

QSAR STUDY OF PYRAZOLE-UREA HYBRID COMPOUNDS AS ANTIMALARIAL AGENT VIA PROLYL-tRNA SYNTHETASE INHIBITION

I. G. A. Sumartha, T.A. Yuniarta[✉] and D. Kesuma

Department of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Surabaya,
Surabaya, Indonesia 60293

[✉]Corresponding Author: tegar.achsendo@staff.ubaya.ac.id

ABSTRACT

Aminoacyl-tRNA synthetase is one of the emerging potential targets for various tropical diseases such as malaria. One of which, prolyl-tRNA synthetase, has known to possess diverse binding modes such as dual inhibitor with febrifugine and the recently elucidated allosteric site. A recent study showed that pyrazole-urea hybrid analogues are a potential selective inhibitor agent against this enzyme. Here, we have created a quantitative structure-activity relationship (QSAR) model using multiple linear regression (MLR) and a partial least square (PLS) regression approach. Due to the limited amount of available data, the double cross-validation (DCV) method was employed to overcome a bias which resulted due to some data points being held out for validation. Several optimizations were performed to obtain the best result, namely different data splitting strategies (random and rational) and different correlation threshold values (0.8 and 0.9). Finally, the two best equations obtained from MLR and PLS approaches were evaluated by checking their applicability domain. It is found that the MLR output model yielded the best result, which indicated the significance of Geary autocorrelation and CATS-2D-based descriptors in representing the correlation between pyrazole-urea hybrid analogues and their enzymatic activity.

Keywords: Malaria, Aminoacyl-tRNA synthetase, QSAR, Allosteric, Pyrazole-urea

RASAYAN J. Chem., Vol. 15, No.2, 2022

INTRODUCTION

Malaria is one of the most prevalent infectious diseases which caused a significant impact on developing countries situated in the tropical and subtropical regions.¹ *Plasmodium falciparum* is known to be the major cause of the disease, accounting for nearly all cases in Africa and a considerable portion in Southeast Asia and Western Pacific². Despite the effectiveness of several antimalarial therapies as per international guidelines, numerous cases of drug resistance have occurred.^{3,4} Therefore, it is necessary to accelerate the discovery and development of novel drug candidates which is not only potent and safe but could also tackle the issue of drug resistance. Aminoacyl-tRNA synthetases are important enzymes for protein biosynthesis, which help the attachment of amino acids to their cognate tRNA *via* an esterification reaction. Studies have indicated the importance of these enzymes as the main target for developing novel antimalarial compounds.⁵ One of the enzymes, prolyl-tRNA synthetase (PRS), is the target of febrifugine-based compounds which are known to possess very potent antimalarial activity.⁶⁻⁹ Based on protein crystallographic data, it is observed that the majority of this chemotype interacts with *Plasmodium falciparum* PRS (PfPRS) in the same manner as observed in its human orthologue.¹⁰ Aside from the previous example, the PRS enzyme also contains several other druggable binding sites.⁵ One of them is the allosteric binding pocket that resides adjacent to the ATP binding site. Studies conducted by Hewitt *et al.*¹¹, suggest that compounds possessing two hydrophobic moieties bridged by the urea functional group will act as a potential ligand in this mechanism. Unlike febrifugine-like compound which acts as a dual inhibitor mimicking L-proline and tRNA 3'-A76 moieties⁷, this class of compound disrupts the interaction between ATP and enzyme by inducing a conformational change in the TXE protein loop. In addition, this scaffold also possesses distinct selectivity towards PfPRS which makes it an interesting target for the novel selective antimalarial drug candidates.^{11,12} Quantitative Structure-Activity Relationship (QSAR) is one of the tools commonly used in the process of drug discovery and development. This approach allows one to correlate quantifiable structure parameters with their respective biological activities using a mathematical model. The established model

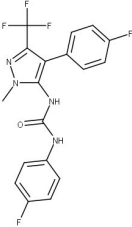
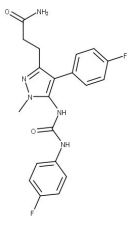
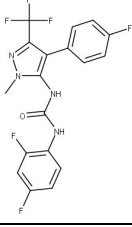
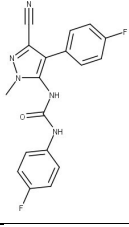
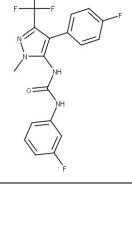
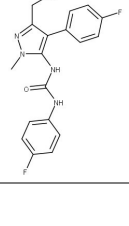
can be used to predict the activity of a compound based on its physicochemical descriptors. It also provides insight into the biological activity of the compound and which parameters are of importance to it.¹³ In its inception, QSAR modeling could be carried out using every compound with available bioactivity data, as long as each of them is derived from the same chemical scaffold.¹⁴⁻¹⁶ Nowadays, it is more accustomed to the utilization of multiple compounds (*e.g.* $n > 50$), either with a similar scaffold or various chemical diversity. In addition, various parameters have been imposed to ensure that a meaningful and statistically valid model is obtained.¹⁷ However, often times this will hinder researchers who are unable to build a decent QSAR model, due to the lack of available experimental datasets. Recently, the trend of small dataset QSAR modeling has resurged and found its application in various fields such as nanoparticles and catalyst development.^{17,18} Several adjustments in algorithm and parameters have also been created to ensure the validity is comparable to the QSAR model with a larger database.¹⁹ The aim of this study was to develop 2D QSAR models to describe *in vitro* activity of several pyrazole-urea analogues towards the PfPRS enzyme.¹¹ Various 2D and 3D based molecular descriptors were generated using various descriptors and analyzed statistically using multiple linear regression (MLR) and partial least square (PLS). The models obtained indicate essential descriptors which play important role in both antiparasitic and enzymatic activity of the series of compounds.

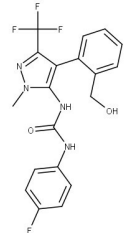
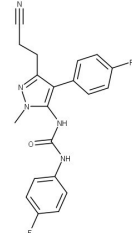
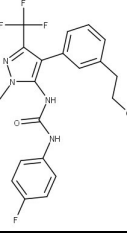
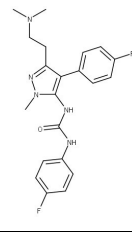
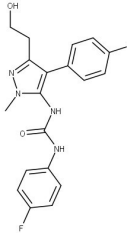
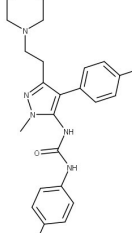
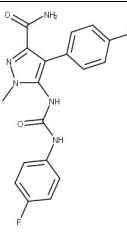
EXPERIMENTAL

Dataset Preparation

13 analogues of the pyrazole-urea hybrid were obtained from¹¹ (Table-1). Compound that has a definite IC₅₀ value of enzymatic activity and was included in the development of the respective QSAR models. SMILES strings of selected compounds were then curated and converted into 3D structures using CORINA.²⁰ This process was carried out in OChem webserver.²¹

Table-1. List of Pyrazole-Urea Analogue Used in the Dataset and its Activity against PfPRS Enzyme

No	Structure	Activity (pIC ₅₀)		No	Structure	Activity (pIC ₅₀)
1		5.30		8		5.10
2		5.60		9		5.30
3		5.10		10		5.22

4		4.70		11		5.04
5		4.89		12		5.40
6		4.74		13		5.52
7		4.85				

Molecular Descriptor Selection

A total of 5305 unique molecular descriptors were generated using alvaDesc 1.0.20 module²², ranging from constitutional to CATS-based descriptors.²³ The value was normalized and then reduced by eliminating descriptors with less than two unique values, large absolute values, and less than 0.01 covariance. Two inter-correlation coefficient values (0.8 and 0.9) were used as collinearity threshold. Should there be two or more descriptors possessing larger values than the respective cut-off values, they will be considered collinear and excluded from model building. This process was also conducted in OChem webserver.²¹

Dataset Division

The dataset was then split into the training set and a test set with the proportion of 80%:20%, which corresponds to 3 compounds in the test set and 10 compounds in the training set to develop the QSAR model. In order to obtain a highly predictive model, several data splitting strategies have been implemented, namely:

- In the y-based algorithm, where the compound was sorted based on their pIC₅₀ value (y), then every Z count, the selected compound was put in the test set.²⁴ Here, the data was sorted both in ascending and descending order.
- Kennard-Stone algorithm²⁵
- Euclidean distance-based algorithm
- Activity/property-based algorithm
- Random selection

Splitting algorithms b, c, and d were performed using Dataset Division Tool 1.2.²⁶ The model was also developed using the whole dataset to produce a broader applicability domain.²⁷

QSAR Model Development

Developing a robust and predictive QSAR model is relatively challenging with small data set. Generally, cross-validation by using an external validation set is the most common method to evaluate those characteristics.^{17,28} However, this statistical technique could pose a problem for this study since it is possible to create a biased result when some compounds are being ‘held out’ for validation purposes, especially in the small dataset.^{17,29} Instead, the double cross-validation (DCV) method was employed here to avoid such issues^{30,31}, where the process is carried out in two loops (inner and outer). Firstly, n number of compounds is allocated from the whole dataset as the basis for the training set. These compounds then split repeatedly into r number of compounds in the validation set and $n-r$ number of compounds in the calibration set. Every possible combination of validation and calibration set (k) is then used to generate the QSAR model and it was assessed for its predictive quality. Ultimately, the best model obtained is then evaluated using the test set compound.^{19,31} In this study, 20% of the total compounds in the dataset (3 compounds) were used as a validation set, and the rest was placed in a calibration set.³² This workflow was fully implemented in Small Dataset Modeler 1.0.0 software.¹⁹ Here, the QSAR model was developed using two approaches: Genetic algorithm-based multiple linear regression (GA-MLR) and partial least square (GA-PLS). After 10 repetitive runs, a model with the lowest MAE value for each replicate was collected and re-evaluated based on several statistical parameters.

Statistical Validation

Statistical evaluation for the inner loop process was performed using various internal validation parameters such as determination coefficient (R^2), adjusted determination coefficient (R^2 adj.), cross-validated correlation coefficient (Q^2), average R_m^2 LOO, ΔR_m^2 LOO, and mean absolute error for fitting after removal of 5% data (MAE_{95%}). A determination coefficient is a measure to determine the variance in the dependent variable which is interpretable by a set of independent variables. Ideally, the value should be equal to 1. A high value indicates the QSAR model fits well with a particular data set. Adjusted determination coefficient has the same purpose, only its value has been adjusted for the numbers of predictors in the model.¹³ The cross-validated correlation coefficient was calculated as a mean to evaluate the robustness of the model against slight change in data set, wherein this instance was performed using leave one out (Q^2 LOO) and leave two out (Q^2 L2O).²⁸ In addition, R_m^2 -based metrics, and MAE_{95%} were also evaluated to verify the predictivity of the model.^{33,34} Meanwhile, the outer loop process was subjected to the calculation of external validation parameters, namely Q^2 F1, Q^2 F2, average R_m^2 test set, ΔR_m^2 LOO test set, CCC, and mean absolute error for fitting after removal of 5% test set data (MAE_{95%} test set). Q^2 F1 and Q^2 F2 are statistical measures regarding external predictivity of the QSAR model, which differ from Q^2 only in the divisor part (Equation 1-3). Q^2 uses TSS (Total Sum of Squares) which sums squared deviation value from the entire dataset mean, while Q^2 F1 and Q^2 F2 refer to the total sum of squared deviation value of the external data set calculated by means of the training set mean and the external set mean, respectively.³⁵⁻³⁷ Ultimately, the concordance correlation coefficient (CCC) is to evaluate the similarity of the observed and predicted value by measuring how far the observation data is situated from the regression line and the deviation of the regression line).^{28,38} Calculation was performed, and generated in Small Dataset Modeler 1.0.0 software.¹⁹

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

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n(ext)} (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_{tr})^2} \quad (2)$$

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n(ext)} (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_{ext})^2} \quad (3)$$

Afterward, the Y-randomization test was performed against the QSAR model which complies with the statistical parameters above. The purpose was to ensure that the obtained model was not from chance correlation.³⁹ During the process, the dependent variable of the training set was scrambled randomly while retaining the value of the independent variables. After 50 permutations, the value of R^2 and Q^2 was observed. The resulting model should possess lower R^2 and Q^2 values than the selected model. Another parameter used as consideration is $^{\circ}\text{Rp}^2$, where the ideal value should be more than 0.5.^{13,40} The calculation

was performed using MLR Y-Randomization 1.2 software.⁴¹ Ultimately, the applicability domain is calculated to define a boundary for a QSAR model in which activity prediction of chemical compounds is applicable.^{13,17} In this study, we examined the reliability of our model using MAE_{95%} LOO and MAE_{95%} LOO + 3 SD criteria as proposed by Roy *et al.*^{31,42} QSAR model with ‘Good’ or ‘Moderate’ quality was selected and then verified its robustness by applying it in the whole unspotted dataset.²⁷ This function is also fully integrated with Small Dataset Modeler 1.0.0¹⁹. In addition, the applicability domain of the model was evaluated using a probability-oriented distance-based approach (AD_{prob-dist}).¹⁸

RESULTS AND DISCUSSION

Variable reduction is one of the most important steps in the QSAR process. This process involves the removal of descriptors with missing values, constant values, and those which are intercorrelated. There are various interpretations of the threshold limit for the latter since no specific rule of thumb regarding this problem exists and each case is known to use a different collinearity threshold with satisfactory results nonetheless.^{42,43} Here, two threshold values were applied to evaluate which workflows produced the better result. From this pre-selection step, a total of 202 and 475 molecular descriptors for the inter-correlation threshold of 0.8 and 0.9, respectively. As of the previous step, dataset division also plays an important role in developing a highly predictive model. The whole dataset is divided into an external set and a modeling set, which is further divided into a training set and a test set. Several approaches can be devised to obtain the best result.^{24,32} In this study, we employed both rational division and random division to yield an acceptable QSAR model. Due to the small dataset utilized (13 compounds), it is in effect not advisable to perform external validation to avoid information loss.^{17,27} However, the external dataset was still allocated in order to better validate the model predictivity. The splitting ratio of the modeling set-external set and training set-test set was kept at 80:20 as shown in several examples.^{27,32} Afterwards, Small Dataset Modeller generated 10 MLR and PLS models for each split. A genetic algorithm (GA) was employed to select the best variables. GA is an evolutionary-based technique through an iterative process mimicking Darwin’s evolutionary theory to develop a better solution for various computational problems.⁴⁴ It is one of the most popular methods in feature selection and has been applied in many QSAR practices.⁴⁵ The resulting QSAR equations were limited to contain only two descriptors, in concordance with the Topliss-Costello rule.⁴⁶ Upon model assessment through various statistical criteria (Table-2) for the both modeling set and the external set, a total of three MLR and two PLS models are deemed acceptable to be processed further. It is noteworthy that all selected equation was obtained using 0.9 as the intercorrelation threshold. We argued that by setting lower threshold values, there might be several important descriptors eliminated and unavailable for model building.⁴⁷ Furthermore, all acceptable QSAR models were obtained *via* the y-based algorithm dataset splitting method, which includes ascending, descending, and software aided activity-based data splitting. This finding is in line with previous studies which showed the advantage of rational dataset splitting over random splitting³², particularly when using a y-based algorithm under GA conditions for variable selection.²⁴ The following process of Y-randomization showed that proposed models possess low Q² and R² values, and °Rp² higher than 0.5. The predictivity of the QSAR models was judged using MAE_{95%} LOO and MAE_{95%} LOO + 3 SD criteria. This method is an error-based metric that has an advantage over the more conventional RMSE or MAE, since it considers the range value used in the data set and has a definite threshold value. There are three categories in which a QSAR model will be classified: Good, Moderate, and Bad. The conclusion will be drawn based on the value of MAE and SD of the 95% dataset, and the pIC₅₀ range value in the training set.^{34,42} Using this metric, only two QSAR equations comply with the criteria among the five. This model consists of one MLR (Equation 4) and one PLS (Equation-5) equation with descriptors such as polarizability-weighted autocorrelation (GATS1p), CATS 2D of a hydrogen donor, and lipophilic pharmacophore points (CATS2D_04_DL), P_VSA-like on Sanderson electronegativity scale (P_VSA_e_5), and spectral moment from augmented edge adjacency matrix weighted by bond order (SM15_AEA(bo)).^{23,48} All statistical parameters of the equations are shown in Table-3. In addition, the closeness of the predicted and observed pIC₅₀ value for both models is depicted in the scatter plot in Fig.-1.

$$pIC_{50} = 5.442(\pm 0.058) + 0.503(\pm 0.109) GATS1p - 1.005(\pm 0.098) CATS2D_04_DL \quad (4)$$

$$pIC_{50} = 5.171 - 0.498 P_VSA_e_5 + 0.395 SM15_AEA(bo) \quad (5)$$

Table-2: Statistical Criteria Implemented for Selection of Acceptable QSAR Models

Parameter	Criteria	Reference
R^2	> 0.7	49
Q^2 LOO; Q^2 LMO	> 0.6	
$R^2 - Q^2$ LOO	< 0.1	
CCC (95%)	> 0.85	
average R_m^2 LOO	> 0.65	42
Q^2F1 ; Q^2F2 (95%)	> 0.7	
ΔR_m^2 LOO	< 0.2	33

Table-3: Statistical Value of Selected QSAR Models

Equation	MLR	PLS
R2	0.943	0.911
R2 adj	0.926	
Q2 LOO	0.867	0.837
Q2 L2O	0.839	
Q2F1 95%	0.972	0.995
Q2F2 95%	0.969	0.994
R2-Q2 LOO	0.076	0.074
R2 adj.-Q2LOO	0.059	
CCC 95%	0.983	0.997
Sc. Avg rm2 LOO	0.828	0.784
Sc. Δ rm2 LOO	0.063	0.008
Sc. Avg rm2 Test 95%	0.987	0.993
Sc. Δ rm2 Test 95%	0	0
R2 yrand.	0.244	0.211
Q2 yrand.	-0.836	-0.665
cRp2	0.834	0.823
MAE+3*SD Train	0.224 (Moderate)	0.2 (Moderate)
MAE+3*SD Test	0.197 (Moderate)	0.083 (Good)

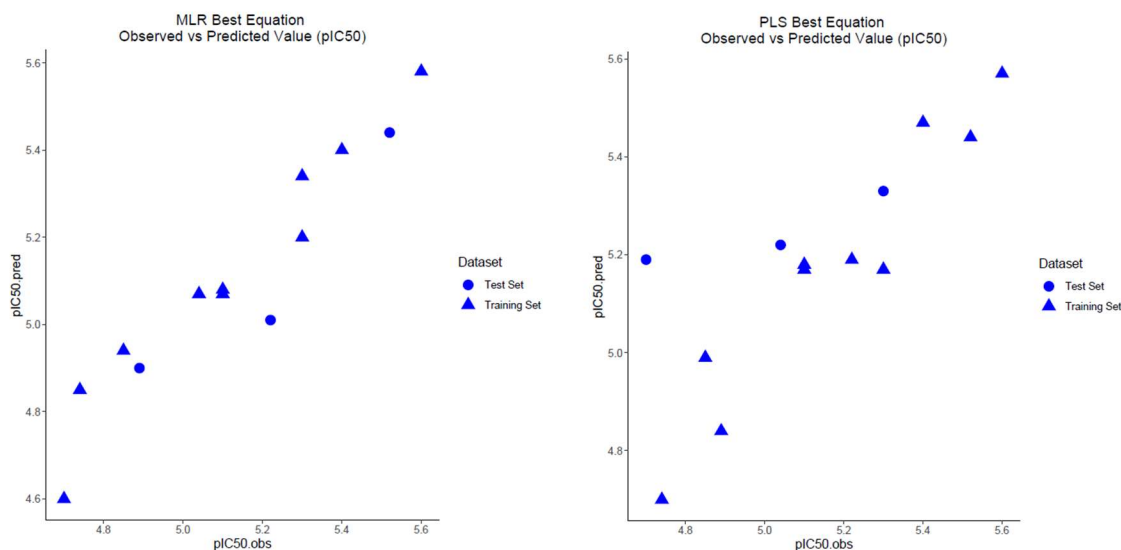


Fig.-1: Scatter Plot of Observed and Predicted IC50 Values of MLR and PLS Model against Pfprs Enzyme

The applicability domain (AD) of the two models was then subsequently examined in order to define the Physico-chemical structural or biological space boundary for which the interpolation of the test set compound is possible. There are various approaches that can be utilized to define the applicability domain of the QSAR model⁵⁰, one of which is the leverage method.¹⁷ Leverage method is based on the distance-to-centroid principle, which describes the distance of a compound from the centroid area of training set data

with the aid of the Williams plot. Despite it being commonly used in many research publications¹⁷, there is a major limitation in which its value is affected by the size of the dataset. Applying the leverage method to small data set (< 20 compounds) might fail to produce data with normal distribution, hence possibly giving an incorrect conclusion.¹⁸ Therefore, in this study we implement the $AD_{prob-dist}$ method as proposed by Gajewicz¹⁸ since it is well suited for evaluating AD in the small dataset. Based on the result (Fig.-2), it can be observed that all training set compounds in both models are within the green zone. This means that they can be categorized as reliable based on a 95% confidence interval from Student's t-test. However, one test set compound in each model is situated outside the green zone. In equation (4), compound 10 is in the orange elliptical curve. It means that special attention should be taken to assess its reliability since the orange zone represents a 95-99% confidence interval boundary. On the other hand, equation (5) depicts one compound from the test set (Compound 4) that falls outside the elliptical curve. This means compound 4 is considered to be highly unreliable, in terms of its predictability. To further verify the robustness of the model, we included the data in the test set and create the QSAR model using the exact same descriptors.²⁷ The resulting equation shows minimum impact with the previous model (Eqn.-6,7).

$$pIC_{50} = 5.462 (\pm 0.065) + 0.511 (\pm 0.093) GATS1p - 1.008 (\pm 0.109) CATS2D_04_DL$$

$$R^2 = 0.912; R^2_{adj} = 0.894; Q^2 = 0.859 \quad (6)$$

$$pIC_{50} = 5.128 - 0.457 p_VSA_e_5 + 0.431 SM15_AEA(bo)$$

$$R^2 = 0.904; Q^2 = 0.855 \quad (7)$$

We have developed MLR and PLS QSAR models for pyrazole-urea-based analogues against the PfPRS enzyme. According to OECD principles, mechanistic interpretation of the QSAR model is preferable in order to better understand the correlation between descriptors and activity.¹³ Based on the overall validation result, we found that equation (4) is more suitable than equation (5), especially upon examining their applicability domain. Therefore, it is argued that descriptors GATS1p and CATS2D_04_DL play important role in correlating the enzymatic activity against PfPRS. The former will help increase the activity of compounds against PfPRS, while the latter descriptor will decrease the bioactivity.

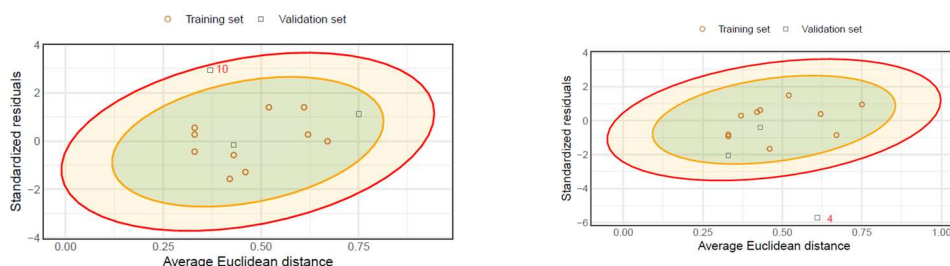


Fig.-2: Applicability Domain of Obtained MLR Model (Equation-4) (Left) and PLS Model (Equation-5) (Right) Using $AD_{prob-Dist}$ Approach

GATS1p is a Geary autocorrelation lag one weighted by atomic polarizability. It belongs to spatial autocorrelation type descriptors, which have been applied extensively in the field of geography and ecology.^{48,51} This concept is adopted into the QSAR field as a measure of the strength of a relationship between observations as a function of space separation, where atoms in a molecule are counted as discrete points and their properties as functions evaluated at those points.²³ GATS1p is calculated from a molecular graph containing hydrogen using the following equation (Eqn.-8).

$$GATS1p = \frac{\frac{1}{2\Delta_k} \sum_{i=1}^N \sum_{j=1}^N (w_i - w_j)^2 \delta(d_{ij}; k)}{\frac{1}{N-1} \sum_{i=1}^N (w_i - \bar{w})^2} \quad (8)$$

Where W_i represents atomic polarizability, $\bar{w}$ is the mean value of atomic polarizability of all N atoms in the molecule, k is the lag value which equals 1, Δ_k is the number of node pairs in a molecular graph at distance equal to 1, d_{ij} refers to the topological distance between atom i and atom j, $\delta(d_{ij}; k)$ is the Kronecker delta which equals to 1 if the value of the interatomic topological distance of i and j is 1, otherwise equals to 0.^{23,52} CATS2D_04_DL belongs to a descriptor family known as CATS (Chemically Advanced Template

Search), which is based on two-dimensional topology. These descriptors classify atoms in the molecule into pharmacophore points such as hydrogen donor/acceptor, positive/negative charge, lipophilic^{23,53}, and aromatic for the more recent version⁵⁴. Any other atom which does not fall into this category will not be considered, for example, C attached to heteroatom or halogen functional group.^{23,53,54} CATS-based descriptors calculated how many combinations of two pharmacophore points were situated at certain topological distances from 0 to 9 chemical bonds.²³ For this instance, this variable counts how many hydrogen bond donor (D) and lipophilic (L) pharmacophore pairs exist in a molecule if the distance between them is 4 chemical bonds.⁵⁵ From the above explanation, we argued that electronic properties play important role in PfPRS activity as represented by GATS1p. Examination of GATS1p values of the dataset indicates that this descriptor yields a high value for the compound substituted with an alkyl chain or bulky group attached with polar atoms, as shown by compounds 12 and 13 which possess high GATS1p scores. Both compounds' substituents are located in the R1 position (Fig.-3), in which additional substituent is favorable due to possible interaction with Tyr746.¹¹ On the other hand, the sum amount of hydrogen bond donor and lipophilic pharmacophore point could also play a role in decreasing PfPRS bioactivity, as represented by CATS2D_04_DL. From the dataset, it can be observed that compounds possessing hydrogen donor point (eg. -OH, -NH) coupled with longer alkyl chain (eg. Compound 4,5,6) have higher CATS2D_04_DL, which in turn potentially lowers their activity against PfPRS enzyme (Table-4).

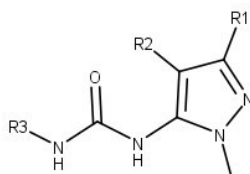


Fig.-3: Parent Structure of Pyrazole-Urea Hybrid with Three Substituent Positions

Table-4: Standardized Descriptor Value of Compounds in Dataset

Compound	GATS1p	CATS2D_04_DL	Compound	GATS1p	CATS2D_04_DL
1	0,29	0,25	8	0,258	0,5
2	0,285	0	9	0,018	0,25
3	0,29	0,5	10	0,145	0,5
4	0,321	1	11	0,267	0,5
5	0,412	0,75	12	0,91	0,5
6	0,321	0,75	13	1	0,5
7	0	0,5			

CONCLUSION

We have developed validated 2D-QSAR models for pyrazole-urea hybrid compounds against the PfPRS enzyme. Using various dataset splitting strategies, we found that rational dataset splitting based on response value has successfully generated the best model, which is in accord with the previous observation.^{24,32} It is also worth noting that all models were obtained using a higher collinearity threshold (0.9 rather than 0.8). The statistical result shows acceptable predictivity on various metrics so far.^{32,34,40,41,49} Judging by their applicability domain¹⁸, MLR obtained model is considered to be generally more reliable than the PLS one. It is argued that electronic parameters such as autocorrelation-based atomic polarizability (GATS1p) and the sum amount of lipophilic and hydrogen bond donor (CATS2D_04_DL) plays important role in predicting the pIC50 value of the pyrazole-urea compound. Nevertheless, more data point is still necessary to build a QSAR model with a wider applicability domain by the addition of compounds with wider pIC50 value and more chemically diverse compound.

ACKNOWLEDGEMENT

The authors would like to thank Prof. Agnieszka Gajewicz-Skretna (University of Gdańsk) for providing the explanation and example of AD_{prob-dist} calculation. This research was funded by Lembaga Penelitian dan Pengabdian Masyarakat University of Surabaya (Letter: 045/ST-Lit/LPPM-01/FF/V/2020).

REFERENCES

1. J. Sachs and P. Malaney, *Nature*, **415**, 680(2002), <https://doi.org/10.1038/415680a>
2. World Health Organization, World Malaria Report 2019. Geneva: World Health Organization (2019).
3. D. Menard and A. Dondorp, *Cold Spring Harbor Perspective in Medicine*, **7(7)**, (2017), <https://doi.org/10.1101/cshperspect.a025619>
4. S. Rout and R. K. Mahapatra, *Chemical Biology & Drug Design*, **93(5)**, 737(2019), <https://doi.org/10.1111/cbdd.13484>
5. M. Yogavel, R. Chaturvedi, P. Babbar, N. Malhotra, V. Jain, and A. Sharma, *Drug Discovery Today*, **23(6)**, 1233(2018), <https://doi.org/10.1016/j.drudis.2018.01.050>
6. T. L. Keller, D. Zocco, M. S. Sundrud, M. Hendrick, M. Edenius, J. Yum, Y-J, Kim, H-K, Lee, J. F. Cortese, D. F. Wirth, J. D. Dignam, A. Rao, C-Y Yeo, R. Mazitschek, and M. Whitman, *Nature Chemical Biology*, **8**, 311(2002), <https://doi.org/10.1038/nchembio.790>
7. J. D. Herman, L. R. Pepper, J. F. Cortese, G. Estiu, K. Galinsky, V. Zuzarte-Luis, E. R. Derbyshire, U. Ribacke, A. K. Lukens, S. A. Santos, V. Patel, C. B. Clish, W. J. Sullivan Jr., H. Zhou, S. E. Bopp, P. Schimmel, S. Lindquist, J. Clardy, M. M. Mota, T. L. Keller, M. Whitman, O. Wiest, D. F. Wirth, and R. Mazitschek, *Science Translational Medicine*, **7(288)** (2015), <https://doi.org/10.1126/scitranslmed.aaa3575>
8. V. Jain, M. Yogavel, Y. Oshima, H. Kikuchi, B. Touquet, M-A. Hakimi, and A. Sharma, *Structure* **23(5)**, 819 (2015), <https://doi.org/10.1016/j.str.2015.02.011>
9. V. Jain, M. Yogavel, H. Kikuchi, Y. Oshima, N. Hariguchi, M. Matsumoto, P. Goel, B. Touquet, R. S. Jumani, F. Tacchini-Cottier, K. Harlos, C. D. Huston, M-A. Hakimi, and A. Sharma, *Structure* **25(10)**, 1495(2017), <https://doi.org/10.1016/j.str.2017.07.015>
10. V. Jain, H. Kikuchi, Y. Oshima, A. Sharma, and M. Yogavel, *Journal of Structural and Functional Genomics*, **15(4)**, 181(2014), <https://doi.org/10.1007/s10969-014-9186-x>
11. S. N. Hewitt, D. M. Dranow, B. G. Horst, J. A. Abendroth, B. Forte, I. Hallyburton, C. Jansen, B. Baragaña, R. Choi, K. L. Rivas, M. A. Hulverson, M. Dumais, T. E. Edwards, D. D. Lorimer, A. H. Fairlamb, D. W. Gray, K. D. Read, A. M. Lehane, K. Kirk, P. J. Myler, A. Wernimont, C. Walpole, R. Stacy, L. K. Barrett, I. H. Gilbert, and W. C. Van Voorhis, *ACS Infectious Diseases*, **3(1)**, 34(2017), <https://doi.org/10.1021/acsinfecdis.6b00078>
12. D. W. Nyamai and Ö Tastan Bishop, *International Journal of Molecular Sciences*, **21(11)**, 3803 (2020), <https://doi.org/10.3390/ijms21113803>
13. K. Roy, S. Kar, and R. N. Das, *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*, Springer, New York (2015).
14. C. Hansch and T. Fujita, *Journal of the American Chemical Society*, **86(8)**, 1616(1964), <https://doi.org/10.1021/ja01062a035>
15. Y. Rifai, F. Permadi, M. N. Amir, and M. Aswad, *Rasayan Journal Chemical*, **14(3)**, 2062(2021), <http://doi.org/10.31788/RJC.2021.1436343>
16. Siswandono, R. Widyowati, A. Suryadi, T. Widiandani, and D. Prismawan, *Rasayan Journal Chemical*, **13(4)**, 2559(2020), <https://doi.org/10.31788/RJC.2020.1345749>
17. P. Gramatica, *QSAR & Combinatorial Science*, **26(5)**, 694(2007), <https://doi.org/10.1002/qsar.200610151>
18. A. Gajewicz, *Environmental Science: Nano*, **5(2)**, 408(2018), <https://doi.org/10.1039/C7EN00774D>
19. P. Ambure, A. Gajewicz-Skretna, M. N. D. S. Cordeiro, and K. Roy, *Journal of Chemical Information and Modeling*, **59(10)**, 4070 (2019), <https://doi.org/10.1021/acs.jcim.9b00476>
20. J. Sadowski, J. Gasteiger, and G. Kleber, *Journal of Chemical Information and Computer Sciences*, **34(4)**, 1000(1993), <https://doi.org/10.1021/ci00020a039>
21. I. Sushko, S. Novotarskyi, R. Körner, A. K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V. Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I. I. Baskin, V. A. Palyulin, E. V. Radchenko, W. J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q-Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, and I. V. Tetko, *Journal of Computer-Aided Molecular Design*, **25(6)**, 533(2011), <https://doi.org/10.1007/s10822-011-9440-2>

22. A. Mauri, 2020, *alvaDesc*: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints, in: K. Roy (Eds.), *Ecotoxicological QSARs*, Humana Press, New York, 801, https://doi.org/10.1007/978-1-0716-0150-1_32
23. R. Todeschini and V. Consonni, 2009, *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical listing/Volume II: Appendices, References*, in: R. Mannhold, H. Kubinyi, G. Folkers (Eds.), *Methods & Principles in Medicinal Chemistry Vol. 41*, John Wiley & Sons, New Jersey.
24. T. Puzyn, A. Mostrag-Szlichtyn, A. Gajewicz, M. Skrzyński, and A. P. Worth, *Structural Chemistry*, **22**(4), 795 (2011), <https://doi.org/10.1007/s11224-011-9757-4>
25. R. W. Kennard and L. A. Stone, *Technometrics*, **11**(1), 137(1969), <https://doi.org/10.1080/00401706.1969.10490666>
26. Dataset Division version 1.2. 2017, Retrieved from http://teqip.jdvu.ac.in/QSAR_Tools/
27. J. C. Dearden, M. Hewitt, and P. H. Rowe, *SAR and QSAR in Environmental Research*, **26**(6), 439 (2015), <https://doi.org/10.1080/1062936X.2015.1046189>
28. P. Gramatica and A. Sangion, *Journal of Chemical Information and Modeling*, **56**(6), 1127(2016), <https://doi.org/10.1021/acs.jcim.6b00088>
29. D. M. Hawkins, S. C. Basak, and D. Mills, *Journal of Chemical Information and Computer Sciences*, **43**(2), 579 (2003), <https://doi.org/10.1021/ci025626i>
30. D. Baumann and K. Baumann, *Journal of Cheminformatics*, **6**, 47(2014), <https://doi.org/10.1186/s13321-014-0047-1>
31. K. Roy and P. Ambure, *Chemometrics and Intelligent Laboratory Systems*, **159**, 18(2016), <https://doi.org/10.1016/j.chemolab.2016.10.009>
32. T. M. Martin, P. Harten, D. M. Young, E. N. Muratov, A. Golbraikh, H. Zhu, and A. Tropsha, *Journal of Chemical Information and Modeling*, **52**(10), 2570 (2012), <https://doi.org/10.1021/ci300338w>
33. P. K. Ojha, I. Mitra, R. N. Das, and K. Roy, *Chemometrics and Intelligent Laboratory Systems*, **107**(1), 194 (2011), <https://doi.org/10.1016/j.chemolab.2011.03.011>
34. K. Roy, R. N. Das, P. Ambure, and R. B. Aher, *Chemometrics and Intelligent Laboratory Systems*, **152**, 18 (2016), <https://doi.org/10.1016/j.chemolab.2016.01.008>
35. L. M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R. M. Blair, W. S. Branham, S. L. Dial, C. L. Moland, and D. M. Sheehan, *Journal of Chemical Information and Computer Sciences*, **41**(1), 186(2001), <https://doi.org/10.1021/ci000066d>
36. G. Schüürman, R-U. Ebert, J. Chen, B. Wang, and R. Kühne, *Journal of Chemical Information and Modeling*, **48**(11), 2140 (2008), <https://doi.org/10.1021/ci800253u>
37. V. Consonni, D. Ballabio, and R. Todeschini, *Journal of Chemical Information and Modeling*, **49**(7), 1669 (2009), <https://doi.org/10.1021/ci900115y>
38. L. I-K. Lin, *Biometrics*, **45**(1), 255 (1989), <https://doi.org/10.2307/2532051>
39. C. Rücker, G. Rücker, and M. Meringer, *Journal of Chemical Information and Modeling*, **47**(6), 2345 (2007), <https://doi.org/10.1021/ci700157b>
40. I. Mitra, A. Saha, and K. Roy, *Molecular Simulation*, **36**(13), 1067(2010), <https://doi.org/10.1080/08927022.2010.503326>
41. MLR Y-Randomization version 1.2. 2015, Retrieved from http://teqip.jdvu.ac.in/QSAR_Tools/
42. L-T. Qin, Y-H. Chen, X. Zhang, L-Y. Mo, H-H. Zeng, and Y-P. Liang, *Chemosphere*, **198**, 122 (2018), <https://doi.org/10.1016/j.chemosphere.2018.01.142>
43. V. M. Alves, A. Golbraikh, S. J. Capuzzi, K. Liu, W. I. Lam, D. R. Korn, D. Pozefsky, C. H. Andrade, E. N. Muratov, and A. Tropsha, *Journal of Chemical Information and Modeling*, **58**(6), 1214(2018), <https://doi.org/10.1021/acs.jcim.8b00124>
44. J. H. Holland, *Adaptation In Natural And Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The MIT Press, Cambridge (1992).
45. M. Goodarzi, B. Dejaegher, and Y. V. Heyden, *Journal of AOAC International*, **95**(3), 636(2012), https://doi.org/10.5740/jaoacint.SGE_Goodarzi
46. J. G. Topliss and R. J. Costello, *Journal of Medicinal Chemistry*, **15**(10), 1066(1972), <https://doi.org/10.1021/jm00280a017>

47. A. Rácz, D. Bajusz, and K. Héberger, *Molecular Informatics*, **38(8-9)**, 18800154(2019), <https://doi.org/10.1002/minf.201800154>
48. R. C. Geary, *The Incorporated Statistician*, **5(3)**, 115(1954), <https://doi.org/10.2307/2986645>
49. N. Chirico and P. Gramatica, *Journal of Chemical Information and Modeling*, **51(9)**, 2320(2011), <https://doi.org/10.1021/ci200211n>
50. D. Gadaleta, G. F. Mangiatordi, M. Catto, A. Carotti, and O. Nicolotti, *International Journal of Quantitative Structure-Property Relationships*, **1(1)**, 45(2016), <https://doi.org/10.4018/IJQSPR.2016010102>
51. A. Getis, *Regional Science and Urban Economics*, **37(4)**, 491(2007), <https://doi.org/10.1016/j.regsciurbeco.2007.04.005>
52. X. Yu, *Ecotoxicology and Environmental Safety*, **190**, 110146(2020), <https://doi.org/10.1016/j.ecoenv.2019.110146>
53. G. Schneider, W. Neidhart, T. Giller, and G. Schmid, *Angewandte Chemie International Edition*, **38(19)**, 2894(1999), [https://doi.org/10.1002/\(SICI\)1521-3773\(19991004\)38:19<2894::AID-ANIE2894>3.0.CO;2-F](https://doi.org/10.1002/(SICI)1521-3773(19991004)38:19<2894::AID-ANIE2894>3.0.CO;2-F)
54. M. Reutlinger, C. P. Koch, D. Reker, N. Todoroff, P. Schneider, T. Rodrigues, and G. Schneider, *Molecular Informatics*, **32(2)**, 133(2013), <https://doi.org/10.1002/minf.201200141>
55. E. Örüçü, G. Tugcu, and M. T. Saçan, *SAR and QSAR in Environmental Research*, **25(12)**, 983 (2014), <https://doi.org/10.1080/1062936X.2014.976266>

[RJC-6811/2021]