JOTCS A
Journal of the Turkish Chemical Society
Section: A
Chemistry
TURKISH CHEMICAL SOCIETY

Supervised Machine Learning-Graph Theory Approach for Analyzing the Electronic Properties of Alkane

Zubainun Mohamed Zabidi[1], Nurul Aimi Zakaria[1], Ahmad Nazib Alias[1*]

[1]Faculty of Applied Sciences, Universiti Teknologi MARA Perak Branch Tapah Campus, 35400 Tapah Road, Malaysia

Abstract: The combination of advanced scientific computing and quantum chemistry improves the existing approach in all chemistry and material science fields. Machine learning has revolutionized numerous disciplines within chemistry and material science. In this study, we present a supervised learning model for predicting the HOMO and LUMO energies of alkanes, which is trained on a database of molecular topological indices. We introduce a new moment topology approach has been introduced as molecular descriptors. Supervised learning utilizes artificial neural networks and support vector machines, taking advantage of the correlation between the molecular descriptors. The result demonstrate that this supervised learning model outperforms other models in predicting the HOMO and LUMO energies of alkanes. Additionally, we emphasize the importance of selecting appropriate descriptors and learning systems, as they play crucial role in accurately modeling molecules with topological orbitals.

Keywords: Supervised machine learning, molecular descriptor, topological indices, electronic properties.

Submitted: August 24, 2022. **Accepted:** October 28, 2023.

Cite this: Zabidi ZM, Zakaria NA, Alias AN. Supervised Machine Learning-Graph Theory Approach For Analyzing the Electronic Properties of Alkane. JOTCSA; 11(1): 137-48.

DOI: https://doi.org/10.18596/jotcsa.1165158.

***Corresponding author. E-mail:** ahmadnazib111@uitm.edu.my

1. INTRODUCTION

Chemical graph theory is a multidisciplinary field combining graph theory and knowledge of chemistry. In chemistry, graph theory is used to analyze various chemical phenomena such as chemical compound composition and classification (1). Chemical graph theory uses the set of points connected by lines to determine structure-property relationships. The molecular structure represented a graph **G**, a set of mathematical structures consisting of several vertices and edges. In chemical graph theory, the molecular structure is typically a suppressed hydrogen with the carbon atom skeleton representing the covalent bond between carbon-carbon atoms (2). The molecular structure is associated with the topological index in chemical graph theory. Topological index is a numerical number invariant for each molecule based on the criteria set using graph theory. Topological indices have gained much attention in various areas of biology and chemistry (3). These are due to the topology indices' most important use in quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR) (4). Wiener invented the first topology index, known as the Wiener distance index. The Wiener index has been improved by Randic, known as the hyper-Wiener index (5). The improvement of the Wiener index is still ongoing to tailor the application requirement (6). There are various classes of topological indices, such as distance-based, vertex-degree-based, and spectrum-based topological indices (7).

Supervised learning has become a prominent approach in machine learning. It involves learning the input-output relationship of a system established from an input-output training sample. The input-output training sample is known as labeled training data or supervised data. This system is also called learning with a 'teacher' (8). The main

purpose of supervised learning is to develop an artificial system that can learn from examples such as the input-output training sample and make predictions of the output based on the input not given in the training set (test set) (9). Various techniques or algorithms exist for conducting supervised learning, including the Decision Tree algorithm, Random Forest (RF), Naïve Bayes algorithm (NB), Support Vector Machine (SVM) and artificial neural network (ANN) (10).

Modelling molecular structure-properties relationships using ANN and SVM as supervised learning starts with a selection topology index to represent the structure. This work aims to look into the applicability of the topology index to the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). Here, we implement a modification of the moment index. The moment topological index was initially introduced by Dalfo et al. (11). However, the definition of ρ for its further applications is not clearly stated. Chang et al. have applied this index to biphenyl and polycyclic hydrocarbons and extremal polyphenyl chains (12). In this paper, we define the new value of ρ-moment based on the degree of vertices. Using the neural network-Graph theory and Support Vector Machine-Graph theory approaches, the relationship between the moment topology and electronic properties has been a new understanding of electronic properties in alkanes.

2. TOPOLOGY INDEXES AS MOLECULAR DESCRIPTORS

2.1 Molecular Descriptor using Degree Indexes

The Quantitative Structure Properties Relationship (QSPR) model is an important tool for chemical and biological disciplines. It assists in analyzing the physical and chemical properties of molecular structure. Quantitative structure-property relationship analysis (QSPR) is a method that relates the chemical, biological, and physical activities of a molecular compound. The topological index is an example of a molecular descriptor that uses graph theory. Vertex points represent the atoms in the chemical structure. At the same time, the chemical bonding is described by the edge (13). The degree based on the vertex is the most widely employed as a chemical descriptor (14, 15). If *G*(*V, E*) is represented as the molecular graph with vertex and edge set, the connectivity index is given by Equation 1.

$$\chi(G)=\sum_{v,u\in V(G)}\frac{1}{\sqrt{d_v d_u}} \qquad \text{(Eq. 1)}$$

where d_v and d_u are the degrees of vertex *u* and *v,* respectively. Zhou and Trinajstić modified the connectivity index by replacing the multiplication product with the summation product (16). This index is also known as the sum connectivity index. The sum connectivity index is given by Equation 2.

$$\chi^{+}(G)=\sum_{v,u\in V(G)}\frac{1}{\sqrt{d_v+d_u}} \qquad \text{(Eq. 2)}$$

Additionally, Estrada has changed the connectivity index by taking into account the degree of the vertex and edge. The equation that describes this index, which is also known as the atom–bond connectivity (ABC), is as follows: (Equation 3) (17).

$$\chi^{\mathrm{ABC}}(G)=\sum_{v,u\in V(G)}\sqrt{\frac{d_v+d_u-2}{d_v d_u}} \qquad \text{(Eq. 3)}$$

The geometry-arithmetic index, or the GA index, is another vertex degree-based topological index. Its definition may be found as Equation 4 (18).

$$GA(G)=\sum_{v,u\in V(G)}\frac{\sqrt{d_v d_u}}{(d_v+d_u)/2} \qquad \text{(Eq. 4)}$$

where $\sqrt{d_v d_u}$ represents the geometry means, and the denominator $(d_v+d_u)/2$ represents the arithmetic mean of end-vertex degrees of the edge.

2.2 Molecular Descriptor using Distance Indexes

A distance-based molecular topology index is another method for analyzing topological molecular structures. The Wiener index is the earliest distance index that has been introduced. The Wiener index of graph G is defined as the sum of all distances between pairs of the graph's vertices given by Equation 5.

$$W(G)=\frac{1}{2}\sum_{v\in V(G)}d(u,v) \qquad \text{(Eq. 5)}$$

where d (u,v) is the shortest distance in G. The Wiener index has been improved, known as the hyper-Wiener index, and its definition is as follows: (Equation 6)(19)

$$WW(G)=\frac{1}{2}\sum_{v\in V(G)}\left(d(u,v)+\left[d(u,v)\right]^2\right) \qquad \text{(Eq. 6)}$$

The reciprocal of the distance between vertex *u* and *v* also has been introduced and defined in an Equation 7.

$$H(G)=\sum_{v\in V(G)}\frac{1}{d(u,v)} \qquad \text{(Eq. 7)}$$

The index in Equation 7 is also known as the Harary index (20). In endeavoring to relate graphic

structures with the chemical structure, Parikh word representable graphs (PWRGs) were developed (21). These graphs were based on the Wiener or Harary index calculations.

2. Three Combinations between Degree and Distance indexes

The degree and distance indexes can be combined into new topological indexes. The adjacency (*A*), degree (*v*), and distance matrices have been employed in the establishment of the Molecular Topological index (MTI index), which is based on matrix algebraic operations. The index is simplified using the following mathematical equation (22):

$$MTI = \sum v(D+A) \qquad \text{(Eq. 8)}$$

The Balaban index promises to be an extremely helpful molecular descriptor with appealing properties (23). Balaban index, *J = J* [*G* (*V, E*)], is calculated using the average-distance sum connectivity and defined as Equation 9:

$$J = q\sum_{ij} \frac{1}{\left(D_i D_j\right)^{1/2}} \qquad \text{(Eq. 9)}$$

where q is the number q of vertex adjacencies, and D_i is the distance sum of *G* (*V, E*) (23). While Ren has combined the distance and degree of the molecular graph to create a new index known as the Xu index (24). Xu index is defined as Equation 10.

$$Xu = n\log\left(\sum_{n}^{i} v_i s_i^2 / \sum_{n}^{i} v_i s_i\right) \qquad \text{(Eq. 10)}$$

where s_i is the distance sum of *G* (*V, E*) and v_i is the sum vertex-degree matrix of *G* (*V, E*).

3. METHOD OF CALCULATION

3.1 Topology Indices As The Input Data

3.1.1.Topology I: Moment Wiener Index

Dalfo et al. (11) have defined a moment topology index. In this paper, the value of *ρ* (moment constant) determines the weights between the vertices. We define a new moment topological index based on the degree of vertices. The moment Wiener indices is defined as follows:

$$DD1(G) = \frac{1}{2}\sum_{ij\hat{I}G} (u_i + u_j)(d_{ij}) \qquad \text{(Eq. 11a)}$$

$$DD2(G) = \frac{1}{2}\sum_{ij\hat{I}G} (u_i \cdot u_j)(d_{ij}) \qquad \text{(Eq. 11b)}$$

$$DD3(G) = \frac{1}{2}\sum_{ij\hat{I}G} |u_i - u_j|(d_{ij}) \qquad \text{(Eq. 11c)}$$

$$DD4(G) = \frac{1}{2}\sum_{ij\hat{I}G} \sqrt{u_i^2 + u_j^2}(d_{ij}) \qquad \text{(Eq. 11d)}$$

where d_{ij} is the shortest distance between vertices *i* and *j*. The numerical value of u_i and u_j is the degree of vertex *i* and *j*.

3.1.2 Topology II: Moment Harary indices

The Harary index of a graph *G (V, E)* is based on reciprocal distance and can be attained as the half-sum of all reciprocal distance elements (25). A new moment Harary indices is given by equation (12).

$$HH1(G) = \frac{1}{2}\sum_{ij\hat{I}G} (u_i + u_j)(d_{ij}^{-1}) \qquad \text{(Eq. 12a)}$$

$$HH2(G) = \sum_{ij\hat{I}G} (u_i \cdot u_j)(d_{ij}^{-1}) \qquad \text{(Eq. 12b)}$$

$$HH3(G) = \frac{1}{2}\sum_{ij\hat{I}G} |u_i - u_j|(d_{ij}^{-1}) \qquad \text{(Eq. 12c)}$$

$$HH4(G) = \frac{1}{2}\sum_{ij\hat{I}G} \sqrt{u_i^2 + u_j^2}(d_{ij}^{-1}) \qquad \text{(Eq. 12d)}$$

3.1.3 Topology III: Moment Balaban indices.

The Balaban index is also called the average-distance sum connectivity (23). The moment Balaban indices is defined as the moment of average-distance sum connectivity, that is:

$$JJ1(G) = q\sum_{ij\hat{I}G} \frac{1}{\sqrt{D1_i D1_j}} \qquad \text{(Eq. 13a)}$$

$$JJ2(G) = q\sum_{ij\hat{I}G} \frac{1}{\sqrt{D2_i D2_j}} \qquad \text{(Eq. 13b)}$$

$$JJ3(G) = q\sum_{ij\hat{I}G} \frac{1}{\sqrt{D3_i D3_j}} \qquad \text{(Eq. 13c)}$$

$$JJ4(G) = q\sum_{ij\hat{I}G} \frac{1}{\sqrt{D4_i D4_i}} \qquad \text{(Eq. 13d)}$$

where *q* is the number *q* of vertex adjacencies. The value of *D*1, *D*2, *D*3 and *D*4 correspond to the average row for the moment distance $(u_i + u_j)d_{ij}, (u_i \cdot u_j)d_{ij}, |u_i - u_j|d_{ij}, \text{and}\left(\sqrt{u_i^2 + u_j^2}\right)d_{ij}$,respectively. The computational algorithm for calculating Moment-Wiener, Harary and Balaban index is given in the appendix.

3.2 Machine learning modeling

In the supervised machine learning-Graph theory approach, the molecular descriptors were normalized according to the equation (14).

$$I_i = \frac{I_x - I_{min}}{I_{max} - I_{min}} \quad \text{(Eq. 14)}$$

where I_x unnormalized input data, I_{max} is the maximum value of the sample, and I_{min} is the minimum value of value of the sample. After that, the main dataset is split into training and test set. The main dataset is then subjected to a machine learning model: artificial neural network and support vector machine. The machine learning model performance was measured using RMSE, which reflects the model's absolute fit and how near the predicted values are to the actual data points. It provides an objective depiction of the model's predicted accuracy. RMSE was determined by applying the Equation 15.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y' - y)^2} \quad \text{(Eq. 15)}$$

Where *n* is the number of data, *y* is the experimental value, and y' is the prediction (calculated) value. The average relative error is calculated as the average of the prediction's absolute divergence from the actual value divided by the actual value.

$$\mathrm{RE} = \frac{|y' - y|}{y} \quad \text{(Eq. 16)}$$

The correlation coefficient *r* was computed using Equation 17:

$$r = \frac{n\sum(y \cdot y') - (\sum y)(\sum y')}{\sqrt{n\sum(y^2) - (\sum y)^2 - (\sum y')^2}} \quad \text{(Eq. 17)}$$

where *y* and *y'* is the experimental and predicted value, respectively.

3.2.1 Artificial neural network (ANN)

Artificial neural networks (ANN) are a machine-learning approach that the models of biological neural networks inspire. In an artificial neural network, the information processing element consists of several artificial neurons. In the present work, the feedforward forward neural network has been used. ANN utilizes supervised learning methods during the learning or training process. The learning process occurs when each target point is used in the training set. The architecture of this work is given in Figure 1. The architecture of ANN consists of an input layer, a hidden layer, a bias unit, and an output layer. The input layer is the input numerical data from the topological index. The hidden layer is the intermediate layer between the input and layer. The hidden layer analogy in an artificial neural network can be compared to a collection of neurons. The activation function used to train the ANN is applied to the hidden layer. In this study, we use the sigmoidal function 1 / (1 + e^{-x}) for this calculation. The bias units were attached to a hidden layer. The final layer of the ANN architecture is referred to as the output layer or output nodes. The conditional mean of output requires the knowledge of the joint probability density function of the random variables output and input layers. The learning rate in this calculation is 0.01.

3.2.2 Support Vector Machine (SVM)

The SVM method is a learning algorithm tool used for classification and regression. A non-linear function will transform the input data into high-dimensional feature space. Then, the samples will be separated by drawing a decision boundary (hyperplane) as a linear classifier (26). The linear classifier can be used to distinguish between "positive" and "negative" attributes from the independent variable. The training and test data are split using the split data operator. The training data set is exploited as a targeted point in the learning process. The type of kernel function parameter in this work uses the inner dot product. The machine learning calculation was calculated using Rapidminer Studio.

3.3 Molecular electronic properties (Learning Input)

The learning data or input is needed for the 'learning process' The learning and testing data consist of the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO). The electronic properties of alkanes HOMO and LUMO were calculated using the semi-empirical self-consistent molecular MOPAC 2016. The detailed method for this calculation can be found in the literature (27).

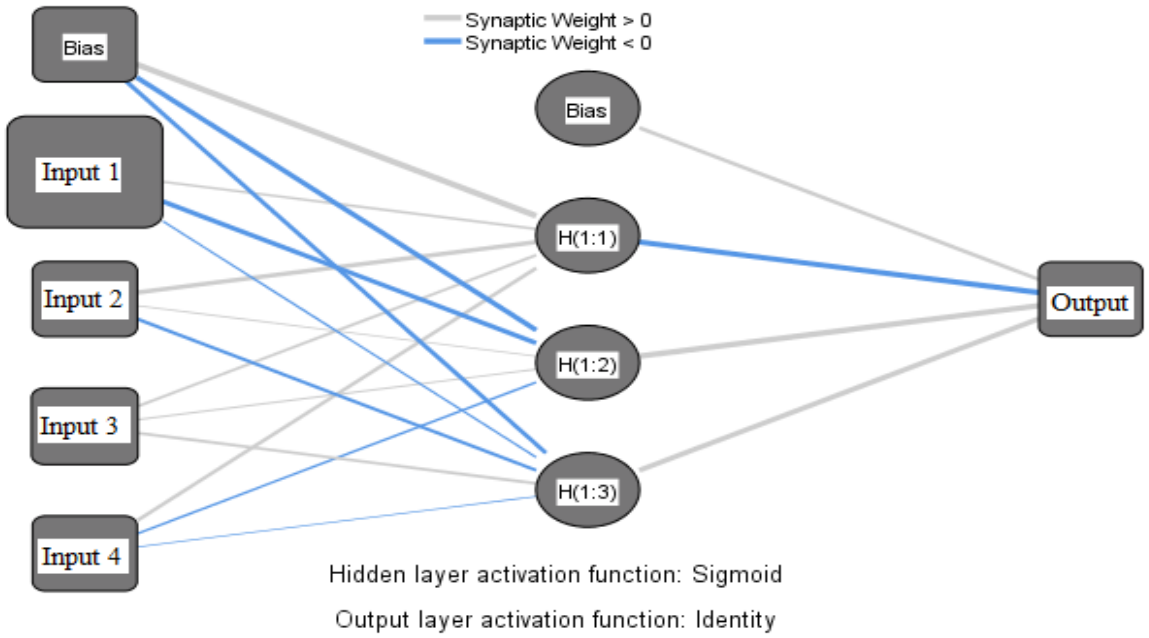


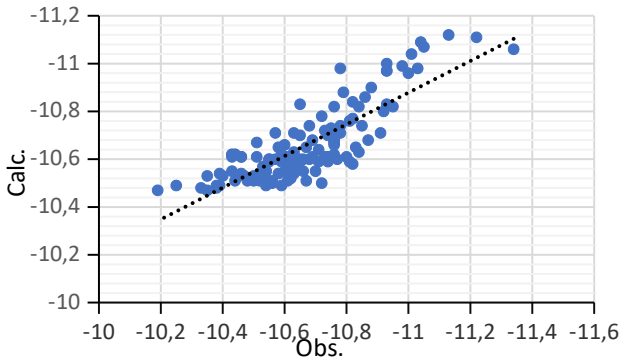
Figure 1. The architecture of ANN.

4. RESULTS

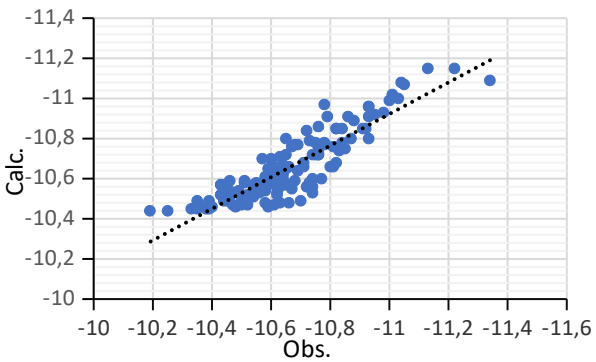
4.1 The artificial neural network-Graph theory approach

The neural network-Graph theory approach extracts complex patterns of molecular structure and relationships from large data sets to predict the electronic properties of alkanes. The moment topology index for 139 alkanes structure is given in Appendix B. The moment topology indices in 2D structural descriptors were more effective than in 3D descriptors (28). In 'normal' electronic calculation, the usual procedure is to find an orbital wave function suitable to the Schrödinger equation. In the supervised learning approach, the relation of molecular descriptors created a statistically optimized relationship with HOMO and LUMO. The relations of the moment topology index with the HOMO and LUMO energies as given in the supplementary table. The correlation plots for the calculated (predicted multilayer perceptron (MLP)) and observed combination topology index are given in Figure 2. The function of ANN is to extract classical quantum chemistry calculations to perform molecular orbital calculations efficiently, incorporating molecular position with the relation of the topology index.

To evaluate whether the accuracy of our models is sufficient for the electronic application, the data was split into training and test sets. Table 1 presents the accuracy results of ANN training and test sets for HOMO and LUMO. The highest root mean square error (RMSE) and relative error for HOMO are the moment Wiener topology indices. These were followed by the moment Harary and Balaban indices. The result also shows the same pattern for LUMO. This indicates that moment Balaban indices are the most stable descriptor for alkanes' electronic properties. This is plausible due to the moment Balaban indices as descriptors can explain the molecular orbital basis of the saturated hydrocarbon (29). Furthermore, adding molecules or atoms distorts the topology from the linear curve.



(a)



(b)

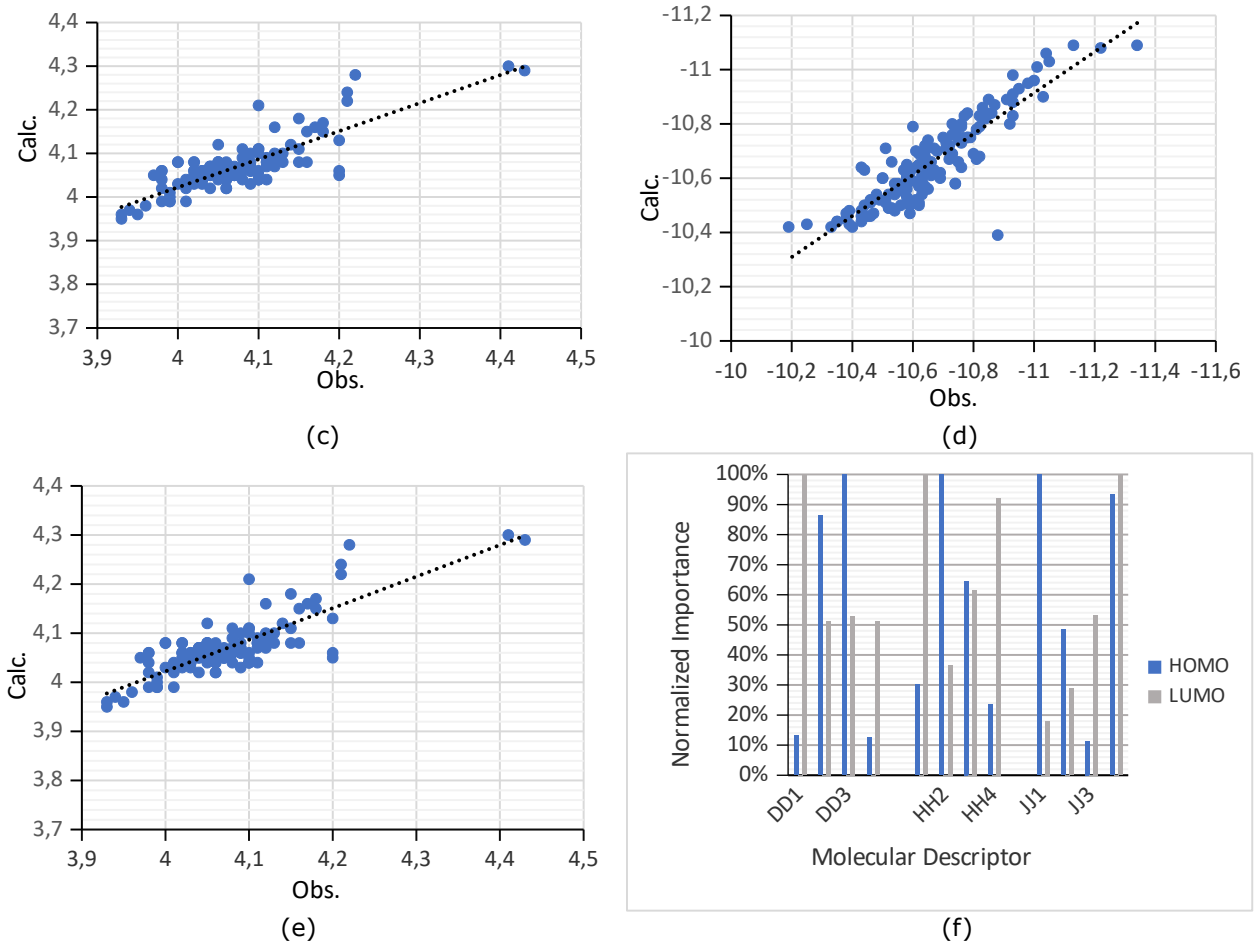


Figure 2: Correlation plots for observed versus calculated for (a) HOMO, (b) LUMO Moment Wiener Indices ; (c) HOMO, (d) LUMO Moment Harary Indices and (e) HOMO, (f) LUMO, Moment Balaban Indices respectively using ANN**.**

Table 1: Results of neural network modeling using moment Wiener, Harary, and Balaban indices.

	HOMO[a)]						LUMO					
	Training set			Test set			Training set			Test set		
	r	RMSE	RE	*r*	RMSE	RE	*r*	RMSE	re	*r*	RMSE	RE
Moment Wiener	0.785	0.122	0.920	0.778	0.117	0.910	0.749	0.054	0.96	0.663	0.057	1.17
Moment Harary	0.889	0.091	0.640	0.834	0.097	0.770	0.841	0.048	0.91	0.711	0.057	1.14
Moment Balaban	0.947	0.067	0.470	0.900	0.077	0.54	0.883	0.041	0.75	0.729	0.049	0.89

[a)]RMSE, Root mean square error and RE, relative error.

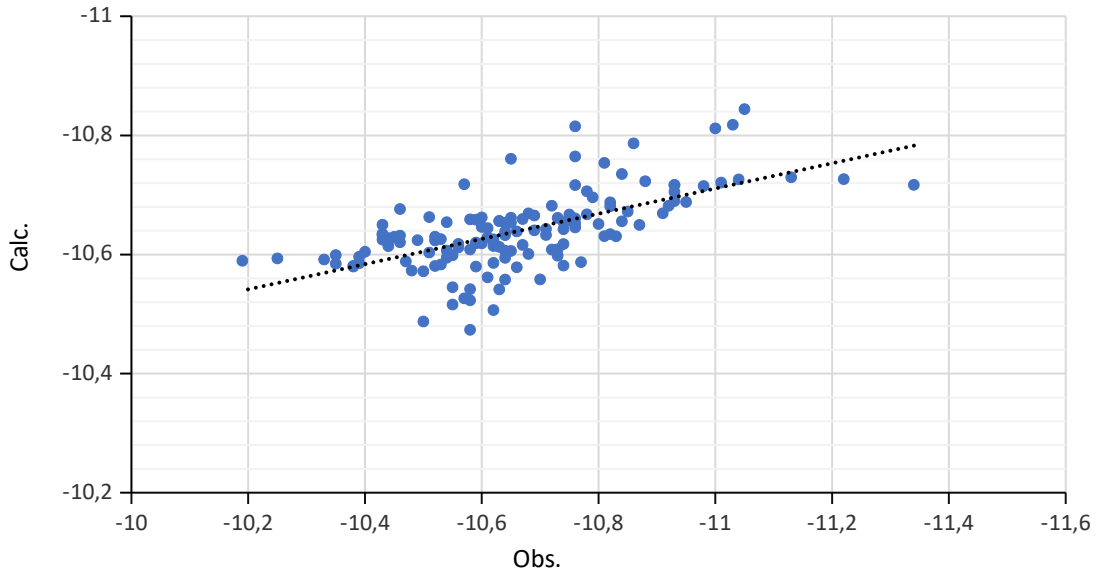


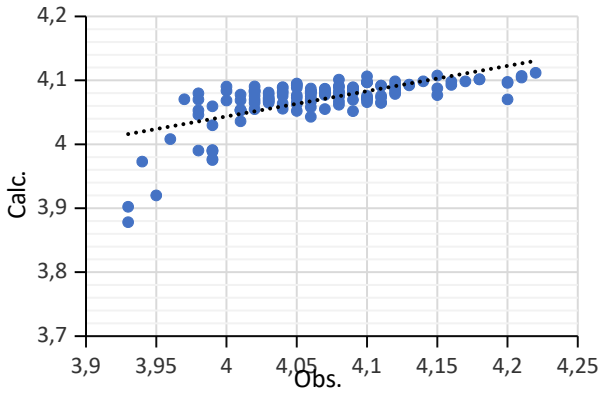
Figure 3: Graph of molecular descriptor versus independent variable importance.

Figure 3 shows each molecular descriptor's contribution to HOMO and LUMO energies. DD3 shows the highest value of HOMO, followed by DD2. However, DD1 and DD4 show a contribution of less than 15%. The contribution for DD1 and DD4 is 13.2% and 12.5%, respectively. This contradicts with LUMO energy, where DD1 contribute 100%. While DD2, DD3 and DD4 contribute more than 50%. In moment Harary indices, HH2 shows the highest contribution for HOMO, followed by HH2, HH3 and HH4. HH1 and HH4 significantly contribute more than 90% of LUMO energy. While HH3 contribute 61.6%. HH2 contributes less than 40%, which is 36.6%. Meanwhile, for the moment, Balaban indices JJ1 and JJ2 show a contribution of more than 90%. While JJ2 contributes 48.5% and JJ contributes 11.5%. JJ4 contributes 100% for LUMO energy, followed by JJ3 (53.3%), JJ2 (28.9%) and JJ1 (18%).

4.2 The Support Vector Machine-Graph Theory Approach

The Support Vector Machine-Graph theory approach is to model complex non-linear relationships of molecular structure from suitable kernel function. The model produced by the support vector depends on the subset of training data to predict the electronic properties of the alkanes molecule. Our model consists of a support vector machine task trained on electronic properties. The optimized relation of the molecular descriptor with HOMO and LUMO energy was created statistically via the applied test model. The relations of the moment topology index with the HOMO and LUMO energies as given in the supplementary table. The correlation plots for the calculated and observed HOMO and LUMO energy for the combination of topology indices is given in Figure 3. The correlation plots assess the performance of the molecular graph represented from the perspective of projection to a very high-dimensional space via the linear kernel classification. This modelling is capable of appertaining the molecular structure with electronic properties.

SVM training and applying the model resulted in models showing slightly higher prediction accuracy than the ANN (Table 2). The lowest root means square error (RMSE) and relative error (RE) for HOMO and LUMO is the moment Balaban topology indices. These are followed by the moment Harary and Wiener indices. This indicates that moment Balaban index shows the highest stability. The difference between the RMSE and RE values for each of these indices is due to the classification of the dataset via SVM from the molecular indices. The smallest value for all molecule descriptors indices indicates that the SVM can correlate with HOMO and LUMO energy.



(a)

$$GA(G)=\sum_{v,u\in V(G)}\frac{\sqrt{d_v d_u}}{(d_v+d_u)/2}$$

(b)

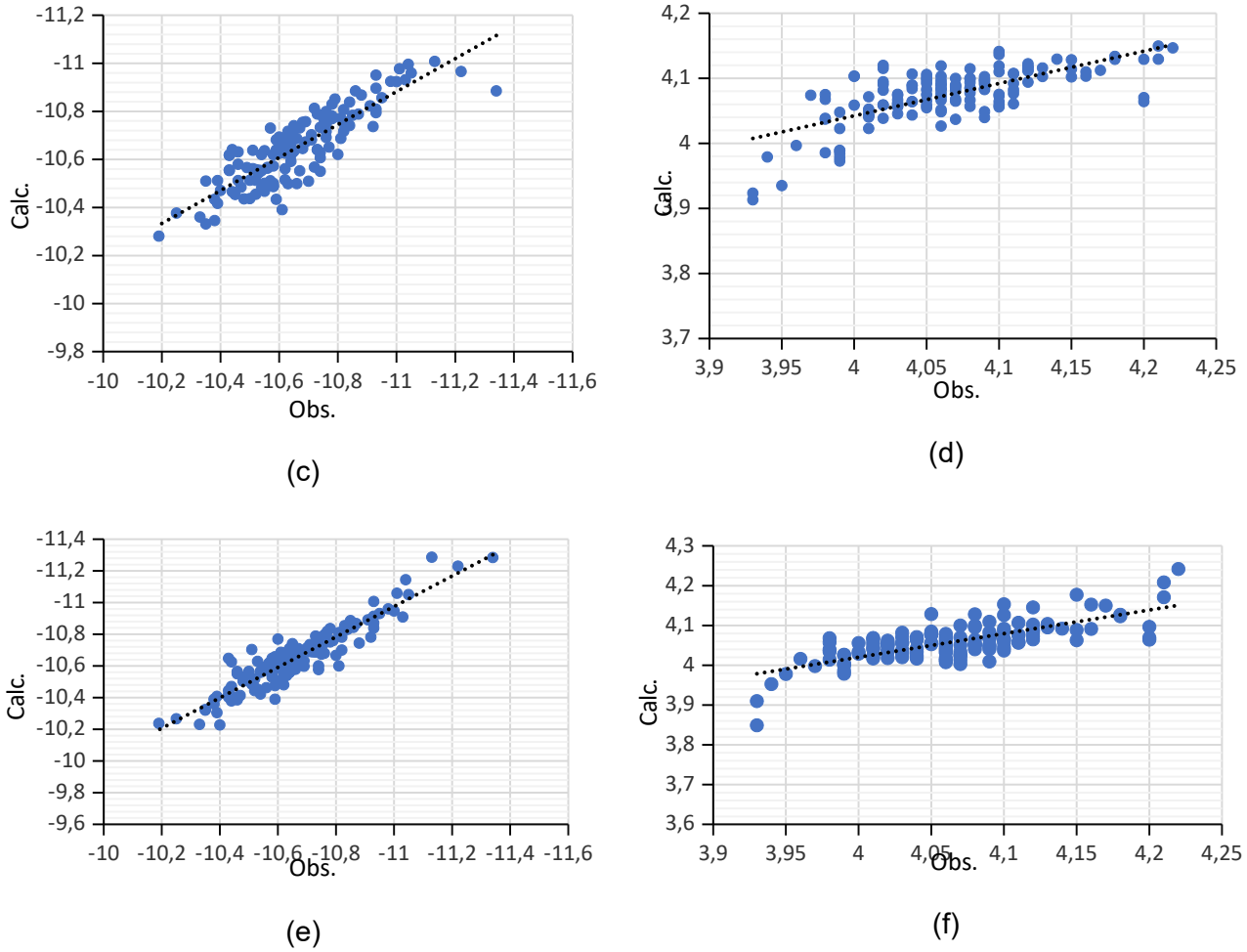


Figure 3: Correlation plots for observed versus calculated for (a) HOMO, (b) LUMO Moment Wiener Indices ; (c) HOMO, (d) LUMO Moment Harary Indices, and (e) HOMO, (f) LUMO Moment Balaban Indices respectively using SVM.

Table 2: Results of Support Vector Machine Using Moment Wiener, Harary, and Balaban Indices.

	HOMO[a)]						LUMO					
	Training set			Test set			Training set			Test set		
	r	RMSE	RE	*r*	RMSE	RE	*r*	RMSE	RE	*r*	RMSE	RE
Moment Wiener	0.69	0.14	1.03	0.641	0.139	0.019	0.640	0.061	0.95	0.703	0.046	0.92
Moment Harary	0.89	0.089	0.60	0.821	0.100	0.80	0.768	0.051	0.8	0.75	0.043	0.91
Moment Balaban	0.947	0.062	0.42	0.904	0.081	0.58	0.830	0.044	0.79	0.728	0.044	0.87

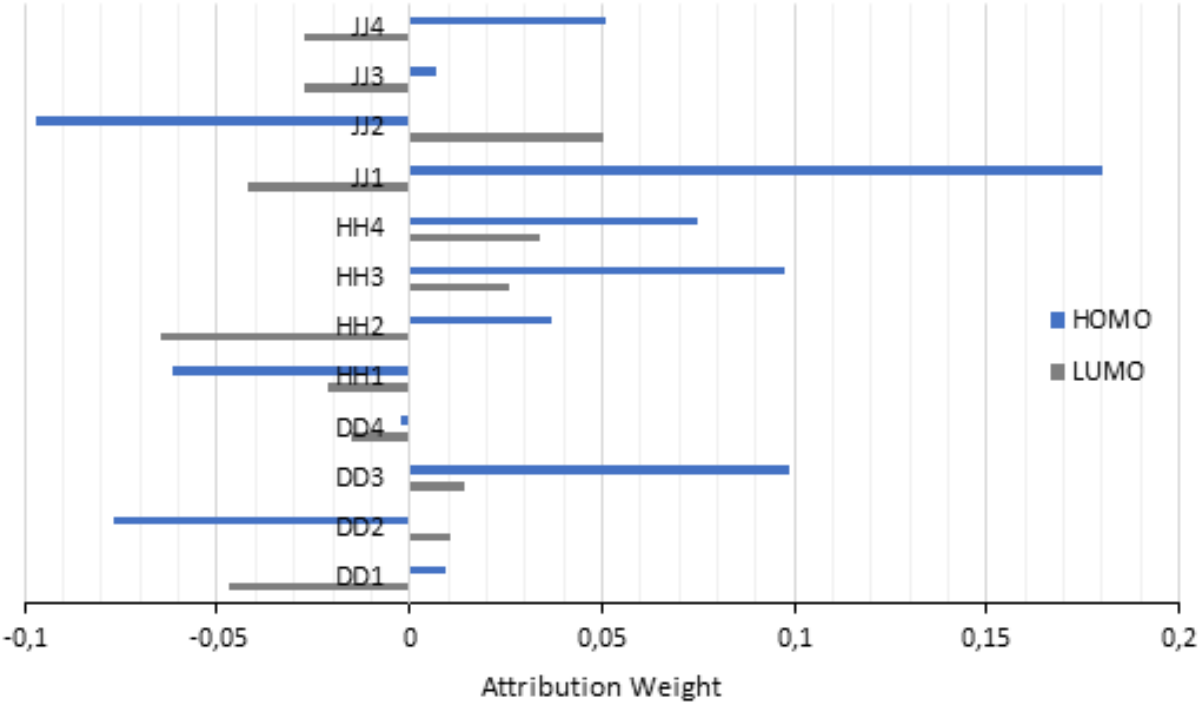


Figure 4: Molecular descriptor Independent variable attribution weight.

The graph of molecular descriptor independent variable attribution weight to HOMO and LUMO energy is shown in Figure 4. DD3 shows the positive and the highest value for the HOMO energy of moment Wiener. This is followed by DD1. DD2, and DD4 showing negative attribution weight. The positive attribution weight for LUMO energy is DD2 and DD3. At the same time, DD1 and DD4 have negative attribution weights. In moment Harary indices, HH2, HH3 and HH4 positively contribute to HOMO. Contrary to HH1, which shows negative attribution weight. HHH1 and HH2 show a negative contribution. Finally, Balaban indices JJ1, JJ2 and JJ3 show the positive attribution for HOMO energy and vice versa for JJ2. At the same time, only JJ2 shows positive attribution for LUMO energy. The variation of attribution weight indicators to classification (30). The effects of attribution weight are obtained from the similarity of the data features of the molecular topology indices.

5. DISCUSSION

The HOMO and LUMO energies reflect the electronic properties of saturated hydrocarbon of alkanes. Alkane is composed of that sigma (σ)-bonds which explain the behavior of electrons in the molecular structure. This bonding is responsible for C – H bond and basic framework C – C bonds. The electronic properties of the molecular structure suggest an interpretation in terms of the localization of electrons. Each molecule has its appropriate energy based on the position of atoms to form the molecules. In the topological approach, the indices are calculated from suppress-hydrogen, which can be related to the sharing of the electron-induced by the σ-bonds. The topology indices lead to the fundamental concept, which is well-known in chemistry, that is, the octet rule. The octet rule refers to the tendency of atoms to combine so that each atom has eight electrons in the valence shell. However, we need to consider the correlation of electrons in electronic calculation. Therefore, the 'moment' topological approach has been introduced with the value ρ weight between the vertices gives all considerations from the molecular calculation. In the moment topology approach, both properties of graph molecule, that is, degree and distance, have been considered. The distance and derivative of the distance matrix represented the molecule. While ρ weight of $u_i + u_j$, $u_i \cdot u_j$ and $|u_i - u_j|$ are analogies to σ-conjugation, orbital overlapping or Coulomb descriptor(31-33). In comparison, $\sqrt{(u_u^2+u_v^2)}$ is an analogy with elementary geometry (using Euclidean metrics) (7). The supervised learning can correlate between localized electrons through the whole molecule using topological indices as a molecular descriptor. In this work, the inductive supervise learning approach requires the functional relationship of topology indices that to be modeled. The supervised learning approach provides an improvement in molecular orbital calculation via the data splitting or partition in systematical modeling (learning experience by the machine) (34, 35). The sampling training data was applied to test the data set developing high accuracy in HOMO and LUMO energies (see Tables 1 and 2). SVM training resulted

in models showing slightly higher prediction accuracy than the ANN. The ANN system, in some cases, fails in non-linear classification. Therefore, in this finding, the chosen descriptors and learning system are vital; it is the set of how molecular has been assigned to be modeled with the topological orbital. The electronic properties of alkanes based on Zagreb and Sombor descriptors show low accuracy compares with current works (36). We also find that the Zagreb and Sombor indices are inadequate to assign as descriptors to the electronic structures of alkanes.

6. CONCLUSION

We have introduced a supervised learning model for predicting HOMO and LUMO energies of alkanes based on training artificial neural networks and support vector machines. A new moment topology approach has been introduced as molecular descriptors by taking consideration from the molecular structure perspective. The sampling training data and applied to test the data set, developing high accuracy in electronic properties. The low sum of square error and relative error shows the outperformance supervise learning modelling to the HOMO and LUMO energies of alkanes. SVM training resulted in models showing slightly higher prediction accuracy than the ANN systems. We also find that the chosen descriptors and learning system are of the vital importance, it is the set of how molecular has been assigned to be model with the topological orbital.

7. CONFLICT OF INTEREST

The authors declare no conflict of interest in this paper.

8. ACKNOWLEDGMENTS

The authors would like to thank Dr. James J. P. Stewart from MOPAC Inc. for his permission to use the MOPAC software.

9. APPENDIX

Algorithm for calculating Molecular topology index.

Step 1: Start
Step 2: Define distance Matrix D
Step 3: Calculate the adjacent Matrix A
  3.1 $a_{ij} = 1$ if $d_{i,j} = 1$
  3.2 $a_{ij} = 0$ if else
Step 4: Calculate the degree matrix $V \leftarrow \sum A^{<i>}$
Step 5: Calculate elements of k_i matrix
  5.1 $k_1 \leftarrow v_i + v_j$
  5.2 $k_2 \leftarrow v_i \cdot v_j$
  5.3 $k_3 \leftarrow |v_i - v_j|$
  5.4 $k_4 \leftarrow [v_i^2 + v_j^2]^{1/2}$
Step 6: Get Moment Wiener and Balaban Index
  6.1 Multiply elements k_{ij} with elements D_{ij}
  6.2 Sum all elements divided by 2
  6.3 Output DD1, DD2, DD3 and DD4 index
  6.4 Calculate the average-distance sum connectivity from DD1, DD2, DD3 and DD4 matrix
  6.5 Output JJ1, JJ2, JJ3 and JJ4 index
Step 7: Define inverse matrix H
  7.1 $h_{ij} = 1/ d_{ij}$
  7.2 $h_{ii} = d_{ii}$
Step 8: Get Moment Harary Index
  8.1 Multiply elements k_{ij} with elements H_{ij}
  8.2 Sum all elements divided by 2
  8.3 Output HH1, HH2, HH3 and HH4 index
Step 9: Stop

10. REFERENCES

1. Takata M, Lin BL, Xue M, Zushi Y, Terada A, Hosomi M. Predicting the acute ecotoxicity of chemical substances by machine learning using graph theory. Chemosphere. 2020 Jan;238:124604. Available from: <URL>.

2. Tamilarasi C, others. QSPR analysis of novel indices with priority polycyclic aromatic hydrocarbons (PAHs). Turkish Journal of Computer and Mathematics Education (TURCOMAT). 2021;12(10):3992–9.

3. Kirmani SAK, Ali P, Azam F. Topological indices and QSPR / QSAR analysis of some antiviral drugs being investigated for the treatment of COVID -19 patients. Int J of Quantum Chemistry. 2021 May 5;121(9):e26594. Available from: <URL>.

4. Sporns O. Graph theory methods: applications in brain networks. Dialogues in Clinical Neuroscience. 2018 Jun 30;20(2):111–21. Available from: <URL>.

5. Randić M. Novel molecular descriptor for structure—property studies. Chemical Physics Letters. 1993 Aug;211(4–5):478–83. Available from: <URL>.

6. Boguñá M, Bonamassa I, De Domenico M, Havlin S, Krioukov D, Serrano MÁ. Network geometry. Nature Reviews Physics. 2021 Jan 29;3(2):114–35. Available from: <URL>.

7. Rada J, Rodríguez JM, Sigarreta JM. General properties on Sombor indices. Discrete Applied Mathematics. 2021 Aug;299:87–97. Available from: <URL>.

8. Van Engelen JE, Hoos HH. A survey on semi-supervised learning. Machine Learning. 2020 Feb;109(2):373–440. Available from: <URL>.

9. Prezhdo OV. Advancing Physical Chemistry with Machine Learning. Journal of Physical Chemistry Letters. 2020 Nov 19;11(22):9656–8. Available from: <URL>.

10. Pavithra M, Kumar PP, Divya P, Manjubala P, Jayalakshmi S. The significance of learning in data analytics: Supervised learning techniques. Global Journal of Internet Interventions and IT Fusion. 2021;4(1–2021).

11. Dalfó C, Fiol MA, Garriga E. Moments in graphs. Discrete Applied Mathematics. 2013 Apr;161(6):768–77. Available from: <URL>.

12. Chang C, Ren H, Deng Z, Deng B. The ρ --moments of vertex-weighted graphs. Applied Mathematics and Computation. 2021 Jul;400:126070. Available from: <URL>.

13. Cao J, Ali U, Javaid M, Huang C. Zagreb connection indices of molecular graphs based on operations. Complexity. 2020 Mar 30;2020:1–15. Available from: <URL>.

14. Chu YM, Julietraja K, Venugopal P, Siddiqui MK, Prabhu S. Degree- and irregularity-based molecular descriptors for benzenoid systems. European Physical Journal Plus. 2021 Jan;136(1):78. Available from: <URL>.

15. Kumar KA, Basavarajappa N, Shanmukha M. QSPR analysis of alkanes with certain degree based topological indices. Malaya Journal of Mathematik. 2020;8(1):314–30.

16. Zhou B, Trinajstić N. On a novel connectivity index. Journal of Mathematical Chemistry. 2009 Nov;46(4):1252–70. Available from: <URL>.

17. Estrada E, Torres L, Rodriguez L, Gutman I. An atom-bond connectivity index: modelling the enthalpy of formation of alkanes. Indian Journal of Chemistry. 1998;37:849-55. Available from: <URL>.

18. Vukičević D, Furtula B. Topological index based on the ratios of geometrical and arithmetical means of end-vertex degrees of edges. Journal of Mathematical Chemistry. 2009 Nov;46(4):1369–76. Available from: <URL>.

19. Shahni Karamzadeh N, Darafsheh MR. Topological Indices of Certain Graphs. Iranian Journal of Mathematical Chemistry [Internet]. 2022 Sep [cited 2023 Nov 27];13(3): 167-74. Available from: <URL>.

20. Alqesmah A, Alloush KAA, Saleh A, Deepak G. Entire Harary index of graphs. Journal of Discrete Mathematical Sciences and Cryptography. 2022 Nov 17;25(8):2629–43. Available from: <URL>.

21. Thomas N, Mathew L, Sriram S, Nagar AK, Subramanian KG. Certain Distance-Based Topological Indices of Parikh Word Representable Graphs. Cangul IN, editor. Journal of Mathematics. 2021 May 25;2021:1–7. <URL>.

22. Gutman I. On degree-and-distance-based topological indices. 2021;66(2):119-23.

23. Das KC. On the Balaban index of chain graphs. Bulletin of the Malaysian Mathematical Sciences Society. 2021;44:2123–38.

24. Ren B. A New Topological Index for QSPR of Alkanes. Journal of Chemical Information and Computer Sciences. 1999 Jan 25;39(1):139–43. Available from: <URL>.

25. Zhou B, Cai X, Trinajstić N. On Harary index. Journal of Mathematical Chemistry. 2008 Aug;44(2):611–8. Available from: <URL>.

26. Zhou ZH. Support Vector Machine. In: Machine Learning [Internet]. Singapore: Springer Singapore; 2021 [cited 2023 Nov 27]. p. 129–53. Available from: <URL>.

27. Alias AN, Zabidi ZM, Zakaria NA, Mahmud ZS, Ali R. Biological Activity Relationship of Cyclic and Noncyclic Alkanes Using Quantum Molecular Descriptors. Open Journal of Applied Sciences. 2021;11(08):966–84. Available from: <URL>.

28. Brown RD, Martin YC. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. Journal of Chemical Information and Computer Sciences. 1997 Jan 1;37(1):1–9. Available from: <URL>.

29. Herndon WC, Ellzey ML, Raghuveer KS. Topological orbitals, graph theory, and ionization potentials of saturated hydrocarbons. Journal of the American Chemical Society. 1978 Apr;100(9):2645–50. Available from: <URL>.

30. Liu Z, Shao J, Xu W, Meng Y. Prediction of rock burst classification using the technique of cloud models with attribution weight. Natural Hazards. 2013 Sep;68(2):549–68. Available from: <URL>.

31. Woon KL, Chong ZX, Ariffin A, Chan CS. Relating molecular descriptors to frontier orbital energy levels, singlet and triplet excited states of fused tricyclics using machine learning. Journal of Molecular Graphics and Modelling. 2021 Jun;105:107891. Available from: <URL>.

32. Chou SH, Voss J, Bargatin I, Vojvodic A, Howe RT, Abild-Pedersen F. An orbital-overlap model for minimal work functions of cesiated metal surfaces. Journal of Physics: Condensed Matter. 2012 Nov 7;24(44):445007. Available from: <URL>.

33. Dewar MJS. σ-Conjugation and σ-Aromaticity. Bulletin des Soc Chimique. 1979 Jan;88(12):957–67. Available from: <URL>.

34. Li Z, Omidvar N, Chin WS, Robb E, Morris A, Achenie L, et al. Machine-Learning Energy Gaps of Porphyrins with Molecular Graph Representations. Journal of Physical Chemistry A. 2018 May 10;122(18):4571–8. Available from: <URL>.

35. von Lilienfeld OA. Quantum Machine Learning in Chemical Compound Space. Angewandte Chemie International Edition. 2018 Apr 9;57(16):4164–9. Available from: <URL>.

36. Zabidi ZM, Alias AN, Nurul AZ, Zaidatul SM, Rosliza A, Muhamad KY. Machine Learning Predictor Models in the Electronic Properties of Alkanes Based on Degree Topology Indices. Unpublished work. 2021;N/A.