https://doi.org/10.26565/2220-637X-2019-32-02

УДК 544.169+519.237.5

A NEW APPROACH IN TOPOLOGICAL DESCRIPTORS USAGE. ITERATED LINE GRAPHS IN THE THEORETICAL PREDICTION OF PHYSICO-CHEMICAL PROPERTIES OF SATURATED HYDROCARBONS

A.B. Zakharov[a] and V.V. Ivanov[b]

V.N. Karazin Kharkiv National University, School of Chemistry, 4 Svobody sqr., 61022 Kharkiv, Ukraine

a) ✉ abzakharov@karazin.ua, https://orcid.org/0000-0003-1577-7261
b) ✉ vivanov@karazin.ua, https://orcid.org/0000-0002-9120-8469

A new look on the problem of the molecular systems index description is presented. The capabilities of iterated line (edge) graphs in characterization of saturated hydrocarbons properties were investigated. It was demonstrated that single selected molecular (graph-theoretical (topological) or informational) descriptor calculated for the sequence of nested line graphs provides quite reliable progressive set of regression equations. Hence, the problem of descriptor set reduction is solved in the presented approach at list partially. Corresponding program complex (QUASAR) has been implemented with Python 3 program language. As the test example physico-chemical properties of octane isomers have been chosen. Among the properties under investigation there are boiling point, critical temperature, critical pressure, enthalpy of vaporization, enthalpy of formation, surface tension and viscosity. The corresponding rather simple linear regression equations which include one, two or three parameters correspondingly have been obtained. The predictive ability of the equations has been investigated using internal validation tests. The test by *leave-one-out* (*LOO*) validation and Y-scrambling evaluate the obtained equations as adequate. For instance, for the regression model for boiling point the best equation characterizes by determination coefficients $R^2 = 0.943$, with *LOO* procedure – $Q^2 = 0.918$, while for the Y-scrambling test $Q^2_{Y-scr} < 0.3$ basically.

It is shown that all the abovementioned molecular properties in iterated line graph approach can be effectively described by commonly used topological indices. Namely almost every randomly selected topological index can give adequate equation. Effectiveness is demonstrated on the example of Zagreb group indices. Also essential effectiveness and rather universal applicability of the so-called “forgotten” index (ZM3) was demonstrated.

Keywords: topological descriptors, line (edge) graph, regression analysis, determination coefficient, leave-one-out cross-validation, Y-scrambling, “forgotten” index.

Introduction

Development and investigation of new materials are strongly connected with building of corresponding mathematical models for target properties. Such a model can be based on either rigorous physical conception (*e.g.* quantum theory, statistical physics) or statistical (chemoinformatics) interpretation of available experimental data. The latter is usually based on the formal description of the molecular structure with large numbers of molecular parameters – descriptors. Such parameters describe different aspects of molecular system. Among them the physico-chemical data (lipophilicity, refractivity, *etc*) or pure mathematical values which are not connected directly with observed molecular properties. The subsequent usage of wide arsenal of statistical and mathematical methods provides possibility to obtain corresponding equations for prediction of desired properties or make a classification of molecular system according to certain criterion. In general, such tasks designated by widely known acronym QSAR – quantity structure-activity relationships (QSPR – quantity structure-property relationships) [1,2].

The central QSAR problem is the selection of minimal set of descriptors which guarantee reliable (adequate) description of desired properties. Nowadays for such selection it is worth mentioning factor analysis and different methods based on regularization technique: LASSO – Least Absolute Selection and Shrinkage Operator, LARS – Least Angle Regression and Shrinkage, *etc.* [3]. However, today the problem of suitable descriptors selection is still the one of the biggest problems in QSAR. The problem is essentially connected with the large number of available descriptors. For instance, in the popular computational QSAR software DRAGON [4] there are more then five thousand descriptors.

During the long history of QSAR investigations, the large set of so-called topological descriptors (TDs) based at chemical graph theory and information theory has been developed. Since the first Wiener index [5] (1947 !) thousands of indices were proposed for description of different molecular properties for various classes of chemical compounds. Among them popular Randić index [6] and corresponding set of generalizations [7], large number of theoretically-informational indices [8,9], so-called Zagreb group indices [10,11] *etc*. For general description of TDs see reviews [8,9,12].

It should be emphasized that development of brand new TDs is connected not only with pure mathematical fancy but also with necessity to investigate the properties which cannot be described by using compact (small enough) set of known descriptors. For instance, important hydrocarbon property – octane numbers (ON) still cannot be described with rather simple, low parametric equation based on TDs.

Recently we proposed a new graph-theoretical approach based on nested chains of line graphs [13]. Namely we build regular molecular (vertex) graph ($G^{(0)}$), then we build line (edge) graph[*)] ($G^{(1)}$), and then a sequence of graphs where each next graph is line graph for the previous one ($G^{(2)}$, $G^{(3)}$, …, $G^{(N)}$). Subsequent calculation of chosen descriptor for vertex and all line graphs of molecule forming a predictor set for regression analysis. Effectiveness of our approach was demonstrated in description of ON for saturated hydrocarbons. In the presented article we continue the investigation of line graph concept in QSAR/QSPR modeling. As the example we use the series of different physico-chemical properties of octane isomers [14].

Line graphs in regression model building

According to our approach, for the molecular system with vertex graph $G^{(0)}$ the iterative construction of a line graph sequence can be described symbolically in the following way:

$$G^0 = V(mol), \tag{1}$$

$$G^{(1)} = E\left(G^{(0)}\right) = E\left(V\left(mol\right)\right), \tag{2}$$

$$G^{(2)} = E\left(G^{(1)}\right) = E\left(E\left(V\left(mol\right)\right)\right), \tag{3}$$

$$G^{(N)} = E\left(G^{(N-1)}\right) = E\left(...V\left(mol\right)\right). \tag{4}$$

Here by $V(mol)$ we designate procedure of molecular vertex graph building, while $G^{(k+1)} = E\left(G^{(k)}\right)$ corresponds to building of line graph from previous one.

Adjacency matrix for such a sequence can be easily calculated using well-known matrix expression:

$$A_{k+1} = B_k^+ B_k - 2I, \tag{5}$$

where A_{k+1} is adjacency matrix for graph $G^{(k+1)}$, B_k – is the incidence matrix of current $G^{(k)}$ graph, and I is the identity matrix.

As an example the graph sequence for methyl derivatives of cyclopropane is presented in Figure 1. From the picture one can notice that while first line graph ($G^{(1)}$) describes the connections between edges through the vertexes, the next line graph ($G^{(2)}$) describes participation of vertices in connection between edges, *etc.* For the first employment of the line graphs in chemometrics see [15-17].

In the present article we use the sequence of graphs for building QSAR models namely regression equations. In particular the regression model of target property (Y) according to our approach is represented with the equation of the following form:

$$Y = a + a_0 X^{(0)} + a_1 X^{(1)} + a_2 X^{(2)} + ... = a + \sum_{i=0}^{n} a_i X^{(i)}, \tag{6}$$

where $X^{(0)}$, $X^{(1)}$, $X^{(2)}$, … are values of the selected descriptor for graphs $G^{(0)}$, $G^{(1)}$, $G^{(2)}$, … correspondently, n is the number of parameters.

---

*) *In contemporary mathematical literature for the edge graph there are several terms. Among them the covering graph, the edge-to-vertex dual, the interchange graph, the adjoint graph,* etc. *In the present article we are using probably most popular among them – the line graph.*

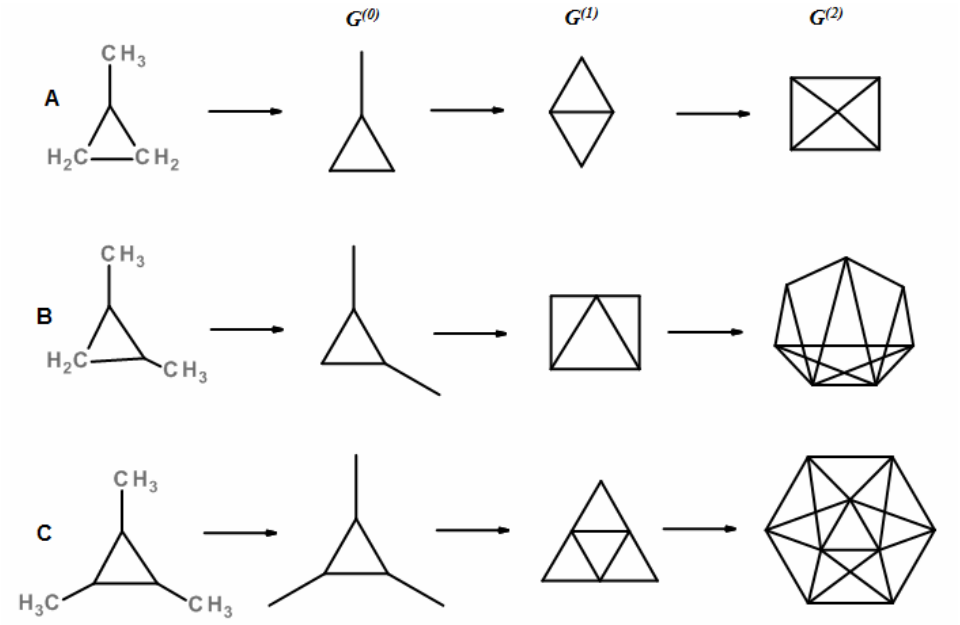

Figure 1. Methylcyclopropane (A), 1,2–dimethylcyclopropane (B), 1,2,3–trimethylcyclopropane (C), their vertex graphs ($G^{(0)}$) and line graphs ($G^{(1)}$, $G^{(2)}$).

To evaluate the prognostic ability of obtained equations we use standard coefficients of determination R^2 and corresponding values obtained via well known leave-one-out (*LOO*) cross-validation procedure Q^2:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad Q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_{i/i})^2}{\sum_i (y_i - \bar{y})^2}. \tag{7}$$

where y_i is experimental, $\hat{y}_i$ calculated and $\hat{y}_{i/i}$ predicted values via *LOO* procedure for i^{th} molecule. $\bar{y}$ is mean value for the training sample.

In the present article we describe regression models of different properties for the set of octane isomers as a test example. The rather small set of isomer molecules (18 molecules, see appendix Table A1) is difficult problem for regression model building. We choose to study seven properties of the isomers (Table 1).

As the descriptor set we use more than 30 indices from different types of topological and informational descriptors. From the large set of obtained successful equations here we will describe the results given by several selected indices (Table 2). Especially the group of Zagreb indices attracted our attention since the latter systematically use vertex degree graph concept. Among them the so-called "forgotten" index, $ZM3$, which is of great interest in contemporary literature [11].

Table 1. Physico-chemical properties of octane isomers under QSAR investigation

	property	designation	dimension
1	boiling point	*BP*	K
2	critical pressure	*CPr*	Pa
3	critical temperature	*CT*	K
4	enthalpy of vaporization	*HV*	KJ/Mol
5	enthalpy of formation	*HF*	KJ/Mol
6	surface tension	*ST*	N/m
7	viscosity	*Vs*	Pa·s

For all indices the set of equations (6) has been obtained for n = 0, 1, 2 (due to small training set the value n is restricted to two). Information about predictive ability of regression equations based on *ZM1, ZM2,* and *ZM3* indices is collected in the Table 3.

Table 2. Most important molecular descriptors in the present investigation. In this table d_i is i^{th} vertex degree, t_{ij} is graph distance between i and j vertices, N is total number of vertices in corresponding graph.

	descriptor	designation
1	First Zagreb Index	$ZM1 = \sum_i d_i^2$
2	Second Zagreb Index	$ZM2 = \sum_{(i,j)} d_i d_j$
3	"Forgotten" Index	$ZM3 = \sum_i d_i^3$
4	Sum of logarithms of row for distance matrix	$LPRS = \sum_i \log \sum_j t_{ij}$
5	Normalized quadratic index	$Qind = 3 - 2N + ZM1/2$

It is worth noting that the equations based on "forgotten" index (*ZM3*) demonstrate good prognostic ability for all presented cases. The equations for best models where R^2 and Q^2 are maximal and $R^2 \sim Q^2$ enumerated below.

$$BP = 436.14 - 0.971 \cdot ZM3^{(0)} + 0.197 \cdot ZM3^{(1)} \quad (8)$$

$$CPr \cdot 10^{-3} = 2856 - 13.6 \cdot ZM3^{(0)} + 7.58 \cdot ZM3^{(1)} - 0.431 \cdot ZM3^{(2)} \quad (9)$$

$$CT = 627.2 - 1.532 \cdot ZM3^{(0)} + 0.396 \cdot ZM3^{(1)} \quad (10)$$

$$HV = 48.86 - 0.514 \cdot ZM1^{(1)} - 0.048 \cdot ZM1^{(1)} + 0.017 \cdot ZM1^{(2)} \quad (11)$$

$$HF = -180.37 - 0.757 \cdot ZM3^{(0)} + 0.195 \cdot ZM3^{(1)} - 0.007 \cdot ZM3^{(2)} \quad (12)$$

$$ST \cdot 10^6 = 2.733 - 1.737 \cdot 10^{-2} \cdot ZM3^{(0)} + 5.526 \cdot 10^{-3} \cdot ZM3^{(1)} - 1.916 \cdot 10^{-4} \cdot ZM3^{(2)} \quad (13)$$

$$Vs \cdot 10^5 = 4.739 + 3.794 \cdot 10^{-2} \cdot ZM2^{(0)} \quad (14)$$

As an alternative to Eqs (9-12, 14) we demonstrate below several additional equations which are the best, based on *LPRS* and *Qind* (see Table 2) along with corresponding determination coefficients.

$$C\,\mathrm{Pr} \cdot 10^{-6} = -9.681 + 1.415 \cdot LPRS^{(0)} - 1.133 \cdot LPRS^{(1)} + 0.01552 \cdot LPRS^{(2)},\ R^2 = 0.980,\ Q^2 = 0.973 \quad (15)$$

$$CT = -1657 + 227.7 \cdot LPRS^{(0)} - 168.3 \cdot LPRS^{(1)} - 2.605 \cdot LPRS^{(2)},\quad R^2 = 0.930,\ Q^2 = 0.885 \quad (16)$$

$$HV = 34.73 - 0.9631 \cdot Qind^{(0)} - 0.09515 \cdot Qind^{(1)} + 0.03312 \cdot Qind^{(2)},\quad R^2 = 0.950,\ Q^2 = 0.913 \quad (17)$$

$$HF = -770.0 + 59.13 \cdot LPRS^{(0)} - 43.96 \cdot LPRS^{(1)} - 1.554 \cdot LPRS^{(2)},\quad R^2 = 0.926,\ Q^2 = 0.877 \quad (18)$$

$$Vs \cdot 10^6 = 226.5 - 13.92 \cdot LPRS^{(0)} + 8.502 \cdot LPRS^{(1)} + 0.1045 \cdot LPRS^{(2)},\quad R^2 = 0.987,\ Q^2 = 0.951 \quad (19)$$

It should be stressed that in most cases only complete chain of descriptors shows satisfactory equation, i.e. if $n = 2$ and for example $G^{(1)}$ is eliminated, the given solution will result in worse prognostic ability. Corresponding determination coefficients for the model based at *ZM3* index are given in the Table 4.

As soon as the training set is quite small we can not select a set for testing. Another way to evaluate effectiveness of obtained equations (except *LOO* procedure) is internal Y-scrambling test. This test is based on the random permutations of Y-column (target property) without corresponding transposition of predictors. Comparison of determination coefficients from Table 3 (and Table 4) with those obtained via Y-scrambling test gives information about causality effects in the regression models. For the above mentioned models (7-13) we obtained pretty close results for Y-scrambling test [18]. Thus we will not describe it completely but only for single general case. Namely for the Eq. (7) (see also corresponding row in Table 3) thousand times Y-scrambling procedure was performed. In case when $n = 1$, 98.8 % of random samples have *LOO* value $Q^2_{Y-scr} < 0.3$ while for the case when $n = 2$, – 98.7 %.

These values significantly less then corresponding data from Table 3 (*i.e. ZM3*, $n = 2$, $Q^2 = 0.799 > Q^2_{Y-scr}$). Hence the chosen model can be treated as statistically adequate.

Table 3. Determination coefficients (first value is R^2, second value is Q^2) for different regression models based on Zagreb indices with n parameters in equation (6).
The best validation parameters are given in bold.

	ZM1			*ZM2*			*ZM3*		
n	0	1	2	0	1	2	0	1	2
BP	0.519	0.889	0.901	0.251	0.368	0.728	0.497	**0.943**	0.945
	0.373	0.838	0.833	0.087	0.032	-0.167	0.341	**0.918**	0.799
CPr	0.524	0.839	0.862	0.817	0.885	0.934	0.510	0.894	**0.976**
	0.429	0.791	0.738	0.780	0.831	0.803	0.411	0.815	**0.922**
CT	0.000	0.795	0.795	0.083	0.233	0.653	0.0	**0.938**	0.953
	-0.256	0.727	0.714	-0.120	-0.171	-0.425	-0.258	**0.885**	0.871
HV	0.810	0.924	**0.950**	0.587	0.637	0.859	0.782	0.925	0.947
	0.727	0.865	**0.913**	0.498	0.434	0.478	0.686	0.888	0.872
HF	0.631	0.846	0.851	0.340	0.559	0.723	0.629	0.904	**0.916**
	0.565	0.791	0.774	0.208	0.356	0.211	0.557	0.774	**0.793**
ST	0.081	0.797	0.800	0.001	0.265	0.650	0.078	0.946	**0.980**
	-0.137	0.730	0.685	-0.195	-0.099	-0.304	-0.146	0.893	**0.962**
Vs	0.848	0.849	0.925	**0.874**	0.882	0.945	0.805	0.807	0.919
	0.776	0.757	0.658	**0.811**	0.796	0.768	0.724	0.697	0.609

Table 4. Demonstration of determination coefficients decay when incomplete sequence of graphs is employed for *ZM3* index example.

	$G^{(0)}, G^{(1)}, G^{(2)}$		$G^{(0)}, G^{(2)}$		$G^{(1)}, G^{(2)}$	
	R^2	Q^2	R^2	Q^2	R^2	Q^2
BP	0.945	0.799	0.895	0.707	0.450	-0.466
CPr	0.976	0.922	0.719	0.416	0.717	0.578
CT	0.953	0.871	0.712	0.235	0.176	-0.893
HV	0.947	0.872	0.946	0.913	0.747	0.357
HF	0.916	0.793	0.821	0.605	0.463	-0.061
ST	0.980	0.962	0.693	0.219	0.091	-0.883
Vs	0.919	0.609	0.814	0.691	0.919	0.686

Conclusion

In the present article we demonstrated ability of iterated line graphs approach in building of QSAR regression equations. In contrary to standard approach, where the combination of descriptors has to be generated by different statistical approaches (like factor analysis, *etc*), we use simple stepwise approach for single selected descriptor. Subsequently we calculate the nested line graphs, the chosen descriptor for it, and then corresponding regression equation. The simple comparison of determination coefficients allows to identify the best equation, and evaluate its prognostic ability.

Another aspect of the article concerned to so-called "forgotten" index, $ZM3$. It was observed before that usually $ZM3$, can not give good predictive ability itself however in combination with other indices it can give quite adequate equation [11,19]. We argued with this issue and demonstrated that $ZM3$ calculated for a sequence vertex and line graphs gives regression equation with good yet rather universal predictivity.

Acknowledgment

The work was performed as part of a research project of the Ministry of Education and Science of Ukraine.

Appendix

Table A1. Physico-chemical properties of octane isomers
(acronyms according to the Table 1)

name	*BP*	*CPr*	*CT*	*HV*	*HF*	*ST*	*Vs*
	K	10^{-6} Pa	K	KJ/mol	KJ/mol	10^{6} N/m	10^{5} Pa·s
2,2-dimethylhexane	379.99	2.53	549.8	32.37	-224.6	1.92	5.91
2,3-dimethylhexane	388.76	2.63	563.4	33.18	-213.8	2.05	5.89
2,4-dimethylhexane	382.58	2.56	553.5	32.48	-219.2	1.96	5.91
2,5-dimethylhexane	382.26	2.49	550.0	32.73	-222.5	1.93	5.84
3,3-dimethylhexane	385.12	2.65	562.0	32.64	-220.0	2.02	5.94
3,4-dimethylhexane	390.88	2.69	568.8	33.31	-212.7	2.12	5.91
3-ethyl-2-methylpentane	388.81	2.71	567.0	32.97	-212.8	2.11	5.96
3-ethylhexane	391.69	2.61	565.4	33.69	-210.7	2.11	5.83
3-ethyl-3-methylpentane	391.42	2.81	576.5	32.81	-214.9	2.15	6.00
2-methylheptane	390.80	2.49	559.6	33.44	-215.4	2.02	5.71
3-methylheptane	392.08	2.55	563.7	34.04	-212.5	2.08	5.76
4-methylheptane	390.86	2.54	561.7	33.88	-212.0	2.05	5.77
octane	398.83	2.49	568.8	34.77	-208.8	2.11	5.48
2,2,3,3-tetramethylbutane	379.60	2.87	567.8	31.42	-225.9	2.02	6.20
2,2,3-trimethylpentane	383.00	2.73	563.5	32.16	-220.0	2.02	6.05
2,2,4-trimethylpentane	372.39	2.57	544.0	31.02	-224.0	1.83	6.05
2,3,3-trimethylpentane	387.92	2.82	573.5	32.40	-218.5	2.11	6.05
2,3,4-trimethylpentane	386.62	2.73	566.3	32.62	-217.3	2.07	6.01

References

1. Kubinyi H. *QSAR: Hansch analysis and related approaches*. VCH: **1993**.
2. Roy K., Kar S., Das N.R. *A Primer on QSAR/QSPR Modeling. Fundamental Concepts*. Springer: **2015**.
3. Hastie T., Tibshirani R., Wainwright M. *Statistical Learning with Scarcity. The Lasso and Generalizations*. CRC Press: **2015**.
4. Talete SRL Homepage [online] http://www.talete.mi.it/products/dragon_description.htm (accessed March 10, 2019)
5. Wiener H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, 69(1), 17-20.
6. Randić M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, 97(23), 6609-6615.
7. Randić M. On Characterization of Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1997**, 37(4), 672-672.
8. Todeschini R., Consonni V. *Handbook of Molecular Descriptors*. New York, Wiley-VCH Verlag: **2000**, 667.
9. Stankevic M.I., Stankevic I.V., Zefirov N.S. Topologicheskie indeksy v organicheskoy khimii. *Uspechi khimii*. **1988**, 57(3), 337-366.
10. Ali A., Trinajstic A. Novel/Old Modification of the first Zagreb Index. *Mol. Inf.* **2018**, 37, 1800008.
11. Furtula B, Gutman I. A forgotten topological index. *J. Math. Chem*. **2015**, 53, 1184-1190.
12 Pogliani L. From Molecular Connectivity Indices to Semiempirical Connectivity Terms: Recent Trends in Graph Theoretical Descriptors. *Chem. Rev.* **2000**, 100, 3827-3858.
13. Zakharov A.B., Dyachenko A.V., Ivanov V.V. Topological characteristics of iterated line graphs in QSAR problem: Octane numbers of saturated hydrocarbons // *Journal of Chemometrics*. **2019**, e3169, 1-10.
14. Yaws C.L.. *Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds: Physical, Thermodynamic and Transport Properties for 5,000 Organic Chemical Compounds*. McGraw-Hill: **2003**.

15. Estrada E. Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 1. Definition and Applications to the Prediction of Physical Properties of Alkanes. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 844-849.
16. Diudea M.V., Horvath D., Graovac A. Molecular Topology. 15. 3D Distance Matrixes and Related Topological Indices. *J. Chem. Inf Comput. Sci.* **1995**, 35, 129-135.
17. Diudea M.V., Horvath D., Bonchev D. Molecular Topology. 14. Molord Algorithm and Real Number Subgraph Invariants. *Croat. Chem. Acta.* **1995**, 68, 131-148.
18. Tropsha A., Gramatica P., Gombar V.K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR. *QSAR Comb. Sci.* **2003**, 22, 69-77.
19. Hosamani S., Perigidad D., Jamagoud S. QSPR Analysis of Certain Degree Based Topological Indices // *J. Stat. Appl. Pro.* **2017**, 6(2), 361-371.

Надіслано до редакції 17 квітня 2019 р.

А.Б. Захаров, В.В. Иванов. Новый подход в использовании топологических дескрипторов. Итерированный реберный граф в проблеме физико-химических свойств насыщенных углеводородов.

Харьковский национальный университет имени В.Н. Каразина, химический факультет, пл. Свободы, 4, Харьков, 61022, Украина

В данной статье представлен новый взгляд на проблему описания молекулярных систем. Изучены возможности итерированных реберных графов при описании свойств насыщенных углеводородов. Продемонстрировано, что единственный выбранный молекулярный (графико-теоретический (топологический) или информационный) дескриптор рассчитанный для последовательности вложенных реберных графов позволяет получить достаточно надежные регрессионные уравнения. Таким образом, проблема сокращения дескрипторного набора решена в данном подходе по крайней мере частично. Разработан и реализован соответствующий программный пакет (QUASAR) с использованием языка программирования Python 3. В качестве тестового примера избранны физико-химические свойства изомеров октана. Среди изученных свойств - температура кипения, критическая температура, критическое давление, энтальпия испарения, энтальпия образования, сила поверхностного натяжения, а также вязкость. Получены соответствующие достаточно простые линейные регрессионные уравнения включающие один, два и три параметра соответственно. Предсказательная способность уравнений изучена с использованием процедур внутренней валидации. По процедурам leave-one-out (*LOO*) и Y-scrambling доказана адекватность полученных уравнений. Например, для регрессионной модели, полученной для температуры кипения, лучшие уравнения характеризуются коэффициентами детерминации R^2 = 0.943 и Q^2 = 0.918 (процедура *LOO*), в то время как по процедуре Y-scrambling $Q^2_{Y-scr} < 0.3$.

Также показано, что вышеуказанные молекулярные свойства в подходе вложенных реберных графов могут быть описаны с использованием общепринятых топологических дескрипторов. В общем почти любой избранный топологический дескриптор может давать адекватные уравнения. Эффективность продемонстрирована на примере индексов Загребской группы. «Забытый индекс» (ZM3) зарекомендовал себя как достаточно универсальный индекс при описании вышеуказанных свойств.

Ключевые слова: топологический дескриптор, реберный граф, регрессионный анализ, коэффициент детерминации, leave-one-out валидация, Y-scrambling, «забытый» индекс.

А.Б. Захаров, В.В. Іванов. Новий підхід у використанні топологічних дескрипторів. Ітерований реберний граф у проблемі фізико-хімічних властивостей насичених вуглеводнів.

Харківський національний університет імені В.Н. Каразіна, хімічний факультет, майдан Свободи, 4, Харків, 61022, Україна

В даній статті представлено новий погляд на проблему опису молекулярних систем. Вивчено здатності ітерованих реберних графів при описі властивостей насичених вуглеводнів. Продемонстровано, що єдиний обраний молекулярний (граф-теоретичний (топологічний) або інформаційний) дескриптор розрахований для послідовності вкладених реберних графів дозволяє достатньо отримати достатньо надійні регресійні рівняння. Таким чином, проблема скорочення дескрипторного набору вирішена в даному підході принаймні частково. Розроблено та реалізовано відповідний програмний пакет (QUASAR) із використанням мови програмування Python 3. У якості тестового прикладу обрані фізико-хімічні властивості ізомерів октану. Серед вивчених властивостей - температура кипіння, критична температура, критичний тиск, ентальпія пароутворення, ентальпія утворення, поверхневий натяг а також в'язкість. Отримані відповідні достатньо

прості лінійні регресійні рівняння що включають один, два та три параметри відповідно. Передбачувальна здатність рівнянь вивчена із використанням процедур внутрішньої валідації. За процедурою leave-one-out (*LOO*) та Y-scrambling доведена адекватність отриманих рівнянь. Наприклад, для регресійної моделі, що отримано для температури кипіння, найкращі рівняння характеризуються коефіцієнтами детермінації $R^2 = 0.943$ та $Q^2 = 0.918$ (процедура *LOO*), в той час як за процедурою Y-scrambling $Q^2_{Y-scr} < 0.3$.

Також показано, що вищевказані молекулярні властивості у підході вкладених реберних графів можуть бути описані із використанням загально вживаних топологічних дескрипторів. Взагалі майже кожен обраний топологічний дескриптор може давати адекватні рівняння. Ефективність продемонстровано на прикладі індексів Загребської групи. «Забутий» індекс (ZM3) зарекомендував себе як достатньо універсальний індекс при описанні вищевказаних властивостей.

Ключові слова: топологічний дескриптор, реберний граф, регресійний аналіз, коефіцієнт детермінації, leave-one-out валідація, Y-scrambling, «забутий» індекс.

Kharkiv University Bulletin. Chemical Series. Issue 32 (55), 2019