



Prediction of Photooxidation Reaction of Aromatic Micropollutants Using the Genetic Algorithm - Support Vector Machine

Eslam Pourbasheer¹✉ | Maryam Nouri majd¹ | Reza Mahmoudzadeh Laki¹ |
Suraj N. Mali² | Mohammad Reza Ganjali³

1. Department of Chemistry, Faculty of Science, University of Mohaghegh Ardabili, Ardabil, Iran.

2. School of Pharmacy, D.Y. Patil University, Navi Mumbai 400706, India.

3. Center of Excellence in Electrochemistry, School of Chemistry, College of Science, University of Tehran, Tehran, Iran.

Article Info

Article type:
Research Article

Article history:
Received: 13 Jun 2026
Revised: 31 Jul 2026
Accepted: 02 Oct 2026

Keywords:
QSAR
Support Vector
Machine
Aromatic
Micropollutants
Photooxidation

ABSTRACT

The quantitative structure-activity relationship (QSAR) models were used to predict the photooxidation reaction of aromatic micro-pollutants (AMPs) using the multiple linear regressions (MLR) and support vector machine (SVM). The dataset includes of 30 compounds, divided into the training and test subsets by hierarchical clustering method. A genetic algorithm (GA) was employed to select the most relevant molecular descriptors, and multiple linear regression (MLR) and support vector machine (SVM) methods were subsequently used to establish linear and nonlinear models, respectively. The GA-MLR model based on three descriptors provided $R^2_{\text{train}}=0.837$, $\text{RMSE}_{\text{train}}=0.184$, $R^2_{\text{test}}=0.781$, and $\text{RMSE}_{\text{test}}=0.231$. The nonlinear GA-SVM model showed improved predictive performance, with $R^2_{\text{train}}=0.955$, $\text{RMSE}_{\text{train}}=0.098$, $R^2_{\text{test}}=0.862$, and $\text{RMSE}_{\text{test}}=0.171$. Cross-validation, Y-randomization, external hold-out evaluation, and applicability-domain analysis were used to assess model reliability. The selected descriptors, ESP-DPSA-1, SPH, and R3m+, indicate that electrostatic surface imbalance, molecular compactness, and mass-weighted three-dimensional atomic distribution contribute to the photooxidation behavior of AMPs. Within its defined chemical domain, the developed model may be used as a screening tool for prioritizing structurally related aromatic micropollutants for experimental photooxidation studies.

Cite this article: Pourbasheer, E., Nouri Majd, M., Mahmoudzadeh Laki, R., Mali, S.N., & Ganjali, M.R. (2026). Prediction of Photooxidation Reaction of Aromatic Micropollutants Using the Genetic Algorithm - Support Vector Machine. *Pollution*, 12(4), 1199-1212.

<https://doi.org/10.22059/poll.2026.416376.3382>



© The Author(s).

Publisher: The University of Tehran Press.

DOI: <https://doi.org/10.22059/poll.2026.416376.3382>

INTRODUCTION

Aromatic micropollutants (AMPs) encompass a wide array of organic substances featuring one or more benzene rings, typically modified with substituents like hydroxyl (-OH), amino (-NH₂), nitro (-NO₂), carboxyl (-COOH), or halogen atoms (-Cl, -Br). These contaminants, which include phenols (e.g., phenol, bisphenol A, chlorophenols), nitroaromatics (e.g., nitrobenzene, nitrophenols), halogenated benzenes (e.g., chlorobenzene), polycyclic aromatic hydrocarbons (PAHs) (e.g., naphthalene, benzo[a]pyrene), drug residues (e.g., carbamazepine, diclofenac), and agrochemicals (e.g., atrazine, diuron), are commonly found at low concentrations in surface waters, aquifers, potable water supplies, and effluents globally (Grandclément et al., 2017; Luo et al., 2014; Miao et al., 2023). AMPs stem from multiple origins, such as discharges from pharmaceutical, chemical, and textile industries, sewage treatment facilities, farming activities, and airborne fallout from burning fossil fuels (Luo et al., 2014; Ma et al., 2018). Their structural traits, including delocalized π -electrons and resonance effects, grant them

*Corresponding Author Email: e.pourbasheer@uma.ac.ir

exceptional durability against breakdown in nature^{¶¶}

The implications of AMPs for the environment and public health are profound. Their longevity, tendency to accumulate in organisms, and ease of transport through water networks create ongoing threats to biodiversity and well-being. Links have been established between AMP exposure and issues like hormonal interference, fertility problems, DNA damage, cancer risk, nervous system disorders, and weakened immunity (Ma et al., 2018; Włodarczyk-Makuła et al., 2023; Younis & Moustafa, 2017). For example, certain PAHs such as benzo[a]pyrene are deemed likely carcinogens in humans, whereas compounds like bisphenol A, can disrupt endocrine functions, causing growth issues in marine species (Wilkinson et al., 2022; Włodarczyk-Makuła et al., 2023). In aquatic settings, AMPs generate additional contaminants via reactions, such as bicyclic alkoxy radicals (BARs) produced from interactions with atmospheric ozone (O₃) and hydroxyl radicals (•OH), which intensify acid rain formation (Miklos et al., 2018).

Standard purification methods in wastewater plants frequently fall short in eliminating these stubborn pollutants, prompting the use of sophisticated techniques like catalytic light-driven processes (e.g., using TiO₂, ZnO, or graphene materials), ozone treatment, and iron-based oxidations (Ge et al., 2018; Phang et al., 2021). Nevertheless, light-induced breakdown of AMPs varies greatly with structure: groups that donate electrons (e.g., -OH, -NH₂) promote complex formation for charge transfer and boost reaction rates in visible light, while electron-accepting moieties (e.g., -NO₂, -COOH) reinforce stability and slow down the process (Awfa et al., 2018; Qutob et al., 2024). Such differences highlight the value of methods that forecast breakdown rates from chemical makeup^{¶¶}

In order to understanding the AMPs catalytic efficiency and behavior of their photochemical reaction, the study on relationship of structures and photooxidation of AMPs is essential. The reaction rate constant is an important physical quantity and index for determining the mechanism of reaction and for assessing the pollutant removal efficiency. Recently, the reaction kinetics parameters of AMPs in Photo graphene oxide photocatalysis system have been reported by Wang et al. (Wang et al., 2023).

Quantitative structure-activity relationship (QSAR) frameworks offer a strong analytical tool to links structures with their chemical, biological, properties, or ecological traits, allowing estimates for novel molecules (Sun et al., 2024; Wang et al., 2023; Yan et al., 2021). For constructing the QSAR models, different machine learning methods can be use. Machine learning methods span simple regressions, like multiple linear regressions (MLR) to advance nonlinear ones, such as support vector machines (SVM) (Alimohammadi et al., 2019; Pourbasheer et al., 2014) Combined variable selection strategies, like GA-linked MLR or SVM, integrate selection and fitting for better precision (Wold et al., 2001).

Although Wang et al. developed a linear QSAR model for the graphene oxide-induced photooxidation of aromatic micropollutants, the possibility of nonlinear relationships between molecular structure and the observed photooxidation rate remained insufficiently investigated (Wang et al., 2023). In addition, further evaluation of descriptor selection, model robustness, and the chemical domain within which predictions can be considered reliable is required. The present study does not introduce a new experimental dataset; rather, it provides a nonlinear QSAR extension and reanalysis of the previously reported kinetic data. A genetic algorithm was used to identify a compact and chemically interpretable descriptor subset, after which GA-MLR and GA-SVM models were developed using the same descriptors. This design enabled a direct comparison between linear and nonlinear structure-activity relationship models. The models were assessed using cross-validation, Y-randomization, an external test set, and applicability-domain analysis. Particular attention was also given to interpreting the selected descriptors in relation to molecular electrostatic properties, three-dimensional shape, atomic mass distribution, and their possible roles in graphene oxide-mediated photooxidation. Therefore, the principal contribution of this study is the development of a validated nonlinear

screening model and the extraction of additional structure activity information from a compact set of molecular descriptors.

MATERIAL AND METHODS

Data set

The molecular structures of 30 aromatic micropollutants and their experimentally measured pseudo-first-order photooxidation rate constants (k_{obs}) were obtained from the study of Wang et al. (Wang et al., 2023). The response variable used for QSAR modeling was the logarithm of the observed rate constant ($\lg k_{\text{obs}}$). The compounds were divided into a training set of 24 compounds and an external test set of six compounds, corresponding to an 80:20 ratio. Hierarchical clustering was used to support dataset division while considering both structural diversity and the distribution of experimental ($\lg k_{\text{obs}}$) values. The training set was used for descriptor selection, model development, cross-validation, and SVM parameters optimization, whereas the test set was reserved for the final evaluation of predictive performance. The chemical structures and their activities are presented in Table 1.

Descriptors calculation and selection

The two-dimensional molecular structures were drawn using HyperChem software and were subsequently subjected to molecular geometry optimization. Initial optimization was performed using the (MM+) molecular mechanics force field (Glaser, 2009), followed by further optimization using AMPAC software (Laxmi & Priyadarshy, 2002); (Grippio & Lucidi, 1997). Molecular descriptors belonging to structural, topological, geometrical, electrostatic, charged partial surface area, quantum-chemical, thermodynamic, and molecular orbital categories were calculated using CODESSA software (Katritzky et al., 2002).

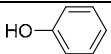
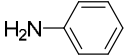
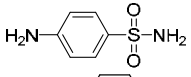
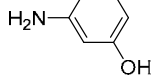
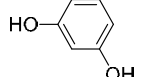
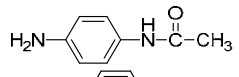
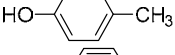
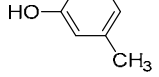
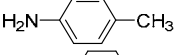
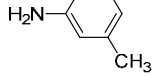
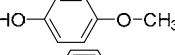
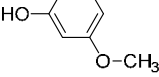
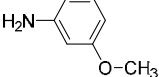
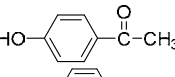
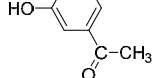
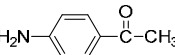
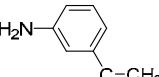
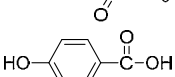
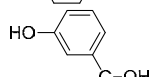
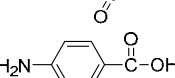
Descriptor preprocessing was performed before model construction. Constant and near-constant descriptors were initially removed because they contained insufficient information for differentiating the compounds. Highly intercorrelated descriptors were subsequently examined, and one descriptor from each strongly correlated pair with ($r > 0.90$) was removed to reduce redundancy and model instability. After preprocessing, 103 descriptors remained available for genetic algorithm-based variable selection. The genetic algorithm was then applied to the training data to identify a parsimonious descriptor subset with high predictive relevance and limited intercorrelation.

Model validation and applicability-domain assessment

The reliability of the developed QSAR models was assessed using complementary internal and external validation procedures. Internal validation of the linear model was performed using leave-one-out and leave-group-out cross-validation. The coefficients (Q^2_{LOO}) and (Q^2_{LGO}) were used to determine whether the model retained its predictive performance after the systematic removal of training compounds. Y-randomization was additionally performed by randomly permuting the response values and rebuilding the models. The significantly lower (R^2) and (Q^2) values for the random models were considered as evidence that the initial structure activity relationship was not obtained by chance. The predictive performance of the models was also evaluated using the coefficient of determination (R^2), root mean square error (RMSE), and F value for both the training and test sets.

The applicability domain of the descriptor-based model was assessed using the Williams plot (Kar et al., 2018), in which standardized residuals are plotted against leverage values. The leverage of each compound was calculated from the molecular descriptor matrix and the warning leverage was determined according to the following equation:
$$h^* = 3(p+1)/n$$

Table 1. Chemical structure and corresponding experimental and predicted $\lg k_{\text{obs}}$ from GA-MLR and GA-SVM methods.[†]

No.	Chemicals	Structures	Exp. ($\lg K_{\text{obs}}$)	Predicted (GA-MLR)	Predicted GA-SVM
1	Phenol		-1.45	-1.50	-1.46
2	Aniline		-0.95	-1.01	-0.96
3	Sulfanilamide		-1.58	-1.64	-1.59
4	3-Aminophenol		-0.73	-1.08	-0.86
5	Resorcinol		-1.37	-1.63	-1.57
6	4-Aminoacetanilide		-0.58	-0.60	-0.56
7	4-Methylphenol		-1.38	-1.02	-1.18
8	3-Methylphenol		-1.39	-1.10	-1.37
9	4-Toluidine		-0.4	-0.55	-0.41
10	3-Toluidine		-0.62	-0.64	-0.63
11	4-Methoxyphenol		-0.76	-1.14	-1.11
12	3-Methoxyphenol		-1.39	-1.24	-1.37
13	3-Anisidine		-0.85	-0.80	-0.83
14	4-Hydroxyacetophenone		-1.71	-1.65	-1.72
15	3-Hydroxyacetophenone		-1.71	-1.56	-1.69
16	4-Aminoacetophenone		-1.34	-1.27	-1.35
17	3-Aminoacetophenone		-1.14	-1.16	-1.16
18	4-Hydroxybenzoic acid		-1.82	-1.95	-1.83
19	3-Hydroxybenzoic acid		-1.94	-1.86	-1.92
20	4-Aminobenzoic acid		-1.88	-1.74	-1.85

Conitued Table 1. Chemical structure and corresponding experimental and predicted lg k_{obs} from GA-MLR and GA-SVM methods.[†]

No.	Chemicals	Structures	Exp. (lg K _{obs})	Predicted (GA-MLR)	Predicted GA-SVM
21	3-Aminobenzoic acid		-1.69	-1.49	-1.64
22	4-Methylsulfonylphenol		-1.96	-2.10	-1.94
23	4-Methylsulphonylaniline		-1.59	-1.67	-1.60
24	Methyl 4-hydroxybenzoate		-1.7	-1.42	-1.63
25 ^a	Methyl 3-hydroxybenzoate		-1.5	-1.43	-1.54
26 ^a	Methyl 4-aminobenzoate		-1.29	-1.24	-1.31
27 ^a	Methyl 3-aminobenzoate		-1.14	-1.09	-1.10
28 ^a	4-Chloroaniline		-0.92	-0.59	-0.95
29 ^a	4-Aminophenetole		-0.09	-0.36	-0.43
30 ^a	3-Ethoxyaniline		-0.78	-0.41	-0.54

^a Used as test set

where (p) is the number of descriptors and (n) is the number of training compounds. Compounds with leverage values greater than (h*) were considered influential structural observations, while compounds with standardized residuals outside the $\pm 3s$ range were considered response outliers. Model predictions were considered reliable only for compounds that were within both the structural and response boundaries of the applicability domain.

RESULTS AND DISCUSSION

Genetic algorithm multiple linear regression method

Genetic algorithm has been used to select the most appropriate descriptors from the remaining (Alhijawi & Awajan, 2024). The details of genetic algorithms have been reported in the literatures and in our previous works (Nouri Majd et al., 2025).

Three descriptors PPSA1-PNSA1, SPH and R3m+ which are play a main role in the photooxidation activity, have been selected using the genetic algorithm technique. The variance inflation factor (VIF)(Pourbasheer et al., 2014) calculation was carried out for the three selected descriptors to check the multicollinearity for each descriptor as follows:

$$\text{VIF} = 1/1-r^2 \quad (1)$$

In this formula, the r value (multiple of correlation coefficient) indicates the strength of the

Table 2. Linear model based on three parameters selected using the GA-MLR method.

Descriptor	Chemical meaning	MF ^a	VIF ^b
Constant	Intercept	-	-
ESP-DPSA-1	Difference in Partially Charged Surface Areas: PPSA1-PNSA1	0.242	1.055
SPH	spherosity	0.921	1.128
R3m+	R maximal autocorrelation of lag 3 / weighted by atomic masses	-0.163	1.085

^a Mean effect^b Variation inflation factors

linear relationship between the independent and dependent variables in a regression model. If the VIF value is equal to 1, the correlation between the variables is not significant and if it is between 1 and 5, the model is statistically acceptable and if it is more than 10, the model is not statistically acceptable and indicates severe multicollinearity, it also needs to revise the model.

The calculated VIF values for the three selected descriptors are recorded in Table 2. As the results show, all variables have VIF values below 1.2. This not only confirms the absence of multicollinearity, but also strengthens the statistical stability and validity of the final model.

The QSAR model was developed on the basis of data from the training set and then its predictive capability was assessed using the test set. This approach ensures the feasibility of the model's performance in predicting unknown compounds. A predictive QSAR model with three descriptors was created using GA-MLR with the following equation:

$$\lg k_{\text{obs}}^{\text{R3m+}} = -4.017(\pm 1.039) + 0.008(\pm 0.001)(\text{ESP-DPSA-1}) + 2.627(\pm 0.810)\text{SPH} - 9.642(\pm 5.706) \quad (2)$$

$$N_{\text{train}} = 24, R^2_{\text{train}} = 0.837, \text{RMSE}_{\text{train}} = 0.184, F_{\text{train}} = 34.44, Q^2_{\text{LOO}} = 0.780, Q^2_{\text{LGO}} = 0.756, Q^2_{\text{adj}} = 0.813, N_{\text{test}} = 6, R^2_{\text{test}} = 0.781, \text{RMSE}_{\text{test}} = 0.231, F_{\text{test}} = 2.259$$

In this analysis, some statistical parameters were used to assess the quality of the model; N (number of compounds in training and test sets), Q^2_{LOO} (Leave one out of the cross-validation coefficient), Q^2_{LGO} (Leave group out of the cross-validation coefficient), R^2 (coefficient of determination), R^2_{adj} (adjusted R^2) and F (Fisher statistic).

The developed GA-MLR model with these parameters showed a good performance; the R^2 values in both training and test sets indicate a high agreement between predicted and experimental data. The Q^2_{LOO} and Q^2_{LGO} values of 0.780 and 0.756 indicate high reliability of the model.

The high F values and low error ($\text{RMSE}_{\text{train}} = 0.184$ and $\text{RMSE}_{\text{test}} = 0.231$) confirm the predictive ability of the model. The predicted photocatalytic activity ($\lg k_{\text{obs}}^{\text{test}}$) for all molecules is presented in Table 1 and the comparison graph of the experimental and predicted values is shown in Figure 1.

To assess the robustness of the model, a Y-randomization test was performed. In this method, the $\lg k_{\text{obs}}$ values were randomly permuted and new models were built based on the random data. The results of this test in Table 3 show that the R^2 and Q^2_{LOO} values of the random models are below 0.35. This confirms that the performance of the original model is not due to chance and has a solid statistical basis.

A Williams plot (cross-validation standardized residuals versus leverage plot) was used to examine outliers in the dataset. This plot is presented in Figure 2. Based on the results obtained from this plot, all the compounds are within the range of the warning level value ($h^* = 0.500$), indicating that there are no unusual influential points in the model. The standardized residual values for all compounds (training and test sets) are less than 3 standard deviation units (3δ). These results confirm that there is no outlier in the developed model and the model predictions are statistically reliable and valid.

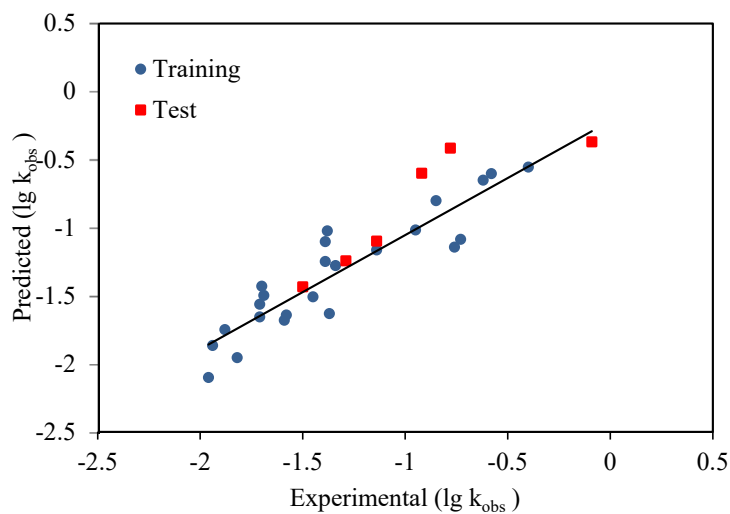


Fig. 1. Experimental versus predicted $\lg K_{\text{obs}}$ values obtained using the GA-MLR model for the training and test sets.

Table 3. R^2_{train} and Q^2_{LOO} values after multiple Y- randomization.[†]

Iteration	R^2_{train}	Q^2_{LOO}
1	0.149	0.015
2	0.186	0.002
3	0.138	0.000
4	0.106	0.013
5	0.008	0.064
6	0.146	0.004
7	0.190	0.011
8	0.060	0.168
9	0.030	0.355
10	0.029	0.042

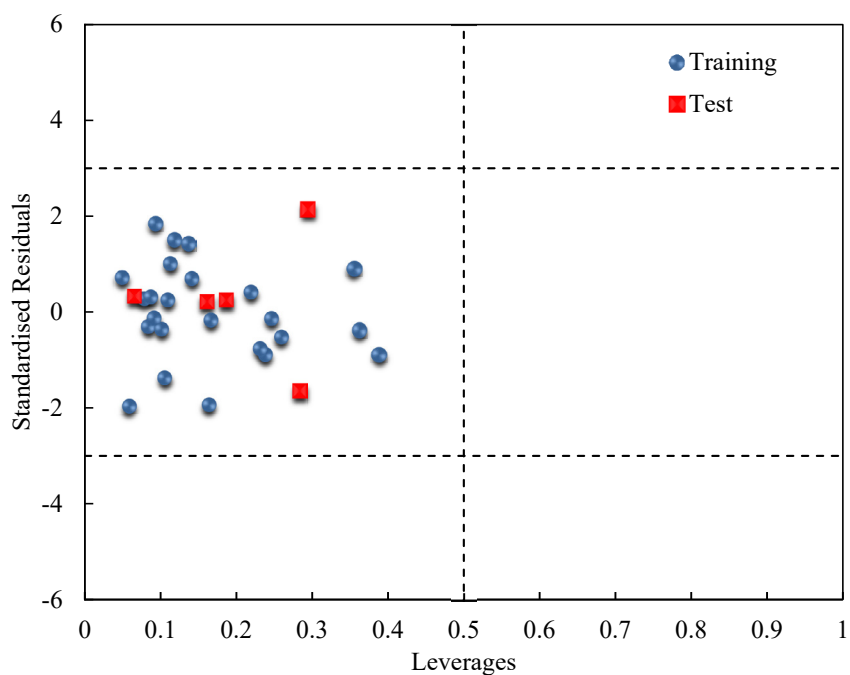


Fig. 2. William's plot of GA-MLR model for training and test sets.

Genetic algorithm- support vector machine method

After developing the linear GA-MLR model, the support vector machine (SVM) method (Auria & Moro, 2008; Huang et al., 2018; Suthaharan, 2016) was used to construct a non-linear model based on the same selected descriptors. SVM regression is dependent on several parameters such as; the type of kernel function, the capacity parameter (C) that controls the balance between overfitting and underfitting, ϵ (epsilon) in the ϵ -insensitive loss function that determines the amount of error allowed, and specific kernel parameters. These parameters were optimized to develop the SVM model with the highest accuracy and the lowest error according to the leave one out cross validation.

The type of kernel function in SVM method determines how to map data to higher dimensional feature space. Therefore, selecting and declaring the type of kernel function is of great importance. In this study, the radial basis function (RBF) was selected as the kernel function due to its high ability in modeling nonlinear relationships and flexibility in setting parameters (Pourbasheer et al., 2019; Pourbasheer et al., 2011). The formula of RBF function is defined as follows:

$$\text{Exp}(-\gamma * |u-v|^2) \quad (3)$$

In this formula, u and v are independent variables, and γ (gamma) is one of the key parameters of the kernel function of the RBF (Pourbasheer et al., 2011; Suthaharan, 2016). This parameter adjusts the extent of the influence of the training data on the model and has a direct influence on the performance of the SVM and time taken to train the model. To achieve the best performance, the parameter γ needs to be optimized. A network search on the training set, was performed using leave-one-out cross-validation, and the range of γ was considered from 0.1 to 5 with an incremental step of 0.1. At each iteration, the root mean square error (RMSE) obtained from the cross-validation was calculated. Figure 3 shows the relationship between the values of γ and RMSE. Based on this analysis, the optimal value of γ was determined to be 2, which produces the lowest RMSE in the validation process.

The ϵ -insensitive parameter in SVM regression prevents the model from being too strict on the training data by creating a margin of error. This parameter allows for better adaptation of the model to noisy data by allowing some freedom in solving the dual Problem (Pourbasheer et al., 2019). Optimal value of ϵ depends on noise level in data; the more noise there is, the larger ϵ is needed to absorb unnecessary variations. The range of ϵ from 0.001 to 0.14 was examined

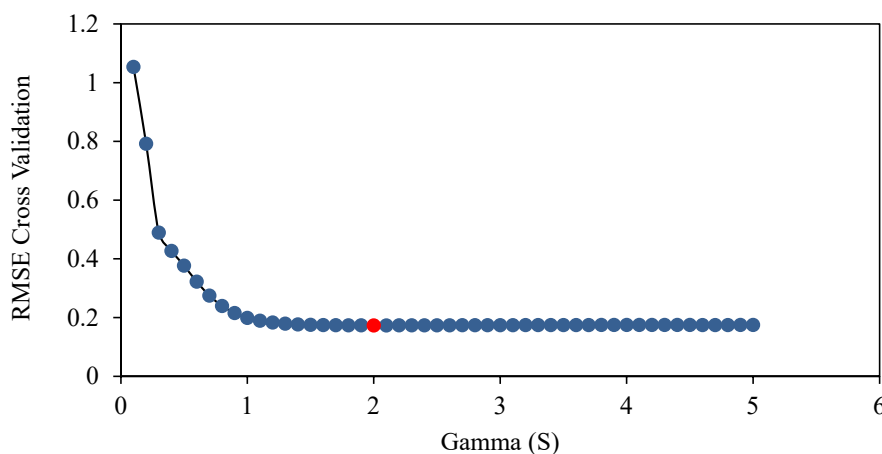


Fig. 3. Gamma(γ) versus RMSE for training set. Optimal Gamma value = 2

in increments of 0.001, and the root mean square error (RMSE) of the cross-validation was calculated for each value of ϵ . Figure 4 shows the relationship between $\epsilon \times 10^3$ values and RMSE. Based on this analysis, the optimal value of ϵ was determined to be 0.014, which produces the lowest RMSE.

The capacity parameter (C) controls the trade-off between model complexity and the magnitude of errors tolerated during SVM training (Pourbasheer et al., 2019; Pourbasheer & Laki, 2025). An excessively small (C) value may lead to underfitting, whereas an excessively large value may produce an overly complex model with reduced generalization ability. To determine the optimal value, (C) was varied from 1 to 5 using an increment of 1, and the cross-validated RMSE was calculated at each step. As shown in Figure 5, the minimum RMSE was obtained at (C=6). Accordingly, the final optimized parameters of the RBF-SVM model were ($\gamma = 2$), ($\epsilon = 0.014$), and (C=6).

After optimization of the SVM hyperparameters, the final GA-SVM model was developed using the three descriptors selected by the genetic algorithm. The predicted $\lg K_{\text{obs}}$ values are presented in Table 1, and the relationship between the experimental and predicted values is

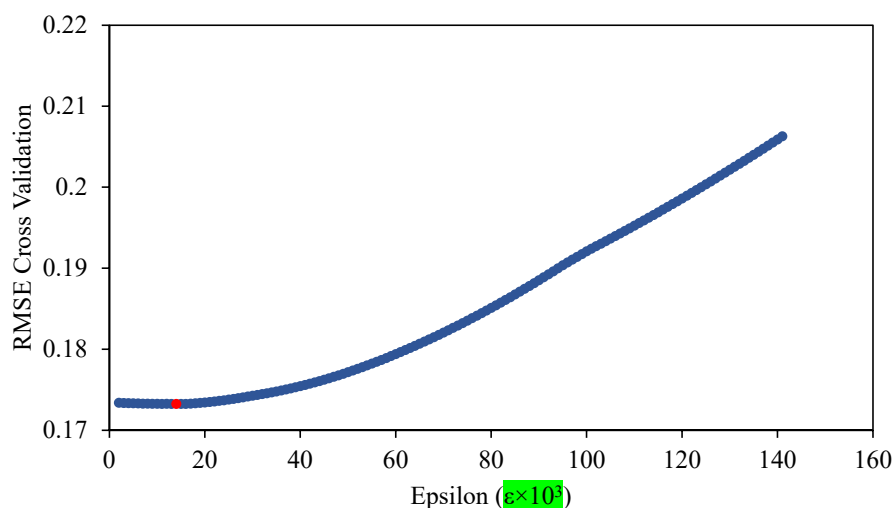


Fig. 4. Epsilon (ϵ) versus RMSE for training set. Optimal Epsilon (ϵ) value = 0.014

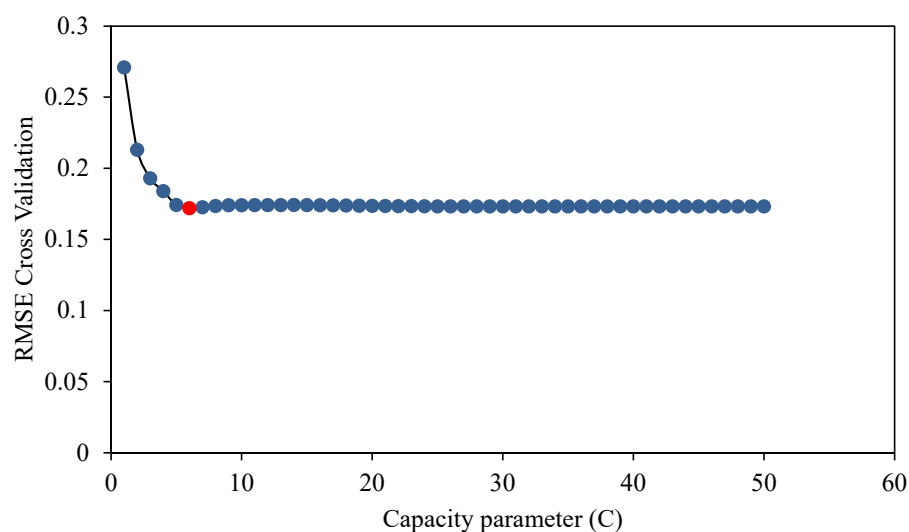


Fig. 5. Capacity parameter (C) versus RMSE for training set. Optimal C value = 6

illustrated in Figure 6. The GA-SVM model provided ($R^2_{\text{train}}=0.955$), ($\text{RMSE}_{\text{train}}=0.098$), and ($F_{\text{train}}=129.643$) for the training set. For the test set, the model produced ($R^2_{\text{test}}=0.862$), ($\text{RMSE}_{\text{test}}=0.171$), and ($F_{\text{test}}=3.521$). The lower training and test errors of GA-SVM compared with GA-MLR indicate that a nonlinear relationship exists between the selected descriptors and the photooxidation rate constants. The following are the statistical results obtained from the GA-SVM model.

$N_{\text{train}}=24$, $R^2_{\text{train}}=0.955$, $\text{RMSE}_{\text{train}}=0.098$, $F_{\text{train}}=129.643$, $Q^2_{\text{LOO}}=0.589$, $Q^2_{\text{LGO}}=0.589$, $Q^2_{\text{adj}}=0.949$, $N_{\text{test}}=6$, $R^2_{\text{test}}=0.862$, $\text{RMSE}_{\text{test}}=0.171$, $F_{\text{test}}=3.521$

The performance of the GA-SVM model was compared with the results obtained from the GA-MLR model of this study and the SW-MLR model of the previous study (Wang et al., 2023), and a summary of the results of all methods is presented in Table 4.

As shown in Table 4, the GA-SVM model achieved higher (R^2) values and lower RMSE values than the GA-MLR model for both the training and test sets. These results indicate that the nonlinear SVM method captured structure–activity relationships that were not completely represented by the linear model. The GA-SVM model also showed a higher training-set (R^2) and a lower training-set RMSE than the previously reported SW-MLR model of Wang et al. (Wang et al., 2023). However, a direct comparison of external predictive performance with the SW-MLR model is not possible because corresponding test-set statistics were not reported for that model. Therefore, the results demonstrate the improved performance of GA-SVM relative to the linear models evaluated using the available data.

Physicochemical and mechanistic interpretation of the selected descriptors

By interpreting the selected molecular descriptors and their correlation with photooxidation activity, valuable chemical insights can be gained into the photo degradation mechanism of

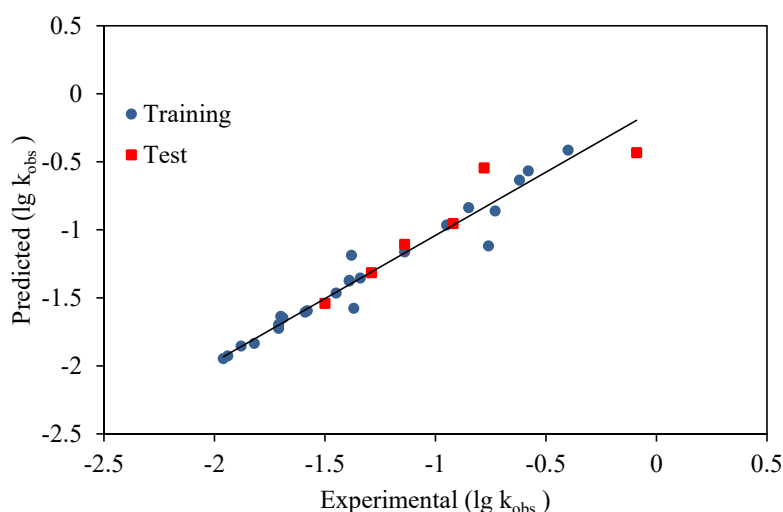


Fig. 6. Experimental versus predicted $\lg k_{\text{obs}}$ values obtained using the GA-SVM model for the training and test sets.

Table 4. Statistical results of the various QSAR models^a

	Training				Test				Ref.
	n	R^2	RMSE	F	n	R^2	RMSE	F	
GA-MLR	24	0.837	0.184	34.442	6	0.781	0.231	2.259	This work
GA-SVM	24	0.955	0.098	129.643	6	0.862	0.171	3.521	This work
SW-MLR	23	0.887	0.190	49.920	7	-	-	-	(Wang et al., 2023)

Table 5. Matrix of correlation coefficients of selected descriptors by GA-MLR

	ESP-DPSA-1	SPH	R3m+
ESP-DPSA-1	1	-	-
SPH	-0.354	1	-
R3m+	-0.330	-0.507	1

AMPs compounds in graphene oxide (GO)-based AOP systems.

ESP-DPSA-1 represents the difference between the positively and negatively charged partial surface areas of a molecule and therefore reflects the imbalance in its electrostatic surface distribution. The positive coefficient of ESP-DPSA-1 in Equation 2 indicates that an increase in this descriptor is associated with an increase in ($\lg k_{\text{obs}}$). In a graphene oxide-mediated system, molecular electrostatic properties may affect adsorption orientation and interactions with oxygen-containing functional groups, including hydroxyl, epoxy, carbonyl, and carboxyl on the graphene oxide surface (Khine et al., 2022). A pronounced electrostatic surface imbalance may promote favorable intermolecular orientation or closer contact between the micropollutant and reactive surface regions. Such interactions can potentially facilitate interfacial charge transfer or improve the accessibility of the molecule to photogenerated reactive species (Wang et al., 2023). Nevertheless, ESP-DPSA-1 alone cannot identify the dominant reactive oxygen species or prove a specific electron-transfer mechanism; it only indicates that molecular surface charge distribution is statistically relevant to the observed photooxidation rate.

The second selected descriptor, SPH, represents molecular sphericity and describes the degree to which the three-dimensional shape of a molecule approaches a compact spherical geometry. The positive coefficient of SPH indicates that, within the chemical space of the present dataset, more compact molecular structures tend to exhibit higher ($\lg k_{\text{obs}}$) values. Molecular shape may influence the orientation and contact area of an aromatic micropollutant on the graphene oxide surface, as well as the exposure of reactive aromatic sites to photogenerated oxidizing species. A compact molecular geometry may permit favorable adsorption configurations or reduce steric restriction around reactive regions of the molecule. However, SPH is a geometrical descriptor and should not be directly interpreted as a measure of hydrophobicity, electrostatic potential, or electron density. Its contribution primarily indicates the importance of molecular three-dimensional shape and compactness in the modeled photooxidation process.

R3m+ is a three-dimensional GETAWAY descriptor corresponding to the maximal autocorrelation at a topological lag of three, weighted by atomic masses (Consonni, Todeschini, & Pavan, 2002). It contains information about the distribution of atomic masses within the molecular geometry and the relationship between heavier atoms and their three-dimensional arrangement (Consonni, Todeschini, et al., 2002). The negative coefficient of R3m+ in Equation 2 indicates that increasing values of this descriptor are associated with decreasing, $\lg k_{\text{obs}}$. This relationship may reflect the influence of heavy substituents, nonuniform mass distribution, steric effects, or increased structural complexity on the interaction of the molecule with graphene oxide and photogenerated reactive species. Molecules with a greater mass-weighted spatial distribution may adopt less favorable adsorption orientations or may possess reduced accessibility of reactive aromatic sites. The negative contribution of R3m+ should therefore be interpreted as an inverse association between mass-weighted three-dimensional organization and the observed photooxidation rate.

The combined contributions of ESP-DPSA-1, SPH, and R3m+ suggest that the photooxidation behavior of the investigated compounds cannot be explained solely by the presence of an individual functional group. Electrostatic surface distribution can affect intermolecular interactions and adsorption orientation, molecular compactness can influence geometrical accessibility, and mass-weighted atomic arrangement can represent steric and structural-

complexity effects. The experimental dataset also shows that several compounds containing amino or alkoxy substituents have relatively high photooxidation rates, whereas compounds containing carboxyl or methylsulfonyl groups generally exhibit lower rates. This trend is consistent with the possible effects of electron-donating and electron-withdrawing substituents on aromatic electron density and interfacial charge-transfer processes (Awfa et al., 2018; Qutob et al., 2024; Wang et al., 2023). Nevertheless, the behavior of positional isomers indicates that substituent identity alone is insufficient and that its position, molecular geometry, and overall electrostatic distribution should also be considered. Confirmation of these proposed relationships will require complementary experiments, such as adsorption measurements, radical-scavenging studies, spectroscopic characterization, or quantum-chemical investigation of charge-transfer pathways.

Environmental relevance of the model

The developed GA-SVM model, within its defined application domain, can be used as a tool for initial screening and prioritization of structure-related aromatic micropollutants. For example, the model can support the identification of compounds predicted to undergo relatively fast or slow photooxidation, guide the selection of target compounds for experimental kinetic studies, and provide initial estimates when experimental data are not available. Such computational prioritization can reduce the number of compounds that require time-consuming laboratory evaluation in the early stages of photocatalytic treatment research.

CONCLUSION

This study successfully developed robust and predictive QSAR models for the photocatalytic oxidation of aromatic micropollutants (AMPs). By combining genetic algorithm-based descriptor selection with both MLR and SVM modeling approaches on a dataset of AMPs, reliable correlations between molecular structure and pseudo-first-order rate constants (k_{obs}) were established. The GA-SVM model clearly outperformed the GA-MLR model, achieving excellent statistical performance and demonstrating superior generalization ability on the external test set. Rigorous validation procedures, including leave-one-out and leave-group-out cross-validation, Y-randomization tests, and William's plot analysis, confirmed the absence of chance correlation, overfitting, or influential outliers, thereby verifying the robustness, stability, and wide applicability domain of the developed models. Moreover, analysis of the three selected descriptors revealed that parameters related to the partial charge of surface areas, surface properties of molecules, and atomic mass play a crucial role in photooxidation reaction of aromatic micro-pollutants. Within its defined applicability domain, the GA-SVM model may assist in screening and prioritizing structurally related aromatic micropollutants for further experimental investigation.

AUTHOR CONTRIBUTIONS

The manuscript was prepared by all authors. The E.P., corresponding author, performed the calculations and supervised the research and M.N., was the writer and compilers of the manuscript. All authors reviewed and approved the final version.

ACKNOWLEDGMENTS

This research was carried out with the support of Mohaghegh Ardabili University (Research grant number: 1402.d.9.23916).

CONFLICT OF INTERESTS

The authors declare that there is not any conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy has been completely observed by the authors.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

- Alhijawi, B., & Awajan, A. (2024). Genetic algorithms: Theory, genetic operators, solutions, and applications. *Evol. Intell.*, 17(3), 1245-1256.
- Alimohammadi, S., Hamidi, A., Pargolghasemi, P., Nourani, N., & Hoseininezhad-Namin, M. S. (2019). QSAR study of antiproliferative drug against A549 by GA-MLR and SW-MLR methods. *Chem. Rev. Lett.*, 2(4), 193-198.
- Auria, L., & Moro, R. A. (2008). Support Vector Machines (SVM) as a technique for solvency analysis. *SSRN Electron. J.*
- Awfa, D., Ateia, M., Fujii, M., Johnson, M. S., & Yoshimura, C. (2018). Photodegradation of pharmaceuticals and personal care products in water treatment using carbonaceous-TiO₂ composites: A critical review of recent literature. *Water Res.*, 142, 26-45.
- Consonni, V., Todeschini, R., & Pavan, M. (2002). Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comput. Sci.*, 42(3), 682-692.
- Consonni, V., Todeschini, R., Pavan, M., & Gramatica, P. (2002). Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J. Chem. Inf. Comput. Sci.*, 42(3), 693-705.
- Ge, L., Dong, Q., Halsall, C., Chen, C.-E. L., Li, J., Wang, D., Zhang, P., & Yao, Z. (2018). Aqueous multivariate phototransformation kinetics of dissociated tetracycline: Implications for the photochemical fate in surface waters. *Environ. Sci. Pollut. Res.*, 25(16), 15726-15732.
- Glaser, R. (2009). AMPAC 9. *J. Am. Chem. Soc.*, 131(37), 13564.
- Grandclément, C., Seyssiecq, I., Piram, A., Wong-Wah-Chung, P., Vanot, G., Tiliacos, N., Roche, N., & Doumenq, P. (2017). From the conventional biological wastewater treatment to hybrid processes, the evaluation of organic micropollutant removal: A review. *Water Res.*, 111, 297-317.
- Grippo, L., & Lucidi, S. (1997). A globally convergent version of the Polak-Ribiere conjugate gradient method. *Math. Program.*, 78(3), 375-391.
- Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics*, 15(1), 41-51.
- Kar, S., Roy, K., & Leszczynski, J. (2018). Applicability domain: A step toward confident predictions and decidability for QSAR modeling. (In O. Nicolotti (Ed.), *Computational toxicology: Methods and protocols* (pp. 141-169). New York: Springer.)
- Katritzky, A. R., Petrukhin, R., Yang, H., & Karelson, M. (2002). CODESSA PRO. User's manual. (Gainesville, FL: University of Florida.)
- Khine, Y. Y., Wen, X., Jin, X., Foller, T., & Joshi, R. (2022). Functional groups in graphene oxide. *Phys. Chem. Chem. Phys.*, 24(43), 26337-26355.
- Laxmi, D., & Priyadarshy, S. (2002). HyperChem 6.03. *Biotech Softw. Internet Rep.*, 3(1), 5-9.
- Luo, Y., Guo, W., Ngo, H. H., Nghiem, L. D., Hai, F. I., Zhang, J., Liang, S., & Wang, X. C. (2014). A review on

- the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. *Sci. Total Environ.*, 473-474, 619-641.
- Ma, X. Y., Li, Q., Wang, X. C., Wang, Y., Wang, D., & Ngo, H. H. (2018). Micropollutants removal and health risk reduction in a water reclamation and ecological reuse system. *Water Res.*, 138, 272-281.
- Miao, M., Lu, Q., Wang, X., Zhang, Y., Wong, N. H., Sunarso, J., Xiao, C., & Li, N. (2023). Removal of micro-organic contaminants from wastewater: A critical review of treatment technology. *Next Mater.*, 1(2), 100016.
- Miklos, D. B., Remy, C., Jekel, M., Linden, K. G., Drewes, J. E., & Hübner, U. (2018). Evaluation of advanced oxidation processes for water and wastewater treatment - A critical review. *Water Res.*, 139, 118-131.
- Nouri Majd, M., Pourbasheer, E., & Mahmoudzadeh Laki, R. (2025). Quantitative structure activity relationship study of p38 α MAP kinase inhibitors. *Chem. Rev. Lett.*, 8(5), 930-940.
- Phang, S. J., Goh, J. M., Tan, L.-L., Lee, W. P. C., Ong, W.-J., & Chai, S.-P. (2021). Metal-free n/n-junctioned graphitic carbon nitride (g-C₃N₄): A study to elucidate its charge transfer mechanism and application for environmental remediation. *Environ. Sci. Pollut. Res.*, 28(4), 4388-4403.
- Pourbasheer, E., Riahi, S., Ganjali, M. R., & Norouzi, P. (2011). Prediction of solubility of fullerene C₆₀ in various organic solvents by genetic algorithm-multiple linear regression. *Fullerenes Nanotubes Carbon Nanostruct.*, 19(7), 585-598.
- Pourbasheer, E., Aalizadeh, R., Ganjali, M. R., Norouzi, P., Shadmanesh, J., & Methenitis, C. (2014). QSAR study of Nav1.7 antagonists by multiple linear regression method based on genetic algorithm (GA-MLR). *Med. Chem. Res.*, 23(5), 2264-2276.
- Pourbasheer, E., Aalizadeh, R., & Ganjali, M. R. (2019). QSAR study of CK2 inhibitors by GA-MLR and GA-SVM methods. *Arab. J. Chem.*, 12(8), 2141-2149.
- Qutob, M., Alshehri, S., Shakeel, F., Alam, P., & Rafatullah, M. (2024). An insight into the role of experimental parameters in advanced oxidation process applied for pharmaceutical degradation. *Environ. Sci. Pollut. Res.*, 31(18), 26452-26479.
- Sun, Y., Wang, D., Li, L., Ning, R., Yu, S., & Gao, N. (2024). Application of remote sensing technology in water quality monitoring: From traditional approaches to artificial intelligence. *Water Res.*, 267, 122546.
- Suthaharan, S. (2016). Support Vector Machine. (In S. Suthaharan (Ed.), *Machine learning models and algorithms for big data classification: Thinking with examples for effective learning* (pp. 207-235). New York: Springer.)
- Wang, H., Zou, Y., Wang, W., Zhang, Y., Mailhot, G., Li, J., Wu, F., & Luo, L. (2023). Quantitative structure-activity relationship for the photooxidation of aromatic micro-pollutants induced by graphene oxide in water. *Chemosphere*, 315, 137781.
- Wilkinson, J. L., Boxall, A. B. A., Kolpin, D. W., Leung, K. M. Y., Lai, R. W. S., Galbán-Malagón, C., Adell, A. D., Mondon, J., Metian, M., Marchant, R. A., Bouzas-Monroy, A., Cuni-Sanchez, A., Coors, A., Carriquiriborde, P., Rojo, M., Gordon, C., Cara, M., Moermond, M., Luarte, T., . . . Teta, C. (2022). Pharmaceutical pollution of the world's rivers. *Proc. Natl. Acad. Sci. U.S.A.*, 119(8), e2113947119.
- Włodarczyk-Makula, M., Myszograj, S., & Włodarczyk, M. (2023). Removal of organic micro-pollutants from wastewater in electrochemical processes-Review. *Energies*, 16(15), 5591.
- Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, 58(2), 109-130.
- Yan, T., Ping, Q., Zhang, A., Wang, L., Dou, Y., & Li, Y. (2021). Enhanced removal of oxytetracycline by UV-driven advanced oxidation with peracetic acid: Insight into the degradation intermediates and N-nitrosodimethylamine formation potential. *Chemosphere*, 274, 129726.
- Younis, S. A., & Moustafa, Y. M. (2017). Synthesis of urea-modified MnFe₂O₄ for aromatic micro-pollutants adsorption from wastewater: Mechanism and modeling. *Clean Technol. Environ. Policy*, 19(2), 527-540.