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

Subhabrata Majumdar, Subhash C. Basak, Gregory D. Grunwald, Mircea V. Diudea and Claudiu Lungu

Abstract: TBD later

Keywords: TBD later

**1. INTRODUCTION**

Hazard assessment of chemicals is often carried out in data poor situations (National Research Council, 1984). The Toxic Substances Control Act (TSCA) Inventory, maintained by the United States Environmental Protection Agency (USEPA), currently has about 85,000 entries (Toxic Substances Control Act (TSCA) Inventory). A large fraction of these chemicals has very little or no data needed for their hazard estimation (Auer, Nabholz, & Baetcke, 1990). 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 (DiMasi, Grabowski, & Hansen, 2016).

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 (Benigni, 2003). 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, the development, computation and usage of chemical descriptors has 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 (Basak, Harriss, & Magnuson, POLLY v2.3, 1988; Stewart, 1990; Sybyl Version 6.2, 1995; Basak, Grunwald, & Balaban, TRIPLET, 1993) consist of those developed and used by Basak and coworkers over the past decades towards effective QSAR model formulation (Basak, Gute, & Grunwald, A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters, 1999; Basak, Mills, Gute, & Hawkins, 2007; Basak & 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: Volume 1, 2016; Majumdar & Basak, Beware of external validation! – A Comparative Study of Several Validation Techniques used in QSAR Modelling, 2018; Basak, Magnuson, Niemi, Regal, & Veith, 1987). The second descriptor set has been developed by Diudea and coworkers using Schrodinger and in-house software TopoCluj [16,17,18]. 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 use cases that practitioners are likely to enter 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 (Debnath, Debnath, Shusterman, & Hansch, 1992). 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* (Debnath, Debnath, Shusterman, & Hansch, 1992)*.*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 (Soderman, 1982), the response variable in this dataset is the 0/1 Ames mutagenicity status of the chemical compounds. 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* (Basak, Mills, Gute, & Hawkins, 2007; Basak, Gute, & Grunwald, A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters, 1999; Majumdar & Basak, Beware of external validation! – A Comparative Study of Several Validation Techniques used in QSAR Modelling, 2018) and Diudea *et al* [21].

*Table 1: Chemical classes of samples in the 508 compounds 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 |

**2.2. Descriptors**

For this study we have used two collections of molecular descriptors. One set of descriptors, used frequently by the Cluj team of Diudea and collaborators, were calculated by the programs Schrodinger and TopoCluj. More detailed references about these descriptors are given in Supplementary Tables 1 and 4. For the 95 and 508 data descriptors were calculated by Diudea lab.

The second set of molecular descriptors, used frequently by Basak *et al*, were calculated by the software POLLY (Basak, Harriss, & Magnuson, POLLY v2.3, 1988), MolConnZ (MolconnZ v4.05, 2003), Triplet (Basak, Grunwald, & Balaban, TRIPLET, 1993), and MOPAC (Stewart, 1990). For the 95 and 508 chemical sets, 275 and 307 descriptors were calculated for this paper by the above software.

**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 (Basak S. C., Mathematical Structural Descriptors of Molecules and Biomolecules: Background and Applications, 2015). 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 (Basak, Magnusson, Niemi, & Regal, 1988; Lauria, Ippolito, & Almerico, 2009; Majumdar & Basak, Exploring intrinsic dimensionality of chemical spaces for robust QSAR model development: A comparison of several statistical approaches, 2016). In this paper, we build predictive models using the following two dimensions reduction methods:

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

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 Majumdar and Basak (Majumdar & Basak, Exploring intrinsic dimensionality of chemical spaces for robust QSAR model development: A comparison of several statistical approaches, 2016) and apply a robust PCA procedure (Majumdar S. , Robust estimation of principal components from depth-based multivariate rank covariance matrix, 2015) to obtain the PC loadings. Following this, we use the transformed data matrix 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 [ref] 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 Fan and Li (Fan & Li, 2001), 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 them is often difficult [refs]. 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 (Svetnik, Liaw, Tong, & others, 2003; Polishchuk, Muratov, Artemenko, & others, 2009; Kuz'min, Polishchuk, Artemenko, & Andronati, 2011).

*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 (Svetnik, Wang, Tong, & others, 2005; Sheridan, Wang, Liaw, & others, 2016).

**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 (Xu & Liang, 2001),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 (Xu & Liang, 2001; Zhang & Yang, 2015). 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 analyzed (Hawkins, Basak, & Mills, 2004; Basak & 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: Volume 1, 2016; Majumdar, Basak, & Grunwald, 2013; Basak & Majumdar, Prediction of Mutagenicity of Chemicals from Their Calculated Molecular Descriptors: A Case Study with Structurally Homogeneous versus Diverse Datasets, 2015).

**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 (R Core Team, 2015).

**3.1. Output of predictive models**

*Table 2: Average and standard deviations (in brackets) of Area Under Curve (AUC) for different methods applied on the 508 compounds heterogeneous dataset*

|  |  |  |  |
| --- | --- | --- | --- |
| **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) |

It is evident from the results in Table 2 above 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).

*Table 3: Median and mean absolute deviations (in brackets) of Mean Square Prediction Error (MSPE) for different methods applied on the 95 amines dataset*

|  |  |  |  |
| --- | --- | --- | --- |
| **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) |

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 the Basak group indices only came out as close second ones.

For all descriptor sets, PLS has the best performance among all methods, while Boosting performs the best for the 95 amines data. The good performance of PLS indicates that there are low-dimensional subspaces in the predictor spaces that are predictive of the responses.

Methods that depend directly on sparse linear combinations of predictors: Lasso, SCAD do not perform well in either case. 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.

PLS performs well in both cases.

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 PCA procedure are presented in Tables 4 and 5 and the below discussion.

*Table 4: Top PC loadings of the 95 amines data*

|  |  |  |  |
| --- | --- | --- | --- |
| **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

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. The four PCs represent important features of chemical graph. Since PCs are orthogonal to each other, higher order PCs do not contain characteristics other than those encoded in the first four PCs. Most of the toxic properties of these 95 compounds are related to their solubility surface and membrane permeability capacity. High loading observed in case of PC1 energetic variable (E\_ele - 0.69) advises that toxicity in case of 95 amines it is correlated with an energetic dependent process.

*Table 5: Top PC loadings of the 508 compounds data*

|  |  |  |  |
| --- | --- | --- | --- |
| **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) |

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 compounds are high with respect to intrinsic physico-chemical properties (related variables such as E\_tor (0.96), vsurf\_DW23 (0.95), density (-0.93), GCUT\_SlogP\_0 (0.96), i.e- ADME features.

**4. Discussion**

(To be completed later after results are written)

**5. CONFLICT OF INTEREST**

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

**6. ACKNOWLEDGEMENTS**

**7. SUPPLEMENTARY MATERIAL**

Supplementary material (Supplementary tables 1-4) is available on the publisher’s web site along with the published article.

# **References**

Auer, C. M., Nabholz, J. V., & Baetcke, K. P. (1990). 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*, 183-197.

Basak, S. C. (2015). Mathematical Structural Descriptors of Molecules and Biomolecules: Background and Applications. In S. C. Basak, G. Restrepo, & J. L. Villaveces (Eds.), *Advances in Mathematical Chemistry and Applications, volume 1* (pp. 3-23). Bentham eBooks, Bentham Science Publishers and Elsevier.

Basak, S. C., & Majumdar, S. (2016). 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: Volume 1. In *Advances in Mathematical Chemistry and Applications* (pp. 251-281). Bentham e-Books.

Basak, S. C., & Majumdar, S. (2015). 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*, 117-123.

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

Basak, S. C., Harriss, D. K., & Magnuson, V. R. (1988). *POLLY v2.3.* Copyright of the University of Minnesota.

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

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

Basak, S., Grunwald, G., & Balaban, A. (1993). *TRIPLET.* Copyright of the Regents of the University of Minnesota.

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

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

Debnath, A., Debnath, G., Shusterman, A., & Hansch, C. (1992). 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*, 37-52.

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

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

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

Small-Molecule Drug Discovery Suite 2009, Schrödinger, LLC, New York, NY, 2009

O. Ursu and M. V.Diudea. TOPOCLUJ software program. Babes –Bolay University, Cluj, 2005

Molecular Operating Environment (MOE),; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2004.

Lungu, C.N.; Diudea, M.V.; Putz, M.V.; Grudziński, I.P. Linear and Branched PEIs (Polyethylenimines) and Their Property Space. Int. J. Mol. Sci. **2016**, 17, 555.

Claudiu N. Lungu, C-C Chemokine receptor type 3 inhibitors: Bioactivity prediction using local vertex invariants based on thermal conductivity layer matrix , Studia UBB Chemia, LXIII,1, 2018 177-188

Claudiu Lungu, Sara Ersali, Beata Szefler, Atena Pirvan-Moldovan, Subhash Basak, Mircea V Diudea, Dimensionality of big data set explored by cluj descriptors, Studia UBB Chemia , LXII, 3, 2017, 197-204

Kuz'min, V. E., Polishchuk, P. G., Artemenko, A. G., & Andronati, S. A. (2011). Interpretation of QSAR Models Based on Random Forest Methods. *Mol. Inform.* *, 30* (6-7), 593-603.

Lauria, A., Ippolito, M., & Almerico, A. M. (2009). 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), 387-395.

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

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

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

Majumdar, S., Basak, S. C., & Grunwald, G. D. (2013). 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*, 463-471.

*MolconnZ v4.05.* (2003). Quincy, MA: Hall Ass. Consult.

National Research Council. (1984). *Toxicity Testing Strategies to Determine Needs and Priorities.* Washington, DC: National Academy Press.

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

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

R Core Team. (2015). R: A Language and Environment for Statistical Computing version 3.3.2.

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

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

Stewart, J. (1990). *MOPAC Version 6.00, QCPE #455.* Frank J. Seiler Research Laboratory: US Air Force Academy, CO.

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

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

*Sybyl Version 6.2.* (1995). St. Louis, MO: Tripos Associates, Inc.

*Toxic Substances Control Act (TSCA) Inventory.* (n.d.). Retrieved 4 11, 2018, from https://19january2017snapshot.epa.gov/tsca-inventory/about-tsca-chemical-substance-inventory\_.html

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

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *J. Econometrics* *, 187*, 95-112.