41). The VIF values of all descriptors are fewer than 5, as shown in table 6, implying that the correlations between these descriptors were poor and that there were no

multicollinearities among these six descriptors. The same table indicates also that the regression coefficients are significant at the 95% confidence level (p-value < 0.5).

The correlation matrix was also plotted to further verify the selection of descriptors in the final model (Table 7). These six descriptors had poor Pearson's correlation coefficient values (< 0.6), indicating that the descriptors in the constructed model were independent of one another.

To see if the modelling work contains a chance correlation, the Y-Randomization test was performed. In this technique, the activity column entries are scrambled, and new QSAR models are developed using the same set of descriptors as present in the original model. In the present case, 50 random trials were run, and a squared mean correlation of the randomized model cR_p^2 was estimated. Table 8 demonstrates the robustness of the proposed QSAR model and the absence of chance correlation during the modelling process (cR_p^2 >0.5).

Figure 1 presents radar plots comparing observed and predicted values estimated by model (3).The level of overlap between blue and brown lines in the radar graphs depicts the difference between observed and predicted activities for the training and test sets individually. The graphic demonstrates that the real activities and the activities predicted by model 3 overlaps for most compounds, which confirms that this prediction is appropriate.

3.2. Molecular descriptors and their significance

Table 9 lists the definitions and categories for the six descriptors that were chosen.

TDB9v, which is topological distance-based autocorrelation – lag 9/weighted by van der Waals volumes, has the biggest contribution with a negative influence on the biological activity, as seen in table 9. This descriptor is obtained by summing up the products of certain properties of two atoms located at given topological distances. The topological distance represents the minimum number of bonds between two atoms. Similarly, spatial autocorrelation considers properties on the molecular surface separated by a given Euclidean distance (42).

PNSA-1 is the partial negatively charged molecular surface area. Charged partial surface area descriptors encoding characteristics were responsible for polar molecular interactions. The molecular depiction used here considers that a molecule has a surface defined by the overlap of the hard sphere, as defined by the atoms' van der Waals radii (43). The descriptor contributes to the biological activity in a negative way.

MoRSEU15 and Mor06 are a 3D-MoRSE (3D- molecular representation of structure based on electron diffraction) un-weighted descriptors and encode information of structural features like mass and number of branching in substituted benzene derivative (44). The positive sign in these descriptors' regression coefficients indicates that adding substituents to the aromatic ring increases the value of activity.

Mor04m is a 3D-MoRSE – signal 4/weighted by atomic masses. The negative sign in the regression coefficient of this descriptor negatively reflects the role of atomic masses on the activity of this derivatives.

RDFU5 is an RDF (radial distribution function) descriptor. The RDF descriptors are based on the distance distribution in the molecule. The radial distribution function of an ensemble of n atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of radius R (45). The importance of the descriptor RDFU5, which has a negative effect on biological activity, is shown in Table 9.

3.3. Applicability domain

Figure 2 shows the 1.3.4-Thiadiazole derivatives dataset's Williams plot. in which the standardized residuals for each molecule in the dataset were plotted against their leverage values to identify possible outliers and standout compounds in the models. As illustrated in

Fig. 2 none of the 33 compounds in the model are outside the range of ±3 standard deviation units. Moreover, a value of 3 for standardized residual is typically employed as a cut-off value for accepting predictions, because data that is normally distributed is covered by points that are 3 standardized residuals from the mean (46). As shown in this plot, all of the compounds' h values are less than the leverage threshold h*. This means that no compound can be classified as an outlier in terms of chemical structure.

3.4. Drug-likeness evaluation

The use of computational tools to identify novel drug candidates reduces the number of experimental investigations and increases the success rate. As an initial screening step for oral bioavailability, we employed Lipinski's rule of five for drug-likeness.

Table 10's results indicate that the ten compounds have good results and follow these criteria. As a result, it appears that all of the chosen compounds have satisfactory oral medicine bioavailability. Furthermore, all of these compounds have a modest water solubility. Because the log S value is between 6 and 2, it may help with oral adsorption.

CONCLUSION

In this study, we have developed an MLR-regression-based QSAR model from 33 compounds having defined anti-proliferate activity against the A549 lung cancer cell lines to investigate the structural requirements or molecular properties essential for the anti-proliferate activity.

The most relevant descriptors were chosen using a variety of variable selection strategies, and the final model was built using the multiple linear regression technique. The model was thoroughly validated using both internal and external validation metrics, with the results demonstrating the generated model's reliability and utility ($Q^2 = 0.751$, R^2_{pred} or $Q^2_{F1} = 0.812$ and $Q^2_{F2} = 0.812$).

CONFLICT OF INTEREST

The authors declare that no conflict of interest, financial or otherwise.

REFERENCES

1. Global Burden of Disease Cancer Collaboration. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2016: A Systematic Analysis for the Global Burden of Disease Study. JAMA Oncol. 2018;4(11):1553–1568. DOI: 10.1001/jamaoncol.2018.2706.
2. World Health Organization, The International Agency for Research on Cancer (IARC) Report, World Health Organization, Geneva, Switzerland, 2018. URL: https://www.iarc.who.int/wp-content/uploads/2018/09/pr263_E.pdf
3. Szeliga, M. Thiadiazole derivatives as anticancer agents. Pharmacol. Rep. 2020;72**:**1079–1100. DOI: 10.1007/s43440-020-00154-7
4. Li Y, GENG J, LIU Y, Yu S, Zhao G. Thiadiazole—A promising structure in medicinal chemistry. Chem. Med. Chem. 2013;8(1):27-41. DOI: 10.1002/cmdc.201200355.
5. Poroikov VV. Computer-aided drug design: from discovery of novel pharmaceutical agents to systems pharmacology. Biomed Khim. 2020;66(1):30-41. DOI: 10.18097/PBMC20206601030

6. Dastmalchi S, Hamzeh-mivehroud M, Sokouti B. Quantitative structure–activity relationship: a practical approach. CRC Press. 2018:37-35. DOI: doi.org/10.1201/9781351113076.
7. Chtita S, Aouidate A, Belhassan A, Ousaa A, Taourati AI., Elidrissi B, et al. QSAR study of N-substituted oseltamivir derivatives as potent avian influenza virus H5N1 inhibitors using quantum chemical descriptors and statistical methods. New J. Chem. 2020;44(5):1747-1760. DOI: doi.org/10.1039/C9NJ04909F.
8. Todeschini R, Consonni V, Handbook of Molecular Descriptors, Method and Principles in Medicinal Chemistry. John Wiley & Sons. 2008;11.
9. Shaker B, Ahmad S, Lee J, Jung C, Na D. In silico methods and tools for drug discovery. Comput Biol Med. 2021;137:104851. DOI: 10.1016/j.compbiomed.2021.104851.
10. Oyewole RO, Oyebamiji AK, Semire B. Theoretical calculations of molecular descriptors for anticancer activities of 1, 2, 3-triazole-pyrimidine derivatives against gastric cancer cell line (MGC-803): DFT, QSAR and docking approaches. Heliyon. 2020;6(5):e03926. DOI: 10.1016/j.heliyon.2020.e03926.
11. Alov P, Tsakovska I, Pajeva I. Hybrid Classification/Regression Approach to QSAR Modeling of Stoichiometric Antiradical Capacity Assays' Endpoints. Molecules. 2022;27(7):2084. DOI: 10.3390/molecules27072084.
12. Hegde G, Bowen RC. Machine-learned approximations to Density Functional Theory Hamiltonians. Sci Rep. 2017;7:42669. DOI: 10.1038/srep42669.
13. Shi Y. Support vector regression-based QSAR models for prediction of antioxidant activity of phenolic compounds. Sci Rep. 2021;11(1):8806. DOI: 10.1038/s41598-021-88341-1.
14. Raj V, Aboumanei MH, Rai A, Verma SP, Singh AK, Keshari, AK, Saha S. Pharmacophore and 3d-Qsar Modeling of new 1, 3, 4-Thiadiazole Derivatives: Specificity to Colorectal Cancer. Pharm. Chem. J.(2020);54(1):12-25. DOI: doi.org/10.1080/10406638.2021.1970586.
15. Chukwuemeka PO, Umar HI, Iwaloye O, Oretade OM, Olowosoke CB, Oretade OJ,et al. Predictive hybrid paradigm for cytotoxic activity of 1,3,4-thiadiazole derivatives as CDK6 inhibitors against human (MCF-7) breast cancer cell line and its structural modifications: rational for novel cancer therapeutics. J. Biomol. Struct. Dyn. 2021:1-20. DOI: 10.1080/07391102.2021.1913231.
16. Farooqi SI, Arshad N, Channar PA, Perveen F, Saeed A, Larik FA, et al. Synthesis, theoretical, spectroscopic and electrochemical DNA binding investigations of 1, 3, 4-thiadiazole derivatives of ibuprofen and ciprofloxacin: Cancer cell line studies. J. Photochem. Photobiol B. 2018;189:104-118. DOI: 10.1016/j.jphotobiol.2019.111733.
17. Wassel MMS, Ammar YA, Elhag Ali GAM, Belal A, Mehany ABM, Ragab A. Development of adamantane scaffold containing 1,3,4-thiadiazole derivatives: Design, synthesis, anti-proliferative activity and molecular docking study targeting EGFR. Bioorg Chem. 2021;110:104794. DOI: 10.1016/j.bioorg.2021.104794.
18. Cascioferro S, Petri G. L, Parrino B, Carbone D, Funel, N, Bergonzini C, et al. Imidazo [2, 1-b][1, 3, 4] thiadiazoles with antiproliferative activity against primary and gemcitabine-resistant pancreatic cancer cells. Eur. J. Med. Chem. 2020;189:112088. DOI: 10.1016/j.ejmech.2020.112088;
19. Zhang S, Liu XJ, Tang R, Wang HX, Liu HY, Liu YM, Chen BQ. Design, Synthesis and Antiproliferative Evaluation of Novel Disulfides Containing 1,3,4-Thiadiazole Moiety. Chem. Pharm. Bull (Tokyo). 2017;65(10):950-958. DOI: 10.1248/cpb.c17-00485.

20. Chem3D,PerkinElmer Informatics, 2010, http://www.cambridgesoft.com.
21. MOPAC2016, James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, URL: OpenMOPAC.net .
22. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J. Cheminform. 2011
23. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 2011;32(7):1466-74. DOI: 10.1002/jcc.21707
24. Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. J Cheminform. 2018;10(1):4. DOI: 10.1186/s13321-018-0258-y.
25. Cao DS, Xu QS, Hu QN, Liang YZ. ChemoPy: freely available python package for computational biology and chemoinformatics. Bioinformatics. 2013;29(8):1092-4. DOI: 10.1093/bioinformatics/btt105.
26. PEDREGOSA, Fabian, VAROQUAUX, Gaël, GRAMFORT, Alexandre, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 2011;12:2825-2830.
27. QSAR tools, DTC laboratory, India, 2015; Software available at URL: dtclab.webs.com/software-tools.
28. Kumar P, Kumar A, Sindhu J. Design and development of novel focal adhesion kinase (FAK) inhibitors using Monte Carlo method with index of ideality of correlation to validate QSAR. SAR QSAR Environ. Res. 2019;30(2):63-80 DOI: 10.1080/1062936X.2018.1564067
29. Roy K, Mitra I. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. Comb. Chem. High Throughput Screen. 2011;14(6):450-74. DOI: 10.2174/138620711795767893.
30. Chirico N, Gramatica P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J. Chem. Inf. Model. 2011;51(9):2320-35. DOI: 10.1021/ci200211n.
31. Zhou P, Liu Q, Wu T, Miao Q, Shang S, Wang H, Chen Z, Wang S, Wang H. Systematic Comparison and Comprehensive Evaluation of 80 Amino Acid Descriptors in Peptide QSAR Modeling. J Chem Inf Model. 2021;61(4):1718-1731. DOI: 10.1021/acs.jcim.0c01370.
32. Roy K, Kar S, Das RN. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic press. 2015.
33. OECD, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, 2014.
34. Wang Z, Chen J, Hong H. Developing QSAR Models with Defined Applicability Domains on PPARγ Binding Affinity Using Large Data Sets and Machine Learning Algorithms. Environ Sci Technol. 2021;55(10):6857-6866. DOI: 10.1021/acs.est.0c07040.
35. Gadaleta D, Mangiatordi GF, Catto M, et al. Applicability domain for QSAR models: where theory meets reality. IJQSPR. 2016;1(1):45-63. DOI: 10.4018/IJQSPR.2016010102
36. Liu H, Yang X, Lu R. Development of classification model and QSAR model for predicting binding affinity of endocrine disrupting chemicals to human sex hormone-binding globulin. Chemosphere. 2016;156:1-7. DOI: 10.1016/j.chemosphere.2016.04.077.
37. Mouchlis VD, Afantitis A, Serra A, Fratello M, Papadiamantis AG, Aidinis V, Lynch I, Greco D, Melagraki G. Advances in de Novo Drug Design: From Conventional to Machine Learning Methods. Int J Mol Sci. 2021 Feb 7;22(4):1676. DOI: 10.3390/ijms22041676.
38. Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics,

drug-likeness and medicinal chemistry friendliness of small molecules. Sci. Rep. 2017;7:42717. DOI: 10.1038/srep42717.

39. Roskoski R Jr. Properties of FDA-approved small molecule protein kinase inhibitors: A 2021 update. Pharmacol Res. 2021;165:105463. DOI: 10.1016/j.phrs.2021.105463.
40. Chtita S, Belhassan A, Bakhouch M, Taourati AI, Aouidate A, Belaidi S, Moutaabbid M, Belaaouad S, Bouachrine M, Lakhlifi T. QSAR study of unsymmetrical aromatic disulfides as potent avian SARS-CoV main protease inhibitors using quantum chemical descriptors and statistical methods. Chemometr. Intell. Lab. Syst. 2021;210:104266. DOI: 10.1016/j.chemolab.2021.104266.
41. Alam S, Nasreen S, Ahmad A, Darokar MP, Khan F. Detection of Natural Inhibitors against Human Liver Cancer Cell Lines through QSAR, Molecular Docking and ADMET Studies. Curr Top Med Chem. 2021;21(8):686-695. doi: 10.2174/1568026620666201204155830.
42. Klein CT, Kaiser D, Ecker G. Topological distance based 3D descriptors for use in QSAR and diversity analysis. J. Chem. Inf. Comput. Sci. 2004;44(1):200-9. DOI: 10.1021/ci0256236
43. Stanton DT, Jurs PC. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. Anal. Chem. 1990;62(21):2323-2329. DOI: 10.1021/ac00220a013
44. Devinyak O, Havrylyuk D, Lesyk R. 3D-MoRSE descriptors explained. J. Mol. Graph. Model. 2014;54:194-203. DOI: 10.1016/j.jmgm.2014.10.006.
45. González MP, Terán C, Teijeira M, Helguera AM. Radial distribution function descriptors: an alternative for predicting A2 A adenosine receptors agonists. Eur. J. Med. Chem. 2006;41(1):56-62. DOI: 10.1016/j.ejmech.2005.08.004.
46. Alisi I, Uzairu A, Abechi SE, IDRİS SO. Development of predictive antioxidant models for 1, 3, 4-oxadiazoles by quantitative structure activity relationship. Journal of the Turkish Chemical Society Section A: Chemistry. 2019;6(2):103-114. doi.org/10.18596/jotcsa.406207 DOI: 10.18596/jotcsa.406207.

Table 1. SMILES of the 1,3,4-thiadiazole derivatives used in the QSAR study

Name	SMILES	IC_{50}
1	CCC(C)SSc1nnc(SCC(=O)Nc2ccccc2)s1	29.52
2	CCC(C)SSc1nnc(SCC(=O)Nc2ccccc2Cl)s1	28.49
3	CCC(C)SSc1nnc(SCC(=O)Nc2cccc(Cl)c2)s1	26.72
4	CCC(C)SSc1nnc(SCC(=O)Nc2ccc(Cl)cc2)s1	25.9
5	CCC(C)SSc1nnc(SCC(=O)Nc2cccc([N+](=O)[O-])c2)s1	25.86
6	CCC(C)SSc1nnc(SCC(=O)Nc2ccc([N+](=O)[O-])cc2)s1	25.47
7	CCC(C)SSc1nnc(SCC(=O)Nc2ccccc2OC)s1	24.42
8	CCC(C)SSc1nnc(SCC(=O)Nc2ccc(OC)cc2)s1	23.27
9[a]	CCC(C)SSc1nnc(SCC(=O)Nc2ccc(F)cc2)s1	22.12
10 [a]	CCC(C)SSc1nnc(SCC(=O)Nc2ccc(C(F)(F)F)cc2)s1	20.02
11[a]	CCC(C)SSc1nnc(SCC(=O)Nc2ccc(C)cc2)s1	19.9
12 [a]	CCCCSSc1nnc(SCC(=O)Nc2ccccc2)s1	17.63
13 [a]	CCCCSSc1nnc(SCC(=O)Nc2ccccc2Cl)s1	17.29
14	CCCCSSc1nnc(SCC(=O)Nc2cccc(Cl)c2)s1	12.76
15 [a]	CCCCSSc1nnc(SCC(=O)Nc2ccc(Cl)cc2)s1	12.65
16	CCCCSSc1nnc(SCC(=O)Nc2cccc([N+](=O)[O-])c2)s1	12.19
17	CCCCSSc1nnc(SCC(=O)Nc2ccc([N+](=O)[O-])cc2)s1	9.8
18 [a]	CCCCSSc1nnc(SCC(=O)Nc2ccccc2OC)s1	9.5
19 [a]	CCCCSSc1nnc(SCC(=O)Nc2ccc(OC)cc2)s1	9.31
20	CCCCSSc1nnc(SCC(=O)Nc2ccc(F)cc2)s1	8.71
21 [a]	CCCCSSc1nnc(SCC(=O)Nc2ccc(C(F)(F)F)cc2)s1	8.71
22	CCCCSSc1nnc(SCC(=O)Nc2ccc(C)cc2)s1	8.69
23 [a]	CC(C)CSSc1nnc(SCC(=O)Nc2ccccc2)s1	8.56
24	CC(C)CSSc1nnc(SCC(=O)Nc2ccccc2Cl)s1	8.51
25	CC(C)CSSc1nnc(SCC(=O)Nc2cccc(Cl)c2)s1	8.26
26	CC(C)CSSc1nnc(SCC(=O)Nc2ccc(Cl)cc2)s1	8.17
27	CC(C)CSSc1nnc(SCC(=O)Nc2cccc([N+](=O)[O-])c2)s1	8.02
28	CC(C)CSSc1nnc(SCC(=O)Nc2ccc([N+](=O)[O-])cc2)s1	8
29 [a]	COc1ccccc1NC(=O)CSc1nnc(SSCC(C)C)s1	7.84
30	COc1ccc(NC(=O)CSc2nnc(SSCC(C)C)s2)cc1	7.71
31	CC(C)CSSc1nnc(SCC(=O)Nc2ccc(F)cc2)s1	7.7
32	CC(C)CSSc1nnc(SCC(=O)Nc2ccc(C(F)(F)F)cc2)s1	6.53
33	Cc1ccc(NC(=O)CSc2nnc(SSCC(C)C)s2)cc1	3.62

[a] compound used in test set

Table 2. Developed MLR models for 1,3,4-thiadiazole derivatives.

Model	Equation
1	IC_{50}= -3.313 + 1.029TDB8e - 11.081RDF130v + 7.071Mor06 - 1.758RDFU5
2	IC_{50}= 9.625 + 1.285TDB8e - 2.715RDF130s + 8.271Mor06 - 21.561GATSm4 - 2.006RDFU5
3	IC_{50}= 79.852 - 0.042TDB9v + 0.061PNSA-1 + 11.889Mor06 - 3.533Mor04m - 1.603RDFU5 + 17.994MoRSEU15

Table 3. Internal validation results for the three models.

Parameters	Model 1	Model 2	Model 3
R^2	0.802	0.803	0.869
R^2_{adj}	0.756	0.741	0.816
Q^2_{LOO}	0.676	0.685	0.751
Standard Error of Estimate	3.927	4.043	3.406
F-value (95%)	17.258	13.035	16.560
$\overline{r^2_{m(LOO)}}$	0.594	0.604	0.681
$\Delta r^2_{m(LOO)}$	0.065	0.051	0.025

Table 4. External validation results for the three models.

Parameters	Model 1	Model 2	Model 3
R^2	0.602	0.675	0.819
Q^2_{f1}	0.601	0.668	0.812
Q^2_{f2}	0.601	0.668	0.812
RMSEP	4.941	4.506	3.388

Table 5. Golbraikh and Tropsha acceptable model criteria for the three models

Parameter	Threshold value	Model 1	Model 2	Model 3
Q^2_{LOO}	Q^2_{LOO}> 0.5	0.676	0.685	0.751
R^2_{test}	R^2_{test} > 0.6	0.602	0.675	0.819
| r20 –r'^2_0 |	|r20 – r'^2_0| < 0.3	0.239	0.178	0.070
k	[0.85<k<1.15 and ((r^2- r^2_0)/ r^2)<0.1] OR [0.85<k'<1.15 and ((r^2-r'^2_0)/ r^2)<0.1]	1.013	1.033	0.989
k’		0.899	0.897	0.969
(r^2 – r^2_0)/ r^2		0.000	0.004	0.007
(r^2 – r'^2_0)/ r^2		0.398	0.268	0.093

Table 6. The descriptors name and their statistical parameters

	Coefficient	Std. error	t test	sig	VIF
Intercept	79.852	21.269	3.754	0.002	
TDB9v	-0.042	0.011	-3.656	0.002	3.884
PNSA-1	0.061	0.022	2.793	0.014	1.337
Mor06	11.889	2.251	5.281	0.000	1.827
Mor04m	-3.534	0.802	-4.406	0.001	2.013
RDFU5	-1.604	0.519	-3.088	0.007	1.595
MoRSEU15	17.994	4.209	4.275	0.001	1.461

Table 7. Correlation matrix among descriptors and between each descriptor and activity value

	TDB9v	PNSA-1	Mor06	Mor04m	RDFU5	MoRSEU15	ACTIVITY
TDB9v	**1**	0.385	0.538	-0.571	-0.501	0.501	0.560
PNSA-1		**1**	0.217	-0.046	-0.368	0.192	0.441
Mor06			**1**	0.007	-0.132	0.340	0.579
Mor04m				**1**	0.368	-0.252	-0.465
RDFU5					**1**	-0.039	-0.461
MoRSEU15						**1**	0.593

Table 8. Random Models Parameters

Average R	0.434
Average R^2	0.203
Average Q^2_{LOO}	-0.559
cR^2_p	0.731

Table 1. Contribution of the model descriptors

Descriptor	Degree of contribution	Percentage of contribution
TDB9v	-0.674	21.9
PNSA-1	0.302	9.8
Mor06	0.667	21.7
Mor04m	-0.585	19.0
RDFU5	-0.365	11.9
MoRSEU15	0.483	15.7

Table 2. Prediction of molecular properties of descriptors for the 10 compounds.

Compound	Lipsinki's parameters					number of violations	solubility	
Molecule	MW	HBA	HBD	Log P	TPSA		Log S	Class
M26	406.01	3	1	2.80	159.02	1	-5.24	Moderately soluble
M21	439.56	6	1	3.17	159.02	1	-5.38	Moderately soluble
M25	406.01	3	1	2.80	159.02	1	-5.24	Moderately soluble
M20	389.55	4	1	2.67	159.02	1	-4.68	Moderately soluble
M10	439.56	6	1	3.17	159.02	1	-5.49	Moderately soluble
M28	416.56	5	1	1.31	204.84	1	-4.70	Moderately soluble
M24	406.01	3	1	2.80	159.02	1	-5.24	Moderately soluble
M15	406.01	3	1	2.80	159.02	1	-5.12	Moderately soluble
M17	416.56	5	1	1.31	204.84	1	-4.59	Moderately soluble
M32	439.56	6	1	3.17	159.02	1	-5.50	Moderately soluble
Threshold	MW ≤ 500	HBA ≤ 10	HBD ≤ 5	Log P ≤ 5	TPSA ≤ 140	N.viol≤1	Log S ≥ -6	

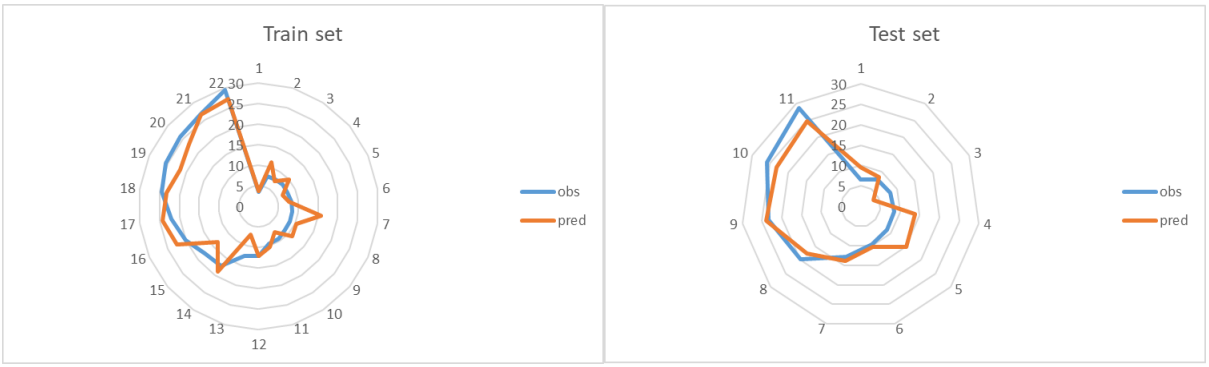

Fig. (1). A radar plot depicting the observed (Obs) and predicted (Pred) behavior of the train set and the test set using the model (3).

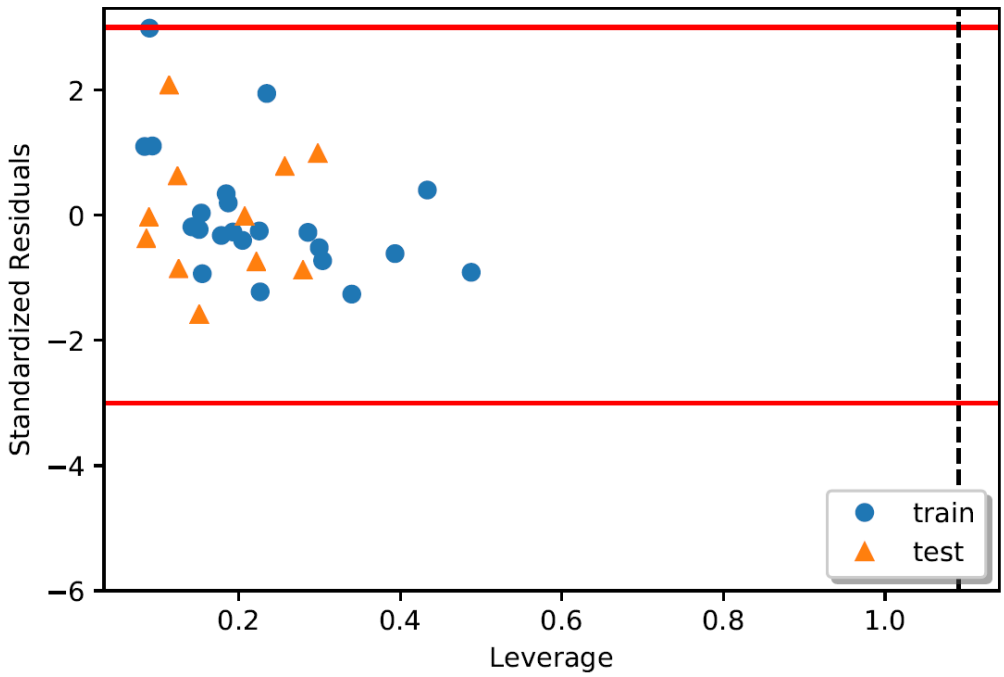

Fig. (1). William's Plot for 1.3.4-thiadiazole derivatives