ORIGINAL SCIENTIFIC PAPER

CROATICA CHEMICA ACTA

Croat. Chem. Acta **2021**, *94*(1), 47–58
Published online: December 31, 2021
DOI: 10.5562/cca3807

HKD 1926

Automorphism Groups of Alkane Graphs

Kurt Varmuza,[1,2,*] Matthias Dehmer,[3,4] Frank Emmert-Streib,[5,6] Peter Filzmoser[1]

[1] Institute of Statistics and Mathematical Methods in Economics / Computational Statistics, Vienna University of Technology, Vienna, Austria
[2] Institute of Chemical, Environmental and Bioscience Engineering, Vienna University of Technology, Vienna, Austria
[3] Institute for Bioinformatics and Translational Research, UMIT, Eduard Wallnöfer Zentrum, Hall in Tyrol, Austria
[4] Swiss Distance University of Applied Sciences, Department of Computer Science, Brig, Switzerland
[5] Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland
[6] Institute of Biosciences and Medical Technology, Tampere University, Tampere, Finland
* Corresponding author's e-mail address: kurt.varmuza@tuwien.ac.at

RECEIVED: June 24, 2021 * REVISED: September 3, 2021 * ACCEPTED: September 10, 2021

Abstract: The complete set of 618047 isomers of the alkanes with 4 to 20 carbon atoms has been created. From all isomers the automorphism groups have been calculated and evaluated in terms of size, number of asymmetric carbon atoms, and numeric properties of atom and bond orbits. The presence of a symmetric bond is related to the number of atom and bond orbits. Molecular descriptors based on automorphism data have been studied, including the known symmetry index, an entropy measure and the root of an orbit polynomial. These descriptors are closely related to the presence of symmetric substructures. The prediction performance of QSPR models for three molecular properties of alkanes and using binary substructure descriptors, is improved by adding descriptors based on automorphism data.

Keywords: alkane isomers, automorphism groups, molecular descriptors, symmetry, orbit polynomial, QSPR models, PLS, repeated double cross validation.

1. INTRODUCTION

THE concept of constitutionally equivalent atoms and constitutionally equivalent bonds is an essential subject in chemoinformatics[1] and in mathematics in chemistry, in particular for applications of the graph theory to chemical structures.[2,3] Constitutionally (topologically) equivalent atoms (bonds) have exactly the same neighborhood in terms of connectivity as described by atom types (elements) and bond types - considering the whole molecular structure. Constitutionally equivalent atoms, e. g., give approximately the same shifts of NMR signals. Other applications of this concept are in synthesis design, canonical numbering of atoms, isomer generation, determination of maximum common substructures, or characterization of the molecular symmetry.

The complete information about constitutionally equivalent atoms and bonds is contained in the data of the automorphism group of a graph representing the molecular structure.[4–6] Automorphism means mapping of a graph onto itself while preserving the connectivity (not cutting any bonds). Asymmetric structures have only a single mapping - the trivial identity mapping - while for highly symmetric structures several (sometime many) mappings onto itself exist. The size of the automorphism group is the number of possible mappings of a graph onto itself; it is a symmetry measure for the graph. For general graphs, the group size is unknown.

As commonly used, we represent chemical structures by colored graphs[3] with the atoms as vertices and the bonds as edges. In this work, however, only alkane structures are considered corresponding to uncolored, undirected graphs of the type trees. We use hydrogen-depleted graphs. The vertex degree - which is the number of bonds (edges) per atom (vertex) - is in alkane graphs between one and four.

The complete sets of isomeric alkane structures for 4 to 20 carbon atoms - comprising in total 618047 structures - is created by the isomer generator program MOLGEN[7–9] with output in the Molfile format (SDF-files).[1] The complete automorphism group of each isomer has been determined by the software SubMat[10,11] applying a

This work is licensed under a Creative Commons Attribution 4.0 International License.

substructure search with the molecular structure itself as substructure and determination of all positions of the substructure. For the evaluation of automorphism data a set of functions in the programming environment R[12] has been developed. Note other software products available for the application of graph theory concepts, like nauty and Traces[13] or igraph.[14]

An extension of this work towards colored graphs (molecular structures with hetero atoms and various bond types), together with the definition of topological descriptors based on automorphism data, is in progress.

2. THEORY AND METHODS

The chemical structure of 2,2,3-trimethyl-butane (Figure 1) serves for a demonstration of automorphism data and graph properties derived thereof as used in this work. The seven vertices (carbon atoms) are arbitrarily denoted by **1** to **7**, the six edges (bonds) by **a** to **f**. Table 1 shows in the upper part the automorphism mappings of this graph. The first row is the identity mapping with the atom identifiers **1** to **7** (left part of the table, 'Atom mappings') and bond identifiers **a** to **f** (right part, 'Bond mappings'). In row i = 2 a non trivial mapping is defined with the atoms **6** and **7** exchanged, and consequently the bonds **e** and **f** exchanged. This mapping can be considered as the result of a substructure search with the molecular structure used as the substructure. In total 12 such mappings are possible (including the identity mapping) and thus defining the complete **automorphism group** with size α = 12, also denoted as $|\mathrm{Aut}(G)|$, the order of the automorphism group of a graph G. For this simple graph α can be easily calculated from the numbers of permutations at the vertices; 3! for atom **2** and 2! for atom **5**, giving $\alpha = 3! \times 2! = 12$. This structure is one of the nine isomers of C_7H_{16}, and has the maximum α among these isomers.

The size of the automorphism group is 1 for asymmetric graphs and is an even number for other graphs. The parity property is based on the fact that in the automorphism mappings the atom positions are permutated and the number of permutations is a factorial always containing the factor 2. Actually, for alkane graphs with $\alpha > 1$, α is always a product of the factors 2!, and/or 3! and/or 4!. Possible values for $\alpha > 1$ are therefore 2, 4, 6, 8, 12, 16, 24, 32, 36, 48, 64, 72, 96, 128, 144, 192, 216, 256, 288, 384, 432, 512, 576, 768, 1024, and so on. The limited number of different values of α restricts the use of molecular descriptors that are based only on α. Theoretical aspects of the size of automorphism groups of simple graphs are discussed by Krasikov.[15]

An **atom (vertex) orbit** is the set of constitutionally equivalent atoms. The column for atom **1** in Table 1 shows that atom **1** can be replaced by atoms **3** or **4**, consequently

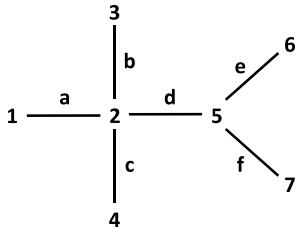

Figure 1. Graph of the C_7-alkane isomer with highest symmetry (size of automorphism group is 12, the maximum in this isomer set), representing the chemical structure of 2,2,3-trimethyl-butane, C_7H_{16}. Atoms (vertices) are denoted by **1** to **7**, bonds (edges) by **a** to **f** (with arbitrary assignments).

Table 1. Automorphism data for structure (graph) of 2,2,3-trimethyl-butane shown in Figure 1.

i	Atom mappings							Bond mappings					
1	1	2	3	4	5	6	7	a	b	c	d	e	f
2	1	2	3	4	5	7	6	a	b	c	d	f	e
3	1	2	4	3	5	6	7	a	c	b	d	e	f
4	1	2	4	3	5	7	6	a	c	b	d	f	e
5	3	2	1	4	5	6	7	b	a	c	d	e	f
6	3	2	1	4	5	7	6	b	a	c	d	f	e
7	4	2	1	3	5	6	7	c	a	b	d	e	f
8	4	2	1	3	5	7	6	c	a	b	d	f	e
9	3	2	4	1	5	6	7	b	c	a	d	e	f
10	3	2	4	1	5	7	6	b	c	a	d	f	e
11	4	2	3	1	5	6	7	c	b	a	d	e	f
12	4	2	3	1	5	7	6	c	b	a	d	f	e
Sets	1	2	1	1	5	6	6	a	a	a	d	e	e
	3		3	3		7	7	b	b	b		f	f
	4		4	4				c	c	c			
Orbits	A	C	A	A	D	B	B	X	X	X	Z	Y	Y

atoms **1**, **3** and **4** form an atom orbit (given in the lower part of Table 1, denoted by "Sets"). The same result appears in the columns for atoms **3** and **4**. The orbit containing atoms **1**, **3**, and **4** is (arbitrarily) denoted by **A** (last row in Table 1). Furthermore, atoms **6** and **7** form an atom orbit **B**; orbits **C** and **D** consist of only one atom (**2** and **5**, respectively). In summary this graph has four atom orbits (**A** to **D**) with sizes 3, 2, 1, and 1, respectively.

In the same way **bond (edge) orbits** - consisting of constitutionally equivalent bonds - are defined. In this example three bond orbits exist: bond orbit **X** (bonds **a**, **b**, **c**; size 3); **Y** (bonds **e**, **f**; size 2), and **Z** (bond **d**; size 1).

Molecular descriptors based on automorphism data can be derived from the size of the automorphism group and from the distributions of the atom and bond orbits. These graph invariants are candidates for symmetry criterions of chemical structures, to be used for molecular

Table 2. Automorphism data for the alkane structures shown in Figures 1 to 3.

Structure in Figure	1	2	3
Molecular formula	C_7H_{16}	C_7H_{16}	$C_{17}H_{36}$
n_A	7	7	17
n_B	6	6	16
n_{ASYM}	0	1	0
$f_{ASYM} = n_{ASYM} / n_A$	0	0.143	0
α	12	1	31104
k_A	4	7	3
k_B	3	6	2
a_i (i = 1 ... k_A)	3, 2, 1, 1	1 in all	12, 4, 1
b_i (i = 1 ... k_B)	3, 2, 1	1 in all	12, 4
E_A	1.842	2.807	1.086
E_B	1.459	2.585	0.811
S_A	4.550	0.000	17.926
S_B	4.711	0.000	18.114
δ_A	0.393	0.143	0.717
δ_B	0.544	0.167	0.909

n_A, number of atoms; n_B, number of bonds;
n_{ASYM}, number of asymmetric carbon atoms;
f_{ASYM}, fraction of asymmetric carbon atoms;
k_A, k_B, number of atom and bond orbits, resp.;
a_i (i = 1 ... k_A), size of atom orbit i;
b_i (i = 1 ... k_B), size of bond orbit i;
α, size of automorphism group;
E_A, E_B, entropy measure for atom and bond orbits, resp.;
S_A, S_B, symmetry index for atom and bond orbits, resp.;
δ_A, δ_B, positive real root of orbit polynomial for atom and bond orbits, resp ...

descriptors in models for QSP(A)R - quantitative structure property (activity) relationships - as well as for structure similarity searches or cluster analyses of structures. A selection of these descriptors is described here and values are given in Table 2 for the structures shown in Figures 1 to 3. Applications to QSPR models are given in section 3.4.

The **size of the automorphism group**, α, may be normalized by the number of atoms, n_A, and/or number of bonds, n_B, giving for the example structure in Figure 1 the descriptors $\alpha_A = \alpha/n_A = 12/7 = 1.71$; $\alpha_B = \alpha/n_B = 12/6 = 2$; and $\alpha_{AB} = \alpha/(n_A + n_B) = 12/(7+6) = 0.92$. Because α spans a wide range of values, $\alpha_{LOG} = \log_{10}(\alpha) = 1.08$ may be an appropriate descriptor.

Asymmetric carbon atoms can be recognized from the automorphism data as follows: The number of (single) bonds must be three (with one carbon-hydrogen bond) or four (quaternary carbon atom), and all bonds must belong to different bond orbits. As descriptors are suggested the absolute number of asymmetric carbon atom, n_{ASYM}, and the fraction $f_{ASYM} = n_{ASYM} / n_A$.

A number of measures have been defined for characterizing the distribution of the sizes of atom and bond orbits. Let a_i be the number of atoms in atom orbit i (i = 1 ... k_A), with k_A for the **number of atom orbits**. For the structure in Figure 1 we have $k_A = 4$ (orbits **A** to **D**; i = 1 to 4) with $a_1 = 3$, $a_2 = 2$, $a_3 = 1$ and $a_4 = 1$. Analogously, we have for the $k_B = 3$ bond orbits (**X**, **Y**, **Z**) the number of bonds per bond orbit $b_1 = 3$, $b_2 = 2$, and $b_3 = 1$.

The number of atom orbits, k_A (bond orbits, k_B) of any graph with n_A atoms and n_B bonds is between one and n_A (n_B). The maximum number of orbits is reached for **asymmetric graphs** with each atom (bond) in a separate orbit. The smallest asymmetric alkane is 2-ethyl-pentane, C_7H_{16}, shown in Figure 2. Each of the seven atoms is in a separate atom orbit; each of the six bonds is in a separate bond orbit; the size of the automorphism group is one. In summary we obtain for asymmetric trees $k_A = n_A$; all $a_i = 1$; α = 1. The alkane isomers $C_{4\text{-}20}$ contain 28597 (4.63 %) asymmetric graphs.

The high symmetry extreme with all atoms in one atom orbit is not possible for alkane trees; however, graphs for rings, a tetrahedron, a cube, and others have all vertices in a single orbit. Also complete graphs (each vertex is connected to all other vertices - not relevant for chemical structures) have only one vertex orbit; in these graphs is $\alpha = n!$, the maximum for graphs with n vertices. In the 618047 alkane isomers considered in this work the structure shown

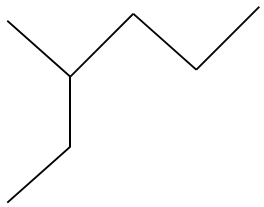

Figure 2. The smallest asymmetric tree has seven vertices and represents the chemical structure of 2-ethyl-pentane, C_7H_{16}. Each atom (vertex) is in a separate atom orbit; each bond (edge) in a separate bond orbit; the size of the automorphism group is one.

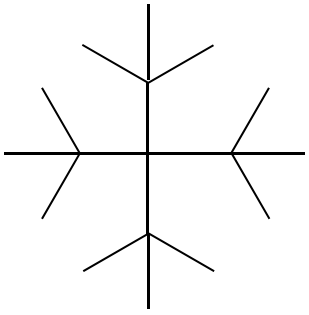

Figure 3. The largest automorphism group in the $C_{4\text{–}20}$ alkane isomers has tetra-tert-butyl-methane, $C_{17}H_{36}$, with α = 31104 mappings. This number can be calculated from the number of permutations of the four methyl groups (each 3!) and of the central carbon atom (4!) by 3! × 3! × 3! × 3! × 4!.

in Figure 3 has the largest automorphism group with α = 31104 mappings. It is the highly symmetric structure tetra-tert-butyl-methane, $C_{17}H_{36}$.

An **entropy** measure (structural information content) for atom orbits has been defined by Dehmer et al.[16] as

$$E_A = -\sum_{i=1}^{k_A} \left[(a_i / n_A) \log_2 (a_i / n_A) \right] \quad (1)$$

The so-called **symmetry index** for atom orbits has been defined by Mowshowitz et al.[17] as

$$S_A = (1 / n_A) \sum_{i=1}^{k_A} \left[a_i \log_2 (a_i) \right] + \log_2 (\alpha) \quad (2)$$

Analogous measures can be defined for the bond orbits as E_B and S_B. Table 2 contains the values of E and S for the structures shown in Figures 1 to 3. Note that for asymmetric graphs $E_A = 0$, and $S_A = \log_2(n_A)$.

The **orbit polynomial** has been defined by using the frequencies of the different orbit sizes as coefficients, and the root of this polynomial has been suggested as a molecular descriptor characterizing the structural symmetry.[18] Let's denote the different orbit sizes of a structure by g_j with j = 1 ... n_g (n_g is the number of different orbit sizes), and the frequencies of the different orbit sizes by h_j (j = 1 ... n_g). We obtain the orbit polynomial by

$$1 - \sum_{j=1}^{n_g} h_j z^{g_j} \quad (3)$$

We have to solve the equation $1 - \sum h_j z^{g_j} = 0$ to obtain the real and positive root δ. Investigation of the mathematical properties of δ proved that δ is less or equal one.[18] The structure in Figure 1 has four atom orbits with sizes 3, 2, 1, and 1; the maximum orbit size is n_g = 3; for the orbit sizes (g_j) 1, 2, 3 we have the frequencies (h_j) 2, 1, 1. The corresponding orbit polynomial, and the resulting equation, is therefore

$$1 - (2z^1 + 1z^2 + 1z^3) = 0 \quad (4)$$

We use the notation δ_A for atom orbit data and δ_B for bond orbit data. For the structure in Figure 1 is δ_A = 0.393, equivalent to the root of polynomial (4); for the bond orbit data we obtain δ_B = 0.544. The roots of polynomials have been calculated by the function **polyroot** provided in the programming environment R.[12]

The minimum of δ_A for a structure with n_A atoms appears for asymmetric graphs (atoms in separate orbits; n_g = 1; g_1 = 1, $h_1 = n_A$); the orbit polynomial is $n_A z^1 = 1$; the root yields to $\delta_A = 1/n_A$. An example is the asymmetric structure in Figure 2 with n_A = 7, δ_A = 1/7 = 0.143. Bounds of δ_A for distinct classes of graphs have been reported.[19]

The maximum of δ_A appears if all atoms are in a single orbit with n_g = 1; $g_1 = n_A$; h_1 = 1; the orbit polynomial is $1z^{n_A} = 1$; the root δ_A = 1. Such structures are not among the alkanes as discussed above. In the 618047 alkane isomers $C_{4–20}$ the maximum of δ_A is 0.826, appearing for tetramethylbutane, C_8H_{18}, a highly symmetric and compact structure. The maximum value 1 for δ_B (all C–C bonds topologically equal) is present in one of the alkane isomers, in 2-methylpropane (isobutane, C_4H_{10}).

3. RESULTS

3.1. Size of Automorphism Groups

An overview of the used alkane isomers and the size of their automorphism groups is given in Table 3. The numbers of isomers, n_{ISO}, are identical to already published data,[3] reaching 366319 for $C_{20}H_{42}$. For alkane isomers with ≥7 carbon atoms (n_C) at least one has an asymmetric structure, the smallest is 2-ethyl-pentane, shown in Figure 2. For example, among the isomers of $C_{20}H_{42}$, $n(\alpha_{MIN})$ = 15641 are asymmetric, which is 4.3 % of the isomers in this set. For n_C 4 to 20, the maximum size of the automorphism groups, α_{MAX}, is between 6 and 31104; the largest value of α has tetra-tert-butyl-methane, $C_{17}H_{36}$, shown in Figure 3. In general, α_{MAX} appears only for one structure in an isomer set; the exception are the $C_{10}H_{22}$ isomers with two structures possessing α_{MAX}.

The number of different values of α in the isomer sets, u_α, is between 2 and 36, demonstrating a low discrimination ability (uniqueness).[20] For example, the 4347 isomers of $C_{15}H_{32}$ have values for α between 1 and 1296, however, only u_α = 20 unique values (1, 2, 4, 6, 8, 12, 16, 24, 32, 36, 48, 64, 72, 96, 128, 144, 288, 432, 576, 1296). The frequency distributions of α are highly asymmetric with medians, α_{MED}, between 2 and 8. Figure 4 shows the histograms of the frequencies of α for the isomer sets with 10, 15, and 20 carbon atoms.

3.2. Descriptors Based on the Distribution of Orbit Sizes

The **correlation** coefficients (Pearson) between the automorphism based descriptors entropy (E_A, E_B), symmetry index (S_A, S_B), root of orbit polynomial (δ_A, δ_B) and size of automorphism group as $\log_{10}(\alpha)$ are given in Table 4. The descriptors from atom orbit data are very highly correlated with the corresponding descriptors from bond orbit data; consequently, we omit E_B, S_B and δ_B from the further discussion.

The absolute correlation coefficients between E_A, S_A, δ_A and $\log(\alpha)$ are between 0.566 for S_A vs. δ_A and 0.998 for S_A vs. $\log_{10}(\alpha)$, indicating that different aspects of symmetry are characterized by these descriptors. The values in Table 4 have been calculated from all isomers and are dominated by the large sets with 19 and 20 carbon atoms. In Figure 5 selected correlation coefficients are shown separately for the isomer sets with 8 to 20 carbon atoms. We see only

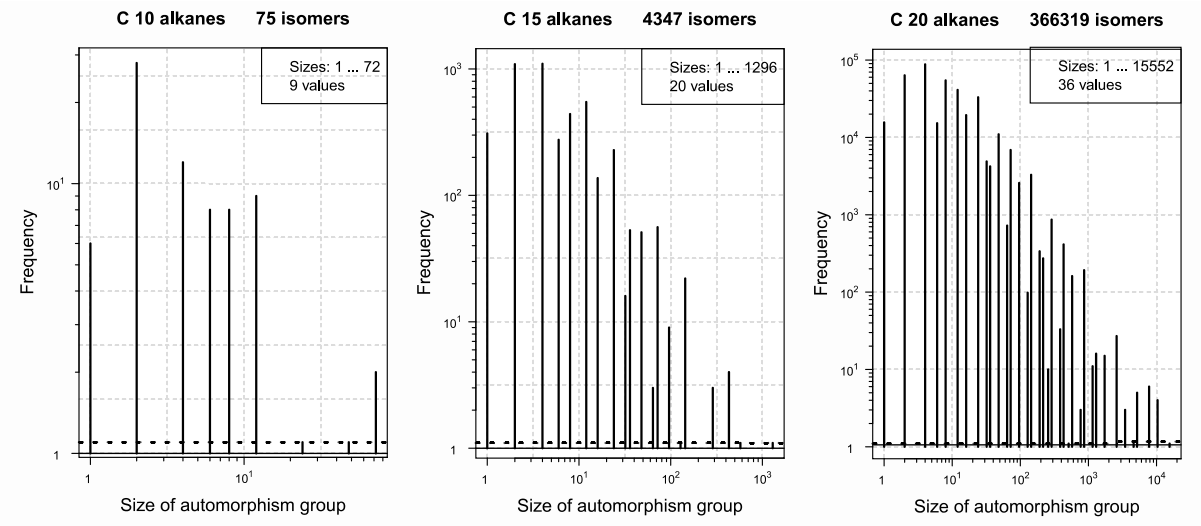

Figure 4. Histograms of the frequencies of α, the size of the automorphism groups, for isomer sets with 10, 15, and 20 carbon atoms. Both axes are logarithmic scaled; for better visualization, frequency 1 is plotted with value 1.2 (appearing 2, 3, and 3 times in the sets $C_{10}H_{22}$, $C_{15}H_{32}$ and $C_{20}H_{42}$, respectively).

weak dependence of the Pearson correlation coefficients on the size of the alkane molecules, thus indicating a stable relation between the descriptors for varying structure size. Remarkably high correlation coefficients between S_A and $\log_{10}(\alpha)$ for the considered tree graphs can be explained by small contributions of the first term in equation (2) for S_A, compared to the second term $\log_2(\alpha)$.

Table 3. Size of the automorphism groups for alkane isomers with 4 to 20 carbon atoms.

n_C	n_{ISO}	α_{MIN}	α_{MED}	α_{MAX}	$n(\alpha_{MIN})$	[%]	$n(\alpha_{MAX})$	u_α
4	2	2	4	6	1	[50.0%]	1	2
5	3	2	2	24	2	[66.7%]	1	2
6	5	2	2	8	3	[60.0%]	1	3
7	9	1	4	12	1	[11.1%]	1	6
8	18	1	3	72	1	[5.6%]	1	7
9	35	1	4	72	3	[8.6%]	1	9
10	75	1	4	72	6	[8.0%]	2	9
11	159	1	4	144	15	[9.4%]	1	12
12	355	1	4	144	29	[8.2%]	1	13
13	802	1	4	1296	67	[8.4%]	1	15
14	1858	1	4	1296	139	[7.5%]	1	18
15	4347	1	4	1296	309	[7.1%]	1	20
16	10359	1	4	2592	661	[6.4%]	1	24
17	24894	1	4	31104	1462	[5.9%]	1	27
18	60523	1	4	10368	3187	[5.3%]	1	28
19	148284	1	6	10368	7076	[4.8%]	1	33
20	366319	1	8	15552	15641	[4.3%]	1	36

n_C, number of carbon atoms; n_{ISO}, number of isomers; α, size of automorphism group, with α_{MIN}, α_{MED}, α_{MAX} for minimum, median, and maximum of α, resp., $n(\alpha_{MIN})$, number of isomers with α_{MIN}, [%] in percent of n_{ISO}; $n(\alpha_{MAX})$, number of isomers with α_{MAX}; u_α, number of different values of α in the isomer sets with n_C carbon atoms.

The **distributions** of the automorphism based descriptors are characterized by boxplots in Figure 6. The values for the entropy (E_A) increase with an increasing number of carbon atoms, while the symmetry index (S_A), the root of the orbit polynomial (δ_A), and the logarithm of the size of the automorphism group ($\log_{10}(\alpha)$) show only a small or no dependence on the number of carbon atoms. All distributions exhibit a tailing, E_A on the low value side, the others on the high value side.

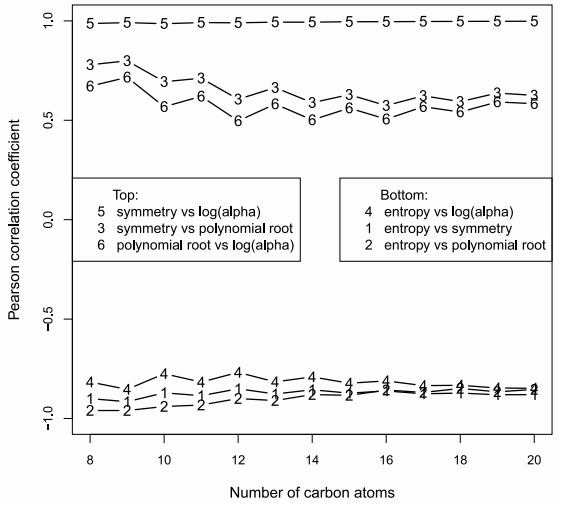

Figure 5. Pearson correlation coefficients between descriptors based on automorphism data for sets with the alkane isomers with 8 to 20 carbon atoms. The descriptors considered are E_A (entropy), S_A (symmetry index), δ_A (root of orbit polynomial), $\log_{10}(\alpha)$ (logarithm of the size of the automorphism group).

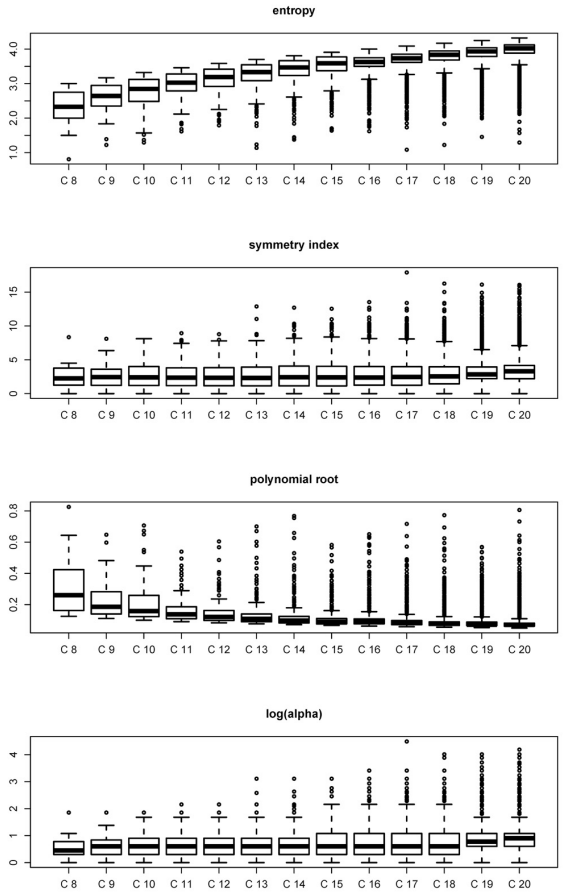

Figure 6. Distribution of the automorphism based descriptors entropy (E_A), symmetry index (S_A), root of the orbit polynomial (δ_A), and size of the automorphism group (number of mappings, as $\log_{10}(\alpha)$). Boxplots C8 to C20 are for the alkane isomer sets with 8 to 20 carbon atoms.

For the alkane isomer sets C_4 to C_{20} Table 5 contains the lowest (*L*) and highest (*H*) values of the number of atom orbits, k_A, and of the descriptors E_A, S_A and δ_A. Columns *u* contain the number of unique values of the descriptors in an isomer set, indicating an only low discrimination power. E. g., the 4347 isomers of $C_{15}H_{32}$ exhibit only 43 different values for E_A, 92 for S_A, and 48 for δ_A; k_A is between 4 and 15 with 12 unique values. In general the highest uniqueness is reached by the symmetry index (S_A). Only 1247 different values appear (rounded to 5 decimals) for the total of 618047 isomers.

The **number of** atom **orbits**, k_A, and the number of bond orbits, k_B, are closely related. The complete set of 618047 $C_{4–20}$ alkane isomers contains 878 structures with $k_A = k_B$; all these structures have an even number of carbon atoms (n_C) and a highly symmetric shape. For all other alkanes $k_A = k_B + 1$. These relations are discussed together with the concept of a **symmetric bond** (also named symmetric edge or exceptional line) by Lygeros et al.[21] based on previous work by Harary[22,23] and Read.[24] A symmetric bond is present if the removal of the bond cuts the graph into two isomorphic subgraphs. An equivalent definition is "the atoms (vertices) forming a symmetric bond (edge) must be topological equivalent" - hence must belong to the same atom (vertex) orbit. Within the alkane graphs a symmetric bond can only be present if n_C is even and it has been claimed that $k_A = k_B$.[21] The present numerical investigations show that alkane structures may contain a symmetric bond if k_A is not equal to k_B. In Figure 7 two of the 18 isomers of C_8H_{18} are shown having a symmetric bond; for structure **V** $k_A = k_B = 3$, while for structure **W** $k_A = 5$ and $k_B = 4$. Actually, six isomers of C_8H_{18} have a symmetric bond but in only four of them is $k_A = k_B$. Table 6 contains the results for the alkane isomers with 4, 6, … , 20 carbon atoms. All considered structures with $k_A = k_B$ have a symmetric bond, however, additional structures also posses a symmetric bond. For instance in the 366319 C_{20}-alkane isomers 2115 (0.58 %) have a symmetric bond, with 507 of them exhibiting $k_A = k_B$.

The smallest alkane with an **asymmetric carbon atom** (see column n_{ASYM}) is 2-ethyl-pentane, C_7H_{16}, Figure 2. The maximum number of eight asymmetric carbon atoms in a structure appears in one of the isomers of $C_{20}H_{42}$; it is 3,4,5,6,7,8,9,10-octamethyl-dodecane with $\alpha = 2$, $k_A = k_B = 10$, where the asymmetric carbon atoms are pairwise topologically equal.

Table 4. Pearson correlation coefficients between descriptors based on automorphism data for alkane graphs with 4 to 20 carbon atoms (n = 618047 structures).

	E_A	E_B	S_A	S_B	δ_A	δ_B	$\log(\alpha)$
E_A	1.000	1.000	–0.774	–0.776	–0.853	–0.863	–0.736
E_B	1.000	1.000	–0.773	–0.774	–0.848	–0.860	–0.785
S_A	–0.774	–0.773	1.000	1.000	0.566	0.574	0.998
S_B	–0.776	–0.774	1.000	1.000	0.568	0.576	0.997
δ_A	–0.853	–0.848	0.566	0.568	1.000	0.990	0.917
δ_B	0.863	–0.860	0.574	0.576	0.990	1.000	0.526
$\log(\alpha)$	–0.736	–0.785	0.998	0.997	0.917	0.526	1.000

E_A, E_B, entropy measure for atom and bond orbits, resp.; S_A, S_B, symmetry index for atom and bond orbits, resp.; δ_A, δ_B, positive real root of orbit polynomial for atom and bond orbits, resp.; α, size of automorphism group.

Table 5. Descriptors based on automorphism data for alkane graphs with 4 to 20 carbon atoms.

n_C	n_{ISO}	n_{ASYM}	k_A			E_A			S_A			δ_A		
			L	*H*	*u*	*L*	*H*	*u*	*L*	*H*	*u*	*L*	*H*	*u*
4	2	0	2	2	1	0.81	1.00	2	2.00	3.77	2	0.68	0.71	2
5	3	0	2	4	3	0.72	1.92	3	1.40	6.18	3	0.30	0.72	3
6	5	0	2	5	4	0.92	2.25	5	1.33	4.67	5	0.24	0.79	5
7	9	1	3	7	5	1.38	2.81	7	0.00	4.55	8	0.14	0.59	7
8	18	2	2	8	7	0.81	3.00	11	0.00	8.36	13	0.12	0.83	11
9	35	2	3	9	7	1.22	3.17	12	0.00	8.12	16	0.11	0.65	12
10	75	2	3	10	8	1.30	3.32	18	0.00	8.12	24	0.10	0.71	19
11	159	3	4	11	8	1.62	3.46	20	0.00	8.94	33	0.09	0.54	21
12	355	4	4	12	9	1.79	3.58	24	0.00	8.80	41	0.08	0.61	26
13	802	4	3	13	11	1.14	3.70	32	0.00	12.90	56	0.08	0.70	34
14	1858	4	3	14	12	1.38	3.81	38	0.00	12.72	72	0.07	0.77	43
15	4347	5	4	15	12	1.64	3.91	43	0.00	12.56	92	0.07	0.58	48
16	10359	6	4	16	13	1.62	4.00	55	0.00	13.55	122	0.06	0.65	65
17	24894	6	3	17	15	1.09	4.09	62	0.00	17.93	158	0.06	0.72	75
18	60523	6	3	18	16	1.22	4.17	69	0.00	16.29	189	0.06	0.77	90
19	148284	7	4	19	16	1.46	4.25	83	0.00	16.13	253	0.05	0.57	107
20	366319	8	3	20	18	1.30	4.32	99	0.00	16.09	303	0.05	0.81	131
All	618047	8	2	20	19	0.72	4.32	530	0.00	17.93	1247	0.05	0.83	625

n_C, number of carbon atoms; n_{ISO}, number of isomers; n_{ASYM}, maximum number of asymmetric carbon atoms in a structure; k_A, number of atom orbits in a structure; E_A, entropy (information content), equation (1) for atom orbits; S_A, symmetry index, equation (2) for atom orbits; δ_A, root of atom orbit polynomial, equation (3). *L* and *H* stand for lowest and highest value in an isomer set; *u* is the number of unique values (rounded to 5 decimals) in an isomer set.

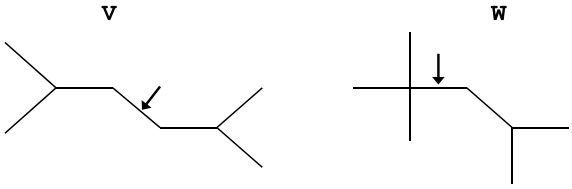

Figure 7. Two isomers of the alkanes C_8H_{18} containing a symmetric bond (marked by an arrow). In structure **V** the numbers of atom orbits (k_A) and bond orbits (k_B) are both three; in structure **W**, $k_A = 5$ and $k_B = 4$.

3.3. Relation With the Presence of Substructures

Relations between the values of automorphism based descriptors (section 3.2) and the presence of certain substructures are discussed for the 4347 isomers of the C_{15}-alkanes. The substructures considered are the trees with 3 to 8 carbon atoms (equivalent to the 38 isomers of the corresponding alkanes). Regarding a particular substructure, the 4347 molecular structures can be divided into a class **1** if the substructure is present, and a class **0** if absent. If the distributions of a descriptor are different for class **1** and **0**, the descriptor is characteristic for the presence/ absence of the substructure.

In Figure 8 two substructures, each with eight carbon atoms are considered: substructure **S1** (2,2,3,3-tetramethylbutane) is symmetric and compact, and is present in 12.4 % of the C_{15}-alkanes; substructure **S2** (n-octane) is present in 87.6 %. The distributions for three descriptors in class **0** and **1** are presented as boxplots: entropy (E_A), symmetry index (S_A), and root of the orbit polynomial (δ_A). In general the distributions of class **1** and **0** are well separated with *p*-value of Mann-Whitney-u-tests < 0.001 in all cases, with some outliers appearing. For the compact substructure **S1** the values for S_A and δ_A are significantly higher in class **1** (substructure present) than in class **0**. In contrary, E_A is smaller in class **1** than in class **0** - this corresponds to the negative correlation coefficients between E_A and S_A or δ_A with values of –0.774 and –0.863, respectively (considering all 618047 structures, see Table 4). The chain substructure

Table 6. Symmetric bonds in alkane isomers.

n_C	n_{ISO}	n_{SYM}	n_{SYM} %	n_{EQU}
4	2	1	50.00	1
6	5	3	60.00	2
8	18	6	33.33	4
10	75	17	22.67	8
12	355	40	11.27	17
14	1858	114	6.14	39
16	10359	290	2.80	89
18	60523	801	1.32	211
20	366319	2115	0.58	507

n_C, number of carbon atoms; n_{ISO}, number of isomers; n_{SYM}, number of isomers with a symmetric bond; n_{SYM} %, n_{SYM} in percent of n_{ISO}; n_{EQU}, number of isomers with a symmetric bond and equal numbers of atom orbits and bond orbits ($n_{EQU} \le n_{SYM}$).

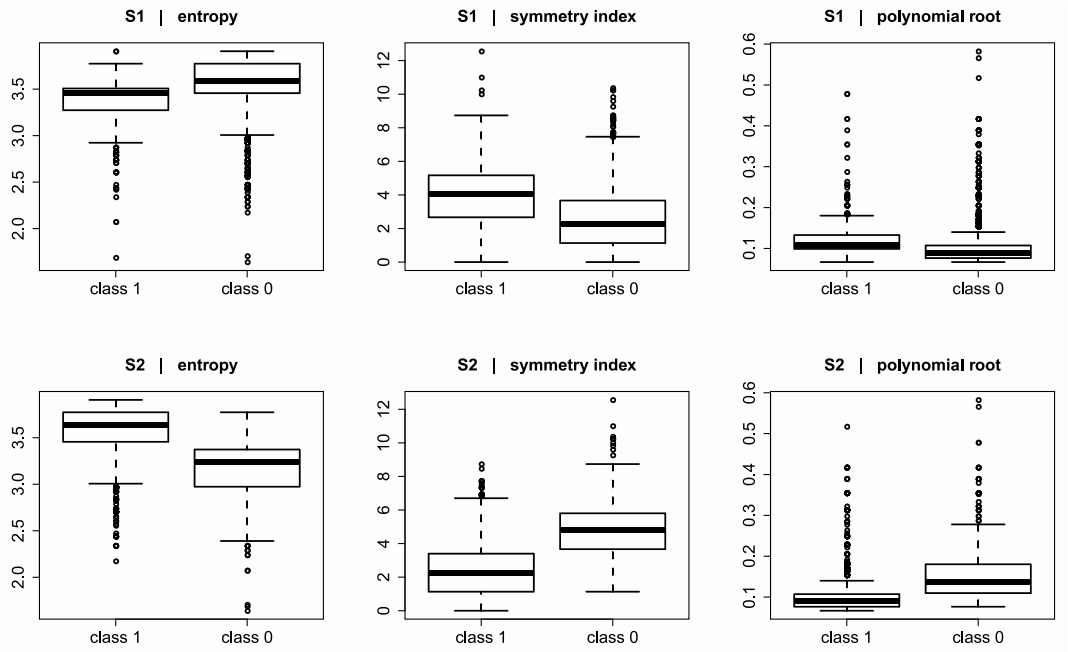

Figure 8. The 4347 isomers of the C_{15}-alkanes are divided into class **1** and **0** according to presence or absence of a substructure. The boxplots show for three molecular descriptors the distributions for class **1** and **0**. Substructure **S1** is 2,2,3,3-tetramethyl-butane, **S2** is n-octane. The molecular descriptors are entropy (E_A), symmetry index (S_A), and root of the orbit polynomial (δ_A).

S2 shows an opposite behavior with lower values for S_A and δ_A in class **1**, and higher values for E_A. The significant discrimination of class **0** and **1** by the descriptors E_A, S_A and δ_A appears in the alkane isomer sets $C_{12–20}$ with p-values < 0.001 in u-tests. These results demonstrate a relationship between the values of the three descriptors and the presence/absence of substructures **S1** and **S2** which mainly differ in their compactness.

In the considered 38 substructures we have 12 containing a quaternary carbon atom; presence/absence of these substructures have a marked influence on the values of the descriptors E_A, S_A and δ_A. For example, the smallest of the 12 substructures is tetramethyl-methane, **S3**. For all isomer sets $C_{12–20}$ we obtain E_A being significantly smaller in class **1** (**S3** present) than in class **0**. In contrast S_A and δ_A, are significantly larger in class **1** than in class **0**. The p-values of u-tests applied for E_A, S_A and δ_A using the 12 substructures are < 0.001 in about 95 % of the 108 cases (12 substructures times 9 isomer sets). This result again reflects the close relationship between the described molecular descriptors - based on automorphism data - and the symmetry of chemical structures.

3.4. Application in QSPR Models

Linear, multivariate models for QSPR (quantitative structure-property relationships) have been made by using two groups of descriptors (x-variables): group ***A*** containing eight descriptors based on automorphism data (as described in section 2), and group ***B*** containing 38 binary substructure descriptors defined by the alkane isomers with 3 to 8 carbon atoms (as used in section 3.3). Group ***A*** consists of the following x-variables: k_A, number of atom orbits; n_{ASYM} and f_{ASYM}, number and fraction of asymmetric carbon atoms; α and log(α), size and its decadic logarithm of the automorphism group; S_A, symmetry index; E_A, entropy; and δ_A, root of the orbit polynomial; the last three descriptors for atom orbits. We compare the variable sets ***A***, ***B*** and ***A*** together with ***B***. The four chemical structure sets used are random samples with 1000 alkane isomers from each of the sets with 14 to 17 carbon atoms.

The three molecular properties modeled by the x-variables are: y_1, the approximate surface area of the molecule (Angström^2, code *ASA*); y_2, the solubility in water (logarithm of mol/L, code log S); and y_3, the octanol/water partition coefficient (logarithm of the concentration ratio, code log P). These properties are estimated by specific methods from chemoinformatics as implemented in the software CORINA Symphony.[25] The calculation of *ASA* is based on the geometry of partially overlapping van der Waals surfaces of the atoms of a molecule.[26] For log S the approximated 3D molecular structures and a set of eight physicochemical descriptors are used with multiple linear regression and neural networks.[27,28] The property log P has been derived by summing appropriate contributions of the atoms together with suitable correction factors.[29] The strategy used here for creating QSPR models is based on standard chemometrics as follows:[30–33]

(1) The applied variable selection consists of two procedures: First, variables are deleted that are constant or almost constant (meaning the same value in all but a maximum of ten objects). Second, a stepwise selection in forward and backward direction is applied using the Bayes information criterion (BIC) as performance measure.[32,34]

(2) Partial least-squares (PLS) regression with repeated double cross validation (rdCV) is applied to the autoscaled matrix $\boldsymbol{X}$ (variables are mean-centered and scaled by the standard deviations). The rdCV approach[35–37] estimates the optimum model complexity (given by the number of PLS components, A_{OPT}) separately from estimating the prediction performance for new objects.[38] Furthermore, the variability of the performance criteria is characterized by repeated random splits into calibration and test sets. The essential parameters used for rdCV are as follows: the numbers of segments in outer and inner loop are 3 and 7, respectively; the number of repetitions is 50, resulting in 50 test-set predictions $\hat{y}$ for each object, and 150 estimations of the optimum number of PLS components, with the most frequent value taken as the final A_{OPT}.

(3) Estimations of the prediction performance are derived from the prediction errors (residuals) $e_i = y_i - \hat{y}_i$ from test-set predictions during the rdCV. Because these residuals are approximately normally distributed with a mean (bias) near zero, the standard deviation of e_i (often called standard error of prediction, SEP) is a useful measure with $\hat{y}_i \pm 2$ SEP defining a 95 % tolerance interval for predictions. For the comparison of models with different numbers of variables and different magnitudes of y, the adjusted squared correlation coefficient between y and $\hat{y}$ is appropriate[32]

$$ {}_{ADJ}R^2 = 1 - (n-1)(1-R^2) / (n-m-1) \qquad (5)$$

with n and m for the number of objects and variables, respectively, and R^2 for the squared Pearson correlation coefficient between y and $\hat{y}$. The measure ${}_{ADJ}R^2$ is independent from the units of y and penalizes models with large m.

First, results of modeling the property *ASA* with data from the C_{15}-alkane isomers are discussed (A), and then a summary is given of the properties and performances of the models made for all three properties and all four isomer sets defined in this section (B).

(A) A random sample with n = 1000 structures from the 4347 isomers of the C_{15}-alkanes is selected. The y-values *ASA* of this set are between 431.0 and 442.3, with a standard deviation of 6.07; note that only 20 different values for y (rounded to 3 decimals) appear.

Using the variable set ***AB***, containing 8 automorphism-based descriptors and 38 binary substructure descriptors, we have m_0 = 46 variables and a matrix $\boldsymbol{X}_0$ (1000 × 46). The first step of the variable selection eliminates 10 variables that are constant or almost constant. In the following stepwise variable selection m = 14 variables are retained for PLS modeling. Ten of the selected variables are binary substructure descriptors (alkane structures with 7 or 8 carbon atoms), and four are based on automorphism data (n_{ASYM}, log(α), S_A, δ_A).

The matrix $\boldsymbol{X}$ (1000 × 14) is autoscaled and the strategy PLS-rdCV gives an estimated optimum number of PLS components of five, and a SEP of 1.88 (equal to 31 % of the standard deviation of y). For a characterization of the prediction performance we consider ${}_{ADJ}R^2$ = 0.903 between the given y and the medians of 50 test-set predicted $\hat{y}$-values (right-hand side plot in Figure 9). The performance of QSPR models from using only automorphism based descriptors (***A***), is poor with ${}_{ADJ}R^2$ = 0.616 (left-hand side plot), from only binary substructure descriptors (***B***), mid plot, is better with ${}_{ADJ}R^2$ = 0.748, however, is clearly enhanced to 0.903 by combining ***A*** and ***B***.

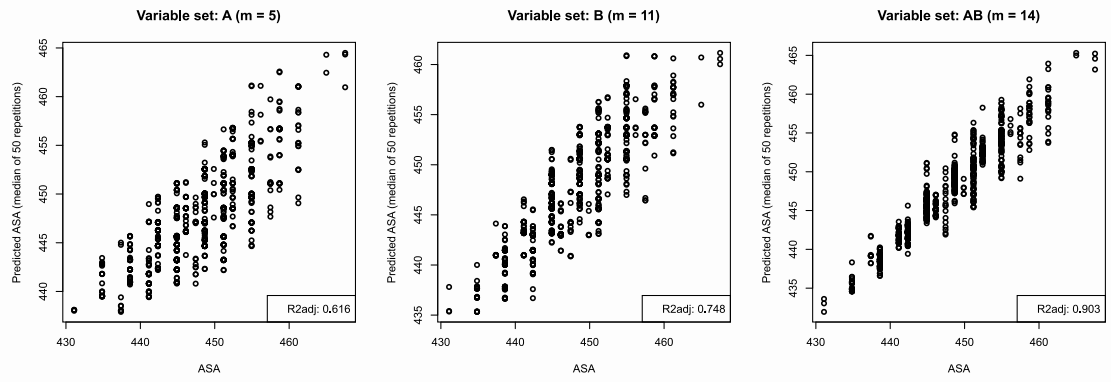

Figure 9. QSPR models for the property *ASA* (y) resulting from the method PLS-rdCV. Data set is a random sample with n = 1000 structures from the C_{15}-alkane isomers. Three variable sets are compared: ***A*** with 8 automorphism based descriptors; ***B*** with 38 binary substructure descriptors; and ***AB***, both together. After variable selection m = 5, 11, and 14 variables, respectively, remain for PLS modeling. The plots show the given *ASA* value versus the predicted value (median of test-set predictions in 50 repetitions in the rdCV procedure).

Table 7. QSPR models.

n_C	y	A			B			AB		
		$_{ADJ}R^2$	m	A_{OPT}	$_{ADJ}R^2$	m	A_{OPT}	$_{ADJ}R^2$	m	A_{OPT}
14	ASA	0.633	5	5	0.792	14	2	0.921	14	5
14	logP	0.429	4	3	0.804	12	3	0.817	17	3
14	logS	0.418	5	3	0.816	13	3	0.831	17	3
15	ASA	0.616	5	5	0.748	11	2	0.903	14	5
15	logP	0.425	4	3	0.744	10	2	0.797	15	5
15	logS	0.435	4	4	0.754	12	2	0.814	18	5
16	ASA	0.614	5	5	0.718	13	3	0.892	13	7
16	logP	0.429	4	3	0.710	11	3	0.782	17	7
16	logS	0.417	4	3	0.724	12	3	0.794	18	7
17	ASA	0.621	4	4	0.666	9	2	0.888	13	6
17	logP	0.425	4	2	0.674	11	2	0.757	12	6
17	logS	0.413	4	2	0.694	11	2	0.777	15	7

n_C, number of carbon atoms of the used alkane isomers; y, property modeled by the descriptors; A, B, AB, descriptor (variable) set; $_{ADJ}R^2$, adjusted squared Pearson correlation coefficient between y and $\hat{y}$; m, number of variables after variable selection and used for PLS regression; A_{OPT}, optimum number of PLS components obtained by the PLS-rdCV strategy.

(B) Table 7 summarizes the results obtained for the three considered properties and the four alkane isomer sets. The measure $_{ADJ}R^2$ is always higher for the descriptor set ***B*** than for the descriptor set ***A***; however, combining ***A*** and ***B*** is always better than ***B***. The prediction performance decreases from using C_{14} alkane isomers to C_{17} isomers; for instance for descriptor set ***AB*** and property *ASA*, the values for $_{ADJ}R^2$ range from 0.921 (C_{14}) to 0.888 (C_{17}). The number of variables (descriptors), m, after variable selection is between 12 and 18, the number of optimum components in PLS regression, A_{OPT}, is between 2 an 7. Modeling the property *ASA* is fairly good, however, for log*P* and log*S* only semi-quantitative PLS models are possible with the used variables.

Finally, we discuss which descriptors are mostly selected from the set ***AB*** in the 12 jobs (3 properties times 4 groups of alkane isomers). Two binary substructure descriptors are selected in all 12 jobs: both are highly symmetric: 2,2,3,3-tetramethyl-butane (substructure **S1** in section 3.3), and 2,3,4-trimethyl-pentane. The importance of binary variables j is connected to their information entropy $H_j = -p_j \log_2(p_j) - (1\text{-}p_j) \log_2(1-p_j)$ with p_j for the probability of a descriptor value '1'. Actually, the nine descriptors selected in > 60 % of the 12 jobs have an entropy between 0.599 an 0.999. On the other hand, a high entropy is not always connected with a frequent selection.

From the eight descriptors based on automorphism data, the number of asymmetric carbon atoms, n_{ASYM}, is always selected; the decadic logarithm of the size of the automorphism group, log(α), and the root of the orbit polynomial, δ_A, are selected in 83 %; and the symmetry index, S_A, in 50 %. The selection of automorphism based descriptors - in the presence of binary substructure descriptors - demonstrates their potential utility in QSPR models.

4. SUMMARY

For alkanes with 4 to 20 carbons atoms all isomers are created, ranging from 2 isomers for C_4H_{10} to 366319 isomers for $C_{20}H_{42}$, in total n_{ALL} = 618047 chemical structures. The atom-bond connectivity of these structures is represented in terms of graph theory by uncolored trees with vertex degrees between 1 and 4. For all isomers the complete automorphism groups are computed and evaluated.

The size of the automorphism group (α, number of mappings of the graph onto itself) is one or an even number; the maximum 32104 is present for the highly symmetric structure of tetra-isobutyl-methane, $C_{17}H_{36}$. The uniqueness of α is low with only 37 different values in the n_{ALL} isomers. The number of asymmetric graphs (α = 1) is 28597 (4.63 % of n_{ALL}). Most of the alkane structures contain at least one asymmetric carbon atom (97.6 %); the maximum per structure is eight, present in one isomer of $C_{20}H_{42}$.

The relation between the number of atom orbits, k_A, and the number of bond orbits, k_B, is $k_A = k_B + 1$ in 99.86 % of the n_{ALL} alkanes, and $k_A = k_B$ for the rest of 878 structures. All 878 structures with $k_A = k_B$ contain a symmetric bond; however, additional 2509 structures with $k_A = k_B + 1$ also have a symmetric bond.

The logarithm of α is highly correlated with the symmetry index S_A and with the polynomial root δ_A, exhibiting Pearson correlation coefficients (R) for all n_{ALL} alkanes of 0.998 and 0.917, respectively. Entropy E_A and symmetry measure S_A are negatively correlated (R = −0.774). The correlation coefficients between the symmetry descriptors have only a weak dependence of the size of the alkane molecules.

The uniqueness of the descriptors E_A, S_A and δ_A is low with the numbers of different values (rounded to five decimals) in the n_{ALL} alkanes of 530, 1247, and 625 (0.086, 0.202, and 0.101 % of n_{ALL}), respectively.

The descriptors E_A, S_A and δ_A discriminate well the presence or absence of certain substructures in alkanes. For instance the presence of substructure $(CH_3)_3C–$ gives low values for E_A, and high values for S_A and δ_A, compared to alkanes not containing this substructure.

The descriptors n_{ASYM}, log(α), S_A and δ_A were successfully applied in QSPR models - together with binary substructure descriptors - for the prediction of molecular properties of alkanes by linear PLS regression.

For the studied set of alkanes, we conclude that descriptors based on complete automorphism data are useful complements to other descriptors, for instance for QSPR models or structure similarity searches. An extension of this concept for general chemical structures - represented by fully colored graphs - is in work.

Acknowledgments. The authors thank A. Kerber, R. Laue (University Bayreuth) and M. Meringer (German Aerospace Center, DLR) for providing the software MOLGEN. The work was supported by the Austrian Science Fund FWF, project P30031.

Appendix (symbols)

α	size of automorphism group (|Aut(G)|), number of automorphism mappings
a_i	number of atoms in atom orbit i (i = 1 … k_A)
A_{OPT}	optimum number of PLS components
b_i	number of bonds in bond orbit i (i = 1 … k_B)
δ	positive real root of orbit polynomial; δ_A for atom orbits, δ_B for bond orbits
E	entropy (information content); E_A for atom orbits, E_B for bond orbits
G	graph
h_j	frequency of orbits with size j (j = 1 … n_g)
k	number of orbits; k_A for atom orbits, k_B for bond orbits
n_A	number of atoms (vertices)
n_{ASYM}	number of asymmetric carbon atoms
n_B	number of bonds (edges)
n_C	number of carbon atoms
n_{EQU}	number of structures (within a set of isomers) with $k_A = k_B$
n_{ISO}	number of isomers
n_g	maximum orbit size in a structure (graph) G
n_{SYM}	number of structures (within a set of isomers) containing a symmetric bond
$_{ADJ}R^2$	adjusted squared Pearson correlation coefficient
S	symmetry index; S_A for atom orbits, S_B for bond orbits
SEP	standard error of prediction
u	number of unique values in an isomer set; e. g., of E, S, δ
z	variable in the orbit polynomial

REFERENCES

[1] *Chemoinformatics - Basic concepts and methods* (Eds.: T. Engel, J. Gasteiger), Wiley VCH, Weinheim, Germany, **2018**.

[2] Chemical graph theory - Introduction and fundamentals (Eds.: D. Bonchev, D. H. Rouvray), Abacus Press, Amsterdam, The Netherlands, **1991**.

[3] N. Trinajstić, *Chemical graph theory*, CRC Press, Boca Raton, FL, USA, **1992**.

[4] K. Balasubramanian, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 621–626. https://doi.org/10.1021/ci00019a022

[5] S. Bohanec, M. Perdih, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 719–726. https://doi.org/10.1021/ci00015a010

[6] M. Razinger, K. Balasubramanian, M. E. Munk, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 197–201. https://doi.org/10.1021/ci00012a003

[7] A. Kerber, *MATCH Commun. Math. Comput. Chem.* **2018**, *80*, 733–744.

[8] M. Meringer, H. J. Cleaves, *Phil. Trans. R. Soc. A* **2017**, *375*, 20160344. https://doi.org/10.1098/rsta.2016.0344

[9] *MOLGEN*, *Molecular structure generator*, version 3.5., Benecke C., Grüner T., Grund R., Hohberger R., Kerber A., Laue R., Wieland T., University of Bayreuth, http://www.molgen.de (accessed Aug 31, 2021), Bayreuth, Germany, **1997**.

[10] *SubMat*, *Software for substructure searches in chemistry*, Scsibrany H., Varmuza K., Vienna University of Technology, http://www.lcm.tuwien.ac.at (accessed Aug 31, 2021), Vienna, Austria, **2004**.

[11] K. Varmuza, W. Demuth, M. Karlovits, H. Scsibrany, *Croat. Chem. Acta* **2005**, *78*, 141–149.

[12] *R*, A language and environment for statistical computing, R Development Core Team, Foundation for Statistical Computing, http://www.r-project.org (accessed Aug 31, 2021), Vienna, Austria, **2021**.

[13] B. D. McKay, A. Piperno, *J. Symb. Comput.* **2014**, *60*, 94–112. https://doi.org/10.1016/j.jsc.2013.09.003

[14] G. Csardi, T. Nepusz, V. Traag, S. Horvat, F. Zanini, *igraph reference manual*, https://igraph.org/c/doc/igraph-docs.pdf (accessed Aug 31, 2021), Cambridge, MA, USA, **2020**.

[15] I. Krasikov, A. Lev, D. T. Bhalchandra, *Discrete Math.* **2002**, *256*, 489–493. https://doi.org/10.1016/S0012-365X(02)00393-X

[16] M. Dehmer, A. Mowshowitz, *Inf. Sci.* **2011**, *181*, 57–78. https://doi.org/10.1016/j.ins.2010.08.041

[17] A. Mowshowitz, M. Dehmer, *Symmetry: Culture and Science* **2010**, *21*, 321–327.

[18] M. Dehmer, Z. Chen, F. Emmert-Streib, A. Mowshowitz, K. Varmuza, L. Feng, H. Jodlbauer, Y. Shi, J. Tao, *IEEE Access* **2020**, *8*, 36100–36112. https://doi.org/10.1109/ACCESS.2020.2970059

[19] M. Dehmer, F. Emmert-Streib, A. Mowshowitz, A. Ilić, Z. Chen, G. Yu, L. Feng, M. Ghorbani, K. Varmuza, J. Tao, *Appl. Math. Comput.* **2020**, *380*, 1–14. https://doi.org/10.1016/j.amc.2020.125239

[20] M. V. Diudea, A. Ilić, K. Varmuza, M. Dehmer, *Complexity* **2011**, *16*, 32–39. https://doi.org/10.1002/cplx.20363

[21] N. Lygeros, P.V. Marchand, M. Massot, *J. Symb. Comput.* **2005**, *40*, 1225–1241. https://doi.org/10.1016/j.jsc.2004.04.009

[22] F. Harary, *Graph theory*, Addison-Wesley Publishing Company, Reading, MA, USA, **1969**. https://doi.org/10.21236/AD0705364

[23] F. Harary, R.Z. Norman, *Proc. Am. Math. Soc.* **1960**, 11, 332–334. https://doi.org/10.1090/S0002-9939-1960-0111699-6

[24] R. C. Read in *Graph theory and applications* (Eds.: Y. Alavi, D. R. Lick, A. T. White), Springer-Verlag, Berlin, Germany, **1972**, pp. 243–259. https://doi.org/10.1007/BFb0067377

[25] C*ORINA-Symphony*, *Chemoinformatics program package*, Molecular Networks GmbH & Altamira LLC, http://www.mn-am.com (accessed Aug 31, 2021), Nuremberg, Germany, **2021**.

[26] P. Labute, *J. Mol. Graphics Modell.* **2000**, 18, 464–477. https://doi.org/10.1016/S1093-3263(00)00068-1

[27] A. Yan, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429–434. https://doi.org/10.1021/ci025590u

[28] A. Yan, J. Gasteiger, M. Krug, S. Anzali, *J. Comput.-Aided Mol. Des.* **2004**, *18*, 75–87. https://doi.org/10.1023/B:jcam.0000030031.81235.05

[29] R. Wang, Y. Gao, L. Lai, *Perspec. Drug Discovery Des.* **2000**, *19*, 47–66. https://doi.org/10.1023/A:1008763405023

[30] R. G. Brereton, *Chemometrics for pattern recognition*, Wiley, Chichester, United Kingdom, **2009**. https://doi.org/10.1002/9780470746462

[31] B. G. M. Vandeginste, D. L. Massart, L. C. M. Buydens, S. De Jong, J. Smeyers-Verbeke, *Handbook of chemometrics and qualimetrics: Part B*, Elsevier, Amsterdam, The Netherlands, **1998**.

[32] K. Varmuza, P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics*, CRC Press, Boca Raton, FL, USA, **2009**.

[33] R. Wehrens, *Chemometrics with R*, Springer, Heidelberg, Germany, **2011**. https://doi.org/10.1007/978-3-642-17841-2

[34] K. Varmuza, P. Filzmoser, M. Dehmer, *Comput. Struct. Biotechnol. J.* **2013**, 5:e201302007, 1–10. https://doi.org/10.5936/csbj.201302007

[35] P. Filzmoser, B. Liebmann, K. Varmuza, *J. Chemom.* **2009**, *23*, 160–171. https://doi.org/10.1002/cem.1225

[36] K. Varmuza in *Chemoinformatics - Basic Concepts and Methods* (Eds.: T. Engel, J. Gasteiger), Wiley-VCH, Weinheim, Germany, **2018**, pp. 399–437. https://doi.org/10.1002/9783527816880.ch11_01

[37] K. Varmuza, P. Filzmoser in *Current applications of chemometrics* (Eds.: M. Khanmohammadi), Nova Science Publishers, New York, NY, USA, **2015**, pp. 15–31.

[38] E. Gurian, A. Di Silvestre, E. Mitri, D. Pascut, C. Tiribelli, M. Giuffrè, L. Saveria Crocè, V. Sergo, A. Bonifacio, *Anal. Bioanal. Chem.* **2021**, 413, 1303–1312. https://doi.org/10.1007/s00216-020-03093-7