

Bond Additive Modeling 6. Randomness vs. Design

Damir Vukičević

Faculty of Natural Sciences and Mathematics, University of Split, HR-21000 Split, Croatia

(Received March 4, 2010)

Abstract

In this paper we analyze predictions of properties of octane isomers given by International Academy of Mathematical Chemistry (IAMC). Recently it was shown that one-parameter linear models based on the Adriatic descriptors have higher coefficient of determination r^2 for 10 out of 16 properties than the analogous models based on the benchmark set of descriptors given by IAMC. However, it is possible that some of these results are the results of the pure chance. Here, we propose the series of five additional tests to test whether these results are obtained by chance or if they are the result of insightful design. It is indicated that the predictions of the melting point (not even analyzed in [1]), density (result commented as doubtful in [1]), molar volume (also results commented as doubtful in [1]), heat capacity and octanol-water partition coefficient at P constant may be the results of pure chance and it seems that predictions of heat capacity at V constant, enthalpy of vaporization, standard enthalpy of vaporization, molar octane number and total surface area are the results of the insightful design.

1. Introduction

The molecular descriptor is the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment [2]. Molecular descriptors have been shown to be useful in modeling many physico-chemical properties in numerous QSAR and QSPR studies [3-5].

Recently, International Academy of Mathematical Chemistry (IAMC) [6] proposed four sets of benchmark descriptors and chemical properties for four classes of molecules [7]. In this

paper, we shall restrict ourselves only to observing octane isomers. IAMC proposed 102 descriptors and 16 properties for this class of molecules. In addition, recently the set of 148 discrete Adriatic properties has been proposed (see [1] and for further studies of the Adriatic indices see [8-12]). They have shown good predictive properties in the one-parametric linear models and outperform benchmark descriptors proposed by IAMC in several cases.

It is of interest to check if these descriptors have good predictive properties, because of their insightful design or their predictive properties having purely random occurrence. In order to test this theory the series of six tests is performed. We assume that predictive property is not the result of the chance if the following conditions are met:

- 1) Coefficient of determination r^2 is higher for the best linear model based on the Adriatic index than the best linear model based on the benchmark descriptor;
- 2) Leave-one-out cross-validation of the whole set of Adriatic descriptors should produce better prediction than leave-one-out cross-validation of the whole set of benchmark descriptors and results for the Adriatic descriptors are statistically significant at significance level of 99.9%;
- 3) Coefficient of determination r^2 is higher for the best linear model based on the Adriatic index than for 99.999% linear models based on the random vectors with uniform distribution;
- 4) Coefficient of determination r^2 is higher for the best linear model based on the Adriatic index than for 99.999% linear models based on the random vectors with normal Gaussian distribution;
- 5) Coefficient of determination r^2 is higher for the best linear model based on the Adriatic index than the best linear model based on the Adriatic index when values of the observed property are randomly permuted in at least 95% of cases;
- 6) Coefficient of determination r^2 is higher for the best linear model based on the Adriatic index than the best linear model based on the Adriatic index when values of the observed property are randomly permuted in such a way that molecules with the same number of pendant vertices are mapped to each other in at least 95% of cases.

It is shown that predictions of the heat capacity at V constant, enthalpy of vaporization, standard enthalpy of vaporization, molar octane number and total surface area satisfied all 6 tests, while prediction of the remaining properties failed at least one of these tests.

2. Main results

The first test based on the comparison of the correlation coefficient is very natural and it does not require any special comments. All the calculations, but for the boiling point are given in [1].

property	Adriatic indices	benchmark descriptors	test passed
boiling point	0.73300	0.77616	-
melting point	0.77317	0.75709	+
heat capacity at V constant	0.76011	0.50484	+
heat capacity at P constant	0.63990	0.59383	+
entropy	0.91241	0.91629	-
density	0.91180	0.59367	+
enthalpy of vaporization	0.90709	0.88606	+
standard enthalpy of vaporization	0.96778	0.92005	+
enthalpy of formation	0.79300	0.83238	-
standard enthalpy of formation	0.59514	0.66918	-
motor octane number	0.95690	0.92741	+
molar refraction	0.93385	0.97937	-
acentric factor	0.99044	0.99229	-
total surface area	0.77615	0.71685	+
octanol-water partition coefficient	0.36487	0.29410	+
molar volume	0.89733	0.54827	+

Table 1. Comparison of r^2 values

Let us describe the second test. Let X be the set of descriptors (in our case X is either set of 102 benchmark descriptors or the set of 148 Adriatic indices) and let P be the observed

property. We repeat the following procedure for the each molecule m in the set of 18 octane isomers M . We observe the set of $M \setminus \{m\}$ molecules and find the descriptor $D \in X$ which linear model $D(\mu) \approx a \cdot P(\mu) + b$ has the highest coefficient of determination r^2 in the set $M \setminus \{m\}$ among all descriptors in X . Then, we estimate $P'(m) = a \cdot m + b$. Further, we calculate:

$$r^2 = \max \left\{ 1 - \frac{\sum_{\mu \in M} (P(\mu) - P'(\mu))^2}{\sum_{\mu \in M} \left(P(\mu) - \frac{\sum_{\mu \in M} P(\mu)}{\text{card}(M)} \right)^2}, 0 \right\},$$

where $\text{card}(M)$ denotes cardinality of M , i.e. the number of elements in M . Finally, we compare the results for the Adriatic indices and the benchmark descriptors. We consider that descriptor is significant if it is better then benchmark descriptor and it has the level of significance at least 99.9%, i.e. that

$$\frac{\sqrt{r^2} \cdot \sqrt{n-2}}{\sqrt{1-r^2}} \geq 2.98.$$

We argue that it is better to perform this test then to compare using leave-one-out cross validation just of the best descriptor, since in the reality we would not be able to say which descriptor is the best if the value of $P(m)$ is not given. The results of calculations are presented in the following table:

property	Adriatic indices	benchmark descriptors	test passed
boiling point	0.23788	0.64588	-
melting point	0	0.04701	-
heat capacity at V constant	0.70288	0	+

heat capacity at P constant	0	0.15822	-
entropy	0.79053	0.82797	-
density	0.20680	0	-
enthalpy of vaporization	0.79086	0.77471	+
standard enthalpy of vaporization	0.94668	0.83854	+
enthalpy of formation	0.74565	0.78828	-
standard enthalpy of formation	0.46210	0.54012	-
motor octane number	0.94615	0.80501	+
molar refraction	0.90991	0.95307	-
acentric factor	0.98873	0.99046	-
total surface area	0.59613	0.51290	+
octanol-water partition coefficient	0	0.13027	-
molar volume	0.25171	0	-

Table 2. Comparison of r^2 values (leave-one-out)

The third and fourth tests (namely, comparing with random) are also quite standard tests. Let us just comment on the value 99.999%. We assume that the number of molecular descriptors in the use today is about several thousands. We would argue that the most predictive ones might already be included in the benchmark set. Hence, if we want to outperform the benchmark set of descriptors, Adriatic indices should be as good as the best of several thousand descriptors. In statistics, it is customary to require 95% level of the significance of the results. Hence, we should observe about 20 times more random descriptors than the number of the descriptors that are used. We get that this value is approximately 100 000 and therefore we require the value 99.999%. In our tests, we make 10 000 000 trials and we expect less than 100 better predictions.

property	better or equal results	worse results	test passed
boiling point	25	9999975	+
melting point	15	9999985	+
heat capacity at V constant	21	9999979	+

heat capacity at P constant	1040	9998960	-
entropy	0	10000000	+
density	0	10000000	+
enthalpy of vaporization	0	10000000	+
standard enthalpy of vaporization	0	10000000	+
enthalpy of formation	18	9999982	+
standard enthalpy of formation	2758	9997242	-
motor octane number	0	10000000	+
molar refraction	0	10000000	+
acentric factor	0	10000000	+
total surface area	16	9999984	+
octanol-water partition coefficient	82977	9917023	-
molar volume	0	10000000	+

Table 3. Comparison of the best Adriatic descriptor with random uniformly distributed vectors

property	better or equal results	worse results	test passed
boiling point	59	9999941	+
melting point	325	9999675	-
heat capacity at V constant	26	9999974	+
heat capacity at P constant	666	9999334	-
entropy	0	10000000	+
density	0	10000000	+
enthalpy of vaporization	0	10000000	+
standard enthalpy of vaporization	0	10000000	+
enthalpy of formation	11	9999989	+
standard enthalpy of formation	1760	9998240	-
motor octane number	0	10000000	+
molar refraction	0	10000000	+
acentric factor	0	10000000	+
total surface area	13	9999987	+

octanol-water partition coefficient	79177	9920823	-
molar volume	0	10000000	+

Table 4. Comparison of the best Adriatic descriptor with random normally distributed vectors

The fifth test is also relatively standard test. Permuting of the values of some property may be better then the comparing with the random since, in this way, particularities of the distribution of the observed property are preserved. In our tests we observe 10 000 random permutations (out of $18! \approx 6.4 \cdot 10^{15}$)

property	better or equal results	worse results	test passed
boiling point	10	9990	+
melting point	995	9005	-
heat capacity at V constant	6	9994	+
heat capacity at P constant	23	9977	+
entropy	0	10000	+
density	1	9999	+
enthalpy of vaporization	0	10000	+
standard enthalpy of vaporization	0	10000	+
enthalpy of formation	4	9996	+
standard enthalpy of formation	28	9972	+
motor octane number	0	10000	+
molar refraction	0	10000	+
acentric factor	0	10000	+
total surface area	6	9994	+
octanol-water partition coefficient	933	9067	-
molar volume	1	9999	+

Table 5. Comparison with the random permutations

The last test is up to author's knowledge a new test and may be of interest as more sensitive then fifth test. It is well known that many properties are highly influenced by the branching of the molecules. One of the simplest values related to branching is the number of pendant

vertices. Also, many molecular descriptors are correlated with the number of pendant vertices. Hence, one may expect that random permutation may decrease the correlation coefficient of some property and the descriptor. This can hide the fact that descriptor does not have good predictive properties. Hence, we propose as additional test the usage of only those permutations that preserve the number of pendant vertices. In the set of 18 octane isomers, there is 1 molecule with two pendant vertices, 4 molecules with three pendant vertices, 8 molecules with four pendant vertices and 1 molecule with 6 pendant vertices, hence there are $4! \cdot 8! \cdot 4! \approx 2.3 \cdot 10^7$ such permutations. In our test we randomly select 10 000 of them.

property	better or equal results	worse results	test passed
boiling point	248	9752	-
melting point	2764	7236	-
heat capacity at V constant	0	10000	+
heat capacity at P constant	12	9988	+
entropy	236	9764	-
density	3	9997	+
enthalpy of vaporization	75	9925	+
standard enthalpy of vaporization	5	9995	+
enthalpy of formation	56	9944	+
standard enthalpy of formation	31	9969	+
motor octane number	0	10000	+
molar refraction	1	9999	+
acentric factor	0	10000	+
total surface area	0	10000	+
octanol-water partition coefficient	1543	8457	-
molar volume	3	9997	+

Table 6. Comparison with the random permutations that preserve the number of pendant vertices

3. Discussion and conclusions

The results of tables 1)-6) are compactly presented in the table 7

property	test 1	test 2	test 3	test 4	test 5	test 6
boiling point	-	-	+	+	+	-
melting point	+	-	+	-	-	-
heat capacity at V constant	+	+	+	+	+	+
heat capacity at P constant	+	-	-	-	+	+
entropy	-	-	+	+	+	-
density	+	-	+	+	+	+
enthalpy of vaporization	+	+	+	+	+	+
standard enthalpy of vaporization	+	+	+	+	+	+
enthalpy of formation	-	-	+	+	+	+
standard enthalpy of formation	-	-	-	-	+	+
motor octane number	+	+	+	+	+	+
molar refraction	-	-	+	+	+	+
acentric factor	-	-	+	+	+	+
total surface area	+	+	+	+	+	+
octanol-water partition coefficient	+	-	-	-	-	-
molar volume	+	-	+	+	+	+

Table 7. The results of the tests 1)-6)

One can see that leave-one-out cross validation was the most important test in this case. In addition, the novel test was more important than conventionally used test 5. Therefore, we propose the test 6 as novel, but potentially significant, improvement of the test 5.

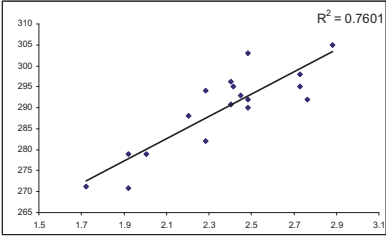
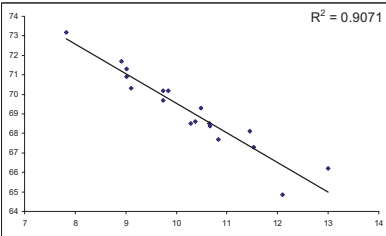
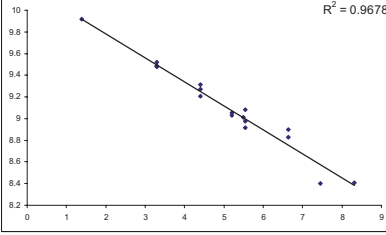
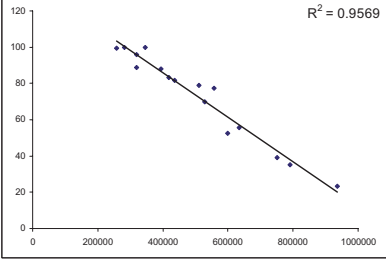
Hence, the results indicate that predictions of one-parameter linear model for the heat capacity at T constant, enthalpy of vaporization, standard enthalpy of vaporization, molar octane number and total surface area are not the result of the pure chance, but of the insightful design.

However, the results for the melting point, heat capacity at P constant, density, octanol-water partition coefficient and molar volume may be the results of the pure chance. Let us comment in more details the results for these properties:

- 1) **melting point** – It is well known that these kind of descriptors do not capture well information about melting point, moreover melting point has been excluded from the analyses of paper [1] in which Adriatic indices have been defined.
- 2) **density and molar volume** – Both of these properties have the value for 2,2,3,3-tetramethylbutane as outlier. Hence, distribution of both properties drastically differs from the normal distribution. In this case linear correlations do not have significance – namely in both cases we have two clusters: one cluster consisting of 2,2,3,3- tetramethylbutane and the other cluster consisting of the remaining 17 molecules. In paper [1] it was commented that these results should not be taken as significant.
- 3) **heat capacity at P constant** – Note that this property has two lowest values for octane and 2,2,3,3- tetramethylbutane. These two properties are two opposite extremal graphs when branching is considered and it seems that Adriatic indices do not perform well on these kind of properties.
- 4) **octanol-water partition coefficient** - r^2 for the best Adriatic index is only 0.365 which (although higher than benchmark descriptor) is very low. Hence, this result does not give a good prediction. It failed all further tests.

As comment 3) suggests the Adriatic indices did not predict well heat capacity at P constant. It would be interesting to collect more chemical properties that have values for the octane and 2,2,3,3- tetramethylbutane as highest or the lowest values and to try to develop descriptors that would perform well for these kind of properties.

Here, we preformed a series of tests and we have found that predictions of four properties can be considered important. Figures [1] corresponding to these properties and respective indices [1] are presented in the following Table:

<p>heat capacity at T constant</p> <p>predicted by</p> <p>Randić type lodeg index:</p> $\sum_{uv \in E(G)} \ln(d_u) \cdot \ln(d_v)$	
<p>enthalpy of vaporization</p> <p>predicted by</p> <p>max-min rodeg index:</p> $\sum_{uv \in E(G)} \frac{\max\{\sqrt{d_u}, \sqrt{d_v}\}}{\min\{\sqrt{d_u}, \sqrt{d_v}\}} = \sum_{uv \in E(G)} \sqrt{\frac{\max\{d_u, d_v\}}{\min\{d_u, d_v\}}}$	
<p>standard enthalpy of vaporization</p> <p>predicted by</p> <p>misbalance lodeg index:</p> $\sum_{uv \in E(G)} \ln d_u - \ln d_v $	
<p>molar octane number</p> <p>predicted by</p> <p>Randić type sdi index:</p> $\sum_{uv \in E(G)} (D_x^2 D_y^2)$	

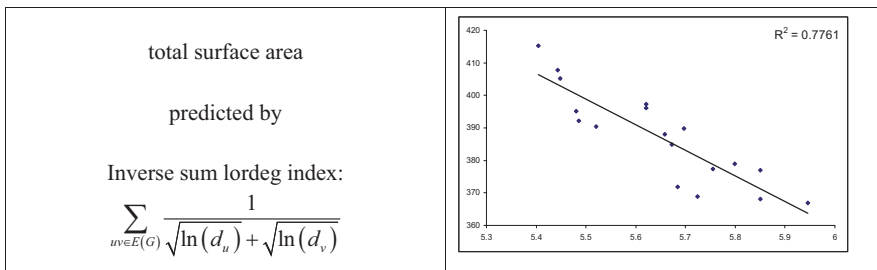


Table 8. Adriatic descriptors that satisfied all 6 tests

4. Acknowledgment

The partial support of Croatian Ministry of Science, Education and Sport (grants no. 177-0000000-0884 and 037-0000000-2779) is gratefully acknowledged. Useful comments and help from Boris Furtula and anonymous referee are gratefully acknowledged.

5. References

1. D. Vukičević, M. Gašperov, Bond additive modeling 1. Adriatic indices, *Croat. Chem. Acta*, accepted for publication.
2. R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
3. N. Trinajstić, *Chemical Graph Theory*, CRC Press, Boca Raton, 1992.
4. J. Devillers, A. T. Balaban (Eds.), *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach, Amsterdam, 1999.
5. M. Karelson, *Molecular Descriptors in QSAR/QSPR*, Wiley-Interscience, New York, 2000.
6. <http://www.moleculardescriptors.eu/dataset/dataset.htm>
7. <http://www.iamc-online.org/>
8. D. Vukičević, Bond additive modeling 2. Mathematical properties of max-min rodeg index, *Croat. Chem. Acta*, accepted for publication.
9. D. Vukičević, N. Trinajstić, Bond-additive modeling. 3. Comparison between the product-connectivity index and sum-connectivity index, *Croat. Chem. Acta*, accepted for publication.
10. D. Vukičević, Bond additive modeling 4. QSPR and QSAR studies of the variable Adriatic indices, *Croat. Chem. Acta*, submitted.
11. D. Vukičević, Bond additive modeling 5. Mathematical properties of the variable sum exdeg index, *Croat. Chem. Acta*, submitted.
12. D. Vukičević, Bond additive modeling. Adriatic indices – Overview of the results, in: I. Gutman, B. Furtula (Eds.), *Novel Molecular Structure Descriptors – Theory and Applications II*, Univ. Kragujevac, Kragujevac, 2010, pp. 269–302.