Mathematical structural descriptors and mutagenicity assessment: A study with congeneric and diverse data sets

Subhabrata Majumdar[a]*, Subhash C. Basak[b], Claudiu N. Lungu[c], Mircea V. Diudea[c] and Gregory D. Grunwald[d]

[a]*University of Florida Informatics Institute, 432 Newell Dr, CISE Bldg E251, Gainesville FL 32611, USA;*

[b]*Department of Chemistry and Biochemistry, University of Minnesota, 246 Chemistry Building, 1039 University Drive, Duluth MN 55812, USA;*

[c]*Department of Chemistry, Babes-Bolyai University, Strada Arany János 11, Cluj-Napoca 400028, Romania;*

[d]*Natural Resources Research Institute, University of Minnesota, 5013 Miller Trunk Highway, Duluth MN 55811, USA.*

*Correspondence e-mail: smajumdar@ufl.edu

Abstract:

Quantitative bioactivity and toxicity assessment of chemical compounds plays a central role in drug discovery as it saves a substantial amount of resources. To this end, high-performance computing has enabled researchers and practitioners to leverage hundreds, or even thousands, of computed molecular descriptors for activity prediction of candidate compounds. In this paper, we evaluate the utility of two large groups of chemical descriptors in such predictive modelling, as well as chemical structure discovery, through empirical analysis. We use a suite of commercially available and in-house software to calculate molecular descriptors for two sets of chemical mutagens- a homogeneous set of 95 amines, and a diverse set of 508 chemicals. Using calculated descriptors, we model the mutagenic activity of these compounds using a number of methods from the statistics and machine learning literature, and use robust principal component analysis to investigate the low-dimensional subspaces that characterize these chemicals. Our results suggest that combining different sets of descriptors is likely to result in a better predictive model- but that depends on the compounds being modelled and the modelling technique being used.

Keywords: molecular descriptors; variable selection; dimension reduction; machine learning; two-deep cross validation; quantitative structure-activity relationship (QSAR)

1. Introduction

Hazard assessment of chemicals is often carried out in data poor situations [1]. The Toxic Substances Control Act (TSCA) Inventory, maintained by the United States Environmental Protection Agency (USEPA), currently has about 85,000 entries [2]. A large fraction of these chemicals has very little or no data needed for their hazard estimation [3]. The assessment of chemical mutagenicity is important both for environmental protection and drug discovery. Identification of potential mutagenicity for industrial chemicals and environmental pollutants is prerequisite to the protection of human and ecological health. For drug discovery, early mutagenicity detection for drug candidates can help in the effective allocation of resources in drug design protocol which costs on the average over US $2 billion [4].

Laboratory testing of mutagenicity for all possible candidate chemicals can be very expensive. Therefore, assessment of potential mutagenicity of chemicals from Quantitative Structure-Activity Relationship (QSAR) models has been accepted for evaluation of chemicals in lieu of experimental mutagenicity data [5]. Such models carry out property/ bioactivity/ toxicity assessment in silico, i.e. without actually performing the experiments, and using quantitative modelling techniques instead that predict properties of compounds using molecular descriptors. In the early stages of QSAR during the middle of the 20th century, the effectiveness of such approaches was limited by the handful of descriptors that could be calculated using the limited computational resources available. This situation has drastically changed in the past two or three decades. High-performance computing has enabled researchers to calculate hundreds or even thousands of descriptors using various software [6-11] in a reasonable amount of time, thus generating a vast amount of information to potentially build effective models for chemical activity prediction. For this reason, development, computation and usage of molecular descriptors have a central role in the present landscape of QSAR research.

In this paper, we consider two approaches towards QSAR descriptor calculation and present an evaluation of their utility. The first set of descriptors [8-11] consists of those developed and used by Basak and co-workers over the past decades towards effective QSAR model formulation [7, 12-15]. The second descriptor set is computed by Diudea and co-workers using Schrodinger and in-house software TopoCluj [16-18]. Examples of their effectiveness in mapping the chemical activity landscape include [19, 20]. Keeping the above in mind, the goal of this paper is two-fold: (1) Present a comparison of the two predictor sets (separately and combined) for the mutagenicity assessment of two data sets, viz., a homogeneous set of 95 aromatic and heteroaromatic amines and a structurally diverse set of 508 chemicals using a battery of various statistical and machine learning approaches, and (2) use robust principal component analysis to explore how the combined set of descriptors map the underlying low-dimensional subspace of chemical properties.

The rest of the paper is organized as follows. Section 2 presents the details of our methodology- data, descriptors, and techniques used for model building and validation. In section 3, we present and elaborate on the findings from our analysis. We conclude the paper with a discussion in section 4.

2. Materials and methods

2.1. Data

The two datasets used in this paper represent two different type of scenarios that practitioners are likely to encounter while doing QSAR analysis. The first data consists of the mutagenic activities of 95 congeneric amines on bacterial samples from the TA98

S. typhimurium strain [21]. The response variable, measured as the log number of revertants per nmol when a chemical compound is applied to the S. typhimurium test cultures, were studied in the original study by Debnath et al [21]. While the compounds in this dataset are very similar to each other in chemical structure, our second dataset consists of data on 508 chemical compounds from several different chemical classes. Table 1 summarizes this classification of the chemical compounds (note that a compound can belong to two or more classes). Collected from the CRC Handbook of Identified Carcinogens and Non-carcinogens [22], the response variable in this dataset is the 0/1 mutagen or non-mutagen status of the chemical compounds as determined by the Ames test of mutagenicity. In total the data contains 256 mutagens and 252 non-mutagens. For each set of compounds, we calculated their corresponding descriptors sets using two sets of software, previously used by Basak et al [12-14] and Diudea et al [23].

<Insert Table 1 here>

2.2. Descriptors

For this study we use two collections of molecular descriptors. One set of descriptors, used frequently by the Cluj team of Diudea and collaborators, are calculated by the programs Schrodinger [16] and TopoCluj [17]. More detailed references about these descriptors are given in Supplementary Tables 1 and 4. For the 95 and 508 data sets, 185 and 201 such descriptors are calculated.

The second set of molecular descriptors, used frequently by Basak et al, is calculated by the software POLLY [8], MolConnZ [24], Triplet [11], and MOPAC [9]. For the 95 and 508 chemical sets, 275 and 307 descriptors are calculated by these software, respectively. Among them, 18 descriptors are common with the Cluj descriptors for the 95 amines data, while for the 508 compounds diverse dataset there are 22 common descriptors. Thus, the number of descriptors in the union of these two sets are 442 and 486, respectively.

Collectively, the combined set of descriptors consist of important classes of topological descriptors like connectivity indices, valence connectivity indices, electrotopological indices, information-theoretic neighbourhood complexity indices of various order developed by Basak et al [25], Triplet indices, electrotopological state indices and a variety of shape and size indices used in numerous QSAR studies by various authors.

2.3. Statistical and machine learning methods

We use three types of methods to build our predictive models.

2.3.1. Dimension reduction

The hundreds of descriptors generally used in chemometric analysis generally have a high degree of correlation among them [25]. For this reason, dimension reduction techniques, such as Principal Component Analysis (PCA) or Partial Least Squares (PLS) have seen widespread use in QSAR model building [26-28]. In this paper, we build predictive models using the following two dimension reduction methods:

Principal Component Regression (PCR): We transform descriptor matrix X, we transform it by multiplying with a principal component loading matrix Γ:

$$T_x = X\Gamma$$

where the number of columns in Γ denote the minimum number of principal components (PCs) that explain 95% of the total underlying variation. We follow the analysis of [28] and apply a robust PCA procedure [29] to obtain the PC loadings. Following this, we use the transformed data matrix T(X) as the matrix of predictors in linear and logistic regression models to predict activities in the 95 and 508 compound datasets, respectively.

Partial Least Squares (PLS): Another popular method in QSAR literature, PLS uses latent variables to model the correlation between predictors and the response variables. Mainly used to build models used in prediction purposes, PLS obtains a sequence of linear regression coefficients by successively regressing orthogonal components in the data matrix on those in the response vector.

2.3.2. Variable selection

Since the datasets we are dealing with are inherently high-dimensional (i.e. large number of predictors that can potentially be more than the number of samples), we use sparse regression methods for variable selection.

Least Absolute Shrinkage and Selection Operator (LASSO): In a linear or generalized linear model, the lasso method [30] obtains sparse estimates of the coefficient estimates by setting some entries to exactly zero. In the QSAR context, this means some predictors will have zero effect on the response variable. Thus, the lasso method is able to perform simultaneous variable selection and model building.

Smoothly Clipped Absolute Deviation penalty (SCAD): Proposed by [31], SCAD is another penalization method that selects sparser models than lasso, i.e. models where more entries in the coefficient vector are set at 0, without compromising on the predictive capability of the model.

2.3.3. Machine learning

Our goal in this paper is to assess and compare the predictive capabilities of different descriptor sets. Machine learning methods are known to produce models with high predictive performance, even though interpreting such ‘black box’ models is often difficult [32]. For this reason, we use the following two methods in our study.

Random Forest (RF): This method trains multiple decision trees on a dataset, each based on a randomly selected subset of total features. The final prediction in a regression problem is taken as the average of individual predictions from all the trees, while in classification problem the final class prediction is done by majority voting. Previous examples of the use of RF models in QSAR include [33-35].

Gradient Boosting Machine (GBM): Gradient boosting attempts to fit the data using multiple ‘weak learners’, which are simple models that work slightly better than random guessing. At first a weak learner is trained on the data, residuals are obtained from that model and those are again fit using weak learners. Boosting methods have proven to be very useful in predictive model building since their proposal. Examples of boosting in the QSAR scenario include [36, 37].

2.4. Validation

We use a ‘two-deep’ multi-split cross validation scheme to evaluate our predictive methods. Multi-split means we consider multiple random train-test splits of the data, build a model on the train partition, evaluate them on the test partition, and compare different methods using the average values of a metric (e.g. Root Mean Squared Error, Area Under Curve etc.) across all such test sets. This has been referred in the QSAR literature as Monte-Carlo Cross Validation [38], and ensures that the true underlying components in a model (e.g. important predictors or principal components) are more and more likely to be recovered accurately as sample size increases [38, 39]. The phrase ‘two-deep’ means we repeat the dimension reduction/ tuning parameter selection steps of the method being implemented. This ensures that information from the test samples are not used while training the model, and gives a more accurate picture of the predictive capability of the technique being analysed [7, 40-42].

3. Results

In this section, we state and discuss the outputs from our analysis. Section 3.1 is concerned about the predictive models and the comparison of outputs across different methods and predictor sets, while in Section 3.2 we list the top principal components for evaluating the effects of the new Cluj descriptors with respect to previous findings on the same datasets. All data analyses were done using the statistical software R v3.3.2 [43].

3.1. Details of predictive models

For PCR on the 95 amines data, it took 32, 17 and 40 PCs respectively to explain 95% of the total variance in the Basak, Diudea and combined descriptor sets, respectively. These numbers are 32, 95 and 39 respectively, for the 508 compounds dataset.

We use the maximum possible number of latent components while training PLS models on the 95 amines data. This sets the number of components to 75, considering the sample size is 76 for each training set in our 5-fold cross-validation. For the logistic PLS models on the 508 data we use 10 latent variables, since computation becomes very slow for higher number of PLS components due to high sample size and a generalized linear model setup.

The Lasso models on all the three sets of descriptors (Basak, Diudea and combined) for

95 amine data contained 75 descriptors, while the SCAD models had 13, 0 and 26 descriptors, respectively. On the other hand, for the 508 compound diverse data, Lasso selects 21 and 25 descriptors from the Basak and combined set respectively, while SCAD selects 15 and 18 descriptors. Both methods select no descriptors when only the Diudea set is analysed.

Each random forest model was built from an ensemble of 500 trees. We used the default setting in the R software for the depth of each tree- which means that minimum node size was 5 for the 95 amines data (as it was a regression problem), and was 1 for the 508 compound data (since it was a classification problem).

Gradient Boosting models depend on a number of tuning parameters- number of trees, and interaction depth, shrinkage parameter and minimum node size in each tree. We selected them from a grid of values using 5-fold cross-validation. For each descriptor and data setting, the selected values are given in Table 2.

<Insert Table 2 here>

3.2. Predictive performance of models

<Insert Table 3 here>

It is evident from the results in Table 3 that the PLS method with the combined set of descriptors gives the best predictive model (AUC = 0.86) while the model based on Basak group index only is also close second (AUC = 0.85).

<Insert Table 4 here>

For the congeneric set of 95 aromatic and heteroaromatic amines, the PLS, RF and GBM methods yield the best predictive models when the combined sets of descriptors were used. Results derived using the Basak group indices only come out as close second ones.

Methods that depend directly on sparse linear combinations of predictors: Lasso, SCAD do not perform well in either case. In some cases (e.g. lasso on 95 amines on Diudea descriptors) they even select a null model, i.e. models with no non-zero descriptor effects. This means there is high degree of nonlinearity among the relationship between the responses and predictors, and activities of compounds are more dependent on lower-dimensional subspaces in the predictor space than individual predictors. These subspaces are often predictive of the response, as corroborated by the good performance of PLS for both the datasets. PLS has the best performance among all methods on the combined set of predictors in the 95 amines data, and performs competitively in the more diverse 508 compounds dataset.

Interestingly, for the 508 compounds dataset, more predictors do not always equate to better prediction. A reason for this can be the fact that this dataset is composed of chemical compounds from diverse classes. In comparison, the homogeneous 95 compound dataset always gives better prediction with the combined set of predictors than either group of predictors alone.

3.2. Principal Component Analysis of descriptor sets

The results obtained using the robust PCA procedure are presented in Tables 5 and 6 and the below discussion. Interestingly, even though the Diudea set of descriptors do not perform well in prediction by themselves, they come up almost completely as the ones with top loadings in the robust PCA results (all non-bold descriptors).

<Insert Table 5 here>

For the 95 amines data set, first four PCs explained 36.8% of the variance in the data, and it needs 28 PCs to explain 90% of the total variance. These PCs represent important features of chemical graph. Since PCs are orthogonal to each other, higher order PCs do not contain characteristics encoded in the first four PCs. The energetic variable E_ele has very high loadings across all PCs (Highest for PC1 to 3, third highest for PC4), advising that the activity of 95 amines is possibly guided by an energetic dependent process, i.e. interaction with a potential target. These loadings are positive for PC1 and PC4, while negative for the other two: which is justified by the dynamic nature of the interaction. These values reflect distinct structural states, possibly conformational changes in compounds by virtue of interaction with a certain target.

<Insert Table 6 here>

For the 508 structurally diverse set, first four PCs explained 50.3% of the variance in data, and 12 PCs were needed to explain 90% of total variance. Loadings in case of this group of compounds are high with respect to intrinsic physicochemical properties. The top 4 PCs have highest loadings for the following relevant variables: E_tor (0.96), vsurf_DW23 (0.95), density (-0.93), GCUT_SlogP_0 (0.96), which are part of the so-called absorption-distribution-metabolism-excretion (ADME) features.

4. Conclusion

In this paper we explored the utility of two large sets of chemical descriptors (separately and combined) using a case study on two sets of chemical compounds. Our findings highlight the fact that a higher number of descriptors is often beneficial for predictive model building- but that is contingent on the specific set of compounds being modelled, as well as the use of proper modelling and validation techniques. Both for the 95 and 508 mutagens, results in tables 3 and 4 show that the combined set of descriptors gave better predictive models as compared to one set of descriptors only.

We analysed the intrinsic dimensionality of the descriptor spaces for the two data sets using PCA (Tables 5 and 6). For the combined set of descriptors, the set of 95 amines needed first 28 PCs whereas 12 PCs were needed to explain the same level of variance for the diverse 508 data set. Such results are in line with the earlier studies of

Basak et al [15, 26] with 90 calculated topological indices and a structurally diverse set of 3,692 chemicals where first ten PCs explained 92.6% of the variance in the data.

In the realm of predictive quality of models developed by different statistical and machine learning methods, no one method emerged as the best one. Whereas for the 508 mutagens PLS gave the best results, GBM and RF were superior to PLS for the 95-congeneric set. However, GBM and RF have consistently competitive performance across the methods used and subsets of descriptors (single or combined). This underscores the utility of 'black box' machine learning models when prediction, not interpretation of descriptor effects on the activity, is the only goal of a QSAR model building exercise.

In the PC loadings, for both the data sets, physicochemical, conformational, and ADME related properties dominate the first four PCs. The Diudea group of descriptors seem to be controlling a high proportion of overall variance, as seen in the PCA loadings in both datasets. However, they do not perform well in the predictive models across different methods (tables 3 and 4). This hints at the need for a more detailed exploration of chemical subspaces that are possibly guided by the specific activity being modelled. To this end, methods like supervised PCA [44-46] can be useful. We intend to pursue this in future research.

From a mechanistic point of view, the mutagenicity of the 95 amine mutagens were well predicted by the set of computed descriptors consisting of quantifiers of molecular shape, size, and electronic characters. Such factors are also related to hydrophobicity which was found by Debnath et al [21] to be a critical factor underlying the mutagenicity of this set of chemicals. The case of the 508 diverse set is more complex. Due to its heterogenous composition, this dataset has to be decomposed into different smaller structural classes and QSARs need to be developed for class-specific mechanistic interpretation. Such studies are in progress which will be published subsequently.

Conflict of Interest

We confirm that there is no conflict of interest on the content of this paper.

Supplementary material

Supplementary material (Supplementary tables 1-4) is available on the publisher's web site along with the published article. Supplementary files 1, 2 and 4 give more details about the descriptors used, and file 3 gives individual split results for all methods and both datasets. All data and code are available in GitHub: https://github.com/shubhobm/Mutagenicity-assessment.

Acknowledgement

The research of SM is supported by George Michailidis.

References

[1] National Research Council, *Toxicity Testing Strategies to Determine Needs and Priorities*, National Academy Press, Washington, DC, 1984.

[2] Toxic Substances Control Act (TSCA) Inventory. Available at https://19january2017snapshot.epa.gov/tsca-inventory/about-tsca-chemical-substance-inventory_.html. [Accessed 11-4-2018].

[3] C. M. Auer, J. V. Nabholz and K. P. Baetcke, *Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5*, Environ. Health, Persp. 87 (1990), pp. 183-197.

[4] J. A. DiMasi, H. G. Grabowski and R. W. Hansen. *Innovation in the pharmaceutical industry: new estimates of R&D costs*, J. Health Econ. 47 (2016), pp. 20-33.

[5] R. Benigni, *Quantitative structure-activity relationship (QSAR) Models for Mutagens and Carcinogens*, CRC Press, Boca Raton, FL, 2003.

[6] E. L. Piparo and A. Worth, *Review of QSAR Models and Software Tools for Predicting Developmental and Reproductive Toxicity*, JRC Scientific and Technical Reports EUR 24522 EN, Ispra, Italy, 2010.

[7] S. C. Basak and S. Majumdar, *Current landscape of hierarchical QSAR modeling and its applications: Some comments on the importance of mathematical descriptors as well as rigorous statistical methods of model building and validation*, in *Advances in Mathematical Chemistry and Applications Vol. 1*, Bentham e-Books, 2016, pp. 251-281.

[8] S. C. Basak, D. K. Harriss and V. R. Magnuson, *POLLY v2.3*, Copyright of the University of Minnesota, 1988.

[9] J. Stewart, *MOPAC Version 6.00, QCPE #455*, Frank J. Seiler Research Laboratory, US Air Force Academy, CO, 1990.

[10] *Sybyl Version 6.2*, Tripos Associates, Inc, St. Louis, MO, 1995.

[11] S. Basak, G. Grunwald and A. Balaban, *TRIPLET*, Copyright of the Regents of the University of Minnesota, 1993.

[12] S. C. Basak, B. D. Gute and G. D. Grunwald, *A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters*, in *Topological Indices and Related Descriptors in QSAR and QSPR*, J. Devillers and A. T. Balaban, eds., Gordon and Breach Science Publishers, Amsterdam, The Netherlands, 1999, pp. 675-696.

[13] S. C. Basak, D. Mills, B. D. Gute and D. M. Hawkins, *Predicting Mutagenicity of Congeneric and Diverse Sets of Chemicals Using Computed Molecular Descriptors: A Hierarchical Approach*, in *Quantitative structure-activity relationship (QSAR) models of mutagens and carcinogens*, R. Benigni, ed., CRC Press, Boca Raton, FL, 2007, pp. 215-242.

[14] S. Majumdar and S. C. Basak, *Beware of external validation! – A Comparative Study of Several Validation Techniques used in QSAR Modelling*, Curr. Comput. Aided Drug Des. 14 (2018), in press.

[15] S. Basak, V. Magnuson, G. Niemi, R. Regal and G. Veith, *Topological indices: their nature, mutual relatedness, and applications*, Math. Modelling 8 (1987), pp. 300-305.

[16] *Small-Molecule Drug Discovery Suite 2009*, Schrödinger LLC, New York, NY, 2009.

[17] O. Ursu and M. V. Diudea, *TOPOCLUJ software program*, Babes-Bolyai University, Cluj, Romania, 2005.

[18] *Molecular Operating Environment (MOE)*, Chemical Computing Group ULC, Montreal, QC, Canada, 2004.

[19] C. Lungu, M. Diudea, M. Putz and I. Grudziński, *Linear and Branched PEIs (Polyethylenimines) and Their Property Space*, Int. J. Mol. Sci. 17 (2016), pp. 555.

[20] C. N. Lungu, *C-C Chemokine receptor type 3 inhibitors: Bioactivity prediction using local vertex invariants based on thermal conductivity layer matrix*, Stud. Univ. Babes-Bolyai Chem. LXIII-1 (2018), pp. 177-188.

[21] A. Debnath, G. Debnath, A. Shusterman and C. Hansch, *A QSAR Investigation of the Role of Hydrophobicity in Regulating Muagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in Salmonella typhimurium TA98 and TA100*, Environ. Mol. Mutagen. 19 (1992), pp. 37-52.

[22] J. V. Soderman, *CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database*, CRC Press, Boca Raton, FL, 1982.

[23] C. Lungu, S. Ersali, B. Szefler, A. Pirvan-Moldovan, S. C. Basak and M. V. Diudea, *Dimensionality of big data set explored by cluj descriptors*, Stud. Univ. Babes-Bolyai Chem. LXII-3 (2017), pp. 197-204.

[24] *MolconnZ v4.05*, Hall Ass. Consult., Quincy, MA, 2003.

[25] S. C. Basak, *Mathematical Structural Descriptors of Molecules and Biomolecules: Background and Applications*, in *Advances in Mathematical Chemistry and Applications Vol. 1*, S. C. Basak, G. Restrepo and J. L. Villaveces, eds., Bentham eBooks, Bentham Science Publishers and Elsevier, 2015, pp. 3-23.

[26] S. C. Basak, V. R. Magnusson, G. J. Niemi and R. R. Regal, *Determining structural similarity of chemicals using graph-theoretic indices*, Discrete Appl. Math. 19 (1988), pp. 17-44.

[27] A. Lauria, M. Ippolito and A. M. Almerico, *Combined Use of PCA and QSAR/QSPR to Predict the Drugs Mechanism of Action. An Application to the NCI ACAM Database*, Mol. Inform. 28-4 (2009), pp. 387-395.

[28] S. Majumdar and S. C. Basak, *Exploring intrinsic dimensionality of chemical spaces for robust QSAR model development: A comparison of several statistical approaches*, Curr. Comput. Aided Drug Des. 12-4 (2016), pp. 294-301.

[29] S. Majumdar, *Robust estimation of principal components from depth-based multivariate rank covariance matrix*, preprint (2015). Available at http://arxiv.org/abs/1502.07042.

[30] R. Tibshirani, *Regression Shrinkage and Selection via the Lasso*, J. R. Statist. Soc. B, 58 (1996), pp. 267-288.

[31] J. Fan and R. Li, *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties*, J. Amer. Statist. Assoc. 96 (2001), pp. 1348-1360.

[32] J. B. O. Mitchell, *Machine learning methods in chemoinformatics*, WIREs Comput. Mol. Sci. 4 (2014), pp. 468-481.

[33] V. Svetnik, A. Liaw, C. Tong and others, *Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling*, J. Chem. Inf. Model., 43-6 (2003), pp. 1947-1958.

[34] P. G. Polishchuk, E. N. Muratov, A. G. Artemenko and others, *Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity*, J. Chem, Inf. Model., 49-11 (2009), pp. 2481-2488.

[35] V. E. Kuz'min, P. G. Polishchuk, A. G. Artemenko and S. A. Andronati, *Interpretation of QSAR Models Based on Random Forest Methods*, Mol. Inform. 30 (2011), pp. 593-603.

[36] V. Svetnik, T. Wang, C. Tong and others, *Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling*, J. Chem. Inf. Model. 45-3 (2005), pp. 786-799.

[37] R. P. Sheridan, W. M. Wang, A. Liaw and others, *Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships*, J. Chem. Inf. Model. 56-12 (2016), pp. 2353-2360.

[38] Q.-S. Xu and Y.-Z. Liang, *Monte Carlo cross validation*, Chemom. Intell. Lab. Syst. 56 (2001), pp. 1-11.

[39] Y. Zhang and Y. Yang, *Cross-validation for selecting a model selection procedure*, J. Econometrics 187 (2015), pp. 95-112.

[40] D. Hawkins, S. Basak and D. Mills, *QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics*, Environ. Toxicol. Pharmacol. 16 (2004), pp. 37-44.

[41] S. Majumdar, S. C. Basak and G. D. Grunwald, *Adapting interrelated two-way clustering method for quantitative structure-activity relationship (QSAR) modeling of mutagenicity/non-mutagenicity of a diverse set of chemicals*, Curr. Comput. Aided Drug Des. 9-4 (2013), pp. 463-471.

[42] S. C. Basak and S. Majumdar, *Prediction of Mutagenicity of Chemicals from Their Calculated Molecular Descriptors: A Case Study with Structurally Homogeneous versus Diverse Datasets*, Curr. Comput. Aided Drug. Des. 11-2 (2015), pp. 117-123.

[43] R Core Team, *R: A Language and Environment for Statistical Computing version 3.3.2*, 2015.

[44] E. Bair, T. Hastie, D. Paul and R. Tibshirani, *Prediction by Supervised Principal Components*, J. Amer. Statist. Assoc. 101-473 (2006), pp. 119-137.

[45] E. Barshan, A. Ghodsi, Z. Azimifar and M. Z. Jahromi, *Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds*, Pattern Recognit. 44-7 (2011), pp. 1357-1371.

[46] X. Chen, L. Wang, J. D. Smith and B. Zhang, *Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes*, Bioinformatics 24-21 (2008), pp. 2474-2481.

Table 1: Chemical classes of samples in the 508 compound diverse dataset

Chemical class	Number of compounds
Aliphatic alkanes, alkenes, alkynes	124
Monocyclic compounds	260
Monocyclic carbocycles	186
Monocyclic heterocycles	74
Polycyclic compounds	192
Polycyclic carbocycles	119
Polycyclic heterocycles	73
Nitro compounds	47
Nitroso compounds	30
Alkyl halides	55
Alcohols, thiols	93
Ethers, sulfides	38
Ketones, ketenes, imines, quinones	39
Carboxylic acids, peroxy acids	34
Esters, lactones	34
Amides, imides, lactams	36
Carbamates, ureas, thioureas, guanidines	41
Amines, hydroxylamines	143
Hydrazines, hydrazides, hydrazones, traizines	55
Oxygenated sulfur and phosphorus	53
Epoxides, peroxides, aziridines	25

Table 2: Tuning parameters for GBM models

Parameter	**95 amines data**		
	Combined	**Basak lab**	**Diudea lab**
No. of trees	1000	1000	100
Interaction depth	2	2	2
Shrinkage	0.01	0.01	0.001
Min. node size	1	2	5
Parameter	**508 compounds data**		
	Combined	**Basak lab**	**Diudea lab**
No. of trees	1000	1000	100
Interaction depth	2	2	2
Shrinkage	0.01	0.01	0.001
Min. node size	5	1	2

Table 3: Average and standard deviations (in brackets) of Area Under Curve (AUC) over 100 random splits for different methods applied on the 508 compounds heterogeneous dataset. See supplementary table 3 for individual split results.

Method	**Descriptor set used**		
	Combined	**Basak lab**	**Diudea lab**
PCR	0.59 (0.055)	**0.78 (0.038)**	0.58 (0.057)
PLS	**0.86 (0.035)**	0.85 (0.033)	0.79 (0.038)
Lasso	0.72 (0.048)	**0.75 (0.045)**	0.63 (0.06)
SCAD	0.57 (0.061)	0.58 (0.059)	**0.62 (0.063)**
RF	**0.81 (0.036)**	0.80 (0.042)	0.79 (0.040)
GBM	0.80 (0.04)	**0.82 (0.04)**	0.75 (0.042)

Table 4: Median and mean absolute deviations (in brackets) of Mean Square Prediction Error (MSPE) over 100 random splits for different methods applied on the 95 amines dataset. See supplementary table 3 for individual split results.

Method	**Descriptor set used**		
	Combined	**Basak lab**	**Diudea lab**
PCR	**29.11 (13.79)**	57.08 (93.829)	76.02 (24.72)
PLS	**18.86 (6.03)**	19.86 (7.464)	75.70 (24.689)
Lasso	**26.85 (9.049)**	28.72 (8.825)	72.75 (17.998)
SCAD	**25.81 (8.962)**	31.77 (21.442)	74.94 (18.322)
RF	**17.25 (6.498)**	18.98 (6.587)	84.59 (21.735)
GBM	**14.79 (5.836)**	18.03 (6.296)	74.78 (17.426)

Table 5: Top PCs and their loadings (in brackets) of the 95 amines data. Brackets in headings indicate percentage of variance explained by each PC.

PC1 (12.5%)	**PC2 (10.3%)**	**PC3 (7.1%)**	**PC4 (6.9%)**
E_ele (0.69)	E_ele (-0.42)	E_ele (-0.35)	E_vdw (-0.73)
vsurf_EWmin1 (-0.43)	vsurf_EWmin1 (-0.27)	vsurf_EWmin1 (-0.33)	E_nb (-0.48)
vsurf_EWmin2 (-0.4)	vsurf_EWmin2 (-0.25)	E_vdw (-0.31)	E_ele (0.34)
vsurf_EWmin3 (-0.3)	vsurf_DW13 (-0.21)	vsurf_EWmin2 (-0.31)	vsurf_EWmin1 (0.19)
vsurf_DW13 (-0.17)	vsurf_EWmin3 (-0.18)	E_nb (-0.27)	vsurf_EWmin2 (0.18)
E_nb (0.11)	E_nb (-0.17)	vsurf_EWmin3 (-0.23)	vsurf_DW13 (-0.14)
vsurf_HB6 (0.08)	E_vdw (-0.17)	vsurf_DW13 (0.21)	vsurf_EWmin3 (0.13)
vsurf_W6 (0.08)	DN2Z2 (0.16)	**DN2Z2 (-0.16)**	GCUT_SlogP_0 (0.06)
vsurf_HL2 (0.06)	SlogP_VSA9 (0.13)	GCUT_SMR_0 (–0.11)	GCUT_SMR_0 (0.06)
vsurf_CW6 (0.06)	**ASZ2 (0.12)**	**ASZ2 (-0.1)**	**DN2Z2 (0.04)**

Bold = Triplet descriptors calculated by Basak group

Table 6: Top PCs and their loadings (in brackets) of the 508 compounds data. Brackets in headings indicate percentage of variance explained by each PC.

PC1 (15%)	**PC2 (14.1%)**	**PC3 (12%)**	**PC4 (9.3%)**
E_tor (0.96)	vsurf_DW23 (0.95)	density (-0.93)	GCUT_SlogP_0 (0.96)
vsurf_DW23 (-0.25)	E_tor (0.23)	GCUT_SlogP_0 (-0.19)	density (-0.19)
GCUT_SlogP_0 (0.06)	GCUT_SlogP_0 (0.15)	vsurf_ID2 (0.12)	vsurf_DW23 (-0.14)
vsurf_ID3 (-0.04)	density (-0.07)	vsurf_ID3 (0.12)	E_tor (-0.09)
vsurf_ID4 (-0.04)	vsurf_ID2 (-0.06)	vsurf_ID4 (0.11)	vsurf_CP (0.05)
vsurf_ID2 (-0.04)	vsurf_ID1 (-0.05)	vsurf_ID1 (0.1)	vsurf_ID2 (0.04)
vsurf_ID6 (-0.03)	vsurf_ID3 (-0.05)	vsurf_IW1 (0.1)	vsurf_ID4 (0.02)
vsurf_ID1 (-0.03)	vsurf_ID4 (-0.04)	vsurf_ID5 (0.09)	vsurf_ID3 (0.02)
vsurf_ID5 (-0.03)	vsurf_CW4 (-0.04)	BCUT_SMR_2 (-0.09)	vsurf_ID1 (0.02)
vsurf_ID7 (-0.03)	vsurf_ID6 (-0.03)	vsurf_ID6 (0.07)	vsurf_ID7 (0.02)