QSPR Using MOLGEN-QSPR: The Challenge of Fluoroalkane Boiling Points

Christoph Rücker,* Markus Meringer, and Adalbert Kerber

Department of Mathematics, Universität Bayreuth, D-95440 Bayreuth, Germany

Received September 2, 2004

By means of the new software MOLGEN-QSPR, a multilinear regression model for the boiling points of lower fluoroalkanes is established. The model is based exclusively on simple descriptors derived directly from molecular structure and nevertheless describes a broader set of data more precisely than previous attempts that used either more demanding (quantum chemical) descriptors or more demanding (nonlinear) statistical methods such as neural networks. The model's internal consistency was confirmed by leave-one-out cross-validation. The model was used to predict all unknown boiling points of fluorobutanes, and the quality of predictions was estimated by means of comparison with boiling point predictions for fluoropentanes.

INTRODUCTION

Recent innovative use of poly- or perfluorinated alkanes has fostered interest in the physical properties of fluoroalkanes.[1] A compound's normal boiling point (bp) is a fundamental thermodynamic property, interesting both in its own right and as a basis for the calculation of other properties.[2] Studies of the boiling points of haloalkanes (including fluoroalkanes and mixed fluorohaloalkanes) were published by several authors[3−5] and recently by ourselves.[6] Fluoroalkanes (containing C, H, and F only) are a subset of haloalkanes, and a description of a subset can be reasonable if higher precision is desired at the expense of some loss of generality. Often a subset is more easily described than a broader compound set. In sharp contrast, we noticed that the largest residuals in our modeling of haloalkane bps were due to fluoroalkanes.[6]

The difficulties with fluoroalkane bps should be caused by the particular properties of the fluorine atom compared to the other halogens. Fluorine is the most electronegative and has the lowest mass of all halogens, and an F atom can be considered a smaller and "harder" sphere than other halogen atoms. A C−Hal bond is a strong dipole, and these local dipoles interact intra- and intermolecularly even if the resulting dipole moment of an isolated molecule is zero. While this is true for all C−Hal bonds, attractive interactions between F atoms (negative end of a strong dipole) and H atoms (positive end of a weak dipole) seem to be stronger than corresponding interactions involving atoms of higher halogens, due to the high partial charge and the small radius of a bonded F atom. A consequence is the decidedly nonlinear dependence of fluoroalkane boiling points on the number of fluorine atoms, resulting in a bp maximum about halfway between an alkane and the corresponding perfluoroalkane (CH_4 −161.5 °C, CH_3F −78.5, CH_2F_2 −51.6, CHF_3 −82, CF_4 −128).[3a]

For this reason previous researchers used inherently nonlinear methods such as neural networks for the description of fluoroalkane boiling points.[3b,7] A manual graphical fitting method was used to describe the bps of fluoroethanes or, separately, of fluoropropanes and allowed some predictions.[8] Portability, parsimony, and predictive power of these methods, however, are less than desirable. Alternatively, multilinear regression (MLR) models required six mostly electrostatic and quantum chemical descriptors for fitting the bps of 42 fluoroalkanes to a standard error of $s = 6.3$ °C.[5] Accepting fluoroalkane boiling points as a challenge we were curious on the viability of a model of the simplest kind for the boiling points of this class of compounds, i.e., a multilinear regression model based on simple descriptors that are directly derived from molecular structure. This is by no means an easy task, as illustrated by the fact that for fluorobutanes alone boiling points between −2 and +110 °C have been reported, and even among the restricted set of isomeric tetrafluorobutanes bps vary between 17 and 110 °C.

METHODS

Data. Boiling points at normal pressure of C_1−C_4 fluoroalkanes (herein given in °C) were taken from the literature[3,8,9−11] and checked against the Beilstein database in order to avoid fitting wrong data. A few compounds/boiling points found in Beilstein only were also included. In cases of marginal divergence between boiling points reported in the sources the average was taken. In cases of major divergence boiling points were excluded. This resulted in a data set of 83 C_1−C_4 fluoroalkanes with boiling points, i.e., all 42 possible fluoromethanes through fluoropropanes plus 41 (out of 116 possible, ignoring stereoisomerism) fluorobutanes. We did not normally examine the primary sources given in Beilstein and thus cannot exclude the remote possibility that a few bp values accepted here may be calculated rather than experimental values.[12] This is the most comprehensive collection of lower fluoroalkane bps of which we are aware.

Boiling points were attributed reliability classes as in our earlier studies:[6,13] Boiling points appearing in the Beilstein database only once are in reliability class 0, those measured at least twice by independent researchers with a difference

* Corresponding author phone: +49 921 553386; fax: +049 921 553385; e-mail: christoph.ruecker@uni-bayreuth.de.

10.1021/ci0497298 CCC: $30.25
Published on Web 12/22/2004

of at most 4 °C are in class 1, and those measured at least four times by independent authors and differing no more than 2 °C are in reliability class 2. According to this measure, the data used in the present study are on average of lower quality than those in our previous studies, with many fluoroalkane bps reported only once.

Information on reliability, though not used in the calculations, proved useful for identification of dubious bp data and in the selection of reference data.

The bp reported for 1,2-difluorobutane, $FH_2C-CHF-CH_2-CH_3$, 39 °C, reliability class 0, is obviously unreasonable and was therefore excluded,[14] whence the number of fluoroalkanes to be treated in this study was reduced to 82. For 1,2-difluoroethane, FH_2C-CH_2F, in our main study we used the bp value 10.5 °C, as was done in most previous studies.[4,7,8,15]

Descriptor Calculation. We used our software MOLGEN-QSPR that combines structure generation with calculation of many molecular descriptors and with data treatment by various statistical methods.[6,16] For a set of compounds structurally rather homogeneous such as the fluorinated alkanes, substructure and fragment counts could be anticipated to be valuable descriptors.[17] In fact, in previous studies such descriptors had been extensively used.[4,8] So, in addition to the descriptors offered routinely by MOLGEN-QSPR,[16,18] we included in this study substructure counts and fragment counts.[19]

Substructure and fragment counts are useful in QSPR studies since they may work even if a descriptor adequately describing a physical phenomenon is not at hand. They do not describe the magnitude of a physical phenomenon in any particular compound, but in a series of compounds they describe the *variation* of the phenomenon's influence due to the variation of the occurrence number of the structural element deemed responsible for it.

A substructure is a part of the hydrogen-suppressed molecular graph, and substructure counts are automatically provided by MOLGEN-QSPR. For example, when requesting counts of all substructures of one to four bonds, the following substructures were found and their occurrences counted: C−C, C−F, C−C−C, F−C−F, C−C−F, C−C−C−C, C−C(C)−C, F_2C-C, F−C−C−F, F_3C, C−C−C−F, C−CF−C, F_3C-C, F_2C-C-F, C−CF−C−C, C−C−C−C−F, C−CF(C)−C, C−C(C)−C−F, F−C−C−C−F, F−C−CF−C, F_2C-C-C, $C-CF_2-C$. MOLGEN rule for substructure counts: Two embeddings of a substructure in a structure are counted separately if they differ in at least one bond, i.e., if they differ in at least one non-H atom.

In contrast, a fragment is a part of the full molecular graph, it may therefore include hydrogen atoms. A fragment is defined by the user and searched for by the system on request. For example, counts of the following fragments were used in this study: H−C−F, H−C−C−F (according to Woolf[8b]); CHF_2, CH_2F (according to Carlton[4]); and CH_3, $C-CH_2-C$, C−CH(C)−C, C−CHF−C, FH_2C-CHF, FHC−CHF. MOLGEN rule for fragment counts: Two embeddings of a fragment in a structure are counted separately if they differ in at least one non-H atom. For example, by this rule 1,1,1,2,3-pentafluoropropane, $F_3C-CHF-CH_2F$, contains 2 H−C−F and 5 H−C−C−F fragments.

In Woolf's study[8b] embeddings of a fragment are counted separately if they differ in at least one atom (which may be an H atom). By this rule $F_3C-CHF-CH_2F$ contains 3 H−C−F fragments and 6 H−C−C−F fragments. Counts of H−C−F and H−C−C−F fragments obtained by this rule (HCF_{man} and $HCCF_{man}$) were manually added to our data.

Further descriptors based on physical concepts were included in the descriptor pool: bip = $n(CHF_2) + n(CH_2F)$ (bipolar groups, see ref 4); tbip = bip + n(C−CHF−C) (total bipolar groups); xsF = $N_F - n(CH_2F) - N_H + n(CHF_2)$, unless this number is negative, in which case xsF is set to 0 (excess exterior fluorine atoms[4]); and $n(CF_3)^2$, $n(CH_3)^2$, $n(CF_3)\cdot n(CH_3)$ (see ref 8b).

Finally, descriptors $(relN_F)^2 = (N_F/\text{number of all atoms})^2$ and $(Frate)^2 = (N_F/(2N_C+2))^2$ were included to account for the bp maximum mentioned in the Introduction.

Descriptor Selection for MLR. For finding best or near-best subsets of *k* molecular descriptors out of a large descriptor pool the step-up method was used. In this method to each of the currently best *n* sets of descriptors another descriptor is added, and the best *n* such sets are collected. This procedure is repeated until the best set of *k* descriptors is found. The better of two descriptor sets is the one leading to the higher r^2 (lower *s*) value in MLR. Since it does not exclude, from the beginning, certain combinations of descriptors, this method is more likely to find a very good subset of descriptors than the methods used in CODESSA,[20] but it still does not guarantee to find the very best subset. The models reported in this paper were obtained with parameter *n* set to 1000. Routinely, MOLGEN-QSPR allows to display all *n* models, so that the user may consider the second-best, third-best, etc. model found along with the best one. In the present paper best models only are mentioned. The quality of the final models was assessed via leave-one-out cross-validation, characterized by r^2_{cv} and s_{cv} values.

RESULTS AND DISCUSSION

Preliminary Study. For comparison with the results of Ivanciuc et al.,[5] we initially treated their data using our descriptors and statistical procedures. For this purpose compound structures and boiling points were taken from Balaban et al.,[3a] as described in ref 5. Unfortunately, Ivanciuc et al. reported results for a set of 42 C_1-C_4 fluoroalkanes, whereas Balaban's data[3a] include 43 such fluoroalkanes. So one compound is missing in Ivanciuc's study, but we do not know which one. For the 43 fluoroalkanes MOLGEN-QSPR found the best 6-descriptor MLR model as follows. (In the text we characterize a MLR model by the descriptors involved and by its r^2, *s*, *F*, r^2_{cv}, and s_{cv} values, the latter two refer to leave-one-out cross-validation. For full models see Table 1.)

Xu^m, ${}^0\chi^v$, ${}^4\chi_p$, S(ssssC), n(F−C−F), n(F−C−C−F) (model 0)

$r^2 = 0.9883$, $s = 4.624$, $F = 508$, $r^2_{cv} = 0.9826$, $s_{cv} = 5.652$, $N = 43$.[21,22]

Xu^m is the modified Xu index,[18,23] S(ssssC) is the sum of

Table 1. Full MLR Models for the Boiling Points of C_1-C_4 Fluoroalkanes

$bp = 83.2226 \cdot Xu^m - 25.1841 \cdot {}^0\chi^v - 12.2045 \cdot {}^4\chi_p - 5.21304 \cdot S(ssssC) - 34.247 \cdot n(F-C-F) - 9.2969 \cdot n(F-C-C-F) - 41.3515$ (model 0)

$bp = 67.3463 \cdot Xu^m - 1.13857 \cdot {}^2TC^v + 0.092004 \cdot {}^3TM1 + 2.35049 \cdot HCCF_{man} - 81.9815 \cdot (relN_F)^2 + 8.5278 \cdot n(FH_2C-CHF) - 75.3646$ (model 1)

$bp = 64.6756 \cdot Xu^m - 5.51401 \cdot {}^2P - 2.05454 \cdot S(ssssC) - 18.0512 \cdot n(F-C-F) - 3.88042 \cdot n(F-C-C-F) - 53.7732 \cdot (relN_F)^2 + 7.49879 \cdot n(FH_2C-CHF) - 72.4514$ (model 2)

electrotopological state indices of carbon atoms bearing no hydrogen,[24] and n(F−C−F) and n(F−C−C−F) are counts of the respective substructures.

This result is to be compared to the previously best 6-descriptor MLR model, which comprises one topological, two electrostatic, and three quantum-chemical descriptors:

$r = 0.989$ ($r^2 = 0.9781$), $s = 6.3$, $F = 267$, $N = 42$.[5]

So even without any quantum-chemical descriptor results better than those obtained in ref 5 are achievable.

Main Study. For the larger set of 82 C_1-C_4 fluoroalkane bps the best multilinear 6-descriptor model found by the step-up procedure is

Xu^m, ${}^2TC^v$, ${}^3TM1$, $HCCF_{man}$, $(relN_F)^2$, $n(FH_2C-CHF)$ (model 1)

$r^2 = 0.9845$, $s = 5.140$, $F = 793$, $r^2_{cv} = 0.9812$, $s_{cv} = 5.679$, $N = 82$

${}^2TC^v$ and ${}^3TM1$ are Bonchev overall indices,[25] $HCCF_{man}$ is the number of HCCF fragments as counted by Woolf,[8b] and $n(FH_2C-CHF)$ is the count of fragment FH_2C-CHF.

The best 7-descriptor model found is

Xu^m, ${}^2P$, S(ssssC), n(F−C−F), n(F−C−C−F), $(relN_F)^2$, $n(FH_2C-CHF)$ (model 2)

$r^2 = 0.9872$, $s = 4.701$, $F = 815$, $r^2_{cv} = 0.9840$, $s_{cv} = 5.252$, $N = 82$

${}^2P$ is the number of paths of length 2 in the H-suppressed molecular graph.

Figure 1 is a plot of experimental vs calculated bps (by model 2) and of the corresponding bps obtained by leave-one-out cross-validation.

Table 2 lists experimental and calculated bps (by model 2) and residuals.

Table 2. Experimental and Calculated Boiling Points (by Model 2), Residuals and Structures of 82 C_1-C_4 Fluoroalkanes Included in This Study

	bp	calc	residual	structure		bp	calc	residual	structure
1	−15	−18.461	3.4608	$F_3C-CF_2-CH_3$	42	12	14.324	−2.324	$F_3C-CH(CF_3)_2$
2	12	8.8309	3.1691	$F_3C-CH(CH_3)_2$	43	−0.3	−2.8219	2.5219	$F_3C-CF(CF_3)_2$
3	24.5	34.612	−10.112	$F_3C-CH_2-CH_2-CF_3$	44	77	72.755	4.2453	$FH_2C-CH_2-CH_2-CH_2F$
4	32.5	33.345	−0.84489	$F_2HC-CHF-CF_2-CF_3$	45	31	24.239	6.7612	$H_3C-CF_2-CH_2-CH_3$
5	35	36.325	-1.3245	$F_2HC-CF_2-CHF-CF_3$	46	10.45	13.729	−3.2793	FH_2C-CH_2F
6	26.5	28.468	−1.9683	$FH_2C-CF_2-CF_2-CF_3$	47	25	24.475	0.52489	$F_2HC-CF_2-CH_2F$
7	−51.6	−57.615	6.015	CH_2F_2	48	40	39.381	0.61881	$F_3C-CH_2-CF_2-CH_3$
8	41	38.948	2.0525	$FH_2C-CH_2-CH_2F$	49	17	27.505	−10.505	$H_3C-CF_2-CF_2-CH_3$
9	21	24.266	−3.2656	$F_3C-CHF-CH_2F$	50	34.3	28.713	5.587	$F_3C-CHF-CHF-CF_3$
10	−78.5	−74.602	−3.8977	CH_3F	51	32	28.758	3.2418	$FH_2C-CHF-CH_3$
11	11	12.337	−1.3371	$F_2HC-CF_2-CHF_2$	52	66	61.716	4.2836	$FH_2C-CHF-CH_2F$
12	7.5	6.8761	0.62385	$F_2HC-CH_2-CH_3$	53	18.7	15.386	3.3139	$FH_2C-CF_2-CH_3$
13	6	8.5232	−2.5232	$F_3C-CHF-CHF_2$	54	27	35.177	−8.1773	$FH_2C-CF_2-CH_2F$
14	−9.7	−7.9039	−1.7961	$H_3C-CHF-CH_3$	55	39	37.214	1.7859	$F_2HC-CH_2-CHF_2$
15	−13	−11.543	−1.4566	$F_3C-CH_2-CH_3$	56	22	16.857	5.1425	$F_2HC-CHF-CH_3$
16	−19	−15.058	−3.9422	$F_3C-CHF-CF_3$	57	55	47.842	7.1577	$F_2HC-CHF-CH_2F$
17	−78	−80.728	2.7278	F_3C-CF_3	58	40.5	32.565	7.9347	$F_2HC-CHF-CHF_2$
18	−82	−80.509	−1.4914	CHF_3	59	−0.8	6.7237	−7.5237	$F_2HC-CF_2-CH_3$
19	−128	−124.53	−3.4678	CF_4	60	15	17.468	−2.4684	$F_2HC-CH_2-CF_3$
20	−2.5	5.2042	−7.7042	$FH_2C-CH_2-CH_3$	61	−0.5	−5.8336	5.3336	$F_3C-CHF-CH_3$
21	−38	−40.317	2.3172	$F_3C-CF_2-CF_3$	62	−48.5	−47.81	−0.69021	F_3C-CHF_2
22	−0.5	−8.2093	7.7093	$H_3C-CF_2-CH_3$	63	17	22.904	−5.904	$F_3C-CH_2-CH_2-CH_3$
23	29.4	19.518	9.8823	$FH_2C-CH_2-CF_3$	64	0	−1.6153	1.6153	$F_3C-CF_2-CH_2F$
24	−0.8	−1.8548	1.0548	$F_3C-CH_2-CF_3$	65	45	38.784	6.216	$F_2HC-CH_2-CH_2F$
25	−17	−14.227	−2.7734	$F_3C-CF_2-CHF_2$	66	14.5	13.168	1.3321	$F_3C-CF_2-CF_2-CH_3$
26	4	8.7202	−4.7202	FH_2C-CHF_2	67	110	104.63	5.3736	$FH_2C-CHF-CHF-CH_2F$
27	32.5	40.107	−7.6066	$FH_2C-CH_2-CH_2-CH_3$	68	78	77.922	0.078154	$F_2HC-CH(CHF_2)-CH_2F$
28	25	26.239	−1.2392	$H_3C-CH_2-CHF-CH_3$	69	56.5	59.7	−3.2	$F_2HC-CH_2-CHF-CH_3$
29	18	20.385	−2.3849	$F_3C-CH_2-CF_2-CF_3$	70	57	57.088	−0.087709	$F_2HC-CH_2-CF_2-CH_3$
30	44	40.696	3.3041	$FH_2C-CF_2-CF_2-CHF_2$	71	46.5	48.489	−1.9886	$F_2HC-CHF-CH_2-CH_3$
31	15	16.08	−1.0797	$F_2HC-CF_2-CF_2-CF_3$	72	90	90.153	−0.15337	$F_2HC-CHF-CHF-CH_2F$
32	−1.7	−8.222	6.522	$F_3C-CF_2-CF_2-CF_3$	73	64	74.296	−10.296	$F_2HC-CF_2-CHF-CH_2F$
33	−21	−18.188	−2.8124	$F_2HC-CHF_2$	74	57.5	58.762	−1.2616	$F_2HC-CF_2-CHF-CHF_2$
34	−25	−28.634	3.6339	F_2HC-CH_3	75	53.5	52.786	0.71424	$F_2HC-CF_2-CF_2-CH_2F$
35	−26.2	−30.827	4.6274	F_3C-CH_2F	76	59	58.031	0.96883	$F_2HC-CH(CF_3)-CH_2F$
36	−37.5	−32.05	−5.4501	H_3C-CH_2F	77	53.5	54.186	−0.68606	$F_3C-CH_2-CH_2-CH_2F$
37	−47	−46.807	−0.19339	H_3C-CF_3	78	45.5	41.217	4.2827	$F_3C-CH_2-CHF-CH_3$
38	22	26.457	−4.4568	$FH_2C-CH(CH_3)_2$	79	23	27.411	−4.4108	$F_3C-CF(CH_2F)-CF_3$
39	12	8.0471	3.9529	$FC(CH_3)_3$	80	42	48.571	−6.5708	$F_3C-CHF-CF_2-CH_2F$
40	21.5	17.311	4.1892	$F_3C-CH(CH_3)-CF_3$	81	43.5	41.032	2.4682	$F_3C-CF_2-CH_2-CH_2F$
41	40	38.896	1.1044	$F_3C-CH(CF_3)-CH_2F$	82	41	41.473	−0.47348	$H_3C-CHF-CHF-CH_3$

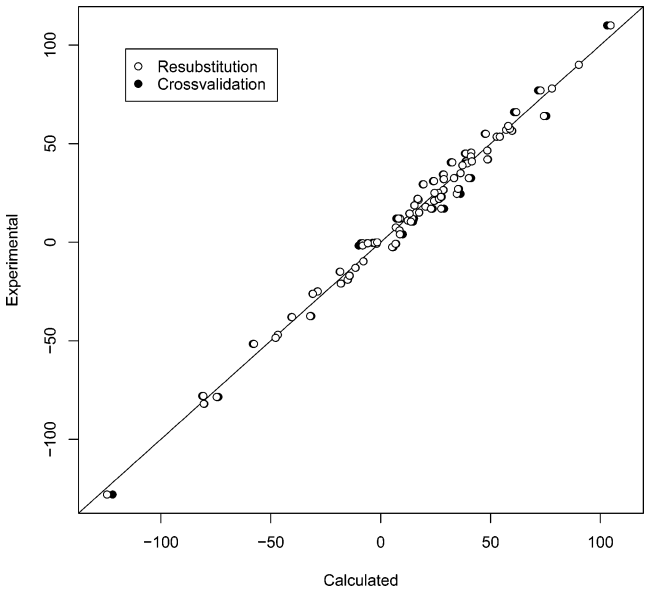

Figure 1. Plot of calculated (by model 2, white disks, and by leave-one-out cross-validation, black disks) vs experimental boiling points of 82 fluoromethanes through fluorobutanes. Note that most black disks are eclipsed by the corresponding white disks.

Figure 2 visualizes the results separately for fluorinated methanes, ethanes, propanes, and butanes. By inspection, the quality of data description is similar in all four subpopulations. In fact, the average absolute error for the 4 fluoromethanes, 9 fluoroethanes, 29 fluoropropanes, and 40 fluorobutanes is 3.72, 3.13, 4.02, and 3.40 °C, respectively.

Note that models 1 and 2 use both (electro)topological indices and substructure and fragment counts. In fact, (electro)topological indices alone or substructure and fragment counts alone led to best 6-descriptor models of $s = 5.83$ or 6.09, respectively. As in our earlier study,[6] geometrical indices did not qualify to appear in the best models. In contrast to expectation, the descriptors based on physical concepts[4,8b] also did not qualify for the best models.

Prediction. By means of model 2 we predicted the normal boiling points of those 76 fluorobutanes whose bps are either not at all known or were excluded from our data set e.g. for contradicting reports in Beilstein. Table 3 shows the predictions.

Of course, the quality of these predictions cannot be assessed at present. We expect the average absolute error to be a bit larger than, but similar to that of the fit for fluorobutanes above, 3.40 °C. An independent and rather conservative estimation was obtained as follows: The data for model 2 lead one to suspect that bp predictions for fluoropentanes by model 2 might not be senseless. Since these are extrapolations, their average error should be larger than that of the predictions for fluorobutanes, which are interpolations. So we predicted by model 2 the bps of all those fluoropentanes that have reliably known experimental bps (reliability classes 1 or 2, eighteen compounds). Results are shown in Table 4 and Figure 3. The average absolute error of these predictions is 4.91 °C. The predictions for the fluorobutanes therefore should on average be better than that.

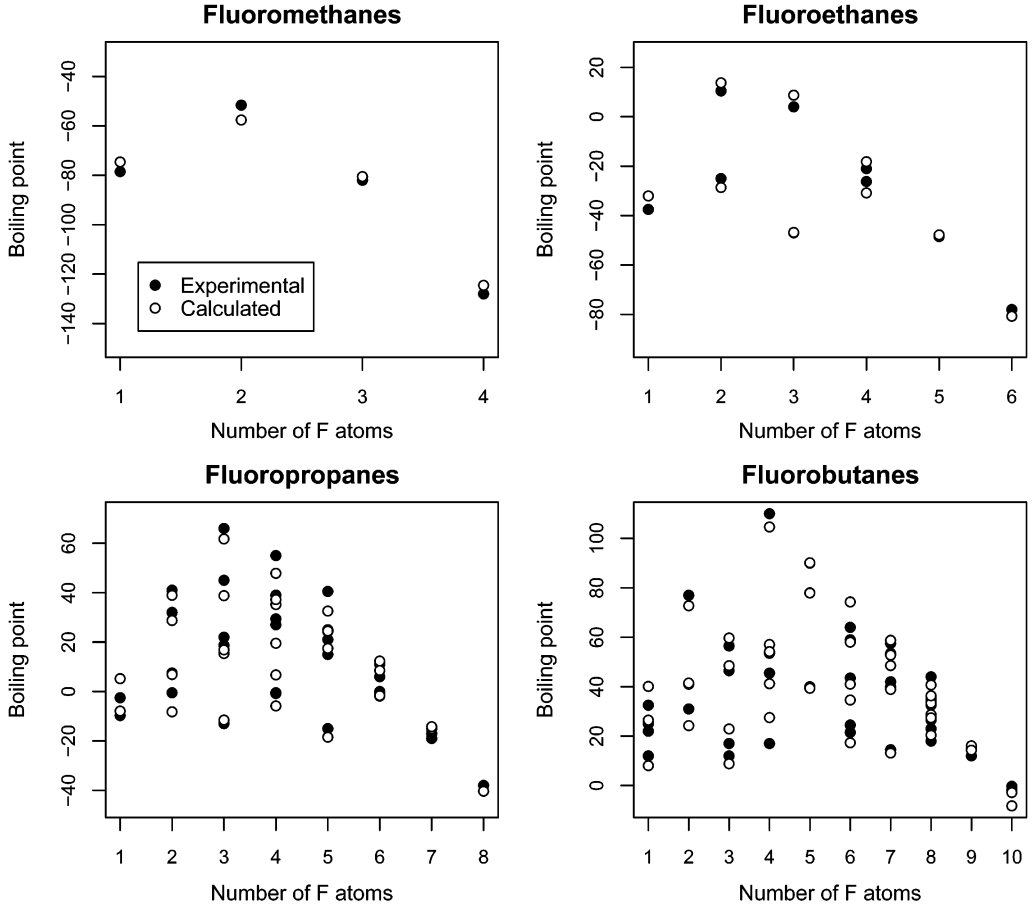

Figure 2. Plot of calculated (by model 2, white disks) and experimental (black disks) boiling points of 82 C_1-C_4 fluoroalkanes vs number of F atoms in the molecule.

Table 3. Predicted Boiling Points (by Model 2) for 76 Fluorobutanes that Were Not Included in This Study

	pred	N_F	structure		pred	N_F	structure
1	60.655	2	$FH_2C-CHF-CH_2-CH_3$	39	74.77	5	$F_2HC-CH_2-CF_2-CH_2F$
2	57.084	3	$FH_2C-CH_2-CF_2-CH_3$	40	71.455	5	$F_3C-CH_2-CHF-CH_2F$
3	56.527	3	$FH_2C-CH(CH_3)-CHF_2$	41	53.27	5	$F_3C-CH_2-CH_2-CHF_2$
4	74.234	3	$FH_2C-CHF-CHF-CH_3$	42	76.202	5	$F_2HC-CHF-CH_2-CHF_2$
5	84.396	3	$HC(CH_2F)_3$	43	32.82	6	$F_3C-CHF-CF_2-CH_3$
6	91.5	3	$FH_2C-CH_2-CHF-CH_2F$	44	21.114	6	$F_3C-CF_2-CHF-CH_3$
7	72.891	3	$FH_2C-CH_2-CH_2-CHF_2$	45	41.266	2	$F_2HC-CH_2-CH_2-CH_3$
8	13.334	4	$F_3C-CF(CH_3)_2$	46	30.614	6	$F_2HC-CF(CH_3)-CF_3$
9	57.735	4	$FH_2C-CF(CH_3)-CHF_2$	47	36.945	6	$F_2HC-CF_2-CF_2-CH_3$
10	55.436	4	$FH_2C-CF_2-CHF-CH_3$	48	56.465	6	$F_3C-CF(CH_2F)_2$
11	37.226	4	$F_2HC-CF_2-CH_2-CH_3$	49	63.937	6	$FH_2C-CF_2-CF_2-CH_2F$
12	27.303	2	$F_2HC-CH(CH_3)_2$	50	56.525	6	$F_3C-CH_2-CF_2-CH_2F$
13	67.773	4	$FH_2C-CHF-CF_2-CH_3$	51	70.555	6	$FH_2C-CF(CHF_2)_2$
14	37.525	4	$FH_2C-CH(CH_3)-CF_3$	52	63.883	6	$F_2HC-CF_2-CH_2-CHF_2$
15	26.182	4	$F_3C-CHF-CH_2-CH_3$	53	70.716	6	$F_2HC-CHF-CF_2-CH_2F$
16	54.727	4	$F_2HC-CH(CH_3)-CHF_2$	54	56.962	6	$F_3C-CH_2-CHF-CHF_2$
17	61.012	4	$F_2HC-CHF-CHF-CH_3$	55	53.199	6	$F_3C-CHF-CH_2-CHF_2$
18	84.728	4	$FC(CH_2F)_3$	56	33.949	3	$F_2HC-CF(CH_3)_2$
19	76.046	4	$FH_2C-CF_2-CH_2-CH_2F$	57	67.226	6	$F_3C-CHF-CHF-CH_2F$
20	81.767	4	$FH_2C-CH(CH_2F)-CHF_2$	58	72.746	6	$HC(CHF_2)_3$
21	78.127	4	$FH_2C-CH_2-CHF-CHF_2$	59	74.796	6	$F_2HC-CHF-CHF-CHF_2$
22	90.422	4	$FH_2C-CHF-CH_2-CHF_2$	60	10.139	7	$H_3C-CF(CF_3)_2$
23	57.286	2	$H_3C-CH(CH_2F)_2$	61	34.269	7	$F_3C-CHF-CH_2-CF_3$
24	72.219	4	$F_2HC-CH_2-CH_2-CHF_2$	62	49.162	7	$F_3C-CF_2-CHF-CH_2F$
25	12.765	5	$H_3C-CH_2-CF_2-CF_3$	63	38.604	7	$F_3C-CF_2-CH_2-CHF_2$
26	36.563	5	$FH_2C-CF(CH_3)-CF_3$	64	45.348	7	$F_3C-CH_2-CF_2-CHF_2$
27	46.804	5	$FH_2C-CF_2-CF_2-CH_3$	65	48.594	7	$F_3C-CF(CH_2F)-CHF_2$
28	51.909	5	$F_2HC-CF(CH_3)-CHF_2$	66	61.602	7	$FC(CHF_2)_3$
29	45.708	5	$F_2HC-CF_2-CHF-CH_3$	67	35.561	3	$H_3C-CF_2-CHF-CH_3$
30	54.435	5	$F_2HC-CHF-CF_2-CH_3$	68	52.567	7	$F_3C-CH(CHF_2)_2$
31	35.614	5	$F_2HC-CH(CH_3)-CF_3$	69	51.596	7	$F_3C-CHF-CHF-CHF_2$
32	38.588	5	$F_3C-CHF-CHF-CH_3$	70	39.393	8	$F_3C-CF(CHF_2)_2$
33	78.235	5	$F_2HC-CF(CH_2F)_2$	71	33.102	8	$F_2HC-CH(CF_3)_2$
34	59.576	2	$FH_2C-CH_2-CHF-CH_3$	72	11.224	9	$F_3C-CF_2-CHF-CF_3$
35	85.333	5	$FH_2C-CF_2-CHF-CH_2F$	73	17.926	9	$F_2HC-CF(CF_3)_2$
36	66.052	5	$F_2HC-CF_2-CH_2-CH_2F$	74	62.477	3	$H_3C-CF(CH_2F)_2$
37	62.093	5	$F_3C-CH(CH_2F)_2$	75	45.854	3	$FH_2C-CF_2-CH_2-CH_3$
38	55.365	5	$F_3C-CHF-CH_2-CH_2F$	76	37.003	2	$FH_2C-CF(CH_3)_2$

Table 4. Experimental (with Reliability) and Predicted Boiling Points (Extrapolations by Model 2), Residuals and Structures of 18 Fluoropentanes

	bp	reliab	pred	residual	structure
1	45	1	39.974	5.0256	$FC(CH_3)_2-CH_2-CH_3$
2	47	1	55.931	−8.9313	$F_3C-CH_2-CH_2-CH_2-CH_3$
3	55	1	60.057	−5.0569	$H_3C-CHF-CH_2-CH_2-CH_3$
4	47	1	55.92	−8.9197	$H_3C-CF_2-CF_2-CH_2-CH_3$
5	75	1	75.356	−0.35596	$H_3C-CF_2-CH_2-CF_2-CH_3$
6	56	1	57.845	−1.8451	$FH_2C-CH(CH_3)-CH_2-CH_3$
7	39.5	1	41.699	−2.1994	$F_3C-CF_2-CF_2-CH_2-CH_3$
8	82	2	79.563	2.437	$F_2HC-CF_2-CF_2-CF_2-CH_2F$
9	63	2	73.321	−10.321	$FH_2C-CH_2-CH_2-CH_2-CH_3$
10	47	1	40.543	6.4565	$F_3C-CH_2-CF(CF_3)_2$
11	68	1	67.426	0.57367	$F_2HC-CF_2-CF_2-CF_2-CHF_2$
12	45	2	44.043	0.95673	$F_3C-CF_2-CF_2-CF_2-CHF_2$
13	54	1	60.32	−6.3204	$FH_2C-CH_2-CH(CH_3)_2$
14	29	2	20.863	8.1366	$F_3C-CF_2-CF_2-CF_2-CF_3$
15	30	2	21.463	8.5366	$F_3C-CF_2-CF(CF_3)_2$
16	41	1	38.573	2.4265	$FH_2C-C(CH_3)_3$
17	61	1	53.602	7.3975	$H_3C-CH_2-CF_2-CH_2-CH_3$
18	60	1	57.546	2.4541	$H_3C-CF_2-CH_2-CH_2-CH_3$

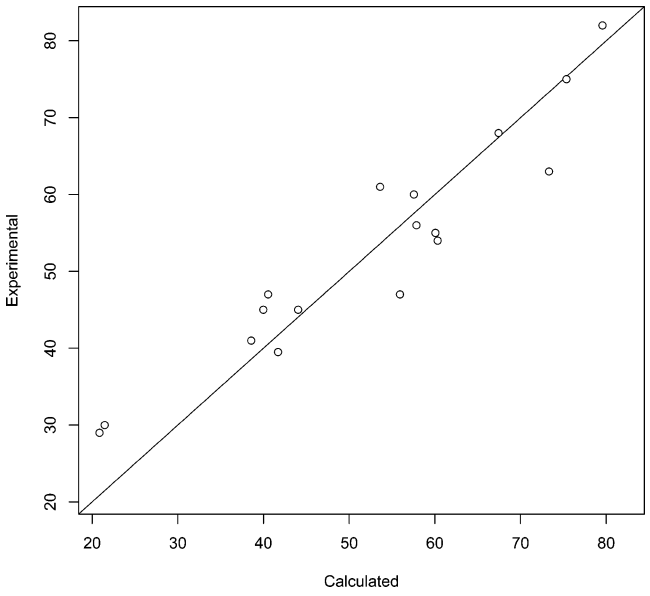

Figure 3. Plot of calculated (by model 2) vs experimental boiling points for 18 fluoropentanes. Note that all these calculations are extrapolations.

CONCLUSION

The combination of global descriptors such as topological and electrotopological indices on one hand and substructure and fragment counts on the other again proved its value for QSPR modeling.[6,12,17] Substructure and fragment counts offer the scientist some flexibility to react whenever in a QSPR model several compounds containing a particular structural element have similar residuals. In fact, in this study fragment count n(FH_2C-CHF) was included in the descriptor pool after preliminary QSPR models had resulted in large positive residuals for compounds $FH_2C-CHF-CH_3$, $FH_2C-CHF-CH_2F$, $F_2HC-CHF-CH_2F$, and $FH_2C-CHF-CHF-CH_2F$.

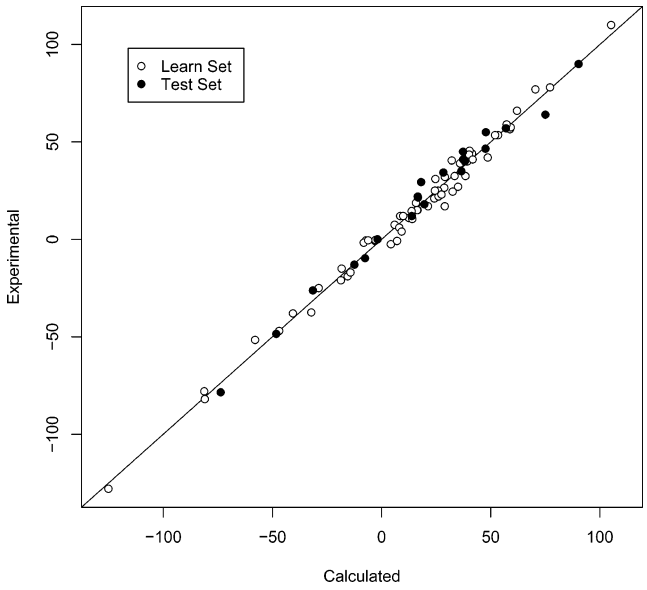

Figure 4. Plot of calculated (by model 2a) vs experimental boiling points for the training set (white disks) and the test set (black disks) as described in the Appendix.

In contrast to the usual observation that property description becomes easier when the compound class is restricted, fluoroalkane bps are more difficult to fit than those of haloalkanes in general. Nevertheless, we found a model that, using descriptors more easily obtained, describes the bps of nearly twice as many fluoroalkanes to a higher precision than did a previous attempt, by the same statistical method and the same number of descriptors.

APPENDIX

In response to a detailed request by a reviewer, we carried out the following additional test of the descriptors appearing in model 2. The 82 C_1-C_4 fluoroalkanes were partitioned into a training set and a test set. A random number between zero and one was attributed to each compound, and the fluoromethane with the lowest such number, the 2 fluoroethanes, the 7 fluoropropanes, and the 10 fluorobutanes with lowest random numbers were set aside as a test set. The test set contained compounds #5, 8, 10, 14, 15, 23, 29, 35, 40, 41, 42, 50, 57, 62, 64, 65, 70, 71, 72, and 73 (compound numbers as in Table 2); it was representative for the 82 fluoroalkanes with respect not only to the number of C atoms but also to the number of F atoms and, most important, to the bp, as was confirmed by t-tests. For the remaining 62 compounds (training set) the multilinear regression using the seven descriptors from model 2 was calculated, resulting in

$$\begin{aligned} \mathrm{bp} = {} & 61.9784\cdot \mathrm{Xu}^{\mathrm{m}} - 4.54262\cdot {}^{2}\mathrm{P} - 1.91339\cdot \mathrm{S(ssssC)} \\ & -18.6994\cdot \mathrm{n(F-C-F)} - 3.94542\cdot \mathrm{n(F-C-C-F)} \\ & - 50.4978\cdot (\mathrm{relN_F})^2 + 8.09058\cdot \mathrm{n(FH_2C-CHF)} \\ & - 71.6826 \qquad \text{(model 2a)} \end{aligned}$$

$$r^2 = 0.98768,\ s = 4.68230,\ F = 618,\ N = 62$$

Model 2a was then used to predict the bps of the test set, resulting in $r^2 = 0.98370$, $s = 6.49742$, $F = 103$, $N = 20$. Figure 4 is a plot of experimental vs calculated (by model 2a) boiling points for the training set (open symbols) and the test set (closed symbols).

Whatever the significance of these numbers may be, the following should be noticed. (i) Such numbers may tell us something about the quality of model 2a but not of model 2. (ii) There is a very large number of possible partitions into a training and a test set. To obtain information on model 2, the entirety of these should be considered, which is obviously impossible for practical reasons. (iii) In building model 2a or each of its cousins, the information obtainable from 25% of the observations is discarded. (iv) The procedure leading to model 2a is problematic since descriptor selection for the training set is dictated by the data of all compounds including the test set.

The tool of cross-validation should avoid all these disadvantages. Various methods of validation were discussed recently.[26,27]

REFERENCES AND NOTES

(1) Tzschucke, C. C.; Markert, C.; Bannwarth, W.; Roller, S.; Hebel, A.; Haag, R. Modern Separation Techniques for Efficient Workup in Organic Synthesis. *Angew. Chem., Int. Ed.* **2002**, *41*, 3964−4000.

(2) (a) Reid, C. R.; Prausnitz, J. M.; Sherwood, J. K. *The Properties of Gases and Liquids*, 4th ed.; McGraw-Hill: New York, 1987. (b) Horvath, A. L. *Molecular Design*; Elsevier: Amsterdam, 1992.

(3) (a) Balaban, A. T.; Joshi, N.; Kier, L. B.; Hall, L. H. Correlations between Chemical Structure and Normal Boiling Points of Halogenated Alkanes $C_1 - C_4$. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 233−237. (b) Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. D. Correlation between Structure and Normal Boiling Points of Haloalkanes C_1-C_4 Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1118−1121.

(4) Carlton, T. S. Correlation of Boiling Points with Molecular Structure for Chlorofluoroethanes. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 158−164.

(5) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Quantitative Structure−Property Relationship Study of Normal Boiling Points for Halogen-/Oxygen-/Sulfur-Containing Organic Compounds Using the CODESSA Program. *Tetrahedron* **1998**, *54*, 9129−9142.

(6) Rücker, C.; Meringer, M.; Kerber, A. QSPR Using MOLGEN-QSPR: The Example of Haloalkane Boiling Points. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2070−2076.

(7) Gakh, A. A.; Gakh, E. G.; Sumpter, B. G.; Noid, D. W.; Trowbridge, L. D.; Harkins, D. A. Estimation of the Properties of Hydrofluorocarbons by Computer Neural Networks. *J. Fluorine Chem.* **1995**, *73*, 107−111.

(8) (a) Woolf, A. A. Boiling Point Relations in the Halogenated Ethane Series, *J. Fluorine Chem.* **1990**, *50*, 89−99. (b) Woolf, A. A. Predicting Boiling Points of Hydrofluorocarbons. *J. Fluorine Chem.* **1996**, *78*, 151−154.

(9) Horvath, A. L. Boiling Points of Halogenated Organic Compounds. *Chemosphere* **2001**, *44*, 897−905.

(10) The NIST Chemistry Webbook, http://webbook.nist.gov.

(11) Burdon, J.; Garnier, L.; Powell, R. L. Fluorination of Propane over Cobalt(III) trifluoride and Potassium tetrafluorocobaltate(III). *J. Chem. Soc., Perkin Trans. 2* **1996**, 625−631.

(12) Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. MOLGEN-QSPR, A Software Package for the Study of Quantitative Structure−Property Relationships. *MATCH Commun. Math. Comput. Chem.* **2004**, *51*, 187−204.

(13) Rücker, G.; Rücker, C. On Topological Indices, Boiling Points, and Cycloalkanes. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 788−802.

(14) By obviously unreasonable we understand, e.g., a bp value that violates the fundamental rule of a considerable increase in bp on enlarging a molecule by a CH_2 group. Consider the following sequences (all bps in reliability classes 1 or 2): CH_3-CH_2F −37.5, $CH_3-CHF-CH_3$ −9.7, $CH_3-CHF-CH_2-CH_3$ 25; CH_3-CHF_2 −25, $CH_3-CF_2-CH_3$ −0.5, $CH_3-CF_2-CH_2-CH_3$ 31; FH_2C-CH_3 −37.5, $FH_2C-CH_2-CH_3$ −2.5, $FH_2C-CH_2-CH_2-CH_3$ 32.5. From these values the bp increment for a CH_2 group can be estimated to be 25−35 °C (for additional examples see ref 14a). Now consider FH_2C-CH_2F, bp 10.45 or 27 (reliability class unattributable[15]), $FH_2C-CHF-CH_3$ 32 (reliability class 0), $FH_2C-CHF-CH_2-CH_3$ 39 (reliability class 0). There

is logic in this sequence only if for FH_2C-CH_2F 10.45, and for $FH_2C-CHF-CH_3$ 32 are essentially correct. For $FH_2C-CHF-CH_2-CH_3$ 39 °C[14b] then is far too low, this compound/bp therefore was excluded. (a) Balaban, A. T.; Basak, S. C.; Mills, D. Normal Boiling Points of 1,ω-Alkanedinitriles: The highest Increment in a Homologous Series. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 769. (b) Baklouti, A.; El Gharbi, R. *J. Fluorine Chem.* **1979**, *13*, 297−314.

(15) The bp of FH_2C-CH_2F is somewhat problematic: An experimental bp of 10−11 °C was reported only once,[15a] while there are several more recent reports of bp values between 25 and 31 °C for this compound, though not all of these are independent.[15b−h] However, the former number fits much better into various bp models. In the NIST Chemistry Webbook, 10.45 °C is the only bp value given for FH_2C-CH_2F.[10] (a) 10−11 °C, Henne, A. L.; Renoll, M. W. *J. Am. Chem. Soc.* **1936**, *58*, 889. (b) 30.7 °C, Edgell, W. F.; Parts, L. *J. Am. Chem. Soc.* **1955**, *77*, 4899. (c) 26−26.2 °C, Titow, *Dokl. Akad. Nauk SSSR* **1957**, *113*, 358. (d) 31 °C, Klaboe, P.; Nielsen, J. R. *J. Chem. Phys.* **1960**, *33*, 1764−1774. (e) 26 °C, Schiemann, G.; Cornils, B. *Chem. Ber.* **1965**, *98*, 3418−3435. (f) 26.5 °C, Abraham, R. J.; Kemp, R. H. *J. Chem. Soc. B* **1971**, 1240−1245. (g) 25−27 °C, Middleton, W. J. *J. Org. Chem.* **1975**, *40*, 574−578. (h) 30.7 °C, Nappa, N. J.; Sievert, A. C., *J. Fluorine Chem.* **1993**, *62*, 111−118.

(16) Braun, J.; Kerber, A.; Meringer, M.; Rücker, C. Similarity of Molecular Descriptors: The Equivalence of Zagreb Indices and Walk Counts. *MATCH Commun. Math. Comput. Chem.*, in press.

(17) Zefirov, N. S.; Palyulin, V. A. Fragmental Approach in QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1112−1122.

(18) Rücker, C.; Braun, J.; Kerber, A.; Laue, R. The Molecular Descriptors Computed with MOLGEN. http://www.mathe2.uni-bayreuth.de/molgenqspr.

(19) (a) Rücker, G.; Rücker, C. Automatic Enumeration of All Connected Subgraphs. *MATCH Commun. Math. Comput. Chem.* **2000**, *41*, 145−149. (b) Rücker, G.; Rücker, C. On Finding Nonisomorphic Connected Subgraphs and Distinct Molecular Substructures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 314−320.

(20) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure−Property Relationship. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28−41.

(21) The bp used in the preliminary study for 1,2-difluoroethane is 26 °C (as, presumably, in ref 5), and this leads to the by far highest residual, 16.2 °C. If 1,2-difluoroethane is excluded, for the remaining 42 compounds the best 6-descriptor model found is Xu, ${}^0\chi^v$, ${}^4\chi_p$, S(ssssC), n(F−C−F), n(F−C−C−F) (model 0a) $r^2 = 0.9931$, $s = 3.580$, $F = 839$, $r^2_{cv} = 0.9897$, $s_{cv} = 4.384$, $N = 42$. When instead fluoromethane was excluded (for this compound the flexibility index Φ is undefined, which may have been a reason for exclusion), then the best model we found is Xu, ${}^0\chi^v$, ${}^4\chi_p$, S(ssssC), n(F−C−F), n(F−C−C−F) (model 0b) $r^2 = 0.9874$, $s = 4.667$, $F = 456$, $r^2_{cv} = 0.9807$, $s_{cv} = 5.771$, $N = 42$.

(22) In the preliminary study fragment counts and the additional descriptors such as bip and xsF were not included in the descriptor pool. If they are included, then models of even slightly better r^2 and s values can be found, which however exhibit worse r^2_{cv} and s_{cv} values.

(23) Ren, B. Atomic-Level-Based AI Topological Descriptors for Structure−Property Correlations. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 161−169.

(24) Kier, L. B.; Hall, L. H. *Molecular Structure Description. The Electrotopological State;* Academic Press: San Diego, 1999.

(25) (a) Bonchev, D.; Trinajstic, N. Overall Molecular Descriptors. 3. Overall Zagreb Indices. *SAR QSAR Environ. Res.* **2001**, *12*, 213−236. (b) Bonchev, D. Overall Connectivity − A next Generation Molecular Connectivity. *J. Mol. Graphics Model.* **2001**, *20*, 65−75.

(26) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Model.* **2002**, *20*, 269−276.

(27) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.

CI0497298