

Development of QSAR Model of Caffeic Acid Phenethyl Ester as Anti-Cancer HT-29

Pengembangan Model HKSA Senyawa *Caffeic Acid Phenethyl Ester* (CAPE) sebagai Antikanker HT-29

Uripto Trisno Santoso^{1,2,*}, Samsul Hadi³, Devi Eka Pratama³

¹Computational Chemistry Laboratory, Laboratory of Mathematics and Natural Sciences, Lambung Mangkurat University, Jl. A. Yani KM. 36 Banjarbaru 70714, Kalimantan Selatan

²Chemistry Department, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, Jl. A. Yani KM. 36 Banjarbaru 70714, Kalimantan Selatan

³Pharmacy Department, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, Jl. A. Yani KM. 36 Banjarbaru 70714, Kalimantan Selatan

*Email: utsantoso@ulm.ac.id

ABSTRACT

Caffeic Acid Phenethyl Ester (CAPE) compounds are potentially colorectal anticancer drugs. HKSA (Quantitative Structure-Activity Relationship) research on the CAPE compound class has been carried out, but the model in the previous study did not meet the goodness of fit criteria. The development of the CAPE compound HKSA model as a colorectal anticancer was carried out to obtain a model that meets the goodness of fit criteria and is valid. Eighteen CAPE compounds were used to build the HKSA model using the Multiple Linear Regression (MLR) technique. The descriptor selection was carried out using the backward elimination method and the validation test using the internal leave one out (LOO) cross-validation. The results showed that the HKSA model with four descriptors, namely MDEC22, MDEC23, JG11, and molecular weight (BM), met the goodness of fit and $Q^2(LOO)$ criteria. The development of the HKSA model by adding the LogP descriptor resulted in the HKSA 5 descriptor model with higher goodness of fit level than the HKSA model without the LogP descriptor. Both of these HKSA models have the potential to be used as predictors in the development of a new class of CAPE compounds as anticancer agents against HT-29 cells.

Keywords: Caffeic Acid Phenethyl Ester (CAPE), Quantitative Structure-Activity Relationship (QSAR), Multiple Linear Regression, Internal Validation.

ABSTRAK

Senyawa Caffeic Acid Phenethyl Ester (CAPE) memiliki potensi sebagai obat antikanker kolorektal. Penelitian HKSA (Hubungan Kuantitatif Struktur-Aktivitas) tentang golongan senyawa CAPE telah dilakukan tetapi model pada penelitian sebelumnya tidak memenuhi kriteria goodness of fit. Pengembangan model HKSA senyawa CAPE sebagai antikanker kolorektal dilakukan untuk mendapatkan model yang memenuhi kriteria goodness of fit dan bersifat valid. Sebanyak 18 senyawa CAPE digunakan untuk membangun model HKSA dengan menggunakan teknik Regresi Linier Berganda (RLB). Pemilihan deskriptor dilakukan dengan metode eliminasi backward dan uji validasinya menggunakan validasi silang internal leave one out (LOO). Hasil penelitian menunjukkan bahwa model HKSA dengan 4 deskriptor, yaitu MDEC22, MDEC23, JG11, dan berat molekul (BM) memenuhi kriteria goodness of fit dan $Q^2(LOO)$. Pengembangan model HKSA dengan menambahkan deskriptor LogP menghasilkan model HKSA 5 deskriptor dengan tingkat goodness of fit yang lebih baik daripada model HKSA tanpa deskriptor LogP. Kedua model HKSA ini berpotensi untuk dijadikan prediktor dalam pengembangan golongan senyawa CAPE yang baru sebagai antikanker terhadap sel HT-29.

Kata Kunci: Caffeic Acid Phenethyl Ester (CAPE), Hubungan Kuantitatif Struktur-Aktivitas (HKSA), Regresi Linier Ganda, Validasi Internal.

Received: January 27, 2022; **Accepted:** June 29, 2022; **Available online:** July 31, 2022

1. INTRODUCTION

Cancer is the second leading cause of death in the world. The number of new cases in 2020 shows the three most extensive cancers, namely lung cancer, breast cancer, and colon cancer. Asia has notable cases, 49.3% of the total 19,292,789 people. Asia's cancer mortality rate in 2020 was 58.3% of the total 9,958,133 people (IARC, 2020). Colon cancer is the third largest cancer in the world and the fourth most recent case in Indonesia. Data from IARC (2020) states that there are 10% of the latest cases worldwide and 8.6% of the newest cases in Indonesia. The risk of developing colorectal cancer, according to the American Cancer Society (2020), is 1 in 23 (4.3%) in men and 1 in 25 (4.0%) in women.

Radiotherapy and chemotherapy are frequently employed cancer treatments. Radiation therapy is generally remarkably effective, but there is a risk of damaging normal cells and tumor cells developing radio resistance. The development of radio resistance causes patients' cancer to recur with a more aggressive phenotype. There are several chemotherapy drugs for colorectal cancer, namely 5-fluorouracil, capecitabine, irinotecan, and oxaliplatin (Katzung *et al.*, 2013). Hand-foot syndrome is a side effect of chemotherapy drugs such as capecitabine or 5-Fluorouracil (American Cancer Society, 2020).

5-Fluorouracil has toxic effects such as nausea, mucositis, diarrhea, bone marrow depression, and neurotoxicity (Katzung *et al.*, 2013). Therefore, new compounds for treating colorectal cancer with antitumor action that has high effectiveness and selectivity of cancer cells, as well as low toxicity of normal cells, are urgently needed. Recently, interest in natural compounds has increased significantly as some compounds exhibit significant cytotoxic, antiproliferative, and proapoptotic effects to inhibit cancer cell growth (Kabała-Dzik *et al.*, 2017).

Recent studies have shown that caffeic acid phenethyl ester (CAPE), a component of honeybee propolis, is a natural compound with a strong chemo preventive effect, including cell cycle inhibition and proapoptotic action. These CAPE

compounds have potential as colorectal anticancer drugs (Wadhwa *et al.*, 2016). Analysis of the HKSA (Quantitative Structure-Activity Relationship) of CAPE compounds as colorectal anticancer can be used to learn more about the structural parameters that influence the activity of these compounds as colorectal anticancer.

Ketabforoosh *et al.* (2013) synthesized several CAPE compounds and their derivatives and evaluated their inhibitory activity on HeLa, SK-OV-3, and HT-29 cells. Evaluation of cytotoxic activity showed that the compound had the potential to inhibit HT-29 cancer cells, which are colorectal cancer cells. Using electronegativity molecular descriptors, topological indices, and steric factors, Ketabforoosh *et al.* (2013) analyzed CAPE compounds HKSA. The results of the reported HKSA model show that this model has a correlation coefficient value (R) = 0.66, a determination coefficient (R^2) = 0.44, and a Leave-One-Out or Q^2 (LOO) cross-validation value = 0.44. According to Golbraikh *et al.* (2003), a model may have good predictive power if the model meets several criteria, including $R^2 > 0.6$ and $Q^2(\text{LOO}) > 0.5$. Thus, this model cannot be said to have good predictive power.

This study describes the development of the CAPE compound HKSA model as a colorectal anticancer using topological and physicochemical descriptors to obtain a model that meets the goodness-of-fit criteria and the validity of Q^2 (LOO). The modeling was performed by using the multiple linear regression techniques, and descriptor selection was made by the backward elimination method. Given that one of the factors that may affect the absorption of a drug compound in the target tissue is the level of polarity or hydrophobicity factor, the descriptor selection is conducted while maintaining the presence of a hydrophobic descriptor in each selected model.

2. MATERIALS AND METHODS

2.1. Materials

The materials used were chemical structure data and pIC₅₀ values of 18 Caffeic Acid Phenethyl Ester (CAPE) compounds, the lead compound was shown in Fig. 1. The pIC₅₀ value is the negative value of log IC₅₀, or pIC₅₀ = -log(IC₅₀). The IC₅₀ value is the concentration at which a drug or active compound can inhibit certain biological processes with an inhibition level of 50%. Because the nature of this potential inhibitory value is logarithmic, the decrease in potential from the micromolar to nanomolar level is a logarithmic change, not a linear change, so for linear regression studies, the use of the pIC₅₀ value as the dependent variable will be more appropriate than the IC₅₀ value. Data on the inhibitory activity of HT-29 cancer cells with CAPE compounds were obtained from a molar sample (Ketabforoosh et al., 2013) and converted to pIC₅₀ (Table 1).

The tools used in this study consisted of hardware and software. The hardware was the Asus ZenBook UX305 laptop with specifications: Processor type Intel(R) Core(TM) M3-6Y30 CPU @ 0.90GHz 1.51 GHz, 8 GB Random Access Memory (RAM). The software used was the Windows 10 Home operating system, Marvin Beans 20.19.0, Microsoft Excel 2010, Mordred Descriptor web UI, and HyperChem 8.0.10. The Modred UI web server can be accessed for free at the link

<https://modred.phs.osaka-u.ac.jp>.

2.2. Methodos

Eighteen structures of CAPE compounds obtained from the research of Ketabforoosh et al. (2013) were drawn using HyperChem software. Each structure was then optimized using the Polak-Ribiere algorithm with a convergence limit of 0.1 kcal/(Åmol) and the semi-empirical calculation method Recief Model 1 (RM 1). The structural image files optimized by HyperChem were saved in the file type (*.HIN) format. Furthermore, with the help of the Marvin Beans program, this file type was converted into *.SMI or *.SMILES format. The calculations of descriptor value were performed using three software, HyperChem, Marvin Beans, and Modred Descriptor web UI. The modeling was conducted using the multiple linear regression techniques and the descriptor selection was done by the backward elimination method. The selected model was validated using the Leave-One-Out (LOO) cross-validation technique using equation (1).

$$Q^2(LOO) = 1 - \frac{\sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2} = 1 - \frac{PRESS}{\sum(Y-\bar{Y})^2} \quad (1)$$

Y is the value of the experimental activity, \hat{Y} is the predictive activity value, and \bar{Y} is the average value of the experimental activity. A model is declared valid if it has a value of $Q^2 > 0,5$.

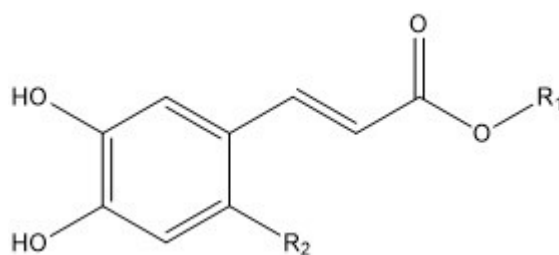
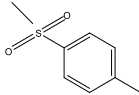
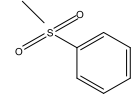
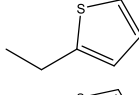
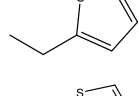
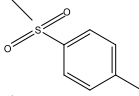
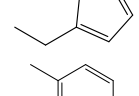
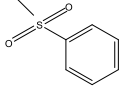


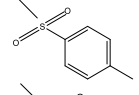

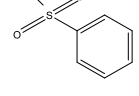


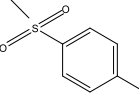

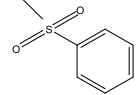
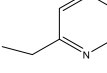
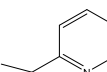
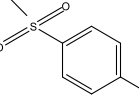
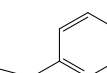
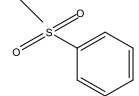


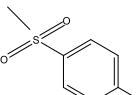

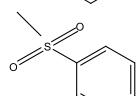


Figure 1. Structure of caffeic acid phenethyl ester (CAPE) with R₁ and R₂ substituent

Table 1. The structure and pIC₅₀ value of CAPE compounds

Compound	R ₁	R ₂	pIC ₅₀ HT-29
1	-	-	4.600
2	-		4.288
3	-		4.314
4		-	4.427
5			4.075
6			4.270
7		-	3.728
8			3.619
9			3.489
10		-	3.896
11			3.678
12			3.740
13		-	4.365
14			4.130
15			3.740
16		-	4.242
17			4.017
18			4.135

Source: Data structure and pIC₅₀ values from Ketabforoosh et al. (2013) processed.

3. RESULTS AND DISCUSSION

3.1. Calculation and Selection of Descriptors

The method of calculating the electronic structural properties used in this study was RM1 (Recife Model 1). In general, the RM 1 method is superior to other semi-empirical methods seen from the average error in calculating the heat of formation parameter, dipole moment, ionization potential, and interatomic distance. The compounds used in the RM1 parameterization include various compounds containing C, H, N, O, P, S, F, Cl, Br, and I atoms (Rocha et al., 2006) so that the RM 1 method will be suitable if used as a method of calculating the structural descriptor of CAPE and its derivative compounds consisting of C, H, O, N and S atoms. Three applications (software) were used to calculate the descriptor value: HyperChem, Modred UI, and Marvin Beans.

The selection of descriptors in this study begins with a multicollinearity test. The multicollinearity test aims to avoid correlation or multicollinearity between descriptors so that all selected descriptors are not correlated with each other. The descriptor from Marvin Beans had a high correlation, so this descriptor was not used for further testing. The results of the multicollinearity test on 47 descriptors from Modred UI and 11 descriptors from HyperChem obtained ten descriptors that did not show high multicollinearity (Table 2). Therefore, these descriptors were chosen for the trial development of the HKSA model.

The steric parameter is an effect that correlates with the spatial arrangement of molecules in three-dimensional space. It is important to monitor the binding of chemicals to biological receptors (Roy et al., 2015). The steric properties affect the molecular energy, the reaction and conformational pathways, the reaction rate and equilibrium, the binding affinity between the ligand and the receptor, and other thermodynamic properties (Todeschini & Consonni, 2000). The steric constant of the substituents can be measured based on the appearance of groups and the effect of the groups on drug contact with adjacent receptor sites (Siswandono, 2016). The descriptors MDEC22, MDEC23, and molecular weight (BM) in Table 2 are the steric parameters evaluated in this study.

Topological descriptors were calculated based on the graphical representation of the molecule. They do not require estimating physicochemical properties or the rigorous computations involved in deriving quantum chemical descriptors (Roy et al., 2015). The topological charge index can evaluate charge transfer between pairs of atoms (Todeschini & Consonni, 2000). The ability to describe the charge distribution of a molecule is determined by relating it to the dipole moment of a heterogeneous set of hydrocarbons, the boiling temperatures of alkanes and alcohols, and the enthalpies of evaporation of alkanes (Galvez et al., 1994). Based on Table 2, the descriptors associated with the topological load index in this study are JGI1, JGI8, and JGI10.

The descriptor related to the hydrophobicity factor (lipophilicity) often used in HKSA is the logarithm of the partition coefficient (logP). The compound's hydrophobicity represents how likely it is for the compound to enter through the cell membrane causing damage and the ability of the compound to interact with its receptors. The logP value can describe the distribution of the drug in the body. If the logP value is positive, the compound tends to be in a non-polar phase (hydrophobic), and if the logP value is negative, the compound tends to be in a polar phase (hydrophilic). SlogP is an octanol-water partition coefficient developed by Wildman and Crippen (1999) to overcome several problems in calculating the logP value. SlogP and LogP in Table 2 are descriptors representing hydrophobic properties.

Electronic descriptors can affect how easily a drug can pass through a cell membrane or how strongly a drug can bind to a receptor. Electronic descriptors also affect the drug distribution process and the penetration of biological membranes, which is strongly influenced by the solubility of the drug in fat/water as well as in the structure-activity relationship and how strong these effects can interact between drugs and receptors (Siswandono, 2016). The groups with a dipolar function (with a dipole moment) are the ester group and CAPE, which is included in the ester group. The hydration energy and dipole moment in Table 2 are descriptors that represent electronic properties.

Table 2. List of selected descriptors that are not correlated

No	Name	Group	Description
1	MDEC22*	Steric	The length of the molecular distance edge connecting the secondary C atoms
2	MDEC23*	Steric	The length of the molecular edge connecting the secondary and tertiary C atoms
3	SlogP*	Hydrophobic	LogP Wildman-Crippen
4	JGI1*	Topology	1-Ordered Mean Topological Charge Index (1-Ordered Mean Topological Charge)
5	1JGI8*	Topology	8 order average topological charge index
6	JGI10*	Topology	10 order topological charge index
7	Hydration Energy**	Electronics	The energy released when one mole of ions is hydrated
8	BLogP**	Hydrophobic	Logarithm of partition coefficient between octanol and water
9	BM**	Steric	Relative molecular weight
10	dipole moment**	Electronics	A vector quantity used to express the polarity of a molecule

* = Descriptors in Modred UI

** = Descriptors in HyperChem

3.2. HKSA Model Development

Ten selected descriptors (Table 2) were tested for developing the HKSA model using multiple linear regression techniques. The best model was explored by selecting the appropriate descriptor through the backward elimination method. Based on the trial, we

obtained the best eight models (Model 1-8) with a coefficient of determination (R^2) $\geq 0,6$ and a $F_{ratio} > 1$, as listed in Table 3. Models 9 and 10 did not meet the goodness of fit criteria because of the $R^2 < 0,6$ and $F_{ratio} < 1$.

Table 3. Linear regression statistical data for ten HKSA equation as the model candidate

Model	F_{ratio}	R^2	R	SE	Adj R^2
1 ^a	1.991	0.912	0.955	0.148	0.786
2 ^b	2.696	0.911	0.955	0.138	0.812
3 ^c	3.527	0.910	0.954	0.131	0.830
4 ^d	4.137	0.901	0.949	0.131	0.831
5 ^e	4.080	0.934	0.873	0.141	0.804
6 ^f	4.499	0.853	0.924	0.145	0.792
7 ^g	3.682	0.783	0.885	0.170	0.716
8 ^h	2.424	0.635	0.797	0.212	0.556
9 ⁱ	0.611	0.231	0.480	0.298	0.128
10 ^j	0.480	0.119	0.345	0.309	0.064

^aDescriptors: MDEC22, MDEC23, SlogP, JGI1, JGI8, JGI10, Energy of Hydration, LogP, BM, Dipole Moment^bDescriptors: MDEC22, MDEC23, JGI1, JGI8, JGI10, Energy of Hydration, LogP, BM, Dipole Moment^cDescriptors: MDEC22, MDEC23, JGI1, JGI8, JGI10, LogP, BM, Dipole Moment^dDescriptors: MDEC22, MDEC23, JGI1, JGI8, LogP, BM, Dipole Moment^eDescriptors: MDEC22, MDEC23, JGI1, JGI8, LogP, BM^fDescriptors: MDEC22, MDEC23, JGI1, LogP, BM^gDescriptors: MDEC22, MDEC23, JGI1, BM^hDescriptors: MDEC22, MDEC23, JGI1ⁱDescriptors: MDEC22, MDEC23^jDescriptors: MDEC23

F_{ratio} value > 1 can be achieved if the value of $F_{count} > F_{table}$. The F-test, in this case, is the overall significance test used to evaluate whether the regression model

provides higher goodness of fit when compared to models that do not contain independent variables. Regression models that do not contain predictors are also known

as intercept-only models. The hypothesis for this significance test is as follows:

- Null hypothesis: there is no significant difference in the level of goodness of fit between the model containing only the intercept (without the descriptor) and the proposed model (containing the selected descriptor),
- Alternative hypothesis: there is a significant difference in the goodness of fit between the model without descriptors and the proposed model.

The data in Table 3 shows that models 1 to 8 have a $F_{\text{ratio}} > 1$ while models 9 and 10 have a $F_{\text{ratio}} < 1$. Based on this F_{ratio} value, only models 1 to 8 have the null hypothesis rejected. Thus, only models 1 to 8 have descriptors that significantly affect the goodness of fit as a whole.

The goodness of fit in this model can be seen from the coefficient of determination (R^2) value. This value shows the proportion of the value of the variation of biological activity that can be explained by the predicted results of the model (OECD, 2007). The higher value of R^2 (closer to 1 or -1) can be obtained if the data distribution gets closer to the trend line. In linear regression, it indicates the stronger the linear relationship between the dependent variable and the independent variable. Thus, the value of R^2 can be a requirement that the model meets the goodness of fit in multiple linear regression. The criteria for the model selected based on the value of R^2 is a model with a value of $R^2 \geq 0.6$. The value of $R^2 \geq 0.6$ can be interpreted that the model can explain 60% of the variation in biological activity. Judging from the R^2 value, only models 1 to 8 meet these criteria (model 9 and model 10 do not meet the criteria).

However, the value of R^2 , in general, tends to increase when the number of independent variables (descriptors) increases, even though it does not contribute a significant effect. Therefore, another parameter is needed for correction, namely adjusted R^2 (R^2_{adj}). R^2_{adj} is the value of R^2 that has been adjusted or corrected concerning the number of descriptors. This value will decrease if additional descriptors have no significant effect (OECD, 2007). The difference between the accepted values of R^2 and R^2_{adj} is less than 0.3 (Veerasingam et al., 2011). All models (1-8) have a difference

R^2 value and R^2_{adj} smaller than 0.3. Therefore, the number of descriptors used in the model is still acceptable.

A p-value analysis was carried out to see the descriptors that had a significant effect on the compound's activity, and the results are presented in Table 4. Based on Table 4, each descriptor model 6, model 7, model 9, and model 10 has a p-value descriptor < 0.05 . Because model 9 and model 10 have a value of $R^2 \geq 0.6$, if the p-value requirements are combined with the R^2 value requirements, only model 6 and model 7 meet the R^2 value and p-value criteria.

Models 6 and 7 of this study, compared with the previous model (Ketabforoosh et al., 2013), show that these two models have a better level of goodness of fit (Table 5). Models 6 and 7 have $R^2 \geq 0.6$, while the previous model has $R^2 \leq 0.6$. Model 6 uses MDEC22, MDEC23, JGI1, LogP, and BM descriptors. Model 7 uses MDEC22, MDEC23, JGI1, and BM descriptors, while the previous research model uses GATS1e and GATS3v. It indicates that the selection of the descriptor type can affect the quality of the obtained HKSA model. The goodness of fit level is very closely related to the significance of the descriptor effect as the independent variable and activity as the dependent variable. The higher fit in models 6 and 7 compared to the previous model indicates that the descriptor used in models 6 and 7 has a more significant effect than the descriptor in the previous model.

Table 4. Regression coefficient and p-value data from ten HKSA equation models

Model	1		2		3		4		5	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Intercept	-9.940	0.211	-9.275	0.164	-10.984	0.005	-8.952	0.002	-8.441	0.003
MDEC22	0.525	0.024	0.501	0.005	0.523	0.001	0.452	0.000	0.433	0.000
MDEC23	-0.789	0.039	-0.759	0.018	-0.816	0.001	-0.669	0.000	-0.633	0.000
SlogP	-0.063	0.842	-	-	-	-	-	-	-	-
JGI1	83.418	0.065	78.601	0.024	86.361	0.000	83.469	0.000	80.362	0.000
JGI8	262.317	0.250	260.266	0.221	229.392	0.194	78.026	0.123	64.409	0.217
JGI10	207.046	0.395	204.134	0.368	167.397	0.360	-	-	-	-
Hydrating Energy	0.058	0.750	0.054	0.750	-	-	-	-	-	-
LogP	-0.575	0.211	-0.532	0.155	-0.442	0.046	-0.376	0.061	-0.417	0.051
BM	0.009	0.138	0.008	0.051	0.008	0.032	0.006	0.025	0.007	0.023
Dipole Moment	0.022	0.546	0.025	0.377	0.026	0.328	0.038	0.126	-	-

Model	6		7		8		9		10	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Intercept	-9.639	0.001	-6.548	0.005	-4.982	0.054	4.289	6.27E-09	4.530	4.6E-10
MDEC22	0.478	0.000	0.352	0.000	0.319	0.001	0.053	1.60E-01	-	-
MDEC23	-0.691	0.000	-0.603	0.000	-0.419	0.001	-0.054	5.21E-02	-0.025	1.6E-01
SlogP	-	-	-	-	-	-	-	-	-	-
JGI1	89.760	0.000	66.870	0.000	59.376	0.001	-	-	-	-
JGI8	-	-	-	-	-	-	-	-	-	-
JGI10	-	-	-	-	-	-	-	-	-	-
Hydrating Energy	-	-	-	-	-	-	-	-	-	-
LogP	-0.463	0.033	-	-	-	-	-	-	-	-
BM	0.008	0.006	0.008	0.011	-	-	-	-	-	-
Dipole Moment	-	-	-	-	-	-	-	-	-	-

Table 5. Results of statistical analysis of model 6, model 7, and previous research

Statistical Parameters	Model 6	Model 7	Previous Model*	Criteria
F _{ratio}	4,499	3,682	1,447	>1
R ²	0,853	0,783	0,415	≥ 0,6
R ² _{adj}	0,792	0,716	0,337	-
(R ² - R ² _{adj})	0,061	0,067	0,078	< 0,3

Note: *Model according to Ketabforoosh *et al.* (2013)

Based on Table 5 and Table 4, two models of HKSA equations that meet these criteria can be drawn up: Equation (2) and Equation (3).

$$pIC_{50} = -9,639 + 0,478MDEC22 - 0,691MDEC23 + 89,760JGI1 - 0,463logP - 0,008BM \quad (2)$$

$$pIC_{50} = -6,548 + 0,352MDEC22 - 0,603MDEC23 + 66,870JGI1 - 0,008BM \quad (3)$$

The HKSA model in Equation 2 contains a logP descriptor, while the HKSA model in Equation 3 does not contain a logP descriptor. The logP coefficient in Equation 2 is negative. If the logP is positive, the greater the logP value will result in a lower pIC₅₀ activity value. On the other hand, if the logP is negative, the larger the logP value, the higher the pIC₅₀ activity value. A positive logP value indicates the compound tends to be in a non-polar phase (hydrophobic). Conversely, a negative logP value indicates that the compound tends to be more soluble in a polar phase (hydrophilic). Thus, for non-polar compounds, there is a tendency that increasing the logP value will decrease the activity (pIC₅₀) of CAPE derivatives. On the other hand, for polar compounds, there is a tendency that increasing the logP value (which is negative) will increase the activity (pIC₅₀) of CAPE derivatives. Thus, this logP parameter becomes particularly important in the design or development of CAPE derivative compounds, considering that CAPE compounds can be developed into polar or non-polar compounds (Hashimoto *et al.*, 2021).

The HKSA model of equation (2) and equation (3) show that the JGI1 descriptor is the descriptor that has the most positive influence on the pIC₅₀ value than the other descriptors. It indicates that the 1-Ordered Mean Topological Charge index is also important to consider in the design of CAPE-

derived compounds with high potential as colorectal anticancer HT-29 cells.

3.3. Model Validation

The estimation of the two HKSA models' predictive power (equation (2) and equation (3)) was carried out using the LOO cross-validation technique test. The validity criterion used is that a model is said to be valid if the value of Q²(LOO) > 0.5 (Veerasamy *et al.*, 2011). The validation test results on Model 6 and Model 7 show that Model 6 has a Q²(LOO) value of 0.702 while Model 7 has a Q²(LOO) value of 0.569. Thus, these two models can be declared valid based on the Q²(LOO) criteria.

In addition to estimating how robust the model predictions are for new compounds that are not in the data, the Q² value can also be used as a reference to assess the possibility of overfitting. This value can also be an indication of overfitting in the model, which is characterized by the difference between R² and Q² > 0.3 (Veerasamy *et al.*, 2011). Based on Q² and R² values, it can be seen that model 6 has a Q² and R² difference of 0.151, while model 7 has a Q² and R² difference of 0.214. Therefore, it can be declared that there is no overfitting in model 6 and model 7.

Based on the Q²(LOO) internal cross-validation test, it appears that model 6 and model 7 meet the criteria as valid models. However, if viewed from the value of the F_{ratio} and R², it appears that model 6 is better than model 7. In addition, model 6 has a PRESS value of 0.516 while model 7 was 0.745. It indicates that the addition of the hydrophobicity descriptor (logP) can improve the quality of the CAPE compound HKSA model as an anticancer HT-29. The comparison of the quality of model 6 (Equation (2)) and model 7 (Equation (3)) can be seen in the scatter diagram of the predicted activity value versus the

experimental activity value in Fig. 2. Fig. 2 shows that the goodness of fit level of model 6 is better than model 7.

Fig. 3 shows a comparison diagram of the predicted activity values of model 6 and model 7, which are very close to the experimental values. Fig. 3 also shows that model 6 has a predictive activity value closer to the experimental activity value than model 7. Overall, it can be stated that model 6 and model 7 are models that meet the $Q^2(\text{LOO})$ validity criteria, but model 6 has highest goodness of fit level than model 7. Based on Table 3, model 6 has a correlation coefficient value (r) of 0.924 while model 7 has a correlation coefficient value of 0.885. Thus, Equation (2) (derived from model 6) tends to have better predictive ability than Equation (3) (from model 7).

Based on Equation (2), to obtain CAPE-derived compounds with higher colorectal anticancer activity in HT-29 cells, new compounds can be developed by: increasing the value of the MDEC22 descriptor and increasing the value of JGI1 or decreasing the value of the MDEC23 descriptor. It should be noted that the derivatization in the production of new compound will generally result in a larger molecular weight even though the effect is relatively small, so the addition of a new group or side chain should not be too large. The design of new CAPE-derived compounds must also pay attention to molecular polarity. For non-polar compounds, a decrease in the $\log P$ value will tend to increase activity. On the other hand, for polar compounds, an increase in the $\log P$ value will increase the activity (pIC_{50}) of CAPE derivatives.

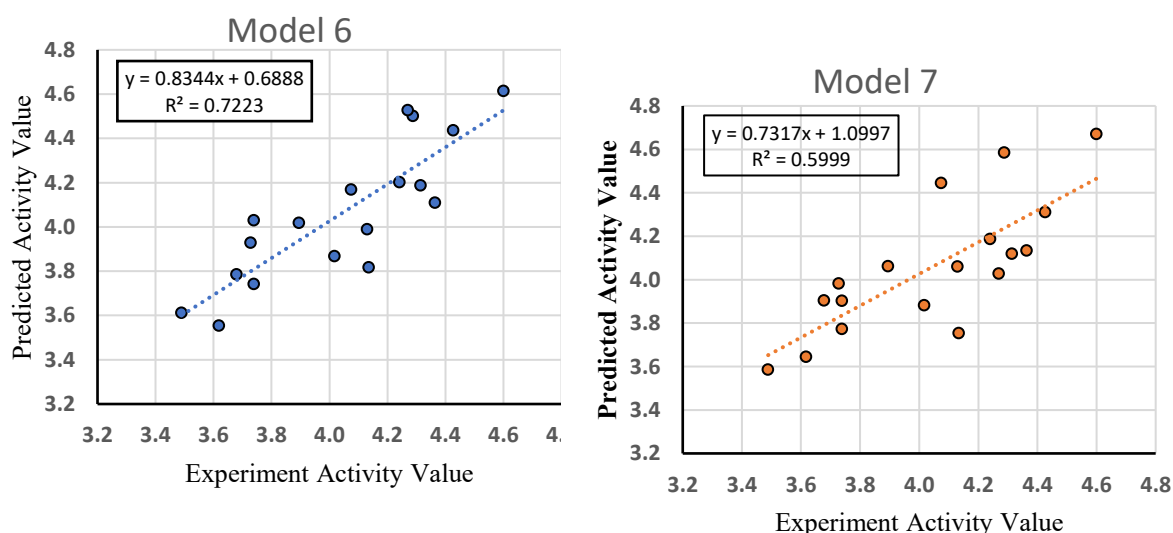


Figure 2. Scatter diagram of the relationship between the experimental activity value vs. the predicted activity value from the HKSA model Equation (2) and Equation (3).

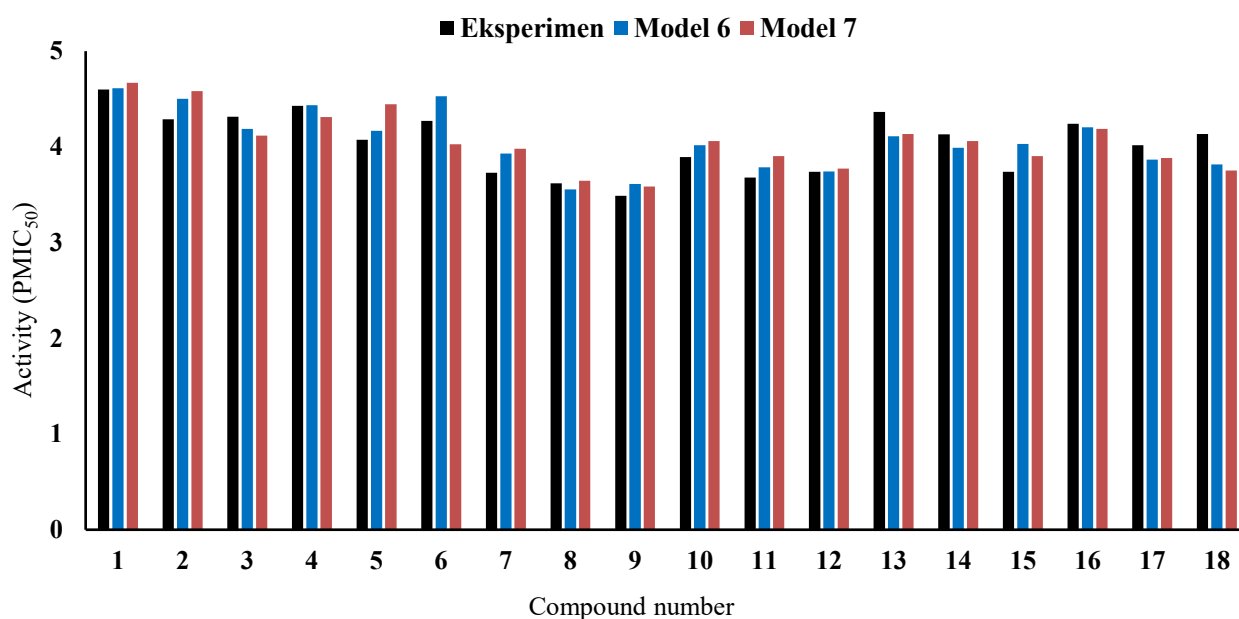


Figure 3. The predictive activity value diagram of the HKSA model Equation 2 and the HKSA model Equation 3 are compared with the experimental activity values.

4. CONCLUSIONS

Based on the data from the research and discussion described, it can be concluded that by using four descriptors, namely MDEC22, MDEC23, JGI1, and BM, the HKSA model meets the validity criteria of $Q^2(\text{LOO})$. The addition of the LogP hydrophobic factor descriptor also resulted in an HKSA model that met the $Q^2(\text{LOO})$ validity criteria. The development of the HKSA model by adding the LogP descriptor resulted in the 5 descriptor HKSA model with higher goodness of fit level than the HKSA model without the LogP descriptor. Based on the goodness of fit and the validity of $Q^2(\text{LOO})$, these two HKSA models have the potential to be used as predictors or aids in the development of a new class of CAPE compounds as anticancer against HT-29 cells with higher activity.

LIST OF REFERENCES

- American Cancer Society. (2020). *General Counsel American Cancer Society, Inc.* Atlanta, Georgia.
- Galves, J., R. Garcia, M. T. Salabert, R. Soler. (1994). Charge Indexes New Topological Descriptors. *Journal of Chemical Information and Modeling*, 34(3): 520-525.

- Golbraikh A., Shen M., Xiao Z., Xiao Y.D., Lee K.H. & Tropsham A. (2003). Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design*, 17, 241–253.
<https://doi.org/10.1023/A:1025386326946>

- Hashimoto R, Lai H, Fujita R, Hanaya K, Higashibayashi S, Inoue H, & Sugai T. (2021). *Global Cancer Observatory: Cancer Today*. Chemoenzymatic semisynthesis of caffeic acid β -phenethyl ester, an antioxidative component in propolis, from raw coffee bean extract. *Bioscience, Biotechnology, and Biochemistry*, 85 (3), 476-480.
<https://doi.org/10.1093/bbb/zbaa077>

- Kabała-Dzik, A., Rzepecka-Stojko, A., Kubina, R., Jastrzebska-Stojko Z., Stojko, R., Wojtyczka, R.D., and Stojko, J. (2017). Comparison of Two Components of Propolis: Caffeic Acid (CA) and Caffeic Acid Phenethyl Ester (CAPE) Induce Apoptosis and Cell Cycle Arrest of Breast Cancer

- Cells MDA-MB-231. *Molecules*, 22(1554), 1-15.
- Katzung, B. G., S. B. Masters & A. J. Trevor. (2013). *Farmakologi Dasar & Klinik*. EGC, Jakarta.
- Ketabforoosh S.H.E., Amini M, Vosooghi M, Shafiee A, Azizi E. & Kobarfard F. (2013). Synthesis, evaluation of anticancer activity and QSAR study of heterocyclic esters of caffeic acid. *Iranian Journal of Pharmaceutical Research*, 12(4), 705–719.
- OECD. (2007). *Guidance Document On the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models*. Organisation de Coopération et de Développement Economiques, France.
- Roy, K., S. Kar & R. N. Das. (2015). *A Primer on QSAR/QSPR Modeling Fundamental Concepts*. Springer, New York.
- Rocha, G. B., R. O. Freire, A. M. Simas & J. P. Stewart. (2006). RM1: A Reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br and I. *Wiley InterScience*, 27(10),1101-1111.
- Siswandono. (2016). *Kimia Medisinal Jilid 1*. Airlangga Universitas Press, Jakarta.
- Todeschini, R. & V. Consonni. (2000). *Handbook of Molecular Descriptor*. Wiley-VCH, Germany.
- Veerasamy, R., H. Rajak, A. Jain, S. Sivadasan, C. P. Varghesel & R. K. Agrawal. (2011). Validation of QSAR Models Strategies and Importance. *International Journal of Drug Desain and Discovery*, 2, 511-519.
- Wadhwa, R., Nigam, N., Bhargava, P, Dhanjal, J.K., Goyal, S, Grover, A., Sundar, D., Ishida, Y., Terao, K., Kau, S. (2016). Molecular characterization and enhancement of anticancer activity of caffeic acid phenethyl ester by γ cyclodextrin. *Journal of Cancer*, 7(13), 1755–1771. <https://doi.org/10.7150/jca.15170>
- Wildman, S. A. & G. M. Crippen. (1999). Prediction of Physicochemical Parameters by Atomic Contributions. *J.Chem.Inf.Comput.Sci*, 39(5), 868-873. <https://doi.org/10.1021/ci990307l>