Mol Divers (2010) 14:821–827
DOI 10.1007/s11030-009-9186-0

SHORT COMMUNICATION

QSAR modelling of the toxicity to *Tetrahymena pyriformis* by balance of correlations

A. A. Toropov · A. P. Toropova · E. Benfenati · A. Manganaro

Received: 1 April 2009 / Accepted: 26 July 2009 / Published online: 14 August 2009

Abstract Balance of correlations is an approach to build up quantitative structure–property/activity relationships (QSPR/QSAR). This approach is based on a split into the subtraining, calibration and test sets instead of classic split into training and test sets. The function of the calibration set is the preliminary check up of the model. In other words, the calibration set is like a preliminary test set. Computational experiments (with the Monte Carlo method) have shown that the statistical characteristics of the prediction for the toxicity to *Tetrahymena pyriformis* (the 50% growth inhibition concentration, IGC_{50}) based on the balance of correlations are better than the statistical characteristics of the prediction based on the classic scheme.

Keywords Toxicity · QSAR · *Tetrahymena pyriformis* · SMILES · Optimal descriptor

Abbreviations

QSPR	Quantitative structure–property relationships
QSAR	Quantitative structure–activity relationships
SMILES	Simplified molecular input line entry system
InChI	International chemical identifier
IGC_{50}	The 50% growth inhibition concentration
DCW	Descriptor of the correlation weights

Electronic supplementary material The online version of this article (doi:10.1007/s11030-009-9186-0) contains supplementary material, which is available to authorized users.

A. A. Toropov · A. P. Toropova
Institute of Geology and Geophysics, Khodzhibaev St.49, 100041 Tashkent, Uzbekistan

A. A. Toropov (✉) · A. P. Toropova · E. Benfenati · A. Manganaro
Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy
e-mail: aatoropov@yahoo.com

Introduction

Quantitative structure–property/activity relationships (QSPR /QSAR) are tools to estimate physico-chemical and biochemical parameters [1–10]. The number of databases available via the Internet with representation of the molecular structure by SMILES is gradually increasing. Under such circumstances, the searching of approaches for QSPR/QSAR analysis which are based on SMILES becomes logical [7–10].

The main problem in any QSPR/QSAR analysis is the evaluation and control of the predictive ability of the developed model. Well-known procedures to secure validation of QSPR/QSAR models (e.g., leave-one-out, leave-many-out) are faced with serious criticism [11–13]. The recently suggested correlation balance of SMILES-based optimal descriptors is an attempt to increase the robustness of SMILES-based QSAR models [14,15].

The toxicity of phenols to *Tetrahymena pyriformis* is an important biochemical and ecologic parameter [16–20]. Scopus[1] gives 2,943 items for query "*Tetrahymena pyriformis*". The main relevance of this endpoint is to characterize aquatic toxicity. According to [20], the IGC_{50} is largest amount of aqueous toxicity information.

The aim of the present study is to estimate the ability of the correlation balance of SMILES-based optimal descriptors for

[1] Scopus is the largest abstract and citation database of research literature and quality web sources. Scopus is available at www.scopus.com.

QSAR modelling of the toxicity of phenols to *Tetrahymena pyriformis*.

Method

The split into the training set ($n = 200$) and the test set ($n = 50$), the numerical data on toxicity of phenols to *Tetrahymena pyriformis* (the 50% growth inhibition concentration, IGC_{50} was expressed in logarithmic units $\log(IGC50)^{-1}$ or $pIGC_{50}$), and the SMILES notations were taken from [16]. The training set from [16] has been split into a subtraining and calibration set of similar size. Two criteria have been used to split the training set into subtraining ($n = 105$) and calibration ($n = 95$): (a) it has been done randomly and (b) we verified that ranges of the endpoint the subtraining and calibration were similar. In fact, the calibration is a preliminary test set. By taking into account statistical status of the model for the calibration set, one can avoid having excellent statistics for the training set and poor statistics for the external test set (overtraining).

The descriptors used in this study have been calculated with the SMILES attributes. The SMILES attribute is a combination of SMILES elements. The majority of SMILES elements used in this study contain one symbol, but 'Cl' and 'Br' are represented by two symbols. The modelling approach examined in this study includes three steps [14,15]:

Step 1

Preparation of the list of SMILES attributes for every SMILES notation. Each SMILES attribute is a string of 12 symbols. This string is separated into three zones. The first four symbols as the zone-1; the second four symbols as the zone-2; and the third four symbols as the zone-3.

There are three categories of the SMILES attributes. The first category refers to attributes (${}^1S_k$) containing sole SMILES element positioned in the zone-1; the second category includes attributes (${}^2S_k$) containing two SMILES elements positioned in zone-1 and zone-2; and the third category includes attributes (${}^3S_k$) containing three SMILES elements positioned in zone-1, zone-2 and zone-3. Table 1 contains an example of the preparation of a list of the attributes for a SMILES notation.

In order to avoid a situation when two different SMILES attributes are representing the same molecular fragments, for instance the '(N' and the 'N(', the elements for the ${}^2SA_k$ and ${}^3SA_k$ are ranged according to their ASCII codes. Furthermore, the symbol ')' is replaced by '(', because these are representation of the same phenomenon (i.e. branch in molecular skeleton).

Step 2

Preparation of the completed list of the SMILES attributes which take place in the work set (i.e. totally in the

Table 1 An example of the preparation of a list of the attributes for a SMILES notation; vacant places are indicated by dots

${}^1S_k$ zone-1 zone-2 zone-3	${}^2S_k$ zone-1 zone-2 zone-3	${}^3S_k$ zone-1 zone-2 zone-3
c..........		
1..........	c...1.....	
(..........	1...(.....	c...1...(...
C..........	C...(.....	C...(...1...
C..........	C...C.....	C...C...(...
(..........	C...(.....	C...C...(...
=..........	=...(.....	C...(...=...
O..........	O...=.....	O...=...(...
(..........	O...(.....	=...O...(...
O..........	O...(.....	O...(...O...
(..........	O...(.....	(...O...(...
c..........	c...(.....	c...(...O...
c..........	c...c.....	c...c...(...
c..........	c...c.....	c...c...c...
(..........	c...(.....	c...c...(...
O..........	O...(.....	c...(...O...
(..........	O...(.....	(...O...(...
c..........	c...(.....	c...(...O...
c..........	c...c.....	c...c...(...
1..........	c...1.....	c...c...1...

SMILES="c1(CC(=O)O)ccc(O)cc1";CAS=156-38-7

Table 2 Statistical characteristics models obtained by the classic scheme: training set–test set and by the correlation balance (subtraining set–calibration set–test set)

Probe	Training set, $n = 200$			No calibration set			Test set, $n = 50$		
	r^2	s	F	r^2	s	F	r^2	s	F
Classic scheme									
1	0.7855	0.386	725				0.7247	0.455	126
2	0.7850	0.387	723				0.7295	0.445	129
3	0.7824	0.389	712				0.7203	0.449	124
Average	0.7843	0.387	720				0.7248	0.450	126
Probe	Subtraining set, $n = 105$			Calibration set, $n = 95$			Test set, $n = 50$		
	r^2	s	F	r^2	s	F	r^2	s	F
Correlation balance									
1	0.7662	0.409	338	0.7662	0.397	305	0.7611	0.406	153
2	0.7694	0.406	344	0.7707	0.393	313	0.7687	0.404	160
3	0.7648	0.410	335	0.7650	0.397	303	0.7611	0.410	153
Average	0.7668	0.408	339	0.7673	0.396	307	0.7637	0.407	155

subtraining/training, calibration and test sets). The correlation weights of all SMILES attributes are installed as equal to 1.

Step 3

The optimization of the correlation weights has been done using the Monte Carlo method. The algorithm of the Monte Carlo optimization [15] has been used in two versions. The first is the traditional classic scheme: correlation weights which produce correlation coefficient as large as possible between the DCW and $pIGC_{50}$ on the training set are calculated [10,14,15].

The second scheme, i.e. the balance of correlations is as follows: available data were split into subtraining, calibration and test set (test set taken from [16]). The target function [14,15] of the optimization for this scheme is calculated as

$$CB = Rs + Rc - ABS\,(Rs - Rc) * 0.1 \quad (1)$$

where Rs and Rc are the correlation coefficients between the DCW and $pIGC_{50}$ for the subtraining and calibration set, respectively. The calibration set plays the role of a preliminary test set.

The SMILES-based descriptor is calculated as follows:

$$DCW = \prod CW\left({}^{1}S_k\right) CW\left({}^{2}S_k\right) CW\left({}^{3}S_k\right) \quad (2)$$

where $CW({}^{1}S_k)$, $CW({}^{2}S_k)$, and $CW({}^{3}S_k)$ are the correlation weights for the above mentioned SMILES attributes.

The algorithm of the Monte Carlo optimization is as follows:

1. The regular order of number of attributes i (i.e. 1, 2, 3, 4, 5,…) is replaced by a random sequence k (e.g. 3, 1, 5, 2, 4,...); k:=0;
2. k:=k+1; Calculation of TF1 // Target function (TF) before modify of the $CW({}^{m}SA_k)$
3. $\Delta CW({}^{m}SA_k) := 0.01*CW({}^{m}SA_k)$; $Eps: = 0.1*\Delta CW({}^{m}SA_k)$; // m=1,2,3
4. $CW({}^{m}SA_k) := CW({}^{m}SA_k) + \Delta CW({}^{m}SA_k)$;
5. Calculation of TF2, after modify of the $CW({}^{m}SA_k)$
6. If TF2 > TF1 then TF1:=TF2; go to 4

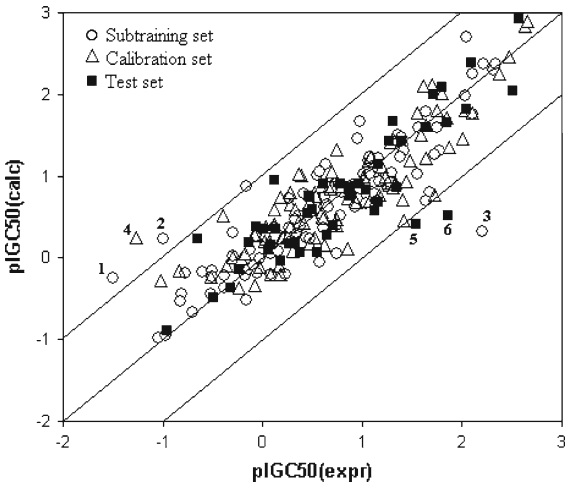


Fig. 1 Plot of experimental versus calculated using Eq. 3 pIGC50 values. Digits 1–6 indicate outliers of the model (Table 2: Probe 1, balance of correlations)

7. $CW(^{m}SA_k) := CW(^{m}SA_k) - \Delta CW(^{m}SA_k)$;
8. $\Delta CW(^{m}SA_k) := -0.5 * \Delta CW(^{m}SA_k)$;
9. If absolute value of the $\Delta CW(^{m}SA_k) > Eps$ then go to 4
10. if $k < N_{SA}$ then go to 2 // N_{SA} is number of the SMILES attributes

The steps of 1–10 is a epoch of the training. In this study 10 epochs have been used for each model.

Table 3 Outliers of model calculated using Eq. 3 (see also Fig. 1)

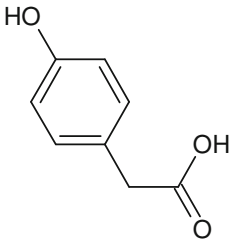
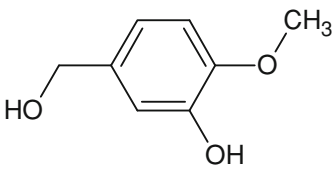
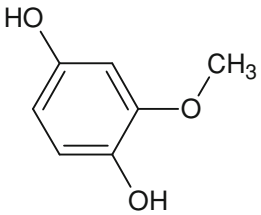
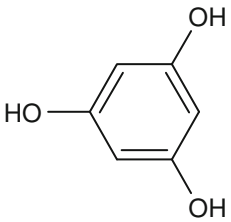
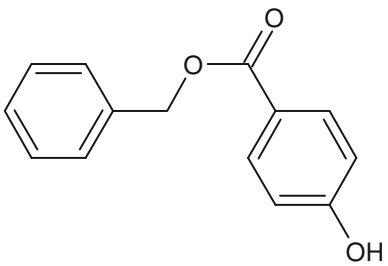
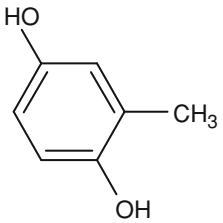
No.	CAS	Structure	Mechanism
1	156-38-7		Polar narcotics
2	4383-06-6		Polar narcotics
3	824-46-4		Pre-electrophiles
4	108-73-6		Polar narcotics
5	94-18-8		Polar narcotics
6	95-71-6		Pre-electrophiles

Table 4 Square correlation coefficients (r^2) of the model calculated using Eq. 3 and model described in [16] for subset of phenols according to mechanism of the activity

	Polar narcotics $n = 138$	Pre-electrophiles $n = 22$	Soft-electriphiles $n = 22$	Pro-redox cyclers $n = 3$	Respiratory uncouplers $n = 15$
Eq. 3	0.8472	0.5374	0.3737	0.1350	0.7997
Ref 16	0.84	0.39	0.25	0.06	0.75

Table 5 Statistical characteristics of SMILES-based model for pre-electrophiles and soft-electrophiles

Probe	Training set, $n = 42$			No calibration set			Test set, $n = 12$		
	r^2	s	F	r^2	s	F	r^2	s	F
Classic scheme									
1	0.7304	0.326	108				0.7284	0.402	27
2	0.7332	0.324	110				0.7186	0.418	26
3	0.7321	0.325	109				0.7140	0.413	25
Average	0.7319	0.325	109				0.7203	0.411	26
Probe	Subtraining set, $n = 24$			Calibration set, $n = 18$			Test set, $n = 12$		
	r^2	s	F	r^2	s	F	r^2	s	F
Correlation balance									
1	0.7249	0.366	58	0.7588	0.585	50	0.8082	0.334	42
2	0.7251	0.366	58	0.7429	0.507	46	0.8200	0.335	46
3	0.7243	0.367	58	0.7375	0.525	45	0.8112	0.329	43
Average	0.7248	0.366	58	0.7464	0.539	47	0.8132	0.333	44

Results and discussion

Table 2 shows the statistical characteristics of the model for the toxicity obtained by the classic scheme (training set–test set) and the statistical characteristics of the model obtained by the balance of correlations (subtraining set–calibration set–test set). One can see that the correlation balance gave a higher r^2 on the test set, which gives information on the predictability of the model. To ensure this, a t test at a significance level of 0.05 has been performed on the two series of resulting r^2, obtaining a *p*-value of 0.000453. This supports that the difference between the mean r^2 values of the two methods is statistically significant, and the values obtained from the balance of correlation method are better. This test has been performed using the STATISTICA 7.1 software.[2]

The first probe of the Monte Carlo optimization (by balance of correlations, Table 2) gives the following model:

$$\begin{aligned} \mathrm{pIGC50} = & -8.6605(\pm 0.0447) \\ & + 5.8200\,(\pm 0.0273) * \mathrm{DCW} \end{aligned} \quad (3)$$

$n = 105, r^2 = 0.7662, s = 0.409, F = 338$ (subtraining set)
$n = 95, r^2 = 0.7662, s = 0.397, F = 305$ (calibration set)
$n = 50, r^2 = 0.7611, s = 0.406, F = 153$ (test set)

The *Supplementary materials* section contains numerical data on the correlation weights for calculation of DCW used in Eq. 3; an example of the DCW calculation; and experimental and calculated values of the $pIGC_{50}$ using Eq. 3. Figure 1 shows this model (calculated using Eq. 3) graphically. Six outliers of this model are indicated in Fig. 1 and listed in Table 3.

The best model described in [16] is characterized by $n = 200, r^2 = 0.71, s = 0.45$ (training set) and $n = 50, r^2 = 0.73$, s=0.44 (test set). The best model described in [17] (after removing of outliers) is characterized by $n = 185, r^2 = 0.83$, $s = 0.34, F = 128$ (training set) and $n = 46, r^2 = 0.78, s = 0.40$, $F = 164$ (test set). Thus, comparing these values with those present in Table 2, it results that our model perform similarly or better than those in [16] and [17] without removing outliers. Outliers have been mainly associated to specific model of actions in [16] and[17]. Thus, we better analysed this issue. Table 4 contains the comparison of the statistical quality of the model calculated using Eq. 3 and model described in [16] for different kinds of chemicals (i.e. polar narcotics,

[2] STATISTICA 7.1 (2006), http://www.statsoft.com.

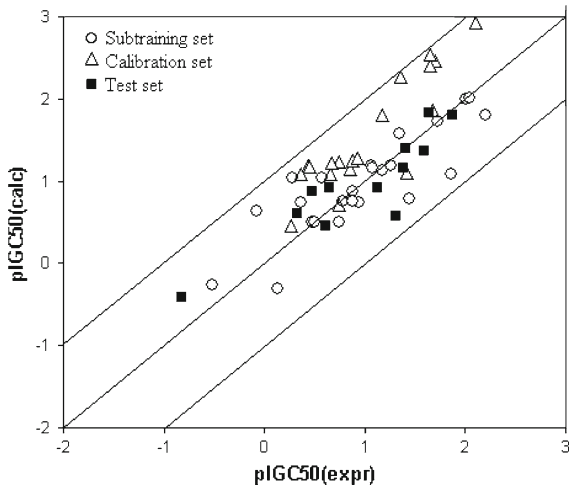


Fig. 2 Plot of experimental versus calculated using Eq. 4 pIGC50 values of model for pre-electrophiles and soft-electrophiles (Table 5: Probe 1, balance of correlations)

pre-electrophiles, soft-electrophiles, etc. [16]). Thus, the Eq. 3 gives prediction of the toxicity with more accuracy.

Table 4 shows that prediction for pre-electrophiles and soft-electrophiles (indicated in [16]) is poor for model calculated using Eq. 3 and for the model described in [16]. The attempt to build up model for a subset that contains these compounds, using the same approach used for the whole set, shows that the model focused on these compounds is satisfactory (Table 5). A t test at a significance level of 0.05 has been performed on the two series of resulting r^2 for the test set, obtaining a *p*-value of 0.000074. This ensures that the difference between the mean r^2 values of the two methods is statistically significant, and the values obtained from the balance of correlation method are better. The test has been performed using the STATISTICA 7.1 software.[2] In particular, the model obtained in the first probe of the balance of correlations for the pre-electrophiles and soft-electrophiles are as follows:

$$\begin{aligned} \text{pIGC50} = & -7.0756(\pm 0.2100) \\ & +5.4057(\pm 0.1376) * \text{DCW} \end{aligned} \quad (4)$$

$$\begin{aligned} n &= 24,\ r^2 = 0.7249,\ s = 0.366,\ F = 58 \text{ (subtraining set)} \\ n &= 19,\ r^2 = 0.7588,\ s = 0.585,\ F = 50 \text{ (calibration set)} \\ n &= 12,\ r^2 = 0.8082,\ s = 0.334,\ F = 42 \text{ (test set)} \end{aligned}$$

In case of the traditional training-test system, the statistical characteristics of the prediction are worse (Table 5). This model in detail is presented in the *Supplementary materials* section. Figure 2 shows this model graphically.

Thus, the applicability domain for the Eq. 3 refers to phenols which are polar narcotics, whereas the applicability domain for Eq. 4 refers to phenols which are pre-electrophiles and soft-electrophiles [16].

Conclusions

SMILES notations can be used for QSAR modelling of toxicity of phenols to *Tetrahymena pyriformis*. The correlation balance scheme (i.e. subtraining set–calibration set–test set) gave a statistically significant improvement for the model in comparison with the classic scheme (training set–test set). The development of models for specific kinds of phenols (i.e. polar narcotics, pre-electrophiles, soft-electrophiles, etc. [16]) can be more robust than the universal model oriented to all 250 compounds which have different mechanisms of toxicity to *Tetrahymena pyriformis*.

Acknowledgements The authors thank the Marie Curie Fellowship (the contract ID 39036, CHEMPREDICT) for financial support.

References

1. Marrero-Ponce Y, Castillo-Garit JA, Castro EA, Torrens F, Rotondo R (2008) 3D-chiral (2.5) atom-based TOMOCOMD-CARDD descriptors: Theory and QSAR applications to central chirality codification. J Math Chem 44: 755–786. doi:10.1007/s10910-008-9386-3
2. Duchowicz PR, Talevi A, Bruno-Blanch LE, Castro EA (2008) New QSPR study for the prediction of aqueous solubility of drug-like compounds. Bioorg Med Chem 16: 7944–7955. doi:10.1016/j.bmc.2008.07.067
3. Roy PP, Roy K (2008) On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci 27: 302–313. doi:10.1002/qsar.200710043
4. Afantitis A, Melagraki G, Sarimveis H, Igglessi-Markopoulou O, Kollias G (2009) A novel QSAR model for predicting the inhibition of CXCR3 receptor by 4-N-aryl-[1,4] diazepane ureas. Eur J Med Chem 44: 877–884. doi:10.1016/j.ejmech.2008.05.028
5. Puzyn T, Mostrag A, Suzuki N, Falandysz J (2008) QSPR-based estimation of the atmospheric persistence for chloronaphthalene congeners. Atmos Environ 42: 6627–6636. doi:10.1016/j.atmosenv.2008.04.048
6. Melagraki G, Afantitis A, Sarimveis H, Igglessi-Markopoulou O, Alexandridis A. (2006) A novel RBF neural network training methodology to predict toxicity to *Vibrio fischeri* Mol Divers. 10: 213–221. doi:10.1007/s11030-005-9008-y
7. Vidal D, Thormann M, Pons M (2005) LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. J Chem Inf Model 45: 386–393. doi:10.1021/ci0496797
8. Vidal D, Thormann M, Pons M (2006) A novel search engine for virtual screening of very large databases. J Chem Inf Model 46: 836–843. doi:10.1021/ci050458q
9. Toropov AA, Benfenati E (2007) SMILES in QSPR/QSAR modelling: results and perspectives. Curr Drug Discov Technol 4: 77–116. doi:10.2174/157016307781483432
10. Toropov AA, Toropova AP, Benfenati E (2009) QSAR modelling for mutagenic potency of heteroaromatic amines by optimal SMILES-based descriptors. Chem Biol Drug Des 73: 301–312. doi:10.1111/j.1747-0285.2009.00778.x
11. Doweyko AM (2004) 3D-QSAR illusions. J Comput Aided Mol Des 18: 587–596. doi:10.1007/s10822-004-4068-0
12. Doweyko AM (2008) QSAR: dead or alive. J Comput Aided Mol Des 22: 81–89. doi:10.1007/s10822-007-9162-7

13. Johnson SR (2008) The trouble with QSAR (or how i learned to stop worrying and embrace fallacy). J Chem Inf Model 48: 25–26. doi:10.1021/ci700332k
14. Toropov AA, Rasulev BF, Leszczynski J (2008) QSAR modelling of acute toxicity by balance of correlations. Bioorg Med Chem 16: 5999–6008. doi:10.1016/j.bmc.2008.04.055
15. Toropov AA, Toropova AP, Benfenati E (2009) Simplified molecular input line entry system-based optimal descriptors: Quantitative structure-activity relationship modelling mutagenicity of nitrated polycyclic aromatic hydrocarbons. Chem Biol Drug Des 73: 515–525. doi:10.1111/j.1747-0285.2009.00802.x
16. Enoch SJ, Cronin MTD, Schultz TW, Madden JC (2008) An evaluation of global QSAR models for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*. Chemosphere 71: 1225–1232. doi:10.1016/j.chemosphere.2007.12.011
17. Cronin MTD, Aptula AO, Duffy JC, Netzeva TI, Rowe PH, Valkova IV, Schultz TW (2002) Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*. Chemosphere 49: 000–1221. doi:10.1016/S0045-6535(02)00508-8
18. Cronin MTD, Gregory BW, Schultz TW (1998) Quantitative structure-activity analyses of nitrobenzene toxicity to *Tetrahymena pyriformis.*. Chem Res Toxicol 11: 902–908. doi:10.1021/tx970166m
19. Cronin MTD, Schultz TW (1997) Validation of *Vibrio fisheri* acute toxicity data: Mechanism of action- based QSARS for non-polar narcotics and polar narcotic phenols. Sci Total Environ 204: 75–88. doi:10.1016/S0048-9697(97)00179-4
20. Duchowicz PR, Ocsachoque MA (2009) Quantitative structure-toxicity models for heterogeneous aliphatic compounds. QSAR Comb Sci 28: a281–295. doi:10.1002/qsar.200860057