Prediction of Dielectric Constant in Series of Polymers by Quantitative Structure-Property Relationship (QSPR)

Estefania Ascencio Medina *, Shan He, Amirreza Daghighi, Kweeni Iduoku, Gerardo M. Casanola-Martin, Sonia Arrasate, Humerto Gonzalez-Diaz, Bakhtiyor Rasulev

Posted Date: 13 August 2024

doi: 10.20944/preprints202408.0884.v1

Keywords: dielectric constant; polymers; QSPR; Gradient Boosting Regressor; Accumulated Local Effect

 doi:10.20944/preprints202408.0884.v1

Article

Prediction of Dielectric Constant in Series of Polymers by Quantitative Structure-Property Relationship (QSPR)

Estefania Ascencio [1,2,3] Shan He [1,2,3], Amirreza Daghighi [1,4], Kweeni Iduoku [1,4], Gerardo M. Casanola-Martin [1] Sonia Arrasate [2], Humberto González-Díaz [2,5] and Bakhtiyor Rasulev [1,4,*]

[1] Department of Coatings and Polymeric Materials, North Dakota State University, Fargo, ND 58102, USA
[2] Department of Organic and Inorganic Chemistry, Faculty of Science and Technology, University of The Basque Country (UPV/EHU), P.O. Box 644, 48080,
[3] IKERDATA S.L., ZITEK, University of The Basque Country (UPVEHU), Rectorate Building, 48940 Leioa, Spain.
[4] Biomedical Engineering Program, North Dakota State University, Fargo, ND 58105, USA.
[5] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Biscay, Spain.
* Correspondence: bakhtiyor.rasulev@ndsu.edu

Abstract: The dielectric constant (ε) reflects the ability of a material to align and orient electrical dipoles within its structure in response to an externally applied electric field; the greater the polarizability of molecules, the greater the value of dielectric constant. In this study a data set of 86 polymers was investigated to develop a structure-property quantitative relations (QSPR) model to predict the dielectric constant in polymers. An initial set of 1273 descriptors was used to select a best set of descriptors and to construct two machine learning models with Gradient Boosting Regressor (GB_A and GB_B). The best performing model (GB_A) with 8 descriptors, exhibited a performance of (R^2_{train}) = 0.938 and (R^2_{test}) = 0.802. The models were internally validated by 5 folds cross-validation, demonstrating robustness. Additionally, using the Accumulative Local Effect (ALE) technique, we analyzed the relationship between the 8 descriptors involved and the impact of these descriptors to dielectric constant of polymers.

Keywords: dielectric constant; polymers; QSPR; gradient boosting regressor; accumulated local effect

1. Introduction

Dielectric permittivity is a fundamental electrical property that characterizes the response of a material when subjected to an electric field [1] This measurement is intrinsically related to the dielectric constant (ε) and reflects the material's ability of the alignment and orientation of electrical dipoles within its structure in response to an externally applied electric field. It can be observed that the greater the polarizability of molecules, the greater the (ε) value [2]. This property is a fundamental molecular characteristic and commonly used to predict other electrical properties of polymers [3–5]. It applies to materials physics, chemistry, electrical engineering, and polymer science [1]. It is implementations are evident in high energy density capacitors [6], high voltage cables [6], microelectronics [7] and photovoltaic devices [8,9].

However, calculating the dielectric constant in polymers theoretically presents a multifaceted challenge. This inherently nonlinear property requires considering factors such as temperature, frequency, polymer structure, composition, sample morphology, impurities, loads, plasticizers, and other additives [4,10]. Furthermore, each application demands a specific range for the polymer's dielectric constant (ε), tailored to the unique requirements of the problem at hand [5]. This comprehension is crucial for designing new materials. Therefore, given the inherent complexity of many substances, there is a significant demand for machine learning (ML) models to efficiently predict these properties, optimizing both time and resources.

 doi:10.20944/preprints202408.0884.v1

In the field of materials science and cheminformatics, the Quantitative Structure-Property Relationship (QSPR) methodology stands out as an important machine learning-based approach. This methodology relies on machine learning models to forecast or elucidate compound properties by leveraging distinct chemical descriptors [11]. The efficacy of the model's predictions and its capacity to unveil the relationships between a material's molecular or other microscopic physical properties and the targeted properties being modeled are significantly influenced by the careful selection of descriptors [11]. In this sense the QSPR approach has proven to be effective in predicting various properties, including glass transition (Tg) in polymers [12,13] and (Tg) in polymer coating materials [14]. Several QSPR models have also been developed for predicting dielectric permittivity in polymers [1,15–17,]

using different datasets, feature-representation methods, variable selection procedures and so on for instance, Liu et al. [17] developed a QSPR model to predict dielectric permittivity using a small dataset of 22 polyalkenes. The resulting model, built utilizing multiple linear regression analysis (MLRA), had a high (R^2_{train}) value of 0.907 and standard error (s) of 0.001 on the training set. Three quantum descriptors were selected: ELUM (energy of the lowest unoccupied molecular orbital), q- (minimum negative atomic charge) and S (configurational entropy of the system). The authors thoroughly explored the physical significance of these descriptors, linking them to polymer polarizability and charge separation capability.

In subsequent studies in 2016, Wu et al. [16]developed a model to predict the dielectric constant of 58 polymers. They employed Partial Least Squares (PLS) regression as the modeling technique, incorporating the Infinite Chain Descriptors (ICD) 2D, TAE and GAP_inf3_inv. The model trained on the training dataset showed (R^2_{train}) of 0.91 and a Root Mean Square Error (RMSE) of 0.11. Additionally, when evaluating the model on an external test set, it achieved strong predictive capabilities, reaching an (R^2_{test}) of 0.96 and an RMSE of 0.11 in both cases. Finally, in a more recent study, Yevhenii et al. [1] used a data set of 71 polymer samples. They applied genetic algorithms (GA) and multiple linear regression analysis (MLRA) to select optimal descriptors and develop models (QSAR). Two models were created: The first model used five descriptors, achieving an (R^2_{train}) of 0.842 and a standard error (s) of 0.187. The second model incorporated eight descriptors, demonstrating improved results with (R^2_{train}) of 0.905 and s of 0.151, both models exhibited robust predictive skills when externally validated, (R^2_{test}) of 0.829 and 0.81 respectively.

Although all of these earlier publications report on QSAR/QSPR studies to predict dielectric permittivity of different polymers, they have certain limitations. First of all, not all models use a separate set of tests to validate model predictions and the size of published data sets is small or limited, restricting the applicability domain of the model.

In this work, a QSAR model was developed using a Gradient Boosting (GB) method, a sequential method that improves predictive accuracy through iterative combination and adjustment of weak models. The method is powerful since it is updating the weights after each iteration, influencing precise models in the sequence for continuous improvement of overall accuracy over time [18,19]. Thus, GB has been successfully used in QSAR models to predict bandgap [20] and glass transition temperature [21] in polymers, with predictive capacity of R^2_{train} above 0.90 in both cases, where high prediction quality was achieved even with many descriptors without overfitting [22].

The model in this study was built with a data set of 86 polymers, where two models (GB_A and GB_B) were evaluated by cross-validation and external data sets. The optimization of the models involved the use of eight descriptors, and six descriptors, respectively. Several parameters were adjusted for this model using a grid search technique. The optimized model demonstrated an effective prediction of the dielectric constant in various types of polymers. Also, in this study the Cumulative Local Effect (ALE) approach was used to facilitate the visualization of the individual impact of each descriptor on dielectric permittivity predictions. ALE graphs serve as effective tools for both visualizing and quantifying the individual influence of each input on prediction [23].

To our best knowledge, to date only one study has utilized the ALE method to elucidate the mechanistic relationship of a nonlinear QSAR models related to toxicity (log LD50) discussed in work

 doi:10.20944/preprints202408.0884.v1

[23]. However, no previous studies have been identified that apply this approach to investigate dielectric permittivity.

2. Materials and Methods

2.1. Experimental Data Collection

In this study, we examined a set of 86 polymers (Supplementary Material, Table S1) compiled from diverse public sources [1,4,24,25].The dataset encompasses various polymer types, including polyvinyles, polyethylenes, polyoxides, polystyrenes, polyethers, polysulfones, polyacrylonitrile, polyamides, polyacrylates, poly-siloxanes, polyxylylenes, polycarbonates, polyisoprenes, polymethylene, aromatic polymers, fluorinated polymers and norbornene polymer.

All experiments were performed at a temperature of 298 K, using frequencies of 60, 100, 1000, 10000 and 1000000 Hz. The dielectric constant values obtained at these frequencies were extrapolated to 1 Hz to unify the data under the equal conditions. The quality of the fits was guaranteed by a coefficient of determination (R^2) greater than 0.90 (see Supporting Information, Figure S1).

The SMILES notations (structure information) for each polymer were obtained from PubChem [26]and ChemDraw [27]. The molecules were optimized by Avogadro Software [28] with Universal Force Fields (UFF). UFF is a general force field designed to optimize minimal energy conformation that works for chemical structures based on all possible elements. It determines parameters based on the element, its hybridization, and connectivity; this force field is a big advantage over other force fields that usually only work in specific cases depending on the available parameters [29].

2.2. Generation of Descriptors

Molecular descriptors are mathematical representations of the molecular properties generated by specific algorithms based on mathematical equations [30]. The descriptors were generated using alvaDesc [31] . The program calculates more than 5000 descriptors, 0-dimensional, 1-dimensional, 2-dimensional, and 3-dimensional, GETAWAY descriptors, among others [31]. Highly correlated descriptors (R>0.9), constant and near constant descriptors (std<0.1) were removed during pre-processing. After eliminating correlated, constant, and near-constant descriptors, about 1273 descriptors were used for further QSPR analysis.

Additionally, given the higher molecular weight of the polymers, the influence of terminal groups on the overall polymer structure is minimal. Consequently, we can disregard the contribution of terminal structures. In this context, we based our calculations of structural features/descriptors on the repeating polymer units' structures. [1,17,32].

2.3. Model Assembly

For model construction and QSPR evaluation, the data was organized in ascending order and randomly partitioned into an 80% training set and a 20% test set. The preliminary phase, illustrated in the data distribution, consists of identifying and excluding 4 atypical structures from the dataset using a histogram [33]. Subsequently, a lower limit (lower limit) was established by subtracting three times the standard deviation (σ)of the mean (χ) : $Lower\ Limit = \chi - 3x\sigma$ and an upper limit (upper limit) by adding three times the standard deviation of the mean: $Lower\ Limit = \chi + 3x\sigma$. This approach was in line with the empirical standard in a normal distribution [33]Several models, including Multilinear Regressor (MLR), Support Vector Machine (SVM), Random Forests (RFR), Decision Tree (DT), K-Nearest Neighbors (KNN), and Gradient Boosting (GB) were built for further evaluation and identification of the best model. These models were generated with the coefficient of determination in the training data set (R^2_{train}) and the validation data set (R^2_{test}) parameters. For model acquisition were performed using python (3.6.3) and implemented in the Scikit-learn package [34]. The selection of variables was made with Genetic Algorithm (GA), a robust tool for search and optimization in predictive modeling [35,36]. The variable selection process using Genetic Algorithms (GA) begin with an initial population of 1000 random models. The evolutionary phase involved 9000 iterations, and a mutation probability of 20% was applied.

 doi:10.20944/preprints202408.0884.v1

2.4. Gradient Boosting Regressor Model Modeling and Validation

The Gradient Boosting Regressor model used several performance metrics, including the coefficient of determination (R^2) Equation (A.1), Root Mean Square Error (RMSE) (Equation (A.2), and Mean Absolute Error (MAE) Equation (A.3). These metrics are commonly employed to evaluate the effectiveness of the model [1,11,23,37].Here, y_i^{obs} and y_i^{pred} represent the observed and predicted values for the ith compound and $\mathcal{Y}_i^{obs}$ is the mean of the observed values. To assess the model's stability, we computed the Mean Absolute Error of cross-validation (MAE_{CV}) in each iteration based on Equation (A.4). Similarly, we used the Concordance Correlation Coefficient (CCC, Equation (A.5)) to gauge the goodness of fit. Additionally, other metrics were incorporated to obtain a more comprehensive and precise estimation of the models' predictive capacity. The external predictability of the model was assessed using metrics such as Q^2F1, Q^2F2, k, k' [38].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{obs} - y_i^{pred}\right)^2}{\sum_{i=1}^{n}(y_i^{obs} - \mathcal{Y}_i^{obs})^2} \quad (A.1)$$

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n}\left(y_i^{obs} - y_i^{pred}\right)^2}}{n} \quad (A.2)$$

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

$$MAECV = \frac{1}{n}\sum_{k=1}^{n}[\, y_i^{obs} - y_i^{predcv}] \quad (A.4)$$

$$\mathrm{CCC} = \frac{2\sum_{i=1}^{N\,obs}\left(y_i^{obs} - y^{-obs}\right)\left(y_i^{pred} - y^{-pred}\right)}{\sum_{i=1}^{n}(y_i^{obs} - y_i^{-obs})^2 + \sum_{i=1}^{n}\left(y_i^{pred} - y^{-pred}\right)^2 + n(y^{-obs} - y^{-pred})^2} \quad (A.5)$$

2.5. Analysis of Descriptors in Models

The research attempts at overcoming the challenge of interpreting non-linear models by employing the ALE approach. This approach proves effective in comprehending the impact of descriptors on the target variable [23,39]. Additionally, data normalization was performed to ensure consistency in interpreting ALE effects, thereby ensuring a precise understanding of how each descriptor influences the model's predictions. The scikit-learn package [34] was utilized for normalizing the descriptors, and ale-python package [23]. to generate graphical representations that visualize the cumulative effects of the descriptors on the predictions of the GB_A and GB_B models.

3. Results and Discussion

3.1. Exploratory Data Analysis

Data visualization through histograms is a crucial initial step in quantitative data analysis [40]. In our study, this graphical representation was applied to illustrate the distribution pattern of the dielectric permittivity for dataset. The Figure 1A show the original database, consisting of 86 polymers, this graph reveals that the majority of data points cluster around the center of the distribution, with fewer points at the extremes.

Additionally, there is a noticeable trend of points extending towards the right, indicating a right-skewed distribution [40]. The mean value (mean): 3.1488 serves as a central point around which the distribution is centered . In addition, histograms play a key role in identifying unusual values, biases and other distinguishing features within the data distribution [41]. In this context a lower limit: -0.5459, was determined by subtracting three times the standard deviation (dev): 1.2316 from the mean value (mean). Conversely, the upper limit: 6.8436, was calculated by adding three times the standard deviation to the mean value. Consequently, any data point falling above or below these limits was flagged as a potential outlier. Noteworthy instances include Fumaronitrile (8.5), Vinyl Fluoride (8.5), Vinylidene Fluoride (8.4), and Methylcellulose (6.8). After the previous step, the data was reduced to a total of 82 points. In this graph (Figure 1B) a significant improvement in the accuracy and reliability

 doi:10.20944/preprints202408.0884.v1

of our data analysis is observed, ensuring a more precise representation of the dataset and achieving a normal distribution.

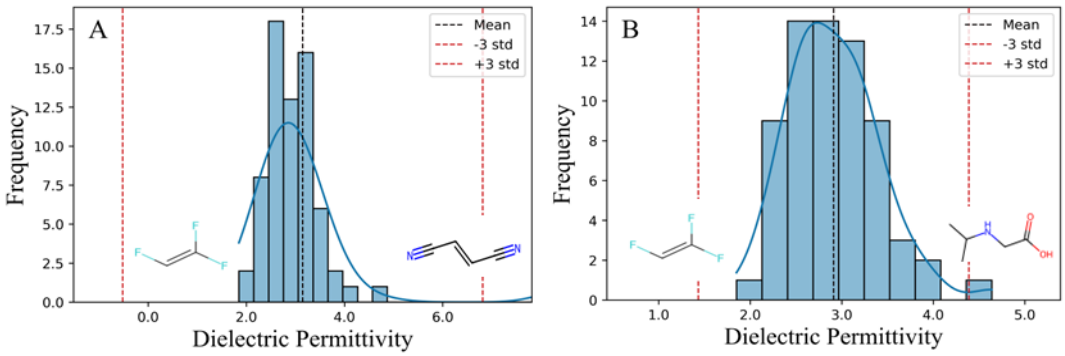

Figure 1. Histogram of a dataset. (A) The original database comprises dielectric permittivity values for 86 polymers, (B) The database is reduced to dielectric permittivity values in 82 polymers after removing outliers.

3.2. Ensemble Model

After an initial preprocessing phase, a dataset of 82 polymers were split into training and test data set, containing 66 and 16 polymers respectively. A total of 1273 descriptors were generated for this data set. Using these descriptors, different models were developed using the following algorithms: Multi-Linear Regressor (MLR), Support Vector Machine (SVM), Random Forests (RF), Decision Tree (DTR), K-Nearest Neighbors (KNN), and Gradient Boosting (GB) (Figure 2). When analyzing several models of the predictive performance was assessed using the coefficient of determination (R^2), aiming for a determination coefficient close to 1 [37].However, overall, all values fell below 0.6 and additionally there was not consistent relationship between the model performance, when evaluated with training data (R^2_{train}) and external data (R^2_{test}) (Figure 2). Nevertheless, the GB model proved effective in predicting dielectric permittivity, surpassing the other ML models. Two options were chosen that performed better for the Gradient Boosting (GB) model. The first model (GB_A) consisted of 8 descriptors and the second model (GB_B) consisted of 6 descriptors (Table 2). Additionally, a hyperparameter optimization was conducted for each model (Table 1). This optimization was crucial to significantly improve the predictive performance of the model while reducing the risk of overfitting by simplifying its complexity [41]. Statistical parameters of the model are presented in Table 3.

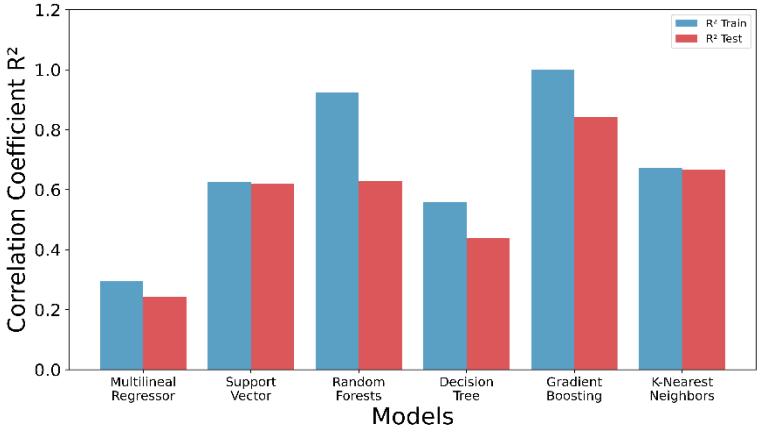

Figure 2. This graphic shows the comparison of models in the prediction of dielectric permittivity for polymers.

Table 1. Runtime parameters for Gradient Boosting Regressor.

Model	Common values	Unique values

 doi:10.20944/preprints202408.0884.v1

Type		
Gradient Boosting Regressor_A	alpha: 0.9; ccp_alpha: 0.0; criterion:friedman_mse; init: None; learning_rate: 0.2; loss: squared_error; max_features: None; max_leaf_nodes: None; min_impurity_decrease: 0.0; min_samples_leaf: 1;	max depth: 4; n estimators: 10
Gradient Boosting Regressor_B	min_samples_split: 2; min_weight_fraction_leaf: 0.0; n_iter_no_change: None; random_state: 42; subsample: 1.0; 'tol': 0.0001; validation_fraction: 0.1; verbose: 0; warm_start: False.	max depth': 2; n estimators: 13

Table 2. Descriptors involved in the GBR model and the corresponding definition.

Descriptor	GBR_A	GBR_B	Definition and Scope	Descriptor Type
N%	X		percentage of N atoms	Constitutional Indices
J_Dz(p)	X		Balaban-like index from Barysz matrix weighted by polarizability	2D matrix-based descriptors
P_VSA_e_3	X		P_VSA-like on Sanderson electronegativity, bin 3	P_VSA-like descriptors
P_VSA_i_1	X		P_VSA-like on ionization potential, bin 1	P_VSA-like descriptors
AVS_Coulomb	X		Average vertex sum from Coulomb matrix	3D matrix-based descriptors
TDB09m	X	X	3D Topological distance-based descriptors lag 9 weighted by mass	3D autocorrelations
HATS2p	X		leverage-weighted autocorrelation of lag 2 /weighted by polarizability	GETAWAY descriptors
MLOGP2	X	X	squared Moriguchi octanol-water partition coeff. (logP^2)	Molecular properties
GATS2s		X	Geary autocorrelation of lag 2 weighted by I-state	2D autocorrelations
Eig08_AEA (ri)		X	Eigen value n. 8 from augmented edge adjacency mat. weighted by resonance integral	Edge adjacency indices
RTs+		X	R maximal index / weighted by I-state	GETAWAY descriptors
WHALES60_Rem		X	WHALES Remoteness (Rem) (percentile 60)	WHALES descriptors

Table 3. Statistical parameters of Gradient Boosting model.

Model	R^2 (train)	RMSE (train)	MAE (train)	MAE_{CV}	R^2 (test)	RMSE (test)	MAE (test)	CCC (test)	$Q2_{F1}$	$Q2_{F2}$	k	k'
GBR_A	0.938	0.123	0.100	0.261	0.802	0.256	0.212	0.869	0.805	0.802	1.035	0.961
GBR_B	0.822	0.208	0.155	0.273	0.704	0.313	0.213	0.787	0.710	0.704	0.101	0.980

 doi:10.20944/preprints202408.0884.v1

In the model GB_A (R^2_{train}) shows a good performance, indicating the model's ability to capture and explain 93.77% of the variations in the training data, showcasing its effective adaptation and precise predictions within this specific set. As for the test set, the R^2_{test} value of 0.801 illustrates a very good model's performance on the external set, which was not used during the model training. This high performance in both training and test sets highlights the model's robustness, supporting its ability to generalize and provide accurate predictions in future scenarios. In contrast, the GB_B model showed slightly lower prediction ability for both the training set (R^2_{train}): 0.822 and the test set (R^2_{test}) 0.708. Therefore, we could assume that having more descriptors might allow the model to capture more details and subtle relationships in the data, potentially improving the accuracy of predictions, however this improvement could introduce a higher complexity to the model. In Figure 3, the relationship between predicted and experimental values for the dielectric constant is illustrated, comparing models GB_A (Figure 3A) and GB_B (Figure 3B). It can be observed that residual errors are small for model GB_A, in contrast to model GB_B. Additionally, the black line represents the regression line associated with the data points, where residual errors are evident.

Without prejudice to the statistics discussed above, other important performance parameters that should also be considered for the selection of good predictive models are the mean root quadratic error (RMSE) and Mean Absolute Error (MAE) [37]. In our study, when comparing the GB_A and GB_B models, the highest R^2 values for training and test sets were consistently correlated with lower RMSE and MAE values. Thus, model GB_A highlights the ability of the model to make better predictions that are quite close to real values. Furthermore, the data were assessed in a cross-validation set (CV) for Models A and B, resulting in a similar MAEcv of 0.2605 and 0.2725, respectively. This indicates, that the models' predictions during cross-validation have an average absolute error of around 0.27 units compared to the actual values

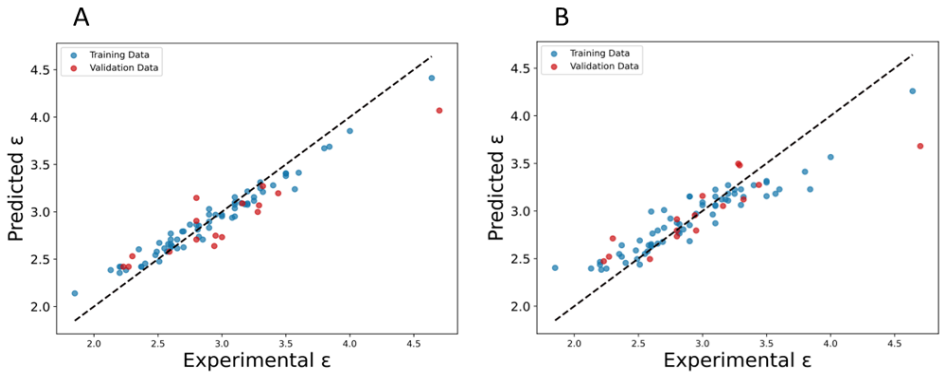

Figure 3. Plots of experimental vs predicted values of the dielectric constant, for GBR_A and GBR_B models.

Previous research [1] reported a QSPR model that utilized GA-MLR analysis. The models developed were generated from 71 polymers, achieving an (R^2_{train}) of 0.905 on the training set and an external (R^2_{test}) of 0.812 on the test set. This dataset served as the main starting point for our study, to which an additional set of polymers was added. It is crucial to highlight that, despite our study employing a gradient-boosting model, the consistency of the results between this model and the one developed earlier suggests the robustness of this methodology. This approach demonstrates its effective ability to capture the relationship between predictor variables and the response variable, even when considering an expanded dataset with the inclusion of additional polymers.

In a similar study Bicerano [4]crafted a QSPR model achieving an (R^2) of 0.958 and (s) of 0.087, aiming to establish a correlation between (ε) and 32 descriptors related to the structure of 61 polymers. However, this model's complexity initiates from an abundance of descriptors, potentially leading to issues like overfitting. The decision to augment the descriptor count may have enhanced results, yet it introduces complexity affecting the model's reliability. Moreover, the model in discussed paper lacks an external validation, i.e. no test set utilized. In a similar way Xu et al. [15] employed a dataset comprising 57 polymers. Instead of using simple repetitive units, they utilized

 doi:10.20944/preprints202408.0884.v1

cyclic dimers to represent polymer structures, providing a more accurate capture of the chemical environment's impact. In total, nine descriptors related to composition, connectivity, charges, and topological indices were selected in the model discussed. The QSPR model yielded (R^2_{train}) of 0.938 and standard error (s) of 0.0873, using MLRA method. Furthermore, the model underwent the external validation on a test set of 12 polymers, achieving notable results with an (R^2_{test}) of 0.969.

While these studies have produced high predictions, the limited dataset limits polymer diversity and the scope of predictions. The Gradient Boosting model provides greater flexibility compared to the Multiple Linear Regression (MLR) model. This model refines an additive model by optimizing regression trees and minimizing the loss function. The "additive nature" means that the model gradually builds complexity, improving its accuracy progressively. The tree-based approach involves constructing decision trees, allowing the model to capture non-linear relationships between input and output variables, and adept at handling complex patterns in the data. This approach offers versatility in modeling various aspects, including interactions between variables, capturing discontinuities, and effectively handling non-monotonous effects present in the dataset [18].

3.3. ML-QSPR Models Explanation

Molecular descriptors play a fundamental role in cheminformatics, representing chemical information in a computationally interpretable format [42]. According to Table 2 the GB_A model comprises of 8 descriptors, and GB_B model comprises of 6 descriptors. These models share two descriptors: TDB09m (spatial 3D molecular geometry and atomic properties of polymeric structures [43] and MLOGP2 (squared Moriguchi octanol–water partition coefficient, a descriptor of lipophilicity indicating a molecule's affinity for non-polar environments based on molecular characteristics such as hydrophobicity, ring structures, hydrogen bonds, etc.) [44] . This descriptor could suggest that polymers based on repeating units with low MLOGP2 values (highly polar) are likely to exhibit high dielectric permittivity. This could be due to the fact that molecular chains distort and orient easily in response to an electric field. The presence of these descriptors could indicate that they have a significant impact on the objective variable studied. Additionally, GB_A model includes the descriptor HATS2p, and GB_B model includes the descriptor RTs+, where both belong to GETAWAY type, this type of descriptors is related to 3D molecular geometry using atomic weights, like atomic mass, polarizability, van der Waals volume, electronegativity, and unit weights [45].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2024 doi:10.20944/preprints202408.0884.v1

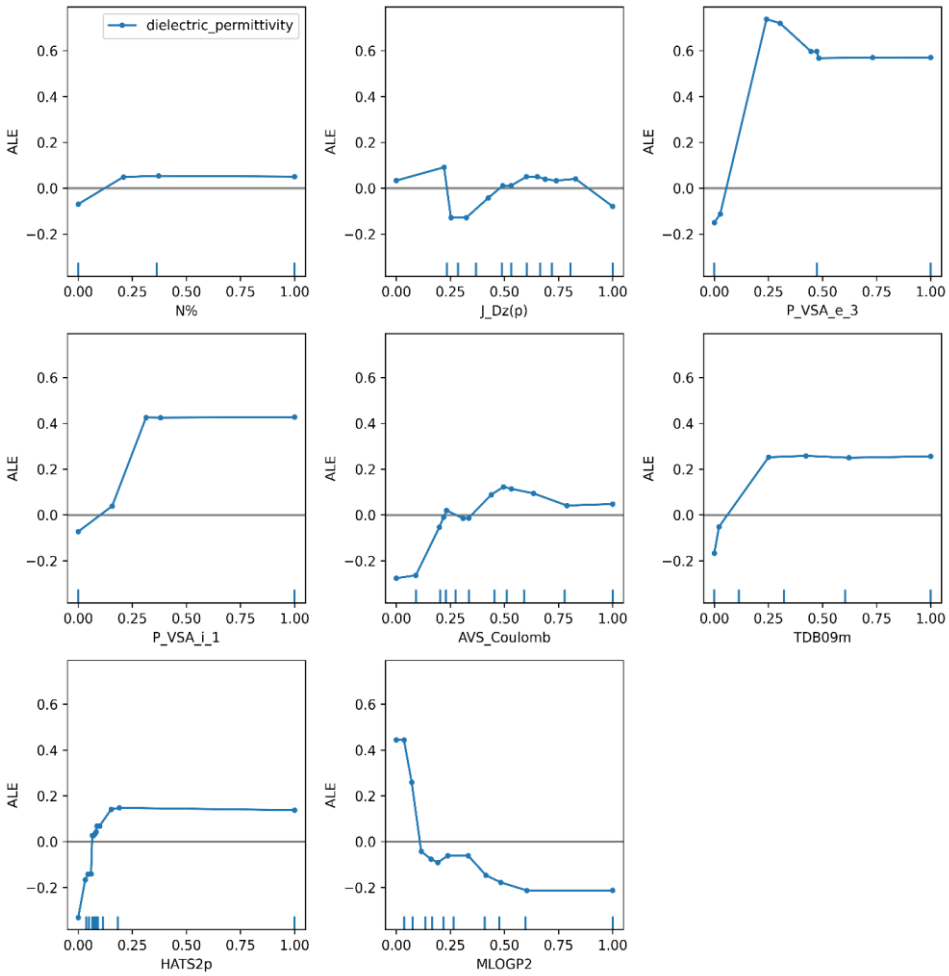

Figure 3. GB_A model. Descriptors represented by ALE plot for the Gradient Boosting model A.

Among the selected descriptors in each model, GB_A model has descriptors related to the type of constitutional indices. For example, N% (percentage of nitrogen atoms) quantifies the proportion of nitrogen atoms in the polymers of the data [43], the presence of functional groups in polymers, such as amino (-NH_2) or cyanide (-CN), which could determine the the behavior of this descriptor towards the property in a positive correlation, where more functional groups lead to higher permittivity.

The descriptors P_VSA_e_3 and P_VSA_i_1 is directly related to the van der Waals surface area (VSA), showing a specific characteristic in a defined area [46,47] where the contribution of these interactions is influenced by Sanderson electronegativity for the first descriptor and ionization potential for the second descriptor, showing a positive trend in both cases, until the descriptor values reach their medium values. Another descriptor is J_Dz(p), which belongs to the 2D type descriptors based on topological representation. It represents a Balaban type index of the polarity-weighted Barysz matrix [43,48]. Finally AVS Coulomb provides a measure of the mean electrostatic interactions between atoms in a 3D molecular structure, taking into account both repulsion and nuclear charge effects, which require 3D coordinates for all atoms, including hydrogen atoms [49].

The GB_B model includes also GATS2s descriptors, which are capturing the similarity between pairs of atoms in the molecule separated by a certain topological distance or lag [50]. This descriptor is related to important properties in dielectric permittivity, such as electronegativity and polarizability, and includes effects of atomic mass and volume, for fragments that are having 2 or more bonds (lag2). Another descriptor, Eig08_AEA (ri), belongs to the Edge Adjacency Indices type, based on the edge adjacency matrix of a graph, providing the sum of all edge entries in the graph's adjacency matrix [51]. Lastly, the descriptor WHALES60_Rem belongs to the WHALES type.

 doi:10.20944/preprints202408.0884.v1

The Figure 3 shows that the descriptors HATS2p and N% do not show a significant effect on dielectric permittivity. However, P_VSA_e_3 and P_VSA_i_1, up to values close to 0.25, have a positive influence on model predictions. The first descriptor relates to Sanderson's electronegativity while P_VSA_i_1 is associated with ionization potential (Table 2). These effects are directly linked to polarization, a crucial factor in dielectric permittivity [4] Electronegativity could lead to a more significant charge distribution in polymers, affecting their ability to respond to electric fields and thus increasing dielectric permittivity [4,52,53].

However, with the increase of values above 0.25 for this descriptor, the influence on the target variable remains constant. The descriptor TDB09m shows similar behavior to the two previous descriptors. We could assume that the high mobility of atoms in the polymer chain could facilitate their ability to align against the electric field, thus obtaining greater permittivity. On the other hand, the descriptor MLOGP2 shows a negative effect until values close to 0.65. As for the descriptor AVS_Coulomb, it shows a significant positive effect within values of the approximate range of 0.25 to 0.50. Finally, the descriptor J_Dz(p) also shows a subtle negative trend around the values close to 0.25, this descriptor potentially correlates with dielectric permittivity, as it captures aspects of molecular structure associated with polarity and electronic distribution [43,48]. In general, when analyzing the trends of several descriptors for the GB_A model, we could infer that values close to 0.25 can mean a remarkable turning point in how descriptors influence the prediction of dielectric permittivity.

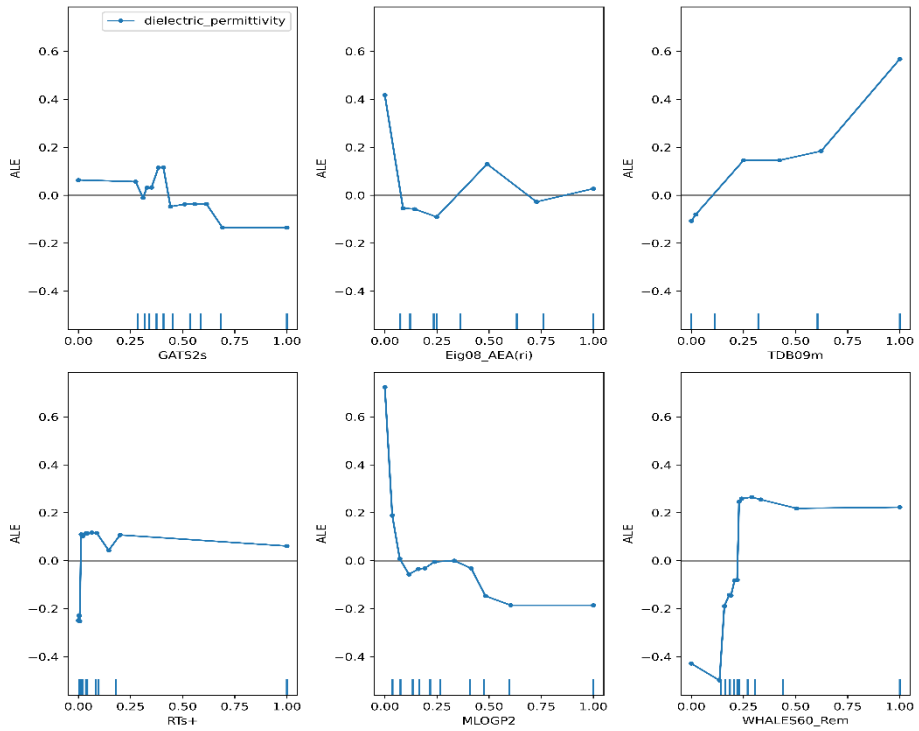

Figure 4. GB_B model. Descriptors represented by ALE plot for the Gradient Boosting model B.

The ALE graphs for the GB_B model provide important information on the factors influencing the dielectric constant in investigated polymers. The descriptor RTs+, does not have a significant impact on dielectric permittivity for most of the part of values, except smallest ones, close to 0. Nevertheles for the two shared descriptors, in the two TDB09m and MLOGP2 models, it should be noted that they behave similarly, showing a strong positive and negative trends, respectively. Therefore, we can conclude that these descriptors play a crucial role in determining dielectric permittivity in our models. However, in the GB_B model, the descriptor TDB09m has a more pronounced positive effect when its values increase beyond 0.65. In the same way, the descriptor WHALES60_Rem also shows a positive effect when its values are around 0.25, but for higher values, a constant behavior is observed. This descriptor belongs to the WHALES type of descriptors, and

 doi:10.20944/preprints202408.0884.v1

based on 3D structure considering all atoms and bonds, along with distances between them and other important properties, such as electronegativity [54].

4. Conclusions

A model was developed to predict the dielectric constants (ε) for various polymers providing a detailed explanation from a mechanistic perspective. The study introduced QSPR models developed by applying Gradient Boosting algorithm. The GB_A model, having 8 descriptors, showed better performance with (R^2_{train}) = 0.938 and (R^2_{test}) = 0.802, while the GB_B model, which has 6 descriptors, showed (R^2_{train}) = 0.822 and (R^2_{test}) = 0.704. The validity of the models was additionally ensured by various statistical verification methods, such as MAE and RMSE. The contribution of each descriptor to dielectric permittivity was discussed by applying the Acumulative Local Effect (ALE) approach. This approach worked well in analyzing the individual influence of each descriptor on dielectric permittivity predictions. The QSPR-GBR models have 5 descriptors in total that showed strong positive effects on dielectric permittivity, while one common descriptor (MLOGP2) showed a negative effect. It is important to note that TDB09m was also involved in these two models, having a positive effect. In conclusion, this study demonstrated an appropriate approach to guide the prediction of dielectric constants in a wide range of polymers, using non-linear models. The ability to predict the dielectric constant through models, with relationship-related interpretations in ALE plots, not only optimizes the design of polymers with specific electrical properties, but also accelerates the development of polymeric materials for practical applications, reducing the need for costly and lengthy experiments.

Supplementary Information: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Table S1. Experimental data representing the dielectric constant of the polymers used in the experiments; Figure S1. Machine Learning prediction of polymers at different frequencies.

Author contributions: Conceptualization, B.R.; Methodology, E.A., S.H., A.D., K.I and G.C.; Validation, E.A.; Formal Analysis, E.A., G.C. and B.R.; Investigation, E.A.; Resources, G.C.; Data Curation, E.A. and G.C.; Writing – Original Draft Preparation, E.A.; Writing – Review & Editing, G.C. and B.R.; Visualization, E.A. and G.C.; Supervision, S.A., H.G.D. and B.R.; Project Administration, B.R.; Funding Acquisition, S.A., H.G.D. and B.R. All authors have read and agreed to the published version of the manuscript.

Funding: The work used resources of the Center for Computationally-Assisted Science and Technology (CCAST) at North Dakota State University (Fargo, ND USA), which was made possible in part by the U.S. National Science Foundation (NSF) MRI Award No. 2019077. A.D. and B.R. also thank the Biomedical Engineering Program for financial support in the form of GRA funding and DOE DE-SC0021287 for partial support (for method development). The authors thank Prof. Paola Gramatica for generously providing a free license for the QSARINS software. Moreover, this work was supported in part by Grant IKERDATA 2022/IKER/000040 funded by NextGenerationEU funds of European Commission.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhuravskyi, Y.; Iduoku, K.; Erickson, M. E.; Karuth, A.; Usmanov, D.; Casanola-Martin, G.; Sayfiyev, M. N. D.; Ziyaev, A.; Smanova, Z.; Mikolajczyk, A.; Rasulev, B. Quantitative Structure Permittivity Relationship Study of a Series of Polymers. *ACS Mater. Au* **2024**, 4, 195–203.
2. Zahidul, M. D.; Fu, Y.; Deb, H.; Khalid, M. D.; Dong, Y.; Shi, S. Polymer-based low dielectric constant and loss materials for high-speed communication network: Dielectric constants and challenges. *Eur. Polym. J* **2023**, 200, 112543.
3. Afantitis, A.; Melagraki, G.; Makridima, K.; Alexandridis, A.; Sarimveis, H.; Iglessi-Markopoulou, O. Prediction of high weight polymers glass transition temperature using RBF neural networks. *J. Mol. Struct. THEOCHEM* **2004**, 716, 193–198.
4. Bicerano, J. Prediction of Polymer Properties, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2002; pp. 1-784.

 doi:10.20944/preprints202408.0884.v1

5. Chen, L.; Kim, C.; Batra, R.; Lightstone, J. P.; Wu, C.; Li, Z.; Deshmukh, A. A.; Wang, Y. Frequency-dependent dielectric constant prediction of polymers using machine learning. *NPJ Comput. Mater* 2020 , 6, 61.
6. Ma, R.; Baldwin, A. F.; Wang, C.; Offenbach, I.; Cakmak, M.; Ramprasad, R.; Sotzing, G. A. Rationally designed polyimides for high-energy density capacitor applications. *ACS Appl. Mater. Interfaces* **2014**, 6, 10445–10451.
7. Maier, G. Low dielectric constant polymers for microelectronics. *Prog. Polym.* Sci. **2001**, 26, 3–65.
8. Dang, M. T.; Hirsch, L.; Wantz, G. P3HT: PCBM, best seller in polymer photovoltaic research. *Advanced Materials* **2011**, 23, 3597–3602.
9. Facchetti, A. π-Conjugated polymers for organic electronics and photovoltaic cell applications. *J. Mater. Chem* **2011**, 23, 733–758.
10. Kim, J. H.; Kim, S. Y.; Moore, J. A.; Mason, J. F. Dielectric Properties of Poly(enaminonitrile)s. *Polym. J* 2000, 32, 57–61.
11. Le, T.; Epa,V. C.; Burden, F.R.; Winkler, D.A. Quantitative structure-property relationship modeling of diverse materials properties. *Chem. Rev* **2012**,112, 2889–2919.
12. Chen, M.; Jabeen, M. F.; Rasulev, B.; Ossowski, M.; Boudjouk, P. A computational structure–property relationship study of glass transition temperatures for a diverse set of polymers. *J. Polym. Sci* **2018**, 56, 877-885.
13. Karuth, A.; Alesadi, A.; Xia,W.; Rasulev,B. Predicting glass transition of amorphous polymers by application of cheminformatics and molecular dynamics simulations. *Polym. J* **2021**, 218, 123495.
14. Petrosyan,L. S.; Sizochenko, N.; Leszczynski, J.; Rasulev, B. Modeling of Glass Transition Temperatures for Polymeric Coating Materials: Application of QSPR Mixture-based Approach. *Mol Inform* **2019**, 38, 8–9.
15. Xu, J.; Wang, L.; Liang, G.; Wang, L.; Shen, X. A general quantitative structure-property relationship treatment for dielectric constants of polymers. *Polym Eng Sci* **2011**,51,2408–2416.
16. Wu, K.; Sukumar, N.; Lanzillo, N. A.; Wang, C.; Ramamurthy, R.; Ma, R.; Baldwin, A. F.; Sotzing, G.; Breneman, C. Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: Toward optimized dielectric polymeric materials. *J. Polym. Sci* **2016**, 54, 2082–2091.
17. Liu,A.; Wang, X.; Wang, L.; Wang, H.; Wang, H. Prediction of dielectric constants and glass transition temperatures of polymers by quantitative structure property relationships. *Eur. Polym. J* **2007**, 43,989–995.
18. Guillen, M. D.; Aparicio, J.; Esteve, M. Gradient tree boosting and the estimation of production frontiers. *Expert Syst Appl* **2023**, 214, 119134.
19. Sipper, M.; Moore, J. H. AddGBoost: A gradient boosting-style algorithm based on strong learners. *Mach. Learn. Appl* **2021**, 7, 100243.
20. Goh, K.L.; Goto, A.; Y. Lu. LGB-Stack: Stacked Generalization with LightGBM for Highly Accurate Predictions of Polymer Bandgap. *ACS Omega* **2022**, 7, 29787–29793.
21. Tao, L.; Varshney, V.; Li, Y. Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature. *J. Chem. Inf. Model* **2021**, 61, 5395–5413.
22. Malashin, I. P.; Tynchenko, V. S.; Nelyub, V. A.; Borodulin, A. S.; Gantimurov, A. P. Estimation and Prediction of the Polymers. Physical Characteristics Using the Machine Learning Models. *Polymers* **2023**,16. 115.
23. Daghighi, A.; Casanola-Martin, G.M.; Timmerman, T.; Milenković, D.; Lučić, B.; Rasulev, B. In Silico Prediction of the Toxicity of Nitroaromatic Compounds: Application of Ensemble Learning QSAR Approach. *Toxics* **2022**, 10, 746.
24. Zha, J. W.; Zheng, M. S.; Fan, B. H.; Dang, Z. M. Polymer-based dielectrics with high permittivity for electric energy storage: A review. *Nano Energy* **2021**, 89, 106438.
25. Ho, J. S.; Greenbaum, S. G. Polymer Capacitor Dielectrics for High Temperature Applications, *ACS Appl. Mater. Interfaces* **2018**, 10, 29189–29218.
26. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han,L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, Bo; Zhang, J.; Bryant, S. H. PubChem substance and compound databases. *Nucleic Acids Res* **2016**, 44.
27. Cousins, K. R.; ChemDraw Ultra 9.0. CambridgeSoft, 100 CambridgePark Drive, Cambridge, MA 02140. www. cambridgesoft.com. *J. Am. Chem. Soc* **2005**, 127, 4115– 4116.
28. Hanwell, M.D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J Cheminform* **2012**, 4,17.
29. Jász, Á.; Rák, Á.; Ladjánszki, I.; Cserey, G. Optimized GPU implementation of Merck Molecular Force Field and Universal Force Field. *J Mol Struct* **2019**, 1188, 227–233.
30. Zhao, Y.; Mulder, R. J.; Houshyar, S.; Le, T. C.; A review on the application of molecular descriptors and machine learning in polymer design. *Polym. Chem* **2023**, 14, 3325–3346.
31. Mauri, A. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. In Ecotoxicological QSARs, 1st ed.; Roy, K., Ed.; Springer Nature: 2020; pp. 1-851.

32. Sun, L.; Zhou, L.; Yu, Y.; Lan, Y.; Li, Z. QSPR study of polychlorinated diphenyl ethers by molecular electronegativity distance vector (MEDV-4). *Chemosphere* **2007**, 66, 1039–1051.
33. Witte, R. S.; Witte, J. S. Statistics, 11th ed.; Wiley: Hoboken, NJ, USA, 2021; pp. 1-496.
34. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn* **2011**,12, 2825–2830.
35. Katoch, S S.; Chauhan, S.; Kumar, V. A review on genetic algorithm: past, present, and future. *Multimed Tools Appl* **2021**, 80, 8091–8126.
36. Gad, A. F. PyGAD: An Intuitive Genetic Algorithm Python Library. *Multimed Tools Appl* **2024**,83, 58029–58042.
37. Gramatica, P.; Sangion, A. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology *J. Chem. Inf. Model* **2016**, 56, 1127–1131.
38. Apley, D. W.; J. Zhu. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *J. R. Stat. Societ Methodol* **2020**, 82, 1059–1086.
39. Wand, M. P. Data-Based Choice of Histogram Bin Width, *American Statistician* **1997**,51,59–64.
40. Boels, L.; Bakker, A.; Van Dooren, W.; Drijvers, P. Conceptual difficulties when interpreting histograms: A review. *Educ. Res. Rev* **2019**, 28, 100291.
41. Bardenet, R.; Brendel, M.; Kégl, B.; Sebag, M. Collaborative hyperparameter tuning. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16-21 June 2013; PMLR: 2013; Vol. 28, pp. 199-207.
42. Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb Chem High Throughput Screen* **2000**,3, 363-72.
43. Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors; WILEY-VCH: Weinheim, Germany, 2000; pp. 154–196.
44. Khan, K.; Kumar, V.; Colombo, E.; Lombardo, A.; Benfenati, E.; Roy, K. Intelligent consensus predictions of bioconcentration factor of pharmaceuticals using 2D and fragment-based descriptors. *Environ Int* **2022**, 170, 107625.
45. 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, 682–692.
46. Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell* **2000**,18, 464-477.
47. Guha R. Willighagen E. A Survey of Quantitative Descriptions of Molecular Structure. *Curr Top Med Chem* **2012**,18, 1946-56.
48. Sun, G.; Fan, T.; Sun, X.; Hao, Y.; Cui, X.; Zhao, L.; Ren, T.; Zhou, Y.; Zhong, R.; Peng, Y. In Silico Prediction of O^6-Methylguanine-DNA Methyltransferase Inhibitory Potency of Base Analogs with QSAR and Machine Learning Methods. *Molecules* **2018**,23, 2892.
49. Huoyu, R.; Zhiqiang, Z.; Zhanggao, L.; Zhenzhen, X. QSPR models for the critical temperature and pressure of cycloalkanes. *Chem Phys Lett* **2022** ,808, 140088.
50. Velázquez-Libera, J. L.; Caballero, J.; Toropova, A. P.; Toropov, A. A. Estimation of 2D autocorrelation descriptors and 2D Monte Carlo descriptors as a tool to build up predictive models for acetylcholinesterase (AChE) inhibitory activity. *Chemom Intell Lab Syst* **2019**, 184, 14-21.
51. Dehmer, M.; Emmert-Streib, F.; Tripathi, S. Large-scale evaluation of molecular descriptors by means of clustering. *PLoS One* **2013**, 8, 83956.
52. Qiu, J.; Gu, Q.; Sha, Y. ; Huang, Y.; Zhang, M.; Z. Luo. Preparation and application of dielectric polymers with high permittivity and low energy loss: A mini review. *J. Appl. Polym. Sci* **2022**,139,52367.
53. Wang, Q.; Che, J.; Wu, W.; Hu, Z.; Liu, X.; Ren, T.; Chen, Y.; Zhang, J. Contributing Factors of Dielectric Properties for Polymer Matrix Composites. *Polymers* **2023**, 15, 590.
54. Grisoni, F.; Merk, D.; Byrne, R.; Schneider, G. Scaffold-Hopping from Synthetic Drugs by Holistic Molecular Representation. *Sci Rep* **2018**,8, 16469.