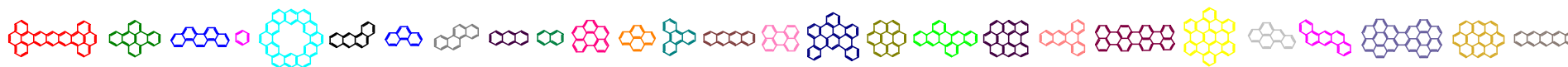


Relação comprimento-ordem de ligação para hidrocarbonetos benzenóides: Uma abordagem quimiométrica

Rudolf Kiralj (PQ), Márcia Miguel Castro Ferreira (PQ)

rudolf@igmp.unicamp.br, marcia@igmp.unicamp.br

Instituto de Química, Universidade Estadual de Campinas, Campinas 13083-970, SP, Brasil



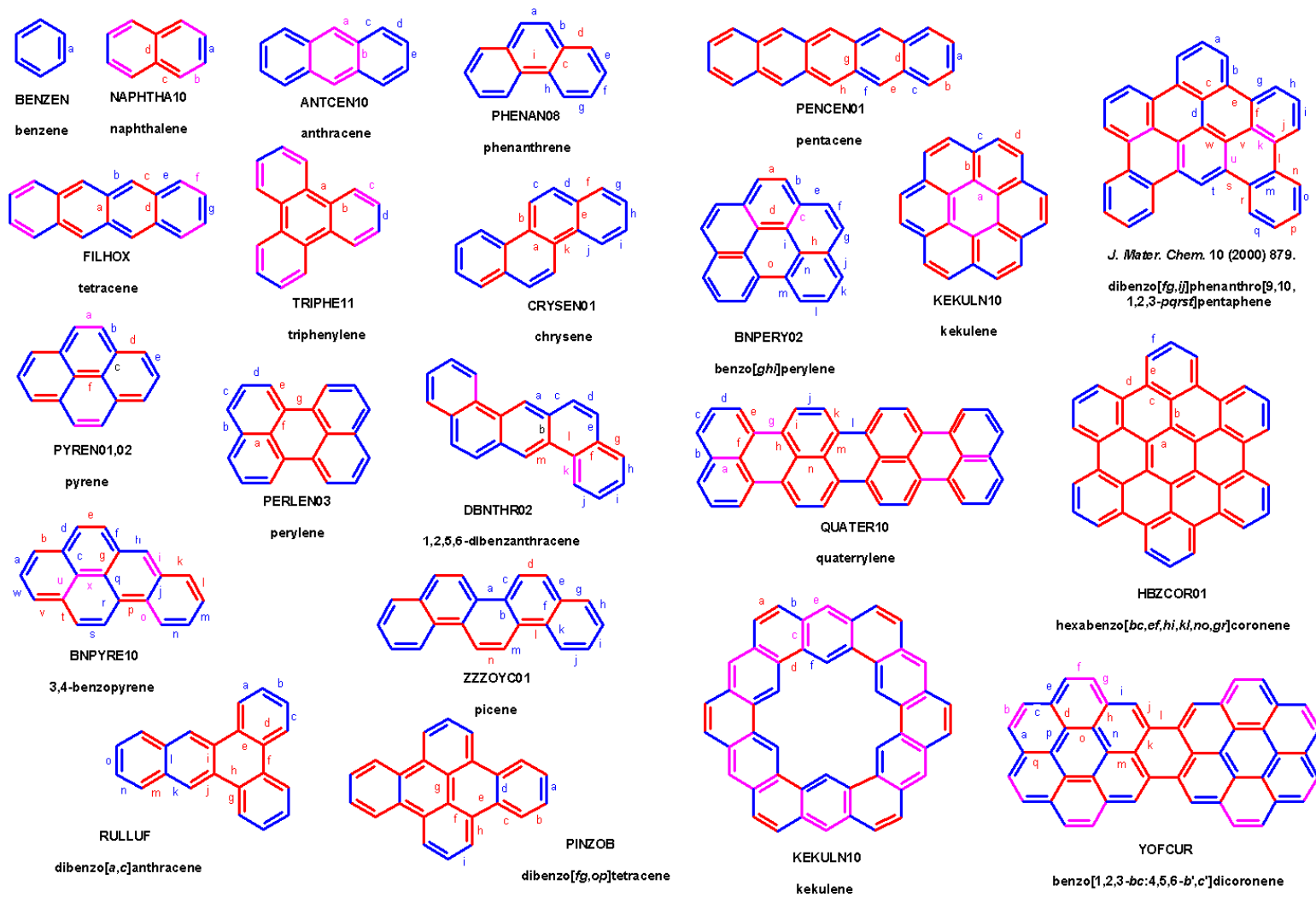


Figure 1. Molecules of the training/validation set represented by Kekulé structures, bond numeration, crystal packing effects, IUPAC names and literature sources.

Crystal packing effects:
 — bond shortening
 — no change
 — bond lengthening

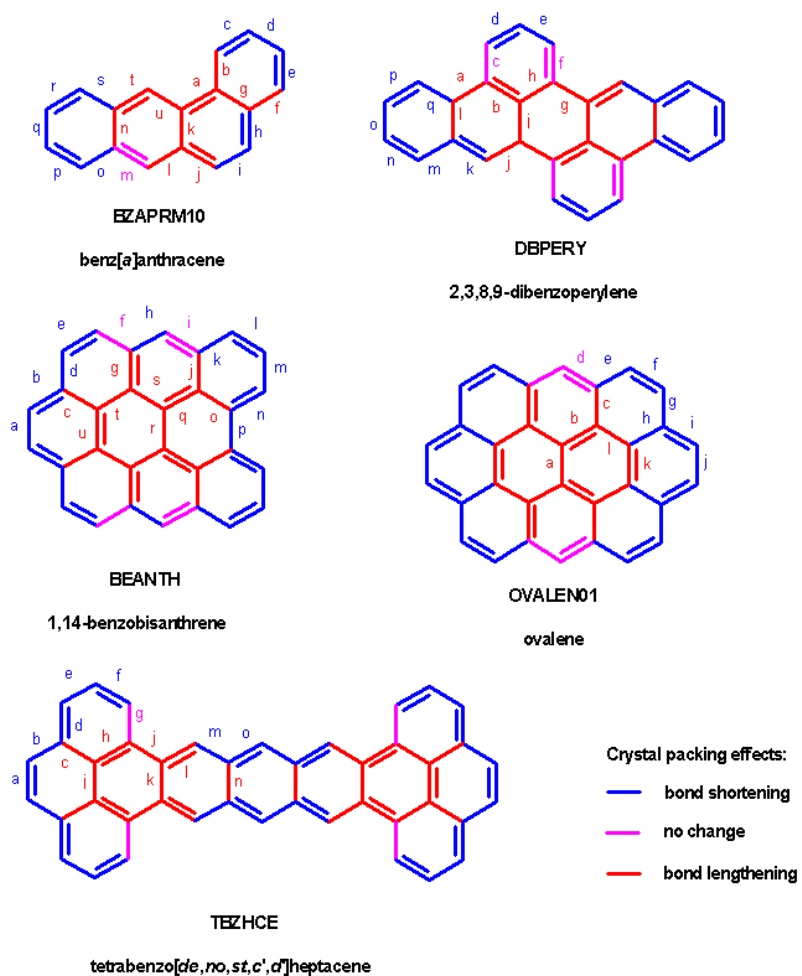


Figure 2. Molecules of the predicted set represented by Kekulé structures, bond numeration, crystal packing effects, IUPAC names and literature sources.

INTRODUCTION

PROBLEM 1.

Chemical bond lengths d are usually 1-2 Å (with relative experimental error $\approx 0.1\%$). Carbon-carbon d 's are 1.1-1.6 Å. For **planar and nearly planar benzenoids (benzenoid polycyclic aromatic hydrocarbons, PB-PAHs)** the d range is 1.34-1.48 Å (10% variation). Changes in d usually follow charge redistribution and can be useful in structural or QSAR studies. Questions to answer:

- Which properties, molecular descriptors (in fact, **bond descriptors**), define d ?
- Is d one- or multi-dimensional phenomenon?
- How to predict d rapidly and easily, with enough accuracy?

PROBLEM 2.

PB-PAHs, their derivatives, fragments and heterocyclic analogues are widely abundant in synthetic and natural substances, especially in biological systems: organic solvents, environmental cancerogens and mutagens, DNA-intercalators, constitutive parts of various drugs, nucleic acids and proteins, vitamins and coenzymes etc. etc. Intra- and intermolecular **(hetero)aromatic-(hetero)aromatic stacking interactions** of PB-PAHs and analogue fragments stabilize chemical and biological systems, especially in crystal phase. Question to be considered:

- How then to rationalize aromatic C-C, C-N, C-O and other d 's if not starting with PB-PAHs?

PROBLEM 3.

Many papers studied d of PAHs (relatively heterogeneous sets with not much accurate experimental data) as uni-dimensional functions of the Pauling π -bond orders p_P , $\log p_P$, rational function of p_P or other bond orders or topological indices.

Here we extend previous studies [1,2] updating the set of experimental d 's for PB-PAHs, use multivariate versus univariate techniques to classify and predict the d 's and to answer the questions like:

- Which analysis, univariate or multivariate is preferred?
- Is highly accurate prediction of d possible?
- Is fast and approximate prediction of d and p_P possible without complicated procedures?

METHODOLOGY

PHASE 1. DATA MINING AND PREPARATION FOR ANALYSIS

The search for the best crystal structures (crystallographic $R < 0.07$, other criteria as defined in [1,2]) in Cambridge Structural Database (CSD) [3], December 2000 release, and in the most recent literature (1999-2001) was performed. The d values with their estimated standard deviations σ found in CSD&literature were averaged over maximal molecular symmetry in the free state.

PHASE 2. PREPARATION OF BOND DESCRIPTORS

The bond descriptors were calculated without computer assistance:

- p_P bond orders
- p_{cr} bond orders: p_P corrected to crystal packing effects [4]
- n number: the number of C atoms around the bond
- m number: the number of hexagons around the bond
- l number: the number of C atoms around those atoms counted for n number
- k number: the number of C atoms around those atoms counted for l number

The n, m, l, k numbers can be considered **topological descriptors (indices)**.

PHASE 3. CLASSIFICATION OF THE BONDS

Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) were performed on autoscaled data (k number excluded) [5].

PHASE 4. REGRESSION MODELS FOR PREDICTION OF d AND p_P

A: Univariate regressions for d using p_P and p_{cr}

B: Multivariate regression models – Partial Least Squares (PLS) with two Principal Components (PC's) and Multivariate Linear Regression (MLR) [5,6] were used to predict d using $p_P, p_{cr}, n, m, l, n^2, m_{cr} = m + a m^2, l_{cr} = l + a l^2 + b l^3 + c l^4 + d l^5 + e l^6$. The prediction of p_P also was performed.

All the analysis was carried out by using Pirouette 3.01 [7] and Matlab 5.4 [8].

RESULTS&DISCUSSION I

1. EXPERIMENTAL DATA STATISTICS

A - There were found 22 high precision crystal structures containing 22 PB-PAHs (Figure 1), what comprised 223 symmetrically independent aromatic C-C bond lengths (Table 1). This dataset was treated as the training/validation set.

B - Five low precision crystal structures of 5 PB-PAHs (Figure 2) with 86 symmetrically independent C-C bonds were used as the prediction set.

C - The training/validation set consists of 44.4% shortened, 45.3% lengthened and 10.3% unchanged experimental bond lengths with respect to those predicted from $d - p_P$ linear relationship (Figure 1). The same ratio for shortened, lengthened and unchanged d 's was kept for the predicted set, following some additional rules [4].

2. THE BOND LENGTH-BOND DESCRIPTOR CORRELATIONS

A - The d values showed these correlation coefficients r with the bond descriptors:

p_P :	-0.895	p_P, p_P^2 :	0.898	$\log(1 + p_P)$:	-0.898	f_P :	-0.896
p_{cr} :	-0.929	p_{cr}, p_{cr}^2 :	0.931	$\log(1 + p_{cr})$:	-0.931	f_{cr} :	-0.927
n :	0.735	m :	0.689	l :	0.502	k :	0.250
n, n^2 :	0.741	m, m^2 :	0.748	l, l^2 :	0.506	l, \dots, l^6 :	0.630

where $f_P = 1.84 p_P / (0.84 p_P + 1)$ and $f_{cr} = 1.84 p_{cr} / (0.84 p_{cr} + 1)$. The relationships marked with red are used in the further analysis.

B - The number k and was not used in further analysis due to low correlation. The r 's for n , l , k decrease linearly with the corresponding **topological distances** (the number of bonds between the counted atoms and the studied bond) 1, 2, 3, respectively. This shows that the neighbourhood with the distance 4 has practically no influence on the bond under the consideration.

C - There is a **degeneration of data** in the sense of $d - p_P$ relationship as already noticed [2]: there are more d values (with differences out of experimental errors σ) with the same p_P values due to structural variations of the chemical bond, packing forces, unknown experimental errors and other effects. This denegeration is more pronounced as the set of data increases [2].

RESULTS&DISCUSSION II

3. THE BOND LENGTH CLASSIFICATION

PCA ANALYSIS:

The training/validation set described by (p_P, p_{cr}, n, m, l) showed that:

A – the aromatic carbon-carbon bond length in PB-PAHs is at least two-dimensional phenomena, what is in accordance with the early observations of Dewar and Gleisher [9]:

PC	% Variance	% Cumul. Var.	p_P	p_{cr}	n	m	l
PC1	74.08	74.08	0.414	0.430	-0.482	-0.482	-0.424
PC2	22.38	96.46	-0.567	-0.527	-0.206	-0.292	-0.522
PC3	2.55	99.01	0.128	0.057	0.844	-0.432	-0.286
PC4	0.85	99.87	-0.147	-0.088	-0.101	-0.704	0.682
PC5	0.13	100.00	-0.685	0.725	0.058	0.021	-0.024

B – all the bond descriptors are important for PC1 and PC2 (p_P and p_{cr} have higher loadings for PC1 while n, m, l have more for PC2). PC1 discriminates well cluster (p_P, p_{cr}) from (n, m, l) .

C – there are twelve C-C bond classes (lines I-XII in Figure 3), in general characterized well by the number $n + m + l$ which increases with PC1. It is $n = 2$ for classes I-III, $n = 3$ for IV and V, $n = 4$ for IX-XII, and VI-VIII disturb the regularity. The neighbourhood of particular C-C bond becomes denser as PC1 increases and PC2 decreases (Figure 4).

HCA ANALYSIS:

Bond variables form two clusters, (p_P, p_{cr}) and (n, m, l) . In general, C-C bonds are grouped in the classes in a similar way as in PCA (Figure 5).

RESULTS&DISCUSSION III

4. THE REGRESSION MODELS (Table 2)

UNIVARIATE REGRESSION MODELS:

The prediction of d (training/validation set) using bond orders reaches $r=0.94$ when p_{cr} are used in $\log x$ form. The use of squares of p_P or p_{cr} (parabolic regression) slightly improves the model. The most reasonable choice are linear models, among which the model 2 should be preferred.

MULTIVARIATE REGRESSION MODELS:

A - The prediction of d (training/validation set) reaches $r=0.96$ in models 1-10. The best and the simplest model we propose (Table 2) is model 8 presented in Figure 6. The calculated and predicted d values for this model are in Table 1.

The regression vector for this model shows almost equal contribution of the bond orders, and less contribution of the topological indices in linear decreasing order $n - m - l$:

$$p_P: -0.357 \quad p_{cr}: -0.385 \quad n: 0.157 \quad m: 0.114 \quad l: 0.006$$

B - The approximate prediction of d from the n, m, l numbers (models 3-6) is of course less accurate than the multivariate models including the bond orders, but the model 5 (containing all the information from the original variables) we can recommend as the fast, easy and approximate prediction of d 's of PB-PAHs.

C - The approximate predictions of p_P and p_{cr} from n, m, l (models 11-18) behave as the analogous models for prediction of d . The models proposed are those containing the maximum information, models 14 and 18.

D - The regression models exhibit that non-linear forms of m and l , *i.e.* m_{cr} and l_{cr} , are needed whenever the bond orders are not included. In such a case all the information is required, as observed when the corresponding MLR and PLS are compared.

E - The multivariate prediction of d , the models 7-10, reach up-to-date experimental precision $\approx 0.005 \text{ \AA}$ (comparable to $\langle \Delta \rangle$ values), and are relatively far from the limit $\langle \Delta / \sigma \rangle = 2.58$.

Table 1. Carbon-carbon bond descriptors* for planar benzenoid PAHs.

No.	molecule	bond	p_P	f_P	p_{cr}	n	m	l	k	$d_{exp}/\text{\AA}$	$d_{calc}/\text{\AA}$	$\sigma/\text{\AA}$
1	benzene	a	0.500	0.648	0.527	2	1	2	0	1.390	1.384	0.009
2	naphthalene	a	0.333	0.479	0.353	2	1	2	2	1.407	1.405	0.002
3		b	0.667	0.787	0.667	2	1	3	2	1.371	1.365	0.002
4		c	0.333	0.479	0.313	3	2	4	1	1.422	1.418	0.002
5		d	0.333	0.479	0.313	4	2	4	0	1.420	1.424	0.002
6		e	0.250	0.380	0.267	2	1	2	2	1.418	1.416	0.005
7	anthracene	a	0.500	0.648	0.500	3	2	5	3	1.395	1.396	0.003
8		b	0.250	0.380	0.250	4	2	4	2	1.432	1.433	0.003
9		c	0.250	0.380	0.267	3	2	4	2	1.428	1.426	0.003
10		d	0.750	0.847	0.787	2	1	3	2	1.353	1.353	0.004
11		e	0.250	0.380	0.267	2	1	2	2	1.418	1.416	0.005
12		a	0.800	0.880	0.839	2	1	4	4	1.338	1.347	0.005
13		b	0.200	0.315	0.215	3	2	4	3	1.422	1.432	0.007
14		c	0.400	0.551	0.377	4	2	5	2	1.413	1.416	0.007
15		d	0.400	0.551	0.377	3	2	4	2	1.414	1.410	0.004
16	e	0.600	0.734	0.631	2	1	3	2	1.349	1.372	0.008	
17	f	0.400	0.551	0.423	2	1	2	2	1.381	1.397	0.009	
18	g	0.600	0.734	0.631	2	1	3	3	1.376	1.372	0.004	
19	h	0.400	0.551	0.423	3	2	5	3	1.391	1.407	0.007	
20	i	0.200	0.315	0.185	4	3	6	2	1.454	1.444	0.006	
21	tetracene	a	0.200	0.315	0.185	4	2	4	4	1.442	1.440	0.002
22		b	0.400	0.551	0.423	3	2	5	4	1.405	1.407	0.002
23		c	0.600	0.734	0.569	3	2	5	3	1.388	1.386	0.002
24		d	0.200	0.315	0.185	4	2	4	2	1.441	1.440	0.002
25		e	0.200	0.315	0.215	3	2	4	2	1.434	1.432	0.002
26		f	0.800	0.880	0.800	2	1	3	2	1.349	1.349	0.002
27		g	0.200	0.315	0.215	2	1	2	2	1.415	1.422	0.002
28	triphenylene	a	0.111	0.187	0.100	4	3	6	4	1.469	1.455	0.008
29		b	0.444	0.595	0.419	4	2	6	4	1.411	1.411	0.009
30		c	0.444	0.595	0.419	3	2	5	4	1.405	1.405	0.009
31		d	0.556	0.697	0.556	2	1	3	3	1.385	1.379	0.010
32		e	0.444	0.595	0.469	2	1	2	2	1.390	1.391	0.011
33	chrysene	a	0.500	0.648	0.473	4	2	6	4	1.402	1.404	0.002
34		b	0.250	0.380	0.233	3	2	5	5	1.437	1.428	0.002
35		c	0.750	0.847	0.787	2	1	4	5	1.331	1.353	0.002
36		d	0.250	0.380	0.267	3	2	4	3	1.417	1.426	0.002
37		e	0.375	0.525	0.353	4	2	5	2	1.417	1.419	0.002
38		f	0.375	0.525	0.353	3	2	4	2	1.415	1.413	0.002
39		g	0.625	0.754	0.657	2	1	3	2	1.361	1.369	0.002
40		h	0.375	0.525	0.397	2	1	2	2	1.392	1.400	0.002
41		i	0.625	0.754	0.657	2	1	3	3	1.370	1.369	0.002
42		j	0.375	0.525	0.397	3	2	5	3	1.409	1.410	0.002
43		k	0.250	0.380	0.233	4	3	6	3	1.453	1.438	0.002
44		pyrene	a	0.833	0.902	0.833	2	1	4	4	1.347	1.345
45	c		0.333	0.479	0.353	4	3	6	3	1.417	1.425	0.003

46		d	0.500	0.648	0.473	3	2	4	3	1.400	1.398	0.004
47		e	0.500	0.648	0.527	2	1	3	3	1.380	1.384	0.005
48		f	0.333	0.479	0.313	4	4	8	2	1.423	1.432	0.003
49	perylene	a	0.333	0.479	0.313	4	2	6	3	1.426	1.424	0.002
50		b	0.333	0.479	0.353	3	2	4	3	1.411	1.415	0.002
51		c	0.667	0.787	0.701	2	1	3	3	1.359	1.363	0.003
52		d	0.333	0.479	0.353	2	1	3	4	1.393	1.405	0.002
53		e	0.667	0.787	0.633	3	2	5	5	1.384	1.378	0.002
54		f	0.333	0.479	0.313	4	3	7	4	1.429	1.428	0.002
55		g	0.000	0.000	-0.007	4	3	6	6	1.471	1.468	0.002
56	1,2,5,6-dibenz- anthracene	a	0.500	0.648	0.527	3	2	5	5	1.391	1.394	0.002
57		b	0.333	0.479	0.313	4	2	5	4	1.426	1.424	0.002
58		c	0.167	0.269	0.181	3	2	4	4	1.438	1.436	0.002
59		d	0.833	0.902	0.874	2	1	4	4	1.338	1.343	0.002
60		e	0.167	0.269	0.181	3	2	4	3	1.436	1.436	0.002
61		f	0.417	0.568	0.393	4	2	5	2	1.413	1.414	0.002
62		g	0.417	0.568	0.393	3	2	4	2	1.412	1.408	0.002
63		h	0.583	0.720	0.614	2	1	3	2	1.360	1.378	0.002
64		i	0.417	0.568	0.441	2	1	2	2	1.403	1.395	0.002
65		j	0.583	0.720	0.614	2	1	3	3	1.373	1.374	0.002
66		k	0.417	0.568	0.417	3	2	5	3	1.406	1.406	0.002
67		l	0.167	0.269	0.153	4	3	6	3	1.455	1.448	0.002
68		m	0.500	0.648	0.473	3	2	5	5	1.397	1.398	0.002
69	picene	a	0.308	0.450	0.327	4	3	6	4	1.429	1.429	0.007
70		b	0.462	0.612	0.488	4	2	6	4	1.388	1.405	0.009
71		c	0.231	0.356	0.247	3	2	5	5	1.412	1.428	0.009
72		d	0.769	0.860	0.731	2	1	4	5	1.367	1.355	0.008
73		e	0.231	0.356	0.247	3	2	4	3	1.410	1.428	0.010
74		f	0.385	0.535	0.362	4	2	5	2	1.414	1.418	0.009
75		g	0.385	0.535	0.362	3	2	4	2	1.414	1.412	0.008
76		h	0.615	0.746	0.647	2	1	3	2	1.367	1.370	0.009
77		i	0.385	0.535	0.408	2	1	2	2	1.366	1.399	0.010
78		j	0.615	0.746	0.647	2	1	3	3	1.409	1.370	0.009
79		k	0.385	0.535	0.408	3	2	5	3	1.394	1.409	0.010
80		l	0.231	0.356	0.215	4	3	6	3	1.469	1.440	0.008
81		m	0.308	0.450	0.289	3	2	5	5	1.430	1.421	0.010
82		n	0.692	0.805	0.657	2	1	4	6	1.406	1.365	0.008
83	3,4-benzopyrene	a	0.444	0.595	0.469	2	1	3	3	1.378	1.391	0.013
84		b	0.556	0.697	0.527	3	2	4	3	1.401	1.391	0.012
85		c	0.333	0.479	0.353	4	3	6	3	1.414	1.425	0.012
86		d	0.111	0.187	0.122	3	2	4	4	1.434	1.443	0.012
87		e	0.889	0.936	0.846	2	1	4	4	1.342	1.341	0.013
88		f	0.111	0.187	0.122	3	2	4	5	1.447	1.443	0.012
89		g	0.222	0.344	0.206	4	3	6	5	1.444	1.441	0.011
90		h	0.667	0.787	0.701	3	2	5	5	1.361	1.373	0.012
91		i	0.333	0.479	0.333	3	2	5	4	1.419	1.417	0.012
92		j	0.333	0.479	0.353	4	2	5	3	1.410	1.421	0.011
93		k	0.333	0.479	0.313	3	2	4	3	1.425	1.418	0.012
94		l	0.667	0.787	0.633	2	1	3	2	1.374	1.368	0.014
95		m	0.333	0.479	0.353	2	1	2	2	1.397	1.405	0.014

96		n	0.667	0.787	0.701	2	1	3	3	1.364	1.363	0.013
97		o	0.333	0.479	0.333	3	2	5	3	1.419	1.417	0.012
98		p	0.333	0.479	0.313	4	3	6	4	1.435	1.428	0.012
99		q	0.444	0.595	0.469	4	3	7	5	1.395	1.412	0.011
100		r	0.222	0.344	0.238	3	2	5	6	1.423	1.429	0.011
101		s	0.778	0.866	0.817	2	1	4	5	1.352	1.349	0.012
102		t	0.222	0.344	0.206	3	2	4	4	1.441	1.431	0.012
103		u	0.333	0.479	0.333	4	3	6	3	1.418	1.427	0.011
104		v	0.444	0.595	0.419	3	2	4	3	1.412	1.404	0.012
105		w	0.556	0.697	0.585	2	1	3	3	1.376	1.377	0.014
106		x	0.333	0.479	0.333	4	4	8	3	1.419	1.431	0.011
107	pentacene	a	0.167	0.269	0.181	2	1	2	2	1.428	1.426	0.005
108		b	0.833	0.902	0.792	2	1	3	2	1.355	1.348	0.006
109		c	0.167	0.269	0.181	3	2	4	2	1.434	1.436	0.005
110		d	0.167	0.269	0.153	4	2	4	2	1.445	1.444	0.005
111		e	0.667	0.787	0.657	3	2	5	3	1.387	1.376	0.005
112		f	0.333	0.479	0.353	3	2	5	4	1.412	1.415	0.005
113		g	0.167	0.269	0.153	4	2	4	4	1.458	1.444	0.005
114		h	0.500	0.648	0.473	3	2	5	4	1.412	1.398	0.005
115	dibenzo[a,c]-	a	0.538	0.682	0.567	2	1	3	3	1.382	1.379	0.001
116	anthracene	b	0.462	0.612	0.488	2	1	2	2	1.389	1.389	0.001
117		c	0.538	0.682	0.567	2	1	3	3	1.373	1.379	0.001
118		d	0.462	0.612	0.436	3	2	5	4	1.412	1.402	0.001
119		e	0.462	0.612	0.436	4	2	6	4	1.410	1.408	0.001
120		f	0.077	0.133	0.067	4	3	6	4	1.469	1.459	0.001
121		g	0.462	0.612	0.436	3	2	5	4	1.410	1.402	0.001
122		h	0.077	0.133	0.067	4	3	6	5	1.469	1.459	0.001
123		i	0.308	0.450	0.289	4	2	6	6	1.436	1.427	0.001
124		j	0.692	0.805	0.657	3	2	6	6	1.387	1.375	0.001
125		k	0.385	0.535	0.408	3	2	5	4	1.409	1.409	0.001
126		l	0.308	0.450	0.327	4	2	4	2	1.420	1.424	0.001
127		m	0.308	0.450	0.289	3	2	4	2	1.426	1.421	0.001
128		n	0.692	0.805	0.727	2	1	3	2	1.364	1.360	0.001
129		o	0.308	0.450	0.327	2	1	2	2	1.412	1.408	0.001
130	dibenzo[fg,op]-	a	0.450	0.601	0.475	2	1	2	2	1.391	1.390	0.005
131	tetracene	b	0.550	0.692	0.521	2	1	3	3	1.397	1.382	0.005
132		c	0.450	0.601	0.425	3	2	5	4	1.438	1.404	0.005
133		d	0.450	0.601	0.475	4	2	6	4	1.386	1.407	0.005
134		e	0.100	0.170	0.089	4	3	6	3	1.477	1.456	0.005
135		f	0.400	0.551	0.377	4	3	7	6	1.416	1.420	0.005
136		g	0.200	0.315	0.185	4	4	8	6	1.457	1.448	0.005
137		h	0.500	0.648	0.473	3	2	5	5	1.416	1.398	0.005
138		i	0.500	0.648	0.527	2	1	3	4	1.383	1.384	0.005
139	benzo[ghi]-	a	0.643	0.768	0.610	2	1	4	4	1.399	1.371	0.008
140	perylene	b	0.357	0.505	0.378	3	2	4	4	1.396	1.412	0.007
141		c	0.429	0.580	0.429	4	3	6	5	1.406	1.415	0.007
142		d	0.286	0.424	0.268	4	4	8	6	1.438	1.438	0.006
143		e	0.214	0.334	0.230	3	2	4	4	1.446	1.430	0.008
144		f	0.786	0.871	0.825	2	1	4	4	1.341	1.348	0.008
145		g	0.214	0.334	0.230	3	2	4	4	1.430	1.430	0.007

146		h	0.357	0.505	0.336	4	3	6	4	1.438	1.425	0.006
147		i	0.286	0.424	0.304	4	4	8	5	1.419	1.436	0.006
148		j	0.429	0.580	0.453	3	2	4	3	1.389	1.403	0.007
149		k	0.571	0.710	0.601	2	1	3	3	1.376	1.375	0.008
150		l	0.429	0.580	0.453	2	1	3	4	1.394	1.393	0.007
151		m	0.571	0.710	0.571	3	2	5	5	1.385	1.387	0.007
152		n	0.357	0.505	0.378	4	3	7	5	1.411	1.423	0.006
153		o	0.071	0.123	0.061	4	3	6	6	1.484	1.460	0.006
154	coronene	a	0.300	0.441	0.300	4	4	8	6	1.424	1.435	0.005
155		b	0.400	0.551	0.377	4	3	6	5	1.420	1.420	0.005
156		c	0.300	0.441	0.319	3	2	4	4	1.414	1.419	0.005
157		d	0.700	0.811	0.665	2	1	4	4	1.372	1.364	0.005
158	dibenzo[fg,ij]-	a	0.500	0.648	0.527	2	1	3	4	1.381	1.384	0.003
159	phenanthro[9,10,	b	0.500	0.648	0.527	3	2	5	5	1.386	1.394	0.003
160	1,2,3-pqrst]-	c	0.400	0.551	0.377	4	3	7	6	1.424	1.420	0.002
161	pentaphene	d	0.200	0.315	0.215	4	4	8	8	1.433	1.446	0.003
162		e	0.100	0.170	0.089	4	3	6	6	1.463	1.456	0.002
163		f	0.500	0.648	0.473	4	3	7	6	1.411	1.408	0.003
164		g	0.400	0.551	0.423	3	2	5	5	1.402	1.407	0.003
165		h	0.600	0.734	0.631	2	1	3	4	1.377	1.372	0.003
166		i	0.400	0.551	0.423	2	1	3	4	1.366	1.397	0.003
167		j	0.600	0.734	0.569	3	2	5	5	1.401	1.386	0.003
168		k	0.300	0.441	0.300	4	3	7	6	1.425	1.431	0.002
169		l	0.100	0.170	0.089	4	3	6	5	1.455	1.456	0.003
170		m	0.350	0.497	0.371	4	2	6	4	1.413	1.419	0.002
171		n	0.550	0.692	0.521	3	2	5	4	1.413	1.392	0.003
172		o	0.450	0.601	0.475	2	1	3	3	1.360	1.391	0.003
173		p	0.550	0.692	0.521	2	1	2	2	1.392	1.381	0.003
174		q	0.450	0.601	0.475	2	1	3	3	1.378	1.391	0.003
175		r	0.550	0.692	0.521	3	2	5	4	1.396	1.392	0.003
176		s	0.100	0.170	0.089	4	3	6	6	1.471	1.456	0.002
177		t	0.400	0.648	0.527	3	2	6	8	1.391	1.395	0.002
178		u	0.400	0.551	0.400	4	3	7	8	1.408	1.419	0.003
179		v	0.200	0.315	0.185	4	4	8	7	1.449	1.448	0.002
180		w	0.400	0.551	0.377	4	4	8	9	1.424	1.424	0.002
181	quaterrylene	a	0.333	0.479	0.333	4	3	6	3	1.420	1.427	0.004
182		b	0.333	0.479	0.353	3	2	4	3	1.417	1.415	0.004
183		c	0.667	0.787	0.701	2	1	3	3	1.367	1.363	0.004
184		d	0.333	0.479	0.353	2	1	3	4	1.401	1.405	0.004
185		e	0.667	0.787	0.633	3	2	5	5	1.382	1.378	0.004
186		f	0.333	0.479	0.313	4	3	7	4	1.431	1.428	0.004
187		g	0.000	0.000	0.000	4	3	6	6	1.468	1.468	0.004
188		h	0.333	0.479	0.313	4	3	7	6	1.431	1.428	0.004
189		i	0.667	0.787	0.633	3	2	5	6	1.383	1.378	0.004
190		j	0.333	0.479	0.353	2	1	4	6	1.383	1.405	0.004
191		k	0.667	0.787	0.633	3	2	5	6	1.389	1.378	0.004
192		l	0.000	0.000	0.007	4	3	6	6	1.462	1.467	0.004
193		m	0.333	0.479	0.313	4	3	7	6	1.429	1.428	0.004
194		n	0.333	0.479	0.313	4	4	8	6	1.431	1.432	0.004
195	hexabenzo[bc,ef,	a	0.400	0.551	0.377	4	4	8	10	1.417	1.424	0.002

196	<i>hi,kl,no,qr</i>]-	b	0.200	0.315	0.185	4	4	8	8	1.446	1.448	0.002
197	coronene	c	0.400	0.551	0.377	4	3	7	6	1.417	1.420	0.002
198		d	0.100	0.170	0.089	4	3	6	6	1.458	1.456	0.002
199		e	0.500	0.648	0.473	3	2	5	5	1.398	1.398	0.002
200		f	0.500	0.648	0.527	2	1	3	4	1.376	1.384	0.002
201	kekulene	a	0.850	0.912	0.809	2	1	4	4	1.350	1.346	0.002
202		b	0.150	0.245	0.163	3	2	4	4	1.442	1.438	0.002
203		c	0.350	0.498	0.350	4	2	5	4	1.418	1.420	0.002
204		d	0.150	0.245	0.137	4	3	6	4	1.456	1.450	0.002
205		e	0.500	0.648	0.500	3	2	5	4	1.395	1.396	0.002
206		f	0.500	0.648	0.527	3	2	6	6	1.386	1.395	0.002
207	benzo[1,2,3-	a	0.300	0.441	0.319	3	2	4	4	1.417	1.419	0.002
208	<i>bc:4,5,6-</i>	b	0.700	0.811	0.700	2	1	4	4	1.364	1.361	0.002
209	<i>b',c'</i>]diconene	c	0.300	0.441	0.319	3	2	4	4	1.422	1.419	0.002
210		d	0.400	0.551	0.377	4	3	6	5	1.415	1.420	0.002
211		e	0.300	0.441	0.319	3	2	4	4	1.419	1.419	0.002
212		f	0.700	0.811	0.700	2	1	4	4	1.365	1.361	0.002
213		g	0.300	0.441	0.300	3	2	4	5	1.424	1.421	0.002
214		h	0.400	0.551	0.377	4	3	6	6	1.412	1.420	0.002
215		i	0.300	0.441	0.319	3	2	5	6	1.413	1.420	0.002
216		j	0.700	0.811	0.665	3	2	6	7	1.379	1.374	0.002
217		k	0.300	0.441	0.281	4	3	7	7	1.432	1.432	0.002
218		l	0.000	0.000	-0.007	4	3	6	8	1.478	1.468	0.002
219		m	0.400	0.551	0.377	4	4	8	8	1.420	1.424	0.002
220		n	0.300	0.441	0.319	4	4	8	8	1.421	1.434	0.002
221		o	0.300	0.441	0.281	4	4	8	6	1.429	1.436	0.002
222		p	0.300	0.441	0.319	4	4	8	6	1.422	1.434	0.002
223		q	0.400	0.551	0.377	4	3	6	5	1.422	1.420	0.002
224	benz[<i>a</i>]anthracene	a	0.143	0.235	0.130	4	3	6	3	1.483	1.471	0.011
225		b	0.428	0.579	0.404	3	2	5	3	1.401	1.404	0.012
226		c	0.571	0.710	0.601	2	1	3	3	1.400	1.356	0.011
227		d	0.428	0.579	0.452	2	1	2	2	1.392	1.383	0.014
228		e	0.571	0.710	0.601	2	1	3	2	1.393	1.356	0.014
229		f	0.428	0.579	0.404	3	2	4	2	1.418	1.403	0.011
230		g	0.428	0.579	0.404	4	2	5	2	1.442	1.414	0.012
231		h	0.143	0.235	0.156	3	2	4	3	1.396	1.453	0.013
232		i	0.857	0.917	0.899	2	1	4	4	1.322	1.302	0.011
233		j	0.143	0.235	0.130	3	2	4	4	1.429	1.455	0.012
234		k	0.286	0.424	0.268	4	2	5	4	1.434	1.439	0.012
235		l	0.571	0.710	0.571	3	2	5	4	1.384	1.375	0.011
236		m	0.428	0.579	0.428	3	2	5	3	1.431	1.401	0.013
237		n	0.286	0.424	0.268	4	2	4	2	1.397	1.439	0.013
238		o	0.286	0.424	0.306	3	2	4	2	1.436	1.426	0.012
239		p	0.714	0.821	0.750	2	1	3	2	1.323	1.329	0.014
240		q	0.286	0.424	0.306	2	1	2	2	1.444	1.409	0.015
241		r	0.714	0.821	0.750	2	1	3	2	1.360	1.329	0.013
242		s	0.286	0.424	0.306	3	2	4	2	1.428	1.426	0.013
243		t	0.428	0.579	0.404	3	2	5	4	1.422	1.404	0.011
244		u	0.571	0.710	0.541	3	2	6	5	1.364	1.378	0.012
245	2,3,8,9-dibenzo-	a	0.200	0.315	0.185	4	3	6	4	1.458	1.461	0.032

246	perylene	b	0.400	0.551	0.377	4	3	7	5	1.384	1.425	0.032
247		c	0.400	0.551	0.400	3	2	5	4	1.422	1.407	0.032
248		d	0.600	0.734	0.631	2	1	3	4	1.387	1.350	0.032
249		e	0.400	0.551	0.423	2	1	3	4	1.381	1.388	0.032
250		f	0.600	0.734	0.600	3	2	5	5	1.394	1.370	0.032
251		g	0.000	0.000	-0.007	4	3	7	7	1.478	1.497	0.032
252		h	0.400	0.551	0.377	4	3	5	5	1.454	1.425	0.032
253		i	0.200	0.315	0.185	4	3	6	6	1.479	1.461	0.032
254		j	0.800	0.880	0.761	3	2	7	7	1.406	1.337	0.032
255		k	0.200	0.315	0.215	3	2	5	5	1.409	1.442	0.032
256		l	0.400	0.551	0.377	4	2	3	3	1.379	1.418	0.032
257		m	0.400	0.551	0.423	3	2	3	3	1.413	1.404	0.032
258		n	0.600	0.734	0.631	2	1	2	2	1.384	1.350	0.032
259	o	0.400	0.551	0.423	2	1	2	2	1.399	1.388	0.032	
260	p	0.600	0.734	0.631	2	1	3	3	1.412	1.350	0.032	
261	q	0.400	0.551	0.423	3	2	3	3	1.403	1.404	0.032	
262	1,14-benzobis- anthrene	a	0.533	0.677	0.562	2	1	2	2	1.40	1.363	0.02
263		b	0.467	0.617	0.467	3	2	4	4	1.39	1.394	0.02
264		c	0.400	0.551	0.377	4	3	6	6	1.420	1.425	0.02
265		d	0.133	0.220	0.145	3	2	4	4	1.46	1.454	0.02
266		e	0.867	0.923	0.969	2	1	4	4	1.35	1.295	0.02
267		f	0.133	0.220	0.145	3	2	4	4	1.47	1.454	0.02
268		g	0.233	0.359	0.217	4	3	6	6	1.44	1.455	0.02
269		h	0.633	0.760	0.666	3	2	5	5	1.37	1.361	0.02
270		i	0.367	0.516	0.367	3	2	5	5	1.40	1.413	0.02
271		j	0.300	0.441	0.281	4	3	6	6	1.42	1.443	0.02
272		k	0.333	0.479	0.353	3	2	4	4	1.43	1.417	0.02
273		l	0.667	0.787	0.701	2	1	3	3	1.37	1.338	0.02
274		m	0.333	0.479	0.353	2	1	3	3	1.43	1.401	0.02
275		n	0.667	0.787	0.701	3	2	5	5	1.36	1.354	0.02
276		o	0.300	0.441	0.281	4	3	7	7	1.43	1.443	0.02
277	p	0.033	0.059	0.041	4	3	6	6	1.49	1.490	0.02	
278	q	0.400	0.551	0.377	4	4	8	8	1.40	1.431	0.02	
279	r	0.133	0.220	0.121	4	4	8	8	1.47	1.479	0.02	
280	s	0.467	0.617	0.441	4	4	8	8	1.41	1.419	0.02	
281	t	0.300	0.441	0.281	4	4	8	8	1.43	1.449	0.02	
282	u	0.300	0.441	0.281	4	4	8	8	1.44	1.449	0.02	
283	ovalene	a	0.00	0.315	0.185	4	4	8	10	1.435	1.467	0.006
284		b	0.400	0.551	0.377	4	4	8	8	1.415	1.431	0.004
285		c	0.300	0.441	0.281	4	3	6	6	1.424	1.443	0.004
286		d	0.500	0.648	0.500	3	2	6	5	1.400	1.388	0.004
287		e	0.200	0.315	0.215	3	2	4	5	1.441	1.442	0.004
288		f	0.800	0.880	0.839	2	1	4	4	1.356	1.313	0.004
289		g	0.200	0.315	0.215	3	2	4	4	1.429	1.442	0.004
290		h	0.400	0.551	0.423	4	3	6	5	1.450	1.421	0.004
291		i	0.400	0.551	0.423	3	2	4	4	1.413	1.404	0.004
292		j	0.600	0.734	0.631	2	1	4	4	1.365	1.351	0.006
293		k	0.300	0.441	0.281	4	4	8	6	1.413	1.449	0.006
294		l	0.300	0.441	0.281	4	4	8	5	1.411	1.449	0.004
295	tetrabenzo[<i>de,no</i> ,	a	0.809	0.886	0.849	2	1	4	4	1.35	1.311	0.02

296	<i>st,c',d'</i>]heptacene	b	0.191	0.303	0.206	3	2	4	4	1.45	1.443	0.02
297		c	0.182	0.290	0.168	4	3	4	4	1.42	1.464	0.02
298		d	0.427	0.578	0.451	3	2	3	3	1.39	1.399	0.02
299		e	0.573	0.712	0.603	2	1	3	3	1.39	1.356	0.02
300		f	0.427	0.578	0.451	2	1	4	4	1.39	1.383	0.02
301		g	0.573	0.712	0.573	3	2	5	5	1.38	1.375	0.02
302		h	0.382	0.532	0.360	4	3	5	5	1.44	1.428	0.02
303		i	0.236	0.362	0.220	4	4	4	4	1.42	1.460	0.02
304		j	0.045	0.079	0.036	4	3	6	6	1.48	1.489	0.02
305		k	0.227	0.351	0.211	4	2	6	6	1.44	1.450	0.02
306		l	0.727	0.831	0.691	3	2	6	6	1.37	1.350	0.02
307		m	0.273	0.409	0.291	3	2	5	5	1.42	1.428	0.02
308		n	0.227	0.351	0.211	4	2	4	4	1.42	1.450	0.02
309		o	0.500	0.648	0.527	3	2	4	4	1.38	1.385	0.02

*The bonds are numbered according to molecular graphs in Figure 1. The bonds 1-223 are for the training/validation set, and 224-309 for the predicted set. The molecular descriptors are: p_p – Pauling π -bond order, $f_p = 1.84p_p / (0.84p_p + 1)$ – modified Pauling π -bond order, p_{cr} – Pauling π -bond order including corrections for the crystal packing effects, n – the number of neighbouring carbon atoms around the bond, m – the number of benzenoid rings around the bond, l – the number of neighbouring carbon atoms around those atoms counted for n , k – the number of neighbouring carbon atoms around those counted for l , d_{exp} – the bond lengths from X-ray or neutron structure determinations (in Å), d_{calc} – (the model 8 was used) calculated (for the training/validation set) or predicted (for the predicted set) d 's in Å, σ – estimated standard deviations for d_{exp} (in Å).

Table 2. The regression models.*

No.	Y	X 's	Method	R	Q	SEV	$\langle \Delta \rangle$	$\langle \Delta / \sigma \rangle$
1 [#]	$d / \text{Å}$	p_P	LR	0.895	0.893	0.014	0.010	3.345
2 [#]	$d / \text{Å}$	p_{cr}	LR	0.929	0.929	0.011	0.008	2.344
3 [#]	$d / \text{Å}$	n, m, l	MLR	0.836	0.830	0.017	0.014	4.492
4	$d / \text{Å}$	n, m, l	PLS	0.820	0.814	0.018	0.014	4.603
5 [#]	$d / \text{Å}$	$n, m, l, m_{crd}, l_{crd}$	MLR	0.848	0.839	0.016	0.013	4.142
6 ^{&}	$d / \text{Å}$	$n, m, l, m_{crd}, l_{crd}$	PLS	0.838	0.830	0.017	0.013	4.195
7	$d / \text{Å}$	p_P, p_{cr}, n, m, l	MLR	0.959	0.957	0.009	0.006	1.823
8	$d / \text{Å}$	p_P, p_{cr}, n, m, l	PLS	0.940	0.938	0.011	0.007	2.167
9	$d / \text{Å}$	$p_P, p_{cr}, n, m, l, m_{crd}, l_{crd}$	MLR	0.960	0.957	0.009	0.006	1.749
10	$d / \text{Å}$	$p_P, p_{cr}, n, m, l, m_{crd}, l_{crd}$	PLS	0.943	0.941	0.010	0.007	2.058
11 [#]	p_P	n, m, l	MLR	0.795	0.787	0.115	0.091	-
12	p_P	n, m, l	PLS	0.765	0.757	0.122	0.098	-
13	p_P	$n, m, l, m_{crpp}, l_{crdpp}$	MLR	0.800	0.789	0.114	0.089	-
14 ^{&}	p_P	$n, m, l, m_{crpp}, l_{crdpp}$	PLS	0.779	0.767	0.119	0.094	-
15 [#]	p_{cr}	n, m, l	MLR	0.815	0.809	0.111	0.092	-
16	p_{cr}	n, m, l	PLS	0.789	0.782	0.117	0.099	-
17	p_{cr}	$n, m, l, m_{crpcr}, l_{crdpcr}$	MLR	0.822	0.812	0.110	0.089	-
18 ^{&}	p_{cr}	$n, m, l, m_{crpcr}, l_{crdpcr}$	PLS	0.802	0.792	0.115	0.095	-

*The Linear Regression (LR), Multiple Linear Regression (MLR) and Partial Least Squares (PLS) models for prediction of d , p_P and p_{cr} . The corrected variables have the form $m_{cr} = m + a m^2$ and $l_{cr} = l + a l^2 + b l^3 + c l^4 + d l^5 + e l^6$, where the coefficients a to e are found in the polynomial fitting to d , p_P and p_{cr} (indexes of the corrections are marked with crd, crpp and crpcr, respectively). R and Q are the prediction and validation correlation coefficients, SEV is the standard error of validation, $\langle \Delta \rangle$ is the average absolute deviation of predicted d 's from experimental, $\langle \Delta / \sigma \rangle$ is the average Δ / σ ratio where σ is the experimental estimated standard deviation on d 's. SEV and $\langle \Delta \rangle$ are in Å when referred to bond lengths d . [#]All the regression coefficients are greater than their statistical errors more than 2.58 times. The model 5 hardly satisfy this condition. [&]The models with 3 PC's used. Other PLS models are performed with 2 PC's.

CONCLUSIONS

The aromatic carbon-carbon bond in PB-PAHs from crystal structures:

I – is at least two-dimensional phenomenon

II – depends both on Pauling bond orders and topological indices describing the bond neighbourhood

III – can be classified in twelve classes (all theoretically possible classes for PB-PAHs) with distinguished n and $n + m + l$ numbers

IV – is better predicted if crystal effects are introduced (as for example p_{cr} or its function)

V – is better predicted with multivariate models based on PC's than with univariate ones (avoiding highly correlated descriptors and two-dimensional descriptors like p_{cr} or functions of p_P or p_{cr}). The model recommended here is the PLS model with p_P, p_{cr}, n, m, l as bond descriptors

VI – contains some information difficult to rationalize, as crystal structures cannot be predicted up-to-date, what limits the full rationalization of crystal effect corrections. Besides, our prediction, compared to experimental errors, is accurate enough for sophisticated studies.

VII – can be, as well as p_P and p_{cr} predicted approximately using regression models with topological indexes, avoiding this way the use of complicated algorithms and computer assistance (the recommended models are MLR or PLS models with all the PC's, including with n, m, l, m_{cr}, l_{cr} as bond descriptors)

VIII – is predicted for the prediction set satisfactorily well ($R = 0.835$, $SEP = 0.016 \text{ \AA}$, $\langle \Delta \rangle = 0.021 \text{ \AA}$, $\langle \Delta / \sigma \rangle = 1.634$)

IX – is a good starting point for the study of other aromatic bonds in different molecular classes

X – shows that the bond orders (p_P, p_{cr}) and the topological indices (n, m, l, m_{cr}, l_{cr}) are not mutually orthogonal but exhibit moderate correlation ($r = 0.33-0.64$)

REFERENCES

- [1] R. Kiralj, B. Kojić-Prodić, M. Žinić, S. Alihodžić, N. Trinajstić, *Acta Cryst.*, **B52** (1996) 823-837.
- [2] R. Kiralj, B. Kojić-Prodić, S. Nikolić, N. Trinajstić, *J. Mol. Struct. (Theochem)*, **427** (1998) 25-37.
- [3] F. K. Allen, O. Kennard, *Chem. Design Autom. News*, **8** (1993) 131-137.
- [4] M. M. C. Ferreira, R. Kiralj, *unpublished results*.
- [5] M. A. Sharaf, D. L. Illman, B. P. Kowalski: *Chemometrics*, J. Wiley&Sons, New York, 1986.
- [6] K. R. Beebe, B. P. Kowalski, *Anal. Chem.*, **59** (1987) 1007A-1017A.
- [7] Pirouette 3.01, Infometrix, Inc., Seattle, WA, 2001
- [8] Matlab 5.4, The MathWorks, Inc., Natick, MA, 2001.
- [9] M. S. J. Dewar, G. J. Gleisher, *J. Chem. Phys.*, **44** (1966) 759-773.

ACKNOWLEDGEMENTS

The authors acknowledge FAPESP for the financial support.

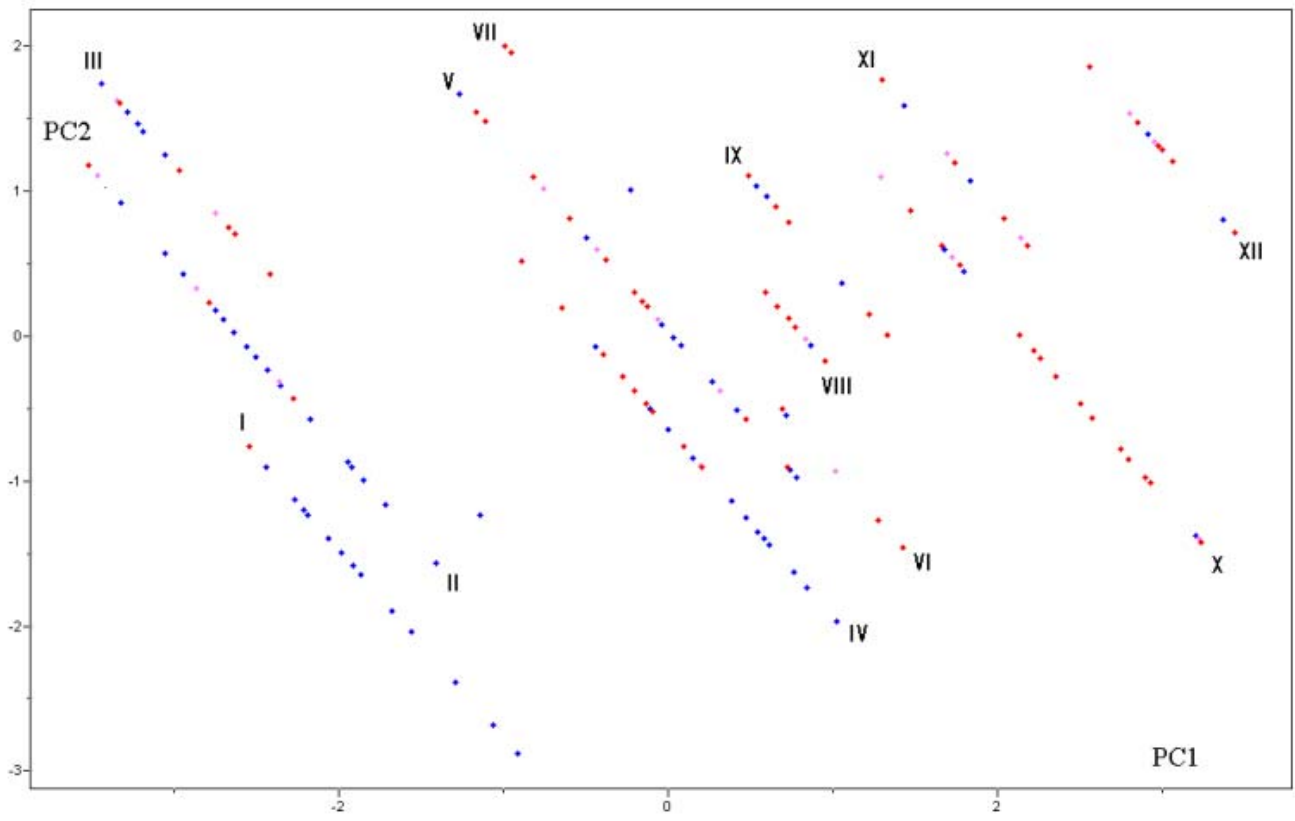


Figure 3. The PCA plot of the samples grouped into twelve groups and colored analogously to Fig. 4.

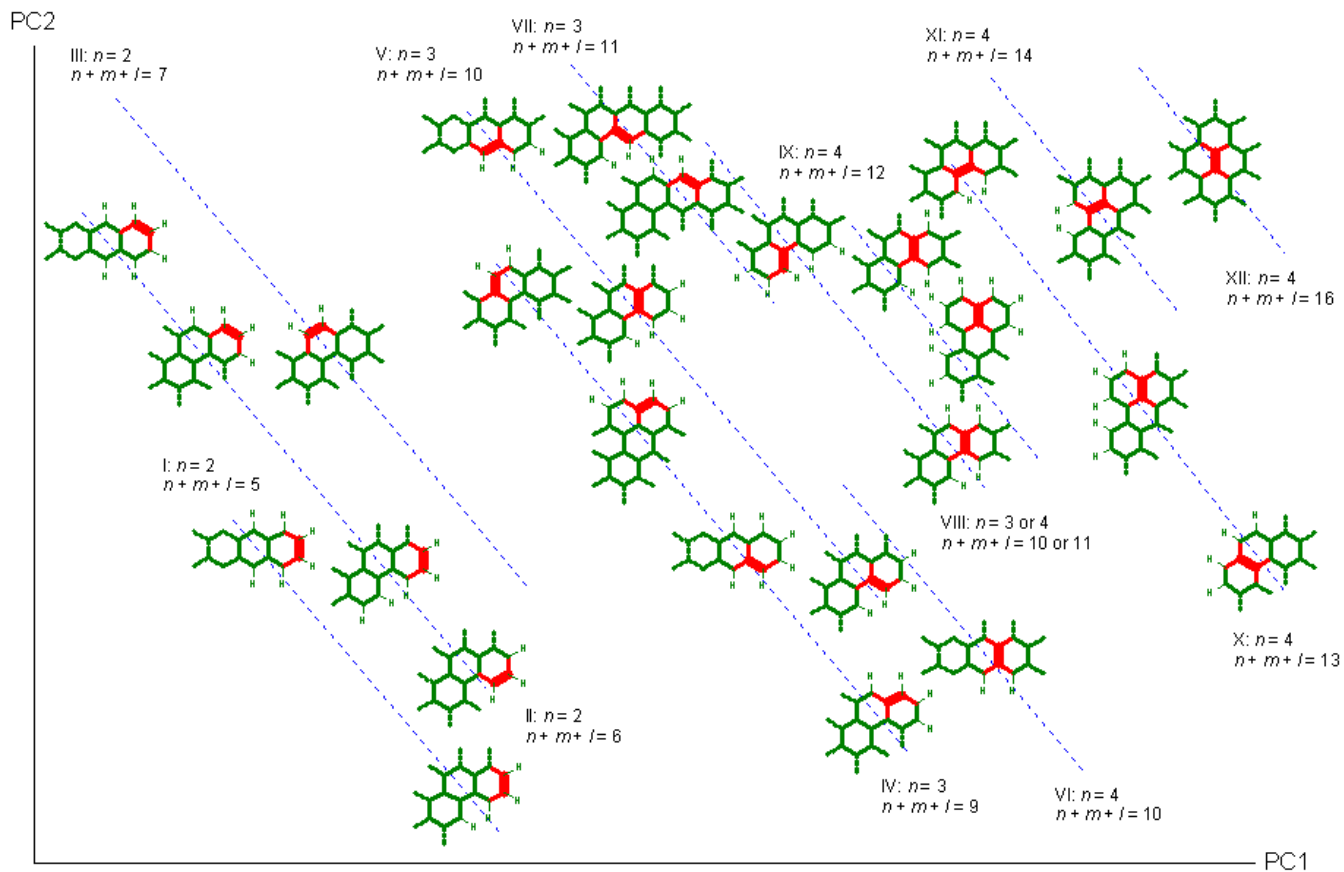


Figure 4. The PCA classes of C-C bonds. The dominant common fragments and bond descriptors are shown.

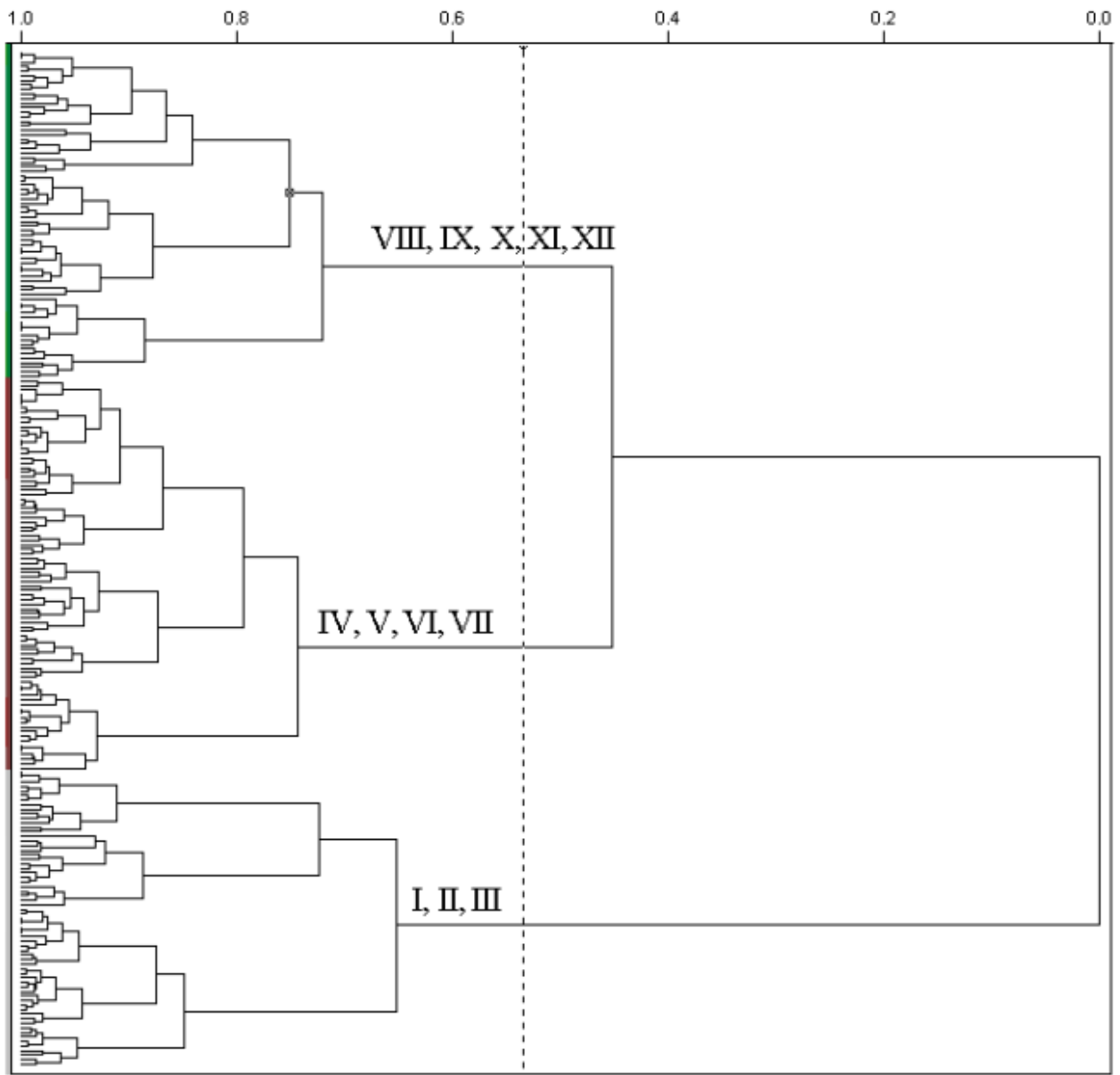


Figure 5. The HCA dendrogram showing the distribution of classes I-XII.

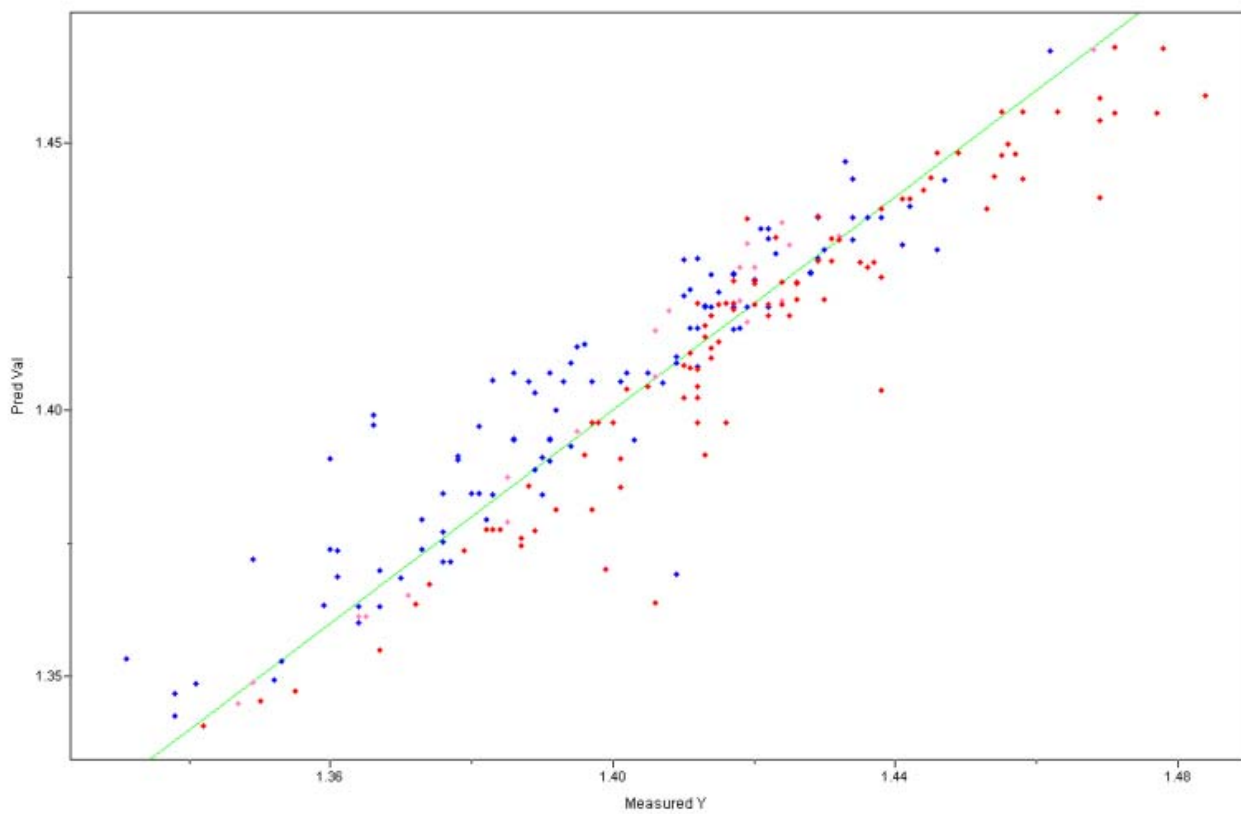


Figure 6. The PLS plot for model 8 colored in the same way as in Figures 1-3.