



REGRESSION-BASED STATISTICAL ANALYSIS OF STRUCTURE-PROPERTY RELATIONSHIPS USING KERNEL AND CLASSICAL PREDICTION MODELS

Mohammed Sadiq Pachapuri

Assistant Professor, Mechanical Engineering, KLE Technological University Dr M S Sheshgiri Campus Belagavi
4th Railway Gate, Angol Road Udyambag, Belagavi, 590008. Email: Pachapuri@gmail.com. ORCID: 0000-0001-6340-4529

Corresponding author: Pachapuri@gmail.com

ARTICLE INFO

Received: 03/10/2025
Revised: 07/12/2025

Accepted: 15/01/2026
Published: 29/01/2026

DOI: <https://doi.org/10.69980/nx5sxf77>
Volume:12, Issue:01, January 2026

Abstract

The study explored quantitative structure–property relationship (QSPR) for aquatic toxicity using regression-based statistical approaches and kernel-based prediction models. The dataset comprised 908 chemical compounds with LC50 [-log(mol/L)] treated as the continuous response variable. Descriptive statistics, Pearson correlation analysis, multicollinearity assessment, multiple linear regression, regularized regression, support vector regression, and kernel ridge regression were applied to characterize descriptor–response relationships and compare predictive performance. Model performance was assessed through 10-fold cross-validation and a held-out test set using R^2 , root mean squared error (RMSE), mean absolute error (MAE), and mean squared error (MSE). Multiple linear regression explained 57.7% of the variability in LC50 [-log(mol/L)], while MLOGP showed the strongest positive bivariate association with the response. Among the evaluated models, radial basis function support vector regression achieved the strongest predictive performance, with test $R^2 = 0.605$, RMSE = 0.978, MAE = 0.677, and MSE = 0.957. Ridge, LASSO, and Elastic Net produced performance like conventional linear regression, indicating that regularization alone provided limited predictive improvement. Permutation-based importance identified MLOGP and SM1_Dz(Z) as the most influential predictors, followed by CIC0, NdsCH, GATS1i, and NdssC. The findings indicate that nonlinear kernel modelling can provide modest but meaningful gains over classical regression for medium-sized molecular descriptor datasets. Combining interpretable linear models with flexible kernel methods offers a practical framework for quantitative structure–property analysis, toxicity prediction, comparative model evaluation, and computational chemical screening. The approach enables simultaneous interpretation of descriptor effects and nonlinear predictive patterns, thereby strengthening evidence obtained from complementary modelling strategies.

Keywords: QSAR; aquatic toxicity; support vector regression; kernel methods; molecular descriptors

1. Introduction

The quantitative structure–property relationship (QSPR) modelling has emerged as an important computational methodology used for the connection between molecular structure and measurable physicochemical, biological and toxicological properties. These can be used to transform chemical structures into numerical descriptors which allow statistical relationships between molecular features and experimentally observed responses to be established. These models may also decrease the need for comprehensive lab testing, speed up compound screening, and aid in chemical design and hazard assessment. With the development of machine learning, the predictive power of the classic QSAR and QSPR has been enhanced, allowing for more complex and nonlinear descriptor–response relationships to be represented (Mao et al., 2021). Meanwhile, key challenges in molecular modelling continue to be the accuracy, interpretability, and generalizability of predictive models (Martinez-Mayorga et al., 2024).

Classical regression methods remain useful for structure–property analysis as they give interpretable coefficients and allow direct statistical analysis of the molecular descriptors. Multiple linear regression, Ridge regression, LASSO, and Elastic Net can be used to quantify the effects of the descriptors and to deal with different correlations and model complexity. Dai et al. (2020) showed that data-driven statistical methods can be applied to discover structure–property relationships that are useful in materials design. Gao et al. (2021) demonstrated that the molecular descriptors are related to chemically relevant response variables by means of QSAR and machine-learning methods. But higher-order interactions and nonlinearities are often found in chemical systems, which may be not represented by a strictly linear formulation.

An alternative way of modelling such nonlinear relationships is by using kernel-based learning. Support vector regression represents the predictor information in higher dimensions using kernel functions, allowing for the estimation of nonlinear response surfaces without having to explicitly write the mathematical functions. The kernel ridge regression is a technique that incorporates kernel transformations with regularization and has received more attention in modern QSAR applications. Mahmood et al. (2025) showed the importance of kernel methods in chemical prediction by presenting modified kernel ridge regression, which achieved better QSAR performance. In general, the authors of the study, Haywood et al. (2022), demonstrated that kernel methods can yield valuable predictive ability for complicated chemical effects.

Structural information can be directly extracted from molecular graph, image, or chemical languages representation by deep learning, representation-learning methods (Tang et al., 2023). Pang et al. (2023) discussed the growing importance of deep learning for predicting molecular properties, and Das et al. (2024) noted the role of chemical language models in predicting molecular and reaction properties. Zeng et al. (2022) also showed the feasibility of accurate molecular property and drug-target predictions using self-supervised representation learning. Regression-based methods are still appealing for medium-sized datasets, where only a small number of informative descriptors is available, since they enable a direct comparison of the interpretability of the model with its flexibility in describing the data.

The ability to predict the toxicity is especially relevant in toxicity modelling which can help with chemical prioritization and environmental risk assessment. In fact, descriptor-based models were shown to be valuable for aquatic toxicology by using machine learning to predict toxicity towards fathead minnow as done by Chen et al. (2020). The *in silico* method for predicting acute aquatic toxicity developed by Gajewicz-Skretna et al. 2021 was also developed for a variety of organic chemicals. Nath et al. (2023) also demonstrated the applicability of the QSAR techniques for assessing the aquatic toxicity of agrochemicals. Gasser et al. (2024) explored the prediction of fish acute mortality using machine learning, and Sun et al. (2024) highlighted the importance of explainable models for pesticide toxicity. Conventional QSAR approaches were expanded by Zubrod et al. (2024) who incorporated chemical and biological data to enhance ecotoxicity predictions.

Belfield et al. (2023) emphasized the need for meticulous validation, model transparency, and proper performance evaluation in toxicological QSAR studies using machine learning. Cronin et al. (2025) also highlighted model predictivity, model validation, and model generalization in assessing toxicity QSARs.

Therefore, the current study aims to explore these structure–property relationships by means of molecular descriptors, and to compare between the classical, regularized, and kernel based regression methods for prediction of a continuous toxicity response. Both statistical association analysis as well as predictive performance are analyzed and evaluated to see if there is any measurable improvement when using nonlinear kernel modelling compared to standard regression models while maintaining interpretable information about the contributions of descriptors.

2. Materials and Methods

2.1 Study Design and Dataset

A quantitative structure–property secondary dataset was used to explore the relationships between molecular descriptors and the continuous toxicity response (Ballabio et al., 2015). A dataset of 908 chemical compounds was used and each compound was characterized by six molecular descriptors: CIC0, SM1_Dz(Z), GATS1i, NdsCH, NdssC, and MLOGP. The dependent variable was LC50 [-log(mol/L) the toxicity response]. Prior to statistical modelling, the data set was checked with regard to completeness and duplicate observations. There were no missing values found and one duplicated record was found. The entire data set of 908 observations was kept for the main analysis.

2.2 Descriptive Statistical Analysis

The distributions of the six molecular descriptors and LC50 [-log(mol/L)] were described using descriptive statistics. The mean, standard deviation, median, minimum, maximum, interquartile range and skewness of each variable were calculated. The distributions and variability of the molecular descriptors and toxicity response were analysed visually using histograms.

2.3 Correlation and Multicollinearity Analysis

Pearson correlation coefficient (r) was used to assess the pairwise relationships between six molecular descriptors and LC50 [-log(mol/L)]. The strength of the correlations and the direction were analyzed to find linear relationships between structural properties and toxicity. Corresponding p -values were used to determine statistical significance. To obtain an overview of the total correlation structure between the variables, a correlation heatmap was built.

The variance inflation factor (VIF) was used to check for possible multicollinearity between the molecular descriptors. The linear dependence between the molecular descriptors was quantified by calculating the VIF value for each of the descriptors.

2.4 Multiple Linear Regression

A multiple linear regression model has been fitted, with LC50 [-log(mol/L)] as the dependent variable and CIC0, SM1_Dz(Z), GATS1i, NdsCH, NdssC, and MLOGP as simultaneous independent variables. The model performance was compared by the coefficient of determination (R^2), adjusted R^2 and overall F -statistic. The estimates of the unstandardized regression coefficient (B), the standard error, the t -statistic, the p -value, and the corresponding 95% confidence interval were calculated for individual predictors.

2.5 Classical, Regularized, and Kernel-Based Regression Models

The eight regression specifications were tested including multiple linear regression, Ridge regression, LASSO regression, Elastic Net regression, linear support vector regression (SVR), polynomial SVR, radial basis function support vector regression (RBF-SVR) and kernel ridge regression.

The data set consisted of 80% training set and 20% test set. The hyperparameter selection and model fitting procedures were performed on the training data set. In case of RBF-SVR, the following specifications were selected: $C = 1$, $\epsilon = 0.1$, and gamma as determined by the scale-based formulation.

2.6 Cross-Validation and Predictive Performance Assessment

The performance and stability of the models were tested using 10-fold cross-validation. The results of the cross validation were summarized by R^2 and root mean squared error (RMSE). After the development of the models, their predictive abilities were tested on the test set that remained out.

The coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and mean squared error (MSE) were used to quantify test-set performance. The eight regression models were compared in terms of their predictive accuracy using these measures. Two new plots were also created for observed and predicted values for multiple linear regression and the best performing RBF-SVR model.

2.7 Residual Analysis

The model that performs best on the test set is the RBF-SVR model, and residual analysis was performed for it. The residuals from the prediction were plotted against the predicted LC50 [-log(mol/L)] values and a reference line with an error of zero was drawn to investigate the distribution and direction of the prediction error.

2.8 Permutation-Based Descriptor Importance

Permutation importance was used to measure the importance of each molecular descriptor to the predictive ability of the RBF-SVR model. The descriptors were randomly permuted to measure the increase in prediction RMSE for each random permutation. However, an increase in RMSE after permutation suggested a higher contribution to the prediction from the permuted descriptor. The mean increase in RMSE and standard deviation were used to summarize the importance of the descriptors, and the descriptors were ranked by their contribution to the prediction error.

2.9 Ethical Considerations

In this study, a publicly available secondary dataset was used (UCI Machine Learning Repository) and human and animal subjects were not directly involved (Ballabio et al., 2015). The dataset contained no personally identifiable or confidential information. Thus, formal ethical approval and informed consent were not necessary. The dataset was used only for academic research and analysis, as per the terms of use.

3. Results

3.1 Dataset Characteristics and Descriptive Statistics

The final dataset consisted of 908 chemical compounds with the following continuous toxicity endpoint LC50 [-log(mol/L)] and six molecular descriptors (CIC0, SM1_Dz(Z), GATS1i, NdsCH, NdssC and MLOGP). There were no missing observations detected. A data audit was performed on the data supplied, and one duplicate record was identified, so that all 908 observations were included in the main analysis. There was significant variation in both molecular structure descriptors and toxicity measurements (Table 1).

Table 1. Descriptive statistics of molecular descriptors and LC50 [-log(mol/L)] (n = 908)

Variable	Mean	SD	Median	Minimum	Maximum	IQR	Skewness
CIC0	2.898	0.756	2.934	0.667	5.926	1.060	0.045

SM1_Dz(Z)	0.628	0.428	0.570	0.000	2.171	0.670	0.695
GATS1i	1.294	0.394	1.241	0.396	2.920	0.612	0.723
NdsCH	0.229	0.605	0.000	0.000	4.000	0.000	3.401
NdssC	0.486	0.861	0.000	0.000	6.000	1.000	2.239
MLOGP	2.109	1.433	2.127	-2.884	6.515	1.896	-0.035
LC50 [-log(mol/L)]	4.064	1.456	3.988	0.053	9.612	1.756	0.252

The LC50 [-log(mol/L)] response showed a mean of 4.064 ± 1.456 , with observed values extending from 0.053 to 9.612. The count-based NdsCH and NdssC descriptors were clearly skewed to the right, but the CICO and MLOGP descriptors were roughly symmetric. The skewness for NdsCH was the highest (3.401), indicating that most of the observations were clustered around zero. The distributions of the descriptors and outcome are shown in Figure 1.

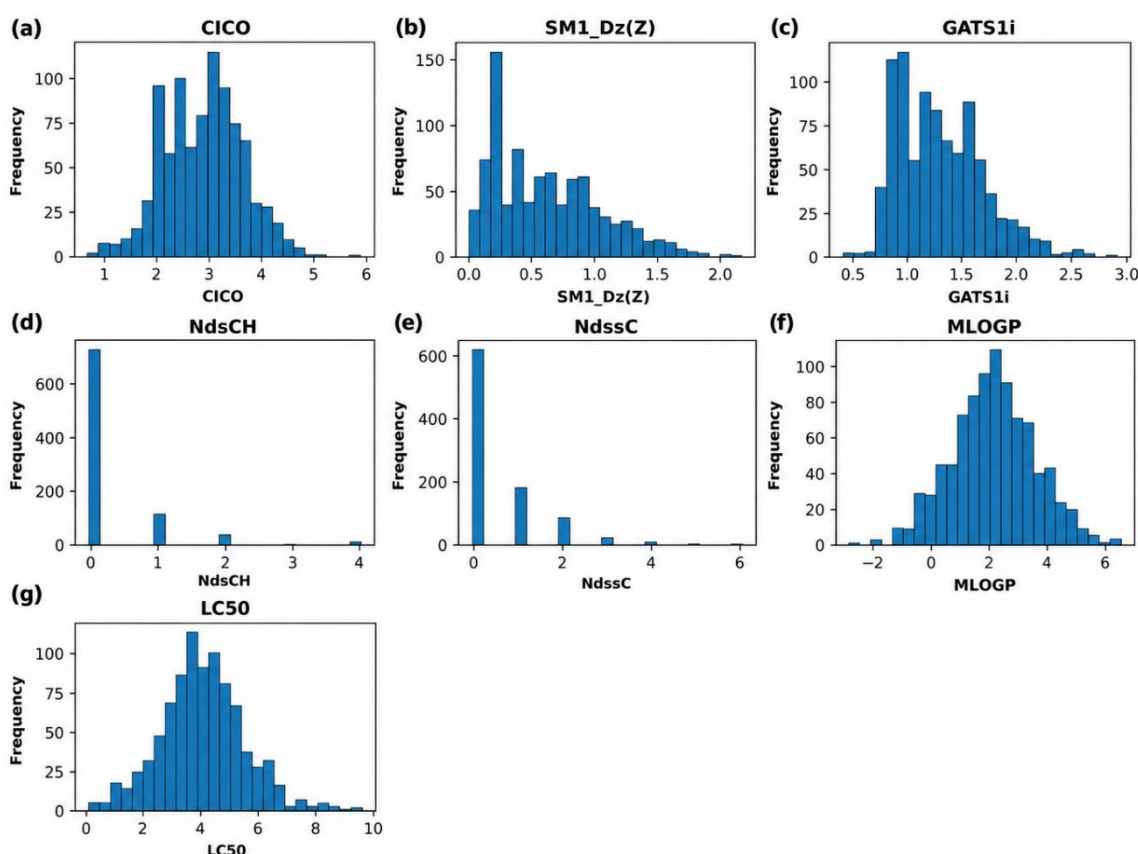


Figure 1. Distributions of molecular descriptors and LC50 [-log(mol/L)] across the analyzed compounds (a) CICO, (b) SM1_Dz(Z), (c) GATS1i, (d) NdsCH, (e) NdssC, (f) MLOGP, and (g) LC50 [-log(mol/L)]

3.2 Correlation and Structure–Property Relationships

Table 2 shows that there are several statistically significant correlations between the molecular descriptors and LC50 [-log(mol/L)] values obtained from the correlation analysis. The best bivariate correlation was found for MLOGP ($r = 0.652$, $p < .001$). Also, SM1_Dz(Z) showed a significant positive correlation with LC50 [-log(mol/L)] ($r = 0.411$, $p < .001$), while GATS1i exhibited a significant inverse correlation ($r = -0.398$, $p < .001$). The correlation was less pronounced, but significant, for CICO; $r = 0.292$, $p < .001$. NdsCH and NdssC showed relatively low and positive correlation with the toxicity response.

Table 2. Pearson correlations between molecular descriptors and LC50 [-log(mol/L)]

Descriptor	Pearson r	p-value	Direction
MLOGP	0.652	<.001	Positive

SM1_Dz(Z)	0.411	<.001	Positive
GATS1i	-0.398	<.001	Negative
CIC0	0.292	<.001	Positive
NdssC	0.172	<.001	Positive
NdsCH	0.172	<.001	Positive

Complete correlation structure is illustrated in Figure 2. The pattern suggests that the response is not due to a single structural feature, but due to several molecular properties.

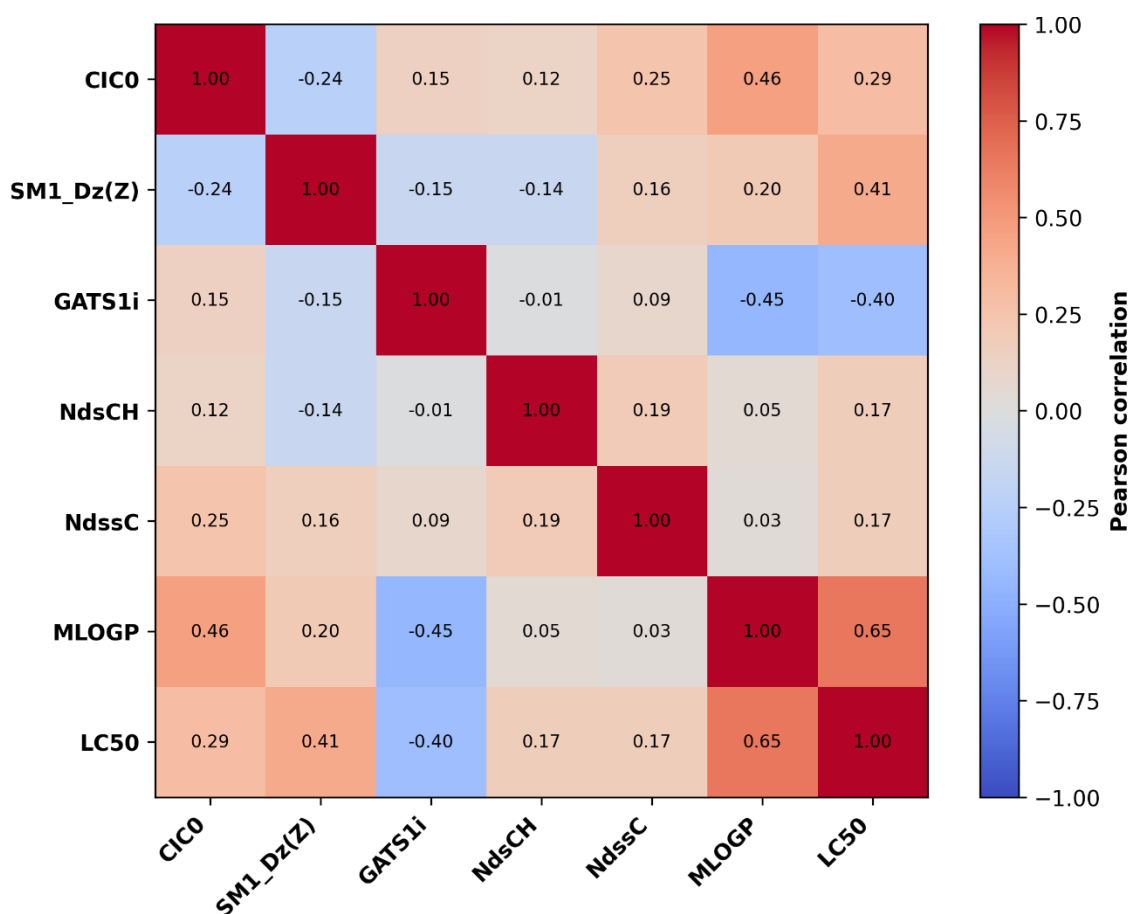


Figure 2. Pearson correlation heatmap of molecular descriptors and LC50 [-log(mol/L)]

A check for multicollinearity, however, revealed significant correlations between certain structural descriptors. The VIF values were 25.60 for CIC0, 14.48 for GATS1i, 7.43 for MLOGP, 3.46 for SM1_Dz(Z), 1.54 for NdssC, and 1.22 for NdsCH. These results justified the use of regularized and kernel-based models, besides the traditional multiple linear regression.

3.3 Multiple Linear Regression Analysis

Multiple linear regression revealed that the six molecular descriptors could account for a significant amount of variance in LC50 [-log(mol/L)]. The overall model was statistically significant, $F(6, 901) = 204.97$, $p < .001$, with $R^2 = 0.577$ and adjusted $R^2 = 0.574$. The linear combination of the six descriptors thus explained about 57.7% of the variance at the toxicity endpoint. Table 3 shows estimated regression coefficients and inferential statistics.

Table 3. Multiple linear regression coefficients for prediction of LC50 [-log(mol/L)]

Predictor	B	SE	t	p-value	95% CI
Intercept	2.175	0.181	12.000	<.001	1.819 to 2.530
CIC0	0.386	0.061	6.333	<.001	0.266 to 0.505

SM1_Dz(Z)	1.256	0.087	14.430	<.001	1.085 to 1.426
GATS1i	-0.746	0.101	-7.365	<.001	-0.945 to -0.548
NdsCH	0.414	0.054	7.644	<.001	0.307 to 0.520
NdssC	0.064	0.041	1.583	.114	-0.015 to 0.144
MLOGP	0.390	0.034	11.555	<.001	0.324 to 0.456

The other descriptors were adjusted simultaneously, and CIC0, SM1_Dz(Z), NdsCH, and MLOGP became significant positive predictors of LC50 [-log(mol/L)], whereas the remaining descriptor, GATS1i, remained significant negative predictor. NdssC was not statistically significant ($p = .114$). The statistical significance of SM1_Dz(Z), combined with that of MLOGP and GATS1i, led to the conclusion that significant multivariable structure–property relationships exist.

3.4 Comparative Performance of Classical and Kernel Regression Models

For an evaluation of predictive performance of eight regression specifications, they were split into 80% training data and 20% testing data, with hyperparameter tuning performed on the training data. Ten-fold cross validation was then performed to evaluate the stability of the model. Multiple linear regression, Ridge, LASSO, Elastic Net, linear support vector regression (SVR), polynomial SVR, radial basis function (RBF)-SVR and kernel ridge regression models were used. Table 4 summarizes the performance of the eight models over the two sets of data (the cross-validation and the held-out test set).

Table 4. Cross-validation and held-out test-set performance of regression models

Model	CV R ²	CV RMSE	Test R ²	Test RMSE	Test MAE	Test MSE
RBF-SVR	0.631	0.852	0.605	0.978	0.677	0.957
Linear SVR	0.554	0.934	0.567	1.024	0.731	1.048
Linear Regression	0.557	0.931	0.564	1.027	0.740	1.054
LASSO	0.557	0.931	0.564	1.027	0.740	1.054
Elastic Net	0.557	0.931	0.564	1.027	0.740	1.054
Ridge	0.557	0.931	0.564	1.027	0.741	1.055
Kernel Ridge	0.588	0.906	0.543	1.051	0.712	1.105
Polynomial SVR	0.158	1.286	0.126	1.454	1.037	2.114

RBF-SVR showed the best predictive results with a CV R² of 0.631 and test-set R² of 0.605 out of all the evaluated approaches. It also had the lowest test RMSE of 0.978 and test MAE of 0.677 of all the models evaluated. $C = 1$, $\epsilon = 0.1$ and gamma was scaled according to the RBF-SVR specification.

Classical linear model yielded test R² of 0.564, and RMSE of 1.027. The test performance results were nearly identical across the three regularization methods of ridge, LASSO and Elastic Net, indicating that the regularization didn't offer much benefit over ordinary linear regression for these six predictors. Linear SVR gave comparable results with R² = 0.567. The overall performance of RBF-SVR is better than linear regression model, as shown in the observed versus predicted plot in Figure 3. However, there were deviations for compounds at the ends of the LC50 [-log(mol/L)] range.

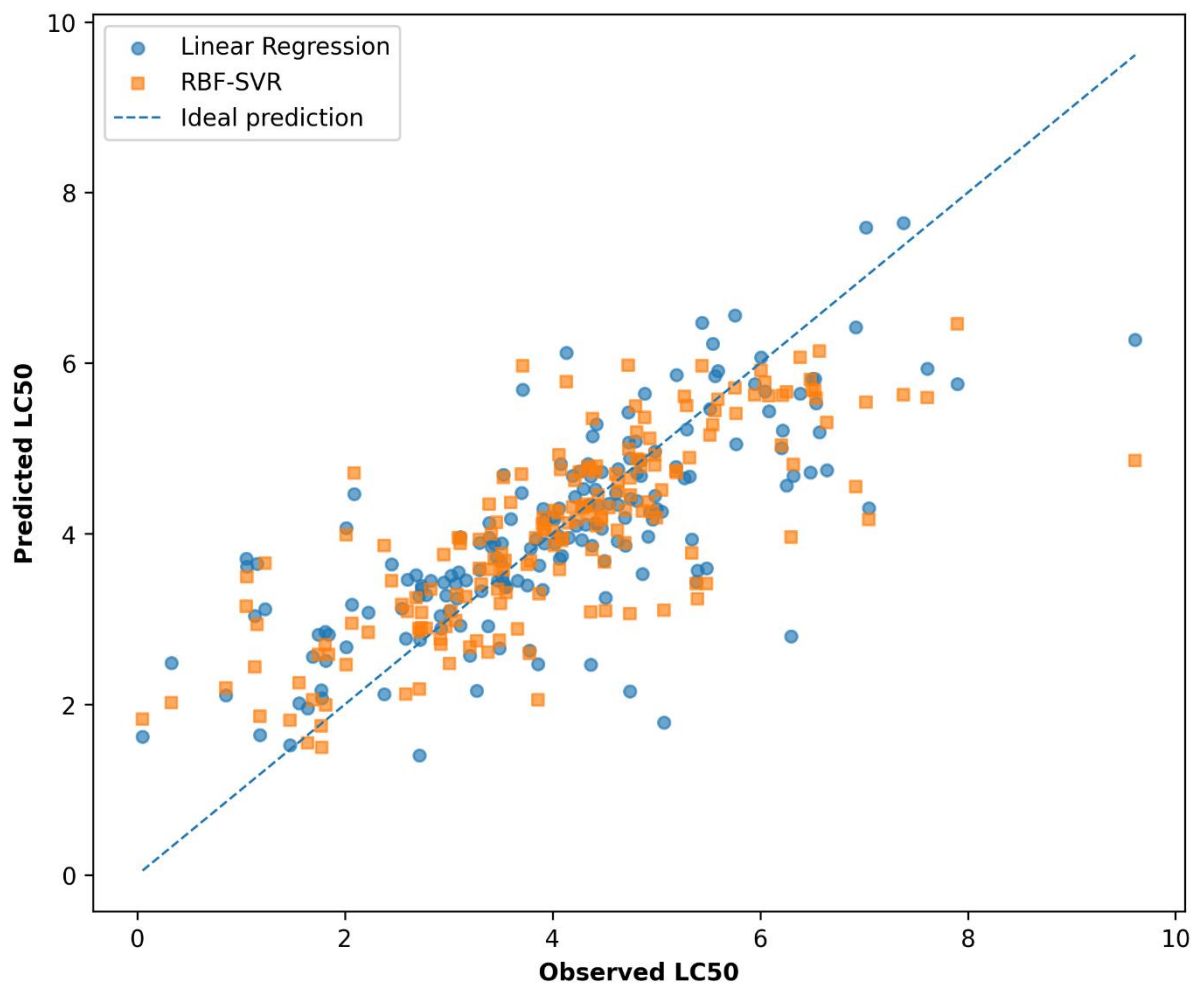


Figure 3. Observed versus predicted LC50 [-log(mol/L)] values for linear regression and RBF-SVR

The kernel ridge regression model had slightly better cross validation results than the classical models, however this did not translate to a better held out test sample performance. The least good predictions were obtained using polynomial SVR, which yielded test $R^2 = 0.126$ and $RMSE = 1.454$. The error profile for all models is compared in Figure 4.

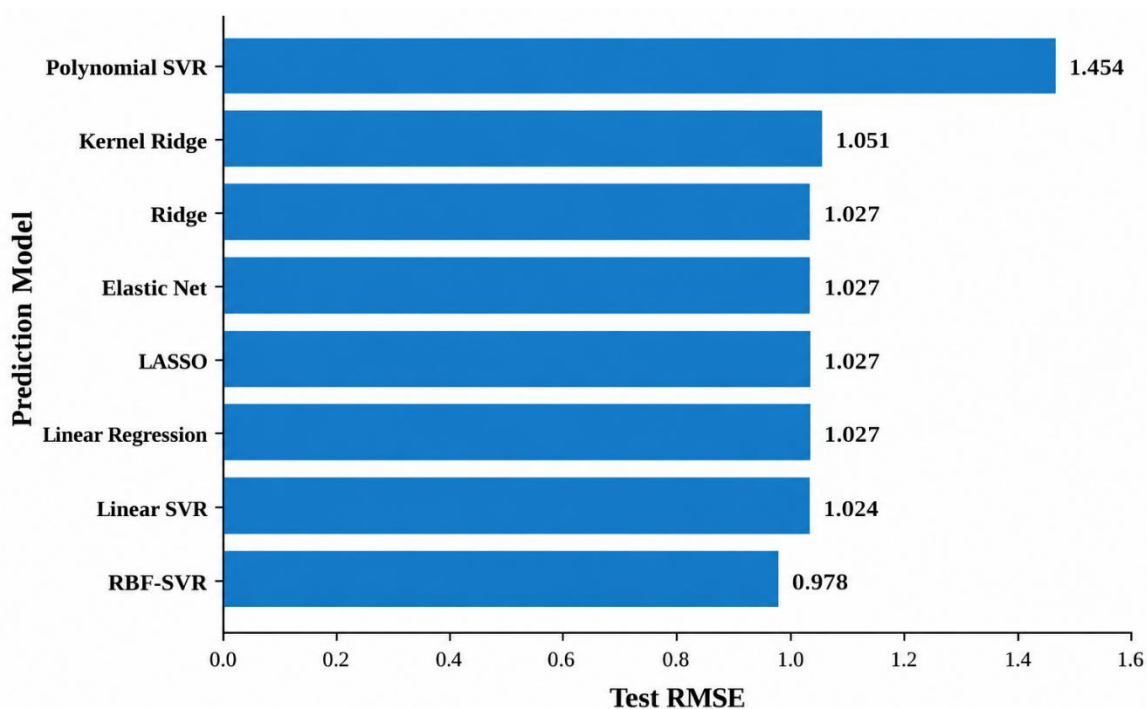


Figure 4. Comparison of test-set RMSE across the evaluated regression models

3.5 Residual Behaviour of the RBF-SVR Model

Since the RBF-SVR model had the lowest held-out test error, the focus of the residual analysis was on this model. The plot of residuals showed errors scattered about the zero-error reference line, with some larger positive and negative errors visible at some predicted LC50 $[-\log(\text{mol/L})]$ values (Figure 5). The model demonstrated its ability to capture the main nonlinearities in the descriptor space as there was no single dominant direction of error in the predictions. The six descriptors, however, do not fully account for the residual dispersion, as seen in the remaining dispersion.

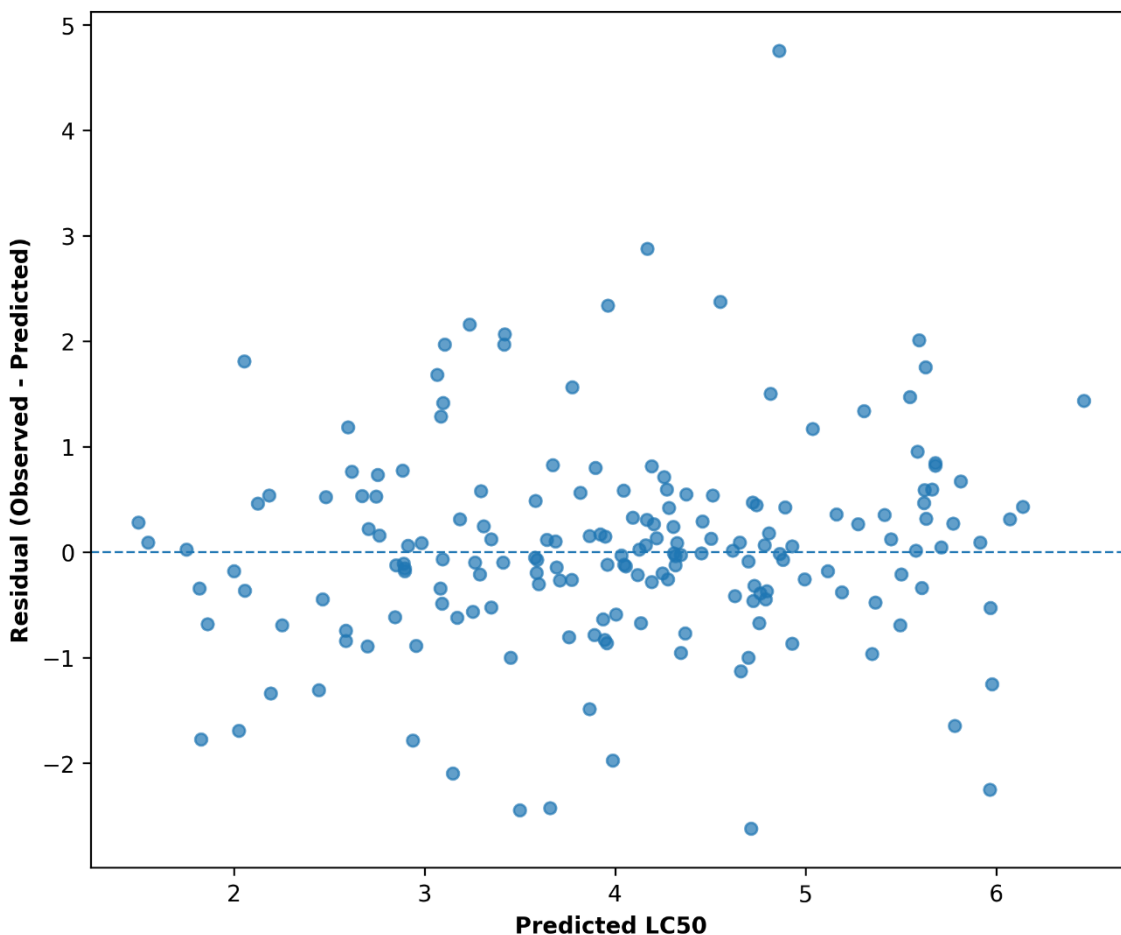


Figure 5. Residual distribution of the RBF-SVR model across predicted LC50 $[-\log(\text{mol/L})]$ values

3.6 Molecular Descriptor Contribution to Nonlinear Prediction

The contribution of each descriptor to out-of-sample predictive accuracy was quantified using permutation-based importance analysis of the RBF-SVR model. Two most influential predictors were MLOGP and SM1_Dz(Z), followed by CIC0, NdsCH, GATS1i and NdssC (Table 5).

Table 5. Permutation importance of molecular descriptors in the RBF-SVR model

Rank	Descriptor	Mean increase in RMSE	SD
1	MLOGP	0.318	0.035
2	SM1_Dz(Z)	0.300	0.037
3	CIC0	0.162	0.030
4	NdsCH	0.131	0.030
5	GATS1i	0.114	0.023
6	NdssC	0.032	0.014

The descriptors containing particularly important predictive information were MLOGP (with prediction error increase of about 0.318 RMSE units) and disruption of SM1_Dz(Z) (RMSE increase of about 0.300). There was also input from CIC0

and NdsCH for prediction. NdssC, however, resulted in a very small loss of performance when permuted, which is consistent with its relatively small independent contribution in the multiple regression model. The relative contribution of the six molecular descriptors is shown in Figure 6.

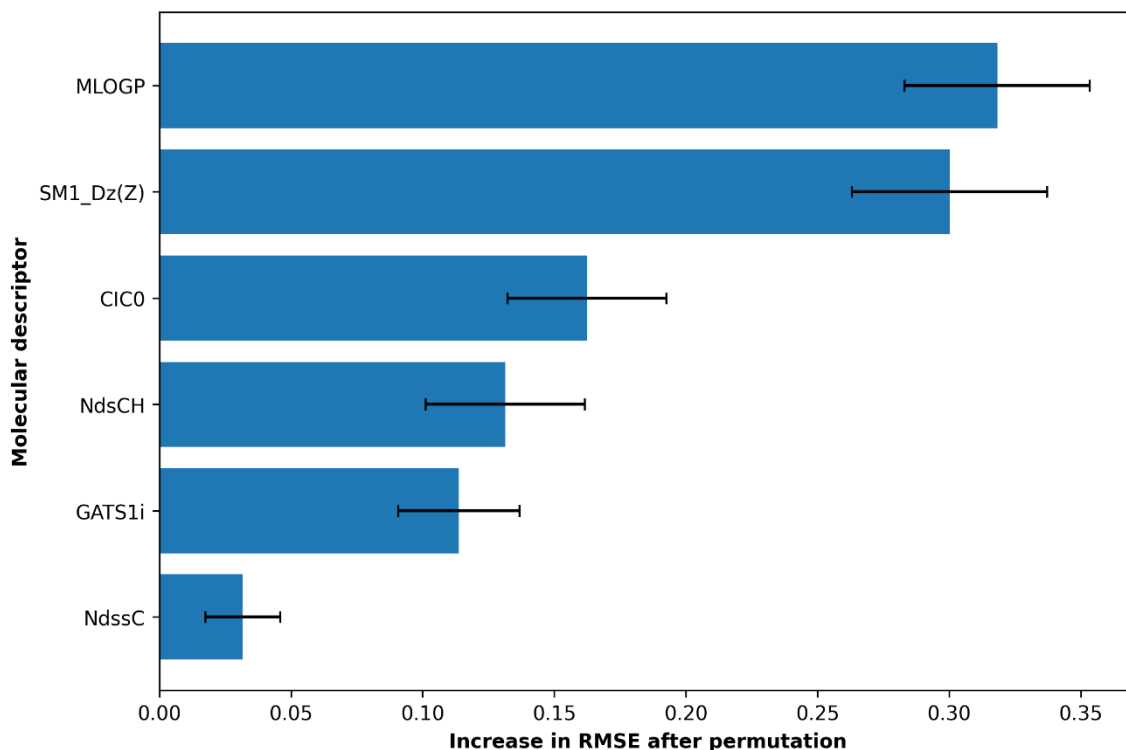


Figure 6. Permutation-based importance of molecular descriptors in the RBF-SVR model

3.7 Overall Structure–Property Modelling Pattern

The combination of the analyses allows one to elucidate a measurable and statistically significant correlation between molecular structural descriptors and LC50 [-log(mol/L)]. The overall variance was explained by classical multiple regression (57.7%), with the highest held-out test R^2 of 0.605 and the lowest prediction error among the eight models tested, being obtained by the RBF-SVR model. The improvement achieved by RBF-SVR demonstrates that the relationship between the descriptors and the response has a nonlinear part, which is not included in a typical additive linear model. In both inferential and predictive analyses, MLOGP was a consistent strong descriptor, whereas NdssC was comparatively weak in both its adjusted linear coefficient and permutation importance. In general, the results are encouraging for the use of kernel-based modelling as a complementary method to classical regression for quantitative structure–property relationship.

4. Discussion

This study showed that molecular descriptors held predictive information for LC50 [-log(mol/L)] and model form had a role in out-of-sample performance. Multiple linear regression accounted for 57.7% of the variance while RBF-SVR had the highest test performance with $R^2 = 0.605$ and RMSE = 0.978. This improvement suggests that the structure–toxicity relationship was not exclusively linear and that there is potential for nonlinear transformations to add predictive value. This result is consistent with the general QSAR literature, where machine-based learning has increasingly been coupled with traditional regression to model complex chemical responses (Mao et al., 2021; Martinez-Mayorga et al., 2024). When the effects of the descriptors cannot be well described by additivity, the superiority of RBF-SVR over linear regression, Ridge, LASSO, Elastic Net, linear SVR, polynomial SVR and kernel ridge supports nonlinear modelling is demonstrated. Support vector regression was also shown to be suitable for QSAR modelling by Shi (2021) and kernel methods to encode complex chemical relationships by Haywood et al. (2022). The obtained results are consistent with Mahmood et al. (2025) who found kernel ridge approaches useful in QSAR. However, in the present work, kernel ridge was not found to be superior to RBF-SVR. This difference highlights the fact that the predictive performance may vary depending on the choice of kernel, regularization and the size of the dataset and the structure of the descriptors.

There was little difference when comparing Ridge, LASSO and Elastic Net with ordinary linear regression, indicating that shrinking the coefficients alone did not yield significant predictive benefits. There was multicollinearity, mostly in the presence of CIC0 and GATS1i, but regularization did not change test accuracy. It has been demonstrated that statistical structure–property modelling is still useful when the descriptors maintain their relationship with molecular responses, which can be interpreted (Dai et al., 2020; Gao et al., 2021). Similarly, Abbod et al. (2024) showed that the multiple linear regression can be competitive when using descriptor selection or machine-learning approaches. The observations indicate that classical regression be kept as a simple yet interpretable baseline model, not as an indication of more robust prediction. Bivariate correlation analysis indicated that MLOGP had the highest positive correlation with LC50 [-log(mol/L)], and permutation analysis revealed that MLOGP and SM1_Dz(Z) were the most influential descriptors. The pattern highlights

the need of physicochemical and structural data in toxicity modelling. Similarly, Chen et al. (2020), Yu (2020) and Tinkov et al. (2020) have found that aquatic toxicity can be predicted by molecular structure derived properties. Gajewicz-Skretna et al. (2021) have shown that similarity based molecular information can augment the prediction of aquatic acute toxicity. Nath et al. (2023) noted that descriptor-based QSAR approaches are more suitable for estimating the toxicity of agrochemicals, while Yu and Zeng (2022) demonstrated the classification of pesticide toxicity to fish using machine-learning models.

It is also noteworthy that the contribution of $NdssC$ is comparatively weak. The regression coefficient of this was not significant and its permutation importance was low; this indicates that it did not contain useful information or part of it was contained in other descriptors. This is an example of the importance of interpretation of the descriptors taking into account both univariate associations and multivariable predictive contribution. Belfield et al. (2023) highlighted transparent interpretation and stringent validation when constructing toxicological QSARs, while Cronin et al. (2025) underscored the aspects of predictivity, applicability and generalization. Finally, Fuadah et al. (2024) showed that consensus and multivariate methods are useful tools for chemical hazard modelling, and that complementary methods should be compared.

Results are also relevant to the exploding development of data driven molecular prediction. Shilpa et al. (2023) discussed machine-learning applications for molecular properties and chemical reaction outcomes, while Tang et al. (2023) highlighted the use of graph-based message passing approaches. Pang et al. (2023) further examined advanced deep-learning methods for molecular property prediction. Chemical language models have also been increasingly applied to molecular and reaction prediction, as discussed by Das et al. (2024). Dobbelaere et al. (2024) investigated geometric deep-learning approaches across chemical space, whereas Sultan et al. (2024) reviewed transformer-based methods for molecular property prediction. Self-supervised molecular representation learning was demonstrated by Zeng et al. (2022), while Marimuthu and McGuire (2025) presented an automated machine-learning pipeline for molecular property prediction.

The results are relevant to computational screening/prioritization from an ecotoxicological perspective. In fish acute-mortality prediction, Gasser et al. (2024) pointed out the importance of interpretability and regulatory relevance whereas Sun et al. (2024) focused on interpretability for pesticide toxicity modelling. Zubrod et al. (2024) revealed that combining both chemical and biological data can increase the robustness of ecotoxicity predictions and Shi et al. (2024) illustrated the power of ensemble-based aquatic toxicity tools. Machine learning and QSAR models also have been widely applied in toxicological and chemical endpoints. Collins and Barton-Maclaren (2022) demonstrated their application in endocrine-disruption prediction, while Wu et al. (2024) used chemometric QSAR and machine-learning approaches for acute oral toxicity assessment. Quadri et al. (2021) further highlighted the broader utility of QSAR and artificial neural network methods for predicting chemical interaction and coordination behaviour. The results indicated that RBF-SVR model yielded the highest prediction accuracy among the analyzed models, and classical regression was also useful in the understanding of the relationship between the descriptors and the responses.

5. Conclusion

In this study regression-based modelling proved to be a useful tool for characterising and predicting structure–property relationships related to aquatic toxicity. The classical multiple linear regression model accounted for a large amount of variance in $LC50 [-\log(\text{mol/L})]$, $R^2=0.577$, thus the descriptors selected possessed meaningful statistical information. The RBF-SVR model had the highest predictive accuracy ($R^2 = 0.605$ and $RMSE = 0.978$) suggesting that the nonlinear relationships added more predictive information than the conventional linear models. Regularization alone was not sufficient for this set of descriptors, as the performance of Ridge, LASSO and Elastic Net was similar to that of ordinary linear regression.

The molecular descriptors with the largest predictive values were $MLOGP$ and $SM1_Dz(Z)$, and the remaining descriptors, $CIC0$, $NdsCH$ and $GATS1i$, also had some influence on the observed structure–toxicity relationship. Contrast this with the comparatively small independent predictive value of $NdssC$. The results presented here justify the application of a hybrid method based on interpretable classical regression and flexible kernel-based methods in quantitative structure–property modelling. Overall, RBF-SVR gave the best predictive accuracy for the current dataset, and linear regression was useful to understand the effects of individual descriptors. In future, larger and chemically diverse data sets and external validation samples should be used to validate those results.

References

1. Abbod, M., Safaie, N., & Gholivand, K. (2024). Genetic algorithm multiple linear regression and machine learning-driven QSTR modeling for the acute toxicity of sterol biosynthesis inhibitor fungicides. *Heliyon*, 10(16).
2. Ballabio, D., Cassotti, M., Consonni, V., & Todeschini, R. (2015). QSAR fish toxicity [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5JG7B>
3. Belfield, S. J., Cronin, M. T., Enoch, S. J., & Firman, J. W. (2023). Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs). *PLoS One*, 18(5), e0282924.
4. Chen, X., Dang, L., Yang, H., Huang, X., & Yu, X. (2020). Machine learning-based prediction of toxicity of organic compounds towards fathead minnow. *RSC advances*, 10(59), 36174-36180.
5. Collins, S. P., & Barton-Maclaren, T. S. (2022). Novel machine learning models to predict endocrine disruption activity for high-throughput chemical screening. *Frontiers in toxicology*, 4, 981928.
6. Cronin, M. T., Basiri, H., Chrysochoou, G., Enoch, S. J., Firman, J. W., Spînu, N., & Madden, J. C. (2025). The

- predictivity of QSARs for toxicity: Recommendations for improving model performance. *Computational Toxicology*, 33, 100338.
7. Dai, D., Liu, Q., Hu, R., Wei, X., Ding, G., Xu, B., ... & Zhang, H. (2020). Method construction of structure-property relationships from data by machine learning assisted mining for materials design applications. *Materials & Design*, 196, 109194.
8. Das, M., Ghosh, A., & Sunoj, R. B. (2024). Advances in machine learning with chemical language models in molecular property and reaction outcome predictions. *Journal of Computational Chemistry*, 45(14), 1160-1176.
9. Dobbelaere, M. R., Lengyel, I., Stevens, C. V., & Van Geem, K. M. (2024). Geometric deep learning for molecular property predictions with chemical accuracy across chemical space. *Journal of Cheminformatics*, 16(1), 99.
10. Fuadah, Y. N., Pramudito, M. A., Firdaus, L., Vanheusden, F. J., & Lim, K. M. (2024). QSAR Classification modeling using machine learning with a consensus-based approach for multivariate chemical hazard end points. *ACS omega*, 9(51), 50796.
11. Gajewicz-Skretna, A., Furuhashi, A., Yamamoto, H., & Suzuki, N. (2021). Generating accurate in silico predictions of acute aquatic toxicity for a range of organic chemicals: Towards similarity-based machine learning methods. *Chemosphere*, 280, 130681.
12. Gao, Y., Zhong, S., Torralba-Sanchez, T. L., Tratnyek, P. G., Weber, E. J., Chen, Y., & Zhang, H. (2021). Quantitative structure activity relationships (QSARs) and machine learning models for abiotic reduction of organic compounds by an aqueous Fe (II) complex. *Water research*, 192, 116843.
13. Gasser, L., Schür, C., Perez-Cruz, F., Schirmer, K., & Baity-Jesi, M. (2024). Machine learning-based prediction of fish acute mortality: implementation, interpretation, and regulatory relevance. *Environmental Science: Advances*, 3(8), 1124-1138.
14. Haywood, A. L., Redshaw, J., Hanson-Heine, M. W., Taylor, A., Brown, A., Mason, A. M., ... & Hirst, J. D. (2022). Kernel methods for predicting yields of chemical reactions. *Journal of Chemical Information and Modeling*, 62(9), 2077-2092.
15. Mahmood, S. W., Basheer, G. T., & Algamal, Z. Y. (2025). Quantitative structure-activity relationship modeling based on improving kernel ridge regression. *Journal of Chemometrics*, 39(5), e70027.
16. Mao, J., Akhtar, J., Zhang, X., Sun, L., Guan, S., Li, X., ... & Wang, G. (2021). Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. *Iscience*, 24(9).
17. Marimuthu, A. N., & McGuire, B. A. (2025). Machine learning pipeline for molecular property prediction using chemxploreML. *Journal of Chemical Information and Modeling*, 65(11), 5424-5437.
18. Martinez-Mayorga, K., Rosas-Jiménez, J. G., Gonzalez-Ponce, K., López-López, E., Neme, A., & Medina-Franco, J. L. (2024). The pursuit of accurate predictive models of the bioactivity of small molecules. *Chemical Science*, 15(6), 1938-1952.
19. Nath, A., Ojha, P. K., & Roy, K. (2023). QSAR assessment of aquatic toxicity potential of diverse agrochemicals. *SAR and QSAR in Environmental Research*, 34(11), 923-942.
20. Pang, C., Tong, H. H., & Wei, L. (2023). Advanced deep learning methods for molecular property prediction. *Quantitative Biology*, 11(4), 395-404.
21. Quadri, T. W., Olasunkanmi, L. O., Fayemi, O. E., Akpan, E. D., Verma, C., Sherif, E. S. M., ... & Ebenso, E. E. (2021). Quantitative structure activity relationship and artificial neural network as vital tools in predicting coordination capabilities of organic compounds with metal surface: a review. *Coordination Chemistry Reviews*, 446, 214101.
22. Shi, X. X., Wang, Z. Z., Wang, Y. L., Wang, F., & Yang, G. F. (2024). AquaticTox: A web-based tool for aquatic toxicity evaluation based on ensemble learning to facilitate the screening of green chemicals. *Environment & Health*, 2(4), 202.
23. Shi, Y. (2021). Support vector regression-based QSAR models for prediction of antioxidant activity of phenolic compounds. *Scientific reports*, 11(1), 8806.
24. Shilpa, S., Kashyap, G., & Sunoj, R. B. (2023). Recent applications of machine learning in molecular property and chemical reaction outcome predictions. *The Journal of Physical Chemistry A*, 127(40), 8253-8271.
25. Sultan, A., Sieg, J., Mathea, M., & Volkamer, A. (2024). Transformers for molecular property prediction: Lessons learned from the past five years. *Journal of Chemical Information and Modeling*, 64(16), 6259-6280.
26. Sun, T., Wei, C., Liu, Y., & Ren, Y. (2024). Explainable machine learning models for predicting the acute toxicity of pesticides to sheepshead minnow (*Cyprinodon variegatus*). *Science of The Total Environment*, 957, 177399.
27. Tang, M., Li, B., & Chen, H. (2023). Application of message passing neural networks for molecular property prediction. *Current Opinion in Structural Biology*, 81, 102616.
28. Tinkov, O. V., Grigorev, V. Y., Razdolsky, A. N., Grigoryeva, L. D., & Dearden, J. C. (2020). Effect of the structural factors of organic compounds on the acute toxicity toward *Daphnia magna*. *SAR and QSAR in Environmental Research*, 31(8), 615-641.
29. Wu, S., Li, S. X., Qiu, J., Zhao, H. M., Li, Y. W., Feng, N. X., ... & Li, Q. X. (2024). Accurate prediction of rat acute oral toxicity and reference dose for thousands of polycyclic aromatic hydrocarbon derivatives based on chemometric QSAR and machine learning. *Environmental Science & Technology*, 58(34), 15100-15110.
30. Yu, X. (2020). Quantitative structure-toxicity relationships of organic chemicals against *Pseudokirchneriella subcapitata*. *Aquatic Toxicology*, 224, 105496.
31. Yu, X., & Zeng, Q. (2022). Random forest algorithm-based classification model of pesticide aquatic toxicity to fishes. *Aquatic Toxicology*, 251, 106265.

32. Zeng, X., Xiang, H., Yu, L., Wang, J., Li, K., Nussinov, R., & Cheng, F. (2022). Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11), 1004-1016.
33. Zubrod, J. P., Galic, N., Vaugeois, M., & Dreier, D. A. (2024). Bio-QSARs 2.0: Unlocking a new level of predictive power for machine learning-based ecotoxicity predictions by exploiting chemical and biological information. *Environment International*, 186, 108607.

EPH – International Journal of Mathematics and Statistics

<https://ejjms.com/index.php/ms/index>



This work is licensed under a Creative Commons Attribution 4.0 International License. The authors retain ownership of the copyright for their article, You are free to: Share — copy and redistribute the material in any medium or format for any purpose, even commercially. To see the complete license contents,

please visit <http://creativecommons.org/licenses/by/4.0/>.