Extended Wiener indices. A new set of descriptors for quantitative structure-property studies

Ernesto Estrada,[a] Ovidiu Ivanciuc,[b] Ivan Gutman,[c] Amauri Gutierrez[a] and Lissette Rodríguez[a]

[a] *Centro de Bioactivo Químicos, Universidad Central de Las Villas, Santa Clara 54830, Villa Clara, Cuba.*
[b] *Faculty of Chemistry, Polytechnic University of Bucharest, 77206 Bucharest, Romania*
[c] *Faculty of Science, University of Kragujevac, P. O. Box 60, YU-34000 Kragujevac, Yugoslavia*

Higher order analogues of the Wiener number are defined, providing rather precise regression models for a number of physico-chemical properties of alkanes, which certainly are much better than the analogous models based solely on the Wiener number. The new structure descriptors are defined on the basis of the Wiener–Hosoya polynomial $H_G(x)$: the kth extended Wiener index ^kW is equal to the kth derivative of $H_G(x)$, evaluated at $x = 1$, in which case ^1W is just the original Wiener number.

An important field of research in contemporary chemistry is the modeling and prediction of physico-chemical and biological properties of molecules.[1,2] This kind of study is based on the paradigm that physico-chemical and biological properties are dependent on molecular structure. As a consequence, one of the most important points in such research is the selection of adequate descriptors containing the information stored in the molecular structure. Owing to the complexity of the molecular structure,[3,4] it seems to be impossible to expect that a single set of descriptors would contain all the relevant structural information. This is the main reason why the search for novel molecular structure descriptors continues. However, this search should not be done at random; it should follow some regular procedure based on the desired attributes that a molecular structure descriptor needs to possess.

The graph-theoretical approach to quantitative structure-property and structure-activity relationships (QSPR and QSAR, respectively) is based on a well-defined mathematical representation of the molecular structure.[5,6] The molecular descriptors derived therefrom are commonly named topological indices.[7,8] These indices are, in general, numbers containing relevant information about the steric structure of the molecule. Most of the measured physico-chemical properties are steric properties, and consequently they may be reasonably well-described by topological indices. However, in some cases these indices also contain structural information related to the electronic and/or dipolar features of molecules.[9–11]

The first topological index reported in the chemical literature is the so-called Wiener W number.[12–14] It was proposed[12] by H. Wiener 50 years ago as a molecular descriptor to model physico-chemical properties of alkanes. After several years, Hosoya[15] pointed out that the W number can be computed from the topological distance matrix of the graph representing the hydrogen-depleted molecule. As a consequence, W is a graph-theoretical or topological index. The Hosoya formula for the Wiener index of a (molecular) graph G is given by[15] $W = W(\mathrm{G}) = (1/2) \sum_{ij} d_{ij}$ where d_{ij} is the topological distance, that is the number of bonds separating the atoms i and j.

More recently, Hosoya[16] introduced a graph polynomial, which he named the Wiener polynomial of the graph: $H_G(x) = \sum_{i=1}^{l} \eta_i x^i$ where η_i is the number of pairs of vertices at distance i and l is the longest distance in the graph. We prefer to name $H_G(x)$ the Wiener–Hosoya polynomial and in the future it may, perhaps, be referred to as the Hosoya polynomial. The Wiener–Hosoya polynomial has an important property from the point of view of our main objective. This property, also recognized by Hosoya,[16] is that the first derivative of $H_G(x)$ evaluated at $x = 1$ is identical to the Wiener number: $H'_G(1) = [\mathrm{d}H_G(1)/\mathrm{d}x] = \sum_{i=1}^{l} i\eta_i = W(\mathrm{G})$.

What we propose here is the use of $H_G(x)$ for generating a series of Wiener-type topological indices, complementing W in the study of structure-property relationships. As will be shown, these extended Wiener indices increase significantly the quality and the predictive ability of QSPR models for several physico-chemical properties.

Extended Wiener Indices

It is a desired attribute of topological indices that they can be extended to series of 'higher' analogues.[17] This extension is necessary in order to complement the simple index in such correlations in which its use as a single variable does not produce sufficiently good results. The best example of the utility of this approach is given by the higher order molecular connectivity indices put forward by Kier and Hall,[18,19] which are extensions of the classical connectivity index of Randic.[20]

There have been several attempts to extend the Wiener index to series of analogous descriptors in order to complement it in QSPR and QSAR studies. For instance, Lukovits[21–23] has decomposed the Wiener index into contributions coming from different types of bonds in the molecule. Klein and coworkers[24] also analyzed several ways in which this number can be extended by algebraic manipulations of graph-theoretical matrices. Estrada *et al.* proposed a vector-matrix-vector multiplication procedure to develop series of Wiener-type indices.[25,26] However, a series of 'higher' analogues, in which the original W number would be the first member, has not been conceived so far. In order to develop such a series we propose here the use of the higher order derivatives of the Wiener–Hosoya polynomial.

Let G be a (molecular) graph and let $H_G(x)$ be its Wiener–Hosoya polynomial. Then we call $H'_G(1)$ the first-order Wiener index and denote it by $^1W(\mathrm{G})$. By using the higher order derivatives of $H_G(x)$, we introduce the following list of extended Wiener indices: $^2W(\mathrm{G}) = H''_G(1)$, $^3W(\mathrm{G}) = H'''_G(1)$, *etc.* For $k = 1$, 2, 3, *etc.*, the quantity $^kW(\mathrm{G}) = H^k_G(1)$ will be named the kth-order Wiener index or Wiener number. Of course, these indices can be calculated without the use of

$H_G(x)$, by means of the following expressions:

$$^2W = \sum_{i=2}^{l} i(i-1)\eta_i, \quad ^3W = \sum_{i=3}^{l} i(i-1)(i-2)\eta_i, \quad \ldots,$$

$$^kW = \sum_{i=k-1}^{l} i(i-2)(i-2) \cdots [i-(k-1)]\eta_i$$

The above formulas permit also a better understanding of the meaning of the novel (extended) Wiener indices. As can be seen from them, the second-order index contains no information about the simple distances in the molecule; that is, the counting of distances begins from two. In the third-order index, both the contributions coming from pairs of atoms separated by distances of one and two are ignored. In general, the kth-order index takes into account only the contributions coming from pairs of vertices separated by distances longer than or equal to k, and of course all contributions coming from pairs of vertices separated by distances shorter than k are disregarded. In addition to this, as the order of the index increases, the weights of the contributions that were taken into consideration also increase. For instance, the weight of the contribution of a pair of vertices at distance i taken into account in ^1W is i, in ^2W it is $i(i-1)$, in ^3W it is $i(i-1)(i-2)$ and in ^kW it is $i(i-1)(i-2) \cdots [i-(k-1)]$. This means that the higher order indices give more and more importance to the contributions coming from pairs of the most distant atoms in the molecule, ignoring each time those coming from the closest ones.

One of the main drawbacks of the original Wiener index is its poor discrimination of isomers: the index has the same value for different isomeric compounds. It is well-known that this degeneracy increases when the number of atoms in the molecule increases, even for simple molecules such as alkanes.[27] The capacity of one index to discriminate isomers can be measured by using a discrimination index, D.[27] It can be calculated as the number of isomers having different values of the index, divided by the total number of isomers. This index has been calculated for series of isomers. Obviously, D is equal to unity whenever there are no two isomers with the same value of the topological index. In Fig. 1 we show the discrimination ability of the Wiener index and its first two higher order analogues, as well as of the Wiener–Hosoya polynomial. The Wiener–Hosoya polynomial has the greatest isomer-discriminating power in this series of Wiener-type structure descriptors. As can be seen, the discrimination ability of ^1W is poor, especially for larger molecules. This discrimination is significantly increased for the higher order Wiener indices. The Wiener index as well as the Wiener–Hosoya polynomial produce regular variations of D, depending on the parity of the number of carbon atoms. In particular, the discrimination of isomers is better when the number of carbon atoms is even, relative to what is found for isomers having an odd number of carbons. This observation

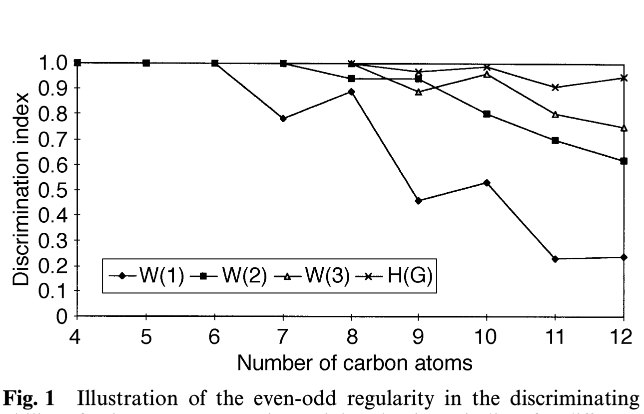

Fig. 1 Illustration of the even-odd regularity in the discriminating ability of Wiener–Hosoya polynomial and Wiener indices for different series of alkane isomers, compared to other extended Wiener indices

was first made by Razinger *et al.*[27] for the Wiener index. In the case of ^3W this regularity is only partially obeyed, whereas for ^2W a continuous decrease of D is observed when the number of atoms increases beyond nine.

A rationalization of the odd-even alternation of the isomer-discriminating power of the Wiener number and the Wiener–Hosoya polynomial is given in the Appendix.

Quantitative Structure-Property Studies

The most important aspect of the development of novel molecular descriptors is their applicability to the quantitative description of experimentally measured properties. In order to test the quality of the extended series of Wiener indices introduced here, we propose the study of several physico-chemical properties of alkanes. The properties that are studied here are: boiling point (°C), heat capacity (J mol^{-1} $grad^{-1}$) at 25 °C, density (kg m^{-33}) at 25 °C, Gibbs energy (kJ mol^{-1}) at 25 °C, enthalpy (kJ mol^{-1}) and refractive index at 25 °C. These properties were collected by Gakh *et al.*[28] for 134 alkanes from 'the most reliable sources of information to avoid experimental errors'. They divided this data set into two subsets, one containing 109 and the other 25 compounds. The first was used as a training set for the development of quantitative models by using neural networks and the second was used as an external prediction set. Here we use the same training and prediction sets proposed by Gakh *et al.*[28] For the development of the quantitative models we use multivariate linear regression analysis instead of neural networks.

In Table 1 are given the results of the stepwise procedure for the development of the QSPR models. In this table we show the statistical parameters: R (correlation coefficient), s (standard deviation) and % error (average error, calculated as $s/\bar{y}$, where $\bar{y}$ is the average of the respective experimental property), for both the training and the prediction series. For the training series we also included the values of the standard deviation and the percentage error obtained from the leave-one-out cross-validation procedure. These values are given in parentheses together with the corresponding values for the training set. Only the models obtained by using one, two and three parameters are reported, together with the best model found by using the present approach. The selection of the best model was carried out by considering all the statistical parameters studied for both training and prediction sets. The case of the heat capacity is exceptional: here the best model found was based on only two variables (^1W and ^2W) and the inclusion of any other higher analogue of the Wiener index resulted in an increase of the standard deviation in the prediction set.

As can be seen from Table 1, three properties (density, Gibbs energy and refractive index) are not well-described by the original Wiener index. The linear regressions pertaining to these properties when using solely ^1W explain less than 50% ($R^2 \leqslant 0.5$) of the variance of the experimental properties. However, inclusion of only the second-order Wiener index produces a significant improvement in these models. The percentage of the variance explained by the two-variable models is greater than 85% (for each of the above-mentioned three properties). The best models found explain more than 95% of the variances of these physico-chemical properties.

For all the studied properties the inclusion of the extended indices produced significant improvements in models, with the only exception being the enthalpy. The inclusion of a second variable represented improvements in the standard deviations of the training sets greater than 30%. These improvements were greater than 50% for the final models found: boiling point (52%), heat capacity (43%), density (74%), Gibbs energy (72%), refractive index (72%). However, the improvement in the standard deviation for the final model describing the enthalpy was only 8%. These improvements are also similar for the leave-one-out cross-validation of the training set, as can be

Table 1 Statistical results for the regression models describing physico-chemical properties of alkanes in the training and prediction sets

Index	Training set[a]			Prediction set		
	R	s	% error	R	s	% error
Boiling point/°C						
^1W	0.9110	10.95 (11.11)	7.81 (7.92)	0.9588	7.94	5.81
^1W, ^2W	0.9622	7.27 (7.43)	5.18 (5.30)	0.9752	6.20	4.53
^1W, ^2W, ^5W	0.9693	6.59 (6.80)	4.70 (4.85)	0.9751	6.20	4.53
^1W, ^2W, ^4W, ^5W, ^6W, ^7W	0.9812	5.25 (5.48)	3.74 (3.91)	0.9866	4.56	3.33
Heat capacity/J mol^{-1} grad^{-1} at 25 °C						
^1W	0.9432	8.36 (8.47)	3.86 (3.91)	0.9388	9.58	4.50
^1W, ^2W	0.9821	4.75 (4.84)	2.19 (2.24)	0.9929	3.32	1.56
Density/kg m^{-3} at 25 °C						
^1W	0.6662	19.73 (19.93)	2.71 (2.73)	0.8094	15.27	2.12
^1W, ^2W	0.9286	9.86 (10.04)	1.35 (1.38)	0.9278	10.59	1.33
^1W, ^2W, ^3W	0.9523	8.14 (8.34)	1.12 (1.14)	0.9486	8.23	1.13
^2W, ^4W, ^5W, ^6W, ^7W	0.9819	5.11 (5.32)	0.70 (0.73)	0.9780	5.42	0.74
Gibbs energy/kJ mol^{-1} at 25 °C						
^1W	0.6113	12.11 (12.18)	39.49 (39.72)	0.7715	7.78	25.38
^1W, ^2W	0.9242	5.87 (5.99)	19.14 (19.53)	0.9207	4.77	15.56
^1W, ^2W, ^5W	0.9582	4.36 (4.54)	14.22 (14.81)	0.9691	3.01	9.82
^1W, ^2W, ^4W, ^5W, ^6W	0.9763	3.38 (3.49)	11.02 (11.38)	0.9792	2.48	8.19
Enthalpy/kJ mol^{-1}						
^1W	0.9619	1.19 (1.21)	3.13 (3.17)	0.9712	1.04	2.71
^1W, ^4W	0.9622	1.19 (1.21)	3.13 (3.19)	0.9741	0.98	2.57
^1W, ^4W, ^7W	0.9629	1.19 (1.22)	3.13 (3.20)	0.9743	0.98	2.57
^1W, ^2W, ^4W, ^5W, ^7W	0.9791	1.09 (1.13)	2.87 (2.98)	0.9744	0.97	2.56
Refractive index at 25 °C						
^1W	0.6986	9.78 (9.88)	0.69 (0.70)	0.8383	7.62	0.54
^1W, ^2W	0.9334	4.93 (5.03)	0.35 (0.36)	0.9334	5.02	0.35
^1W, ^2W, ^4W	0.9582	3.94 (4.09)	0.28 (0.29)	0.9512	4.31	0.31
^2W, ^4W, ^5W, ^6W, ^7W	0.9801	2.76 (2.91)	0.20 (0.21)	0.9514	4.30	0.31

[a] Standard deviations and the % error for the leave-one-out cross-validation of the training set are given in parentheses together with the corresponding values for the model.

observed from the values of the standard deviation given in parentheses in Table 1. Another interesting feature is the fact that final models describing density and refractive index do not contain the original ^1W index. It can be observed from Table 1 that nearly all statistical parameters of the prediction set are significantly better than those corresponding to the training set. This is also observed in the results obtained by Gakh *et al.*[28] who obtained average deviations between 1.3–2.7% for the training set while this parameter in the prediction set was always less than 2%. We recall that we are using here exactly the same data sets as reported by these authors.

All these observations, the improvements in the quality of the models and the lack of inclusion of ^1W in two of them, are in complete agreement with the nature of our approach to extend the Wiener index. We may assume that some properties are more dependent on local structural features than others are. For instance, it is a well-established fact that the enthalpy is one of these properties, that is it is well-described in terms of additive bond properties. The procedure carried out here to extend the Wiener index 'truncates' the contributions coming from the less distant pairs of atoms in the molecule. On the other hand, the contributions that are taken into account in these extended indices have greater weights each time; the contributions in ^kW are multiplied by a factor that is greater than the one used in ^{k-1}W. As a consequence, the higher order Wiener indices can be considered as more 'global' molecular descriptors than the preceeding ones. Of course, ^1W is also a 'global' molecular descriptor, but it contains information coming from the atoms at distance one, which can be considered as local, bond-defined, information. These contributions are ignored in higher order indices, which give more and more importance to those coming from the most distant atoms in the molecule.

Considering the structural information contained in the extended Wiener indices we can say that some of the studied physico-chemical properties, such as the density and the refraction index, are not dependent on the local contributions contained in ^1W but are mainly explained by the global structural features of molecules, as measured by the distance between the most distant atoms. Other properties, such as the enthalpy, are more dependent on local contributions, as are the atom-atom contributions, and consequently there is no improvement in the quality of the models intended to describe these properties when higher order indices are introduced. There are properties, such as the Gibbs energy, for which the contributions coming from local structural features are very small and the use of higher order Wiener indices produces dramatic improvement in the quality of the models. Other cases, such as the boiling point and heat capacity, have important contributions from the atom-atom terms contained in ^1W. In these cases the use of higher order indices introduces complementary structural information and results in a further increase in the quality of the QSPR models.

It is clear that there are several possible extensions of known topological indices that produce significant improvements in the quality of the models obtained from them, as well as in the structural interpretation of such models. The extension of Wiener indices carried out in the present work constitutes one of these examples. Consequently, the extended Wiener indices are good candidates to be used in QSPR and QSAR studies as a complement of the original W index. In addition, one of the most desirable attributes for topological indices, the extension to 'higher' order analogues, has been fulfilled for the Wiener index. The extended W indices represent these 'higher' order analogues and their use is justified in the frequently occurring cases in which the original W

number does not suffice for constructing a predictive model of satisfactorily high accuracy.

Appendix: An Even-Odd Regularity for the Wiener–Hosoya Polynomial

Consider a (connected) bipartite graph G, possessing a vertices of one color (say, black) and b vertices of another color (say, white). Thus G possesses $n = a + b$ vertices. Recall that trees, and therefore molecular graphs of alkanes, are bipartite.

Any path in G goes alternately through black and white vertices. For instance, a shortest path starting and ending at a black vertex has the form black–white–black–white– $\cdots$ – black–white–black and is therefore necessarily of even length. Consequently the distance between any two black vertices is even. Similarly, the distance between any two white vertices is also even, whereas the distance between a black and a white vertex is odd. This implies that in the graph G there are $\binom{a}{2} + \binom{b}{2}$ pairs of vertices at even distances and ab pairs of vertices at odd distances.

As before, we denote the number of vertex pairs at distance i by η_i. The above conclusions are then rewritten as:

$$\eta_2 + \eta_4 + \eta_6 + \cdots = \binom{a}{2} + \binom{b}{2}$$
$$\eta_1 + \eta_3 + \eta_5 + \cdots = a \cdot b \qquad (1)$$

If the number of vertices, n, is even, then either both a and b are even or both a and b are odd. Therefore the left-hand side of eq. (1) may assume either even or odd values. If, however, the number of vertices is odd, then either a is even and b is odd or vice versa, but the product ab is necessarily even. Consequently, the sum of the odd coefficients of the Wiener–Hosoya polynomial must assume only even values. This means that for bipartite graphs of odd n, there exist fewer possible combinations for these coefficients, pointing towards a lower isomer-discriminating power than in the cases $n - 1$ and $n + 1$.

In the case of the Wiener number the situation is analogous. The contribution to W coming from vertex pairs at even distances, namely $2\eta_2 + 4\eta_4 + 6\eta_6 + \cdots$, is necessarily even. On the other hand, the contribution coming from vertex pairs at odd distances, namely $\eta_1 + 3\eta_3 + 5\eta_5 + \cdots$ has the same parity as $\eta_1 + \eta_3 + \eta_5 + \cdots$ that is, the same parity as ab. Therefore, if n is even, then the Wiener number may be either even or odd. If, however, n is odd, then the Wiener number must be even. Therefore, in the case of bipartite graphs with an odd number of vertices there are fewer possible values that W may assume than in the case of similar-sized graphs with an even number of vertices. Hence, for odd n, the isomer-discriminating power of W could be expected to be smaller than for $n - 1$ or $n + 1$.

No restriction in parity of the above mentioned type exists for ${}^2W(\mathrm{G})$, ${}^3W(G)$, *etc.*, which is in agreement with the finding that the isomer-discriminating power of these topological indices does not oscillate with the parity of the number of vertices [*cf.* Fig. (1)].

References

1 M. Randić and N. Trinajstić, *New J. Chem.*, 1994, **18**, 179.
2 M. Randić, *New J. Chem.*, 1996, **20**, 10019.
3 M. Randić, *J. Math. Chem.*, 1991, **7**, 155.
4 M. Randić, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 672.
5 P. G. Seybold, M. May and V. A. Bagal, *J. Chem. Educ.*, 1987, **64**, 189.
6 Z. Mihalić and N. Trinajstić, *J. Chem. Educ.*, 1992, **69**, 701.
7 A. Sablijć and N. Trinajstić, *Acta Pharm. Jugosl.*, 1981, **31**, 189.
8 P. J. Hansen and P. C. Jurs, *J. Chem. Educ.*, 1988, **65**, 574.
9 L. H. Hall, B. Mohney and L. B. Kier, *Quant. Struct.-Act. Relat.*, 1991, **10**, 43.
10 L. H. Hall, B. Mohney and L. B. Kier, *J. Chem. Inf. Comput. Sci.*, 1991, **31**, 76
11 J. Galvez, R. Garcia, M. T. Salabert, R. Soler, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 520.
12 H. Wiener H., *J. Am. Chem. Soc.*, 1947, **69**, 17.
13 I. Gutman, Y.-N. Yeh, S.-L. Lee and Y.-L. Luo, *Indian J. Chem., Sect. A*, 1993, **32**, 651.
14 S. Nikolić, N. Trinajstić and Z. Mihalić, *Croat. Chem. Acta*, 1995, **68**, 105.
15 H. Hosoya, *Bull. Chem. Soc. Jpn.*, 1971, **44**, 2332.
16 H. Hosoya, *Discrete Appl. Math.*, 1988, **19**, 239.
17 M. Randić, *J. Math. Chem.*, 1991, **7**, 155.
18 L. B. Kier and L. H. Hall, in *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976.
19 L. B. Kier and L. H. Hall, in *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, Letchworth, UK, 1986.
20 M. Randić, *J. Am. Chem. Soc.*, 1975, **97**, 6609.
21 I. Lukovits, *J. Chem. Soc., Perkin Trans. 2*, 1988, 1667.
22 I. Lukovits, *Quant. Struct.-Act. Relat.*, 1990, **9**, 227.
23 I. Lukovits, *Int. J. Quantum. Chem., Quantum Biol. Symp.*, 1992, **19**, 217.
24 H.-Y. Zhu, D. J. Klein and I. Lukovits, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 420.
25 E. Estrada, L. Rodríguez and A. Gutiérrez, *Commun. Math. Chem. (MATCH)*, 1997, **35**, 145.
26 E. Estrada and L. Rodríguez, *Commun. Math. Chem. (MATCH)*, 1997, **35**, 157.
27 M. Razinger, J. R. Chretien and J. E. Duboois, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 23.
28 A. A. Gakh, E. G. Gakh, B. G. Sumpter and D. W. Noid, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 832.

Received in Montpellier, France, 15th December 1997; *Paper* 7/09255E