QSPR Using MOLGEN-QSPR: The Example of Haloalkane Boiling Points

Christoph Rücker,* Markus Meringer, and Adalbert Kerber

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

Received June 17, 2004

MOLGEN-QSPR is a software newly developed for use in quantitative structure property relationships (QSPR) work. It allows to import, to manually edit, or to generate chemical structures, to detect duplicate structures, to import or to manually input property values, to calculate the values of a broad pool of molecular descriptors, to establish QSPR equations (models), and using such models to predict unknown property values. In connection with the molecule generator MOLGEN, MOLGEN-QSPR is able to predict property values for all compounds in a predetermined structure space (inverse QSPR). Some of the features of MOLGEN-QSPR are demonstrated on the example of haloalkane boiling points. The data basis used here is broader than in previous studies, and the models established are both more precise and simpler than those previously reported.

INTRODUCTION

Halogenated hydrocarbons find many uses as solvents, blowing agents, anesthetics, in refrigerating systems, fireproof coatings, etc., and therefore they finally end up in the environment. There they are of much concern (greenhouse gases, agents damaging the ozone layer). One fundamental physical property ruling the spread of a compound in the environment is its volatility, which in turn is easily characterized by its boiling point at normal pressure (bp). The boiling points of many halogenated hydrocarbons are known; however, there is still a need for predictions. For example, the number of possible acyclic saturated compounds containing in their molecules one through four carbon atoms, at least one F, Cl, or Br atom, and no elements other than C, H, F, Cl, and Br is 28600 by exhaustive and redundance-free construction (neglecting stereoisomers). In the Beilstein database no more than 1264 of these compounds appear, and for 988 of these some information on a boiling point is given (not always the boiling point at normal pressure).

Several authors analyzed the boiling points of halogenated hydrocarbons using various statistical methods. Balaban et al. in 1992 correlated the boiling points of 532 halo- and polyhaloalkanes C_1-C_4 with topological indices by multilinear regression (MLR).[1] Essentially the same set of compounds was treated by MLR using a broader set of descriptors including electrostatic and quantum chemical indices.[2] The boiling points of more restricted and of more general series of compounds were correlated with simple arithmetic and topological descriptors using artificial neural networks,[3] MLR,[4,5] or the k nearest neighbors method.[6] Boiling points of 240 haloalkanes and haloalkenes recently compiled by Horvath[7] (including several bps not reported earlier) were studied by Öberg by principal component analysis and partial-least squares regression (PLSR).[8]

In some of the studies mentioned the software package CODESSA was used.[2,5] While we in an earlier study used the general-purpose statistics package SAS,[9] in the meantime we developed our own QSPR software called MOLGEN-QSPR,[10] which combines structure generation with calculation of many molecular descriptors of various types and with data treatment by various statistical methods.

METHODS

Data Verification. The canonizer[11] built-in in MOLGEN-QSPR detected duplicates in all haloalkane samples taken from the literature. Removing duplicates may either improve or (more frequently) worsen the statistics of a model, depending on how well the duplicate end points fit. In any case, the model without duplicates is certainly the more correct.

The boiling points at normal pressure of all compounds (herein given in °C) were checked against the Beilstein database in order to avoid fitting wrong data. 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. Further, we excluded obviously unreasonable boiling points.[12] We did not examine the primary sources given in Beilstein and thus cannot exclude the possibility that a few bp values accepted here are calculated rather than experimental values.[10]

Boiling points were attributed to reliability classes as in our earlier work:[9] 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 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. Information on reliability, though not used in the calculations, proved very useful for identification of dubious bp data and in the selection of reference data.

Descriptor Calculation. For all compounds the values of various molecular descriptors were calculated by MOLGEN-QSPR. These include arithmetic descriptors (number of atoms, of atoms of specific elements, molecular weight, number of bonds, etc.), topological indices (Wiener index,

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

10.1021/ci049802u CCC: $27.50
Published on Web 10/14/2004

Table 1. Full Models[a]

$$bp = -171.719 \cdot relN_F + 64.3147 \cdot relN_I + 43.8371 \cdot {}^1\chi^s + 0.0161573 \cdot G_2(topo.dist) - 75.1416 \cdot \chi_T - 0.446177 \cdot {}^2TC^v - 10.4681 \quad (m1)$$

$$bp = -169.681 \cdot relN_F + 70.67 \cdot relN_I + 44.594 \cdot {}^1\chi^s + 0.0156703 \cdot G_2(topo.dist) - 73.0697 \cdot \chi_T - 0.450738 \cdot {}^2TC^v - 13.1558 \quad (m2)$$

$$bp = -154.835 \cdot relN_F + 62.126 \cdot relN_{Br} + 189.584 \cdot relN_I + 49.1184 \cdot {}^1\chi^s - 93.33 \cdot \chi_T - 0.643605 \cdot {}^2TC^v + 0.0989034 \cdot {}^2TM1 - 1.13343 \quad (m3)$$

$$bp = -146.86 \cdot relN_F + 52.6505 \cdot {}^1\chi^s + 111.825 \cdot SCA1 - 23.9133 \cdot slogP - 0.957835 \cdot {}^2TC^v + 0.615364 \cdot {}^3TC_c - 234.709 \quad (m4)$$

$$bp = -143.075 \cdot relN_F + 55.9487 \cdot {}^1\chi^s + 100.113 \cdot SCA1 - 20.0704 \cdot slogP - 0.874827 \cdot {}^2TC^v + 1.31124 \cdot {}^4TC_c - 224.676 \quad (m5)$$

$$bp = 247.274 \cdot relN_H + 164.776 \cdot relN_{Cl} + 123.72 \cdot {}^1\chi^s - 100.128 \cdot {}^1\chi^v - 133.73 \cdot \chi_T - 2.72462 \cdot S(sF) - 130.907 \quad (m6)$$

$$bp = 16.3798 \cdot \Phi - 4.73024 \cdot {}^4\chi_{pc} + 9.20771 \cdot n(C-C-F) + 3.68688 \cdot n(C-C-Cl) + 6.88957 \cdot sMR - 0.435968 \cdot {}^2TC^v - 127.846 \quad (m7)$$

$$bp = -80.7738 \cdot relN_F - 59.5275 \cdot relN_= + 17.2007 \cdot {}^2\kappa_\alpha - 4.02546 \cdot AI(sCH_3) + 6.86834 \cdot n(C-C-F) + 6.25861 \cdot sMR - 0.148077 \cdot {}^3TC^v - 103.396 \quad (m8)$$

$$bp = -133.284 \cdot relN_F + 23.3129 \cdot {}^0\chi^v + 12.3372 \cdot \Phi + 18.8312 \cdot FRB + 1.96738 \cdot S(ssssC) - 2.16586 \cdot n(C-C-C-F) - 62.2489 \quad (m9)$$

$$bp = -132.647 \cdot relN_F - 79.7429 \cdot relN_{Cl} + 65.2762 \cdot {}^1\kappa_\alpha - 7.46179 \cdot mwc2 - 7.53387 \cdot S(sCH_3) - 1.4616 \cdot S(sF) + 4.08375 \cdot n(C-C-F) - 120.115 \quad (m10)$$

$$bp = -165.437 \cdot relN_F + 44.2922 \cdot relN_{Br} + 155.035 \cdot relN_I + 55.2678 \cdot {}^1\chi^s - 163.328 \cdot SCA2 - 0.484855 \cdot {}^2TC^v + 14.5579 \quad (m11)$$

$$bp = -161.81 \cdot relN_F + 62.5918 \cdot relN_{Br} + 189.91 \cdot relN_I + 48.6681 \cdot {}^1\chi^s - 93.5478 \cdot \chi_T - 0.600777 \cdot {}^2TC^v + 0.0874494 \cdot {}^2TM1 + 0.45034 \quad (m12)$$

$$bp = -155.376 \cdot relN_F + 60.6623 \cdot relN_{Br} + 49.138 \cdot {}^1\chi^s - 93.462 \cdot \chi_T - 0.625984 \cdot {}^2TC^v + 0.0898582 \cdot {}^2TM1 - 0.624743 \quad (m13)$$

$$bp = -153.251 \cdot relN_F + 73.1663 \cdot relN_{Br} + 53.3144 \cdot {}^1\chi^s + 100.227 \cdot SCA1 - 16.7507 \cdot slogP - 0.828538 \cdot {}^2TC^v + 1.12749 \cdot {}^4TC_c - 223.678 \quad (m14)$$

[a] For explanation of the descriptors involved see Table 2.

connectivity, and valence connectivity indices χ and χ^v of various order, solvation connectivity indices, eccentric connectivity and total connectivity index, κ and κ_α shape indices, Balaban's J and Hosoya's Z index, Basak's information content indices, molecular walk counts, molecular path counts, gravitational indices, topological charge indices, principal eigenvalues of the adjacency and the distance matrix, χ and χ^v indices of subgraphs of type path, cluster, and path-cluster, etc.), electrotopological state and AI indices, geometrical indices (van der Waals volume, van der Waals surface, solvent-accessible surface etc. of the lowest-energy conformation, as determined by a built-in molecular mechanics module), counts of all individual substructures of a predefinable range of numbers of bonds, counts of user-defined fragments, Bonchev's overall topological indices,[13] and Crippen's slogP and sMR.[14] For more information on the descriptors and for references to the original literature see a recent book[15] and the MOLGEN-QSPR documentation.[16]

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, 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 the display of 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

Balaban's Data. Balaban et al. found as the best 6-regressor MLR model for the boiling points of 532 haloalkanes C_1-C_4 the following:[1]

$$[{}^1\chi^v - {}^0\chi^v],\ {}^D\chi^0,\ {}^1\chi,\ N_{Br},\ N_I,\ [{}^2\chi^v - {}^1\chi^v]$$

$$r^2 = 0.97,\ s = 10.94,\ F = 2953$$

(In the text we describe a MLR model by the descriptors involved and by its r^2, *s*, F, r^2_{cv}, and s_{cv} values. For full models see Tables 1 and 2.) MOLGEN-QSPR now reproduced this result. However, thanks to the broad descriptor pool available in MOLGEN-QSPR, for the same data set the following best 6-descriptor MLR model was now found ($N = 532$).

$$relN_F,\ relN_I,\ {}^1\chi^s,\ G_2(topo.dist),\ \chi_T,\ {}^2TC^v \quad (m1)$$

$$r^2 = 0.9853,\ s = 7.820,\ F = 5869,\ r^2_{cv} = 0.9848,\ s_{cv} = 7.952$$

Thus an improvement of more than 3 °C in the *s* value was achieved. Note that the six descriptors in model m1 are all simple arithmetic and topological descriptors. Despite this, model m1 is better than the best previously found 6-descriptor MLR model containing electrostatic and quantum chemical descriptors, $s = 8.6$.[2]

The original data from ref 1 were treated here for the single purpose to allow comparisons such as those just given. Otherwise, to obtain more correct results, the input data set was scrutinized and found to require modification. The canonizer built-in in MOLGEN-QSPR detected two cases of duplicate structures in Balaban's data.[17a] Due to missing or conflicting bp data in Beilstein, 22 boiling points had to be excluded,[17b] and others required more or less severe changes. One boiling point in the original data is obviously unreasonable and was therefore excluded.[17c] For this modified data set ($N = 507$) the best 6-descriptor MLR model found is

$$relN_F,\ relN_I,\ {}^1\chi^s,\ G_2(topo.dist),\ \chi_T,\ {}^2TC^v \quad (m2)$$

$$r^2 = 0.9862,\ s = 7.435,\ F = 5968,\ r^2_{cv} = 0.9857,\ s_{cv} = 7.573$$

Table 2. Descriptors Appearing in the Models[15,16]

$relN_H$, $relN_F$, $relN_{Cl}$, $relN_{Br}$, $relN_I$: number of H (F, Cl, Br, I) atoms divided by the number of all atoms
$relN_{=}$: number of double bonds diveded by the number of all bonds between non-hydrogen atoms
${}^0\chi^v$, ${}^1\chi^v$: Kier and Hall valence chi indices of zeroth and first order
Φ: Kier and Hall molecular flexibility index
FRB: number of freely rotatable bonds
$S(sCH_3)$, S(ssssC), S(sF): sum of Kier and Hall electrotopological state indices for all methyl C (tetrasubstituted C, fluorine) atoms
n(C−C−C−F), n(C−C−F), n(C−C−Cl): occurrence number of the respective substructure
${}^1\kappa_\alpha$, ${}^2\kappa_\alpha$: Kier and Hall alpha-modified shape indices of first and second order
mwc2: molecular walk count of order 2
G_2(topo.dist): gravitational index, $\Sigma_{bonds} w_i w_j$, where w_i is the atomic weight of atom i, and the summation includes all bonds between non-hydrogen atoms
χ_T: total chi index
${}^4\chi_{pc}$: chi index for path-cluster subgraphs of four bonds
sMR and slogP: molecular refraction and logP calculated by the Wildman and Crippen method[14]
${}^2TC^v$, ${}^3TC^v$, ${}^2TM1$, ${}^3TC_c$, ${}^4TC_c$: Bonchev's overall topological indices:[13] overall valence connectivity index of subgraphs of two and of three bonds, overall first Zagreb index of subgraphs of two bonds, and overall connectivity index of cluster subgraphs of 3 or 4 bonds, respectively
$AI(sCH_3)$: Ren's AI index of methyl groups
${}^1\chi^s$: solvation connectivity index of first order
SCA1: sum of coefficients of principal eigenvector of the adjacency matrix
SCA2: SCA1 divided by the number of non-hydrogen atoms

The best 7-descriptor model is

$$relN_F,\ relN_{Br},\ relN_I,\ {}^1\chi^s, \chi_T,\ {}^2TC^v,\ {}^2TM1 \qquad (m3)$$

$r^2 = 0.9876$, $s = 7.067$, $F = 5670$, $r^2_{cv} = 0.9871$, $s_{cv} = 7.207$

Earlier for a subset of the 532 haloalkanes, 276 chloro-, fluoro-, and chlorofluoro(hydro)carbons, a neural network (5 topological descriptors as input, 5−10−1 architecture, 61 adjustable parameters) resulted in $s = 8.5$.[3] We now obtained as the best 6-descriptor MLR model for all chloro-, fluoro-, and chlorofluoro(hydro)carbons from ref 1 ($N = 278$ after removal of duplicates), using the bp data originally reported:

$$relN_F,\ {}^1\chi^s,\ SCA1,\ slogP,\ {}^2TC^v,\ {}^3TC_c \qquad (m4)$$

$r^2 = 0.9891$, $s = 7.307$, $F = 4117$, $r^2_{cv} = 0.9885$, $s_{cv} = 7.529$

For the modified data (changes made as described in note 17, $N = 269$)

$$relN_F,\ {}^1\chi^s,\ SCA1,\ slogP,\ {}^2TC^v,\ {}^4TC_c \qquad (m5)$$

$r^2 = 0.9904$, $s = 6.759$, $F = 4495$, $r^2_{cv} = 0.9898$, $s_{cv} = 6.976$

For almost the same sample (267 chloro-, fluoro- and chlorofluoro(hydro)carbons) Basak et al. tested the model-free method of k nearest neighbors and found the best $s = 23.7$ °C ($r^2 = 0.8705$) for $k = 5$, when 8 principal components composed of 59 descriptors were used.[6] We now revisited these bps, this time using MLR. After data verification as described above[18] MOLGEN-QSPR proposed the best 6-descriptor MLR model ($N = 257$):

$$relN_H,\ relN_{Cl},\ {}^1\chi^s,\ {}^1\chi^v,\ \chi_T,\ S(sF). \qquad (m6)$$

$r^2 = 0.9896$, $s = 6.415$, $F = 3957$, $r^2_{cv} = 0.9888$, $s_{cv} = 6.645$

For comparison with the above k-nearest neighbors result, we calculated for the same compound sample ($N = 257$) k-nearest neighbors fits for $k = 2, 3, 4, 5, 6$, using the 6 descriptors from model m6 without any principal component analysis.[19a] The r^2 values obtained are 0.9785, 0.9679, 0.9648, 0.9626, 0.9601, respectively, corresponding to $s = 12.83$, 13.58, 13.41, 13.38, 13.52 and $s_{cv} = 23.53$, 18.35, 17.30, 16.56, 16.77. At least for this data set the k-nearest neighbors method, though less powerful than MLR, is not as bad as it seemed from ref 6, provided a good set of descriptors is found.

Horvath's Data. An independent set of 240 haloalkanes and haloalkenes not containing iodine was recently extracted by Öberg from Horvath's compilation.[7] Öberg described the bps by a PLSR model in which six latent variables composed of 511 descriptor variables were used, and after exclusion of several outliers and partition into a calibration set and a test set $s_{fit} = 4.90$ and $s_{predict} = 6.17$ were obtained.[8]

For reasons of parsimony and portability we now undertook to treat the Horvath data using MLR. In the beginning, MOLGEN-QSPR detected many duplicates and even triplicates in that data set, so that the 240 compounds treated in ref 8 are in fact not more than 203.[20a] Again, we had to exclude several bps for contradictory reports in Beilstein and to change others[20b] and excluded two boiling points as obviously unreasonable.[20c] The best 6-descriptor MLR model for the modified Horvath data set ($N = 185$) is

$$\Phi,\ {}^4\chi_{pc},\ n(C-C-F),\ n(C-C-Cl),\ sMR,\ {}^2TC^v \qquad (m7)$$

$r^2 = 0.9857$, $s = 5.644$, $F = 2038$, $r^2_{cv} = 0.9843$, $s_{cv} = 5.908$

and the best 7-descriptor model is

$$relN_F,\ relN_{=},\ {}^2\kappa_\alpha,\ AI(sCH_3),\ n(C-C-F),\ sMR,\ {}^3TC^v \qquad (m8)$$

$r^2 = 0.9879$, $s = 5.207$, $F = 2057$, $r^2_{cv} = 0.9863$, $s_{cv} = 5.530$

For the same data and a somewhat restricted pool of descriptors (no Bonchev overall topological indices, no Crippen slogP or sMR) we compared a few statistical procedures. MLR yielded the best 6-descriptor model

$$\mathrm{relN_F},\ {}^0\chi^v,\ \Phi,\ \mathrm{FRB},\ \mathrm{S(ssssC)},\ \mathrm{n(C-C-C-F)} \quad \text{(m9)}$$

$r^2 = 0.9841,\ s = 5.934,\ F = 1841,\ r^2_{cv} = 0.9825,\ s_{cv} = 6.226$

and the best 7-descriptor model

$$\mathrm{relN_F},\ \mathrm{relN_{Cl}},\ {}^1\kappa_\alpha,\ \mathrm{mwc2},\ \mathrm{S(sCH_3)},\ \mathrm{S(sF)},\ \mathrm{n(C-C-F)} \quad \text{(m10)}$$

$r^2 = 0.9862,\ s = 5.542,\ F = 1813,\ r^2_{cv} = 0.9848,\ s_{cv} = 5.819$

An artificial neural network (6−2−1 architecture, using the six descriptors from model m9 as input) resulted in $r^2 = 0.9870$, $s = 5.535$, $F = 796$.[19b] Regression trees[21] or support vector machines,[22] alternative methods offered by MOLGEN-QSPR, did not result in improvement.

Combined Data. The haloalkanes from refs 1, 4, 5, and 7 were combined into one set containing (due to overlap) 606 compounds, 573 of which have usable boiling point data. The best 6-descriptor MLR model found for this combined sample ($N = 573$) is

$$\mathrm{relN_F},\ \mathrm{relN_{Br}},\ \mathrm{relN_I},\ {}^1\chi^s,\ \mathrm{SCA2},\ {}^2\mathrm{TC}^v \quad \text{(m11)}$$

$r^2 = 0.9851,\ s = 7.448,\ F = 6239,\ r^2_{cv} = 0.9846,\ s_{cv} = 7.565$

and the best 7-descriptor model is

$$\mathrm{relN_F},\ \mathrm{relN_{Br}},\ \mathrm{relN_I},\ {}^1\chi^s,\ \chi_T,\ {}^2\mathrm{TC}^v,\ {}^2\mathrm{TM1} \quad \text{(m12)}$$

$r^2 = 0.9866,\ s = 7.067,\ F = 5950,\ r^2_{cv} = 0.9861,\ s_{cv} = 7.191$

For an application to be discussed below we need a QSPR equation for the bps of iodine-free C_1-C_4 haloalkanes. This subset of the $N = 606$ set includes 507 compounds with usable bp data. The best 6-descriptor MLR model found for these is

$$\mathrm{relN_F},\ \mathrm{relN_{Br}},\ {}^1\chi^s,\ \chi_T,\ {}^2\mathrm{TC}^v,\ {}^2\mathrm{TM1} \quad \text{(m13)}$$

$r^2 = 0.9879,\ s = 6.875,\ F = 6787,\ r^2_{cv} = 0.9875,\ s_{cv} = 6.980$

and the best 7-descriptor model is

$$\mathrm{relN_F},\ \mathrm{relN_{Br}},\ {}^1\chi^s,\ \mathrm{SCA1},\ \mathrm{slogP},\ {}^2\mathrm{TC}^v,\ {}^4\mathrm{TC_c} \quad \text{(m14)}$$

$r^2 = 0.9888,\ s = 6.607,\ F = 6304,\ r^2_{cv} = 0.9884,\ s_{cv} = 6.737$

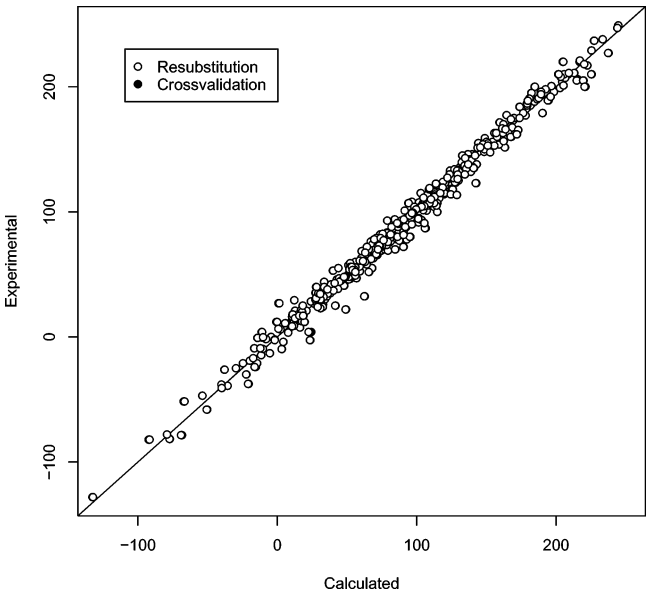

Figure 1. Plot of calculated (by model m14, white disks, and by leave-one-out cross-validation, black disks) vs experimental boiling points of the 507 iodine-free haloalkanes contained in model m14. Note that most black disks are eclipsed by the corresponding white disks.

Figure 1 is a plot of calculated vs experimental boiling points for model m14.

Descriptor Intercorrelation. Some of the models above contain at least one pair of highly intercorrelated descriptors. Worst in this respect is the 7-descriptor model m10, $s = 5.542$, where the coefficient of linear correlation $r({}^1\kappa_\alpha, \mathrm{mwc2})$ is 0.9861, $r(\mathrm{S(sF)},\mathrm{mwc2}) = 0.9287$, and $r(\mathrm{S(sF)},\mathrm{n(C-C-F)}) = 0.9493$. We therefore left out each descriptor in turn from model m10, producing seven 6-descriptor models. The s values of these are 9.142 ($\mathrm{relN_F}$ left out), 9.086 ($\mathrm{relN_{Cl}}$ left out), 16.618 (${}^1\kappa_\alpha$ left out), 10.099 (mwc2 left out), 6.039 ($\mathrm{S(sCH_3)}$ left out), 7.485 (S(sF) left out), and 6.353 (n(C−C−F) left out). Interestingly, s increases most on elimination of either of the most intercorrelated descriptors. We conclude that each descriptor in model m10, including the highly intercorrelated ones, catches important structural information not covered by the others. Had highly intercorrelated pairs of descriptors been excluded from model building, model m10 would not have been found.[23]

Limitation. We observed that the bps of the fluorohydrocarbons (not containing another halogen) are the most difficult to fit. Thus in all models above 1-fluorobutane, 1-fluoropropane, 1-fluoro-2-methylpropane, fluoroethane, and 2-fluorobutane have large negative residuals ($bp_{exp} - bp_{calc} < 0$), while difluoromethane, 1,2-difluoroethane, 1,1,2-trifluoroethane, 1,1,1,3-tetrafluoropropane, 1,1,1,3,3,3-hexafluoropropane, and 1,1,1,3,3,3,4,4,4-nonafluoroisobutane have large positive residuals. This seems to be due to the lack, in the present version of MOLGEN-QSPR, of descriptors taking account of strong bond dipoles. While this should be a problem in the other haloalkanes as well, in simple fluoroalkanes it may not be compensated for by other effects.[24]

Prediction. In several cases of boiling points excluded from the models for conflicting reports in Beilstein (see notes 17b, 18, and 20b), prediction by model m14 was helpful in deciding which reports are probably erroneous (not shown). For large scale prediction see the next section.

Table 3. Internal Validation of the Inside/Outside Range Classification by Model M14[a]

	calculated bp		
experimental bp	within range	outside range	total
within range	19	8	27
outside range	8	472	480
total	27	480	507

[a] Target range is 130−140 °C.

Inverse QSPR. Finally, we demonstrate the solution to a hypothetical problem of inverse QSPR. The inverse QSPR problem is to propose all compounds within a defined structure space that have a prescribed value of a particular property.[25] We assume here (quite arbitrarily) to require all acyclic haloalkanes C_1-C_4 that contain no elements other than C, H, F, Cl, and Br and have a bp between 130 and 140 °C. As mentioned in the Introduction, there are 28600 compounds in the given structure space (stereoisomerism neglected, generated by MOLGEN). Using equation m14, MOLGEN-QSPR predicted their bps, and as a result 655 compounds are predicted to boil between 130 and 140 °C at normal pressure.

Of course, the quality of most of the more than 28000 predictions cannot be evaluated, since nobody knows the corresponding experimental bps.[26] However, for the 507 compounds on which model m14 was built we can compare calculated and experimental bps (internal validation). The results are condensed in Table 3, from which the misclassification rate is calculated as 16/507 = 0.0316. Original data are shown in Table 4.

For an external validation we envisaged those compounds/bps contained in the Beilstein database but not included in Balaban's, Carlton's, or Horvath's data sets. As mentioned in the Introduction, there is some boiling point information in Beilstein for 988 iodine-free C_1-C_4 haloalkanes. After exclusion of all compounds treated above, there remained 455 iodine-free haloalkanes whose boiling points were never before correlated with their structures to the best of our knowledge. Many of these compounds have bp data at pressures quite different from normal only. After these and those with conflicting or obviously unreasonable bp data had been removed, 223 compounds with usable bps remained. Unfortunately, it turned out that these bps are of far lower reliability than those entered into model m14. Thus, of the 507 bps used for model m14, 151 (30%) are in reliability class 1 and 165 (32.5%) in reliability class 2. In contrast, of the 223 usable boiling points in the external validation set, 35 (16%) are in reliability class 1 and a single one (0.4%) is in reliability class 2.[27]

In our opinion, it does not make much sense to check the quality of predictions by comparison with experimental data that in themselves are not reliably known. Therefore, we predicted, using model m14, the bps of the 36 compounds in the external validation set that have experimental bps of reliability classes 1 and 2. The results are presented in Table 5, the misclassification rate is 2/36 = 0.0556. Table 6 shows the original data. For these data, $r^2(bp_{experimental},bp_{predicted}) = 0.9859$, to be compared to $r^2 = 0.9888$ for model m14.[28]

Results for the *complete* external validation set ($N = 223$) are shown in Table 7; original data are in Table 8. The misclassification rate now is 12/223 = 0.0538, while the correlation now is lower than above, $r^2(bp_{experimental},bp_{predicted}) = 0.9778$, presumably due to erroneous experimental bp values among the mostly unreliable data in this sample.

Table 4. Internal Validation of Model M14

	bp_{exp}	bp_{calc}	residual
Part A. Listed Are All Compounds in the $N = 507$ Set That Have a $bp_{experimental}$ between 130 and 140 °C			
Cl_3C-CH_2Cl	130.0	131.15	−1.15
$BrF_2C-CBr(CH_3)-CH_3$	130.0	123.69	6.31
$ClH_2C-CClF-CH_2Cl$	130.0	134.58	−4.58
BrH_2C-CH_2Br	131.5	130.35	1.15
$ClF_2C-CH_2-CCl_3$	132.0	138.76	−6.76
$F_3C-CBrCl-CCl_2F$	132.0	131.43	0.57
$ClH_2C-CCl(CH_3)-CF_2Cl$	132.0	138.76	−6.76
$H_3C-CHCl-CHCl_2$	133.0	123.98	9.02
$H_3C-CHCl-CH_2-CH_2Cl$	133.0	129.01	3.99
$H_3C-CH_2-CH_2-CCl_3$	133.5	138.40	−4.90
$BrF_2C-CBrF-CBrF_2$	133.6	130.52	3.08
$ClH_2C-CHBr-CClF_2$	134.0	139.89	−5.89
$H_3C-CH_2-CHBr_2$	134.0	127.11	6.89
$ClF_2C-CClF-CF_2-CCl_2F$	134.0	132.94	1.06
$ClF_2C-CClF-CClF-CClF_2$	134.0	132.94	1.06
$Cl_2FC-CCl_2-CHF_2$	134.6	135.54	−0.94
$H_3C-CCl_2-CCl_2F$	135.0	132.40	2.60
$ClH_2C-CHBr-CH_2-CF_3$	135.0	141.08	−6.08
Br_2CCl_2	135.0	140.79	−5.79
$ClH_2C-CHBrCl$	136.0	131.22	4.78
$FH_2C-CH_2-CH_2-CH_2Br$	136.0	140.11	−4.11
$ClH_2C-CH(CH_3)-CH_2Cl$	136.0	133.84	2.16
Cl_2HC-CH_2Br	137.0	134.25	2.75
Cl_3C-CCl_2F	138.0	138.10	−0.10
$BrH_2C-CF_2-CH_2Br$	138.0	142.99	−4.99
$FH_2C-CHBr-CH_2Cl$	138.0	136.94	1.06
$H_3C-CClF-CCl_3$	139.6	132.40	7.20
Part B. Listed Are All Compounds in the $N = 507$ Set That Have a $bp_{calculated}$ between 130 and 140 °C			
$F_3C-CHCl-CCl_3$	125.1	130.06	−4.96
BrH_2C-CH_2Br	131.5	130.35	1.15
$BrF_2C-CBrF-CBrF_2$	133.6	130.52	3.08
Cl_3C-CH_2Cl	130.0	131.15	−1.15
$ClH_2C-CHBrCl$	136.0	131.22	4.78
$F_3C-CBrCl-CCl_2F$	132.0	131.43	0.57
$H_3C-CClF-CCl_3$	139.6	132.40	7.20
$H_3C-CCl_2-CCl_2F$	135.0	132.40	2.60
$H_3C-CH_2-CBr_2-CH_3$	145.0	132.78	12.22
$ClF_2C-CClF-CClF-CClF_2$	134.0	132.94	1.06
$ClF_2C-CClF-CF_2-CCl_2F$	134.0	132.94	1.06
$ClH_2C-CH(CH_3)-CH_2Cl$	136.0	133.84	2.16
Cl_2HC-CH_2Br	137.0	134.25	2.75
$ClH_2C-CClF-CH_2Cl$	130.0	134.58	−4.58
$H_3C-CHCl-CCl_2-CH_3$	143.0	134.74	8.25
$Cl_2FC-CCl_2-CHF_2$	134.6	135.54	−0.94
$Cl_2HC-CHCl_2$	146.0	136.79	9.21
$FH_2C-CHBr-CH_2Cl$	138.0	136.94	1.06
$ClH_2C-CH_2-CH_2Br$	142.0	138.06	3.94
Cl_3C-CCl_2F	138.0	138.10	−0.10
$Br_2HC-CBrF_2$	144.0	138.30	5.70
$H_3C-CH_2-CH_2-CCl_3$	133.5	138.40	−4.90
$ClH_2C-CCl(CH_3)-CF_2Cl$	132.0	138.76	−6.76
$ClF_2C-CH_2-CCl_3$	132.0	138.76	−6.76
$H_3C-CHBr-CH_2Br$	141.0	139.24	1.76
$Br_2FC-CHBrF$	146.0	139.76	6.24
$ClH_2C-CHBr-CClF_2$	134.0	139.89	−5.89

Table 5. External Validation of the Inside/Outside Range Classification by Model M14[a]

	predicted bp		
experimental bp	within range	outside range	total
within range	5	0	5
outside range	2	29	31
total	7	29	36

[a] Target range is 130−140 °C. Results for the 36 compounds with bps of reliability class 1 or 2.

Table 6. External Validation of Model M14[a]

	bp_{exp}	bp_{pred}	residual
$(H_3C)_2CF-CH_2Cl$	72.0	74.1	−2.1
$H_3C-CH_2-CH_2-CF_3$	17.0	32.7	−15.7
$(H_3C)_2CH-CF_3$	12.0	23.7	−20.7
$H_3C-CH_2-CF_2-CH_2Cl$	83.0	80.4	2.6
$ClH_2C-CHF-CH_2Cl$	128.0	115.0	13.0
$ClH_2C-CHCl-CH_2F$	119.5	116.7	2.8
$ClH_2C-CH_2-CH_2-CF_3$	86.0	81.5	4.5
$H_3C-CH_2-CF_2-CHCl_2$	111.0	106.1	4.9
$ClH_2C-CH_2-CCl_2F$	118.0	122.1	−4.1
$F_3C-CH_2-CH_2-CF_3$	24.5	37.3	−12.8
$BrH_2C-CH_2-CH_2-CH_2Cl$	176.0	170.6	5.4
$BrH_2C-CHF-CH_2Cl$	148.0	138.3	9.7
H_3C-CCl_2Br	99.0	97.7	1.3
$H_3C-CHCl-CF_2-CHClF$	105.0	107.5	−2.5
$ClH_2C-CHBr-CH_2Cl$	177.5	170.4	7.1
$H_3C-CHCl-CCl_2-CH_2Cl$	180.0	185.6	−5.6
$BrClHC-CHClF$	125.0	119.2	5.8
BrH_2C-CCl_2F	110.5	113.5	−3.0
$F_2HC-CHF-CF_2-CF_3$	32.5	28.2	4.3
$F_2HC-CF_2-CHF-CF_3$	35.0	31.8	3.2
$FH_2C-CF_2-CF_2-CF_3$	26.5	24.7	1.8
BrH_2C-CCl_3	152.0	157.0	−5.0
$H_3C-CH_2-CHBr-CH_2Br$	166.0	168.9	−2.9
$F_3C-CCl_2-CF_3$	36.0	36.3	−0.3
$BrClHC-CCl_2F$	137.0	137.8	−0.8
$F_3C-CHF-CF_2-CF_2Cl$	44.5	45.0	−0.5
$F_2HC-CF_2-CClF-CF_2Cl$	83.0	83.1	−0.1
Br_2FC-CF_3	46.5	50.1	−3.6
$BrCl_2C-CBrF_2$	139.0	137.0	2.0
$BrClFC-CBrClF$	140.0	137.0	3.0
$(H_3C)_2CBr-CHBr_2$	207.0	191.6	15.4
$H_3C-CHBr-CBr_2-CH_3$	207.0	188.1	18.9
$(F_3C)_2CCl-CCl_3$	134.0	130.7	3.3
$F_3C-CCl_2-CCl_2-CF_3$	134.0	132.9	1.1
$F_3C-CHBr-CHBr-CF_3$	118.0	132.7	−14.7
$Cl_3C-CF_2-CF_2-CCl_3$	208.5	203.4	5.1

[a] Results for the 36 compounds with bps of reliability class 1 or 2.

Table 7. External Validation of the Inside/Outside Range Classification by Model M14[a]

	predicted bp		
experimental bp	within range	outside range	total
within range	9	6	15
outside range	6	202	208
total	15	208	223

[a] Target range is 130−140 °C. Results for all 223 compounds in the external set.

Obviously, the inside/outside range classification as well as the experimental vs predicted correlation will improve if a QSPR model better than m14 is found and used. For a given model the classification success depends on the ratio of *s* and the target range width. Thus the hit rate will improve or worsen if the desired range is made broader or narrower, respectively.

CONCLUSION

From the statistics obtained we conclude that our MLR models describe the haloalkane bp data surprisingly well and are of useful predictive power. We did not exclude any compound as outlier, and all MLR models given here contain no more than 6 or 7 simple indices, that is no more than 7 or 8 adjustable parameters. All descriptors used in the present study are arithmetic, topological, or geometric indices, i.e., they are of the simplest available types, obtained directly from the chemical structure. Of these, notably, the geometric indices did not qualify to be contained in the final models. The models given here are of the simplest possible type (MLR) and thus conform to the parsimony principle.[29,30] Nevertheless the results obtained are better (in terms of *s* values) than those of previous attempts found in the literature. The success is due in part to critical evaluation of input data and in part to a very broad pool of available descriptors including walk counts, substructure counts, and overall topological indices. For the data treated here, the variety of statistical methods offered by MOLGEN-QSPR was of no advantage compared to MLR. For other data, the picture may be quite different.

Table 8. External Validation of Model M14

	bp_{exp}	bp_{pred}	residual
Part A. Listed Are All Compounds in the $N = 223$ Set That Have a $bp_{experimental}$ between 130 and 140 °C			
$ClFHC-CF_2-CClF-CBrF_2$	130.0	132.49	−2.49
$BrH_2C-C(CH_3)Br-CF_3$	131.0	142.65	−11.65
$BrH_2C-CH_2-CBrF-CF_3$	131.0	141.95	−10.95
$ClH_2C-C(CH_3)F-CH_2Cl$	133.0	126.63	6.37
$FH_2C-CH_2-CClF-CHClF$	133.5	126.21	6.79
$(F_3C)_2ClC-CCl_3$	134.0	130.69	3.31
$F_3C-CCl_2-CCl_2-CF_3$	134.0	132.94	1.06
$ClF_2C-CCl_2-CF_2-CClF_2$	134.2	132.94	1.26
$BrH_2C-CH_2-CCl_2F$	136.0	145.40	−9.40
$BrH_2C-CH_2-CF_2-CBrF_2$	136.0	141.95	−5.95
$BrClHC-CCl_2F$	137.0	137.81	−0.81
$Br_2ClC-CClF_2$	137.0	136.96	0.04
$BrCl_2C-CBrF_2$	139.0	136.96	2.04
$Cl_3C-CH_2-CF_2-CH_3$	139.2	133.81	5.39
$BrClFC-CBrClF$	140.0	136.96	3.04
Part B. Listed Are All Compounds in the $N = 223$ Set That Have a $bp_{predicted}$ between 130 and 140 °C			
$BrH_2C-CH_2-CF_2-CHBrF$	129.0	130.49	−1.49
$(F_3C)_2ClC-CCl_3$	134.0	130.69	3.31
$ClFHC-CF_2-CClF-CBrF_2$	130.0	132.49	−2.49
$F_3C-CHBr-CHBr-CF_3$	118.0	132.74	−14.74
$FCl_2C-CF_2-CF_2-CCl_2F$	127.5	132.94	−5.44
$F_3C-CCl_2-CF_2-CCl_2F$	127.0	132.94	−5.94
$ClF_2C-CCl_2-CF_2-CClF_2$	134.2	132.94	1.26
$F_3C-CCl_2-CCl_2-CF_3$	134.0	132.94	1.06
$Cl_3C-CH_2-CF_2-CH_3$	139.2	133.81	5.39
$Br_2ClC-CClF_2$	137.0	136.96	0.04
$BrCl_2C-CBrF_2$	139.0	136.96	2.04
$BrClFC-CBrClF$	140.0	136.96	3.04
$BrClHC-CCl_2F$	137.0	137.81	−0.81
$BrH_2C-CHF-CH_2Cl$	148.0	138.27	9.73
$H_3C-CF_2-CCl_2-CH_2Cl$	141.0	138.76	2.24

Supporting Information Available: A list of the 507 compounds in models m13/m14 with ID numbers, experimental boiling points, reliability of boiling point values, calculated (by model m14) boiling points, residuals, and structures. ID numbers are Bjkh*nnn*, Car*nn*, and Hor*nnn*, where *nnn* or *nn* is a natural number, and Bjkh*nnn,* Car*nn,* and Hor*nnn* represent compound number *nnn* or *nn* in refs 1, 4, and 7, respectively. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) 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.

(2) 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.

(3) 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) 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.

(6) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Estimation of the Normal Boiling Points of Haloalkanes Using Molecular Similarity. *Croat. Chem. Acta* **1996**, *69*, 1159−1173.

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

(8) Öberg, T. Boiling Points of Halogenated Aliphatic Compounds: A Quantitative Structure−Property Relationship for Prediction and Validation. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 187−192.

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

(10) 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.

(11) Braun, J.; Gugisch, R.; Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. MOLGEN-CID − A Canonizer for Molecules and Graphs Accessible through the Internet. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 542−548.

(12) By obviously unreasonable we understand a bp value that violates a fundamental rule such as the considerable increase in bp on enlarging a molecule by a CH_2 group or on substituting an H by a Cl or (even more) a Br atom, or that violates one of the rules given in ref 1.

(13) (a) Bonchev, D. The Overall Wiener Index − A New Tool for Characterization of Molecular Topology. *J. Chem. Inf Comput. Sci.* **2001**, *41*, 582−592. (b) Bonchev, D. Overall Connectivity − A next Generation Molecular Connectivity. *J. Mol. Graphics Modell.* **2001**, *20*, 65−75.

(14) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(15) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.

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

(17) (a) Compounds #53 (bp 179 °C) and #242 (bp 193 °C) are identical, as are #86 (bp 88 °C), and #291 (bp 88 °C). (b) For this reason the following were excluded: ##39, 45, 46, 49, 100, 123, 160, 165, 191, 221, 298, 330, 345, 354, 363, 438, 441, 465, 473, 491, 529, 532. (c) The bp of compound #387, $ClF_2C-CH_2-CFCl-CH_3$, 129 °C (reliability class 0), was excluded as unreasonable: The four reliably known bps of lower homologues $C_3H_3F_3Cl_2$, compounds #98, #279, #280, #282 (ref 1 numbering), are $CF_3-CCl_2-CH_3$ 49 °C, $FCl_2C-CF_2-CH_3$ 60 °C, $CF_3-CH_2-CHCl_2$ 72 °C, and $CF_3-CHCl-CH_2Cl$ 76.5 °C, all in reliability class 1. Adding to these values the contribution of a terminal methyl group, we expect for compound #387 a bp of at most slightly above 100 °C. Compare also the bp of compound #281, $ClF_2C-CFCl-CH_3$, 55.6 °C (reliability class 0). Further, there is an isomer of compound #387 in the data, #389, $ClF_2C-CH_2-CHF-CH_2Cl$ with bp 118.5 °C (reliability class 0). Structures #387 and #389 are related in the same manner as are #104 and #94, $H_3C-CFCl-CH_3$ and $H_3C-CHF-CH_2Cl$ (bp 35.2 °C and 68.5 °C, respectively, both in reliability class 2). Accordingly, #387 should exhibit a bp well below 118.5 °C.

(18) Two cases of duplicates were detected and removed: In ref 6 compounds #48 (bp 165.5 °C) and #52 (bp 153 °C) are identical, as are #71 (bp 108 °C) and #261 (bp 104 °C). Seven boiling points were excluded for conflicting reports in Beilstein, those of compounds #35, 41, 42, 92, 177, 246, and 254.

(19) (a) For this purpose, all descriptor values were autoscaled by MOLGEN-QSPR. (b) Result without descriptor preprocessing. When the descriptors were subjected to autoscaling, $r^2 = 0.9880$, $s = 5.315$, $F = 864$ was obtained.

(20) (a) In ref 7 structures ##25, 271, and 366 are identical, as are the following: #43 = #367, #47 = #147, #58 = #62, #70 = #390, #76 = #84, #104 = #307 = #404, #110 = #311 = #407, #112 = #315 = #409, #126 = #321 = #416, #131 = #300, #137 = #299 = #393, #146 = #309 = #310 = #405 = #406 (stereoisomers not distinguished), #172 = #489, #174 = #490, #272 = #368, #282 = #380, #288 = #388, #293 = #389, #296 = #391, #314 = #408, #317 = #417, #319 = #414, #320 = #413, #324 = #418, #325 = #419, #327 = #421, and #330 = #426. (b) For this reason we could not use the boiling points of compounds ##29, 36, 39, 45, 49, 70, 79, 82, 89, 117, 126, 131, 146, 211, 296, and 319 (numbering from ref 7). (c) An obviously unreasonable bp is listed in ref 7 for compound #308, $CH_3-CF_2-CCl_2Br$ (35.5 °C). Compare the reliably known bps of $CH_3-CF_2-CH_2Cl$ (55 °C), $CH_3-CF_2-CHCl_2$ (79.2 °C), $CH_3-CF_2-CCl_3$ (102 °C), $CH_3-CF_2-CHClBr$ (101 °C), and HCF_2-CCl_3 (73 °C). Likewise the bp listed for compound #415, $CH_2Br-CF_2-CHBr_2$ (139 °C), is unreasonable, compared to the reliably known bp of #322, $CH_2Br-CF_2-CH_2Br$ (138 °C).

(21) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth International: Belmont, CA, 1984.

(22) Vapnik, V. *The Nature of Statistical Learning*; Springer: New York, 1995.

(23) (a) Randić, M. Orthogonal Molecular Descriptors. *New J. Chem.* **1991**, *15*, 517−525. (b) Randić, M. On Characterization of Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672−687.

(24) When 1-fluorobutane, 1-fluoropropane, 1-fluoro-2-methylpropane, difluoromethane, 1,2-difluoroethane, 1,1,2-trifluoroethane, and 1,1,1,3-tetrafluoropropane were excluded from model m14, the best 7-descriptor model still contained the descriptors of model m14, but the statistics improved to $r^2 = 0.9905$, $s = 6.049$, $F = 7328$, $r^2_{cv} = 0.9901$, $s_{cv} = 6.174$ ($N = 500$). For a treatment of the bps of C_2 and C_3 fluoroalkanes, see: Woolf, A. A. Predicting Boiling Points of Hydrofluorocarbons. *J. Fluorine Chem*. **1996**, *78*, 151−154, and references therein.

(25) (a) Skvortsova, M. I.; Baskin, I. I.; Slovokhotova, O. L.; Palyulin, V. A.; Zefirov, N. S. Inverse Problem in QSAR/QSPR Studies for the Case of Topological Indices Characterizing Molecular Shape (Kier Indices). *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 630−634. (b) Kier, L. B.; Hall, L. H. The Generation of Molecular Structures from a Graph-Based QSAR Equation. *Quant. Struct.-Act. Relat.* **1993**, *12*, 383−388. (c) Kvasnička, V.; Pospichal, J. Simulated Annealing Construction of Molecular Graphs with Required Properties. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 516−526.

(26) Predicted bps resulting from extrapolation should not be trusted. For example, bps are predicted for perbromobutane and perbromoisobutane, C_4Br_{10}, to be 516.6 °C and 514.4 °C, respectively. These are severe extrapolations since the compounds entered into model m14 contain 4 Br atoms at most and exhibit bps of not higher than 250 °C. Bps higher than about 250 °C should not be trusted anyway, since at temperatures that high many compounds will decompose.

(27) As was to be expected therefrom, the bps in the external validation set are more difficult to fit by our methods than those treated above.

(28) The calculated bp values given in Tables 5−8 are predictions in the sense that the corresponding experimental bps were not used for establishing the model. This is a legitimate use of the term "prediction", along with its use for the calculation of values when experimental data are not known to anybody. On the other hand, some authors use the term for calculation of values by a model when the corresponding experimental values had been used for establishing the model. The latter use in our opinion is a misuse.

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

(30) Lučić, B.; Nadramija, D.; Bašic, I.; Trinajstić, N. Toward Generating Simpler QSAR Models: Nonlinear Multivariate Regression versus Several Neural Network Ensembles and Some Related Methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1094−1102.

CI049802U