ACADEMIA ROMÂNĂ
*Revue Roumaine de Chimie*
http://web.icf.ro/rrch/

*Rev. Roum. Chim.,*
**2010**, *55*(11-12), 941-946

*Dedicated to the memory of*
*Professor Ioan Silaghi-Dumitrescu (1950 – 2009)*

# MULTIPLE LINEAR REGRESSION AND PARTIAL LEAST SQUARES QSAR MODELING APPLIED TO A SERIES OF ANTIPSYCHOTIC SERTINDOLE DERIVATIVES

Luminița CRIȘAN,[a] Alina BORA,[a] Ludovic KURUNCZI,[b*] Vicențiu VLAIA[b] and Zeno SIMON[a]

[a]Romanian Academy, Institute of Chemistry Timişoara, Mihai Viteazul 24, 300223 Timişoara, Roumania
[b]University of Medicine and Pharmacy "Victor Babes" Timişoara, Faculty of Pharmacy, E. Murgu 2, 300041 Timişoara, Roumania

*Received March 30, 2010*

The aim of the current paper is to make a contribution to the understanding of the molecular features of the 5-substituted derivatives of sertindole that influence their blocking potencies against the α1-adrenergic receptor by applying Multiple Linear Regression (MLR) and a Partial Least Squares (PLS) methodology. These approaches were applied to a series of 33 α1-adrenoceptor antagonists derived from the antipsychotic sertindole. A comparison of the MLR and PLS methods revealed the importance of polar interactions, molecular flexibility, and steric fit in the binding pocket which includes the substituent from position 5 of the indol ring.

## INTRODUCTION

The development of new antipsychotic drugs and their relative efficacy is an important ongoing field of research. Usually, antipsychotics are divided into two groups, the typical or first-generation antipsychotics classified according to their chemical structure and the atypical or second-generation antipsychotics that are classified according to their pharmacological properties.[1] In the last ten years, the discovery of new atypical antipsychotic drugs have revolutionized the pharmacologic treatment of schizophrenia including both negative and positive symptoms, non-schizophrenic psychoses, depression, anxiety, hypertension and other related disorders.[2] All currently discovered atypical antipsychotic drugs such as sertindole (1), olanzapine, risperidole, asenapine, aripiprazole which are available on the market manifest similar or partially blocking effect and particularly high affinities for dopamine D2, serotonin 5-HT2 and α1-adrenergic receptors.[3] The α1-adrenergic receptors are members of the seven-transmembrane domain G protein-coupled receptor superfamily and are subdivided into the α1A, α1B and α1C adrenergic receptor subtypes. Pharmacological, structural and molecular cloning data indicate significant heterogeneity within this receptor family.[4] The non-sedating atypical antipsychotic sertindole (1) is a specific inhibitor of α1A-adrenoreceptor subtype in rat small arteries that binds with nanomolar affinity to the α1A-adrenergic receptor and lower affinity to the α1B and α1D adrenergic receptors.[5] The phenylindole skeleton of sertindole (1) was selected as a template for development of a new class of centrally acting α1-adrenoceptor antagonists.[6] Based on the previous studies,[6, 7] replacement of the 5-chloro atom in sertindole with polar substituents such as heteroaryl, carbamoyl,

* Corresponding author: dick@acad-icht.tm.edu.ro; click.kurunczi@gmail.com

aminomethyl groups afforded a new class of 39 selective α1-adrenergic receptor antagonists. The pre-requisite of developing QSAR equations is the availability of a wide range of molecular structures and their complementary activities.[8] A QSAR equation is a mathematical equation that correlates the biological activity to a wide variety of physico-chemical parameters.[9] Generally, these type of studies were performed on compounds which contained a common skeleton, usually a rigid one, with structural variation limited to functional group changes at specific positions.[10] Even if there are limits in this approach, it permits complex biological systems to be modeled successfully using simple structural parameters and to predict substituent effects for a series of biologically active compounds.[11] In the present, most QSAR studies are rather confined to small data sets and are using both, the classical quantitative structure-activity relationship (2D-QSAR)[12] and also 3D-QSAR approaches[13] giving rise to enhanced knowledge also about antipsychotics drugs and their interactions with different membrane receptors. Molecular models developed in the 3D-QSAR approach give the full structural information, taking into account the nature of atoms in composition, topology, spatial atomic distribution and shape of the molecule. Either 2D, or 3D, these models can be used to improve interpretation of pharmacological data and to predict novel biologically active compounds in the series.[14] Building a bridge between 2D and 3D QSAR, in the last period a great number of 3D-structure dependent, so called whole molecular structure descriptors have been introduced. In the attempt to find the best relation between these and the biological activity, the MLR and the PLS methodology were involved. The advantages of the PLS method are that this enables the employment of collinear descriptors, leads to stable, robust models assuring a better prediction for the unmeasured activities.[15]

The goal of the present paper is to make a contribution to the understanding of the molecular features of the 5-substituted derivatives of sertindole (1) that influence the affinity for the α1-adrenergic receptor by applying MLR and PLS procedures. In the mean time this parallel approach gives the opportunity to compare the quality of the results supplied by the two methodologies.

## METHODS AND MATERIALS

**Experimental data.** From the 39 selective α1-adrenergic receptor antagonists mentioned above, for this study 33 sertindol derivatives with known biological activities were selected (Balle *et all*),[7] and displayed in Table 1. The biological activity for sertindole derivatives were expressed as negative logarithm of the inhibition constant (pKi). The general template of sertindol analogues is depicted in Figure 1.

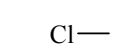
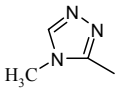
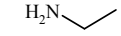
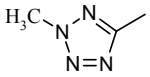
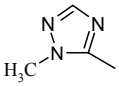
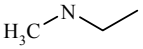
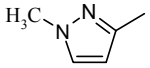
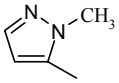
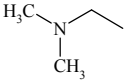
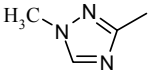
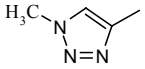
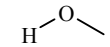
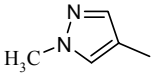
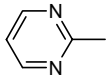
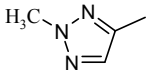
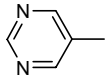
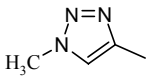
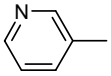
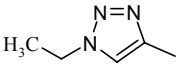
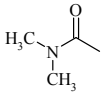
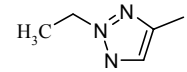
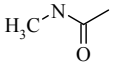
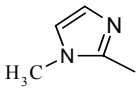
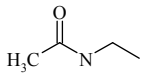
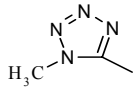
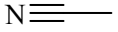
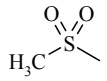
**Structural calculations.** First of all, the 33 investigated molecules were pre-optimized by means of the Molecular Mechanics Force Field (MM+) included in HyperChem version 7.52[16] package. After that, the resulted minimized structures were further refined using the semiempirical AM1 Hamiltonian implemented also in HyperChem. We chose a gradient norm limit of 0.01kcal/Å for the geometry optimization. In order to have the "real" spatial orientation of the subtituents of the indol moiety, a conformational search for all the flexible lateral chains of this rigid skeleton was performed, and only the low energy conformations were retained (0.5 kcal/mole above the lowest). The superposition (see Fig.2) of these conformers on the 3D structure of the most active compounds (26) revealed that the 3D descriptors which will be calculated, will display mainly the structural variation in the substituent position 5. As RMS fit criterion for the superposition three atoms from the rigid indol skeleton were used: C2, C3a, C5 (conventional IUPAC numbering).

**Molecular descriptors calculation.** Constitutional, topological and molecular descriptors were calculated with the DRAGON[17] software. A set of 383 molecular descriptors of different kinds was used to describe the chemical diversity of the 33 compounds. The resulted descriptor classes are: 30 constitutional descriptors, 18 molecular walk-counts, 28 geometrical descriptors, 99 Weighted Holistic Invariant Molecular (WHIM) descripstors, 196 Geometry, Topology, and Atom-Weights Assembly (GETAWAY) descriptors, 9 Functional groups and 3 properties.

**MLR method.** Because the great number of calculated descriptors (K = 383) in comparison with the number of compounds (N = 33) a reliable variable selection method is imperative. The MobyDigs[18] package which uses a Genetic Algorithm approach to sort out the pertinent descriptors determining the variation in the biologic activity was involved in this work. The analysis was restricted to construct only 3 descriptors containing equations. As model selection criterion the squared predicted regression coefficient, $Q^2$, usually employed in the cross-validation procedure[19] was used.

*Table 1*

The structure and biological activity of sertindol derivatives

| No. | R | $pK_i$ | No. | R | $pK_i$ | No. | R | $pK_i$ |
|---|---|---|---|---|---|---|---|---|
| 1 |  | 8.85 | 12 |  | 8.17 | 23 |  | 9.3 |
| 2 |  | 8.74 | 13 |  | 9.35 | 24 |  | 9.34 |
| 3 |  | 8.52 | 14 |  | 8.52 | 25 |  | 9.24 |
| 4 |  | 8.6 | 15 |  | 8.92 | 26 |  | 9.74 |
| 5 |  | 7.65 | 16 |  | 8.69 | 27 |  | 8.89 |
| 6 |  | 8.77 | 17 |  | 8.49 | 28 |  | 9.20 |
| 7 |  | 8.02 | 18 |  | 8.27 | 29 |  | 8.82 |
| 8 |  | 7.96 | 19 |  | 8.41 | 30 |  | 9.00 |
| 9 |  | 8.36 | 20 |  | 8.6 | 31 |  | 8.70 |
| 10 |  | 8.92 | 21 |  | 9.34 | 32 |  | 8.52 |
| 11 |  | 8.89 | 22 |  | 9.34 | 33 |  | 9.24 |

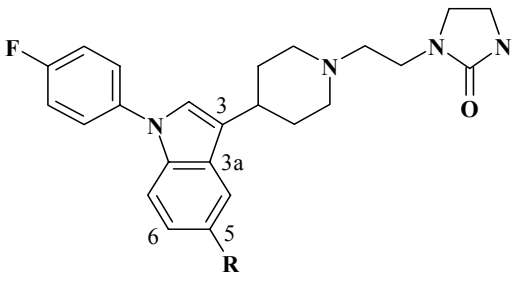



Fig. 1 – The template of sertindole analogues.

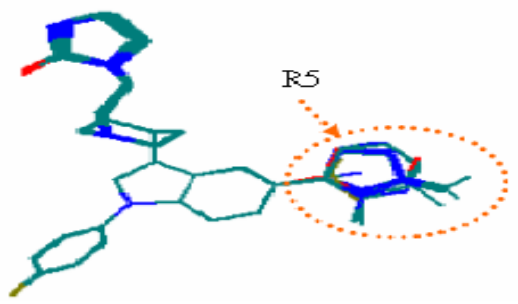



Fig. 2 – The superposition of the 33 sertindole derivatives on the most active compound.

**PLS method.** In the PLS procedure, the relationship between the chemical descriptors and the biological test data are described as a linear model in some latent variables.[20] The QSAR matrix was submitted to the SIMCA P 9.0 package[21] to perform initially a PCA (Principal Component Analysis),[22] and afterwards a PLS analysis. In order to interpret the results of PLS, the equation in latent variables was transformed in function of the original $X_{ij}$ (i =1, 2,..., N; j=1, 2,..., K) variables, resulting Eq. (1) (with $\hat{Y}_i$ the calculated dependent variable, i.e. the calculated pKi value, and $b_j$ the PLS coefficients):

$$\hat{Y}_i = b_0 + b_1X_{i1} + b_2X_{i2} + ... + b_jX_{ij} + ... + b_KX_{iK} \quad (1)$$

The main advantage of PLS in this case is its ability to handle collinearity among the independent variables.

## RESULTS AND DISSCUTION

**MLR analysis.** Several MLR models, (the first five in decreasing order of the squared crossvalidation correlation coefficient presented in Table 2), were suggested by the variable selection procedure. Some statistical parameters concerning the fitting and predictive (cross validation and bootstrapping) performances of the models characterize the selected QSAR equations. WHIM (G1v - the first component symmetry of the atomic volume distribution along the longest latent axis in the space of coordinates weighted by atomic van der Waals volumes; G3m - the third component symmetry directional WHIM index/weighted by atomic masses) and GETAWAY (R2e+ – R maximal autocorrelation of lag 2/weighted by atomic Sanderson electronegativities; H3u - H autocorrelation of lag 3/unweighted; H7m - H autocorrelation of lag 7/weighted by atomic masses; HTe - H total index /weighted by atomic Sanderson electronegativities)[23] descriptors dominate in the equations, in accordance with the findings of Consonni *et al.*[24] concerning the adrenergic blocking potencies of a series of N,N-dimethyl-2-halo-phenethylamines. Besides these, the mean atomic van der Waals volume (Mv), the number of rotatable bonds and rotatable bond fraction (RBN, RBF) and the sum of geometrical distances between the N and F atoms (G(N..F)) seems to be important from the point of view of these models. The difference between the $R^2$ and $R^2_{adj}$ values suggests that some explanatory variable(s) are missing, and other extraneous predictors are added to the model. Also the statistical performances presented in Table 2 do not justify a real choice for a "best" equation, and for the pertinent descriptors. On that account the PLS approach has been applied.

*Table 2*

The first five identified MLR equations in decreasing order of the selection criterion ($Q^2$)*

| Descriptors | R | $R^2$ | $R^2_{adj}$ | $Q^2$ | $Q^2_{boot}$ | SDEP | SDEC | F |
|---|---|---|---|---|---|---|---|---|
| G3m, H3u, R2e+ | 0.818 | 0.669 | 0.635 | 0.582 | 0.635 | 0.298 | 0.265 | 19.55 |
| G3m, G1v, HTe | 0.797 | 0.635 | 0.597 | 0.534 | 0.513 | 0.315 | 0.279 | 16.81 |
| Mv, RBN, H7m | 0.781 | 0.610 | 0.569 | 0.534 | 0.493 | 0.315 | 0.288 | 15.09 |
| Mv, RBF, H7m | 0.780 | 0.608 | 0.567 | 0.528 | 0.489 | 0.317 | 0.289 | 14.97 |
| G(N..F), G3m, H3u | 0.785 | 0.616 | 0.577 | 0.524 | 0.489 | 0.318 | 0.285 | 15.54 |

*R – correlation coefficient, $Q^2$ squared crossvalidation correlation coefficient, $Q^2$ boot – squared bootstrapping correlation coefficient, $R^2_{adj}$ – adjusted $R^2$, SDEP – standard deviation error in prediction, SDEC – standard deviation error in calculation, F – Fischer test.

**PLS analysis.** In the first step of the analysis, a PCA model (M1) was constructed for the whole X matrix (N=33 rows/compounds, and K= 383 columns/descriptors). A 16 principal component model have resulted, but the first two components already explain 64% of the information content of the QSAR matrix. Also the first component distinguish very well between the most active and the most inactive compounds, putting them in opposite senses against the zero value of this component axis.

The first PLS model constructed using the same matrix has led to the M2 model, presented in Table 3 using some statistical characteristics. The separation between the active and inactive compounds in the latent variable space mentioned above (PCA) was well preserved in the PLS model. The diagnostic value analysis for the quality of the model (normal probability plot for Y standardized residuals, and the distance to the model in X-space) have proved that in the limit of the statistical significance level adopted (0.05) there are no meaningful outliers in the series of compounds. Still, the great difference between the $R^2_Y$ and $Q^2$ values for this model proves an overfitting in the regression. Therefore a variable selection was performed, preserving in the new model only the descriptors significantly different from zero in M2 (elimination of noise). In this way M3 was obtained (see Table 3), which presents the greatest $Q^2$ value obtained hereunto, and nearer to the corresponding $R^2_Y$ value, even if the last one is smaller in comparison with that in M2. Otherwise the quality of the model concerning the serious outliers was preserved, and so M3 was considered the model with the best performances obtained. The following analysis refers to this model.

*Table 3*

Statistical characteristics of the deduced PLS models*

| PLS Model | $R^2_X$(CUM) | R | $R^2_Y$(CUM) | $Q^2$(CUM) | RMSEE | A | K |
|---|---|---|---|---|---|---|---|
| M2 | 0.746 | 0.904 | 0.817 | 0.549 | 0.214 | 4 | 383 |
| M3 | 0.715 | 0.853 | 0.727 | 0.610 | 0.257 | 3 | 55 |

* R is the equivalent of regression coefficient from Table 2; $R^2_X$(CUM) and $R^2_Y$(CUM) are the cumulative sum of squares of all the X and Y values, respectively, explained by all extracted principal components; $Q^2$(CUM) is the fraction of the total variation of the Y values that can be predicted for all the A extracted principal components in the cross validation procedure (7 rounds) used to establish the number of significant principal components, i.e. A; RMSEE is the equivalent of SDEC value from Table 2.

*Table 4*

The first ten $b_j$ coefficients from eq. (1) in order of descending VIP* values for model M3

| Variable | VIP* | $b_j$ eq. (1), j = 1, 2, ..., K | Variable | VIP* | $b_j$ eq. (1), j = 1, 2, ..., K |
|---|---|---|---|---|---|
| PSA | 1.314 | 0.005 | L3u | 1.185 | 0.617 |
| RBF | 1.307 | 1.947 | H7p | 1.115 | -1.062 |
| G1v | 1.265 | -16.637 | R8u+ | 1.113 | -4.353 |
| nHDon | 1.264 | 0.093 | H7v | 1.108 | -1.175 |
| RBN | 1.246 | 0.030 | H3v | 1.106 | -0.128 |

* VIP is the Variable Influence on Projections; for descriptors indentifiers: see text.

In PLS a measure of the importance of an $X_j$ variable for both modeling of X and Y is the VIP value: variable importance for the projection. The descriptors with VIP >1 are the most relevant for a model.

In Table 4 the VIP values for M3 are presented in descending order for the first ten most important variables such as (nHDon - number of H-bond donors; PSA - polar surface area; L3u - the third eigenvalue of the PCA performed on the unweighted coordinates; H7p - H autocorrelation of lag 7/weighted by atomic polarizabilities; H7v - H autocorrelation of lag 7/weighted by atomic van der Waals volumes; H3v - H autocorrelation of lag 3/weighted by atomic van der Waals volumes; R8u+ - R maximal autocorrelation of lag 8/unweighted)[23], together with the corresponding $b_j$ coefficients from equation (1).

**Interpretation.** The comparison of the descriptors present in the MLR equations and the M3 PLS model ascertains that G3m, Mv and G(N...F) in PLS were eliminated as variables with $b_j$ insignificantly different from zero. RBF, RBN and G1v are preserved between the first ten descriptors from the PLS equation enforcing their influence on the activity values explanation (H3u is the 13$^{th}$, and HTe the 22$^{th}$ in the VIP order).

The most important variable in M3, PSA, *i.e.* accounts for favorable polar interactions ($b_{PSA} > 0$) between the 5-substituent pocket and the ligand for increasing the activity. The same message is carried by nHDon variable. RBF and RBN (proportional with, or number of rotatable bonds) suggest that some flexibility of a smaller substituent is needed to realize the above mentioned favorable interactions.

The effect of G1v, shows that a greater asymmetry, *i.e* a better filling of the binding pocket will augment the blocking potency of the compound. L3u descriptor is proportional with the molecular size corresponding with the third latent axis. Seeing the positive sign of the coefficient corresponding to this descriptor (see Table 4), its presence in eq.(1) explains the same need for binding space filling as G1v.

The GETAWAY descriptors[25] (H7p, H7v, H3v, R8u+) match 3D-molecular geometry and atom relatedness by molecular topology, with chemical information by using different atomic weightings (as in the case of WHIM descriptors). The H descriptors represent the degree of accessibility to interactions of an atom with other one, the last being situated at a certain topologic distance. The H7 descriptors increase with the molecular dimension, and as can be seen from the sign of the corresponding $b_j$ coefficients (Table 4), concomitantly the activity of the studied series decreases, thus demonstrating that the binding pocket for the analyzed substituents is limited in dimension. R8u+ has a high dependence on conformational changes, and thus it is expected to be related also with the steric complementarity between the ligand and the binding site.

## CONCLUSIONS

The present study emphasizes that often the MLR variable selection method, among the suitable

variables, selects also irrelevant descriptors, probably because the restrictions imposed to eliminate the presence of correlated variables. The PLS methodology, by the adequate handling of the intercorrelated variables, is able to manage this insuperable problem in the MLR. Thus it seems more plausible to consider pertinent descriptors those appearing in both methods. As a result polar interactions, augmented in the case of the small and slender substituents by chain flexibility, and the need of steric complementarity can be identified as important factors influencing the α1-adrenergic blocking ability of the analyzed sertindol derivatives.

*Acknowledgement:* We thank Dr. Erik Johansson (Umetrics, Sweden) for kindly providing the SIMCA program package and to Prof.Dr. M. Mracec for the access to the HyperChem package.

## REFERENCES

1. J. Horacek, V. Bubenikova-Valesova and M. Kopecek , *CNS Drugs*, **2006**, *20,*389-409.
2. L. R. Bryan , D. Sheffler and S. G. Potkinm, *Clinical Neuroscience Research*, **2003**, *3,* 108-117.
3. M. J. Millan, K. Bervoets and F. C. Colpaert, *J. Pharmacol. Exp. Ther.,* **1991***, 256*, 973-982.
4. J. P. Hieble, D. B. Bylund, D. E. Clarke, D. C. Eikenburg, S. Z. Langer, R. J. Lefkowitz, K. P. Minnermann and R. R. Ruffolo Jr., *Pharmacol. Rev.*, **1995**, *47*, 267-260.
5. M. Ipsen, Y Zhang, N. Dragsted, C. Han and M.J. Mulvany, *Eur. J. Pharmacol*., **1997**, *336*, 29-35.
6. T. Balle, J. Perregaard, A. K. Larsen, M.T.Ramirez, K. Krøjer Søby, T. Liljefors and K. Andersen, *Bioorg. Med. Chem*., **2003**, *11*, 1065-1078.
7. T. Balle, J. Perregaard , M.T.Ramirez, A.K. Larsen, K. Krøjer Søby, T. Liljefors and K. Andersen, *J. Med.Chem*., **2003**, *46,* 265-283.
8. P. K. Naik, Sindhura, T. Singh and H. Singh, *SAR and QSAR in Environmental Research*, **2009**, *20,* 551-566.
9. L. M. Shi, Y. Fan, T. G. Myers, and J. N. Paul, *J. Chem. Inf. Comput. Sci.,* **1998**, *38*, 189-199.
10. R. D. Cramer, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, **1988**, *110*, 5959-5967.
11. http://www.netsci.org/Science/Compchem/feature19.html
12. G. Campiani, S. Butini, C. Fattorusso, F. Trotta, S. Gemma, B. Catalanotti, V. Nacci, I. Fiorini, A. Cagnotto, I. Mereghetti, T. Mennini, P. Minetti, M. A. Di Cesare, M. A. Stasi, S. DiSerio, O. Ghirardi, O. Tinti and P. Carminati, *J. Med. Chem*., **2005**, *48*, 1705-1708.
13. V. E. Kuz'min, A. G. Artemenko, P. G. Polischuk, E. N. Muratov, A. I. Hromov, A. V. Liahovskiy, S. A. Andronati and S. Yu. Makan, *J. Mol. Model*, **2005**, *11*, 457-467.
14. A. Ganjee and X. Lin, *J. Med. Chem*., **2005**, *48*, 1448-1469.
15. A. Höskuldsson, *J. Chemometrics.,* **1988**, *2*, 211-228.
16. Hyperchem 7.52 release for Windows; HyperCube, Inc., Gainesville, Florida, USA, http://www.hyper.com
17. Dragon v 3.0, Milano Chemometrics and QSAR Research Group, **2002**.
18. MobyDigs v.1.1. is available from Talete SRL, via V. Pisani, 13-20124, Milano, Italy, http://www.talete.mi.it.
19. D. M. Hawkins, S. C. Basak, and D. Mills, *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 579-586.
20. S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intel .Lab. Syst.*, **2001**, *58*, 109-130.
21. SIMCA P, version 9.0; Umetrics AB: Umea, Sweden. http://www.umetrics.com.
22. M. Daszykowski, K. Kaczmarek, Y. Vander Heyden and B. Walczak, *Chemom. Intel .Lab. Syst.*, **2007**, *85*, 203-219.
23. V. Consonni and R. Todeschini, "Molecular Descriptors for Cheminformatics", Vol. I-II, WILEY, New York, **2009**.
24. V. Consonni, R. Todeschini, M. Pavan and P. Gramatica, *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 693-705.
25. V. Consonni, R. Todeschini and M. Pavan, *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 682-692.