Scientific paper

Variable Connectivity Model for Determination of pK$_a$ Values for Selected Organic Acids

Matevž Pompe*,[1] and Milan Randić[2]

[1] Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva 5, 1000 Ljubljana, Slovenia,
Fax: ++386 1 2419 220

[2] National Chemical Institute, Hajdrihova 19, 1000 Ljubljana;
3225 Kingman Rd. Ames, IA 50014, USA;

** Corresponding author: E-mail: matevz.pompe@guest.arnes.si*

Received: 06-02-2007

Abstract

A variable anti-connectivity topological index was optimized for the modeling of pK$_a$ values. The variable anti-connectivity index of order one showed superior modeling capabilities compared to ordinary variable connectivity index of the same order because it is capable to account for the combination of positive and negative contributions for the molecular descriptor in structure-property-relationship. Additionally we examined functional dependence of individual bond contributions on modeling property by varying also the exponent appearing in connectivity index. Such variation did not make significant improvements of the calculated results.

Keywords: Anti-connectivity, variable connectivity index, pK$_a$, prediction

1. Introduction

The proton transfer reactions constitute an important class of chemical reactions and are crucial for studying chemical processes in the solution and therefore also in biological systems. For instance many biological systems use proton-transfer reactions to perform intracellular or extracellular communication. The rate of the proton-transfer reactions depends also on the degree of dissociation of the involved compounds. It is known that using *ab initio* molecular orbital calculations, the gas-phase acidities can be obtained with good accuracy, but the calculation of acidities in the solution phase is still associated with large uncertainties.[1] This deficiency is due to the lack of a complete and precise solvation model. Several procedures are already developed for the description of processes in the aqueous phase. Initial polarized continuum model[2,3] for the calculation solvation free energy was improved by new parameterization of the Langevin dipole model.[4] The overall performance of the model is comparable or slightly better than the polarized continuum model. However, the main advantage of new model is the simplified representation of solvent molecules, so one can gain a clearer insight into the molecular origin of different solvent effects. The mentioned model was successfully applied for the calculation of biologically relevant chemical reactivity.[5] Nevertheless, only very limited numbers of studies are available for the prediction of dissociation constant in the solution phase.[6–9]

Quantitative structure – property relationship (QSPR) modeling, which uses molecular descriptors to represent molecules, that is, topological, electrostatic, geometric and quantum chemical descriptors, can represent alternative to *ab initio* calculations of pK$_a$ values or other molecular property in the solution phase. Such models give especially good predictions when created for family of similar compounds where the same structural feature are influencing the modeled property.[10,11] For instance, empirical atom charge descriptors were used in the combination with multiple linear-regression model for the prediction of pKa values for 1122 aliphatic carboxylic acids and 288 alcohols.[12] However, it should be mentioned that usually the good prediction ability of such models is traded for the lack of structural interpretation of the obtained multiple linear correlation models. The modeling of dissociation is far from straightforward because of opposing influences of individual structural feature on pK$_a$. It is known that an increase of the number of carbon atoms will increase pK$_a$ values, while the presence of strong electron acceptors like halogens will increase dissociation and therefore decrease pK$_a$ value. In order to account for both effects one must construct structural descriptor that is able to capture the positive as well as the negative contributions of critical structural factors within the same molecule. Unfortunately available molecular descriptors are not able to simultaneously account for the presence and role of atoms or fragments that can exert positive and negative contributions to molecular properties. Only recently, however, a modification of the variable connectivity index was suggested, which takes into account besides the positive also possible negative contributions of atoms or bonds in structure-property-activity relationship.[13] The negative portion, that is one the presence of which decreases the overall molecular property was named 'anti-connectivity' region. The modified variable connectivity index of the order zero was used for the modeling of the FID (the flame ionization detector) response factors. Models using anti-connectivity yield considerably smaller calculation error than is the case when only positive additive contributions in construction of the connectivity indices were allowed. In addition to improved calculation ability of the model involving anti-connectivity offers novel structural interpretation, because it gives an information about which part of the molecule is an enhancer and which a suppressor of the modeled property.

In this work we have extended the definition of the anti-connectivity index of zero order to the anti-connectivity indices of the first order and higher orders. We also show that the 'anti-connectivity' phenomenon may be more widespread in structure-property relationship than may have been hitherto anticipated. We will illustrate capabilities of generalized variable connectivity indices on calculation of the proton donor affinity expressed as pK_a value for a selection of organic acids.

2. Calculation of the Modified Variable Connectivity Indices

We will refer to the variable connectivity indices in which the domain of the variables has been extended so that they may introduce negative contributions that characterize the anti-connectivity phenomenon as the variable anti-connectivity indices. The zero order variable anti-connectivity index[13] was developed from the variable connectivity index[14,15] in which in order to differentiate between heteroatoms the adjacency matrix was augmented by inclusion of variables to replace zero diagonal matrix elements. Many early topological indices,[16–18] including the connectivity index[19,20] and other widely used topological indices like the Wiener,[21] the Zagreb,[22] the Balaban,[23] and the Hosoya[24] index did not differentiate heteroatoms. Although that the later developed weighted indices differentiated between heteroatoms,[25–27] they were still not able to account for negative and positive contribution of individual atoms to the modeled property within the same molecule. The modification of the zero order variable connectivity index was introduced to solve this deficiency (Equation 1).

$$ {}^{0}\chi^{f} = \sum_{j=1}^{m} \pm {}^{0}\chi_{j}^{f} = \sum_{j=1}^{m} \pm (\delta_{j}^{f})^{-0.5} \qquad (1) $$

Here m is the number of vertices, ${}^{0}\chi_{j}^{f}$ is contribution of atom j to the connectivity index, while δ_{j}^{f} is the row sum of the augmented adjacency matrix. The zero order connectivity index is especially suitable for modeling atom additive properties and reflects the size of a molecule. In cases where beside size also the molecular branching plays significant role in determining a particular property, the higher order connectivity indices must be used. The expression (1) for zero order anti-connectivity index can be extended and in the case of the first order variable connectivity index becomes:

$$ {}^{1}\chi^{f} = \sum_{j=1}^{m} \pm {}^{1}\chi_{j,i}^{f} = \sum_{j,i=1}^{m} \pm (\delta_{j}^{f} \cdot \delta_{i}^{f})^{-0.5} \qquad (2) $$

The two expressions differ in replacement of the single factor summation terms by two factors terms belonging to connected atoms. Further generalization gives the following expression:

$$ {}^{k}\chi^{f} = \sum_{j=1}^{m} \pm {}^{k}\chi_{j}^{f} = \sum_{j=1}^{m} \pm \prod_{i=1}^{k+1} (\delta_{i}^{f})_{j}^{-0.5} \qquad (3) $$

where Π combines contributions of atoms forming paths of length k to the variable connectivity index. In equations 1–3 the individual contribution becomes negative if at least one of the atoms shows suppressive influence to the property considered.

The prediction abilities of the final model were tested using the leave-one out cross-validation procedure of the whole modeling procedure. It was already shown that just the cross-validation of the final models overestimates the prediction capabilities of the same models.[28] Therefore, individual compounds were omitted from the training set before the optimization procedure. Afterwards, the obtained model was used for the prediction of the omitted compounds. The reported results represent an average of all obtained models generated during leave-one-out cross-validation procedure.

3. Results and Discussion

We report here results obtained for 31 carboxylic and halogenated carboxylic organic acids in which the presence of halogen atoms show 'anti-connectivity' effect on pK_a values. The search for optimal parameters that characterize carbon, oxygen, fluorine, chlorine and bromine atoms starts by selecting at random the initial values for these parameters. Afterwards the variables were optimized using the Simplex optimization algorithm, which was already shown to give good optimization results in cases of variable connectivity indices[29–32] or generalized topological indices.[33, 34] In order to give the lowest RMS error for the linear regression models: $pK_a = b^{*0}\chi^f + a$, $pK_a = b^{*1}\chi^f + a$ or $pK_a = b^{*2}\chi^f + a$, that is simple regression using the zero the first and the second order variable connectivity index. Initially all three indices were restricted only to be composed from positive contributions. The RMS errors found for the three models were 0.904, 0.899, and 0.600, respectively. It is interesting that when individual contributions were restricted to positive values the connectivity index of order two offered the best calculation results. However, the variable anti-connectivity models using ${}^0\chi^f$ and ${}^2\chi^f$ gave slightly worse results compared to ${}^1\chi^f$. At the same time they didn't offer any additional information about the structural features that are influencing pK_a values, that is why, we continued our study with the models using the variable connectivity index of order one.

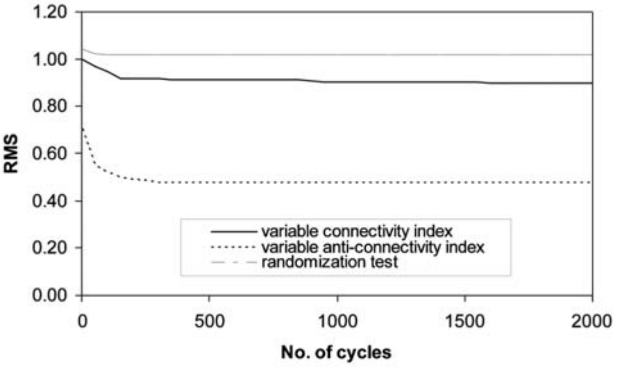

Fig. 1. Optimization of variable connectivity index ${}^1\chi^f$ for the modeling of pK_a values

In Fig. 1 we illustrate the changes of RMS error during optimization of ${}^1\chi^f$. We can see from the figure that the optimization process did not improve the RMS significantly: the RMS error of 1 associated with random parameters was reduced to 0.8986 when optimal state was reached. At this stage the negative anti-connectivity contribution were not considered. We found then that the optimal values for the variables for fluorine and chlorine atoms achieved maximum possible valued of 10^8, selected as the limit for the particular computer program. On one side this means that optimal variables make contributions of the corresponding atoms practically zero and on the other side this signals that fluorine and chlorine are likely candidates for presence of 'anti-connectivity' phenomenon. Subsequently these variables were allowed to enter the 'anti-connectivity' region, the RMS error dropped significantly from 0.899 to 0.581. Moreover, at that moment also the variable for bromine atom reached maximum the possible limiting value of 10^8, which indicates that also bromine atoms lowers the pK_a value and have anti-connectivity character for the considered property. Finally, we allowed all variables representing contributions of halogens to enter the 'anti-connectivity' region. The RMS error now reached the minimum value of 0.473 (Fig. 1). The values of the corresponding variables for carbon, oxygen, fluorine, chlorine and bromine atoms all received large but finite values: $2.819 \cdot 10^5$, $5.846 \cdot 10^3$, (–) $2.0934 \cdot 10^3$, (–) $2.452 \cdot 10^3$, and (–) $5.939 \cdot 10^3$, respectively. The minus sign in the brackets indicates that these variables are in 'anti-connectivity' region. The best regression model was:

$$pK_a = 3.281 \cdot 10^4\ {}^1\chi^f + 2.680. \tag{5}$$

No obvious outliers were detected, but some degeneracy of the calculated values can be observed, that is, the model failed to differentiate position isomers. In general in order to find a good prediction model one must identify the correct influences of the individual structural features on modeled property. Using the variable connectivity index of order one, we have successfully solved this problem. The model correctly predicts the relative influence along the halogen series, that is, fluorine has the strongest effect, than chlorine and bromine.

Further improvement of the regression model is possible if the one considers variation of the connectivity exponent of –0.5. [35,36] We tested this additional modification of variable index, which takes the form:

$${}^1\chi^{f^\lambda} = \sum_{j=1}^{m} \pm {}^1\chi^{f^\lambda}_{j,i} = \sum_{j,i=1}^{m} \pm (\delta_j^f \cdot \delta_i^f)^\lambda \tag{5}$$

where λ is an additional variable. We selected the interval ($-2 < \lambda < +2$) to test the above model. A complete optimization of diagonal elements of augmented connectivity matrix was performed for each pre-selected λ. The variation of λ did not change significantly the quality of the regression model, which appears to be insensitive to λ over large intervals. When λ was changing from –0.2 to –2, the RMS error varied less than 0.3%. The significant increase in RMS error was detected when λ was selected close to zero, that is, when ($-0.1 < \lambda < + 0.1$), the maximum RMS error being when λ is zero (Fig 2.). In this special case the variable connectivity index reduces to the counts of paths of length one. When λ entered positive region we find the smallest calculation error, but again there are no significant changes of RMS over a large interval of λ. It appears that the best results were obtained when λ was +0.4, when the RMS error was reduced to 0.460. One can see from table 1 that new model differentiate between positional isomers, however estimated influence of the halogen atoms at a position is considerably to low, that is error being almost 1 pH unit.

In order to test significance of the obtained models a randomization test was performed as well as leave-one-out cross-validation test of the whole optimization

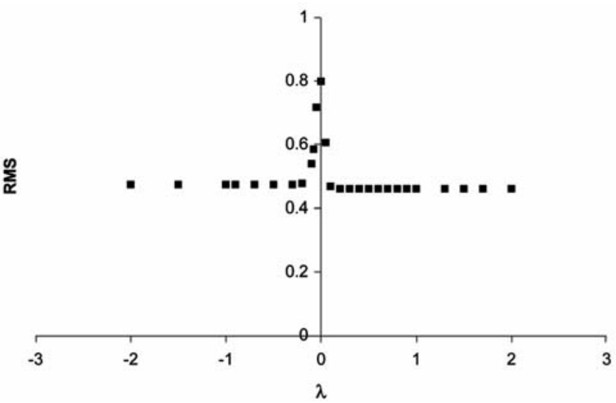

Fig. 2. Changes of RMS error due to the variation of the exponent λ

procedure. After randomization of the pK$_a$ values the optimization procedure failed to reduce error of the model from the initial random value of RMS close to 1 (Fig. 1). Therefore the probability that proposed optimization procedure would find a random model when the RMS error is significantly reduced is low. The final model was validated using leave-one-out cross-validation procedure to further test its significance. The corresponding average variables for carbon, oxygen, fluorine, chlorine and bromine atoms were 1.80, 3543, (–) 1.267 · 10^3, (–) 1.169 · 10^3, and (–) 4.552 · 10^2, respectively. The accompanied final regression model was:

$$pK_a = 3.902 \cdot 10^{-2} * {}^1\chi^{f^{\lambda}} + 0.368. \quad (6)$$

The results of the cross-validated pK$_a$ values are shown in Table 1 and Fig. 3. The cross-validation RMS error was 0.463. The small difference in retrieved and cross-validated RMS error, that is, less than 1%, again points to the fact that the obtained model is significant. At the end it we should stressed that although we have low sensitivity of the exponent λ in the range of –2 to 2 always the best results were obtained when variables for carbon and oxygen were optimized in the connectivity region, that is when their contributions are positive, and when the variables for halogens entered the anti-connectivity region.

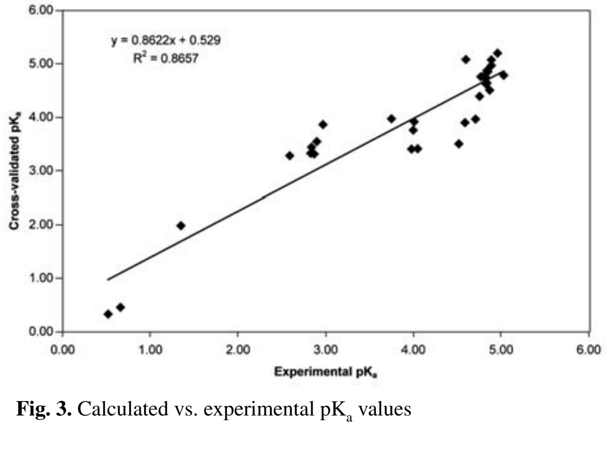

Fig. 3. Calculated vs. experimental pK$_a$ values

Table 1. Experimental and cross-validation pK$_a$ for listed 31 organic acids for both models, that is, pKa = f($^1\chi^f$) and pKa = f($^1\chi^{f^\lambda}$)

ID	Compound name	Experi-mental	pK$_a$ = f($^1\chi^f$)	pK$_a$ = f($^1\chi^{f^\lambda}$)
1	Formic acid	3.75	4.30	4.12
2	Trichloroacetic acid	0.66	0.67	0.46
3	Trifluoroacetic acid	0.52	0.36	0.33
4	Dichloroacetic acid	1.35	1.92	1.98
5	Bromoacetic acid	2.90	3.61	3.67
6	Chloroacetic acid	2.87	3.17	3.31
7	Fluoroacetic acid	2.59	3.06	3.28
8	Acetic acid	4.76	4.41	4.39
9	3-Bromopropanoic acid	4.00	3.73	3.76
10	2-Chloropropanoic acid	2.83	3.28	3.33
11	3-Chloropropanoic acid	3.98	3.28	3.40
12	Propanoic acid	4.87	4.53	4.51
13	2-Chlorobutanoic acid	2.84	3.40	3.44
14	3-Chlorobutanoic acid	4.05	3.40	3.41
15	4-Chlorobutanoic acid	4.52	3.40	3.50
16	Butanoic acid	4.83	4.65	4.62
17	2-Methylpropanoic acid	4.84	4.65	4.64
18	Pentanoic acid	4.83	4.76	4.74
19	2-Methylbutanoic acid	4.80	4.76	4.76
20	3-Methylbutanoic acid	4.77	4.76	4.76
21	2,2-Dimethylpropanoic acid	5.03	4.76	4.79
22	Heptanoic acid	4.89	4.99	4.97
23	Hexanoic acid	4.85	4.88	4.85
24	4-Methylpentanoic acid	4.84	4.88	4.87
25	Octanoic acid	4.89	5.11	5.07
26	2-Propylpentanoic acid	4.60	5.11	5.08
27	Nonanoic acid	4.96	5.23	5.20
28	2-Bromopentanoic acid	2.97	3.96	3.95
29	3-Bromopentanoic acid	4.01	3.96	3.92
30	4-Bromopentanoic acid	4.59	3.96	3.90
31	5-Bromopentanoic acid	4.71	3.96	3.97

The degree of dissociation of organic acids is related to the free energy difference contribution of the proton affinity of the anion, which dominates in the gas phase, and hydration component of acid and anion that become important in the aqueous phase. While in the series of aliphatic carboxylic acids (formic, acetic, propionic ...) the hydration energies are the most important component rather than in vacuo proton affinities, the presence of the strong electron acceptors like halogens increases the importance of the later mentioned component. Therefore any model that predicts pK$_a$ values for halogenated organic acids should encode both mentioned contributions, which is the case also in the developed validated variable connectivity model. The information about both contributions is encoded in the weights and regression parameters of the obtained model, so we are not able to partition this information into above mentioned contributions. However, since we are using bond-additive model we are able to partition the overall effects to the individual functional groups. Such an approach is quite common in QSAR/QSPR studies, for instance, group philicity was used for the modeling of pK_a values of the series of carboxylic acids, various substituted phenols, anilines, phosphoric acids, and alcohols.[37] It was already shown that that such group contribution is valid if we are able to decompose the changes of free energy,[38,39] which govern the mentioned processes. Just recently the conditions of such decomposition have been presented.[40]

4. Conclusions

The aim of this work was to show the application of variable anti-connectivity indices on the modeling of pK_a values for the selected organic and halogenated carboxylic acids. These recently suggested indices are capable to account for a combination of positive and negative contributions of individual atoms and bonds in QSPR studies. Among the three anti-connectivity indices the index of order one gave the best calculation results. The optimization of variables for carbon, oxygen, fluorine, chlorine and bromine atoms reduced significantly the RMS error from around 1 to 0.473. The variables of all three halogenids were optimized in anti-connectivity region, while the contributions for oxygen and carbon atoms remained positive. The variation of the exponent did not improve calculation results substantially. The RMS error was reduced just by another 3% when λ was set to 0.4. The obtained RMS_{CV} error was 0.463. It must be stressed that the most important advantage of variable anti-connectivity indices is, besides improving the calculation model, the ability of structural interpretation of the model. From the optimization of suggested variable indices one can gain information about which part of the molecule enhances and which suppresses the modeled property. However, the currently described index still lack of differentiation between position isomers.

5. Acknowledgments

The authors acknowledge the financial support by the Ministry of Education, Science and Sport of the Republic of Slovenia (Grant P1-0153).

6. References

1. B. J. Smith, L. Radom, *J. Phys Chem.* **1991**, *95*, 10549–10551.
2. S. Miertus; E. Scrocco, Tomasi, *J. Chem. Phys.* **1981**, *55*, 117.
3. S. Miertus, Tomasi, *J. Chem. Phys.* **1982**, *65*, 239.
4. J. Florian, A. Warshel, *J. Phys. Chem. B* **1997**, *101*, 5583–5595.
5. A. Kranjc, J. Mavri, *J. Phys. Chem. A* **2006**, *110*, 5740–5744.
6. G. Schrüümann, M. Cossi, V. Barone, J. Tomasi, *J. Phys. Chem. A* **1998**, *102*, 6706–6712.
7. C. O. Silva, E.C. da Silva, M.A.C Nascimento, *J. Phys. Chem. A* **1999**, *103*, 11194–11199.
8. C. O. Silva, E. C. da Silva, M. A. C. Nascimento, *J. Phys. Chem. A* **2000**, *104*, 2402–2409.
9. M. Namazian, H. Heidary, *J. Mol. Struct. (Theo.)* **2003**, *620*, 257–263.
10. E. Soriano, S. Cerdán, P. Ballesteros, *Mol. Struct. (Theo.)* **2004**, *684*, 121–128.
11. E. Estrada, G. A. Diaz, E. J. Degado, *J Comput. Aid. Mol. Des.* **2006**, *20* 539–548.
12. J. Zhang, T. Kleinöder, J. Gasteiger, *J.Chem. Inf. Mod.* **2006**, 46, 2256–2266.
13. M. Pompe, *Chem. Phys Lett.* **2005**, *404*, 296–299.
14. M. Randić, *Chemometrics Intel. Lab. Systems* **1991**, *10*, 213–227.
15. M. Randić, *J. Comput. Chem.* **1991**, *12*, 970–980.
16. N. Trinajstić, Chemical Graphs Theory, 2nd revised ed. CRC Press, Boca Raton, **1992**.
17. J. Devillers, A.T. Balaban, Eds. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach, Amsterdam, **1999**.
18. R. Todeschini, V. Consonni, The Handbook of Molecular Descriptors, in the Series of Methods and Principles in Medicinal Chemistry, Vol. 11 (Eds: R. Mannhold, H. Kubinyi, H. Timmerman); Wiley-VCH, New York, **2000**.
19. M. Randić, *J.Am.Chem.Soc.* **1975**, *97* 6609–6615.
20. L. B. Kier, L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, London, **1976**.
21. H. Wiener, *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
22. I. Gutman, N. Trinajstić, *Chem. Phys. Lett.* **1972**, *17*, 535–538.
23. A. T. Balaban, *Chem. Phys. Lett.* **1982**, *89*, 399–404.
24. H. Hosoya, *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332–2339.
25. O. Ivanciuc, Rev. Roum. Chim. **2000**, 45, 289–301.
26. O. Ivanciuc, Rev. Roum. Chim. **2001**, 46, 543–552.
27. A. Perdih, F. Perdih, Acta Chim. Slov. **2006**, 53, 180–190.
28. D. M. Hawkins, S. C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579–586.
29. M. Randić, S. C. Basak, M. Pompe, M. Novic, Acta Chim. Slov. **2001**, *48,* 169–180.
30. M. Pompe, M. Veber, M. Randić, A. T. Balaban, Molecules **2004**, *9,* 1160–1176.
31. M. Randić, M. Pompe, D. Mills, S. C. Basak, Molecules **2004**, *9,* 1177–1193.
32. M. Pompe, M. Randić, J. Chem. Inf. Model. **2006**, *46,* 2–8.
33. A. R Matamala, E. Estrada, Chem. Phys. Lett. **2005**, *410,* 343–347.
34. A. R Matamala, E. Estrada, J. Phys. Chem. A **2005**, *109,* 9890–9895.
35. B. Lučić, A. Miličević, S. Nikolić, N. Trinajstić, *Indian J. Chem. Sect A* **2003**, *42*, 1279–1282.
36. A. Miličević, S. Nikolić, *Croat. Chem. Acta* **2004**, *77*, 97–101.
37. R. Parthasarathi, J. Padmanabhan, M. Elango, K. Chitra, V. Subramanian, and P. K. Chattaraj, *J. Phys. Chem. A* **2006**, *110*, 6540–6544.

38. S. Boresch and M. Karplus, *J. Mol. Biol.* **1995**, *254*, 801–807.
39. U. Bren, V. Martínek, J. Florián, *J. Phys. Chem. B* **2006**, *110*, 10557–10566.
40. M. Bren, J. Florián, J. Mavri, U. Bren, *Theor Chem Acc* **2007**, *117*, 535–540

Povzetek

Za modeliranje pK_a vrednosti smo uporabili optimiziran variabilni anti-povezovalni indeks. Variabilni anti-povezovalni indeks prvega reda je pokazal bistveno boljše rezultate pri modeliranju pK_a vrednosti v primerjavi z navadnim variabilnim povezovalnim indeksom, saj dovoljuje pozitivne kot tudi negativne vplive posameznih delov molekule na modelirano lastnost pri študiju povezave med kemijsko strukturo in njihovimi lastnostmi. Dodatno smo raziskali funkcijsko odvisnost posameznih vezi na modelirano lastnost tako, da smo spreminjali tudi koeficient, ki nastopa v izrazu za povezovalni indeks. Te spremembe niso bistveno izboljšale kvalitete modela.