

CHARACTERISTIC POLYNOMIAL (CHARACT-POLY)

Lorentz JÄNTSCHI¹, Sorana D. BOLBOACĂ²

¹ Technical University of Cluj-Napoca

² Iuliu Hațieganu University of Medicine and Pharmacy Cluj-Napoca

DEFINITION

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 and atom types then adjacencies are simply stored with 0 and 1 in the vertex adjacency matrix ([Ad]) and the identities are stored with 0 and 1 into the identity matrix ([Id]). 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] \leftrightarrow it exists $[v] \neq 0$ eigenvector such that $\lambda \cdot [v] = [Ad] \cdot [v] \rightarrow (\lambda \cdot [Id] - [Ad]) \cdot [v] = 0$; since $v \neq 0 \rightarrow [\lambda \cdot Id - Ad]$ is singular $\rightarrow \det([\lambda \cdot Id - Ad]) = 0$

$$\text{ChP} \stackrel{\text{def}}{=} |\lambda \cdot \text{Id} - \text{Ad}|$$

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.

KEYWORDS:

Topological theory of aromaticity; Structure-resonance theory; Quantum chemistry; Counts of random walks; Eigenvectors; Eigenvalues

HISTORICAL ORIGIN(S)

First reports relating to the use of the characteristic polynomial in relation with the chemical structure appears shortly after the discovery of wave-based treatment of microscopic level in (Hückel 1931¹). The Hückel's method of molecular orbitals it is actually the first extension of the Charact-poly definition. It uses the 'secular determinant', the determinant of a matrix which is decomposed as $[E \cdot Id - Ad]$, standing with the energy of the system (E in the place of λ), for approximate treatment of π electron systems in organic molecules. In this approximate treatment of the Schrödinger's (Schrödinger 1926²) equation ($E\psi = \hat{H}\psi$), the wavefunction (ψ) of the system configuration is defined as a linear combination (c_i stands for unknown, to be determined, coefficients) of the π electrons (p_i , each assigned to an atom), $\psi = \sum_i c_i p_i$ and the components of the molecular Hamiltonian (\hat{H}) are identified based on the orthogonal states of the electrons ($\langle p_i | p_j \rangle = \delta_{ij}$; $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = 0$ when $i \neq j$): $H_{ij} = \langle p_i | \hat{H} | p_j \rangle$ when $H_{i,i} = \langle p_i | \hat{H} | p_i \rangle = \alpha$ (if $i = j$, the same for all atoms) and $H_{i,j} = \langle p_i | \hat{H} | p_j \rangle = \beta$ if $[Ad]_{i,j} = 1$ and $H_{i,j} = \langle p_i | \hat{H} | p_j \rangle = 0$ if $[Ad]_{i,j} = 0$. The roots of this extended version of Charact-poly

are assigned to the individual electronic energies (ϵ_i). For further details please see (Coulson 1940³, Coulson 1940⁴, and Coulson 1950⁵).

Going in a different direction with the approximation of the wavefunction treatment, Hartree (Hartree 1928a⁶b⁷) and Fock (Fock 1930a⁸b⁹) finds the same eigenvector-eigenvalue problem (§20 in Laplace 1776¹⁰; T1 in Cauchy 1829¹¹) in the Slater's treatment (Slater 1929¹²; Hartree & Hartree 1935¹³). Here is the second extension of the Charact-poly, the eigenproblem (finding of eigenvalues and eigenvectors) being involved to any Hessian (Sylvester 1880¹⁴) matrix [A] ([Ad] \rightarrow [A]).

The Charact-poly it is related with the matching polynomial (Godsil & Gutman 1981¹⁵), because both polynomials degenerates to same expression for forests (disjoint union of trees). Adapting (Godsil 1995¹⁶) for molecules, a k-matching in a molecule is a matching with exact k bonds between different atoms (each set containing a single edge is also an independent edge set; the empty set should be treated as a independent edge set with zero edges - this set is unique; also due to the constraint of connecting different atoms the matching may involve no more than $\lfloor n/2 \rfloor$ bonds, where n is the number of atoms - see §3.1 & §3.3 in Diudea et al. 2001¹⁷). It is possible to count the k-matches (Ramaraj & Balasubramanian 1985¹⁸) - but nevertheless it is a hard problem (Curticapean 2013¹⁹), as well as to express the derived Z-counting polynomial (Hosoya 1971²⁰) and matching polynomial (both are defined using $m(k)$ as the k-matching number of the selected molecule):

Table 1. Polynomials derived from k-matching

Z-counting polynomial	Matching polynomial (where n the number of atoms)
$\sum_{k \geq 0} m(k) \cdot \lambda^k$	$\sum_{k \geq 0} (-1)^k \cdot m(k) \cdot \lambda^{n-2k}$

NANO-SCIENTIFIC DEVELOPMENT(S)

There are many methods (algorithms) for calculation of the characteristic polynomial and of its roots. A method with complexity of $O(n^4)$ extracts the coefficients one by one applying the Newton's identities (see Figure 1). Please note that the algorithm given below works only for the classical version of Charact-poly (ChP $\equiv |\lambda \cdot \text{Id} - \text{Ad}|$), where Trace(\cdot) function sums the elements on the main diagonal.

Input data: adjacency matrix ([Ad]) [Bx] \leftarrow [Ad] $c_0 \leftarrow 1$ For each k from 1 to n-1 do $c_k \leftarrow \text{Trace}([Bx])$ $c_k \leftarrow c_k \cdot (-1)/k$ [Bx] \leftarrow [Bx] - $c_k \cdot [\text{Id}]$ [Bx] \leftarrow [Ad] \times [Bx] End for $c_n \leftarrow \text{Trace}([Bx])$ $c_n \leftarrow c_n \cdot (-1)/n$ Output data: the series of the coefficients (c_k) $_{0 \leq k \leq n}$ for Charact-poly, ChP = $\sum_{0 \leq k \leq n} c_k \cdot \lambda^{n-k}$
--

Figure 1. Tracing the coefficients of the Charact-poly from adjacencies

When using the previous given algorithm, in order to avoid the lost of the precision with increasing of the number of atoms, someone should use the arbitrary-precision integer libraries for calculation (note that all coefficients of the Charact-poly are integers) of the arithmetic's given in the algorithm such as is bcmath (Morris & Cherry 1975²¹)

Nelson 1991²²).

It is possible to reduce the complexity of the calculation of Charact-poly by taking the advantage of its symmetry ($[Ad]_{i,j} = [Ad]_{j,i}$). Budde's method (Givens 1957²³) is one alternative. Following (Rehman & Ipsen 2011²⁴), the Budde's method requires first a tridiagonalization (Householder 1958²⁵) of the adjacency matrix (let us call it Td) and is requested a number of significant (of the leading degrees) coefficients. Figure 2 provides the algorithm.

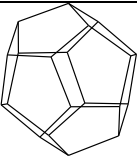
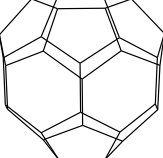
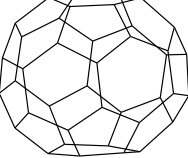
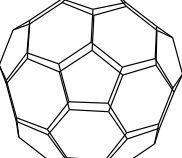
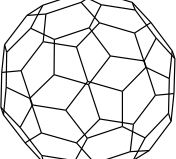
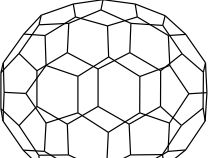
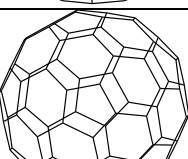
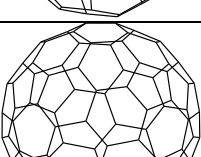
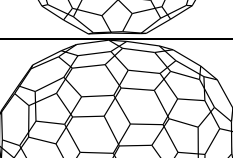
The main inconvenient of the previous given method (Budde's method) is that it requires the tridiagonalization of the adjacency, which it means that a series of operations including divisions are involved and the resulted matrix no more contains only integers, and therefore is lost the feature to work with arbitrary-precision integers and to extract the exact values of the coefficients. The results may come only as floating point numbers and the precision strongly depends by the number of operations involved, therefore by the number of atoms (n).

On the other hand, the using of arbitrary-precision integer libraries for calculation in conjunction with the algorithm given in Figure 1 its expected to increase the complexity of the calculation. Indeed, as was resulted from a study conducted on a series of fullerenes, the complexity becomes of order $O(n^4 \cdot \ln(n))$ - see Table 2, where some additional information is provided too, containing the Total strain energy (from continuum elasticity) in eV (Tománek 2014²⁶).

Input data: tridiagonalized matrix ([Td], as $(\alpha_j)_{1 \leq j \leq n}$ and $(\beta_j)_{2 \leq j \leq n}$)	
<pre> c₀ ← 1 c_{1,1} ← - α₁ c_{1,2} ← c_{1,1} - α₂ c_{2,2} ← α₁·α₂ - β₂·β₂ For each i from 3 to k do c_{1,i} ← c_{1,i} - α_i c_{2,i} ← c_{2,i} - α_i·c_{1,i-1} - β_i·β_i For each j from 3 to i-1 do c_{i,j} ← - α_i·c_{i-1,i-1} - β_i·β_i·c_{i-2,i-2} End for c_{i,i} ← - α_i·c_{i-1,i-1} - β_i·β_i·c_{i-2,i-2} End for For each i from k+1 to n do c_{1(i)} ← c_{1(i-1)} - α_i If k > 2 then c_{2,i} ← c_{2,i-1} - α_i·c_{1,i-1} - β_i·β_i For each j from 3 to i-1 do c_{i,j} ← - α_i·c_{i-1,i-1} - β_i·β_i·c_{i-2,i-2} End for End if End for c_{0,n} ← 1 Return (c_{i,n})_{0 ≤ i ≤ k} </pre>	$[Td] = \begin{bmatrix} \alpha_1 & \beta_2 & 0 & \dots & 0 \\ \beta_2 & \alpha_1 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta_n \\ 0 & \dots & 0 & \beta_n & \alpha_n \end{bmatrix}$
Output data: partial series of the coefficients, $(c_i)_{0 \leq i \leq k}$ for Charact-poly $ChP = \sum_{0 \leq i \leq n} c_i \cdot \lambda^{n-i}$	

Figure 2. Budde's first coefficients of the Charact-poly from adjacencies

Table 2. Calculation times of the Charact-poly on fullerenes

Fullerene	Additional information	Calculation times (s)
	Molecular formula: C ₂₀ Molecular symmetry: I _h Total strain energy: 24.204 (the only one topology) Isomers: none	Run 1: 0 Run 2: 1 Average: 0.5 Estimated: 0.2
	Molecular formula: C ₃₀ Molecular symmetry: C _{2v} Total strain energy: 25.204 (smallest value among isomers) Isomers: 3	Run 1: 2 Run 2: 2 Average: 2.0 Estimated: 1.7
	Molecular formula: C ₄₀ Molecular symmetry: C _{2v} Total strain energy: 25.684 (smallest value among isomers) Isomers: 40	Run 1: 8 Run 2: 7 Average: 7.5 Estimated: 7.0
	Molecular formula: C ₅₀ Molecular symmetry: D _{5h} Total strain energy: 25.474 (smallest value among isomers) Isomers: 271	Run 1: 20 Run 2: 20 Average: 20.0 Estimated: 19.7
	Molecular formula: C ₆₀ Molecular symmetry: I _h Total strain energy: 24.849 (the only one topology) Isomers: none	Run 1: 43 Run 2: 48 Average: 45.5 Estimated: 45.6
	Molecular formula: C ₇₀ Molecular symmetry: D _{5h} Total strain energy: 26.486 (the only one topology) Isomers: none	Run 1: 88 Run 2: 95 Average: 91.5 Estimated: 91.7
	Molecular formula: C ₈₀ Molecular symmetry: D _{5h} Total strain energy: 26.274 (smallest value among isomers) Isomers: 6	Run 1: 166 Run 2: 170 Average: 168.0 Estimated: 167.2
	Molecular formula: C ₉₀ Molecular symmetry: C ₂ Total strain energy: 30.066 (smallest value among isomers) Isomers: 46	Run 1: 282 Run 2: 283 Average: 282.5 Estimated: 283.0
	Molecular formula: C ₁₀₀ Molecular symmetry: D ₅ Total strain energy: 30.446 (smallest value among isomers) Isomers: 450	Run 1: 455 Run 2: 449 Average: 452.0 Estimated: 452.0

For reproducibility of the study, the calculation was conducted on a 2.33 Ghz dual core computer by running a single core tasked program (a PHP implementation) for the data and the results given in Table 2. The variability among execution times can be assigned to the multitasking operating system (which runs in background other tasks too) as well as to the CPU's cache memory (cached in 2 levels, 2x2x32 kBytes in the first and 4096 in the second).

The estimated times from Table 2 are from the best (among alternatives) fit, as is given in Table 3 (where independent variable was tenth part of the number of atoms, $x = n_c/10$, and the dependent variable was $y = \text{time}$, in seconds).

Table 3. Calculation complexity of the Charact-poly on fullerenes

Model	Coefficients & significances	Statistics	Remarks
$\hat{y} = a \cdot x^4 + b$	$a = 0.045$ ($t_a = 58$); $b = -7.98$ ($t_b = 2.39$)	$r^2 = 0.99794$; see = 7.6	$p_b \approx 5\%$
$\hat{y} = a \cdot x^5 + b$	$a = 0.0045$ ($t_a = 58$); $b = 7.45$ ($t_b = 2.20$)	$r^2 = 0.99765$; see = 8.1	$p_b \approx 6\%$
$\hat{y} = a \cdot x^4 \cdot \ln(x) + b$	$a = 0.02$ ($t_a = 1134$); $b = 0.26$ ($t_b = 1.53$)	$r^2 = 0.99999$; see = 0.4	$p_b \approx 17\%$
$\hat{y} = a \cdot x^4$	$a = 0.044$ ($t_a = 59$)	$r^2 = 0.996$; see = 9.6	see below
$\hat{y} = a \cdot x^5$	$a = 0.0046$ ($t_a = 58$)	$r^2 = 0.996$; see = 9.9	
$\hat{y} = a \cdot x^4 \cdot \ln(x)$	$a = 0.01963$ ($t_a = 1346$)	$r^2 = 0.999996$; see = 0.4	

As can be concluded from the values given in Table 3, all the models with intercept ($\hat{y} = a \cdot x^4 + b$; $\hat{y} = a \cdot x^5 + b$; $\hat{y} = a \cdot x^4 \cdot \ln(x) + b$) are susceptible to have the intercept (the b coefficient) not significantly different from zero - actually it is the expected result, since a fullerene with a zero size requires no calculations. Thus, is perfectly justified to try and use the models without intercept. Before to proceed, it is something else which it should keep our attention. Under assumption that exist intercept, then this should be seen as a small time required by the program implementing the algorithm for initializations and for displaying the results. But looking at the models, only one of them proposes a small value for that time, namely $b = 0.26$ s in the model $\hat{y} = a \cdot x^4 \cdot \ln(x) + b$, which it means that if it is a trustable model without intercept, this should be the one. Indeed, by conducting the analysis without intercept, the results sustain the hypothesis. The standard error of estimate (see) remains almost unchanged when the intercept is removed only for this model. Therefore, the best guess for the approximation of the complexity of the algorithm for the calculation of the Charact-poly with arbitrary-precision integers given in Figure 1 is of $O(n^4 \cdot \ln(n))$. For reproducibility of the study, the calculation was conducted on a 2.33 Ghz dual core computer by running a single core tasked program (a PHP implementation) for the data and the results given in Table 2.

NANO-CHEMICAL APPLICATION(S)

In classical molecular topology the atoms are considered undistinguishable and are represented as vertices, the bonds are considered unweighted and are represented as edges and the obtained molecular graphs are unweighted and unoriented. In this context, the set of the bonds (edges) is a subset of the Cartesian product of the set of the atoms (vertices) by itself and the molecules (graph) is defined as the collection of the set of vertices and of the set of edges (see Table 4).

Table 4. Classical molecular graph

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

There is something to consider when discuss calculations on molecular graphs. Thus, with

the increasing of the simplification in the molecular graph representation (neglecting type of the atom, bond orders, geometry in the favour of topology) increases the degeneration of the whole pool of possible calculations (descriptors) on the graph structure - existing more and more molecules possessing same representation as molecular graph. This consequence is favourable for the problems seeking for similarities and is unfavourable for the problems seeking for dissimilarities.

A necessary step to accomplish a better coverage of the similarity vs. dissimilarity dualism is to build and use a family of molecular descriptors, large enough to be able to provide answers for the all (by its individuals) when is feed with molecules datasets. On the natural way a such kind of family should posses a 'genetic code' - namely a series of variables of which values to (re)produce a (one by one) molecular descriptor, all descriptors being therefore obtained on same way (being breed in the family). All individuals of the family should be independent of the numbering of the atoms in molecule (should be molecular invariants).

Since these are all restrictions applying, may seem a simple construction, but it is in same time a complex one. First, is obviously that such construction - the family of molecular descriptors - can be build only with the help of the computer, because requires a great number of operations repeatedly (with different augments) applied on the same molecular structure. The molecular geometry should be considered too, and for this reason (of obtaining of the models for the molecular geometry) this subject will be continued in a later section.

In order to reflect the topology of a graph structure, three adjacency matrices can be built. If we store the full graph (each pair of vertices stored twice, in both ways) then the rectangular matrices reflects 1:1 the graph (these matrices are more convenient when we do matrix operations). The matrices of vertex adjacency and of edges adjacency are square matrices and the enumerating twice the edges is reflected in symmetry of the matrix relative to its main diagonal, which can be rebuild in the absence of the representation, by having only the lists of vertices and edges (see Table 4).

An extremely important problem in chemistry is to identify uniquely a chemical compound. If the visual identification (looking on the structure) seems simple, for compounds of large size this alternative is no more viable. The data of the structure of the compounds stored into the informational space may provide the answer to this problem. Together with the storing of the structure of the compound other issue is raised, namely the arbitrary in the numbering of the atoms. Namely for a chemical structure with N atoms stored as a (classical molecular) graph exists exactly $N!$ possibilities of different numbering of the atoms. Unfortunately, storing the graphs as lists of edges (and eventually of vertices) does not provide a direct tool to check this arbitrary differentiation due to the numbering. The same situation applies on the adjacency matrices.

Therefore, seeking for graph invariants is perfectly justified: an invariant (graph invariant) does not depend on numbering. The adjacency matrix is not a graph invariant (and very simple examples may be created instantly to proof this). The ideal situation is that the invariant to be unique assigned to each (and any) structure, but this kind of invariants are very hardly to be found.

A procedure to generate an no degenerated invariant is proposed by IUPAC as the international chemical identifier (InChI) which converts the chemical structure to a table of connectivity expressed as a unique and predictable series of characters (McNaught 2006²⁷).

An important class of graph invariants are the graph polynomials. To this category belongs the characteristic polynomial, a graph invariant encoding important properties of the graph. Unfortunately, does not represent a bijective image of the graph, existing different graphs with same characteristic polynomial (cospectral graphs), and smallest cospectral graphs occurs for 5 vertices (Von Collatz & Sinogowitz 1957²⁸). In order to count the cospectral graphs, one should

compare A000088 (Sloane 1996²⁹) and A082104 (Weisstein 2003³⁰) integer sequences.

Let's take a chemical compound, namely hexamine. (Pubchem CID: 4101). Hexamine ($C_6H_{12}N_4$) it uses in the production of powdery or liquid preparations of phenolic resins and phenolic resin moulding compounds. It has been proposed that hexamethylenetetramine could work as a molecular building block for self-assembled molecular crystals (Markle 2000³¹). It has a cage-like structure similar to adamantane and its representation is given in Figure 3.

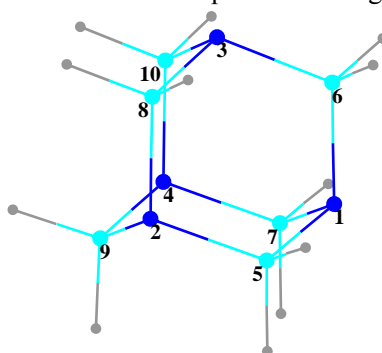


Figure 3. Hexamine

In Figure 3 the hydrogen atoms are represented with grey (and are not numbered), carbon with light blue, and nitrogen with blue.

Let's take the representation in a matrix form by its adjacency matrix by taking it conventionally without attached hydrogen atoms as well as by its distance matrix in two scenarios: topological and geometrical distances. Please note that this simple case of a molecule, but even here the geometrical distance is with a totally different meaning than the topological distance. The resulted matrices are given in the Figure 4.

Ad	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	$3^{1/46}$	$3^{1/46}$	$3^{1/46}$	0	0	0
2	0	0	0	0	$3^{1/46}$	0	0	$3^{1/46}$	$3^{1/46}$	0
3	0	0	0	0	0	$3^{1/46}$	0	$3^{1/46}$	0	$3^{1/46}$
4	0	0	0	0	0	0	$3^{1/46}$	0	$3^{1/46}$	$3^{1/46}$
5	$3^{1/46}$	$3^{1/46}$	0	0	0	0	0	0	0	0
6	$3^{1/46}$	0	$3^{1/46}$	0	0	0	0	0	0	0
7	$3^{1/46}$	0	0	$3^{1/46}$	0	0	0	0	0	0
8	0	$3^{1/46}$	$3^{1/46}$	0	0	0	0	0	0	0
9	0	$3^{1/46}$	0	$3^{1/46}$	0	0	0	0	0	0
10	0	0	$3^{1/46}$	$3^{1/46}$	0	0	0	0	0	0

Ad	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	$3^{1/46}$	$3^{1/46}$	$3^{1/46}$	0	0	0
2	0	0	0	0	$3^{1/46}$	0	0	$3^{1/46}$	$3^{1/46}$	0
3	0	0	0	0	0	$3^{1/46}$	0	$3^{1/46}$	0	$3^{1/46}$
4	0	0	0	0	0	0	$3^{1/46}$	0	$3^{1/46}$	$3^{1/46}$
5	$3^{1/46}$	$3^{1/46}$	0	0	0	0	0	0	0	0
6	$3^{1/46}$	0	$3^{1/46}$	0	0	0	0	0	0	0
7	$3^{1/46}$	0	0	$3^{1/46}$	0	0	0	0	0	0
8	0	$3^{1/46}$	$3^{1/46}$	0	0	0	0	0	0	0
9	0	$3^{1/46}$	0	$3^{1/46}$	0	0	0	0	0	0
10	0	0	$3^{1/46}$	$3^{1/46}$	0	0	0	0	0	0

bonds represented undistinguishable (with 1)	bonds represented from geometrical distances (inverse of the distance in Å)
--	---

Figure 4. Different adjacency matrices representing hexamine

The unity (or identity) matrix stores 1 on the main diagonal and is easy to be extended to store a atomic property (such as something in relation with atomic mass, electronegativity, partial charge or even the number of attached hydrogen atoms, when also 0 is allowed) when the new matrices continues to have all non-null values on the main diagonal, but can be different from one and different one to each other depending now from the atom type. The result is exemplified in Figure 5 (where electronegativity is taken from Pauling scale and was divided by 4). The general idea when the weights was chosen in the identity matrices exemplified in Figure 5 is to have (almost everywhere) subunitary numbers, because when on these matrices is applied the procedure of calculation of the characteristic polynomial, then for large molecules numbers greater than 1 rapidly produces big numbers as coefficients as well as outcomes of the evaluation

of the polynomial.

	Ad	1	2	3	4	5	6	7	8	9	10		Ad	1	2	3	4	5	6	7	8	9	10	
	1	1	0	0	0	0	0	0	0	0	0		1	.75	0	0	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0	0		2	0	.75	0	0	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0	0		3	0	0	.75	0	0	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	0	0	0		4	0	0	0	.75	0	0	0	0	0	0	0
	5	0	0	0	0	1	0	0	0	0	0		5	0	0	0	0	.625	0	0	0	0	0	0
	6	0	0	0	0	0	1	0	0	0	0		6	0	0	0	0	0	.625	0	0	0	0	0
	7	0	0	0	0	0	0	1	0	0	0		7	0	0	0	0	0	0	.625	0	0	0	0
	8	0	0	0	0	0	0	0	1	0	0		8	0	0	0	0	0	0	0	.625	0	0	0
	9	0	0	0	0	0	0	0	0	1	0		9	0	0	0	0	0	0	0	0	.625	0	0
	10	0	0	0	0	0	0	0	0	0	1		10	0	0	0	0	0	0	0	0	0	0	.625

atoms represented with 1 (undistinguishable) | atoms represented by electronegativity (distinguishable)

Figure 5. Different identity matrices representing hexamine

Based on the modified forms of the adjacency and identity matrices, the extension of the formula of the characteristic polynomial is immediate:

$$P_{\varphi, A_P, M_O}(\lambda) = P_{\varphi}(\lambda, G) = |\lambda \cdot \text{Id}(A_P) - \text{Ad}(M_O)|$$

where M_O is a certain metric operator (as were exemplified in Figure 4) and A_P is a certain atomic property (as were exemplified in Figure 5). For a single molecule it results a series of the polynomial formulas (given in the next for a clear reading as determinants) which can be evaluated for different values of the argument (X). The next figure exemplifies the calculation for hexamine.

	M_O from classical topology	M_O from geometry ($a = -31/46$)
A_P from classical topology	$\begin{vmatrix} \lambda & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 & -1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & \lambda & 0 & 0 & -1 & 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & \lambda & 0 & 0 & -1 & 0 & -1 & -1 \\ -1 & -1 & 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 & \lambda & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 0 & \lambda & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & \lambda & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & \lambda \end{vmatrix}$	$\begin{vmatrix} \lambda & 0 & 0 & 0 & a & a & a & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 & a & 0 & 0 & a & a & 0 \\ 0 & 0 & \lambda & 0 & 0 & a & 0 & a & 0 & a \\ 0 & 0 & 0 & \lambda & 0 & 0 & a & 0 & a & a \\ a & a & 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ a & 0 & a & 0 & 0 & \lambda & 0 & 0 & 0 & 0 \\ a & 0 & 0 & a & 0 & 0 & \lambda & 0 & 0 & 0 \\ 0 & a & a & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & a & 0 & a & 0 & 0 & 0 & 0 & \lambda & 0 \\ 0 & 0 & a & a & 0 & 0 & 0 & 0 & 0 & \lambda \end{vmatrix}$
A_P from electronegativities ($b = 5/8; c = 3/4$)	$\begin{vmatrix} c \cdot \lambda & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & c \cdot \lambda & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & c \cdot \lambda & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & c \cdot \lambda & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & b \cdot \lambda & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & b \cdot \lambda & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & b \cdot \lambda & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & b \cdot \lambda & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & b \cdot \lambda & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & b \cdot \lambda \end{vmatrix}$	$\begin{vmatrix} c \lambda & 0 & 0 & 0 & a & a & a & 0 & 0 & 0 \\ 0 & c \lambda & 0 & 0 & a & 0 & 0 & a & a & 0 \\ 0 & 0 & c \lambda & 0 & 0 & a & 0 & a & 0 & a \\ 0 & 0 & 0 & c \lambda & 0 & 0 & a & 0 & a & a \\ a & a & 0 & 0 & b \lambda & 0 & 0 & 0 & 0 & 0 \\ a & 0 & a & 0 & 0 & b \lambda & 0 & 0 & 0 & 0 \\ a & 0 & 0 & a & 0 & 0 & b \lambda & 0 & 0 & 0 \\ 0 & a & a & 0 & 0 & 0 & 0 & b \lambda & 0 & 0 \\ 0 & a & 0 & a & 0 & 0 & 0 & 0 & b \lambda & 0 \\ 0 & 0 & a & a & 0 & 0 & 0 & 0 & 0 & b \lambda \end{vmatrix}$

Figure 6. Different characteristic-like polynomials for the chemical structure of hexamine

The outcome of the regression analysis is a model with a certain explanatory power. This explanatory power is influenced by the number of the coefficients included in the model (n_c), as well as by the number of independent variables (n_d) used to explain the association when the model was feed with a certain number of molecules (m), and therefore the adjusted value (r^2_{adj}) of the correlation coefficient (r^2) provides a ordering of the explanatory powers:

$$r_{\text{adj}}^2 = r^2 - (1 - r^2) \frac{n_d}{m - n_c}$$

The use of the extended characteristic polynomial is exemplified on a series of 45 C₂₀ fullerene congeners which were obtained by replacing the carbon atom with nitrogen and boron by a certain pattern which is illustrated in Figure 7 (S1 to S4 are shells; on each shell are atoms of same type).

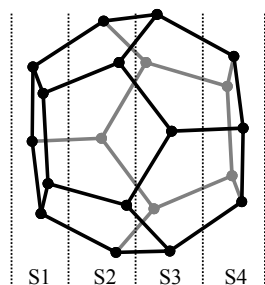


Figure 7. Pattern for generation of C₂₀ fullerene congener structures

Are 4³ (64) possible arrangements of Carbon, Nitrogen and Boron in the sites defined by the shells S1 to S4 in Figure 7, but some of them defines identical molecules as long as the structure is free to move (and rotate). Due to this fact, there are only 45 different structures. The structures were drawn and stored in separate files. The geometries were built at HF 6-31G level of theory and a series of calculated properties were collected for them and are given in the Table 5 along with the file name (named accordingly to the design from Figure 7, where Homo: highest occupied molecular orbital energy, in eV; Lumo: lowest unoccupied molecular orbital energy, in eV; Pola: polarizability, in 10⁻³⁰ m³, DipM: dipole moment, in Debye).

Table 5. Selected properties from HF 6-31G calculations

Mol	Homo	Lumo	Pola	DipM	Mol	Homo	Lumo	Pola	DipM
bbbb	-0.2461	-0.0385	60.469	0.003541	cbnn	-0.3918	0.0903	55.059	8.237555
bbbn	-0.2546	-0.095	61.208	1.189393	ccbb	-0.2697	-0.0317	61.306	6.250461
bbcn	-0.2988	-0.0764	60.081	9.954801	ccbc	-0.3159	0.0574	59.911	0.713889
bbnb	-0.2959	-0.0608	60.272	1.381921	ccbn	-0.3237	0.074	57.661	5.714922
bbnn	-0.2747	-0.0603	58.602	7.083818	cccb	-0.2589	-0.0696	61.359	1.960172
bcbb	-0.2471	-0.0647	62.309	2.179006	cccc	-0.3843	0.1622	58.487	0.000832
bcbn	-0.3083	-0.0407	59.187	6.785335	cccn	-0.3122	0.1522	57.091	7.334984
bccb	-0.2036	-0.1014	62.714	0.002733	ccnb	-0.3265	0.0326	58.372	3.414014
bccn	-0.2994	-0.039	59.196	9.030429	ccnc	-0.3363	0.1803	56.784	1.457706
bcnb	-0.3023	-0.0469	59.881	5.572056	ccmn	-0.3565	0.1456	54.976	10.170182
bcnn	-0.2578	-0.1307	58.182	0.680023	cnbb	-0.3004	-0.087	59.378	3.2186
bnbn	-0.3545	-0.0214	54.813	6.939198	cnbn	-0.3617	0.1144	54.798	8.288544
bncn	-0.3513	-0.006	56.543	4.948867	cncb	-0.3145	-0.0319	59.111	3.170551
bnnb	-0.3466	-0.0157	57.038	0.008881	cncn	-0.3547	0.171	54.802	6.884452
bnnn	-0.3903	-0.0527	54.572	6.824007	cnnb	-0.3394	-0.0048	56.511	3.000425
cbbb	-0.2621	-0.0498	61.333	3.939499	cnnc	-0.334	0.1663	54.992	0.001746
cbbc	-0.2475	0.026	60.198	0.003295	cnnn	-0.3883	0.1159	53.051	9.415604
cbbn	-0.221	-0.0178	59.654	3.793689	nbbn	-0.2199	-0.0619	57.786	0.019205
cbcb	-0.2836	-0.006	61.228	0.128928	nbnn	-0.4057	0.0636	52.938	0.768984
cbcn	-0.3008	0.0269	58.345	8.600025	ncbn	-0.3241	0.0499	55.991	2.259238
cbnb	-0.3269	0.0075	58.474	0.04766	nccn	-0.3292	0.1694	54.967	0.003738
cbnc	-0.3749	0.1273	56.843	0.89975	ncnn	-0.3882	0.1361	52.906	1.953068
					nmmn	-0.454	0.0624	51.084	0.002512

The calculations for the extended characteristic polynomial were conducted diversifying the atomic property in 8 levels, as given in Table 6.

Table 6. Atomic properties included in the extension of the Charact-poly

'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)

The calculations for the extended characteristic polynomial were conducted diversifying the adjacency in 3 levels, as well as were used the distance matrix in place of the adjacency when the diversification were produced in 6 levels, as given in Table 7.

Table 7. Adjacency weights included in the extension of the Charact-poly

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

A FreePascal program were build to split the work in parallel depending on the number of processors available splitting the job by properties (no more than 8 parallel tasks can be produced). A huge file containing the descriptors names as well as calculated values of the polynomials for the series of the molecules is produced. The descriptors are named as is described in the Table 8.

Table 8. Names of the descriptors calculated using the extended characteristic polynomial

Variable	Description
L ₁ L ₂ L ₃ L ₄ d ₁ d ₂ d ₃ d ₄	8 characters, first 4 being letters, last 4 being digits
d ₁ d ₂ d ₃ d ₄	$\overline{d_1d_2d_3d_4}$ ranges from 0000 to 1000; is evaluated $P_{\varphi, L_2, L_3}(\pm \overline{d_1d_2d_3d_4} / 1000)$
L ₁	is 'T' when the evaluated polynomial value is unchanged ($f(x)=x$); is 'R' when reciprocal ($f(x)=1/x$) value of the evaluated polynomial is calculated; is 'L' when logarithm ($f(x)=\ln(x)$) value of the evaluated polynomial is calculated;
L ₂	encodes the atomic property used to diversify the identity matrix (see Table 6)
L ₃	encodes the metric operator used to diversify the adjacency matrix (see Table 7)
L ₄	encoding for negative (N) or positive (P) argument of the polynomial

It is expected that a diversification like the one considered here to produce degenerations, namely identical values of the descriptors for different descriptor names, which is the case in the dataset considered here. For instance, one degeneration is immediate, namely between the classical topological calculation when 1 is encoded in the adjacency matrix and the diversified calculation in which the inverse of the bond order is considered, because all bonds in the congener series are single bonds (see Figure 7). This is the reason for which, before regression analysis involving the properties from HF 6-31G calculations, a filtering program was designed to look for degenerations and to reduce the pool of descriptors eliminating them. Because also this program works in parallel, without channelling between tasks, may still contain few degenerations, and another program were designed to sort the data (acting in parallel, but all tasks are writing in a common shared output file) to clean the descriptors pool by repeated series of identical values.

The next stage is to test the descriptors in regard to their ability to make simple linear

regressions, but very different square deviations when the data are normalized have no meaning when a linear relation with a measured property is desired. Therefore, in a stage the values are normalized and compared the two distributions (of the property and of the descriptor) and the association is rejected for a departure with a probability of association less than 1% (this procedure should be called normalization) and in the last stage simple linear associations are obtained (when some descriptors are removed when possess more than 100 times or less than 100 times variance than the observed).

The Table 9 contains the statistics of the descriptors generated, where stage refers the stage applied on the pool of descriptors.

Table 9. Statistics from preliminary treatment of the descriptors pool

Stage	Number of descriptors	Remarks
Generating	272124	always using the defined configuration
Filtering	235530	degeneration depends on the complexity of the dataset
Sorting	230450	degeneration depends on the parallelization level
Normalization	140447	for Homo
Normalization	147071	for Lumo
Normalization	139902	for Pola
Normalization	146134	for DipM
Simple linear regression	114936	for Homo
Simple linear regression	115747	for Lumo
Simple linear regression	119249	for Pola
Simple linear regression	93299	for DipM

A remark is immediate about the results given in Table 9: even if the design of the diversification admits the degeneracy (see Table 8) as well as the dataset is a simple pattern on which the congener series were constructed (see Figure 7) the level of degeneracy is very low (the descriptors pool size is reduced from 272124 to 230450 till the sorting stage included, which is about 84.7% of its initial size; a reduction to 40.7% in average for the last stage of simple linear regressions) and when is compared with other families of descriptors, such as MDF, for similar sample sizes the reduction is much less (in [32] on 40 compounds from 787968 descriptors after a similar filtering remained 70943 which is about 9%).

Regression was conducted with one and two dependent variables taking into account additive and/or multiplicative effects among them. The analysis produced more than one possible outcome for each case, and were selected the best candidate regressions with the highest explanatory power on the molecules subject to investigation for the property with which the model was feed. The equations are given in the Table 10.

Table 10. Regressions with highest explanatory power

No	Property	Model	Descriptors	Coefficients	r^2	r^2_{adj}
1	Dipole moment	$\hat{Y} = a \cdot X_1 + d$	$X_1 = \text{REtN0841}$	$a = -0.1410$ (t = 5.74) $d = 3.742$ (t = 9.78)	0.434	0.420
2	Dipole moment	$\hat{Y} = c \cdot X_1 \cdot X_2 + d$	$X_1 = \text{RDtP0066}$ $X_2 = \text{IDtP0087}$	$c = 4.221$ (t = 8.85) $d = 1.555$ (t = 4.04)	0.651	0.635
3	Dipole moment	$\hat{Y} = a \cdot X_1 + b \cdot X_2 + d$	$X_1 = \text{LFgN0609}$ $X_2 = \text{IDcP0908}$	$a = 5.328 \cdot 10^1$ (t = 7.10) $b = 5.434 \cdot 10^1$ (t = 8.89) $d = 7.544$ (t = 15.1)	0.689	0.675
4	Dipole moment	$\hat{Y} = a \cdot X_1 + b \cdot X_2 + c \cdot X_1 \cdot X_2 + d$	$X_1 = \text{LFgN0612}$ $X_2 = \text{RDtN0065}$	$a = -3.103 \cdot 10^1$ (t = 3.78) $b = -5.126 \cdot 10^1$ (t = 8.79) $c = -6.851 \cdot 10^2$ (t = 8.59) $d = 1.791$ (t = 4.54)	0.725	0.711
5	HOMO energy	$\hat{Y} = a \cdot X_1 + d$	$X_1 = \text{LFgP0454}$	$a = -0.04027$ (t = 8.79) $d = -0.4028$ (t = 36.6)	0.643	0.634
6	HOMO energy	$\hat{Y} = c \cdot X_1 \cdot X_2 + d$	$X_1 = \text{IDTP0633}$ $X_2 = \text{IDTP0653}$	$c = -7.434 \cdot 10^{-7}$ (t = 11.5) $d = -2.755 \cdot 10^{-1}$ (t = 50.6)	0.758	0.747

No	Property	Model	Descriptors	Coefficients	r^2	r^2_{adj}
7	HOMO energy	$\hat{Y} = a \cdot X_1 + b \cdot X_2 + d$	$X_1 = \text{RGTN0155}$ $X_2 = \text{LBgN0874}$	$a = -0.01407$ (t = 6.49) $b = -0.1827$ (t = 8.26) $d = -0.9522$ (t = 12.1)	0.808	0.799
8	HOMO energy	$\hat{Y} = a \cdot X_1 + b \cdot X_2 + c \cdot X_1 \cdot X_2 + d$	$X_1 = \text{RGCN0182}$ $X_2 = \text{LBgP0165}$	$a = 1.060 \cdot 10^{-1}$ (t = 4.80) $b = 6.9852 \cdot 10^{-2}$ (t = 11.0) $c = 2.4080 \cdot 10^{-2}$ (t = 5.51) $d = -6.947 \cdot 10^{-1}$ (t = 19.7)	0.830	0.822
9	LUMO energy	$\hat{Y} = a \cdot X_1 + d$	$X_1 = \text{LBgN0566}$	$a = 1.214$ (t = 8.79) $d = 0.1019$ (t = 8.33)	0.643	0.634
10	LUMO energy	$\hat{Y} = c \cdot X_1 \cdot X_2 + d$	$X_1 = \text{IHGN0132}$ $X_2 = \text{IHGN0157}$	$c = 5.972 \cdot 10^{-11}$ (t = 13.7) $d = -5.854 \cdot 10^{-2}$ (t = 7.15)	0.816	0.808
11	LUMO energy	$\hat{Y} = a \cdot X_1 + b \cdot X_2 + d$	$X_1 = \text{LGGN0094}$ $X_2 = \text{LBgN0584}$	$a = 7.458 \cdot 10^{-2}$ (t = 7.27) $b = 1.008$ (t = 10.0) $d = 6.230 \cdot 10^{-1}$ (t = 8.72)	0.830	0.821
12	LUMO energy	$\hat{Y} = a \cdot X_1 + b \cdot X_2 + c \cdot X_1 \cdot X_2 + d$	$X_1 = \text{IHGN0150}$ $X_2 = \text{IHGN0131}$	$a = 4.098 \cdot 10^{-6}$ (t = 10.9) $b = 4.103 \cdot 10^{-6}$ (t = 11.6) $c = 6.936 \cdot 10^{-11}$ (t = 12.9) $d = -6.481 \cdot 10^{-2}$ (t = 7.97)	0.840	0.832
13	Polarizability	$\hat{Y} = a \cdot X_1 + d$	$X_1 = \text{IATN0079}$	$a = -38.50$ (t = 18.9) $d = 500.4$ (t = 21.4)	0.893	0.890
14	Polarizability	$\hat{Y} = c \cdot X_1 \cdot X_2 + d$	$X_1 = \text{LFTN0225}$ $X_2 = \text{LBGP0631}$	$c = 25.09$ (t = 24.9) $d = 71.64$ (t = 126)	0.937	0.934
15	Polarizability	$\hat{Y} = a \cdot X_1 + b \cdot X_2 + d$	$X_1 = \text{LBgN0419}$ $X_2 = \text{LGGN0488}$	$a = -2.896$ (t = 26.1) $b = 0.6533$ (t = 12.7) $d = 37.82$ (t = 43.1)	0.959	0.957
16	Polarizability	$\hat{Y} = a \cdot X_1 + b \cdot X_2 + c \cdot X_1 \cdot X_2 + d$	$X_1 = \text{LDGN0394}$ $X_2 = \text{LDGN0402}$	$a = 760.6$ (t = 14.9) $b = 735.4$ (t = 13.8) $c = -1.499$ (t = 5.76) $d = -58.45$ (t = 3.34)	0.963	0.961

In all cases the best model selections were with both additive and multiplicative effects included ($\hat{Y} = a \cdot X_1 + b \cdot X_2 + c \cdot X_1 \cdot X_2 + d$), and therefore someone can say that the association between the structure as can be described by the characteristic polynomial and the dipole moment, HOMO and LUMO energies and polarizability is hardly to be considered as being purely linear. Also in all cases linear models ($\hat{Y} = a \cdot X_1 + b \cdot X_2 + d$) were selected as the second best alternative in all cases in disfavour of the multiplicative effects models ($\hat{Y} = c \cdot X_1 \cdot X_2 + d$), suggesting that however the linear component is the predominant one. Test results for significant differences among explanatory powers of the models (by using of Fisher Z transformation) are given in Table 11.

Table 11. Z values for comparison of explained variances for models from Table 10
 $z(r^2_{adj}) = \text{arctanh}(\sqrt{r^2_{adj}})$; $\sigma(r^2_{adj}) = 1/\sqrt{(45-3)}$; $z_{i,j} = (z_i - z_j)/\sigma\sqrt{2}$; $z(2.5\%) = 1.96$

Dipole moment				HOMO energy				LUMO energy				Polarizability							
$z_{i,j}$	1	2	3	4	$z_{i,j}$	5	6	7	8	$z_{i,j}$	9	10	11	12	$z_{i,j}$	13	14	15	16
1					5					9					13				
2	9.78				6	6.82				10	11.6				14	8.21			
3	12.0	2.20			7	10.8	4.02			11	12.8	1.19			15	15.0	6.77		
4	14.1	4.37	2.17		8	12.9	6.09	2.07		12	13.9	2.27	1.07		16	16.5	8.30	1.53	

Comparison of the explanatory powers of the models reveals only three cases in which the differences are not significant, namely for LUMO energy when multiplicative or additive effects are used to explain it, and when additive of both additive and multiplicative effects are used to explain it, and for polarizability when additive or both additive and multiplicative effects are used to explain it. In these cases, using of the larger models (with more coefficients) are not fully justified statistically based on what may come from a by chance association.

For the Dipole moment (see Table 10) using of a model with multiplicative effects only in

association with the structure of the compounds selects as a best pair a reciprocal of a characteristic polynomial value (RDtP0066) and a directly proportional one (IDtP0087) while using a model with additive effects only selects as a best pair a logarithm of a characteristic polynomial value (LFgN0609) and a directly proportional one (IDcP0908). Interesting is the fact that when a full additive and multiplicative effects model is used, it is kept the transformation of the descriptors from both partial effects models, being selected reciprocal (RDtN0065) and a logarithmic (LFgN0612) transformed descriptors, when not only the transformation is kept, it is kept also the polynomial formula ("Dt" in RDtP0066 and in RDtN0065; "Fg" in LFgN0609 and in LFgN0612). The only difference is at values in which the polynomials are evaluated (RDtN0065 evaluates the "Dt" polynomial in $-65/1000$ while RDtP0066 evaluates it in $66/1000$; LFgN0612 evaluates the "Fg" polynomial in $-612/1000$ while LFgN0609 evaluates it in $-609/1000$). The model reveals an association of the dipole moment with the first ionization potential dependent on geometry and solid state density dependent on topology, being able to explain about 71.1% of the variability by these factors.

Looking at HOMO energy (see Table 10) the additive effects model and the full multiplicative and additive effects model have the same composition of descriptors (RGTN0155 and LBgN0874 for additive effects only; RGCN0182 and LBgP0165 for the full effects one). While looking to the descriptors values in the original files in which were placed the evaluation results, one can find that actually RGCN0155 and RGTN0155 provides same series of values for the molecules on which the calculation were applied. Therefore, in this case, the additive effects model and the full effects model possess the same characteristic polynomials formulas ("GC" and "Bg") and the same operations ("R" and "L") on it, the polynomials being only evaluated in different points (RGCN0182 evaluates the "GC" polynomial in $-182/100$ while RGCN0155 evaluates it in $-155/1000$; LBgP0165 evaluates "Bg" polynomial in $165/1000$ while LBgN0874 evaluates it in $-874/1000$). Since no change in the composition of the selected model is observed in this case of HOMO energy, it can be said that the multiplicative effects have a minor influence on the values of HOMO energy when this is related with the structure with two characteristic polynomials. The obtained model shows an association of HOMO energy with the melting points and the geometry of the molecules, being able to explain with them about 79.9% of the variability.

For the LUMO energy the situation is totally reversed than for the HOMO energy. The full model borrows its composition from the model with multiplicative effects only (IHGN0132 and IHGN0157 descriptors in multiplicative model; IHGN0131 and IHGN0150 descriptors in full model). Therefore since no change in the composition of the selected model is observed in this case of LUMO energy, it can be said that the additive effects have a minor influence on the values of LUMO energy when this is related with the structure with two characteristic polynomials. The obtained model shows that for LUMO energy the attached hydrogen atoms and the geometry of the molecule plays an important role, and the model is able to explain at least 80.8% of the total variability using these factors.

When looking at polarizability, all additive, multiplicative and full models keeps the same operation (logarithm, "L" letter at the beginning of the descriptors names) while the composition is subject to change from one model to another. It is also interesting that the full model selects the same polynomial for the both descriptors ("DG" in both LDGN0394 and LDGN0402) 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. The association between the solid state densities as atomic properties ("D" letter) and polarizability as estimated property requires further studies designed to see more about. The model based on solid state density and geometry of the molecule is able to explain

about 96.1% of the total variability.

MULTI-/TRANS- DISCIPLINARY CONNECTION(S)

Other one connected polynomial with the Charact-poly is the Laplacian polynomial which use a modified form of the adjacency matrix ([Ad]), the Laplacian matrix ([La]), calculated as $[La] = [Dg] - [Ad]$, where [Dg] simply counts on the main diagonal the number of atom's bonds (the rest of its elements are null; for convenience with the graph theory related concept were noted [Dg] - from vertex degree). The Laplacian polynomial is the Charact-poly of the Laplacian matrix:

$$LaP \stackrel{\text{def}}{=} |\lambda \cdot Id - La| = |\lambda \cdot Id - Dg + Ad|$$

The Laplacian matrix is often used in the analysis of electrical networks. The roots of the Laplacian polynomial uses too, under the name of Laplacian spectra.

OPEN ISSUES

Based on the conducted study the extension of the characteristic polynomial to take into account the type of the atom when is counted as a vertex in its classical approach and to take into account the type of the bond when is counted as a adjacency in its classical approach, as well as the alternative use of the distance matrix, computed by topologies as well as by geometries are fruitful extensions. The case study reveals that useful information may be bring out from the structure-property and structure-activity study when the extended characteristic polynomial is used. Some disappointments can be recorded as well, one of them being the relatively low (when compared with other family-based derived descriptors) as well as the inconvenience of the calculation of the polynomial for values of the argument outside of the [-1,1] interval, when results of the calculation goes outside of the precision of calculation for any reasonable sized molecule.

RELATED LIST OF ABBREVIATIONS

The term secular function has been used for what is now called characteristic polynomial (in some literature the term secular function is still used). The term comes from the fact that the characteristic polynomial was used to calculate secular perturbations (on a time scale of a century, i.e. slow compared to annual motion) of planetary orbits, according to Lagrange's theory of oscillations.

Sachs graphs (Sachs 1962³³) is a possible enumeration of what the characteristic polynomial counts as authors of (Graovac et al. 1972³⁴) observed.

REFERENCES AND FURTHER READING

¹ Huckel, E.; 1931. Quantentheoretische Beiträge zum Benzolproblem, Z. Phys. 70: 204-286.

² Schrödinger, E.; 1926. An undulatory theory of the mechanics of atoms and molecules. Physical Review 28(6): 1049-1070.

-
- ³ Coulson C.A.; 1937. The evaluation of certain integrals occurring in studies of molecular structure. *Mathematical Proceedings of the Cambridge Philosophical Society* 33(1): 104-110.
- ⁴ Coulson C.A.; 1940. On the calculation of the energy in unsaturated hydrocarbon molecules. *Mathematical Proceedings of the Cambridge Philosophical Society* 36(2): 201-203.
- ⁵ Coulson C.A.; 1950. Notes on the secular determinant in molecular orbital theory. *Mathematical Proceedings of the Cambridge Philosophical Society* 46(1): 202-205.
- ⁶ Hartree, D.R.; 1928. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Math. Proc. Cambridge* 24(1): 89-110.
- ⁷ Hartree, D.R.; 1928. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part II. Some Results and Discussion. *Math. Proc. Cambridge* 24(1): 111-132.
- ⁸ Fock, V.A.; 1930. Approximation method for solving the quantum mechanical many-body problem (In German). *Zeitschrift für Physik* 61(1-2): 126-148.
- ⁹ Fock, V.A.; 1930. "Self consistent field" with exchange for sodium (In German). *Zeitschrift für Physik* 62(11-12): 795-805.
- ¹⁰ Laplace, P.S.; 1776. Recherches sur le calcul intégral et sur le système du monde. *Mémoires de l'Académie des Sciences de Paris* 2: 47-179.
- ¹¹ Cauchy, A.; 1829. Sur l'équation à l'aide de laquelle on détermine les inégalités séculaires des mouvements des planètes. *Exercices de mathématique* 4: 140-160.
- ¹² Slater, J.C.; 1929. The Theory of Complex Spectra. *Phys. Rev.* 34(10):1293-1295.
- ¹³ Hartree, D.R.; Hartree, W.; 1935. Self-Consistent Field, with exchange, for Beryllium. *Proc. R. Soc. London* 150(869): 9-33.
- ¹⁴ Sylvester, J.J.; 1880. On the theorem connected with Newton's rule for the discovery of imaginary roots of equations. *Messenger of Mathematics* 9: 71.84.
- ¹⁵ Godsil, C.D.; Gutman, I.; 1981. On the theory of the matching polynomial. *Journal of Graph Theory* 5(2): 137-144.
- ¹⁶ Godsil, C.D.; 1995. Algebraic matching theory. *The electronic journal of combinatorics* 2: #R8 (14p).
- ¹⁷ Diudea, M.V.; Gutman, I.; Jäntschi, L.; 2001. *Molecular Topology*. New York: Nova Science.
- ¹⁸ Ramaraj, R.; Balasubramanian, K.; 1985. Computer Generation of Matching Polynomials of Chemical Graphs and Lattices. *J. Comput. Chem.* 6: 122-141.
- ¹⁹ Curticapean, R.; 2013. Counting Matchings of Size k Is $\# W[1]$ -Hard. *Lecture Notes in Computer Science* 7965: 352-363.
- ²⁰ Hosoya, H.; 1971. Topological Index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Japan* 44: 2332-2339.
- ²¹ Morris, R.; Cherry, L.; 1975. bc - Version 6 Unix. 6th Edition Unix bc source code available at <http://minnie.tuhs.org/cgi-bin/utree.pl?file=V6/usr/source/s1/bc.y>
- ²² Nelson, P.A.; 1991. bc - an arbitrary precision calculator language. Documentation of version 1.06 available at: https://www.gnu.org/software/bc/manual/html_mono/bc.html
- ²³ Givens, W.B.; 1957. The characteristic value-vector problem, *J. Assoc. Comput. Mach.* 4: 298-307.
- ²⁴ Rehman, R.; Ipsen, I.C.F.; 2011. La Budde's Method for Computing Characteristic Polynomials. arXiv:1104.3769 (URL: <http://arxiv.org/format/1104.3769>).
- ²⁵ Householder, A.S.; 1958. Unitary Triangularization of a Nonsymmetric Matrix. *Journal of the ACM* 5(4): 339-342.
- ²⁶ Tománek, D.; 2014. Supplementary Information of: Guide through the Nanocarbon Jungle: Buckyballs, Nanotubes, Graphene, and Beyond. San Rafael: Morgan & Claypool Publishers.
- ²⁷ Alan McNAUGHT, 2006. The IUPAC International Chemical Identifier: InChI - A New Standard for Molecular Informatics. *Chem Int* 28(6): 12-15
- ²⁸ Von Collatz, L; Sinogowitz, U; 1957. Spectra of finite graphs (In German). *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg* 21(1): 63-77
- ²⁹ Sloane, N.J.A.; 1996. Number of graphs on n unlabeled nodes. OEIS(A000088): <http://oeis.org/A000088>

-
- ³⁰ Weisstein, W.E.; 2003. Number of unique characteristic polynomials among all simple undirected graphs on n nodes. OEIS(A082104): <http://oeis.org/A082104>
- ³¹ Markle, R.C.; 2000. Molecular building blocks and development strategies for molecular nanotechnology. *Nanotechnology* 11(2): 89-99.
- ³² Jäntschi, L; Bolboacă, S.D.; 2006. Modelling the Inhibitory Activity on Carbonic Anhydrase IV of Substituted Thiadiazole- and Thiadiazoline- Disulfonamides: Integration of Structure Information. *Electronic Journal of Biomedicine* 2: 22-33.
- ³³ Sachs, H.; 1962. Über selbstkomplementäre Graphen. *Publ. Math.* 9: 270-282.
- ³⁴ Graovac, A.; Gutman, I.; Trinajstić, N.; Živković, T; 1972. Graph Theory and Molecular Orbitals: Application of Sachs Theorem. *Theoretica chimica acta* 26: 67-78.