Liquid Density and Critical Properties of Hydrocarbons Estimated from Molecular Structure

ARTICLE *in* JOURNAL OF CHEMICAL & ENGINEERING DATA · APRIL 2002

Impact Factor: 2.04 · DOI: 10.1021/je010308l

William A. Wakeham
Imperial College London
363 PUBLICATIONS 7,509 CITATIONS
SEE PROFILE

Georgi St. Cholakov
University of Chemical Technology and …
54 PUBLICATIONS 366 CITATIONS
SEE PROFILE

ROUMIANA PETROVA STATEVA
Bulgarian Academy of Sciences
93 PUBLICATIONS 911 CITATIONS
SEE PROFILE

ELSEVIER

Fluid Phase Equilibria 163 (1999) 21–42

FLUID PHASE EQUILIBRIA

www.elsevier.nl/locate/fluid

Estimation of normal boiling points of hydrocarbons from descriptors of molecular structure

Georgi St. Cholakov [a], William A. Wakeham [b,*], Roumiana P. Stateva [c]

[a] Department of Petroleum and Solid Fuels Processing Technology, University of Chemical Technology and Metallurgy, Sofia 1156, Bulgaria

[b] Department of Chemical Engineering, Imperial College of Science, Technology and Medicine, London SW7 2BY, UK

[c] Institute of Chemical Engineering, Bulgarian Academy of Sciences, Sofia 1113, Bulgaria

Received 12 January 1999; accepted 19 April 1999

Abstract

Correlations for estimation of thermophysical properties are needed for the design of processes and equipment related to phase equilibria. The normal boiling point (NBP) is a fundamental characteristic of chemical compounds, involved in many correlations used to estimate important properties. Modern simulation packages usually require the NBP and a standard liquid density from which they can estimate all other necessary properties and begin the design of particular processes, installations and flowsheets. The present work contributes a correlation between the molecular structure and the normal boiling point of hydrocarbons. Its main features are the relative simplicity, sound predictions, and applicability to diversified industrially important structures, whose boiling points and numbers of carbon atoms span a wide range. An achievement of particular interest is the opportunity revealed, for reducing the number of the compounds required for the derivation (the learning set), through multivariate analysis and molecular design. The high accuracy achieved by the correlation opens up a possibility for systematic studies of chemical engineering applications in which the effects of small changes are important. This also defines a path towards the more general problem of the influence of uncertainties in calculated thermophysical parameters on the final outcome of computer aided simulation and design. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Molecular simulation; Model; Normal boiling point; Hydrocarbons

1. Introduction

Correlations for estimation of thermophysical properties are an important tool for design of processes and equipment, environmental impact assessment, HAZOP studies, and other important

[*] Corresponding author. Tel.: +44-171-594-5005; fax: +44-171-594-8802; e-mail: w.wakeham@ic.ac.uk

PII: S0378-3812(99)00207-1

chemical engineering problems related to phase equilibria. Consequently, large commercial databases of miscellaneous properties are compiled, but have to be populated with new compounds within the limits of interpolation of the available experimental data [1]. On the other hand, methods for extrapolation of existing data are needed for assessment of compounds not yet synthesized and/or high molecular compounds for which the experimental determination is unreliable or impossible because of degradation [2].

The physical properties of chemical compounds are described by a large group of structure related characteristics, such as normal boiling point (NBP) and critical parameters. Most of these have been targeted by different correlations and approaches [3–6]. However, thermophysical properties are interrelated and an efficient strategy is to identify a suitable number of independently determined primary target parameters, which are connected to the largest possible number of properties and can be used for their computational estimation [2].

The normal boiling point (NBP) is a fundamental characteristic of chemical compounds. It is involved in many correlations used to estimate thermophysical properties. Modern computer simulation packages usually require the NBP and a standard liquid density from which they can estimate all necessary properties and begin the design of particular processes, installations and flowsheets for their realization.

The analysis of prior work, recently reviewed by Katritzky et al. [7], shows that historically two types of empirical correlations have been developed — correlations, aimed at molecules with the widest possible variation of functional groups and heteroatoms, and — correlations concentrating on molecules within homologous series. The former follow the success of the first group contribution methods [4], and the most recent ones apply electronic and graph topological descriptors [1,2]. A common feature of these correlations is that the dependent variable is a function of estimated contributions of diversified structural features, even when only one complex descriptor is incorporated in the final model [7]. They will be further referred to here as ‘‘contribution’’ models.

Correlations developed for homologous series usually employ the total number of C atoms or the molecular mass of the compounds with adjustable constants [5,8]. Gasem et al. [9] recently suggested the abbreviation ABC — Asymptotic Behavior Correlations for such models. Marano and Holder [10] proposed a generalization for all ABCs and developed such correlations for a wide number of thermophysical properties of several homologous series [11]. It has been shown also that ABCs can be developed with graph topological indices [12], and molecular energy descriptors [6]. Theoretical explanations have been suggested to relate quantum chemical descriptors to the thermodynamic properties of polar molecules [2,13]. The lattice fluid model [14–16] and the cell model [17] have been used to explain ABCs [11]. A common feature of ABCs is that the dependent variable is a non-linear function with several adjustable constants describing the relations between repeated segments of the molecules (mers) and empty ‘‘holes’’ (lattice-fluid models) or mers and free volume (cell models). They will be further referred to as ‘‘mers’’ models.

The advantages and disadvantages of the two approaches have been well documented by the respective authors. From a practical point of view, there is clearly a need for a compromise between the high accuracy but limited functionality of the ‘‘mers’’ models, and the low accuracy and widely varied functionality of the ‘‘contribution’’ models. The present work is an attempt to find such compromise. Furthermore, it is devoted to the investigation of the correlation power of molecular descriptors estimated with conventional programmes for computer simulation of molecular mechanics. These are considered as a potential tool for enhancing the capabilities of the simulation packages widely used nowadays for computer aided chemical engineering design. A third objective of the present work is to explore opportunities to reduce the size of the data set upon which the derivation of the correlation is based (the learning set), since databases employed in contribution models are becoming increasingly larger. Finally, an object of the study was the evaluation of the extrapolation predictive power of such correlations for outlying molecules of industrial importance.

2. Methodology

The development of any correlation relies on a *database* including the objects of interest (molecular structures in the present context), and relevant known properties of these objects (*descriptors* of the molecular structures). Independent variables defined from the database have to be correlated to a set of dependent characteristics of functional interest (NBPs) with the help of a suitable *modelling technique*. The predictive power of the correlations usually is confined to the space defined by the constraints of its derivation, although in the specific case of molecular modelling some extrapolation to structurally related outlying molecules might be possible at the cost of higher error.

Experimental values for low and moderate NBPs of industrially important compounds are usually available from many sources. Higher boiling points are determined in vacuum, and may be recalculated for normal conditions if a pressure–temperature relation suitable for the particular group of compounds is available. For many compounds, however, the latter relations have not been studied, and the amount of experimental data even at reduced pressure is limited.

2.1. Database

The design of the database of relevant compounds is perhaps the most important step in the derivation of statistical correlations. The weighting of different groups presented in the database directly influences the subsequent modelling [18]. The database should contain all relevant structural features of the modelled groups of compounds, but it should be emphasized again that the relative representation of those groups influences the uniformity of the prediction for the different groups of objects.

Several features were sought from the database used in the present study, in order to achieve representation of the main structures, and the possibility for extrapolation of the predictions towards the three industrially important high molecular hydrocarbons with unknown NBPs — lycopene, β-carotene and 1,2-benzo [*a*] pyrene, chosen as an example. These are:

- systematic change of properties within several homologous series, since any compound may be viewed as a member of some appropriate series;
- presence of series of branched hydrocarbons with increasing numbers of double bonds, cycloalkanes and terpenoids with known NBPs, which might be extrapolated towards high molecular terpenoids;
- presence of series of hydrocarbons differing by one aromatic ring, which might be extrapolated towards benzopyrenes;
- presence of a control set of compounds with complex molecular structure, estimated by other authors, to be used for comparison with the present study.

Table 1
Hydrocarbons included in the database

No.	Name	No.	Name	No.	Name
1[b]	ethane	48	3,3-dimethylpentane	96	1-dodecene
2	propane	50	2,2,3-trimethylbutane	97	1-tridecene
3	*n*-butane	51	2-methylheptane	98	1-tetradecene
4	*n*-pentane	52	3-methylheptane	99	1-pentadecene
5	*n*-hexane	53	4-methylheptane	100	1-hexadecene
6	*n*-heptane	54	2,2-dimethylhexane	101	1-heptadecene
7	*n*-octane	55	2,3-dimethylhexane	102	1-octadecene
8	*n*-nonane	56	2,4-dimethylhexane	103	1-nonadecene
9	*n*-decane	57	2,5-dimethylhexane	104	1-eicosene
10	*n*-undecane	58	3,3-dimethylhexane	105	1-heneicosene
11	*n*-dodecane	59	3,4-dimethylhexane	106	1-docosene
12	*n*-tridecane	60	3-ethylhexane	107	1-tricosene
13[a]	*n*-tetradecane	61	2,2,3-trimethylpentane	108	1-tetracosene
14	*n*-pentadecane	62	2,2,4-trimethylpentane	109	1-pentacosene
15	*n*-hexadecane	63[a]	2,3,3-trimethylpentane	110	1-hexacosene
16	*n*-heptadecane	64	2,3,4-trimethylpentane	111	1-heptacosene
17	*n*-octadecane	65	2-methyl-3-ethylpentane	112	1-octacosene
18	*n*-nonadecane	66	3-methyl-3-ethylpentane	113	1-nonacosene
19	*n*-eicosane	67	2,2,3-trimethylhexane	114[b]	1-triacontene
20	*n*-heneicosane	68	2,2,4-trimethylhexane	115	1,3-butadiene
21	*n*-docosane	69	2,2,5-trimethylhexane	116	*c*-2-butene
22	*n*-tricosane	70[b]	3,3-diethylpentane	117	*t*-2-butene
23	*n*-tetracosane	71	2,2,3,3-tetramethylpentane	118	*i*-butene
24	*n*-pentacosane	72	2,2,3,4-tetramethylpentane	119	isoprene
25	*n*-hexacosane	73[a]	2,2,4,4-tetramethylpentane	120	2,3-dimethyl-1-butene
26	*n*-heptacosane	74[b]	2,3,3,4-tetramethylpentane	121	2,3-dimethyl-2-butene
27	*n*-octacosane	75	2-methyloctane	122	2-ethyl-1-butene
28	*n*-nonacosane	76	2-methylnonane	123	*c*-2-hexene
29	*n*-triacontane	77	3,3,5-trimethylheptane	124	*t*-2-hexene
30	*n*-dotriacontane	78	2,2,3,3-tetramethylhexane	125	2-methyl-1-pentene
31	*n*-pentatriacontane	79	2,5-dimethyldecane	126	4-methyl-1-pentene
32	*n*-hexatriacontane	80	2,5,-dimethyldodecane	127	2,4,4-trimethyl-1-pentene
33	*n*-tetracontane	81	2,6,10-trimethyldodecane	128	2,4,4-trimethyl-2-pentene
34	*n*-tetratetracontane	82	2,6,10-trimethyltetradecane	129	2-methyl-1-butene
35[b]	*n*-hexacontane	83[a]	pristane	130	2-methyl-2-butene
36	*i*-butane	84[a]	phytane	131	3-methyl-1-butene
37	2-methylbutane	85[a]	squalane	132	2,3,-dimethyl-butadiene
38[b]	2,2-dimethylpropane	86[b]	lycopane	133	3,3-dimethyl-1-butene
39	2-methylpentane	87[b]	propylene	134	2-methyl-2-pentene
40[a]	3-methylpentane	88[a]	1-butene	135	3-methyl-1-pentene
41	2,2-dimethylbutane	89	1-pentene	136	1,5-hexadiene
42	2,3-dimethylbutane	90	1-hexene	137	limonene
43	2-methylhexane	91	1-heptene	138[b]	α-pinene
44	3-methylhexane	92	1-octene	140	lycopene
45	2,2-dimethylpentane	93	1-nonene	141	β-carotene
46	2,3-dimethylpentane	94	1-decene	142	cyclopropane
47	2,4-dimethylpentane	95	1-undecene	143[b]	cyclobutane

Table 1 (continued)

No.	Name	No.	Name	No.	Name
144[a]	cyclopentane	183	cyclopentene	222[b]	mesitylene
145	cyclohexane	184	cyclohexene	223	1,2,3-trimethylbenzene
146	cycloheptane	185	1,3-cyclohexadiene	224	1,2,4-trimethylbenzene
147	cyclooctane	186	5-methyl-1,3-cyclopentadiene	225	1,2,3,4-tetrahydronaphtalene
148	methylcyclohexane	187	1,3-cyclopentadiene	226	*t*-butylbenzene
149	ethylcyclohexane	188[b]	benzene	227	*p*-cymene
150	propylcyclohexane	189	toluene	228	*m*-diethylbenzene
151	butylcyclohexane	190	ethylbenzene	230	*i*-butylbenzene
152	methylcyclopentane	191	propylbenzene	231	*m*-diisopropylbenzene
153	ethylcyclopentane	192	butylbenzene	232	diphenylmethane
154	propylcyclopentane	193	*o*-xylene	233	*m*-ethyltoluene
155	butylcyclopentane	194	*m*-xylene	234	*s*-butylbenzene
156	pentylcyclopentane	195	*p*-xylene	235	*p*-diethylbenzene
157	hexylcyclopentane	196	1-methyl-3-ethylbenzene	236	*p*-diisopropylbenzene
158	heptylcyclopentane	197	pentylbenzene	237	diphenyl
159	octylcyclopentane	198	hexylbenzene	238	1,1-diphenylethane
160	nonylcyclopentane	199	heptylbenzene	239[b]	1,2-diphenylethane
161	decylcyclopentane	200[a]	octylbenzene	240[b]	naphtalene
162	undecylcyclopentane	201	nonylbenzene	241	antracene
163	dodecylcyclopentane	202	decylbenzene	242	phenanthrene
164	tridecylcyclopentane	203	undecylbenzene	243[b]	*m*-terphenyl
165	tetradecylcyclopentane	204	dodecylbenzene	244	*p*-terphenyl
166	pentadecylcyclopentane	205[b]	tridecylbenzene	245	1,2-benzo [*a*] pyrene
167	hexadecylcyclopentane	206	tetradecylbenzene	246	pyrene
168[b]	heptadecylcyclopentane	207	pentadecylbenzene	247	chrysene
169	octadecylcyclopentane	208	hexadecylbenzene	248[a]	*o*-terphenyl
170	nonadecylcyclopentane	209	heptadecylbenzene	249[a]	triphenylmethane
171	eicosylcyclopentane	210	octadecylbenzene	250[a]	acenaphtylene
172	heneicosylcyclopentane	211	nonadecylbenzene	251	acenaphtene
173	docosylcyclopentane	212	eicosylbenzene	252[a]	1,1,2,2-tetraphenylethane
174	tricosylcyclopentane	213	heneicosylbenzene	253[a]	4-methyloctane
175	tetracosylcyclopentane	214	docosylbenzene	254[a]	2,2,3,3-tetramethylbutane
176[b]	pentacosylcyclopentane	215	tricosylbenzene	255[a]	2-ethylhexene
177	*c*-1,2-dimethylcyclohexane	216[b]	tetracosylbenzene	256[a]	adamantane
178	*t*-1,2-dimethylcyclohexane	217	styrene	257[a]	1,5-cyclooctadiene
179	*c*-1,3-dimethylcyclohexane	218	α-methylbenzene	258[a]	2,5-methyl-1,5-hexadiene
180	*t*-1,3-dimethylcyclohexane	219	cumene	259[a]	*c*-1-propenylbenzene
181	*c*-1,4-dimethylcyclohexane	220	*o*-ethyltoluene	260[a]	1-phenylnaphtalene
182	*t*-1,4-dimethylcyclohexane	221	*p*-ethyltoluene	261[a]	indane

[a] Members of the control set.

[b] Members of the designed learning set of 20 hydrocarbons.

The names of the compounds selected for the database used in this study are presented in Table 1, with their published NBPs listed in Table 6. The objects have been limited only to hydrocarbons in order to achieve a reasonable presentation of the functional groups of these fundamental compounds. The homologous series included allow a ‘‘mers’’ influence also to be expressed in the modelling.

This approach follows from one of the objectives of the present work — to find a compromise between the breadth of the functionality of molecular structures and the precision achieved in their estimation. It has been also justified by recent prior work [7,19].

The present database of 261 hydrocarbons was compiled from several sources [4,19–24]. The data for the normal alkanes with more than 30 carbon atoms were calculated by a ‘‘mers’’ correlation [25]. Three hydrocarbons with unknown NBPs were included in the database as an illustration of the case when objects of industrial importance have to be evaluated as outliers. Such molecules are often referred to as ‘‘hypotheticals’’. Lycopene and β-carotene are industrially important constituents of natural products, 1,2-benzo [*a*] pyrene is a carcinogenic hydrocarbon often used as reference in ecological studies. Most of the hydrocarbons are identical with those used in the most recent correlation for description of NBPs of hydrocarbons [19]. The values for some of the hydrocarbons, mainly in the control set were recalculated for normal conditions from vacuum data, which were considered more reasonable.

The limits for the main hydrocarbon series, and structures, included in the database, which determine also the boundaries for the predictive ability of the derived models may be assessed from Tables 1 and 6, but are more clearly outlined by the total carbon atoms dependence of the predicted points (Fig. 2), and the scatterplot of the first two principle components (Fig. 3). NBPs are varied in the widest practical range from 184.5 to 877.5 K. The total number of carbon atoms spans from 2 to 60 for the *n*-alkanes, from 3 to 40 — for the series finishing with β-carotene, and — to 30 for the rest of the homologous series.

2.2. Descriptors

Two types of descriptors were employed in the present investigation.

Molecular energy descriptors were evaluated with a conventional computer programme for molecular mechanics simulation, based on the MMX modification of the MM2 method [26]. In such programmes a structure is considered a collection of atoms held together by elastic (harmonic) forces-bonds, which constitute the force field. The calculations start with a structure with relevant default values of parameters and its optimized geometry is found by iterational minimization of its total steric energy. Further refinement of the energy contributions may be achieved by assigning more accurate values for the starting force constants and/or applying several programmes with different sophistication for gradual assessment of the more intimate structural elements or specific programmes, designed to target particular structural features [1,2,6]. Such refinement of the molecular energy descriptors used in the present study has been deliberately avoided.

For the practical purposes of the present study, the minimised molecular energy models of all 261 molecules were obtained with a conventional programme for molecular mechanics simulation, and the contributions of different energies in the minimized models were tested as descriptors. An illustration of the molecular energy descriptors for adamantane is presented in Fig. 1. The names and codes of the descriptors are given in Table 2.

Carbon atom descriptors of various levels of sophistication can be used. The highest level of sophistication presently available comprises the graph topological indices, derived from the adjacency and distance matrices of a chemical structure [12]. More than 120 such indices have been suggested. The latest versions can evaluate 3D structural information [27], and many of them have been involved in correlations with thermophysical properties and characteristics [12].

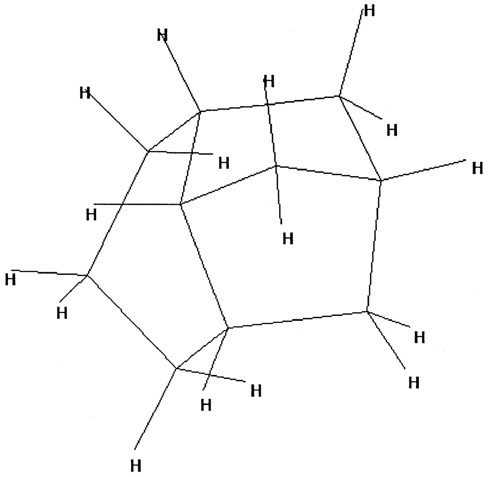

Fig. 1. The minimized energy model and the 15 molecular descriptors of adamantane. (Dimensions are given as estimated by the molecular mechanics simulation programme — in kcal mol^{-1}, A^3 mol^{-1}, etc.). Total energy (E_{tot}): 34.616; Stretch energy (E_{str}): 1.188; Bond energy (E_{bnd}): 10.273; Stretch–bend energy (E_{s-b}): -0.292; Torsion energy (E_{tor}): 16.524; Van der Waals energy (E_{vdw}): 6.924; Dipole-charge interaction energy (E_{dch}): 0.000; Electric dipole moment (DM): 0.000; Standard enthalpy (H_f): -14.00; Strain energy (E_{ste}): 27.30; Van der Waals volume (V_{vdw}): 254 A^3, Molar volume (V_M): 152 cm^3; Total van der Waals surface (S_{tot}): 174.82 A^2, Saturated van der Waals surface (S_{sat}):174.82 A^2, Unsaturated van der Waals surface: 0.00 A^2 (S_{unsat}).

The lowest level of sophistication of carbon atom descriptors is to use the numbers of atoms engaged in specific groups (atom counts). These are generically related to the group contributions, which multiply the particular number of atoms by empirically assigned constants. In the present

Table 2
Descriptors from simulated molecular mechanics

No.	Description	Code
1	Total energy	E_{tot}
2	Stretch energy	E_{str}
3	Bond energy	E_{bnd}
4	Stretch–bend energy	E_{s-b}
5	Torsion energy	E_{tor}
6	Van der Waals energy	E_{vdw}
7	Energy of ‘‘dipole-charge’’ interaction	$E_{d\text{-}ch}$
8	Electrostatic dipole moment	DM
9	Standard heat of formation	H_f
10	Strain energy	E_{st}
11	Van der Waals volume	V_{vdw}
12	Molar volume	V_M
13	Total Van der Waals surface	S_t
14	Saturated Van der Waals surface	S_{sat}
15	Unsaturated Van der Waals surface	S_{unsat}

Table 3
Carbon atom descriptors and molecular mass

No.	Name	Code
16	Total number of C atoms	C_{tot}
17	Number of C atoms in CH_3 groups	N_{CH_3}
18	Number of C atoms in aliphatic CH_2 groups	$N^{a}_{CH_2}$
19	Number of C atoms in aliphatic CH groups	N^{a}_{CH}
20	Number of C atoms in aliphatic C groups	N^{a}_{C}
21	Number of C atoms in aliphatic $CH_2{=}CH_2$ groups	$N^{a}_{DCH_2}$
22	Number of C atoms in aliphatic CH=CH groups	N^{a}_{DCH}
23	Number of C atoms in aliphatic C= groups	N^{a}_{DC}
24	Total number of C atoms in aliphatic double bonds	DBA
25	Number of C atoms in cyclic CH_2 groups	$N^{c}_{CH_2}$
26	Number of C atoms in cyclic CH groups	N^{c}_{CH}
27	Number of C atoms in cyclic C groups	N^{c}_{C}
28	Number of C atoms in cyclic CH= groups	N^{c}_{DCH}
29	Number of C atoms of cyclic C= groups	N^{c}_{DC}
30	Total number of C atoms in cyclic double bonds	DBC
31	Molecular mass	M

investigation we have chosen the carbon atom descriptors, presented in Table 3. For the most part they coincide with the descriptors in the recent Joback group contribution model [4].

Because of the success of prior work with topological indices, a limited number of them is also tested in the present work. This number includes: the Wiener Index, (W); the Balaban index, the Bonchev and Trinajstic information content and mean information content of the unit distances, information content and mean information content of the distances' distribution indices, (known respectively as IWD, IWDM, IED and IEDM); the cyclomatic number, (μ) and the Randic path connectivity indices (CHI) up to third order terms. The meaning and methods of calculation of these indices have been extensively reviewed elsewhere [12,28]. An additional descriptor — the gravitation index, (G_I), successfully employed lately for the evaluation of NBPs [7,29] was also tested. Where applicable, the descriptors were calculated not only with unit distances, but also with distances between bonded atoms, obtained from the structures minimized by molecular mechanics simulation.

The total number of descriptors, including the topological ones, as well as molecular mass, amounted to 59. This is a relatively low number as compared to the most recent description of NBPs of databases including heterocompounds [7], in which more than 800 descriptors are tested, or the study of Wessel and Jurs [19], confined only to hydrocarbons selecting among 81 descriptors. It should be emphasized though, that one of the objectives of the present investigation was to keep the methods for calculation and the meaning of descriptors as simple as possible. That is why electronic descriptors and complex functions of one or more descriptors were deliberately avoided.

2.3. Modelling

A conventional ‘‘stepwise’’ multiple regression procedure [30] was employed to select the most influential variables from the 59 descriptors and determine their optimal number. This procedure is a subjective molecular feature selection [19], in which the dependent variable is used to develop models in the form:

$$\mathrm{NBP}_j = b_o + \sum_{i=1}^{n} (b_i X_{ij}) \tag{1}$$

where NBP_j is the normal boiling point of the compound j, b_o is the intercept term, and b_i is the coefficient for descriptor X_{ij}.

A linear contribution of the structural descriptors was adopted for all variables, except for a nonlinear ''mers''-type independent variable (total number of carbon atoms, C_{tot}). The latter is successfully used in ABC correlations for homologous series [11]. We assume that each molecule may be considered a member of some homologous series. The boiling points of the molecules would then lie on a family of curves, different for each series, but asymptotically dependent on the ''mers'' variable, C_{tot}. The distances between the curves in the family would be then accounted for by the linearly contributing independent variables, which would reflect the specific features of the particular

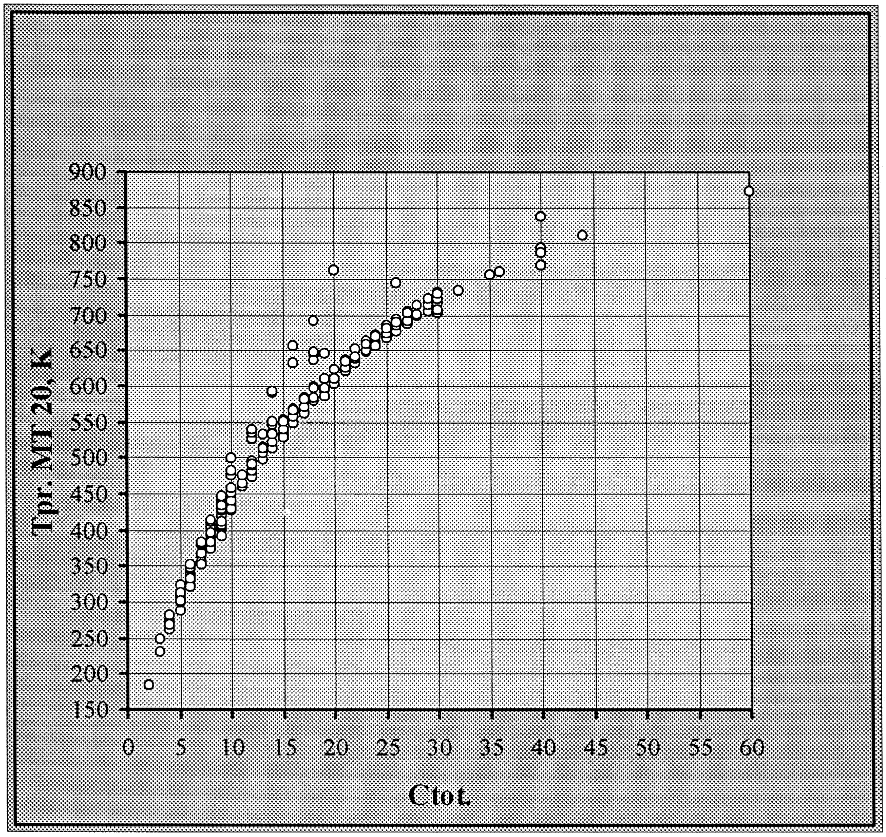

Fig. 2. Asymptotic dependence of the normal boiling points predicted by the M-20 model (T_{pred} MT 20, K) on the total number of C atoms (C_{tot}).

mers of molecules belonging to different series. These assumptions are illustrated in Fig. 2, which presents the boiling points, predicted by one of our models, as a function of the total number of C atoms. The experimental temperatures obey the same dependence. In a later section of the paper we shall show that this concept for the structure of the model is successful.

The algorithm, used also in this paper, for obtaining the best models with an increasing number of independent variables has been described in detail elsewhere [29].

The targeted representation of the published NBPs was set as a mean standard deviation of relative errors of 2.1%. This target follows from the observation that the experimental uncertainties in the DIPPR database for relevant molecules are around 2.1% [1]. Thus, we use the DIPPR estimated uncertainty as a reasonable figure to aim for in the representation. Ref. [1] suggests also that one descriptor of a pair with a pairwise correlation $\geq \pm 0.95$ should be discarded. Later work has gone much below that limit, but for hydrocarbons especially this does not seem to be practical. That is why, for the present work the limit of pairwise correlations was set at ± 0.85.

As in other similar studies, the compounds in the database were devided into a learning set, and a control set (Table 1). The compounds in the control set were not used in the derivation of the model. They were chosen mainly from the latest and most successful work on NBPs of hydrocarbons [19]. An attempt was made to predict the boiling points of nearly all compounds from [19], which were reported to be difficult for prediction. Compounds with triple bonds, which are not present in our database, were omitted. Our control set includes also three terpenoids, the boiling points for which were among the few obtained only from the work of Bogomolov et al. [23], and could not be compared with other sources.

3. Results and discussion

The model derived from the learning set with 235 hydrocarbons (M-235) is presented in Table 4. It can be seen from this table, and the predictions of M-235, given in Table 6, that the discrepancies

Table 4
Model (M-235), derived from the learning set with 235 molecules. $N = 235$; Standard residual error = 4.95 K. Coefficient of Multiple correlation: 0.999. Calculated Fisher's Criterion: 24279.2

Independent variables	Coefficients	Standard deviation (±)	*F* criterion for removal from model
$C_{tot}^{0.678}$	142.51096	1.22788	9999.99
$N_{CH_2}^{a}$	4.96750	0.19820	628.15
N_{C}^{a}	−8.24519	1.06216	60.26
N_{DCH}^{c}	−3.55218	0.34924	103.45
E_{tot}	1.45273	0.10757	182.38
E_{bnd}	−2.05695	0.19640	109.70
V_{vdw}	−1.04224	0.01463	5075.83
S_{unsat}	−0.37320	0.02613	203.99
Intercept	46.97246	2.20274	–

from the published values fall even below the average relative deviation targeted in the present study. In the context of the uncertainties of the DIPPR database [19], and from the point of view of practical application of this correlation, such precision has no particular merit. However, this success of the description allows us to address the third objective of the study, which is to reduce the number of the compounds in the learning set, because a reasonable loss of precision can be tolerated.

In the first attempt to reduce the number of the compounds in the learning set an approach widely used in similar studies (for instance in [1]) was adopted. The compounds left in the learning set were selected to give a fair representation of the main groups and structures in the database. This approach, however, cannot be based on anything more than a general perception of homologous series. As such it might be inadequate for complex hybrid structures, which cannot be assigned to a particular group. A model was derived from half of the learning set (116 compounds), chosen following the above approach. It showed a 1.02 K higher standard residual error, but the same set of variables was selected. The analysis of this model indicated that the key to successful description is the uniformity of the distribution of the compounds of the learning set over the space occupied by the database, rather than their number. Our next task then was to find a tool for the statistically based selection of a learning set, and to determine the minimum number of its members. This task has been addressed by multivariate analysis of the database and statistical molecular design of the learning set.

3.1. Multivariate analysis and molecular design of the learning set

The input data of the database consist of columns (descriptors) and rows (objects-molecules). With the help of multivariate analysis the information, contained in the database may be connected to the dependent variable (NBP) in two ways — either with the columns as independent variables, or — with the principal components (factors) as independent latent variables [31]. The principal components of the data provide several additional opportunities for data manipulation. When used as latent variables in the modelling, problems with collinearity between the original variables can be solved. Partial least square (PLS) regression may be applied with a smaller number of latent variables. The latter are usually factors with eigenvalues higher than 1, if cross validation shows that they can account for a sufficient portion of the variance in the data. The objects (molecules in our case) may be projected onto the plane formed by the first two most influential principal components, and the points on the resulting scatterplot will represent the molecules of the database. This suits our above defined task, since the placement of the molecules of the learning set can be chosen directly from this scatter plot.

Information about the principles of multivariate analysis and its application in characterization of quantitative structure-properties relationships can be found elsewhere [32,33]. Its application for statistical molecular design is presented in Ref. [34].

Fig. 3 shows the scatterplot of the molecules of the whole database, and the selection of the molecules in the new learning set. It depicts also the molecules in the control set. It may be seen from the figure that the database is not ideally balanced. The points are situated approximately in a triangle. The homologous series are placed on linear loci. The three hypotheticals can be classified as close outliers, situated close to the right hand side of the triangle, which is formed, however, by only two molecules with known temperatures. The learning set, selected from the scatterplot, is distributed to cover the sides and the center of the triangle.

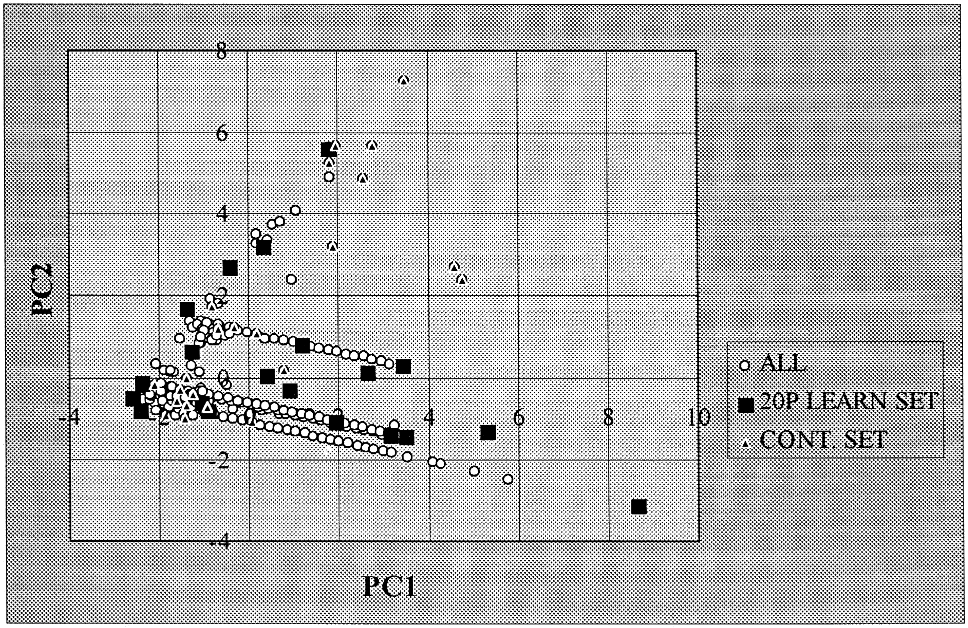

Fig. 3. Selection of the learning set with 20 hydrocarbons (20P LEARN SET) from the scatter plot of the first two principal components of all molecules (ALL). The molecules in the control set (CONT. SET) are also shown. The numbers and the names of the molecules, selected for the learning and the control sets are given in Table 1.

The learning set presented in Fig. 3 consists of 20 molecules. This may be considered the minimum number of members of the set as selected by statistical molecular design.

In the derivation of the M-235 model the descriptors were used as independent variables. The factor analysis of the matrix consisting of the values for the independent variables selected for this model, showed that three factors have eigenvalues higher than 1. Together with a fourth factor with an eigenvalue close to one they account for 95.8% of the variation in the database. If a model with 4 latent variables is constructed through PLS regression, the minimal number of experimental data follows the $2^k = 16$ points requirement of statistical designs (where k is the number of factors), plus at least four additional points [34]. When applied to our present task these rules determine the minimum number of points in the learning set at twenty. As already explained, the points should be evenly distributed to cover the factor space. In molecular design exact distribution according to a statistical protocol can rarely be achieved, because data for particular structures may not be available or may not be accessible experimentally, as also illustrated by Fig. 3.

A model was further derived with the new learning set of 20 molecules. This model (M-20) is presented in Table 5, and its predictions for the original 235 learning set — in Table 6. It is not really a new model, but a variation of the M-235 model. Its independent variables are the same as in M-235, and only its coefficients are derived from the set of 20 molecules. These coefficients will be less sensitive to the distribution of data in the database, than the coefficients of M-235. Its standard residual error is 5.46 K, which compares favourably with the model derived from 116 molecules (standard residual error of 6.07 K).

Table 5
Model derived from the designed learning set with 20 molecules (M-20). $N = 20$; Standard residual error: 5.46 K. Coefficient of Multiple correlation: 0.999. Calculated Fisher's Criterion: 3221.25

Independent Variables	Coefficients	Standard Deviation	F criterion for removal from model
$N_C^{0.675}$	141.66807	2.83713	2493.36
$N_{CH_2}^a$	3.78832	0.44837	71.39
N_C^a	−9.94330	4.29798	5.35
N_{DCH}^c	−3.89729	0.99589	15.31
E_{tot}	1.46310	0.24684	35.13
E_{bnd}	−2.36292	0.32912	51.54
V_{vdw}	−0.92884	0.03353	767.43
S_{unsat}	−0.33671	0.08398	16.08
Intercept	40.41170	4.93330	–

It seems appropriate to repeat again here that M20 is only an improvement of the original M235 model. The success of this attempt, however, illustrates an approach, which may be explored in the future for establishing correlations from a limited amount of experimental data. This opportunity is very important since the databases of the latest similar investigations seem to become larger and larger. Katrizky et al. [7], for instance employ 612 compounds for presenting heterostructures, Wessel and Jurs [19] used 356 compounds in their hydrocarbons only study.

3.2. Prediction of the control set and hypotheticals

The last object of the present study was to evaluate the extrapolation predictive power of the above described models.

The values for the NBPs predicted by M235 and M20 for the control set and the ''hypothetical'' molecules are presented in Table 7. Values for the boiling points of the hypotheticals, estimated by asymptotic behavior correlations [10] are also included. The table shows the published NBPs of all control compounds, and NBPs calculated from published boiling points of the same compounds under reduced pressure. The latter were considered more reliable for two reasons. First, some of the experimental results have been obtained more than 50 years ago with the technique and expertise then available. Secondly, for components whose NBPs are higher than the decomposition temperature the measurement must have been performed at reduced pressure, and recalculated.

As seen from Table 7, most of the NBPs of the ''difficult'' compounds are predicted by M235 and M20 with a relative error of less than 5% from the published value, if the calculated boiling points for several particular compounds are chosen. There are two compounds, which are predicted with a distinctively higher error. One of them — adamantane, has a very complex structure (Fig. 1), which obviously is not well represented in the database. The other one — 1,1,2,2-tetraphenylethane, even with a recalculated NBP falls closer to the aliphatic homologous series (Fig. 2), rather than to the non-alkylated aromatics, as might be expected. Its published NBP is drastically different from the predicted values.

Table 6
Predictions of the M-235 and M-20 models for the original learning set of 235 hydrocarbons
Statistics for the predictions of the M-235 model: Mean standard deviation of absolute errors = 4.86 ± 0.35 K; Min absolute error = −13.4 K; Max absolute error = +15.1 K (7 points out of a ±11 K error range); Mean standard deviation of relative errors = 1.15 ± 0.07%; Min relative error = −4.35%; Max relative error = +3.69% (9 points out of a ±2.4% error range). Statistics for the predictions of the M-20 model: Mean standard deviation of absolute errors = 5.51 ± 0.36 K; Min absolute error = −15.1 K; Max absolute error = +26.3 K (10 points out of a ±11 K error range); Mean standard deviation of relative errors = 1.27 ± 0.08%; Min relative error = −3.84%; Max relative error = +3.92% (13 points out of a ±2.4% error range).

No. of compound	T_{publ}, (K)	T_{pred} M-235, (K)	T_{pred} M-20, (K)	Absolute Error M-235, (K)	Absolute Error M-20, (K)	Relative Error M-235, (%)	Relative Error M-20, (%)
1	184.6	184.6	184.9	−0.1	−0.3	0.0	−0.2
2	231.1	230.7	230.9	0.3	0.1	0.1	0.1
3	272.7	270.4	270.4	2.2	2.2	0.8	0.8
4	309.2	307.1	306.7	2.1	2.5	0.7	0.8
5	341.9	339.0	338.4	2.9	3.5	0.8	1.0
6	371.6	369.0	368.0	2.6	3.6	0.7	1.0
7	398.8	395.6	394.3	3.2	4.5	0.8	1.1
8	424.0	421.2	419.7	2.7	4.3	0.6	1.0
9	447.3	444.1	442.3	3.2	5.0	0.7	1.1
10	469.0	462.3	460.7	6.7	8.3	1.4	1.8
11	489.4	483.4	481.5	6.1	7.9	1.2	1.6
12	508.6	499.1	497.5	9.5	11.1	1.9	2.2
14	543.8	534.8	532.8	9.0	10.9	1.7	2.0
15	559.9	551.9	549.7	8.1	10.2	1.4	1.8
16	575.3	568.0	565.7	7.3	9.6	1.3	1.7
17	589.9	583.5	581.0	6.4	8.9	1.1	1.5
18	603.8	598.3	595.7	5.4	8.0	0.9	1.3
19	616.9	613.4	610.5	3.5	6.4	0.6	1.0
20	629.7	624.9	622.1	4.8	7.5	0.8	1.2
21	641.8	635.8	633.2	6.0	8.6	0.9	1.3
22	653.4	650.4	647.5	2.9	5.9	0.4	0.9
23	664.4	664.6	661.4	−0.2	3.0	0.0	0.5
24	675.0	670.6	667.8	4.5	7.2	0.7	1.1
25	685.4	680.6	677.9	4.8	7.5	0.7	1.1
26	695.3	691.1	688.4	4.1	6.9	0.6	1.0
27	704.8	702.4	699.5	2.3	5.2	0.3	0.7
28	713.9	709.1	706.6	4.8	7.4	0.7	1.0
29	722.9	718.6	716.0	4.3	6.9	0.6	0.9
30	736.7	736.5	733.9	0.2	2.8	0.0	0.4
31	758.8	759.7	757.3	−0.9	1.5	−0.1	0.2
32	765.6	764.2	762.1	1.4	3.5	0.2	0.5
33	790.8	795.7	793.4	−4.9	−2.6	−0.6	−0.3
34	812.7	814.2	812.8	−1.5	−0.1	−0.2	0
35	877.5	872.3	873.9	5.2	3.6	0.6	0.4
36	261.4	261.2	263.5	0.2	−2.1	0.1	−0.8
37	301.0	298.8	300.6	2.2	0.4	0.7	0.1
38	282.6	287.4	288.4	−4.8	−5.8	−1.7	−2.1
39	333.4	331.8	333.2	1.6	0.2	0.5	0.1

Table 6 (continued)

No. of compound	T_{publ}, (K)	T_{pred} M-235, (K)	T_{pred} M-20, (K)	Absolute Error M-235, (K)	Absolute Error M-20, (K)	Relative Error M-235, (%)	Relative Error M-20, (%)
41	322.9	324.1	324.3	−1.3	−1.5	−0.4	−0.5
42	331.1	328.7	331.8	2.5	−0.6	0.7	−0.2
43	363.2	363.9	364.7	−0.7	−1.5	−0.2	−0.4
44	365.0	366.1	366.7	−1.1	−1.7	−0.3	−0.5
45	352.3	353.2	353.1	−0.9	−0.8	−0.2	−0.2
46	362.9	360.7	363.2	2.3	−0.2	0.6	−0.1
47	353.6	355.2	358.1	−1.6	−4.5	−0.4	−1.3
48	359.2	358.6	357.9	0.6	1.3	0.2	0.3
49	366.6	368.4	368.7	−1.8	−2.1	−0.5	−0.6
50	354.0	351.0	352.5	3.1	1.5	0.9	0.4
51	390.8	388.4	389.2	2.4	1.6	0.6	0.4
52	392.1	392.7	393.0	−0.6	−0.9	−0.1	−0.2
53	390.9	390.7	391.3	0.2	−0.4	0.1	−0.1
54	380.0	379.7	379.5	0.3	0.5	0.1	0.1
55	388.8	387.3	389.6	1.5	−0.8	0.4	−0.2
56	382.6	389.0	391.4	−6.4	−8.8	−1.7	−2.3
57	382.3	384.2	386.8	−1.9	−4.5	−0.5	−1.2
58	385.1	384.8	383.9	0.3	1.2	0.1	0.3
59	390.9	390.7	392.7	0.2	−1.8	0.1	−0.5
60	391.7	398.9	398.8	−7.2	−7.1	−1.8	−1.8
61	383.0	383.7	384.6	−0.7	−1.6	−0.2	−0.4
62	372.4	377.9	379.3	−5.5	−6.9	−1.5	−1.9
64	386.6	385.9	389.6	0.7	−3.0	0.2	−0.8
65	388.8	390.2	392.1	−1.4	−3.3	−0.4	−0.8
66	391.4	393.6	395.1	−2.2	−3.7	−0.6	−1.0
67	406.8	406.3	407.2	0.5	−0.4	0.1	−0.1
68	399.7	406.8	407.6	−7.1	−7.9	−1.8	−2.0
69	397.2	400.2	401.5	−3.0	−4.3	−0.8	−1.1
70	419.3	423.2	420.5	−3.9	−1.1	−0.9	−0.3
71	413.4	405.9	405.2	7.5	8.2	1.8	2.0
72	406.2	403.7	406.2	2.5	0.0	0.6	0.0
74	414.7	405.6	407.8	9.1	6.9	2.2	1.7
75	416.4	413.0	413.6	3.4	2.8	0.8	0.7
76	440.2	435.8	436.2	4.4	4.0	1.0	0.9
77	428.9	432.9	433.1	−4.1	−4.3	−1.0	−1.0
78	433.3	427.8	427.0	5.5	6.3	1.3	1.5
79	471.3	472.4	474.4	−1.1	−3.1	−0.2	−0.7
80	506.8	511.9	513.3	−5.1	−6.5	−1.0	−1.3
81	526.2	525.8	528.7	0.4	−2.6	0.1	−0.5
82	558.2	559.1	561.6	−0.9	−3.5	−0.2	−0.6
86	769.0	758.6	770.5	10.4	−1.5	1.4	−0.2
87	225.4	231.6	232.1	−6.2	−6.7	−2.7	−3
89	303.1	304.9	305.2	−1.8	−2.1	−0.6	−0.7
90	336.6	339.9	339.6	−3.3	−3.0	−1.0	−0.9
91	366.8	370.0	369.3	−3.2	−2.6	−0.9	−0.7
92	394.4	396.4	395.5	−1.9	−1.1	−0.5	−0.3

Table 6 (continued)

No. of compound	T_{publ}, (K)	T_{pred} M-235, (K)	T_{pred} M-20, (K)	Absolute Error M-235, (K)	Absolute Error M-20, (K)	Relative Error M-235, (%)	Relative Error M-20, (%)
93	420.0	422.8	421.5	−2.8	−1.5	−0.7	−0.4
94	443.8	445.4	444.0	−1.6	−0.2	−0.4	0.0
95	465.8	466.9	465.3	−1.1	0.5	−0.2	0.1
96	486.5	485.4	483.8	1.1	2.7	0.2	0.5
97	505.9	509.5	507.2	−3.6	−1.3	−0.7	−0.3
98	524.3	525.0	522.8	−0.7	1.4	−0.1	0.3
99	541.5	542.1	539.8	−0.6	1.7	−0.1	0.3
100	558.0	559.5	557.0	−1.5	1.0	−0.3	0.2
101	573.5	573.3	570.9	0.1	2.5	0.0	0.4
102	588.0	586.8	584.4	1.2	3.6	0.2	0.6
103	601.7	599.5	597.2	2.2	4.5	0.4	0.7
104	615.5	612.4	610.1	3.1	5.4	0.5	0.9
105	628.2	627.9	625.4	0.2	2.8	0.0	0.4
106	640.4	642.3	639.5	−1.9	0.9	−0.3	0.1
107	652.0	652.6	649.9	−0.6	2.1	−0.1	0.3
108	663.2	658.8	656.7	4.3	6.5	0.6	1.0
109	674.3	679.0	675.9	−4.8	−1.6	−0.7	−0.2
110	684.3	689.8	686.6	−5.5	−2.3	−0.8	−0.3
111	694.3	695.4	692.7	−1.2	1.5	−0.2	0.2
112	702.2	704.7	702.1	−2.6	0.1	−0.4	0.0
113	713.2	717.9	714.9	−4.7	−1.7	−0.7	−0.2
114	720.9	726.4	723.5	−5.5	−2.5	−0.8	−0.4
115	268.7	275.1	275.8	−6.4	−7.1	−2.4	−2.6
116	276.9	267.6	268.8	9.2	8.1	3.3	2.9
117	274.0	267.6	268.8	6.4	5.2	2.3	1.9
118	266.3	267.2	268.4	−1.0	−2.2	−0.4	−0.8
119	307.2	310.0	311.1	−2.8	−3.8	−0.9	−1.3
120	328.8	332.2	334.7	−3.4	−5.9	−1.0	−1.8
121	346.4	333.6	335.4	12.8	11	3.7	3.2
122	337.8	339.2	339.4	−1.4	−1.6	−0.4	−0.5
123	342.0	337.0	337.4	5.0	4.6	1.5	1.4
124	341.0	336.8	337.3	4.2	3.7	1.2	1.1
125	335.3	337.1	337.7	−1.9	−2.5	−0.6	−0.7
126	327.0	330.1	332.2	−3.1	−5.2	−1.0	−1.6
127	374.6	379.8	380.6	−5.2	−6	−1.4	−1.6
128	378.1	373.9	375.2	4.2	2.8	1.1	0.7
129	304.3	305.1	305.9	−0.8	−1.6	−0.3	−0.5
130	311.7	301.6	302.8	10.1	8.9	3.3	2.9
131	293.2	297.9	300.4	−4.7	−7.1	−1.6	−2.4
132	341.9	337.4	339.4	4.6	2.5	1.3	0.7
133	314.4	321.1	322.2	−6.7	−7.8	−2.1	−2.5
134	340.5	335.3	336.4	5.2	4.1	1.5	1.2
135	327.3	332.8	334.6	−5.4	−7.3	−1.7	−2.2
136	332.6	340.0	340.2	−7.4	−7.6	−2.2	−2.3
137	449.2	447.2	449.0	2.0	0.1	0.4	0.0
138	429.3	437.4	431.0	−8.1	−1.7	−1.9	−0.4

Table 6 (continued)

No. of compound	T_{publ}, (K)	T_{pred} M-235, (K)	T_{pred} M-20, (K)	Absolute Error M-235, (K)	Absolute Error M-20, (K)	Relative Error M-235, (%)	Relative Error M-20, (%)
139	708.5	699.1	708.0	9.4	0.4	1.3	0.1
142	240.4	250.8	249.6	−10.4	−9.2	−4.3	−3.8
143	285.7	287.8	283.3	−2.1	2.4	−0.7	0.8
145	353.9	349.9	351.2	4.0	2.7	1.1	0.8
146	391.9	382.5	383.7	9.4	8.2	2.4	2.1
147	424.3	410.6	411.5	13.7	12.8	3.2	3.0
148	374.1	377.6	379.3	−3.5	−5.3	−0.9	−1.4
149	405.0	402.9	404.6	2.0	0.3	0.5	0.1
150	429.9	429.6	430.9	0.3	−1.0	0.1	−0.2
151	454.1	452.4	453.5	1.7	0.6	0.4	0.1
152	345.0	351.9	352.7	−6.9	−7.8	−2.0	−2.2
153	376.6	381.3	381.8	−4.7	−5.2	−1.2	−1.4
154	404.1	409.9	410.0	−5.8	−5.9	−1.4	−1.5
155	429.8	435.5	435.3	−5.7	−5.5	−1.3	−1.3
156	453.7	454.2	454.2	−0.5	−0.5	−0.1	−0.1
157	476.0	476.5	476.3	−0.5	−0.3	−0.1	−0.1
158	497.0	495.5	495.3	1.5	1.7	0.3	0.3
159	516.7	517.5	516.7	−0.8	0.0	−0.2	0.0
160	535.2	536.3	535.4	−1.1	−0.2	−0.2	0.0
161	552.5	554.2	553.0	−1.7	−0.5	−0.3	−0.1
162	568.9	570.2	568.9	−1.3	0.0	−0.2	0.0
163	584.4	585.4	584.1	−1.0	0.3	−0.2	0.1
164	599.0	600.8	599.4	−1.8	−0.4	−0.3	−0.1
165	613.2	613.5	612.1	−0.3	1.1	0.0	0.2
166	625.9	626.6	625.2	−0.7	0.7	−0.1	0.1
167	639.3	640.1	638.6	−0.8	0.7	−0.1	0.1
168	650.3	654.0	652.5	−3.7	−2.2	−0.6	−0.3
169	661.9	665.5	663.8	−3.6	−1.9	−0.5	−0.3
170	673.1	673.5	672.2	−0.4	0.9	−0.1	0.1
171	683.1	688	686.3	−4.9	−3.2	−0.7	−0.5
172	693.1	695.9	694.4	−2.8	−1.3	−0.4	−0.2
173	703.1	706.5	704.9	−3.4	−1.8	−0.5	−0.3
174	712.0	716.8	715.2	−4.8	−3.2	−0.7	−0.5
175	720.4	725.4	723.9	−5.0	−3.5	−0.7	−0.5
176f	729.3	732.8	731.6	−3.5	−2.3	−0.5	−0.3
177	402.9	405.3	407.4	−2.4	−4.4	−0.6	−1.1
178	396.6	402	404.4	−5.5	−7.8	−1.4	−2
179	393.2	406.6	408.3	−13.4	−15.1	−3.4	−3.8
180	397.6	404.5	406.6	−6.9	−9.0	−1.7	−2.3
181	397.5	402.3	404.6	−4.8	−7.2	−1.2	−1.8
182	392.5	396.6	399.5	−4.1	−7.0	−1.0	−1.8
183	317.4	315.4	312.9	2.0	4.5	0.6	1.4
184	356.1	346.1	345.8	10.1	10.3	2.8	2.9
185	353.5	350.6	348.8	2.9	4.6	0.8	1.3
186	345.9	336.3	332.4	9.7	13.6	2.8	3.9
187	314.7	307.8	303.1	6.8	11.6	2.2	3.7

Table 6 (continued)

No. of compound	T_{publ}, (K)	T_{pred} M-235, (K)	T_{pred} M-20, (K)	Absolute Error M-235, (K)	Absolute Error M-20, (K)	Relative Error M-235, (%)	Relative Error M-20, (%)
188	353.3	355.9	352.4	2.6	0.9	−0.7	0.2
189	383.8	385.5	383	−1.8	0.8	−0.5	0.2
190	409.3	413.8	410.9	−4.5	−1.6	−1.1	−0.4
191	432.4	436.4	433.5	−4.0	−1.2	−0.9	−0.3
192	456.5	461.8	458.4	−5.3	−2.0	−1.2	−0.4
193	417.6	411.7	409.6	5.9	8.0	1.4	1.9
194	412.3	406.9	405.0	5.4	7.3	1.3	1.8
195	411.5	408.6	406.9	2.9	4.6	0.7	1.1
196	434.5	430.4	428.4	4.1	6.1	1.0	1.4
197	478.6	480.5	477.2	−1.9	1.3	−0.4	0.3
198	499.3	500.4	497.1	−1.2	2.2	−0.2	0.4
199	519.2	517.9	514.6	1.2	4.6	0.2	0.9
201	555.2	554.9	551.1	0.3	4.1	0.1	0.7
202	571.0	570.3	566.5	0.7	4.6	0.1	0.8
203	586.4	586.4	582.4	−0.1	3.9	0.0	0.7
204	600.8	600.8	596.8	0.0	4.0	0.0	0.7
205	614.4	616.2	612	−1.8	2.5	−0.3	0.4
206	627.0	628.8	624.6	−1.8	2.4	−0.3	0.4
207	639.3	638.4	634.5	0.8	4.7	0.1	0.7
208	650.9	646.1	642.6	4.9	8.3	0.7	1.3
209	662.0	664.0	659.9	−1.9	2.2	−0.3	0.3
210	673.2	674.8	670.7	−1.7	2.4	−0.3	0.4
211	683.2	685.5	681.4	−2.4	1.7	−0.3	0.3
212	693.2	695.5	691.4	−2.3	1.7	−0.3	0.2
213	702.0	707.7	703.4	−5.6	−1.4	−0.8	−0.2
214	710.9	718.5	714.1	−7.5	−3.1	−1.1	−0.4
215	719.3	726.8	722.6	−7.6	−3.3	−1.1	−0.5
216	727.0	734.1	729.5	−7.1	−2.5	−1.0	−0.3
217	418.3	415.6	413.2	2.8	5.1	0.7	1.2
218	438.7	443.1	440.8	−4.4	−2.2	−1.0	−0.5
219	425.6	430	429.1	−4.4	−3.5	−1.0	−0.8
220	438.3	440.1	438.1	−1.8	0.2	−0.4	0.0
221	435.2	436.8	435.3	−1.6	−0.1	−0.4	0.0
222	437.9	431.7	431.1	6.2	6.8	1.4	1.5
223	449.3	440.4	439.6	8.9	9.7	2.0	2.2
224	442.5	436.7	436.3	5.8	6.2	1.3	1.4
225	480.8	478.4	476.1	2.4	4.7	0.5	1.0
226	442.3	445.9	443.7	−3.6	−1.4	−0.8	−0.3
227	450.3	452.0	452.4	−1.7	−2.2	−0.4	−0.5
228	454.3	457.2	455.5	−2.9	−1.2	−0.6	−0.3
229	456.6	461.6	459.8	−5.0	−3.1	−1.1	−0.7
230	445.9	441.5	441.2	4.4	4.7	1.0	1.1
231	476.3	487.5	489.2	−11.1	−12.9	−2.3	−2.7
232	537.4	539.3	534.2	−1.9	3.3	−0.3	0.6
233	434.6	436.2	434.4	−1.7	0.2	−0.4	0.0
234	446.5	452.6	451.5	−6.1	−5.1	−1.4	−1.1

Table 6 (continued)

No. of compound	T_{publ}, (K)	T_{pred} M-235, (K)	T_{pred} M-20, (K)	Absolute Error M-235, (K)	Absolute Error M-20, (K)	Relative Error M-235, (%)	Relative Error M-20, (%)
235	456.9	460.5	458.8	−3.6	−1.8	−0.8	−0.4
236	483.7	491.1	492.9	−7.5	−9.2	−1.5	−1.9
237	528.4	531.7	526.3	−3.3	2.1	−0.6	0.4
238	545.8	550.0	546.8	−4.2	−1.0	−0.8	−0.2
239	553.7	557.0	551.6	−3.3	2.0	−0.6	0.4
240	491.1	489.2	483.7	1.9	7.4	0.4	1.5
241	612.9	597.8	590.6	15.1	22.3	2.5	3.6
242	612.6	600.4	593.1	12.2	19.5	2.0	3.2
243	638.0	651.4	644.5	−13.4	−6.5	−2.1	−1
244	649.0	643.9	637.6	5.1	11.4	0.8	1.8
246	668.0	666.4	657.6	1.6	10.4	0.2	1.6
247	714.1	701.3	691.9	12.8	22.2	1.8	3.1
251	550.5	548.1	540.5	2.4	10.0	0.4	1.8

In order to illustrate further the problem with the reliability of published NBPs we have tried to follow up the original source of the NBP of 1,1,2,2-tetraphenylethane. In our database this value (358–362°C) was cross-referenced from two papers citing as reference the DIPPR database [19], and the Beilstein database [22]. Reference of the original Beilstein handbook revealed that this compound has been included in the main work (Hauptwerke) of the series [35]. Two NBPs are recommended — 358–362°C (uncorrected) and — 379–383°C (corrected). The original experimental determination of the boiling point of 1,1,2,2-tetraphenylethane was done by Biltz [36], who synthesized and purified the compound himself in the year of 1897.

The M-20 model provides somewhat better predictions for most of the published NBPs. The values predicted for the hypotheticals are close to values estimated by ABCs, which are developed for rather different homologous series, but are the only other alternative. It has to be underlined that the three hypotheticals are outliers on the principal components plot (Fig. 3), so the error in the estimation of their boiling points is expected to be higher than that within the boundaries of the models.

It should be noted also that separate models for the different groups of hydrocarbons may be developed following the principles suggested in the present work. Figs. 2 and 3 indicate that ‘‘pseudohomologous’’ series, including a greater number of available data may be created. For instance, alkylaromatics, behave as aliphatic hydrocarbons, above a particular aliphatic chain length.

Care should be taken when ascribing physical significance to statistically derived correlations for the contribution of particular descriptors. The latter heavily relies on the particular design of the database. For instance, the gravitational index, which managed to describe NBPs of hydrocarbons in the Katritzky et al. work [7] as a sole independent variable, did not prove to be useful with the database of the present investigation. Neither were the variations of any of the tested topological descriptors, although they carry at least part of the molecular information of some of the descriptors successfully employed by Wessel and Jurs [19].

Our observations suggest that more work is necessary to establish which descriptors might be the real determinants of the NBPs, and which are only surrogates for more fundamental features of the molecules.

Table 7
Predictions of the M-235 and M-20 models for the control set and ‘‘hypothetical’’ molecules
Statistics for the predictions of the M-235 model: Mean standard deviation of absolute errors = 18.64 ± 3.89 K; Min absolute error = −53.6 K; Max absolute error = +24.8 K (7 points out of a ±11 K error range); Mean standard deviation of relative errors = 3.23 ± 0.67%; Min relative error = −9.4%; Max relative error = +3.8% (2 points out of a ±5.0% error range).
Statistics for the predictions of M-20 model: Mean standard deviation of absolute errors = 16.34 ± 3.41 K; Min absolute error = −48.5 K; Max absolute error = +20.3 K (10 points out of a ±11 K error range); Mean standard deviation of relative errors = 2.89 ± 0.60%; Min relative error = −8.7%; Max relative error = +3.2% (2 points out of a ±5.0% error range).

No.	Name	T_{calc}, (K)	T_{publ}, (K)	T_{pred} M-235, (K)	T_{pred} M-20, (K)	Absolute Error M-235, (K)	Absolute Error M-20, (K)	Relative Error M-235, (%)	Relative Error M-20, (%)
13	*n*-tetradecane		526.7	516.9	515.2	9.8	11.5	1.9	2.2
40	3-methylpentane		336.4	333.9	335.1	2.5	1.3	0.7	0.4
63	2,3,3-trimethylpentane		387.9	382.9	383.7	5.0	4.2	1.3	1.1
73	2,2,4,4-tetramethylpentane		395.4	393.7	393	1.7	2.4	0.4	0.6
83	pristane		604.3	582.9	587.1	21.6	17.2	3.6	2.8
84	phytane		625.6	601.9	605.3	23.7	20.3	3.8	3.2
85	squalane		720.0	695.2	702.9	24.8	17.1	3.4	2.4
88	1-butene		266.9	270.3	270.8	−3.4	−3.9	−1.3	−1.5
140	lycopene	795.8[a]	–	766.0	788.6				
141	β-carotene	831.4[b]	–	818.4	839.2				
144	cyclopentane		322.4	324.9	325.0	−2.5	−2.6	−0.8	−0.8
200	*n*-octylbenzene		537.5	536.5	533.0	1.0	4.5	0.2	0.8
245	1,2-benzo [*a*] pyrene	778.28[c]	–	773.5	763.6				
248	*o*-terphenyl	637.5[d]	610.6	654.6	648.0	−9.2	−2.6	−1.4	−0.4
249	triphenylmethane		632.1	651.8	646.4	−19.7	−14.3	−3.1	−2.3
250	acenaphtalene	552.0[d]	543.1	548.1	537.0	−0.1	11.0	0.0	2.0
252	1,1,2,2-tetraphenylethane	697.7[d]	633.1	751.5	746.4	−53.6	−48.5	−7.7	−7.0
253	4-methyloctane		415.6	412.1	412.8	3.5	2.8	0.8	0.7
254	2,2,3,3-tetramethylbutane		379.4	374.3	374.4	5.1	5.0	1.3	1.3
255	2-ethyl-1-hexene		393.1	397.2	397.0	−4.1	−3.9	−1.0	−1.0
256	adamantane		461.0	504.6	501.2	−43.6	−40.2	−9.5	−8.7
257	1,5-cyclooctadiene	409.9[d]	423.3	415.7	414.6	−5.8	−4.7	−1.4	1.1
258	2,5-dimethyl-1,5-hexadiene		387.4	380.8	383.2	6.6	4.2	1.7	1.1
259	*cis*-1-propenylbenzene	441.9[d]	452	436.9	434.5	5.0	7.4	1.1	1.7
260	1-phenylnaphtalene	621.0[d]	607.1	643.6	634	−22.6	−13.0	−3.6	−2.1
261	indane		451.1	451.4	447.1	−0.3	4.0	−0.1	0.9

[a] Calculated with an Asymptotic Behaviour Correlation [10]. The original correlation is for $n-1$-alkenes.
[b] Calculated with an Asymptotic Behaviour Correlation [10]. The original correlation is for alkylcyclohexanes.
[c] Calculated from a fit of the NBPs of the rings only aromatic hydrocarbons as a function of $C_{tot}^{0.70}$.
[d] Calculated from published boiling temperatures at reduced pressure. Considered more reliable and used for the determination of errors.

4. Conclusions

The present work contributes a correlation to the very challenging and important investigations of the quantitative relation between the molecular structure and the functional properties of chemical compounds, which has been a fundamental task of chemistry and chemical engineering for many years.

Its main features, as perceived by the present authors, are its relative simplicity, its reliable predictions of the NBPs, and its applicability to diversified industrially important hydrocarbon structures within a widely spanned range of NBPs and number of carbon atoms.

An achievement of particular interest in the present work is the revealed opportunity for the limitation of the learning set through multivariate analysis and molecular design.

The molecular mechanics simulation employed in this study is viewed by the present authors as a potential tool for incorporation in future chemical engineering simulators. It will significantly enhance the capabilities of the latter for designing processes with chemical reactions and is especially suitable for optimisation of the composition of the additive products of the chemical industry [37]. Furthermore, it allows straightforward, but correct input of complex chemical structures by drawing them directly on the monitor. However, from the point of view of the chemical engineer as the user of such programmes, the benefits of the sophistication cannot be easily appreciated. On the one hand, the sophistication requires an in depth knowledge of the quantum chemistry of the particular structures, which is not so readily available for the structures targeted by chemical engineers. On the other hand, in many cases of engineering importance the sophistication and high accuracy may not be justified, since simple group contribution correlations still work successfully for particular problems. The appropriate level of sophistication for many of the common chemical engineering applications will be different and can only be determined by systematic studies of the influence of uncertainties on key parameters.

The high accuracy achieved by the correlation opens up a possibility for systematic studies of chemical engineering applications in which the effects of small changes are important. This also outlines a path towards the more general problem of the influence of uncertainties in calculated thermophysical parameters on the final solution of computer aided simulation and design.

Acknowledgements

The present authors acknowledge with gratitude the financial support of The Royal Society.

References

[1] L.M. Egolf, M.D. Wessel, P.C. Jurs, J. Chem. Inf. Comput. Sci. 34 (1994) 947–956.

[2] S.J. Grigoras, Comput. Chem. 11 (1990) 593–610.

[3] W.J. Lyman, W.F. Reehl, D.H. Rosenblatt, Handbook of Chemical Property Estimation Methods, McGraw-Hill, New York, 1982.

[4] R.C. Reid, J.M. Prauznitz, B.E. Poling, Properties of Gases and Liquids, 4th edn., McGraw-Hill, New York, 1987.

[5] C.H. Fisher, J. Am. Oil Chem. Soc. 67 (1990) 101–102.

[6] R.C. Mebane, C.D. Williams, T.R. Rybolt, Fluid Phase Equilibria 124 (1996) 111–122.

[7] A.R. Katritzky, V.S. Lobanov, M. Karelson, J. Chem. Inf. Comput. Sci. 38 (1998) 28–41.
[8] A. Kreglewski, B.J. Zwolinski, J. Phys. Chem. 65 (1961) 1050–1052.
[9] K.A. Gasem, C.H. Ross, R.L. Robinson Jr., Can. J. Chem. Eng. 77 (1993) 805–816.
[10] J.J. Marano, G.D. Holder, Ind. Eng. Chem. Res. 36 (1997) 1887–1894.
[11] J.J. Marano, G.D. Holder, Ind. Eng. Chem. Res. 36 (1997) 1895–1907.
[12] A.L. Horvath, Molecular Design, Elsevier, Amsterdam, 1992.
[13] M. Karelson, Adv. Quant. Chem. 28 (1997) 141–157.
[14] M. Kurata, S. Ishida, J. Chem. Phys. 23 (1955) 1126–1131.
[15] I.C. Sanchez, R.H. Lacombe, J. Phys. Chem. 80 (1976) 2352–2362.
[16] I.C. Sanchez, R.H. Lacombe, Macro-molecules 11 (1978) 1145–1156.
[17] P.J. Flory, R.A. Orwoll, A. Vrij, J. Am. Chem. Soc. 86 (1964) 3507–3514.
[18] A. Vetere, Fluid Phase Equilibria 124 (1996) 15–29.
[19] M.D. Wessel, P.C. Jurs, J. Chem. Inf. Comp. Sci. 35 (1995) 68–76.
[20] J. Buckingham, S.M. Donaghy (Eds.), Dictionary of Organic Compounds, 5th ed., Chapman and Hall, New York, 1982.
[21] API Technical data Book — Petroleum Refining, 4th edn. American Petroleum Institute, Washington DC, 1983.
[22] Iu.V. Pokonova, A.A. Gaile, V.G. Spirkin, Chemistry of Petroleum, Himia, Leningrad, 1984 (in Russian).
[23] A.I. Bogomolov, A.A. Gaile, V.V. Gromova, Chemistry of Petroleum and Gas, Himia, Leningrad, 1989 (in Russian).
[24] TRC (Thermodynamic Research Center). TRC Thermodynamic Tables–Hydrocarbons, The Texas A&M University, College Station, TX, USA, 1997 revision.
[25] A.S. Teja, R.J. Lee, D. Rosenthal, M. Anselme, Fluid Phase Equilibria 56 (1990) 153–169.
[26] PCMODEL, 5th edn., Serena Software, Bloomington, IN, USA, 1992.
[27] M. Randic, B. Jerman-Blazic, N. Trinajstic, Comput. Chem. 14 (1990) 237–246.
[28] J.K. Labanowski, I. Motoc, R.A. Damkoehler, Comp. Chem. 15 (1) (1991) 47–53.
[29] A.R. Katritzky, L. Mu, V.S. Lobanov, M. Karelson, J. Phys. Chem. 100 (1996) 10400–10407.
[30] STATGRAPHICS for DOS 7th edn., STSC, Inc. and Manugistics, Inc., Rockville, MD, USA.
[31] P. Geladi, M.-L. Tosato, in: W. Karcher, J. Devillers (Eds.), Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology, Kluwer Acad. Publ., Dodrecht, 1990, pp. 170–179.
[32] S. Wold, K. Esbensen, P. Geladi, Chemometrics and Intelligent Laboratory Systems 2 (1987) 37–52.
[33] P. Geladi, B. Kowalski, Anal. Chim. Acta 185 (1986) 1–17.
[34] M.-L. Tosato, P. Geladi, in: W. Karcher, J. Devillers (Eds.), Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology, Kluwer Acad. Publ., Dodrecht, 1990, pp. 317–341.
[35] Beilsteins Handbuch der Organischen Chemie, 4th edn., H, bd. V, Springer Verlag, 1933, p. 739.
[36] H. Biltz, Liebigs Annalen der Chemie 296 (1897) 221.
[37] G.S. Cholakov, K.G. Stanulov, P.A. Devenski, H.A. Iontchev, Wear 216 (2) (1998) 194–201.