JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Predicting Toxicities of Diverse Chemical Pesticides in Multiple Avian Species Using Tree-Based QSAR Approaches for Regulatory Purposes

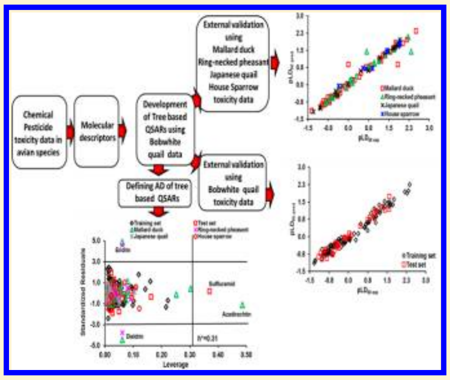
Nikita Basant,[†] Shikha Gupta,[‡] and Kunwar P. Singh*[,‡]

[†]Kan Ban Systems Pvt. Ltd., Laxmi Nagar, Delhi 110092, India

[‡]Environmental Chemistry Division, CSIR-Indian Institute of Toxicology Research, Post Box 80, Mahatma Gandhi Marg, Lucknow, Uttar Pradesh 226 001, India

S *Supporting Information*

ABSTRACT: A comprehensive safety evaluation of chemicals should require toxicity assessment in both the aquatic and terrestrial test species. Due to the application practices and nature of chemical pesticides, the avian toxicity testing is considered as an essential requirement in the risk assessment process. In this study, tree-based multispecies QSAR (quantitative-structure activity relationship) models were constructed for predicting the avian toxicity of pesticides using a set of nine descriptors derived directly from the chemical structures and following the OECD guidelines. Accordingly, the Bobwhite quail toxicity data was used to construct the QSAR models (SDT, DTF, DTB) and were externally validated using the toxicity data in four other test species (Mallard duck, Ring-necked pheasant, Japanese quail, House sparrow). Prior to the model development, the diversity in the chemical structures and end-point were verified. The external predictive power of the QSAR models was tested through rigorous validation deriving a wide series of statistical checks. Intercorrelation analysis and PCA methods provided information on the association of the molecular descriptors related to MW and topology. The S36 and MW were the most influential descriptors identified by DTF and DTB models. The DTF and DTB performed better than the SDT model and yielded a correlation (R^2) of 0.945 and 0.966 between the measured and predicted toxicity values in test data array. Both these models also performed well in four other test species ($R^2 > 0.918$). ChemoTyper was used to identify the substructure alerts responsible for the avian toxicity. The results suggest for the appropriateness of the developed QSAR models to reliably predict the toxicity of pesticides in multiple avian test species and can be useful tools in screening the new chemical pesticides for regulatory purposes.


Chemical Pesticide toxicity data in avian species
Molecular descriptors
Development of Tree based QSARs using Bobwhite quail data
External validation using Mallard duck Ring-necked pheasant Japanese quail House Sparrow toxicity data
External validation using Bobwhite quail toxicity data
Defining AD of tree based QSARs
Mallard duck
Ring-necked pheasant
Japanese quail
House sparrow
Training set
Test set
Standardized Residuals
Leverage
Sulfuramid
Azadirachtin
Dieldrin
h*=0.31

1. INTRODUCTION

The worldwide usage of chemical pesticides has increased exponentially during the last few decades, as these find enormous applications in various crop protection and disease vector control programs. More specifically, the developing and agriculture based countries consume much higher quantities of these chemicals. Chemical pesticides are the designed toxic molecules for specific purposes, and their unscientific application practices may cause hazards to nontarget species, including humans and ecosystems.[1] Due to the persistent, trans-boundary, and semivolatile characteristics of synthetic pesticides, a significant fraction of their applied dose can be detected over long periods as residues on the crops and soils, as well as in the atmosphere. Consequently, their high concentrations have been reported in vegetation, crops, and other edible products[2] and cause exposure to humans and animals. Long-term exposure to these chemicals is reported to adversely affect the nervous, endocrine, immune, reproductive, renal, cardiovascular, and respiratory systems in humans.[3] Moreover exposure to these chemicals has been reported to be the reason for the extinction of several avian and other animal species during the past decades.[1]

In view of the above, various regulatory agencies have been stressing for the toxicity evaluation of the existing as well as new chemical pesticides. Thus, for ensuring safety of the chemicals, there is a need for comprehensive toxicity testing that includes both the toxicity data on aquatic and terrestrial test species. However, most research efforts in the past have been directed to the aquatic toxicity testing, and only a few studies have been reported on the avian toxicity assessment of the chemicals.[4−7] The avian toxicity tests are required for the registration of the plant protection product active substances and in some regions also their formulated products. Data from these tests can be required for classification, labeling, and/or risk assessment.[8] For avian species, oral intake is considered as the most significant route of exposure, and, hence, the test for oral toxicity is considered important in determining the toxicological significance of any compound under investigation on avian species. In general, the northern Bobwhite quail (Colinus *virginianus*), Mallard duck (Anas *platyrhynchos*), Japanese quail, Ring-necked pheasant, and House sparrow are the common avian test species recommended by OECD[9] and EPA.[10]

Received: March 12, 2015
Published: July 9, 2015

In the case of the avian toxicity test, the 50% lethality from oral dose (LD_{50}) is considered appropriate. Although standard protocols for the experimental avian toxicity testing (LD_{50}) have been evolved,[11,12] these tests are expensive, cumbersome, and unethical. In terms of the animal numbers, long-term avian studies are the most animal intensive. The computational approach in chemical toxicology continues to be an attractive viable approach to reduce the amount of efforts and cost of experimental toxicity assessment[13] and provide suitable methods for early evaluation in the development of new chemicals.[14,15] The European Union (EU) regulation "Registration, Evaluation, Authorization and Restriction of Chemicals (REACH)"[16] advocates the use of nonanimal testing methods and in particular quantitative structure-toxicity/activity relationship (QSTR/QSAR) approaches in order to decrease the number and costs of animal testing. Moreover, the OECD[17] has provided a set of guidelines for development of QSARs.

Recently, a few studies have reported the qualitative[4,5] and quantitative[6,7] SARs for estimating the avian toxicities of the pesticides. However, all of these considered single individual avian test species for toxicity prediction of pesticides, and a comprehensive study of the mechanism of avian oral toxicity establishing its relationships with the chemical structures is still missing.[4] For a comprehensive risk assessment program, toxicity estimation of a chemical in a single test species will be inadequate, and the computational methods capable of predicting the expected toxicity of a compound in multiple test species (model developed using toxicity data in one test species and capable of predicting toxicity in other test species) can be a useful screening tool. To date no efforts have been made to develop QSARs for the estimation of toxicity of pesticides in multiple avian species.

In this study, we established tree-based QSAR models for predicting the toxicity of structurally diverse chemical pesticides in multiple avian test species following the OECD guidelines on the validation of QSAR Models.[17] Accordingly, the single decision tree (SDT), decision tree forest (DTF), and decision tree boost (DTB) regression QSAR models were constructed to predict the toxicity end-point (pLD_{50}) of the structurally diverse chemical pesticides in avian test species using a set of selected molecular descriptors as the estimators. The predictive and generalization abilities of the QSAR models constructed here were evaluated using several statistical criteria parameters, and the predictive power of these models was tested using external data sets, including those of other avian test species. The ChemoTyper[18] was applied to identify the privileged substructures that might be responsible for the avian toxicity.

2. MATERIALS AND METHODS

Here, QSAR models were established for the prediction of the toxicity of pesticides in five different avian species in accordance with the OECD principles.[17] Accordingly, the OECD guidelines were followed for the selection of a definite data set with defined end-point (principle 1) in multiple avian test species, an explainable model building strategy in view of the nature (linear, nonlinear) of the selected data set (principle 2), a defined applicability domain of the constructed models (principle 3), appropriate validation strategies corresponding to the goodness of fit, robustness, and predictivity of the QSARs (principle 4), and finally offering possible mechanistic interpretation of the developed models (principle 5).

2.1. Data Set. Oral toxicity data (LD_{50} mg/kg body weight) of chemical pesticides on different avian species were collected from the OPP Pesticide Ecotoxicity Database.[19] This database contained avian toxicities of 4768 compounds. For a given compound, if it has several data points for the same avian species, the most toxic one, namely the smallest LD_{50} value, was kept. Here, we considered toxicity data on five different species tested in 14-day oral gavage. In order to get a high quality data set, a rigorous screening process was applied here. All the mixtures, duplicates, salts, and the compounds that have only qualitative end-point values or have too little information to find the structure were removed. Finally, a total of 131 pesticides (rodenticide, miticide, microbicide, insecticide, herbicide, fungicide, avicide, piscicide, nematicide, growth regulators, etc.) for Bobwhite quail (Colinus *virginianus*) were retained for the QSAR analysis. The compounds in toxicity data set were further screened out, and chemicals that were uncommon in other species were selected. Accordingly, 46 pesticides in Mallard duck (Anas *platyrhynchos*), 27 pesticides in Ring-necked pheasant (Phasianus *colchicus*), 19 pesticides in Japanese quail (Coturnix *japonica*), and 9 pesticides in House sparrow (Passer *domesticus*) were retained and used as the external data set for multiple species QSAR modeling. Further, the avian toxicity data analysis revealed that eight of the compounds were common in all four external data sets. Prior to the QSAR analysis, the toxicity values were converted into the negative logarithmic scale ($-\log LD_{50}$, mmol/kg bw). For different test species, the end-point toxicity (pLD_{50}) values ranged between −1.32 and 2.37 (Bobwhite quail), −1.36 and 2.51 (Mallard duck), −0.84 and 2.33 (Ring-necked pheasant), −1.29 and 1.92 (Japanese quail), and −0.36 and 1.93 (House sparrow), respectively (SI Table 1, Supporting Information).

Table 1. Statistics of Selected Descriptors in the Bobwhite Quail Toxicity Data Set

descriptor	range	mean	SD[a]
MW	91.11−665.01	279.05	102.76
Nhet	0.00−21.00	6.11	3.03
Nsulph	0.00−9.00	2.44	1.57
PC4	1.00−109.00	34.65	21.24
TIAC	13.77−119.99	52.65	19.52
ISIZ	28.53−498.00	156.75	77.71
TPSA	0.00−157.33	49.34	25.41
S36	0.00−31.88	5.96	6.50
ETA_Beta_ns	0.00−18.00	6.24	3.71

[a] SD standard deviation.

The chemical pesticides toxicity data in multiple avian test species were analyzed statistically by generating the Box-Whiskers diagrams (Figure 1). These diagrams summarize each

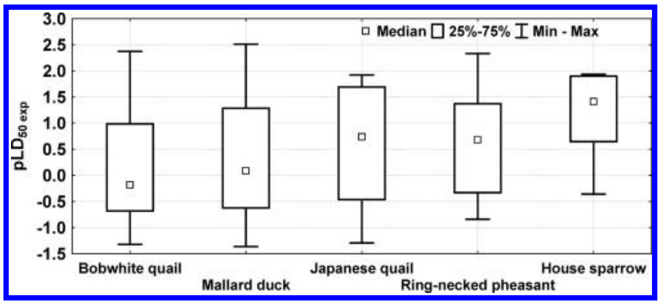


Figure 1. Box-whisker plots of the chemical pesticides toxicity values in five different avian test species.

toxicity data by a central point to indicate central tendency (median); a box to indicate variability around this central tendency (25th and 75th percentiles); and whiskers around the box to indicate the range of the data.

For developing a reliable QSAR model for the toxicity prediction of pesticides in multiple test species, the toxicity data in different species must be statistically different. A non-parametric Kruskal−Wallis (K−W) test was used to analyze the difference in the experimental toxicity data pertaining to five different test species considered here. The K−W test compares the residual value of each toxicity end-point among different test species with the significance level set at $p < 0.05$, which is suitable for comparison between different data sets with uneven sample numbers.[20] The K−W test estimates the value of χ^2 statistics for the experimental end-points and, therefore, evaluates the probability that samples were drawn from statistically different groups. When the value of χ^2 exceeds the critical value for the selected significance level, the two groups are deemed significantly different.[21]

2.2. Molecular Descriptors, Feature Selection, and Data Processing. Theoretical molecular descriptors for 232 pesticides were calculated by the ChemDes program.[22] A total of 316 2D molecular descriptors were calculated including the constitutional, topological, molecular properties, and E-state indices. The E-state indices are also a type of topological descriptor. These descriptors were calculated by 2D structures of the molecules, which were taken in the form of SMILES (simplified molecular input line entry system). SMILES and molecular weight (MW) for the pesticides were obtained using Chemspider.[23] Further, 43 ETA (Extended Topochemical Atom) descriptors[24,25] were also calculated using the freeware PaDEL descriptor software.[26]

The relevant features for QSAR analysis were selected using the model-fitting approach.[27] The descriptors exhibiting a constant or near constant value (variance <1) and those found to be correlated pairwise ($R > 0.95$) were excluded to decrease the redundant information in a preselection step. Multiple linear regression (MLR) and tree-based modeling were performed to further select the most relevant features. Prior to the model fitting, the Bobwhite quail data were split into the training (75%) and test (25%) subsets using the random distribution approach. The random approach provides an estimate of the variability of the model prediction accuracy. Such test sets (when defined prior to analysis) come close to the external validation set, which are commonly accepted as the gold standard to assess real predictivity.[28] The QSAR models were constructed with the training data using all the descriptors left (113) in the pool. The optimal values of the model parameters were determined using the respective scoring functions (mean squared error, MSE) to rank the contribution of features in the current data (training) by a 5-fold cross-validation (CV). The lowest ranked features (contribution <10%) were then removed in the successive steps. Finally a set of nine descriptors for QSAR analysis was considered in this study. Basic statistics of the selected descriptors are shown in Table 1.

2.3. Structural Diversity and Nonlinearity. Structural diverse compounds in the model building phase ensure the model generalizability. The structural diversity of the pesticides was assessed by the Tanimoto similarity index (TSI) based on the molecular descriptors. TSI is calculated using Tanimoto similarity between the fingerprint of a chemical and a consensus fingerprint, which is 1024 bit fingerprint (Toxmatch, Ideaconsult Ltd.). Smaller TSI (<0.8) means compounds have good diversity.[29] A histogram showing the distribution of the TSI values of the considered pesticides is given in Figure 2, which shows a sufficiently high diversity among the considered pesticides.

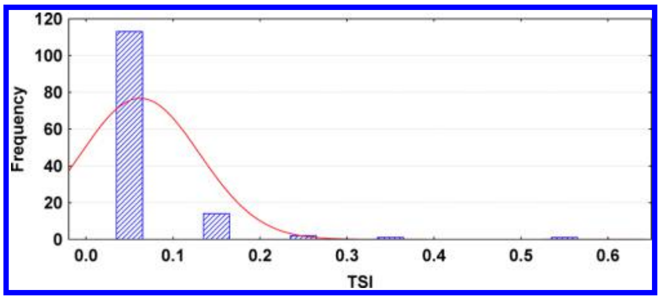


Figure 2. Histogram of the TSI values of complete toxicity data in Bobwhite quail species.

Nonlinearity in the experimental toxicity data was tested using the Brock-Dechert-Scheinkman (BDS) statistics.[30] It tests the null hypothesis of independent and identically distributed data against an unspecified alternative. If the computed BDS statistics exceed the critical value at the conventional level, the null hypothesis of linearity is rejected, revealing the nonlinear structure in the data.[31] In the present study, the BDS statistics exceeded the significance level ($p < 0.05$), thus suggesting for severe nonlinearity in data, and hence, a QSAR based on a nonlinear modeling approach will be required. The nonlinearity in data was also checked by constructing a linear QSAR model based on a multiple linear regression (MLR) approach, which establishes a linear relationship between the estimators and the end-point property.

2.4. Tree Based QSAR Modeling Methods. Three different decision tree based methods including the SDT, DTF, and DTB were used for QSAR model development for avian multispecies toxicity prediction. These methods have shown good performance in QSAR studies from previous works.[27,29,32−36] A brief account of these methods is provided here.

2.4.1. SDT-QSAR. SDT analysis can be used for regression problems and has significant features, including the ability to deal with collinear data, to exclude insignificant variables, and to allow asymmetrical distribution of samples.[37] It is comprised of nodes which represent a set of records from the original data set. Nodes may further be divided into the interior nodes and terminals or leaf nodes. SDT algorithms considered here use the minimum variance (least-squares criteria) within nodes for regression tree (line and function fitting). Overfitting of the training data may be prevented by the pruning technique that removes some branches of the tree after the tree is constructed. The maximum depth of tree and terminal (leaf) nodes are the method parameters which need to be optimized.

2.4.2. DTF-QSAR. A DTF is an ensemble of SDTs, in which a large number of independent trees are grown in parallel, and they do not interact until after all of them have been built. Variety in bagging is derived by using bootstrapped replicas of the original data. Different training subsets are drawn at random with replacement from the training data set. Separate models are produced and used to predict the entire data from aforesaid subsets. Then various estimated models are aggregated by using the mean for regression problems.[38] The

DTFs gaining strength from the bagging technique use the out of bag data rows for model validation. This provides an independent test set without requiring a separate data set or holding back rows from the tree construction. The number of trees in random forest and depth of individual trees are the method's parameter which needs to be adjusted for optimal model selection.

2.4.3. DTB-QSAR. DTB, an ensemble of SDTs, is considered as the most accurate modeling technique. The DTB combines the strengths of the regression tree and boosting algorithms. Boosting improves the accuracy of a prediction by applying a function repeatedly in a series and combining the output of each function with weighting and minimizes the total error.[39] The DTB algorithm uses randomization during the tree creations. Initially, a certain tree population is selected, and the first tree is fitted to the data. The residuals from the first tree are then fed into the second tree which attempts to reduce the error. This process is repeated through a chain of successive trees. The final predicted value is formed by adding the weighted contribution of each tree. The optimal size of the tree is decided using the criteria of minimal cross-validation error. The number and depth of the trees are the method's parameter which can be adjusted for a data set at hand. It controls the maximum allowed level of interaction between the variables in the model.

2.5. Model Validation and Applicability Domain Analysis. According to the OECD principles (3 and 4), the model validation and applicability domain (AD) analysis are important issues. Here, the regression QSAR models were developed using the training set (Bobwhite quail), while keeping the test data for external validation of the constructed models. The optimal architecture, model parameters of the QSAR models (SDT, DTF, DTB), and the number of relevant descriptors were determined following the V-fold CV procedure[27] using the criterion of minimum MSE values.[29] The advantage of this method (CV) is that it performs reliable and unbiased testing on the data set.[34]

Adequacy of the selected descriptors in the constructed QSARs was checked through the Y-randomization test. It determines the possibility of chance of correlation during the descriptor selection procedure.[29] In Y-randomization, the dependent variable (pLD_{50}) was randomly shuffled, and new models were built using the original independent variables. The procedure was repeated a number of times, and CV statistics were computed. The performance (R^2) of the new QSAR was compared with the original models. If the new models have a lower R^2 values for several trials, then the given model is not due to chance correlation.

Further, QSAR validation using an external data set provides information about the predictive ability of the trained model for the unknown data.[35] For external validation, a separate subset of the data was used which was kept out during the training process.[27] The predictive power of the constructed QSAR model for a new set was evaluated using various validation criteria parameters.[40] The concordance correlation coefficient (CCC) for external validation[41] without involving training data information measures the precision and accuracy and is considered a true external validation measure independent of the sampled chemical space. Q^2_{F1} uses the average of the training data, instead of that of the prediction set.[42] Criterion of Q^2_{F2} for external validation[43] differs from the earlier one because the average value at the denominator is calculated using the prediction set instead of the training one. In another criterion[44] based on Q^2_{F3} the denominator is calculated on the training set, and both numerator and denominator are divided by the number of corresponding elements. The results obtained by Q^2_{F3} are independent of the prediction set distribution and sample size.[44] The metric[45] r^2_m is calculated based on the correlations between the observed and predicted values with intercept (r^2) and (r^2_o) without an intercept for the least-squares lines. This metric avoids overestimation of the quality of prediction due to a wide response range (Y-range). The performance of the proposed QSAR model in predicting the toxicity of pesticides was also assessed by calculating R^2 and the root mean squared error (RMSE) for the training, test, and external data arrays.

Leverage is one of the standard methods for the analysis of the applicability domain (AD) of the model. The leverage value, h_i for each i^{th} pesticide, is calculated from the descriptor ($i \times j$) matrix ($\mathbf{X}$) as $h_i = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$, where $\mathbf{x}_i$ is a raw vector of molecular descriptors for a particular i^{th} compound. The limit of X-outliers is determined by their critical leverage value (h^*) calculated[46] as $h^* = (3(p + 1))/n$, where p is the number of variables used in the model, and n is the number of training compounds. The value of $h_i > h^*$ indicates that the structure of the compound substantially differs from those used for the calibration. Therefore, the compound is located outside the optimum prediction space.

3. RESULTS AND DISCUSSION

The mean rank values of 103.77 (Bobwhite quail), 121.22 (Mallard duck), 135.47 (Japanese quail), 137.56 (Ring-necked pheasant), and 174.44 (House sparrow) yielded in K−W statistics ($p < 0.05$) much higher than the critical χ^2 value (15.82), which suggest that the toxicity end-point values in five different avian populations are significantly different. This may be due to the fact that none of the compounds in the Bobwhite quail data was common with any of the other four avian species data. Moreover, among the four data sets (Mallard duck, Japanese quail, Ring-necked pheasant, House sparrow), only eight compounds were common, which (except dieldrin) have close toxicity values on four different avian species. Zhang et al.[5] also reported similar toxicities of most chemicals on different avian species.

A nonlinear dependence, structural diversity of the pesticides considered in five different toxicity data sets, and results of the K−W test suggest that the selected toxicity data in five avian test species are suitable for the development of reliable and robust predictive models for estimating the toxicities of new untested pesticides for regulatory purpose.

3.1. QSAR Modeling and Evaluation. External validation of the constructed QSAR models was performed using an independent test set of the model building species (Bobwhite quail) and the toxicity data pertaining to four other avian test species. In CV, the average MSE in the training and CV data sets were 0.09, 1.16 (SDT), 0.08, 0.71 (DTF), and 0.02, 0.76 (DTB), respectively. Further, a low R^2 value of 0.005 (SDT), 0.005 (DTF), and 0.001 (DTB) in Y-randomization revealed that the original tree-based QSAR models are unlikely to arise as a result of chance correlation. These results indicate that the constructed models showed no overfitting of the data.

The optimal SDT model has maximum depth of the tree, total number of group splits, and terminal (leaf) nodes of 7, 26, and 27, respectively. The DTF and DTB models have a total number of trees in series, maximum depth, and the average number of group splits in each tree: 322, 570; 20, 6; and 60.7, 98.9, respectively. The finally selected QSAR models (SDT, DTF, and DTB) in training captured 80.46%, 87.07%, and 95.89% of the data variance, respectively. The proportion of the variance captured by the model is a measure of the closeness of the predicted and actual values of the response. The variance explained by a model can be calculated by the regression sum of squares (SS_{reg}) of the data. The corresponding R^2 is the ratio of the explained variance to the total variance in data. The three tree-based QSAR models in training and test data yielded RMSE and R^2 values of 0.43, 0.805 and 0.48, 0.760 (SDT); 0.35, 0.935 and 0.33, 0.945 (DTF); and 0.20, 0.972 and 0.17, 0.966 (DTB), respectively. The model is considered acceptable when the value of R^2 in external set exceeds 0.6.[47] The RMSE is a quadratic scoring rule which measures the average magnitude of error.

The appropriateness of the selected nonlinear modeling methods in this study was further verified through constructing the MLR based QSAR model using the Bobwhite quail toxicity data and a set of six descriptors (Table 2). The MLR approach identified a set of descriptors different (except S36) from those of the tree methods. The optimal MLR model yielded the R^2 and RMSE values of 0.405, 0.75 (training) and 0.714, 0.48 (test), respectively. This suggests that there exists a nonlinear relationship between the descriptors and the end-point toxicity. These results are consistent with those of the BDS statistics. Such a mode of relationship between the end-point and the descriptors may represent the results of numerous physiological processes in biological systems including kinetic control of transport of substances, equilibrium control of their distribution, stearic factor, different pharmacokinetics, metabolism, solubility, and other factors. From this viewpoint, nonlinear relationships should be considered as a common platform during QSAR modeling.[48]

Table 2. Selected Descriptors in Bobwhite Quail Toxicity Data Set for MLR Analysis

descriptors	range	description
radiust	2.00−9.00	radius based on topology
Smin	[−8.81]−1.31	the minimal Estate value in all atoms
Shev	13.72−166.92	the sum of the EState indices over all non-hydrogen atoms
S34	0.00−26.62	sum of E-state of atom type: sOH
Smin37	0.00−13.62	minimum of E-state value of specified atom type
S36	0.00−31.88	sum of E-state of atom type: ssO

Further, the robustness of the nonlinear QSARs is better understood through the linear graphical representation between the actual and model predicted end-point values of the chemical pesticides in training and test data (Figure 3). The linear graphical representation shows the extent of variation between the actual and predicted end-point values of the data set. A closely followed pattern of variation by the measured and model predicted toxicities of the chemical pesticides by the constructed QSARs in the training and test data suggest that all three models performed reasonably well.

Further, according to the OECD guidelines (principle 4), a predictive QSAR model should be rigorously tested for its robustness using a set of stringent statistical tests, so that it could be applied to new untested data. The goodness of fit of the developed QSAR models here was judged by the quality metric R^2, and the predictivity for new chemicals was verified using the external validation metrics, such as CCC, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, and r^2_m. The values of these test coefficients for the training and the test data arrays are presented in Table 3.

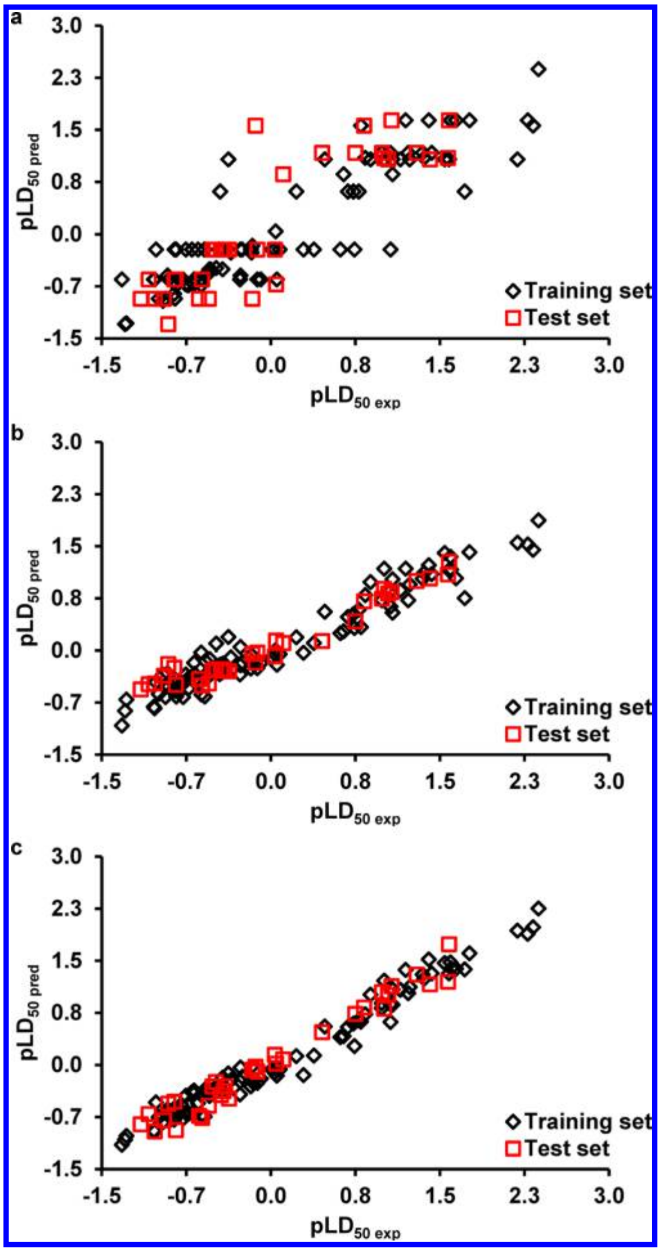


Figure 3. Plot of the measured and model predicted values of the pesticides toxicities in (a) SDT-QSAR, (b) DTF-QSAR, and (c) DTB-QSAR models in the training and test sets.

A threshold criteria of 0.85 for CCC, 0.7 for Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}, and 0.65 for r^2_m as an indicator for the acceptability of the QSAR models have been proposed in the literature.[49] From Table 3, it is evident that all the validation metrics for the developed QSAR models bore values that lie within the acceptable limits (except Q^2_{F1} and Q^2_{F2} in SDT). Identical values for the R^2 metrics in training and test sets indicate that the test set selected for the QSAR model development had a similar distribution of responses as the training set. Eriksson et al.[50] suggested that for an acceptable QSAR model, the difference between R^2 values in the training and test data should be within 0.3. In our case, these values were well within the suggested limit. Thus, the predictive potential of the developed QSAR

Table 3. Performance Parameters for the QSAR Models for Bobwhite Quail Toxicity Data

data set/model	R^2	RMSE	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC	r^2_m
SDT-QSAR							
training set	0.805	0.43					
test set	0.760	0.48	0.677	0.672	0.753	0.858	0.667
DTF-QSAR							
training set	0.935	0.35					
test set	0.945	0.33	0.848	0.846	0.884	0.894	0.808
DTB-QSAR							
training set	0.972	0.20					
test set	0.966	0.17	0.959	0.958	0.968	0.977	0.905

Table 4. Performance Parameters of the QSAR Models Applied to the External Data Sets

data set/model	species	R^2	RMSE	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC	r^2_m
DTF-QSAR								
external set-I	Mallard duck	0.924	0.41	0.857	0.850	0.822	0.898	0.754
external set-II	Ring-necked pheasant	0.918	0.36	0.881	0.853	0.859	0.902	0.695
external set-III	Japanese quail	0.962	0.31	0.934	0.923	0.899	0.952	0.782
external set-IV	House sparrow	0.956	0.30	0.944	0.834	0.905	0.881	0.447
DTB-QSAR								
external set-I	Mallard duck	0.962	0.22	0.959	0.957	0.949	0.976	0.919
external set-II	Ring-necked pheasant	0.940	0.24	0.949	0.937	0.939	0.966	0.881
external set-III	Japanese quail	0.994	0.10	0.993	0.992	0.989	0.996	0.963
external set-IV	House sparrow	0.976	0.15	0.985	0.956	0.975	0.974	0.816

models in terms of internal and external validation tests is reflected in the acceptable values of the metrics.

From the statistical test results, it may be noted that the three tree-based QSAR models performed well; however, the DTF and DTB performed better than SDT, because they are combinatorial models. It may be attributed to the fact that DTF and DTB models incorporate bagging and stochastic gradient boosting algorithms, respectively.[51] In bagging and boosting multiple version of SDTs are formed by making bootstrapped replicas of the learning set and subsequently using these as new learning sets. These models can inherit almost all advantages of tree based models while overcoming their primary problem of inaccuracy.[52,53] A better performance of these models in toxicity prediction of chemicals has also been reported earlier.[27,33−36] A rigorous external validation made in our models is an additional indication that these bear good predictive abilities. Therefore, both the DTF and DTB models developed using the Bobwhite quail toxicity data were further applied to predict the toxicities of diverse chemical pesticides in four other avian test species. These models can guide the researchers to design new pesticide molecules by increasing or decreasing the values of the descriptors for new molecules in the desired direction. Further, it may be mentioned that since the optimal architectures of the proposed models have been established, these can be re-established by the user and applied in toxicity predictions.

3.2. Multispecies QSAR Modeling. For a comprehensive safety assessment, toxicity estimation of chemicals in different test species has been advocated by various regulatory agencies. A QSAR model developed using toxicity data of a single particular test species and capable of predicting the toxicities of structurally diverse chemicals in other test species would be considered as an important tool as it could reduce the cost and efforts. The DTF and DTB models developed here using the Bobwhite quail toxicity data of pesticides were applied to predict the toxicities of pesticides in other avian test species (Mallard duck, Ring-necked pheasant, Japanese quail, and House sparrow). The real usefulness of QSAR models is not just their ability to reproduce known data, verified by their fitting power (R^2), but mainly their possibility of predictive application in unknown data. The average TSI values for the chemical pesticides in these data sets (Mallard duck, Ring-necked pheasant, Japanese quail, and House sparrow) were 0.063, 0.043, 0.048, and 0.046, respectively. These values suggest that the chemical moieties in these four data sets were structurally diverse. The DTF and DTB models applied to these species (Mallard duck, Ring-necked pheasant, Japanese quail, and House sparrow) yielded a low RMSE and high correlations (R^2) between the experimental and predicted toxicity values of 0.41, 0.924 and 0.22, 0.962 (Mallard duck); 0.36, 0.918 and 0.24, 0.940 (Ring-necked pheasant); 0.31, 0.962 and 0.10, 0.994 (Japanese quail); and 0.30, 0.956 and 0.15, 0.976 (House sparrow), respectively. Further, the external validation metrics derived for these species using two QSAR models are provided in Table 4. It is evident that the values of all the coefficients for both the models in all the species (except r^2_m for DTF model in House sparrow) were within their acceptable limits. The measured and model predicted toxicity values of the chemical pesticides in multiple avian test species by the constructed QSAR models (Figure 4) suggest that both the models (DTF, DTB) performed well.

Further analysis of the QSAR modeling (DTB) results revealed that the predicted toxicity (pLD_{50}) values of none of the compounds in Bobwhite quail and only two compounds (dieldrin, endrin) in Mallard duck and Ring-necked pheasant exhibited residuals of >0.5. Dieldrin and endrin are both organochlorine insecticides. Insecticides cause break down of neuronal activity causing brain death or affect the motor system through paralysis, convulsions, hyperactivity, and spasms.[54] Organochlorines are distributed more easily in air than in water, and hence they are more toxic to wetland bird (Mallard duck). The anomalous behavior of these compounds could be due; the

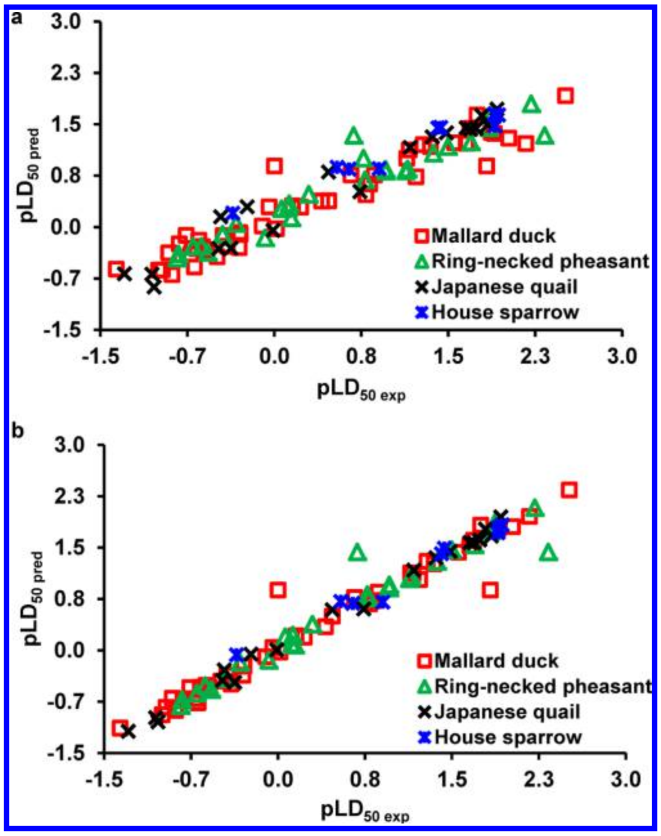


Figure 4. Plot of the measured and model predicted values of the pesticides toxicities in four different avian test species using (a) DTF-QSAR and (b) DTB-QSAR models.

descriptors selected do not capture some relevant structural features present in these compounds; and their biological mechanism is different from the remaining pesticides.[55]

3.3. Interpretation of the Descriptors.

Here, the tree-based QSAR models for predicting the avian toxicities of pesticides were constructed using nine constitutional, topological, E-state, and ETA descriptors (MW, Nsulph, Nhet, S36, PC4, TIAC, ISIZ, TPSA, ETA_Beta_ns). The principle 5 of the OECD guidelines emphasizes that a QSAR model should be capable of providing mechanistic interpretation establishing association between the descriptors used in the model and the end-point being investigated. The inter-relationships among the descriptors were investigated using the intercorrelations and principal components analysis (PCA) approaches. It is evident (Table 5) that all the descriptors exhibited significant ($p < 0.05$) (except S36 with PC4, TPSA, and ETA_Beta_ns) positive correlations. Descriptors (MW, PC4, ISIZ, TIAC) which are related with the molecular composition had high significant positive intercorrelations, whereas other descriptors (Nhet, Nsulph, and TPSA) related with surface topology exhibited high significant intercorrelations. The PCA was performed on a standardized data matrix, and three significant principal components (PCs) together explained >80% of the data variance. PC1 (39.0%) has high positive loadings on MW, PC4, TIAC, and ISIZ and moderate positive loadings on ETA_Beta_ns. PC2 (19.0%) exhibited high positive loadings on S36, whereas in PC3 (22.1%), TPSA has high positive loadings, and both the Nhet and Nsulph showed moderate positive loadings. A 3D plot of the PCs (Figure 5) revealed visibly separate groupings of the descriptors as observed in the intercorrelation study.

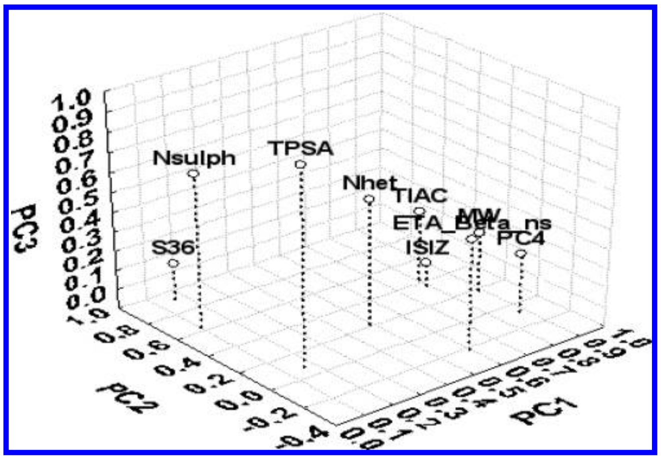


Figure 5. A 3D plot of the PCA loadings.

All the selected descriptors exhibited a significant ($p < 0.05$) (except PC4, TPSA) positive correlation (except ETA_Beta_ns) with pLD_{50} values of the considered chemical pesticides (Bobwhite quail). Since the most toxic compounds are characterized by low LD_{50} (and respectively high pLD_{50}) values, the eight descriptors having a positive correlation with end-point increase the adverse effect (toxicity). ETA_Beta_ns, which exhibited a negative correlation with pLD_{50}, is an extended topochemical atom descriptor. It is calculated from all the π bonds and electron loan pairs in the molecule and is a measure of the electron richness (unsaturation) of the molecule.[56] A negative correlation of ETA_Beta_ns with pLD_{50} suggests that the electron richness in a molecule causes a decrease in the toxicity. The contributions of each of the

Table 5. Correlation Matrix of Selected Descriptors in Bobwhite Quail Data Set[a]

descriptors	ETA_Beta_ns	MW	nhet	nsulph	PC4	TIAC	ISIZ	TPSA	S36
ETA_Beta_ns	1.000								
MW	**0.525**	1.000							
Nhet	**0.291**	**0.727**	1.000						
Nsulph	**0.285**	**0.358**	**0.481**	1.000					
PC4	**0.587**	**0.811**	**0.532**	**0.249**	1.000				
TIAC	**0.489**	**0.799**	**0.602**	**0.499**	**0.649**	1.000			
ISIZ	**0.388**	**0.673**	**0.290**	**0.326**	**0.593**	**0.893**	1.000		
TPSA	**0.433**	**0.295**	**0.507**	**0.643**	**0.285**	**0.478**	**0.327**	1.000	
S36	0.088	**0.340**	**0.312**	**0.650**	0.126	**0.496**	**0.376**	0.130	1.000

[a] Bold values are significant at $p < 0.05$.

selected descriptors in three QSAR models constructed here are shown in Figure 6. It is evident that in both the DTF and DTB models the MW and S36 were the most significant descriptors and the Nsulph as the least. MW, a molecular bulkiness descriptor, represents the molecular size, which is an important factor influencing the diffusion in biological membrane and continuous fluid media.[57] Since it has a positive correlation with toxicity (pLD_{50}) in Bobwhite quail, a compound with higher MW would render high biological activity. Low MW weight compounds would render lower toxicity due to their volatile nature. S36, a sum of E-state of atom type (−O−) descriptors, are topological indices. These indices are sensitive to one or more structural features of the molecule, such as size, shape, symmetry, branching, and cyclicity and can also encode chemical information concerning atom type and bond type multiplicity.[58] A positive correlation of S36 with toxicity (pLD_{50}) suggests for increased toxicity of a compound with increasing value of this descriptor. TPSA is defined as the part of the surface area of the module associated with N, O, S, and H bonded to any of these atoms. It has been shown to correlate with the molecular transport properties, such as collections of molecules or whole virtual combinatorial libraries.[59] TIAC and ISIZ are topological information index which represent the atomic composition and molecular size. These indices have several advantages, such as unique representation of the compound and high discriminating power (isomer discrimination). These descriptors have successfully been used in biological activity modeling.[60,61] Nsulph and Nhet are constitutional descriptors and represent count of sulfur and heteroatoms in the molecule. PC4 is also a constitutional descriptor which accounts for the molecular path count of length 4. Constitutional descriptors capture property of the molecule that is related to element constituting its structure. The molecular descriptors used for developing the avian toxicity prediction QSARs here encode structural information and thus implicitly account for determining the biological activity of molecules.

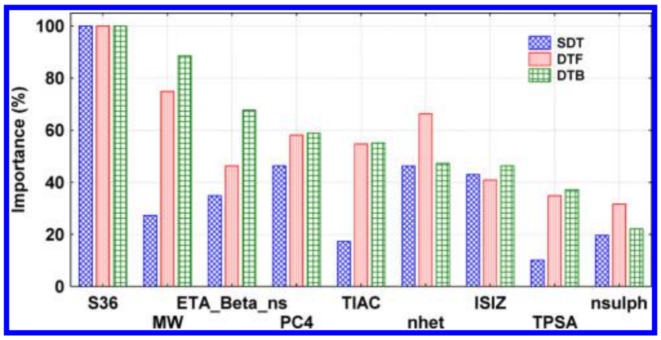


Figure 6. Plot of the contributions of the input variables in tree-based QSAR models for avian toxicity prediction.

3.4. Applicability Domain Analysis.

The ADs of the developed QSAR models were defined using the leverage approach. The graph which has been considered here for representing the AD of the QSAR models is referred to as the Williams plot (Figure 7).

Analysis of the Williams plots revealed that one compound (Sulfluramid) in Bobwhite quail test set and one compound (Azadirachtin) in the Mallard duck data exceeded the critical leverage value (0.31). The numerical value of leverage has certain advantages: the value is always greater than 0; the lower the value, the higher is the confidence in the prediction. A leverage value greater than a critical value indicates that the predicted response is the result of substantial extrapolation of the model.[62] On the other hand, in the SDT model, five compounds (bromethalin, methiocarb, 1,3-dichlopropene, coumaphos, and triclopyr butoxyethyl ester) exhibited high standardized residual (>3) in Bobwhite quail data. Bromethalin, a single dose rodenticide, is classified as very highly toxic if absorbed through the skin. In DTB, two compounds (dieldrin and endrin) each in Mallard duck and Ring-necked pheasant data were found to exceed the standardized residual value. However, in DTF, none of the compounds exceeded the standardized residuals limit. The outliers here are mainly chlorinated compounds, carbamates, and organophosphate. These compounds typically have several functional groups and

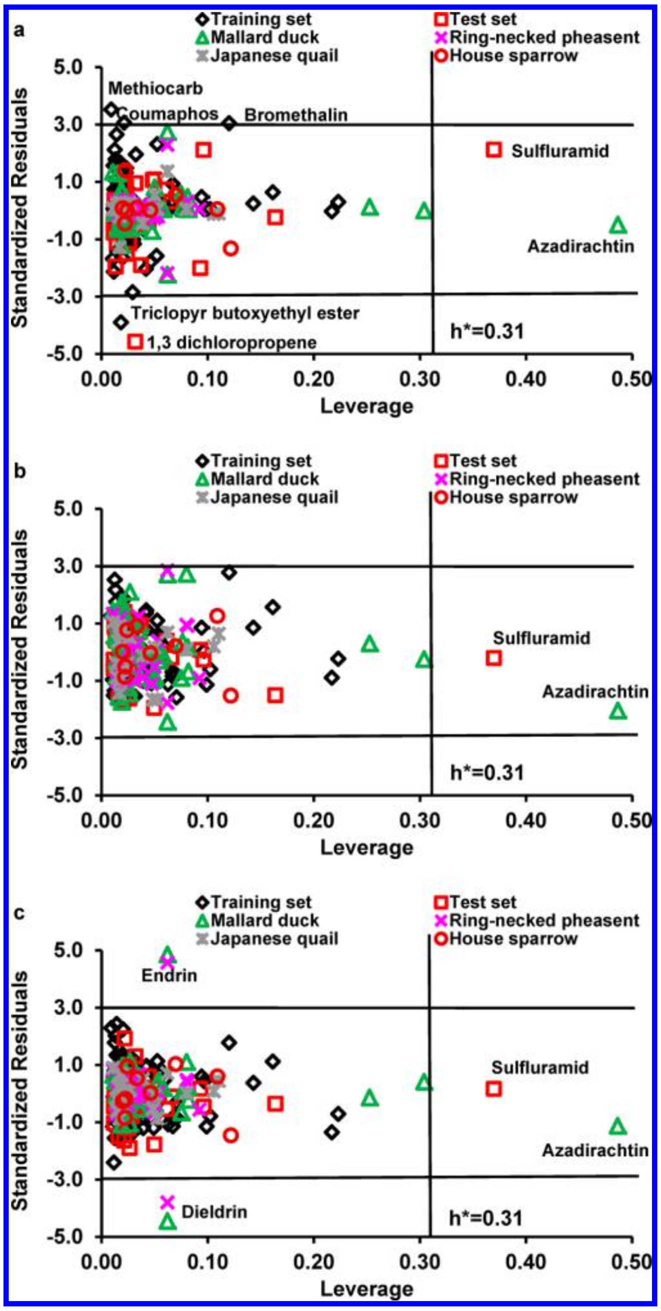


Figure 7. Williams plot for the (a) SDT-QSAR, (b) DTF-QSAR, and (c) DTB-QSAR models.

are quiet active on living organisms, since pesticides are intended to have adverse efforts on a certain target. These pesticides produce their effect through a series of biological mechanisms, which increases the complexity of the modeling tasks. Further, the metabolism and transformation of these compounds in the environment are very complex.[63] We have also further tried to check the AD of the developed QSAR models by the Euclidean distance (ED) approach. It is based on the distance scores calculated by the Euclidean distance norms. The calculated distance values were normalized between 0 and 1. Plots of normalized mean distance and toxicity of compounds both for the training and test sets are shown in Figure 8. It may be noted that all the compounds are inside the domain/area covered by the training set compounds (except azadirachtin, sulfluramid, kepone, and tributyltin maleate). A comparison of the William's and ED plots shows that the same compounds (azadirachtin and sulfluramid) lie out of the range in both cases. Hence, both the methods for the AD analysis employed here are fully capable of identifying the influential compounds.

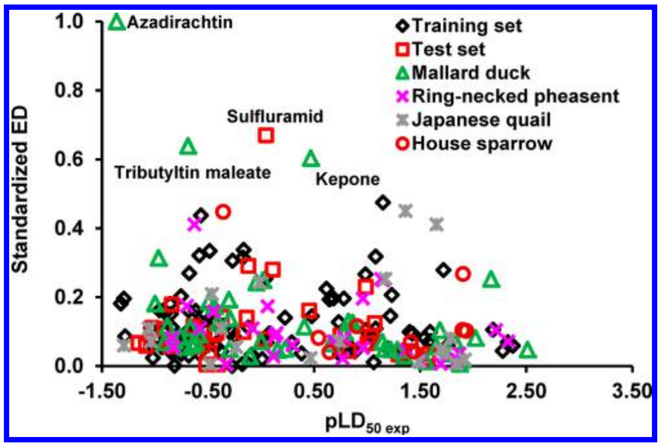


Figure 8. Plot of the Euclidean distance for the chemical pesticides with experimental toxicity values in multiple avian test species.

3.5. Analysis of Toxic Substructural Alerts. Analysis of the structural alerts facilitates the clear interpretability of the results. Although conventional QSAR models based on several molecular descriptors and nonlinear approaches can successfully capture complex dependences, the structural alerts analysis further identifies potentially toxic compounds. Accordingly, the substructures were treated as the alert substructures of the toxic compounds in our database. ChemoTyper[18] was used to identify the toxic substructures in compounds pertaining to the toxicity data sets in all five test species. A chemotype is defined as a structural fragment encoded for connectivity and also, when desirable, for physicochemical properties of atoms, bonds, fragments, electron systems, and even a whole molecule.[64] The toxic substructures are defined as molecular functional groups that make compounds toxic, which are hence used as substructural alerts. Substructural alerts were derived directly from mechanistic knowledge, so they are important to predict toxicity.[65] In all five toxicity data sets, carboxamide (C(=O)N), carbonyl (C=O), aminocarbonyl (NC=O), aromatic amine (CN), aromatic halides (Ph−X), halide (X), nitro (N=O), sulfide (CS), phosphate thioate (P=O), oxophosphorus (P=O), aromatic alkane (Ph−C1), aromatic benzene, aliphatic ether (COC), and alkane linear (C2) were the common major substructures responsible for the avian toxicity. Compounds with unsaturated carbonyl are bis-electrophiles that may interact with electron-rich biological macromolecules. In addition to the carbon in the carbonylic functionality, the β-carbon is positively polarized because of conjugation with the carbonyl group and becomes the preferred site of nucleophilic attack.[66] The unsaturated carbonyl compounds can undergo different interactions with DNA, which lead to different genotoxic and mutagenic responses. Aromatic amines have the ability to induce mutation and cancer. These compounds are believed to be biochemically transformed to a hydroxylamine intermediate, which is then activated to give an electrophilic nitrogen species.[67] The halide and nitro functional groups show bimolecular nucleophilic substitution and form covalent adducts with DNA.[68] Sulfur containing pesticides play roles in the biochemical interactions between small molecules and the biological system.[4] Pesticides containing the P=O group bind to the AcHE active site and inhibit the enzyme activity causing neurological disorders.[69] The chemical containing the ether, aromatic hydrocarbon, and sulfide are known to cause baseline toxicity.[70] Aliphatic hydrocarbons cause skin and eye irritation, neuropathy, and narcosis.[71] The identified substructure alerts are very important in ecological risk assessment and can help us to find toxic compounds. The structural alerts are not individually considered to constitute a QSAR model; they embody elements of QSAR and have proven useful for screening and prioritizing.[64]

4. CONCLUSIONS

In this study, tree-based QSAR models (SDT, DTF, and DTB) for predicting avian toxicity of diverse pesticides in multiple test species were developed using nine 2D molecular descriptors following the OECD guidelines. Accordingly, the toxicity data in five different avian species (Bobwhite quail, Mallard duck, Ring-necked pheasant, Japanese quail, and House sparrow) were considered. The results of TSI, BDS, and K−W tests established structural diversity of pesticides, nonlinear dependence, and significantly different populations of avian toxicity data sets. The QSAR models were developed using the Bobwhite quail toxicity data and identified the S36 and MW as the most influential descriptors. Intercorrelation analysis and PCA methods provided information on association of the molecular descriptors. Several statistical checks applied to ensure the external predictability of the models revealed high confidence. Among the three models, DTF and DTB performed clearly better and were subsequently applied to predict toxicities in four other test species. Both the DTF and DTB models performed well in these test data and established their robustness for applications in new untested data. AD analysis performed using leverage and ED approaches further verified as which of these compounds are well or not well predicted by the constructed models. We also used ChemoTyper to identify the structural alerts which can be used to recognize chemical toxicity on avian species. This study provided a powerful tool for prediction of chemical avian toxicity, and the methods used here could be promoted to other toxicity end-points or balanced data set.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Toxicity data of chemical pesticides in avian species. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00139.

AUTHOR INFORMATION

Corresponding Author

*Phone: 0091-522-2476091. Fax: 0091-522-2628227. E-mail: kpsingh_52@yahoo.com, kunwarpsingh@gmail.com.

Funding

No funding for this work was received from any source, except Senior Research Fellowship provided by the Council of Scientific and Industrial Research (CSIR), New Delhi, India to one of the author (S.G.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the Director, CSIR-Indian Institute of Toxicology Research, Lucknow (India) for his keen interest in this work and providing all necessary facilities.

ABBREVIATIONS

AD, applicability domain; BDS, Brock-Dechert-Scheinkman; CCC, concordance correlation coefficient; CV, cross-validation; DTB, decision tree boost; DTF, decision tree forest; ED, Euclidean distance; EPA, Environmental Protection Agency; ETA_Beeta_ns, a measure of electron richness of the molecule; EU, European Union; ISIZ, total information index on molecular size; K−W, Kruskal−Wallis; LD_{50}, 50% lethality from oral dose; MSE, mean squared error; MW, molecular weight; Nsulph, count of sulfur atoms; Nhet, count of hetero atoms; OECD, Organization for Economic Cooperation and Development; PCA, principal components analysis; PC4, molecular path counts of length 4; QSAR, quantitative structure−activity relationship; QSTR, quantitative structure-toxicity relationship; R^2, squared correlation coefficient; r_o^2, correlations between the observed and predicted values without intercept for the least-squares line; REACH, Registration, Evaluation, Authorization and Restriction of Chemicals; RMSE, root mean squared error; SD, standard deviation; SDT, single decision tree; SMILES, simplified molecular input line entry system; S36, sum of E-state of atom type (−O−); TIAC, total information index on atomic composition; TPSA, topological polarity surface area; TSI, Tanimoto similarity index; χ^2, chi-squared; 2D, two-dimensional

REFERENCES

(1) Singh, K. P.; Gupta, S.; Basant, N.; Mohan, D. QSTR modeling for qualitative and quantitative toxicity predictions of diverse chemical pesticides in honey bee for regulatory purposes. *Chem. Res. Toxicol.* **2014**, *27*, 1504−1515.

(2) Aktar, M. W.; Sengupta, D.; Chowdhury, A. Impact of pesticides use in agriculture: their benefits and hazards. *Interdiscip. Toxicol.* **2009**, *2*, 1−12.

(3) Mostafalou, S.; Abdollahi, M. Pesticides and human chronic diseases: Evidences, mechanisms, and perspectives. *Toxicol. Appl. Pharmacol.* **2013**, *268*, 157−177.

(4) Mazzatorta, P.; Cronin, M. T. D.; Benfenati, E. A QSAR study of avian oral toxicity using support vector machines and genetic algorithms. *QSAR Comb. Sci.* **2006**, *25*, 616−628.

(5) Zhang, C.; Cheng, F.; Sun, L.; Zhuang, S.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In silico prediction of chemical toxicity on avian species using chemical category approaches. *Chemosphere* **2015**, *122*, 280−287.

(6) Toropov, A. A.; Benfenati, E. QSAR models of quail dietary toxicity based on the graph of atomic orbitals. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1941−1943.

(7) Amaury, N.; Benfenati, E.; Boriani, E.; Casalengo, M.; Chana, A.; Chaudhry, Q.; Chretien, J. R.; Cotterill, J.; Lemke, F.; Piclin, N.; Pintore, M.; Porcelli, C.; Price, N.; Roncaglioni, A.; Toropov, A. Results of DEMETRA models. Chapter 7. In Benfenati, E. *Quantitative structure-activity relationship (QSAR) for pesticide regulatory purposes*; Elsevier B.V.: 2007; pp 201−282.

(8) Maynard, S. K.; Edwards, P.; Wheeler, J. R. Saving two birds with one stone: using active substance avian acute toxicity data to predict formulated plant protection product toxicity. *Environ. Toxicol. Chem.* **2014**, *33*, 1578−1583.

(9) OECD; *Test No. 223: Avian Acute Oral Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, Effects on Biotic Systems*; OECD Publishing: Paris, France, 2010. DOI: 10.1787/9789264090897-en.

(10) EPA; *US Environmental Protection Agency, Ecological Effects Test Guidelines, OCSPP 850.2100: Avian Acute Oral Toxicity Test, Office of Chemical Safety and Pollution Prevention (7101)*; Washington, DC, EPA 712-C-025, 2012.

(11) Organization for Economic Co-operation and Development (OECD); *Report of the SETAC/OECD workshop on avian toxicity testing*; OECD/GD(96)166; Paris, France, 1996.

(12) US Environmental Protection Agency (EPA); *Guideline 71-1: Avian single-dose oral LD_{50} test*. In *Pesticide Assessment Guidelines, Subdivision E—Hazard Evaluation*; Wildlife and Aquatic Organisms; EPA 540/9-82/024; Washington, DC, 1982; pp 33−36.

(13) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **2008**, *48*, 766−784.

(14) Cronin, M. T. D.; Jaworska, J. S.; Walker, J. D.; Comber, M. H. I.; Watts, C. D.; Worth, A. P. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. *Environ. Health Persp.* **2002**, *111*, 1391−1401.

(15) Jaworska, J. S.; Comber, M.; Auer, C.; Van Leeuwen, C. J. Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environ. Health Persp.* **2003**, *111*, 1358−1360.

(16) REACH. http://ee.europa.eu/environment/chemicals/reach/reach_intro.htm (accessed November 10, 2014).

(17) OECD; Environment Health and Safety Publications Series on Testing and Assessment No. 69. Guidance Document On The Validation Of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models; 2007. Accessed from http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en (accessed September 15, 2014).

(18) ChemoTyper community. Website: https://chemotyper.org/ (accessed November 8, 2014).

(19) OPP Pesticide Ecotoxicity Database. 2014. Available at http://www.ipmcenters.org/ecotox/ (accessed October 13, 2014).

(20) Zhao, L.; Dong, Y. H.; Wang, H. Residues of veterinary antibiotics in manures from feedlot livestock in eight provinces of China. *Sci. Total Environ.* **2010**, *408*, 1069−1075.

(21) Tcheslavski, G.; (Louis) Beex, A. A. Effects of smoking, schizotypy, and eyes open/closed conditions on the γ1 rhythm phase synchrony of the electroencephalogram. *Biomed. Signal Proces* **2010**, *5*, 164−173.

(22) ChemDes. www.scbdd.com/chemdes/ (accessed October 30, 2014).

(23) ChemSpider. www.chemspider.com (accessed October 28, 2014).

(24) Roy, K.; Ghosh, G. Exploring QSARs with extended topochemical atom (ETA) indices for modeling chemical and drug toxicity. *Curr. Pharm. Des.* **2010**, *16*, 2625−2639.

(25) Roy, K.; Das, R. N. On some novel extended topochemical atom (ETA) parameters for effective encoding of chemical information and modelling of fundamental physicochemical properties. *SAR QSAR Environ. Res.* **2011**, *22*, 451−472.

(26) Yap, C. W. PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−1474.

(27) Singh, K. P.; Gupta, S.; Kumar, A.; Mohan, D. Multispecies QSAR Modeling for Predicting the Aquatic Toxicity of Diverse Organic Chemicals for Regulatory Toxicology. *Chem. Res. Toxicol.* **2014**, *27*, 741−753.

(28) Benigni, R.; Netzeva, T. I.; Benfenati, E.; Bossa, C.; Franke, R.; Helma, C.; Hulzebos, E.; Marchant, C.; Richard, A.; Woo, Y. P.; Yang, C. The expanding role of predictive toxicology: An update on the (Q)SAR models for mutagens and carcinogens. *J. Environ. Sci. Health Part C* **2007**, *25*, 53−97.

(29) Singh, K. P.; Gupta, S.; Basant, N. In silico prediction of cellular permeability of diverse chemicals using qualitative and quantitative SAR modeling approaches. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 61−72.

(30) Broock, W. A.; Dechert, W.; Scheinkman, J. A.; LeBaron, B. A test for independence based on the correlation dimension. *Economet. Rev.* **1996**, *15*, 197−235.

(31) Anoruo, E. Testing for linear and nonlinear causality between crude oil price changes and stock market returns. *Int. J. Econ. Sci. Appl. Res.* **2011**, *4*, 75−92.

(32) Singh, K. P.; Gupta, S.; Rai, P. Predicting acute aquatic toxicity of structurally diverse chemicals in fish using artificial intelligence approaches. *Ecotoxicol. Environ. Saf.* **2013**, *95*, 221−233.

(33) Singh, K. P.; Gupta, S. In silico prediction of toxicity of non-congeneric industrial chemicals using ensemble learning based modeling approaches. *Toxicol. Appl. Pharmacol.* **2014**, *275*, 198−212.

(34) Singh, K. P.; Gupta, S. Nano-QSAR modeling for predicting biological activity of diverse nanomaterials. *RSC Adv.* **2014**, *4*, 13215−13230.

(35) Singh, K. P.; Gupta, S.; Basant, N. Predicting toxicities of ionic liquids in multiple test species − An aid in designing of green chemicals. *RSC Adv.* **2014**, *4*, 64443−64456.

(36) Singh, K. P.; Gupta, S.; Basant, N. QSTR modeling for predicting aquatic toxicity of pharmacological active compounds in multiple test species for regulatory purpose. *Chemosphere* **2015**, *120*, 680−689.

(37) Coops, N. C.; Waring, R. H.; Beier, C.; Roy-Jauvin, R.; Wang, T. Modeling the occurrence of 15 coniferous tree species throughout the Pacific Northwest of North America using a hybrid approach of a generic process-based growth model and decision tree analysis. *Appl. Veg. Sci.* **2011**, *14*, 402−414.

(38) Pino-Mejias, R.; Jimenez-Gamero, M. D.; Cubiles-de-la-Vega, M. D.; Pascual-Acosta, A. Reduced bootstrap aggregating of learning algorithms. *Pattern Recogn. Lett.* **2008**, *29*, 265−271.

(39) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189−1232.

(40) Chirico, N.; Gramatica, P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J. Chem. Inf. Model.* **2011**, *51*, 2320−2335.

(41) Lin, L. I. Assay validation using the concordance correlation coefficient. *Biometrics* **1992**, *48*, 599−604.

(42) Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Model.* **2001**, *41*, 186−195.

(43) Schuurmann, G.; Ebert, R.; Chen, J.; Wang, B.; Kuhne, R. External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean. *J. Chem. Inf. Model.* **2008**, *48*, 2140−2145.

(44) Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q^2 parameter for QSAR validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669−1678.

(45) Pratim Roy, P.; Paul, S.; Mitra, I.; Roy, K. On two novel parameters for validation of predictive QSAR models. *Molecules* **2009**, *14*, 1660−1701.

(46) Netzeva, T. I.; Worth, A. P.; Aldenberg, A.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klpoman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patliwicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. P.; van de Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationship. *ATLA, Altern. Lab. Anim.* **2005**, *33*, 155−173.

(47) Tropsha, A.; Golbraikh, A.; Cho, W. J. Development of kNN QSAR models for 3-arylisoquinoline antitumor agents. *Bull. Korean Chem. Soc.* **2011**, *32*, 2397−2404.

(48) Raevsky, O. A.; Liplavskaya, E. A.; Yarkov, A. V.; Raevskaya, O. E.; Worth, A. P. Linear and nonlinear QSAR models of acute intravenous toxicity of organic chemicals for mice. *Biochemistry-Moscow* **2011**, *5*, 213−225.

(49) Chirico, N.; Gramatica, P. Real external predictivity of QSAR models: Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J. Chem. Inf. Model.* **2012**, *52*, 2044−2058.

(50) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361−1375.

(51) Grunwald, S.; Daroub, S. H.; Lang, T. A.; Diaz, O. A. Tree-based modeling of complex interactions of phosphorus loadings and environmental factors. *Sci. Total Environ.* **2009**, *407*, 3772−3783.

(52) Chou, J. S.; Chiu, C. K.; Farfoura, M.; Al-Taharwa, I. Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data mining techniques. *J. Comput. Civil Eng.* **2011**, *25*, 242−253.

(53) Erdal, H. I.; Karakurt, O. Advancing monthly stream flow prediction accuracy of CART models using ensemble learning paradigms. *J. Hydrol.* **2013**, *477*, 119−128.

(54) Sánchez-Bayo, F. Insecticides Mode of Action in Relation to Their Toxicity to Non-Target Organisms. *J. Environ. Analytic. Toxicol.* **2012**, *S4:002*, 1−9.

(55) Fatemi, M. H.; Izadiyan, P. Cytotoxicity estimation of ionic liquids based on their effective structural features. *Chemosphere* **2011**, *84*, 553−563.

(56) Das, R. N.; Roy, K. Predictive in silico Modeling of Ionic Liquids toward Inhibition of the Acetyl Cholinesterase Enzyme of Electrophorus electricus: A Predictive Toxicology Approach. *Ind. Eng. Chem. Res.* **2014**, *53*, 1020−1032.

(57) Hou, T. J.; Zhang, W.; Xia, K.; Qiao, X. B.; Xu, X. J. ADME Evaluation in Drug Discovery. 5. Correlation of Caco-2 Permeation with Simple Molecular Properties. *J. Chem. Inf. Model.* **2004**, *44*, 1585−1600.

(58) Choudhary, M.; Sharma, B. K. QSAR rationales for the 5-HT6 antagonistic activity of Epiminocyclohepta[*b*]indoles. *Der Pharma Chemica* **2014**, *6*, 321−330.

(59) Zakeri-Milani, P.; Tajerzadeh, H.; Islambolchilar, Z.; Barzegar, S.; Valizadeh, H. The relation between molecular properties of drugs and their transport across the intestinal membrane. *DARU J. Pharm. Sci.* **2006**, *14*, 164−171.

(60) Afantitis, A.; Melagraki, G.; Sarimveis, H.; Igglessi-Markopoulou, O.; Kollias, G. A novel QSAR model for predicting the inhibition of CXCR3 receptor by 4-N-aryl-[1,4] diazepaneureas. *Eur. J. Med. Chem.* **2009**, *44*, 877−884.

(61) Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. Investigation of substituent effect of 1-(3,3-diphenylpropyl)—piperidinylphenylacetamides amides on CCR5 binding affinity using QSAR and virtual screening techniques. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 83−95.

(62) Karbakhsh, R.; Sabet, R. Application of different chemometric tools in QSAR study of azoloadamantanes against influenza a virus. *Res. Pharm. Sci.* **2011**, *6*, 23−33.

(63) Zitko, V. Chlorinated pesticides: Aldrin, DDT, Endrin, Dieldrin, Mirex. In *The Handbook of Environmental Chemistry*; Fiedler, H., Ed.;

Part O Persistent Organic Pollutants. Springer-Verlag; Berlin, Heidelberg, 2003; Vol. *3*, Chapter 4, pp 47−90.

(64) Yang, C.; Tarkhov, A.; Marusczyk, J.; Bienfait, B.; Gasteiger, J.; Kleinoeder, T.; Magdziarz, T.; Sacher, O.; Schwab, C. H.; Schwoebel, J.; Terfloth, L.; Arvidson, K.; Richard, A.; Worth, A.; Rathman, J. New Publicly Available Chemical Query Language, CSRML, To Support Chemotype Representations for Application to Data Mining and Modeling. *J. Chem. Inf. Model.* **2015**, *55*, 510−528.

(65) Sun, L.; Zhang, C.; Chen, Y.; Li, X.; Zhuang, S.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In silico prediction of chemical aquatic toxicity with chemical category approaches and substructural alerts. *Toxicol. Res.* **2015**, *4*, 452−463.

(66) Koleva, Y. K.; Madden, J. C.; Cronin, M. T. D. Formation of categories from structure activity relationships to allow read across for risk assessment: toxicity of *α*, *β*-unsaturated carbonyl compounds. *Chem. Res. Toxicol.* **2008**, *21*, 2300−2312.

(67) Benigni, R.; Bossa, C. Mechanisms of chemical carcinogenicity and mutagenecity: A review with implications for predictive toxicology. *Chem. Rev.* **2011**, *111*, 2507−2536.

(68) Enoch, S. J.; Cronin, M. T. D. Development of new structural alerts suitable for chemical category formation for assigning covalent and non covalent mechanisms relevant to DNA binding. *Mutat. Res., Genet. Toxicol. Environ. Mutagen.* **2012**, *743*, 10−19.

(69) Elersek, T.; Filipic, M. Organophosphorus pesticides − Mechanisms of their toxicity. In *Pesticides − the impacts of pesticides exposure*; Stoytcheva, M., Ed.; InTech Publishers: Croatia, 2011.

(70) Franks, N.; Lieb, W. Mechanism of general anesthesia. *Environ. Health Perspect.* **1990**, *87*, 199−205.

(71) Astill, B. D. Structure-activity relationships within and between chemical classes. In *Methods for assessing the effects of mixtures of chemicals*; Vouk, V. B., Butler, G. C., Upton, A. C., Parke, D. V., Asher, S. C., Eds.; Wiley: Chichester, 1987; pp 209−223.