Development and Evaluation of a QSPR Model for the Prediction of Diamagnetic Susceptibility

Antreas Afantitis[a, b, c, d]*, **Georgia Melagraki**[a, b, c, d], **Haralambos Sarimveis**[a], **Panayiotis A. Koutentis**[d], **John Markopoulos**[e] **and Olga Igglessi-Markopoulou**[a]

[a] School of Chemical Engineering, National Technical University of Athens, Athens, Greece
[b] Department of ChemoInformatics, NovaMechanics Ltd, Cyprus
[c] Cyano Research Corporation Ltd, P. O. Box 28670, 2081 Nicosia, Cyprus
[d] Department of Chemistry, University of Cyprus, P. O. Box 20537, 1678 Nicosia, Cyprus
[e] Department of Chemistry, University of Athens, Athens, Greece
* afantiti@mail.ntua.gr

Keywords: Aromaticity, Diamagnetic susceptibility, QSAR, QSPR

Received: June 20, 2007; Accepted: July 30, 2007

DOI: 10.1002/qsar.200730083

Abstract

A novel QSPR model is developed and evaluated for the prediction of diamagnetic susceptibility. The model was produced using the Multiple Linear Regression (MLR) technique on a database that consists of 406 organic compounds involving a diverse set of chemical structures. The accuracy of the QSPR model ($R^2 = 0.88$) is illustrated using various evaluation techniques, such as leave-one-out procedure ($Q^2 = 0.87$) and validation through an external test set ($R^2_{\mathrm{pred}} = 0.89$). The study leads to the conclusion that three physical–topological descriptors affect significantly the diamagnetic susceptibility: Polar Surface Area (PSAr), Principal Moment of Inertia X (PMIX), and Diameter (Diam).

1 Introduction

The design of new materials with optimal thermophysical, mechanical, and optical properties is a challenge for computational chemistry. Novel materials are typically developed using a trial and error approach, which is costly and time-consuming [1]. An alternative strategy is to model the material properties as functions of the molecular structure using the so-called Quantitative Structure–Property Relationships (QSPR) [2–5]. Application of QSPR methodologies in material design has the potential to decrease considerably the time and effort required to improve the material properties in terms of their efficacy or to discover new materials with desired properties.

The diamagnetic susceptibility (χ) of organic compounds is a very important physicochemical property due to the provision of structural information on resolving various existing structural controversies in structural chemistry [6]. According to the definition in the CRC Handbook of Chemistry and Physics [7], when a material is placed in a magnetic field H, a magnetization M is induced in the material which is related to H by $M = \kappa M$, where κ is dimensionless. Usually molar susceptibility ($\chi_{\mathrm{m}} = \kappa V_{\mathrm{m}} = \kappa M/\varrho$) is used where V_{m} is the molar volume of the substance, M the molar mass, and ϱ the mass density. Compounds without any unpaired electrons are called diamagnetic and they have negative values of χ_{m}.

In the past, several attempts have been made for the prediction of diamagnetic susceptibility applying QSPR and semiempirical models [8–9]. The latest QSPR models were presented by Estrada *et al.* [8] and Zhokhova *et al.* [9] using TOSS-MODE and fragment based approaches, respectively. In those studies an extensive bibliographical research was made and the references herein are omitted for brevity. Diamagnetic susceptibility (χ_{m}) is the only measurable property, which is uniquely associated with aromaticity since compounds which exhibit significantly exalted diamagnetic susceptibility are aromatic [10]. An attempt to apply QSPR models for the quantification of aromaticity was made by Duchowicz and Castro [11].

In this work, we present a new linear QSPR model for predicting diamagnetic susceptibility, which contains only three descriptors. Three particular variables were selected among 30 candidates by using the Elimination Selection Stepwise Regression (ES-SWR) variable selection method. The produced model was validated using several strategies: cross-validation, Y-randomization, and external validation using division of the entire dataset into training and test sets. Furthermore, the calculation of the domain of applicability defines the area of reliable predictions.

Supporting information for this article is available on the WWW under www.qcs.wiley-vch.de

2 Materials and Methods

2.1 Dataset

The experimental values of diamagnetic susceptibility of 406 common organic compounds were used for this study. All values refer to room temperature and atmospheric pressure and to the physical form that is stable under these conditions [7]. Diamagnetic susceptibility units are given in the CGS system (multiplied by 4π to convert them into SI). In order to model and predict diamagnetic susceptibility, 30 physicochemical constants, topological and structural descriptors for each chemical structure (Table 1), were considered as possible input candidates to the model. All the descriptors were calculated using ChemSar which is included in the ChemOffice (CambridgeSoft Corporation) suite of programs [12]. Before the calculation of the descriptors, all structures were fully optimized using CS Mechanics and more specifically, MM2 force fields and the Truncated-Newton – Raphson optimizer, which provide a balance between speed and accuracy (ChemOffice Manual).

2.2 Multiple Linear Regression (MLR) Model Development – Variable Selection

The first objective was to determinate the best variables which produce the most significant linear QSPR models linking the structure of compounds with their diamagnetic susceptibility. The ES-SWR algorithm was used on the training dataset to select the most appropriate descriptors. ES-SWR is a popular stepwise technique [13] that combines Forward Selection (FS-SWR) and Backward Elimination (BE-SWR).

2.3 Model Validation

The accuracy of the proposed MLR model was illustrated using the following evaluation techniques: leave-one-out and leave-five-out cross-validation procedures, validation through an external test set, and Y-randomization.

2.4 Cross-Validation Test

Cross-validation is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified datasets are created by deleting in each case one or a small group (leave-some-out) of objects. For each dataset, an input – output model is developed, based on the utilized modeling technique. The model is evaluated by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been used in the development of the model) [14, 15].

2.5 Validation Through the External Validation Set

According to the Tropsha group [16, 17] a QSAR model is considered predictive, if the following conditions are satisfied:

$$R^2_{\text{pred}} > 0.6 \qquad (1)$$

$$\frac{(R^2 - R_0^2)}{R^2} \text{ or } \frac{(R^2 - R_0'^2)}{R^2} \quad \text{islessthan0.1} \qquad (2)$$

$$k \text{ or } k' \text{ is close to } 1 \qquad (3)$$

In Eqs. (2 and 3) R^2 is the coefficient of determination between experimental values and model prediction on the training set. Mathematical definitions of R_0^2, $R_0'^2$, k, and $k'$are based on regression of the observed activities against the predicted activities and *vice versa* (regression of the

Table 1. Physicochemical constants, topological and structural descriptors.

ID	Description	Notation	ID	Description	Notation
1	Molar refractivity	MR	16	Diameter	Diam
2	Partition coefficient (octanol water)	ClogP	17	Molecular topological index	TIndx
3	Principal moment of inertia Z	PMIZ	18	Number of rotatable bonds	NRBo
4	Principal moment of inertia Y	PMIY	19	Polar surface area	PSAr
5	Principal moment of inertia X	PMIX	20	Radius	Rad
6	Connolly accessible area	SAS	21	Shape attribute	ShpA
7	Connolly molecular area	MS	22	Shape coefficient	ShpC
8	Total energy	TotE	23	Sum of valence degrees	SVDe
9	LUMO energy	LUMO	24	Total connectivity	TCon
10	HOMO energy	HOMO	25	Total valence connectivity	TVCon
11	Balaban index	BIndx	26	Wiener index	WIndx
12	Dipole length	DPLL	27	Electronic energy	ElcE
13	Repulsion energy	NRE	28	Connolly solvent-excluded volume	SEV
14	Ovality	Ovality	29	Cluster count	ClsC
15	Sum of degrees	SDeg	30	Molecular weight	MW

predicted activities against observed activities). The definitions are presented clearly in Ref. [18] and are not repeated here for brevity.

2.6 Y-Randomization Test

This technique ensures the robustness of a QSPR model [16–18]. The dependent variable vector (diamagnetic susceptibility) is randomly shuffled and a new QSPR model is developed, using the given modeling algorithm. The procedure is repeated several times and the new QSPR models are expected to have low R^2 and Q^2 values. If the opposite happens then an acceptable QSPR model cannot be obtained for the specific modeling method and data.

2.7 Defining Model Applicability Domain

In order for a QSAR model to be used for screening new compounds, its domain of application [16–18] must be defined and predictions for only those compounds that fall into this domain may be considered reliable. *Extent of Extrapolation* [16] is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage h_i [19] for each chemical, where the QSPR model is used to predict its property.

$$h_i = x_i(X^T X)^{-1} x_i^T \qquad (4)$$

In Eq. (4), x_i is the row vector containing the k model parameters of the query compound and X is the $n \times k$ matrix containing the k model parameters for each one of the n training compounds. A leverage value greater than $3k/n$ is considered large. It means that the predicted response is the result of a substantial extrapolation of the model and may not be reliable.

3 Results and Discussion

First, the dataset of 406 organic compounds was partitioned into a training set of 203 compounds, and a validation set of 203 compounds. The dataset was partitioned to provide a representative training set (Supplementary Material Table 1) and at the same time a diverse test set (Supplementary Material Table 2) in terms of molecular structure [20]. More specifically, the selection of the compounds formulating the training set was made according to the structure and the scale of the investigated property, so that representatives of a wide range of structures (with respect to different substituents, atoms, and values of diamagnetic susceptibility) were included. Additional effort was taken to achieve similar distributions of property values in the training and validation sets. According to Golbraikh and Tropsha [21] this approach is correct since representative compounds in the test set must be close to those of the training set and *vice versa*. The validation data (Supplementary Material Table 2) were not involved by any means in the process of selecting the most appropriate descriptors or in the development of the QSPR model. They were considered as a completely unknown external set of data, which was used only to test the accuracy of the produced model.

For the selection of the most important descriptors, the aforementioned stepwise multiple regression technique was used on the training dataset. The procedure was automated using a software that has been developed in our laboratory on the MatLab platform, which realizes the ES-SWR algorithm. The result was the following four-parameter (three descriptors and the intercept) MLR QSAR equation:

$$-\chi_m{*}10^{-6}(\mathrm{CGS}) = 25.8 - 0.31\mathrm{PSAr} + 0.08\mathrm{PMIX} + 9.84\mathrm{Diam} \qquad (5)$$

RMS = 10.57, $R^2 = 0.88$, $F = 500.03$, $Q^2 = 0.87$, $S_{\mathrm{Press}} = 11.06$, $n = 203$

From the above equation, it can be concluded that the most significant descriptors according to the ES-SWR algorithm are Polar Surface Area (PSAr), Principal Moment of Inertia X (PMIX), and Diameter (Diam). Table 2 presents the correlation matrix, where it is clear that the three selected descriptors are completely uncorrelated. It should be mentioned here that the produced model does not preclude any causal relationship between the diamagnetic susceptibility and the three selected descriptors.

A brief explanation of the descriptors that were selected is as follows:

PSAr [13] is defined as the part of the surface area of the module associated with oxygen, nitrogen, sulfur, and the atoms of hydrogen bonded to any of these atoms.

The Principal Moments of Inertia (PMI) (g/mol Å^2) are physical quantities related to the rotational dynamics of a module [13]. The PMIs are the moments of inertia corresponding to that particular unique orientation of the axes for which one of the three moments has a maximum value (PMIZ), another a minimum value (PMIY), and the third (PMIX) is either equal to one or intermediate between the other two [13]. In this case, the products of inertia tensor matrix are zero and the three diagonal elements, PMIX, PMIY, and PMIZ correspond to the moments of inertia about the X, Y, and Z axes of the module. The ES-SWR algorithm identifies PMIX as a significant descriptor for modeling the diamagnetic susceptibility.

Table 2. Correlation matrix of the three selected descriptors.

	Diam	PSAr	PMIX
Diam	1		
PSAr	0.223	1	
PMIX	0.184	−0.039	1

Diam is the maximum such value for all atoms and is held by the most outlying atom(s). In terms of graph theory, diameter is defined as the largest vertex eccentricity in the graph [13].

Equation (5) was used to predict the diamagnetic susceptibility for the validation examples. The results are presented in the last columns of Table 2 of Supplementary Material and correspond to the following statistics: $R^2_{\text{pred}} = 0.89$ and RMSE = 10.27. A small percentage (5%) of the test set falls outside the domain of the model (warning leverage limit 0.059). The results illustrated once more that the linear MLR technique combined with a successful variable selection procedure is adequate to generate an efficient QSPR model for predicting the diamagnetic susceptibility of a large diverse set of compounds.

The proposed model (Eq. 5) passed all the tests related to the predictive ability (Eqs. 1–3).

$$R^2_{\text{pred}} = 0.89 > 0.6$$

$$\frac{(R^2 - R^2_0)}{R^2} = -0.07 < 0.1$$

$$k = 0.98$$

For a more exhaustive testing of the predictive power of the model, validation of the model was also carried out using the LOO and the L5O cross-validation techniques on the training set of compounds. The L5O method was implemented by selecting randomly groups of five compounds from the available training data. Each group was left out and that group was predicted by the model developed from the remaining observations. Three thousand random groups of five compounds were selected for the implementation of the L5O cross-validation test. It should be emphasized that the procedure for developing the QSPR models included the selection of the best descriptors. Therefore, each time one (LOO) or five (L5O) compounds were excluded from the training set, the modeling procedure selected the best descriptors and developed an MLR model based only on the remaining observations. The excluded compounds were not involved by any means in the development of the model. It was important that the model was stable to the inclusion/exclusion of compounds. The results produced by the LOO ($Q^2 = 0.87$) and the L5O ($Q^2_{\text{L5O}} = 0.83$) cross-validation tests illustrated the quality of the obtained model.

The model was further validated by applying Y-randomization. Several random shuffles of the Y vector were performed and the low R^2 and Q^2 values that were obtained showing that the good results in the original model use not due to a chance correlation or structural dependency of the training set. It should be noted that for each random permutation of the **Y** vector, the complete training procedure was followed for developing the new QSPR model, including the selection of the most appropriate descriptors. The results of the Y-randomization test are presented in Table 3.

According to the proposed QSPR model, high values of PMIX and the diameter of the module correspond to increased diamagnetic susceptibility. PMIX gives information about how the product of mass and distance influence the value of diamagnetic susceptibility.

On the other hand, a high value of the PSAr contributes negatively to the diamagnetic susceptibility. The introduction of groups with high diameter (Diam) is recommended to increase the diamagnetic susceptibility and hence the aromaticity of the module [10]. In contrast, the presence of atoms of oxygen, nitrogen, sulfur, and the atoms of hydrogen bonded to any of these atoms increases the PSAr value and should be avoided.

The proposed method, due to the high predictive ability [22], could be a useful aid to the costly and time consuming experiments for determining the diamagnetic susceptibility. The method can also be used to screen existing databases or virtual chemical structures to identify organic compounds with desired diamagnetic susceptibility. In this case, the applicability domain will serve as a valuable tool to filter out "dissimilar" chemical structures.

4 Conclusions

A novel QSPR model was developed that can predict diamagnetic susceptibility using molecular descriptors. Using a dataset of 406 common organic compounds and a rigorous variable selection method, three descriptors were chosen out of the 30 different descriptors that were examined. Several validation techniques illustrated the accuracy of the produced model, not only by calculating its fitness on sets of training data but also by testing the predicting abilities of the model. The encouraging results showed that the QSPR methodology overcomes several of the limitations experienced by empirical models. The physicochemical constants, quantum, topological and structural descriptors used in QSPR encode information about the structure of the module and thus implicitly account for cooperative effects between functional groups and charge re-

Table 3. R^2 and Q^2 values after several Y-randomization test.

Iteration	R^2	Q^2
1	0.15	0.02
2	0.16	0.04
3	0.05	0.00
4	0.35	0.12
5	0.08	0.01
6	0.28	0.10
7	0.19	0.08
8	0.09	0.01
9	0.30	0.13
10	0.27	0.09

distribution. The selected descriptors offer clear physical meaning and help the researcher to design novel chemistry driven molecules with desired diamagnetic susceptibility.

Acknowledgements

G. M. and A. A. would like to thank the Cyprus Research Promotion Foundation (grants no. KINHT/0505/03, PENEK/ENISX/10603/05 and PLYPH/0506/25).

References

[1] K. V. Camarda, C. D. Maranas, *Ind. Eng. Chem. Res.* **1999**, *38*, 1884 – 1892.

[2] X. Yu, X. Wang, H. Wang, X. Li, J. Gao, *QSAR Comb. Sci.* **2005**, *2*, 151 – 161.

[3] J. Xu, L. Liu, W. Xu, S. Zhao, D. Zuo, *J. Mol. Graph. Mod.* **2007**, *26*, 352 – 359.

[4] A. P. Toropova, A. A. Toropov, S. K. Maksudov, *Chem. Phys. Lett.* **2006**, *428*, 183 – 186.

[5] A. Afantitis, G. Melagraki, H. Sarimveis, P. A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, *Polymer* **2006**, *47*, 3220 – 3248.

[6] R. R. Gupta, R. Swaroop, M. Kumar Kishan, *J. Am. Chem. Soc.* **1984**, *106*, 4378 – 4380.

[7] D. R. Lide (Ed.), *CRC Handbook of Chemistry and Physics* (85th Ed.), CRC Press, Boca Raton, FL **2005**.

[8] E. Estrada, Y. Gutierrez, H. Gonzalez, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1386 – 1399.

[9] N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov, N. S. Zefirov, *J. Struct. Chem.* **2004**, *45*, 626 – 635.

[10] P. R. Schleyer, H. Jiao, *Pure Appl. Chem.* **1996**, *68*, 209 – 218.

[11] P. R Duchowicz, E. A Castro, *J. Theor. Comput. Chem.* **2004**, *3*, 145 – 153.

[12] CambridgeSoft Corporation, Chemoffice Suite.

[13] R. Todeschini, V. Consonni, R. Mannhold, H. Kubinyi, H. Timmerman (Eds.), *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim **2000**, p. 112, 330, 352.

[14] D. W. Osten, *J. Chemom.* **1998**, *2*, 39 – 48.

[15] B. Efron, *J. Am. Stat. Assoc.* **1983**, *78*, 316 – 331.

[16] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 1 – 9.

[17] M. Shen, C. Beguin, A. Golbraikh, J. Stables, H. Kohn, A. Tropsha, *J. Med. Chem.* **2004**, *47*, 2356 – 2364.

[18] A. Golbraikh, A. Tropsha, *J. Mol. Graph. Mod.* **2002**, *20*, 269 – 276.

[19] A. C. Atkinson, *Plots, Transformations and Regression*, Clarendon Press, Oxford, UK **1985**, p. 282.

[20] A. Afantitis, G. Melagraki, H. Sarimveis, P. A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, *QSAR Comb. Sci.* **2006**, *25*, 928 – 935.

[21] A. Golbraikh, A. Tropsha, *Mol. Div.* **2000**, *5*, 231 – 243.

[22] A. O. Aptula, N. G. Jeliazkova, T. W. Schultz, M. T. D. Cronin, *QSAR Comb. Sci.* **2005**, *24*, 385 – 396.