Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 7, No. 2, May, Pp. 149-160

QSAR study on aromatic disulfide compounds as SARS-CoV Mpro inhibitor using genetic algorithm-support vector machine

Rizki Amanullah Hakim[1], Annisa Aditsania[2], Isman Kurniawan[3*]
School of Computing, Telkom University, Indonesia[1]
Research Center of Human Centric Engineering, Telkom University[2,3]

Article Info

Keywords:
Main Protease (Mpro), Quantitative Structure-Activity Relationship (QSAR), Genetic Algorithm, Support Vector Machine (SVM)

Article history:
Received: March 24, 2022
Accepted: April 14, 2022
Published: May 31, 2022

Cite:
R. A. Hakim, A. . Aditsania, and I. Kurniawan, "QSAR Study on Aromatic Disulfide Compounds as SARS-CoV Mpro Inhibitor Using Genetic Algorithm-Support Vector Machine", *KINETIK*, vol. 7, no. 2, May. 2022. https://doi.org/10.22219/kinetik.v7i2.1428

*Corresponding author.
Isman Kurniawan
E-mail address:
ismankrn@telkomuniversity.ac.id

Abstract

COVID-19 is a type of pneumonia caused by the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). This virus causes severe acute respiratory syndrome and 2 million active cases of COVID-19 have been found worldwide. A new strain of the SARS-CoV-2 virus emerged that proved to be more virulent than its predecessor. Regarding the design of a new inhibitor for this strain, SARS-CoV Main Protease (Mpro) was used as the target inhibitor. In the in silico development, the Quantitative Structure-Activity Relationship (QSAR) method is commonly used to predict the biological activity of unknown compounds to improve the process of drug design of a disease, including COVID-19. In this study, we aim to develop a QSAR model to predict the activity of aromatic disulfide compounds as SARS-CoV Mpro inhibitors using Genetic Algorithm (GA) – Support Vector Machine (SVM). GA was used for feature selection, while SVM was used for model prediction. The used dataset is set of features of aromatic disulfide compounds, along with information on the toxicity activity. We found that the best SVM model was obtained through the implementation of the polynomial kernel with the value of R^2_{train} and R^2_{test} scores are 0.952 and 0.676, respectively.

1. Introduction

Since its first appearance in November 2019 in Hubei City, Wuhan Province, China. Coronavirus disease (COVID-19) is known as a real threat and challenge to the modern world. This coronavirus is known to have significant consequences in infecting humans, including bronchiolitis and pneumonia. It can even be involved in otitis, asthma, diarrhea, and neurological diseases [2][3][4]. The 'pandemic' virus causes severe acute respiratory syndrome and more than 2 million active cases are found worldwide. According to many experts, this virus is very dangerous and able to cause one type of pneumonia that can threaten the life of sufferers, namely Severe Acute Respiratory Syndrome or abbreviated SARS. SARS is a highly infectious respiratory disease and is caused by the family of SARS coronavirus (SARS-CoV) [2][5][6][7]. Health organizations in various countries are particularly concerned and wary of the rate of spread of the COVID-19 coronavirus that becomes a respiratory pathogen in humans, coupled with the discovery of a vaccine that has proven strong in controlling the virus that continues to mutate [2][5][6].

Although SARS-CoV-1 was successfully controlled in 2003, the potential risk of stronger strain of SARS-CoV is inevitable. There is even a new strain of the SARS Coronavirus 2 (SARS-CoV-2) proven to be more virulent than before [8]. SARS-CoV Mpro or main protease (main protease) is one of the enzymes that have a key role in processing polyproteins and is active in dimetric form. This enzyme is a therapeutic target shaped like 3-chymotrypsin cysteine protease (3CLpro) or called the main protease (Mpro). These major proteases have a major influence in reducing the risk of drug resistance mutations and displaying antiviral activity with a broad spectrum [10]. There are various experimental data on the inhibition activity in SARS-CoV-2 against components of organic compounds [8].

There, have been many reports that some aromatic disulfide compounds may exhibit antiviral activity in some cases. For example, decreased antiviral activity in NSC4492 compounds used in in vitro research, in the synthesis of Junín Virus RNA (JUNV) which is one type of arenavirus [9]. In the experiment, more than 99.0% decreased the titer of the virus in incubation with virion at 37 °C for 90 minutes. The potential perspective of NSC4492 as an aromatic disulfide compound that can be inactivating, as well as being able to be used as an arenavirus pathogen is very likely to be discussed in future research. For the record, the antiviral ability in arenaviruses will be different from coronavirus. Binding activity in disulfide compounds has a key role in the selection of aspects of proper bioactive protein folding [9].

To support the selection of aspects of these compounds, a model is needed that can provide the results of predicting the activity of compounds to be used as drug candidates, one such method is a quantitative structure-activity relationship (QSAR). The use of QSAR was able to produce a model that could show the relationship between individual compounds and their biological activity, by applying mathematical calculations based on the biological activity of a compound with structural characteristics of the compound[10].

One example of QSAR use is in a study conducted in 2020, where Andrey A. Toropova and colleagues conducted a study to find potential inhibitor compounds in Mpro SARS-CoV using QSAR. In the study, asymmetric aromatic disulfide compounds were used. Molecular docking is used to test the inhibition effect of molecules that have been designed on QSAR; because calculations performed on the strength of molecular bonds can correlate with potential inhibition [12], [13]. From the study, obtained a "scoring" result from each method Molecular Docking, ReRank, and Van der Waals; obtained R2 scores reached 95.6%, 94.4%, and 91.7% respectively. The results showed there was a relationship and uniformity of the results of the QSAR model. Confirming that the way the target is taken resulted in an accuracy rate of 50% on the predicted activity (IC50) inhibitors of SARS-CoV Mpro [13]. From the study, asymmetric aromatic disulfide compounds were very potent in showing antiviral activity against Mpro SARS-CoV [14]. It is also supported by research conducted on derivatives of aromatic disulfide compounds substituted with strong electrons and the volume and size of appropriate substituents can improve the inhibitory ability of these compound components[15]. This suggests aromatic disulfides, especially asymmetrical forms, and their derivatives, could be used as a basis in developing more effective SARS-CoV inhibitors.

In 2020, Azmi et al conducted a QSAR study on fusidic acid as a Malaria inhibitor agent [24]. In the study, they took a data processing approach based on statistical analysis and genetic algorithmic in processing molecular descriptors. Processing is done by reducing descriptors using a method of removal based on standard deviation, along with the removal of weakly correlated descriptors, and does not correlate with the main response target [24]. Then the implementation of the genetic algorithm as the determinant of the best descriptor combination, which will be used as the best descriptor solution.

The main goal of this research was to develop a QSAR model to predict the activity of aromatic disulfide compounds as SARS-CoV Mpro inhibitors using Genetic Algorithm (GA) – Support Vector Machine (SVM). GA was used for feature selection, while SVM was used for predicting accurate outputs for impairing SARS-CoV Mpro and can be useful in the development of SARS-CoV-2 antiviral agents. GA was used in the feature selection because it is frequently used for reducing the number of feature as part of the QSAR in-silico research [24] and SVM was used as the model prediction since the model is known as one of machine learning model that is capable to producing an accurate prediction [24].

2. Research Method

2.1 Dataset

In this study, the used data set contains 40 data of asymmetric aromatic disulfide compounds with their activity values against SARS-CoV-2 calculated by in vitro studies [14]. We calculated the molecular descriptors by using the Padel program to obtain the shape of the structure, topology, and electrostatic properties. The descriptor data will be used as the main feature in predicting the compound's bioactivity value. With the total number of descriptors being 1444, we determined pLC50 as the response target. In which the pLC50 represents the lethal concentration value, which can impair 50% of the response target. The histogram of the pLC50 that shows the distribution of values is shown in Figure 1.

In the first stage, the number of molecular descriptors that will be used for each compound will be reduced by performing a feature selection. The process is carried out using statistical analysis on compounds by selecting features that have a deviation above 0.5 [25]. This process reduces the number of features from 1444 descriptors, become 436 features. Then, normalization is performed on each feature. Then, the data set was divided into two sub-datasets i.e., train and test set by looking at the pLC50 feature as the response target, with a ratio of 7:3 respectively [15].

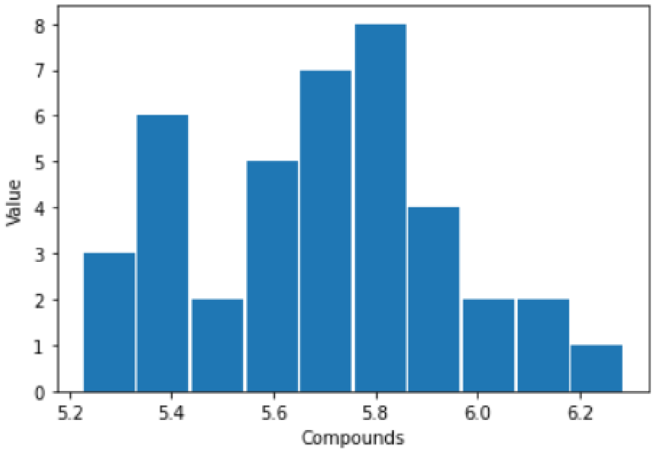


Figure 1. The Distribution of pLC50 Values

2.2 Genetic Algorithm (GA)

Genetic Algorithm is a stochastic optimization method governed by the rules of biological evolution which have been inspired by the principle of evolution [31]. GA investigates many possible solutions simultaneously and each explores a different area in the parameter space, such as probabilities of crossover (Pc), probabilities of mutation (Pm), and stopping criteria [32]. In its implementation, an individual is generated based on a binary representation at random, which forms a population of N individuals, where each individual has the cost of each individual in the current generation. Shown in Figure 2.

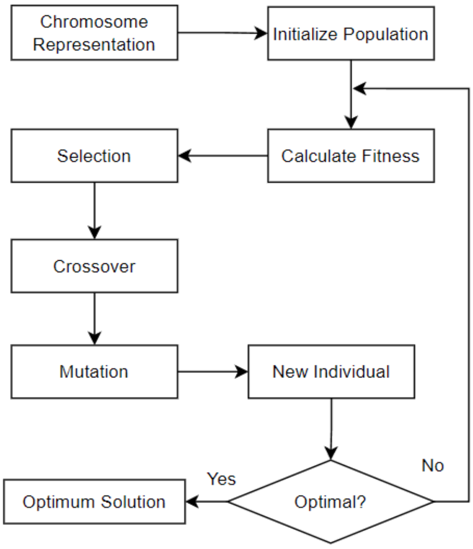


Figure. 2. Genetic Algorithm Flow

Parents are selected based on their respective *fitness* scores; Fitness Functions are Formulated as follows Equation 1.

$$f(x) = \frac{n}{\sum(y_i - \hat{y}_i)^2} \tag{1}$$

Where *n* is the total number of individuals in the population, the value of y_i is the actual value to be calculated and $\hat{y}_i$ is the predicted value of the individual, adjusting for the case of regression in this study. From the calculation of each individual's fitness, it will produce a small portion of the next generation of children with crossovers and the rest with mutations. After that, a new parent is selected for the next generation. In this way, the new offspring bear the characteristics of the parents. This is done continuously and will stop when the overall optimal result is obtained from the determination of the stopping criteria that have been determined from the start. The speed of GA in finding a wide range of possible solutions, and poor initial initiation generation capability that does not affect the final solution, makes GA very attractive to be used as a selection of dominant compound descriptors, in drug discovery models, where each problem is highly specialized due to lack of prior knowledge about functional relationships and generalizability are very difficult [31][32].

Feature selection is done by combining various descriptors using a Genetic Algorithm. This method is done by generating various solutions as a set of integers in the chromosome, and the integer numbers in the chromosome are used as the basis for selecting the index descriptor. At this stage, the cross-entropy loss is also used as an objective function in the feature selection process. The parameters used in feature selection using the genetic algorithm are listed in Table 1.

Table 1. Parameter for Feature Selection using GA [24]

Parameter	Value
Generation	50
Population	10
Mutation Probability	20%
Parent Selection	Roulette Wheel
Selection Criterion	Fitness-based selection

2.3 Support Vector Machine (SVM)

Support Vector Machine is a two-class classification method and is a supervised learning method. The support vector machine detects or predicts a pattern using kernel functions to map the input data to a high-dimensional space and finds the optimal hyperplane to separate the data into two classes [33]. To get the optimal hyperplane that separates two different classes in the vector space. A classification problem is a problem to categorize an entity into a certain group, for example, are binary classification which classifies an entity into a group of True (+1) and False (-1) [34]. This algorithm Equation 2 looks for the function $f: X \rightarrow \{-1, +1\}$ which minimizes empirical risk [34].

$$f^* = \min \frac{1}{n} \sum_{i=1}^{n} L(\mathrm{f}(x_i), y_i) \tag{2}$$

Where *L (a, b)* = $1_{a \neq b}$ it's a loss function.

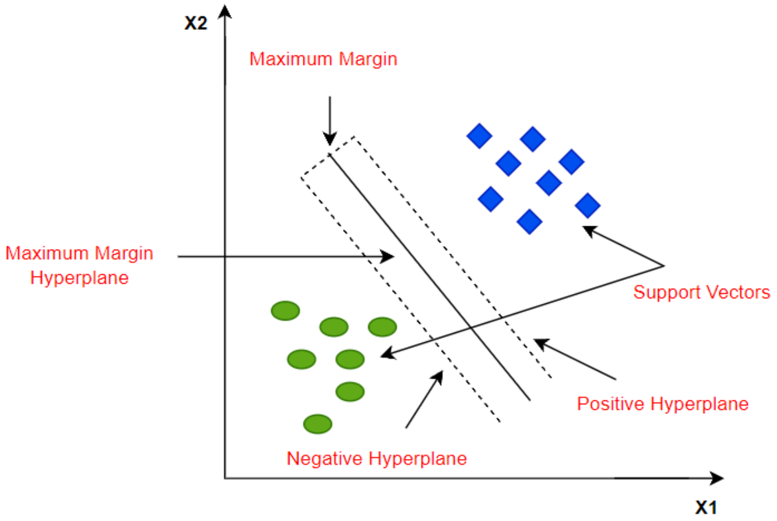


Figure. 3. Hyperplane on SVM

SVR model is an SVM algorithm development model that contains noise and nonlinearity components that were originally used for classification problems. The dependent variable in the SVR algorithm has a range of values in the form $y_t \in \mathbb{R}$. The main tool that needs to be understood and mastered to understand SVM is the hyperplane. A hyperplane is a generalization of straight lines in two-dimensional space on a plane. Basically, in the two-dimensional (2D) field, the equation of the line ax + by + c = 0 can be converted into an Equation 3 that can cover a multidimensional space as a hyperplane or called a decision boundary [34]. By changing the notation of these variables and constants such as x into x_1, y into x_2, a into w_1, and b into w_2:

$$\mathrm{w1x1} + \mathrm{w2x2} + B = 0 \tag{3}$$

when it is in dimension *d* > 1, the following Equation 4.

$$\sum_{n=1}^{d} w_n x_n + B = 0 \tag{4}$$

As for the writing of the equation above when in vector notation. Become the following Equation 5.

$$\langle w, x \rangle + B = 0 \tag{5}$$

where w, $x \in \mathbb{R}^d$ dan $\langle w, x \rangle = w^T x$ (dot product). In general, *the hyperplane* Equation 6.

$$W^T . X + B = 0 \tag{6}$$

Where *W* is the weight of the data, *X* is the input variable and *C* is the scalar variable which is a negative, zero, or positive number.

Equation 6 will look for the weight of the value on the sample as the basis for finding the hyperplane with the best margin [24]. However, the main problem in the regression model used is how to determine the hyperplane $f(x_0\theta) = \langle w, x \rangle + c$ that meets the specified value limits e so that $\{y_t - \hat{y}_t < e\}$. The specified margin is stated in the following Equation 7.

$$\frac{1-b-(-1-b)}{|w|} = 2 \text{ x } \frac{1}{|w|} \tag{7}$$

Several kernel functions are often used in SVM to handle the nonlinearity that can be found in data. The kernel function is the function of the *inner product* in the feature space, the SVM kernels used and their equations represented in Table 2 [24]. The application of the SVM algorithm for both regression and classification models require set-up parameters known as hyperparameters. Similar to the classification model, the regression model is also optimized using regularization (C) and gamma values. And independent parameters can be applied, such as epsilon (ϵ), and degree [33]. The regularization parameter (C) assigns and/or adds a penalty weight to the SVM model if the data point prediction output is wrong. The smaller the value of C applied, the smaller the penalty weight for missing predictions from an SVM model, and the resulting output has a large margin and stays away from the decision boundary.

Table 2. Kernel Functions in Support Vector Machine

Kernel	Formula
Linear Kernel	$K(x_i , x_j) = xi\ T\ .\ xj + C$
Radial Basis Function Kernel	$K(X_i, X_j) = \exp\ (-y\ \lVert Xi, Xj \rVert)^2$
Polynomial Kernel	$K\left(X_i, X_j\right) = (yX_i^T X_j)^p, y > 0$

On the other hand, the larger the value of C applied, the SVM model will minimize the number of erroneous prediction data points, which has an impact on the prediction output with a small margin and close to the decision boundary. In its implementation, the penalty weight for each missing data point depends on the distance between the predicted data points and the decision boundary [37]. The Gamma parameter is useful for setting the distance of influence from a training point. With a low gamma value, the radius of similarity level will be large, and many data points can be included in it. Meanwhile, with a low gamma value, the radius of the similarity level will be smaller, and each data point needs to have high proximity to be included in it [37].

2.4 Hyperparameter Tuning

After selecting the features, the model is built based on the SVM method. The SVM model uses training data and test data with selected features. To predict the model, it is necessary to tune the hyperparameter which aims to improve the performance of the designed model. This stage is done by trying all combinations of parameters and comparing the results of the combination to determine the best one. Grid Search cross-validation is used to tune these hyperparameters. The following are the tuned SVM parameters described in Table 3.

Table 3. Range value for Hyperparameter Tuning

Parameter	Range Value [25]
C	[0.001,0.01,0.1,1,10,100,1000]
Epsilon	[0.001,0.01,0.1,1,10,100,1000]
Degree	1,2,3,4,5
Gamma	auto, scale

2.5 Model Validation

To validate the prediction model, internal and external validation methods are used, by calculating certain parameters. For internal parameters, correlation coefficient (R^2_{train}), cross-validation (Q^2_{loo}), *leave-one-out* (LOO), and y-randomization test (${}^cR_p^2$) were used using training data. While for external parameters, correlation coefficient (R^2_{test}) is used using test data. The validity of the model can be said to be fulfilled if the parameter values are by their respective thresholds. In addition, several validation methods were also carried out to confirm that the model was acceptable. The following Equation 8 – Equation 20 are the validation parameters along with the threshold [24].

$$R^2_{train} = 1 - \frac{\sum(\mathrm{y_{train}} - \hat{\mathrm{y}}_{\mathrm{train}})^2}{\sum(\mathrm{y_{train}} - \hat{\mathrm{y}}_{\mathrm{train}})^2} \quad (R^2 > 0.6 \tag{8}$$

$$Q^2_{loo} = 1 - \frac{\sum(\mathrm{y_{train}} - \hat{\mathrm{y}}_{\mathrm{loo}})^2}{\sum(\mathrm{y_{train}} - \hat{\mathrm{y}}_{\mathrm{train}})^2} \qquad (\mathrm{Q}^2 > 0.5) \tag{9}$$

$$\mathrm{k}' = \frac{\sum(\mathrm{y} \times \hat{\mathrm{y}})}{\sum(\mathrm{y})^2} \qquad (0.85 > \mathrm{k}' > 1.15) \tag{10}$$

$$\mathrm{r}^2 = \frac{[\sum(\mathrm{y} - \bar{\mathrm{y}})(\mathrm{y} - \hat{\mathrm{y}})]^2}{\sum(\mathrm{y} - \bar{\mathrm{y}})^2 \times \sum(\mathrm{y} - \hat{\mathrm{y}})^2} \tag{11}$$

$$\mathrm{r}_0^2 = 1 - \frac{\sum(\mathrm{y} - \mathrm{k} \times \hat{\mathrm{y}})^2}{\sum(\mathrm{y} - \bar{\mathrm{y}})^2} \tag{12}$$

$$\mathrm{r}_0'^2 = 1 - \frac{\sum(\hat{\mathrm{y}} - \mathrm{k}' \times \mathrm{y})^2}{\sum(\hat{\mathrm{y}} - \bar{\hat{\mathrm{y}}})^2} \tag{13}$$

$$|\mathrm{r}_0^2 - \mathrm{r}_0'^2| \qquad (|\mathrm{r}_0^2 - \mathrm{r}_0'^2| < 0.3) \tag{14}$$

$$\mathrm{r}_\mathrm{m}^2 = \mathrm{r}^2 \times \left(1 - \sqrt{\mathrm{r}^2 - \mathrm{r}_0^2}\right) \qquad (\ \mathrm{r}_\mathrm{m}^2 > 0.5) \tag{15}$$

$$\mathrm{r}_\mathrm{m}'^2 = \mathrm{r}^2 \times \left(1 - \sqrt{\mathrm{r}^2 - \mathrm{r}_0'^2}\right) \tag{16}$$

$$\overline{\mathrm{r}_\mathrm{m}^2} = \frac{(\mathrm{r}_\mathrm{m}^2 + \mathrm{r}_\mathrm{m}'^2)}{2} \qquad (\overline{\mathrm{r}_\mathrm{m}^2} > 0.5) \tag{17}$$

$$\Delta\mathrm{r}_\mathrm{m}^2 = |\mathrm{r}_\mathrm{m}^2 - \mathrm{r}_\mathrm{m}'^2| \qquad (\Delta\mathrm{r}_\mathrm{m}^2 < 0.2) \tag{18}$$

$${}^cR_p^2 = R \text{ x} \sqrt{R^2 - R^2} \qquad ({}^cR_p^2 > 0.5) \tag{19}$$

$$R^2_{test} = 1 - \frac{\sum(\mathrm{y_{test}} - \hat{\mathrm{y}}_{\mathrm{test}})^2}{\sum(\mathrm{y_{test}} - \hat{\mathrm{y}}_{\mathrm{test}})^2} \qquad (\mathrm{R}^2 > 0.6) \tag{20}$$

The application of the Applicability Domain (AD) of each model was also carried out to confirm that the data points remained in the domain of the model [26]. By using the leverage method in determining AD, which is formulated as follows Equation 21.

$$\mathrm{H} = \mathrm{X}(\mathrm{X}^\mathrm{T}\mathrm{X})^{-1}\mathrm{X}^\mathrm{T} \tag{21}$$

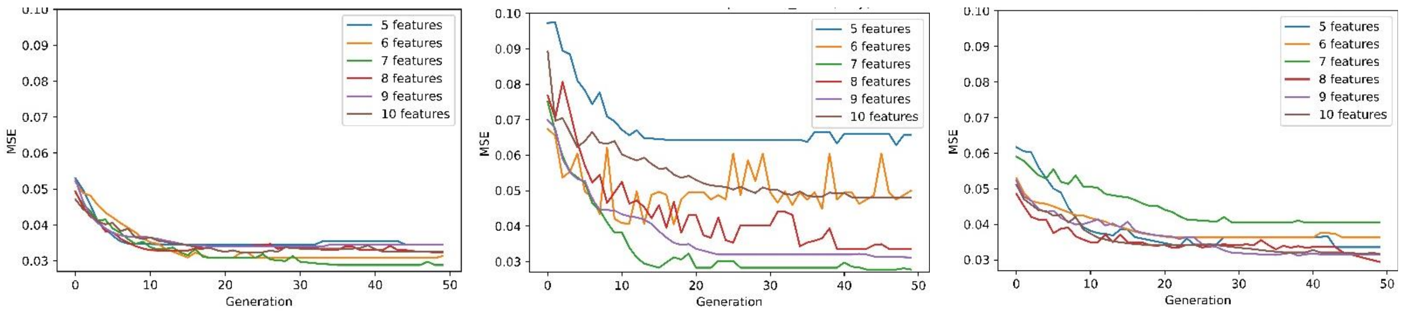
Where *X* is a representation of the matrix value obtained from the PLSR procedure. Where critical influence is defined as 3p/n, with p and n respectively being the number of attributes and datasets, which are used in the model training process. Then, the results of the model are displayed through the William Plot after the Applicability Domain (AD) results are obtained [34].

3. Results and Discussion

3.1 Feature Selection

To get the best descriptor, molecular descriptors will be reduced to remove descriptors that are considered less capable or weak in representing the compound's ability through one of the statistical analysis methods. First, descriptor removal is done by calculating the variance of each descriptor. Descriptors with deviations below 0.5 will be unused. It is intended to obtain descriptor columns that have good data in representing the ability of the compound. After being reduced, feature selection is carried out using GA to obtain optimal results. The representation of individuals in the selection model is a compound descriptor. In achieving these results and considering the limited resources in this study, the number of combinations determined only between 5, 6, 7, 8, 9, and 10 combinations of descriptors in the GA selection algorithm. The number of selected descriptor combinations is based on the best fitness value.

The fitness value in question is a value based on the value of the cross-validation regression model that is applied to individuals in a population. Where the individual in the population is a representation of several combinations of compound descriptors. After the feature selection process has been carried out on a model, the resulting output is a combination of the most dominant descriptors. But in its implementation, GA's random concept in the initial generation resulted in the accuracy and number of descriptors being different in each iteration of the process. To maintain the random concept of GA, 50 iterations of GA generation were performed, and comparisons based on MSE were performed in each generation iteration. And for the selection of the best combination is based on the lowest MSE value generated. This study uses the SVM kernel function to find the contribution of each function, including linear, polynomial, and RBF. Optimal results are sought, determined by calculating the MSE value of each kernel function. Mean Square Error (MSE) can indicate the success of the QSAR model prediction on GA. The following is the MSE distribution generated for each number of descriptors for the GA generation, shown in Figure 5.



(a) (b) (c)

Figure 5. MSE Distribution on Each Number of Feature Descriptors Against the Generation in the Kernel (a) Linear, (b) Polynomial, and (c) RBF

As for linear kernel, we found that fluctuations in the MSE value are not extreme. From each number of features, there is a significant downward trend in the MSE value for generations. Also, optimal results are obtained on a combination of 7 features with an MSE of 0.0286. Slightly different in the RBF kernel, although it has the same decreasing trend until the end of the generation, the decrease in the MSE value concerning generation is not as large as that found in the linear kernel. From this model, the best MSE is 0.273 with the best combination of 8. While in the polynomial kernel model, although there is also a downward trend in MSE values, significant fluctuations are found. Fluctuations were found in the number of features 6 and 8, and the smallest MSE was found at 0.0276 with a combination of 7 features.

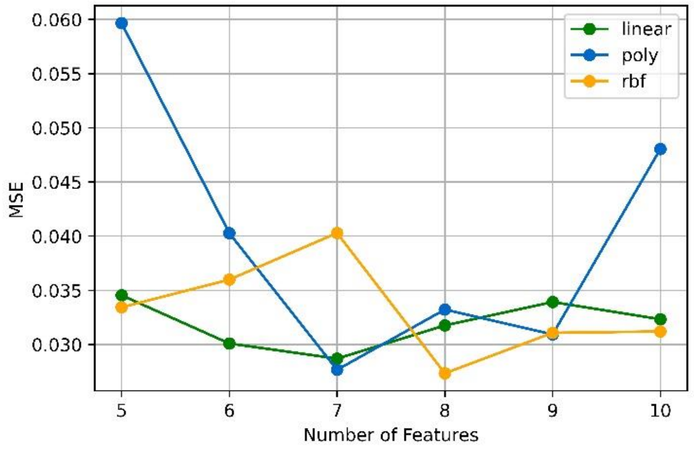


Figure. 6. MSE Values Against the Number of Features

The graph of the tendency of each kernel function towards each number of descriptors is shown in Figure 6. It can be concluded that the MSE value distribution trend of the polynomial kernel model is quite volatile compared to the linear kernel and RBF which do not encounter extreme fluctuations. However, the optimal number of descriptors obtained in linear and polynomial kernels is 7, while in the RBF kernel it is 8. The optimal number of descriptors is chosen by determining the number of descriptors with the lowest MSE value in each kernel. In Table 4, the minimum and average MSE values of each kernel are obtained for the number of descriptors. The smallest average is obtained in the linear kernel of 0.323 with the number of descriptors 7. The lowest MSE value of the whole kernel that also considered as the optimum model, obtained by the RBF kernel model with its MSE 0.276 with the number of descriptors is 7.

Table 4. Obtained Minimum and Average MSE Values

Kernel	Optimum Number of Features	Minimum MSE	Average MSE
Linear	7	0.02869	0.3230
Polynomial	7	0.02765	0.04489
RBF	8	0.02733	0.03512

3.2 Hyperparameter Tuning

After obtaining the optimal descriptor for each model, Hyperparameter Tuning is performed to obtain parameters that can be used to improve the performance of the SVM model. This process uses the GridSearchCV method and K-Fold Cross Validation with a value of K = 10. From each kernel, the best tuning results for each kernel are shown in Table 5, as well as a comparison of performance before and after tuning based on the least mean square, shown in Figure 7. We found after hyperparameter tuning, each kernel gains a significant increase in R^2 score. As of linear kernel, we found that there is a significant gain of 0.495 from before. Then on the polynomial kernel, the performance gain is 0.677, and in the RBF kernel we found it has the most significant gain difference is 0.846.

Table 5. Hyperparameter Tuning result

Parameter	Kernel		
	Linear	Polynomial	RBF
C	10 (1.0)	1000 (1.0)	10 (1.0)
Degree	1 (3)	3 (3)	1 (3)
Epsilon	0.01 (0.1)	0.001 (0.1)	0.1 (0.1)
Gamma	auto (scale)	auto (scale)	scale (scale)

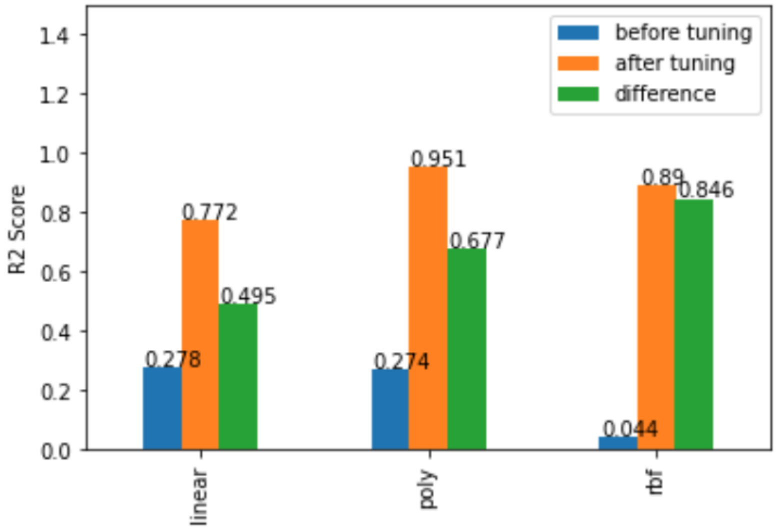
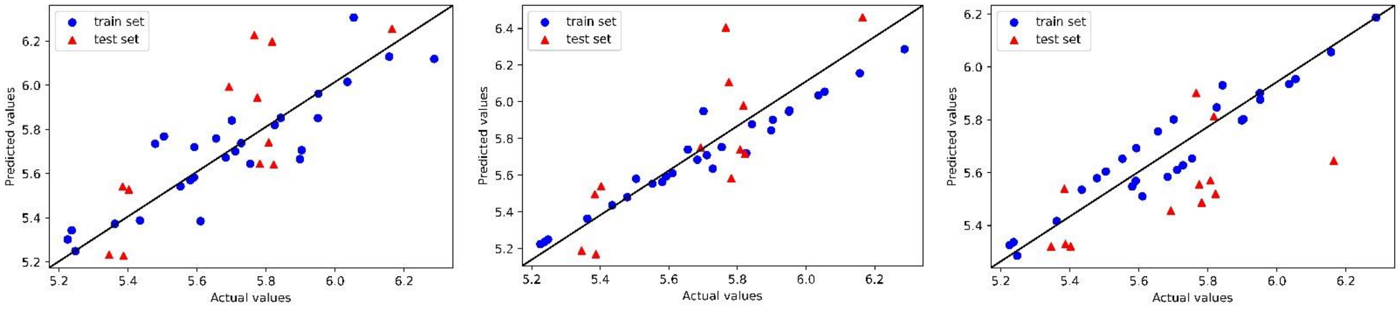


Figure. 7. Before and After Hyperparameter Tuning Performance

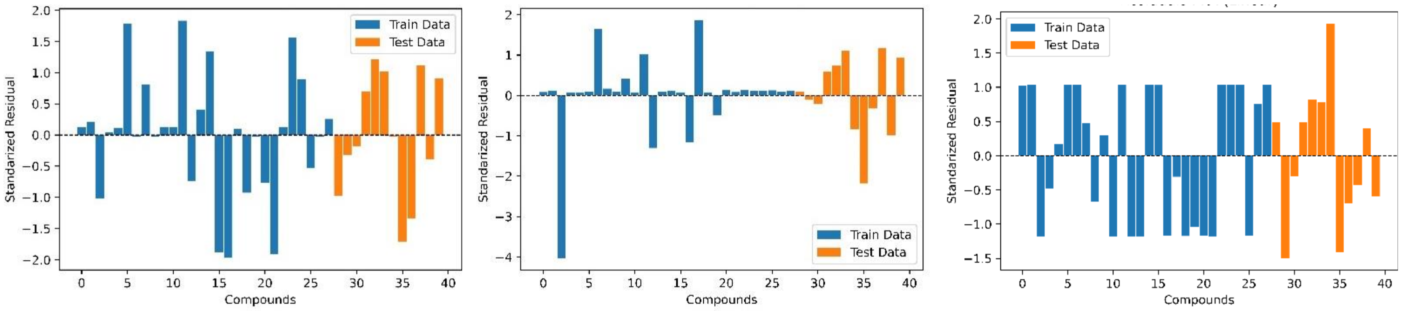
3.3 Model Validation

After building the model using training data, evaluation and validation of the model will be carried out, using both training data and test data. The prediction model is developed using a model that has been optimized in the previous tuning. The results of the regression between the predicted and actual values are shown in Figure 8. In SVM, a diagonal line is used which serves as a reference for the match between the predicted value and the actual value. If there is a deviation at a data point with a diagonal line, then it shows the degree of error in each model prediction. From the validation process, it was found that the deviation between the data points and the diagonal line in the linear kernel, polynomial, and RBF is quite acceptable. To confirm the performance of the model, a plot of the residuals between the predicted pLC values is generated in each kernel in Figure 9. From the figure, we found the difference in magnitude generated and value of error for each sample indicating there is no systematical error found in the model.



(a) (b) (c)

Figure. 8. Regression Plot between the Predicted Value and the Actual Value of the SVM on the Kernel (a) Linear, (b)Polynomial, and (c) RBF.



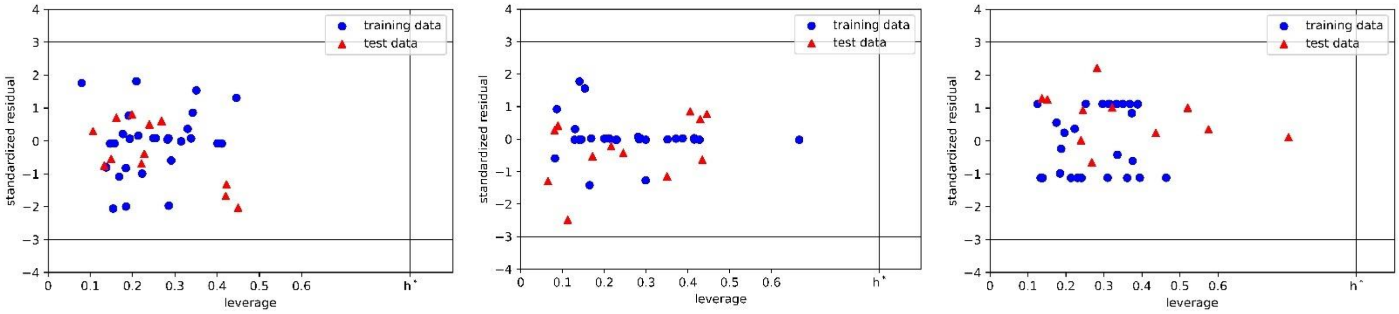
(a) (b) (c)

Figure. 9. The Plot of Standardized Residuals between the Actual Value and the Predicted Value in the Kernel (a) Linear, (b) Polynomial, and (c) RBF

Based on the results obtained in Table 6, a model can be accepted or validated if the calculation results of the parameters meet the predetermined threshold. As for the linear kernel, we find that the kernel satisfies all parameters. This is indicated by the entire measurement score of each parameter is above the threshold. The same result is also found in the polynomial kernel. But on the RBF kernel, we found values of r_m^2 and $\overline{r_m^2}$ score which are below the predetermined threshold and considered invalid. As for comparison, we found the R^2_{train} and the R^2_{test} value found in the linear kernel, both values are 0.778 and 0.642 that above the threshold. Also, in the polynomial kernel we found the R^2_{train} is 0.952 and R^2_{test} score is 0.676, both values exceed the threshold. For overall measurement based on the R^2_{test} between the linear and polynomial kernel, we found that the polynomial kernel was superior when compared with the linear kernel. Thus, making the polynomial kernel is best model based on internal and external validation of SVM models performance value in predicting SARS-CoV main protease inhibitors on the used dataset.

Table 6. Parameter Validation Result

Validation Parameter	*Kernel*						*Threshold [26]*
	Linear		*Poly*		*RBF*		
	Train Set	*Test Set*	*Train Set*	*Test Set*	*Train Set*	*Test Set*	
R^2	0.778	0.642	0.952	0.676	0.909	0.623	> 0.6
Q^2_{LOO}	0.483	-	0.521	-	0.453	-	> 0.5
k'	0.937	1.164	1.006	1.041	0.988	0.828	0.85 > k' > 1.15
$\lvert r_0^2 - r_0'^2 \rvert$	0.003	0.054	0.000	0.0276	0.007	0.107	< 0.3
r_m^2	0.704	0.588	0.941	0.608	0.905	0.236	> 0.5
$\overline{r_m^2}$	0.698	0.535	0.932	0.576	0.867	0.271	> 0.5
Δr_m^2	0.013	0.105	0.018	0.063	0.075	0.069	< 0.2
${}^cR_p^2$	0.922	-	0.711	-	0.542	-	> 0.5



(a) (b) (c)

Figure. 10. Applicability Domain in the Kernel (a) Linear, (b) Polynomial, and (c) RBF

Overall, we found that SVM with the polynomial kernel performed better than compared to other kernel models. This is due to the large value of the C parameter, and the automatic gamma value, following the optimal value obtained. When the value of C obtained is large, the SVM model will reduce the number of data misses, this is due to the high error load for each data miss. That way, the polynomial model minimizes the prediction data output with a high error rate, resulting in a small margin of data points and tends to be close to the decision boundary as shown in Figure 8. Based on the model performance output, which is marked by the best model score value based on the internal validation parameters in the training data obtained an R^2_{train} score of 0.952 and external validation on the training data R^2_{test} score of 0.676 by a polynomial kernel. As for the underperforming model, it is considered to be the linear kernel. This is caused because the smaller C parameter value generated, the larger the data misses as a result to the lower load penalty for each data miss. As for the results in the applicability model to the dataset, it can be confirmed that both the model built is applicable to the used dataset. This is indicated by the distribution of data points are lied inside the region of the domain boundary, both in test data and training data shown in the Figure 10.

4. Conclusion

Based on the results, we found that by using a combination of the Genetic Algorithm feature selection method and SVM regression prediction model used to predict the activity of aromatic disulfide compounds as potential main protease inhibitors (Mpro) in the SARS-CoV virus. The first step is to select the descriptor feature of the compound that has a deviation < 0.5, and then proceed with feature selection using a Genetic Algorithm based on the smallest mean square error (MSE) using the linear kernel, polynomial, and RBF to get the most optimal combination of descriptors. Then, we improved the model by performing hyperparameter tuning procedure based on the predetermined input constraints. The ability of GA-SVM in predicting the activity of aromatic disulfide compounds to be used as potential inhibitors is generally acceptable. We found that SVM with polynomial kernel produce the best result with the value of R^2 of train and test are 0.952 and 0.676, respectively. From the AD capabilities of each model, all of which are in the application domain, we found that all model is acceptable according to applicability domain analysis.

References

[1] C. Drosten, W. Preiser, S. Günther, H. Schmitz, and H. W. Doerr, "severe acute respiratory syndrome: Identification of the etiological agent," *TrendsMol. Med.*, vol. 9, no. 8, pp. 325–327, 2003. https://doi.org/10.1016/s1471-4914(03)00133-3

[2] A. Remuzzi and G. Remuzzi, "COVID-19 and Italy: what next?" *Lancet*,vol. 395, no. 10231, pp. 1225–1228, 2020. https://doi.org/10.1016/S0140-6736(20)30627-9

[3] H. Yang *et al.*, "Design of wide-spectrum inhibitors targeting coronavirus main proteases," *PLoS Biol.*, vol. 3, no. 10, 2005. https://doi.org/10.1371/journal.pbio.0030324

[4] W. W. C. Topley and S. G. S. Wilson, "Topley and Wilson's Microbiologyand Microbial Infections, 8 Volume Set, 10th Edition," J. Infect., vol. 38, no. 2, p. 3500, 1999.

[5] Worldometers, "No Title," 2020.

[6] S. A. Amin, S. Bhargava, N. Adhikari, S. Gayen, and T. Jha, "Exploring pyrazolo[3,4-d]pyrimidine phosphodiesterase 1 (PDE1) inhibitors: a predictive approach combining comparative validated multiple molecular modelling techniques," *J. Biomol. Struct. Dyn.*, vol. 36, no. 3, pp. 590–608,2018. https://doi.org/10.1080/07391102.2017.1288659

[7] S. Jain, S. A. Amin, N. Adhikari, T. Jha, and S. Gayen, "Good and bad molecular fingerprints for human rhinovirus 3C protease inhibition: identification, validation, and application in designing of new inhibitors through Monte Carlo-based QSAR study," *J. Biomol. Struct. Dyn.*, vol. 38, no. 1, pp. 66–77, 2020. https://doi.org/10.1080/07391102.2019.1566093

[8] Y. Yang *et al.*, "The deadly coronaviruses: The 2003 SARS pandemic andthe 2020 novel coronavirus epidemic in China," *J. Autoimmun.*, vol. 109, no. February, p. 102434, 2020. https://doi.org/10.1016/j.jaut.2020.102434

[9] L. Wang *et al.*, "Discovery of unsymmetrical aromatic disulfides as novelinhibitors of SARS-CoV main protease: Chemical synthesis, biological evaluation, molecular docking and 3D-QSAR study," *Eur. J. Med. Chem.*, vol. 137, pp. 450–461, 2017. https://doi.org/10.1016/j.ejmech.2017.05.045

[10] A. Golbraikh and A. Tropsha, "Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection," *J. Comput. Aided. Mol. Des.*, vol. 16, no. 5–6, pp. 357–369,2002. https://doi.org/10.1023/a:1020869118689

[11] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, “Docking and scoring in virtual screening for drug discovery: Methods and applications,” *Nat. Rev. Drug Discov.*, vol. 3, no. 11, pp. 935–949, 2004. https://doi.org/10.1038/nrd1549

[12] J. M. Halperin *et al.*, “Training Executive, Attention, and Motor Skills: AProof-of-Concept Study in Preschool Children With ADHD,” *J. Atten. Disord.*, vol. 17, no. 8, pp. 711–721, 2013, doi:. https://doi.org/10.1177/1087054711435681

[13] A. A. Toropov, A. P. Toropova, A. M. Veselinović, D. Leszczynska, and J.Leszczynski, “SARS-CoV Mpro inhibitory activity of aromatic disulfide compounds: QSAR model,” *J. Biomol. Struct. Dyn.*, pp. 1–7, 2020, doi:. https://dx.doi.org/10.1080%2F07391102.2020.1818627

[14] L. Wang *et al.*, “Discovery of unsymmetrical aromatic disulfides as novelinhibitors of SARS-CoV main protease: Chemical synthesis, biological evaluation, molecular docking and 3D-QSAR study,” *Eur. J. Med. Chem.*, vol. 137, pp. 450–461, 2017. https://doi.org/10.1016/j.ejmech.2017.05.045

[15] A. Gholamy, V. Kreinovich, and O. Kosheleva, “Why 70 / 30 or 80 / 20 Relation Between Training and Testing Sets : A Pedagogical Explanation,” pp. 1–6.

[16] S. Chtita *et al.*, “QSAR study of unsymmetrical aromatic disulfides as potent avian SARS-CoV main protease inhibitors using quantum chemical descriptors and statistical methods,” *Chemom. Intell. Lab. Syst.*, vol. 210, no. February 2021. https://doi.org/10.1016/j.chemolab.2021.104266

[17] E. Pourbasheer, R. Aalizadeh, and M. R. Ganjali, “QSAR study of CK2 inhibitors by GA-MLR and GA-SVM methods,” *Arab. J. Chem.*, vol. 12, no. 8, pp. 2141–2149, 2019. http://dx.doi.org/10.1016/j.arabjc.2014.12.021

[18] E. Pourbasheer, S. Vahdani, D. Malekzadeh, R. Aalizadeh, and A. Ebadi, “Qsar study of 17β-HSD3 inhibitors by genetic algorithm-support vector machine as a target receptor for the treatment of prostate cancer,” *IranianJournal of Pharmaceutical Research*, vol. 16, no. 3. pp. 966–980, 2017. https://doi.org/10.22037/ijpr.2017.2096

[19] E. Pourbasheer, S. Riahi, M. R. Ganjali, and P. Norouzi, “Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity,” *Eur. J. Med. Chem.*, vol. 44, no. 12, pp. 5023–5028, 2009. https://doi.org/10.1016/j.ejmech.2009.09.006

[20] M. H. Fatemi and S. Gharaghani, “A novel QSAR model for prediction ofapoptosis-inducing activity of 4-aryl-4-H-chromenes based on support vector machine,” *Bioorganic Med. Chem.*, vol. 15, no. 24, pp. 7746–7754,2007. https://doi.org/10.1016/j.bmc.2007.08.057

[21] R. Burbidge, M. Trotter, B. Buxton, and S. Holden, “Drug design bymachine learning: Support vector machines for pharmaceutical data analysis,” *Comput. Chem.*, vol. 26, no. 1, pp. 5–14, 2001. https://doi.org/10.1016/S0097-8485(01)00094-8

[22] S. Abe, “Support Vector Machines for Pattern Classification My Research History on NN, FS , and SVM,” *Sci. Technol.* http://dx.doi.org/10.1109/TCYB.2013.2279167

[23] F. Liu, C. Cao, and B. Cheng, “A quantitative structure-property relationship (QSPR) Study Of aliphatic alcohols by the method of dividingthe molecular structure into substructure,” *Int. J. Mol. Sci.*, vol. 12, no. 4, pp. 2448–2462, 2011. https://doi.org/10.3390/ijms12042448

[24] H. F. Azmi, K. M. Lhaksmana, and I. Kurniawan, “QSAR Study of FusidicAcid Derivative as Anti-Malaria Agents by using Artificial Neural Network-Genetic Algorithm,” *2020 8th Int. Conf. Inf. Commun. Technol. ICoICT 2020*, pp. 3–6, 2020. https://doi.org/10.1109/ICoICT49345.2020.9166158

[25] F. Rahman, K. M. Lhaksmana, and I. Kurniawan, “Implementation of Simulated Annealing-Support Vector Machine on QSAR Study of Fusidic Acid Derivatives as Anti-Malarial Agent,” *6th Int. Conf. Interact. Digit. Media, ICIDM 2020*, no. Icidm, pp. 8–11, 2020. https://doi.org/10.1109/ICIDM51048.2020.9339632

[26] I. Kurniawan, M. S. Fareza, and P. Iswanto, “Comfa, molecular docking and molecular dynamics studies on cycloguanil analogues as potent antimalarial agents,” *Indones. J. Chem.*, vol. 21, no. 1, pp. 66–76, 2021. https://doi.org/10.22146/ijc.52388

[27] Y. Yuliana, “Corona virus diseases (Covid-19): Sebuah tinjauan literatur,” *Wellness Heal. Mag.*, vol. 2, no. 1, pp. 187–192, 2020, doi: https://doi.org/10.30604/well.95212020

[28] M. D. Christian Drosten, M.D., Stephan Günther, M.D., Wolfgang Preiser, M.D., Sylvie van der Werf, Ph.D., Hans-Reinhard Brodt, M.D., Stephan Becker, Ph.D., Holger Rabenau, Ph.D., Marcus Panning, M.D., Larissa Kolesnikova, Ph.D., Ron A.M. Fouchier, Ph.D., Annema, “Identification ofa Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome,”pp. 1967–1976, 2020. https://doi.org/10.1056/nejmoa030747

[29] PDPI, “Panduan Praktik Klinis: Pneumonia COVID-19,” *J. Am. Pharm.Assoc.*, vol. 55, no. 5, pp. 1–67, 2020.

[30] J. Ivanov *et al.*, “Quantitative structure−activity relationship machine learning models and their applications for identifying viral 3Clpro- And RDRP-targeting compounds as potential therapeutics for Covid-19 and related viral infections,” *ACS Omega*, vol. 5, no. 42, pp. 27344–27358, 2020.: https://doi.org/10.1021/acsomega.0c03682

[31] J. H. Holland, *Adaption in natural and artificial systems*. Michigan: The University of Michigan Press, Ann Arbor, MI, 1975. https://dl.acm.org/doi/10.5555/531075

[32] C. HM, *Applications of artificial intelligence in chemistry*. Oxford: OxfordUniversity Press, Oxford, 1993.

[33] M. F. Asshiddiqi, Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI. Universitas Telkom, 2020. Accessed: Nov. 30, 2020.

[34] I. Aydin, M. Karakose, and E. Akin, “A multi-objective artificial immune algorithm for parameter optimization in support vector machine,” *Appl. SoftComput. J.*, vol. 11, no. 1, pp. 120–129, 2011. https://doi.org/10.1016/j.asoc.2009.11.003

[35] M. Ghifari, “Gif’s note Support Vector Machines: Penjelasan Matematisdan Intuitif,” pp. 1–13, 2021.

[36] Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., and Todeschini, R., 2012, Comparison of different approaches to define the applicability domain of QSAR models, Molecules, 17 (5), 4791–4810. https://doi.org/10.3390/molecules17054791

[37] S. Yildirim, “Support Vector Machine - Explained,” 2020.