processes

MDPI

Article

Machine Learning-Driven QSAR Modeling for Predicting Short-Term Exposure Limits of Hydrocarbons and Their Derivatives

Jingjie Shi, Cheng Wang, Linli Ni, Wei Zhao and Xiongjun Yuan *

School of Safety Science and Engineering, Changzhou University, Changzhou 213164, China; shijingjie@cczu.edu.cn (J.S.); wangcheng1302@163.com (C.W.); nilinliczdx@163.com (L.N.); zhaoweicczu@163.com (W.Z.)
* Correspondence: yxj@cczu.edu.cn

Abstract

The scarcity of reliably determined STELs for numerous chemicals severely impedes occupational health risk assessment. To address this gap, this study establishes and validates a suite of robust quantitative structure–activity relationship (QSAR) models to efficiently predict STELs for hydrocarbons and their derivatives. A dataset of 60 compounds was partitioned using Affinity Propagation clustering, and the validity of this division was verified using Tanimoto similarity analysis and Uniform Manifold Approximation and Projection (UMAP). Four optimal molecular descriptors, indicative of molecular size and spatial configuration, were identified using a genetic algorithm. These descriptors served as inputs for one linear model—multiple linear regression (MLR)—and three nonlinear models: support vector machine (SVM), back-propagation artificial neural network (BP-ANN), and extreme gradient boosting (XGBoost). All models were rigorously validated according to OECD principles. The results demonstrated that the XGBoost model achieved superior performance, with key metrics (R^2, Q^2_{loo}, Q^2_{ext}) all exceeding 0.9. Interpretability analysis using SHAP (SHapley Additive exPlanations) revealed that molecular size and symmetry descriptors (*E3u*, *G2m*) positively correlate with STEL, while the degree of unsaturation (*n* = *CHR*) shows a significant negative influence, providing novel mechanistic insights into the structure–toxicity relationship. Notably, 96% of the predictions fell within the defined applicability domain, confirming the model's reliability. This study therefore serves as a rapid, accurate, interpretable, and reliable computational tool, with the potential to significantly inform and enhance occupational health and safety decision-making, especially for novel or data-poor chemicals.

Keywords: quantitative structure–activity relationship (QSAR); short-term exposure limit (STEL); hydrocarbons and their derivatives; molecular descriptors; multiple linear regression (MLR); support vector machine (SVM); back-propagation artificial neural network (BP-ANN); extreme gradient boosting (XGBoost)

check for updates

Academic Editor: Gade Pandu Rangaiah

Received: 5 November 2025
Revised: 9 December 2025
Accepted: 10 December 2025
Published: 12 December 2025

Citation: Shi, J.; Wang, C.; Ni, L.; Zhao, W.; Yuan, X. Machine Learning-Driven QSAR Modeling for Predicting Short-Term Exposure Limits of Hydrocarbons and Their Derivatives. *Processes* **2025**, *13*, 4025. https://doi.org/10.3390/pr13124025

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Hydrocarbons and their derivatives, essential feedstocks in the petrochemical and energy sectors, pose significant risks of occupational exposure [1,2]. Compounds like benzene, alkanes, and aromatics can enter the body through inhalation or dermal contact, where short-term, high-concentration exposure may cause acute neurotoxicity, organ damage, or

fatality [3–5]. The Short-Term Exposure Limit (STEL)—the time-weighted average concentration not to be exceeded in any 15-min period—is a critical metric for evaluating such risks [6]. However, reliable STEL values are lacking for a wide range of industrially relevant substances. For example, Song et al. [7] reported the absence of STEL data in up to 65% of chemical leakage incidents involving alkanes and aromatic derivatives, forcing emergency responses to rely on operational experience rather than scientific evidence. This gap is further highlighted by Zheng et al. [8] and Ismail et al. [9], who documented insufficient occupational exposure limits (OELs) and elevated health risks in settings like LPG retail and gasoline stations. Field data from aromatics production [10] also show that workplace benzene concentrations often exceed safety thresholds. Current approaches for determining STELs remain challenging and can be broadly categorized into several methodological streams, each with inherent limitations. One stream relies on epidemiological studies and field monitoring, investigating the health impacts of short-term exposure to various pollutants such as airborne particulate matter [11], radon [12], and nitrogen dioxide [13]. While valuable, these approaches are often resource-intensive and not universally applicable for deriving STELs for specific industrial chemicals. Consequently, many researchers continue to depend on controlled animal experimentation to establish short-term exposure guidelines, as evidenced by recent toxicological studies on volatile pesticides [14], PM2.5 [15], pharmaceuticals [16], and nanomaterials [17]. However, these methods are associated with lengthy cycles, high costs, and significant ethical concerns, limiting their practicality for high-throughput screening.

To circumvent the limitations of experimental approaches, some scholars have pursued analytical and computational methods. For instance, instrumental detection techniques have been employed to establish STELs for specific compounds like diacetyl in workplace environments [18]. Physiologically based pharmacokinetic (PBPK) modeling has also been utilized to simulate internal exposure and estimate STELs for chemicals like toluene [19]. While promising, these methods often remain targeted at individual or limited classes of substances. Notably, the field of Quantitative Structure–Activity Relationship (QSAR) modeling, which offers the potential for high-throughput prediction, has seen limited application. The pioneering work of Russell et al. [20] represents a foundational effort, establishing a correlation model between RD_{50} values and STELs. Nevertheless, this model has two critical constraints that hinder its general utility: firstly, its applicability is confined to chemicals with available RD_{50} data, excluding numerous compounds of interest; secondly, it does not directly leverage the fundamental molecular structural characteristics that intrinsically govern toxicity, potentially limiting its mechanistic insight and contributing to its suboptimal predictive accuracy (R^2 = 0.75). Therefore, a significant gap exists for a universal, accurate, and structurally based predictive QSAR model capable of efficiently predicting STELs for a broad range of chemicals without relying on specialized bioassay data.

Given the vast chemical space and the practical constraints of data availability, this study strategically focuses on hydrocarbons and their derivatives. This class of compounds represents a critically important and ubiquitous group in the petrochemical, energy, and manufacturing sectors, where short-term exposure risks are frequently encountered [1,2,7,10]. By concentrating on this chemically coherent and high-priority domain, we aim to develop a highly reliable and fit-for-purpose predictive tool, establishing a robust methodological foundation that can be extended to other chemical classes in the future.

To address this gap, this study develops robust QSAR models for predicting the STELs of hydrocarbons and their derivatives. Our work introduces several key advancements beyond the existing literature: (i) Development of a Generalizable Model: We predict STELs

directly from fundamental molecular structures, eliminating the dependency on specialized bioassay data and significantly broadening the model's applicability domain. (ii) Advanced Data-Splitting Strategy: The Affinity Propagation (AP) clustering algorithm, coupled with Tanimoto similarity analysis and UMAP visualization, ensures the representativeness of the training set and the structural independence of the test set. (iii) Comprehensive Model Benchmarking: We conduct a systematic comparison of diverse machine learning algorithms—from multiple linear regression (MLR) to support vector machine (SVM), back-propagation artificial neural network (BP-ANN), and extreme gradient boosting (XGBoost)—to identify the optimal modeling approach. (iv) Rigorous Validation and Defined Applicability: Adherence to OECD principles is ensured through internal and external validation, complemented by a leverage-based applicability domain analysis, guaranteeing prediction reliability and transparency. Consequently, this study establishes a reliable and well-characterized strategy for the rapid prediction of STELs. Furthermore, by employing both univariate correlation and multivariate modeling techniques, we aim to dissect the individual and collective contributions of molecular descriptors, providing a more nuanced understanding of the structure–STEL relationship. The research roadmap is illustrated in Figure 1.

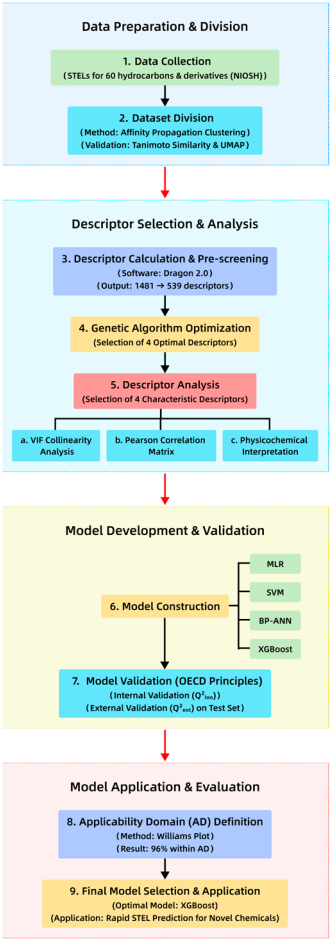


Figure 1. Workflow for the Development and Validation of QSAR Models Predicting STELs of hydrocarbons and their derivatives.

2. Principal Theoretical Approaches

2.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning method with a solid theoretical foundation. It is fundamentally based on statistical learning theory, particularly the

Vapnik–Chervonenkis (VC) theory and the principle of structural risk minimization [21]. Its core mechanism involves transforming nonlinear problems from low-dimensional input spaces into linearly separable ones in high-dimensional feature spaces. This capability allows the SVM to effectively classify data that are not separable in the original space. The method performs robustly in high-dimensional settings and under small-sample conditions, exhibiting strong generalization ability [22,23]. Owing to these advantages, SVM has been widely applied in computational toxicology and quantitative structure–activity relationship (QSAR) modeling in drug discovery and development [24–26]. Accordingly, the relationship between STEL and molecular feature descriptors was examined using a Support Vector Machine (SVM) model to capture its inherent nonlinearity.

Consider a training set D comprising n samples, where each sample is a pair (x_i, y_i), with x_i being the input vector and y_i the corresponding output. The objective of the regression is to learn a function $f(x)$ that maps the inputs to the outputs. The regression function is formulated as presented in Equation (1).

$$f(x) = \omega^T \phi(x) + b \tag{1}$$

In the given equation, $f(x)$ is the regression function, while ω, $\phi(x)$, and b represent the weight vector, the feature mapping function, and the bias term, respectively. The weight vector ω and the bias term b are the model parameters to be determined. The regression formulation allows a maximum deviation of ς between the predicted value $f(x)$ and the true value y. Thus, the learning problem is framed as a constrained optimization task, which aims to maximize the margin while respecting this deviation tolerance. The minimization of the loss function is described in Equation (2).

$$\begin{cases} \underset{w,b,\varepsilon_i,\hat{\varepsilon}}{min_i} = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\varepsilon_i + \hat{\varepsilon}_i) \\ f(x_i) - y_i \le \varepsilon_i + \varsigma, \\ y_i - f(x_i) \le \hat{\varepsilon}_i + \varsigma, \\ \varepsilon_i \ge 0, \hat{\varepsilon}_i \ge 0, \ i = 1, 2, \ldots m. \end{cases} \tag{2}$$

In Equation (2), C denotes the regularization constant, ε_i and $\hat{\varepsilon}_i$ the slack variables, and ς is the width of the ε-insensitive band.

The original optimization problem is transformed into its dual form via the Lagrange multiplier method. Solving this dual problem yields the support vector regression function, given in Equation (3).

$$\begin{cases} f(x) = \sum_{i=1}^{m}(\alpha \hat{} i + \alpha_i)k(x, x_i) + b \\ k(x, x_i) = \varphi(x_i)^T \varphi(x_i) \end{cases} \tag{3}$$

In this formulation, $k(x, x_i)$ denotes the kernel function, and α_i and $\hat{\alpha}_i$ denote the Lagrange multipliers derived from the dual problem. The radial basis function (RBF) kernel was selected for its ability to implicitly map data into a high-dimensional space, thereby enhancing linear separability and enabling the construction of an effective nonlinear regression model. The support vector machine (SVM) demonstrates excellent generalization performance and a notable ability to address nonlinear problems [27]. Consequently, the model's efficacy is highly dependent on its parameter settings. Therefore, the appropriate selection of the kernel parameters and the regularization constant C is critical for optimizing model performance.

2.2. Back-Propagation Artificial Neural Network (BP-ANN)

The Back-Propagation Artificial Neural Network (BP-ANN) is a widely used model, renowned for its well-established learning algorithm, ease of implementation, and proficiency in nonlinear mapping, function approximation, and robust performance. These attributes have led to its extensive application in diverse fields such as combinatorial optimization [28], image processing [29], and pattern recognition [30]. The Back-Propagation Artificial Neural Network (BP-ANN) is a multi-layer feedforward network renowned for its proficiency in nonlinear mapping, function approximation, and pattern recognition [31]. The network architecture comprises an input layer, one or more hidden layers, and an output layer. The model processes information through forward propagation, where input signals are transformed by weighted connections and activation functions to generate an output.

(1) Forward Propagation

Each neuron in the hidden and output layers receives the weighted sum of its inputs, which is then passed through an activation function. The hyperbolic tangent sigmoid (tansig) and linear (purelin) functions were employed as the activation functions for the hidden and output layers, respectively, as defined in Equations (4) and (5).

$$tansig(x) = \frac{2}{1 + exp(-2x)} - 1 \tag{4}$$

$$\text{purelin}(x) = x \tag{5}$$

The discrepancy between the network's predicted output and the actual target is quantified by a loss function. The Mean Squared Error (MSE), given in Equation (6), was used as the loss function in this study.

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \tag{6}$$

(2) Back-propagation

The core of the BP-ANN learning algorithm is error back-propagation. The calculated error is propagated backward through the network, and the connection weights (w) and biases (b) are iteratively adjusted using gradient-based optimization to minimize the error function. The Levenberg–Marquardt algorithm (Equation (7)) was utilized as the training function due to its fast convergence for medium-sized networks.

$$W_{k+1} = W_k - (J^T J + \mu I)^{-1} J^T e \tag{7}$$

In Equation (7), J denotes the Jacobian matrix, e is the error vector, and μ is the adaptive damping factor. For the weight update process, the gradient descent with momentum method (learngdm) was applied, as defined in Equations (8) and (9).

$$\Delta W_{ij}(k) = \alpha \cdot W_{ij}(k-1) - \eta \cdot \frac{\partial E}{\partial W_{ij}} \tag{8}$$

$$W_{ij}(k+1) = W_{ij}(k) + \Delta W_{ij}(k) \tag{9}$$

In these equations, $w_{ij}(k)$ denotes the connection weight from neuron i to neuron j at the k-th iteration, $\Delta w_{ij}(k)$ is the weight update value, α is the momentum coefficient (typically set to 0.9), η is the learning rate, and $\frac{\partial E}{\partial w_{ij}}$ is the gradient of the error function E with respect to the weight w_{ij}.

The number of neurons in the hidden layer (L) was determined empirically using Equation (10), where N and M are the number of neurons in the input and output layers, respectively, and λ is a scaling constant ranging from 1 to 10.

$$L = \sqrt{(N+M)} + \lambda \tag{10}$$

The iterative process continues until the network's output error falls below a specified threshold or the maximum number of epochs is reached.

2.3. Extreme Gradient Boosting (XGBoost)

XGBoost is an efficient, parallel ensemble learning algorithm capable of accurately processing high-dimensional and large-scale datasets, while exhibiting broad applicability across diverse application scenarios. This method is developed based on the principle of Boosting in ensemble learning [32].

Initially, a simple model is constructed using the training set data as a weak learner. Subsequently, by analyzing the prediction residuals of this model and minimizing the objective function, an additional weak learner is iteratively generated. Through repeated iterations, multiple weak learners are sequentially produced and integrated, ultimately resulting in a strong learner with enhanced predictive accuracy. Given its inherent advantages in modeling complex data and capturing nonlinear relationships, this study utilizes XGBoost to further develop a predictive model suitable for characterizing the nonlinear relationship between STELs and molecular feature descriptors. XGBoost employs the second-order Taylor expansion to approximate the objective function and prevents overfitting by incorporating a regularization term that controls model complexity. The objective function is formulated as follows:

$$\begin{cases} Obj = \sum_{i=1}^{n} (y_i, \hat{y}_t^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant \\ \Omega(f_t) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \end{cases} \tag{11}$$

Following a second-order Taylor expansion, the objective function becomes:

$$Obj^t = \sum_{i=1}^{n} \left[l(y_i, \hat{y}_t^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t) + \text{constant} \tag{12}$$

Among them, $\sum_{i=1}^{n} l\left(y_i, \hat{y}_t^{(t-1)}\right)$ is a fixed value and can be omitted along with the constant term in the optimization objective function. The equation thus simplifies to:

$$Obj^t = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{13}$$

The k-th decision tree maps the input sample to a leaf node, denoted as $f_k(x)$, where the output value at the leaf node corresponding to the index $q(x)$ is represented by $w_{q(x)}$; thus, $f_k(x) = w_{q(x)}$.

$$\begin{cases} Obj^t = \sum_{i=1}^{n} \left[g_i w_{q(x_i)} + \frac{1}{2} h_i {w_{q(x_i)}}^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ = \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) {w_j}^2 \right] + \gamma T \\ = \sum_{j=1}^{T} \left[G_i w_j + \frac{1}{2} (H_i + \lambda) {w_j}^2 \right] + \gamma T \end{cases} \tag{14}$$

The above equation represents a quadratic function of w_j. By taking its derivative, the optimal value of w_j can be determined as:

$$\begin{cases} w_j^* = -\dfrac{G_j}{H_j + \lambda} \\ Obj^* = -\dfrac{1}{2}\sum\limits_{j=1}^{T} \dfrac{G_j^2}{H_j + \lambda} + \gamma T \end{cases} \quad (15)$$

3. Model Construction Preliminaries

3.1. Determination of the Sample Set

To ensure consistency in data sources, all STEL data for the 60 hydrocarbons and their derivatives were obtained from the NIOSH Pocket Guide to Chemical Hazards (NPG) [33]. This database draws on policy documents, standard guidelines, and recognized references in industrial hygiene and toxicology, ensuring its authority and reliability. The natural logarithm of STEL (ln(STEL)) was used as the dependent variable for subsequent model development. A comprehensive listing of all 60 compounds, including their names, CAS numbers, descriptor values, experimental ln(STEL), and predicted values from all models, is provided in Table S1 of the Supplementary Information. To ensure a scientifically robust partition of the sample set, we employed the Affinity Propagation (AP) clustering algorithm to classify the compounds [34]. The AP algorithm was executed with the following parameters: similarity matrix (Euclidean distance), preference (median similarity), damping factor (0.5), and maximum iterations (200). Convergence was achieved after 101 iterations, defining nine cluster centers (exemplars). The dataset was then stratified by these clusters and split into a 4:1 ratio, resulting in a training set (n = 48) for descriptor selection and model development, and a test set (n = 12) for external validation of predictive performance and robustness [35].

To validate the rationality of this cluster-based partitioning, the calculation of Tanimoto similarities (E_j) serves as a quantitative measure to evaluate the rationality of randomly partitioning the training and test sets. This metric effectively reflects the degree of structural similarity between the test set and the training set within the molecular feature space [36]. Figure 2 illustrates the distribution characteristics of E_j for the test set samples. The histogram exhibits a trend toward a nearly normal and highly concentrated distribution. Following Gaussian fitting, the distribution parameters were determined as a mean of 0.371 and a standard deviation of 0.078. Although there is a minor numerical discrepancy in the statistical data (mean = 0.4094, standard deviation = 0.1005), the overall distribution trend of the fitting results is in high agreement, further supporting the concentrated nature of the E_j distribution.

The statistical analysis presented in Table 1 indicates that the median Ej value across the 12 test set samples is 0.3868, with the 25th and 75th percentiles recorded at 0.3646 and 0.4255, respectively. The values range from a minimum of 0.2857 to a maximum of 0.6500, predominantly falling within the low to moderate similarity intervals. Overall, the distribution exhibits a moderate level of dispersion. The aforementioned distribution characteristics demonstrate that the training set exhibits comprehensive coverage of the molecular chemical space, while the test set maintains a structurally representative relationship with the training set, evidenced by an average similarity value approaching 0.41. Furthermore, the standard deviation of E_j is 0.1005, indicating a relatively low level of dispersion. This implies that the similarity distribution within the test set remains stable with minimal variation, which is consistent with the statistical principles underlying random sampling. In conclusion, the acceptable range and central tendency of sample similarity indicate that the partitioning of the training and test sets exhibits sufficient independence and

representativeness, thereby establishing a robust foundation for subsequent model training and the evaluation of generalization performance.

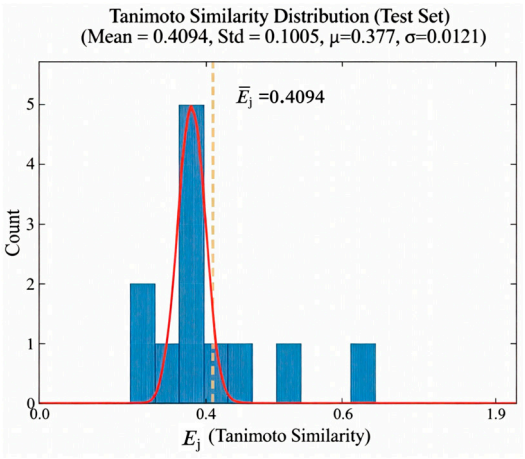


Figure 2. Distribution of Tanimoto similarity coefficients between test set compounds and their nearest neighbors in the training set.

Table 1. Tanimoto Similarity Degree Statistics.

Sample Count	Median	25% Quantile	75% Quantile	Min	Max
12	0.3868	0.3646	0.4255	0.2857	0.65

The visualization of the Morgan fingerprint feature space of chemical compounds in the training and test sets using UMAP (Figure 3) reveals overlapping regions in sample distribution, indicative of shared molecular characteristics and structural similarities, as well as distinct and well-separated clusters specific to each dataset. This pattern supports the conclusion that the training and test sets are structurally related while retaining unique, dataset-specific features. The spatial distribution pattern is consistent with the conclusion drawn from the Tanimoto similarity analysis, which indicates that "the test set and the training set are both related and independent." This agreement suggests that the dataset partitioning not only preserves the intrinsic relationships among molecular features but also ensures a sufficiently diverse sample basis for reliable model evaluation, thereby supporting the rationality and validity of the data splitting strategy. The rational partitioning of the sample dataset establishes a reliable foundation for the subsequent selection of feature molecular descriptors, as well as for model development and evaluation. This integrated workflow—combining AP clustering, Tanimoto similarity analysis, and UMAP visualization—was implemented with particular rigor due to the moderate sample size (n = 60). In QSAR modeling, a limited dataset increases the risk of over-optimistic performance metrics if the training and test sets are not structurally representative and independent [37]. The approach adopted here proactively mitigates this risk by ensuring that the chemical space is comprehensively sampled for training while maintaining a test set that is both representative of the overall domain and structurally distinct enough to provide a stringent assessment of generalizability. Consequently, the external validation coefficient (Q^2_{ext}) and other performance metrics reported in this study can be considered a robust and conservative estimate of the model's predictive power for new, previously unseen compounds.

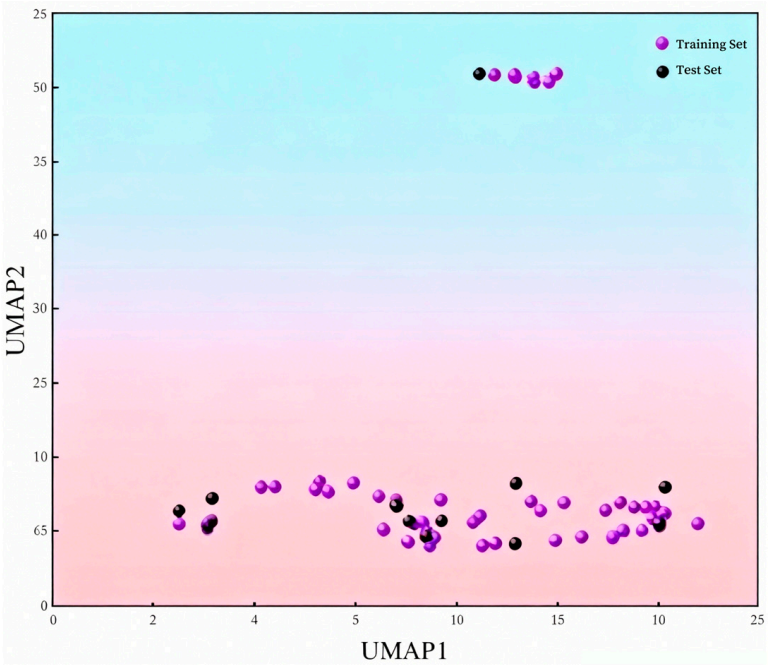


Figure 3. UMAP visualization of the chemical space for training and test sets, demonstrating structural representativeness and independence.

3.2. Determination of Characteristic Molecular Descriptors

Molecular descriptors are numerical representations of molecular structures and play a crucial role in elucidating physicochemical properties, predicting biological activities, and supporting drug discovery research [24]. In this study, the molecular structures of hydrocarbons and their oxygen-containing derivatives were initially constructed using ChemDraw23.1.1. These structures were subsequently imported into Hyperchem professional 8.0 software for structural optimization using the MM+ molecular mechanics and PM3 geometry optimization methods [38]. Subsequently, the Dragon2.0 software was employed to transform the chemical information contained within molecular structures into numerical descriptors. A total of 1481 molecular descriptors spanning 18 categories were generated and underwent preliminary screening. During this screening phase, descriptors exhibiting constant or near-constant values, as well as those with correlation coefficients of 0.95 or above, were removed. As a result, a refined set of 539 molecular descriptors was obtained. However, at this stage, the number of descriptor variables was still excessive, and substantial multicollinearity continued to exist among the independent variables. The GFA module in Materials Studio 2024 software was employed to screen molecular descriptors using a genetic algorithm, leading to the identification of four characteristic molecular descriptors for model development. The genetic algorithm was configured with the following parameter settings. The Friedman LOF served as the fitness function, with a smoothing parameter α set to 0.5. The initial equation length was set to 5, and the maximum equation length was limited to 10. The population size was fixed at 50, the maximum number of generations was set to 500, and the mutation probability was specified as 0.1. Under these conditions, four characteristic molecular descriptors were successfully identified: *Mor28m*, *E3u*, *G2m*, and *n* = *CHR*. The names, types, definitions, and corresponding variance inflation factor (VIF) values of the four characteristic molecular descriptors are presented in Table 2.

Table 2. The names, types, definitions, and corresponding variance inflation factor (VIF) values of the four characteristic molecular descriptors.

Name	Type	Definition	VIF
Mor28m	3D-MoRSe	signal 28/weighted by atomic masses	1.237
E3u	WHIM	3rd component accessibility directional WHIM index/unweighted	1.4
G2m	WHIM	2nd component symmetry directional WHIM index/weighted by atomic masses	1.211
n = CHR	functional group	number of secondary C (sp2)	1.047

As shown in Table 2, the four descriptors fall into three distinct categories—3D-MoRSE, WHIM, and functional group descriptors—each of which captures molecular structural features from a unique perspective. *Mor28m* is a 3D-MoRSe descriptor primarily used to characterize the three-dimensional spatial configuration and mass-weighted atomic distribution obtained from electron diffraction data, thereby effectively capturing molecular structural features in three-dimensional space. Both *E3u* and *G2m* are WHIM descriptors. Specifically, *E3u* denotes the accessibility-based WHIM index for the third principal component derived from the unweighted covariance matrix of atomic coordinates, reflecting molecular size. In contrast, *G2m* refers to the symmetry-based WHIM index for the second principal component obtained from the covariance matrix of atomic coordinates weighted by atomic mass, offering insight into the distribution of atomic sizes within the molecule. *n = CHR* is a functional group descriptor frequently employed to denote the number of secondary carbon atoms within a molecular structure and to characterize its geometric configuration [39,40]. To ensure the validity of the selected descriptors, a multicollinearity analysis was conducted for each variable using the variance inflation factor (VIF) as an indicator. It is generally accepted that the absence of multicollinearity among independent variables is indicated when the VIF falls within the range of 0 to 10 [41,42]. As presented in Table 2, the VIF values for the characteristic molecular descriptors range from 1 to 2, which are substantially below the commonly accepted threshold of 10. This suggests that multicollinearity among the characteristic molecular descriptors, as well as between these descriptors and the STELs, is not a significant concern. Moreover, all four characteristic molecular descriptors demonstrate a statistically significant association with STELs.

To assess the contribution of the molecular descriptors, Figure 4 displays the Pearson correlation coefficient matrix. The Pearson correlation matrix is a square matrix derived from the Pearson correlation coefficients [43], serving as a fundamental tool in multivariate analysis. It systematically quantifies the strength and direction of linear relationships between pairs of variables, offering a comprehensive overview of inter-variable associations within a dataset. The correlation coefficient's sign indicates the direction of the relationship (positive or negative), while its magnitude represents the strength, with values near ± 1 denoting strong correlations and those near 0 weak ones. This matrix not only offers a visual depiction of the correlation patterns between each descriptor and the target property (ln(STEL)), as well as among the descriptors, but also provides indirect support for the multicollinearity analysis results derived from the VIF approach. Based on the correlation coefficients with ln(STEL) (Figure 4), the molecular descriptors were ranked as follows: *G2m* > *E3u* > *n = CHR* > *Mor28m*. This initial ranking, derived from bivariate analysis, identifies descriptors with the strongest isolated linear relationship with the target property. It is important to note, however, that this perspective does not account for potential interdependencies among the descriptors themselves, which will be further explored in the subsequent multivariate modeling phase.

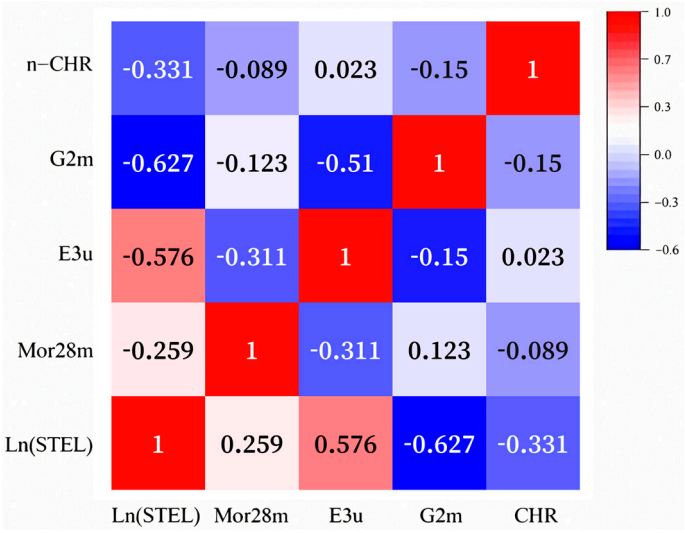


Figure 4. Pearson correlation matrix showing linear relationships between ln(STEL) and the four selected molecular descriptors.

The selection of these four descriptors provides valuable insight into the key structural properties governing the STELs of hydrocarbons and their derivatives. The fact that three of the descriptors (*Mor28m*, *E3u*, *G2m*) belong to the 3D-MoRSE and WHIM categories underscores the paramount importance of a molecule's three-dimensional size, shape, and spatial configuration. These volumetric and spatial properties are critical as they directly influence a compound's volatility, respiratory tract deposition efficiency, and diffusion rate across the alveolar-capillary membrane—all primary factors determining the internal dose following short-term inhalation exposure [44,45]. For instance, larger, bulkier molecules tend to have lower volatility, reducing their airborne concentration, and may deposit more efficiently in the upper airways, limiting their reach to the deep lung. In contrast, the functional group descriptor $n = CHR$, which represents the number of sp^2 hybridized secondary carbons, relates to molecular unsaturation and electron density. This electronic characteristic may affect the compound's chemical reactivity, its propensity to interact with biological nucleophiles, or its participation in metabolic activation pathways, thereby modulating acute toxic effects such as narcosis or tissue irritation [46]. Collectively, these descriptors encode a complementary set of features encompassing molecular volume, accessibility, symmetry, and key functional groups. This holistic representation not only provides a statistical correlation but also aligns with well-established physiochemical and toxicokinetic principles governing acute inhalation toxicity.

3.3. Performance Evaluation Metrics of the Model

A robust model must be validated and assessed in compliance with the Organisation for Economic Co-operation and Development (OECD)'s five principles for QSAR models [47]. This study systematically evaluated the model performance through both internal and external validation approaches, supported by a comprehensive set of performance metrics. The goodness-of-fit on the training set was assessed using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The internal predictive robustness and potential for overfitting were evaluated via the leave-one-out cross-validation coefficient (Q^2_{loo}). Furthermore, the true external predictive capability of the finalized model was rigorously verified using the external validation coefficient (Q^2_{ext}) on the hold-out test set. Additionally, residual analysis plots were employed to diagnose model behavior and ensure the absence of systematic error across all validation stages.

4. Results and Discussion

4.1. Model Construction

4.1.1. MLR Model

A multiple linear regression (MLR) model was constructed using IBM SPSS Statistics 26 to examine the relationship between the four characteristic molecular descriptors (independent variables) and ln(STEL) (dependent variable). The derived MLR equation is presented in Equation (16), and the corresponding model performance metrics are summarized in Table 3. While the statistical parameters for each variable included in the model are detailed in Table 4.

$$\ln(STEL) = 4.338 + 5.49 \times X_{E3u} + 6.88 \times X_{Mor28m} - 2.216 \times X_{n-CHR} - 2.313 \times X_{G2m} \quad (16)$$

Table 3. Test Results of the MLR Model.

Test Variable	R^2	RMSE	SD	*p*-Value	F
Result	0.8043	0.8707	1.213	<0.001	20.381
Inspection Standards	>0.6	the smaller, the better	the smaller, the better	<0.05	$>F_{theory}$

Table 4. The statistical parameters for each variable included in the MLR model.

Characteristic Molecular Descriptors	Regression Coefficient	Standardized Coefficient	Standard Error	*t*-Value	*p*-Value
Constant	4.338		0.497	8.735	<0.001
E3u	5.490	0.527	1.070	5.129	<0.001
Mor28m	6.880	0.460	1.444	4.766	<0.001
n = CHR	−2.216	−0.339	0.582	−3.811	<0.001
G2m	−2.313	−0.360	0.613	−3.771	<0.001

To assess the reliability of the constructed MLR model, its goodness of fit, overall significance of the regression equation, and statistical significance of individual variables were evaluated using mathematical statistical methods, as presented in Tables 2 and 3. The coefficient of multiple determination (R^2) is 0.8043, which exceeds the commonly accepted threshold of 0.6 [48], indicating a relatively strong model fit. The root mean square error (RMSE) is 0.8707, and the standard deviation (SD) of the observed ln(STEL) values is 1.213. The fact that the RMSE is notably lower than the SD suggests that the model captures meaningful structural trends in the data, rather than merely approximating the mean. This indicates a high degree of predictive accuracy and minimal estimation error in the regression model. The obtained value was 20.381, which is greater than the *F* critical value (4, 48) at the 0.05 significance level. Moreover, the *p*-value is significantly less than 0.001. These findings collectively suggest that the equation is statistically significant. The absolute values of the *t*-value for each variable exceed 2, and the corresponding significance levels are substantially below 0.001, indicating that each variable has a statistically significant effect on the regression model. Furthermore, to evaluate the magnitude of the influence exerted by each independent variable on the dependent variable, this study utilized standardized coefficients to assess the relative importance of each variable's association with the dependent variable. The absolute value of the standardized coefficient indicates the relative importance of each descriptor in predicting the ln(STEL) value, with a higher absolute value signifying a greater contribution. As shown in Table 4, the independent variables are ranked in descending order of their contribution to ln(STEL) as follows: *E3u* > *Mor28m* > *G2m* > *n = CHR*. This ranking, based on standardized regression

coefficients within the multivariate model, offers a perspective distinct from the simple Pearson correlation coefficients (*G2m* > *E3u* > *n* = *CHR* > *Mor28m*, Figure 4). The discrepancy is anticipated and insightful. The genetic algorithm (GA) was optimized to select descriptors that maximize the collective predictive power of the model, not necessarily a set of perfectly independent variables. A compact set of descriptors with low-to-moderate intercorrelations (as confirmed by the low VIF values in Table 2) can often capture complementary structural information more effectively than orthogonal ones. The fact that *G2m* exhibits the highest simple correlation but a lower standardized coefficient indicates that a portion of its predictive information is shared with other descriptors in the model. Conversely, the elevated importance of *E3u* and *Mor28m* highlights the unique explanatory power they contribute to the multivariate context. This phenomenon underscores the value of the GA selection and the multivariate modeling approach in elucidating complex structure–property relationships.

Figure 5 presents a comparison between the ln(STEL) values predicted by the MLR model and the corresponding experimental data to evaluate their consistency. As observed, most data points are well aligned along the diagonal line, with only a few showing noticeable deviations, indicating satisfactory predictive accuracy of the MLR model. Moreover, the random scattering of the data points on both sides of the diagonal suggests the absence of systematic bias, further affirming the model's reliability. Thus, the MLR model represents a feasible and reliable tool for predicting the ln(STEL) values of hydrocarbons and their derivatives. Nonetheless, to further investigate potential nonlinear relationships between the ln(STEL) and the molecular structure, a nonlinear modeling approach was also developed in this study with the aim of enhancing predictive performance.

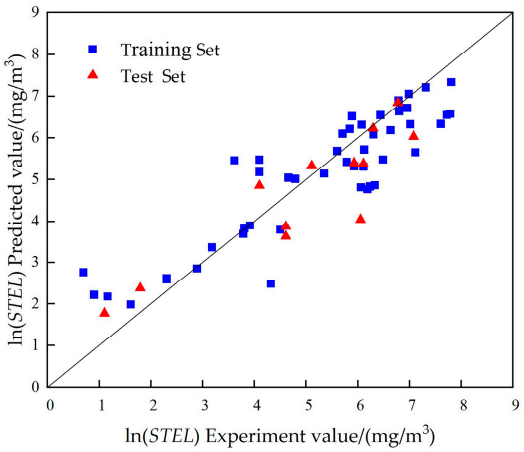


Figure 5. Predicted versus experimental ln(STEL) values for the MLR model.

4.1.2. SVM Model

The libSVM toolbox in MATLAB 2024a software was employed to construct the SVM model based on the characteristic molecular descriptors and ln(STEL). The same set of four characteristic molecular descriptors was used as input variables, with the radial basis function (RBF) selected as the kernel function. A grid search approach was applied to identify the optimal parameter combination across the ranges $C \in (0.001, 1000]$, $\varepsilon \in (0.001, 1]$, and $\gamma \in (0.001, 1]$, thereby establishing the SVM model with the best predictive performance. Following the normalization of both independent and dependent variables, the optimal parameters for the Support Vector Machine (SVM) model were identified as follows: a penalty coefficient (C) of 256, a kernel function bandwidth (γ) of 0.16612, and an insensitivity loss function parameter (ε) of 0.011049. Utilizing the aforementioned optimal parameters, the SVM-based QSAR model was developed. The model demonstrated strong

performance, achieving coefficients of determination (R^2) of 0.8089 for the training set and 0.8257 for the test set, respectively. A plot comparing the predicted and experimental values is provided in Figure 6.

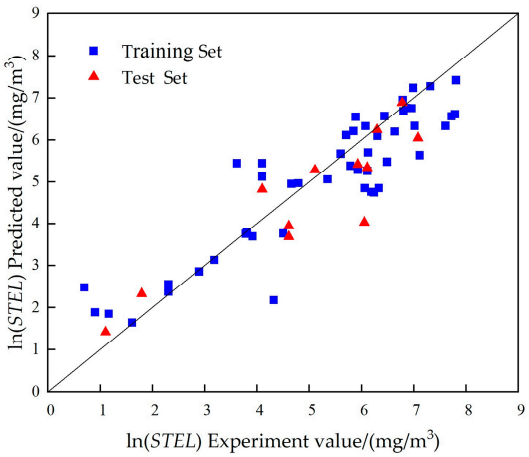


Figure 6. Predicted versus experimental ln(STEL) values for the SVM model.

As shown in Figure 6, the SVM model exhibits a data distribution pattern highly similar to that of the MLR model, with both delivering satisfactory and consistent predictions. Given the comparable performance of these linear and shallow nonlinear approaches, a BP-ANN model was employed to further probe the potential deeper nonlinear relationships between STEL values and molecular structures.

4.1.3. BP-ANN Model

Based on the theoretical framework described in Section 2.2, a Back-Propagation Artificial Neural Network (BP-ANN) model was implemented in MATLAB to predict ln(STEL) using the four selected molecular descriptors as inputs. The network architecture was configured with these descriptors as input neurons and ln(STEL) as the output neuron.

The number of neurons in the hidden layer was optimized within the range of 3 to 13, guided by the empirical formula (Equation (10)). After systematic evaluation, a configuration with five hidden neurons was identified as optimal, yielding the lowest Root Mean Square Error (RMSE = 0.3344). Consequently, a multilayer perceptron (MLP) model with a 4-5-1 architecture was adopted. The model was trained using the Levenberg–Marquardt algorithm, with a learning rate of 0.001, a performance goal (tolerance) of 0.01, and a maximum of 3000 epochs. The hyperbolic tangent sigmoid (tansig) and linear (purelin) functions were employed as activation functions for the hidden and output layers, respectively. The training process converged after 58 iterations, achieving a coefficient of determination (R^2) of 0.8396 for the training set and 0.8824 for the test set.

A comparison between the predicted values generated by the BP-ANN model and the experimentally obtained values is presented in Figure 7. As shown, the data points of the BP-ANN model are more closely clustered around the diagonal than those of the SVM and MLR models, indicating its superior predictive accuracy. This enhanced performance implies a strong nonlinear relationship between STEL values and molecular structure. However, the presence of some outliers indicates potential for further model refinement. Consequently, the XGBoost algorithm was employed to delve deeper into the underlying nonlinear relationships.

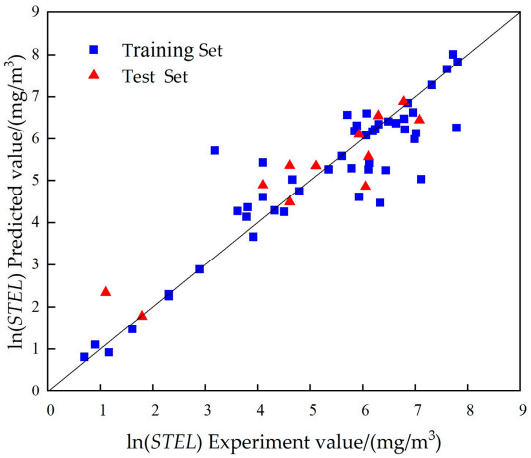


Figure 7. Predicted versus experimental ln(STEL) values for the BP-ANN model.

4.1.4. XGBoost Model

The XGBoost model based on characteristic molecular descriptors and ln(STEL) was developed using MATLAB software. The XGBoost model was developed using MATLAB software. Hyperparameter optimization was conducted, and the final optimal booster configuration was determined as follows: the number of boosting rounds was 120, the maximum tree depth was 5, the learning rate (eta) was 0.2, with L2 and L1 regularization terms (lambda and alpha) set to 1 and 0, respectively. All other parameters were maintained at their default values. With this optimal configuration, the XGBoost model achieved superior performance. It yielded high coefficients of determination, with R^2 values of 0.9445 for the training set and 0.9152 for the test set. The corresponding RMSE values were 0.3608 and 0.6815, and the MAE values were 0.1302 and 0.4645 for the training and test sets, respectively. A comparison between the predicted and experimental values of the XGBoost model is presented in Figure 8.

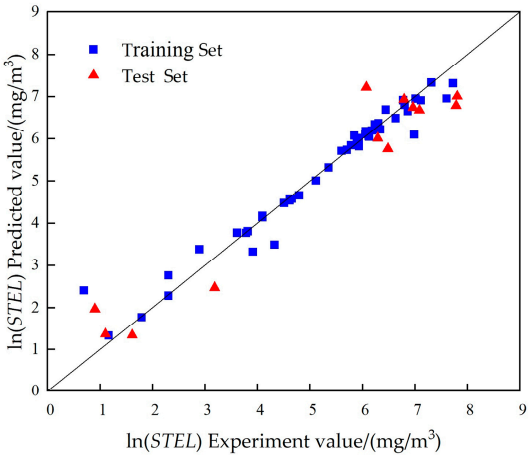


Figure 8. Predicted versus experimental ln(STEL) values for the XGBoost model.

As illustrated in Figure 8, the distribution pattern of the XGBoost model exhibits significant differences compared to those of the MLR, SVM, and BP-ANN models. Specifically, in contrast to the other models, the sample points in the XGBoost model are predominantly clustered closely around the diagonal line, with the majority lying directly on it and only a small fraction deviating noticeably from it. This indicates a high degree of consistency between the predicted and experimental values, reflecting superior predictive performance. In summary, the XGBoost model demonstrates the strongest predictive performance among

all models evaluated, suggesting a robust nonlinear relationship between the STEL values and molecular structure.

4.2. Model Evaluation and Validation

4.2.1. Statistical Quality and Validation Analysis

A systematic evaluation was conducted to compare the performance of the four developed models. Their goodness-of-fit, robustness, and predictive capability were assessed by analyzing key performance metrics and examining residual plots. The primary performance metrics of the MLR, SVM, BP-ANN, and XGBoost predictive models for hydrocarbons and their derivatives are presented in Table 5.

Table 5. Main performance parameters of MLR, SVM, BP-ANN, and XGBoost prediction models for ln(STEL) of hydrocarbons and their derivatives.

Performance Parameters	Model							
	MLR		SVM		BP-ANN		XGBoost	
	Training Set	Test Set	Training Set	Test Set	Training Set	Test Set	Training Set	Test Set
R^2	0.8043	0.8095	0.8089	0.8257	0.8396	0.8824	0.9445	0.9152
RMSE	0.8707	0.8595	0.854	0.8314	0.7672	0.649	0.3608	0.6815
MAE	0.6598	0.6994	0.6356	0.6545	0.496	0.5069	0.1302	0.4645
Q^2_{loo}	0.7891	-	0.7971	-	0.8363	-	0.9532	-
Q^2_{ext}	-	0.7785	-	0.7928	-	0.8737	-	0.9291

As presented in Table 5, the coefficients of determination (R^2) for the training and test sets of the MLR model were 0.8043 and 0.8095, respectively. The corresponding values for the SVM model were 0.8089 and 0.8257, while the BP-ANN model achieved R^2 values of 0.8396 and 0.8824. The XGBoost model yielded the highest R^2 on the training set (0.9445) and maintained a high value on the test set (0.9152). All models demonstrated strong predictive performance, with R^2 values exceeding 0.8, which is above the acceptable threshold of 0.6. The root mean square error (RMSE) values for the MLR model on the training and test sets were 0.8707 and 0.8595, respectively. The SVM model yielded RMSEs of 0.8540 and 0.8314, while the BP-ANN model achieved values of 0.7672 and 0.6490. In contrast, the XGBoost model recorded RMSEs of 0.3608 and 0.6815. All models demonstrated relatively low RMSE values. Similarly, the mean absolute error (MAE) values for the MLR model were 0.6598 and 0.6994 for the training and test sets, respectively. The SVM model yielded MAEs of 0.6356 and 0.6545, while the BP-ANN model recorded values of 0.4960 and 0.5069. The XGBoost model achieved the lowest MAE on the training set (0.1302), with a test set MAE of 0.4645. All models demonstrated low prediction errors, with MAE values consistently below 1. The collective evidence from the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) indicates that all four models possess satisfactory fitting performance. The internal (Q^2_{loo}) and external (Q^2_{ext}) validation coefficients for the models were as follows: MLR (0.7891, 0.7785), SVM (0.7971, 0.7928), BP-ANN (0.8363, 0.8737), and XGBoost (0.9532, 0.9291). All values clustered well above 0.75, surpassing the accepted threshold of 0.6, which collectively confirms the models' robust predictive performance.

A comparative analysis of the performance metrics across the four models reveals that all constructed models exhibit satisfactory fitting capability and predictive accuracy. Although the performance indicators of all models meet the predefined evaluation criteria, the XGBoost nonlinear model consistently outperforms the MLR linear model as well as the SVM and BP-ANN nonlinear models in predicting the STEL of hydrocarbons and their derivatives. Specifically, the SVM and BP-ANN models show marginally better

performance than the MLR model in most parameters. These results suggest a strong nonlinear relationship between the STEL values of hydrocarbons and their derivatives and their molecular structures, with the XGBoost model demonstrating the best overall performance. The superior performance of XGBoost can be attributed to its inherent algorithmic strengths in handling complex structure–toxicity relationships. Unlike MLR, which is confined to linear associations, XGBoost is an ensemble method that sequentially builds decision trees, with each tree correcting the residuals of its predecessors. This boosting mechanism, combined with built-in regularization, enables it to capture intricate nonlinearities and, most importantly, complex descriptor interactions while minimizing overfitting—a crucial advantage for small datasets [49]. The earlier observed discrepancy between univariate correlation and multivariate regression importance rankings is a clear indicator of such underlying interactions, which the XGBoost model is uniquely positioned to leverage. To further investigate the robustness of the developed models, an in-depth error analysis was conducted.

4.2.2. Comparative Error Analysis

Figure 9 displays the residual plots for the four models. In all cases, the residuals of all samples are randomly distributed within the range of (−3, 3), with the majority of data points evenly scattered on both sides of the baseline. Only a small number of points deviate relatively far from the baseline, and most residuals exhibit only minor fluctuations. This random distribution without systematic patterns indicates the absence of systematic error and suggests good model stability. Furthermore, it is evident that the XGBoost model adheres more closely to the baseline compared to the MLR, SVM, and BP-ANN models, showing the smallest residual fluctuations and the least deviation, which reflects its superior performance. Overall, both the internal validation coefficients and the residual plots demonstrate that all models possess strong robustness. Nevertheless, the XGBoost model outperforms the others, with more data points lying near the baseline and the smallest residual values, confirming its optimal predictive capability.

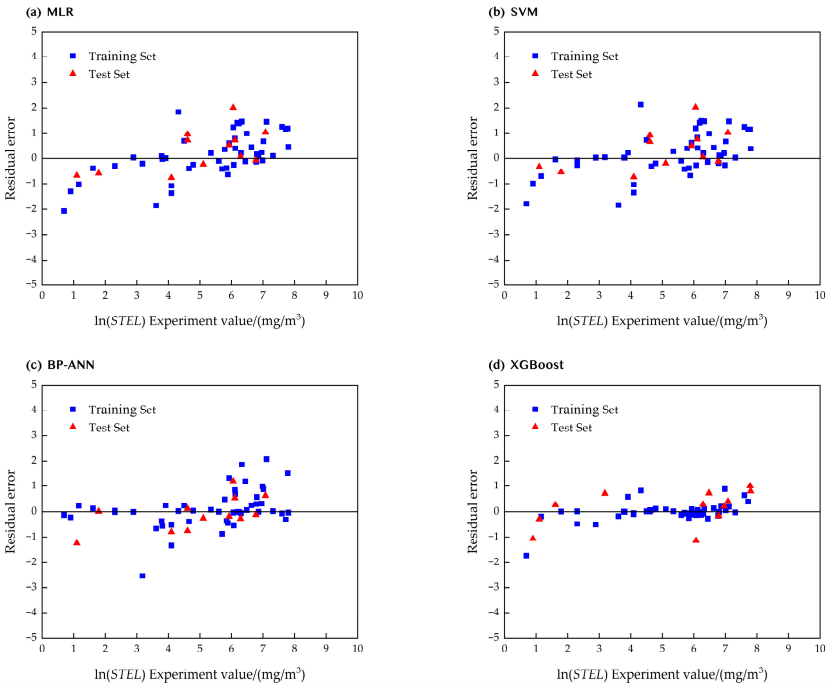


Figure 9. Residual plots comparing prediction errors for (**a**) MLR, (**b**) SVM, (**c**) BP-ANN, and (**d**) XGBoost models.

For a more in-depth error analysis, the error distributions of both the training and test sets were examined. Regarding the training set, the error distribution of the MLR model was relatively dispersed, with a peak count of 14 and a notable number of samples falling outside the [−1, 1] interval, indicating limited fitting accuracy. The SVM model exhibited a more concentrated error distribution, with a peak count of 15 and most errors clustered within [−1, 1], though a small number of samples still lay outside this range, reflecting an improvement over MLR. The BP-ANN model showed further concentration in its error distribution, reaching a peak count of 18, with the vast majority of sample errors confined to the [−1, 1] interval, demonstrating even better fitting accuracy. In contrast, the XGBoost model displayed the most concentrated error distribution, achieving a peak count of 24 and having nearly all errors located within [−1, 1], with very few outliers. This result signifies that XGBoost offers significantly superior fitting accuracy on the training set compared to the MLR, SVM, and BP-ANN models, highlighting its enhanced capability in capturing training data patterns. This error distribution is visually summarized in Figure 10.

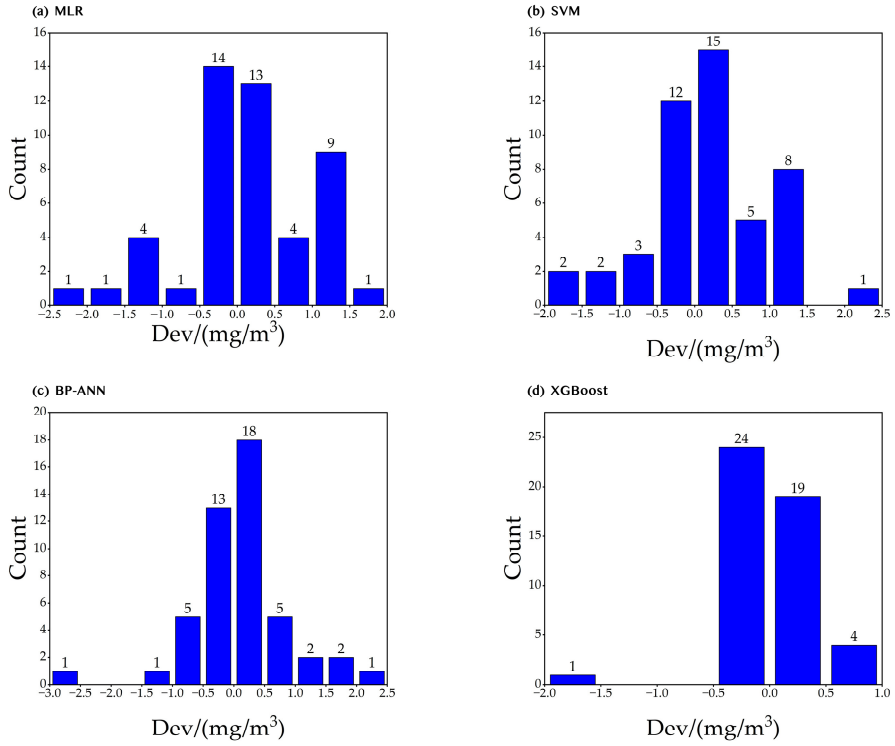


Figure 10. Distribution of training set prediction errors for (**a**) MLR, (**b**) SVM, (**c**) BP-ANN, and (**d**) XGBoost models.

Regarding the test set error distribution, the MLR model exhibited a relatively dispersed error spread within the range of 1.0 to 1.5, without a distinct concentration trend. The SVM model showed a moderate degree of error concentration, though the overall dispersion remained considerable. In comparison, the BP-ANN model demonstrated improved error concentration, with most errors lying between 0.5 and 1.0. The XGBoost model displayed the most concentrated error distribution, with the majority of errors clustered near 0.0 and 0.5, and almost no samples in the outer intervals. This indicates a superior predictive accuracy and error control capability on the test set, markedly outperforming the MLR, SVM, and BP-ANN models. The corresponding error distribution for the test set is presented in Figure 11. Based on the above analysis, it can be concluded that the XGBoost model exhibits the best performance in establishing the relationship between the STELs of benzene and its derivatives and their molecular structures.

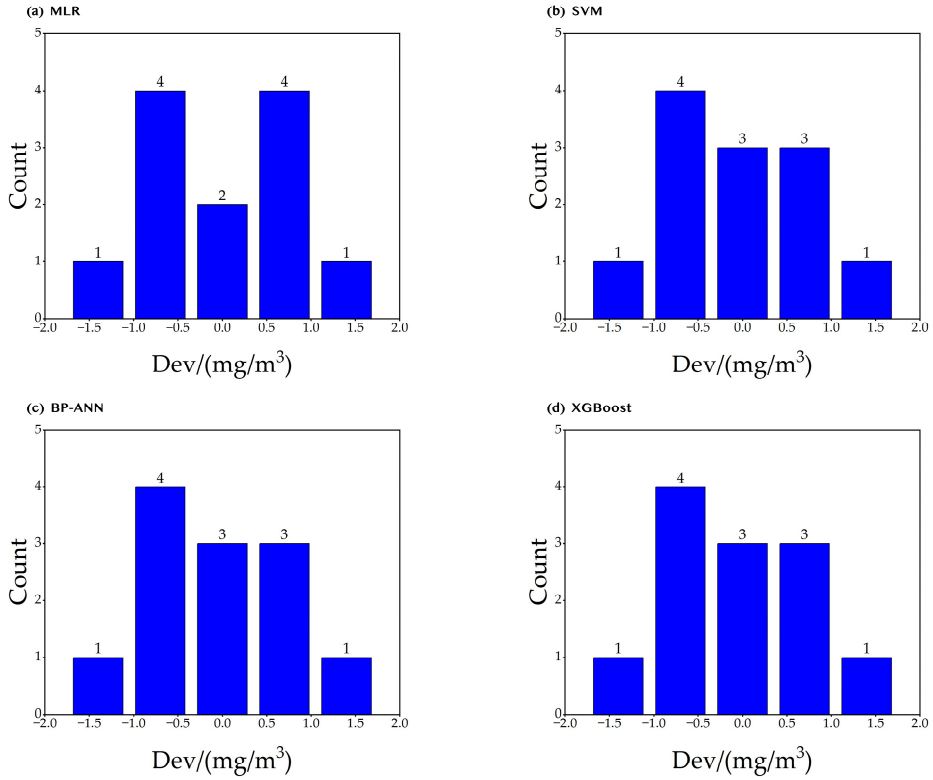


Figure 11. Distribution of test set prediction errors for (**a**) MLR, (**b**) SVM, (**c**) BP-ANN, and (**d**) XGBoost models.

4.2.3. Applicability Domain Evaluation

To evaluate the applicability of the four models, Williams plots were employed to validate model effectiveness. In these plots, the horizontal axis represents the leverage value of each sample, while the vertical axis indicates the standardized residuals. Samples were assessed based on whether they fell within or on the boundaries of the rectangular region defined by ±3 standardized residuals and the critical leverage value (h*). Potential outliers identified through this method were subsequently analyzed for further investigation [17]. The Williams plots for the MLR, SVM, BP-ANN, and XGBoost models are presented in Figure 12.

As illustrated in Figure 12, approximately 96% of the samples from both the training and test sets fall within the rectangular region bounded by standardized residuals of −3 to +3 and the critical leverage (h* = 0.3125). This distribution indicates that the four developed models possess a broad applicability domain, coupled with robust predictive capability and generalizability. Furthermore, these results validate the reliability of the models and the representativeness of the database employed in this study.

Despite their broad applicability domains, all four models exhibited a limited number of outliers, as summarized in Table 6. Di(2-ethylhexyl) phthalate was identified in the MLR, SVM, BP-ANN, and XGBoost models as lying within the ±3 standardized residual range but outside the critical leverage value (h*). This pattern suggests that certain specific structural features of this compound may not be fully captured by the selected molecular descriptors, leading to predictions based on model extrapolation. According to previous studies, such data points can be regarded as "benign outliers" [50,51], which contribute positively to the stability and robustness of QSAR models. In addition, isopropylamine in the BP-ANN model and oxalic acid in the XGBoost model were detected as outliers outside the ±3 standardized residual range yet within the critical leverage threshold (h*). These outliers may originate from biased experimental data, potentially rendering their

predictions less reliable. It is important to note that minor deviations in experimental measurements are inevitable due to various interfering factors; thus, the presence of very few such instances remains acceptable. Overall, these findings confirm that the developed models exhibit wide applicability and are statistically reliable for practical use.

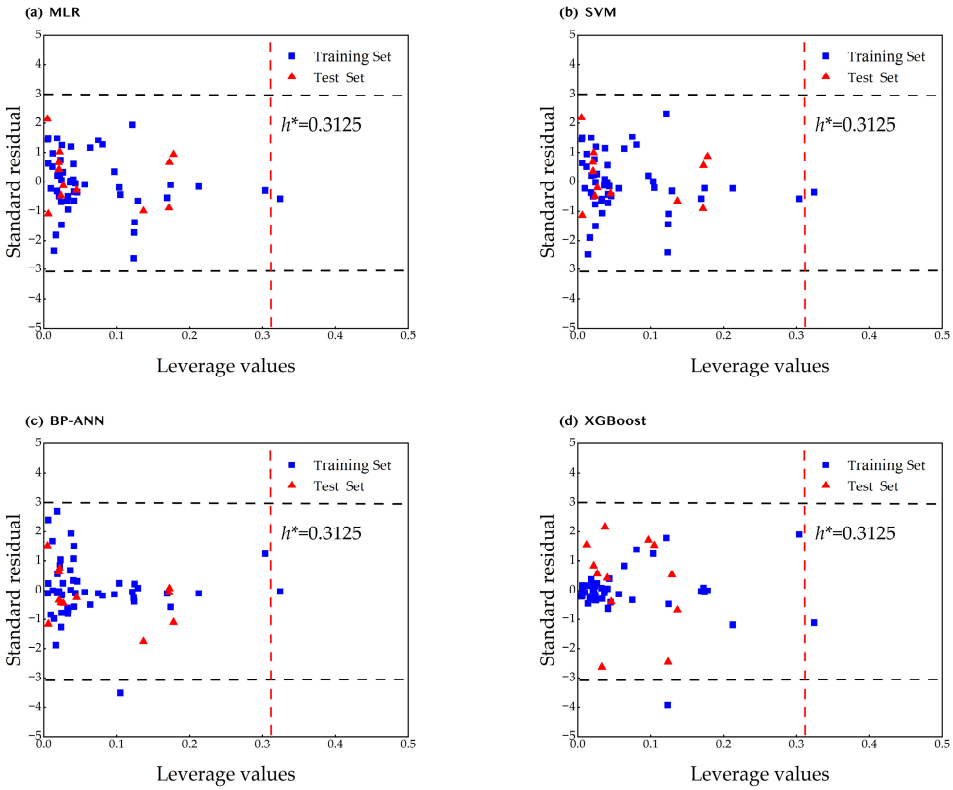
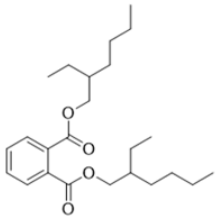
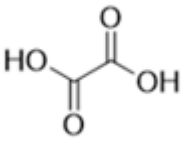


Figure 12. Williams plots defining the applicability domain for (**a**) MLR, (**b**) SVM, (**c**) BP-ANN, and (**d**) XGBoost models.

Table 6. Identified outliers and their molecular structures.

CAS No.:	Name:	Experimental Value ln(STEL):	Molecular Structure
Sample with leverage values above the critical threshold (h*)			
117-81-7	di(2-ethylhexyl) phthalate	2.3	
Samples outside the ±3 standardized residual range			
75-31-0	isopropylamine	3.17	
144-62-7	oxalic acid	0.69	

To further contextualize the identified outliers within the modeling framework, it is pertinent to note their distribution relative to the training and test sets. All three outlier compounds—di(2-ethylhexyl) phthalate (a consistent high-leverage point), isopropylamine, and oxalic acid (residual outliers in specific models)—were exclusively contained within the training set. The presence of di(2-ethylhexyl) phthalate as a high-leverage point delineates the boundary of the chemical space covered during model training, and its inclusion likely contributed to the model's robustness by mitigating overfitting. The residual outliers, conversely, highlight specific structural features for which the global descriptor set and model form could not achieve optimal fitting. The paramount finding, however, is that the external test set, constructed via cluster-based splitting, contained no outliers. This result robustly affirms that the applicability domain derived from the training data is comprehensive and representative. Consequently, for new molecules structurally analogous to the training set outliers, the model would provide a reliability warning via high leverage; for the vast majority of compounds within the domain, as evidenced by the test set performance, the model delivers predictions of high reliability.

4.2.4. Interpretability Analysis of the Optimal XGBoost Model Using SHAP

While three nonlinear models (SVM, BP-ANN, and XGBoost) were developed in this study, the XGBoost model was conclusively identified as the superior predictor based on a comprehensive evaluation of its performance (as detailed in Sections 4.2.1–4.2.3). To reliably interpret the structure–activity relationships captured by this optimal model, we deconstructed its decision-making process using SHapley Additive exPlanations (SHAP) and Partial Dependence Plots (PDPs). These techniques quantify feature contributions and elucidate their nonlinear influences on the predicted ln(STEL).

The SHAP summary plot (Figure 13E) reveals that *E3u* serves as the core driving feature, followed by *Mor28m* and *n* = *CHR*, while *G2m* exhibits the weakest contribution. The SHAP values for all four descriptors span a wide range (approximately −3 to +3), underscoring their significant and diverse impacts on the model output. Notably, *E3u* and *n* = *CHR* display the most dispersed SHAP value distributions, indicating their dominant role in pushing predictions toward both high and low extremes and highlighting their pronounced nonlinear characteristics. The Partial Dependence Plots (PDP; Figure 13A–D) further delineate the marginal dependence between each descriptor and the predicted ln(STEL). Both *E3u* and *G2m* show a strong positive correlation with the target; increasing their values directly elevates the ln(STEL) prediction, consistent with the high-feature-value, high-SHAP-value pattern observed in the SHAP analysis. In contrast, *Mor28m* and *n* = *CHR* exhibit more complex, negative-influence trends. The PDP for *Mor28m* (Figure 13C) decreases rapidly before plateauing, indicating that in the low-to-medium range, increasing this descriptor significantly lowers the predicted value. This non-monotonic behavior explains its broad SHAP value distribution (−3 to +3), demonstrating its capacity to either strongly suppress or mildly enhance the prediction. The analysis of *n* = *CHR* is particularly insightful. As a count descriptor for sp^2-hybridized secondary carbon atoms, its PDP (Figure 13D) shows a distinct decreasing trend in the predicted ln(STEL) as its value increases. This finding aligns with the negative standardized coefficient (−0.339) for *n* = *CHR* in the MLR model, jointly providing robust evidence that this structural feature is consistently associated with lower short-term exposure limits. Mechanistically, an increase in such unsaturated carbons may reduce acute inhalation toxicity by influencing molecular electron distribution, chemical reactivity, or metabolic pathways.

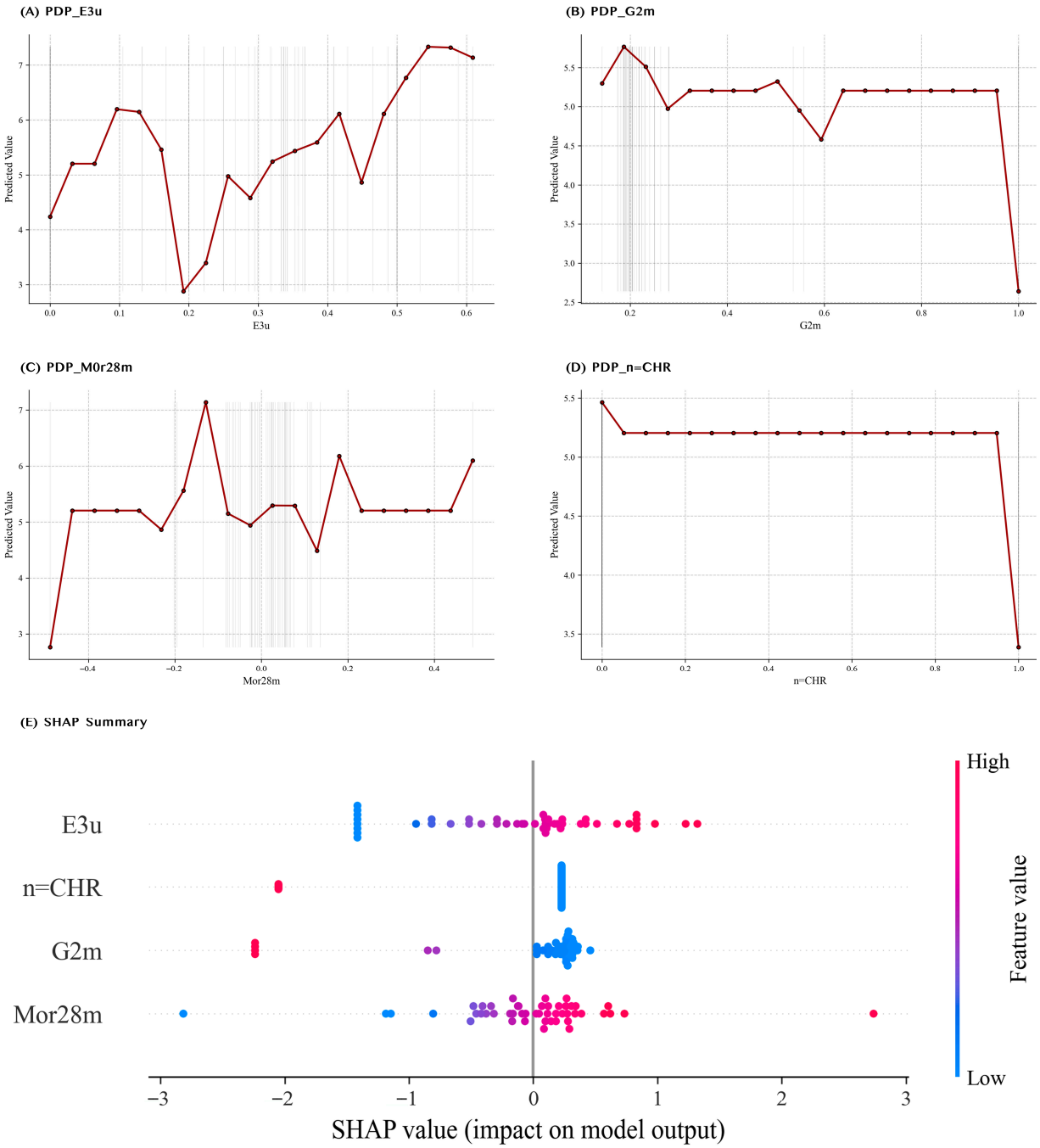


Figure 13. Analysis of Feature Impacts on Model Predictions. (**A**–**D**). Partial dependence plots (PDPs) for key features (*E3u*, *G2m*, *Mor28m*, *n = CHR*), showing the marginal effect of each feature on the model's predicted output. (**E**). SHAP summary plot illustrating the global importance and direction of impact of each feature on the model output. The point color from red to blue represents the feature value from high to low.

Collectively, these results demonstrate that the XGBoost model successfully captures complex nonlinear and interactive descriptor effects that traditional linear models like MLR cannot adequately reveal. The combined application of SHAP and PDP not only enhances model interpretability but also offers novel mechanistic perspectives on the short-term exposure toxicity of hydrocarbons and their derivatives.

5. Conclusions

This study developed and validated a suite of QSAR models for the rapid prediction of STELs of hydrocarbons and their derivatives. Through a systematic benchmarking approach, the extreme gradient boosting (XGBoost) model was identified as the superior predictor, demonstrating exceptional fitting accuracy (R^2_{train} = 0.9445) and generalization capability (Q^2_{ext} = 0.9291). The analysis conclusively revealed a strong nonlinear relationship between STELs and the molecular descriptors, primarily related to molecular size, spatial configuration, and unsaturation. Crucially, by employing SHAP analysis, we transcended mere numerical prediction and elucidated the specific direction and magnitude of each descriptor's influence. For instance, the negative association between the *n = CHR* descriptor and STEL was robustly confirmed, suggesting a potential role of molecular unsaturation in mitigating acute inhalation toxicity. This provides a mechanistically transparent and interpretable modeling framework. Importantly, the comparative analysis between univariate correlations and multivariate model contributions revealed significant descriptor interdependencies, underscoring the necessity of advanced machine learning techniques over simpler modeling approaches for this task. The comprehensive applicability domain analysis, which confirmed that no outliers in the external test set, further underscores the model's reliability for practical predictions on new, structurally analogous compounds. With 96% of predictions residing within the well-defined applicability domain, the proposed XGBoost model provides a reliable and cost-effective computational tool for occupational health risk assessment, especially for data-poor chemicals. This work provides a valuable theoretical foundation and a robust predictive strategy to support the mitigation of health risks associated with short-term chemical exposure in the workplace. Future work will focus on expanding the chemical diversity and size of the dataset by integrating data from additional public toxicological and occupational exposure databases. This will broaden the model's applicability domain and, crucially, enable the development and evaluation of more complex model architectures, such as sophisticated hybrid ensembles and deep learning networks, which require larger data volumes to realize their full potential.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/pr13124025/s1, Table S1: List of studied chemicals, descriptor values, experimental and predicted ln(STEL).

Author Contributions: Conceptualization, J.S. and X.Y.; methodology, J.S.; software, L.N.; validation, W.Z.; formal analysis, J.S. and W.Z.; investigation, J.S. and C.W.; writing—original draft preparation, J.S. and W.Z.; writing—review and editing, J.S. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: Assistance in conducting the experimental and instrumental analyses of this study was provided by Changzhou University, Changzhou, Jiangsu Province, China.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kamal, A.; Malik, R.N.; Fatima, N.; Rashid, A. Chemical exposure in occupational settings and related health risks: A neglected area of research in Pakistan. *Environ. Toxicol. Pharmacol.* **2012**, *34*, 46–58. [CrossRef]
2. Barros, B.; Oliveira, M.; Morais, S. Firefighters' occupational exposure: Contribution from biomarkers of effect to assess health risks. *Environ. Int.* **2021**, *156*, 106704. [CrossRef]

3. Pyatt, J.R.; Gilmore, I.; Mullins, P.A. Abnormal liver function tests following inadvertent inhalation of volatile hydrocarbons. *Postgrad. Med. J.* **1998**, *74*, 747–748. [CrossRef] [PubMed]
4. Boffetta, P.; Jourenkova, N.; Gustavsson, P. Cancer risk from occupational and environmental exposure to polycyclic aromatic hydrocarbons. *Cancer Causes Control* **1997**, *8*, 444–472. [CrossRef] [PubMed]
5. Yu, Y.; Li, Q.; Wang, H.; Wang, B.; Wang, X.; Ren, A.; Tao, S. Risk of human exposure to polycyclic aromatic hydrocarbons: A case study in Beijing, China. *Environ. Pollut.* **2015**, *205*, 70–77. [CrossRef] [PubMed]
6. U.S. Department of Labor, Occupational Safety and Health Administration (OSHA). Air Contaminants. Code of Federal Regulations. 2023. Available online: https://www.ecfr.gov/current/title-29/subtitle-B/chapter-XVII/part-1910/subpart-Z/section-1910.1000 (accessed on 10 October 2025).
7. Song, X. Study on Environmental Control After Chemical Leakage During Storage and Transportation. Ph.D. Thesis, Shanghai Ocean University, Shanghai, China, 2019. [CrossRef]
8. Zheng, S.; Hu, C.; Huang, S.; Xiao, X.; Luo, L. Occupational Exposure Limits of Air Toxic Substances in the GESTIS Substance Database: Current Status. *Chin. J. Ind. Hyg. Occup. Dis.* **2024**, *42*, 417–425. [CrossRef]
9. Ismail, A.U.; Ibrahim, S.A.; Gambo, M.D.; Muhammad, R.F.; Badamasi, M.M.; Sulaiman, I. Impact of differential occupational LPG exposure on cardiopulmonary indices, liver function, and oxidative stress in Northwestern city of Nigeria. *Sci. Total Environ.* **2023**, *862*, 160881. [CrossRef] [PubMed]
10. Wu, S.; Zhang, Y.; Ma, Y.; Zhang, M.; Huang, D. Carcinogenic Risk Assessment of Occupational Exposure to Low-Level Benzene in a Large-Scale Aromatic Plant. *Chin. J. Occup. Med.* **2015**, *42*, 205–207. [CrossRef]
11. Jiang, X.; Wang, R.; Chang, T.; Zhang, Y.; Zheng, K.; Wan, R.; Wang, X. Effect of short-term air pollution exposure on migraine: A protocol for systematic review and meta-analysis on human observational studies. *Environ. Int.* **2023**, *174*, 107892. [CrossRef]
12. Nunes, L.J.R.; Curado, A. Long-term vs. short-term measurements in indoor Rn concentration monitoring: Establishing a procedure for assessing exposure potential (RnEP). *Results Eng.* **2023**, *17*, 100966. [CrossRef]
13. Zhou, Y.; Xu, C.; Zhang, Y.; Zhao, M.; Hu, Y.; Jiang, Y.; Li, D.; Wu, N.; Wu, L.; Li, C.; et al. Association between short-term nitrogen dioxide exposure and outpatient visits for anxiety: A time-series study in Xi'an, China. *Atmos. Environ.* **2022**, *279*, 119122. [CrossRef]
14. Li, S.-S.; Fang, S.-M.; Chen, J.; Zhang, Z.; Yu, Q.-Y. Effects of short-term exposure to volatile pesticide dichlorvos on the olfactory systems in Spodoptera litura: Calcium homeostasis, synaptic plasticity and apoptosis. *Sci. Total Environ.* **2023**, *864*, 161050. [CrossRef]
15. Liu, Y.; Xu, J.; Shi, J.; Zhang, Y.; Ma, Y.; Zhang, Q.; Su, Z.; Zhang, Y.; Hong, S.; Hu, G.; et al. Effects of short-term high-concentration exposure to PM(2.5) on pulmonary tissue damage and repair ability as well as innate immune events. *Environ. Pollut.* **2023**, *319*, 121055. [CrossRef]
16. Dong, Z.; Li, X.; Chen, Y.; Zhang, N.; Wang, Z.; Liang, Y.-Q.; Guo, Y. Short-term exposure to norethisterone affected swimming behavior and antioxidant enzyme activity of medaka larvae, and led to masculinization in the adult population. *Chemosphere* **2023**, *310*, 136844. [CrossRef]
17. Balasch, J.C.; Brandts, I.; Barria, C.; Martins, M.A.; Tvarijonaviciute, A.; Tort, L.; Oliveira, M.; Teles, M. Short-term exposure to polymethylmethacrylate nanoplastics alters muscle antioxidant response, development and growth in Sparus aurata. *Mar. Pollut. Bull.* **2021**, *172*, 112918. [CrossRef]
18. Pengelly, I.; O'Shea, H.; Smith, G.; Coggins, M.A. Measurement of Diacetyl and 2,3-Pentanedione in the Coffee Industry Using Thermal Desorption Tubes and Gas Chromatography-Mass Spectrometry. *Ann. Work Expo. Health* **2019**, *63*, 415–425. [CrossRef]
19. North, C.M.; Rooseboom, M.; Kocabas, N.A.; Synhaeve, N.; Radcliffe, R.J.; Segal, L. Application of physiologically-based pharmacokinetic modeled toluene blood concentration in the assessment of short term exposure limits. *Regul. Toxicol. Pharmacol.* **2023**, *140*, 105380. [CrossRef] [PubMed]
20. Russell, A.J.; Vincent, M.; Buerger, A.N.; Dotson, S.; Lotter, J.; Maier, A. Establishing short-term occupational exposure limits (STELs) for sensory irritants using predictive and in silico respiratory rate depression (RD50) models. *Inhal. Toxicol.* **2024**, *36*, 13–25. [CrossRef] [PubMed]
21. Du, K.-L.; Jiang, B.; Lu, J.; Hua, J.; Swamy, M.N.S. Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions. *Mathematics* **2024**, *12*, 3935. [CrossRef]
22. Jung, T.; Kim, J. A new support vector machine for categorical features. *Expert Syst. Appl.* **2023**, *229*, 120449. [CrossRef]
23. Baldomero-Naranjo, M.; Martínez-Merino, L.I.; Rodríguez-Chía, A.M. A robust SVM-based approach with feature selection and outliers detection for classification problems. *Expert Syst. Appl.* **2021**, *178*, 115017. [CrossRef]
24. Wu, F.; Zhang, X.; Fang, Z.; Yu, X. Support Vector Machine-Based Global Classification Model of the Toxicity of Organic Compounds to Vibrio fischeri. *Molecules* **2023**, *28*, 2703. [CrossRef]

25. Yao, X.J.; Panaye, A.; Doucet, J.P.; Zhang, R.S.; Chen, H.F.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1257–1266. [CrossRef]
26. Sengupta, A.; Singh, S.K.; Kumar, R. Support Vector Machine-Based Prediction Models for Drug Repurposing and Designing Novel Drugs for Colorectal Cancer. *ACS Omega* **2024**, *9*, 18584–18592. [CrossRef] [PubMed]
27. Ibrahim, S.I.; Ghoneim, S.S.M.; Taha, I.B.M. DGALab: An extensible software implementation for DGA. *IET Gener. Transm. Distrib.* **2018**, *12*, 4117–4124. [CrossRef]
28. Li, Z.; Huang, J.; Wang, J.; Ding, M. Comparative study of meta-heuristic algorithms for reactor fuel reloading optimization based on the developed BP-ANN calculation method. *Ann. Nucl. Energy* **2022**, *165*, 108685. [CrossRef]
29. Wei, Z.; Su, X.; Wang, D.; Feng, Z.; Gao, Q.; Xu, G.; Zu, G. Three-dimensional processing map based on BP-ANN and interface microstructure of Fe/Al laminated sheet. *Mater. Chem. Phys.* **2023**, *297*, 127431. [CrossRef]
30. Guo, Z.; Guo, C.; Sun, L.; Zuo, M.; Chen, Q.; El-Seedi, H.R.; Zou, X. Identification of the apple spoilage causative fungi and prediction of the spoilage degree using electronic nose. *J. Food Process Eng.* **2021**, *44*, e13816. [CrossRef]
31. Chen, L.-S.; Chung, W.-H.; Chen, Y.; Kuo, S.-Y.J.I.A. AMC with a BP-ANN scheme for 5G enhanced mobile broadband. *IEEE Access* **2020**, *13*, 124689–124696. [CrossRef]
32. Bauer, E.; Kohavi, R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Mach. Learn.* **1999**, *36*, 105–139. [CrossRef]
33. The National Institute for Occupational Safety and Health (NIOSH). *NIOSH Pocket Guide to Chemical Hazards*; Centers for Disease Control and Prevention: Atlanta, GA, USA, 2020. Available online: https://www.cdc.gov/niosh/npg/npgdcas.html (accessed on 4 October 2024).
34. Frey, B.J.; Dueck, D. Response to Comment on "Clustering by Passing Messages Between Data Points". *Science* **2008**, *319*, 726. [CrossRef]
35. Gramatica, P.; Pilutti, P.; Papa, E. Approaches for externally validated QSAR modelling of Nitrated Polycyclic Aromatic Hydrocarbon mutagenicity. *SAR QSAR Environ. Res* **2007**, *18*, 169–178. [CrossRef]
36. Yang, C.; Chen, J.; Wang, R.; Zhang, M.; Zhang, C.; Liu, J. Density Prediction Models for Energetic Compounds Merely Using Molecular Topology. *J. Chem. Inf. Model.* **2021**, *61*, 2582–2593. [CrossRef]
37. Tropsha, A.; Gramatica, P.; Gombar, V.K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69–77. [CrossRef]
38. Camellini, F.; Crisci, S.; De Magistris, A.; Franchini, G. A line-search based SGD algorithm with Adaptive Importance Sampling. *J. Comput. Appl. Math.* **2026**, *477*, 117120. [CrossRef]
39. Todeschini, R.; Gramatica, P. The Whim Theory: New 3D Molecular Descriptors for Qsar in Environmental Modelling. *SAR QSAR Environ. Res.* **1997**, *7*, 89–115. [CrossRef]
40. Salmina, E.; Haider, N.; Tetko, I. Extended Functional Groups (EFG): An Efficient Set for Chemical Characterization and Structure-Activity Relationship Studies of Chemical Compounds. *Molecules* **2015**, *21*, 1. [CrossRef] [PubMed]
41. Hair, J.F. Multivariate Data Analysis: An Overview. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 904–907.
42. Peng, F.; Lu, Y.; Wang, Y.; Yang, L.; Yang, Z.; Li, H. Predicting the formation of disinfection by-products using multiple linear and machine learning regression. *J. Environ. Chem. Eng.* **2023**, *11*, 110612. [CrossRef]
43. Dufera, A.G.; Liu, T.; Xu, J. Regression models of Pearson correlation coefficient. *Stat. Theory Relat. Fields* **2023**, *7*, 97–106. [CrossRef]
44. Phalen, R.F. *Inhalation Studies: Foundations and Techniques*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2002.
45. Dahl, A.R.; Gerde, P. Uptake and metabolism of toxicants in the respiratory tract. *Environ. Health Perspect.* **1994**, *102*, 67–70. [CrossRef] [PubMed]
46. Enoch, S.J.; Ellison, C.M.; Schultz, T.W.; Cronin, M.T.D. A review of the electrophilic reaction chemistry involved in covalent protein binding relevant to toxicity. *Crit. Rev. Toxicol.* **2011**, *41*, 783–802. [CrossRef]
47. Mansouri, K.; Grulke, C.M.; Judson, R.S.; Williams, A.J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *J. Cheminformatics* **2018**, *10*, 10. [CrossRef] [PubMed]
48. Barber, C.; Heghes, C.; Johnston, L. A framework to support the application of the OECD guidance documents on (Q)SAR model validation and prediction assessment for regulatory decisions. *Comput. Toxicol.* **2024**, *30*, 100305. [CrossRef]
49. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]

50. Nahum, O.E.; Yosipof, A.; Senderowitz, H. A Multi-Objective Genetic Algorithm for Outlier Removal. *J. Chem. Inf. Model.* **2015**, *55*, 2507–2518. [CrossRef] [PubMed]
51. Netzeva, T.I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M.T.; Gramatica, P.; Jaworska, J.S.; Kahn, S.; Klopman, G.; Marchant, C.A.; et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* **2005**, *33*, 155–173. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.