Stacked-based *in-silico* ensemble model for toxicity prediction of the chemical data

Priyanka Shit*, Sidhanta Kumar Balabantaray, Smita Ranee Barik
Department of Computer Science and Engineering
Gandhi Institute for Education and Technology, Baniatangi, Bhubaneswar

KAMALINI SINGH, Department of Computer Science and Engineering, Rajdhani Engineering College, Bhubaneswar

Abstract-The stacked-based ensemble method is based on a meta-learning approach that has been effectively used in various fields of application. In the area of computational chemistry, the prediction of toxicity is a very challenging task. The required problem is solved by different methodologies which have been used in the literature. The ensemble classifier is very efficient to use for any classification problem. In this study, we have used a stacked-based ensemble learning method to classify toxic and non-toxic data. For feature reduction of the chemical data, we have used a hybrid filter-wrapper approach. This approach has given an optimized feature set for our classification problems. For comparing the result, we have taken various pre-existing machine learning methods like artificial neural networks, support vector machines, random forest, etc., and we got improved results with a stacked-based ensemble approach.

Keywords: Toxicity, Feature Reduction, Ensemble method, Machine Learning

Introduction

Chemicalsaretodayanessentialpartofhumanlifeandcomfortandconsequently,alargenumberofindustrialchemicalsaremanufacturedand used every day. However, many of these have been identified aspotentially toxic to the humans. The regulatory agencies have beenemphasizing their safety assessment prior to their manufacture anduse (Casalegno et al., 2005). A 40-h toxicity test, expressed in terms ofthe median growth inhibition concentration (IGC_{50}) value to the freshwaterciliate*Tetrahymenapyriformis*isconsideredappropriatefortoxi-cological testing and safety evaluation of the chemicals (Cronin et al.,2002). *T. pyriformis* is an ubiquitous ciliated protozoan belonging tofreeliving,freshwatergenusandaresensitivetogrowthconditions.Ithas several advantageous characteristics for toxicity studies, such astheyplaykeyroleinthetransferofenergyandmatterwithinthemicro-bial loop. They occupy one of the first trophic levels in aquatic ecosys-tems and thus, are early warning of a toxic danger (Lukacinova et al.,2007). In *T. pyriformis*, the toxicological end-points have usually beenrestrictedtoameasureofgrowthinhibitionratherthanofcellviability (Schultz,1997).Themainrelevanceofthisend-pointistocharacterizeaquatic toxicity. According to Duchowicz and Ocsachoque (2009), theIGC_{50} encapsulates the most aqueous toxicity information. Toxicityevaluationofthehighproductionvolumecompounds,suchasindustri-al chemicals and pesticides is a top global regulatory priority. Accord-ingly,severalexperimentaldatabasesonIGC50ofdifferentgroupsofchemicalshavebeendevelopedbyvariousresearchgroups(Aptulaet al., 2005; Cronin and Schultz, 2001; Cronin et al., 2002, 2004;Devillers, 2004; Deweese and Schultz, 2001; Melagraki et al., 2005;NetzevaandSchultz,2005;Netzevaetal.,2003;Ren,2003;RenandSchultz, 2002; Roy et al., 2005; Schultz, 1999; Schultz et al., 2005;Schuurmann et al., 2003; Serra et al., 2001). Because the experimentaldeterminationoftoxicologicalpropertiesisacostlyandtimeconsumingprocess,itisessentialtodeveloppredictivemathematicalrelationshipsto theoretically quantify toxicity (Melagraki et al., 2006). Quantitativestructure-activity relationships (QSARs) are increasingly being usedasatooltoassessregulatoryagenciesintoxicologicalassessmentofchemicalsubstances(Croninetal.,2002).DesirablequalitiesofQSARin-cludethemodelbeingtransparent,easilyportable,anddevelopedwithinterpretable descriptors. Several quantitative structure-toxicity rela-tionship (QSTR) models, based on multiple linear regression (MLR),partial least squares regression (PLSR), k-nearest neighbor (k-NN), arti-ficialneuralnetworks(ANN),supportvectormachines(SVM),decisiontree(DT)approacheshavebeen proposedbyvariousresearch groups

(Chengetal.,2011;Ivanciuc,2004;Jalali-HeraviandKyani,2008;Melagraki et al., 2006; Schuurmann et al., 2003; Toropov et al., 2010;Zhuetal.,2008)forpredicting*T.pyriformis*toxicityofchemicals.However,themainprobleminanyQSARanalysisistheevaluationand control of the predictive ability of the developed model (Toropovet al., 2010). Moreover, these local models could provide acceptablepredictive accuracy for a very limited chemical domain; they were notapplicable to assess a large diverse set of chemical structures (Chenget al., 2011). Also, all these approaches considered different types ofmoleculardescriptorsasestimators,andselectionandcomputationofrelevantdescriptorstoextractinformationfromcompoundstructuresisthemajorlimitationofthisresearch field.

Inrecentyearsensemblelearning(EL)methods(Snelderetal., 2009) have emerged as unbiased tools for modeling the complex relationships between set of independent and dependent variables andhave been applied successfully in various research areas (Yang et al.,2010). In general, these methods are designed to overcome problemswithweakpredictors(Hancocketal.,2005)andhavetheadvantageofalleviatingthesmallsamplesizeproblembyaveragingandincorporating over multiple classification models to reduce the potential forover-fitting the training data (Dietterich, 2000). Decision tree forest(DTF)anddecisiontreeboost(DTB)implementingbaggingandboosting techniques, respectively are relatively new methods for improvingtheaccuracyofapredictivefunction(Yangetal.,2010).Thesetechniquesareinherentlynon-parametricstatisticalmethodsandmakenoassumptionregardingtheunderlyingdistributionofthevaluesof predictor variables and can handle numerical data that are highlyskewedormulti-modelinnature(MahjoobiandEtemad-Shahidi,2008). To our knowledge, ensemble learning methods have not yetbeenappliedtothetoxicitypredictionmodeling.

Selectionofappropriatemoleculardescriptorsintoxicityprediction is yet another important issue. A large number and variety of such descriptors have been used in several earlier studies, generally derivedthroughhighlycomplicatedsemi-empiricalandempiricalmethodsbased on quantum mechanical calculations (Eroglu et al., 2007; Wangetal.,2010;Inetal.,2012).Hence,itwouldbedesirabletodevelopticologicallyrelevantQSTRsusingsimplepropertiesthatcanbederiveddirectlyfromachemical'sstructure.Moreover,inviewoftheregulatorytoxicology requirements, models discriminating compounds merelybetween toxic and non-toxic classes are not enough and it is verymuchdesirabletohavemoreefficientscreeningtoolscapableofclassi-fying compounds in several toxicity classes, such as highly toxic, toxic,harmful, and non-harmful; as well as capable of predicting the toxicityend-points inaquantitativemanner.

Inthisstudy,thebasicobjectivesweretoconstructtheensemble learningbasedmodels(DTBandDTF)forpredictingthetoxicityofthediversechemicalcompoundsusingsimplemoleculardescriptors,Accordingly, classification and regression models were constructed topredict the toxicity classes and the toxicity end-point (−log IGC_{50}) ofthe diverse chemicals using a set of selected molecular properties/descriptorsasestimators.Thepredictiveandgeneralizationabilitiesofthe DTBandDTFclassificationandregressionmodelsconstructedherewereevaluatedusingseveralstatisticalcriteriaparametersandperfor-mance of these models were tested using external datasets. Moreover,thepredictiveabilityoftheDTBandDTFregressionmodelswascomparedwithkernelpartialleastsquaresregression(KPLSR),abasicmodelingapproach.

Materialsandmethods

Dataset

FordevelopingtheensemblelearningbasedQSTRmodelsfortoxic-ity prediction of chemicals in *T. pyriformis*, data from multiple sourceswere considered (Cheng et al., 2011; Tetko et al., 2008; Xue et al.,2006).Thechemicallyheterogeneousdatasetiscomprisedof1450

chemicals representing different groups. The reported 50% growth-inhibitoryconcentrationvalues(IGC_{50})arebasedonthestandard *T.pyriformis*testprotocol(SchultzandNetzeva,2004).Forthesecompounds, values of 40-h exposure based median growth inhibition con-centration (IGC_{50}) for *T. pyriformis* are reported as −log IGC_{50} ($pIGC_{50}$,mmolL^{-1}),wherethelogarithmistakentocontractthedatasettoacomputationallyefficientrange(Serraetal.,2001).Toxicitylevelgenerallyincreaseswithincreasingvalueof—logIGC_{50}(mmolL^{-1}),andcompounds with positive values are generally considered to be toxicor weakly toxic. Complete dataset of 1450 compounds with endpointvaluesarepresentedinTableS1(SupportingInformation).Since,validation aims to stimulate the predictivity of a model towards new,unknown chemicals, external datasets were collected from literaturefor model validation (Zhang et al., 2010). Accordingly, the externaldatasets (Table S2 of Supporting Information) contained comparativetoxicity (IGC_{50}) data of chemicals to *T. pyriformis*, and median effectiveconcentration(EC_{50})valuesofchemicalsin*Vibriofischeri*(marinebacterium) and *Scenedesmus obliguue* (algae). The EC_{50} values in viewoftheirhighcorrelationreportedwithIGC_{50}valuesindifferentspecies(Zhang etal.,2010), were also considered for external validation oftheconstructed models.

Moleculardescriptorsandfeatureselection

In toxicological studies, the molecular descriptors represent struc-tural and physicochemical properties of compounds. The descriptorswere calculated for each molecule using Toxmatch (Ideaconsult Ltd.).Molecular descriptors (physical, constitutional, geometrical, and topological)werecomputedby2Dstructuresofthemolecules, whichwere taken in the form of SMILES (simplified molecular input lineentry system). A set of 60 different molecular properties of each of thecompound were selected initially. Since, all the molecular propertiesmaynotberelevanttothemodeling;eliminationoflesssignificantdescriptorscanimprovetheaccuracyofprediction,andfacilitatetheinterpretationofthemodelthroughfocusingonthemostrelevantvariables.Initial features were selected by model-fitting approach. EL modelingwas performed. For optimal values of the model parameters, the ELmodels were trained by using the complete set of features computingthe respective scoring functions to rank the contribution of features inthe current set. The lowest ranked features were then removed (Xueet al., 2006). The EL systems were retained by using the remaining setoffeatures,andthecorrespondingpredictionaccuracies(misclassifica-tion rate, and mean squared error of prediction) were computed bymeansof10-foldcrossvalidation.Theselecteddescriptorsarelogarithmicformofoctanol-waterpartitioncoefficient(logP),molecularweight (MW), molecular surface area (MSA), charge partial surfacearea22(CPSA-22),connectivityindexorderone(CIOO),eccentricconnectivityindex(ECI),numberofatoms inlargest chain (NALC),and number ofatomsinlargestpi-system (NALPS). Finally a set ofsix descriptors for classification and five descriptors for regressionmodeling were considered in this study. The basic statistics of theselected descriptors for different datasets are given in Table 1andTableS3(Supportinginformation).

Dataprocessing

Since the aim of present study is to build a robust model capable ofmakingaccurateandreliablepredictionsoftoxicityofnewcompound,themodelderivedfromatrainingsetshouldbevalidated/testedusingnewchemicalmoietiesforcheckingitspredictiveability.Thevalidationstrategieschecktherelibilitiesofmodelsfortheirpossibleapplicationon a new dataset, and confidence in the prediction can thus be judged.For predictive modeling the end point (IGC_{50}) values were expressedas—logIGC_{50}($pIGC_{50}$,mmolL^{-1}).Categorizationofcompoundsas *T. pyriformis* toxic (TPT) and non-toxic to *T. pyriformis* (Non-TPT) wasbasedonthecriteriaofXueetal.(2006).Accordingtothecriteria

Table1
Basicstatisticsoftheselected moleculardescriptorsintoxicitypredictionofchemicalcompounds.

Descriptors	Model	Min	Max	Mean	SD	CV
CPSA.22	C	0.09	1.00	0.43	0.14	31.86
CIOO	R,C,C*	2.56	33.45	12.45	4.48	35.98
ECI	C,C*	2.00	573.00	96.28	57.91	60.15
LogP	R, C,C*	−2.41	6.86	2.08	1.26	60.43
MSA	R	0.00	157.69	38.33	23.79	62.06
MW	R, C,C*	32.03	623.74	153.53	51.47	33.52
NALC	C*	0.00	50.00	11.20	9.54	85.23
NALPS	R,C	0.00	20.00	6.33	4.09	64.54
$pIGC_{50}$	R	−2.67	3.34	0.33	1.03	315.02

SD—standarddeviation,CV—coefficientofvariation;R—regression;C—two-categoryclassification;C*—four-categoryclassification modeling.

(Xueetal.,2006),compoundswith$pIGC_{50}$valueof≤—0.5mmol$L^{-1}$are categorizedasnon-TPTandthosewith$pIGC_{50}$valueof≥—0.5mmolL^{-1}as TPT. In the selected dataset, total 310 compounds were taken as non-TPTandtheremaining1140compoundswereconsideredasTPT.However, for regulatory purpose, classification of chemicals merely in twocategories(TPT,non-TPT)willnotsufficient,andneedsamoreprecisecategorization of chemicals. Therefore, a four category classification sys-tem of chemicals was considered. Four toxicity classes of the compoundsweregeneratedaccordingtotheintervals:$pIGC_{50}$N1.5mmolL^{-1}forclass1(highlytoxic);1.5mmolL^{-1}N$pIGC_{50}$N0.5mmolL^{-1}forclass2(toxic);0.5mmolL^{-1}mmolL^{-1}N$pIGC_{50}$N—0.5mmolL^{-1}forclass3(harmful);$pIGC_{50}$b—0.5mmolL^{-1}forclass4(notharmful).In this toxicity categorization scheme, the cut-off limit for non-TPT compoundswassimilar(pIGC50≤—0.5mmolL^{-1})toproposedbyXueetal.(2006).Thesecriteriarendered191compoundsinclassI,458inclassII,491inclassIII,and310inclassIV.Inthisstudy,forclassificationandre-gression modeling, data were split into training (80%) and test (20%)subsets using the random distribution approach. Such test sets (whendefined prior to analysis) come close to external validation set, whicharecommonlyacceptedasthegoldstandardtoassessrealpredictivity(Benignietal.,2007)

Diversityanalysis

Thediversityofadatasetinvolvesdefiningadiversesubsetofrepresentative compounds and is important for global model development(Zhao et al., 2006). The diversity and similarity of the dataset were exploredthroughradarchartanalysis(Singhetal.,2013a)andTanimotosimilarityanalysis(Johnetal.,1998).Aradarchartdisplaysmultivariatedata in the form of a two-dimensional chart with several quantitativevariables represented on axis starting from the same point. Tanimotosimilarity index (TSI), is an appropriate distance metric for topology-basedchemicalsimilaritystudies.TanimotodistancemethodcalculatesTanimotosimilaritybetweenfingerprintofachemicalandaconsensusfingerprint,whichis1024bitfingerprint(Toxmatch,IdeaconsultLtd.).Thefingerprintgenerationisbasedonthefingerprintimplementationof the open source chemoinformatics library (Steinbeck et al., 2003).Forgivenmoleculeallpossiblepathsforapredefinedlengtharegener-ated,thepathissubmittedtoahashfunctionwhichusesitasaseedtoapseudorandomgenerator,thehashfunctionoutputsasetofbits,whichis addedtothe fingerprint.For amolecule,TSI iscalculated as;

$$TSI_{AB} = 2Z_{AB} / [Z_{AA} + Z_{BB} - Z_{AB}]^{-1} \quad (1)$$

where Z is the similarity matrix, A and B are the two molecules beingcompared.TheTSIrangesfrom0(nosimilarity)to1(pair-wisesimilarity). Smaller TSI means compounds have good diversity (Cheng et al.,2011).Agoodcut-offforbiologicallysimilarmoleculesis0.7or0.8.

Predictivemodeling

Here, we constructed the EL approach based DTB and DTF models forclassification and regression problems to predict the toxicity class ofcompoundsasTPTversusnon-TPT,andamongfourcategories(highlytoxic,toxic,harmful,andnon-harmful);aswellas the toxicity(pIGC$_{50}$)ofthestructurallydiversechemicalsusingasetofnon-quantummechanicalmoleculardescriptorsastheestimators.Allcomputations were performed using the EXCEL 97 and the modelingalgorithms were implemented in MATLAB (Mathworks, Inc, Natwick,MA).BrieftheoreticaldescriptionoftheELbasedmodelingapproachesusedhereisprovidedbelow.

Ensemblelearningapproaches

An ensemble contains a number of base learners (Ishwaran andKogalur, 2010) and their generalization ability is usually much stron-ger. Bagging and stochastic gradient boosting algorithms are consid-ered here for constructing the classification and regression decisiontree (DT) models (DTF and DTB). Bagging uses different perturbeddataandfeaturesetsfortrainingbaseclassifiers, whereas inboosting,diversityisobtainedbyincreasingtheweightsofmisclassifiedsamplesinaniterativemanner(Yanget al., 2010).Thesemethodsusedecisiontreesas base classifiers because DTsaresensitivetosmallchangesonthetrainingset(Dietterich,2000).

Decision TreeBoost. DTB combines the strengths of regression tree andstochastic gradient boosting algorithm. Boosting improves the accu-racy of a predictive function by applying it repeatedly in a series andcombiningtheoutputofeach function with weighting,sothat thetotalerrorofpredictionisminimized(Friedman,2002).Gradientboostingisasequentialstage-wiseforwarditerative algorithm tofindanadditivepredictor(Fig.1a). The DTB algorithm creates atree ensemble and it uses randomization during the tree creations.Thegoalistominimizethelossfunctioninthetrainingset,{x,y}.Aftereachiteration,Frepresentssumofalltreesbuiltsofar:

$$F_m(x) = F_{m-1}(x) + \mathrm{Tree}_m(x) \quad (2)$$

where m is the number of trees in the model. Regardless of the loss-function, the trees fitting the gradient on pseudo residuals are re-gression trees trained to minimize mean squared error (MSE). TheDTB model for classification is essentially the same as for regressionexceptlogit(probability)valuesarefittedratherthanrawtargetvalues. The DTB usesthe Huber M-regression loss function whichmakesithighlyresistanttooutliers(Huber,1964).

Decision tree forest. In DTF, a large number of independent trees aregrown in parallel, and they do not interact until after all of them havebeen built (Fig. 1b). Bootstrap re-sampling method (Efron, 1979) andaggregating, the basis of bagging, are incorporated in DTF. Differenttrainingsub-setsaredrawnatrandomwithreplacementfromthetrain-ingdataset.Separatemodelsareproducedandusedtopredicttheentiredatafromaforesaidsub-sets.Thenvariousestimatedmodelsareaggregatedbyusingthemeanforregressionproblemsormajorityvotingforclassification problems. Bagging can reduce variance when combinedwiththe base learner generation with a good performance (Wanget al., 2011). The DTFs gaining strength from bagging technique usetheoutofbagdatarowsformodelvalidation.Thisprovidesanindependenttestsetwithoutrequiringaseparatedatasetorholdingbackrowsfrom the tree construction. The stochastic element in DTF algorithmmakesithighlyresistant toover-fitting.

Kernelpartialleastsquaresregression(KPLS)

KPLSisanonlinearextensionoflinearPLSinwhichtrainingsamplesare transformed into a feature space via a nonlinear mapping throughthekerneltrick,andthePLSalgorithmisthenimplementedinthe

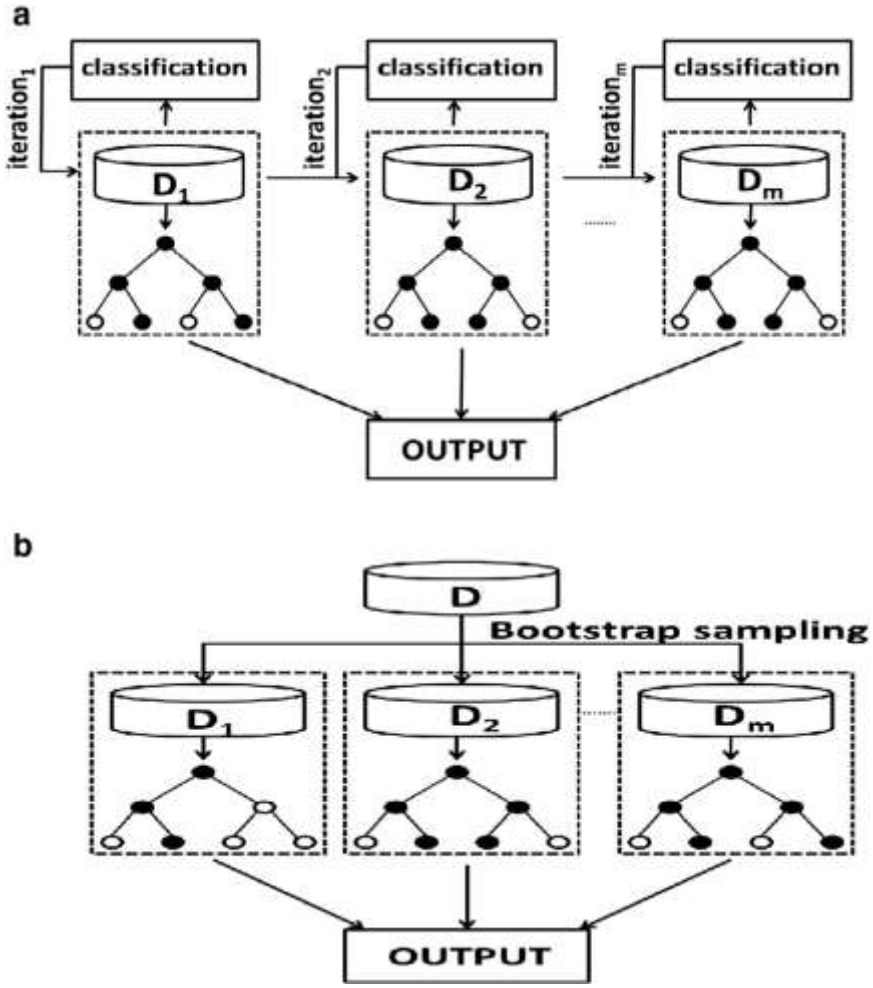


Fig.1.Conceptualdiagram ofthe(a)DTBand(b)DTFmodels.

featurespace.Nonlineardatastructureintheoriginalspaceismostlike-ly to be linear after high-dimensional nonlinear mapping (Singh et al.,2013b).Therefore,KPLScanefficientlycomputelatentvariables(LVs)inthefeaturespacebymeansofintegraloperatorsandnonlinearkernelfunction. Here, we used the radial basis function (RBF) kernel. For theRBF kernel, the most important parameter is the width (σ) of the RBFwhichcontrolstheamplitudeofthekernelfunction.Here,theoptimum valueofσandLVsweredeterminedthroughCVprocedure.

Model validation

Theoptimalarchitecturesandmodelparametersoftheclassificationand regression models constructed here were determined followingboththeinternalandexternalvalidationprocedures.Forinternalvali-dation, a V-fold cross validation (CV) method was adopted. The V-foldCVisthemostcommonprocedurerecommendedtocheckthegeneral-izationabilityofthemodel(Benignietal.,2007).Theadvantageofthismethodisthatitperformsreliableandunbiasedtestingondataset.Forexternalvalidation,aseparatevalidation(test)sub-setofthedatawasused which was kept out during the training process (Singh et al.,2012,2013a).Incaseofthepredictivemodels,validationstepusingex-ternal data set provides information about the predictive ability of thetrained model for the unknown data (Singh et al., 2012). Benigni et al.(2008)pointed out that the prediction reliabilityshould be checkedbymeansofanexternaltestsetwithnewchemicalsnotused inmodel building. Optimal models were selected on the basis of the mis-classificationrate(classification),MSEandthevalueofsquaredcorrelationcoefficient(R^2)betweenthemeasuredandmodelpredicted

response (regression) in the training and validation data (Singh et al.,2011).

The developed classification and regression models were furthervalidated using the Y-randomization test. Y-randomization has beenfrequently used to determine the possibility of chance of correlationduring descriptor selection procedure (Xue et al., 2006; Masand et al.,2010).InY-randomization,thedependentvariable(categoryandpIGC$_{50}$) vectors were randomly shuffled and new models were builtusing the original independent variables (Mahajan et al., 2013). Theprocedure was repeated a number of times and CV statistics werecomputed. The performance (misclassification rate and R^2) of the newmodels were compared with the original models. If the new modelshave higher misclassification rate (classification) and lower R^2 values(regression) for several trials, then the given model is thought to berobust.Thus,Y-randomizationisusefultoavoidanychance-comercor-relationbetweendependentvariablesvectorandindependentvariables(Mahajanetal.,2013).

Prediction verification

Thestatisticalcharacterizationofclassificationmodelisbasedonthe'confusion' matrix. The performance of the classification models wereassessedintermsofthemisclassificationrate(MR),sensitivity,specific-ity,accuracyofprediction,andMatthew'scorrelationcoefficient(MCC)(Chengetal.,2011;Singhetal.,2011).Inaddition,theareaundercurve(AUC)forthereceiveroperatingcharacteristic(ROC)wascalculated.IftheAUCofROCcurveis1,aperfectclassifiercanbefound,andthe

AUC value of 0.5 suggests the classifier has no discriminative power atall(Chengetal.,2011).

Performance of the regression models used here was evaluatedusing different statistical criteria parameters: the root mean squareerror(RMSE),meanabsoluteerror(MAE),andthesquaredcorrelationcoefficient (R^2) between the measured and predicted values of theresponse(Singhetal.,2009a,2010).Eachperformancecriteriatermdescribed above conveys specific information regarding the predictiveperformanceefficiencyofaspecificmodel.Goodnessoffitofthselect-edmodel wasalso checkedthrough theanalysis oftheresiduals.

Results

Basic statistics of the selected molecular descriptors (Table 1) suggestthatthestandarddeviationandthecoefficientofvariation(CV)indicatedhighvariabilitywithintheselecteddescriptors.TheCVvaluesofthe descriptors oscillate between 31.86% (CPSA.22) and 85.23% (NALC).It may be noted that the constitutional descriptors exhibited highestvariability(64.54%–85.23%)followedbygeometrical(31.86%–62.06%), physico-chemical(33.52%–60.43%)andtopological(35.98%–60.15%) descriptors.Awidevariabilityinmolecularpropertiesoftheconsideredchemicalsrevealstheimportanceoftheselecteddescriptorsforproposedmodelingstudieshere.

Pattern of the frequency distribution of the experimental toxicityvalues ($pIGC_{50}$) of the chemicals in training and test sets used for theclassificationandregressionmodelingwasassessedthroughconstructing the histogram (Fig. 2a). Frequencies of the experimental toxicityvalues in various sub-ranges (bins) are represented as vertical bars.Thehistogramshowsanearlynormaldistributionoftheexperimentaltoxicityvaluesfortheselectedsetofchemicals.Theradarchartanalysis(Fig. 2b)showsthatthecompoundsusedinourdatasetcoveredasufficientlylargechemicalspace.Tanimotosimilarityanalysisyieldedsmallvaluesof TSIforthetoxic(0.011),non-toxic(0.013),andcompletedata(0.011)(Fig.2c),indicatingstructurallydiversenatureofchemicals.InTanimotosimilarityassessment,thecut-offvaluesforsimilar/dissimilarmoleculesaregenerallyacceptedbetween 0.85and0.70(Manleyetal.,2010).

Classificationmodeling

Classification modeling was performed to categorize the chemicalsamong two categories (TPT and non-TPT) as well as four categories(highlytoxic,toxic,harmful,andnotharmful)ofthechemicals.Accord-ingly, EL-approach based models (DTB and DTF) were constructed fortwoandfour-categoryclassificationsusingtheselectedmoleculardescriptors (Table 1). Optimal architecture and the model parameterswere determined through the internal and external validation procedures.Internalvalidationwasperformedusingthe10-foldCV,whereas,forexternalvalidation,asubsetofdata(20%)wasused.TheCVresultsboth for two and four-category classifications are given in Table 2. FortheTPTversusnon-TPTclassification,theaverage(tenruns)valuesofMR rendered by DTB and DTF models are between 8.90% and 8.97%,whereas, in the four-category classification, the models yielded theaverageMRvaluesof25.72%and26.69%.Theresultsindicatethatthetoxicity prediction accuracies of both the models are comparable intwo and four-category classifications. The results have also shown noobvious over-fittingofdata.

TheY-randomizationwasperformedbothintwo-andfour-category classificationsofchemicalsusing10-foldCVprocedure.Intwo-categoryclassification, the average MR value of these scrambled DTB and DTFmodelswerefoundtobe16.36%and16.25%,respectively,whereasin four-category classification the respective models yielded MR valuesof36.58%and36.16%,whicharesignificantlyhigherthanthoseoftheoriginal DTB and DTF classification systems (Tables 3a, 3b). This suggests that the original classification models are relevant and unlikelytoariseasaresultofchanceofcorrelation.

Two-categoryclassification

In the selected optimal DTB model, the total number of trees in se-ries, maximumdepth of tree, number of average group splits, andshrinkagefactorvalueswere394,10,1028.1,and0.01,respectivel y.Intwo-categoryclassification,theDTBmodelfitsthelogitvalues(proba-bility).Theshrinkagefactorimprovesthepredictiveaccuracyofa DTBs eries(Friedman,2001).Ontheotherhand,thetotalnumberoftreein series, maximum depth of a tree, and the number of average groupsplitsinoptimalDTFmodelwere188,17,and84.2,respectively.Th eop-timal DTB and DTF models were applied to the test and completedatasets. Contribution of the selected descriptors in classification modelsrangedbetween16.40%–100%(DTB)and17.85%–100%(DTF).Thedis- criminating descriptors in each model were determined in view oftheir importance in corresponding model. The importance of the inde-pendent variables in each model was determined using the differenceinMRcalculatedusingactualdatavaluesofallpredictorsandth osecom-putedthroughrandomlyrearrangementvaluesoftheeachpredictor.

Theperformanceparameters ofDTBandDTFmodelsforthetraining, testandcompletedataarepresentedinTable3a.TheMRvaluesyieldedb y DTB and DTF models were 1.10% and 1.17% in complete data. Theoverallaccuracyofthetrainingandtestsetswere100%and94.48%f orDTB,and100%and94.14%forDTF,whereasthevaluesofMCCforth erespectivemodelsintrainingandtestdatawere1,0.83(DTB)and1, 0.83 (DTF). As shown in the Table 3a, the performance of both themodels was reasonably good. The results showed that the sensitivityand specificity values for DTB and DTF classification were more than96% incompletedata.

Further,theperformanceofaclassificationmodelcanbequantified by calculating the area under the ROC curve (AUC). Both the DTB andDTFmodelsyieldedAUCvaluesof1.0intrainingand0.94invalidati onset. Moreover, the average gain values in DTB and DTF models rangedbetween1.136-1.929and1.209-2.150,respectivelyinvalidationdata.

Four-categoryclassification

The optimal DTB and DTF models with the total number of trees inseries (400, 124), maximum depth of tree (10, 21), number of averagegroupsplits(5699.5,241.9),andshrinkagefactor(DTB,0.01)w ereusedfor four-category toxicity classification of the chemicals. Contribution ofthe selected descriptors in four-category classification models ranged be-tween27.52%–100%(DTB)and43.43%–100%(DTF).Themeansensitivity, specificity,accuracyandMCCvaluesofdifferentmodelsforthetraining,t estandcompletedataaresummarizedinTable3b.BoththeDTBandDTFm odelsyieldedtheMRvalueof3.72%incompletedata.Themeanaccu-racy and MCC values in the training and test sets for both the modelswere100%,90.69%,1.0,and0.76,respectively.Theaveragegain valuesin DTB and DTF models ranged between 1.481-1.975 and 1.482-1.989,respectivelyinvalidationdata.Asevident,theperformanceofth eDTBandDTFmodelsconstructedherewerereasonablygood.

Regressionmodeling

Regressionmodelingwasperformedtopredictthetoxicity($pIGC_{50}$) ofchemicalsusingtheELbasedmodelingmethods(DTB,DTF).Selectedd escriptorsforregressionmodelingaregiveninTable1.Optimalarchi-tecture and the model parameters were determined through the 10-fold CV. A criterion of minimum MSE value (training and validation)was used to determine the optimal model parameters. The averagevalues of MSEs and R^2 in CV and training data for the proposed regres-sion model are presented in Table 4a. It is evident that the values ofMSEs and R^2 in ten different data folds ranged between 0.155-0.279,and 0.712-0.854 (DTB) and 0.173-0.277, and 0.717-0.835 (DTF), respectively and their average values 0.220, 0.793 (DTB), and 0.227,0.786(DTF).Thesevaluesarequitecomparabletotheresultsobtain edwhen establishing the models in the training (DTB 0.042, 0.963; DTF0.047,0.960)andtest(DTB0.128,0.874;DTF0.129,0.876)treatment. Similarly, for thegivencompounds, predictionsare generallyvery

similarinthetenruns.TheY-randomizationresultsforDTBandDTFregression models derived using 10-fold CV procedure yielded R^2 valuesofb0.001inboththecases,whichrevealedthattheoriginalregression modelsarerelevantandunlikelytoariseasaresultofchanceofcorre-lation. These results indicate that both the models herein investigatedarerobust andshowedno over-fitting ofdata.

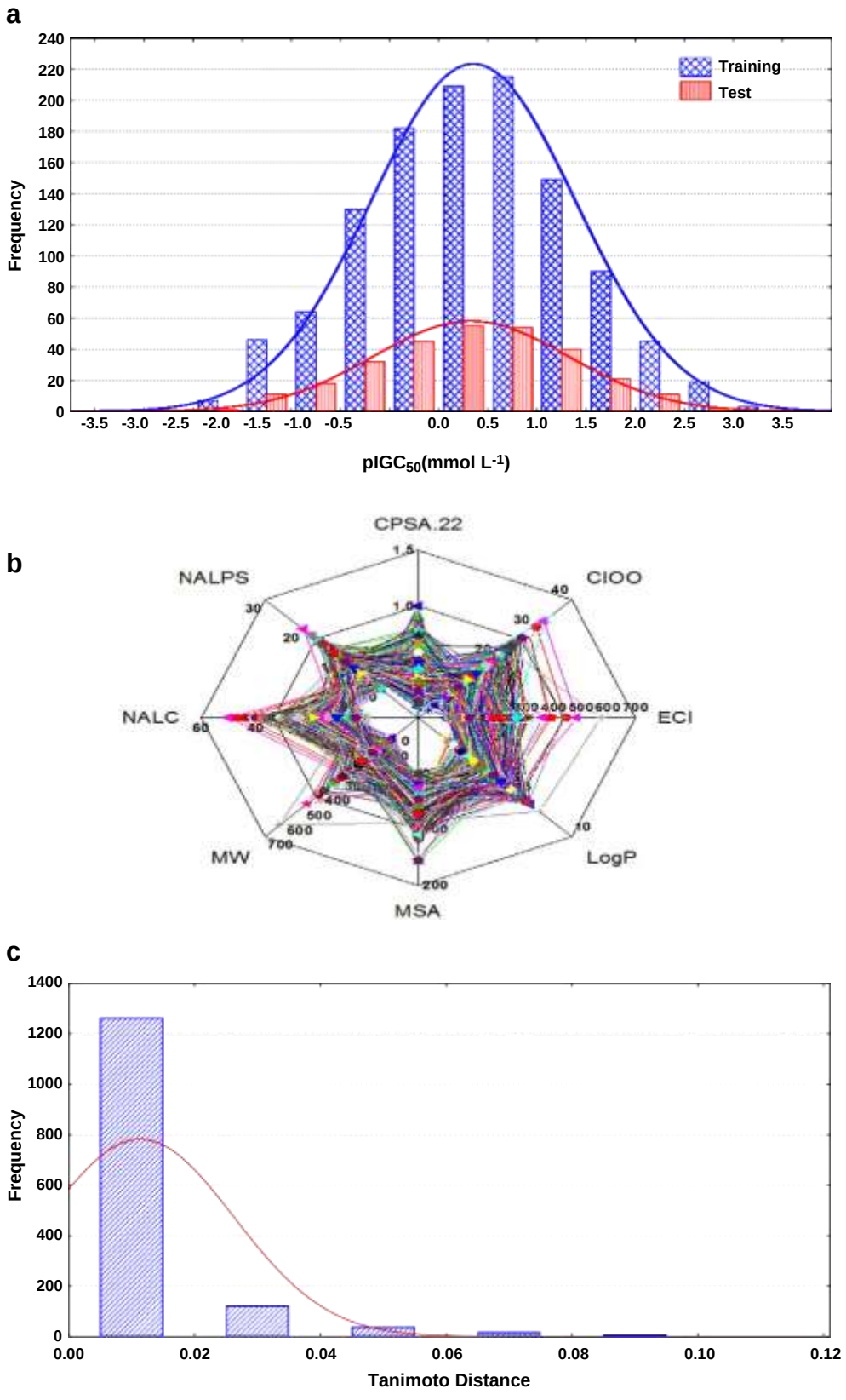


Fig.2. (a)Histogramofthetoxicityvalues(pIGC$_{50}$)of chemicalsin *T. pyriformis*intrainingand testsets,(b)Radarplotof thecompletetoxicitydata(*T. pyriformis*), and(c)HistogramofTanimotodistanceofcompletedataset(*T.pyriformis*).

Table2
Misclassificationrate(%)intendifferentfoldsoftoxicitydataincrossvalidationofclassification(DTB,DTF)models.

Run	2-categoryclassification				4-categoryclassification			
	DTB		DTF		DTB		DTF	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
1	0.15	7.59	0.15	8.28	0.92	24.14	1.00	26.21
2	0.15	9.66	0.15	11.03	0.77	28.28	0.84	31.03
3	0.08	13.79	0.08	13.79	0.77	26.21	0.77	24.83
4	0.15	8.97	0.15	8.28	1.00	26.21	1.30	27.59
5	0.15	11.72	0.08	13.10	0.84	28.28	1.00	31.03
6	0.15	7.59	0.15	7.59	0.54	27.59	0.61	28.28
7	0.15	6.90	0.15	6.90	0.92	27.59	1.00	26.90
8	0.23	7.59	0.15	7.59	0.69	20.00	2.53	22.07
9	0.08	6.90	0.08	6.21	0.84	24.83	0.84	24.14
10	0.23	8.28	0.15	6.90	0.77	24.14	0.84	24.83
Average	0.15	8.90	0.13	8.97	0.80	25.72	1.07	26.69

TheoptimalDTBandDTFregressionmodelshavethetotalnumberof trees in series, maximum depth of tree, and the number of averagegroup splits 435, 190; 10, 21; 973.1, 609.5, respectively. Contributionof the selected descriptors in constructed regression models rangedbetween19.36%–100%(DTB)and24.02%–100%(DTF).Theoptimal DTB and DTF models were applied to the test and complete datasets.TheconstructedDTBandDTFmodelsexplained96.05%,95.60%varianceintraining,87.08%,86.94%varianceintest,and94.37%,93.98%varian cein complete data. Proportion of variance explained by the model vari-ablesisthebestsinglemeasureofhowwellthepredictedvaluesmatch the actual values. A model predicting exactly matching valueswith measured ones would explain 100% variance in data. The twomodels yielded MSE and R^2 values of 0.059, 0.945 (DTB) and 0.064,0.944(DTF)incompletedata.Valuesoftheperformancecriteriapara m-eters(R^2,MAE,andRMSE)yieldedbytheregressionmodelsintraining,testa ndcompletedataarepresented inTable4b.

Further,acloselyfollowedpatternofvariationbythemeasuredand modelpredictedtoxicitiesofchemicalsbytheconstructedDTBandDTFmod els in the training and test phases (Fig. 3) suggest that both themodelsperformedreasonablywell.Plotsofthemodel-predictedresponses(DTBandDTF)andthecorrespondingresidualsforthe train-ing and test sets show almost complete independence and randomdistribution(Fig.4).

Table3a
ClassificationresultsfortoxicitypredictionofchemicalsinTPTandnon-TPTcategoriesbyensemble learningmodels.

Model/sub-sets	Class	Totalcases	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
Trainingset						
DTB	TPT	910	100.00	100.00	100.00	1.00
	Non-TPT	250	100.00	100.00	100.00	1.00
	Total	1160				
DTF	TPT	910	100.00	100.00	100.00	1.00
	Non-TPT	250	100.00	100.00	100.00	1.00
	Total	1160				
Testset						
DTB	TPT	230	96.52	86.67	94.48	0.83
	Non-TPT	60	86.67	96.52	94.48	0.83
	Total	290				
DTF	TPT	230	96.92	84.13	94.14	0.83
	Non-TPT	60	84.13	96.92	94.14	0.83
	Total	290				
Completeset						
DTB	TPT	1140	99.30	97.42	98.90	0.97

Further,thepredictiveperformanceoftheproposedEL-based(DTBand DTF) regression models was compared with the KPLS model. InKPLS, the kernel function (RBF) was selected on the basis of R^2 andRMSE values of the validation set. The kernel function parameter (σ)and number of the LVs in the feature space were determined on thebasis of the minimum CV error value (0.41). The optimum values of σandLVswere0.5and5,respectively.Theresultspertainingtotheper-formance criteria parameters for the KPLS and EL-based regression(DTB, DTF) models are presented in Tables 4a, 4b. The optimal KPLSyielded R^2 and RMSE values of 0.678, 0.59 in training, 0.803, 0.44 intest, and 0.701, 0.56 in complete data. The results suggest that DTBandDTFmodelsperformedrelativelybetterthantheKPLSinpredicting the $pIGC_{50}$ values of the chemicals, which may be attributed to theboostingandbaggingalgorithmsimplementedinDTBandDTFapproac hes.

Testingwithexternaldatasets

TheconstructedDTBandDTFmodelswerealsoappliedtothreedif-ferent external validation datasets collected from literature (Zhanget al., 2010) to further assess their versatility towards using these forpredictivepurposes.Thesedatasetsreportexperimentaltoxicities(IG C_{50}) of chemicals in *T. pyriformis* (TSI = 0.26; n = 15), and EC_{50}valuesin*V.fischeri*(TSI=0.04;n=96),amarinebacterium,and *S.obliguue*(TSI =0.17; n =37), an algae. The *T. pyriformis*data was usedbothforclassificationandregression,whereasothertwodatasetswer eusedforregressionalone.BoththeproposedDTBandDTFmodelsyielded1 00%accuracyofclassification(*T.pyriformis)*intwo-categoryscheme.Theclassificationresultssuggestthattheproposeddiscri mina-tionmodelsperformedwellwithexternaldata.

TheconstructedELregressionmodelswerealsoappliedtothethreediffe renttoxicitydatasets(*T.pyriformis*,*V.fischeri*,*S.obliguue*)withan

Table3b
Classification results for toxicity prediction of chemicals by ensemble learning models infourtoxicitycategories.

Model	Totalcases	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
Trainingset					
DTB	1160	100.00	100.00	100.00	1.00
DTF	1160	100.00	100.00	100.00	1.00
Testset					
DTB	290	83.00	93.64	90.69	0.76

	Non-TPT	310	97.42	99.30	98.90	0.97
	Total	1450				
DTF	TPT	1140	99.38	96.81	98.83	0.97
	Non-TPT	310	96.81	99.38	98.83	0.97
	Total	1450				

DTF	290	83.00	93.61	90.69	0.76
Completeset					
DTB	1450	96.46	98.69	98.14	0.95
DTF	1450	96.48	98.70	98.14	0.95

objective to develop inter-species predictive models. The DTB modelyieldedMSEof0.02,0.53,0.15,andR^2of0.975,0.637and0.741between the experimental and predicted toxicity values in three testspecies, respectively, whereas in DTF model the respective diagnosticvalueswere:0.02,0.51,0.17(MSE)and0.968,0.655,0.691(R^2).Thesultssuggestthatboththe ELmodelsfitwelltothemarinebacteriaandalgaetoxicitydataandcanbeusedasinterspeciesmodelforpredictingtoxicityinthesespecies.

Applicabilitydomainoftheproposedmodels

Theapplicabilitydomain(AD)ofapredictivemodelswastakenintoaccount in order to consider the scope and limitations of the proposedmodels,i.e.the rangeof chemical structures for which the modelsare considered to be applicable (Netzeva et al., 2005). In this study, atwo-dimensional descriptor space is considered; thus, the minimumand maximum descriptor values defined a rectangle in one plane. Theranges ofimplementeddescriptorsin classificationand regressionmodels are presented in Tables 5a, 5b. It is evident that all the com-pounds in all datasetsconsidered here are within theADrange.

Discussion

Inthisstudy,wehavedevelopedandrigorouslyvalidatedEL-basedQSAR models for predicting the toxicity of diverse chemicals in *T.pyriformis*.Accordingly,two-categoryandfour-categorytoxicitycriteriabasedclassificationandregressionmodels(DTB,DTF)werecon-structed.ThefourthOECDprincipalrequiressuitablemeasureofperfor-mances.ValidationofQSARmodelsisanimportantaspectfordeterminatio n of reliability of such models (Tropsha, 2010). There aredifferentapproachesofvalidationincludinginternalandexternalvali-dation. To measure the goodness-of-fit of a model, the R^2 is calculatedbetweenpredictedandexperimentalvalues.Theoptimalarchitecturesof the respective models were determined using 10-fold CV procedureand these were further validated using external datasets. For 10 foldCV, the training data were partitioned into 10 folds and iterations oftrainingandvalidationwereperformedsuchthat,withineachiterationa different fold of data held-out for validation while the remaining 9folds were used for learning and subsequently the learned models areused to make predictions about the data in the validation fold. Thus,eachtime,amodelwasconstructedandtestedwithanunseendataset.Thisprocedurepreventstheover-fittingproblem(Singhetal.,2011).Aclose resemblance between the criteria parameters for classification(MR) and regression (MSE) models in CV and model training phases(Tables 2–4) revealed that the constructed models are robust. Modelswerefurthervalidatedusing20%ofthedataastestset,excludedduringthe model building phase, and predicted by the constructed model(Tables 3a, 3b, 4a, 4b). Y-randomization was performed to determinetheprobabilityofchanceofcorrelationduringdescriptorselectionpro-cessbothinclassificationandregressionmodeling.Thechanceof correlationmayoccurduringdescriptorselectionespeciallyifthenum-ber of descriptors is large (Jouan-Rimbaud et al., 1996). In two- andfour-category classifications, a portion of each category agents wasrandomlyselectedandinterchangedkeepingtheratioofthe classesunchanged.Inregression,valuesoftheresponsevariablewererandomlyselectedandscrambledkeepingtheindependentvariablesunchanged.Process of scrambling of the dataset was repeated 10 times both inclassificationandregressionmodeling.Therandomizationresultssuggested that original classification and regression models are relevantandunlikelytoariseasaresultofchanceofcorrelation.Theclassifica-tionand regressionmodels were also validated using the externaldatasets(*T.pyriformis*,*S.obliguue*,*V.fischeri*).Boththeoptimalclassifi-cationmodels(DTB,DTF)yielded100%accuracyfortheexternaldata(*T.pyriformis*)intwo-categorymode.

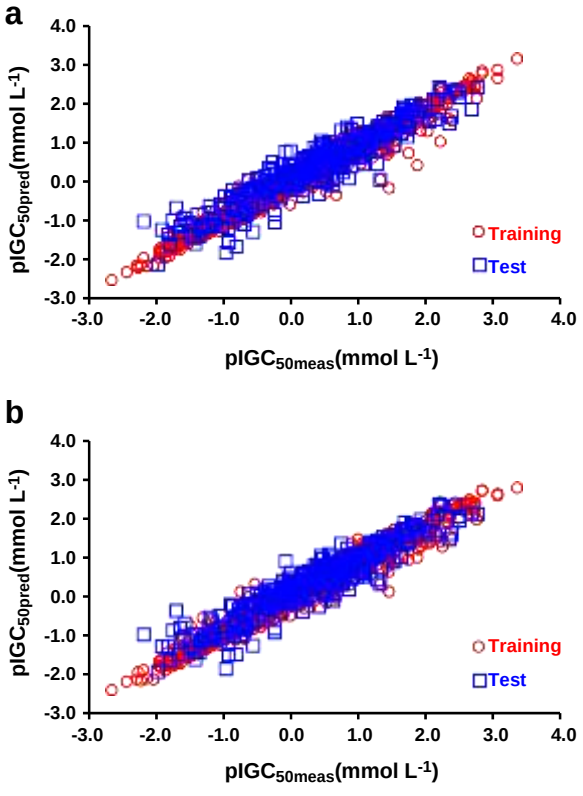
Table4a
Ten-foldcrossvalidationresultsforregressionmodels.

Run	DTB-Training	DTB-Validation	DTF-Training	DTF-Validation

Table4b
Performanceparametersforensemblelearningmodelsintoxicitypredictionofchemicals.

Model	Sub-sets	Mean	*SD	MAE	RMSE	R^2
Measured	Training	0.33	1.04	–	–	–
	Test	0.31	1.00	–	–	–
	Complete	0.33	1.03	–	–	–
DTB	Training	0.33	0.97	0.14	0.21	0.963
	Test	0.35	0.90	0.27	0.36	0.874
	Complete	0.33	0.96	0.17	0.24	0.945
DTF	Training	0.33	0.95	0.15	0.22	0.960
	Test	0.36	0.88	0.27	0.36	0.876
	Complete	0.34	0.93	0.18	0.25	0.944
KPLS	Training	0.33	0.85	0.43	0.59	0.678
	Test	0.35	0.89	0.35	0.44	0.803
	Complete	0.34	0.86	0.41	0.56	0.701

*SD—standarddeviation.



2	0.966	0.037	0.821	0.169	0.960	0.045	0.787	0.202
3	0.968	0.035	0.846	0.155	0.962	0.043	0.828	0.173
4	0.968	0.035	0.776	0.256	0.962	0.042	0.759	0.277
5	0.968	0.034	0.773	0.260	0.964	0.042	0.781	0.255
6	0.968	0.035	0.838	0.218	0.961	0.041	0.835	0.203
7	0.968	0.035	0.854	0.181	0.960	0.044	0.832	0.209
8	0.968	0.035	0.773	0.262	0.962	0.043	0.738	0.261
9	0.968	0.034	0.712	0.279	0.964	0.042	0.717	0.274
10	0.968	0.035	0.810	0.178	0.962	0.043	0.803	0.184
Average	0.968	0.035	0.793	0.220	0.963	0.043	0.786	0.227

Fig. 3. Plot of the measured and model predicted values of the $pIGC_{50}$ for training and testsetsin(a)DTB,and(b) DTFmodel.

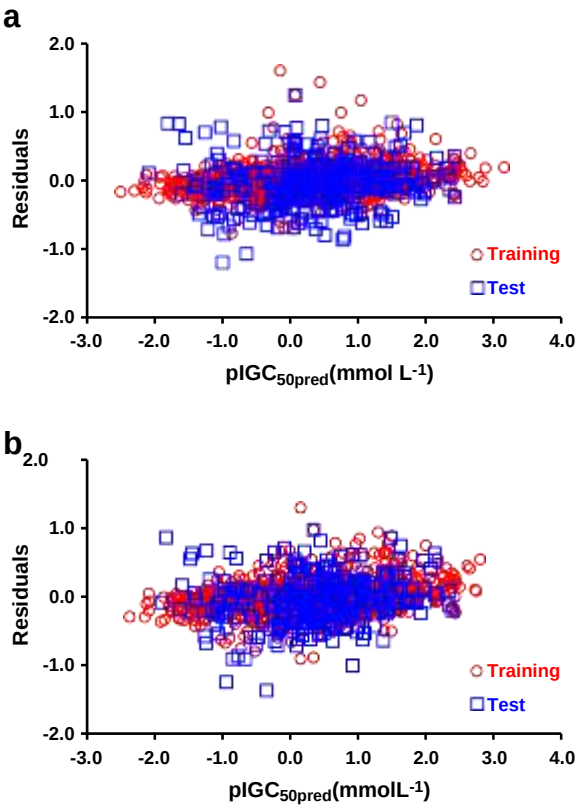


Fig.4.PlotofthemodelpredictedvaluesofthepIGC$_{50}$andresidualsfortrainingandtestsets in(a)DTB,and (b)DTFmodels.

Thepredictivepoweroftheregressionmodelswasevaluatedonexternalvalidationsetsderivingaseriesofcoefficients:R^2(characterizingthecorrelationbetweenpredictedandexperimentalvaluesinthe validation set)and coefficientsQ²(Consonni etal.,2009) and r² (Roy etal.,2008).InQ²the denominatoriscalculatedonthetrainingset, andbothnumeratoranddenominatoraredividedbythenumberof correspondingelements$\left(Q^2 = 1 - \frac{4\sum^{n_{ext}}(y_i - y_i)^2/n_{ext}}{\sum^{n_{Tr}}(y_i - y_{Tr})^2/n_{Tr}}\right)$.Consonni etal.(2009)demonstratedthatresultsobtainedbyQ²areindependentof the prediction set distribution and sample size. Recently, r^2_m metrichas been proposed as an additional validation parameter (Roy et al.,2008). This metric is calculated based on the correlations between theobservedandpredictedvaluewith(R^2)andwithout(R^2_0)interceptfor theleastsquarelines,as $r^2_m = R^2 \times (1 - \sqrt{R^2 - R^2_0})$.Itdoesnotconsider thedifferencesbetweenindividualresponsesandthetrainingsetmean,andthusavoidsover-estimationofthequalityofpredictionduetowide response range (Y-range) (Roy et al., 2009). The results of externalvalidationoftheEL-basedregressionmodels(DTB,DTF)arepresentedinTable6.

ThevalidationthresholdcriteriavaluesforQ^2_{F3}andr^2_mare0.6and0.5,respectively.Thesethresholdsofvalidationmeasuresareappliedtoalldatasets, andthen,foreveryvalidationcriteria,acountingoftheaccept-ed models performed. In our case, considerably high correlations (R^2)between the measured and predicted values of response variable bothintrainingandtestdatawereobtained.Modelisconsideredacceptable when the value of R^2 in external set exceeds 0.5 (Ojha et al., 2011).According to the proposed reference criteria (Eriksson et al., 2003),the difference between $R^2_{ext}$and R^2 should not exceed 0.3. Moreover,R^2valueof≥0.81forinvitroand≥0.64forinvivodatacanberegardedas good (Kubinyi, 1993).Model validation usingthe external data yielded criteriaparameter(Q^2_{F3},r^2_m)values (Table6) wellabove therespective thresholds (except for *V. fischeri*). As our models fulfills thesecriteria and also positively pass internal and external validation, thesecouldbeapplied to predictthetoxicity of new,untestedchemicals.

Both the two- and four category classification models (DTB, DTF)yieldedhighsensitivity,specificity,accuracyandMCCvaluesintraining,testandcompletedataarrays(Table3a,3b).Fjodorovaetal.(2010)suggestedthatthemodelforregulatorypurposeshouldbeconnectedwithhigh sensitivity.Itmaybementionedthatsensitivityisthemostimportantparameterinaclassificationmodel.Infact,thelowsensitivityvalueindicates the low ability of a model to recognize the toxicity of diversecompounds.Thespecificityisanotherimportantindicator.Highspeci-ficity value indicates the high ability of the model to recognize thefalsepositivecompoundsanditcansavetheexperimentalcosts(Cheng et al., 2011). Accuracy represents the total number of activeand inactive compounds correctly predicted among the total numberoftestedcompounds.MCCvalueequalto1isregardedasaperfectpre-diction, whereas, 0 is for a completely random prediction. It is evidentthat both theDTBand DTF modelsyieldedexcellent results.

Thetwo-categoryclassificationmodelsyieldedAUCvaluescloseto unity.HighAUCvaluesindicatetherankingqualityofclassification.It can be viewed as a measure based on pair-wise comparisons betweenclassificationsoftwoclassesandisanestimateoftheprobabilityt hat theclassifierranks arandomlychosenpositiveexamplehigherthana negativeexample.Withaperfectranking,allpositiveexamplesare rankedhigherthanthenegativeonesandAUC=1.AnydeviationfromthisrankingdecreasestheAUC,andtheexpectedAUCvaluefor arandomrankingis0.5(Fawcett,2006).Intwo-andfour-categoryclassifications,theconstructedmodelsyieldedthegainvalues(intraining andvalidation)rangingbetween1.136-2.357,and1.481-2.707,respectively.Thegainvaluesshowhowmuchimprovementthemodelprovides in picking out the best of the cases. The gain of 1 means non-selective targeting (Abdelaal et al., 2010; Singh et al., 2013a). The performancecriteriaparameters(sensitivity,specificity,accuracy,MCC), andvaluesofaveragegainfortheclassificationmodelssuggestthat boththeselectedmodelsarefullycapabletodiscriminatethechemicalsintwoandfourcategories.

Table5a
Applicabilitydomain of theselected descriptors in two-category andfour-category classificationmodels.

Descriptors	Two-categoryclassification				Four-categoryclassification				*T.pyriformis**	
	Training		Test		Training		Test			
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
CPSA.22	0.09	1.00	0.14	0.82	–	–	–	–	0.33	0.75
CIOO	2.56	33.45	6.37	21.67	2.56	33.45	5.33	21.67	6.46	10.63
ECI	2.00	573.00	19.00	196.00	2.00	573.00	14.00	196.00	45.00	123.00
LogP	−2.41	6.86	−1.95	4.93	−2.41	6.86	−1.95	4.93	1.84	4.73
MW	32.03	623.74	70.04	390.82	32.03	623.74	55.04	390.82	108.06	263.85

NALC					0.00	50.00	0.00	32.00	0.00	5.00
NALPS	0.00	20.00	0.00	15.00	–	–	–	–	6.00	12.00

*Externaltestset.

Table5b
Applicability domain oftheselecteddescriptorsinregressionmodels.

Descriptors	Trainingset		Testset		*T.pyriformis**		*S.obliguue**		*V.fischri**	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
CIOO	2.56	33.45	2.85	32.67	6.46	10.63	6.46	11.63	2.64	14.57
LogP	−2.41	6.86	−2.01	6.50	1.84	4.73	1.10	4.73	0.19	5.75
MSA	0.00	157.69	0.00	155.39	0.00	86.28	20.23	106.51	0.00	112.30
MW	32.03	623.74	46.04	483.59	108.06	263.85	93.06	326.79	58.04	281.81
NALPS	0.00	20.00	0.00	16.00	6.00	12.00	7.00	13.00	0.00	13.00

*Externaltestset.

An in-depth investigation of the results revealed that the proposedtwoandfour-categoryELclassificationmodelsincompletedatamisclassified seventeen and fifty four compounds, respectively. Thesecompounds were mainly comprised of phenols, esters, alcohols, aldehydes,acrylates,aromaticamines,andneutralorganics.Fororganicpollutants the mechanism can be distinguishedbased onthe primaryprocessesthatcausetoxicity(Verhaaretal.,1992).Non-polarnarcosisis also called baseline toxicity. This type of action results from ratherinertchemicals,whichactviaanon-specificmodeofactionandfinallyproduce an effect that is characterized as narcosis. The severity of thisnarcosis type effect and the rate at which the effect occurs depend onthehydrophobicityofthecompound.Mainlyalcohols(KolevaandBarzilov, 2010), and esters are known to cause non-polar narcosis,whereas phenols and their derivatives cause polar narcosis. Phenolsare slightly more toxic than compounds causing non-polar narcosis.The capability of these substances to form H-bond probably enhancestheir toxicity (Roex et al., 2000). Neutral organic chemicals are non-ionizableandnon-reactiveandactviasimplenon-polarnarcosis(VeithandBroderius,1990).

EL-basedregressionmodels(DTB,DTF)wereconstructedtopredict the toxicity ($pIGC_{50}$) of chemicals. The performance results (Tables 4a,4b)suggestthatboththemodelsyieldedconsiderablylow RMSE,MAE and high R^2 values in training, test and complete data. RMSE is aquadratic scoring rule which measures the average magnitude of theerror.Itgivesarelativelyhighweighttolargeerrors,hencemostusefulwhenlargeerrorsareparticularlyundesirable.MAEmeasurestheaver-age magnitude of the error in a set of predictions, without consideringtheir direction. It is a linear score which means that all the individualdifferences between predictions and corresponding measured valuesareweightedequallyintheaverage(Singhetal.,2013a).Plotsofthere-sidualsandmodelpredictedvaluesoftheresponsevariableareknowntoprovidemoreinformationregardingthemodelfitnesstoadataset.Arandom distribution of the residuals suggests that the model fits thedata well, whereas, a non-random distribution shows that the modeldoes not fit the data adequately (Singh et al., 2009b). Fig. 4 shows arandomdistributionofresidualsyieldedbyboththeModelsintrainingand test sets, suggesting for the adequacy of constructed models inpredicting thetoxicityofchemicals.

Table6
Performanceparametersoftheensemblelearningmodelsintoxicitypredictionusingexternaldatasets.

Model	Sub-sets	RMSE	R^2	Q^2_{F3}	r^2_m
DTB	Training	0.21	0.963	–	–
	Test	0.36	0.874	0.881	0.790
	Complete	0.24	0.945	0.945	0.915
	T.*pyriformis*	0.13	0.975	0.984	0.975
	S.*obliguue*	0.39	0.741	0.855	0.674
	V.*fischri*	0.73	0.637	0.501	0.561
DTF	Training	0.22	0.960	–	–
	Test	0.36	0.876	0.879	0.771

An investigation of the prediction results of both the EL models revealedthatmostpoorlypredictedcompounds,forwhichthedifferencebetween the experimental and predicted values are more than one logunit included the aromatic amines, phenols, and esters, and neutralorganics. From a mechanistic of action point of view, the phenol compoundsarecommonlyassociatedwiththeweakacidrespiratoryuncoupling mechanism of toxic action (Terada, 1990), which is knownto exert toxicity in excess of narcosis. These are generally bulky andelectronegative compounds produce their toxic effect by disruptingATP synthesis. They induce disruption of hydrogen ion gradient in theinner mitochondrial membrane (Schultz, 1999). Increased acidity ofphenols promotes their ability to uncouple the respiratory chain fromoxidative phosphorylation (Schuurmann et al., 2003). These two toxiceventsarerelatedtotheionizationabilityandunequalchargedistribu-tionprofileofacompound.

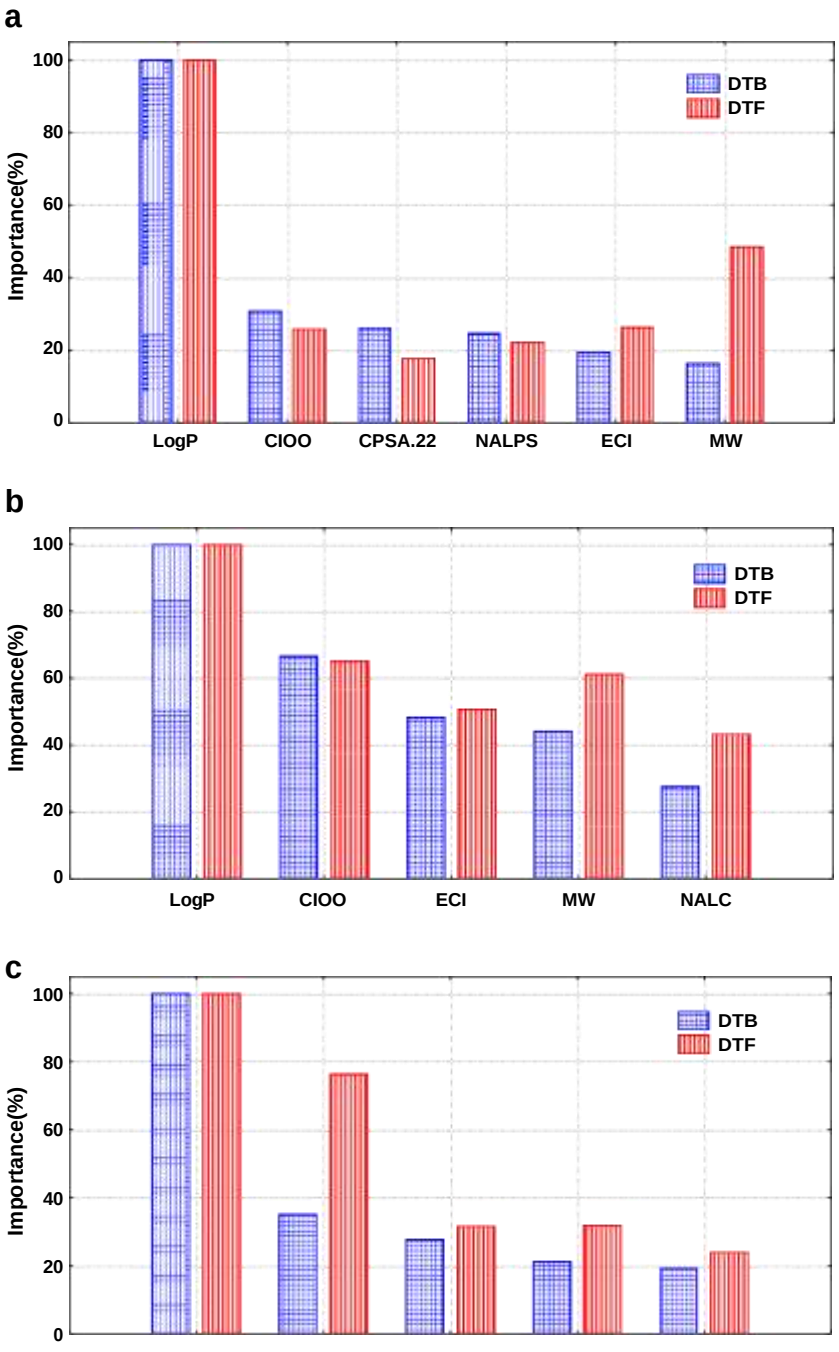
Selectionofmostrelevantmoleculardescriptorstothetoxicity prediction is important for optimizing the prediction models and forelucidatingthemolecularfactorscontributingtotoxicity.Themolecular structuresofthechemicalscontroltheiractivitiesanddescriptorsdirectly encode particular features of molecular structures. Thus, it ispossible to shed light on mechanism of toxic action of the compoundsby interpreting the descriptors in the predictive models. The toxicityprocess are very complicated and involve both toxicant transport tothe target site of interaction and interaction between the toxicant andreceptorsatthesitesandwhichmaynotbefullydescribedbyafewdescriptors,particularlyforalargenumberofheterogeneouscompoundsthan those covered in previous studies (Xue et al., 2006). A set of eightmoleculardescriptors(physico-chemical,topological,geometrical,andconstitutional) are selected here. These describe the molecular bulkproperties (including transport properties such as membrane perme-ability) and those representing chemical reactivity of the substanceunder study. Since the most toxic compounds are characterized bylowIGC_{50}(orhighpIGC_{50})values,alldescriptorshavingpositivecorrelationcoefficients(exceptCPSAandNALC)increasetheadverseeffect,whereas negative correlation coefficients led to a decreased toxicityeffect(Katritzkyetal.,2009).Fromthecontributionsofthedescriptorsinclassificationandregressionmodelshere,itcanbeconcludedthatlog P contributedthemostsignificantly to pIGC_{50}variation (Fig.5).

Thepenetration/solubilitydescriptors(likelogP)reflecttheability ofacompoundtoformnon-covalentinteractionswithitsenvironment,todissolveandpersistinwaterorinalipidicenvironment,ortoperme-atethephaseinterfaces.Generally,largerlogPindicatesastrongerabil- ity of the chemical to permeate the cell membrane of an organism and,therefore, to much more easily interact with its target in the organism(Jiang et al., 2011). The second significant molecular descriptor is theMW,whichrepresentsthebulkeffectandiscorrelatedwithlipophilicityofmolecules.Schultzetal.(2007)alsoreporteditasoneofthemostimportant descriptors in toxicity models. The topological descriptorstreat

the structure of the compound as a graph with atom as verticesand covalent bonds as edges. Both the topological descriptors (CIOOand ECI) considered here showed positive correlations with toxicity.Stearicpropertyofamoleculecanbedescribedbytopologicalde scrip-tors. It canaffect membrane transportand specific interactions atreactive sites with implications to its toxicological profile (Xue et al.,2006).Geometricaldescriptors(MSAandCPSA) reflectfeaturesof the

moleculargeometry.Theseincludedistancebetweenparticularpointsofthemolecularsurface(thetwofarthestpoints,thetwoclosestpoints)anddistancesbetweengivenchemicalgroups.IncreasingtheMSAandthusrevealingthesitesofH-bonddonationshouldalsoincreasehydro-phobicity. The CPSA has a negative contribution to the total toxicity.Morelikelythisdescriptorcanbe related to the hydrophilicity ofthe molecule (Katritzky et al., 2003). The constitutional descriptors(NALC and NALPS) account for the stearic hindrance effect. The sizeand shape of compounds influence their transport properties througha biological system as well as their stearic hindrance at the reactivesite. A larger value of the descriptor indicates the larger stearic hin-drance (Luan et al., 2005). From the above discussion, it can be seenthatthephysico-chemical,constitutional,geometricalandtopological properties of the molecules are likely major factors in the process oftoxicity, and all of the descriptors involved in the model, which haveexplicit physical meanings, may account for the structural featuresresponsible fortoxicologicalproperties ofchemicals.

Defininganappropriateapplicabilitydomain(AD)isalsoveryimportant for the application of QSAR models (Gramatica, 2007). TheADofapredictivemodeldefinestheboundarieswherebythepredicted values can be trusted with confidence. The AD of a QSAR model is thephysico-chemical,structural,orbiologicalspace,knowledgeorinformationonwhichthetrainingsetofthemodelhasbeendeveloped,andforwhich it is applicable to make predictions for new compounds. Ideallythemodelshouldonlybeusedtomakepredictionswithinthatdomainbyinterpolationandnotextrapolation(Nikolova-Jeliazkovaand



Log P MW ClOO NALPS MSA

Fig.5.Plotof thecontributionof the selecteddescriptorsin toxicityprediction(a)two-categoryclassification,(b)four-categoryclassification,and(c)regressionmodels.

Jaworska,2005).OneoftheapproachesofdefiningtheADistoestimatethe training set coverage in the molecular descriptor space. In mathematicalterms,itmeansestimationofinterpolationregionsthemultivariate space of training set, because the interpolated predictionresultsarereliablethanextrapolated.Thisapproachisespecially suitable to those models based on statistical mining techniques (Tanet al., 2010). An interpolation region in one dimensional descriptorspace is the interval between the minimum and maximum values ofthe training dataset. In this study, a two-dimensional descriptor spaceisconsidered; thus, the minimumandmaximumdescriptor valuesdefinedarectangleinoneplane.Therangesofimplementeddescriptorsarepresentedin Tables5a,5b.However,thiskindofglobal,chemometricestimationoftheADdoesnotaddressthechemicalspacecharacterizedbythedescriptorrange,soitmighthappenthatthetargetchemicalfalls in an area poorly represented in the training set. Moreover, sincethe AD relies on the chemical descriptors alone, the property underinvestigationisneglected(Fjodorovaetal.,2010).

Literaturesurveyshowsanumberofstudiesreportingdifferent modeling methods for toxicity prediction in *T. pyriformis*. The resultsof some research groups on classification and regression modeling fortoxicity ofchemicalcompoundsareshowninTables7a,7b.

Adirectcomparisonofourresultswithpreviousstudiesisinappropriate,becausethenumberofchemicalcompounds,natureandnumberofdescriptors,andmodelingapproachesconsideredinthesestudiesdif-fer to a large extent. Among various modeling approaches, MLR, PLSR,ANNs, SVMs have commonly been used in QSAR modeling (Tables 7a,7b). Although, the predictive responses achieved using modeling techniqueshavebeenwithinacceptablerange,thesemethodshavecertainlimitations. MLR and PLSR are linear methods and do not fit the datawith nonlinear structure, a common feature of experimental toxicitydata. ANNs, although a universal nonlinear method, it suffers fromoverfitting in training. SVM uses only a limited data points duringmodelbuilding phase(Singhetal., 2013a).Nevertheless,asimplecomparison of the model statistics could provide some basic informationabouttheaccuracyofvariouspredictionmethodologies.Itmaybenoted that all these studies considered toxicity data of chemicals insingle organism (*T. pyriformis*) only. Moreover, most of these studiesconsideredcomplexandlargenumberofdescriptors(includingthermodynamicandquantummechanical)andinseveralofthese,classificationandregressionaccuracieswerenotsatisfactory,thuslimiting theapplicabilityofthesemodelsfortoxicitypredictioninnewunknownchemicals.Amongthese,thepresentstudyproposedELbasedmodelingapproaches considering the dataset of structurally diverse chemicalsand using smaller number of simple molecular descriptors yielded thebestpredictionaccuracyandcorrelationforthetraining,test,completeandexternaldatasets.ProposedELmethodswithbaggingandstochas-tic gradient boosting techniques improve the prediction accuracy ofweaklearners(Breiman,1996).Thebaggingminimizespredictionvariance by generating bootstrapped replica datasets, whereas, boostingcreates a linear combination out of many models, where each newmodel in dependent on the preceding model (Friedman, 2002). Thepresent study demonstrated applicability of the constructed models(DTB,DTF)inpredictingthetoxicityofdiversechemicalsinmarinebac-teria and algae and thus, suggesting for the appropriateness of thesemethodsfortoxicitypredictionofnewchemicalsindifferentorganismsand can be used as effective tools in risk assessment for regulatorydecisionmaking.

Conclusions

In this study, multi-species ensemble learning approach based robust predictive models, such as DTB and DTF were established using alarge experimental toxicity data containing 1450 structurally diversecompounds,withoutrequiringtheknowledgeofmechanismsand choices of specific molecular descriptors derived from the chemical'sstructure. Constructed classification and regression models were validatedusinginternalandexternalvalidationprocedures,whichweredirectlydevelopedfromthemoleculardescriptors.Thestatisticalresultsproved that developed DTB and DTF models were efficient algorithmsforbuildingtwo-categoryandfour-categorytoxicityclassificationmodelsofTPTprediction.OptimalDTBandDTFmodelsdemonstratedexcellentpredictiveabilitiesindiscriminatingtheTPTandnon-TPT,aswell as finer toxicity classes of chemicals and can be used as a tool inscreening of the new chemicals. The DTB and DTF regression modelsexhibitedexcellentabilityinpredictingthetoxicityofthediversechemicals with the toxicity end points in *T. pyriformis*, marine bacteria(*V.fischeri*),andalgae(*S.obliguue*)testspecies.Anotableadvantageofthe present study may be use of simple non-quantum mechanical de-scriptors as estimators of the toxicity end-point. Excellent predictiveandgeneralizationachievedforproposedELmodelsmaybeattributed

Table7a
Classificationaccuraciesoftoxicitymodelsfrom differentstudies reportedintheliterature.

Model	No.ofcompound	No.ofdescriptors	Typeofdescriptors	Accuracy (%)	Reference
SVM	337	3	PD,OH	71.00	Ivanciuc(2004)
SVM	1129	199	CD,PD,GD,CoD,OH	88.90	Xueetal.(2006)
LR	1129	199	CD,PD,GD,CoD,OH	70.10	Xueetal.(2006)
Adaboost	274	9	ED, StD,ThD	92.80(Test)	Niuetal.(2009)
C4.5	1571	166	MDL	89.20(Test)	Chengetal.(2011)
k-NN	1571	166	MDL	91.60(Test)	Chengetal.(2011)
DTB (2-cat)	1450	6	PD, CD, TD,GD	100.00(Train) 94.48(Test) 98.90 (Complete) 100.0(ES)	Presentstudy
DTF (2-Cat)	1450	6	PD,CD,TD, GD	100.00(Train) 94.14(Test) 98.83 (Complete) 100.0(ES)	Presentstudy
DTB	1450	5	PD, CD,TD	100.00(Train)	Presentstudy

(4-cat)				90.69(Test)	
DTF	1450	6	PD, CD,TD	98.14 (Complete)100.0 0Train)	Presentstudy
(4-Cat)				90.69(Test) 98.14 (Complete)	

PD—Physico-chemicaldescriptor;CD—Constitutionaldescriptors;GD—Geometricaldescriptor;CoD;Connectivitydescriptors;ThD—thermodynamicdescriptor;ED—Electronicdescriptors;StD—Stearic descriptors; TD—topologicaldescriptors;OH—others; ES—T.pyriformisexternal set.

Table7b
Prediction result for toxicity ofchemicalsfromdifferent studiesreportedintheliterature.

Model	No.ofcompounds	No.ofdescriptors	Typeofdescriptors	R^2	Reference
RSM	200	2	PD,QD	0.540	Croninetal.(2002)
kNN	250	173	–	0.646	Guoetal.(2005)
GRNN	202	–	HD,EpD,OH	0.794	Panayeetal.(2006)
RBF-NN	221	25	PD,QD,OH	0.942(Train) 0.882(Val)	Melagrakietal.(2006)
SCR-qQNA	200	–	QNAD	0.685	Lagunietal.(2007)
SVM	983	36	MolconnZ	0.890(Train) 0.830(Val)	Zhuetal.(2008)
NN	250	6	PD,CD,OH	0.710(Train) 0.730(Val)	Enochetal.(2008)
MLR	250	168	PD,CD,GD, QD,OH	0.660(Train) 0.720(Val)	Enochetal.(2008)
MDF-SAR	30	4	TD,GD	0.974	Jäntschietal.(2008)
MLR	313	–	ABSIL	0.730(Train) 0.697(Test)	Castillo-Garitetal.(2009)
Linear-QSTR	35	3	VD	N0.972	Vlaiaetal.(2009)
ANN	392	7	–	0.822(Train) 0.815(Test)	FatemiandMalekzadeh(2010)
PLS	95	58	PD,OH	0.840	Artemenkoetal.(2011)
Stepwise-MLR	250	95	PD,MZD	0.600	Jiangetal.(2011)
DTB	1450	5	PD,CD,GD,TD	0.963(Train) 0.874(Test) 0.945(Complete) 0.975(ES)	Presentstudy
DTF	1450	5	PD,CD,GD,TD	0.960(Train) 0.876(Test) 0.944(Complete) 0.968(ES)	Presentstudy

PD—physico-chemicaldescriptor;TD—Topologicaldescriptors;GD—Geometricaldescriptor;CD—constitutionaldescriptor;QD—quantummechanicaldescriptor;EpD—electrophilicity descriptors;HD—hydrophobicitydescriptors;VD—vanderwaalspacedescriptors;QNAD—Quantitativeneighbor-hoodsofatomsdescriptors;MZD—Molconn-Zdescriptors;ABSIL—atom-based non-stochasticlinearindices;OH—others; ES—*T.pyriformis* external set.

tothe fact thattheymake no assumptionregarding the underlinedistributionofvaluesofpredictorvariables.Thoughtheneedofa nimaltestingcannotbecompletelyignored,theproposedmodelscanbeuse daseffectivetoolsforscreeningthechemicalsfortheirtoxicityprofile.

Conflictofinterest statement

Theauthorsdeclarethattherearenoconflictsofinterest.

Acknowledgments

TheauthorsthanktheDirector,CSIR-IndianInstituteofToxicologyResearch,Lucknow(India)forhiskeenintere stinthisworkandprovid-ing allnecessaryfacilities.

AppendixA.Supplementarydata

Supplementarydatatothisarticlecanbefoundonlineathttp://dx.doi.org/10.1016/j.taap.2014.01.006.

References

Abdelaal, M.M.A., Farouq, M.W., Sena, H.A., Salem, A.B.M., 2010. Applied classification sup-portvectormachineforprovidingsecondopinionofbreastcancerdiagnosis.Online J.Math.Stat.1,1–7.

Aptula,A.,Roberts,D.,Cronin,M.,Schultz,T.,2005.Chemistrytoxicityrelationshipsforthe effects of di- and trihydroxybenzenes to *Tetrahymena pyriformis*. Chem. Res.Toxicol.18,844–854.

Artemenko, A.G., Muratov, E.N., Kuzmin, V.E., Muratov, N.N., Varlamova, E.V., Kuzmina,A.V., Gorv, L.G., Golius, A., Hill, F.C., Leszczynski, J., Tropsha, A., 2011. QSAR analysisof nitroaromatics toxicity in *Tetrahymena pyriformis*: structural factors and possiblemodesofaction.SARQSAREnviron.Res.22,575–601.

Benigni, R., Netzeva, T.I., Benfenati, E., Bossa, C., Franke, R., Helma, C., Hulzebos, E.,Marchant,C.,Richard,A.,Woo,Y.P.,Yang,C.,2007.Theexpandingroleofpredictivetox icology:anupdateonthe(Q)SARmodelsformutagensandcarcinogens. J.Environ.Sci.HealthPartC25,53–97.

Benigni,R.,Bossa,C.,Richard,A.,Yang,C.,2008. Anovelapproach:chemicalrelationaldatabases,andtheroleoftheisscandatabase onassessingchemicalcarcinogenicity.Ann.Ist.Super.Sanita44,48–56.

Breiman,L.,1996.Heuristicsofinstabilityandstabilizationinmodelselection.Ann.Stat. 24,2350–2383.

Casalegno,M.,Benfenati,E.,Sello,G.,2005.Anautomatedgroupcontributionmethodinp redictingaquatictoxicity:thediatomicfragmentapproach.Chem.Res.Toxicol.18,7 40–746.

Castillo-Garit, J.A., Escobar, J., Marrero-Ponce, Y., Torrens, F., 2009. Atom-based quadraticindicestopredictaquatictoxicityofbenzenederivativesto*Tetrahymenap yriformis*.13rdInternationalElectronicconferenceonSyntheticOrganicChemist ry(ECSCO-13),1–30November,pp.1–26.

Cheng, F., Shen, J., Yu, Y., Li, W., Liu, G., Lee, P.W., Tang, Y., 2011. In silico prediction of*Tetrahymena pyriformis* toxicity for diverse industrial chemicals with substruc- turepatternrecognitionandmachinelearningmethods.Chemosphere82,1636– 1643.

Consonni, V., Ballabio, D., Todeschini, R., 2009. Comments on the definition of the Q2parameterforQSARvalidation.J. Chem.Inf. Model.49, 1669–1678.

Cronin, M., Schultz, T., 2001. Development of quantitative structure-activity relationshipsfor the toxicity of aromatic compounds to *Tetrahymena pyriformis*: comparativeassessmentofthemethodologies.Chem.Res. Toxicol.14,1284–1295.

Cronin, M.T.D., Aptula, A.O., Duffy, J.C., Netzeva, T.I., Rowe, P.H., Valkova, I.V., Schultz,T.W.,2002.ComparativeassessmentofmethodstodevelopQSARsforthe predictionofthe toxicityof phenols toTetrahymena pyriformis. Chemosphere49,1201–1221.

Cronin,M.T.D.,Netzeva,T.I.,Dearden,J.C.,Edwards,R.,Worgan,A.D.P.,2004.Assessment andmodelingofthetoxicityoforganicchemicalsto*Chlorellavulgaris*:development ofanoveldatabase.Chem.Res.Toxicol.17,545-554.

Devillers,J.,2004.LinearversusnonlinearQSARmodelingofthetoxicityofphenolderivati vesto*Tetrahymenapyriformis*.SARQSAREnviron.Res.15,237-249.

Deweese,A.D.,Schultz,T.,2001.Structure-activity relationshipsforaquatictoxicityto *Tetrahymena*:halogen-substitutedaliphaticesters.Environ.Toxicol.16,54–60.

Dietterich,T.G.,2000.Ensemblemethodsinmachinelearning.Lect.NotesComput.Sci. 1857,1-15.

Duchowicz, P.R., Ocsachoque, M.A., 2009. Quantitative structure-toxicity models forheterogeneousaliphatic compounds.QSARComb. Sci. 28, 281–295.

Efron,B.,1979.Bootstrapmethods:anotherlookatthejackknife.Ann.Stat.7,1–26.

Enoch, S.J., Cronin, M.T.D., Schultz, T.W., Madden, J.C., 2008. An evaluation of global QSARmodels for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*.Chemosphere71,1225–1232.

Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M., Gramatica, P., 2003.Methods for reliability and uncertainty assessment and for applicability evaluationsofclassification-andregression- basedQSARs.Environ.HealthPerspect.111,1361-1375.

Eroglu, E., Palaz, S., Oltulu, O., Turkmen, H., Ozaydin, C., 2007. Comparative QSTRstudy using semi-empirical andfirst principle methods based descriptors foracute toxicity of diverse organic compounds to the fathead minnow. Int. J. Mol.Sci.8,1265-1283.

Fatemi,M.H.,Malekzadeh,H.,2010.PredictionofLog($1/IGC_{50}$)$^{-1}$forbenzenederivativestociliate *Tetrahymena pyriformis* from their molecular descriptors. Bull. Chem. Soc. Jpn.83,233-245.

Fawcett,T.,2006.AnintroductiontoROCanalysis.PatternRecogn.Lett.27,861-874.

Fjodorova, N., Vracko, M., Novic, M., Roncaglioni, A., Benfenati, E., 2010. New public QSARmodelsforcarcinogenicity.Chem.Cent.J.4,1-15.

Friedman,J.H.,2001.Greedyfunctionapproximation:agradientboostingmachine.Ann. Stat.29,1189-1232.

Friedman,J.H.,2002.Stochasticgradientboosting.Comput.Stat.DataAnal.38,367-378.Gramatica,P.,2007.PrinciplesofQSARmodelsvalidation:internalandexternal.QSAR Comb.Sci.26,694-701.

Guo,G.,Neagu,D.,Cronin,M.T.D.,2005.Astudyonfeatureselectionfortoxicitypredic-tion.Lect.NotesComput.Sci.3614,31-34.

Hancock,T.,Put,R.,Coomans,D.,VanderHeyden,Y.,Everingham,Y.,2005.Aperformancecomparisonofmodernstatisticaltechniquesformoleculardescriptorselectionandretention prediction in chromatographic QSRR studies. Chemom. Intell. Lab. Syst.76,185-196.

Huber,P.,1964.Robustestimationofalocationparameter.Ann.Math.Stat.53,73-101.In,Y.,Lee,S.K.,Kim,P.J.,No,K.T.,2012.PredictionofacutetoxicitytofatheadminnowbyLocalModel basedQSARandglobalQSARapproaches.Bull.KoreanChem.Soc.33, 613-619.

Ishwaran,H.,Kogalur,U.B.,2010.Consistencyofrandomsurvivalforests.Stat.Probab. Lett.80, 1056-1064.

Ivanciuc,O.,2004.Supportvectormachinespredictionofthemechanismoftoxicactionfrom hydrophobicity and experimental toxicity against *Pimephales promelas* and*Tetrahymenapyriformis*.InternetElectron.J.Mol.Des.2004(3),802-821.

Jalali-Heravi, M., Kyani, A., 2008. Comparative structure-toxicity relationship study ofsubstituted benzenes to *Tetrahymena pyriformis* using shuffling-adaptive neurofuzzyinferencesystemandartificialneuralnetworks.Chemosphere72,733-740.

Jäntschi, L., Popescu, V., Bolboacă, S.D., 2008. Toxicity caused by para-substitutedphenolson*Tetrahymenapyriformis*:thestructure-activityrelationships.Electron. J. Biotechnol. 11, 1-12.Jiang,D.X.,Li,Y.,Li,J.,Wang,G.X.,2011.Predictionoftheaquatictoxicityofphenolsto*Tetrahymenapyriformis*frommoleculardescriptors.Int.J.Environ.Res.5,923-938.

John,M.B.,Geoffrey,M.D.,Willet,P.,1998.Chemicalsimilaritysearching.J.Chem.Inf. Comput.Sci.38,983-996.

Jouan-Rimbaud,D.,Massart,D.L.,deNoord,O.E.,1996.Randomcorrelationinvariableselection for multivariate calibration with a genetic algorithm. Chemom. Intell. Lab.Syst.35,213-220.

Katritzky, A.R., Oliferenko, P., Oliferenko, A., Lomaka, A., Karelson, M., 2003. Nitrobenzenetoxicity:QSARcorrelationsandmechanisticinterpretations.J.Phys.Org. Chem.16,811-817.

Katritzky,A.R.,Slavov,S.H.,Stoyanova-Slavova,I.S.,Kahn,I.,Karelson,M.,2009.Quantativestructure-activityrelationship(QSAR)modelingofEC_{50}ofaquatictoxic-ities for*Daphniamagna*.J.Toxicol. Environ.Health 72, 1181-1190.

Koleva,Y.,Barzilov,I.,2010.Comparativestudyofmechanismofactionofallylalcoholsfordifferentendpoints.Sci.Work.Univ.49,55-59.

Kubinyi, H., 1993. Methods and Principles in Medicinal Chemistry. In: Mannhold, R.,Kroogsgard-Larsen,P.,Timmerman,H.(Eds.),vol. 1. VCH,Wiley.

Lagunin, A.A., Zakharov, A.V., Filimonov, D.A., Poroikov, V.V., 2007. A new approach toQSARmodellingofacutetoxicity.SAR QSAREnviron.Res. 18,285-298.

Luan,F.,Zhang,R.S.,Zhao,C.Y.,Yao,X.J.,Liu,M.C.,Hu,Z.D.,Fan,B.T.,2005.Classificationofthecarcinogenicityof*N*-nitrosocompoundsbasedonsupportvectormachinesandlineardiscriminantanalysis .Chem.Res.Toxicol.18,198-203.

Lukacinova,A.,Mojzis,J.,Benacka,R.,Lovasova,E.,Hijova,E.,Nistiar,F.,2007.*Tetrahymenapyriformis*asavaluableunicellularanimalmodelfordeterminationofxenobiotics.Proceedings27thInternationalSymposium"IndustrialToxicology",Bratislava,30.May-1June2007,pp.1-6.

Mahajan,D.T.,Masand,V.H.,Patil,K.N.,Hadda,T.B.,Rastija,V.,2013.IntegratingGUSARandQSARanalysesforantimalarialactivityofsyntheticprodigininesagainstmultidrugresistantstrain.Med.Chem.Res.22,2284–2292.

Mahjoobi, J., Etemad-Shahidi, A., 2008. An alternative approach for the prediction ofsignificant wave heights based on classification and regression trees. Appl. OceanRes.30,172-177.

Manley, P.W., Stiefl, N., Cowan-Jacob, S.W., Kaufman, S., Mestan, J., Wartmann, M.,Wiesmann, M., Woodmanb, R., Gallagher, N., 2010. Structural resemblances andcomparisons of the relative pharmacological properties of imatinib and nilotinib.Bioorg.Med.Chem.18,6977-6986.

Masand,V.H.,Jawarkar,R.D.,Patil,K.N.,Mahajan,D.T.,Hadda,T.B.,Kurhade,G.H.,2010.COMFA analysis and toxicity risk assessment of coumarin analogues as Mao-Ainhibitors:attemptingbetterinsightindrugdesign.Der.Pharm.Lett.2,350-357.

Melagraki,G.,Afantitis,A.,Makridima,K.,Sarimveis,H.,Igglessi-Markopoulou,O.,2005.PredictionoftoxicityusinganovelRBFneuralnetworktraining methodology. J.Mol.Model.8,1-9.

Melagraki,G.,Afantitis,A.,Makridima,K.,Sarimveis,H.,Igglessi-Markopoulou,O.,2006.Prediction of toxicity using a novel RBF neural network training methodology.J.Mol.Model.12,297-305.

Netzeva,T.,Schultz,T.,2005.QSARsfortheaquatictoxicityofaromaticaldehydesfrom *Tetrahymena*data.Chemosphere61,1632-1643.

Netzeva,T.,Schultz,T.,Aptula,A., Cronin,M.,2003.Partialleastsquaresmodellingoftheacute toxicity of aliphatic compounds to *Tetrahymena pyriformis*. SAR QSAR Environ.Res.14,265-283.

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 deSandt, J.J.M., Tong, W., Veith, G., Yang, C., 2005. Current status of methods for definingtheapplicabilitydomainof(quantitative)structure–activityrelationship.ATLA33,155–173.

Nikolova-Jeliazkova, N., Jaworska, J., 2005. An approach to determining applicabilitydomainsforQSARgroupcontributionmodels:ananalysisofSRCKOWWIN.Altern.Lab.Anim.33,461–470.

Niu, B., Jin, Y., Lu, W., Li, G., 2009. Predicting toxic action mechanisms of phenols usingAdaBoostLearner.Chemom.Intell.Lab.Syst.96,43–48.

Ojha,P.K.,Mitra,I.,Das,R.N.,Roy,K.,2011.Furtherexploringr²m metricsforvalidationofQSPRmodels.Chemom.Intell.Lab.Syst.107,194–205.

Panaye, A., Fan, B.T., Doucet, J.P., Yao, X.J., Zhang, R.S., Liu, M.C., Hu, Z.D., 2006. Quantitative structure–toxicity relationships (QSTRs): a comparative study of variousnon linear methods. General regression neural network, radial basis functionneuralnetworkandsupport vector machine in predicting toxicity of nitro-andcyano-aromaticsto *Tetrahymena pyriformis*. SARQSAREnviron. Res.17,75–91.

Ren, S., 2003. Ecotoxicity prediction using mechanism and non-mechanism base QSARs: apreliminarystudy.Chemosphere53,1053–1065.

Ren, S., Schultz, T., 2002. Identifying the mechanism of aquatic toxicity of selectedcompounds by hydrophobicity and electrophilicity descriptors. Toxicol. Lett. 129,151–160.

Roex, E.W.M., Gestel, C.A.M.V., Wezel, A.P.V., Straalen, N.M.V., 2000. ratios between acuteaquatictoxicityandeffectsonpopulationgrowthratesinrelationtotoxicantmode ofaction.Environ.Toxicol.Chem.19,685–693.

Roy,D.,Parthasarathi,R.,Maiti,B.,Subramanian,V.,Chattaraj,P.,2005.Electro-philicity as a possible descriptor for toxicity prediction. Bioorg. Med. Chem. 13,3405–3412.

Roy,P.P.,Leonard,J.T.,Roy,K.,2008.ExploringtheimpactofsizeoftrainingsetsforthedevelopmentofpredictiveQSARmodels.Chemom.Intell.Lab.Syst.90,31–42.

Roy, P.P., Paul, S., Mitra, I., Roy, K., 2009. On two novel parameters for validation ofpredictiveQSARmodels.Molecules29,1660–1701.

Schultz, T.W., 1997. Tetratox: *Tetrahymena pyriformis* population growth impairmentendpoint—asurrogateforfish lethality. Toxicol.Methods7, 289-309.

Schultz,T.W., 1999. Structure–toxicity relationships for benzenes evaluated with *Tetrahymenapyriformis*.Chem.Res.Toxicol.12,1262–1267.

Schultz, T.W., Netzeva, T.I., 2004. Development and evaluation of QSARs for ecotoxic end-points: The benzene response-surface model for *Tetrahymena toxicity*. In: Cronin,M.T.D., Livingstone, D.J. (Eds.), Modeling environmental fate and toxicity. CRC Press,BocaRaton,FL,USA,pp.265-284.

Schultz, T., Netzeva, T., Roberts, D., Cronin, M., 2005. Structure toxicity relationshipsfor the effects to *Tetrahymena pyriformis* of aliphatic, carbonyl-containing, R, a-unsaturatedchemicals.Chem.Res.Toxicol.18,330–341.

Schultz, T.W., Hewitt, M., Netzeva, T.I., Cronin, M.T.D., 2007. Assessing applicabilitydomainsoftoxicologicalQSARs:definition,confidenceinpredictedvalues,andtheroleofmechanismsofaction.QSARComb.Sci.26,238–254.

Schuurmann,G.,Aptula,A.,Kuhne,R.,Ebert,R.,2003.Stepwisediscriminationbetweenfour modes of toxic action of phenols in the *Tetrahymena pyriformis* assay. Chem.Res.Toxicol.16,974–987.

Serra,J.,Jurs,P.,Kaiser,K.,2001.Linearregressionandcomputationalneuralnetworkprediction of tetrahymena acute toxicity for aromatic compounds from molecularstructure.Chem.Res.Toxicol.14,1535–1545.

Singh,K.P.,Basant,A.,Malik,A.,Jain,G.,2009a.Artificialneuralnetworkmodelingoftheriverwaterquality—acasestudy.Ecol.Model.220,888–895.

Singh,K.P.,Ojha,P.,Malik,A.,Gain,J.,2009b.Partialleastsquaresandartificialneuralnetworksmodelingforpredictingchlorophenolremovalfromaqueoussolution.Chemom.Intell.Lab.Syst.99,150–160.

Singh, K.P., Basant, N., Malik, A., Jain, G., 2010. Modeling the performance of "upflowanaerobic sludge blanket" reactor based wastewater treatment plant using linearand nonlinearapproaches—acasestudy. Anal.Chim.Acta658, 1–11.

Singh,K.P.,Basant,N.,Gupta,S.,2011.Supportvectormachineinwaterqualitymanagement.Anal.Chim.Acta703,152–162.

Singh, K.P., Gupta, S., Kumar, A., Shukla, S.P., 2012. Linear and nonlinear modelingapproachesforurbanairqualityprediction.Sci.TotalEnviron.426,244–255.

Singh,K.P.,Gupta,S.,Rai,P.,2013a.Predictingacuteaquatictoxicityofstructurallydiversechemicals in fish using artificial intelligence approaches. Ecotoxicol. Environ. Saf. 95,221–233.

Singh,K.P.,Gupta,S.,Rai,P.,2013b.Predictingdissolvedoxygenconcentrationusingkernelregressionmodelingapproacheswithnonlinearhydrochemicaldata.Environ.Monit.Assess.http://dx.doi.org/10.1007/s10661-013-3576-6.

Snelder,T.H.,Lamouroux,N.,Leathwick,J.R.,Pella,H.,Sauquet,E.,Shanker,U.,2009. Predictivemappingofthenaturalflowregimesof France.J.Hydrol.373,57–67.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E., 2003. TheChemistryDevelopmentKit(CDK)—anopen-sourceJavalibraryforchemo-andbio-informatics.J.Chem.Inf.Model.43,493–500.

Tan, N.X., Li, P., Rao, H.B., Li, Z.R., Li, X.Y., 2010. Prediction of the acute toxicity of chemicalcompoundstofatheadminnowbymachinelearningapproaches.Chemometr.Intell.Lab.Syst.100,66–73.

Terada,H.,1990.Uncouplersofoxidativephosphorylation.Environ.HealthPerspec.87,213–218.

Tetko,I.V.,Sushko,I.,Pandey,A.K.,Zhu,H.,Tropsha,A.,Papa,E.,Oberg,T.,Todeschini,R.,Fourches,D.,Varnek,A.,2008.CriticalassessmentofQSARmodelsofenvironmentaltoxicity against *Tetrahymena pyriformis*: focusing on applicability domain andoverfitting by variable selection.J. Chem. Inf.Model. 48,1733–1746.

Toropov, A.A., Toropova, A.P., Benfenati, E., Manganaro, A., 2010. QSAR modeling of thetoxicityto*Tetrahymenapyriformis*bybalanceofcorrelations.Mol.Divers.14,821–827.

Tropsha,A.,2010.BestpracticesforQSARmodeldevelopment,validation,andexploitation.Mol.Inf.29,476–488.

Veith,G.D.,Broderius,S.J.,1990.Rulesfordistinguishingtoxicantsthatcausetype Iand type I Inarcosissyndromes.Environ.Health Perspect.87,207–211.

Verhaar, H.J.M., Van Leeuwen, C.J., Hermens, J.L.M., 1992. Classifying environmentalpollutants.Chemosphere25,471–491.

Vlaia,V.,Olariu,T.,Ciubotariu, C.,Medeleanu,M.,Vlaia,L.,Ciubota,D.,2009.Moleculardescriptorsforquantitativestructure–toxicityrelationship(QSTR),II.Threeovalitymoleculardescriptorsandtheiruseinmodelingthetoxicityofaliphaticalcoholson*Tetrahymenapyriformis*.Rev.Chim. (Bucuresti)60,1357–1361.

Wang, Y., Zheng, M., Xiao, J., Lu, Y., Wang, F., Lu, J., Luo, X., Zhu, W., Jiang, H., Chen, K., 2010.Usingsupportvectorregressioncoupledwiththegeneticalgorithmforpredicting acutetoxicity tothefatheadminnow.SARQSAREnviron.Res. 21,559–570.

Wang,G.,Hao,J.,Ma,J.,Jiang,H.,2011.Acomparativeassessmentofensemblelearningforcreditscoring.ExpertSyst.Appl.38,223–230.

Xue, Y., Li, H., Ung, C.Y., Yap, C.W., Chen, Y.Z., 2006. Classification of a diverse set of*Tetrahymena pyriformis* toxicity chemical compounds from molecular descriptors bystatisticallearningmethods.Chem.Res.Toxicol.19,1030–1039.

Yang,P.,Yang,Y.H.,Zhou,B.B.,Zomaya, A.Y.,2010.Areviewofensemblemethodsinbioinformatics.Curr.Bioinforma.5,296–308.

Zhang,X.J.,Qin,H.W.,Su,L.M.,Qin,W.C.,Zou,M.Y.,Sheng,L.X.,Zhao,Y.H.,Abraham,M.H.,2010. Interspecies correlations of toxicity to eight aquatic organisms: theoreticalconsiderations.Sci.TotalEnviron.408,4549–4555.

Zhao, C.Y., Zhang, H.X., Zhang, X.Y., Liu, M.C., Hu, Z.D., Fan, B.T., 2006. Application ofsupport vector machine (SVM) for prediction toxic activity of different data sets.Toxicology217,105–119.

Zhu,H.,Tropsha,A.,Fourches,D.,Varnek,A.,Papa,E.,Gramatica,P.,O¨berg,T.,Dao,P.,Cherkasov,A.,Tetko,I.V.,2008.CombinatorialQSARmodelingofchemicaltoxicantstestedagainst*Tetrahymena pyriformis*.J.Chem. Inf.Model.48, 766–784.