

Article

Superior Performance of Extreme Gradient Boosting Model Combined with Affinity Propagation Clustering for Reliable Prediction of Permissible Exposure Limits of Hydrocarbons and Their Oxygen-Containing Derivatives

Jingjie Shi, Zixiang Zhang, Yongde Wei, Wei Zhao and Xiongjun Yuan *

School of Safety Science and Engineering, Changzhou University, Changzhou 213000, China; shijingjie@cczu.edu.cn (J.S.); zhangzixiangcczu@163.com (Z.Z.); 19810619920@163.com (Y.W.); zhaoweicczu@163.com (W.Z.)

* Correspondence: yxj@cczu.edu.cn

Abstract

In order to conveniently and efficiently determine the Permissible Exposure Limits (PELs) of organic chemicals in the workplace, this study employed Quantitative Structure–Activity Relationship (QSAR) modeling to predict properties related to occupational health and safety. The predictive study was conducted by correlating the PELs of 75 hydrocarbons and their oxygen-containing derivatives with the molecular structures of the organic compounds. Meanwhile, this study conducted a comprehensive and in-depth comparative analysis of the four developed predictive models. The sample set was partitioned using the Affinity Propagation (AP) clustering algorithm. Four characteristic molecular descriptors were selected by integrating the Genetic Algorithm (GA) with the variance inflation factor (VIF) value. Subsequently, the Multiple Linear Regression (MLR) model and two nonlinear models, namely the Support Vector Machine (SVM) and the Extreme Gradient Boosting (XGBoost), were developed and used for predictive comparison. Furthermore, the performance of the models was evaluated through both internal and external validation methods, and the Williams plots were constructed to define the model's applicability domain. The results indicated that the XGBoost model achieved high performance, with a coefficient of determination (R^2) of 0.9962 on the training set and 0.8892 on the testing set. The corresponding root mean square errors (RMSE) were 0.1012 and 0.6623 for the training and testing sets, respectively. The internal validation coefficient (Q^2_{loo}) was 0.8975, while the external validation coefficient (Q^2_{ext}) was 0.832. Moreover, the majority of the sample data (approximately 96%) fell within the application domain defined by ± 3 times the standard residue-to-critical arm ratio, where $h^* = 0.2$. This demonstrates that the XGBoost model exhibits excellent fitting capability, stability, and predictive power, thereby uncovering a significant nonlinear relationship between the molecular structure of compounds and the PELs. As outlined above, the utilization of the QSAR method for predicting the PELs of hydrocarbons and their oxygen-containing derivatives constitutes a highly effective approach.



Academic Editor: Nicola Magnavita

Received: 14 September 2025

Revised: 25 October 2025

Accepted: 29 October 2025

Published: 31 October 2025

Citation: Shi, J.; Zhang, Z.; Wei, Y.; Zhao, W.; Yuan, X. Superior Performance of Extreme Gradient Boosting Model Combined with Affinity Propagation Clustering for Reliable Prediction of Permissible Exposure Limits of Hydrocarbons and Their Oxygen-Containing Derivatives. *Appl. Sci.* **2025**, *15*, 11642. <https://doi.org/10.3390/app152111642>

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/>).

Keywords: quantitative structure–activity relationship; permissible exposure limit; affinity propagation; extreme gradient boosting; support vector machine; hydrocarbons and oxygen-containing derivatives; molecular descriptors; Williams plot

1. Introduction

Oxygen-containing hydrocarbon derivatives represent a class of chemically reactive organic substances, with their molecular structures featuring oxygen-functional groups that predispose them to redox reactions [1–3]. These compounds, which include alcohols, phenols, aldehydes, carboxylic acids, and esters, exhibit considerable structural diversity and are widely employed in industrial applications as solvents, polymer precursors, disinfectants, and in the manufacture of resins, plastics, and pharmaceuticals [4–6]. However, their chemical instability, combined with inherent sensitivity and susceptibility to degradation by factors such as photolysis, hydrolysis, and oxidation, makes them significant contributors to occupational health hazards and safety incidents in environments involving hazardous chemicals [7–9]. For example, the photolysis of acetaldehyde results in the formation of peroxyacetyl nitrate (PAN), elevated levels of which can pose serious health risks to humans, including severe irritation to the eyes and respiratory tract, and the potential to cause systemic oxidative stress [10]. Polycyclic aromatic hydrocarbons (PAHs) and their oxygenated derivatives, such as anthraquinone, enter the human body via inhalation and dietary intake, accumulate in adipose tissue, and can induce DNA adduct formation and apoptosis, thereby posing significant adverse effects on the respiratory and reproductive systems [11]. Hou et al. emphasized Oxygenated polycyclic aromatic hydrocarbons (OPAHs), emerging contaminants in the environment, identifying key genes involved in OPAHs-induced IBD [12]. Despite the increasing adoption of improved occupational health and safety management systems by global enterprises through process optimization and the implementation of protective equipment [13–15], occupational health and safety remain a pressing and complex industrial challenge of global importance due to its multifaceted nature [16]. In particular, exposure to hazardous chemicals in industrial environments represents a prevalent and critical risk that necessitates prompt intervention. Oxygen-containing hydrocarbon derivatives serve as a representative example of such hazardous substances. Their volatility and reactivity may lead to the accumulation of adverse health effects among workers via inhalation or dermal exposure [17]. To address these risks, the U.S. Occupational Safety and Health Administration (OSHA) has established the *PELs* as a regulatory benchmark for occupational exposure [18]. The *PELs* defines the maximum allowable airborne concentration of hazardous substances to which workers can be exposed, based on an 8 h time-weighted average (TWA), either within a single workday or over a 40 h workweek, without incurring significant long-term adverse health effects [19,20].

Currently, a substantial body of research has investigated physical factors associated with occupational exposure in workplace environments. Examples of such studies include examinations of the effects of noise exposure *PELs* on auditory threshold levels [21], the impact of optical radiation exposure limits on retinal health [22], and the evaluation of permissible exposure limits for thermal stress in controlled laboratory settings [23]. Additionally, numerous studies have concentrated on incidents resulting from occupational exposure to hazardous chemicals in industrial and occupational settings [24,25]. In 2022, D. Szczesna et al. [26] conducted a comprehensive review of the existing literature concerning the health effects of occupational exposure to inhalation anesthetics and proposed maximum allowable concentration (MAC) values for the volatile anesthetic agents enflurane, desflurane, isoflurane, and sevoflurane. In 2021, Tustin Aaron W. et al. [27] conducted an investigation into three fatal cases associated with occupational exposure to nickel carbonyl, methyl bromide, or styrene through the development of a single-compartment pharmacokinetic (PK) model. The study included an analysis of biological monitoring data concerning acute chemical exposure in industrial settings, aiming to evaluate employer compliance with the Occupational Safety and Health Administration's (OSHA) *PELs* for airborne contaminants. In 2020, Ronald N. Kostoff et al. [28] introduced a streamlined

methodology aimed at enhancing regulatory exposure limits for combined toxic stressors. This approach enabled the normalization of various combinations of stressors by converting exposure doses into their corresponding Toxicity Reference Value (TRV) fractions, which were then divided by the total sum of all such fractions. Established by the U.S. Occupational Safety and Health Administration (OSHA) under the Occupational Safety and Health Act of 1970, the *PELs* serve a dual function: facilitating the evaluation of occupational hygiene conditions and enhancing the regulation and oversight of worker health and safety [29]. These functions highlighted the broad applicability and critical importance of *PELs* within the domain of occupational health [30]. Nevertheless, the existing *PELs* for hazardous chemicals are predominantly derived from animal studies that assess acute inhalation or oral toxicity [31–33]. These conventional experimental approaches are subject to significant limitations, such as substantial costs, prolonged timelines, and notable safety risks. Moreover, ethical concerns, particularly those aligned with the 3Rs principle in animal testing, impose additional constraints on their applicability [34]. Consequently, the *PELs* are insufficient to address the progressively changing requirements of regulatory frameworks and industrial practices [35]. With the ongoing development and introduction of new chemical substances, particularly within the class of hydrocarbons and their derivatives, there is an increasing demand for efficient and practical theoretical prediction methods to complement and overcome the limitations of traditional experimental approaches. Quantitative Structure–Activity Relationship (QSAR) technology utilizes mathematical modeling approaches to construct reliable predictive models by establishing correlations between compound activity data and molecular structures [36]. This feature makes it a highly valuable tool for predicting *PELs*, as supported by internationally accepted OECD guidelines and established QSAR principles [37].

To date, the application of QSAR in predicting the *PELs* remains largely unexplored. However, this work is part of a broader scientific trend towards computational methods for occupational exposure and safety assessment [38–41]. This study employed the QSAR approach to estimate the *PELs* of oxygen-containing hydrocarbon derivatives. The proposed methodology not only addressed the limitations inherent in experimental methods but also offered an alternative strategy for predicting the *PELs* for as-yet unregulated or newly synthesized chemical compounds within this broad class. Furthermore, the AP clustering algorithm was employed to partition the sample set of oxygen-containing hydrocarbon derivatives. This algorithm effectively classified the dataset into distinct clusters, ensuring maximal similarity among data points within the same cluster and maximal dissimilarity between clusters. This clustering approach markedly enhanced the rationality and scientific rigor of the sample division process. Subsequently, the *MLR* model was constructed to correlate the molecular structures of oxygen-containing hydrocarbon derivatives with their corresponding activity data values. In addition, the nonlinear models based on *SVM* and *XGBoost* were also developed. The primary objective of these models is to establish quantitative relationships between four molecular descriptors and the permissible exposure limits (*PELs*) of chemical compounds, with the descriptors serving as input variables and *PELs* as the target output for prediction. Among the modeling approaches, multiple linear regression (*MLR*) provides a clear and direct representation of this relationship through a linear equation, offering high interpretability. In contrast, nonlinear models—such as support vector machines (*SVM*) and extreme gradient boosting (*XGBoost*)—are not inherently opaque “black boxes”; rather, they utilize complex mathematical architectures—such as kernel functions in *SVM* and ensembles of decision trees in *XGBoost*—to capture intricate nonlinear patterns between the descriptors and *PELs*. Although their internal mechanisms are less intuitively transparent than those of *MLR*, these models derive highly accurate predictive rules through training and are capable of delivering reliable *PEL* predictions. Conceptually,

molecular descriptors can be regarded as “chemical fingerprints” of compounds, while the predictive models act as “translators” that decode the quantitative relationship between these fingerprints and their corresponding exposure limits. To ensure the reliability of the models, both internal and external validation methods were implemented to assess their predictive performance [37]. Furthermore, the Williams plot was utilized to delineate the applicability domain of the models [37], thereby improving the interpretability and generalizability of the predictive outcomes. In this study, a schematic illustration depicts the procedural steps involved in the development of the *MLR*, *SVM*, and *XGBoost* models, as shown in Figure 1.

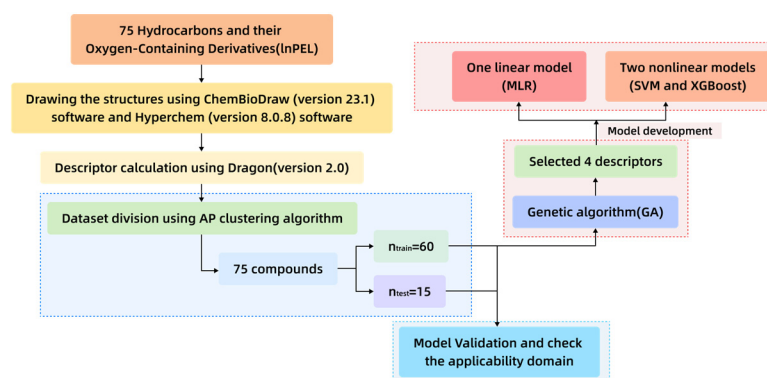


Figure 1. Schematic illustration of the steps involved in the development of QSAR models presented in this study.

2. Fundamental Principles

2.1. Affinity Propagation Clustering Algorithm

The AP clustering algorithm divides a dataset into distinct groups based on pairwise similarity measures derived from Euclidean distances [42]. By iteratively refining the information of attraction and belongingness among data points in the dataset, the proposed method maximizes the similarity of data objects within the same cluster while simultaneously enhancing the dissimilarity between objects in different clusters [43]. The AP algorithm has been widely applied across various domains, including power system evaluation [44], spatial zoning of seawater quality monitoring stations [45], and text clustering analysis [46]. During the clustering process, the algorithm initially treats each data point as a potential exemplar, or cluster center. Through iterative message passing, it progressively refines and determines the most appropriate exemplars and their associated cluster members. A cluster is generally considered valid by default if it contains at least one data point. Compared to conventional clustering algorithms such as K-Means and K-Medoids, the AP algorithm presents several distinct advantages. It does not necessitate the prior specification of the number of clusters, produces clear and interpretable exemplars as cluster centers, exhibits robustness to initial parameter settings, achieves high clustering accuracy and computational efficiency, and consistently generates stable and reproducible results [47]. Therefore, this study utilized the AP algorithm for the initial classification of the sample set. The sample set, sourced from the NIOSH Pocket Guide to Chemical Hazards [48], was partitioned using the Affinity Propagation (AP) clustering algorithm.

Step1: The algorithm initialization involves computing the initial similarity matrix. The Euclidean distance is employed to compute the similarity values $s(i, k)$ among N data points, as expressed in Equation (1). The resulting values are then organized and stored in the similarity matrix S .

$$\begin{cases} s(i, k) = -\|x_i - x_k\| (i \neq k) \\ s(i, k) = P(k) (i = k) \end{cases} \quad (1)$$

$s(i, k)$ denote the similarity between data points i and k . The p -value, which indicates the reference degree, corresponds to the diagonal elements of the similarity matrix, where the row and column indices are the same. It is of crucial importance in determining the appropriate number of clusters.

Step2: Configure the essential parameters, including the reference degree (p -value), damping factor (λ), and maximum iteration count. Subsequently, compute the responsibility values (r) and availability values (a), as formally defined in Equations (2) to (4).

$$r(i, k) = s(i, k) - \max\{a(i, k') + S(i, k')\} (k' \neq k) \quad (2)$$

$$a(i, k) = \min\{0, r(k, k)\} + \sum \{\max\{0, r(i', k)\}\} (i' \neq i, k' \neq k) \quad (3)$$

$$a(k, k) = \sum \max\{0, r(i', k)\} \quad (4)$$

$r(i, k)$ denotes the extent to which data point k is considered suitable as a class representative for data point i . Meanwhile, $a(i, k)$ denotes the membership degree of data point i to the class representative k .

Step3: The attraction value r and membership value a are iteratively updated to determine the high-quality clustering center k , as formally defined in Equations (5) to (7).

$$r_{new}(i, k) = \lambda \times r_{old}(i, k) + (1 - \lambda) \times r(i, k) \quad (5)$$

$$a_{new}(i, k) = \lambda \times a_{old}(i, k) + (1 - \lambda) \times a(i, k) \quad (6)$$

$$k = \operatorname{argmax}\{a(i, k) + r(i, k)\} \quad (7)$$

$r_{new}(i, k)$ and $a_{new}(i, k)$ represent the updated attraction value and membership value, respectively. λ represents the damping coefficient, which governs the convergence speed and iterative stability of the algorithm. k denotes the cluster center. Specifically, if i is equal to k , then i itself serves as a cluster center; otherwise, k represents the cluster center associated with data point i . The iterative process concludes when either the maximum number of iterations is attained or the cluster centers stabilize, at which point the algorithm proceeds back to Step 2.

2.2. Extreme Gradient Boosting

As a highly efficient and parallelizable ensemble learning algorithm, *XGBoost* is capable of accurately and efficiently processing high-dimensional and large-scale datasets, while maintaining wide applicability across diverse application domains [49–51]. The algorithm is constructed upon the Boosting framework, which is a method within the domain of ensemble learning [52]. In the initial phase, a base model is developed using the training dataset and functions as a weak learner [53]. Subsequently, the residuals of the model's predictions are examined, and a new weak learner is created by minimizing a predefined objective function. This iterative process continues until a predetermined number of weak learners is obtained. These weak learners are then integrated to form a highly accurate strong learner for predictive tasks [53,54]. The *XGBoost* further employs a second-order *Taylor* expansion to approximate the objective function and incorporates a regularization term to manage model complexity, thereby effectively mitigating the risk of overfitting [55]. The mathematical formulation of the objective function is presented below.

$$\begin{cases} \text{Obj} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{constant} \\ \Omega(f_t) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \end{cases} \quad (8)$$

The objective function is expanded by means of a second-order *Taylor* series approximation.

$$\text{Obj}^{(t)} = \sum_{i=1}^n \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + \text{constant} \quad (9)$$

Here, $\sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)})$ denotes a fixed value, and together with the constant, both terms can be omitted during the optimization of the loss function. As a result, the aforementioned equation can be simplified as follows.

$$\text{Obj}^t = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (10)$$

The k -th decision tree maps the input sample data to a specific leaf node, denoted as $f_k(x)$. The output value associated with the leaf node indexed by $q(x)$ is represented as $w_{q(x)}$, $f_k(x) = w_{q(x)}$,

$$\begin{cases} \text{Obj}^t = \sum_{i=1}^n \left[g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2 \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \\ = \sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \\ = \sum_{j=1}^T \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \end{cases} \quad (11)$$

Specifically, the aforementioned equation $\sum_{i \in I_j} g_i = G_j$, $\sum_{i \in I_j} h_i = H_j$ represents a univariate quadratic function with respect to w_j . By differentiating this function with respect to w_j , the optimal value can be derived as follows.

$$\begin{cases} w_j^* = -\frac{G_j}{H_j + \lambda} \\ \text{Obj}^t = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \end{cases} \quad (12)$$

3. Sample and Methodology

3.1. Origin and Composition of the Sample Dataset

In QSAR studies, the accuracy of predictive models is largely dependent on the quality and reliability of the underlying data [56]. The Permissible Exposure Limit (PEL) data for all 75 hydrocarbon compounds and their oxygen-containing derivatives were initially collected from the NIOSH Pocket Guide to Chemical Hazards (NPG) [48]. The NPG, developed by the National Institute for Occupational Safety and Health (NIOSH), provides a authoritative and publicly available database of occupational chemical information. The PEL values cited in the NPG are primarily based on the legal standards established by the U.S. Occupational Safety and Health Administration (OSHA), which are derived from extensive toxicological and epidemiological studies, including animal inhalation toxicity experiments and human occupational exposure assessments [10,27]. To ensure consistency in data collection, this study compiled the comprehensive dataset from this source. The specific information of the dataset constructed in this article can be found in the Supplementary Information (SI). This study initially employed the *AP* clustering algorithm to conduct cluster analysis on the sample dataset. Owing to its capacity to automatically determine the optimal number of clusters without requiring prior specification, the *AP* algorithm iteratively updates and

calculates the attraction value and membership value for each data point. The data point enables the efficient identification of cluster centers, particularly for high-dimensional and multi-class data [47]. These cluster centers are actual data points from the original dataset and function as representative exemplars by maximizing overall similarity and minimizing the distances between other data points and their respective class representatives. The *AP* clustering algorithm was implemented using the following parameter configurations: the similarity matrix *S* was determined based on the Euclidean distances between data points; the preference value *P*(*i*) was set to the median value of the similarity matrix *S*; a damping factor λ of 0.5 was applied to mitigate oscillations during the message-passing process; and the maximum number of iterations was capped at 200 to ensure computational efficiency. After 59 iterations, the algorithm ultimately converged to 13 optimal cluster centers. At this stage, the total similarity between all data points and their corresponding class representatives reached its maximum. Subsequently, the *AP* clustering algorithm was applied to classify the sample dataset, and a random partitioning method was employed to divide each category into training and test sets at a ratio of 4:1. The training set consisted of 60 samples, while the test set contained 15 samples. The training set was used to extract molecular descriptors and build the predictive model, whereas the test set was employed to evaluate the model's performance metrics [37].

3.2. Selection of Characteristic Molecular Descriptors

Molecular descriptors, which serve as digital representations of compound molecular structures, play a critical role in elucidating physicochemical properties, predicting biological activities, and supporting drug discovery research [57]. In this study, the molecular structures of hydrocarbons and their oxygen-containing derivatives were initially constructed using the software ChemBioDraw23.1. These structures were subsequently imported into Hyperchem8.0.8 for structural optimization using the MM+ molecular mechanics and PM3 geometry optimization methods [58]. Subsequently, the Dragon 2.0 software was employed to transform the chemical information contained within molecular structures into numerical descriptors [59]. A total of 1481 molecular descriptors spanning 18 categories (constitutional, topological, molecular walk counts, BCUT, Galvez topol. charge indices, 2D autocorrelations, charge, aromaticity indices, Randic molecular profiles, geometrical, RDF, 3D-MoRSE, WHIM, GETAWAY, functional groups, atom-centred fragments, empirical, and properties) 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 [60]. As a result, a refined set of 514 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. Therefore, the genetic function algorithm proposed by Rogers and Hopfinger [61], which utilizes the Lack-of-Fit (LOF) criterion as its fitness function, was employed to further select descriptors. The fitness function was defined using the Friedman LOF metric, with a smoothing parameter α set to 0.5. The algorithm was configured with an initial equation length of 5, a maximum equation length of 10, a population size of 50, a maximum number of generations capped at 500, and a mutation probability of 0.1. As a result, four characteristic molecular descriptors were identified: *X0Av*, *SIC4*, *RDF010e*, and *Dv*. The detailed nomenclature, classification, definitions, and Variance Inflation Factor (VIF) values of these descriptors are provided in Table 1.

Table 1. The nomenclature, classification, definitions, and VIF values of the molecular descriptors.

Nomenclature	Classification	Definition	VIF
<i>X0Av</i>	WHIM descriptors	The average connectivity index chi-0 serves as a quantitative descriptor of molecular branching	1.422
<i>SIC4</i>	topological descriptors	The structural information content at the 4th-order neighborhood symmetry level serves as an indicator of molecular complexity	1.408
<i>RDF010e</i>	RDF descriptors	When weighted by atomic Sanderson electronegativity, the three-dimensional spatial structure reflects the molecular radial distribution	1.240
<i>Dv</i>	topological descriptors	When weighted by atomic van der Waals volume, the total feasibility index D serves as a descriptor of molecular size	1.412

As shown in Table 1, among the four descriptors, three major categories could be distinguished: topological descriptors, RDF descriptors, and WHIM descriptors. Each category captured different aspects of molecular structural features from a unique perspective. Specifically, both *X0Av* and *SIC4* belong to the category of topological descriptors. *X0Av* refers to the average valence connectivity index chi-0, which quantifies the degree of molecular branching. *SIC4* refers to the structural information content derived from the fourth-order neighborhood symmetry, and it functions as a measure of molecular complexity. *RDF010e* is categorized as an RDF descriptor and characterizes the three-dimensional spatial configuration of the molecular radial distribution when weighted by atomic Sanderson electronegativity. *Dv* is classified as a WHIM descriptor and represents the total directional index of D when weighted by atomic van der Waals volume, thereby reflecting the molecular size [62].

To ensure the reliability and validity of the selected descriptors, a multicollinearity assessment was conducted for each variable based on the VIF. A VIF value within the range of (0, 10) is generally interpreted as indicating negligible multicollinearity between independent and dependent variables [63]. As presented in Table 1, the VIF values of the selected molecular descriptors all fall within the range of (1, 2), which is substantially lower than the commonly accepted threshold of 10 [64]. This suggests the absence of significant collinearity between the molecular descriptors and the *PELs*. Consequently, the effects of the four selected molecular descriptors on the *PELs* are statistically significant.

3.3. Model Establishment

3.3.1. MLR Model

The MLR model was constructed using IBM SPSS Statistics 26 software, incorporating four selected molecular descriptors as independent variables and the permissible exposure limit as the dependent variable. The derived MLR model is shown in Equation (13).

$$\ln(PEL) = -16.139 + 21.918 \times X_{X0Av} - 0.568 \times X_{RDF010e} + 18.089 \times X_{Dv} + 4.652 \times X_{SIC4} \quad (13)$$

In the equation, $\ln(PELs)$ denotes the predicted value of MLR for the logarithm of the *PELs*, while X represents the characteristic molecular descriptor.

3.3.2. SVM Model

The SVM model was constructed using the LIBSVM toolbox in the software MATLAB (version 2023), integrating selected molecular descriptors and the *PELs* as key modeling variables. The Radial Basis Function (RBF) was selected as the kernel function, based on the same four characteristic molecular descriptors used as independent variables. To achieve the highest predictive accuracy, the grid search algorithm was applied to identify the optimal parameter combination within the following ranges. The penalty coefficient

$C \in (0.001, 1000]$, the insensitive loss function $\varepsilon \in (0.001, 1]$, and the kernel function width $\gamma \in (0.001, 1]$. Following the normalization of all independent and dependent variables, the optimal SVM parameters were identified as $C = 256$, $\gamma = 0.011$, and $\varepsilon = 0.0731$. Based on these optimized parameters, a QSAR predictive model was constructed using the SVM framework [65].

3.3.3. XGBoost Model

The XGBoost model was developed using MATLAB software to establish a predictive relationship between characteristic molecular descriptors and the PELs. The hyperparameters of the XGBoost model were systematically optimized by adjusting the general parameters, booster parameters, and objective function settings, with the aim of constructing a model exhibiting enhanced predictive performance. The specific parameter configurations are outlined as follows. For the general parameters, the booster type was set to “gbtree”, representing a tree-based *boosting* algorithm. The “nthread” parameter was configured to leverage the maximum number of available computational threads. And the “silent” parameter was set to 0 to enable detailed logging output during the training process. With respect to the objective function configuration, the “binary logistic” function was utilized to perform binary logistic regression, and the Root Mean Square Error (RMSE) was selected as the evaluation metric, aligning with the characteristics of the prediction task. The critical Booster parameters, including the number of decision trees, maximum tree depth, learning rate, and regularization parameters, significantly influenced the model’s performance and required continuous optimization and training. Through iterative optimization and training, the optimal parameter configuration was determined as follows. The number of decision trees (*num_trees*) was set to 120, the maximum tree depth (*depth*) was configured to 5, the learning rate (*eta*) was set to 0.3, and the regularization parameters lambda and alpha were assigned values of 1 and 0, respectively. All remaining parameters were retained at their default settings, under which the model achieved optimal predictive performance [66,67].

3.4. Model Validation and Evaluation

Model validation is conducted to evaluate the fitting capability, robustness, external predictive performance, and generalization ability of the developed model [68]. The validation criteria are established based on the guiding principles and evaluation standards provided by the Organization for Economic Co-operation and Development (OECD) for the assessment of QSAR models [69]. The fitting capability of the model is assessed through statistical metrics, including the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). A higher value of R^2 , accompanied by lower values of RMSE and MAE, indicates enhanced model fitting performance. Model robustness is evaluated using the leave-one-out cross-validation coefficient (Q^2_{loo}), with a higher Q^2_{loo} value reflecting enhanced robustness. External validation is a crucial approach for assessing a model’s predictive performance on independent and previously unseen datasets. To ensure objectivity and reliability in the evaluation process, the external validation dataset should comprise samples that were not involved in the model training stage. The external predictive ability is measured by the external validation coefficient (Q^2_{ext}), where higher values reflect stronger external predictive performance. Furthermore, the applicability domain and generalization ability of the model are evaluated through the Williams plot [37]. In this graphical diagnostic tool, if the majority of sample data points lie within the area confined by three times the standardized residual and the leverage threshold h^* . It indicates that the model is valid and possesses robust predictive and generalization performance. Any data points falling outside this boundary warrant further examination to detect

possible outliers [70]. Detailed calculation procedures for the aforementioned evaluation metrics are presented in the subsequent sections [68].

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (14)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (15)$$

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (16)$$

$$Q_{100}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (17)$$

$$Q_{ext}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (18)$$

$$SD = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 1}} \quad (19)$$

$$F = \frac{\frac{1}{m} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n - m - 1} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (20)$$

In the formula, n denotes the number of samples, i represents index variable representing individual observations (ranging from 1 to n) y_i represents actual observed value of the dependent variable for the i -th observation, $\hat{y}_i$ denotes predicted value of the dependent variable for the i -th observation generated by the model, $\bar{y}$ represents mean value of all actual observed values of the dependent variable across the entire dataset, m represents number of independent variables or features used in the model.

4. Results and Discussion

4.1. Performance of Models

4.1.1. Results and Evaluation of the MLR Model

The statistical validation results of the MLR model are presented in Table 2, and the descriptive statistics of the PELs for hydrocarbons and their oxygenated derivatives across all variables are provided in Table 3.

Table 2. The MLR Model Test Results.

Key Inspection Parameters	R ²	RMSE	SD	<i>p</i>	F
Result	0.8202	0.7441	0.972	<0.001	48.319
Standard	>0.6	A smaller value indicates a more favorable outcome.	A smaller value indicates a more favorable outcome.	<0.05	>F _{theory}

Table 3. Statistical Parameters of the *MLR* Model for Predicting *PELs* of Hydrocarbons and Their Oxygenated Derivatives.

Characteristic Molecular Descriptors	Regression Coefficient	Standardized Coefficient	Standard Error	t-Value	Sig
Constant	16.139		1.911	−8.466	<0.001
<i>X0Av</i>	21.918	0.836	1.956	11.204	<0.001
<i>RDF010e</i>	−0.568	−0.452	0.088	−6.485	<0.001
<i>Dv</i>	18.089	0.427	3.148	5.746	<0.001
<i>SIC4</i>	4.652	0.325	1.064	4.373	<0.001

To evaluate the reliability of the developed *MLR* model, statistical analyses were performed to assess the model's goodness of fit, the overall significance of the regression equation, and the significance of each individual predictor variable, as presented in Tables 2 and 3. The R^2 of 0.8202 indicated a strong overall relationship between the independent variables and $\ln(PELs)$, demonstrating that the model exhibited a satisfactory level of goodness of fit.

The RMSE was 0.7441. Given that the standard deviation of the experimental $\ln(PEL)$ values was 2.03, the RMSE was substantially lower, indicating that the model's prediction error was small relative to the inherent variability in the data. The calculated F-value (F_{true}) was 48.319, whereas the theoretical F-value (F_{theory}) was 2.525, indicating that F_{true} exceeded F_{theory} . Furthermore, the p -value was less than 0.001, suggesting that the regression equation was statistically significant. The absolute values of the t-value for all variables exceeded 4, and the corresponding significance levels (*Sig*) were substantially below 0.001. This indicates that each variable had a statistically significant impact on the regression model ($p < 0.05$). To evaluate the magnitude of each independent variable's impact on the dependent variable, this study utilized standardized coefficients to assess the relative importance of their correlations. The magnitude of the standardized coefficient indicates the strength of the relationship between the descriptor and the *PELs*, with larger absolute values representing stronger associations. Based on the comparison of the standardized coefficients for each independent variable presented in Table 3, it could be concluded that, within this model, the descending order of the contribution of the descriptors to the *PELs* was as follows. $X0Av > RDF010e > Dv > SIC4$.

To evaluate the consistency between the predicted *PELs* and the experimentally observed values, a comparative analysis was performed between the predictions generated by the *MLR* model and the corresponding experimental results, as illustrated in Figure 2. As shown in the figure, the majority of data points closely followed the diagonal line, indicating a strong agreement between the predicted and observed values. A small number of data points, however, deviated noticeably from the line. These findings suggest that the *MLR* model exhibits robust predictive performance. Moreover, the samples were symmetrically and randomly distributed on both sides of the diagonal line, indicating the absence of systematic errors. Therefore, the *MLR* model demonstrated both feasibility and reliability in predicting the *PELs* of hydrocarbons and their derivatives. However, to further explore the nonlinear relationship between the *PELs* and the molecular structures, a nonlinear modeling approach was employed to enhance predictive accuracy.

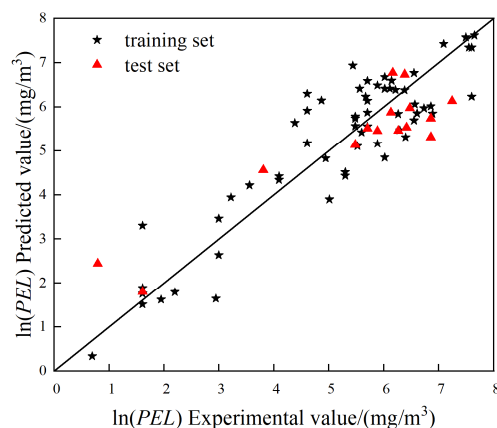


Figure 2. Comparison between the logarithmically transformed predicted values and the experimentally observed values of the *MLR* model.

4.1.2. Results and Evaluation of the *SVM* Model

The *SVM* model demonstrated correlation coefficients of 0.8229 for the training set and 0.843 for the test set, respectively. As illustrated in Figure 3, a comparison is presented between the predicted values generated by the *SVM* model and the experimentally observed values. The distribution pattern of the *SVM* model closely resembled that of the *MLR* model, indicating that both models exhibit robust predictive performance and consistency. Compared to the *MLR* model, the *SVM* model demonstrated a more concentrated data distribution around the diagonal line, with a higher number of substances aligned along it, suggesting improved predictive accuracy. Therefore, it can be concluded that the nonlinear relationship between the *PELs* and the molecular structures plays a more critical role than the linear relationship. Although the *SVM* model showed superior predictive performance, certain substances still deviated notably from the diagonal line. To further investigate the underlying nonlinear relationship between the *PELs* and molecular structures, the *XGBoost* model was employed.

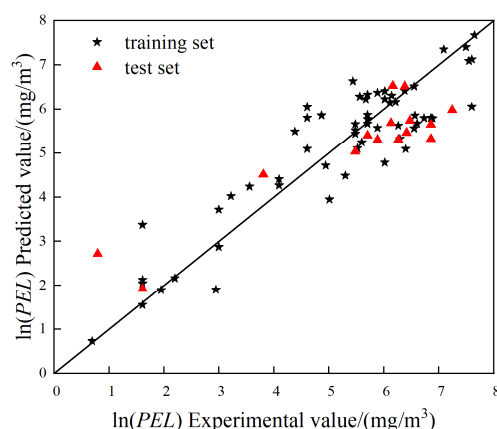


Figure 3. Comparison between the logarithmically transformed predicted values and the experimentally observed values of the *SVM* model.

4.1.3. Results and Evaluation of the *XGBoost* Model

In the *XGBoost* model, the R^2 for the training and test sets were 0.9962 and 0.8892, respectively. The corresponding RMSE were 0.1012 and 0.6623, while the MAE were 0.0102 and 0.4386, respectively. As illustrated in Figure 4, the close agreement between the predicted values and the experimentally observed data visually demonstrates the superior predictive accuracy of the *XGBoost* model. The distribution pattern of the *XGBoost* model showed a significantly closer agreement with the diagonal line, in contrast to the

patterns exhibited by the *MLR* and *SVM* models. Specifically, the majority of data points in the *XGBoost* model were closely clustered around the diagonal line, with a substantial proportion lying directly on it, whereas only a small number exhibited noticeable deviations. This strong agreement between predicted and observed values demonstrates that the *XGBoost* model exhibits the highest predictive accuracy among the evaluated models. Furthermore, the results suggest the presence of a significant nonlinear relationship between the *PELs* and the molecular structures.

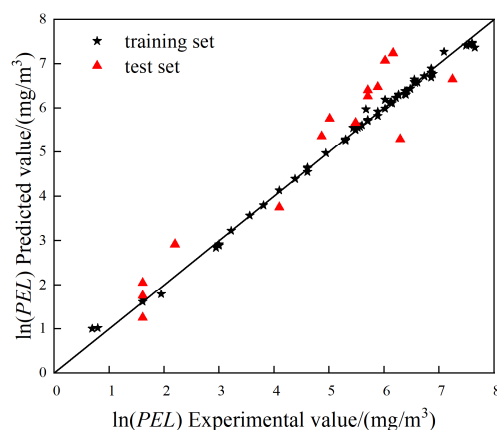


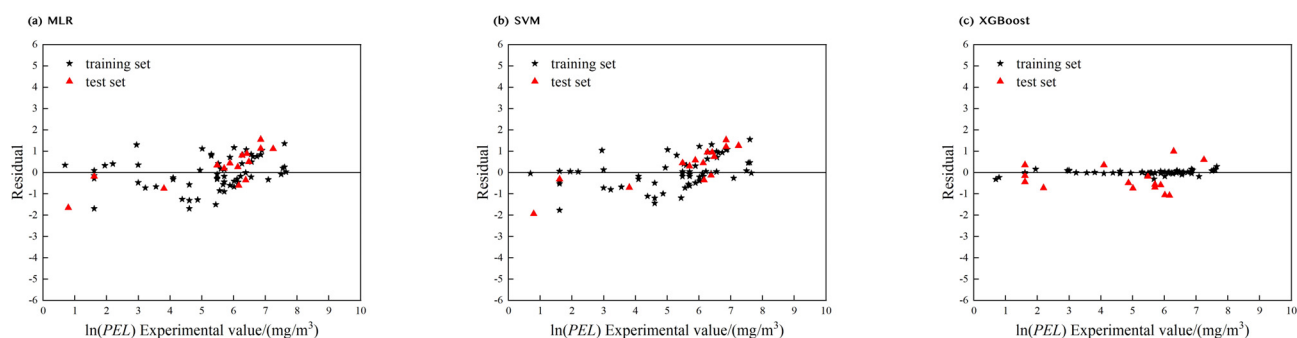
Figure 4. Comparison between the logarithmically transformed predicted values and the experimentally observed values of the *XGBoost* model.

4.2. Model Evaluation and Validation

To systematically evaluate the model's performance in a clear and comprehensive manner, we primarily computed key performance metrics and generated residual plots to visually illustrate the predictive accuracy of the models. The key performance metrics for the three predictive models are presented in Table 4, while the corresponding residual analysis plots for each model are shown in Figure 5. Overall, the *XGBoost* nonlinear model demonstrated consistently superior performance across all evaluation metrics when compared to both the *MLR* linear model and the *SVM* nonlinear model. In most instances, the *SVM* nonlinear model also yielded marginally better results than the *MLR* linear model. These findings suggest the existence of a significant nonlinear relationship between the *PELs* of oxygenated hydrocarbon derivatives and their corresponding molecular structures. From the perspective of performance metrics, the R^2 values of all three models exceed 0.8, which was significantly higher than the generally accepted threshold of 0.6. Additionally, both the RMSE and MAE values were below 1, indicating a relatively low level of prediction error. It can be concluded that the three models exhibit high multiple correlation coefficients, along with low root mean square error and mean absolute error values, reflecting strong fitting performance. Moreover, the cross-validation coefficients Q^2_{loo} for all three models were above 0.8. In the residual analysis plots, the residuals were randomly distributed within the interval $(-3, 3)$, displaying no discernible patterns and remaining close to the baseline. This indicates that the models exhibit a high level of stability. Additionally, the external validation coefficient Q^2_{ext} for all three models exceeded 0.75, further confirming their robust predictive performance on independent external datasets.

Table 4. Performance Evaluation Parameters of Predictive Models.

Performance Parameters	Models					
	MLR		SVM		XGBoost	
	Training Set	Test Set	Training Set	Test Set	Training Set	Test Set
R^2	0.8202	0.8427	0.8229	0.843	0.9962	0.8892
RMSE	0.7441	0.8479	0.7243	0.9324	0.1012	0.6623
MAE	0.6034	0.7163	0.562	0.7906	0.0102	0.4386
Q^2_{loo}	0.8127		0.8225		0.9964	
Q^2_{ext}		0.7936		0.7505		0.8921

**Figure 5.** Residual plots of predicted $\ln(PEL)$ values for the (a) MLR, (b) SVM, and (c) XGBoost models.

Based on a comprehensive evaluation and comparative analysis of the three models, it is evident that the performance parameters of all models meet the established criteria for model construction. Moreover, in terms of predicting the PEL s of oxygenated hydrocarbon derivatives, the XGBoost model consistently outperforms the MLR and SVM models across all evaluated metrics. This suggests a significant nonlinear relationship between the PEL s of these compounds and their molecular structures.

4.3. Evaluation of the Models' Applicability Domain

To further validate the validity and reliability of the model, the Williams plot was utilized to evaluate its applicability domain [37]. The identification of outliers is determined by whether the sample data fall within the rectangular region defined by ± 3 standard deviations of the residuals and the critical leverage value, h^* , or on its boundary. Further analysis and evaluation are then carried out based on these findings. As illustrated in Figure 6, the standardized residuals of most sample data points fall within the rectangular region defined by the interval $(-3, +3)$ and the critical leverage value $h^* = 0.2$. Nevertheless, three outliers were identified. Among these, diisobutyl ketone and dioctyl phthalate fall within the ± 3 standard deviations range of the residuals; however, they exceed the critical leverage value h^* . This observation may be attributed to the relatively unique molecular structures of these compounds, which necessitate the model to generate predictions through extrapolation. Nevertheless, such characteristics may contribute to enhanced model stability and provide a certain level of extrapolative capacity. As a result, these outliers can be categorized as “benign outliers” [71]. In summary, all three models demonstrate broad applicability, as well as strong predictive performance and generalization capabilities.

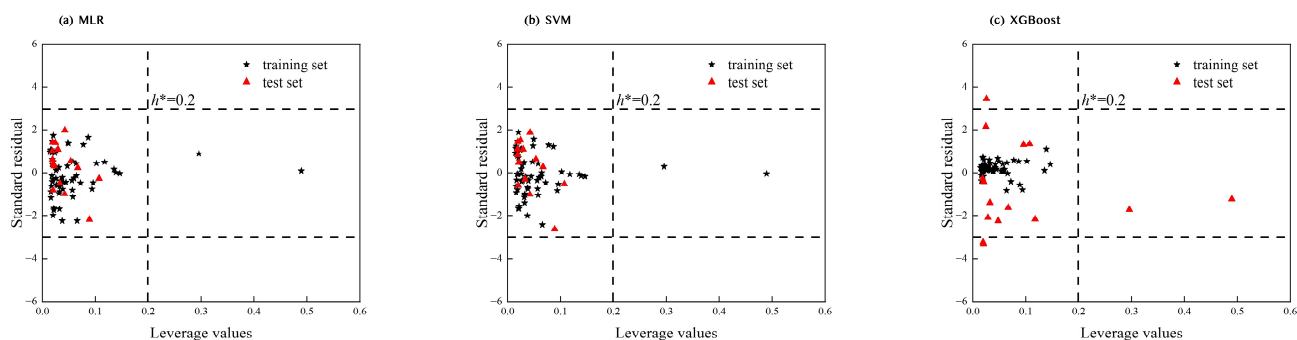


Figure 6. Williams plots for the (a) MLR, (b) SVM, and (c) XGBoost models.

5. Conclusions

This study constitutes the first systematic attempt to establish a correlation between the *PELs* and the molecular structures of organic chemical compounds. Oxygen-containing derivatives of hydrocarbons, chosen for their high reactivity and widespread practical applications, were selected as the target compounds. Both the linear model based on *MLR* and nonlinear models based on the *SVM* and *XGBoost* were constructed to enable a comparative performance analysis. The core value of the entire modeling process lay in translating abstract chemical structural information—represented by four key molecular descriptors—into concrete and quantifiable recommendations for permissible exposure limits (*PELs*). The *MLR* model, which elucidated clear linear relationships, along with the *SVM* and *XGBoost* models, capable of capturing complex nonlinear patterns, collectively functioned as essential translational tools that bridged the gap between “chemical fingerprints” and occupational “safety limits.” To ensure a balanced, scientifically valid, and representative distribution of sample types, the *AP* clustering algorithm was utilized to partition the dataset (sourced from the NIOSH Pocket Guide to Chemical Hazards [48]) into training and testing subsets. Through a comprehensive evaluation of model performance metrics and the construction of the Williams plot, the fitting capability, predictive accuracy, and model stability were thoroughly assessed. The results of the study indicate that all three models demonstrate robust performance, exhibiting strong predictive power and generalization ability. However, the *XGBoost* model in particular exhibited superior predictive performance, as indicated by the high coefficient of determination values of $R^2_{\text{train}} = 0.9962$ and $R^2_{\text{test}} = 0.8892$. These results offer strong empirical support for the presence of a significant nonlinear association between the *PELs* and the molecular structures. This observed nonlinearity implies that minor structural modifications in hydrocarbons and their derivatives could lead to disproportionate shifts in their toxicological profiles, necessitating a move beyond traditional linear models. Consequently, our QSAR model, by capturing these complex relationships, provides a more robust tool for the a priori estimation of *PELs* for untested compounds. This approach paves the way for a more mechanistic understanding of how specific molecular descriptors govern occupational health risks.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app152111642/s1>, Table S1: Table SI Names of hydrocarbons and their oxygen-containing derivatives and predicted results of $\ln(\text{PEL})$ training and test sets.

Author Contributions: Conceptualization, J.S. and X.Y.; methodology, J.S.; software, Z.Z.; validation, W.Z.; formal analysis, J.S. and W.Z.; investigation, J.S. and Y.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.

Institutional Review Board Statement: Not applicable for studies not involving humans or animals.

Informed Consent Statement: Not applicable.

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

- Shen, Z.; Hu, Y.; Li, B.; Zou, Y.; Li, S.; Busser, G.W.; Wang, X.; Zhao, G.; Muhler, M. State-of-the-art progress in the selective photo-oxidation of alcohols. *J. Energy Chem.* **2021**, *62*, 338–350. [\[CrossRef\]](#)
- Yan, B.; Wu, J.; Deng, J.; Chen, D.; Ye, X.; Yao, Q. Recent progress in light-driven direct dehydroxylation and derivation of alcohols. *Chin. J. Org. Chem.* **2023**, *43*, 3055–3066. [\[CrossRef\]](#)
- Wei, D.; Bu, J.; Zhang, S.; Chen, S.; Yue, L.; Li, X.; Liang, K.; Xia, C. Light-driven stepwise reduction of aliphatic carboxylic esters to aldehydes and alcohols. *Angew. Chem. Int. Ed.* **2025**, *64*, e202420084. [\[CrossRef\]](#)
- McDonough, D.; Bezold, E.L.; Wuest, W.M.; Minbiole, K.P.C. The versatile synthesis and biological evaluation of all-alkyl biscationic quaternary phosphonium compounds: Atom-economical and potent disinfectants. *RSC Med. Chem.* **2025**. [\[CrossRef\]](#) [\[PubMed\]](#)
- Tzouras, N.V.; Zorba, L.P.; Kaplanai, E.; Tsoureas, N.; Nelson, D.J.; Nolan, S.P.; Vougioukalakis, G.C. Hexafluoroisopropanol (HFIP) as a multifunctional agent in gold-catalyzed cycloisomerizations and sequential transformations. *ACS Catal.* **2023**, *13*, 8845–8860. [\[CrossRef\]](#)
- Singh, P.; Kumar, R. Critical review of microbial degradation of aromatic compounds and exploring potential aspects of furfuryl alcohol degradation. *J. Polym. Environ.* **2019**, *27*, 901–916. [\[CrossRef\]](#)
- Bouzidi, H.; Laversin, H.; Tomas, A.; Coddeville, P.; Fittschen, C.; El Dib, G.; Roth, E.; Chakir, A. Reactivity of 3-hydroxy-3-methyl-2-butanone: Photolysis and OH reaction kinetics. *Atmos. Environ.* **2014**, *98*, 540–548. [\[CrossRef\]](#)
- Gugumus, F. Contribution to the role of aldehydes and peracids in polyolefin oxidation1. Photolysis and photooxidation of aldehydes in polyethylene. *Polym. Degrad. Stab.* **1999**, *65*, 259–269. [\[CrossRef\]](#)
- Shen, M.; Almallahi, R.; Rizvi, Z.; Gonzalez-Martinez, E.; Yang, G.; Robertson, M.L. Accelerated hydrolytic degradation of ester-containing biobased epoxy resins. *Polym. Chem.* **2019**, *10*, 3217–3229. [\[CrossRef\]](#)
- Zhang, S.; Jia, C.; Gao, H.; Huang, T.; Bai, X.; Suo, H.; Pu, G.; Wang, C.; Chen, H.; Ma, J. Pollution characteristics and potential sources of Peroxyacetyl Nitrate in a petrochemical industrialized city, Northwest China. *Chemosphere* **2025**, *372*, 144104. [\[CrossRef\]](#)
- Zhang, S.; Li, H.; He, R.; Deng, W.; Ma, S.; Zhang, X.; Li, G.; An, T. Spatial distribution, source identification, and human health risk assessment of PAHs and their derivatives in soils nearby the coke plants. *Sci. Total Environ.* **2023**, *861*, 160588. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hou, Y.; Che, Y.; Li, T.; Yan, Z.; Zhao, W.; Lv, S.; Zhang, F.; Zhou, M.; Zhou, Y.; Zhu, Z.; et al. Exploring the mechanisms underlying effects of oxygenated polycyclic aromatic hydrocarbons exposure on inflammatory bowel disease. *Ecotoxicol. Environ. Saf.* **2025**, *304*, 119153. [\[CrossRef\]](#)
- Liu, R.; Liu, Z.; Liu, H.C.; Shi, H. An improved alternative queuing method for occupational health and safety risk assessment and its application to construction excavation. *Autom. Constr.* **2021**, *126*, 103672. [\[CrossRef\]](#)
- Fata, C.M.L.; Giallanza, A.; Micale, R.; Scalia, G.L. Ranking of occupational health and safety risks by a multi-criteria perspective: Inclusion of human factors and application of VIKOR. *Saf. Sci.* **2021**, *138*, 105234. [\[CrossRef\]](#)
- Dizdar, E.N.; Ünver, M. The assessment of occupational safety and health in Turkey by applying a decision-making method: MULTIMOORA. *Hum. Ecol. Risk Assess.* **2020**, *26*, 1693–1704. [\[CrossRef\]](#)
- Caraballo-Ay, Y. Occupational safety and health in Venezuela. *Ann. Glob. Health* **2015**, *81*, 512–521. [\[CrossRef\]](#)
- Boom, Y.J.; Enfrin, M.; Grist, S.; Giustozzi, F. Analysis of possible carcinogenic compounds in recycled plastic modified asphalt. *Sci. Total Environ.* **2023**, *858*, 159910. [\[CrossRef\]](#)
- Rappaport, S.M. The rules of the game: An analysis of Osha's enforcement strategy. *Am. J. Ind. Med.* **1984**, *6*, 291–303. [\[CrossRef\]](#)
- Tang, S.H.; Zhang, C.; Zhou, L.L.; Li, Y.Q.; Xu, S.X.; Wang, Z. An investigation and analysis of an acute occupational methyl acetate poisoning. *Chin. J. Ind. Hyg. Occup. Dis.* **2021**, *39*, 943–946.
- Shen, H.; Xu, S.; Fei, X.; Song, X.; Chang, Q.; Zhu, B. Investigation and Analysis of an Acute Occupational Methanol Poisoning Accident. *J. Environ. Occup. Med.* **2020**, *37*, 818–820.
- Sayapathi, B.S.; Su, A.T.; Koh, D. The impact of different permissible exposure limits on hearing threshold levels beyond 25 dBA. *Iran. Red Crescent Med. J.* **2014**, *16*, e15520. [\[CrossRef\]](#) [\[PubMed\]](#)

22. Jou, J.; Chen, J.; Lin, J.; Cheng, M. An easy-to-apply method for determining permissible exposure limit of retina to light. *Heliyon* **2022**, *8*, e10927. [CrossRef]
23. Uhlemeier, K.V.; Wood, T.B. Laboratory evaluation of permissible exposure limits for men in hot environments. *Am. J. Ind. Med.* **1979**, *40*, 1097–1103.
24. Wu, M. Discussion on the setting of emergency rescue isolation areas based on simulation of benzene tower leakage scenarios. *Occup. Health Emerg. Rescue* **2017**, *35*, 558–560.
25. Yu, L.; Shen, X.; Yang, M.; Xiu, G.; Qian, F.; Wang, J. Simulation study on benzene leakage and exposure risk in the isomerization unit of the aromatic hydrocarbon plant. *Chin. J. Saf. Sci.* **2017**, *27*, 79–84.
26. Szczesna, D.; Kupczewska-Dobacka, M.; Konieczko, K.; Jurewicz, J. P19-08 New values of occupational exposure limits (OELs) of inhalation anesthetics: Enflurane, isoflurane, sevoflurane and desflurane in Poland. *Toxicol. Lett.* **2022**, *368*, S213. [CrossRef]
27. Tustin, A.W.; Cannon, D.L. Analysis of biomonitoring data to assess employer compliance with OSHA's permissible exposure limits for air contaminants. *Am. J. Ind. Med.* **2021**, *65*, 81–91. [CrossRef]
28. Kostoff, R.N.; Aschner, M.; Goumenou, M.; Tsatsakis, A. Setting safer exposure limits for toxic substance combinations. *Food Chem. Toxicol.* **2020**, *140*, 111346. [CrossRef]
29. Occupational Safety and Health Administration. *OSHA History*. Available online: <https://www.osha.gov/history> (accessed on 24 October 2025).
30. Zheng, Y. Occupational Exposure Assessment and Genetic Damage Study of Vinyl Chloride Workers. Master's Thesis, Fudan University, Shanghai, China, 2009. [CrossRef]
31. Pamies, D.; Estevan, C.; Vilanova, E.; Sogorb, M.A. Chapter 7-Alternative methods to animal experimentation for testing developmental toxicity. In *Reproductive and Developmental Toxicology*, 3rd ed.; Academic Press: Lausanne, Switzerland; Elche, Spain, 2022; pp. 107–125.
32. DeSesso, J.M. Future of developmental toxicity testing. *Curr. Opin. Toxicol.* **2017**, *3*, 1–5. [CrossRef]
33. Manganelli, S.; Schilter, B.; Scholz, G.; Benfenati, E.; Piparo, E.L. Value and limitation of structure-based profilers to characterize developmental and reproductive toxicity potential. *Arch. Toxicol.* **2020**, *94*, 939–954. [CrossRef] [PubMed]
34. Zhao, F.; Rogers, W.J.; Sam, M.M. Experimental measurement and numerical analysis of binary hydrocarbon mixture flammability limits. *Process Saf. Environ. Prot.* **2009**, *87*, 94–104. [CrossRef]
35. Rappaport, S.M. Threshold limit values, permissible exposure limits, and feasibility: The bases for exposure limits in the United States. *Am. J. Ind. Med.* **1993**, *23*, 683–694. [CrossRef]
36. 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]
37. Gissi, A.; Tcheremenskaia, O.; Bossa, C.; Battistelli, C.L.; Browne, P. The OECD (Q)SAR Assessment Framework: A tool for increasing regulatory uptake of computational approaches. *Comput. Toxicol.* **2024**, *31*, 100326. [CrossRef]
38. Otto, M.A.; Martin, N.J.; Rous, J.S.; Stevens, M.E. Determination of airborne concentrations of dichlorvos over a range of temperatures when using commercially available pesticide strips in a simulated military guard post. *J. Occup. Environ. Hyg.* **2017**, *14*, D54–D61. [CrossRef] [PubMed]
39. Habschied, K.; Šarić, G.K.; Krstanović, V.; Mastanjević, K. Biomonitoring and human exposure. *Toxins* **2021**, *13*, 113. [CrossRef] [PubMed]
40. Storsjö, T.; Tinnerberg, H.; Sun, J.; Chen, R.; Farbrot, A. Elemental carbon—An efficient method to measure occupational exposure from materials in the graphene family. *NanoImpact* **2024**, *33*, 100499. [CrossRef]
41. Cai, N.; Zhao, Y.; Xu, F.; Jiang, M.; Han, L.; Zhu, B.; Wang, B. Integrated internal and external exposure models for dimethylformamide risk assessment and health risk monetization. *Ecotoxicol. Environ. Saf.* **2025**, *291*, 117890. [CrossRef]
42. Golalipour, K.; Akbari, E.; Hamidi, S.S.; Lee, M.; Enayatifar, R. From clustering to clustering ensemble selection: A review. *Eng. Appl. Artif. Intell.* **2021**, *104*, 104388. [CrossRef]
43. Frey, B.J.; Dueck, D. Response to comment on clustering by passing messages between data points. *Science* **2008**, *319*, 726–727. [CrossRef]
44. Bejar, J.; Paternina, M.R.A.; Mendez, A.Z.; Lugnani, L.; Tellez, E. Power system coherency assessment by the affinity propagation algorithm and distance correlation. *Sustain. Energy Grids Netw.* **2022**, *30*, 100658. [CrossRef]
45. Fang, X.; Luo, C.; Zhang, D.; Zhang, H.; Qian, J.; Zhao, C.; Hou, Z.; Zhang, Y. Pre-selection of monitoring stations for marine water quality using affinity propagation: A case study of Xincun Lagoon, hainan, China. *J. Environ. Manag.* **2022**, *325*, 116666. [CrossRef]
46. Reddy, V.S.; Kinnicutt, P.; Lee, R. Text document clustering: The application of cluster analysis to textual document. In Proceedings of the 2016 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 15–17 December 2016; IEEE: Las Vegas, NV, USA, 2016; pp. 1174–1179.
47. Marc, M. Where Are the Exemplars? *Science* **2007**, *315*, 949–951. [CrossRef]

48. 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 24 October 2025).
49. Lin, M.; Yang, K.; Yu, Z.; Shi, Y.; Chen, C.L.P. Hybrid ensemble broad learning system for network intrusion detection. *IEEE Trans. Ind. Inform.* **2023**, *20*, 5622–5633. [[CrossRef](#)]
50. Yu, Z.; Dong, Z.; Yu, C.; Yang, K.; Fan, Z.; Chen, C.L.P. A review on multi-view learning. *Front. Comput. Sci.* **2025**, *19*, 197334. [[CrossRef](#)]
51. Kang, S.W.; Park, C.H. Effective federated XGBoost learning for multi-class classification in Non-IID environments. *J. Supercomput.* **2025**, *81*, 777. [[CrossRef](#)]
52. Bauer, E.; Kohavi, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Mach. Learn.* **1999**, *36*, 105–139. [[CrossRef](#)]
53. 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; ACM: New York, NY, USA, 2016; pp. 785–794.
54. Mienye, I.D.; Sun, Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access* **2022**, *10*, 99129–99149. [[CrossRef](#)]
55. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
56. De, P.; Kar, S.; Ambure, P.; Roy, K. Prediction reliability of QSAR models: An overview of various validation tools. *Arch Toxicol* **2022**, *96*, 1279–1295. [[CrossRef](#)] [[PubMed](#)]
57. Gutman, I. Degree-based topological indices. *Croat. Chem. Acta* **2013**, *86*, 351–361. [[CrossRef](#)]
58. Coleman, W.F.; Arumainayagam, C.R. HyperChem 5 (by Hypercube, Inc.). *J. Chem. Educ.* **1998**, *75*, 416. [[CrossRef](#)]
59. Hutter, M.C. Molecular descriptors for chemoinformatics (2nd ed.). *ChemMedChem* **2010**, *5*, 306–307. [[CrossRef](#)]
60. Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *MATCH Commun. Math. Comput. Chem.* **2006**, *56*, 237–248.
61. Rogers, D.; Hopfinger, A.J. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866. [[CrossRef](#)]
62. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
63. Cheng, J.; Sun, J.; Yao, K.; Xu, M.; Cao, Y. A variable selection method based on mutual information and variance inflation factor. *Spectrochim. Acta Part A* **2022**, *268*, 120652. [[CrossRef](#)] [[PubMed](#)]
64. Peng, F.; Lu, L.; 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](#)]
65. Wainer, J.; Fonseca, P. How to tune the RBF SVM hyperparameters? An empirical evaluation of 18 search algorithms. *Artif. Intell. Rev.* **2021**, *54*, 4771–4797. [[CrossRef](#)]
66. Lv, C.X.; An, S.; Qiao, B.; Wu, W. Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. *BMC Infect. Dis.* **2021**, *21*, 839. [[CrossRef](#)]
67. Rahman, M.S.; Chowdhury, A.H.; Amrin, M. Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh. *PLoS Glob. Public Health* **2022**, *2*, e0000495. [[CrossRef](#)]
68. Pore, S.; Pelloux, A.; Chatterjee, M.; Banerjee, A.; Roy, K. Machine learning-based q-RASAR predictions of the bioconcentration factor of organic molecules estimated following the organisation for economic co-operation and development guideline 305. *J. Hazard. Mater.* **2024**, *479*, 135725. [[CrossRef](#)] [[PubMed](#)]
69. Organisation for Economic Co-operation and Development (OECD). *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*; OECD Series on Testing and Assessment, No. 69; OECD Publishing: Paris, France, 2014.
70. Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci* **2007**, *26*, 694–701. [[CrossRef](#)]
71. Wang, D.; Yuan, Y.; Duan, S.; Liu, R.; Gu, S.; Zhao, S.; Liu, L.; Xu, J. QSPR study on melting point of carbocyclic nitroaromatic compounds by multiple linear regression and artificial neural network. *Chemom. Intell. Lab. Syst.* **2015**, *143*, 7–15. [[CrossRef](#)]

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.