Extending characteristic polynomial from graphs to molecules

Lorentz JÄNTSCHI

&

Sorana D. BOLBOACĂ

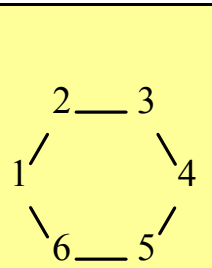
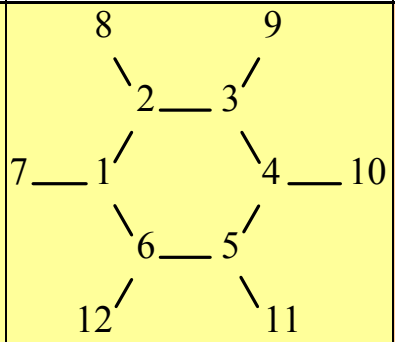
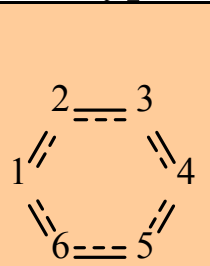
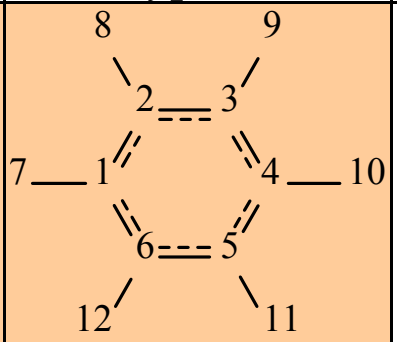
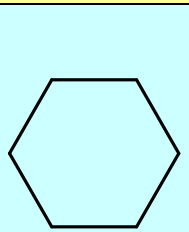
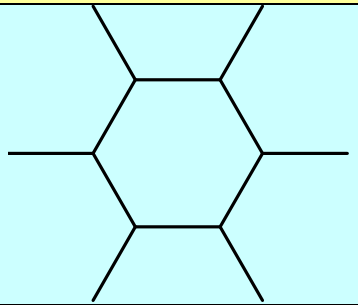
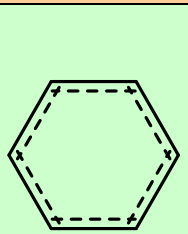
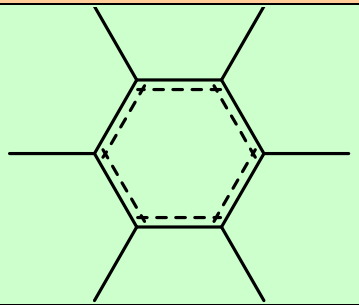
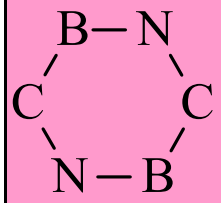
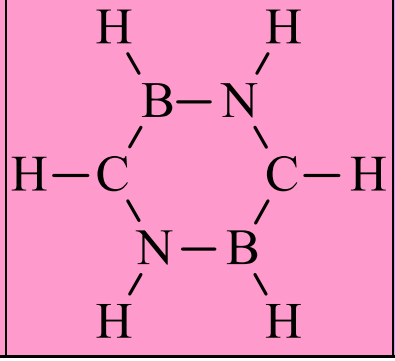
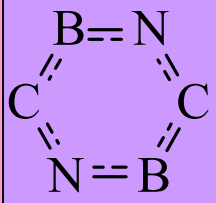
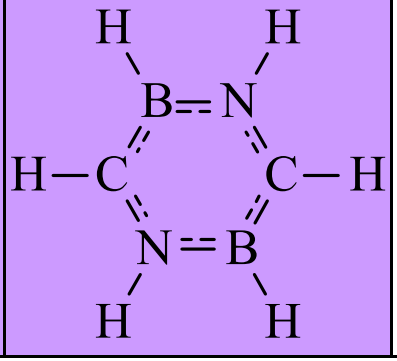
UT & UMF Cluj-Napoca

May 13, 2016

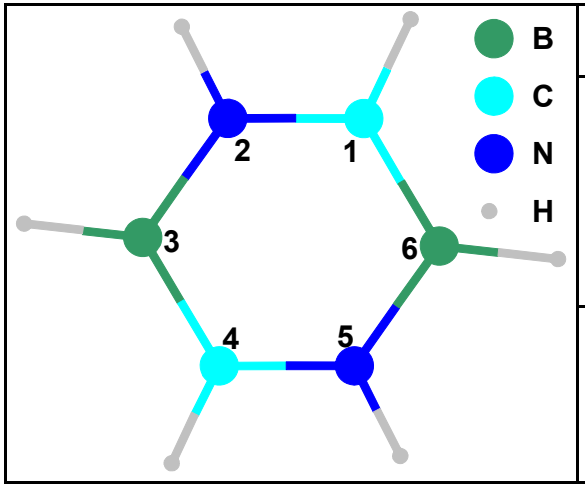
Reference materials

- Jäntschi L, 2006. Characteristic & counting polynomials. Online calculation software. URL: http://l.academicdirect.org/Fundamentals/Graphs/polynomials/
- Bolboacă SD, Jäntschi L, 2007. How good can the characteristic polynomial be for correlations?, International Journal of Molecular Sciences 8(4): 335-345.
- Jäntschi L, Bolboacă SD, Furdui CM, 2009. Characteristic and counting polynomials: modelling nonane isomers properties, Molecular Simulation 35(3): 220-227.
- Jäntschi L, 2012. Structure vs. Properties: Algorithms and Models. Habilitation thesis in Chemistry. Bucharest: CNATDCU (defended in 2013 in Cluj-Napoca: Babeş-Bolyai University). 175 p.
- Jäntschi L, Bolboacă SD, 2016. Characteristic polynomial In: Explicative dictionary of nanochemistry (Ed. M.V. Putz). Oakville(ON): AAP & CRC Press, to appear.

Graphs vs. molecules

&	classical	classical & H	b.-type	b.-type & H	groups
num.					classical molecular topology
un-num.					classical characteristic polynomial numbered weighted graphs isomorphic-free weighted graphs vertex-labeled graphs
a.-type					chemical structural formula (labeled graphs)

Molecular geometry

	$\angle(BCN) = 120.0°$	$\angle(CNB) = 125.7°$	$\angle(NBC) = 114.3°$
	$\overrightarrow{CH} \cdot (\overrightarrow{CB} \times \overrightarrow{CN}) = 0$ $\angle(HCB) = 124.3°$ $\angle(HCN) = 115.7°$	$\overrightarrow{NH} \cdot (\overrightarrow{NB} \times \overrightarrow{NC}) = 0$ $\angle(HNB) = 117.6°$ $\angle(HNC) = 116.7°$	$\overrightarrow{BH} \cdot (\overrightarrow{BC} \times \overrightarrow{BN}) = 0$ $\angle(HBC) = 126.0°$ $\angle(HBN) = 119.7°$
	$\lvert \overrightarrow{BC} \rvert = 1.493$ Å $\lvert \overrightarrow{BH} \rvert = 1.187$ Å	$\lvert \overrightarrow{CN} \rvert = 1.350$ Å $\lvert \overrightarrow{CH} \rvert = 1.075$ Å	$\lvert \overrightarrow{NB} \rvert = 1.448$ Å $\lvert \overrightarrow{NH} \rvert = 1.005$ Å

Challenges:

- Molecules carry both topological and geometrical features
- Atom type & bond type are essential factors in the expression of the measurable properties
- Atoms &/| bonds numbering induces an undesired isomorphism
- Geometry & bond types induces other kinds of isomorphism

Classical molecular graphs

Classical Molecular Topology:
Unlabeled Unweighted Undirected Graphs

Definition	Names (concepts)	Cardinality	Example
V: finite set	V: vertices (atoms)	$\|V\|=N$: atoms	G = "A-B-C"
$E \subseteq V \times V$	E: edges (bonds)	$\|E\|=M$: bonds	V=$\{1(\leftrightarrow A),2(\leftrightarrow B),3(\leftrightarrow C)\}$
$G=G(V,E)$	G: graph (molecule)	$\forall N, V \leftrightarrow 1..N$	E=$\{(1,2),(2,3)\}$

Topological description of a molecule requires storing the adjacencies (the bonds) between the atoms as well as the identities (the atoms). If this problem is simplified at maximum, by disregarding the bond types and atom types then:

- adjacencies are simply stored with 0 and 1 in the vertex adjacency matrix ([Ad])
- the identities are stored with 0 and 1 into the identity matrix ([Id]).

Characteristic polynomial

- The characteristic polynomial (ChP) is the natural construction of a polynomial in which the eigenvalues of the [Ad] are the roots of the ChP as it follows:

λ is an eigenvalue of [Ad] ↔ it exists [v] ≠ 0 eigenvector such that λ·[v] = [Ad]·[v] → (λ·[Id]-[Ad])·[v] = 0;

since v ≠ 0 → [λ·Id - Ad] is singular → det([λ·Id - Ad]) = 0

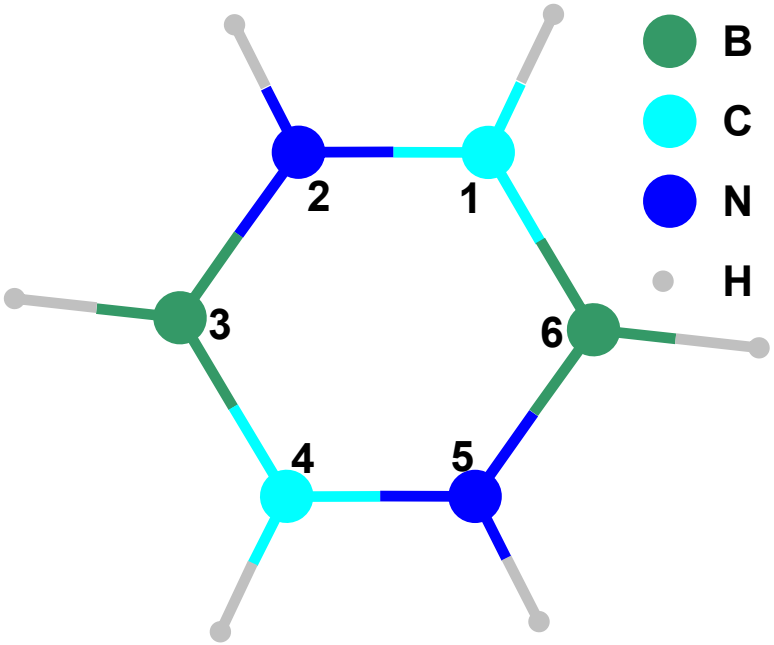
$$\mathrm{ChP} \overset{\mathrm{def}}{=} \left|\lambda \cdot \mathrm{Id} - \mathrm{Ad}\right|$$

- The characteristic polynomial is a polynomial in λ of degree the number of atoms.
- Please note that this definition allows extensions.
  - A natural extension is to store in the identity matrix (instead of unity) non-unity values accounting for the atom types, as well as to store in the adjacency matrix (instead of unity) non-unity values accounting for the bond types.

Short history

- Hückel 1931: First report of using the ChP in relation with the chemical structure; Hückel's method of molecular orbitals it is actually the first extension of the ChP
- Hartree 1928ab, Slater 1929, Fock 1930ab, Hartree & Hartree 1935 give the path of the second extension of ChP (embeded as HF 'theory level' in molecular modeling; geometry of $B_2C_2N_2H_6$ molecule obtained from restricted HF with 3-21G* basis set)
- More general, the eigenproblem (finding of eigenvalues and eigenvectors) may be involved to any Hessian (Sylvester 1880) matrix [A] ([Ad] → [A]; Hessian: $A_{i,j} = A_{j,i}$).

Characteristic polynomial extension



[Ad]	1	2	3	4	5	6
1	0	1	0	0	0	1
2	1	0	1	0	0	0
3	0	1	0	1	0	0
4	0	0	1	0	1	0
5	0	0	0	1	0	1
6	1	0	0	0	1	0

[Id]	1	2	3	4	5	6
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	1	0
6	0	0	0	0	0	1

↓ extension ↓

[A]	1	2	3	4	5	6
1	0	a	0	0	0	c
2	a	0	b	0	0	0
3	0	b	0	c	0	0
4	0	0	c	0	a	0
5	0	0	0	a	0	b
6	c	0	0	0	b	0

[I]	1	2	3	4	5	6
1	d	0	0	0	0	0
2	0	e	0	0	0	0
3	0	0	f	0	0	0
4	0	0	0	d	0	0
5	0	0	0	0	e	0
6	0	0	0	0	0	f

Challenge:

a=?, b=?, c=?, d=?, e=?, f=?

ChPE solution: a family of ChPs!

- There is no determinism in selecting the values of a-f.
- If $a = b = c = d = e = f = 1$ then ChPE ← ChP as in classical molecular topology;
- If $a = b = c = 1.5^{-1}$, then [A] accounts for the (inverse of the) bond order;
- If $a = 1.35^{-1}$, $b = 1.448^{-1}$, and $c = 1.493^{-1}$ then [A] accounts for the (inverse of the) geometrical distance (in ˚A);
- If $d = 12/294$, $e = 14/294$, and $f = 10.8/294$, then [I] accounts for atomic mass relative to Uuo, last element from the 7*th* period of the system of elements;
- If $d = 2267/\rho_{ref}$, $e = 1026/\rho_{ref}$, and $f = 2460/\rho_{ref}$, then [I] accounts for solid state relative density (in $m3/kg$); ρ_{ref} can be fixed to 30000;
- If $d = 2.55/4.00$, $e = 3.04/4.00$, and $f = 2.04/4.00$, then [I] accounts for electronegativity relative to Fluorine when Pauling scale is used;
- If $d = 1086.2/1312$, $e = 1402.3/1312$, and $f = 800.6/1312$, then [I] accounts for first potential of ionization relative to the potential of ionization for Hydrogen;
- If $d = 3820/3820$, $e = 63/3820$, and $f = 2573/3820$, then [I] accounts for melting point relative to diamond's allotrope of Carbon (in K);
- If $d = 1/4$, $e = 1/4$, and $f = 1/4$, then [I] accounts for the number of hydrogen atoms attached relative to the score of CH_4;

Characteristic polynomial extension

[Ad]	1	2	3	4	5	6
1	0	1	0	0	0	1
2	1	0	1	0	0	0
3	0	1	0	1	0	0
4	0	0	1	0	1	0
5	0	0	0	1	0	1
6	1	0	0	0	1	0

[Id]	1	2	3	4	5	6
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	1	0
6	0	0	0	0	0	1

→ extension →

[Di]	1	2	3	4	5	6
1	0	1	2	3	2	1
2	1	0	1	2	3	2
3	2	1	0	1	2	3
4	3	2	1	0	1	2
5	2	3	2	1	0	1
6	1	2	3	2	1	0

[Id]	1	2	3	4	5	6
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	1	0
6	0	0	0	0	0	1

↓ extension ↓ ↓ extension ↓

[A]	1	2	3	4	5	6
1	0	a	0	0	0	c
2	a	0	b	0	0	0
3	0	b	0	c	0	0
4	0	0	c	0	a	0
5	0	0	0	a	0	b
6	c	0	0	0	b	0

[I]	1	2	3	4	5	6
1	d	0	0	0	0	0
2	0	e	0	0	0	0
3	0	0	f	0	0	0
4	0	0	0	d	0	0
5	0	0	0	0	e	0
6	0	0	0	0	0	f

→ extension →

[D]	1	2	3	4	5	6
1	0	a	h	k	g	c
2	a	0	b	g	j	i
3	h	b	0	c	i	l
4	k	g	c	0	a	h
5	g	j	i	a	0	b
6	c	i	l	h	b	0

[I]	1	2	3	4	5	6
1	d	0	0	0	0	0
2	0	e	0	0	0	0
3	0	0	f	0	0	0
4	0	0	0	d	0	0
5	0	0	0	0	e	0
6	0	0	0	0	0	f

Extended characteristic polynomial

$$\mathrm{ChPE} \overset{\mathrm{def}}{=} \left|\lambda \cdot \mathrm{I} - \mathrm{A}\right|$$

[I] ← Atomic properties

'A' - atomic mass (/294.0)	'B' - cardinality (always 1)
'C' - charges (atomic electrostatic charge, ESP)	'D' - solid state density (in kg/m^3, /30000)
'E' - electronegativity (revised Pauling, /4.00)	'F' - first ionization potential (in kJ/mol, /1312.0)
'G' - melting point temperature (in K, /3820.0)	'H' - attached hydrogen atoms (/4)

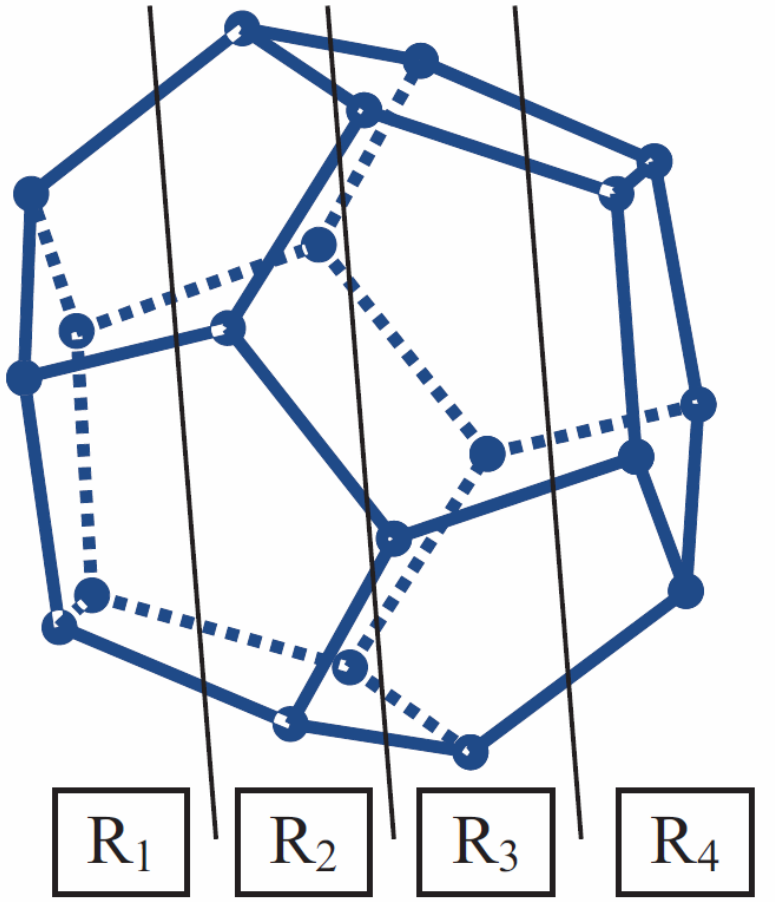
[A] ← Adjacency weights (bonds, distances)

On [Ad]	'g' - 0 or geometrical distance inversed	't' - (0 or 1)	'c' - 0 or inverse of bond order
On [Di]	'G' - geometrical distance inversed	'T' - topological distances	'C' - smallest sum of bond orders inverses

Software implementation

- ChPE is intended to be used to estimate molecular properties
- Let be ChPE = ChPE(A,I; x); A, I – fixed parameterizations; x – the variable. As for ChP(x), ChPE(x) diverges (to ∞) fast with increasing of x > 1. Thus, useful for polynomial evaluation is the [-1, 1] range
- We cannot use the full [-1, 1] interval to do regressions with measured properties but we may use a grid of it → let be an integer (k) ranging from -1000 to 1000 and its fraction k/1000 vary then from -1 to 1 and provides a grid
- Usually biological properties are expressed in log scale and therefore a linearization (L_1) is useful.
- The evaluated ChPE is labeled as follows: $L_1L_2L_3L_4d_1d_2d_3d_4$, with:
  - $L_1 \in \{I, R, L\}$ with $f_I(x) = x$, $f_R(x) = 1/x$, $f_L(x) = \ln(x)$
  - $L_2 \leftarrow$ Atomic property alternative from [I] matrix
  - $L_3 \leftarrow$ Metric of the distance alternative from [D] matrix
  - $L_4 \in \{N, P\}$ with N when k<0 (& x = k/1000 <0) and P otherwise
  - $L_5 \leftarrow |k|$
- Evaluation is made in every point (out of 2001) requiring $O(N^3)$ ops.
- When k ranges from -1000 to 1000 the ChPE family have 288144 individuals and in general when k ranges from –r to r it have 144·(2r+1) individuals.

Numerical case study



C_{20} congeners
Atoms $\in$ {B, C, N}
Stable if all the same in one layer
A simple math gives a count of 45 distinct compounds

The geometries were build at HF 6-31G level of theory and a calculated property were extracted from these calculations – polarizability, in $Å^3$

Molecular property data

Name	Pola	Name	Pola	Name	Pola	Name	Pola	Name	Pola
bbbb	60.469	bcnb	59.881	cbcb	61.228	cccc	58.487	cnnb	56.511
bbbn	61.208	bcnn	58.182	cbcn	58.345	cccn	57.091	cnnc	54.992
bbcn	60.081	bnbn	54.813	cbnb	58.474	ccnb	58.372	cnnn	53.051
bbnb	60.272	bncn	56.543	cbnc	56.843	ccnc	56.784	nbbn	57.786
bbnn	58.602	bnnb	57.038	cbnn	55.059	ccnn	54.976	nbnn	52.938
bcbb	62.309	bnnn	54.572	ccbb	61.306	cnbb	59.378	ncbn	55.991
bcbn	59.187	cbbb	61.333	ccbc	59.911	cnbn	54.798	nccn	54.967
bccb	62.714	cbbc	60.198	ccbn	57.661	cncb	59.111	ncnn	52.906
bccn	59.196	cbbn	59.654	cccb	61.359	cncn	54.802	nnnn	51.084

ChPE usage: Polarizability(Molecule) = f(ChPE(Molecule))

Alternatives for f:

$f(x_1) = ax_1 + b$

$f(x_1,x_2) = ax_1x2 + b$

$f(x_1,x_2) = ax_1 + bx_2 + c$

$f(x_1,x_2) = ax_1 + bx_2 + cx_1x_2 + d$

x_1, x_2 individuals of ChPE family

Results: best candidate regressions

Model	Descriptors	Coefficients	r^2_{adj}
$\hat{y} = a \cdot x_1 + d$	x_1 = IATN0079	a = -38.50 (t = 18.9) d = 500.4 (t = 21.4)	0.890
$\hat{y} = c \cdot x_1 \cdot x_2 + d$	x_1 = LFTN0225 x_2 = LBGP0631	c = 25.09 (t = 24.9) d = 71.64 (t = 126)	0.934
$\hat{y} = a \cdot x_1 + b \cdot x_2 + d$	x_1 = LBgN0419 x_2 = LGGN0488	a = 25.09 (t = 24.9) b = 0.6533 (t = 12.7) d = 37.82 (t = 43.1)	0.957
$\hat{y} = a \cdot x_1 + b \cdot x_2 + c \cdot x_1 \cdot x_2 + d$	x_1 = LDGN0394 x_2 = LDGN0402	a = 760.6 (t = 14.9) b = 735.4 (t = 13.8) c = -1.499 (t = 5.76) d = -58.45 (t = 3.34)	0.961

Comparison based on r_{adj}^2 values (see Fisher Z transformation):

- $z_{1,2} = \mathbf{8.21}$, between $\hat{y} = a \cdot x_1 + d$ and $\hat{y} = c \cdot x_1 \cdot x_2 + d$;
- $z_{1,3} = \mathbf{15.0}$, between $\hat{y} = a \cdot x_1 + d$ and $\hat{y} = a \cdot x_1 + b \cdot x_2 + d$;
- $z_{1,4} = \mathbf{16.5}$, between $\hat{y} = a \cdot x_1 + d$ and $\hat{y} = a \cdot x_1 + b \cdot x_2 + c \cdot x_1 \cdot x_2 + d$;
- $z_{2,3} = \mathbf{6.77}$, between $\hat{y} = c \cdot x_1 \cdot x_2 + d$ and $\hat{y} = a \cdot x_1 + b \cdot x_2 + d$;
- $z_{2,4} = \mathbf{8.30}$, between $\hat{y} = c \cdot x_1 \cdot x_2 + d$ and $\hat{y} = a \cdot x_1 + b \cdot x_2 + c \cdot x_1 \cdot x_2 + d$;
- $z_{3,4} = 1.53$, between $\hat{y} = a \cdot x_1 + b \cdot x_2 + d$ and $\hat{y} = a \cdot x_1 + b \cdot x_2 + c \cdot x_1 \cdot x_2 + d$;

Discussion

- It is important that the full model selects the same polynomial for the both descriptors (”DG” in both L**DG**N0394 and L**DG**N0402) while the evaluation is in two closer points (-394/1000 and -402/1000). This fact suggests that the best model to be used is this model of full effects since is strongly related with the concept of polarization - a charge separation - which usually takes small values relative to the total charge of an atom or an molecule.

Conclusion

- ChPE it is useful for what was intended to to – estimation of the molecular properties?
  - It seems that the answer is “Yes”; statistics suggests and physical interpretation sustain
- ChPE is the ideal extension of ChP?
  - Ideal gas doesn’t actually exist, so the same for ChPE
  - It is its natural extension
  - The scales of the atomic properties were more or less arbitrary selected – here is something to do later
  - The reversed distance seemed to be the best alternative, but also here some search for the best alternative can be deployed

Thank you for your attention!

- Questions?