In Silico Prediction of the Biodegradability of Chlorinated Compounds: Application of Quantitative Structure-Biodegradability Relationship Approach

Meade Erickson[1], Denys Vasyutyn[1,2], Marvellous Ngongang[1], Amirreza Daghighi [1,3], Stephen Szwiec[1,4], Gerardo Casañola-Martin[1], Bakhtiyor Rasulev[1,3,4]

[1] *Department of Coatings and Polymeric Materials, North Dakota State University, Fargo, ND 58105, USA*

[2] *Department of Chemical and Biomolecular Engineering, New York University Tandon School of Engineering, Brooklyn, NY, 11201, USA*

[3] *Biomedical Engineering Program, North Dakota State University, Fargo, ND 58105, USA*

[4] *Materials & Nanotechnology Program, North Dakota State University, Fargo, North Dakota 58108, USA*

KEYWORDS: Machine learning, QSAR, QSBR, biodegradability, chlorinated compounds, polymers

ABSTRACT: Chlorinated compounds are generally known to be non-readily biodegradable. The insight into the structural features that allow chlorinated compounds to readily biodegrade is crucial information that needs to be unveiled. Combined in silico modeling and machine learning approach to predict desirable compound properties has proven to be an effective tool, enabling chemists to save time and resources compared to web lab experimentation. Here we present two machine learning-based quantitative structure – biodegradability relationship (QSBR) models, one for predicting biodegradability values of chlorinated compounds, and the other one for classifying chlorinated compounds as biodegradable or non-biodegradable. The regression models were generated using the Support Vector Regression (SVR) machine learning method. The optimal regression model was a 10 descriptor SVR model with R^2 = 0.925 and R^2_{test} = 0.881. The optimal classification model was a logistic regression classifier model with 5 descriptors. It has a Matthew's Correlation Coefficient of 0.59 for training and 0.55 for test, as well as accuracy of 0.79 for training set, 0.77 for test set. For validation purposes the models were tested on an external data set of chlorinated compounds. In addition, models were further applied to an external test set of monomeric units representing polymers to assess the capability of the model to estimate the biodegradability of polymers, where the models showed statistical robustness. The developed SVR model could be used for accurate prediction of biodegradability of various organic molecules, as well as materials based on organic compounds. The analysis of the influence of descriptors on biodegradability is performed. The classification model showed that the biodegradability of chlorinated compounds is heavily correlated with descriptors that relate to electrotopological descriptors, position of oxygen relative to chlorine.

Introduction

Polymer plastic materials have profoundly shaped the course of the last century, revolutionizing industries and daily life alike. Their lightweight, versatile, and cost-effective nature paved the way for innovations in packaging, manufacturing, and medical devices, fostering economic growth and enhancing human well-being. Plastics assumed a significant role in the global economy with their increased usage. However, as the utilization of plastic escalated, apprehensions regarding the environmental repercussions of post-use plastics arose. Rapid plastic accumulation has shown devastating effects in marine and land environments.[1,2,3]. The environmental impact of traditional petrochemical-based polymers has caused concern over the depletion of nonrenewable resources and waste management.[4,5,6]. In addition to concerns for the carbon footprint and bulk management of these materials, microplastic pollution is a major challenge that could have lasting environmental and human health impacts. [7,8,9,10] Polymer plastics can pose a severe threat to the environment, biodiversity, and human health due to possible buildup where natural functions are inhibited, especially as they degrade into microplastics.[1,3]

Questions regarding approaches of plastic management led the scientific community to research techniques to better re-utilize or to safely degrade plastics.[11] Such techniques include mechanical degradation, thermo-chemical degradation, photodegradation, and biodegradation.[11,12] The complexity of logistics and the expenses associated with mechanical, thermo-chemical, and photodegradation reprocessing have made biodegradation the preferred solution for addressing plastic waste challenges on a large scale.[13,14,15] Biodegradability is the ability of a material to break down into simpler compounds by the action of microorganisms, and the term is generally used as a measure of the rate at which a material will break down. Ultimately, any organic compound is biodegradable, going through processes of deterioration, fragmentation, assimilation, and

mineralization; however, this process may take hundreds of years for certain materials.[16]

The investigation and data analysis of biodegradability of materials is delineated by the various engineering standards applied for determining biodegradability, which include ISO, CEN, and ASTM standards.[17], [18],[19],[20],[21],[22],[23],[24]. Moreover, these standards do not take into account the environmental impact of the material at end-of-life. While some overlap exists between standards, specific implementation of variables such as time frames and percentages degraded differ. Furthermore, no single standard can simulate the process of biodegradation within a natural environment fully. Despite these challenges, investigation into the engineering of enzymes, bacteria, fungi, or algae capable of plastic degradation has yielded promising outcomes for certain extensively-utilized plastic types in recent years. However, owing to the extensive diversity of plastics, the hereto developed microorganism solutions exhibit the capability to degrade only specific plastic varieties.[15]

Clearly, given the engineering and environmental challenges polymer plastics, the development of new polymers must focus on the creation of safely biodegradable or completely compostable materials. Emerging materials, such as Polycaprolactone-collagen hydrolysate blends, have been engineered to be compostable by design.[25] To increase the pace of materials development, it is essential to create a framework of principles for polymer biodegradation to aid in the research of plastic-degrading organisms and to develop more biodegradable materials. Quantitative Structure-Property Relationships (QSPR) is a pivotal concept and a technique in materials design which encompasses computational models that correlate the structural characteristics of materials with their specific properties, enabling informed predictions and tailoring of materials for desired functionalities.[26] Furthermore, an extension of this concept, Quantitative Structure-Biodegradability Relationships (QSBR), applies the QSPR framework specifically to the prediction of material biodegradation behaviors. These QSBR models establish connections between molecular structures and biodegradability, offering valuable insights for the design of environmentally sustainable materials with enhanced degradation properties.[27],[28]

Multiple QSBR models have been developed for biodegradation, providing insight into the impact of chemical structure on the complex biodegradation process.[29],[30],[31] Certain models exhibit strong predictive capabilities for biodegradation yet lack conceptual clarity, whereas others demonstrate a partial accuracy in predicting biodegradability while emphasizing structure-property relationships. The central hurdle in biodegradability prediction centers around the intricate task of formulating precise models for polymers. Constructing such models for discrete chemicals is feasible; however, the complexity arises when dealing with polymers due to the absence of uniform biodegradation testing and data reporting standards, the lack of standardized characterization methods, and the extensive diversity and intricacy inherent to polymers. It is important to note that presence of above-mentioned structures does not guarantee the biodegradability or non-biodegradability of compounds but is highly correlated and can be used as a tool for predictability with either class.

Chlorinated polymers represent a distinctive class of materials characterized by the incorporation of chlorine into their polymer chains, introducing unique properties to the resulting polymers, such as improved flame resistance, thermal stability, chemical resistance, and electrical insulation. As a subset of halogenated polymers, chlorinated polymers find applications in various industries, including construction, electronics, automotive, and aerospace, where their exceptional fire-retardant properties are particularly advantageous. However, concerns have been raised about the environmental impact of chlorinated polymers due to the potential release of toxic halogenated compounds during their production, use, and disposal. As presented by Vorberg et. al., the tendency of halogens in organic compounds making them extremely non-biodegradable, especially the tendency of chlorinated compounds which are estimated to be 15 times more likely to be non-biodegradable than biodegradable seems to be of great interest.[29] Additionally, most chloride substituents are acutely toxic, and could provide a long-term toxicity effect on the environment.[32]

There is an ongoing effort to develop more sustainable alternatives to traditional chlorinated polymers while retaining their desirable attributes. Despite the overall trend of toxicity among the by-products of chlorinated compound biodegradation, exceptions exist, and these exceptions warrant a closer examination to discern underlying factors governing the biodegradation process. In this context, QSBR emerges as a valuable tool, offering the potential to decipher the intricate relationships between molecular structures and biodegradability outcomes. This research endeavors to construct focused models specifically tailored to chlorinated polymer compounds. The objective is to establish a heightened level of precision in predicting the biodegradability of chlorinated materials. By concentrating on this narrower range of chemical compounds, the study seeks to enhance the reliability and applicability of the QSBR model, thereby offering a targeted tool for advancing the understanding of chlorinated compound degradation and guiding the design of environmentally sustainable materials within this distinct chemical context.

2. Materials and Methods

The data for the experiment consisted of two data sets of chlorinated compounds, continuous (I) and binary (II).[33],[31]

Data Set I Collection and Modeling

Biodegradability information for chlorinated compounds in the continuous data set (I) was obtained from Toropov et al.[33] Where dataset unit is originally measured as oxygen consumption following Organization for Economic Co-Operation and Development (OECD) 301C – MITI Biodegradation Test.[34],[35] The data set contained 74 compounds, with a continuous value for biodegradability, ranging from 0 to 1. For validation purposes, the data set was split into training and testing sets, where the training set was used to train models and the test set was used for model validation.

To generate a set of descriptors, 2D molecular structures in MOL format were generated using ChemSketch.[36] The 2D structures were used to obtain SMILES notation by applying OpenBabel software.[37] To obtain a set of molecular

descriptors, the Online Chemical Modeling Environment (OCHEM.EU) was used.[38] The molecules were standardized using the CDK standardizer. After standardization, molecules were neutralized, salts were removed, and structures were cleaned. The OEState Bonds 2D Indices descriptors, ALogPS 2D descriptors, and all subcategories of alvaDesc v.2.0.14 3D descriptors excluding Drug-like indices were calculated. The compounds were optimized with OpenBabel using the Merck Molecular Force Field.[39–41] In total 5856 descriptors were calculated. Highly correlated descriptors (R>0.9), constant and near constant descriptors (std<0.1) were removed from the data set. Descriptors having less than 10 compounds containing non-zero entries were removed due to inability for descriptor to effectively describe the response value. After eliminating highly correlated descriptors, constant, near constant descriptors, and descriptors with low presence, 925 descriptors were present in the data set. All data curation steps were performed using Python (version 3.10.9).[42,43]

The data set was normalized by the standard scale normalization using Scikit-learn package, which normalizes the data through their mean and standard of deviation using the **Equation (1).**[44]

$$x_{ij} = \frac{X_{ij} - \overline{X}_j}{\sqrt{\frac{\sum_1^n \left(X_{ij} - \overline{X}_j\right)^2}{n-1}}} \quad (1)$$

Where X_j is the mean values of the j^{th} descriptor, n is the number of compounds, x_{ij} and X_{ij} are the normalized and original values of the j^{th} descriptor of the i^{th} compound.

The developed QSBR model was subjected to statistical analysis evaluating the squared correlation coefficient (R^2) external validation metrics ($Q^2_{F1}, Q^2_{F2}, Q^2_{F3}$), Mean Absolute Error (MAE), and Concordance Correlation Coefficient (CCC). The following equations were used to calculate the squared correlation coefficient R^2 (**Equation (2)**), the Mean Absolute Error (**Equation (3)**), Concordance Correlation Coefficient CCC (**Equation (4)**)

$$R^2 = 1 - \frac{\sum_{i=1}^n \left(y_i^{obs} - y_i^{pred}\right)^2}{\sum_{i=1}^n \left(y_i^{obs} - \overline{y}^{obs}\right)^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left|y_i^{obs} - y_i^{pred}\right| \quad (3)$$

$$CCC = \frac{2\sum_{i=1}^n \left(y_i^{obs} - \overline{y}^{obs}\right)\left(y_i^{pred} - \overline{y}^{pred}\right)}{\sum_{i-1}^n \left(y_i^{obs} - \overline{y}^{obs}\right)^2 + \sum_{i=1}^n \left(y_i^{pred} - \overline{y}^{pred}\right)^2 + n\left(\overline{y}^{ol}\right.} \quad (4)$$

Where y_i^{obs} and y_i^{pred} are observed and predicted values for i^{th} compound, respectively, n is the number of compounds, and $\overline{y}^{obs}$ and $\overline{y}^{pred}$ are the mean values for observed and predicted values, respectively. Then criteria chosen for assessing QSBR model were chosen in accordance with OECD principal No.4 for developing QSBR models with "appropriate measures of goodness of fit, robustness, and predictivity).

The best model was selected based on the best values for the above criteria for both training and testing sets to avoid overfitting. The R^2 and R^2 test of the best models from each number of variables were comparted to indicate possible over fitting at specific variable number. This allows for ease of understanding the statistically best number of variables for indicating performance of models in both training and test analysis.

To prepare the data for construction and evaluation of QSBR models, the data was sorted based on the biodegradability values from smallest to largest and split into training and test sets in 4:1 ration, with every 5th compound going into the test set. The model was developed based on the training set through Genetic Algorithm (GA) for variable selection. The variable selection process started with a population of 1000 random models and 10000 iterations for evolution, with the mutation probability being 50%.

To further refine the descriptor selection of GA, an iterative descriptor selection process was implemented. Models, based on R^2, were compared after 10000 iterations were conducted to determine the best models. The descriptors of the best models were inputted into a new dataset that included only high performing descriptors. This method was conducted 4 times, labeled as repetitions, to give a dataset containing 94 descriptors. Where after the 4th repetition, models' statistical results were not improved. Therefore the 94 descriptor dataset was used to perform final model development. The process of descriptor elimination allowed the final data set to be a set of better-performing descriptors. The overall model development process is shown in **Figure 1**.

The SVR parameters for the estimator were set to 'rbf' for kernel, C equal to 10, and 'auto' for gamma, which are default Scikit-learn parameters, with the rest of parameters being set to a default state. More information regarding the chosen parameters can be found in Scikit-learn library documentation.[44]

Due to the difficulty in interpretation of descriptors from the SVR continuous data model caused by its nonlinearity, a differing approach is needed to analyze descriptors. The descriptors were analyzed using accumulated local effect (ALE) plots to determine their influence on the original dataset.[45] Briefly, interpretation of ALE plots are conditional on a given value, the relative effect of changing the feature on the prediction is read from the ALE plot. Due to ALE plots being centered at 0, the interpretation is comparable between other ALE plots and each point is the difference to the mean prediction. Finally, the ALE plots show the interaction of the response value and the specific descriptor. This can help describe descriptor influence on the whole data set.

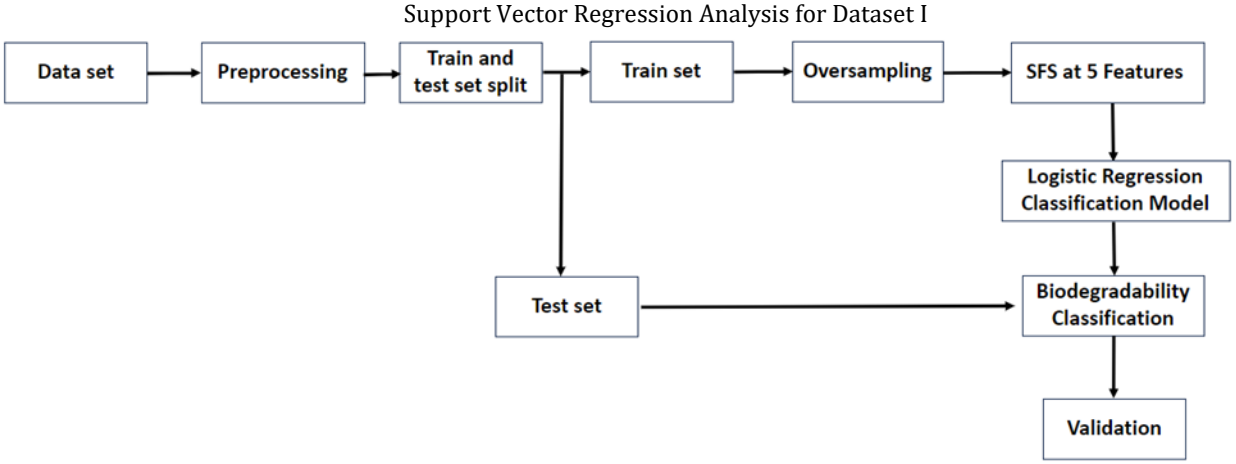

Figure 1. QSBR SVR Model Development Methodology

Data Set II Collection and Modeling

Biodegradability information for chlorinated compounds in the binary data set (II) was obtained from the work of Mansouri et al.[31] The data set contained 236 chlorinated compounds, with values being either 0 or 1 and representing the non-biodegradable or readily biodegradable compounds respectively. Similarly, to the continuous dataset, the oxygen consumption was measured following OECD-301C protocol.[35] In Mansouri et al.'s work, compounds with biochemical oxygen demand (BOD) at or above 60% are classified as readily biodegradable. Any compounds below 60% are classified as non-readily biodegradable.[46] For validation purposes, the data set was split into training and testing sets, where the training set was used to train models and the test set was used for model validation.

To generate a set of descriptors, 2D molecular structures in MOL format were generated using ChemSketch.[36] The 2D structures were used to obtain SMILES notation by applying OpenBabel software.[37] To obtain a set of molecular descriptors, the OCHEM.eu was used.[38] The molecules were standardized using the CDK standardizer. After standardization, molecules were neutralized, salts were removed, and structures were cleaned. The OEState Bonds 2D Indices descriptors, ALogPS 2D descriptors, and all subcategories of alvaDesc v.2.0.14 3D descriptors excluding Drug-like indices were calculated. The compounds were optimized with OpenBabel. In total 5856 descriptors were calculated. Highly correlated descriptors (R>0.9), constant and near constant descriptors (std<0.1) were removed from the data set. Descriptors with less than 10 compounds of them having non-zero entries were removed as well. After eliminating highly correlated descriptors, constant and near constant descriptors, and descriptors with low presence, 992 descriptors were present in the data set. All data curation steps were performed using Python (version 3.10.9).[42] The large difference in range among descriptors makes descriptors with smaller range outweighed by descriptors with larger range. [43]

The data set was normalized by the standard scale normalization using Scikit-learn package.[44] which normalizes the data through their mean and standard of deviation using the (**Equation (1)**).

The developed QSBR model was subjected to statistical analysis evaluating confusion matrix by means of calculating Precision, Recall, F-Score (F1), Accuracy, Specificity, and Mathhew's Correlation Coefficient (MCC). The following equations were used to calculate Precision (**Equation (5)**), Recall (**Equation (6)**), F1 (**Equation (7)**), Accuracy (**Equation (8)**), Specificity (**Equation (9)**), and MCC (**Equation (10)**)

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = 2 \cdot \frac{(Precision \cdot Recall)}{(Precision \cdot Recall)} \quad (7)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Specificity = \frac{TN}{TN + FP} \quad (9)$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN}} \quad (10)$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative predictions, respectively. The MCC was chosen as the main indicator of model robustness since in most instances it is a more reliable indicator of model's performance than accuracy and other criteria. [47]

The best model was selected based on the best values for the above criteria for both training and testing sets to avoid overfitting, with MCC being the main selection factor.

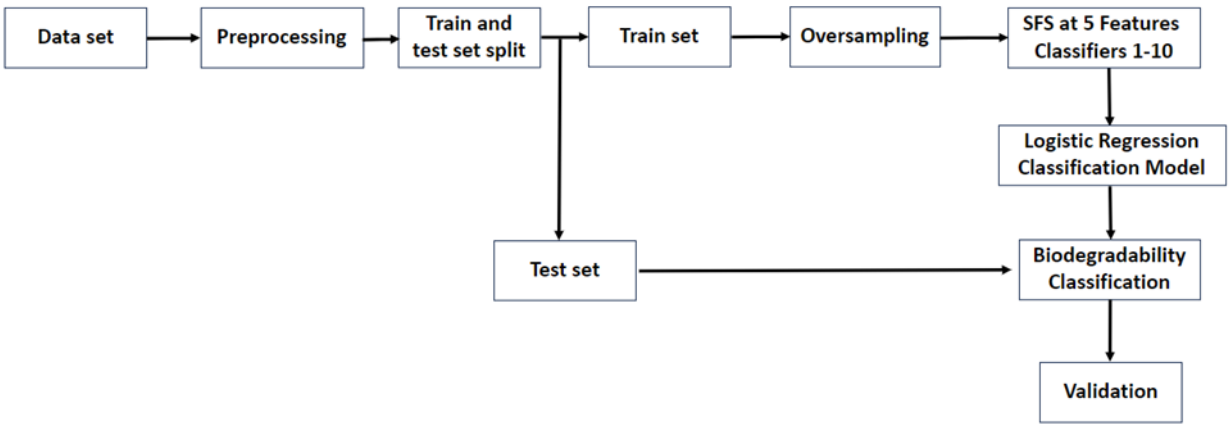

Figure 2. QSBR Logistic Regression Classification Model Development Methodology

To prepare the data for construction and evaluation of QSBR models, the data was oversampled to prevent model from being biased towards the majority class (non-biodegradable), sorted based on the biodegradability values from smallest to largest and split into training and test sets in 4:1 ration, with every 5th compound going into the test set. The oversampling was performed using the Scikit-learn python package.[44] Utilizing 'minority' strategy for random state in a range from 0 to 99, with 10 types of classifiers performing 5 feature Sequential Feature Selection (SFS) for each random state of oversampling. The MCC was set as the scoring metrics for the selection. Ten classifiers were applied to the dataset to determine best classifier for dataset. Rationality of choosing best classifier was conducted based on a combined highest average performing model, out of 100, with ease of descriptor explanation. The ten classifiers Logisic Regression, K-Neighbors, SVC, SVM, Naïve Bayes, Decision Tree, Random Forest, Ada Boost, Quadratic, and MLP. The optimal results were found with Logistic Regression classifier with assuming 5 features for all classifiers. Data of this methodology is shown in supporting information.

All parameters for oversampling, SFS, and classifiers were set to default which could be found in the Scikit-learn library documentation.[44] The feature selection process for the optimal number of features at the set oversampling state utilized the SFS method with Logistic Regression being set as the classifier and MCC being set as the scoring metrics. All other parameters were set to default and could be found in the Scikit-learn library documentation. [44] The process development method is shown in **Figure 2.**

3. Results and Discussion

Continuous Data – SVR Model

After initial data processing the continuous data set contained 925 descriptors. Those descriptors were used to develop multiple QSBR models. The QSBR model with the best statistical characteristics was selected for further analysis. The selected model's statistical performance is listed in **Table 1**, (Model 3). In Table 1 the quality of models was assessed based on the correlation coefficient for training set, leave-one-out coefficient, and external validation coefficient, as well as mean absolute error MAE. The values for robustness of the model for the training set were higher than for the test set, and the values of the errors for the training set were lower than for the test set, indicating consistency.

Table 1. Statistical Parameters of Selected Best SVR Model

Parameters	Model 3
No. of Descriptors	10
R^2 (Training)	0.925
R^2 (Test)	0.881
Q^2F1	0.881
Q^2F2	0.881
Q^2F3	0.894
MAE (Training)	0.060
MAE (Test)	0.067
CCC	0.948
k	0.863
k′	1.070

For visualization of performance, **Figure 3** shows the correlation plot if training and test sets. Where the experimental data and predicted values are compared.

It can be seen in **Table 1.** that the performance of the model is quite accurate and robust of chlorine containing compounds. $R^2_{training}$ is above 0.90 showing high levels of predictability and R^2_{test} is 0.88 shows its predictive capabilities. The relatively low values of MAE (~0.06) show the model's performance error is low.

It is necessary to mention how this data was heavily influenced by a large amount of data with a response value at or near 0. This emphasizes the performance of the model with having a skewed dataset.

The predicted values and the experimental values, shown in **Figure 3**, for biodegradability are relatively close to the diagonal line, which confirms the high performance of the

model As mentioned previously the dataset contains many values near 0 where it skews the dataset which is known to poorly influence the model performance. The Figure shows a large amount of these near 0 values, to be predicted relatively well.

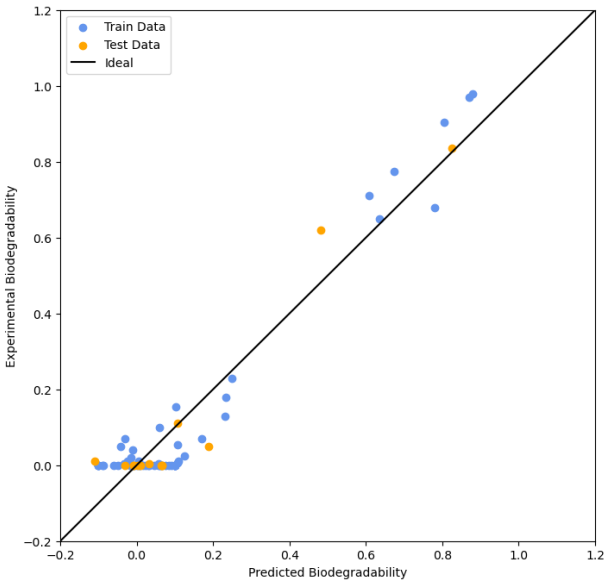

Figure 3. Experimental vs Predicted Biodegradability Values Obtained by the SVR Model

The Williams plot of the model for continuous data is shown in **Figure 4.** The applicability domain ranges from Std. Residual of –3 to +3 along y axis and 0.0 to h(i/i) of the leverage data along the x-axis.

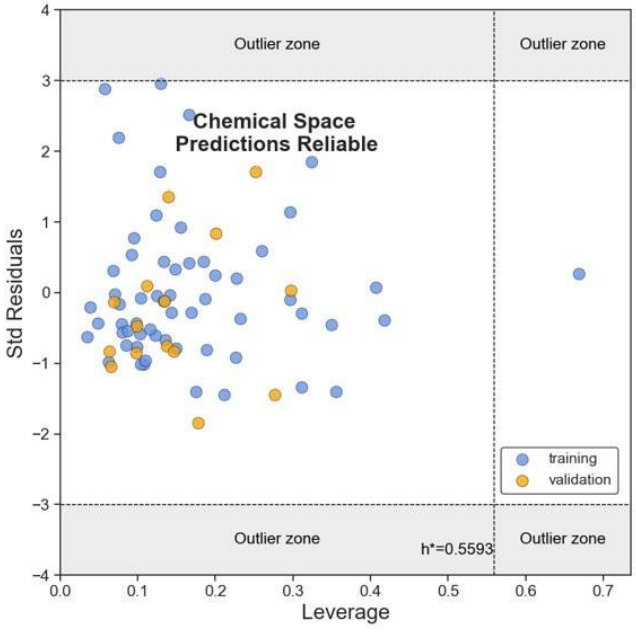

Figure 4. William's Plot – SVR model of Continuous Data

Regarding **Figure 4**, the Williams Plot, all the points of the model fall within three standardized residuals, indicating proximity of predicted and experimental values. The leverages of all molecules are within the threshold, except for one compound in the training set, indicating it being a structural outlier but still able to be predicted. With this information it is shown that most of our dataset falls within the applicability domain.

With model development we can assume some level of influential information being derived from the selected descriptors. Due to the nature of SVR model development we can assume a non-linear correlation between descriptors and biodegradability. With this in mind, the descriptors chosen by GA are shown in **Table 2**, and then analysis of normalized descriptors' values by ALE plots are shown in **Figures 5** and **6.** The process we employ to understand the descriptors and their influence is a knowledge-based analysis.

Due to the complexity of the descriptors and their influences, we take a two-step approach to help discuss their influence. The first step is by grouping them in similar structural informational groups based on their structural influence. The second step is by analyzing them individually by discussing the trends found in the ALE plots.

Groupings are meant to simplify the analysis process by conceptualizing their general information. In the first group we categorize it according to its overall mass and shape of the compounds. The second grouping of descriptors is more focused on atomic volume, distance, and presence of functional groups within the structures. The third group emphasizes the electrotopological information of the structures where electron availability combined with general topology has influence on biodegradation.

Based on the training set ALE plots above, it is shown that all descriptors in the model have a complicated relationship with the response value of biodegradability. Training set ALE plots are analyzed since that was the dataset that trained the model. It was necessary to compare training vs test set ALE plots to highlight their similarity emphasizing the test set representing, with some differences, the training set with little significant outliers.

The interpretation of ALE plots could be described as follows, the x-axis signifies the descriptor value, and the slope of the tangent line at that x-value signifies the impact on biodegradability.[45] All x-values listed on the graph are normalized using standard scaler.

Table 2. SVR Continuous Data Model Descriptors, Their Definitions and Generalized Groupings

Descriptor	Definition	Descriptor Type
Mass and shape information		
P2m	2nd component shape directional WHIM index / weighted by mass	WHIM descriptors
GATS4m	Geary autocorrelation of lag 4 weighted by mass	2D autocorrelations
Atomic volume, distance, and presence of functional groups		
VE1_G/D	coefficient sum of the last eigenvector (absolute values) from distance/distance matrix	3D matrix-based descriptors
RDF030v	Radial Distribution Function – 030 / weighted by van der Waals volume	RDF descriptors
F05[C-Cl]	Frequency of C – Cl at topological distance 5	2D Atom Pairs
Eta_F_A	eta average functionality index	ETA indices
Electrotopological information		
GATS4p	Geary autocorrelation of lag 4 weighted by polarizability	2D autocorrelations
Eig05_AEA(dm)	eigenvalue n. 5 from augmented edge adjacency mat. weighted by dipole moment	Edge adjacency indices
VE1sign_Dz(p)	coefficient sum of the last eigenvector from Barysz matrix weighted by polarizability	2D matrix-based descriptors
SpMaxA_EA(dm)	normalized leading eigenvalue from edge adjacency mat. weighted by dipole moment	Edge adjacency indices

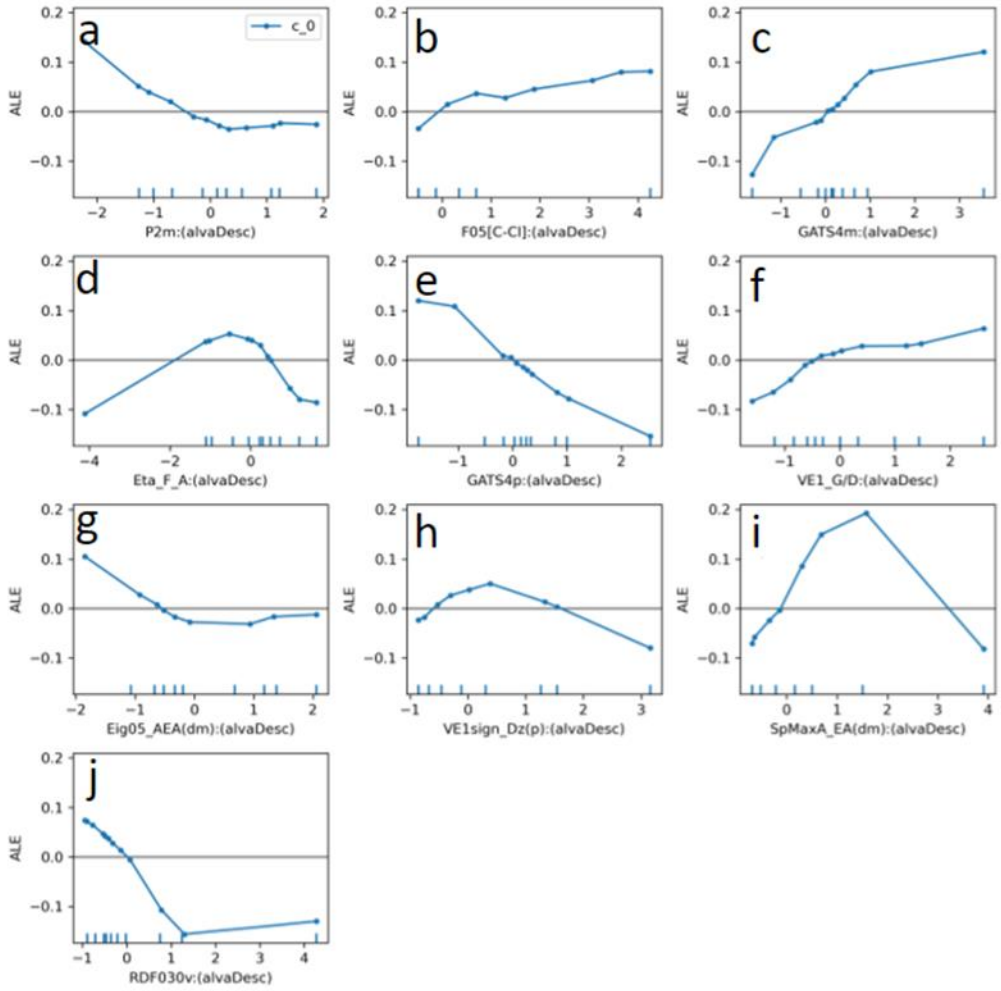

Figure 5. ALE Plots for SVR Model of Continuous Data, Training Set

Descriptors referring to mass and shape information

As seen in **Figure 5a** negative values of P2m negatively impact biodegradation and positive values of P2m barely impact biodegradation, signifying that high shape uniformness of a molecule with respect to mass doesn't impact biodegradability significantly, and that high heterogeneousness of molecule's shape with respect to mass impacts biodegradability negatively.

Shown in **Figure 5c**, the slope of GATS4m is positive for values lower than 1, signifying that high shape uniformness of a molecule with respect to mass affects biodegradability positively. There is only one data point in the ALE plot for GATS4m with value greater than 1, therefore it could not be concluded based of this descriptor whether heterogeneousness of shape with respect to mass would affect the biodegradability positively or negatively.

Descriptors referring to atomic volume, distance, and presence of functional groups

The slope of F05[C-Cl], shown in **Figure 5b**, is mostly positive, signifying that the more C-Cl bonds at topological distance of 5 the molecule has, the higher is the rate of biodegradability. This information is generally against what other works have suggested where Cl containing molecules have generally lower levels of biodegradability. This is a contentious finding but since our work focuses solely on this bond arrangement of Cl containing compounds, there can be further investigation into this influence.

The general trend of Eta_F_A, shown in **Figure 5d**, from –4 to ~ 2, indicates that for larger molecules the rate of biodegradability goes down.

The slope of VE1_G/D, shown in **Figure 5f**, is positive or somewhat positive for all values of the descriptors, being most positive between the values of –1 and 0, and most neutral between the values of 0 and 1. This signifies that VE1_G/D aids biodegradability the most for values between –1 and 0, and aids biodegradability the least between the values of 0 and 1. It also shows that there is a relationship between topological distances within the molecule.

The slope of RDF030v, shown in **Figure 5j**, is negative for values less than 1 and positive for values larger than one. Though there are 2 values larger than one, this trend is signifying that van der Waals volume is positively correlated with the rate of biodegradation, and low van der Waals volume affects biodegradability negatively.

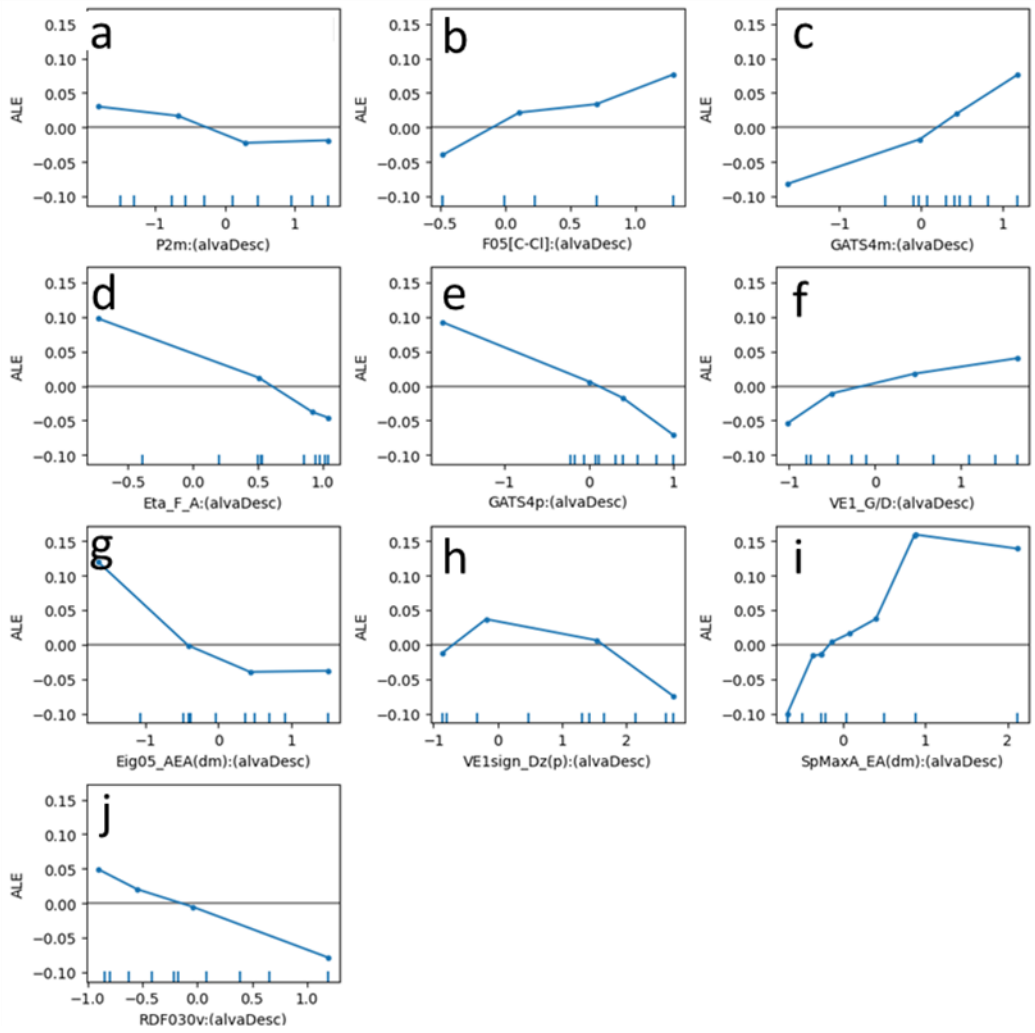

Figure 6. ALE Plots for SVR Model of Continuous Data, Test Set

This is an interesting finding where all compounds within our dataset contain H,C,Cl, and sometimes N or O. The highest two volume atoms being C and Cl, it is believed that comparing this descriptor and F05[C-Cl] suggest a synergistic effect between the two. At the very least we believe they are pointing to some unforeseen insight into the connection between C and Cl arrangement along molecules which positively influence the biodegradability

Figure 5e, shows the slope of GATS4p as negative for all values. Signifying that high uniformness of a molecule with respect to polarizability negatively affects biodegradability. There is only one data point in the ALE plot for GATS4p with value greater than 1, therefore it could not be concluded based of this descriptor whether heterogeneousness of polarizability would affect the biodegradability positively or negatively.

Autocorrelation descriptors compare neighboring sections of molecules to one another, that in combination of the uniformness found with polarizability. We believe that a repetitive molecule, such as a uniform carbon chain may be negatively influential to biodegradability. Though simplifying the material, an example with long aliphatic chains are polymer materials. Where carbon backbone polymers are developed and are to have generally low levels of biodegradability.

Descriptors referring to electrotopological information

The slope of Eig05_AEA(dm), shown in **Figure 5g**, is negative for values less than 0 and neutral for values greater than 0, signifying that Eig05_AEA(dm) is negatively correlated with biodegradability for values less than 0, and does not seem to impact biodegradability for values greater than 0. It also shows that there is a complex relationship between the overall dipole moment of a molecule and biodegradability. More insight is needed to fully describe this descriptor's influence.

Both descriptors including dipole moment have an overall negative trend, if looking at minimal descriptor value to maximum descriptor value along the ALE plot of Eig05_AEA(dm) and SpMaxA_EA(dm) . The dipole moment is known to be the separation of charges throughout a polar molecule. The general negative trend in both Eig05_AEA suggests low levels of dipole moment being a positive trend in biodegradability. Though initial thoughts are increased level of dipole moment would allow for more interactions

to occur between microbial enzymes and molecules, this negative trend may suggest possible toxicity influence of the compounds during biodegradation process. Where reactive compounds can decay the microorganism's degradation capabilities.

The ALE plot of **Figure 5h**, shows that the slope of VE1sign_Dz(p) is positive for values less than 0.5 and negative for values greater than 0.5. Signifying that VE1sign_Dz(p) is negatively correlated with biodegradability for values less than 0.5, and positively correlated with biodegradability for values greater than 0.5. It also shows that there is a relationship between the overall polarizability of a molecule and biodegradability.

Generally polarizability increases as the volume occupied by electrons increases.[48] This is generally thought of as larger atoms have more loosely held electrons in comparison to smaller atoms.[49] The electrotopological information this provides is assumed to combine atomic size, similar to Rdf030V descriptor where volume is an influential manner. The available electrons for dipole moment to occur shows synergy with the Eig05_AEA(dm), SpMaxA_EA(dm), and GATS4p descriptors.

The slope of SpMaxA_EA(dm), as shown in **Figure 5i**, is positive for values less than roughly 1.5 and negative for values greater than that point, which signifies that if value is less than 1.5, SpMaxA_EA(dm) influences biodegradability positively, and if value is greater than that of 1.5 it impacts biodegradability negatively. It also shows there is a relationship between biodegradation and alcohols (which was confirmed by past studies to be positive) and biodegradation and dipole moment (which is a trend that could be studied further).

The incorporation of dipole moment in this descriptor aligns with what was previously mentioned in with the Eig05_AEA(dm) descriptor. Additionally, the inclusion of edge adjacency information in both descriptors gives insight into the bond-atom ratio within a molecule towards biodegradability. Though more investigation is needed to understand the bond-atom ratio influence, it highlights influential connections between structure and biodegradability.

A quick summary of the SVR model of the continuous data is as follows. High frequency of carbon and chlorine atoms at topological distance 5, high uniformness of molecule with respect to mass, and high van der Waals volume are positively correlated with biodegradability. High heterogeneousness of molecule's shape with respect to mass, large size of a molecule, and high uniformness of a molecule with respect to polarizability are negatively correlated with biodegradability. All electrotopological group descriptors have an overall negative relationship to the biodegradability of molecules. Topological distances between various atoms, overall dipole moment of a molecule, overall polarizability of a molecule, and the relationship between alcohols in a molecule and its dipole moment need to be investigated further to determine their relationship to biodegradability.

Binary Data

After initial data processing the classification data set contained 992 descriptors. SFS was used to determine the best performing descriptors. The chosen descriptors were used to develop a logistic regression classification QSBR model. The models were monitored to avoid over-fitting and the highest statistical performing model was chosen for further analysis. To monitor model performance, relative to descriptor number and avoid overfitting, the performance of models relative to number of descriptors is shown in **Figure 7** for both training and test set.

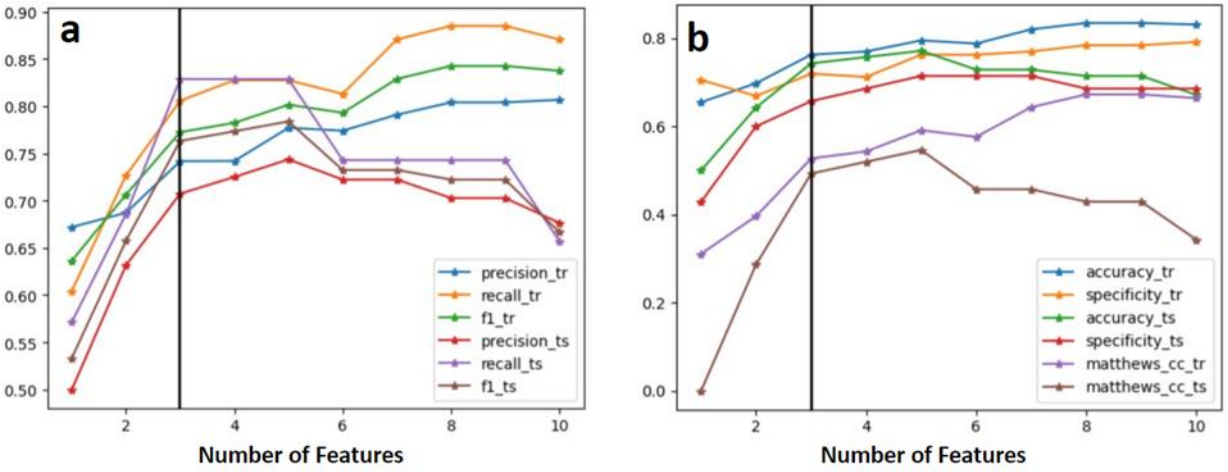

Figure 7. Performance of Models with Respect to Number of Features in Model. Training Set, a, and Test Set, b.

$$ln\left(\frac{P_x}{1-P_x}\right) = -0.2691 - 0.7823(MATS1s) - 0.7173\big(SM06\_AEA(dm)\big) \\ -0.4507(H2s) - 0.1798(B03[O-Cl]) + 0.3228(Se1C1C2s) \quad (11)$$

where P_x is the probability of a compound to be biodegradable and descriptors are presented in **Table 3**.

Table 3. Logistic Regression Classification Model Descriptors, Definitions, Types, and Model Coefficients

Descriptor	Definition	Descriptor Type	Coefficient
Electrotopological Information			
MATS1s	Moran autocorrelation of lag 1 weighted by I-state	2D autocorrelations	–0.7823
SM06_AEA(dm)	spectral moment of order 6 from augmented edge adjacency matrix weighted by dipole moment	Edge Adjacency indices	-0.7174
H2s	H autocorrelation of lag 2 / weighted by I-state	GETAWAY descriptors	-0.4507,
Se1C1C2s	E State	E-state indices	0.3228
Presence of Functional Group Information			
B03[O-Cl]	Presence/absence of O – Cl at topological distance 3	2D atom pairs	-0.1798

Once the best performing number of descriptors was determined, the quality of models was assessed based on MCC. Additionally, we take into consideration accuracy, precision, recall, f1, and specificity. The overall performance of the model shows moderate to high level of predictability while also allowing for interpretability of descriptors. As logistic regression is readily interpretable.[43] The developed logistic regression model is shown in **Equation 11**.

Model performance is shown in **Figure 8** where the MCC shows the values of true positive, true negative, false positive, and false negative. Generally, the ratio of correctly predicted true values relative to incorrectly predicted false values, give insight into the performance of our model. The statistical performance of the model is shown in **Table 4** where the model performs well. High levels of performance for precision, accuracy, and specificity suggest great predictability and classification.

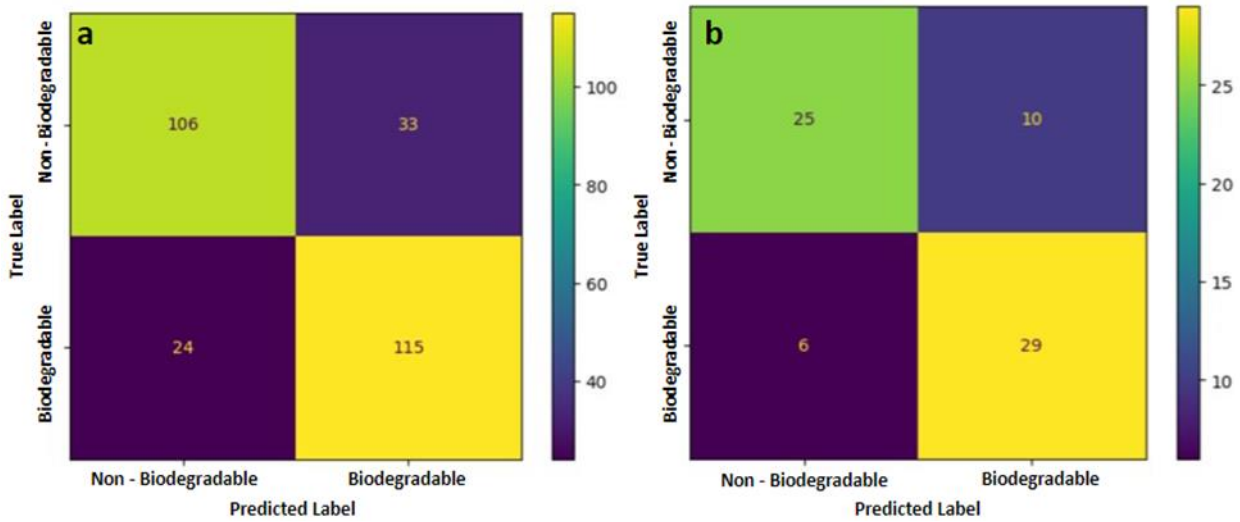

Figure 8. The Correlation Matrix of a, the Training Set, b, Test Set. of the Classification – Logistic Regression Model for Binary Data.

Table 4. - Statistical Performance of the Classification – Logistic Regression 5 Variable Model for Binary Data

Parameter	Training Set	Test Set
Precision	0.78	0.74
Recall	0.83	0.83
F1	0.80	0.78

Accuracy	0.79	0.77
Specificity	0.76	0.71
MCC	0.59	0.55

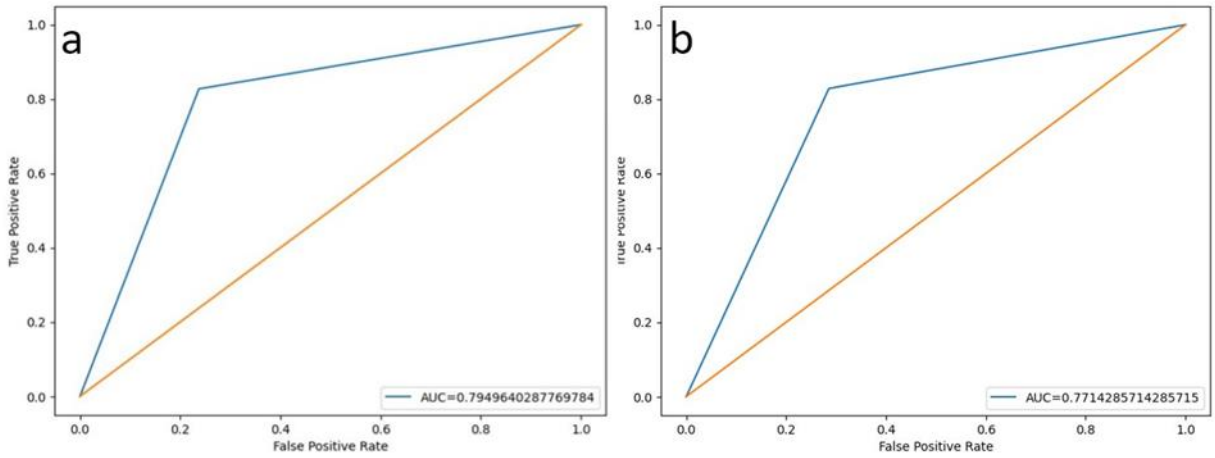

Figure 9. The Area Under the Curve of a, Training and b, Test Sets for Logistic Regression Model.

To view the robustness of the model, area under the curve (AUC) plots were generated and shown in **Figure 9.** The area under the blue curve shows the ability of the model to predict compounds within the applicability domain. If the blue line is aligned with the orange line this indicates random classification (0.5), where the orange line is superimposed onto the figure for comparison purposes.

Classification Model

Discussing **Figure 7**, the plot was made to monitor possible over fitting and determine best number of descriptors based on statistical performance. It can be seen that the model's performance, relative to number of descriptors, stagnates or reduces after the 5-descriptor mark. The best performing model with fewest number of descriptors was chosen.

Before analyzing the descriptors, we can discuss the statistical performance of the logistic regression model. **Figure 8** shows how the model disproportionally predicts false positives relative to the amount of positives (readily biodegradable) within the dataset. This indicates that the model's capability is tailored more towards predicting non-biodegradable compounds. This can occur due to influence of the low amount readily biodegradable compounds vs non readily biodegradable distribution of the dataset. The dataset is heavily weighted by non-readily biodegradable compounds. Though this occurs, influential information can still be found regarding both readily biodegradable and non-readily biodegradable compounds**.**

The ability of the model to distinguish between the two classes of readily biodegradable and non-readily biodegradable is represented as AUC plot in **Figure 9**. The closer the value to one, the better discriminatory power of the binary classification model. An AUC value of 0.79 indicates that the model has reasonably good ability to distinguish between the biodegradable and non-biodegradable compounds. This highlights the ability of our model to distinguish between the two categories even with the dataset heavily favored towards non-readily biodegradable.

The descriptors and their influential information selected by SFS is described using knowledge-based analysis. We grouped the descriptors into 2 categories based on their definition and influence of defined chemical properties. The categories being electrotopological and presence of functional groups.

Descriptors referring to electrotopological information

MATS1s, the Moran autocorrelation of lag 1 weighted by I-state, describes Moran coefficient of a molecular graph with respect to its intrinsic states at topological distance of 1.[50,51,52] The low value of the descriptor indicates random dispersion of free valence electrons, while a high value indicates valence electrons being clustered in one area of a compound.

H2s, the H autocorrelation of lag 2 / weighted by I-state, describes the average leverage of each atom in relation to other atoms at a topological distance 2, with respect to intrinsic state, where higher leverage indicates smaller or linear molecules and lower leverage indicates larger or spherical molecules. The value also increases from linear to more branched molecules.[53]

Se1C1C2s, the E-state index, describes a sum of intrinsic states based on the ratio of Kier-Hall electronegativity to the number of skeletal single bonds of molecule's Carbon 1 and Carbon 2 atoms with available sigma bonds. The intrinsic state value is larger for electronegative atoms or atoms with few skeletal connections and is smaller for atoms with several available sigma bonds.[54]

Descriptors of MATS1s, H2s, and Se1C1C2s all relate to the intrinsic state of the molecules and include some level of topological information. Intrinsic state signifies the amount of weak bonded electrons relative to strongly bonded electrons. It suggests that the electrons distributed along the molecule has influence on the biodegradability. Though it is difficult to ascertain what level of influence electronic information has on biodegradability due to the varying magnitude and sign of the coefficients, we can assume large influence by these factors. Further information on the topology includes the size, shape, and bond types thus combining with electronic information gives us a possible direction for future investigations.

SM06_AEA(dm), the spectral moment of order 6 from augmented edge adjacency matrix weighted by dipole moment, describes a complex non-linear relationship between various atoms of a compound, and in simple terms could be conceptualized as the average amount of bonds that an atom shares with other atoms written in a matrix that is brought to a power of its order (in this case 6), indicating embedding frequencies, with respect to dipole moment. The low value of SM06_AEA could indicate low branching of a molecule to dipole moment ratio, and high values could indicate high branching to dipole moment ratio.[52]

Descriptors referring to presence of functional groups

The descriptor, B03[O-Cl], describes the frequency of chlorine atoms being at a topological distance of 3 in relation to oxygen atoms. Due to the obvious connection of the descriptor and its connection to the topology of the structure allows for ease of understanding, with the chlorine atom being relatively close.

Model

As seen in the model **Equation 11**, the probability of biodegradation is negatively impacted by MATS1s, signifying that dispersed intrinsic states of a molecule aid biodegradation and clustered intrinsic states impact it negatively.

The negative impact on biodegradation by SM06_AEA(dm) indicates that low branching of a molecule to dipole moment ratio aids biodegradation, while high value becomes an obstacle to the process.

The negative impact on biodegradation by H2s indicates that molecules with low average leverage of atoms are more likely to be biodegradable, and molecules with high leverage are less likely. It also indicates that spherical molecules are more likely to be biodegradable than linear molecules, and linear molecules are more likely to be biodegradable than branched molecules.

The negative impact on biodegradation by B03[O-Cl] indicates that presence of oxygen atoms at a topological distance 3 away from chlorine atoms makes the compound less likely to be biodegradable.

The positive impact on biodegradation by Se1C1C2s indicates that compounds with C1 and C2 carbons having fewer skeletal connections are more likely to be biodegradable. The presence of few sigma bonds for those atoms would lower the chance of biodegradability.

4. Conclusions

Chlorinated compounds are generally known to be difficult to degrade due to their toxic products formed when in natural environments. Because of this, we were interested in determining compounds that contain chlorine but are known to be readily biodegradable due to their scarcity and possible use. Two models were developed with the goal of predicting the biodegradability of chlorine containing compounds. One model, being a 10 variable SVR model based on continuous data, was developed using Scikit-learn package. Analysis of the predictive capability and influential explanation of the descriptors was conducted. The statistical performance of $R^2_{Training}$ and R^2_{Test} are 0.925 and 0.881 respectively. It was found that mass, shape, topology, functionality, and electrotopological information of the compounds have influential characteristics relative to the distribution of the descript on biodegradability. Synergistic effects are believed to be expressed through commonalities of the descriptors where RDF030v and F05[C-Cl] indicate atomic volume connecting to C – CL bond arrangement in the topology of the compounds.

The second model, being logistic regression model based on categorical data of readily biodegradable vs. Non readily biodegradable, was developed and analyzed. The MCC was used as the primary indicator of model performance with an precision and accuracy of 0.78 and 0.79 respectively. The model showed higher levels of false positives which we attribute to the skewed dataset containing many more non-readily biodegradable compounds vs readily biodegradable. The descriptors were further analyzed and suggest that electrotopological information, specifical intrinsic state, and topological information have influence on biodegradability. Though it is difficult to determine the direction of influence of the electrotopological information, such as positive or negative towards biodegradability, we can see how available electrons for interatomic interactions shows varying levels of biodegradability. This may be signifying the need for electrons to be easily moved throughout the molecule for degradation to occur while also signifying possible toxic products being formed due to changes in electrotopological structure. The topological descriptor of BO3[C-Cl] suggests C to Cl connections 3 atoms apart negatively affect biodegradation. We believe it describes aliphatic chains nearby Cl atoms, such as found in polymeric materials, deter biodegradation to occur.

Both models give insight into the structural features of biodegradability of chlorine containing compounds where further investigation in the complex nature of electronic and topological information are intertwined in biodegradability. We believe future work in investigating described variations in the electronic structural and topological systems, such as applying ab initio methods, would give needed insight into the complex relationship between structure and biodegradability of chlorine containing compounds.

Acknowledgments

This work used resources of the Center for Computationally Assisted Science and Technology (CCAST) at North Dakota State University, which were made possible in part by NSF MRI Award No. 2019077.

This work is also supported in part by the ND EPSCoR award #IIA-1355466 and by the State of North Dakota. Authors also thank the Extreme Science and Engineering Discovery Environment (XSEDE) for the award allocation (TG-DMR110088). Supercomputing support from CCAST HPC System at NDSU is acknowledged.

References

(1) Kim, M. S.; Chang, H.; Zheng, L.; Yan, Q.; Pfleger, B. F.; Klier, J.; Nelson, K.; Majumder, E. L.-W.; Huber, G. W. A Review of Biodegradable Plastics: Chemistry, Applications, Properties, and Future Research Needs. *Chemical Reviews*. July 20, 2022. https://doi.org/10.1021/acs.chemrev.2c00876.

(2) Stubbins, A.; Law, K. L.; Muñoz, S. E.; Bianchi, T. S.; Zhu, L. *Plastics in the Earth System*.

(3) Thushari, G. G. N.; Senevirathna, J. D. M. Plastic Pollution in the Marine Environment. *Heliyon*. Elsevier Ltd August 1, 2020. https://doi.org/10.1016/j.heliyon.2020.e04709.

(4) Geyer, R.; Jambeck, J. R.; Law, K. L. Production, Use, and Fate of All Plastics Ever Made. *Sci. Adv.* **2017**, *3* (7), 5. https://doi.org/10.1126/sciadv.1700782.

(5) Di, J.; Reck, B. K.; Miatto, A.; Graedel, T. E. United States Plastics: Large Flows, Short Lifetimes, and Negligible Recycling. *Resour. Conserv. Recycl.* **2021**, *167* (October 2020), 105440. https://doi.org/10.1016/j.resconrec.2021.105440.

(6) Hohn, S.; Acevedo-Trejos, E.; Abrams, J. F.; Fulgencio de Moura, J.; Spranz, R.; Merico, A. The Long-Term Legacy of Plastic Mass Production. *Sci. Total Environ.* **2020**, *746*, 141115. https://doi.org/10.1016/j.scitotenv.2020.141115.

(7) Noventa, S.; Boyles, M. S. P.; Seifert, A.; Belluco, S.; Jiménez, A. S.; Johnston, H. J.; Tran, L.; Fernandes, T. F.; Mughini-Gras, L.; Orsini, M.; Corami, F.; Castro, K.; Mutinelli, F.; Boldrin, M.; Puntes, V.; Sotoudeh, M.; Mascarello, G.; Tiozzo, B.; McLean, P.; Ronchi, F.; Booth, A. M.; Koelmans, A. A.; Losasso, C. Paradigms to Assess the Human Health Risks of Nano- and Microplastics. *Microplastics and Nanoplastics* **2021**, *1* (1), 1–28. https://doi.org/10.1186/s43591-021-00011-1.

(8) Bhuyan, M. S. Effects of Microplastics on Fish and in Human Health. *Front. Environ. Sci.* **2022**, *10* (March), 1–17. https://doi.org/10.3389/fenvs.2022.827289.

(9) Shinkevich, A. I.; Kudryavtseva, S. S.; Ershova, I. G. Modelling of Energy Efficiency Factors of Petrochemical Industry. *Int. J. Energy Econ. Policy* **2020**, *10* (3), 465–470. https://doi.org/10.32479/ijeep.9396.

(10) Zughaibi, T. A.; Sheikh, I. A.; Beg, M. A. Insights into the Endocrine Disrupting Activity of Emerging Non-Phthalate Alternate Plasticizers against Thyroid Hormone Receptor: A Structural Perspective. *Toxics* **2022**, *10* (5). https://doi.org/10.3390/toxics10050263.

(11) Idumah, C. I.; Nwuzor, I. C. Novel Trends in Plastic Waste Management. *SN Applied Sciences*. Springer Nature November 1, 2019. https://doi.org/10.1007/s42452-019-1468-2.

(12) Erickson, M.; Han, Y.; Rasulev, B.; Kilin, D. Molecular Dynamics Study of the Photodegradation of Polymeric Chains. *J. Phys. Chem. Lett.* **2022**, *13* (19), 4374–4380. https://doi.org/10.1021/acs.jpclett.2c00802.

(13) Thew, C. X. E.; Lee, Z. Sen; Srinophakun, P.; Ooi, C. W. Recent Advances and Challenges in Sustainable Management of Plastic Waste Using Biodegradation Approach. *Bioresource Technology*. Elsevier Ltd April 1, 2023. https://doi.org/10.1016/j.biortech.2023.128772.

(14) Thakur, S.; Mathur, S.; Patel, S.; Paital, B. Microplastic Accumulation and Degradation in Environment via Biotechnological Approaches. *Water (Switzerland)*. MDPI December 1, 2022. https://doi.org/10.3390/w14244053.

(15) Ali, S. S.; Elsamahy, T.; Al-Tohamy, R.; Zhu, D.; Mahmoud, Y. A. G.; Koutra, E.; Metwally, M. A.; Kornaros, M.; Sun, J. Plastic Wastes Biodegradation: Mechanisms, Challenges and Future Prospects. *Science of the Total Environment*. Elsevier B.V. August 1, 2021. https://doi.org/10.1016/j.scitotenv.2021.146590.

(16) Innocenti, F. D. Biodegradability and Compostability; The International Norms, 2003; pp 33–45.

(17) ISO 11734 - Water Quality - Evaluation of the "Ultimate" Anaerobic Biodegradability of Organic Compounds in Digested Sludge — Method by Measurement of the Biogas Production. International Organization for Standardization 1995, pp 1–13.

(18) ISO 14855 - 2 - Determination of the Ultimate Aerobic Biodegradability of Plastic Materials under Controlled Composting Conditions — Method by Analysis of Evolved Carbon Dioxide. The international Organization for Standardization 2018.

(19) ISO 16929 – Plastics – Determination of Degree of Disintegration of Plastic Materials under Defined Composting Conditions in a Pilot-Scale Test. International Organization for Standardization 2021.

(20) EN 14995 Plastics - Evaluation of Compostability - Test Scheme and Specifications. Commonwealth Standards Network (CSN) 2006.

(21) EN 14046 Evaluation of the Ultimate Aerobic Biodegradability and Disintegration of Packaging Materials under Controlled Composting Conditions - Method by Analysis of Released Carbon Dioxide. German Institute for Standardisation (DIN) 2003.

(22) EN 13432 Packaging. Requirements for Packaging Recoverable through Composting and Biodegradation. Test Scheme and Evaluation Criteria for the Final Acceptance of Packaging. British Standards Institution 2007.

(23) ASTM. ASTM D5338-15 Standard Test Method for Determining Aerobic Biodegradation of Plastic Materials Under Controlled Composting Conditions, Incorporating Thermophilic Temperatures. *ASTM Int. 2021*. https://doi.org/10.1520/D5338-15R21.

(24) ASTM. ASTM D5526-12 Standard Test Method for Determining Anaerobic Biodegradation of Plastic Materials Under Accelerated Landfill Conditions. *ASTM Int.* **2018**, 6. https://doi.org/10.1520/D5526-12.

(25) Seggiani, M.; Altieri, R.; Puccini, M.; Stefanelli, E.; Esposito, A.; Castellani, F.; Stanzione, V.; Vitolo, S. Polycaprolactone-Collagen Hydrolysate Thermoplastic Blends:

Processability and Biodegradability/Compostability. *Polym. Degrad. Stab.* **2018**, *150* (February), 13–24. https://doi.org/10.1016/j.polymdegradstab.2018.02.001.

(26) Anas Karuth Wenjie Xia Bakhtiyor Rasulev, A. A. Predicting Glass Transition of Amorphous Polymers by Application ofCheminformatics and Molecular Dynamics Simulations. *Polymer (Guildf).* **2021**, *218*.

(27) Sabljic, A.; Nakagawa, Y. Biodegradation and Quantitative Structure-Activity Relationship (QSAR). *ACS Symp. Ser.* **2014**, *1174*, 57–84. https://doi.org/10.1021/bk-2014-1174.ch004.

(28) Raymond, J. W.; Rogers, T. N.; Shonnard, D. R.; Kline, A. A.; Rogers;, J. W. R. T. N.; Kline, D. R. S. A. A.; Raymond, J. W.; Rogers, T. N.; Shonnard, D. R.; Kline, A. A. A Review of Structure-Based Biodegradation Estimation Methods. *J. Hazard. Mater.* **2001**, *84* (2–3), 189–215. https://doi.org/10.1016/S0304-3894(01)00207-2.

(29) Vorberg, S.; Tetko, I. V. Modeling the Biodegradability of Chemical Compounds Using the Online Chemical Modeling Environment (OCHEM). *Mol. Inform.* **2014**, *33* (1), 73–85. https://doi.org/10.1002/minf.201300030.

(30) Toropov, A. A.; Toropova, A. P.; Lombardo, A.; Roncaglioni, A.; De Brita, N.; Stella, G.; Benfenati, E. CORAL: The Prediction of Biodegradation of Organic Compounds with Optimal SMILES-Based Descriptors. *Cent. Eur. J. Chem.* **2012**, *10* (4), 1042–1048. https://doi.org/10.2478/s11532-012-0031-4.

(31) Mansouri, K.; Ringsted, T.; Ballabio, D.; Todeschini, R.; Consonni, V. Quantitative Structure-Activity Relationship Models for Ready Biodegradability of Chemicals. *J. Chem. Inf. Model.* **2013**, *53* (4), 867–878. https://doi.org/10.1021/ci4000213.

(32) Kim, J. R.; Thelusmond, J. R.; Albright, V. C.; Chai, Y. Exploring Structure-Activity Relationships for Polymer Biodegradability by Microorganisms. *Science of the Total Environment*. Elsevier B.V. September 10, 2023. https://doi.org/10.1016/j.scitotenv.2023.164338.

(33) Andrey A. Toropov Anna Lombardo, A. P. T.; Alessandra Roncaglioni, N. D. B.; Giovanni Stella, E. B. CORAL: The Prediction of Biodegradation of Organic Compounds with Optimal SMILES-Based Descriptors. *Cent. Eur. J. Chem.* **2012**, *10*, 1042–1048. https://doi.org/10.2478/s11532-012-0031-4.

(34) Inherent Biodegradability: Modified MITI Test (II). 2009.

(35) OECD. OECD 301 - Ready Biodegradability. *OECD Guidel. Test. Chem.* **1992**, *301* (July), 1–62.

(36) Inc. (ACD/Labs). ChemSketch. Inc. (ACD/Labs): Toronto, ON, Canada 2022.

(37) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3* (10). https://doi.org/10.1186/1758-2946-3-33.

(38) Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin II; Palyulin, V. A.; Radchenko, E. V; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q. Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput. Aided. Mol. Des.* **2011**, *25* (6), 533–554. https://doi.org/10.1007/s10822-011-9440-2.

(39) Halgren, T. a. Merck Molecular Force Field. *J. Comput. Chem.* **1996**, *17* (5–6), 490–519.

(40) Halgren, T. A. Molecular Geometries and Vibrational Frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17* (5 & 6), 553–586.

(41) Halgren, T. a. Electrostatic Parameters for Intermolecular Interactions. *J. Comput. Chem.* **2000**, *17* (1996), 520–552.

(42) Van Rossum, G., & Drake, F. L. *Python 3 Reference Manual*; Scotts Valley, CA, 2009.

(43) Daghigi, A.; Casanola-Martin, G. M.; Timmerman, T.; Milenkovic, D.; Lucic, B.; Rasulev, B.; Daghighi, A.; Casanola-Martin, G. M.; Timmerman, T.; Milenković, D.; Lučić, B.; Rasulev, B.; Milenkovic, D.; Lucic, B.; Rasulev, B. In Silico Prediction of the Toxicity of Nitroaromatic Compounds: Application of Ensemble Learning QSAR Approach. *Toxics* **2022**, *10* (12), 14. https://doi.org/10.3390/toxics10120746.

(44) Pedregosa, F. . V. G. . G. A. . M. V. . T. B. . G. O. . B. M. . P. P. . W. R. . D. V. . et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*.

(45) Apley, D. W.; Zhu, J. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. **2016**.

(46) Cheng, F.; Ikenaga, Y.; Zhou, Y.; Yu, Y.; Li, W.; Shen, J.; Du, Z.; Chen, L.; Xu, C.; Liu, G.; Lee, P. W.; Tang, Y. In Silico Assessment of Chemical Biodegradability. *J. Chem. Inf. Model.* **2012**, *52* (3), 655–669. https://doi.org/10.1021/ci200622d.

(47) European Commission Environment Directorate General. Guidance Document on the Validation of (Quantitative)Structure-Activity Relationships [(Q)Sar] Models. **2014**, No. February, 1–154.

(48) EV Anslyn, D. D. Modern Physical Organic Chemistry. *J. Phys. Org. Chem.* **2011**, *24* (9), 743–743. https://doi.org/10.1002/poc.1909.

(49) Schwerdtfeger, P. ATOMIC STATIC DIPOLE POLARIZABILITIES. In *Atoms, Molecules and Clusters in Electric Fields*; PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO., 2006; pp 1–32. https://doi.org/10.1142/9781860948862_0001.

(50) Galvez, J.; Zanni, R.; Galvez-Llompart, M.; Benlloch, J. M. Macrolides May Prevent Severe Acute Respiratory Syndrome Coronavirus 2 Entry into Cells: A Quantitative Structure Activity Relationship Study and Experimental Validation. *J. Chem. Inf. Model.* **2021**, *61* (4), 2016–2025. https://doi.org/10.1021/acs.jcim.0c01394.

(51) Pagar, R. R.; Musale, S. R.; Pawar, G.; Kulkarni, D.; Giram, P. S. Comprehensive Review on the Degradation Chemistry and Toxicity Studies of Functional Materials. *Acs Biomater. Sci. Eng.* **2022**, *8* (6), 2161–2195. https://doi.org/10.1021/acsbiomaterials.1c01304.

(52) R. Todeschini, V. C.; Todeschini, R.; Consonni, V.; R. Todeschini, V. C.; Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; WILEY-VCH, 2000; Vol. 11. https://doi.org/10.1002/9783527613106.

(53) Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (3), 682–692. https://doi.org/10.1021/ci015504a.

(54) Hall, L. H.; Kier, L. B. *Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information*; 1995; Vol. 35.