



# NOVEL SARS-CoV-2 INHIBITORS FROM PHENETHYLTHIAZOLETHIOUREA DERIVATIVES USING HYBRID QSAR MODELS AND DOCKING SIMULATION

Pham Van Tat, Tran Thai Hoa, Au Vo Ky & Pham Nu Ngoc Han

To cite this article: Pham Van Tat, Tran Thai Hoa, Au Vo Ky & Pham Nu Ngoc Han (2021): NOVEL SARS-CoV-2 INHIBITORS FROM PHENETHYLTHIAZOLETHIOUREA DERIVATIVES USING HYBRID QSAR MODELS AND DOCKING SIMULATION, Smart Science, DOI: [10.1080/23080477.2021.1914967](https://doi.org/10.1080/23080477.2021.1914967)

To link to this article: <https://doi.org/10.1080/23080477.2021.1914967>



Published online: 05 May 2021.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



## ARTICLE



# NOVEL SARS-CoV-2 INHIBITORS FROM PHENETHYLTHIAZOLETHIOUREA DERIVATIVES USING HYBRID QSAR MODELS AND DOCKING SIMULATION

Pham Van Tat<sup>a</sup>, Tran Thai Hoa<sup>b</sup>, Au Vo Ky<sup>c</sup> and Pham Nu Ngoc Han<sup>d</sup>

<sup>a</sup>Department of Pharmacy, Faculty of Health Science, Hoa Sen University, Viet Nam; <sup>b</sup>Department of Chemistry, Hue University of Sciences, Hue University, Hue City, Viet Nam; <sup>c</sup>Franklin High School, Elk Grove, USA; <sup>d</sup>Department of Food Technology, Hoa Sen University, Ho Chi Minh City, Viet Nam

**ABSTRACT**

Currently, there are several groups of HIV-1 virus inhibitors that could potentially be used in the treatment of SARS-CoV-2. Particularly, the phenethylthiazolethiourea compounds are capable of inhibiting HIV-1 RT and have been tested by IC<sub>50</sub>. This work contributed to the search for SARS-CoV-2 inhibitors; a group of these compounds was developed to obtain SARS-CoV-2 inhibitors. The hybrid QSAR<sub>GA-ANN</sub> model with I(5)-HL(9)-O(1) architecture used for developing for HIV-1 inhibitors and it successfully predicted the pIC<sub>50</sub> activities of six newly designed compounds. The predicted results of the pIC<sub>50</sub> activity received from the QSAR<sub>GA-ANN</sub> model agreed well with the docking simulation. The C-n6 new molecule that has been bound to the SARS-CoV-2 protein receptors (PDB ID: 6LU7) using docking simulation. It demonstrated a more effective activity against HIV-1 (PDB ID: 1ODW). This compound C-n6 exhibited the binding affinity for the HIV-1 protein (1ODW) is -23.6137 kJ.mol<sup>-1</sup>; for the target protein SARS-CoV-2 (6LU7), its binding affinity is -27.4235 kJ.mol<sup>-1</sup>. The retrosynthesis plan for the most active substance C-n6 1-(2-chloro-5-hydroxy-4-nitrophenethyl)-3- (thiazol-2-yl) thiourea has been successfully constructed. In this research the designed directions for new substances can generate the SARS-CoV-2 inhibitory drugs in a fast and reliable way.

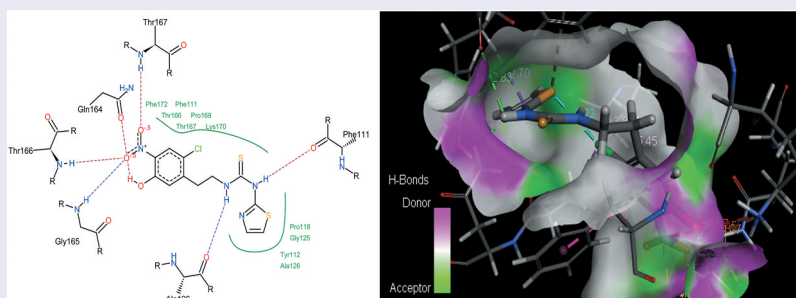
**ARTICLE HISTORY**

Received 27 October 2020

Accepted 19 January 2021

**KEYWORDS**

SARS-CoV-2; hybrid QSAR models; molecular docking simulation; retrosynthesis analysis



## 1. INTRODUCTION

COVID-19, caused by SARS-coronavirus-2 (SARS-CoV-2), has created a worldwide pandemic, affecting hundreds of millions of lives. To share Chinese clinical practices and to provide research in the fierce battle against COVID-19 virus, Dan Zhang et al. provided insight into the clinical applications and basic mechanisms of the proposed CPM against COVID-19 virus [1]. The drugs Azithromycin and Nitazoxanide have been shown to be clinically effective against SARS-CoV-2 [2]; based on pathophysical and pharmacological approaches, the structural and functional relationships may aggregate these drugs. Mina T. Kelleni suggests using

a combination of the two drugs as soon as possible in the clinical course of COVID-19 virus [2]. However, the vaccines against this virus are not currently available, and there are several drugs that are undergoing trials to inhibit the infection and replication of SARS-CoV-2 [3]. Furthermore, Dwight L. McKee et al. has shown that chloroquine and hydroxychloroquine and antiviral drugs, such as remdesivir nucleotide analogues, can be HIV protease inhibitors lopinavir and ritonavir; so far, broad-spectrum antiviral drugs, such as arbidol, favipiravir, and antiviral phytochemicals, have been able to limit the infection of SARS-CoV-2 and mortality from the COVID-19 pandemic [3]. Rucong Yang et al. has

**CONTACT** PHAM Van Tat ✉ [vantat@gmail.com](mailto:vantat@gmail.com) Department of Pharmacy, Faculty of Health Science, Hoa Sen University, 08 Nguyen Van Trang Str., Dist. 01, Ho Chi Minh City, Viet Nam

© 2021 Pham Van Tat, Tran Thai Hoa, Au Vo Ky and Pham Nu Ngoc Han

studied the chemical composition and pharmacological mechanism analysis of Qingfei Paidu Decoction (QFPD), clinically used to treat COVID-19 patients in China [4]. Analysis and synthesis exploring the relationship between the lymphocyte count and the severity of COVID-19 were studied by Qianwen Zhao et al. [5]. The drug discovery and testing efforts for the COVID-19 virus combined experimental methods and computer simulation techniques of the transmission mechanism of human SARS-CoV-2 virus proposed by Jian Shang et al [6]. In the current pandemic, efforts to find broad-spectrum antiviral agents pose a challenge. Using the data analysis and modeling techniques that support and predict SARS-CoV-2 protease inhibitors, Kalyan Ghosh et al. has launched a number of QSAR models, based on Monte Carlo optimization techniques, to screen natural products [7]. Using an *in silico* model for predicting the lysosome and endoscopic accumulation, Ulf Norinder et al. identified 36 compounds that could possibly have antiviral effects against the coronavirus [8]. To determine the photoelectron and spectral properties of molecules that are resistant to the virus SARS-CoV-2, G.W. Ejub et al. have summarized and evaluated the significant role of molecular descriptors to construct models (QSAR) and to design molecules such as electronegativity index, global hardness chemical potential, ionizing potential, electronic affinity [9]. Meenakshi Negi et al. has evaluated the effects of heterocyclic compounds that are closely related to viral infections, AIDS, and cancer [10]. In a more extensive research scope to explore coronavirus inhibitory compounds, Meenakshi Negi et al. proposed a treatment based on the research and development of heterocyclic compounds against MERS-CoV and SARS-CoV-2 by *in vitro*, *in vivo*, and *in silico* approaches [10].

In Vietnam, because there are no cures for COVID-19, respiratory infections caused by the coronavirus still have complications and potential risks. Additionally, in Vietnam, there are many studies that combine experimentally and building *in silico* models. Thuy et al. performed the binding of compounds in garlic essential oil to the SARS-CoV-2 virus by docking simulation [11]. In another study on essential oils of natural *Melaleuca leucadendra* in Vietnam, Ai Nhung et al. carried out the docking simulation to determine the binding energy to express the binding affinity between ligand and protein for the SARS-CoV-2 virus inhibition [12]. In order to perform docking simulations, looking for different types of compounds to prevent SARS-CoV-2 infection, Ai Nhung et al. investigated the use of silver and bis-silver complexes with lighter tetrylene. They demonstrated the potential of using silver-carbene and bis-

silver-carbene complexes to inhibit the SARS-CoV-2 virus infection [13].

There are many reasons to explain the earlier effects in the treatment of SARS-COV-2 patients in Vietnam. Many drugs and regimens used by Vietnamese doctors to treat patients with respiratory tract infections are similar to the drugs used to treat HIV patients. Another treatment has shown that the use of chloroquine and an antibiotic can also treat the SARS-CoV-2 virus. However, all these drugs are still in the research and development phase. Meanwhile, there are also many different drugs belonging to the antibacterial, antifungal and antiviral groups that are being considered and studied. The discovery of a new drug and inhibitory mechanism for the SARS-COV-2 still requires a lot of research.

From the studies on SARS-CoV-2, we realize that there is a need to continue research to build *in silico* models based on the molecular descriptors. Currently, this remains a major research and plays an important role in the development of new drugs to treat the SARS-CoV-2. Furthermore, nitrogen and sulfur heterocyclic compounds have been comprehensively explored for their antimalarial, anti-HIV, anti-inflammatory, anticancer, antibacterial, and antiviral properties. Nitrogen and sulfur heterocyclic compounds phenethylthiazolethiourea with anti-HIV-1 RT inhibition have been suggested by Frank W. Bell et al. (1995) [14]. We found that this is a group of substances with good inhibitory activity against HIV-1 RT. Therefore, in this study, we decided to use the group of phenethylthiazolethiourea. Furthermore, the *in silico* models based on machine learning and simulation techniques are still the main focuses today, as they provide more reliable predictability [15,16].

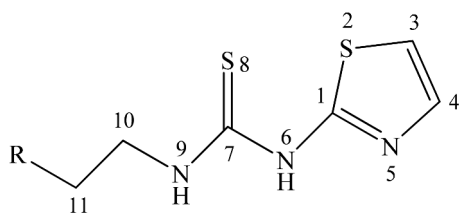
The main objective of this work is to select a group of anti-HIV-1 RT phenethylthiazolethiourea compounds [14] to screen and develop new compounds that can inhibit SARS-CoV-2 infection. We used the molecular descriptors of phenethylthiazolethiourea derivatives filtered by genetic algorithm to develop the QSAR models based on the multivariate linear regression (MLR), Boosted Trees for Regression (BTR) method (this is a machine learning technique), and artificial neural network (ANN). We then designed several new derivatives and predicted the HIV-1 inhibitory activity based on the most reliable QSAR models. The new molecules are explored for their inhibitory activity by docking simulations that bind them to HIV-1 receptors (PDB ID: 1ODW) [17] and SARS-CoV-2 receptors (PDB ID: 6LU7) [18]. To evaluate the binding of new substances on the receptors of HIV-1 and SARS-CoV-2 proteins,

we employed the electrostatic, hydrogen-bond, non-binding and hydrophobic interaction properties. Our team carried out the selection of a novel substance with an energy-based inhibitory activity that binds the substance on the receptors of HIV-1 (PDB ID: 1ODW) and SARS-CoV-2 (PDB ID: 6LU7). The new substance with the strongest activity will be synthesized using retrosynthesis analysis technique.

## 2. CALCULATION METHOD

### 2.1. Research Data

In recent studies, several new types of nonnucleoside have been shown to inhibit HIV-1 [14]. In addition, it was shown that the analogues are selective HIV-1 inhibitors that are not active against any other nucleic acids with the exception of polymerase. These derivatives



**Figure 1.** A general skeleton of the phenethylthiazolethiourea (PETT) derivatives [14].

appear to be the noncompetitive inhibitors of HIV-1 RT [14] and are mostly associated with the same allosteric site of the enzyme. A structural-active (SAR) relationship study of a group of PETT compounds against HIV-1 is explored as a new SARS-CoV-2 inhibitor. The PETT group of compounds includes the molecular structure and the 50% inhibitory concentration ( $IC_{50}$ ,  $\mu M$ ) taken from Frank's work (1995) [14]. This value is the average of at least two trials,  $IC_{50}$  ( $\mu M$ ). The general molecular structure is shown in Figure 1. The  $IC_{50}$  value indicates the inhibitory concentration in unit  $\mu M$ . The structure of phenethylthiazolethiourea derivatives (PETT) with activity  $pIC_{50} = -\log IC_{50}$  is shown in Table 1.

Group of 26 different derivatives of phenethylthiazolethiourea [14] were used in this study to construct the QSAR models. This data group was divided into the 70% training group, the 15% validation group, and the 15% test group, as given in Table 1.

### 2.2. Calculating the Molecular Descriptor

To build QSAR models with effective predictive quality, the molecular descriptors characterizing 2D, 3D structural, and the electrostatic properties of molecules were calculated from the QSARs program [19,20]. These are required for developing the *in silico* QSAR models. To develop the QSAR models,

**Table 1.** The structures of the antiHIV-1 phenethylthiazolethiourea derivatives corresponds to the experimental activities,  $pIC_{50,exp}$  [14] and those from the QSAR<sub>GA-MLR</sub>, QSAR<sub>GA-BTR</sub> and QSAR<sub>GA-ANN</sub> models.

No	R	$pIC_{50,exp}$	QSAR <sub>GA-MLR</sub>		QSAR <sub>GA-BTR</sub>		QSAR <sub>GA-ANN</sub>	
			$pIC_{50,cal}$	SR	$pIC_{50,cal}$	SR	$pIC_{50,cal}$	SR
C-01 <sup>t</sup>	C <sub>6</sub> H <sub>5</sub> -	0.0458	-0.1186	0.0270	0.0748	0.0008	0.0492	0.0000
C-02 <sup>tr</sup>	2-FC <sub>6</sub> H <sub>4</sub> -	1.2218	1.2363	0.0002	1.2870	0.0043	1.2348	0.0002
C-03 <sup>v</sup>	3-FC <sub>6</sub> H <sub>4</sub> -	0.8239	1.0474	0.0499	0.8641	0.0016	0.7918	0.0010
C-04 <sup>tr</sup>	4-FC <sub>6</sub> H <sub>4</sub> -	0.0000	0.6963	0.4848	0.1005	0.0101	0.0227	0.0005
C-05 <sup>tr</sup>	2-CH <sub>3</sub> OC <sub>6</sub> H <sub>4</sub> -	1.3979	0.6966	0.4918	1.0923	0.0934	1.0883	0.0959
C-06 <sup>t</sup>	3-CH <sub>3</sub> OC <sub>6</sub> H <sub>4</sub> -	0.8239	0.6429	0.0328	0.8001	0.0006	0.9139	0.0081
C-07 <sup>tr</sup>	4-CH <sub>3</sub> OC <sub>6</sub> H <sub>4</sub> -	0.4559	0.5116	0.0031	0.5680	0.0126	0.4720	0.0003
C-08 <sup>tr</sup>	2-CH <sub>3</sub> C <sub>6</sub> H <sub>4</sub> -	1.0969	0.9079	0.0357	0.8165	0.0786	1.1011	0.0000
C-09 <sup>tr</sup>	2-N = N-C <sub>6</sub> H <sub>4</sub> -	1.5229	1.0686	0.2064	1.2740	0.0619	1.5658	0.0018
C-10 <sup>tr</sup>	2-NO <sub>2</sub> C <sub>6</sub> H <sub>4</sub> -	0.8239	0.8212	0.0000	0.8998	0.0058	0.8212	0.0000
C-11 <sup>tr</sup>	2-HOC <sub>6</sub> H <sub>4</sub> -	-0.0414	-0.0388	0.0000	-0.0122	0.0009	-0.0638	0.0005
C-12 <sup>v</sup>	2-ClC <sub>6</sub> H <sub>4</sub> -	0.2218	0.8097	0.3456	0.4914	0.0727	0.2045	0.0003
C-13 <sup>tr</sup>	3-C <sub>2</sub> H <sub>5</sub> OC <sub>6</sub> H <sub>4</sub> -	1.2218	0.8345	0.1500	0.8793	0.1173	1.1904	0.0010
C-14 <sup>tr</sup>	3-C <sub>3</sub> H <sub>7</sub> OC <sub>6</sub> H <sub>4</sub> -	0.6990	0.7226	0.0006	0.7451	0.0021	0.6898	0.0001
C-15 <sup>tr</sup>	3-iso-C <sub>3</sub> H <sub>7</sub> OC <sub>6</sub> H <sub>4</sub> -	0.3979	0.7378	0.1155	0.8031	0.1642	0.4025	0.0000
C-16 <sup>tr</sup>	3-C <sub>6</sub> H <sub>5</sub> OC <sub>6</sub> H <sub>4</sub> -	-0.0414	-0.0922	0.0026	0.1613	0.0411	-0.0781	0.0013
C-17 <sup>v</sup>	2,6-di-CH <sub>3</sub> OC <sub>6</sub> H <sub>3</sub> -	1.0458	1.2788	0.0543	0.6673	0.1432	1.0183	0.0008
C-18 <sup>tr</sup>	2,5-di-CH <sub>3</sub> OC <sub>6</sub> H <sub>3</sub> -	0.6990	1.2442	0.2972	0.7109	0.0001	0.7087	0.0001
C-19 <sup>tr</sup>	3-Br-C <sub>6</sub> H <sub>4</sub> OC <sub>6</sub> H <sub>3</sub> -	1.5229	1.5088	0.0002	1.1860	0.1135	1.5305	0.0001
C-20 <sup>tr</sup>	2-F-6-CH <sub>3</sub> OC <sub>6</sub> H <sub>3</sub> -	2.0000	1.8000	0.0400	1.7614	0.0569	1.9533	0.0022
C-21 <sup>v</sup>	2-C <sub>2</sub> H <sub>5</sub> O-6-F-C <sub>6</sub> H <sub>3</sub> -	2.0000	2.0051	0.0000	1.7961	0.0416	1.9981	0.0000
C-22 <sup>tr</sup>	2,6-di-F-C <sub>6</sub> H <sub>3</sub> -	2.0000	2.0736	0.0054	2.0137	0.0002	2.0281	0.0008
C-23 <sup>tr</sup>	2-Cl-6-F-C <sub>6</sub> H <sub>3</sub> -	2.2218	1.8299	0.1536	2.0577	0.0269	2.0995	0.0150
C-24 <sup>tr</sup>	2-pyridyl	0.6990	0.1825	0.2668	0.3011	0.1583	0.6733	0.0007
C-25 <sup>t</sup>	1-CH <sub>3</sub> -pyrrol-2-yl	-0.2788	-0.0991	0.0323	0.0656	0.1186	-0.2873	0.0001
C-26 <sup>t</sup>	2-furyl	0.1871	0.4601	0.0745	0.3011	0.0130	0.1288	0.0034

tr: training set; v: validation set; t: test set; SR: square of residual

220 different descriptors were calculated from the molecular structures and used to validate the relationship between structure and activity  $pIC_{50}$ . A total of 220 molecular descriptors consisting of 9 different groups was calculated. Out of these molecular descriptors, the 2D descriptors consist of five groups: connectivities simple ( $n = 45$ ) describes the connection of a molecule as a graph on the basis of the delta values of the atoms in a certain type of sub-graph. When calculating simple bonds, all non-hydrogen atoms are considered to be of the same type; connectivities valences ( $n = 45$ ) obtained by calculating values for non-carbon atoms different from those calculated for identical bonded carbon atoms; E-state ( $n = 80$ ) a family of atomic-level molecular descriptors that characterize the accessibility of electrons at each structural component (atom or group of hydride, atomic type, bonding, or bonding type) of the molecule; Kappa Shape Indices ( $n = 8$ ) is a family of graph-based structural descriptors designed with the particular goal of encoding relative shape characteristics into multiple computed values for each molecule; Molecular Properties ( $n = 18$ ) includes parameters describing the overall individual molecule (molecular weight, number of various elements such as hydrogen bonding or forming agent, etc.); General 3D Descriptors ( $n = 11$ ) is a set of physical properties that characterize 3D molecular structures; Molecular moment ( $n = 13$ ) gives the absolute values and components of the moment of inertia, the dipole moment, and the quadrupole moment of the molecule. Symbols and property descriptions of the important descriptors are selected (in Table S1) [20,21]. The contribution ability to  $pIC_{50}$  activity of each descriptor (that could be structural descriptors of the E-state electrostatic group) in the QSAR model is important [10,20,21]. These molecular descriptors are specific for bonds  $=C<$ ,  $-NH-$  and  $OH$ . These bonding groups have a great influence on electron-changing properties of molecules by the conjugation effect, especially by changing the substituents at the  $R$ -position of the molecule. The molecular electrostatic potential determines the binding ability when docking the molecular system. Here, nucleophilic or electrophilic interactions may take place.

### 2.3. Building QSAR Model

#### 2.3.1. Variable Selection and QSAR<sub>GA-MLR</sub> Model

In the general case, the coefficients in the multivariable linear regression model

QSAR<sub>GA-MLR</sub> characterize the contribution to  $pIC_{50}$

activity [22]. The general model QSAR<sub>GA-MLR</sub> is shown as follows [23]:

$$y = b_0 + \sum_{i=1}^k b_i x_i \quad (1)$$

The observed values  $y_i$  for a compound are approximately represented by the linear combination of molecular descriptors  $x_i$ . The characteristic coefficients for that association are called the regression coefficients  $b_i$  [22]. The significance of these coefficients in the

QSAR<sub>GA-MLR</sub> model is evaluated based on:

- Standard error is a measure of the accuracy of the estimate with respect to coefficient.

- The Student  $t$ -value is also used in our study to compare with the standard  $t$ -value at the confidence level. 95% linear regression modeling was performed by program MiniTab [24]. The QSAR<sub>GA-MLR</sub> model shows the correlation between the dependent variable and a number of independent variables.

The regression coefficients corresponding to each variable have a certain contribution weight value when predicting a specific  $y_i$  value, providing insight into the influential nature of each set of molecular descriptors as well as assist in interpreting the quality of a QSAR<sub>GA-MLR</sub> model.

As we all know, the selection of molecular descriptors for the QSAR<sub>GA-MLR</sub> model is becoming very important in analysis, molecular design, and activity prediction  $pIC_{50}$ . Indeed, the molecular descriptor dataset used in this study consisted of 220 molecular descriptors. Due to the large amount of molecular descriptors, it is imperative to find and choose the most useful and meaningful molecular descriptor. The variable selection method is used to eliminate variables that contribute insignificantly to the predictive efficiency of the QSAR model. The negative contributing variables can be removed to improve the overall capabilities and efficiency of the QSAR model. The molecular descriptors used for the QSAR<sub>GA-MLR</sub> model are selected by a forward technique based on the genetic algorithm (GA) [20,21,25]. The genetic algorithm (GA) is one of the most widely used and most effective current priority algorithms for selecting molecular descriptor. The method begins with a random set of molecular descriptors. Then, the input configuration determines which molecular descriptor is ignored during the predictability test of the QSAR<sub>GA-MLR</sub> model. During the next generation, the genetic algorithm uses a natural selection process to select molecular descriptors as the superior input configuration type. The genetic algorithm uses this configuration of molecular

descriptors to create a new population. The successive generations (or call each successive step) will provide better configuration (set of molecular descriptors) for the QSAR<sub>GA-MLR</sub> model. In the final step, the best configuration (set of descriptors) is selected. This method is time-consuming, but yields great results for identifying satisfactory molecular descriptors and detecting interdependencies. The QSAR model of multivariate regression is built from selected variables by the genetic algorithm called QSAR<sub>GA-MLR</sub> model.

In this study, we combined genetic algorithm with regression techniques to create hybrid models (QSAR<sub>GA-MLR</sub>) [21,23,25]. Genetic algorithm is one of the most effective algorithms for selecting the molecular descriptors. This method relies on a randomized algorithm to select a globally optimized model using natural genetic mechanisms and biological evolution [20,25].

### 2.3.2. Boosting Trees for Regression (QSAR<sub>GA-BTR</sub>)

Boosted Trees for Regression (BTR) is a machine learning technique used to build regression models [26]. This technique produces a prediction model in the form of a set of prediction models, typically a decision tree [26,27]. The algorithm for the BTR model evolved from an enhanced method for regression trees to perform various studies in chemistry [27,28]. In this study, we built a BTR model limited to 50 regression trees. For each step of the algorithm, a partition of the data is identified, and the deviation of the observed values is calculated. Each small regression tree consists of three nodes. The successive 3-node trees are formed and attached to the remainder to find another partition. This further reduces the remaining variance for the data. The BTR model provides predictive results by combining decisions from a series of underlying models  $f_i(x)$ . This method uses many simple trees for better predictive performance. This model can be written as a series of models:

$$G(x) = f_0(x) + f_1(x) + f_2(x) + \dots + f_i(x) \text{ with } i = 1 - 50(2)$$

Where  $G(x)$  is the sum of  $f_i(x)$  base regression trees corresponding to a simple decision tree.

This is done through SAS JMP Pro [29], which provides suitable regression trees. The BTR tree model [30] includes an initial data entry stage  $\{(x_i, y_i)\}_{i=1}^n$  and a differential loss function  $L(y_i, f(x))$ ; the second stage, initializing the first tree, has only one leaf and the value of the leaf is the average of all initial prediction results using constant values:  $f_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n L(y_i, \gamma)$  [30]:

- Calculate values  $r_{im} = - \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)}$  with  $i = 1, \dots, n(3)$
- The regression tree is matched for the  $r_{im}$  values to create an area  $R_{jm}$  with  $j = 1 \dots j_m$
- With  $j = 1 \dots j_m$  calculate  $\gamma_{im} = \arg \min \sum L(y_i, f_{m-1}(x_i) + \gamma)(4)$
- Update  $f_m(x) = f_{m-1}(x) + \gamma \sum_{j=1}^m I(x \in R_{jm})(5)$

With this second stage, each successive regression tree is constructed. The residue of one successive regression tree is calculated and minimized.

$$y_{i+1} = y_i + \text{learning rate} \times \text{residual.}$$

This process is performed until all 50 regression trees have been completed. These regression trees undergo repeated screening until new regression trees are developed, and further selection is based on the value RMSD [29]. This is done until the stop criterion of the RMSD value is reached. We get the final predicted results.

### 2.3.3. Model QSAR<sub>GA-ANN</sub>

A neural network architecture consists of a number of layers, each consisting of a number of neurons [16,29]. The structure of the artificial neural network is built in accordance with the nature of the phenethylthiazothiourea derivatives. We can apply the artificial intelligence techniques to find the most suitable network architecture for these derivatives. At this stage, the neural network nodes apply an iterative process of the number of molecular descriptors (input variables) to adjust the weights for optimal prediction.

However, the relationships between the molecular descriptors and activity pIC<sub>50</sub> cannot be expressed as clearly as in the traditional models. We use the neural network architecture I(k)-HL(m)-O(1) [31]. Here, I(k) is the input layer with  $k$  molecular descriptors or the number of neurons as defined in the QSAR<sub>GA-MLR</sub> model (13); layer HL(m) is hidden layer with  $m$  neurons; layer O(1) is the output layer with 1 neuron corresponding to pIC<sub>50</sub> activity in the QSAR<sub>GA-MLR</sub> model.

To train the neural network for good results, the sum of squared residuals (SSR) is used to determine the difference between the target values  $t_i$  and the prediction results  $y_i$  [15,30]:

$$SSR = \sum (y_i - t_i)^2 \text{ with } i = 1 - n \text{ (n is the number of training cases)}(6)$$

These techniques may require a smaller number of iterations to train an artificial neural network based on fast convergence rates and smarter search criteria [29–31].

## 2.4. Docking Simulation

Docking is used to predict the preferred direction of two chemical species to interact with each other. Usually, this is the interaction of a small molecule to a protein. Docking is an extremely important tool in the structural drug design process. The docking process involves locating the different forms (potential isomers) of a small molecule relative to a protein molecule and determining the optimal interaction geometry and its binding energy. The docking process begins with a prepared receptor and ligand. The general algorithm is divided into major phases:

1. Prepare protein (CPLX) (ie from PDB).

The first phase of the docking simulation is to prepare protein for input into other protocols. We insert the missing atoms into the incomplete residues of the protein chains, model the missing loop regions, remove alternative structures, remove water, normalize atomic names. The SARS-CoV-2 (6LU7) protein provided by the Protein Data Bank. These are the database of the experimental structures of the SARS-CoV-2 (6LU7) protein provided by the Protein Data Bank [18]. The RCSB PDB data is built on experimental data by generating resources for molecular biology research. The protein structure of HIV-1 (PDB ID: 1ODW) was also obtained through the Protein Data Bank [17]. The native HIV-1 Proteinase was suggested and corrected by Thanki, N. et al. in 2011 based on X-ray technique diffraction data (Classification: Hydrolase/Hydrolase Inhibitor; Organism (s): Human immunodeficiency virus 1) [17].

To prepare the docking proteins, we set the value of the loop parameter 'True' to model the missing loop region with a maximum loop length of 20. The CHARMM force field is used for optimization. The protein dielectric constant is 10. The pH value for protonation is 7.4. The ionic Strength is 0.145. The Energy cutoff is 0.9. The Keep ligands is true. The X, Y, Z coordinates of active location SBD\_Site\_Sphere XYZ are identified as 97.797224, 85.523750, 135.422526, respectively.

2. Prepare the ligands

The molecular structures are built and optimized by a quick energy minimization algorithm (Dreiding). The force field used is Dreiding with the steepest descent minimizer to optimize all. Then small molecules (ligand) are built up in an sdf database. After building and optimizing small molecules, we prepared the parameters needed for the docking process. The second phase is to prepare the ligands for importing into other protocols, perform the tasks of removing duplicates, enumerating

isomers and tautomers, and generating 3D conformations. The Specify Ligands parameter allows you to select a primary ligand source in the sdf data file. Database of selected ligands with change parameters of change Ionization method is at maximum and minimum pH 6.5; The Generate tautomers parameter is true and the maximum Tautomer number is 10; Amides Tautomerization is Tauomerize Only Diamides. Tao Isomers is true; Generate coordinates are 3D.

3. Prepare all the files required for assembling (grid parameter file, map file, parameter docking file).

4. Run docking.

5. Analysis of docking results.

## 2.5. Validation of QSAR Models

The QSAR models and the docking simulations developed here were evaluated using statistical analysis such as multiple correlation coefficient ( $R^2$ ), cross-validation  $Q^2_{LOO}$  (leave-one-out) (Eqs. (7)), adjusted correlation coefficient  $R^2_{adj}$  (Eq. (8)), and significance level (p-value) [21,22,32]. In addition, the VIF value is employed to validate and detect the multicollinearity (Eq. (9)) of the QSAR model [19,20]. The statistical parameters are calculated by the following equations [20–22]:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

Where,  $y_i$  and  $\hat{y}_i$  are the experimental and prediction values

$$R^2_{adj} = 1 - \left[ \frac{(n-1) \times (1-R^2)}{N-k-1} \right] \quad (8)$$

$$VIF = \frac{1}{1-R^2} \quad (9)$$

The predictive quality of  $pIC_{50}$  in the QSAR models are validated by the RMSR error values for phenethylthiazolethiourea derivatives in the training group, validation group, and test group [23,32]. The prediction error RMSR is calculated by the equation (10):

$$RMSR = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (10)$$

The significant percentage contribution of each descriptor for  $pIC_{50}$  can be determined based on the mean percentage contribution,  $APC, \%$  [19,21,23] from the selected QSAR<sub>GA-MLR</sub> models. It is calculated using the following formula (11):

$$APC_{m,n,x_i}, \% = \frac{1}{n} \left( 1 \sum_{j=1}^m \frac{|b_j x_i|}{\sum_{i=1}^k |b_i x_i|} 100\% \right) \quad (11)$$

Additionally, the equilibrium constant ( $K_{eq}$ ) was used to examine the binding affinity between the ligand and the protein receptor during a docking simulation study using equation (12) [33]:

$$\Delta G^0 = -2.303RT \log_{10} K_{eq} \quad (12)$$

Here,  $R$  is a gas constant  $0.00831446 \text{ kJ.K}^{-1}.\text{mol}^{-1}$

## 2.6. Planning Retrosynthesis

After constructing the QSAR models, the  $pIC_{50}$  activity of the newly designed compounds is predicted. Finding and building a viable synthetic pathway for organic compounds remains the focus. This is the path that requires building a complex matrix of retrosynthesis ability [34] which are supported by developed computer technology that quickly builds and selects a synthesis path that meets the requirements of a new drug synthesis [35,36]. The retrosynthesis path will simultaneously show all possible paths from the starting materials. Today, machine learning algorithms have significantly contributed to the work of synthesizing new substances [37].

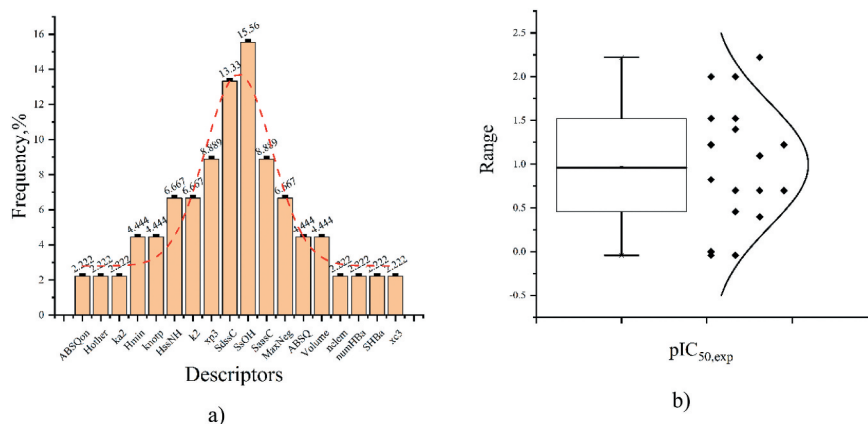
To find a new compound that could rapidly treat the new SARS-CoV-2 virus, a new compound synthesis path is determined from the molecular design and screening process. The retrosynthesis paths are provided using the complex algorithms retrieved from a database of over 100,000 reaction rules [37]. The reaction path construction is quickly and efficiently scanned from millions of data points to identify the available reactions with maximum accuracy [36,37]. Based on the established requirements, we plan the reaction paths for highly bioactive molecules. Our retrosynthesis system begins to design the reaction mode based on the rules of the reaction.

## 3. RESULTS AND DISCUSSION

### 3.1. Building QSAR Model

#### 3.1.1. QSAR<sub>GA-MLR</sub> Model

The usage data set was collected from one source. It shows that the distribution density of  $pIC_{50}$  activity of derivatives is concentrated in the range  $-0.2788$  to  $2.2218$ , as shown in Table 1 [14]. We used the Grubbs statistical test to check the distribution of  $pIC_{50}$  values in the data set with 95% confidence. The test results show the Grubbs statistic value at 95% confidence level  $G_{Obs} = 1.873 < G_{Critical} = 2.841$ ; These  $pIC_{50}$  activity data appear to be the suitable distribution for the construction of QSAR models, as shown in Figure 2b. The dataset for



**Figure 2.** A) Distribution of molecular descriptors in nine QSAR<sub>GA-MLR</sub> models; b) Distribution of  $pIC_{50}$  values in the original dataset.

**Table 2.** Molecular descriptors with range of variation selected for all QSAR<sub>GA-MLR</sub> models with  $k = 1-9$ .

descriptor	notation	min	Max	mean	descriptor	notation	min	max	mean
ABSQ	$x_1$	2.642	4.652	3.090	Nelem	$x_{10}$	4.000	6.000	4.962
ABSQon	$x_2$	0.727	2.776	1.079	numHBa	$x_{11}$	5.000	8.000	6.269
Hmin	$x_3$	0.457	0.937	0.752	SaasC	$x_{12}$	-0.345	3.847	2.336
Hother	$x_4$	8.011	16.052	9.844	SdssC	$x_{13}$	0.381	0.624	0.554
HssNH	$x_5$	3.653	3.800	3.721	SHBa	$x_{14}$	16.988	42.933	26.435
k2	$x_6$	6.667	10.222	8.252	SsOH	$x_{15}$	0.000	9.599	0.715
ka2	$x_7$	6.071	9.067	7.446	Volume	$x_{16}$	208.732	329.190	247.757
knotp	$x_8$	-1.293	-0.490	-0.874	xc3	$x_{17}$	0.697	1.309	0.950
MaxNeg	$x_9$	-1.000	-0.280	-0.362	xp3	$x_{18}$	4.994	7.803	6.085

QSAR modeling was randomly divided into 70% training group, 15% validation group, and 15% test group. The

QSAR<sub>GA-MLR</sub> models were constructed from a training group of 18 phenethylthiazoethiourea derivatives. To construct the QSAR models, the genetic algorithm [20,21,32] was used to select molecular descriptors from a dataset consisting of 220 different molecular descriptors [20]. The molecular descriptors in the selected models are given in Table 2. We selected nine QSAR<sub>GA-MLR</sub> models that are significant with the most influential molecular descriptors based on the statistical values  $R^2$ ,  $R^2_{adj}$ ,  $Q^2_{LOO}$  and SE (standard error), as given in Table 3.

The frequencies of molecular descriptors in all the selected QSAR<sub>GA-MLR</sub> models are used to calculate the contribution level in QSAR<sub>GA-MLR</sub> models as well as in each molecule. The cumulative frequency statistics of each molecular descriptor is shown in Figure 2a.

We found that some 2D and 3D molecular descriptors such as HssNH, k2, xp3, SdssC, SsOH, SaasC, and MaxNeg appear most in the QSAR<sub>GA-MLR</sub> models. In addition, in Figure 2a, we can easily see that the molecular descriptors SsOH and SdssC have the largest frequencies. Thus, these molecular descriptors may have the most significant contribution to pIC<sub>50</sub> activity.

From the nine QSAR<sub>GA-MLR</sub> models shown in Table 3, the significance of the molecular descriptors was determined based on their contribution to the phenethylthiazoethiourea derivatives. The percentage contribution of each descriptor to the pIC<sub>50</sub> can be determined based on the mean percentage contribution,

APC%, of these mine QSAR<sub>GA-MLR</sub> models using Equation (11). The average contribution, APC%, of the molecular descriptors can be used as the basis for determining the relevant molecular structural properties that are related to the molecular descriptors. This may lead to changes in molecular structure when designing new molecules. The quantitative effect on the ability to contribute of each molecular descriptor was clearly shown on each of the phenethylthiazoethiourea molecular structures, as shown in Table 4.

As described in Table 4, the values APC% of the molecular descriptors were evaluated and compared with each other. The molecular descriptor SdssC characterizes the type of bond (Sum of (=C<)) in E-states group) relating to the molecular electrostatic properties. The largest contribution of SdssC molecular descriptor is 50.751%, which includes Atom Type E-State Sum Nonhydrogen Indices; SdssC is largely influenced by the properties of the double bonds (= C <). The molecular descriptor HssNH characterizes the electrostatic properties of the bond – NH – (Sum of ([–NH–]) hydrogen E-States) and contributes 21.538%. The contribution of the xp3 molecular descriptor characterizing the simple 3rd order path chi index is 8.749%. For a newly designed molecule, we can change the R-functional groups on the phenyl ring

Furthermore, the descriptor HssNH has also been shown to be highly influenced by the bonds that have nitrogen atoms ([–NH–]). In addition, the molecular descriptor SsOH also displayed a significant contribution from the – OH bond. Because the electrostatic property of the bond – OH (Sum of (–OH) E-States)

**Table 3.** The statistical values and regression coefficients of selected QSAR<sub>GA-MLR</sub> models with  $k = 1-9$ .

Variable	QSAR <sub>GA-MLR</sub> model with $k = 1-9$								
	1	2	3	4	5	6	7	8	9
$R^2$	0.419	0.545	0.638	0.720	0.778	0.854	0.849	0.896	0.934
$R^2_{adj}$	0.395	0.505	0.588	0.667	0.722	0.808	0.791	0.848	0.897
SE	0.559	0.505	0.461	0.415	0.379	0.315	0.329	0.281	0.231
$Q^2_{LOO}$	0.330	0.451	0.536	0.598	0.701	0.718	0.724	0.821	0.729
Constant	5.110	5.704	7.110	14.440	17.370	–118.300	24.630	–32.600	–76.350
$x_1$						–5.022			–4.414
$x_2$								–3.146	
$x_3$					–1.663	–1.950			
$x_4$								–0.518	
$x_5$						35.280		7.820	22.480
$x_6$						0.493	0.773		2.101
$x_7$									–1.321
$x_8$							–1.242		–2.635
$x_9$						–9.830		–9.740	–10.130
$x_{10}$						–0.545			
$x_{11}$								0.534	
$x_{12}$			0.339	0.858	0.881		0.639		
$x_{13}$	–7.650	–8.590	–12.550	–20.370	–22.600		–33.300		
$x_{14}$							–0.150		
$x_{15}$		–0.103	–0.105	–0.122	–0.111		–0.118	–0.073	–0.078
$x_{16}$								0.033	0.017
$x_{17}$								–2.248	
$x_{18}$				–0.689	–0.774		–1.683		–1.693

**Table 4.** The percent contribution,  $PC\%$  and average contribution percentage,  $APC\%$ , per molecular descriptor on all the selected QSAR<sub>GA-MLR</sub> models for phenethylthiazolethiourea derivatives.

variable	Value $PC\%$ of each descriptor for QSAR <sub>GA-MLR</sub> models with $k = 1 - 9$									$APC\%$
	1	2	3	4	5	6	7	8	9	
$X_1$	0.000	0.000	0.000	0.000	0.000	9.757	0.000	0.000	9.381	2.126
$X_2$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	6.091	0.000	0.677
$X_3$	0.000	0.000	0.000	0.000	6.165	0.925	0.000	0.000	0.000	0.788
$X_4$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	9.282	0.000	1.031
$X_5$	0.000	0.000	0.000	0.000	0.000	82.820	0.000	53.187	57.839	21.538
$X_6$	0.000	0.000	0.000	0.000	0.000	2.563	15.269	0.000	11.938	3.308
$X_7$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	6.774	0.753
$X_8$	0.000	0.000	0.000	0.000	0.000	0.000	2.603	0.000	1.581	0.465
$X_9$	0.000	0.000	0.000	0.000	0.000	2.230	0.000	6.345	2.514	1.232
$X_{10}$	0.000	0.000	0.000	0.000	0.000	1.704	0.000	0.000	0.000	0.189
$X_{11}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	6.096	0.000	0.677
$X_{12}$	0.000	0.000	9.961	11.203	9.831	0.000	3.566	0.000	0.000	3.840
$X_{13}$	100.000	98.609	89.092	64.194	60.612	0.000	44.253	0.000	0.000	50.751
$X_{14}$	0.000	0.000	0.000	0.000	0.000	0.000	9.558	0.000	0.000	1.062
$X_{15}$	0.000	1.391	0.947	0.508	0.391	0.000	0.204	0.087	0.037	0.396
$X_{16}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	15.026	2.839	1.985
$X_{17}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	3.887	0.000	0.432
$X_{18}$	0.000	0.000	0.000	24.095	23.001	0.000	24.546	0.000	7.098	8.749

caused its highest present frequency in the models, it contributes significantly in the QSAR<sub>GA-MLR</sub> model (13). This also demonstrates that the molecular activity is influenced by the molecular structure related to the electrostatic properties of this bond ( $-OH$ ).

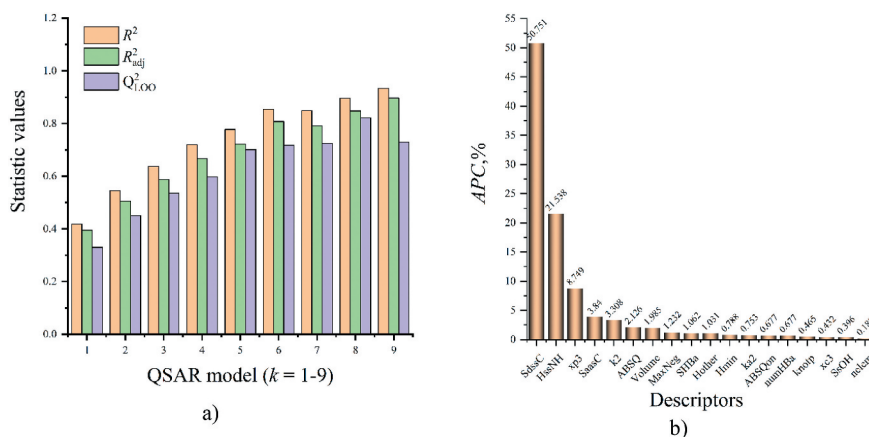
Thus, the molecular descriptors SdssC, SsOH, and HssNH present in the models (Table 3) show the significant contribution of all types of molecular topologies (Table 4). This gives us a very important decision to design molecules. The magnitude of the  $APC\%$  may be an important statistical basis to consider for the substitution group change in the phenyl ring, which produces a molecule with a higher activity  $pIC_{50}$ . The molecular descriptors can be sorted based on the decreasing value of the percentage contribution  $APC\%$ : SdssC > HssNH > xp3 > SaasC > k2 > ABSQ > Volume > MaxNeg > SHBa > Hother > Hmin > ka2 > ABSQon > numHBa > knotp > xc3 > SsOH > Nelem, as seen in Figure 3b.

The quality of training and predictability of the QSAR<sub>GA-MLR</sub> models in Table 3 were compared based on the statistical values  $R^2$  (regression correlation value) and  $Q^2_{LOO}$  (the statistical value assessed by the method leave-one-out). We found that the QSAR<sub>GA-MLR</sub> model with  $k = 5$  seemed to be the most suitable and the QSAR<sub>GA-MLR</sub> model can be chosen for the following studies. This can also be seen in Figure 4a. This QSAR<sub>GA-MLR</sub> model (13) has good predictability as well:

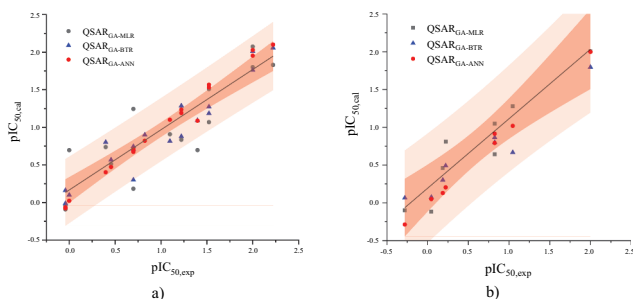
$$pIC_{50} = 17.37 - 22.60 \times SdssC - 0.1108 \times SsOH + 0.881 \times SaasC - 0.774 \times xp3 - 1.663 \times Hmin(13)$$

With  $n = 18$ ;  $R^2 = 0.778$ ;  $R^2_{adj} = 0.722$ ;  $Q^2_{LOO} = 0.701$ ; F-value = 13.99; the values VIF are in range 1.12 to 8.19 and p-value is in range 0.000 to 0.034 at 95% confidence level for molecular descriptors included in QSAR<sub>GA-MLR</sub> model.

The training set can be well described by the QSAR<sub>GA-MLR</sub> regression equation (13), which appeared to



**Figure 3.** A) Quality comparison of QSAR<sub>GA-MLR</sub> models based on  $R^2$ ,  $R^2_{adj}$ ,  $Q^2_{LOO}$  values; b) Percentage of average contribution to  $APC\%$  of each descriptor.



**Figure 4.** Correlation between experimental and calculated  $pIC_{50}$  values from the QSAR models with  $k = 5$ ; a) for training set; b) for validation set and test set.

be statistically significant. The leave-one-out cross-validation (LOO) technique is the basis for choosing the built-in QSAR<sub>GA-MLR</sub> model that can meet the requirement to predict  $pIC_{50}$  value. The QSAR<sub>GA-MLR</sub> model with  $k = 4$  (Table 3) shows less predictive power than the QSAR<sub>GA-MLR</sub> model with  $k = 5$ . Furthermore, we also see that the QSAR<sub>GA-MLR</sub> models with  $k = 6$  to 9 show that the predictability is also not much better than the new model with  $k = 5$ ; this can be seen through the difference value  $Q^2_{LOO}$  between the two models QSAR<sub>GA-MLR</sub> ( $k = 5$  and  $k = 9$ ) is only 2.8%, while the total number of the molecular descriptors of the QSAR<sub>GA-MLR</sub> model with  $k = 9$  is nearly two times more than the number of the molecular descriptors of the QSAR<sub>GA-MLR</sub> model (13) with  $k = 5$ , as shown in Figure 3a.

Furthermore, we find that the largest negative charge on the atoms of the molecule (MaxNeg) also exhibits a significant contribution to  $pIC_{50}$  bioactivity, as shown

in Table 4. Thus, when designing a new molecule, we need to change the functional groups toward increasing the maximum negative charge density (MaxNeg). The functional groups -NH-, -OH, CH<sub>3</sub>O- and bond (=C<) of the carbonyl group are preferred to change. These can alter the electron density of the phenyl ring by the conjugate bonds as well as increase the number of hydrogen bonds with the amino acid groups of the protein. In addition, these also increase the electrostatic and van der Waals bonding interactions between ligand and protein receptors. The biological activity of each molecule depends on the ability of the phenethylthiazolethiourea derivatives to bind to the HIV-1 protein receptor. The SdssC, SsOH and HssNH descriptors in the QSAR<sub>GA-MLR</sub> model (13) are variables that indicate the properties of the molecular structure. The phenethylthiazolethiourea derivatives have been appropriately altered to increase the bioactivity of the molecules during the design process. We rely on the characterization of these molecular descriptors to design new molecules with better response activity, as shown in Table 5.

### 3.1.2. QSAR<sub>GA-BTR</sub> Model

After the QSAR<sub>GA-MLR</sub> model with  $k = 5$  (13) is selected, the QSAR<sub>GA-BTR</sub> model is constructed. The QSAR<sub>GA-BTR</sub> model is built on the basis of machine learning with the molecular descriptors selected in the QSAR<sub>GA-MLR</sub> model (13). The parameters include the number of layers of 50, splits per tree of 3, learning rate of 0.1 and minimum size split of 5 to construct the QSAR<sub>GA-BTR</sub> model. Each layer is a small tree that is used to estimate the  $pIC_{50}$  values. The building process for QSAR<sub>GA-BTR</sub> model was evaluated continuously from layer 1 to layer 50, as shown in Table S2. The QSAR<sub>GA-BTR</sub> model is evaluated for training and predicting ability, which can be based off of RMSR values. For the training, validation, and test group, these values are 0.2295, 0.2545 and 0.1823, respectively. The prediction results and the squares of residuals (SR) of this QSAR<sub>GA-BTR</sub> model for the phenethylthiazolethiourea derivatives in the training, validation, and test group are also shown in Table 1.

In Table S2, the training results of the QSAR<sub>GA-BTR</sub> model showed that the training process is performed continuously by expanding the trees of each layer with corresponding weights. Finally, it can create a great match between predicted values with experimental values, even if the nature of the relationships between the predictor and dependent variables is complicatedly related. Therefore, the BTR method changes with the gradient, which is suitable for extending the corresponding weighted value of simple trees. This demonstrated the potential of the BTR method for use in

**Table 5.** The activity values  $pIC_{50}$  and  $IC_{50}/\mu M$  for HIV-1 proteinase of new derivatives predicted by QSAR<sub>GA-ANN</sub> model.

Derivatives	Molecular structure	$pIC_{50}$	$IC_{50}/\mu M$	Ref.
C-12		-0.2045	0.6245	[14]
C-n1		-0.3102	0.4895	This work
C-n2		-1.1532	0.0703	This work
C-n3		-2.0900	0.0081	This work
C-n4		-2.7751	0.0017	This work
C-n5		-3.0572	0.0009	This work
C-n6		-3.3979	0.0004	This work

a general and powerful machine learning algorithm. Figure 4a and 4b demonstrates the correlation between the experimental and the calculated  $pIC_{50}$  values from the QSAR<sub>GA-BTR</sub> model with  $k = 5$  obtained from the training and prediction. We have found that most of the predicted values (blue triangle points) from the QSAR<sub>GA-BTR</sub> model are in the 95% confidence boundary. These results are much better than those from the QSAR<sub>GA-MLR</sub> model.

### 3.1.3. QSAR<sub>GA-ANN</sub> Model

To conduct

QSAR<sub>GA-ANN</sub> model construction, we used the molecular descriptors of the QSAR<sub>GA-MLR</sub> model (13) as the inputs of the neural network. We build a neural network style with three layers I(5)-HL(9)-O(1), as shown in Figure S1. The input layer I(5) consists of 5 neurons, which are 5 molecular descriptors SdssC, SsOH, SaasC, xp3 and Hmin. The output layer O(1) has 1 neuron, which is the bioactive value  $pIC_{50}$ . This neural network was trained by the Holdback method with the holdback proportion parameter 0.3333 [31]. This multi-layer perceptron network uses an error backpropagation algorithm, transfer function TanH, and learning rate 0.1. The obtained RMSD values are 0.0818, 0.0229 and 0.0538 for the training, validation, and test group, respectively.

To build the QSAR<sub>GA-ANN</sub> model with the structure I(5)-HL( $m$ )-O(1), we need to define the number of hidden layers and the number of required hidden neurons  $m$ . To reduce the complexity and noise in the learning process of a neural network, we construct a neural network model I(5)-HL( $m$ )-O(1) with a hidden layer. The number of neurons  $m$  on the hidden layer HL( $m$ ) can be determined according to the relative rule proposed by Huang (2003) [38,39] as:

$$m = \sqrt{(x+2)N} + \sqrt{N/(x+2)} \quad (14)$$

Here  $x$  output neurons;  $m$  the number of hidden neurons;  $N$  samples are used to train the neural network.

In our study,  $x = 1$ ,  $N = 18$  training samples account for 70% of the data set (in Table 1). Number of neurons  $m$  on the hidden layer determined is 9. The neural network structure uses I(5)-HL(9)-O(1).

The predictability of QSAR<sub>GA-ANN</sub> model for three derivatives groups is much better than the QSAR<sub>GA-MLR</sub> and QSAR<sub>GA-BTR</sub> models, as depicted in Figure 4a and 4b. The prediction values (red solid point) obtained from the QSAR<sub>GA-ANN</sub> model are mostly within and near the 95% confidence boundary. Moreover, the correlation coefficient values  $R^2$  are 0.9873, 0.9058 and 0.7417 for the models

QSAR<sub>GA-ANN</sub>, QSAR<sub>GA-BTR</sub> and QSAR<sub>GA-MLR</sub>, respectively. Thus, the QSAR<sub>GA-ANN</sub> model gives the prediction results at the greatest level of confidence. The QSAR<sub>GA-ANN</sub> model is applicable to predict the  $pIC_{50}$  activity. From the predicted  $pIC_{50}$  activity values for the phenethylthiazolethiourea derivatives of the training, validation and test group, the genetic algorithm showed that it could generate an efficient hybrid QSAR<sub>GA-ANN</sub> model. In particular, the QSAR<sub>GA-ANN</sub> model is used to predict  $pIC_{50}$  activity against newly designed phenethylthiazolethiourea compounds. Meanwhile, the models QSAR<sub>GA-MLR</sub> and QSAR<sub>GA-BTR</sub> failed to predict  $pIC_{50}$  activity of new compounds and had high error rates, as seen in Figure 4.

### 3.1.4. New Derivative Design

The HIV-1 antiviral agents of phenethylthiazolethiourea derivatives containing the altered phenyl rings (Table 1) were suggested by Frank W. Bell et al. (1995) [14]. In the study conducted by Frank W. Bell et al., they demonstrated that the substituent change in the meta position and the substituents in the ortho position, especially, are preferred to displace over the para position. We also saw that fluoro and methoxy were clearly substituted at compounds C-02 and C-04 as well as compounds C-05 and C-07, respectively. Therefore, the electron nature of the substituent group at the ortho position has influenced the anti-HIV activity of HIV-1. From the data in Table 1, the compounds C-02, C-05, C-07 to C-12 show that both types of electron acceptor and donor groups produce the substances with better activity.

The preferred substituents for molecular design that selected are the functional groups fluoro, chloro and methoxy. In addition, the length of the  $m$ -alkoxy substituents can also greatly influence compound activity. This is also shown in Table 1. Thus, the design of a new molecule can be based on the contributions of the SdssC, SsOH and HssNH molecular descriptors, as well as the functional group changes in the phenyl ring in Table 1. So we designed the new molecules by changing the alkoxy and halogen substitution groups, as well as the -OH, -NH<sub>2</sub> and -NO<sub>2</sub> groups. The substance C-12 is chosen as lead substance used to design the different new derivatives. The novel substances designed based on the molecular structure of lead compound C-12 have anti-HIV-1 activity values that are predicted by the QSAR<sub>GA-ANN</sub> model, as shown in Table 5. The prediction RMSR error for  $pIC_{50}$  activity of compound C-12 in the validation group by the QSAR<sub>GA-ANN</sub> model is 0.0173. We found that the activities of the new compounds (Table 5) were stronger than the lead compound

C-12. This also confirms that the orientation for designing new molecules is reliable. The newly designed derivatives are strongly influenced by the – OH and – NH – substituent groups as well as the unsaturated bonds ( $=C<$ ). This is also consistent with our discussion above. The substitution groups were changed in the ortho, meta, and para positions. The bonding types between substituent group and phenyl ring form the conjugate relationship, drastically changing the electron system in the molecular structure.

### 3.2. Docking Simulation

#### 3.2.1. The Docking Process

We performed a docking simulation using the Genetic Optimization for Ligand Docking (GOLD) given by Jones et al. [40]. This is a genetic algorithm to dock the ligand to a binding site on a protein. Predicting how a small molecule bind to a protein is difficult and no method can guarantee success. The best approach is to measure the method reliability as accurately as possible. For this reason, the GOLD method was proposed by Jones et al. [40] and was ranked correctly in the approx 70–80%. The Dock Ligands (GOLD) method provided several parameters for binding the ligand to an active site on the protein. After inputting the receptors on proteins with coordinates X, Y, Z determined, we input the ligands needed for docking. The Genetic Algorithm parameters include the selected Fitness Goldscore function and GA automatic search efficiency. The number of docking is 10. Detect Cavity is True. Early Termination is True and Number of Solutions is 3. The value RMSD is 1.5.

#### 3.2.2. Analysis of Docking Results

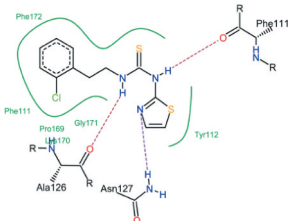
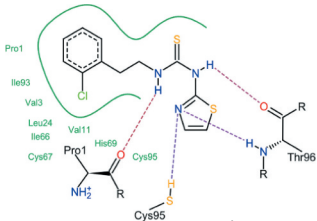
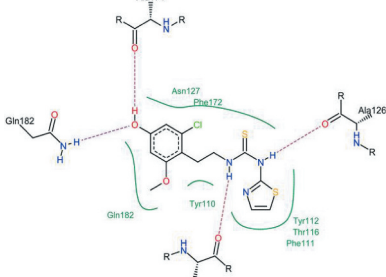
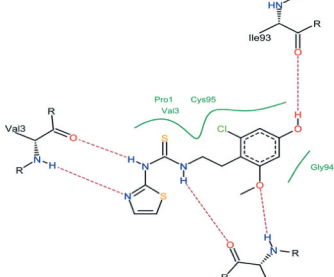
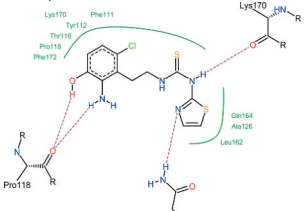
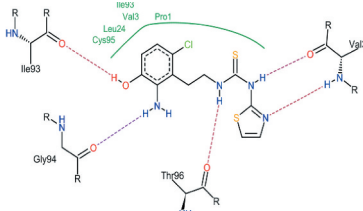
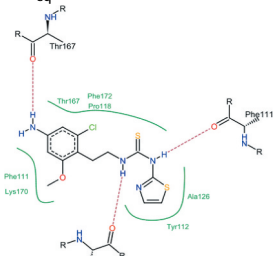
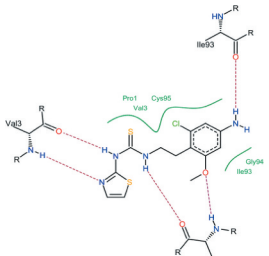
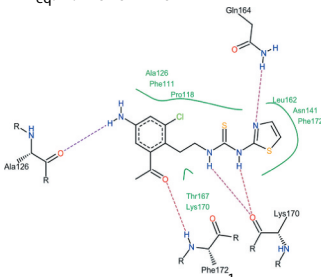
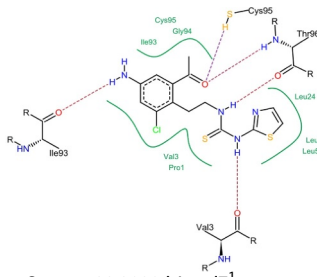
The docking simulation has become an important tool to be used to consider the molecular-binding affinity between phenethylthiazolethiourea derivatives and HIV-1 proteins and SARS-CoV-2. The receptor-inhibitory activity of the newly designed phenethylthiazolethiourea derivative can be determined by introducing the ligand to the active sites of the target protein. The 3D structural models of proteins encoded by the SARS-CoV-2 genome (PDB ID: 6LU7) [18] and HIV-1 (PDB ID: 1ODW) [17] are operated for docking for lead compound C-12 and six new compounds (in Table 5). The binding energy ( $\text{kJ}\cdot\text{mol}^{-1}$ ) of the molecule to the protein receptors during docking simulation is the basis for assessing the activity of a compound. In addition, the biological activity of the new molecules to the proteins SARS-CoV-2 and HIV-1 can also be determined by the magnitude of the equilibrium constant  $K_{\text{eq}}$  of the complex. The equilibrium constant  $K_{\text{eq}}$  can be calculated by

equation (12). The new compounds from C-n1 to C-n6 (shown in Table 5) showed higher predictive anti-HIV-1 activity than the lead compound C-12. In this group of new compounds, the compound C-n6 seems to have much stronger inhibitory activity with HIV-1 than lead compound C-12. These new compounds can be probed for SARS-CoV-2 resistance based on the ability to bind the molecules to the protein-based active site of SARS-CoV-2 as well as HIV-1 by docking simulation. The molecules were docked on the BioSolveIT LeadIt [41] and Discovery Studio Client [42] systems. The maximum result for each iteration step is 1000 and the maximum number of results per ligand fragmentation is 200. The most suitable conformation of each molecule binding to the receptor of the SARS-CoV-2 protein considered is retained by the docking simulation. The molecular conformation is best suited for the lowest docking score ( $\text{Score}/\text{kJ}\cdot\text{mol}^{-1}$ ). This docking score is the total energy consumed for the formation of molecular-binding interactions with the target receptor. This is the basis for identifying an active compound with the ability to bind sustainably with a protein receptor. So it can be said that the docking scores can demonstrate the ability to bind the molecule to the protein receptor. The binding affinity of the molecules to the SARS-CoV-2 protein are proved to be stronger than the HIV-1 protein, as given in Table 6. The C-n6 derivative with the docking score of  $-27.4235 \text{ kJ}\cdot\text{mol}^{-1}$  for the SARS-CoV-2 protein was lower than the docking score of  $-23.6137 \text{ kJ}\cdot\text{mol}^{-1}$  for the HIV-1 protein.

Furthermore, the C-n6 derivative showed that its docking score is much lower than that of the other new substances. Therefore, the C-n6 derivative has much stronger binding affinity than the C-12 lead derivative. This is consistent with the activity obtained from the QSAR<sub>GA-ANN</sub> model (in Table 5). These derivatives turned out that their equilibrium constants  $K_{\text{eq}}$  binding to the protein receptor of SARS-CoV-2 are much greater than the  $K_{\text{eq}}$  constants of the HIV-1 protein. The C-n6 derivative is one of the SARS-CoV-2 inhibitors appears to be a derivative with the greatest potential for inhibiting SARS-CoV-2.

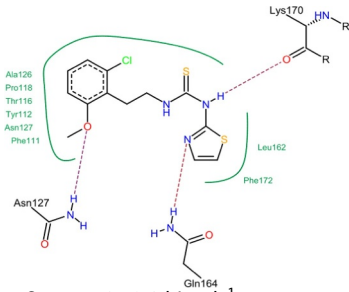
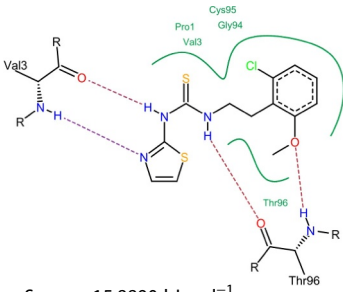
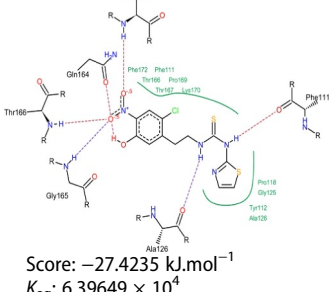
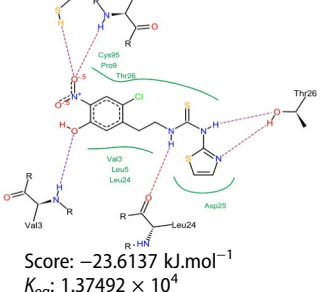
The docking simulation results for the seven compounds in Table 5 were described and analyzed under the systems BioSolveIT LeadIt [41] and Discovery Studio Client [42]. Their receptor-ligand complexes are depicted in Table 6 and Figures 5 and 6. Seven receptor-ligand complexes formed from compounds C-12, C-n1, C-n2, C-n3, C-n4, C-n5 and C-n6 with HIV protein receptors (1ODW) [17], as illustrated in Table 6. The C-12 derivative combined with nine amino acids PRO1, ILE93, VAL3, LEU24, ILE66, CYS67, VAL11, HIS69, CYS95 formed the hydrophobic bonds,

**Table 6.** The binding energies and  $K_{eq}$  equilibrium constants between new derivatives with the receptors of HIV-1 proteinase (PDB ID: 1ODW) and SARS-CoV-2 protein (PDB ID: 6LU7).

Derivatives	SARS-CoV-2 protein (PDB ID: 6LU7).	HIV-1 Proteinase (PDB ID: 1ODW)
C-12	 <p>Score: <math>-16.3802 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 7.42359 \times 10^2</math></p>	 <p>Score: <math>-16.0955 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 6.61791 \times 10^2</math></p>
C-n1	 <p>Score: <math>-16.8894 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 9.11698 \times 10^2</math></p>	 <p>Score: <math>-21.2204 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 5.23425 \times 10^3</math></p>
C-n2	 <p>Score: <math>-23.9224 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 1.55732 \times 10^4</math></p>	 <p>Score: <math>-20.5518 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 3.99653 \times 10^3</math></p>
C-n3	 <p>Score: <math>-20.8173 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 4.44848 \times 10^3</math></p>	 <p>Score: <math>-18.2729 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 1.59334 \times 10^3</math></p>
C-n4	 <p>Score: <math>-22.2446 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 7.91306 \times 10^3</math></p>	 <p>Score: <math>-22.2098 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 7.80272 \times 10^3</math></p>

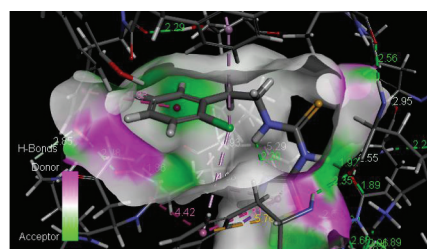
(Continued)

Table 6. (Continued).

Derivatives	SARS-CoV-2 protein (PDB ID: 6LU7).	HIV-1 Proteinase (PDB ID: 1ODW)
C-n5	 <p>Score: <math>-18.4652 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 1.72190 \times 10^3</math></p>	 <p>Score: <math>-15.8890 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 6.08880 \times 10^2</math></p>
C-n6	 <p>Score: <math>-27.4235 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 6.39649 \times 10^4</math></p>	 <p>Score: <math>-23.6137 \text{ kJ.mol}^{-1}</math>  <math>K_{eq}: 1.37492 \times 10^4</math></p>

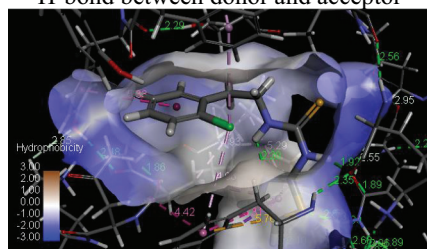
and polar hydrogen bonds are generated with three amino acids THR96, CYS95, PRO1; for the C-n1 derivative, the hydrophobic bonds are formed between four amino acids PRO1, VAL3, CYS95, GLY94 and C-n1, and polar hydrogen bonds are formed by three amino acids ILE93, VAL3, THR96; for the C-n2 derivative, the

hydrophobic bonds are generated with five amino acids PRO1, CYS95, LEU24, VAL3, ILE93, and polar hydrogen bonds are created with four amino acids ILE93, GLY94, THR96, VAL3; the C-n3 derivative combined with five amino acids PRO1, VAL3, CYS95, ILE93, GLY94 to generate the hydrophobic bonds, and polar



a1

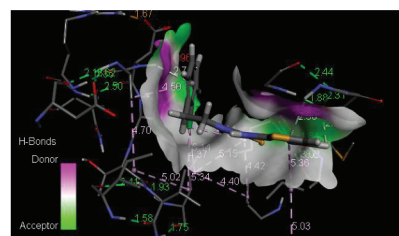
H-bond between donor and acceptor



a2

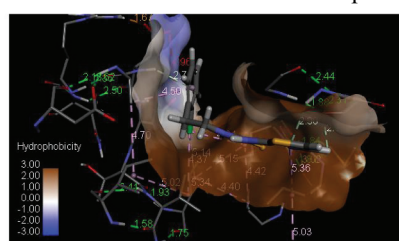
Hydrophobicity

a) SARS-COV-2 (PDB ID: 6LU7).



b1

H-bond between donor and acceptor



b2

Hydrophobicity

b) HIV-1 Proteinase (PDB ID: 1ODW)

Figure 5. Comparison of the interaction properties of the C-12 substance with the protein receptors of a) SARS-COV-2 protein (PDB ID: 6LU7) and b) HIV-1 Proteinase protein (PDB ID: 1ODW).

hydrogen bonds are formed with three amino acids VAL3, ILE93, THR96; the C-n4 derivative generated the hydrophobic bonds by incorporating eight amino acids PRO1, CYS95, GLY94, ILE93, VAL3, LEU24, LEU97, LUE5, and polar hydrogen bonds are combined with four amino acids ILE93, CYS95, THR96, VAL3; the C-n5 derivative exhaled the hydrophobic bonds with five amino acids PRO1, VAL3, CYS95, GLY94, THR96, and polar hydrogen bonds are formed with two amino acids VAL3, THR96; the derivative C-n6 combined with seven amino acids PRO9, CYS95, THR26, VAL3, LEU5, LEU24, ASP25 to create the hydrophobic bonds; and with five amino acids CYS95, VAL3, LEU24, THR26, THR96 generated the polar hydrogen bonds. The binding affinity of compound C-12 and new derivatives C-n1 to C-n6 for HIV-1 target protein (1ODW) resulted in docking points ranging from  $-15.889 \text{ kJ.mol}^{-1}$  to  $-23.6137 \text{ kJ.mol}^{-1}$ , and the equilibrium  $K_{eq}$  constants of the ligand-receptor complexes range from  $6.0888 \times 10^2$  to  $1.3749 \times 10^4$ .

The structures of the seven receptor-ligand complexes produced by the target protein SARS-CoV-2 (6LU7) with C-12, C-n1, C-n2, C-n3, C-n4, C-n5 and C-n6 were analyzed. With the C-12 derivative, six amino acids PRO169, LYS170, GLY171, PHE111, PHE172, TYR112 are involved in the formation of hydrophobic interaction, and three amino acids ALA126, ASN127, PHE111 formed the polar hydrogen bonds; with the C-n1 compound, seven amino acids ASN127, PHE172, GLN182, TYR110, TYR112, THR116 and PHE111 are involved to form the hydrophobic interactions, and the polar hydrogen bonds are made up of four amino acids ALA174, GLN182, PHE111, and ALA126; with the C-n2 compound, nine amino acids PRO118, LYS170, PHE172, THR116, TYR112, PHE111, LEU162, ALA126, and GLN164 formed the hydrophobic interactions with this ligand, and three amino acids PRO118, LYS170 and GLN164 create the polar hydrogen bonds; with the C-n3 derivative, seven amino acids PRO118, THR167, PHE172, PHE111, LYS170, TYR112, and ALA126 have hydrophobic interactions with this ligand; and the polar hydrogen bonds are produced by three amino acids THR167, PHE111 and ALA126; with the C-n4 derivative, eight amino acids PRO118, ALA126, PHE111, LEU162, ASN141, PHE172, THR167, LYS170 formed the hydrophobic interactions with this derivative, and the polar hydrogen bonds created by four amino acids ALA126, GLN164, PHE172, LYS170; with the C-n5 derivative, eight amino acids PRO118, ALA126, THR116, TYR112, ASN127, PHE111, LEU162, PHE172 formed the hydrophobic interactions, and four amino acids ASN127, GLN164, LYS170 formed

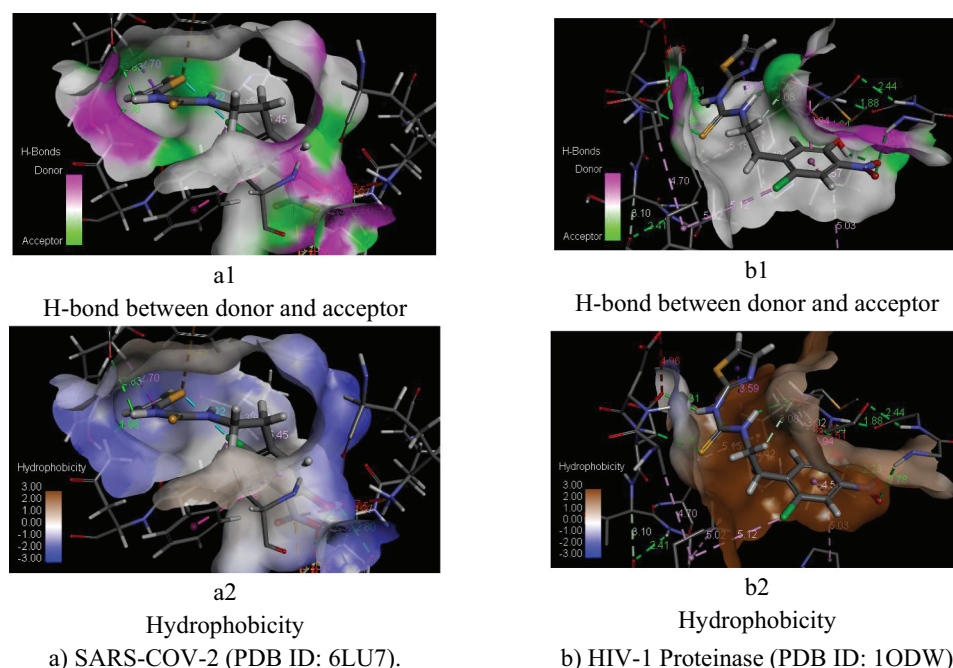
polar hydrogen bonds; with the C-n6 compound, ten amino acids PRO169, PHE111, PHE172, THR166, THR167, LYS170, PRO118, GLY125, TYR112, and ALA126 created the hydrophobic interactions, and six amino acids PHE111, THR167, GLN164, THR166, GLY165 and ALA126 formed the polar hydrogen bond with this ligand. For the SARS-CoV-2 target protein (6LU7), the docking points that demonstrate the binding affinity between the ligand and the protein receptor are in the range  $-16.3802 \text{ kJ.mol}^{-1}$  to  $-27.4235 \text{ kJ.mol}^{-1}$ , and the  $K_{eq}$  equilibrium constants of the ligand-receptor complexes range from  $7.4236 \times 10^2$  to  $6.39648910 \times 10^4$ , as shown in Table 6.

The  $\text{pIC}_{50}$  activities and the docking scores of the newly designed compounds C-n1 to C-n6 resulting from the QSAR<sub>GA-ANN</sub> model (in Table 5) and the docking scores exhibited the binding affinity of the ligands with the receptor, as seen in Table 6. These demonstrated that the new derivatives have a stronger inhibitory ability for HIV-1 (1ODW). This proved the potential for inhibiting effectively SARS-CoV-2 (6LU7).

In addition, the ability of ligands to firmly bind to the active receptors is also somewhat influenced by their electrostatic properties. All new molecules can form more polar hydrogen bonds and hydrophobic interactions with the amino acids of the SARS-CoV-2 protein than with the HIV-1 protein. Also, the electrostatic density of the entire C-n6 ligand changes more when the substituents on this molecule are substituted, as described in Figures 5 and 6. The polar hydrogen bonds and hydrophobic interactions of this ligand are greater than other new ligands.

The C-n6 derivative in (Table 5) gives the strongest predictive activity, which is also reflected in the molecular properties obtained from the docking simulations. The electrostatic potential surrounding this ligand was also calculated, as seen in Figures 5 and 6. From the analysis and comparison of docking results between two substances C-12 and C-n6 on the HIV-1 and SARS-CoV-2 protein receptors, it is demonstrated that there is a clear difference between the binding properties of these compounds.

The hydrophobic and hydrogen bonding properties obtained from the docking simulation are described in Figures 5 and 6. The hydrophobic factor depends on the nature of the molecular structure and the ability to interact in the surrounding environment. The hydrophobic properties of the C-n6 and C-12 derivatives at the receptor site of SARS-CoV-2 differ greatly from them in the receptor of the HIV-1 protein. Hydrophobicity produces a colored surface due to the hydrophobicity of the receptor residue, from blue to hydrophilic to brown to hydrophobic. The compounds



**Figure 6.** Comparison of the interaction properties of the C-n6 substance with the protein receptors of a) SARS-COV-2 protein (PDB ID: 6LU7) and b) HIV-1 Proteinase protein (PDB ID: 1ODW).

C-n6 and C-12 in the SARS-CoV-2 receptor exhibit more hydrophilic properties compared to the HIV-1 protein receptor. This indicates that the molecules are well soluble in water and are more flexible in their receptor spatial configuration. This will induce the stronger binding to the SARS-CoV-2 receptor by forming more binding types in the receptor space. In the receptor of HIV-1 protein, the C-12 and C-n6 derivatives exhibit a stronger hydrophobicity. Thus the ability to dissolve in water is limited. This results in compounds being less flexible and difficult to form many types of bonds in the receptor space.

The same goes for the hydrogen bond between the ligand and the receptor. Hydrogen bonding produces a surface colored in a hydrogen bonding fashion, with the donor group showing green and the acceptor group showing a green-pink color, as shown in Figures 5 and 6. We can also observe that the C-12 and C-n6 ligands

on the SARS-CoV-2 receptor have more hydrogen-binding donor and acceptor groups than on the HIV-1 protein receptor. This also demonstrates that the C-12 and C-n6 ligands are more strongly bound to the SARS-CoV-2 receptor and more stable than the HIV-1 protein receptor binding. The hydrophobic and hydrogen bonding properties also demonstrated that the C-n6 compound has a stronger inhibitory activity on SARS-CoV-2 than to inhibit HIV-1 protein. In general the C-n6 structure indicated the best binding ability on both HIV-1 and SARS-CoV-2 protein receptors. The C-n6 compound illustrated is an effective derivative for SARS-CoV-2 inhibition.

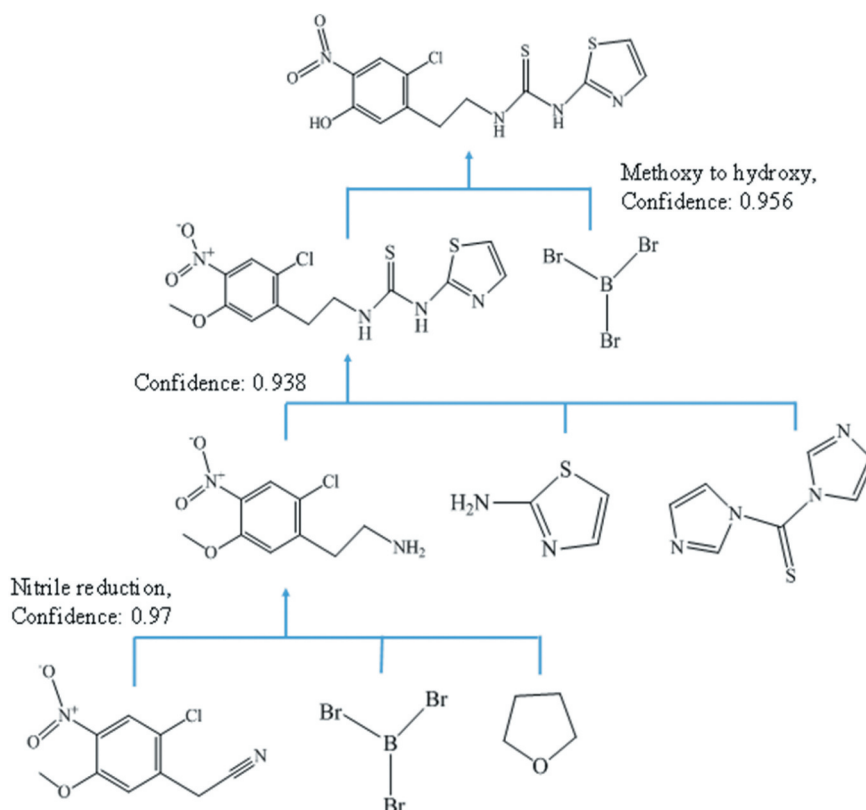
### 3.3. Discussion

Here we discuss the importance of each new derivative for HIV-1 and SARS-CoV-2. The current drug design is

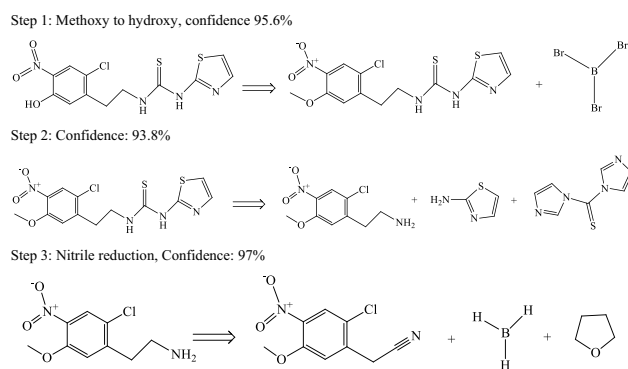
**Table 7.** The values  $pIC_{50}$  and equilibrium constants  $K_{eq}$  of new derivatives binding the receptor of SARS-CoV-2 protein (PDB ID: 6LU7) derived from the QSAR<sub>GA-ANN</sub> model and docking simulation.

Derivatives	$pIC_{50}$ for SARS-CoV-2		docking simulation		docking simulation [43]	
	QSAR <sub>GA-ANN</sub>	2D-QSAR [43]	$\Delta G/kJ.mol^{-1}$	$K_{eq}, cal$	$\Delta G/kJ.mol^{-1}$	$K_{eq}$
C-12	-1.625		-16.380	$7.424 \times 10^2$		
C-n5	-3.594		-18.465	$1.722 \times 10^3$		
C-n6	-3.894		-27.424	$6.397 \times 10^4$		
Pyridine 1	-3.663	-3.035	-33.944	$8.886 \times 10^5$	-33.787	$8.341 \times 10^5$
Thiophene 4	-3.245	-3.446	-34.186	$9.797 \times 10^5$	-32.657	$5.286 \times 10^5$
Thiophene 5	-3.195	-3.463	-31.513	$3.332 \times 10^5$	-29.601	$1.540 \times 10^5$
Thiophene 83	-3.946	-4.565	-25.523	$2.970 \times 10^4$	-23.530	$1.329 \times 10^4$
Pyridine 84	-3.709	-4.607	-20.160	$3.412 \times 10^3$	-19.762	$2.905 \times 10^3$

\* $E(kJ) = 4.184 \times E(kcal)$



**Figure 7.** Diagram of retrosynthesis for new derivative C-n6.



**Figure 8.** Retrosynthesis analysis to convert to C-n6.

aiming at the appropriate design of drugs that bring many activities. In this study, we design anti-SARS-CoV-2 drugs from anti-HIV-1 derivatives combining QSAR modeling technique with docking simulation technique. This will reduce the time it takes to perform the virtual screening from a large database. We only need to focus on enhancing the anti-HIV-1 activity of the phenethylthiazolethiourea derivative group together with the inhibition of the SARS-CoV-2 virus. The new

derivatives C-n1 to C-n6 have been designed, as shown in Table 5. These new compounds exhibited higher anti-HIV-1 activity against the C-12 model derivative as well as their high inhibitory affinity for the SAR-CoV-2 virus. We can have more detailed discussions on the following techniques:

As we all know, a molecule entering a protein receptor can be affected by spatial configuration. To fully evaluate the effect of the molecular structure, we have successfully built a database of 2D and 3D molecular descriptors [20,21]. In some previous studies on development of SARS-CoV-2 inhibitors, 2D descriptors have been used for the purpose of developing a 2D-QSAR model suggested by V. Kumar et al. (2020) [43], T. Bobrowskia et al. (2020) [44] and Sk.A Amin et al. (2020) [45]. The 2D-QSAR models are used to interpret and can be used to quickly predict the SAR-CoV-2 inhibition of a derivative based on a linear regression model (MLR) [43–46]. These 2D-QSAR models have also had some success in predicting and developing the derivatives nPyridines and nThiophenes that inhibit SARS-CoV [43]. The 2D parameters describe the flatness of the molecule. But in fact the molecules can turn around a single bond. This can produce the molecules that include the 3D structural

properties. Therefore, in this study we have investigated a more complete set of molecular descriptors including 2D and 3D descriptors. The genetic algorithms have assisted in selecting a globally optimized set of descriptors. This allowed to build the optimal QSAR<sub>GA-MLR</sub> model from the multivariate linear regression algorithm. Meanwhile, the 2D molecular descriptors in the 2D-QSAR models was developed by Kumar et al. [43], Bobrowskia et al. (2020) [44] and Sk.A Amin et al. (2020) [45] did not determine the contribution effect of the 3D molecular descriptors to SARS-CoV-2 inhibitory activity. This will be difficult to navigate when designing a fully new molecule. In another direction, the QSAR model is constructed from physicochemical parameters related to the structure with Monte Carlo optimization technical assistance given by K. Ghosh et al. [46]. This QSAR model also provides significant results for the classification and prediction of compound activity. This is not enough to support a more rational drug design for SARS-CoV-2 inhibition. In this work, the 2D, 3D molecular descriptors in the QSAR<sub>GA-MLR</sub> model allowed to fully calculate the contribution of each descriptor to the activity. The influence of the set of descriptors on the activity can be ranked (as seen in Figure 3). The design orientation in C-n1 to C-n6 derivatives has yielded better activity results than the lead derivative C-12 (Table 5). Molecules in the phenethyl thiazole thiourea group carry a 2-methylthiazole 5-sided heterocycle and a bonding group >C = S. These are important molecular structural parts that are similar to those studied by Kumar et al. [43]. and Sk.A Amin et al. (2020) [45].

The QSAR<sub>GA-MLR</sub> model has been built and tested for reliability. But the QSAR<sub>GA-MLR</sub> model is just a linear model anyway. Therefore, the predictive ability of the QSAR<sub>GA-MLR</sub> model cannot meet the desired requirements. The predictive quality of the QSAR<sub>GA-MLR</sub> linear model is depicted in Figure 4 and at its statistical values  $R^2 = 0.778$ ,  $R^2_{adj} = 0.722$  and  $Q^2_{LOO} = 0.701$ . We can also see that the 2D-QSAR model was developed by Kumar et al. [43], Bobrowskia et al. (2020) [44] and Sk.A Amin et al. (2020) [45] is also based on multivariate regression techniques. In general, the predictive quality of the 2D-QSAR model is not high because it is simply performed using regression techniques as given by Kumar et al. [43] and Bobrowskia et al. (2020) [44]. To improve the QSAR model quality, Sk.A. Amin et al. [45] and K. Ghosh [46] also built the QSAR models using the genetic algorithms and machine learning. In our study, the QSAR<sub>GA-MLR</sub> model has also improved predictive quality using the machine

learning regression technique Boosted Trees for Regression (QSAR<sub>GA-BTR</sub>) and the neural network (QSAR<sub>GA-ANN</sub>). We have successfully built hybrid QSAR models. The input variables used for machine learning are also a set of 2D and 3D molecular descriptions in the QSAR<sub>GA-MLR</sub> model. The correlation coefficient values of the hybrid QSAR models increased from  $R^2 = 0.7417$  for QSAR<sub>GA-MLR</sub> to  $R^2 = 0.9058$  for QSAR<sub>GA-BTR</sub> and  $R^2 = 0.9873$  for QSAR<sub>GA-ANN</sub>. These hybrid QSAR<sub>GA-BTR</sub> and QSAR<sub>GA-ANN</sub> models built here have better predictability than the corresponding

QSAR<sub>GA-MLR</sub> model. The strong ability of the hybrid QSAR<sub>GA-ANN</sub> (HQSAR<sub>GA-ANN</sub>) model with the architecture I(5)-HL(9)-O(1) has been demonstrated by the correlation coefficient for the training set, the validation set and the test set, as depicted in Figure 4. The correlation coefficient of training set  $R^2$  and validation set  $Q^2$  of the

QSAR<sub>GA-ANN</sub> model is higher than the QSAR models built by Kumar et al. [43], Bobrowskia et al. (2020) [44], Sk.A. Amin et al. [45] and K. Ghosh [46]. The QSAR<sub>GA-ANN</sub> model used to predict pIC<sub>50</sub> of the groups Pyridine and Thiophene in the study of Kumar et al. [43]. Furthermore, the  $K_{eq}$  values of the compounds in the Pyridine and Thiophene groups were also determined from docking simulation, as shown in Table 7. The pIC<sub>50</sub> values of the derivatives C-12, C-n5 and C-n6 for SARS-CoV-2 predicted from the QSAR<sub>GA-ANN</sub> model are also consistent with the stable constants  $K_{eq}$  obtained from docking simulation for SARS-CoV-2, also seen in Table 6. Moreover, we can see that the pIC<sub>50</sub> values of Pyridine and Thiophene compounds obtained from QSAR<sub>GA-ANN</sub> model are also correlated with those obtained from Kumar et al. [43]. But the predictive results for the Pyridine and Thiophene compounds from

QSAR<sub>GA-ANN</sub> model showed that SARS-CoV-2 inhibitory activity is apparently higher than those obtained from the 2D-QSAR model [43]. This is also exhibited by the  $K_{eq}$  values obtained from the docking simulation. The binding energies of these substances to the SARS-CoV-2 receptors are lower than those obtained from docking simulations [43]. In this work, the hybrid QSAR<sub>GA-ANN</sub> model has shown its ability to train and prediction to respond well to different data types. This compound C-n6 has encapsulated the success of the research.

### 3.4. Planning New Derivative Synthesis

The rules for the synthesis of the selected substances are given according to the reaction diagram in Figure 7. The core of the reaction is defined as consisting of atoms or bonds that can be changed by switching from a reactant to a product. This is done by comparing reactants and products. The process expands the core of the reaction to accommodate neighboring atoms or functional groups, as depicted in Figures 7 and 8 [36]. The group of reactive fragments at the core of the reaction can be divided into different groups. The synthesized plan could be generalized into the common rules for each group. With this in mind, the retrosynthesis analysis of the new compound C-n6 aims to present a planning diagram for the synthesis of the C-n6 derivative 1-(2-chloro-5-hydroxy-4-nitrophenethyl)-3-(thiazol-2-yl)thiourea. The new algorithms in retrosynthesis analysis are supported powerfully by computer techniques that the synthesis plans generated are faster and more detailed. This is the method that can support the rapid synthesis of new drugs with promising prospects.

To synthesize the new substance C-n6, 1-(2-chloro-5-hydroxy-4-nitrophenethyl)-3-(thiazol-2-yl)thiourea, the available starting materials can be quickly retrieved. This process has been done with the retrosynthesis tree selection starting with the design of a synthesis plan, as shown in Figures 7 and 8. This retrosynthesis analysis is executed until a suitable synthesis tree is found to gradually convert C-n6 to simpler structures (as shown in Figure 7).

The retrosynthesis analysis for C-n6 is done in three stages: In the first stage, the retrosynthesis analysis will perform the structural analysis at the locations where the bonds can be changed for re-synthesis with 95.6% confidence [36]; in the second stage, the re-synthesis can be performed on a simpler compound that is produced in the first stage. The easily modifiable bonds of this derivative can analyze to simpler substances with 93.8% confidence; in the final stage, the retrosynthesis analysis can be stopped at the starting materials that are available to synthesize for a 97% confidence level [35,37]. A retrosynthesis tree that can assist organic synthesizers has been developed for the target substance C-n6 to drive the ideas of synthesis quickly, as exhibited in Figures 7 and 8.

## 4. CONCLUSION

The bioactive phenethylthiazolethiourea derivatives against HIV-1 RT were used to study the construction of *in silico* models QSAR. The electronic molecular

descriptors were calculated using the theoretical methods for optimal structures. The QSAR models were developed successfully for a group of HIV-1 inhibitory phenethylthiazolethiourea compounds. The QSAR models are the basis for the development and successful prediction of anti-HIV-1 pIC<sub>50</sub> inhibitory activity of novel phenethylthiazolethiourea compounds. The model QSAR<sub>GA-ANN</sub> I(5)-HL(9)-O(1) gave the most reliable prediction results of the compounds. The prediction results of pIC<sub>50</sub> activity of new compounds received from the QSAR<sub>GA-ANN</sub> model agree well with the results obtained from the docking simulation process.

The new molecules from C-n1 to C-n6 were docked successfully to the protein receptors of HIV-1 and SARS-CoV-2 using the docking simulation. It has demonstrated the effective inhibitory activity of novel agents against SARS-CoV-2. The new compounds demonstrated the higher inhibitory activities on HIV-1 than the lead compound C-12, while also exhibiting much higher activity against SARS-CoV-2.

In this research, the design directions of the SARS-CoV-2 inhibitors using the molecular descriptors from theoretical calculations to build the QSAR models are reliable and effective. Moreover, by this theoretical method, we have successfully built the retrosynthesis tree for the new substance C-n6: 1-(2-chloro-5-hydroxy-4-nitrophenethyl)-3-(thiazol-2-yl)thiourea. It is the highest inhibitor for SARS-CoV-2 with the lowest binding energy  $-27.4235$  kJ. mol<sup>-1</sup> and the highest equilibrium constant  $K_{eq}$   $6.39649 \times 10^4$  compared to others. The C-n6 substance provides a new retrosynthesis orientation that meets the requirements for the development of new drugs that are resistant to SARS-CoV-2.

## Acknowledgments

To complete this project, we have received a lot of support from my Prof. Dr. U. K. Deiters at the Institute of Physical Chemistry, Cologne University, Germany. He enthusiastically conveyed me a lot of experience and valuable knowledge in the field of theoretical calculations during my time studying at the Institute of Physical Chemistry. I sincerely thank my Prof. Dr. U. K. Deiters very much.

To complete this project, we have received all the necessary support conditions from Hoa Sen University, Viet Nam. .

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## ORCID

Pham Van Tat  <http://orcid.org/0000-0002-5143-6299>

Tran Thai Hoa  <http://orcid.org/0000-0002-8712-6233>

## References

- [1] Zhang D, Zhang B, Jin-Tao L, et al., The clinical benefits of Chinese patent medicines against COVID-19 based on current evidence., *Pharmacol Res.* **2020**;157:104882.
- [2] Kelleni MT. Nitazoxanide/azithromycin combination for COVID-19: a suggested new protocol for early management. *Pharmacol Res.* **2020**;157:104874.
- [3] McKee DL, Sternberg A, Stange U, et al. Candidate drugs against SARS-CoV-2 and COVID-19. *Pharmacol Res.* **2020**;157:104859.
- [4] Yang R, Liu H, Bai C, et al. Chemical composition and pharmacological mechanism of qingfei paidu decoction and Ma Xing Shi Gan decoction against coronavirus disease 2019 (COVID-19): *in silico* and experimental study. *Pharmacol Res.* **2020**;157:104820.
- [5] Zhao Q, Meng M, Kumar R, et al. Lymphopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: a systemic review and meta-analysis. *Inter J Infect Dis.* **2020**;96:131–135.
- [6] Shang J, Gang Y, Shi K, et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature.* 2020 May 14;581 (7807). DOI:10.1038/s41586-020-2179-y.
- [7] Kalyan G, Amin A, Gayen S, et al. Chemical-informatics approach to COVID-19 drug discovery: exploration of important fragments and data mining based prediction of some hits from natural origins as main protease (Mpro) inhibitors. *J Mol Struct.* 2021 Jan 15; 1224:129026..
- [8] Norinder U, Tuck A, Norgren K, et al. Existing highly accumulating lysosomotropic drugs with potential for repurposing to target COVID-19. *Biomed Pharmacother.* Oct 2020; 130:110582..
- [9] Ejub GW, Fonkem C, Tadjouteu Assatse Y, et al. Study of the structural, chemical descriptors and optoelectronic properties of the drugs hydroxychloroquine and azithromycin. *Heliyon.* Aug 2020 6;(8) e04647.
- [10] Negi M, Chawla PA, Faruk A, et al. Role of heterocyclic compounds in SARS and SARS CoV-2 pandemic. *Bioorg Chem.* Nov 2020; 104:104315..
- [11] Bui Thi Phuong T, Tran Thi Ai M, Nguyen Thi Thanh H, et al. Investigation into SARS-CoV-2 resistance of compounds in garlic essential oil. *ACS Omega.* **2020**;5(14):8312–8320. .
- [12] Tran Thi Ai M, Huynh Thi Phuong L, Nguyen Thi Thanh H, et al. Evaluation of the inhibitory activities of COVID-19 of melaleuca cajuputi oil using docking simulation. *J. ChemistrySelect.* . 2020 Jun 8;5 (21):6312–6320.
- [13] Bui TQ, Huynh Thi Phuong L, Tran Thi Ai M, et al. A density functional theory study on silver and bis-silver complexes with lighter tetraylene: are silver and bis-silver carbenes candidates for SARS-CoV-2 inhibition? insight from molecular docking simulation. *RSC Adv.* **2020**;10:30961–30974.
- [14] Bell FW, Cantrell AS, Marita Hogberg S, et al. Phenethylthiazolethiourea (PEW) compounds, a new class of HIV-1 reverse transcriptase inhibitors. 1. synthesis and basic structure-activity relationship studies of PEW analogs. *J Med Chem.* **1995**;38(25):4929–4936. .
- [15] Roy. K. *Advances in QSAR Modeling Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental.* Gewerbestrasse, Cham, Switzerland: Springer International Publishing AG; **2017**.11, 6330.
- [16] Engel T, Gasteiger. J. *Applied Chemoinformatics: achievements and Future Opportunities.* Weinheim, Germany: Wiley-VCH Verlag GmbH; **2018**.
- [17] Thanki N, Kervinen J, Wlodawer A. Native HIV-1 Proteinase. *Protein Pept Lett.* **1996**;3:399.
- [18] Jin Z, Xiaoyu D, Yechun X, et al.; Show fewer authors. The crystal structure of COVID-19 main protease in complex with an inhibitor N3. *Nature.* **2020**;582 (7811):289–293. .
- [19] Lowell H. Hall, Lemont B. Kier, L. Mark Hall., *QSARIS 1.1.* SciVision. Inc, USA: Statistical Solutions Ltd; **2001**.
- [20] Joseph R. Votano and Scott Lee., *QSARIS Reference Guide: statistical Analysis and Molecular Descriptors.* San Diego, USA: Academic Press; **2000**.
- [21] Tat PV. *Development of QSAR and QSPR.* Hanoi: Publisher of Natural sciences and Technique; **2009**.
- [22] Montgomery DC, Peck EA, Vining CG. *Introduction to Linear Regression Analysis.* Third Edition ed. New York: Wiley-Interscience; **2001**.
- [23] Quang NM, Mau TX, Nguyen Thi Ai N, et al. Novel QSPR modeling of stability constants of metalthiosemicarbazone complexes by hybrid multivariate technique: GA-MLR, GA-SVR and GA-ANN. *J. Molecular Structure.* **2019**;1195:95–109.
- [24] Lee C-F, Lee J, Chang J-R, et al. *Essentials of Excel, Excel VBA, SAS and Minitab for Statistical and Financial Analyses.* Switzerland: Springer International Publishing; **2016**.
- [25] Frances B. *Genetic Algorithms and Machine Learning for Programmers.* Raleigh, North Carolina: Andy Hunt; Jan 2019.
- [26] Shin Y, Application of boosting regression trees to preliminary cost estimation in building construction projects. *Comput Intell Neurosci.* 2015; 2015: (149702):9.
- [27] Foroughi M, Mohammad Hossein Ahmadi A, Kakhki S. Bio-inspired, high, and fast adsorption of tetracycline from aqueous media using Fe<sub>3</sub>O<sub>4</sub>-g-CN@PEI-β-CD nanocomposite: modeling by response surface methodology (RSM), boosted regression tree (BRT), and general regression neural network (GRNN). *J Hazard Mater.* 2020 Apr 15;388:121769..
- [28] Podio NS, Baroni MV, Wunderlin DA. Relation between polyphenol profile and antioxidant capacity of different Argentinean wheat varieties. A boosted regression trees study. *Food Chem.* Oct 2017;232:79–88. . 79-88.
- [29] Martin Porter., *JMP® 14 Fitting Linear Models.* Cary, North Carolina, USA: SAS Institute Inc; **2018**.
- [30] Hastie T, Tibshirani R, Friedman. J. *The Elements of Statistical Learning: data Mining, Inference, and Prediction.* 2nd Edition. Springer-Verlag; New York, 763. Feb 2009.

- [31] Neil Hodgson., JMP® 14 Predictive and Specialized Modeling. Cary, North Carolina, USA: SAS Institute Inc; 2018.
- [32] Dehmer M, Varmuza K, Bonchev D. Statistical Modelling of Molecular Descriptors in QSAR/QSPR. Weinheim, Germany: Wiley-VCH Verlag & Co. KGaA; 2012.
- [33] Peter A, Julio DP. Physical Chemistry. Freeman WH, Sixth Edition, Oxford University Press, 2012 Dec 19.
- [34] Fatahala. SS. Retrosynthesis analysis; a way to design a retrosynthesis map for pyridine and pyrimidine ring. Ann Adv Chem. 2017;1(2):057–060.
- [35] Šunjić V, Peroković. VP. Organic Chemistry from Retrosynthesis to Asymmetric Synthesis. Switzerland: Springer International Publishing ; 2016.
- [36] Olivier P. Retrosynthetic Analysis and Synthesis of Natural Products 1: synthetic Methods and Applications. Hoboken, USA: John Wiley & Sons, Inc; 2019.
- [37] Wei JN, Duvenaud D, Alán A-G. Neural networks for the prediction of organic chemistry reactions. ACS Cent. Sci. 2016;2(10):725–732.
- [38] Stathakis. D. How many hidden layers and nodes?. Int J Remote Sens. 2009 Apr 20; 30(8):2133–2147.
- [39] Huang G-B. Learning capability and storage capacity of two-hidden-layer feedforward networks. IEEE Trans Neural Networks. 2003;14(2):274–281.
- [40] Jones G, Willett P, Glen RC, et al. Development and validation of a genetic algorithm for flexible docking. J Mol Biol. 1997;267(3):727–748.
- [41] Rarey M, Wefing. S. FlexX Protein-Ligand Docker: user & Technical Reference as Part of LeadIT 2.3. An der Ziegelei, Augustin, Germany: BioSolveIT GmbH; 2017.
- [42] BIOVIA, Dassault Systemes, San Diego, Discovery studio modeling environment, 2020..
- [43] Kumar V, Roy. K. Development of a simple, interpretable and easily transferable QSAR model for quick screening antiviral databases in search of novel 3C-like protease (3CLpro) enzyme inhibitors against SARS-CoV diseases. SAR QSAR Environ Res. 2020 Jun 16;31(7):511–526.
- [44] Bobrowski T, Alves V, Melo-Filho CC, et al. Computational models identify several FDA approved or experimental drugs as putative agents against SARS-CoV-2. Chemrxiv. 2020 Apr 22. DOI:10.26434/chemrxiv.12153594.
- [45] Amin SA, Banerjee S, Singh S, et al. First structure–activity relationship analysis of SARS-CoV-2 virus main protease (Mpro) inhibitors: an endeavor on COVID-19 drug discovery. Mol Divers. 2021 Jan 05. DOI:10.1007/s11030-020-10166-3.
- [46] Ghosh K, Amin SA, Gayen S, et al. Chemical-informatics approach to COVID-19 drug discovery: exploration of important fragments and data mining based prediction of some hits from natural origins as main protease (Mpro) inhibitors. J. Molecular Structure. 2021 Jan 15;1224:129026.