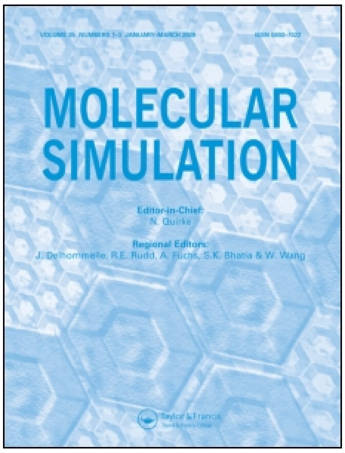


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Predicting thermochemical parameters of oxygen-containing heterocycles using simple QSPR models

N. Adams[a]; J. Clauss[b]; M. Meunier[c]; U. S. Schubert[a]
[a] Laboratory of Macromolecular Chemistry and Nanoscience, Eindhoven University of Technology and Dutch Polymer Institute, Eindhoven, MB, The Netherlands [b] Ticona GmbH, Core Technology, Kelsterbach, Germany [c] Accelrys Ltd, Cambridge, UK

To cite this Article Adams, N. , Clauss, J. , Meunier, M. and Schubert, U. S.(2006) 'Predicting thermochemical parameters of oxygen-containing heterocycles using simple QSPR models', Molecular Simulation, 32: 2, 125 — 134

To link to this Article: DOI: 10.1080/08927020500474300

URL: http://dx.doi.org/10.1080/08927020500474300

Molecular Simulation, Vol. 32, No. 2, 15 February 2006, 125–134

Taylor & Francis
Taylor & Francis Group

Predicting thermochemical parameters of oxygen-containing heterocycles using simple QSPR models

N. ADAMS†, J. CLAUSS‡, M. MEUNIER¶ and U.S. SCHUBERT†*

†Laboratory of Macromolecular Chemistry and Nanoscience, Eindhoven University of Technology and Dutch Polymer Institute, P.O. Box 513, Eindhoven 5600 MB, The Netherlands
‡Ticona GmbH, Core Technology, Professor-Staudinger-Strasse, Kelsterbach 65451, Germany
¶Accelrys Ltd, 334 Cambridge Science Park, Cambridge CB4 0WN, UK

(Received November 2005; in final form November 2005)

Quantitative structure–property relationships for the prediction of standard enthalpies and entropies of formation as well as standard molar heat capacities for small oxygen heterocyclic compounds were developed, using 1D, 2D and 3D descriptors and experimental or computed thermochemical data. To develop the models, the data set was split into test and training sets using D-optimal experimental design to generate a diverse training set. Internal ($R^2_{\text{cross-validated}} = 0.898 - 0.998$) and external ($R^2_{\text{cross-validated}} = 0.847 - 0.996$) validation showed the models to be both stable and highly predictive. Enthalpies of formation were best described by electrotopological, atomic composition and molecular refractivity descriptors, while Kier and Hall χ and κ descriptors as well as the number of rotatable bonds appear frequently in models describing the entropy of formation of these compounds. Heat capacity models often feature the molecular area descriptor as well as the Kier and Hall ${}^0\chi$ descriptor and the number of methyl groups present in the molecule.

Keywords: Thermochemistry; QSPR; Enthalpy of formation; Entropy of formation; Heat capacity; Heterocycles

1. Introduction

The availability of good-quality thermochemical data for small molecules is of great importance for a number of problems in chemistry and chemical engineering. Oxygen heterocycles are key ingredients in a number of industrial processes [1] and their omnipresence in our environment has lead to a significant interest in the way in which they are broken down both in nature as well as in the body [2]. As far as their industrial usage goes, oxygen heterocycles are key ingredients in the manufacture of polyacetals. Polyoxymethylene (POM), for example, is manufactured by polymerizing 1,3,5-trioxane and, if a co-monomer is included, by copolymerizing various 1,3-dioxolanes [3]. Generally, the world market for engineering plastics is growing and, as an example, the demand for POM in China alone was estimated to increase from 140 kt in 2003 to 180 kt in 2005 [4]. Tetrahydrofuran is another important monomer, which has attracted a significant amount of industrial attention recently, with the BASF opening the world's largest polyTHF plant in Caojing (China) in the spring of 2005 [5]. All of this has led to a need to obtain good thermodynamic data for this class of compounds.

A significant amount of effort has been devoted to the development of methodologies for the estimation of enthalpies and entropies of formation as well as molar heat capacities. As early as the 1950s, Benson *et al.* [6–8] published a general method for estimating the thermochemical properties of chemical species on the basis of group additive contributions. The group additive method makes the assumption, that most molecular properties are made up of additive contributions from individual atoms or bonds in the molecule. With the advent of high-performance computing, thermochemical parameters could also be estimated using computational tools, ranging from semi-empirical methods [9] through to DFT [10] and other *ab initio* [11] calculations. Furthermore, Gibbs-ensemble Monte Carlo simulations can also be used to derive thermodynamic properties [12].

In an early study, Lay *et al.* reported thermochemical data for a 34-membered dataset of three-to-six membered oxygen-containing heterocyclic hydrocarbons calculated

*Corresponding author. E-mail: u.s.schubert@tue.nl

Molecular Simulation
ISSN 0892-7022 print/ISSN 1029-0435 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/08927020500474300

using the semi-empirical PM3 method [13,14] and developed a set of group additivity ring corrections for use with Benson's group additivity parameters [15]. The authors later expanded this work and, using a combination of *ab initio* calculations and isodesmic reactions, developed thermochemical and group additive parameters for linear [16] and cyclic alkyl peroxides [17]. In a subsequent study, Shirel and Pulay investigated the stability of oxo- and chloro-substituted trioxanes [18] and Saito and Fuwa conducted an extensive study concerning the thermochemical properties of polychlorinated dibenzo-*p*-dioxins, dibenzofurans and polychlorinated biphenyls using the PM3 Hamiltonian [9]. Notario *et al.* [11] studied dibenzofurans using *ab initio* calculations at the GAUSSIAN-3 G3(MP2)//B3LYP level, albeit with a much smaller compound set. Li *et al.* [19] calculated thermochemical parameters for 76 polybrominated dibenzo-*p*-dioxins using B3LYP/6-31G(d) functional and basis set. To the best of our knowledge, no quantitative structure–property relationships (QSPRs) for thermochemical parameters for small oxygen heterocycles have been developed so far. The present paper aims to fill this gap, using 1D, 2D and 3D descriptors. QSPRs for the prediction of enthalpies of formation were generated on the basis of available experimental data, while models for entropies of formation and heat capacities were, due to the paucity of available experimental data, developed on the basis of validated computed values.

2. Computational procedure

2.1 DFT and semi-empirical calculations

Molecular energies, geometries and vibrational frequencies were determined using DMol^3 [20,21]. Geometry optimizations were performed using general gradient corrected Perdew, Burke, Ernzerhof (PBE), the revised PBE (RPBE) functional [22], the Becke, Lee, Yang, Parr (BLYP) correlation functional [23,24], or the Hamprecht (HCTH) functional [25], using a double numerical basis set including polarization functions (DNP) [20,21]. Optimum structures were confirmed as such by the absence of imaginary vibrations (self-consistent field density convergence: 1×10^{-6} ha). Semi-empirical calculations were carried out using the PM3 method [13,14] as implemented in VAMP [26–28]. Enthalpies and entropies of formation as well as heat capacities were also estimated using a modified version of Benson's group additive method as implemented in an electronic form by the National Institutes of Standards (NIST) [6,29–31].

2.2 QSPR studies

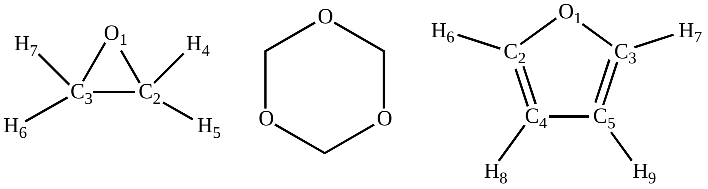
Descriptors were calculated using the MS QSAR 3.2 and TSAR 3.3 software packages [28] and experimental thermochemistry data was taken from the computational chemistry benchmark and comparison database (CCBCD) [32] or the chemistry webbook [29], both maintained by NIST. QSPR equations were developed on the basis of experimental (enthalpies of formation) or computational data (entropies of formation and heat capacities) if insufficient experimental data was available. To develop the QSPR models, D-optimal design was used to split the dataset into a training and a test set. Regression equations were derived using genetic algorithms [33] to select key descriptors. Models were validated by predicting thermochemical properties of the test set molecules.

2.3 Validation of computational data

When comparing the calculated structural data for all combinations of functional and basis set used in this study to experimentally determined values contained in the CCCBD and ref [54] for ethylene oxide (**1**), 1,3,5-trioxane (**2**) and furan (**3**) (figure 1), it could be shown that the latter are reproduced with good to excellent accuracy (table 1).

As expected, the agreement between experimental data and structures computed for the PM3 Hamiltonian is less good. Furthermore, different functionals and basis sets are in good agreement with respect to the computed entropies and heat capacities. All further calculations of entropies and heat capacities using DFT methods were therefore carried out using the PBE functional in connection with a DNP basis set.

In order to determine how accurately DFT methods predict standard entropies of formation as well as molar heat capacities, the geometries of 84 compounds were optimized using the PBE/DNP functional and basis set combination and thermodynamic data were calculated. Although, sufficient experimental data is available, enthalpies of formation were also computed using PM3 and Benson's method. The current commercial implementation of PBE/DNP in DMol^3 is not suitable



(1) **(2)** **(3)**

Figure 1. Ethylene oxide (**1**), 1,3,5-trioxane (**2**) and furan (**3**).

Table 1. Experimentally determined [29,30,54] and calculated geometries for ethylene oxide (**1**), 1,3,5-trioxane (**2**) and furan (**3**).

			BLYP			*RPBE*			*PBE*			*HCTH*			
		Expt.	*DNP*	*DND*	*DN*	*DNP*	*DND*	*DN*	*DNP*	*DND*	*DN*	*DNP*	*DND*	*DN*	*PM3*
(1)	d O1–C2	1.452	1.453	1.452	1.529	1.445	1.445	1.518	1.438	1.437	1.510	1.424	1.424	1.493	1.432
	d C2–C3	1.459	1.475	1.474	1.490	1.475	1.475	1.490	1.438	1.469	1.485	1.461	1.460	1.477	1.484
	d C2–H4	1.084	1.093	1.096	1.091	1.096	1.098	1.093	1.095	1.097	1.092	1.093	1.094	1.090	1.096
	∠ O1–C2–C3	59.203	59.504	59.499	60.846	59.326	59.327	60.615	59.261	59.261	60.542	59.142	59.150	60.357	58.806
	∠ C2–O1–C3	61.594	60.997	61.007	58.318	61.350	61.349	58.786	61.481	61.481	58.923	61.715	61.70	59.283	62.388
	∠ H4–C2–H5	116.750	115.659	115.529	115.898	115.635	115.485	115.919	115.713	115.583	115.901	115.503	115.366	115.551	111.654
	∠ O1–C2–H4	114.704	114.987	115.032	114.042	115.136	115.126	114.183	115.228	115.164	114.334	115.063	115.088	114.355	116.313
	$\Delta S_f^{298.15}$	58.08	59.55	59.55	59.93	59.63	59.60	59.93	59.54	59.52	59.87	59.36	59.31	59.61	58.03
	$C_p^{298.15}$	11.44	11.61	11.61	12.03	11.61	11.83	12.14	11.78	11.71	12.07	11.18	11.08	11.36	11.51
(2)	d O–C	1.422	1.431	1.431	1.477	1.428	1.427	1.471	1.418	1.419	1.463	1.410	1.409	1.451	1.410
	d C–H_{eq}		1.095	1.095	1.089	1.096	1.097	1.091	1.091	1.096	1.090	1.092	1.093	1.087	1.097
	d C–H_{axial}		1.111	1.111	1.107	1.110	1.112	1.108	1.112	1.112	1.108	1.108	1.109	1.104	1.107
	∠ O–C–O	112.2	111.708	111.708	111.127	111.577	111.868	111.268	111.784	111.800	111.211	111.861	111.925	111.297	107.523
	∠ C–O–C	110.3	109.542	109.524	110.124	109.061	109.046	109.773	108.525	108.943	109.620	109.569	109.495	110.585	112.979
	$\Delta S_f^{298.15}$	68.09	71.88	71.88	72.79	71.49	71.51	72.31	71.71	71.36	71.91	74.89	73.75	71.75	70.62
	$C_p^{298.15}$	19.57	20.67	20.67	21.79	20.46	20.41	21.29	20.52	20.23	20.97	20.97	20.68	20.53	21.53
(3)	d O1–C2	1.362	1.382	1.382	1.382	1.378	1.378	1.416	1.371	1.371	1.409	1.362	1.361	1.398	1.378
	d C2–C4	1.361	1.365	1.365	1.382	1.368	1.368	1.374	1.364	1.364	1.370	1.364	1.363	1.368	1.373
	d C4–C5	1.4338	1.438	1.438	1.438	1.437	1.437	1.451	1.431	1.431	1.444	1.426	1.426	1.438	1.441
	d C3–H7	1.0760	1.082	1.082	1.082	1.085	1.085	1.083	1.084	1.084	1.082	1.080	1.082	1.078	1.085
	d C4–H8	1.0760	1.084	1.084	1.084	1.087	1.087	1.086	1.086	1.086	1.085	1.082	1.084	1.081	1.086
	∠ O1–C2–C4	110.700	110.348	110.348	110.348	110.497	110.497	109.677	110.449	110.449	109.603	110.391	110.427	109.683	110.238
	∠ C2–O1–C3	106.60	106.322	106.322	106.322	106.347	106.347	106.171	106.477	106.477	106.309	106.827	106.796	106.850	106.857
	∠ C2–C4–C5	106.00	106.491	106.491	106.491	106.329	106.329	107.237	106.313	106.313	107.242	106.195	106.175	107.042	106.334
	∠ O1–C3–H7	115.90	115.586	115.586	115.586	115.613	115.613	115.461	115.722	115.722	115.572	115.652	115.556	115.570	115.492
	$\Delta S_f^{298.15}$	63.82	65.36	65.36	65.36	65.39	65.39	65.46	65.28	65.28	65.37	64.99	64.95	65.03	64.58
	$C_p^{298.15}$	15.63	15.87	15.87	15.87	15.89	15.89	15.94	15.74	15.74	15.81	15.09	15.03	14.99	15.56

Bond lengths are given in Ångströms (Å), bond angles in degrees (°), $\Delta S_f^{298.15}$ and $C_p^{298.15}$in cal mol^{-1}K^{-1}.

for the calculation of standard enthalpies of formation, as these are calculated using a database of atomic binding energies, which is not currently available for the PBE functional. Tables of experimental and calculated data are given in the supporting information.

PM3 parameters have been optimized to reproduce experimental enthalpies of formation at 298.15 K. Consequently the method performs well when compared to experimental data ($R = 0.984$, $R^2 = 0.968$, SD = 9.21 kcal mol^{-1}). The use of DFT optimized structures and subsequent enthalpy prediction using PM3, did not lead to improved data. The results are in good agreement with those previously obtained by Lay *et al.* [15] with the somewhat higher standard deviation and lower correlation coefficients reflecting the much larger and more diverse set used in the present study. For those cases, for which group additive parameters were available, Benson's method performs well. When comparing experimental and computed data, a correlation coefficient of $R = 0.997$ ($R^2 = 0.993$) and a standard deviation of 3.63 kcal mol^{-1} was determined. *cis*- and *trans*-2,2,4,6-tetramethyl-1,3-dioxin are the only significant outliers in this case and are overestimated by 5.7 and 8.8 kcal mol^{-1}.

Density functional theory, the PM3 Hamiltonian as well as Benson's group additive method (where appropriate) were used to calculate standard entropies of formation. Unfortunately, there is significantly less entropy than enthalpy data available in the literature and, therefore, the calculations could only be validated using a significantly smaller data set (8 datapoints). The risk inherent in such a small dataset is that it could lead to either a serious over- (in cases in which there is a good accidental agreement between experimental and computed data) or underestimation (in case the experimental data is very noisy or there are experimental errors) of the accuracy of the computational methods evaluated here. This also means that any comparison between methods may be affected by a certain amount of uncertainty. In the absence of further data, however, this is the best that can currently be achieved. On this basis, all three methods gave satisfactory results, with DFT giving a slightly better correlation between calculated and experimental values than the other two methods (figure 2).

DFT calculations give rise to a correlation coefficient of $R = 0.963$ ($R^2 = 0.927$) and a standard deviation of 2.72 cal mol^{-1} K^{-1}, whereas, the Benson model gives $R = 0.912$ ($R^2 = 0.832$) and a standard deviation of 3.44 cal mol^{-1} K^{-1}. This is reflected in a certain amount of disagreement between the two models (figure 3).

The major outliers (in addition to the *cis*- and *trans*-2,2,4,6-tetramethyl-1,3-dioxines) are compounds containing an oxetanone or a carbonate motif, probably indicating that the parameterization of the Benson model is not optimal for this type of structures. PM3 delivers results close to those of the DFT calculations ($R = 0.961$, $R^2 = 0.923$, SD = 3.21 cal mol^{-1} K^{-1}).

Regarding the prediction of standard heat capacities, all methods give good to excellent agreement between experimentally determined and calculated heat capacities. Overall, PM3 seems to perform best ($R = 0.979$, $R^2 = 0.958$, SD = 0.70 cal mol^{-1} K^{-1}), followed by Benson's group additive method ($R = 0.968$, $R^2 = 0.938$, SD = 1.93 cal mol^{-1} K^{-1}) and PBE/DNP ($R = 0.935$, $R^2 = 0.875$, SD = 1.24 cal mol^{-1} K^{-1}), although, the differences between the methods are small (figure 4).

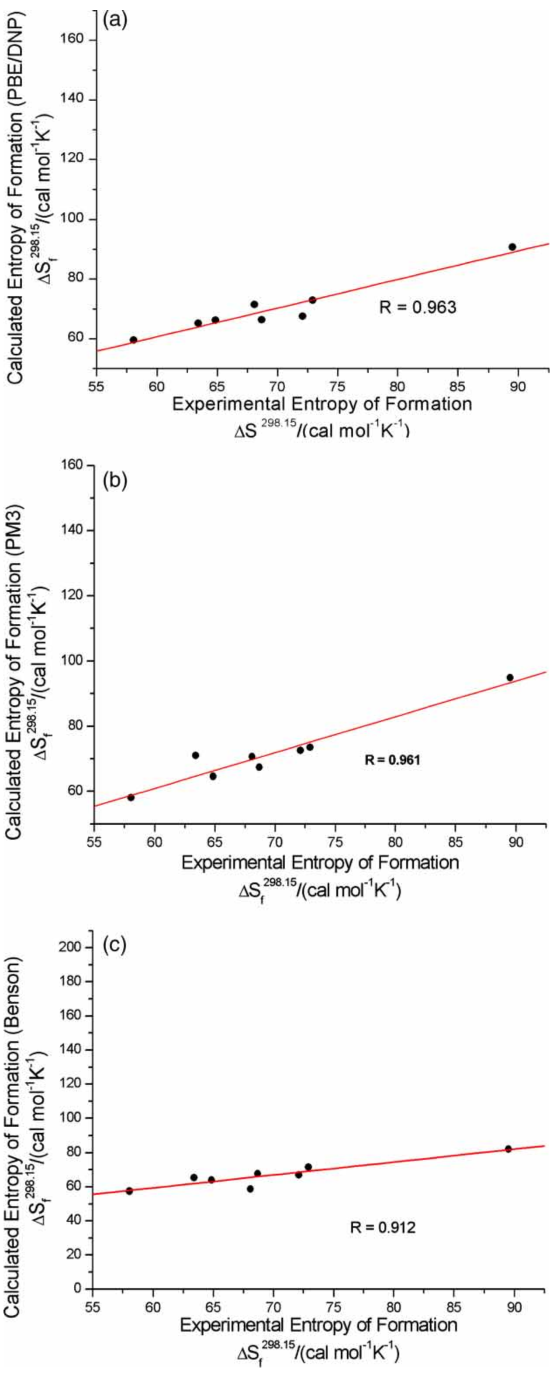


Figure 2. Experimentally determined entropies of formation vs results derived using: (a) DFT (PBE/DNP); (b) PM3 and (c) Benson's group additive method.

Given the fact, that all three different computational methodologies give very good agreement between

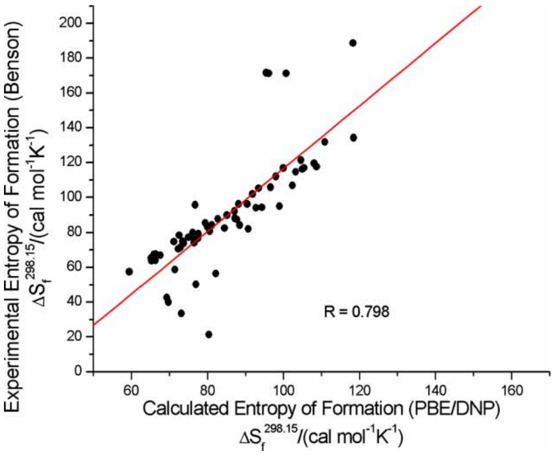


Figure 3. Comparison of entropies of formation calculated using DFT (PBE/DNP) and Benson's group additive method.

predicted and experimentally determined values, it is suggested that these methods provide high-quality data, suitable for the development of QSPRs in the absence of experimental data.

3. Thermochemical data from quantitative structure, property relationships (QSPRs)

3.1 Subset selection

The selection of diverse subsets of molecules for model development is a non-trivial problem and a number of different approaches, such as clustering techniques [34], random selection [35–38], activity sampling [38–40], self-organizing maps [41,42] as well as a number of experimental design approaches [43,44] have been reported in the literature. In a comparative study, Massart *et al.* demonstrated that both D-optimal and Kennard stone designs ultimately led to better models than random sampling or self-organizing maps [36] and other authors have also reported favourable experiences [45–47]. D-optimal designs aim to maximize the determinant of the variance–covariance matrix $|\mathbf{X'X}|$, where $\mathbf{X}$ is the information matrix of independent co-variables. This determinant will be at a maximum for compound sets, which have a maximum variance (i.e. span a large chemical parameter space) and a minimum co-variance (i.e. there is minimum similarity between the molecules) [48]. In a first step, therefore, 126 different 1D, 2D and 3D descriptors were calculated for each compound in the dataset. Subsequent principal component analysis showed that the first 27 principal components explain 99% of the variance in the dataset. The maximum and minimum values of the first seven principal component (80% variance explained) vectors were used as inputs for a D-optimal design, resulting in an ensemble of 46 candidate points in virtual space, representing 55% of the compounds in the dataset (table 2).

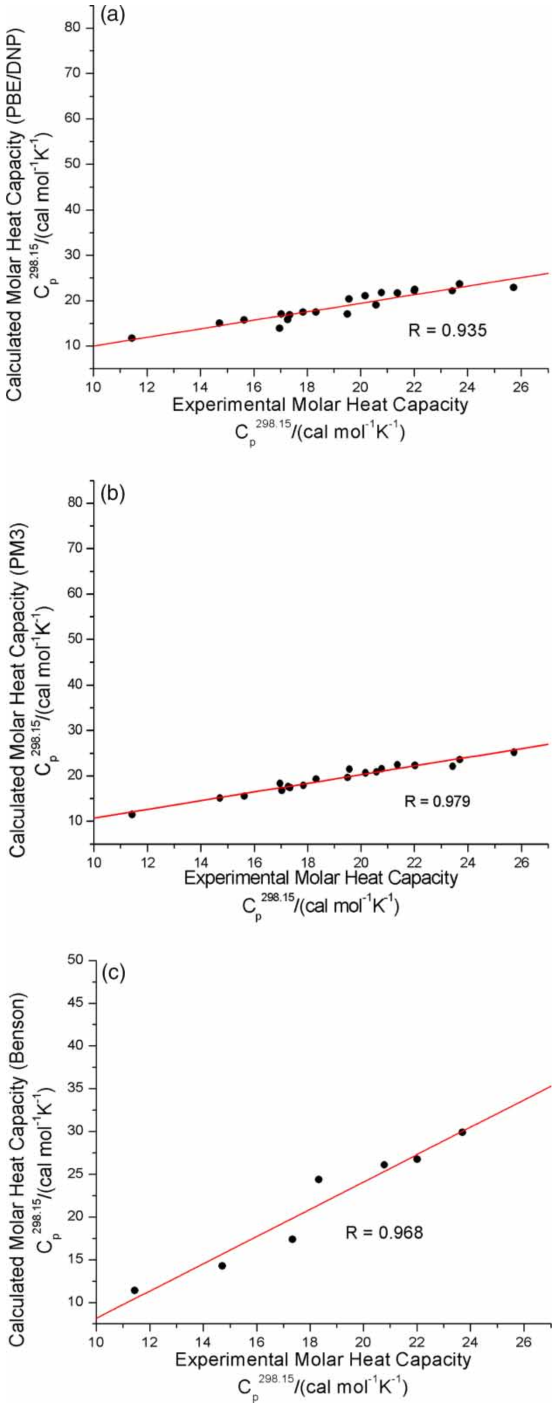


Figure 4. Experimentally determined molar heat capacities vs results derived from: (a) DFT (PBE/DNP); (b) PM3 and (c) Benson's group additive method.

A simple Euclidean distance measure was used to identify those compounds in "real" chemistry space that lie closest to the design points. A visual examination of the scores for the first vs the second principal component shows that the selected training set is diverse and well distributed over the whole dataset (figure 5). Those compounds not included in the training set were used for external validation of the QSPR model (test set).

Table 2. Results from principal component analysis.

Principal component	*Variance explained*	*Cumulative variance*	*Min*	*Max*
1	0.449	0.449	− 1.634	3.841
2	0.100	0.549	− 2.147	2.058
3	0.074	0.624	− 3.526	5.235
4	0.065	0.689	− 2.437	2.985
5	0.048	0.737	− 1.771	2.768
6	0.031	0.768	− 4.603	2.148
7	0.028	0.796	− 2.554	2.722

3.2 QSPR model development

The diverse subset of 46 compounds was used to develop QSPR models for standard enthalpies and entropies of formation and molar heat capacities. Enthalpy models were constructed using experimental data, whereas entropy and heat capacity models were developed using computed data (DFT). Model construction was carried out using genetic algorithm driven linear regression methods [33]. At the beginning of the optimization procedure, 500 equations were randomly selected and evolved until convergence was achieved. To guard against overfitting, the maximum equation length was set to five independent variables in accordance with the recommendation that a regression model with k independent variables and n compounds in the training set, should satisfy the $n > 4k$ criterion (in this study, $n = 46$ and $k = 5$) [49]. Care was taken to penalize equations with both large sum-of-squares errors and large numbers of independent variables [50]. The top five models for standard enthalpies, entropies and molar heat capacities, together with validation data, are given in table 3.

3.2.1 Standard enthalpy of formation. In the case of standard enthalpies of formation, internal as well as external validation suggests that the models are extremely robust: both adjusted (R^2_{adj}) as well as cross-validated (R^2_{cv}) coefficients of determination are not significantly different, indicating that the models are both robust and predictive. The top-performing model (H1) has a coefficient of determination of 0.921 for the training set and 0.852 for the external test set (table 3, figure 6).

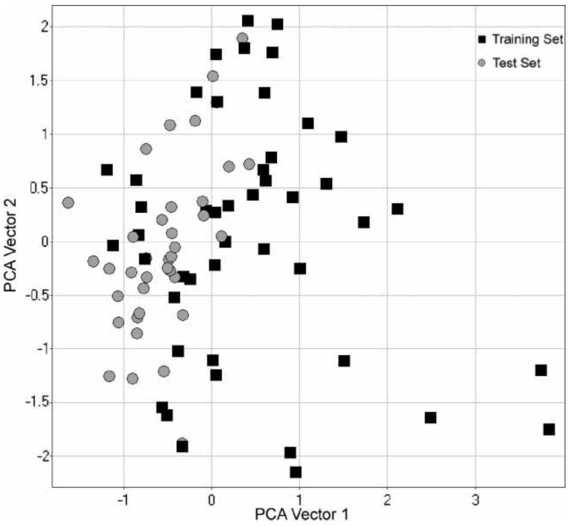


Figure 5. Principal component analysis of all descriptors (PC1–PC2: EV = 55%) for the data set chemicals split into training (■) and test (●) set.

The standard deviation for the training set is 16.75 kcal mol^{-1}. However, it can be seen that a number of outliers are present in the training (O2–O5) and test (O1) sets. All of these, with the exception of O3 (*tert*-butylperoxymethyl oxirane) are furan derivatives and O4–O6 all contain acetoxy-substituents and O6 an additional nitro-group. Removing the outliers from the dataset results in an improved value of R^2 of 0.980 and a standard deviation of 8.45 kcal mol^{-1}. It should be noted that once outliers have been removed from the dataset, the standard deviation is approximately comparable to results obtained from PM3 calculations. Full tables of computed results are given in the supporting information.

The good agreement between experimental and predicted data shows the value of using diverse subsets, such as those generated via D-optimal design for the development of QSPR equations. Examination of the top-performing models shows that a number of descriptors are repeatedly represented. The electrotopological S_ddsN descriptor [51] appears in all five models, closely followed by the S_dssC [51], molecular refractivity and atomic composition descriptors, all at three counts each. Electrotopological descriptors, or E-state indices, were introduced by Kier and Hall [51]. Each atom in a molecular graph is represented by an E-state, which encodes the electronic state of an atom as influenced by the other electronic states of all the other atoms in the molecule, within the context of the molecular graph. The E-state for a given atom, therefore, varies from molecular structure to molecular structure.

The presence of the S_ddsN descriptor and its negative contribution indicates that the presence and number of nitro-groups in the molecule has a significant bearing on the standard enthalpy of formation. Interestingly, three of the six outliers are nitro compounds (O1, O4, O6). The S_dssC descriptor makes a positive contribution to the equations. Again, it is probably not surprising that the descriptor should be present, as presence and number of double bonds in the system can be expected to have a significant bearing on the enthalpy of formation.

Table 3. QSPR Models for standard enthalpies and entropies of formation and molar heat capacities.

		R	R^2	R^2_{adj}			*SD*
	Equation	*(test set)*	*(test set)*	*(test set)*	R^2_{cv}	*F*	*(test set)*
H1	$\Delta H_f^{298.15} = (10.741 \times \text{RB}) - (73.681 \times \text{HA}) - (176.609 \times \text{S\_ddsN}) + (340.444 \times \text{MD}) - 284.461$	0.959	0.921	0.913	0.898	119	16.75
		(0.923)	(0.852)	(0.847)			(13.87)
H2	$\Delta H_f^{298.15} = (74.772 \times \text{HA}) + (14.942 \times {}^3\kappa) - (165.827 \times \text{S\_ddsN}) + (392.964 \times \text{MD}) - 352.086$	0.958	0.918	0.909	0.889	115	17.01
		(0.835)	(0.697)	(0.688)			(18.16)
H3	$\Delta H_f^{298.15} = (7.911 \times \text{MR}) + (32.762 \times \text{SC\_c}) - (13.531 \times \text{AC}) + (43.197 \times \text{S\_dssC}) - (109.266 \times \text{S\_ddsN}) + 7.575$	0.972	0.946	0.938	0.931	138	14.06
		(0.839)	(0.704)	(0.695)			(19.66)
H4	$\Delta H_f^{298.15} = (7.935 \times \text{MR}) + (200.597 \times {}^3\chi) - (13.576 \times \text{AC}) + (43.116 \times \text{S\_dssC}) - (109.465316780 \times \text{S\_ddsN}) + 7.992$	0.972	0.945	0.938	0.930	137	14.17
		(0.834)	(0.696)	(0.687)			(19.86)
H5	$\Delta H_f^{298.15} = (7.935 \times \text{MR}) + (115.815 \times {}^3\chi) - (13.575 \times \text{AC}) + (43.116 \times \text{S\_dssC}) - (109.465 \times \text{S\_ddsN}) + 7.992$	0.972	0.945	0.938	0.930	137	14.17
		(0.834)	(0.696)	(0.687)			(19.86)
S1	$\Delta S_f^{298.15} = (7.773 \times {}^1\kappa) + (23.597 \times {}^3\chi) - (0.004 \times \text{VDM}) - (0.400 \times \text{S\_ssO}) + 45.165$	0.994	0.988	0.987	0.984	826	1.98
		(0.979)	(0.958)	(0.957)			(2.19)
S2	$\Delta S_f^{298.15} = (7.773 \times {}^1\kappa) + (40.8716 \times {}^3\chi) - (0.004 \times \text{VDM}) - (0.400 \times \text{S\_ssO}) + 45.166$	0.994	0.987	0.987	0.984	826	1.98
		(0.979)	(0.958)	(0.957)			(2.19)
S3	$\Delta S_f^{298.15} = (1.883 \times \text{RB}) + (5.346 \times {}^1\kappa) + 53.712$	0.988	0.977	0.975	0.971	876	2.76
		(0.968)	(0.937)	(0.936)			(2.23)
S4	$\Delta S_f^{298.15} = (1.201 \times \text{RB}) + (5.753 \times {}^1\kappa) + (24.050 \times {}^3\chi) + 51.105$	0.991	0.983	0.981	0.978	767	2.39
		(0.976)	(0.953)	(0.952)			(2.04)
S5	$\Delta S_f^{298.15} = (1.201 \times \text{RB}) + (5.753 \times {}^1\kappa) + (13.885 \times {}^3\chi) + 51.105$	0.991	0.983	0.981	0.978	767	2.39
		(0.976)	(0.953)	(0.952)			(2.04)
C1	$\Delta C_p^{298.15} = (1.008 \times {}^0\chi) + (0.951 \times \text{MF}) + (0.198 \times \text{MA}) - 5.529$	0.997	0.994	0.993	0.993	2095	1.08
		(0.982)	(0.964)	(0.963)			(1.46)
C2	$\Delta C_p^{298.15} = (1.179 \times {}^0\chi) + (0.471 \times \text{S\_sCH}_3) + (0.191 \times \text{MA})$	0.997	0.993	0.993	0.992	2005	1.09
		(0.982)	(0.964)	(0.963)			(1.45)
C3	$\Delta C_p^{298.15} = (2.039 \times {}^3\chi) + (0.243 \times \text{MA}) - 6.787$	0.995	0.991	0.990	0.989	2239	1.29
		(0.984)	(0.967)	(0.966)			(1.39)
C4	$\Delta C_p^{298.15} = (0.732 \times {}^1\kappa) + (0.921 \times \text{MF}) + (0.204 \times \text{MA}) - 4.842$	0.997	0.993	0.993	0.992	1944	1.12
		(0.981)	(0.962)	(0.961)			(1.51)
C5	$\Delta C_p^{298.15} = (4.087 \times \text{SC\_p}) - (7.661 \times {}^1\chi) + (0.226 \times \text{MA}) - 6.229$	0.996	0.993	0.992	0.992	1925	1.12
		(0.984)	(0.969)	(0.968)			(1.35)

R^2 = coefficient of determination; R^2_{adjusted} = adjusted coefficient of determination, R^2_{cv} = cross-validated coefficient of determination; RB = number of rotatable bonds; HA = number of hydrogen bond acceptors; S_ddsN = E state keys (sums) = S_ddsN; MD = molecular density; ${}^3\kappa$ = 3-Kappa (Kier and Hall); MR = molecular refractivity; SC_c = Subgraph counts (3): chain; AC = atomic composition; S_dssC = E-state keys (sums): S_dssC; ${}^3\chi$ = 3-Chi (chain) (Kier and Hall); ${}^1\kappa$ = 1-Kappa (atom modified) (Kier and Hall); VDM = Vertex distance/magnitude; S_ssO = E-state keys (sums): S_ssO; A log P = A log P; S_sCH$_3$ = E-state keys (sums): S_sCH$_3$; MA = molecular area; MF = number of methyl groups; SC_p = subgraph counts (0): path.

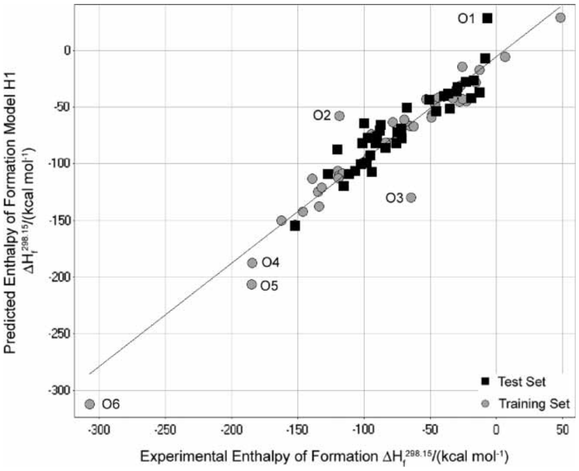


Figure 6. Predicted vs experimental standard enthalpies of formation for both training (●) and test sets (■), using model H1.

Molecular refractivity is defined as

$$\mathrm{MR} = \frac{n^2 - 1}{n^2 + 2} \times \frac{M_\mathrm{w}}{d} \tag{1}$$

where n is the refractive index, M_w the molecular weight and d the density. As n usually does not change significantly, the molecular refractivity is effectively a measure of volume, and therefore, the size of the molecule, albeit coupled to polarizability information [52]. The atomic composition index, finally, is an information content descriptor and the name is programmatic in this context-the descriptor encodes the elemental composition of a molecule. It is intuitively comprehensible, why such a descriptor should encode information about enthalpies of formation.

3.2.2 Standard entropy of formation. Little experimental entropy data is available for the compounds contained in the dataset (8 out of 84 compounds). Furthermore, the available data is tightly clustered and not suitable for the development of high quality models. QSPRs for the standard entropy of formation of small oxygen-containing heterocycles were, therefore, developed using calculated data. As discussed above, density functional methods perform marginally better than PM3. Therefore, data from the PBE/DNP calculations was used to develop the equations. The models appear to be very stable and predictive with $R^2 = 0.988$, $R^2_\mathrm{adjusted} = 0.987$ and $R^2_\mathrm{cross-validated} = 0.984$ in the training and $R^2 = 0.958$ and $R^2_\mathrm{adjusted} = 0.957$ in the test set, using model S1. The standard deviation is 1.99 cal mol^{-1} K^{-1} in the training set (table 3, figure 7). The descriptors which appear most frequently in the top five models (table 3) are the $^1\kappa$ and $^3\chi$ descriptors [53] as well as the rotatable bond count.

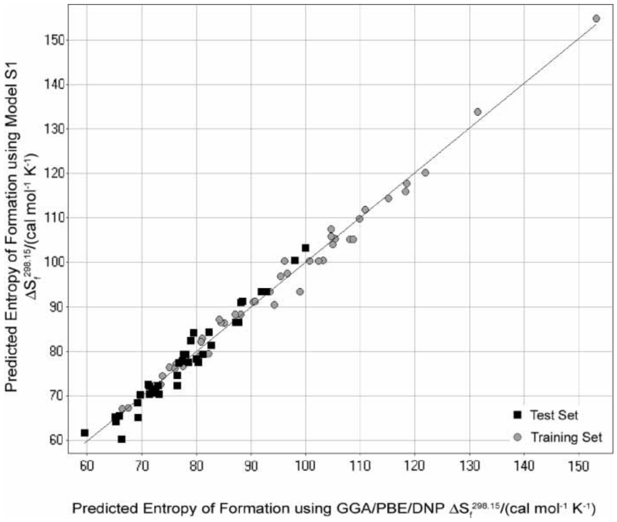


Figure 7. Predicted vs computational standard entropies of formation for both training (●) and test sets (■), using model S1.

3.2.3 Standard heat capacity. Again, very little experimental heat capacity data is available (19 out of 84 compounds) for model development and the data is tightly clustered. As there is no real difference between the heat capacity data computed using DFT and semi-empirical methods, data generated using density functional theory was used to derive the QSPR equations. Again, the models show extremely good performance, both in terms of training and validation sets (table 3, figure 8). The highest performing model C1 gave $R^2 = 0.994$, $R^2_\mathrm{adjusted} = 0.993$ and $R^2_\mathrm{cross-validated} = 0.993$ in the training set and $R^2 = 0.964$ and $R^2_\mathrm{adjusted} = 0.963$ for the validation set. The most frequently observed descriptor here is the molecular area (vdW area) descriptor. Its presence is probably not surprising as it describes the van der Waals area of the molecule and therefore also its size. As the heat capacity is defined as the amount of heat required to change the temperature of a substance by one degree, larger molecules will need more heat than smaller ones, which, in turn explains the correlation with the molecular area descriptor. The only other descriptors appearing multiple times are $^0\chi$ and the methyl group count, both at two counts each. $^0\chi$ describes the immediate bonding environment of atoms in a molecule, while containing

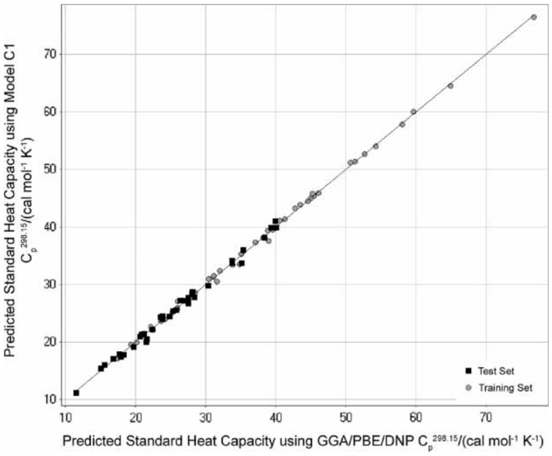


Figure 8. Predicted vs computational standard heat capacities for both training (●) and test sets (■), using model C1.

relative information about the connectivity of the molecular skeleton.

While the QSPRs for both the entropy of formation as well as the heat capacities were developed using computed data, one would have to expect that similar robust models could be developed for experimental data on the basis of the fact that all three different computational methods (see above) are in close agreement with each other and with the available experimental data; i.e. the computed results must be close to the experimental values, were these available.

4. Summary and conclusions

Several computational ways of obtaining thermochemical parameters for small oxygen-containing heterocycles were investigated and compared and QSPR models for the prediction of standard enthalpies and entropies of formation as well as standard heat capacities were developed. Robust and predictive QSPRs were developed for all three thermodynamic parameters on the basis of experimental or validated computed data. It could be shown that QSPR models can be a fast and powerful tool for the prediction of thermodynamic parameters of small oxygen-containing heterocycles.

Acknowledgements

The authors wish to acknowledge the Dutch Polymer Institute (DPI), Project #500, for financial support of this work. Furthermore, Brian Pauw is gratefully acknowledged for help with MatLab programming.

References

[1] S. Penczek. Cationic ring-opening polymerization (CROP) major mechanistic phenomena. *J. Polym. Sci. A*, **38**, 1919 (2000).

[2] A. Baccarelli, P. Mocarelli, D.G. Patterson Jr, M. Bonzini, A.C. Pesatori, N. Caporaso, M.T. Landi. Immunologic effects of dioxin: new results from seveso and comparison with other studies. *Environ. Health Perspect*, **110**, 1169 (2002).

[3] H. Cramail, A. Deffieux. Cationic polymerization. In *Synthesis of Polymers*, A.D. Schlueter (Ed.), p. 231, Wiley-VCH, Weinheim (1999).

[4] E. Brock. Asia—future market for engineering plastics, http://www.ticona.com/ticona/ed_brock_16_6_uk_pdf.pdf (accessed November 2005).

[5] BASF, Press Release. Available online at: http://media.basf.com/en/presse/mitteilungen/pm.htm?pmid=1761&id=c0P1t7gw7bcp0oS (accessed November 2005).

[6] S.W. Benson. *Thermochemical Kinetics*, John Wiley & Sons, New York (1976).

[7] S.W. Benson, J.H. Buss. Additivity rules for the estimation of molecular properties. Thermodynamic properties. *J. Chem. Phys.*, **29**, 546 (1958).

[8] S.W. Benson, F.R. Cruickshank, D.M. Golden, G.R. Haugen, H.E. O'Neal, A.S. Rodgers, R. Shaw, R. Walsh. Additivity rules for the estimation of thermochemical parameters. *Chem. Rev.*, **69**, 279 (1969).

[9] N. Saito, A. Fuwa. Prediction for thermodynamic function of dioxins for gas phase using semiempirical molecular orbital method with PM3 Hamiltonian. *Chemosphere*, **40**, 131 (2000).

[10] C.F. Wilcox, Y.-X. Zhang, S.H. Bauer. The thermochemistry of TNAZ (1,3,3-trinitroazetidine) and related species: G3(MP2)//B3LYP heats of formation. *J. Mol. Struct.: Theochem.*, **538**, 67 (2001).

[11] R. Notario, M.V. Roux, O. Castano. The enthalpy of formation of dibenzofuran and some related oxygen containing heterocycles in the gas phase. *Phys. Chem. Chem. Phys.*, **3**, 3717 (2001).

[12] M.G. Ahunbay, S. Kranias, V. Lachet, P. Ungerer. Prediction of thermodynamic properties of heavy hydrocarbons by Monte Carlo simulations. *Fluid Phase Equilib.*, **224**, 73 (2004).

[13] J.J.P. Stewart. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.*, **10**, 209 (1989).

[14] J.J.P. Stewart. Optimization of parameters for semiempirical methods II. Applications. *J. Comput. Chem.*, **10**, 221 (1989).

[15] T.S. Lay, T. Yamada, P.-L. Tsai, J.W. Bozelli. Thermodynamic parameters and group additivity ring corrections for three-to-six-membered oxygen heterocyclic hydrocarbons. *J. Phys. Chem. B*, **101**, 2471 (1997).

[16] T.S. Lay, J.W. Bozzelli. Enthalpies of formation and group additivity of alkyl peroxides and trioxides. *J. Phys. Chem. A*, **101**, 9505 (1997).

[17] T.S. Lay, J.W. Bozzelli. Enthalpies of formation of cyclic alkyl peroxides: dioxirane, 1,2-dioxetane, 1,2-dioxolane and 1,2-dioxane. *Chem. Phys. Lett.*, **268**, 175 (1997).

[18] M.L. Shirel, P. Pulay. Stability of novel oxo- and chloro-substituted trioxanes. *J. Am. Chem. Soc.*, 121 (1999).

[19] X.-W. Li, E. Shibata, T. Nakamura. Theoretical calculation of thermodynamic properties of polybrominated dibenzo-*p*-dioxins. *J. Chem. Eng. Data*, **48**, 727 (2003).

[20] J. Delley. An all-electron numerical method for solving the local density functional for polyatomic molecules. *J. Chem. Phys.*, **92**, 508 (1990).

[21] J. Delley. From molecules to solids with the DMol3 approach. *J. Chem. Phys.*, **113**, 7756 (2000).

[22] B. Hammer, L.B. Hansen, J.K. Norskov. Improved adsorption energetics within density-functional theory using revised Perdew–Burke Ernzerhof functionals. *Phys. Rev. B*, **59**, 7413 (1999).

[23] A.D.J. Becke. A multicenter numerical integration scheme for polyatomic molecules. *J. Chem. Phys.*, **88**, 2547 (1988).

[24] C. Lee, W. Yang, R.G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, **37**, 785 (1988).

[25] A.D. Boese, N.C. Handy. A new parametrization of exchange-correlation generalized gradient approximation functionals. *J. Chem. Phys.*, **114**, 5497 (2001).

[26] G. Rauhut, T. Clark. Multicenter point charge model for high-quality molecular electrostatic potentials from AM1 calculations. *J. Comput. Chem.*, **15**, 503 (1993).

[27] B. Beck, G. Rauhut, T. Clark. The natural atomic orbital point charge model for PM3: multipole moments and molecular electrostatic potentials. *J. Comput. Chem.*, **15**, 1064 (1994).

[28] Accelrys. homepage, Available online at: http://www.accelrys.com (accessed November 2005).

[29] National Institute of Standards, Chemistry Webbook, Available online at: http://webbook.nist.gov/chemistry (accessed November 2005).

[30] R.C. Reid, J.M. Prausnitz, B.E. Poling. *The Properties of Gases and Liquids*, MacGraw-Hill Book Company, London (1987).

[31] W.J. Lyman, W.F. Reehl, D.H. Rosenblatt. Environmental behaviour of environmental compounds. *Handbook of Chemical Property Estimation Methods*, McGraw-Hill Book Company, London (1982).

[32] R.D. Johnson III. NIST computational chemistry comparison and benchmark database, NIST standard reference database number 101, Available online at: http://srdata.nist.gov/ccbdb (accessed November 2005).

[33] D. Rogers, A.J. Hopfinger. Application of genetic function approximation in quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.*, **34**, 854 (1994).

[34] C.L. Senese, A.J. Hopfinger. A simple clustering technique to improve QSAR model selection and predictivity: application to a receptor independent 4D-QSAR analysis of cyclic urea derived inhibitors of HIV-1 protease. *J. Chem. Inf. Comput. Sci.*, **43**, 2180 (2003).

[35] A. Yasri, D. Hartsough. Toward an optimal procedure for variable selection and QSAR model building. *J. Chem. Inf. Comput. Sci.*, **41**, 1218 (2001).

[36] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, K.A. Prebble. Artificial neural networks in classification of NIR spectral data: design of the training set. *Chemom. Intell. Lab. Syst.*, **33**, 35 (1996).

[37] T. Poetter, H. Matter. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med Chem.*, **41**, 478 (1990).

[38] A. Golbraikh, A. Tropsha. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput. Aided Mol. Des.*, **16**, 357 (2002).

[39] G.V. Kauffman, P.C. Jurs. QSAR and k-nearest neighbour classification analysis of selective cyclooxygenase-2 inhibitors using topologically based numerical descriptors. *J. Chem. Inf. Comput. Sci.*, **41**, 1553 (2001).

[40] B.E. Mattioni, P.C. Jurs. Development of quantitative structure-activity relationship and classification models for a set of carbonic anhydrase inhibitors. *J. Chem. Inf. Comput. Sci.*, **42**, 94 (2002).

[41] R. Guha, J.R. Serra, P.C. Jurs. Generation of QSAR sets with a self-organizing map. *J. Mol. Graph. Model*, **23**, 1 (2004).

[42] E. Bayram, P. Santago, R. Harris, Y.-D. Xiao, A.J. Clauset, J.D. Schmitt. Genetic algorithms and self-organizing maps: a powerful combination for modeling complex QSAR and QSPR problems. *J. Comput. Aided Mol. Des.*, **18**, 483 (2004).

[43] M. Sjostrom, L. Eriksson. Applications of statistical experimental design. In *Chemometrics Methods in Molecular Design*, H. van de Waterbeed (Ed.), p. 63, VCH, Weinheim (1995).

[44] L. Eriksson, E. Johansson. Multivariate design and modeling in QSAR. *Chemom. Intell. Lab. Syst.*, **34**, 1 (1996).

[45] P. Gramatica, E. Papa. QSAR Modeling of bioconcentration factor by theoretical molecular descriptors. *QSAR Comb. Sci.*, **22**, 374 (2003).

[46] P. Gramatica, P. Pilutti, E. Papa. QSAR prediction of ozone tropospheric degradation. *QSAR Comb. Sci.*, **22**, 364 (2003).

[47] E. Marengo, R. Todeschini. A new algorithm for optimal, distance-based experimental design. *Chemom. Intell. Lab. Sys.*, **16**, 37 (1992).

[48] D.C. Montgomery. *Design and Analysis of Experiments*, John Wiley, New York (2001).

[49] A. Tropsha, P. Grammatica, V.K. Gombar. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.*, **22**, 69 (2003).

[50] Accelrys, Accelrys MS modeling 3.2 documentation. (2005).

[51] L.H. Hall, B. Mohney, L.B. Kier. The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.*, **31**, 76 (1991).

[52] C. Hansch, R. Garg, A. Kurup. Searching for Allosteric Effects via QSARs. *Bioorg. Med. Chem.*, **9**, 283 (2001).

[53] L.H. Hall, L.B. Kier. The molecular connectivity chi indexes and kappa shape indexes in structure property modeling. In *Reviews in Computational Chemistry*, K.B. Lipkovitz, D.B. Boyd (Eds.), (1992).

[54] J.-M. Colmont. Assignment method of the rotational spectrum of a slightly asymmetric molecule. *J. Molec. Spec.*, **80**, 166 (1980).