ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

# Comparative Study of Machine Learning-Based QSAR Modeling of Anti-inflammatory Compounds from Durian Extraction

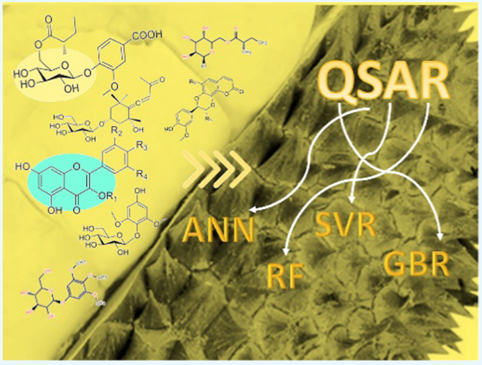
Amphawan Wiriyarattanakul,[∇] Wanting Xie,[∇] Borwornlak Toopradab, Sopon Wiriyarattanakul, Liyi Shi, Thanyada Rungrotmongkol,* and Phornphimon Maitarad*

Cite This: *ACS Omega* 2024, 9, 7817−7826

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Quantitative structure−activity relationship (QSAR) analysis, an *in silico* methodology, offers enhanced efficiency and cost effectiveness in investigating anti-inflammatory activity. In this study, a comprehensive comparative analysis employing four machine learning algorithms (random forest (RF), gradient boosting regression (GBR), support vector regression (SVR), and artificial neural networks (ANNs)) was conducted to elucidate the activities of naturally derived compounds from durian extraction. The analysis was grounded in the exploration of structural attributes encompassing steric and electrostatic properties. Notably, the nonlinear SVR model, utilizing five key features, exhibited superior performance compared to the other models. It demonstrated exceptional predictive accuracy for both the training and external test datasets, yielding $R^2$ values of 0.907 and 0.812, respectively; in addition, their RMSE resulted in 0.123 and 0.097, respectively. The study outcomes underscore the significance of specific structural factors (denoted as shadow ratio, dipole *z*, methyl, ellipsoidal volume, and methoxy) in determining anti-inflammatory efficacy. Thus, the findings highlight the potential of molecular simulations and machine learning as alternative avenues for the rational design of novel anti-inflammatory agents.



## 1. INTRODUCTION

Polyphenols, found abundantly in durian extracts, have gained significant attention and have been extensively studied for their diverse medicinal properties.[1,2] These compounds possess a wide range of therapeutic potentials, including strong anticancer effects by inhibiting the growth and spread of various cancer cell types.[3] Additionally, they play a key role in regulating insulin secretion and utilization, leading to improved blood glucose levels.[4,5] Moreover, polyphenols contribute to the modulation of intestinal microbiota, thereby reducing intestinal inflammation and enhancing the integrity of the intestinal barrier, an important strategy for addressing gastrointestinal inflammation.[6] Among their therapeutic attributes, the antioxidant and anti-inflammatory abilities of polyphenols are particularly noteworthy.[7,8]

Focusing specifically on durian peel, previous studies have highlighted its exceptional nutritional composition and rich content of bioactive compounds. Feng et al. noted the presence of triterpenoids and glycosides in durian peel, which have demonstrated anti-inflammatory effects via the inhibition of lipopolysaccharide-induced nitric oxide production in the RAW 264.7 cell line.[9] Furthermore, extracts from different durian cultivars, such as Monthong and Chanee, have shown varying degrees of antioxidant and anti-inflammatory activities.[10] The antioxidant effects of durian peel are attributed to flavonoids and phenolic compounds, with the coumarin derivative propacin exerting significant inhibitory effects on lipopolysaccharide-induced nitric oxide and prostaglandin E2 (PGE2) release in RAW 264.7 cells.[1]

However, the complex structural arrangements of these natural constituents and their multifaceted anti-inflammatory mechanisms make the elucidation of specific biochemical targets and reaction pathways challenging.[11,12] In response to this complexity, quantitative structure−activity relationship (QSAR) analysis proves to be a valuable in silico strategy. Through the development of accurate predictive models, this approach enables the exploration of the intricate relationship between the molecular structure and activity in natural products.[13,14]

Several QSAR models have been proposed to investigate the antioxidant activity of flavonoids.[15−18] Although linear models have offered useful insights, they face difficulties in explaining the complex relationship between structural factors and antioxidant activity.[19,20] In this regard, machine learning algorithms, such as random forests (RFs), support vector

**Received:** September 25, 2023
**Revised:** November 15, 2023
**Accepted:** December 27, 2023
**Published:** February 7, 2024

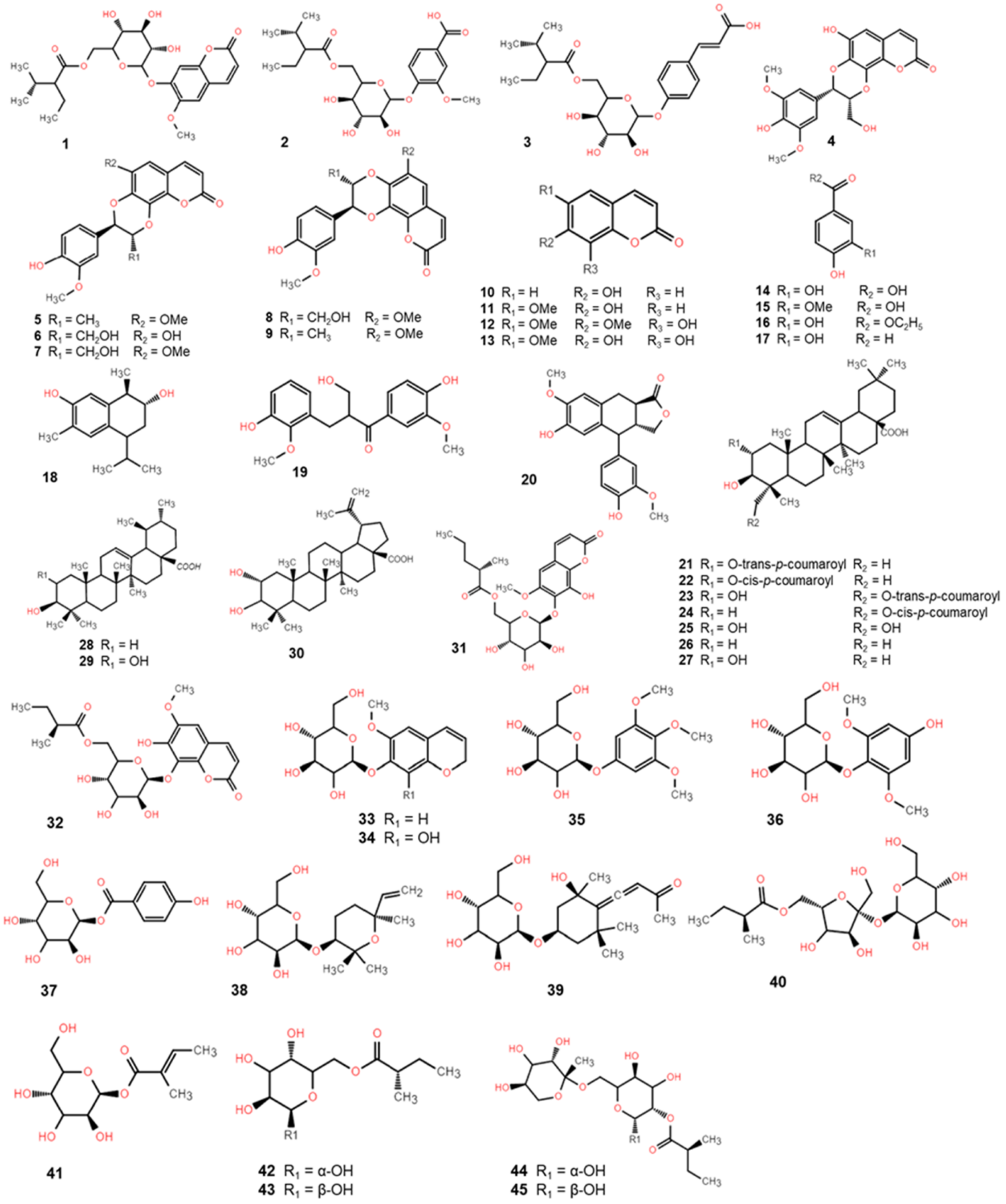


**Figure 1.** Two-dimensional structures of molecules extracted from durian.

regression (SVR),[21] and artificial neural networks (ANNs), will be introduced to offer enhanced predictive capabilities; additionally, the QSAR with machine learning models can help over new information from the existing data.[22−24] For instance, in a study conducted by Li et al., they attained an impressive coefficient of determination ($R^2$) of 0.807 by employing four significant descriptors and the multiple linear regression (MLR) method to forecast the antioxidant activity of polysaccharides. Furthermore, a comparative analysis of results from the multilayer perceptron artificial neural network (MLP-ANN) model revealed even more precise predictions, showcasing an exceptional $R^2$ value of 0.944.[25]

To shed light on the molecular structure−activity relationship, a dataset comprising bioactive chemicals extracted from durian peel was utilized for machine learning-based QSAR analysis in this work. The robustness of these models bodes well for guiding future compound design and predicting the antioxidant activity. The investigations can reveal key structural characteristics impacting the anti-inflammatory activity of these compounds, offering both theoretical directions and the potential to molecularly fine-tune newly designed anti-inflammatory chemicals inspired by nature.

## 2. METHODS

**2.1. Data Sources and Collection.** The NO inhibitory activity of 45 natural bioactive chemicals extracted from durian shells (depicted in Figure 1) was assessed, as documented by Feng et al.[9,26] and presented in Table 1. It is worth noting that

Table 1. Anti-Inflammatory Activity Data of Compounds Extracted from Durian

| | NO inhibition | | | NO inhibition | | | NO inhibition | |
|---|---|---|---|---|---|---|---|---|
| cpd. | $pIC_{50}$ | $IC_{50}$ | cpd. | $pIC_{50}$ | $IC_{50}$ | cpd. | $pIC_{50}$ | $IC_{50}$ |
| 1 | 4.441 | 36.220 | 16 | 5.137 | 7.290 | 31 | 4.000 | >100.000 |
| 2 | 4.419 | 38.070 | 17 | 4.795 | 16.030 | 32 | 4.000 | >100.000 |
| 3 | 4.384 | 41.260 | 18[a] | 4.301 | 50.000 | 33 | 4.000 | >100.000 |
| 4[a] | 4.453 | 35.230 | 19 | 4.483 | 32.910 | 34 | 4.000 | >100.000 |
| 5 | 4.301 | 50.000 | 20 | 4.511 | 30.820 | 35 | 4.000 | >100.000 |
| 6 | 4.664 | 21.700 | 21 | 5.184 | 6.550 | 36 | 5.082 | 8.280 |
| 7 | 4.539 | 28.880 | 22 | 4.610 | 24.560 | 37[a] | 4.680 | 20.900 |
| 8 | 5.449 | 3.560 | 23[a] | 4.576 | 26.570 | 38 | 4.000 | >100.000 |
| 9 | 4.301 | 50.000 | 24 | 4.519 | 30.280 | 39 | 4.000 | >100.000 |
| 10 | 4.519 | 30.280 | 25 | 4.766 | 17.130 | 40 | 4.547 | 28.380 |
| 11 | 4.585 | 26.010 | 26 | 5.023 | 9.480 | 41 | 4.000 | >100.000 |
| 12 | 4.551 | 28.150 | 27 | 4.801 | 15.820 | 42 | 4.498 | 31.750 |
| 13 | 5.432 | 3.700 | 28[a] | 4.967 | 10.800 | 43 | 4.498 | 31.750 |
| 14 | 4.748 | 17.870 | 29 | 5.470 | 3.390 | 44 | 4.634 | 23.230 |
| 15 | 4.501 | 31.530 | 30 | 4.527 | 29.720 | 45 | 4.634 | 23.230 |

[a] Samples serving as the external validation test set in building the QSAR model.

the dataset in this work covered the main natural bioactive chemicals found in the durian shell extraction, including phenols, glycosides, and pentacyclic triterpenoids. Notably, compounds 21−30 fall within the pentacyclic triterpenoid classification; the remaining substances are classified as phenolic compounds and glycosides, with some overlap due to the presence of phenolic hydroxyl groups in certain glycoside aglycones. To facilitate QSAR model manipulation, the NO inhibitory data, initially presented as $IC_{50}$ values, were transformed into $pIC_{50}$ ones.

Based on the QSAR analysis, the Kennard-Stone algorithm was applied to select the external set. There were five compounds (compounds 4, 18, 23, 28, and 37) that served as an external set for the model evaluations, which are shown in Table 1.

**2.2. Structural Optimization and Feature Generation.** The three-dimensional (3D) structures of the investigated molecules were constructed, and their geometry was subsequently optimized at the B3LYP/6-31G(d,p) level of theory using the Gaussian 16 package.[27] The QSAR module in Materials Studio (MS) 8.0[31] was utilized to calculate the structural descriptors (features) based on the optimized structures. 96 valid features, including spatial, electronic, thermodynamic, topological, E-state, fragment, and molecular geometry descriptors, were generated as seen details in Tables S1 and S2.

**2.3. Multicollinearity of Features.** The challenge of a limited dataset coupled with an extensive feature set (45 structures, each with 96 distinct attributes) and the redundancy of the descriptors can be mitigated by selecting descriptors that are the most relevant to the response variable. Multicollinearity describes the state where the independent variables used in a study exhibit a strong relationship with each other; this might pose a problem since the independent variables in a desired model should preferably be independent. Herein, variance inflation factor (VIF) analysis, as one evaluation metric,[28] was conducted to tackle this issue. The VIF represents the ratio of the variance in the presence of multicollinearity between explanatory variables to that in its absence and is calculated as in eq 2.1

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} \quad i = 1, 2, \cdots, k \tag{2.1}$$

where $R_i^2$ refers to the coefficient of determination when the $i$th explanatory variable acts as the explanatory one while the remaining $k - 1$ variables are used to perform linear regression. A large $R_i^2$ value implies that the remaining variables have a high explanatory property along with the $i$th variable. The larger the $R_i^2$ value, the larger the $\mathrm{VIF}_i$ value; for example, when $R_i^2$ equals 0.9, $\mathrm{VIF}_i$ equals 10. It is generally believed that strong multicollinearity exists when VIF > 10, in which case some variables need to be eliminated.

VIF analysis was conducted on all 96 descriptors utilizing the statsmodels[29] toolkit, an open-source Python statistical package. The descriptors were ranked based on computed VIF values, with those above a VIF threshold of 10 being progressively removed. This iterative process was continued until all retained descriptors exhibited VIF values below 10. Additionally, a correlation coefficient analysis was performed among the selected descriptors. The details of VIF analysis codes are shared in Code S1 of the Supporting Information.

**2.4. QSAR Modeling Algorithms.** The QSAR approach was employed to develop a mathematical model, with descriptors as the independent variables ($X$) and $pIC_{50}$ values as the response variable ($Y$), to elucidate the complicated connection between the physicochemical properties (descriptors) and the anti-inflammatory activities of the studied durian compounds. All descriptors' values of $Y$ and $X_1$, $X_2$, $X^3$, ⋯, $X_{96}$ are listed in Table S3. In this work, four different machine learning techniques were applied to generate QSAR models for a comparative study: RFs, gradient boosting regression (GBR), ANNs, and SVR. The Python package Scikit-learn[30] was used to complete the entire modeling procedure. Each of the QSAR models, based on the four algorithms, was subjected to grid search cross-validation hyperparameter fine tuning during the model training procedure. An exhaustive search over the parameter values for every model was conducted using GridSearchCV, which means that the accuracy score metric was employed along with fivefold cross-validation. The best parameter values that maximized the accuracy score on the validation set were selected for further modeling. The details of RF, GBR, SVR, and ANN Python codes of parameter optimizations are shared in Code S1 of the Supporting Information.

### 2.5. QSAR Model Evaluation Methods.

Prior to modeling, a subset of five compounds was manually selected as a test set based on the distribution of $\text{pIC}_{50}$ values. The remaining compounds constituted the training dataset for QSAR model development and internal validation, while the test set was reserved for external validation. The quality of the fit between the model-predicted and experimental values and the suitability of the regression models in terms of their ability to describe the data were measured by using the $R^2$ value of the training set. This value ranges from 0 to 1 (with negative values also possible for poorly fit models); the closer $R^2$ is to 1, the higher the accuracy of the model's predictions.

The cross-validation metric assists in evaluating a model's performance by dividing the data into multiple subsets, training the model on some of these subsets, and evaluating it on the remaining ones. With 10-fold cross-validation, the data will be divided into 10 subsets, with training the model on 9 of them and testing it on the remaining one. An estimate of the model's performance that is more stable could be obtained by averaging over these 10 partitions. This is particularly crucial when data availability is limited.

In this work, the stability of the QSAR models was assessed using the 10-fold cross-validation $R^2(\text{CV})$ and $R^2$ values of the test dataset. The models' accuracy was evaluated using the root-mean-square error (RMSE), that is, the square root of the mean absolute error (MAE); the smaller the RMSE, the smaller the prediction error of the model and the better the model's performance. $R^2$ and the RMSE were calculated as in eqs 2.2 and 2.3, respectively

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\acute{y}_1 - y_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{2.2}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{n} (y_i - \acute{y}_i)^2} \tag{2.3}$$

where $n$ represents the number of molecules used in the model prediction; $y_i$ and $\acute{y}_i$ represent the true and predicted $\text{pIC}_{50}$ values of the $i$th molecule, respectively; and $\overline{y}$ represents the average of the true $\text{pIC}_{50}$ values of the $n$ predicted molecules.

### 2.6. Feature Importance.

Permutation importance and SHAP (Shapley additive explanations) are two commonly used methods for feature importance analysis and can explain the outputs of any machine learning model. Permutation importance is based on model performance and evaluates the importance of a feature by randomly shuffling its values and observing the changes in the model's prediction accuracy or loss function. The advantages of this approach lie in its simplicity, generality, and model-agnostic nature. However, it does present some drawbacks; it is unable to reveal the nature of the relationship (positive or negative) between features and prediction outcomes and ignores the interactions between features. SHAP, on the other hand, is rooted in game theory, quantifying the importance of a feature by calculating its contribution to the prediction outcome. Its strengths include its ability to reflect the impact and polarity of all features within each sample and to consider the interactions between features. Its drawbacks, however, include higher computational complexity and a reliance on specialized toolkits for implementation. In this work, SHAP values were computed by using the SHAP interpretability model Python package.

## 3. RESULTS AND DISCUSSION

### 3.1. Feature Analysis.

The stepwise VIF selection method, outlined in Table 2, identified 11 descriptors that persisted within the dataset. Notably, the methyl and ellipsoidal volume descriptors exhibit comparably high VIF values, suggesting a potential correlation between these features and necessitating further investigation. Among the 11 descriptors, the five exhibiting the strongest correlation with the predictive variable ($\text{pIC}_{50}$ value) are shadow ratio, dipole $z$, methyl, and E-state keys (sums): S_dssC, and Methoxy. Importantly, their respective correlation coefficients (r) with the $\text{pIC}_{50}$ values exceed 0.1, indicating significant contributions to predicting the NO inhibitory activity. Conversely, the $r$ value characterizing the relationship between the ellipsoidal volume and $\text{pIC}_{50}$ is −0.057. Figure 2 depicts the correlation matrix of the 11 molecular descriptors.

**Table 2. Eleven Selected Descriptors with Calculated VIF Values**

| no. | descriptors | VIF |
|---|---|---|
| 1 | E-state keys (sums): S_dssC | 2.017 |
| 2 | E-state keys (sums): S_aasC | 2.295 |
| 3 | E-state keys (sums): S_aaaC | 2.024 |
| 4 | E-state keys (sums): S_ssssC | 1.877 |
| 5 | methoxy | 2.949 |
| 6 | methyl | 9.028 |
| 7 | ellipsoidal volume | 9.991 |
| 8 | shadow ratio | 4.147 |
| 9 | dipole $x$ | 1.130 |
| 10 | dipole $y$ | 1.130 |
| 11 | dipole $z$ | 1.303 |

It is worth noting that Pearson correlation analysis primarily captures linear associations between variables and may not fully elucidate nonlinear relationships. The specific role of the ellipsoidal volume in explaining molecular activity remains uncertain. Thus, the subsequent section delves into a comprehensive analysis and discussion of the QSAR model, drawing insights from the four investigated machine learning algorithms and the permutation importance of the featured attributes within the models.

### 3.2. Regression Models Based on 11 Selected Features.

Among the four algorithms, the SVR model demonstrates the most promising overall performance, achieving a significantly high $R^2$ value of 0.975 in the training set, with $R^2(\text{CV})$ and external validation $R^2$ values of 0.851 and 0.862, respectively (Table 3). In contrast, while the ANN model exhibits exceptional performance in the training set ($R^2 = 0.995$), it displays overfitting issues, with an $R^2(\text{CV})$ of 0.724 and an external validation $R^2$ of 0.450.

Among the ensemble learning algorithms, the GBR model displays an enhanced training performance compared to the RF model. However, the latter exhibits higher prediction accuracy in the external validation set, with an $R^2$ of 0.854, while the former yields an $R^2$ of 0.734. Notably, the minimal disparity between the $R^2(\text{CV})$ values of the GBR and RF models relative to their respective training set $R^2$ values indicates the robustness of the ensemble learning models.

https://doi.org/10.1021/acsomega.3c07386

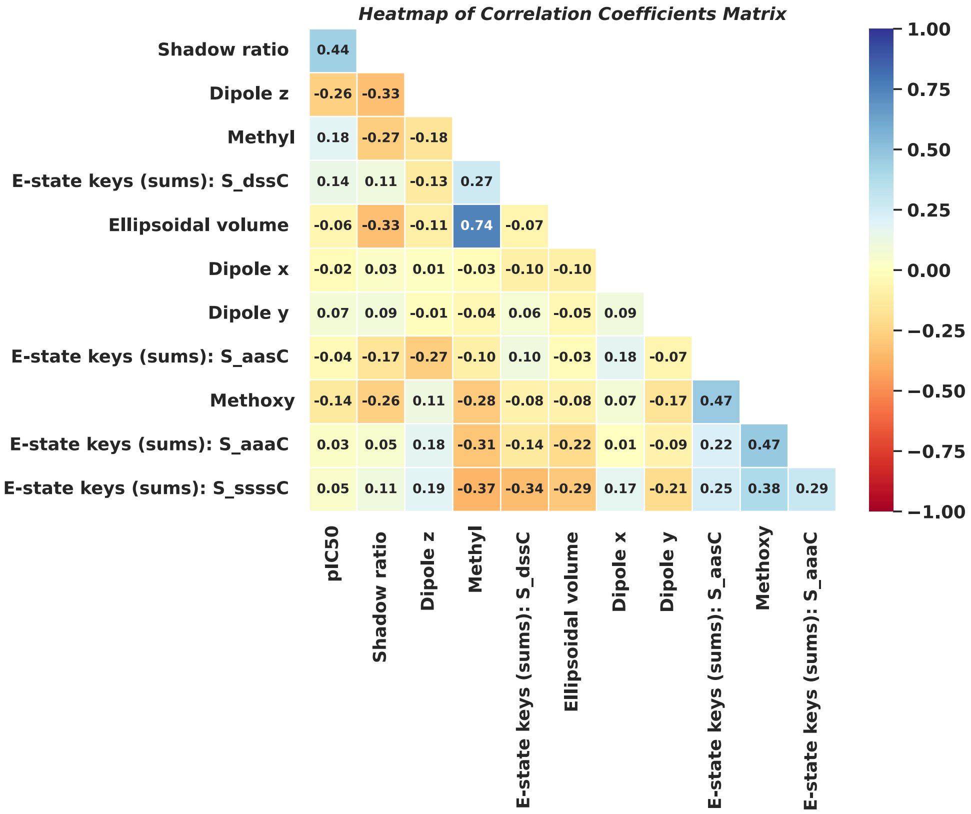


**Figure 2.** Correlation matrix of 11 molecular descriptors.

**Table 3. Statistical Parameters for the Four Developed Models Are Based on 11 Selected Features**

| methods | evaluation | $RF_{11}$ | $GBR_{11}$ | $SVR_{11}$ | $ANN_{11}$ |
|---|---|---|---|---|---|
| full train | RMSE | 0.195 | 0.157 | 0.063 | 0.028 |
| | $R^2$ | 0.766 | 0.849 | 0.975 | 0.995 |
| 10-fold CV | RMSE(CV) | 0.227 | 0.197 | 0.148 | 0.199 |
| | $R^2$(CV) | 0.680 | 0.757 | 0.851 | 0.724 |
| external | RMSE | 0.086 | 0.116 | 0.083 | 0.167 |
| | $R^2$ | 0.854 | 0.734 | 0.862 | 0.450 |

Note: The models are based on 11 molecular features selected from the VIF analytical results.

Moreover, for a comprehensive exploration of the significance of the 11 molecular features regarding anti-inflammatory activity, a detailed permutation feature importance analysis was conducted across the four models. The outcomes are visually represented in Figures 3(a−d), which indicate a trend of consistently high rankings for the shadow ratio and methyl molecular attributes across all models. In particular, within the RF model, the top three features comprise shadow ratio, methyl, and dipole *z*, while the remaining eight exhibit importance values below 0.1.

In the GBR model, ellipsoidal volume assumes a higher position relative to dipole *z*, indicating its relevance. In contrast, within the SVR and ANN models, methoxy exhibits more prominence than E-state keys (sums): S_dssC. This distinction suggests the varying roles of the methoxy descriptor in different algorithm types, with the SVR and ANN models attributing greater importance to it due to their ability to capture complex functional mappings.

**3.3. Simplified Model with Five Key Features.** To enhance clarity and conciseness, a streamlined model incorporating only five essential molecular attributes (shadow ratio, dipole *z*, methyl, E-state keys (sums): S_dssC, and methoxy) was developed (Table 4). The initial selection of these attributes was based on their high correlation coefficients with anti-inflammatory activity. However, given the significance of ellipsoidal volume in the ensemble learning models, an alternative configuration was explored as well, replacing E-state keys (sums): S_dssC with an ellipsoidal volume (Table 4).

The outcomes, as provided in Table 4, indicate that although the models using a reduced feature set exhibit slightly diminished predictive performance, they maintain commendable accuracy. For instance, the RF and GBR models utilizing five attributes display minimal alterations in their evaluation coefficients for the training set with deviations below 0.05. This resilience can be attributed to the relatively lower scores assigned to the discarded molecular features, as highlighted by the permutation importance analysis conducted on the ensemble learning tree models.

The SVR and ANN models, however, demonstrate significantly greater declines in predictive capacity, indicating a sensitivity to feature reduction. Notably, the reduction in the degree of overfitting is evident in the ANN model after including a reduced number of molecular features, as shown in Table 4. Furthermore, from the eight distinct QSAR models in Table 4, the SVR model incorporating the shadow ratio, dipole *z*, methyl, ellipsoidal volume, and methoxy descriptors showcases superior performance, maintaining high prediction accuracy and demonstrating robust generalization capability. The evaluation metrics for this model are as follows: a training set $R^2$ of 0.907 and RMSE of 0.123; an $R^2$(CV) of 0.770 and RMSE(CV) of 0.191; and an external validation $R^2$ of 0.812 and RMSE of 0.097.

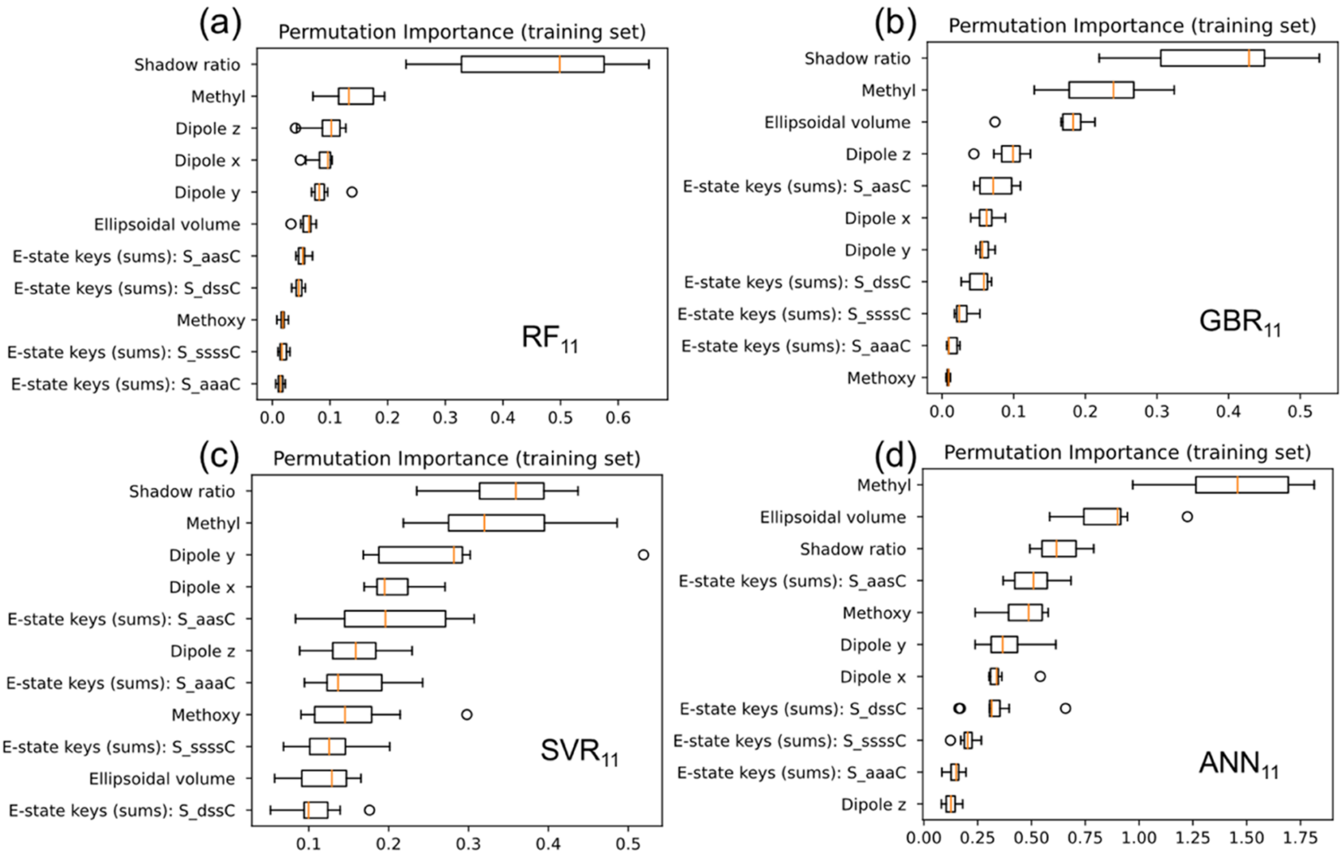


**Figure 3.** Permutation importance of QSAR models, including (a) $RF_{11}$, (b) $GBR_{11}$, (c) $SVR_{11}$, and (d) $ANN_{11}$.

**Table 4. Statistical Parameters for the Four Developed Models Based on Five Selected Features**

| | full train | | 10-fold CV | | external | |
|---|---|---|---|---|---|---|
| model[a] | RMSE | $R^2$ | RMSE(CV) | $R^2$(CV) | RMSE | $R^2$ |
| $RF_{5a}$ | 0.208 | 0.735 | 0.223 | 0.691 | 0.117 | 0.727 |
| $GBR_{5a}$ | 0.159 | 0.846 | 0.190 | 0.769 | 0.126 | 0.686 |
| $SVR_{5a}$ | 0.143 | 0.874 | 0.186 | 0.785 | 0.148 | 0.561 |
| $ANN_{5a}$ | 0.158 | 0.846 | 0.219 | 0.690 | 0.146 | 0.574 |
| $RF_{5b}$ | 0.212 | 0.724 | 0.231 | 0.670 | 0.115 | 0.738 |
| $GBR_{5b}$ | 0.178 | 0.806 | 0.209 | 0.730 | 0.138 | 0.622 |
| $SVR_{5b}$ | 0.123 | 0.907 | 0.191 | 0.770 | 0.097 | 0.812 |
| $ANN_{5b}$ | 0.153 | 0.857 | 0.214 | 0.711 | 0.295 | −0.726 |

[a] Model 5a: The five molecular features used in the models are shadow ratio, dipole *z*, methyl, E-state keys (sums): S_dssC, and methoxy. Model 5b: The five molecular features used in the models are shadow ratio, dipole *z*, methyl, ellipsoidal volume, and methoxy.

### 3.4. Structural Characteristics of Anti-inflammatory Compounds.

The SHAP interpretability model was applied to analyze the relationship between individual molecular features and anti-inflammatory activity within the top-performing SVR model ($SVR_{5b}$) discussed in Section 3.3. The analysis combined feature implications with Shapley values, which quantify each feature's contribution to the model's output.

The SHAP summary graph in Figure 4 offers a comprehensive overview by combining feature importance and positive or negative correlation effects. Each data point corresponds to a sample instance, with the *x*-coordinates determined by Shapley values and the *y*-coordinates influenced by the average of the Shapley absolute values for a given feature across all samples. The color of each point represents the significance of the associated Shapley value, conveying the importance of the features in a broader context. The methyl descriptor is the most significant feature, with the SHAP values indicating a negative relationship between the presence of methyl groups in the molecular structure and anti-inflammatory activity. However, seven significant anomalous samples deviate from this trend, showing a more positive SHAP value contribution. Table 5 presents the definitions of the five molecular features included in Figure 4.

The SHAP dependence graph in Figure 5 provides an overview of each feature's impact on the model's interpretation, helping elucidate how changes in feature values affect predictions and offering insights into the data. Furthermore, Figure 6(a),(b) depict a 3D molecular structure (using compound 8 as an example) along the coordinate axis directions.

Figure 5(a) shows the SHAP dependence graph for the methyl feature, indicating the Shapley contributions linked to different values. The eight samples with the highest Shapley values contain structures with six or seven methyl groups, corresponding to durian-extracted pentacyclic triterpenes with elevated $pIC_{50}$ values. Some molecules with one or two methyl groups have positive but relatively modest Shapley values (below 0.1), suggesting a limited influence on the overall model.

The molecular feature with the second highest average Shapley value is the shadow ratio, as shown in Figure 5(b). Molecules with shadow ratio values exceeding 2 have more significant positive Shapley contributions. This pattern, combined with insights into the molecules' 3D structure,

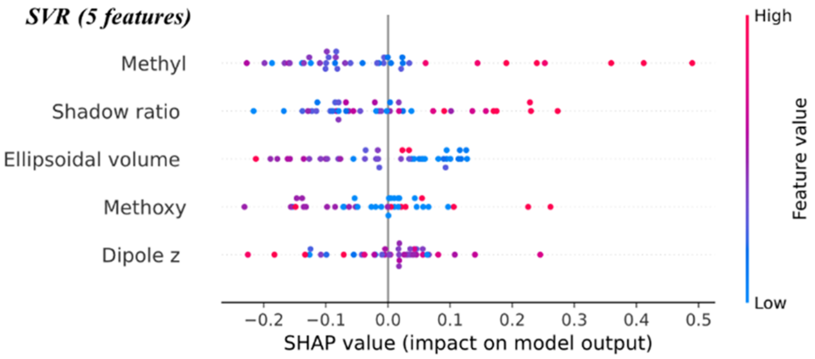


**Figure 4.** SHAP summary diagram of the SVR model.

### Table 5. Definitions of Five Molecular Features

| feature | definition |
|---|---|
| shadow ratio | A spatial Cartesian coordinate system is established based on the moment of inertia of the molecules; generally, the *x*-axis lies along the direction where the distance between molecules is the greatest, and the *z*-axis is in the direction where the distance between molecules is the smallest. The shadow areas of the molecular surface projected onto three mutually perpendicular planes, XY, YZ, and XZ, are calculated, with "shadow ratio" describing the ratio between the largest and smallest dimensions. |
| methyl | The number of methyl groups in the molecular structure |
| dipole *z* | Dipole moment (debye) of the molecule in the *z*-axis direction |
| ellipsoidal volume | Ellipsoidal volume (Å$^3$) of the molecule, describing the volume of an inertial spheroid derived from the inertial tensor of the system |
| methoxy | The number of methoxy groups in the molecular structure |

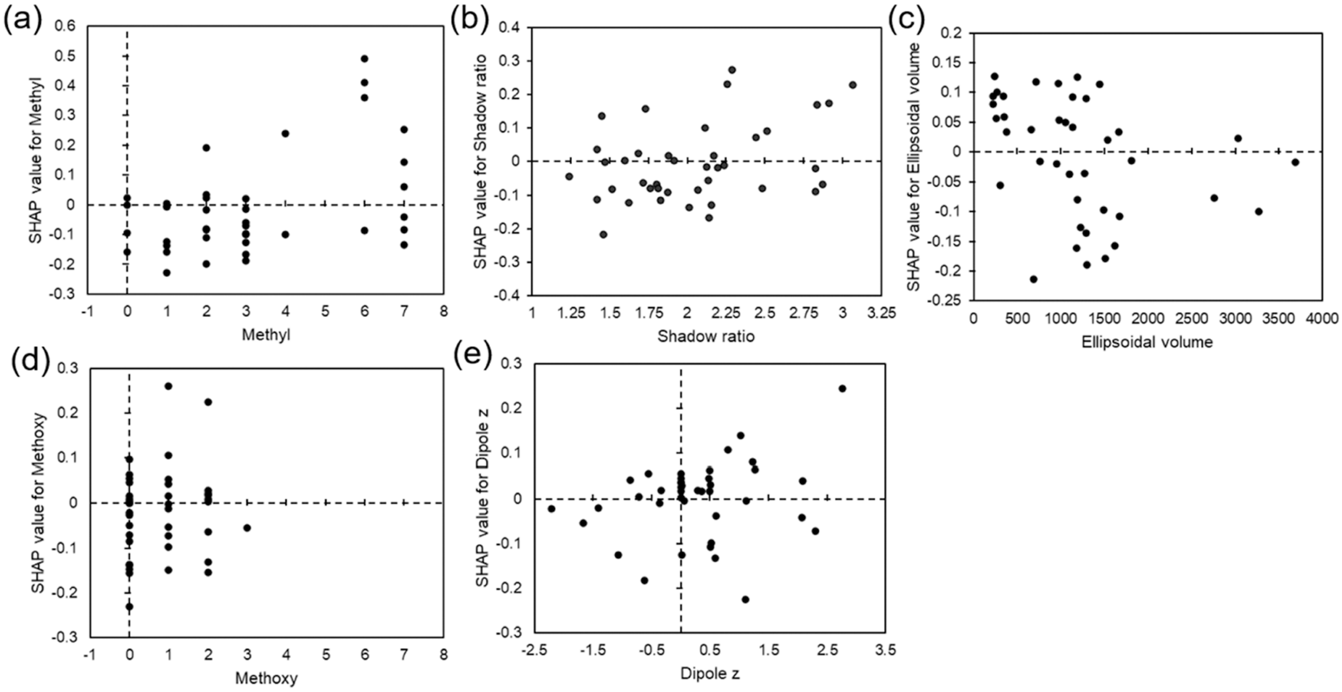


**Figure 5.** SHAP dependence graph of each feature in the SVR$_{5b}$ model. The *x*-axis represents distinct feature values for each sample, while the *y*-axis indicates the Shapley contribution of the corresponding feature within the model.

suggests that a higher shadow ratio corresponds to an overall surface contour resembling an ellipsoid.

The ellipsoidal volume of a molecule is the third pivotal molecular feature that affects its anti-inflammatory activity. When it exceeds 1000 Å$^3$, a noticeable decrease in the Shapley value is observed, indicating a negative correlation. A higher ellipsoidal volume of drug molecules can impede their capacity to engage in biochemical reactions or interact with target binding sites within the cellular matrix. Thus, retaining the ellipsoidal volume feature is valuable for assessing the impact of molecular volume on drug bioavailability.

The presence of methoxy groups within a molecular structure is denoted as methoxy. As portrayed in Figure 5(d), the contribution from Shapley values is more pronounced for molecules containing one or two methoxy groups, as opposed to none. Moreover, most of the methoxy groups contained in structures exhibit a positive contribution to anti-inflammatory activity, while a single group imparts a

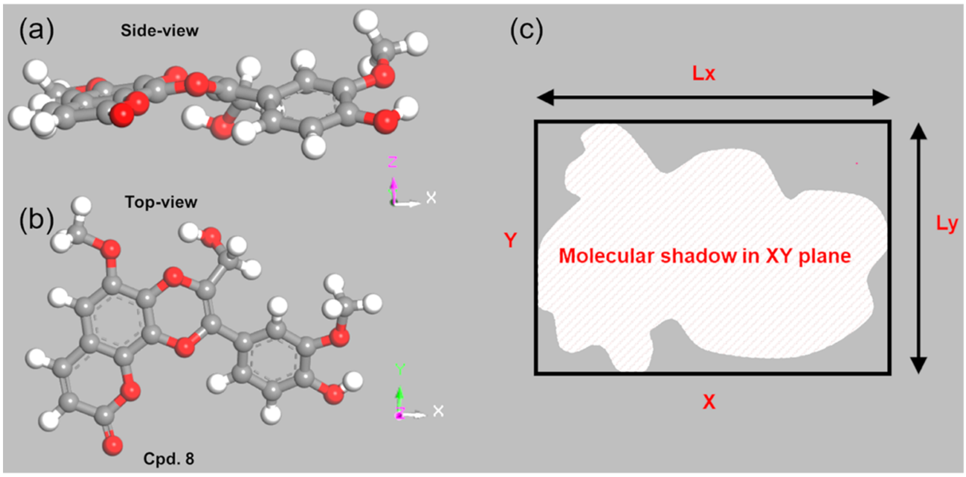


**Figure 6.** (a) Side and (b) top views of the 3D structure of compound 8 (Cpd 8) and (c) schematic diagram of a planar projection of the molecular surface.

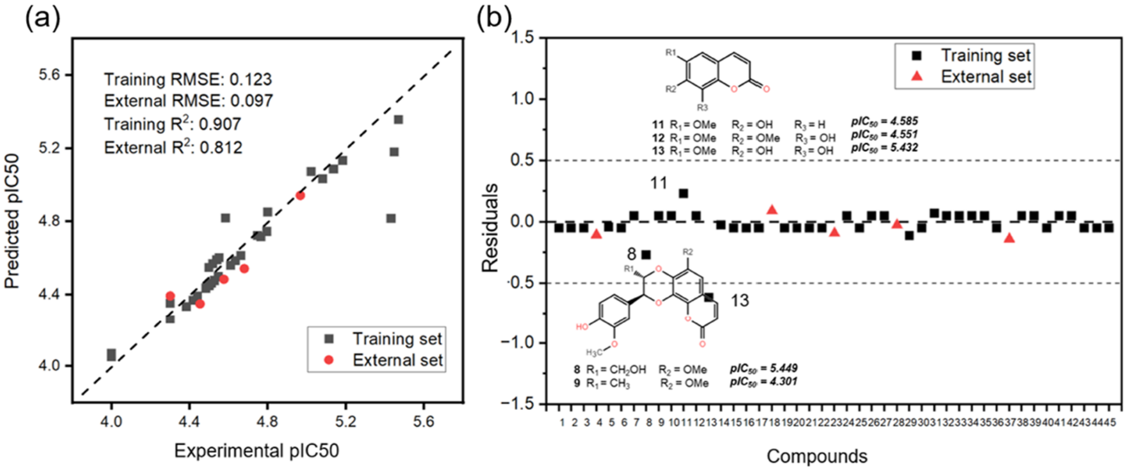


**Figure 7.** (a) Scatter diagram of the relationship between the predicted ($SVR_{5b}$ model) and experimental values of the data and (b) residuals of predicted and experimental anti-inflammatory activity values of 45 durian extracts.

negative influence. The electronegativity of the oxygen atom in the methoxy group results in (1) an electron-withdrawing inductive effect when connected to a $sp^3$-hybridized carbon and (2) conjugation with its two lone pairs of electrons when linked to a $sp^2$-hybridized carbon. Considering these varied electronic effects and the Shapley analytical results together, one can speculate that the distinct impacts on anti-inflammatory activity arise from diverse positional electronic influences.

The dipole $z$ feature quantifies a molecule's polarity along the $z$-axis and results from the magnitude and distance of charge centers. A molecule with a zero dipole moment is nonpolar, while a molecule with a nonzero one is polar, with higher values indicating a higher degree of polarity. Figure 5(e) reveals that dipole $z$ values from −1 to 1 correspond to significant Shapley contributions, implying that substituents in the z-direction affect the anti-inflammatory activity. Groups leading to charge center imbalances amplify the polarity and influence the anti-inflammatory properties of molecules.

**3.5. Activity Cliff.** Analyzing the data of the top-performing model, $SVR_{5b}$, as shown in Table 4, reveals the presence of three samples with notable discrepancies in the predictions for the training set. Figure 7 displays the differences between the SVR-predicted and actual anti-inflammatory values for the 45 extracts, some of which surpass 0.5, revealing significant disparities. From a closer examination of the molecular structures, one can attribute this phenomenon to the existence of a molecular activity cliff within the data of the training set.

Molecular activity cliffs emerge when slight structural changes produce large differences in biological activity among closely related molecules. Compounds 8 and 9, which differ only in the $R_1$ group, exhibit significantly different anti-inflammatory actions, with $pIC_{50}$ values of 5.449 and 4.301, respectively. Similarly, compounds 11, 12, and 13 display this trend, with compound 13 (featuring −OH substituents at both the $R_2$ and $R_3$ positions) exhibiting the highest activity ($pIC_{50}$ = 5.432). This phenomenon highlights the challenge of distinguishing between near-identical compounds due to a lack of relevant information about molecular characteristics in the QSAR model. Thus, forecasting the activity cliff is difficult, even when using the best SVR model in this particular class.

## 4. CONCLUSIONS

The dataset in this work covered the main natural bioactive chemicals found in the durian shell extraction, including phenols, glycosides, and pentacyclic triterpenoids, and there was a normal distribution of activity ranges and amounts of compounds. Therefore, the QSAR analysis with four machine learning algorithms, RF, GBR, SVR, and ANN, was conducted on natural anti-inflammatory compounds extracted from the durian shell. Among the models, the SVR one based on the five molecular features shadow ratio, dipole *z*, methyl, ellipsoidal volume, and methoxy yielded the optimal performance. According to the model analytical results, we speculate that molecules with a prolate ellipsoidal shape exhibit better cell membrane penetration, which can increase the cellular utilization of such natural anti-inflammatory molecules and, thereby, lead to improved anti-inflammatory effects. In addition, the substituted methoxy group on the aromatic ring along the shortest dimension of the molecules (*z*-direction) is also an important factor in enhancing the anti-inflammatory activity. Finally, the analysis of quantitative structure−activity relationships in this article can provide new insights for in-depth future research on the anti-inflammatory activity and molecular design of new anti-inflammatory drugs.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c07386.

> Details of 96 descriptors used for construction of QSAR models, definitions of 96 descriptors, dataset of 45 studied compounds, and Python code examples of the QSAR modeling process (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Authors

**Thanyada Rungrotmongkol** − *Center of Excellence in Structural and Computational Biology, Department of Biochemistry, Chulalongkorn University, Bangkok 10330, Thailand; Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Bangkok 10330, Thailand;* orcid.org/0000-0002-7402-3235; Email: t.rungrotmongkol@gmail.com

**Phornphimon Maitarad** − *Research Center of Nano Science and Technology, College of Sciences, Shanghai University, Shanghai 200444, P. R. China;* orcid.org/0000-0003-0035-0070; Email: pmaitarad@shu.edu.cn

### Authors

**Amphawan Wiriyarattanakul** − *Program in Chemistry, Faculty of Science and Technology, Uttaradit Rajabhat University, Uttaradit 53000, Thailand*

**Wanting Xie** − *Research Center of Nano Science and Technology, College of Sciences, Shanghai University, Shanghai 200444, P. R. China*

**Borwornlak Toopradab** − *Center of Excellence in Structural and Computational Biology, Department of Biochemistry, Chulalongkorn University, Bangkok 10330, Thailand; Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Bangkok 10330, Thailand*

**Sopon Wiriyarattanakul** − *Program in Computer Science, Faculty of Science and Technology, Uttaradit Rajabhat University, Uttaradit 53000, Thailand*

**Liyi Shi** − *Research Center of Nano Science and Technology, College of Sciences, Shanghai University, Shanghai 200444, P. R. China; Emerging Industries Institute, Shanghai University, Jiaxing, Zhejiang 314006, P. R. China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c07386

### Author Contributions

$^{\nabla}$A.W. and W.X. contributed equally.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

A.W. would like to thank the National e-Science Infrastructure Consortium (URL: http://www.e-science.in.th) for computing resources and Thailand Science Research and Innovation. P.M. and L.S. would like to thank the Shanghai Municipal Science and Technology Commission of the Professional and Technical Service Platform for the Designing and Manufacturing of Advanced Composite Materials (16DZ2292100) and the Emerging Industries Research Institute, Shanghai University (Jiaxing, Zhejiang; 20H01612). T.R. acknowledges financial support from the Thailand Research Fund (grant number RSA6280085). This research was partially supported by Chulalongkorn University (CE66_036_2300_008) for T.R.

## ■ REFERENCES

(1) Zhan, Y.-f.; Hou, X.-t.; Fan, L.-l.; Du, Z.-c.; Ch'ng, S. E.; Ng, S. M.; Thepkaysone, K.; Hao, E.-w.; Deng, J.-g. Chemical constituents and pharmacological effects of durian shells in ASEAN countries: A review. *Chinese Herb. Med.* **2021**, *13* (4), 461−471.

(2) Arsa, S.; Wipatanawin, A.; Suwapanich, R.; Makkerdchoo, O.; Chatsuwan, N.; Kaewthong, P.; Pinsirodom, P.; Taprap, R.; Haruenkit, R.; Poovarodom, S.; et al. Properties of different varieties of durian. *Appl. Sci.* **2021**, *11* (12), 5653.

(3) Haruenkit, R.; Poovarodom, S.; Vearasilp, S.; Namiesnik, J.; Sliwka-Kaszynska, M.; Park, Y.-S.; Heo, B.-G.; Cho, J.-Y.; Jang, H. G.; Gorinstein, S. Comparison of bioactive compounds, antioxidant and antiproliferative activities of Mon Thong durian during ripening. *Food Chem.* **2010**, *118* (3), 540−547.

(4) Primarianti, A. U.; Sujono, T. A. Antidiabetic activity of durian (Durio zibethinus Murr.) and rambutan (Nephelium lappaceum L.) fruit peels in alloxan diabetic rats. *Procedia Food Sci.* **2015**, *3*, 255−261.

(5) Muhtadi, M.; Haryoto, H.; Sujono, T. A.; Suhendi, A. Antidiabetic and antihypercholesterolemia activities of rambutan (Nephelium lappaceum L.) and durian (Durio zibethinus Murr.) fruit peel extracts. *J. Appl. Pharm. Sci.* **2016**, *6* (4), 190−194.

(6) Jiang, Q.; Charoensiddhi, S.; Xue, X.; Sun, B.; Liu, Y.; El-Seedi, H. R.; Wang, K. A review on the gastrointestinal protective effects of tropical fruit polyphenols. *Crit. Rev. Food Sci. Nutr.* **2022**, 7197−7223.

(7) Toledo, F.; Arancibia-Avila, P.; Park, Y.-S.; Jung, S.-T.; Kang, S.-G.; Gu Heo, B.; Drzewiecki, J.; Zachwieja, Z.; Zagrodzki, P.; Pasko, P.; Gorinstein, S. Screening of the antioxidant and nutritional properties, phenolic contents and proteins of five durian cultivars. *Int. J. Food Sci. Nutr.* **2008**, *59* (5), 415−427.

(8) Charoenphun, N.; Klangbud, W. K. Antioxidant and anti-inflammatory activities of durian (Durio zibethinus Murr.) pulp, seed and peel flour. *PeerJ* **2022**, *10*, No. e12933.

(9) Feng, J.; Yi, X.; Huang, W.; Wang, Y.; He, X. Novel triterpenoids and glycosides from durian exert pronounced anti-inflammatory activities. *Food Chem.* **2018**, *241*, 215−221.

(10) Chingsuwanrote, P.; Muangnoi, C.; Parengam, K.; Tuntipopipat, S. Antioxidant and anti-inflammatory activities of durian and rambutan pulp extract. *Int. Food Res. J.* **2016**, *23* (3), 939−947.

(11) Rakha, A.; Umar, N.; Rabail, R.; Butt, M. S.; Kieliszek, M.; Hassoun, A.; Aadil, R. M. Anti-inflammatory and anti-allergic potential of dietary flavonoids: A review. *Biomed. Pharmacother.* **2022**, *156*, No. 113945.

(12) Khanna, S.; Bishnoi, M.; Kondepudi, K. K.; Shukla, G. Isolation, characterization and anti-inflammatory mechanism of probiotics in lipopolysaccharide-stimulated RAW 264.7 macrophages. *World J. Microbiol. Biotechnol.* **2020**, *36*, 74.

(13) Ahmadi, S.; Ghanbari, H.; Lotfi, S.; Azimi, N. Predictive QSAR modeling for the antioxidant activity of natural compounds derivatives based on Monte Carlo method. *Mol. Divers.* **2021**, *25*, 87−97.

(14) Spiegel, M.; Kapusta, K.; Kołodziejczyk, W.; Saloni, J.; Żbikowska, B.; Hill, G. A.; Sroka, Z. Antioxidant activity of selected phenolic acids−ferric reducing antioxidant power assay and QSAR analysis of the structural features. *Molecules* **2020**, *25* (13), 3088.

(15) Das, S.; Majumder, T.; Sarkar, A.; Mukherjee, P.; Basu, S. Flavonoids as BACE1 inhibitors: QSAR modelling, screening and in vitro evaluation. *Int. J. Biol. Macromol.* **2020**, *165*, 1323−1330.

(16) Sharma, M.; Ahuja, D. QSAR Studies of Flavonoids Derivatives for Antioxidant and Antimicrobial Activity. *J. Drug Delivery Ther.* **2019**, *9* (4), 765−773.

(17) Tomar, V. A Review on Procedure of QSAR assessment in Organic Compounds as a Measure of Antioxidant Potentiality. *Int. J. Glob. Acad. Sci. Res.* **2022**, *1* (1), 8−18.

(18) Žuvela, P.; David, J.; Yang, X.; Huang, D.; Wong, M. W. Non-linear quantitative structure−activity relationships modelling, mechanistic study and in-silico design of flavonoids as potent antioxidants. *Int. J. Mol. Sci.* **2019**, *20* (9), 2328.

(19) Ray, S.; Sengupta, C.; Roy, K. QSAR modeling of antiradical and antioxidant activities of flavonoids using electrotopological state (E-State) atom parameters. *Open Chem.* **2007**, *5* (4), 1094−1113.

(20) Ray, S.; Sengupta, C.; Roy, K. QSAR modeling for lipid peroxidation inhibition potential of flavonoids using topological and structural parameters. *Open Chem.* **2008**, *6* (2), 267−276.

(21) Shi, Y. Support vector regression-based QSAR models for prediction of antioxidant activity of phenolic compounds. *Sci. Rep.* **2021**, *11* (1), No. 8806.

(22) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-J.; Zhang, H. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, *55*, 12741−12754.

(23) Jia, X.; Wang, T.; Zhu, H. Advancing Computational Toxicology by Interpretable Machine Learning. *Environ. Sci. Technol.* **2023**, *57*, 17690−17706, DOI: 10.1021/acs.est.3c00653.

(24) Zorn, K. M.; Foil, D. H.; Lane, T. R.; Hillwalker, W.; Feifarek, D. J.; Jones, F.; Klaren, W. D.; Brinkman, A. M.; Ekins, S. Comparison of Machine Learning Models for the Androgen Receptor. *Environ. Sci. Technol.* **2020**, *54* (21), 13690−13700.

(25) Li, Z.; Nie, K.; Wang, Z.; Luo, D. Quantitative structure activity relationship models for the antioxidant activity of polysaccharides. *PLoS One* **2016**, *11* (9), No. e0163536.

(26) Feng, J.; Wang, Y.; Yi, X.; Yang, W.; He, X. Phenolics from durian exert pronounced NO inhibitory and antioxidant activities. *J. Agric. Food Chem.* **2016**, *64* (21), 4273−4279.

(27) Frisch, M.; Trucks, G.; Schlegel, H. B.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H. Gaussian 16; Gaussian, Inc.: Wallingford, CT, 2016.

(28) Daoud, J. I. Multicollinearity and regression analysis. *J. Phys.: Conf. Ser.* **2017**, *949*, No. 012009.

(29) Seabold, S.; Perktold, J. *Statsmodels: Econometric and statistical modeling with python*, Proc. of the 9th Python in Science Conf. (SciPy 2010), 2010; p 25080.

(30) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(31) *Materials Studio Modeling; release 8.0*, Accelrys Software Inc.: San Diego, CA, USA, 2014.