In *silico* study for discovering novel thiazole derivatives as anti-breast cancer agents (MCF-7)

Khám phá các dẫn xuất thiazole mới có hoạt tính kháng ung thư vú (MCF-7) bằng phương pháp *In silico*

Cao Thanh Nhan[1], Pham Van Tat[2], Nguyen Minh Quang[1]

[1]Faculty of Chemical Engineering, Industrial University of Ho Chi Minh City, Ho Chi Minh City

[2]Institute of Pharmaceutical Education and Research, Binh Duong University, Thu Dau Mot City

Corresponding author: Nguyen Minh Quang. Email: nguyenminhquang@iuh.edu.vn

Abstract: The work validated the pharmacokinetic characteristics and docking approach, and discovered novel thiazole compounds with MCF-7 breast anti-cancer efficacy using quantitative structure-activity relationships (QSAR) models. Using an experimental dataset of 53 thiazole derivatives with IC_{50} values, QSAR models were developed using multivariate linear regression ($QSAR_{MLR}$) and artificial neural networks ($QSAR_{ANN}$). Eight descriptions, with statistical values of R2train = 0.875 and Q^2_{LOO} = 0.834, comprised the successfully developed $QSAR_{MLR}$ model. Based on the descriptors of the $QSAR_{MLR}$ model, which has statistical values of R^2 = 0.918, Q^2_{test} = 0.934, and Q^2_{CV} = 0.916, the $QSAR_{ANN}$ neural network model with the architectural network of I(8)-HL(9)-O(1) has also been constructed. Using the AD and outliers analysis, the models were utilized to forecast 120 new design derivatives based on the thiazole framework, and the IC_{50} values of 10 new thiazole derivatives were determined. Additionally, the derivatives were assessed for resistance to estrogen-positive breast cancer by docking them onto the Polo-like kinases (Plk1) receptor and screening them for pharmacokinetic features in accordance with Lipinski and Ghose guidelines. Consequently, it was discovered that the thiazole TAZ5 had promising derivative activity against the primary MCF-7 breast cancer cell line.

Keywords: *Anti-breast cancer; docking; MCF-7; QSAR; thiazole*

Tóm tắt: Nghiên cứu tìm kiếm các dẫn xuất thiazole mới có hoạt tính kháng ung thư vú (MCF-7) bằng cách xây dựng mô hình QSAR dựa trên hồi quy tuyến tính đa biến ($QSAR_{MLR}$) và mạng thần kinh nhân tạo ($QSAR_{ANN}$) trên tập dữ liệu thực nghiệm gồm 53 dẫn xuất của thiazole với các giá trị IC_{50} kết hợp với đánh giá tính giống dược và mô phỏng docking. Kết quả mô hình $QSAR_{MLR}$ có tám mô tả được phát triển thành công, với giá trị thống kê của R^2_{train} = 0,875 và Q^2_{LOO} = 0,834. Dựa trên các mô tả của mô hình $QSAR_{MLR}$, mô hình mạng thần kinh $QSAR_{ANN}$ với kiến trúc I(8)-HL(9)-O(1) cũng đã được xây dựng có giá trị thống kê là R^2 = 0,918, Q^2_{test} = 0,934 và Q^2_{CV} = 0,916. Sử dụng phân tích miền ứng dụng và quan sát ngoại biên (AD and outliers), các mô hình đã được sử dụng để dự đoán 120 dẫn xuất thiết kế mới dựa trên khung thiazole, và các giá trị IC_{50} của 10 dẫn xuất thiazole mới đã được xác định thuộc miền AD. Ngoài ra, các dẫn xuất đã được đánh giá khả năng kháng với ung thư vú dương tính với estrogen bằng cách gắn chúng vào thụ thể Polo-like kinase (PLK1) và đánh giá đặc tính dược động học của chúng theo quy tắc của Lipinski và Ghose. Kết quả, chúng tôi đã phát hiện ra dẫn xuấ thiaxole TAZ5 có hoạt tính cao chống lại dòng tế bào ung thư vú MCF-7.

***Từ khóa*:** *Dẫn xuất thiazole; Docking; Kháng ung thư vú; MCF-7; QSAR*

1. Introduction

A heterocyclic molecule with five members, thiazole has a sulphur atom and a nitrogen atom. It has several biological uses, including the ability to fight cancer, and is a flexible building component in organic synthesis. It has been demonstrated that thiazole-based drugs are effective against several cancer cell lines, including breast cancer [1]. Conversely, breast cancer is the most prevalent kind of cancer among women globally. There are several subtypes of this complicated illness, and each has a different biology and reaction to therapy. The well-known breast cancer cell line MCF-7 is frequently utilized in preclinical studies. Because MCF-7 cells express the estrogen receptor (ER+), they are dependent on oestrogen for development and viability [2].

It has been demonstrated that thiazole-based substances impede MCF-7 cell growth and proliferation via a number of different ways. Certain cancer-related proteins, such the estrogen receptor or the epidermal growth factor receptor (EGFR), are the targets of certain thiazole-based drugs. Others cause apoptosis or cell cycle halt. Although thiazole-based compounds are currently in the early phases of research and development, they might represent a novel class of anti-breast cancer medications. There are now many thiazole-based drugs being studied in clinical studies to treat breast cancer. In addition, the thiazole structure was found in over eighteen FDA-approved medications, such as ritonavir, abafungin, dasatinib, tiazofurin, and montelukast, in addition to over seventy other investigational medications [1].

In this study, the target of this study is to find potential thiazole derivatives for anti-breast cancer (MCF-7 cell line) by docking with Polo-like kinase 1 (PLK1) (PDB-3FC2). One significant gene involved in cell division and a recognized target for cancer medications is Polo-like kinase 1 (PLK1), the master mitotic regulator. Numerous distinct cancer forms have been identified to overexpress it, and this tumoral overexpression frequently indicates a bad prognosis for the patient. Polo-like kinases (PLK1) enable tumor cells to proliferate destructively and greatly induce the formation of cell circles, according to research by Sanhaji [3]. Cells can overcome obstacles when PLK1 is overexpressed, leading to genomic ambiguity and inducing changes in mammalian cells. PLK1 is one of the most effective receptors for treating breast cancer. In human breast cancer cells, PLK1 mediates the ER, which controls the overexpression of certain genes [3].

The work computes quantum parameters and optimizes the structure using quantum mechanical computation using the novel semi-experimental approach PM7. The work uses an experimental dataset to construct QSAR models using the multivariate linear regression ($QSAR_{MLR}$)) and artificial neural network approach (QSARANN) techniques, both of which are based on the OECD principles [4]. The research also designed 120 novel thiazole derivatives and used QSAR models to predict their activity. Subsequently, it evaluated the drug-likeness of these compounds and

eliminated any molecules deemed unsuitable. Lastly, promising targeted derivatives operating on the Polo-like kinases (PLK1) receptor are sought for using molecular docking approaches.

2. Methodology

2.1. Data set

This investigation began with an experimental data collection and was conducted on the thiazole derivatives with structural frameworks as seen in Figure 1. The initial step in the research process is data mining, and the first data set was gathered from published experimental results (Table 1). The main structural framework of these studies is thiazole, with substituent variations occurring in different locations along different IC_{50} (μM) values related to breast anti-cancer studies (MCF-7).

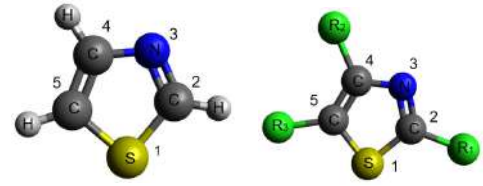


Figure 1. The Thiazole and its derivatives structure in the work

First, a sizable data set was gathered from esteemed publications. The data was then divided into data subsets using the Agglomerative Hierarchical Clustering (AHC) approach. QSAR models were built using data from two datasets: the one with 53 thiazole derivatives for training set and the other with 12 thiazole compounds for external validation (EV) with experimental IC_{50} (μM) values. The information is displayed in Table 1 and Table 5.

Table 1. The experimental $IC_{50,exp}$ values in training dataset for building the models

No	Thiazole derivatives			$IC_{50,exp}$ (μM)	Ref.	No.	Thiazole derivatives			$IC_{50,exp}$ (μM)	Ref.
	R_1	R_2	R_3				R_1	R_2	R_3		
1	$-C_{12}H_{12}N_3O_3S$	$-C_{11}H_{12}N_2O$	-H	10.03	[5]	28	$-C_{10}H_8F_3N_2O_2$	$-CF_3$	$-C_9H_6F_3N_2O$	7.69	[11]
2	$-C_{17}H_{13}N_2O_2S$	$-C_{11}H_{12}N_2O$	-H	26.08	[5]	29	$-C_{10}H_8F_3N_2O_2$	$-CF_3$	$-C_9H_6F_3N_2O$	3.7	[11]
3	$-C_{14}H_9N_2O_2S$	$-C_{11}H_{12}N_2O$	-H	22.35	[5]	30	$-C_{10}H_8F_3N_2O_2$	$-CF_3$	$-C_9H_6F_3N_2O$	6.5	[11]
4	$-C_{27}H_{16}C_{12}N_6O_2$	$-CH_4N$	$-CH_3$	41.37	[6]	31	$-C_{10}H_8F_3N_2O_2$	$-CF_3$	$-C_{10}H_5F_6N_2O$	6.08	[11]
5	$-CH_4N$	$-C_{27}H_{15}Cl_2N_6O_2$	$-CH_3$	41.37	[6]	32	$-C_9H_7F_2N_2O_2$	$-CF_3$	$-C_9H_6F_3N_2O$	8.21	[11]
6	$-C_{14}H_{11}N_5O_2S$	$-C_6H_4NO_2$	-H	57	[7]	33	$-C_9H_7F_2N_2O_2$	$-CF_3$	$-C_9H_6F_3N_2O$	5.93	[11]
7	$-C_{14}H_{11}N_5O_2S$	$-C_{10}H_8$	-H	59	[7]	34	$-C_9H_7F_2N_2O_2$	$-CF_3$	$-C_{10}H_5F_6N_2O$	2.61	[11]
8	$-C_{14}H_{11}N_5O_2S$	$-C_{12}H_{10}$	-H	45	[7]	35	$-C_9H_7F_2N_2O_2$	$-CF_3$	$C_9H_6F_3N_2O$	21.78	[11]
9	$-C_{21}H_{20}BrO_2P$	$-C_{13}H_{11}O$	-H	26.68	[8]	36	$-C_9H_7F_2N_2O_2$	$-CF_3$	$-C_{10}H_5F_6N_2O$	6.85	[11]
10	$-C_{21}H_{20}BrO_2P$	$-C_{14}H_{11}F_3O$	-H	12.49	[8]	37	$-C_9H_8FN_2O_2$	$-CF_3$	$-C_9H_6F_3N_2O$	12.23	[11]
11	$-C_{21}H_{20}BrO_2P$	$-C_{13}H_{10}FO$	-H	17.42	[8]	38	$-C_9H_8FN_2O_2$	$-CF_3$	$-C_9H_6F_3N_2O$	20.54	[11]
12	$-C_{21}H_{20}BrO_2P$	$-C_{13}H_{10}ClO$	-H	15.78	[8]	39	$-C_9H_8FN_2O_2$	$-CF_3$	$-C_9H_6F_3N_2O$	4.98	[11]
13	$-C_{21}H_{20}BrO_2P$	$-C_{13}H_{10}BrO$	-H	14.83	[8]	40	$-C_9H_8FN_2O_2$	$-CF_3$	$-C_{10}H_5F_6N_2O$	6.16	[11]
14	$-C_{21}H_{20}BrO_2P$	$-C_{17}H_{19}O$	-H	25.13	[8]	41	$-C_9H_7Cl_2N_2O_2$	$-CF_3$	$-C_9H_6F_3N_2O$	10.7	[11]
15	$-C_{14}H_{13}N_4O_2$	$-C_6H_5$	-H	25.75	[9]	42	$-C_9H_7Cl_2N_2O_2$	$-CF_3$	$-C_9H_6F_3N_2O$	8.42	[11]
16	$-C_{14}H_{13}N_4O_2$	$-C_6H_4Cl$	-H	22.84	[9]	43	$-C_9H_7Cl_2N_2O_2$	$-CF_3$	$-C_9H_6F_3N_2O$	6.24	[11]
17	$-C_{14}H_{13}N_4O_2$	=O	$-C_8H_8O$	12.68	[9]	44	$-C_9H_8N_3O_4$	$-CF_3$	$-C_9H_6F_3N_2O$	4.39	[11]
18	$-C_{14}H_{13}N_4O_2$	=O	$-C_3H_5O_2$	5.36	[9]	45	$-C_9H_8N_3O_4$	$-CF_3$	$-C_9H_6F_3N_2O$	4.73	[11]
19	$-C_{14}H_{13}N_4O_2$	$-CH_3$	$-C_5H_9NO$	27.63	[9]	46	$-C_9H_7F_2N_2O_2$	$-CF_3$	$-C_8H_6N_3O$	4.81	[11]
20	$-C_{14}H_{13}N_4O_2$	$-CH_3$	$-C_3H_3N_2$	5.84	[9]	47	$-C_9H_7Cl_2N_2O_2$	$-CF_3$	$-C_8H_6N_3O$	6.74	[11]
21	$-C_{13}H_8ClN_4O_2S$	$-C_6H_4Cl$	-H	3.3	[10]	48	$-C_9H_7Cl_2N_2O_2$	$-CF_3$	$-C_8H_6N_3O$	7.51	[11]
22	$-C_{14}H_{11}N_4O_3S$	$-C_6H_4Cl$	-H	2.4	[10]	49	$-C_9H_9N_2O_2$	$-CF_3$	$-C_8H_6N_3O$	15.4	[11]
23	$-C_{13}H_9N_4O_2S$	$-C_7H_7O$	-H	4.9	[10]	50	$-C_9H_7F_2N_2O_2$	$-CH_3$	$-C_9H_6F_3N_2O$	28.15	[11]
24	$-C_{13}H_8ClN_4O_2S$	$-C_7H_7O$	-H	1.9	[10]	51	$-C_9H_7F_2N_2O_2$	$-CH_3$	$-C_{10}H_5F_6N_2O$	11.63	[11]
25	$-C_{14}H_{11}N_4O_2S$	$-C_7H_7O$	-H	4.9	[10]	52	$-C_9H_8N_3O_4$	$-CF_3$	$-C_8H_6N_3O$	21.23	[11]
26	$-C_{14}H_{11}N_4O_3S$	$-C_7H_7O$	-H	1.3	[10]	53	$-C_9H_7F_2N_2O_2$	$-CH_3$	$-C_9H_6F_3N_2O$	7.34	[11]
27	$-C_{10}H_8F_3N_2O_2$	$-C_8H_6Cl_2N_2O$	$-CH_3$	24.52	[11]						

2.2. Descriptors

Using ChemDraw Pro13, the thiazole derivatives were arranged structurally. Afterward, they underwent structural optimization through quantum mechanics, utilizing the semi-experimental method PM7 by using the MOPAC tool [12]. Quantum parameters, including energy, HOMO, LOMO, partial charge, heat of formation, and so on, were computed. Next, Test software is used to generate the 0-3D molecular descriptors.

In the end, the complete set of chemical descriptors needed to construct QSAR models at the next stage was acquired by combining all of the data and eliminating irrelevant factors.

2.3. Estimation of QSAR models

The study used the multiple linear regression method ($QSAR_{MLR}$) and the artificial neural network method (QSARANN) to construct the QSAR model. Which, one of the most basic and often used modeling techniques for QSAR regression models is multiple linear regression (MLR). By creating a linear equation for observational data, MLR aims to create a correlation model between two or more independent variables (descriptors) and a dependent variable (bioactivity) [13]. Every value of an independent variable has a corresponding value for a dependent variable. The regression add-in Excel tool was utilized in the study to create a linear regression model [13]. The choice of variables is the most challenging stage in the construction of this model. The XLSTAT2016 package and the modeling were the foundations for the variable selection in this study.

In the meanwhile, the artificial neural network (ANN) is a paradigm for information processing that emulates the information-processing mechanism of biological neuron systems [14]. The input layer, the output layer, and the hidden layer comprise the three layers of an artificial neural network. As a result, the I(k)-HL(m)-O(n) architecture of a neural network is made up of the input layer I(k), the hidden layer HL(m), and the output layer O(n). In the study, the log-sig and tan-sig transfer functions were utilized, and the Levenberg-Marquardt back-propagation technique was employed to train the artificial neural network models. The input nodes used in the construction of the models are descriptor variables of the MLR models. In this instance, the training dataset was split into three subsets: the training subset (70%), the cross-validation (CV) subset (20%), and the test subset (10%). Matlab2019a was the program used to create the ANN model, along with the 'nntools' function. In addition, this study evaluated the external model using a separate, external dataset.

2.4. Design and selection of new thiazole derivatives

The three places R1, R2, and R3 on the structure of thiazole derivatives (Figure 1), where various substituents were replaced to create the novel compounds, were referenced to the findings of experimental investigations. Allopurinol, vanillin, iso-coumarin, 1,4-dihydropyridine, 1,2,4-triazole, 6-fluorobenzo, and phthalimide derivatives are the attached substituents. The chemical groups were chosen from esteemed papers and demonstrated their prospective use; in particular, they exhibited a variety of biological activities, including antibacterial, antifungal, and anti-cancer properties. The Design Expert tool was utilized to combine the seven substituents and H atom in three spots once the substituent was chosen. Using the constructed QSAR models, the anti-breast cancer activity of the new design thiazole derivatives was predicted by IC_{50} values.

2.5. Evaluation of pharmacokinetics properties and molecular docking

After utilizing QSAR models to estimate the anti-breast cancer activity of the compounds, pharmacokinetic characteristics (drug-likeness) were assessed to identify derivatives with potential use in drug development. When determining if a given molecule is a pharmacological variation, drug-likeness may be seen as a complicated balance between several molecular and structural properties. The work used the Lipinski-5 and Ghose rules [15,16] for this aim. The criteria such as molecular weight (MW<500 Da), high lipophilicity (LogP<5), hydrogen bond donors (HBD<5), and hydrogen bond acceptors (HAD<10) are the factors that determine drug-likeness according to the Lipinski-5 rule [17]. Meanwhile, the criteria of Ghose rule are 160 ≤ MW ≤ 480; -0.4 ≤ WLOGP (lipophilicity) ≤ 5.6; 40 ≤ Molar refractivity (MR) ≤ 130; and 20 ≤ Number of atoms (NAtoms) ≤ 70.

Once the drug-likeness of the novel derivatives was established, they carried out molecular docking using the primary PLK1protease (PDB-3FC2) to confirm that the derivatives could form a bond with the active site using the binding energy (E-binding, kcal.mol-1) and the root mean square deviation (RMSD, Å) values. Undoubtedly, a molecular modelling method called docking is employed to forecast the interactions between ligands and proteins, or enzymes [17]. Using the MOE package and the subsequent three phases, the docking process was simulated: (1) Docking preparation; (2) protein structure and ligand preparation; and (3) analysis of the docking data. If the RMSD value is less than 1.5 Å and the binding energy (E-binding) is less than -7.0 kcal.mol-1, the procedure is considered successful. Additionally, the interaction between docking compounds is used to evaluate their binding.

3. Results and discussion

3.1. $QSAR_{MLR}$ model

Using the Regression programme [13], the $QSAR_{MLR}$ models were constructed using a dataset of 53 IC_{50} experimental values of thiazole derivatives (Table 1) together with a matching structural and quantum descriptor. Two subsets of this training dataset were created: the test subset (20% of chemicals) and the modelling subset (80% of substances). The variables of models are assessed using the P-value (<0.05) and coefficient of determination R^2 (>0.6) [4,13,18]. To regulate the model and make it acceptable, significant changes are made to the indicators used for model assessment, such as R^2_{train}, R^2_{adj}, Q^2_{LOO}, and SE. Consequently, the MLR models with k of 1 to 9 and statistical characteristics listed in Table 2 were constructed.

The values of R^2_{train}, R^2_{adj}, Q^2_{LOO} are proportionate to the number of variables, as Table 3 and Figure 2 demonstrate. This demonstrates that the model improves with an increase in the number of variables accompanied by a notable rise in variables from 8 to 9, even if the previous increases in variables did not significantly alter the situation. Consequently, there are eight variables, suggesting that this is the most promising $QSAR_{MLR}$ model. The eight variables in the model are as follows: xvch5, BELp3, SssO, HOMO, SssNH, BEHp3, BEHv2, and BELm1. Here, xvch5 is valence 5th order chain chi

index and it is the one type of the valence connectivity indices, HOMO is an abbreviation of the phrase “Highest Occupied Molecular Orbital”. It describes the molecular orbital that has the most energy when a molecule's electrons are in their ground state. Reactivity, stability, and bonding are only a few of the features of chemicals in which HOMO is essential. SsNH and AaaO are atom type E-State sum non-hydrogen indices, SssO is sum of (-O-) E-States and SssNH is sum of (-NH-) E-States. The E-State value for a given non hydrogen atom i in a molecule is given by its intrinsic state plus the sum of the perturbations on that atom by all the other atoms in the molecule. The Burden eigenvalue descriptors are determined by solving the following general eigenvalue equation. In which, BELp3, BEHp3, BEHv2 and BEHv2 are the Burden eigenvalue descriptors, BELp3 is lowest eigenvalue n. 3 of Burden matrix and weighted by atomic polarizabilities, BEHp3 is highest eigenvalue n. 3 of Burden matrix and weighted by atomic polarizabilities, BEHv2 is highest eigenvalue n. 2 of Burden matrix and weighted by atomic van der Waals volumes, BELm1 is lowest eigenvalue n. 1 of Burden matrix and weighted by atomic masses. According to the $QSAR_{MLR}$ model findings for $R^2 = 0.875$ as Fig. 2, it showed that 84.9 % of the biological activity variables in the data set are encoded by the model. 82.6 % of the active value variable in the data is encoded by the values $R^2_{adj} = 0.853$ and $Q^2_{LOO} = 0.834$ (Figure 2). This discovery leads to the model's quite accurate prediction results.

Table 2. Selected models $QSAR_{MLR}$ (k of 1 to 9) and statistical values

k	Variables	*SE*	R^2_{train}	R^2_{adj}	Q^2_{LOO}	F_{stat}	*PRESS*
1	x_1	10.900	0.376	0.364	0.278	30.740	7019.64
2	$x_1 \mid x_2$	9.673	0.519	0.499	0.402	26.941	5812.71
3	$x_1\|x_2\|x_3$	8.624	0.625	0.602	0.515	27.225	4717.27
4	$x_1\|x_2\|x_3\|x_4$	7.643	0.712	0.687	0.599	29.597	3898.65
5	$x_1\|x_2\|x_3\|x_4\|x_5$	6.526	0.794	0.772	0.718	36.248	2738.33
6	$x_1\|x_2\|x_3\|x_4\|x_5\|x_6$	6.026	0.828	0.806	0.764	36.949	2294.75
7	$x_1\|x_2\|x_3\|x_4\|x_5\|x_6\|x_7$	5.735	0.848	0.824	0.775	35.785	2186.43
8	$\mathbf{x_1\|x_2\|x_3\|x_4\|x_5\|x_6\|x_7\|x_8}$	**5.248**	**0.875**	**0.853**	**0.834**	**38.607**	**1608.88**
9	$x_1\|x_2\|x_3\|x_4\|x_5\|x_6\|x_7\|x_8\|x_9$	5.048	0.887	0.864	0.851	37.601	1446.02

Notation of molecular descriptors

xvch5	x_1	HOMO	x_4	BEHv2	x_7
BELp3	x_2	SssNH	x_5	BELm1	x_8
SssO	x_3	BEHp3	x_6	MDEC22	x_9

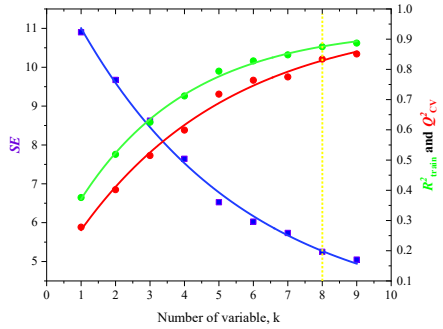


Figure 2. Change of values RMSE, R2train and Q2LOO according to k descriptors

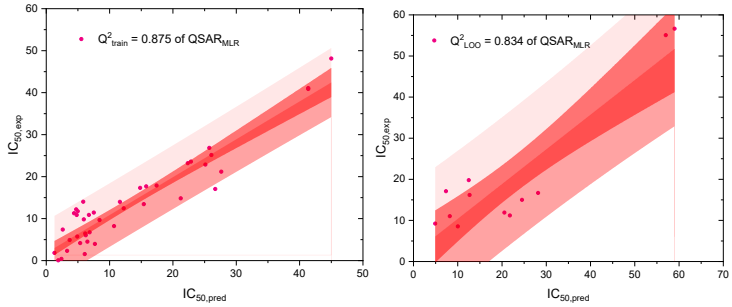


Figure 3. Correlation of experimental vs. predicted values IC_{50} of the training compounds using the $QSAR_{MLR}$ model (with k = 8)

In addition, with statistical values like R^2_{train} = 0.875 and Q^2_{LOO} = 0.834, the regression equation demonstrated a strong association, according to the whole the result of $QSAR_{MLR}$ model (Table 2). A model with an R^2_{adj} value of 0.853 is said to encapsulate 85.3 % of the activity values in the dataset; random errors and off-model factors account for the remaining 5.7 % of values. The model attained 95% dependability, as evidenced by the p-values, which were also considerably less than 0.05.

To observe the contribution of variables, the $GMPx_i$ quantity is used [19]. It is the average value of the MPx_j values [19] of three models of models neighboring the male model selected in the study. The full results of these values are presented in Table 3, it shows that the contribution of the BEHv2 (x_7) is the largest corresponding to a value of 39.6 %, and then the following variables, in the order listed below, also contribute to the $QSAR_{MLR}$ model: BEHv2 > BELm1 > BELp3 > HOMO > xvch5 > SssNH > SssO > MDEC22.

Table 3. The values of contribution ($MPx_{k,i}$ and $GMPx_i$) in $QSAR_{MLR}$ models with k of 7 to 9

Statistics and variables	$QSAR_{MLR}$			$MPx_{k,i}$, %			$GMPx_i$, %
	$k = 7$	$k = 8$	$k = 9$	$k = 7$	$k = 8$	$k = 9$	
R^2_{train}	0.848	0.875	0.887	–	–	–	–
R^2_{adj}	0.824	0.853	0.864	–	–	–	–
Q^2_{LOO}	0.775	0.834	0.851	–	–	–	–
SE	5.735	5.248	5.048	–	–	–	–
constant	-71.339	68.219	4.077	–	–	–	–
x_1	117.642	90.116	81.40	1.364	0.700	0.601	0.889
x_2	83.847	78.208	86.89	20.075	12.544	13.235	15.285

Statistics and variables	$QSAR_{MLR}$			$MPx_{k,i}$, %			$GMPx_i$, %
	$k = 7$	$k = 8$	$k = 9$	$k = 7$	$k = 8$	$k = 9$	
x_3	-0.822	-1.371	-1.222	0.335	0.371	0.314	0.340
x_4	8.483	10.115	11.15	9.919	7.925	8.297	8.714
x_5	3.072	1.914	1.633	1.561	0.653	0.529	0.914
x_6	64.879	25.896	27.82	32.799	8.770	8.948	16.839
x_7	-66.233	-128.397	-124.79	33.948	44.085	40.693	39.575
x_8	–	141.575	162.27	–	24.951	27.160	17.371
x_9	–	–	-0.224	–	–	0.222	0.074

3.2. $QSAR_{ANN}$ model

The network architecture of the artificial neural network model was I(8)-HL(m)-I(1) since it was created utilising MLP networks to build the variables of the $QSAR_{MLR}$ model. First, the network design that satisfies statistical indicators, including identifying the hidden layer with the m value, was surveyed using the training dataset, which consists of 53 empirical IC_{50} values (Table 1). This was the first phase in the building process. Table 3 displays the findings of the survey.

Table 4. The results of the first step for discovering QSPRANN model I(8)-HL(m)-O(1) with statistical parameters

Code	$QSAR_{ANN}$ model	R^2_{train}	Q^2_{test}	Q^2_{CV}	Training error	Test Error	Validation Error	Transfer Function
ANN1	I(8)-HL(8)-O(1)	0.930	0.892	0.931	0.731	0.229	0.964	hyperbolic tangent
ANN2	I(8)-HL(11)-O(1)	0.931	0.946	0.948	0.648	0.009	0.454	log-sigmoid
ANN3	I(8)-HL(12)-O(1)	0.923	0.881	0.928	0.658	0.718	0.848	hyperbolic tangent
ANN4	**I(8)-HL(9)-O(1)**	0.918	0.934	0.916	0.665	0.857	0.750	**hyperbolic tangent**
ANN5	I(8)-HL(10)-O(1)	0.920	0.962	0.918	0.688	0.962	0.177	log-sigmoid

Next, using the Q^2_{ex} and *MARE(%)* indicators, an external assessment set with 12 empirical IC_{50} values (Table 4) was used to find the optimal ANN. The findings indicated that, with $Q^2_{ex} = 0.921$ (>0.5) and *MARE(%)* = 19.767 (Figure 5), the ANN4 model (bold in Table 4) had the best prediction.

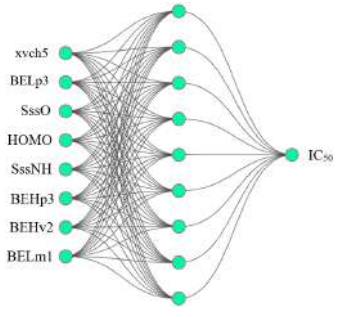


Figure 4. Architecture of neural network I(8)-HL(9)-O(1) model

Consequently, the I(8)-HL(9)-I(1) architecture of the $QSAR_{ANN}$ network was created, with a hyperbolic tangent transfer function. The statistical parameters of the network are as follows: $R^2_{train} = 0.918$, $Q^2_{test} = 0.934$, and $Q^2_{CV} = 0.916$. New thiazole compounds were created using the ANN4 model, which took into account input variables including xvch5, BELp3, SssO, HOMO, SssNH, BEHp3, BEHv2, and BELm1. In addition, the *MARE(%)* value (10.142) of the $QSAR_{ANN4}$ model is lower than one (28.310) of the MLR model, indicating that the $IC_{50,pred}$ values of the ANN4 model are much closer to experimental values than those of the MLR model.

3.3. External validation

External evaluation (EV) is regarded as an assessment meant to confirm the capacity for the prediction of the constructed models. Choose the model that produces forecast results that are most similar to the experiment after that. Twelve chemicals from trials are part of the external assessment data set, which is separate and distinct from the collection used to create the QSAR model. Table 5 provides comprehensive details on these variants.

Table 5. The EV data set and the predicted IC_{50},pred values from the QSAR models

No.	Thiazole derivatives			$IC_{50,exp}$	$IC_{50,pred}$		ref.
	R_1	R_2	R_3		MLR	ANN4	
1	$-C_{12}H_{12}N_3O_2S$	$-C_{11}H_{11}N_2O$	-H	15.880	14.793	13.269	[5]
2	$-C_{16}H_{15}N_2O_2$	$-C_6H_4Br$	-H	20.140	20.381	22.627	[20]
3	$-CH_4N$	$-C_{13}H_6Cl_2N_3OS$	$-CH_3$	32.600	22.175	41.647	[6]
4	$-CH_4N$	$-C_{22}H_{12}Cl_2N_5O_2$	$-CH_3$	34.190	28.183	37.917	[6]
5	$-C_8H_4N_2O_2$	$-CH_3$	$-C_4H_8N_3S$	63.000	35.348	58.364	[7]
6	$-CH_4N$	$-C_{13}H_6Cl_2N_3OS$	$-CH_3$	32.600	22.417	41.071	[6]
7	$-CH_4N$	$-C_{22}H_{12}Cl_2N_5O_2$	$-CH_3$	34.190	28.142	37.747	[6]
8	$-C_6H_3Cl_2N$	$-C_6H_6NO_2S$	$-C_6H_4Cl$	46.720	22.174	40.643	[21]
9	$-C_6H_4BrN$	$-C_6H_6NO_2S$	$-C_6H_4Br$	40.210	22.725	46.559	[21]
10	$-C_8H_9NO$	$-C_8H_6ClN_2$	-H	4.285	3.798	2.991	[22]
11	$-C_6H_6N_3$	$-C_6H_5$	-H	11.000	12.311	5.786	[23]
12	$-C_6H_6N_3$	$-C_6H_4Cl$	-H	4.700	6.484	5.621	[23]

As previously indicated, the MLR and ANN models are constructed using the external evaluation data set. Additionally, as Table 4 illustrates, it is employed to choose the optimal ANN model among the models that were first surveyed. The two parameters that serve as the foundation for the assessment are *MARE(%)* [19] and Q^2EX (>0.5) [18] and Figure 5 presents the results in full. The findings show that the linear regression models satisfy the requirements, and the fourth neural network model ($QSAR_{ANN4}$) is chosen for the project since its *MARE(%)* value is acceptable when compared to the other models and its Q^2EX value of 0.921 is the highest.

Furthermore, the value of Q^2EX of the $QSAR_{MLR}$ model between the observed IC_{50},exp values and the anticipated IC_{50},pred values is displayed in Figure 5 and yields a result of 0.829 and comparable to the $QSAR_{ANN4}$ yields a value of 0.921. Given that two of the aforementioned models produce strong correlation index findings for the external evaluation set, it can be concluded that the evaluation ability is dependable and predictive for a variety of design materials

and the ANN4 model has better predictive ability.

Additionally, when the findings F = 0.0012 < $F_{0.05}$ = 4.3001, a one-way ANOVA analysis of variance revealed that the difference between the experimental and predicted values from the two models - $QSAR_{MLR}$ and $QSAR_{ANN4}$ - was not significant. Thus, the capacity for prediction of the two models is appropriate.

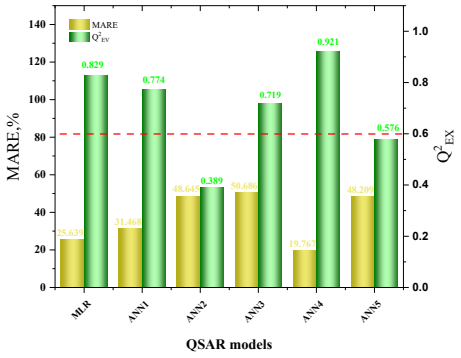


Figure 5. The Q^2EX and *MARE(%)* values of QSAR models

3.4. Design and predict of the new thiazole derivatives

The XLSTAT programme was utilised to screen 120 novel thiazole compounds that were generated for the study in order to forecast the IC_{50} and new values using Cook's D values [4]. Substances outside of the value domain |Cook's D > 1| will be regarded as outliers and deleted, whereas derivatives inside the domain of the training set with the value |Cook's D < 1| will be projected. Consequently, the 10 new derivatives (Figure 6) satisfied the assessment requirements for AD and Outliers, and both the constructed MLR and ANN models will forecast them at the $IC_{50,new}$ Values.

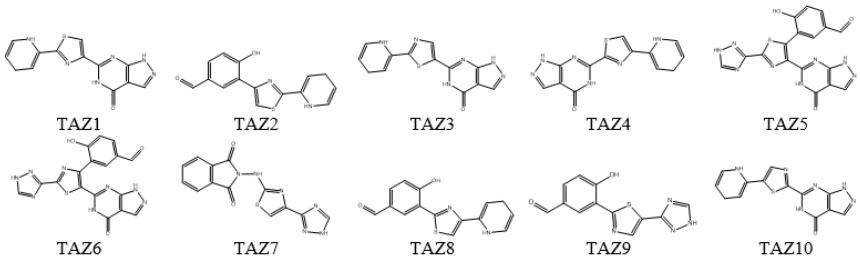


Figure 6. The structures of 10 novel thiazole derivatives

The all of 10 novel derivatives had calculated the $IC_{50,new}$ predictive activity values in both models, according to the prediction results displayed in Figure 7. It was also possible to confirm that the predicted $IC_{50,new}$ values of the 10 new design derivatives will be almost identical to those of subsequent experiments and application as an anti-breast cancer agent (MCF-7) because the built models have confirmed this.

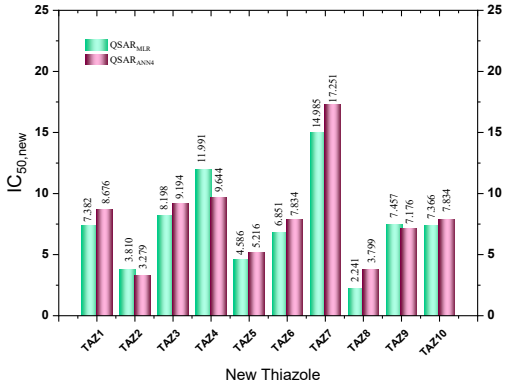


Figure 7. The $IC_{50,new}$ predicted values of 10 novel thiazole from the QSAR models

Findings of two QSAR models, including $QSAR_{MLR}$, and $QSAR_{ANN6}$, that forecast

novel compounds. Analysing variance for findings F = 0.0859 < F0.05 = 4.4138 shows no discernible difference. As a result, two models have dependable and consistent forecasting abilities.

3.4. Drug-likeness and molecular docking

Before being used in pharmacy, the novel compounds need to be assessed for pharmacokinetics properties (drug-likeness). As previously stated, the SwissADME online platform (http://www.swissadme.ch/) was utilized to assess derivatives based on the standards of the Lipinski and Ghose rules [15,16]. Consequently, the results from Table 7 show that every one of the ten compounds met the requirements of Lipinski and Ghose rules.

In this work, ten novel compounds with anti-breast cancer activity (MCF-7) that matched the drug-likeness criteria and had IC_{50} values were determined. To ascertain whether there may be any interaction with PLK1, this investigation is being expanded. The study employed docking simulation methods on their proteins to achieve this.

Table 6. The drug-likeness results by using Lipinski and Ghose rules

Code	Lipinski Rule				Ghose Rule			Drug-likeness
	MW	LogP	HBD	HBA	WLOGP	MR	NAtoms	
TAZ1	298.32	1.33	3	4	1.24	84.26	35	Yes
TAZ2	284.33	2.51	2	3	2.80	83.60	31	Yes
TAZ3	298.32	1.39	3	4	1.24	84.26	35	Yes
TAZ4	298.32	1.31	3	4	1.24	84.26	35	Yes
TAZ5	406.38	0.53	4	8	1.74	103.84	54	Yes
TAZ6	406.38	0.50	4	8	1.74	103.84	54	Yes
TAZ7	312.31	0.68	2	5	0.98	81.79	38	Yes
TAZ8	284.33	2.38	2	3	2.80	83.60	31	Yes
TAZ9	272.28	1.10	2	5	2.11	70.34	35	Yes
TAZ10	298.32	1.46	3	4	1.24	84.26	35	Yes

The RMSD and the binding energy data (E-binding, kcal.mol-1) are used to assess the docking outcomes, as was indicated in the methods section. When the binding energy is as negative as feasible, the RMSD is less than 1.5 Å and the binding energy value is less than 7.0 kcal.mol-1. According to the docking data of Figure 8, we took into consideration the docking point along with the IC_{50} value, and the results showed that there are many thiazole derivatives with RMSD values less than 1.5 Å, but to ensure the standard of energy value less than 7.0 kcal.mol-1, only the TAZ5 derivative meets this standard. Therefore, the docking results can confirm that the TAZ5 derivative is considered the most effective derivative to fight breast cancer.

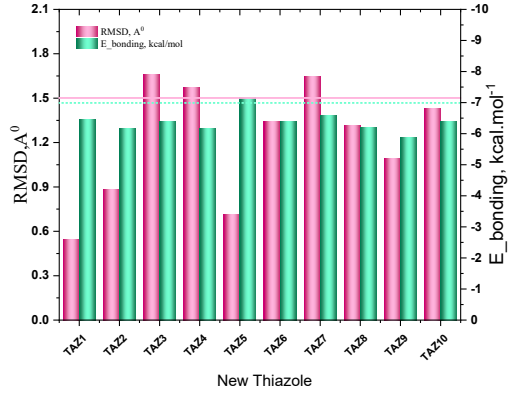
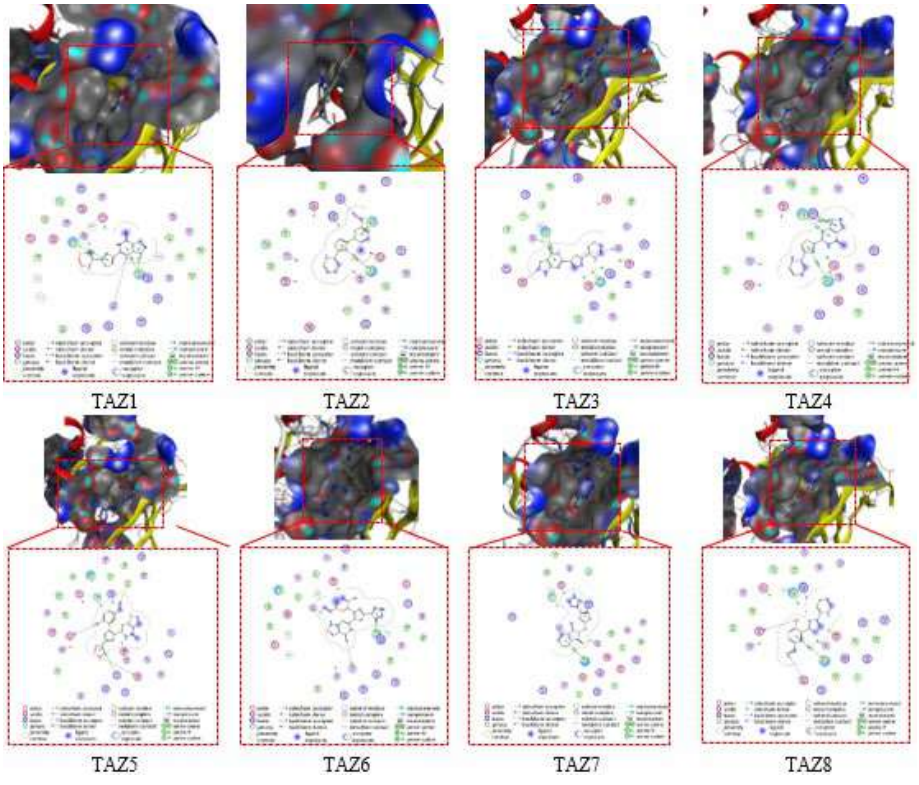


Figure 8. Docking results of 10 novel thiazole on PDB-3FC2

Table 7. Interactions between ligands and receptors in the docking model of the potential TAZ5 thiazole

Ligand	Receptor		Interaction	Distance,	E (kcal.mol^{-1})
5-ring	N	PHE 195 (A)	pi-H	4.16	-1.60
N18	OE1	GLU 101 (A)	H-donor	2.98	-8.20
N20	NZ	LYS 82 (A)	H-acceptor	2.80	-9.10
O38	O	GLU 131 (A)	H-donor	2.85	-2.90
6-ring	CD2	LUE 59 (A)	pi-H	3.72	-0.60
6-ring	CG	ARG 136 (A)	pi-H	3.62	-0.70
C38	OE1	GLU 69 (A)	H-donor	3.56	-0.70
N28	N	CYS 133 (A)	H-acceptor	2.81	-5.20
C1	OE1	GLU 140 (A)	H-donor	3.36	-0.60
O2	OE2	GLU 140 (A)	H-donor	2.45	-2.30
O3	O	GLY 180 (A)	H-donor	2.67	-2.0

The amino acids such as Glu, Cys, and Lys are among the primary binding receptors for the putative TAZ5 thiazole, and N donors are the primary target of bonding, according to the docking model displayed in Figure 9 and the information provided in Table 7. Additionally, 5- and 6-rings of this derivative interact with the Phe, Lue, and Arg amino acids receptors in suitably powerful ways. Furthermore, the ligand that makes hydrogen bonds with enzymes interacts with hydrophobic amino acids through pi-H, H-donor, and H-acceptor interactions.



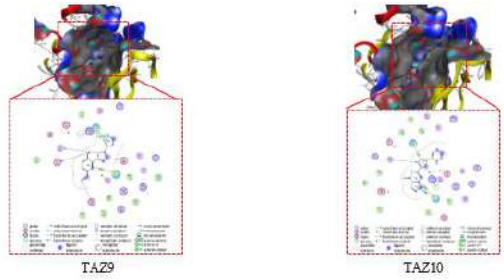


Figure 9. Docking simulation of 10 new thiazole derivatives

4. Conclusion

Quantitative structure-activity relationship models (QSAR) have been effectively constructed via this work by combining machine learning approaches with both linear and nonlinear regression methods. Using experimental datasets containing descriptive variables derived from PM7 semi-empirical quantization, QSAR models were built. Based on statistical measures with high reliability, such as R^2_{train}, Q^2_{LOO}, Q^2_{ex}, SE, and MARE(%), construction QSAR models have been thoroughly assessed. The two models, $QSAR_{MLR}$ and $QSAR_{ANN}$ I(8)-HL(9)-O(1), thereby satisfied the criteria for practical application. The IC_{50} value results of the two models are used to forecast the newly created derivatives of breast anti-cancer activity (MCF-7), and they used Lipinski and Ghose methods to identify the derivatives with drug-like qualities. Additionally, the capacity of the novel design variations to attach to the proteins of the Polo-like kinases (PLK1) virus strain was investigated using molecular docking modelling. Following rigorous screening, the TAZ5 thiazole has been identified as a possible derivative of the PLK1 receptor (PDB-3FC2). The outcomes of this work enable the experimental synthesis of thiazole derivatives with good activity for pharmacological areas to be predicted and directed.

References

[1] I. P. Singh, S. Gupta, and S. Kumar, “Thiazole Compounds as Antiviral Agents: An Update,” Medicinal Chemistry, Vol. 16(1), pp. 4-23, 2020.

[2] E. Güven, “Gene expression analysis of MCF-7 cell lines of breast cancer treated with herbal extract of Cissampelos pareira revealed association with viral diseases”, Gene Reports, Vol. 23, no. 101169, 2021.

[3] M. Sanhaji, N.N. Kreis, B. Zimmer, T. Berg, F. Louwen, and J. Yuan, “p53 is not directly relevant to the response of polo-like kinase inhibitors”, Cell Cycle Vol. 11(3), pp. 543-553, 2012.

[4] R. Kunal, K. Supratik, and N. D. Rudra, “A Primer on QSAR/QSPR Modeling”, 2022.

[5] K. M. Dawood, Taha M. A. Eldebss, S.A. Heba, El-Zahabi, M. H. Yousef, and P. Metz, “Synthesis of some new pyrazole-based 1,3-thiazoles and 1,3,4-thiadiazoles as anticancer agents”, European Journal of Medicinal Chemistry, Vol. 70, 2013.

[6] T. A. Farghaly, G. S. Masaret, Z. A. Muhammad, and M. F. Harras, “Discovery of thiazole-based-chalcones and 4-hetarylthiazoles as potent anticancer agents: Synthesis, docking study and anticancer activity” Bioorganic Chemistry, Vol. 98, 2020.

[7] A. R. Oliveira, F. A. Dos Santos, L. P. De Lima Ferreira, M. G. Da Rocha Pitta, M. V. De Oliveira Silva, M. V. De Oliveira Cardoso, A. F. Pinto, P. Marchand, M. J. Barreto de Melo Rego, and A. Cristina Lima Leite, “Synthesis, anticancer activity and mechanism of action of new phthalimido-1,3-thiazole derivatives”,

Chemico-Biological Interactions, Vol. 347, 2021.

[8] X. Dang, S. Lei, S. Luo, Y. Hu, J. Wang, D. Zhang, D. Lu, F. Jiang, and L. Fu, "Design, synthesis and biological evaluation of novel thiazole-derivatives as mitochondrial targeting inhibitors of cancer cells", Bioorganic Chemistry, Vol. 114, 2021.

[9] A. M. Alqahtani, and A. A. Bayazeed, "Synthesis and antiproliferative activity studies of new functionalized pyridine linked thiazole derivatives", Arabian Journal of Chemistry, Vol. 14, 2020.

[10] S. A. Abdel-Aziz, E. S. Taher, P. Lan, G. F. Asaad, H. A. M. Gomaa, N. A. El-Koussi, and B. G. M. Youssif, "Design, synthesis, and biological evaluation of new pyrimidine-5-carbonitrile derivatives bearing 1,3-thiazole moiety as novel anti-inflammatory EGFR inhibitors with cardiac safety profile", Bioorganic Chemistry Vol. 111, 2021.

[11] H. He, X. Wang, L. Shi, W. Yin, Z. Yang, H. He, and Y. Liang, "Synthesis, antitumor activity and mechanism of action of novel 1,3-thiazole derivatives containing hydrazide-hydrazone and carboxamide moiety", Bioorganic & Medicinal Chemistry Letters, Vol. 26, 2016.

[12] J. J. P Stewart, "Optimization of parameters for semi-empirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters", J Mol Model, Vol. 19, pp. 1-32, 2013.

[13] D. D. Steppan, Joachim Werner, and P. Robert Yeater, "Essential Regression and Experimental Design for Chemists and Engineers', May 23, 2022.

[14] J. Gasteiger, and J. Zupan, "Neural networks in chemistry", Chiw Inr Ed EngI., Vol. 32, pp. 503-521, 1993.

[15] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," Advanced Drug Delivery Reviews, Vol. 23(1), pp. 3-25, 1997.

[16] A. K. Ghose, V. N. Viswanadhan, and J. J. A. Wendoloski, "A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases", J. Comb. Chem.,Vol. 1, pp. 55-68, 1999.

[17] A. H. Santoyo, A.Yair, V. Altuzar, H. V. Cid, and C. M. Barrer, "Protein-Protein and Protein-Ligand Docking", In T. Ogawa (Ed.), Protein Engineering-Technology and Application, InTech, 2013.

[18] A. Golbraikh, and A. Tropsha, "Beware of q2!", J. Mol Graph Model, Vol. 20(4), pp. 269-76, 2002.

[19] P. V. Tat, "Development of QSAR and QSPR", Publisher of Natural sciences and Technique, Ha Noi, 2009.

[20] Zhong-Chang Wang, Fa-Qian Shen, Meng-Ru Yang, Ling-Xia You, Li-Zhi Chen, Hai-Liang Zhu, Ya-Dong Lu, Fan-Lei Kong, and Ming-Hua Wang, "Dihydropyrazothiazole derivatives as potential MMP-2/MMP-8 inhibitors for cancer therapy", Bioorganic & Medicinal Chemistry Letters, Vol. 28, 2018.

[21] S. Mirza, S. Asma Naqvi, K. Mohammed Khan, U. Salar, M. Iqbal Choudhary, and Facile, "Synthesis of Novel Substituted Aryl-Thiazole (SAT) Analogs via One-Pot Multi-component Reaction as Potent Cytotoxic Agents against Cancer Cell lines", Bioorganic Chemistry, Vol. 70, 2016.

[22] M. F. Baig, V. Lakshma Nayak, P. Budaganaboyina, K. Mullagiri, S. Sunkari, J. Gour, and A. Kamal, "Synthesis and biological evaluation of imidazo[2,1-b]thiazolebenzimidazole conjugates as microtubule-targeting agents", Bioorganic Chemistry, Vol. 77, 2018.

[23] T. David dos Santos Silva, L. M. Bomfim, A. C. B. Da Cruz Rodrigues, R. Borges Dias, C. B. Schlaepfer Sales, C. A. G. Rocha, M. B. Pereira Soares, D. P. Bezerra, M. V. De Oliveira Cardoso, A. Cristina Lima Leite, and G. C. Gadelha Militão, "Anti-liver cancer activity in vitro and in vivo induced by 2-pyridyl 2,3-thiazole derivatives:, Toxicology and Applied Pharmacology, Vol. 329, 2017.

Ngày nhận bài: 06/11/2023

Ngày hoàn thành sửa bài: 10/12/2023

Ngày chấp nhận đăng: 13/12/2023