MOLGEN–QSPR, A SOFTWARE PACKAGE FOR THE STUDY OF QUANTITATIVE STRUCTURE PROPERTY RELATIONSHIPS

A. KERBER, R. LAUE, M. MERINGER, C. RÜCKER[1]

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

Received December 10, 2003

Abstract. A new software package MOLGEN–QSPR for the exploration of quantitative structure property relationships is introduced. Practical results obtained using this software are presented.

1. Introduction

Recently we introduced a computer program MOLGEN–COMB [1] that allows to generate virtual combinatorial libraries from a given set of reactants and reactions. Our new software package MOLGEN–QSPR provides methods for the study of quantitative structure property relationships (QSPRs) in combinatorial libraries and the prediction of property values for such virtual libraries. Figure 1 shows a simplified flowchart of QSPR search and application. Algorithmic parts are highlighted in grey. Figure 2 is a screenshot of MOLGEN-QSPR's graphical user interface (GUI).

The input of MOLGEN–QSPR is a set of chemical compounds given as molecular graphs together with values for a continuous target variable representing the physicochemical property under consideration. Examples are the boiling point or the density.

The QSPR search consists of three principal steps:

- structure preprocessing,
- descriptor computation,
- regression analysis.

There exist several opportunities for structure input:

Financial support by the Federal Ministry of Research and Technology is gratefully acknowledged.

[1]Corresponding author e–mail: ChristRckr@aol.com

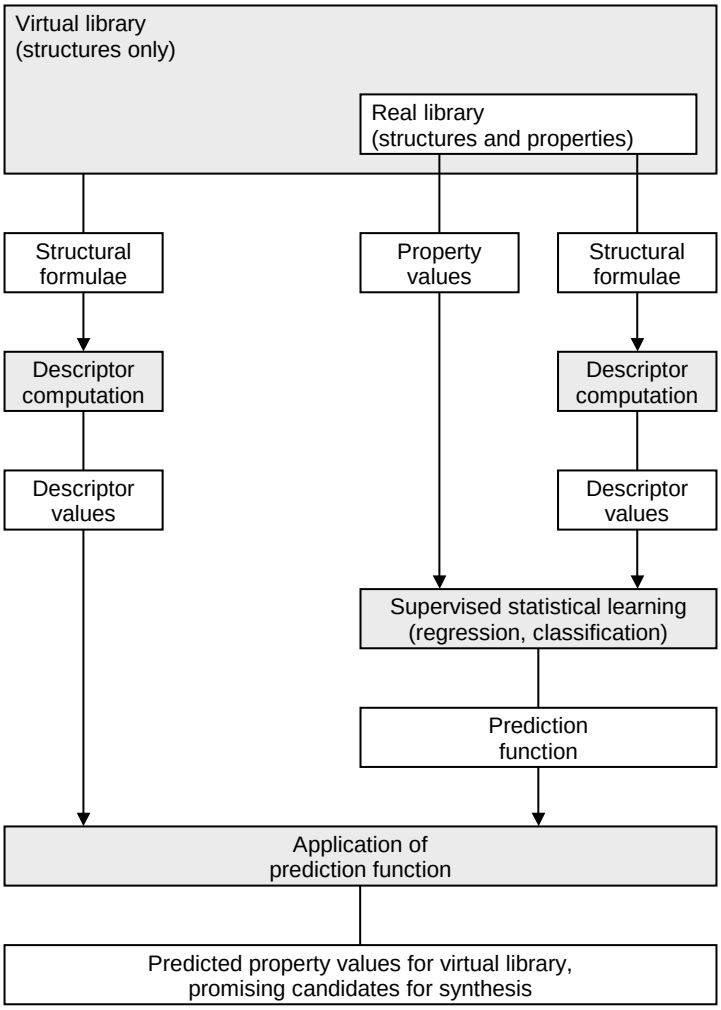

FIGURE 1. Flowchart of QSPR search and application

- structure import as MDL Mol– or SDfile[2],
- structure generation by MOLGEN–COMB,
- manual structure input using the included molecular structure editor MOLED.

Property values can be input from tabulator separated ASCII tables as written by EXCEL or added manually in MOLGEN–QSPR's GUI. Structures *and* corresponding property values can be imported using the format of CODESSA input files [2].

Structure preprocessing includes

- addition of H atoms, which are typically suppressed in electronic representations of molecular graphs,

[2]MDL Mol– and SDfiles are a widespread exchange formats for molecular structures based on connection tables. A detailed specification is available at www.mdl.com

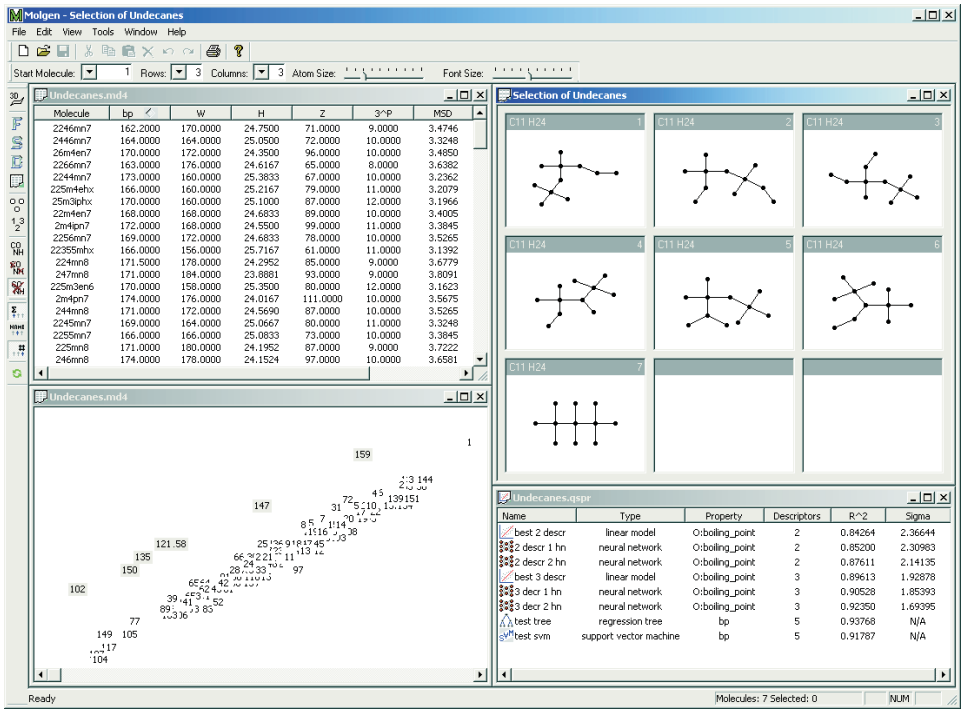

FIGURE 2. Screenshoot of MOLGEN–QSPR's GUI. Upper left window: Table of property and descriptor values. Lower left window: Plot of experimental vs calculated property values with a selection of 7 compounds. Upper right window: Structures of selected compounds. Lower right window: Table of calculated QSPRs.

- identification of aromatic bonds, which are often coded as alternating single and double bonds, and
- computation of a 3D layout using a force field model [3].

The latter is necessary if geometrical descriptors are to be applied.

Molecular descriptors are used in order to map molecular structures onto real numbers. Currently MOLGEN–QSPR provides 705 built–in descriptors of various types:

- arithmetical descriptors (using information coded in the compound's molecular formula),
- topological descriptors (using information coded in the compound's constitution),
- geometrical descriptors (using 3D information coded in the compound's configuration and conformation),
- electrotopological and AI indices [4, 5],
- overall indices [6, 7],
- Crippen indices [8].

A detailed specification of the available indices is provided in [9]. Besides these indices substructure counts can be used as molecular descriptors. MOLGEN–QSPR supplies an algorithm that generates systematically all substructures (with optional lower and upper limits for the number of bonds) that occur in the library. User–defined substructures also can be counted.

Once the descriptor values are calculated, methods of supervised[3] statistical learning [10] are applied in order to find prediction functions that fit the target variable well. Along with ordinary least squares regression (OLS, based on a QR–decomposition of the design matrix defined by the descriptor values) and principal component regression (using the singular value decomposition of the design matrix) there are several sophisticated methods available via an interface to the (freely available) statistical software package R [11]:

- regression trees [12],
- artificial neural networks [13, 14],
- support vector machines [15].

In order to avoid overfitting it is necessary to find small subsets of descriptors that allow the calculation of good prediction functions. For this purpose there is an algorithm included that performs an exhaustive search for best subsets of descriptors for OLS regression. For problems with large numbers of molecules and descriptors and/or big subsets exhaustive search might be too time expensive. In order to handle such problems MOLGEN–QSPR offers an algorithm for stepwise subset selection.

The user may want to calculate several prediction functions for the same QSPR problem. In order to select the best model one needs a method for model assessment. An easy way is resubstitution, i.e. substituting descriptor values of the compounds of the real library into the prediction function and comparing the results with the experimental property values. Usually one calculates the sum of squares of the residuals (RSS) for this purpose.

However, a small RSS does not necessarily imply good predictive power. The predictive ability of a prediction function is best measured by an independent test set. Therefore MOLGEN–QSPR offers the possibility to define disjoint learning and test sets. For cases where the real library is too small for partition into learning and test set, MOLGEN–QSPR provides leave–one–out crossvalidation (based on OLS).

In addition several tools are included in MOLGEN–QSPR that execute elementary statistical tasks. For instance arithmetic mean, standard deviation, value distribution of descriptors and properties can be computed as well as their correlation matrix.

[3] It is called "supervised" learning because the presence of the target variable guides the learning process and acts as a "teacher".

In the following we describe two case studies of QSPRs which were performed using MOLGEN–QSPR. Initially, we intended to use these cases simply as tests for our program. However, the results obtained were unexpected, and in retrospect they reveal typical traps to be avoided in the use and interpretation of QSPR equations.

2. Boiling Points of Undecanes and Dodecanes

Boiling points of alkanes at atmospheric pressure (bp, measured in °C) are a classical proving ground for QSPRs since Wiener's time [16]. We recently collected experimental boiling points of alkanes and (poly)cycloalkanes up to the decanes (saturated hydrocarbons of up to 10 carbon atoms) and produced for them multilinear correlations with topological descriptors [17]. In that work, once more the utmost importance became evident to use in QSPR work experimental property values from highly reliable sources only.

To the best of our knowledge, boiling points of higher alkanes, such as the undecanes and dodecanes, were never before correlated with such descriptors, though a complete compilation of boiling points for all 159 undecanes and all 355 dodecanes exists (stereoisomers not distinguished)[4]. These data come from a source thought to be particularly reliable, the American Petroleum Institute Research Project 44, associated with the name of F.D. Rossini. Therefore, using these data, we established QSPR equations for the bps of undecanes and, separately, of dodecanes. The best 2– and 3–descriptor correlations found are as follows:

Undecanes: $bp = 25.799 MSD + 4.4696 P^{(3)} + 35.823,$
$r^2 = 0.84264,\ s = 2.3664,\ n = 159,$

$bp = 1.7113 W + 35.141 H + 0.51917 Z - 1031.5,$
$r^2 = 0.89613,\ s = 1.9288,\ n = 159,$

where MSD (mean square distance) is a topological index introduced by Balaban [18], $P^{(3)}$ is the count of paths of length 3 [16], W is the Wiener index [16, 19], H is the Harary index [20], and Z is the Hosoya index [19].

Dodecanes: $bp = 25.620 MSD + 4.3113 P^{(3)} + 42.768,$
$r^2 = 0.83502,\ s = 2.5285,\ n = 355,$

[4]Selected Values of Properties of Hydrocarbons and Related Compounds. Supplementary Vol. No. A–78 of the Thermodynamics Research Center (TRC) Hydrocarbon Project Publication, dated April 30, 1979, Tables 45a and 46a. (The TRC Hydrocarbon Project was formerly the American Petroleum Institute Research Project 44.) TRC, formerly located at the Chemical Engineering Division of the Texas Engineering Experiment Station, Texas A&M University, College Station, Texas, USA, is now at the National Institute of Standards and Technology (NIST), Boulder, Colorado, USA.

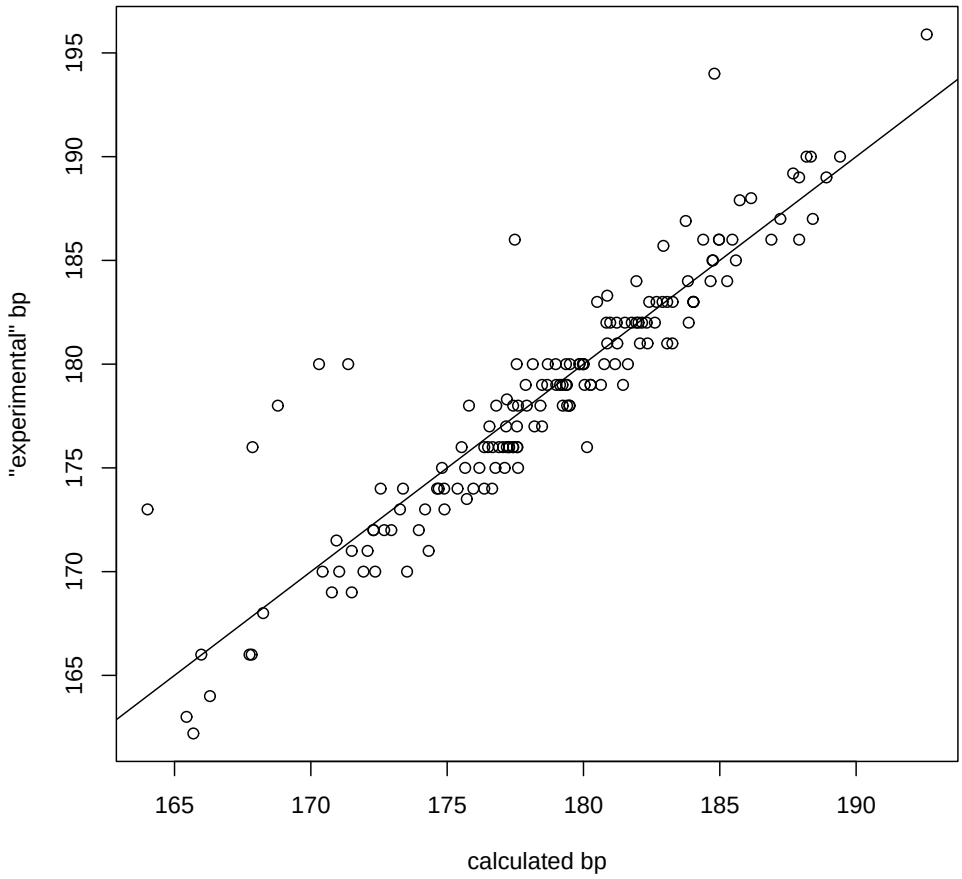

FIGURE 3. Bps of undecanes (2–descriptor model)

$$bp = 59.933MSD + 27.856H + 0.32152Z - 862.76,$$
$$r^2 = 0.88734,\ s = 2.0924,\ n = 355.$$

The quality of these equations, as expressed by the r^2 and s values, is somewhat lower than desirable, this fact is reflected in Figures 3 – 6, scatterplots of experimental bps vs bps calculated by the above equations for undecanes and dodecanes, respectively.

As seen in Figures 3 and 5, there is, in both cases, a minor population of alkanes having experimental bps about 10°C higher than expected from the correlation valid for the majority of compounds. MOLGEN–QSPR allows easy identification of those special compounds, clicking on the symbols in the plot displays the structures. The special undecanes and dodecanes are shown in Figures 7 and 8. Interestingly, all these compounds have one structural feature in common, the 2,2,4,4–tetramethylpentane substructure. None of the majority compounds contains this substructure, and each compound containing this substructure is in the minor population. At this point we believed to have

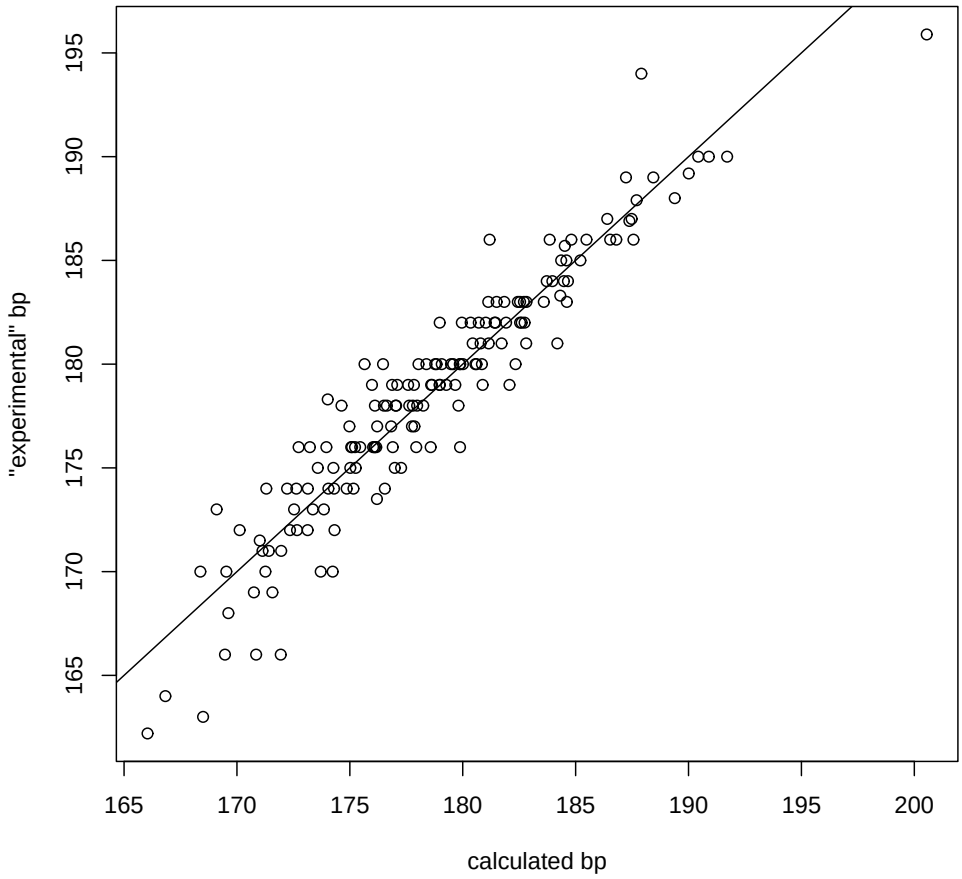

FIGURE 4. Bps of undecanes (3–descriptor model)

detected an interesting effect on the bp somehow associated with this substructure.

However, to be sure that the effect was not an artifact due to insufficient care in compound preparation and isolation, we asked Professor R.C. Wilhoit of the Thermodynamics Research Center (TRC) for more information on the origin of the undecanes and dodecanes bp data. The response was quite unexpected[5], in essence it said that

[5]Thanks are due to R.C. Wilhoit for his kind cooperation. The following is a quote from his e–mail letter to C.R.: "The tables you mention, 45a and 46a on boiling points of C11 and C12 alkanes, have not been revised since 1956. However you will find a Specific Reference sheet for them on pages a–ref–1580 through a–ref–1607. The numbers there shown for each property refer to the list at the end of each table. The entries in that list identify the authors of the original source of data. The complete reference is given in the "General List of References": which is in Volume XIV of the current set of tables. In tables published for about the past 25 years the complete references is given directly in the Specific Reference sheets. In the Specific Reference sheet for table 45a, starting on page a–ref–1580, you will see that numbers 4 and 12 are listed for most boiling points. 12 refers to "American

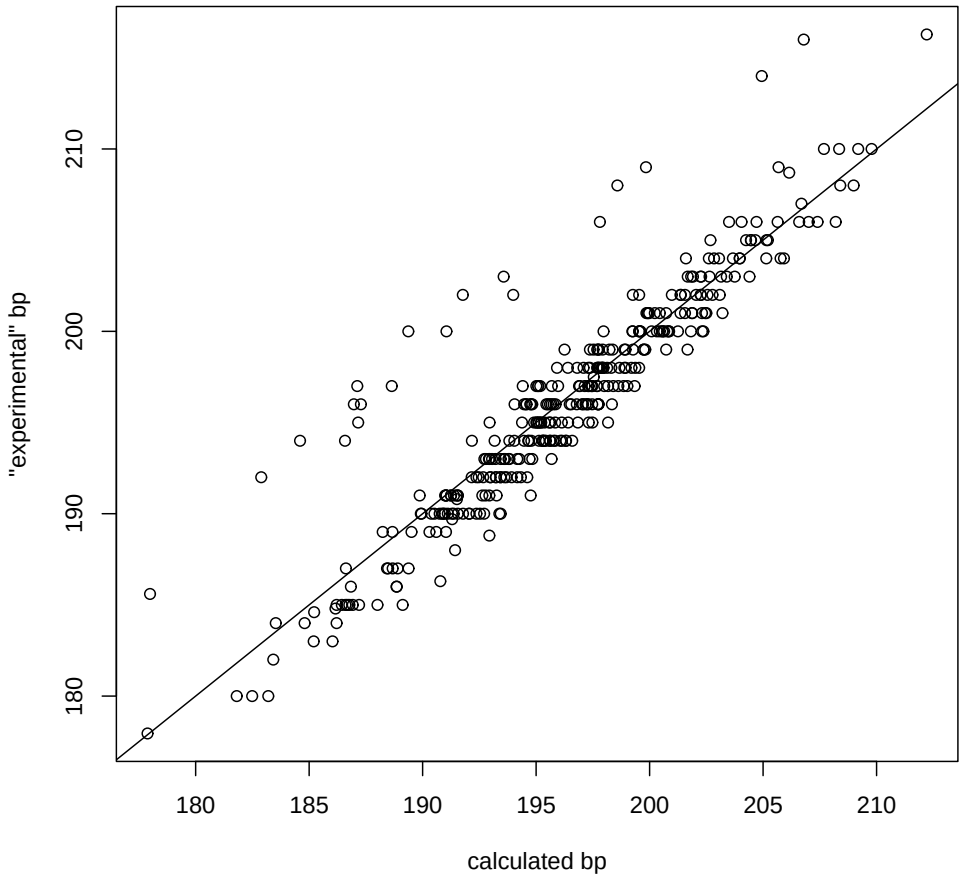

FIGURE 5. Bps of dodecanes (2–descriptor model)

most of the undecane and dodecane bp data in the TRC database[3] are not at all experimental, rather they were calculated using an empirical QSPR equation, as described in a paper by Rossini [21]. Reading that paper one finds that alkane bps were calculated therein according to a Wiener–type equation, but a "corrective term" of +10.6°C was added for exactly those compounds containing the 2,2,4,4–tetramethylpentane substructure, for reasons not given in detail.

The "effect" found in our investigation thus is nothing but an artifact due to the "experimental" bp values in fact being calculated values. The nice aspect of this story is that MOLGEN–QSPR was able

Petroleum Institute Research Project 44". This means that the boiling point was calculated from an empirical correlation. The procedure, number 4, identifies the author, Greenshields–1. If you then look this up in the General List of References you will find "Greenshields, J.B., Thesis, Ohio State ...". Actually this procedure was published by the next entry, Greenshields and Rossini, J. Phys. Chem., 1958, 62, 271. ... Similarly for the C12 alkanes, Specific References for Table 46a, ... Numbers 1,5 identify those obtained by correlation by the same procedure."

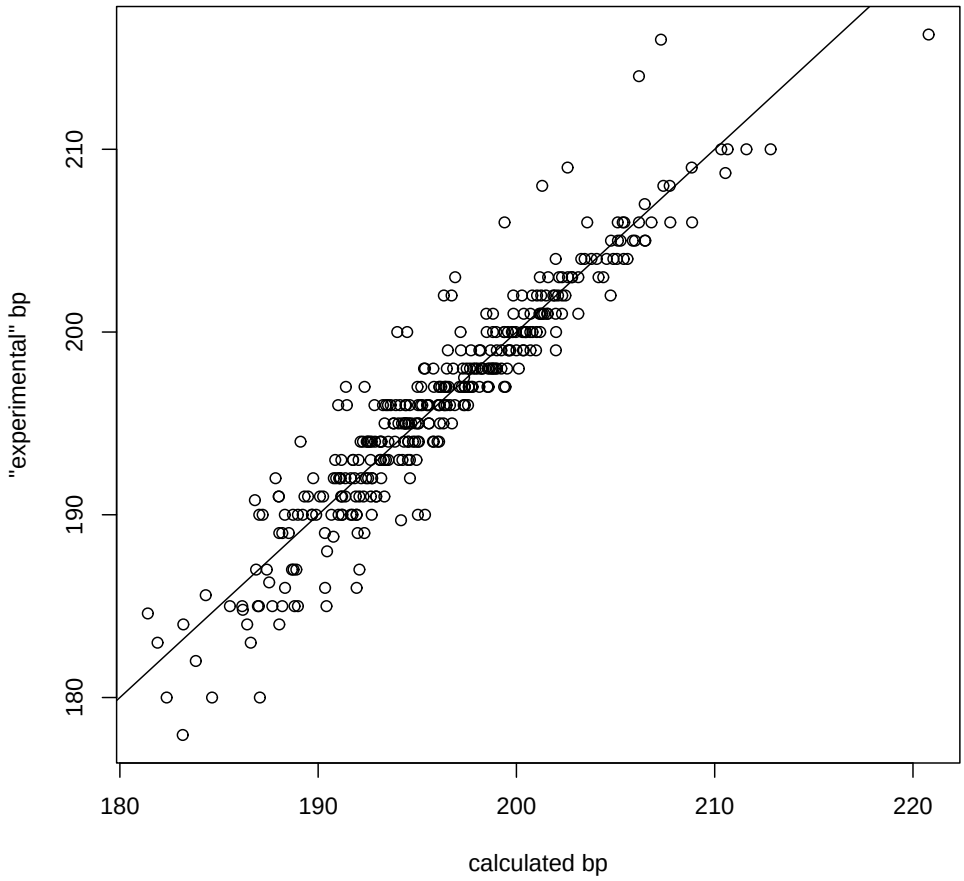

FIGURE 6. Bps of dodecanes (3–descriptor model)

to detect the inconsistency in the data due to this "corrective term", application of which thus seems not to have been a very good idea. The lesson to be learnt for us is that one can never be careful enough in the critical evaluation of experimental data.

In the meantime, the TRC is addressing the problem of both experimental and calculated data being contained in one and the same database [22].

3. ^{17}O NMR Chemical Shifts

In their book "Molecular Structure Description — The Electrotopological State" [4] Kier and Hall wrote

> "Two studies have been reported (Kier and Hall, 1990; Hall et al., 1991) in which the E–State of oxygen atoms in a series of ethers and a series of carbonyl compounds was calculated. These indices were compared to the ^{17}O

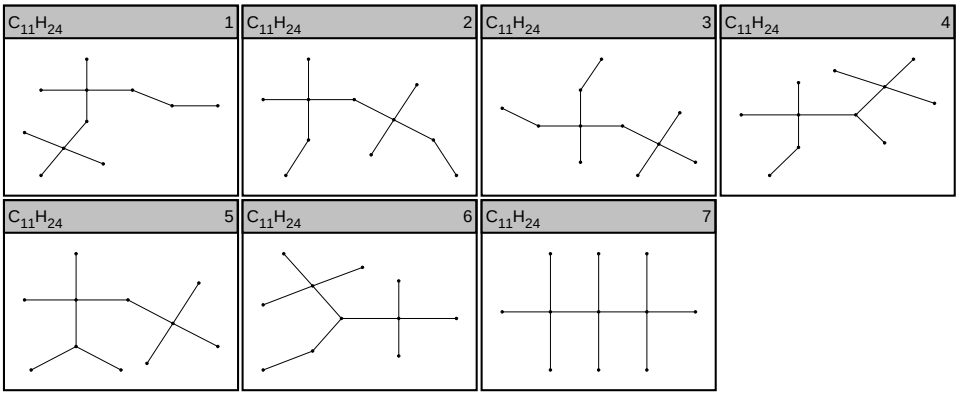

FIGURE 7. Undecanes with unexpected "experimental" bps

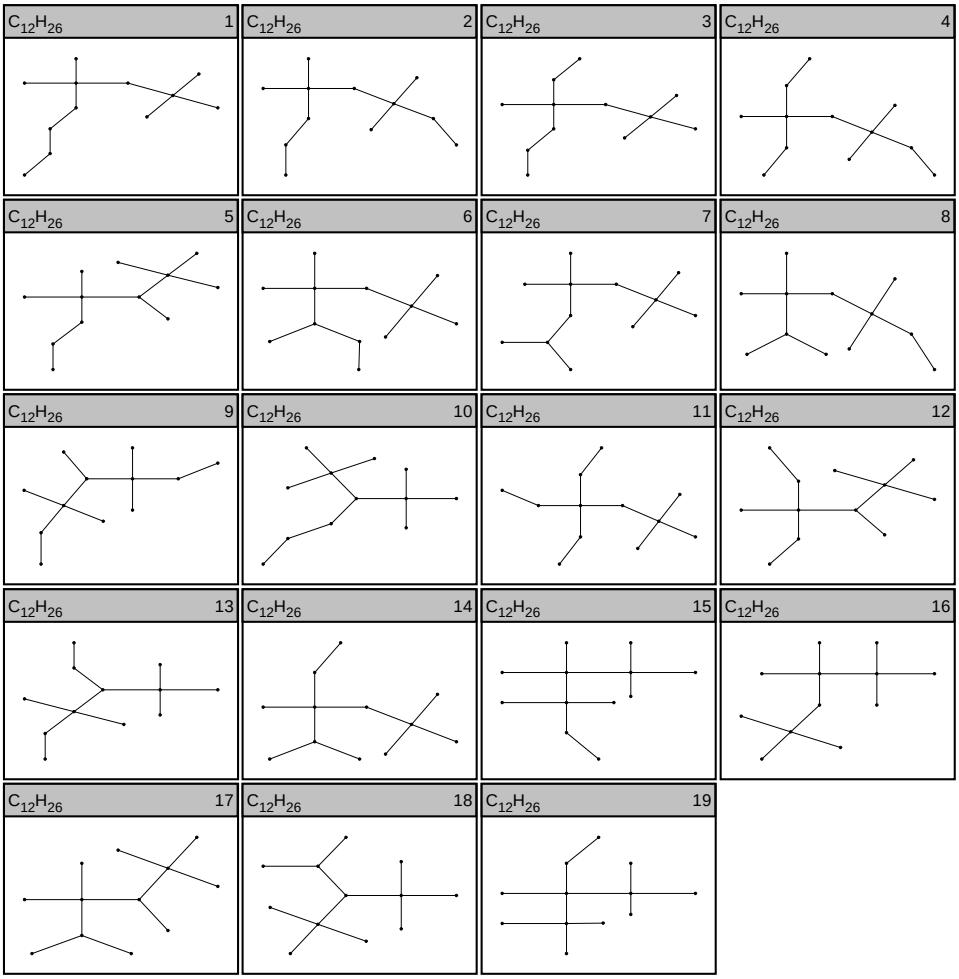

FIGURE 8. Dodecanes with unexpected "experimental" bps

> chemical shifts ... . The correlations between the E–State values and the chemical shifts (Δ) are very close.
>
> For ethers: $\Delta = 92.56S(\text{-O-}) - 441.65$,
> $r^2 = 0.99$, $s = 4.3$, $n = 10$,
>
> For carbonyls: $\Delta = -27.77S(=\text{O}) + 834.48$,
> $r^2 = 0.97$, $s = 3.67$, $n = 9$.
>
> Clearly, the E–State indices encode relevant structure information influencing this property."

Here S(-O-) and S(=O) are the electrotopological state values of an ether oxygen and a carbonyl oxygen atom, respectively.

3.1. **Ethers.** In the original paper [23] Kier and Hall correlated experimental ^{17}O chemical shifts (in ppm units, from a paper by Delseth and Kintzinger [24]) of ten saturated acyclic ethers with the oxygen E–state values, the QSPR equation reads

$$\delta(^{17}\text{O}) = 92.564S(\text{-O-}) - 441.65,$$
$$r = 0.995,\ s = 4.3,\ n = 10.$$

MOLGEN–QSPR was able to reproduce this result, as follows:

$$\delta(^{17}\text{O}) = 95.129S(\text{-O-}) - 454.88,$$
$$r^2 = 0.99005,\ s = 4.2171,\ n = 10.$$

However, this equation is not the best single–descriptor model. Linear regression using the simple Randić index $^1\chi$ gives $r^2 = 0.99365$, $s = 3.3679$, and the Kier and Hall valence connectivity index $^1\chi^v$ gives $r^2 = 0.99539$, $s = 2.8718$.

More importantly, we stumbled over the fact that in the Delseth–Kintzinger work there are experimental ^{17}O shifts given not only for the 10 ethers treated by Kier and Hall, but for 31 saturated acyclic ethers altogether. For the 21 remaining ethers or for the combined sample of all 31 ethers, the above equation or the best model based on S(-O-) for the respective sample is far less satisfying, as are the alternative models using $^1\chi$ or $^1\chi^v$.

Figure 9 is a plot of $\delta(^{17}\text{O})$ vs S(-O-) for all 31 ethers, where the compounds are identified by the ID numbers attributed to them in [24], and the 10 compounds contained in the smaller sample are represented by filled circles.

For the complete ether sample ($n = 31$), the best multilinear single–, 2–, and 3–descriptor models, as found using MOLGEN–QSPR, are:

1 descriptor: S(-O-), $r^2 = 0.5994$, $s = 19.6$,
2 descriptors: $^2\chi$, $^2\chi^v$, $r^2 = 0.9759$, $s = 4.9$,
3 descriptors: S(-O-), $^2\chi$, $^2\chi^v$, $r^2 = 0.9935$, $s = 2.6$.

These models are best as long as the descriptor pool to choose from contains arithmetic, topological and geometrical descriptors only. When substructure counts are additionally included in the descriptor pool,

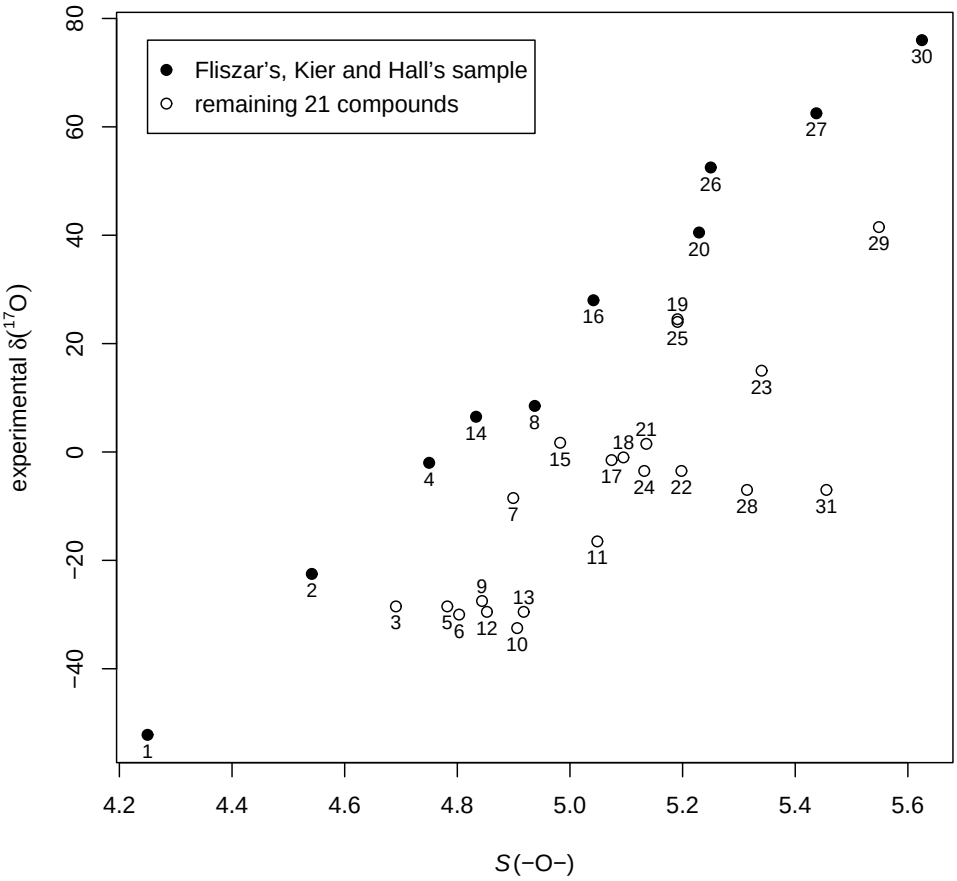

FIGURE 9. $\delta(^{17}\mathrm{O})$ of 31 ethers vs S(-O-). Compounds: 1 *dimethyl ether*, 2 *ethyl methyl ether*, 3 methyl propyl ether, 4 *methyl isopropyl ether*, 5 methyl butyl ether, 6 methyl isobutyl ether, 7 methyl sec–butyl ether, 8 *methyl tert–butyl ether*, 9 methyl pentyl ether, 10 methyl neopentyl ether, 11 methyl sec–pentyl ether, 12 methyl isopentyl ether, 13 methyl neohexyl ether, 14 *diethyl ether*, 15 ethyl propyl ether, 16 *ethyl isopropyl ether*, 17 ethyl butyl ether, 18 ethyl isobutyl ether, 19 ethyl sec–butyl ether, 20 *ethyl tert–butyl ether*, 21 ethyl pentyl ether, 22 ethyl neopentyl ether, 23 ethyl sec–pentyl ether, 24 dipropyl ether, 25 propyl isopropyl ether, 26 *diiso-propyl ether*, 27 *isopropyl tert–butyl ether*, 28 dibutyl ether, 29 di–sec–butyl ether 30 *di–tert–butyl ether*, 31 diisopentyl ether. Compounds in *italics* are contained in the 10–compound sample.

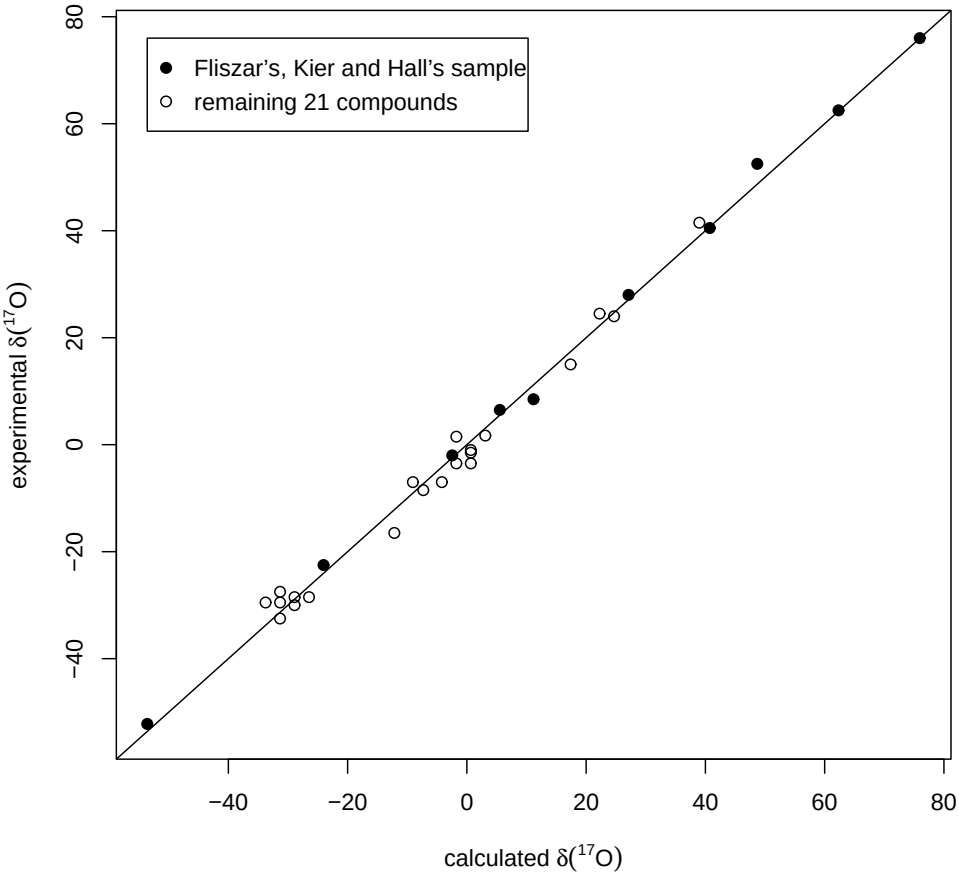

FIGURE 10. $\delta(^{17}\mathrm{O})$ of 31 ethers vs 3–descriptor model

better models are obtained. The best models found for the 31 ethers are:

1 descriptor: n(C-C-O), $r^2 = 0.9645$, $s = 5.8$,
2 descriptors: n(C-C-O), n(C-C(-C)-O), $r^2 = 0.9802$, $s = 4.4$,
3 descriptors: n(C-C-O), n(C-C(-C)-O), $P^{(3)}$, $r^2 = 0.9942$, $s = 2.4$,

where n(C-C-O) is the occurrence number of the substructure C-C-O, and $P^{(3)}$ is the number of paths of length 3 in the molecule. Figure 10 is a plot of experimental vs calculated values for $\delta(^{17}\mathrm{O})$ by the latter model.

The reason why Kier and Hall treated the 10–compound sample only seems to be that they intended to compare their model [25] to a model published earlier by Fliszar [26]. This author had published a correlation of ^{17}O chemical shifts with the net atomic charge of the ether O atom, as obtained by ab initio (STO–3G) calculations, for exactly these 10 ethers, rather than for all 31 ethers with known ^{17}O NMR chemical shifts.

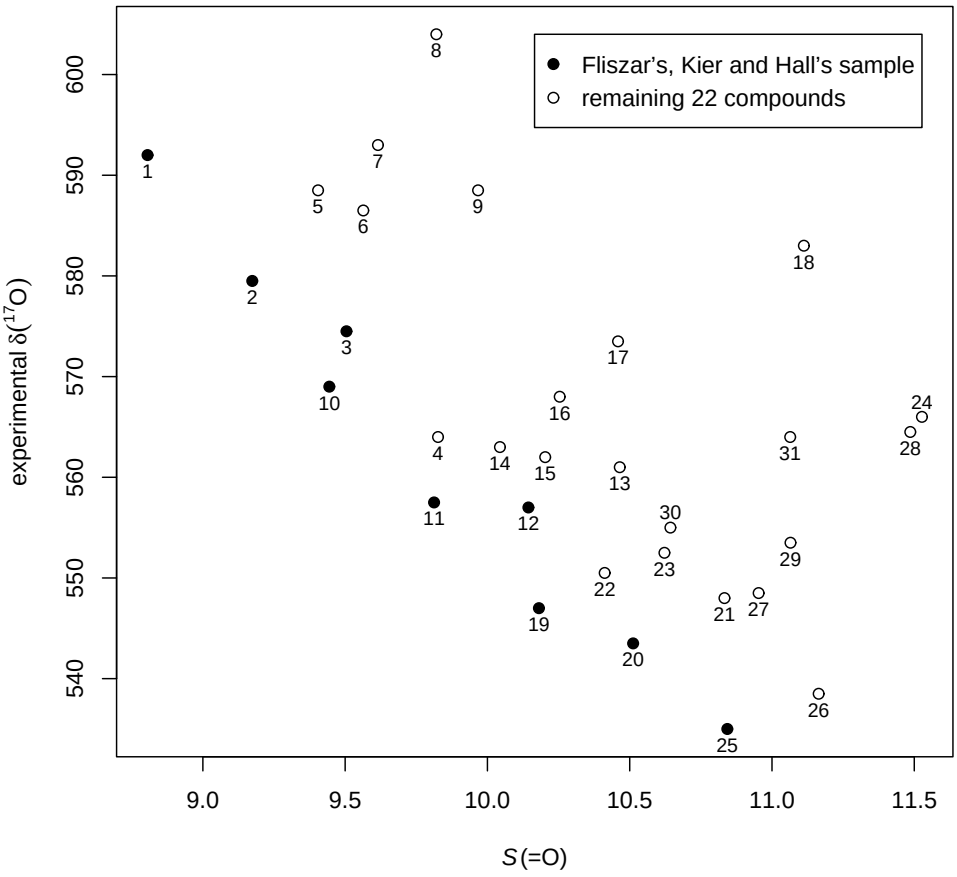

FIGURE 11. $\delta(^{17}\mathrm{O})$ of 31 acyclic saturated aldehydes/ketones vs $S(=\mathrm{O})$. Compounds: 1 *ethanal*, 2 *propanal*, 3 *2–methylpropanal*, 4 2,2–dimethylpropanal, 5 butanal, 6 pentanal, 7 3–methylbutanal, 8 3,3–dimethylbutanal, 9 2–methylpentanal, 10 *propanone*, 11 *butanone*, 12 *3–methylbutanone*, 13 2,2–dimethylbutanone, 14 pentan–2–one, 15 hexan–2–one, 16 4–methylpentan–2–one, 17 4,4–dimethylpentan–2–one, 18 3,3,4,4–tetramethylpentan–2–one, 19 *pentan–3–one*, 20 *2–methylpentan–3–one*, 21 2,2–dimethylpentan–3–one, 22 hexan–3–one, 23 5–methylpentan–3–one, 24 4,4–diethylhexan–3–one, 25 *2,4–dimethylpentan–3–one*, 26 2,2,4–trimethylpentan–3–one, 27 2,5–dimethylhexan–3–one, 28 2,2,4,4–tetramethylpentan–3–one, 29 2,2–dimethylhexan–3–one, 30 heptan–4–one, 31 2,6–dimethylheptan–3–one. Compounds in *italics* are contained in the 9–compound sample.

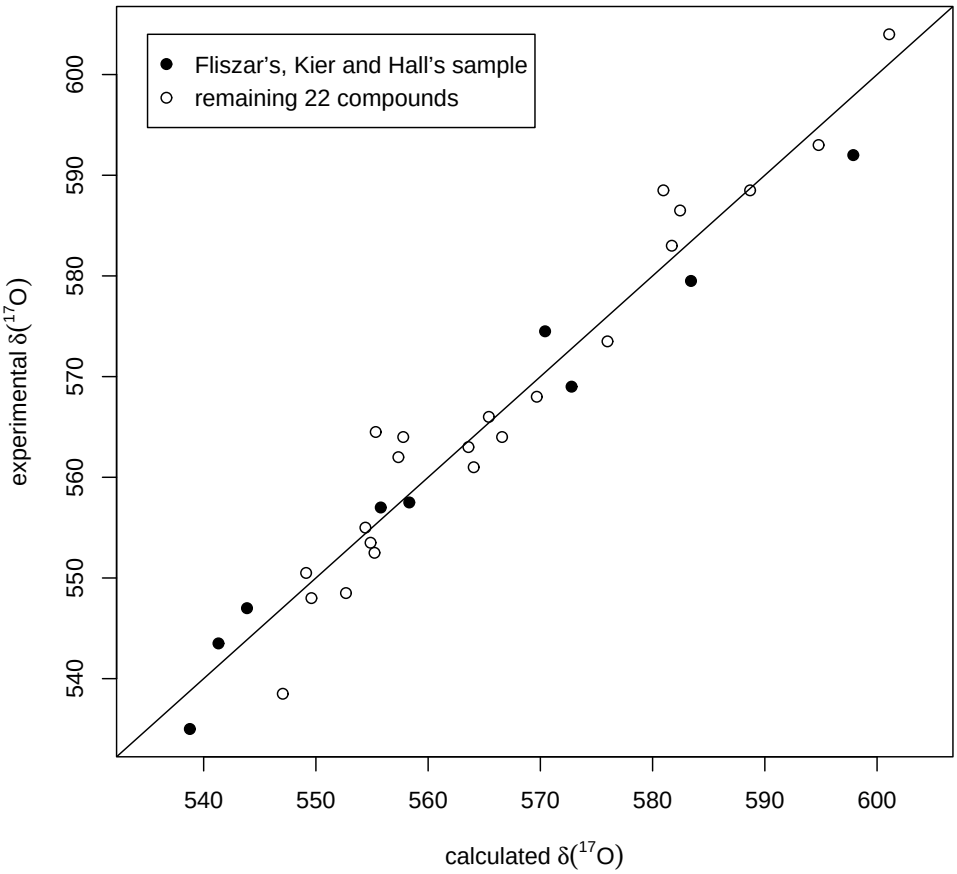

FIGURE 12. $\delta(^{17}\mathrm{O})$ of 31 acyclic saturated aldehydes/ketones vs 3–descriptor model

3.2. **Carbonyl Compounds.** Here things are similar to the ether case: For nine saturated acyclic aldehydes/ketones, in the original paper [27] a correlation between the ^{17}O chemical shifts (in ppm, from another paper by Delseth and Kintzinger [28]) and the E–state of the oxygen atom was found:

$$\delta(^{17}\mathrm{O}) = -27.768 S(=\mathrm{O}) + 834.48,$$
$$r = 0.986,\ s = 3.7,\ n = 9.$$

MOLGEN–QSPR reproduced this result, as follows:

$$\delta(^{17}\mathrm{O}) = -27.772 S(=\mathrm{O}) + 834.51,$$
$$r^2 = 0.96575,\ s = 3.6697,\ n = 9.$$

However, there are experimental ^{17}O chemical shift values provided in reference [27] for no fewer than (by coincidence again) 31 saturated acyclic aldehydes and ketones. For the complete sample the above equation or the corresponding best equation based on S(=O) gives very poor results. Figure 11 is a scatterplot of $\delta(^{17}\mathrm{O})$ vs $S(=\mathrm{O})$ for

all 31 aldehydes/ketones, the 9 compounds contained in the smaller sample are represented by filled circles.

For the 31 aldehydes/ketones, the best single–, 2–, and 3–descriptor models now found are:

1 descriptor: $S(=O)$, $r^2 = 0.3332$, $s = 14.4$,
2 descriptors: $S(=O)$, H, $r^2 = 0.8817$, $s = 6.2$,
3 descriptors: $S(=O)$, $n(C)$, $mwc^{(8)}$, $r^2 = 0.9109$, $s = 5.5$.

where H is the Harary index [20], $n(C)$ is the number of carbon atoms in the molecule, and $mwc^{(8)}$ is the molecular walk count of length 8 [29, 30].

These models are best as long as the descriptor pool to choose from contains arithmetic, topological and geometrical descriptors only. When substructure counts are additionally included in the descriptor pool, better models are obtained. The best models for the 31 aldehydes/ketones are:

1 descriptor: n(C-C=O), $r^2 = 0.5680$, $s = 11.6$,
2 descriptors: $S(=O)$, H, $r^2 = 0.8817$, $s = 6.2$,
3 descriptors: $S(=O)$, n(C-C-C-C=O), n(C-C(-C)-C(=O)-C),
$r^2 = 0.9475$, $s = 4.2$.

Figure 12 is a plot of experimental vs calculated values for $\delta(^{17}O)$ by the latter model.

Again the reason for Kier and Hall to treat 9 compounds only was the opportunity to compare their model [25] to Fliszar's [26], in which the ^{17}O chemical shifts were correlated with the net atomic charges of the carbonyl O atoms from STO–3G calculations for exactly these rather than for all 31 aldehydes/ketones with known ^{17}O NMR chemical shifts.

There are at least three lessons to be recalled here.

(1) The range of validity of a QSPR equation should not be exaggerated ("carbonyl compounds" instead of "otherwise unsubstituted saturated acyclic aldehydes and ketones").
(2) It is a fatal error to use, in a statistical consideration, a selection of cases instead of all available cases.
(3) It is always advisable both for those producing and for those using a QSPR equation to critically examine the primary literature.

References

[1] R. Gugisch, A. Kerber, R. Laue, M. Meringer, and J. Weidinger. *MOLGEN-COMB, a Software Package for Combinatorial Chemistry.* MATCH — Commun. Math. Comput. Chem., 41:189–203, 2000.

[2] A. R. Katritzky, V. S. Lobanov, and M. Karelson. *CODESSA: Reference Manual, Version 2.* University of Florida, 1994.

[3] N. L. Allinger. *MM2. A Hydrocarbon Force Field Utilizing V_1 and V_2 Torsional Terms*. J. Am. Chem. Soc., 99:8127–8134, 1977.

[4] L. B. Kier and L. H. Hall. *Molecular Structure Description. The Electrotopological State*. Academic Press, San Diego and London, 1999.

[5] B. Ren. *Atomic-Level-Based AI Topological Descriptors for Structure–Property Correlations*. J. Chem. Inf. Comput. Sci., 43:161–169, 2003.

[6] D. Bonchev and N. Trinajstić. *Overall Molecular Descriptors. 3. Overall Zagreb Indices*. SAR QSAR Environ. Res., 12:213–236, 2001.

[7] D. Bonchev. *The Overall Wiener Index — A New Tool for Characterization of Molecular Topology*. J. Chem. Inf. Comput. Sci., 41:582–592, 2001.

[8] S. A. Wildman and G. M. Crippen. *Prediction of Physicochemical Parameters by Atomic Contributions*. J. Chem. Inf. Comput. Sci., 39:868–873, 1999.

[9] C. Rücker, J. Braun, A. Kerber, and R. Laue. *The Molecular Descriptors Computed with MOLGEN*. http://www.mathe2.uni-bayreuth.de/molgenqspr, 2003.

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer–Verlag, New York, Berlin, Heidelberg, 2002.

[11] R. Ihaka and R. Gentleman. *R: A Language for Data Analysis and Graphics*. J. Comput. Graph. Stat., 5:299–314, 1996.

[12] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, California, 1984.

[13] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

[14] J. Zupan and J. Gasteiger. *Neural Networks for Chemists*. VCH Verlagsgesellschaft, Weinheim, New York, Basel, Cambridge, Tokyo, 1993.

[15] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer–Verlag, New York, Berlin, Heidelberg, 1995.

[16] H. Wiener. *Structural Determination of Paraffin Boiling Points*. J. Am. Chem. Soc., 69:17–20, 1947.

[17] G. Rücker and C. Rücker. *On Topological Indices, Boiling Points, and Cycloalkanes*. J. Chem. Inf. Comput. Sci., 39:788–802, 1999.

[18] A. T. Balaban. *Topological Indices Based on Topological Distances in Molecular Graphs*. Pure Appl. Chem., 55:199–206, 1983.

[19] H. Hosoya. *Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons*. Bull. Chem. Soc. Jpn., 44:2332–2339, 1971.

[20] B. Lučić, A. Miličević, S. Nikolić, and N. Trinajstić. *Harary Index — Twelve Years Later*. Croat. Chem. Acta, 75:847–868, 2002.

[21] J.B. Greenshields and F.D. Rossini. *Molecular Structure and Properties of Hydrocarbons and Related Compounds*. J. Phys. Chem., 62:271–280, 1958.

[22] Q. Dong, X. Yan, R.C. Wilhoit, X. Hong, R.D. Chirico, V.V. Diky, and M. Frenkel. *Data Quality Assurance for Thermophysical Property Databases–Applications to the TRC SOURCE Data System*. J. Chem. Inf. Comput. Sci., 42:473–480, 2002.

[23] L.B. Kier and L.H. Hall. *An Electrotopological–State Index for Atoms in Molecules*. Pharm. Res., 7:801–807, 1990.

[24] C. Delseth and J. P. Kintzinger. *Carbon-13 and Oxygen-17 Nuclear Magnetic Resonance of Aliphatic Ethers. γ–Effects between Oxygen and Carbon Atoms*. Helv. Chim. Acta, 61:1327–1334, 1978.

[25] L.H. Hall, B. Mohney, and L.B. Kier. *The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs*. J. Chem. Inf. Comput. Sci., 31:76–82, 1991.

[26] S. Fliszar. *Charge Distribution and Chemical Effects*. Springer–Verlag, New York, 1983.

[27] L.H. Hall, B. Mohney, and L.B. Kier. *The Electrotopological State: An Atom Index for QSAR*. Quant. Struct. Act. Relat., 10:43–51, 1991.

[28] C. Delseth and J.-P. Kintzinger. *Oxygen-17 Nuclear Magnetic Resonance. Aliphatic Aldehydes and Ketones: Additivity of Substituent Effects and Correlation with Carbon-13 NMR*. Helv. Chim. Acta, 59:466–475, 1976. Erratum: page 1410.

[29] C. Rücker and G. Rücker. *Counts of All Walks as Atomic and Molecular Descriptors*. J. Chem. Inf. Comput. Sci., 33:683–695, 1993.

[30] I. Gutman, C. Rücker, and G. Rücker. *On Walks in Molecular Graphs*. J. Chem. Inf. Comput. Sci., 41:739–745, 2001.