This is a peer-reviewed, accepted author manuscript of the following article: Alhomaidhi, A., Al-Thukair, F., & Estrada, E. (2021). Double gaussianization of graph spectra. Applied Mathematical Modelling, 93, 134-147. https://doi.org/10.1016/j.apm.2020.12.018

Double Gaussianization of Graph Spectra

Alhanouf Alhomaidhi[1,2], Fawzi Al-Thukair[1], Ernesto Estrada*[3,4]

[1] *Department of Mathematics, King Saud University, Saudi Arabia;* [2] *Department of Mathematics & Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G11XQ, UK;*

[3] *Institute of Mathematics and Applications (IUMA), Universidad de Zaragoza, Pedro Cerbuna 12, E-50009 Zaragoza, Spain;* [4] *ARAID Foundation, Government of Aragón, 50018 Zaragoza, Spain*

Abstract

The graph spectrum is the set of eigenvalues of a simple graph with n vertices. Here we fold this graph spectrum at a given pair of reference eigenvalues and then exponentiate the resulting folded graph spectrum. This process produces double Gaussianized functions of the graph adjacency matrix which give more importance to the reference eigenvalues than to the rest of the spectrum. Based on evidences from mathematical chemistry we focus here our attention on the reference eigenvalues ± 1. They seems to enclose most of the HOMO (highest occupied molecular orbital) and LUMO (lowest unoccupied molecular orbital) of organic molecular graphs. We prove here several results for the trace of the double Gaussianized adjacency matrix of simple graphs–the double Gaussianized Estrada index. Finally we apply this index to the classification of polycyclic aromatic hydrocarbons (PAHs) as carcinogenic or inactive ones. We discover that local indices based on the previously developed matrix function allow to classify correctly 100% of the PAHs analyzed. Such indices reflect the electron population of the HOMO/LUMO and eigenvalues close to them, in the so-called K and L regions of PAHs.

Keywords: matrix functions; mathematical chemistry; polycyclyc aromatic compounds; graph spectra; eigenvalues; HOMO; LUMO; frontier orbitals, HMO

1. Introduction

The use of functions of the adjacency matrix A of graphs has proved to be very useful in setting structure-spectra relations which found applications in many different research fields [1, 2, 3, 4, 5, 6, 7, 8, 9] (for a recent review see [10]). As part of this effort we have previously started the investigation of the spectral region close to the zero eigenvalue [5, 8]. That is, instead of using matrix functions, such as the matrix exponential $\exp(A)$ [1, 2], which prioritize the contribution of the spectral radius λ_1 over the rest of the eigenvalues of A,

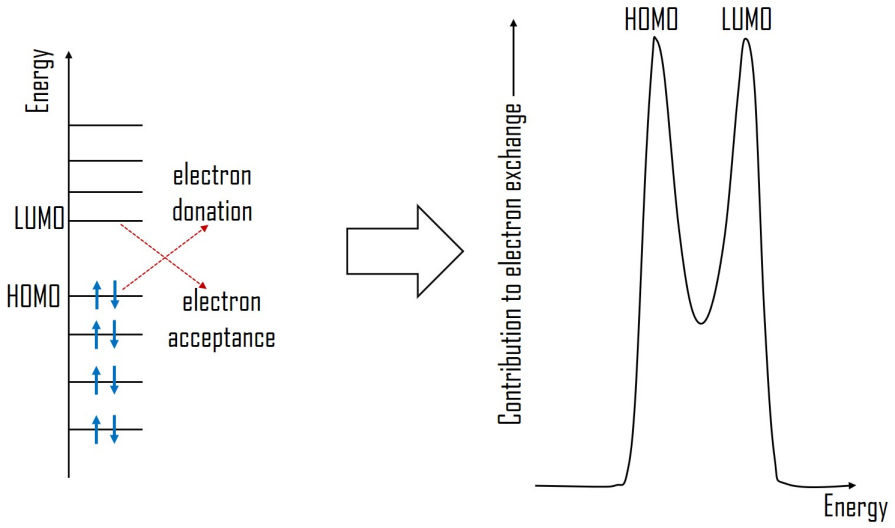

Figure 1: Scheme illustrating the double-Gaussian transformation of the spectrum of a graph representing the energy levels of a molecule.

we focus on general Gaussian function $\tilde{G}(\lambda_{\text{ref}}) = \exp\left[-\left(\lambda_{\text{ref}} I - A^2\right)\right]$ which prioritize λ_{ref} [5, 8]. Then, when $\lambda_{\text{ref}} = 0$ we are centering the matrix function on the nullity of the graph, which has been shown to play an important role, for instance in chemical applications [5]. We extended this study to the analysis of $\lambda_{\text{ref}} = -1$, which revealed important structural patterns hidden in the graph spectra [8].

When studying molecules with the so-called tight-binding Hamiltonians, e.g., the Hückel molecular orbital (HMO) approach [11, 12], there are two eigenvalues of the graph spectra which play a fundamental role in understanding molecular properties. They are known as the highest occupied (HOMO) and the highest unoccupied molecular (LUMO), respectively [13]. These "molecular orbitals" are schematically illustrated in Fig. 1 (left panel) where we indicate their importance as electron donor and acceptor, respectively. Our goal in this work is then to "fold" the graph spectra such that two eigenvalues, like for instance the HOMO and LUMO, have the largest contribution to the corresponding matrix function.

Therefore, in this work we define double Gaussian functions of the graph spectra:

$$\tilde{G}(\lambda_{\text{ref}_1}, \lambda_{\text{ref}_2}) = \exp\left[-\left(\lambda_{\text{ref}_1} I - A\right)^2 \left(\lambda_{\text{ref}_2} I - A\right)^2\right]. \tag{1}$$

The schematic process of the double Gaussianization of the graph spectra is illustrated in Fig. 2. In the case of alternant conjugated molecules with n atoms and graph eigenvalues $\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n$, the HOMO/LUMO correspond to the eigenvalue $\mp\lambda_{n/2}$, respectively. Therefore, here we will focus on the case in which $\lambda_{\text{ref}_1} = -\lambda_{\text{ref}_2}$, but the formulation is general enough as to consider any

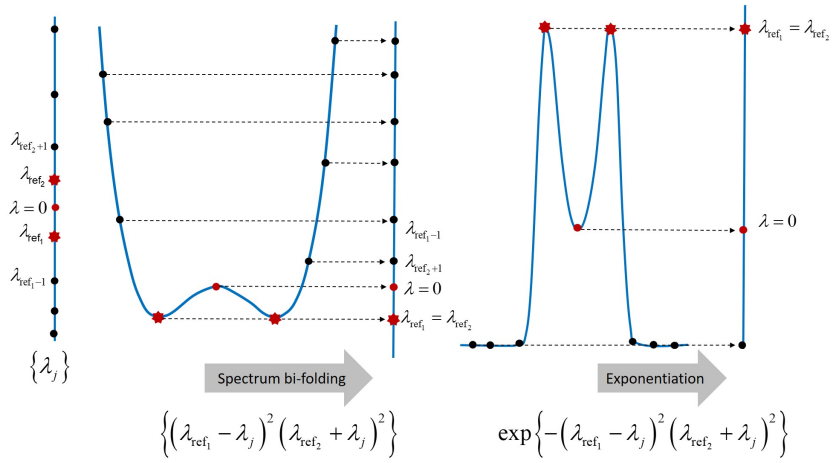

Figure 2: Schematic illustration of the double Gaussianization of the graph spectra. In the extreme left we illustrate the graph spectra where the eigenvalues are represented as dots in a vertical line. In the central panel we illustrate the bifolding of the spectrum where the reference eigenvalues occupy the lowest position in a vertical line. Finally (extreme right) we exponentiate the bifolded spectrum and the reference eigenvalues make the highest contribution to the matrix function.

further case. Additionally, Fowler and Pisanski [14] have called "normal" the molecular graphs for which $+1 \geq \lambda_{HOMO} \geq \lambda_{LUMO} \geq -1$, while the rest of molecular graphs are called "exceptional". The reason for this is that most of molecular graphs have their HOMO and LUMO within the 'chemical triangle' of an HOMO-LUMO map [14]–a scatterplot of the middle eigenvalues of the graph–, with vertices at $(-1,-1)$, $(+1,-1)$, $(+1,+1)$. They proved that all chemical trees lie within the triangle, as do all chemical graphs with up to 12 vertices [14]. Therefore, and for the sake of homogeneity of results, we will focus here on the case $\lambda_{\text{ref}_1}=1$, $\lambda_{\text{ref}_2}=-1$. Then, we study the function

$$\begin{aligned} G(-1,1) &= \exp\left[-\left((-1)I-A^2\right)(I-A)^2\right] \\ &= \exp\left[-\left(A^2-I\right)^2\right], \end{aligned} \tag{2}$$

and in particular the corresponding Estrada index [1] of this function:

$$\begin{aligned} H_{-1,1} &= trG(-1,1) \\ &= \sum_{j=1}^{n} e^{-\left(\lambda_j^2-1\right)^2}, \end{aligned} \tag{3}$$

where tr is the trace of the corresponding matrix.

2. Preliminaries

Here we settle the notations used in this work. We consider here simple, connected graphs $G = (V, E)$ with n nodes (vertices) and m edges. A *walk* of length k in G is a set of nodes $i_1, i_2, \ldots, i_k, i_{k+1}$ such that for all $1 \leq l \leq k$, $(i_l, i_{l+1}) \in E$. A *closed walk* is a walk for which $i_1 = i_{k+1}$. The degree of a vertex is the number of incident edges to it, i.e., the number of nearest neighbors the vertex has. The following types of graphs are used in this work. The complete graph of n vertices K_n is the graph having an edge between every pair of vertices. The complete bipartite graph K_{n_1,n_2} is the graph with the vertex set partitioned into two disjoint subsets of cardinalities n_1 and n_2, respectively, such that every vertex in one set is connected to every vertex in the other set. The star graph is the particular case in which $n_1 = 1$ and $n_2 = n - 1$. The path graph of n vertices P_n is the connected graph in which every vertex has degree 2, but two vertices which have degree one. The cycle C_n is a connected graph in which every vertex has degree 2. A subgraph $G' = (V', E')$ of G is a graph such that $V' \subseteq V$ and $E' \subseteq E$. An induced subgraph is a subgraph formed by a subset of the vertices of the graph and all of the edges connecting pairs of vertices in that subset.

Let A be the adjacency of matrix of the graph. We label the eigenvalues of A in non-increasing order: $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n$. Since A is a real-valued, symmetric matrix, we can decompose A into $A = U\Lambda U^T$ where Λ is a diagonal matrix containing the eigenvalues of A and $U = [\overrightarrow{\psi}_1, \ldots, \overrightarrow{\psi}_n]$ is orthonormal, where $\overrightarrow{\psi}_i$ is an eigenvector associated with λ_i. Because the graphs considered here are connected, A is irreducible and from the Perron-Frobenius theorem we can deduce that $\lambda_1 > \lambda_2$ and that the leading eigenvector $\overrightarrow{\psi}_1$ can be chosen such that its components $\psi_1(u)$ are positive for all $u \in V$.

3. $H_{-1,1}$ index for graphs with all but two eigenvalues equal to ± 1

Three infinite families of connected graphs have been reported [15] to have eigenvalues $r > 1$ and $s < -1$, and all other eigenvalues equal to ± 1. The adjacency matrix and spectra of these families are as follows. Let O be an all-zeros matrix, J an all-ones matrix and I_h the $h \times h$ identity matrix. Let R_{2k} be the adjacency matrix of k copies of K_2, i.e., the disjoint union of k edges.

Theorem 1. *[15] The infinite families of graphs having the following adjacency matrices and spectra are the only ones having all but two eigenvalues different from* ± 1*:*

$\begin{bmatrix} O & J - I_m \\ J - I_m & O \end{bmatrix}$ $(m \geq 3)$ *with spectrum* $\{\pm(m-1), 1^{m-1}, -1^{m-1}\}$*;*

$\begin{bmatrix} J - I_a & J \\ J & R_{2k} \end{bmatrix}$ $(a \geq 1,\, k \geq 2)$ *with spectrum*

$$\left\{ \frac{a}{2} \pm \frac{1}{2}\sqrt{a^2 + 8ak - 4a + 4}, 1^{k-1}, -1^{a+k-1} \right\}.$$

When $a = 1$, the resulted family is the friendship graphs;
$\begin{bmatrix} R_{2\ell} & J \\ J & R_{2m} \end{bmatrix}$ $(\ell \geq m \geq 2)$ *with spectrum* $\left\{1 \pm 2\sqrt{\ell m}, 1^{\ell+m-2}, -1^{\ell+m}\right\}$.

Then we have the following result.

Theorem 2. *The $H_{-1,1}$ index of the above families is presented in the next theorem*

$$H_{-1,1} = 2e^{-m^2(m-2)^2} + 2m - 2. \tag{4}$$

$$H_{-1,1} = e^{-\frac{a^2}{4}(a+4k+b-2)^2} + e^{-\frac{a^2}{4}(a+4k-b-2)^2} + (a+2k) - 2, \tag{5}$$

where $b = \sqrt{a^2 + 8ak - 4a + 4}$.

$$H_{-1,1} = e^{-16\left(\ell m + \sqrt{\ell m}\right)^2} + e^{-16\left(\ell m - \sqrt{\ell m}\right)^2} + 2(\ell + m) - 2. \tag{6}$$

Proof. (i)

$$H_{-1,1} = \sum_{j=1}^{n} e^{-\left(\lambda_j^2 - 1\right)^2} \tag{7}$$

$$= e^{-\left((m-1)^2-1\right)^2} + e^{-\left((1-m)^2-1\right)^2} + m - 1 + m - 1 \tag{8}$$

$$= e^{-m^2(m-2)^2} + e^{-(m-2)^2} + 2m - 2 \tag{9}$$

$$= 2e^{-m^2(m-2)^2} + 2m - 2. \tag{10}$$

(ii) let $b = \sqrt{a^2 + 8ak - 4a + 4}$ for more simplification, then

$$H_{-1,1} = \sum_{j=1}^{n} e^{-\left(\lambda_j^2 - 1\right)^2} \tag{11}$$

$$= e^{-\left(\left(\frac{a}{2} + \frac{1}{2}b\right)^2 - 1\right)^2} + e^{-\left(\left(\frac{a}{2} - \frac{1}{2}b\right)^2 - 1\right)^2} + k - 1 + a + k - 1 \tag{12}$$

$$= e^{-\frac{a^2}{4}(a+4k+b-2)^2} + e^{-\frac{a^2}{4}(a+4k-b-2)^2} + (a+2k) - 2. \tag{13}$$

(iii)

$$H_{-1,1} = \sum_{j=1}^{n} e^{-\left(\lambda_j^2 - 1\right)^2} \tag{14}$$

$$= e^{-\left(\left(1+2\sqrt{\ell m}\right)^2 - 1\right)^2} + e^{-\left(\left(1-2\sqrt{\ell m}\right)^2 - 1\right)^2} + \ell + m - 2 + \ell + m \tag{15}$$

$$= e^{-16\left(\ell m + \sqrt{\ell m}\right)^2} + e^{-16\left(\ell m - \sqrt{\ell m}\right)^2} + 2(\ell + m) - 2. \tag{16}$$

□

4. $H_{-1,1}$ Index for simple graphs

Here we prove some results for simple graphs which may be useful in understanding further structure-spectra relations in general graphs.

Lemma 3. *Let K_n be the complete graph of n nodes. Then*

$$H_{-1,1}(K_n) = n - 1 + e^{-n^2(n-2)^2}. \tag{17}$$

Proof. The spectrum of K_n is $\sigma(K_n) = \left\{[n-1]^1, [-1]^{n-1}\right\}$ so we have

$$\begin{aligned} H_{-1,1}(K_n) =& \sum_{j=1}^{n} e^{-\left(\lambda_j^2-1\right)^2} && (18)\\ =& (n-1)e^0 + e^{-\left((n-1)^2-\right)1^2} && (19)\\ =& n - 1 + e^{-n^2(n-2)^2}. && (20) \end{aligned}$$

□

Lemma 4. *Let K_{n_1,n_2} be the complete bipartite graph of $n_1 + n_2$ nodes. Then*

$$H_{-1,1}(K_{n_1,n_2}) = \frac{n_1+n_2-2}{e} + 2e^{-(n_1n_2-1)^2}. \tag{21}$$

Proof. The spectrum of K_{n_1,n_2} is $\sigma(K_{n_1,n_2}) = \left\{\left[\sqrt{n_1n_2}\right]^1, \left[-\sqrt{n_1n_2}\right]^1, [0]^{n_1+n_2-2}\right\}$ so we have

$$\begin{aligned} H_{-1,1}(K_{n_1,n_2}) =& \sum_{j=1}^{n_1+n_2} e^{-\left(\lambda_j^2-1\right)^2} && (22)\\ =& e^{-(n_1n_2-1)^2} + e^{-(n_1n_2-1)^2} + (n_1+n_2-2)\,e^{-1} && (23)\\ =& \frac{n_1+n_2-2}{e} + 2e^{-(n_1n_2-1)^2}. && (24) \end{aligned}$$

□

Corollary 5. *Let $K_{1,n-1}$ be the star graph of n nodes. Then*

$$H_{-1,1}(K_{1,n-1}) = \frac{n-2}{e} + 2e^{-(n-2)^2}. \tag{25}$$

Lemma 6. *Let P_n be a path having n nodes. Then, asymptotically as $n \to \infty$*

$$H_{-1,1}(P_n) = \frac{n+1}{\pi}\int_0^{\pi} e^{-(2\cos\theta+1)^2}d\theta - e^{-9}. \tag{26}$$

Proof. The spectrum of P_n consists of the numbers $2\cos\frac{j\pi}{n+1}$, $j = 1, 2, \ldots, n$. The angles $\frac{j\pi}{n+1}$ do not cover the entire interval $[0, \pi]$. Therefore when employing an integral approximation we need to compensate for the missing near-zero and near-π contributions. It is done as follows:

$$H_{-1,1}(P_n) = \sum_{j=1}^{n} e^{-\left(\lambda_j^2-1\right)^2} \tag{27}$$

$$= \sum_{j=1}^{n} e^{-\left(4\cos^2\left(\frac{j\pi}{n+1}\right)-1\right)^2} \tag{28}$$

$$= \sum_{j=1}^{n} e^{-\left(1+2\cos\left(\frac{2j\pi}{n+1}\right)\right)^2} \tag{29}$$

$$= \frac{1}{2}\sum_{j=0}^{n} e^{-\left(1+2\cos\left(\frac{2j\pi}{n+1}\right)\right)^2} + \frac{1}{2}\sum_{j=1}^{n+1} e^{-\left(1+2\cos\left(\frac{2j\pi}{n+1}\right)\right)^2} - \frac{1}{2}e^{-9} - \frac{1}{2}e^{-9} \tag{30}$$

$$= \frac{1}{2}\sum_{j=0}^{n} e^{-\left(1+2\cos\left(\frac{2j\pi}{n+1}\right)\right)^2} + \frac{1}{2}\sum_{j=1}^{n+1} e^{-\left(1+2\cos\left(\frac{2j\pi}{n+1}\right)\right)^2} - e^{-9}. \tag{31}$$

Now, when $n \to \infty$ the summation in 31 can be approached by the following integral

$$H_{-1,1}(P_n) = \frac{1}{2}\frac{n+1}{\pi}\int_0^{\pi} e^{-(2\cos\theta+1)^2} d\theta + \frac{1}{2}\frac{n+1}{\pi}\int_0^{\pi} e^{-(2\cos\theta+1)^2} d\theta - e^{-9} \tag{32}$$

$$= \frac{n+1}{\pi}\int_0^{\pi} e^{-(2\cos\theta+1)^2} d\theta - e^{-9}. \tag{33}$$

□

The following theorem gives the value of the H_{-1} index for paths where $H_{-1} = tr(e^{(I+A)^2})$.

Lemma 7. *Let P_n be a path having n nodes. Then, asymptotically as $n \to \infty$*

$$H_{-1}(P_n) = \frac{n+1}{\pi}\int_0^{\pi} e^{-(2\cos\theta+1)^2} d\theta - \frac{1}{2}\left(e^{-1} + e^{-9}\right). \tag{34}$$

Proof.

$$H_{-1}\left(P_n\right) = \sum_{j=1}^{n} e^{-(\lambda_j+1)^2} \tag{35}$$

$$= \sum_{j=1}^{n} e^{-\left(2\cos\left(\frac{j\pi}{n+1}\right)+1\right)^2} \tag{36}$$

$$= \frac{1}{2}\sum_{j=0}^{n} e^{-\left(1+2\cos\left(\frac{j\pi}{n+1}\right)\right)^2} + \frac{1}{2}\sum_{j=1}^{n+1} e^{-\left(1+2\cos\left(\frac{j\pi}{n+1}\right)\right)^2} - \frac{1}{2}e^{-9} - \frac{1}{2}e^{-1}. \tag{37}$$

Now, when $n \to \infty$ the summation can be approached by the following integral

$$H_{-1}\left(P_n\right) = \frac{n+1}{\pi}\int_0^{\pi} e^{-(2\cos\theta+1)^2} d\theta - \frac{1}{2}\left(e^{-1}+e^{-9}\right). \tag{38}$$

□

Remark. In the proofs of Lemmas 6 and 7, i) follows the steps of [16].

Lemma 8. *Let C_n be a cycle having n nodes. Then, asymptotically as $n \to \infty$*

$$H_{-1,1}\left(C_n\right) = \frac{n}{\pi}\int_0^{\pi} e^{-(2\cos\theta+1)^2} d\theta. \tag{39}$$

Proof. Notice that the adjacency matrix of a cycle is a circulant matrix and consequently any function of it and that gives

$$H_{-1,1}(C_n) = \sum_{j=1}^{n} \tilde{G}_{pp}, \text{ for any node } p \tag{40}$$

$$= n\left(\frac{tr\left(e^{-\left(A^2-I\right)^2}\right)}{n}\right) \tag{41}$$

$$= n\left(\frac{1}{n}\sum_{j=1}^{n} e^{-\left(4\cos^2\left(\frac{2\pi j}{n}\right)-1\right)^2}\right) \tag{42}$$

$$= n\left(\sum_{j=1}^{n} \frac{1}{n} e^{-\left(2+2\cos\left(\frac{4\pi j}{n}\right)-1\right)^2}\right) \tag{43}$$

$$= n\left(\sum_{j=1}^{n} \frac{1}{n} e^{-\left(1+2\cos\frac{4\pi j}{n}\right)^2}\right). \tag{44}$$

Now, when $n \to \infty$ the summation in 44 can be approached by the following

integral

$$H_{-1,1}(C_n) = n\frac{1}{4\pi}\int_0^{4\pi} e^{-(1+2\cos\theta)^2}d\theta \tag{45}$$

$$= n\frac{1}{4\pi}(2)\int_0^{2\pi} e^{-(1+2\cos\theta)^2}d\theta \tag{46}$$

$$= n\frac{1}{4\pi}(2)(2)\int_0^{\pi} e^{-(1+2\cos\theta)^2}d\theta \tag{47}$$

where $\theta = \frac{4j\pi}{n}$. Thus, when $n \to \infty$ we have

$$H_{-1,1}(C_n) = \frac{n}{\pi}\int_0^{\pi} e^{-(1+2\cos\theta)^2}d\theta. \tag{48}$$

□

The following theorem gives the value of the H_{-1} index for cycles where $H_{-1} = tr\left(e^{(I+A)^2}\right)$.

Lemma 9. *Let C_n be a cycle having n nodes. Then, asymptotically as $n \to \infty$*

$$H_{-1}(C_n) = \frac{n}{\pi}\int_0^{\pi} e^{-(2\cos\theta+1)^2}d\theta. \tag{49}$$

Proof. Notice that the adjacency matrix of a cycle is a circulant matrix and consequently any function of it and that gives

$$H_{-1}(C_n) = \sum_{j=1}^{n}\tilde{G}_{pp}, \text{ for any node } p \tag{50}$$

$$= n\left(\frac{tr\left(e^{-(A+I)^2}\right)}{n}\right) \tag{51}$$

$$= n\left(\frac{1}{n}\sum_{j=1}^{n} e^{-\left(1+2\cos\left(\frac{2\pi j}{n}\right)\right)^2}\right) \tag{52}$$

Now, when $n \to \infty$ the summation in 52 can be approached by the following integral

$$H_{-1}(C_n) = n\frac{1}{2\pi}\int_0^{2\pi} e^{-(1+2\cos\theta)^2}d\theta \tag{53}$$

$$= n\frac{1}{2\pi}(2)\int_0^{\pi} e^{-(1+2\cos\theta)^2}d\theta \tag{54}$$

where $\theta = \frac{2j\pi}{n}$. Thus, when $n \to \infty$ we have

$$H_{-1}(C_n) = \frac{n}{\pi}\int_0^{\pi} e^{-(2\cos\theta+1)^2}d\theta. \tag{55}$$

□

Corollary 10. *As $n \to \infty$ then:*

1. *$H_{-1,1}(P_n)$ and $H_{-1}(P_n)$ are asymptotically equivalent.*
2. *$H_{-1,1}(C_n)$ and $H_{-1}(C_n)$ are asymptotically equivalent.*
3. *$H_{-1,1}(K_n)$ and $H_{-1}(K_n)$ are asymptotically equivalent.*
4. *$H_{-1,1}(K_{n_1,n_2})$ and $H_{-1}(K_{n_1,n_2})$ are asymptotically equivalent.*
5. *$H_{-1,1}(K_{1,n-1})$ and $H_{-1}(K_{1,n-1})$ are asymptotically equivalent.*

Proof. Let us write the following limits of the ratios of both indices:

1)

$$\lim_{n\to\infty} \frac{H_{-1,1}(C_n)}{H_{-1}(C_n)} = \frac{\frac{n}{\pi}\int_0^\pi e^{-(2\cos\theta+1)^2}d\theta}{\frac{n}{\pi}\int_0^\pi e^{-(2\cos\theta+1)^2}d\theta} = 1.$$

2)

$$\lim_{n\to\infty} \frac{H_{-1,1}(P_n)}{H_{-1}(P_n)} = \frac{\frac{n+1}{\pi}\int_0^\pi e^{-(2\cos\theta+1)^2}d\theta - e^{-9}}{\frac{n+1}{\pi}\int_0^\pi e^{-(2\cos\theta+1)^2}d\theta - \frac{1}{2}\left(e^{-1}+e^{-9}\right)} = 1$$

3)

$$\lim_{n\to\infty} \frac{H_{-1,1}(K_n)}{H_{-1}(K_n)} = \frac{n-1+e^{-n^2(n-2)^2}}{n-1+e^{-n^2}} = 1.$$

4) when $n \to \infty$ we have also $n_1 + n_2 = n \to \infty$ and $n_1 n_2 \to \infty$, thus

$$\begin{aligned}\lim_{n\to\infty} \frac{H_{-1,1}(K_{n_1,n_2})}{H_{-1}(K_{n_1,n_2})} &= \lim_{n\to\infty} \frac{\frac{n_1+n_2-2}{e} + 2e^{-(n_1n_2-1)^2}}{\frac{n_1+n_2-2}{e} + e^{-n_1n_2-1}\left(\frac{e^{2\sqrt{n_1n_2}}+e^{-2\sqrt{n_1n_2}}}{2}\right)} \\ &= \lim_{n\to\infty} \frac{\frac{n_1+n_2-2}{e} + 2e^{-(n_1n_2-1)^2}}{\frac{n_1+n_2-2}{e} + \left(\frac{e^{2\sqrt{n_1n_2}-(n_1n_2+1)}+e^{-2\sqrt{n_1n_2}-(n_1n_2+1)}}{2}\right)} \\ &= 1\end{aligned}$$

5) We proved the general case in (4). □

5. Extremal graphs for $H_{-1,1}$ index

Let us start here by stating a result from Cioabă et al. [15]. Define $\mathcal{G}$ to be the set of connected graphs with eigenvalues $r > 1$ and $s < -1$, and all other eigenvalues equal to ± 1. Then, Cioabă et al. [15] proved the following result.

Lemma 11. *No graph in $\mathcal{G}$ has one of the graphs presented in Fig. 3 as an induced subgraph.*

We calculated the $H_{-1,1}$ index for all 11,117 connected graphs with 8 nodes and determined those with the largest values of the index. These graphs are illustrated in Fig. 4. The largest value of $H_{-1,1}$ is obtained for the complete graph K_8 (not illustrated in the Fig. 4). We have verified that for graphs $G_{n\leq 8}$, $H_{-1,1}(G_n) < H_{-1,1}(K_n)$. Therefore we have the following.

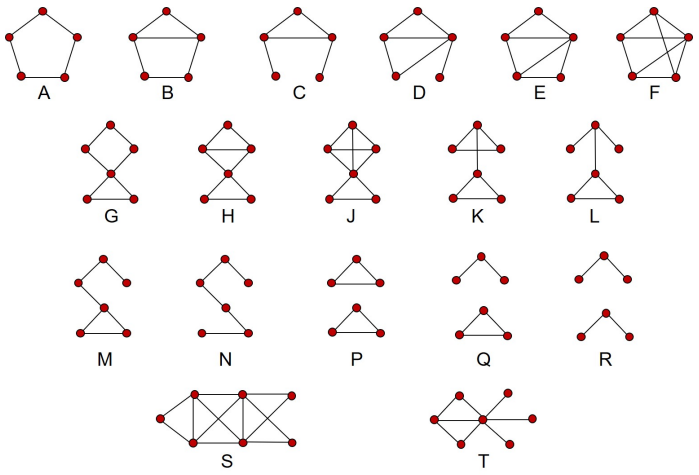

Figure 3: Illustration of the prohibited induced subgraphs found by Cioabă et al. [15]. We use the same labelling as in the paper of Cioabă et al. [15].

Conjecture 12. *Let G be any connected graph of n nodes, then*

$$H_{-1,1}(G) \leq H_{-1,1}(K_n). \tag{56}$$

In addition, none of the graphs in Fig. 4 contain any of the graphs in Fig. 3 as an induced subgraph.

We then explore the graphs with the smallest values of $H_{-1,1}$ among all 11,117 connected graphs with 8 nodes. the 10 ones with the smallest values of this index are illustrated in Fig. 5.

We have calculated the number of each of the prohibited induced subgraphs in these 10 graphs displaying the minimum values of $H_{-1,1}(G)$. We have found that 11 out of the 18 prohibited induced subgraphs appear very frequently in

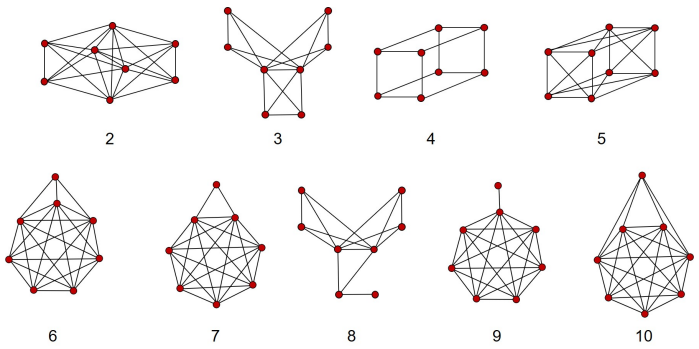

Figure 4: Illustration of the 10 graphs (K_8 is the number 1, which is omitted) with the largest values of $H_{-1,1}(G)$ among all connected graphs with 8 nodes.

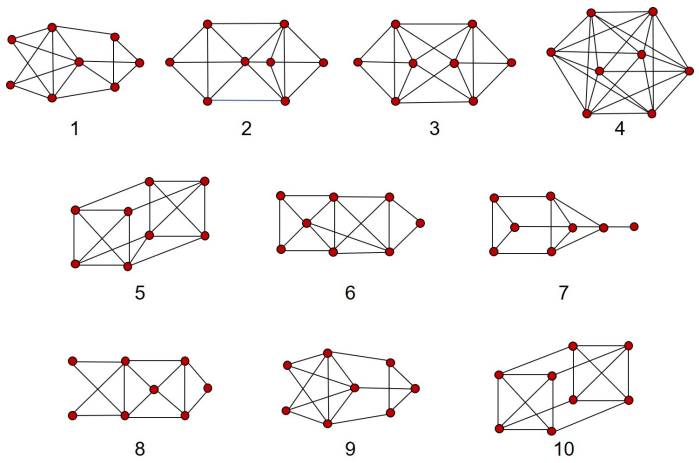

Figure 5: Illustration of the 10 graphs with the minimum values of $H_{-1,1}(G)$ among all connected graphs with 8 nodes.

these 10 graphs. The results are illustrated in Fig. 6. for instance, the graph with the least value of $H_{-1,1}(G)$ has the prohibited induced sugraph B 8 times, D 6 times and L 2 times. Others, like graph 10 in Fig. 6 contains only one prohibited subgraph, i.e., subgraph B 24 times.

Lemma 13. *Let Let G be connected bipartite graph of n nodes, then*

$$H_{-1,1}(G) \leq H_{-1,1}(K_n). \tag{57}$$

Proof. for any graph G we have $\lambda_1 \geq d_{ave}$ where λ_1 is the principal eigenvalue and d_{ave} is the average degree of the graph G. Then we have

$$\begin{aligned} \lambda_1 \geq d_{ave} &= \frac{2m}{n} \\ &\geq \frac{2(n-1)}{n}. \end{aligned} \tag{58}$$

Suppose that $n \geq 5$ (it is easy to check that the statement is true for all graphs of nodes less than 5) then we have $\lambda_1 \geq \frac{2(n-1)}{n} \geq \frac{2(4)}{5} = 1.6$. Thus, $(\lambda_1^2-1)^2 \geq 2.4336$ and that implies $\exp\left[-(\lambda_1^2-1)^2\right] \leq \exp(-2.4336)$. If G is bipartite, then we will have a symmetry in the spectra of G and we get $\lambda_n \leq -1.6$. Following the same steps we end withexp $\left[-(\lambda_n^2-1)^2\right] \leq \exp(-2.4336)$. Now

$$\begin{aligned} H_{-1,1}(G) =& \sum_{j=1}^{n} e^{-\left(\lambda_j^2-1\right)^2} = \sum_{j=2}^{n-1} e^{-\left(\lambda_j^2-1\right)^2} + e^{-\left(\lambda_1^2-1\right)^2} + e^{-\left(\lambda_n^2-1\right)^2} \\ &\leq n-2+2e^{-2.4336} \\ &< n-2+1 = n-1 \leq n-1+e^{-n^2(n-2)^2} = H_{-1,1}(K_n). \end{aligned}$$

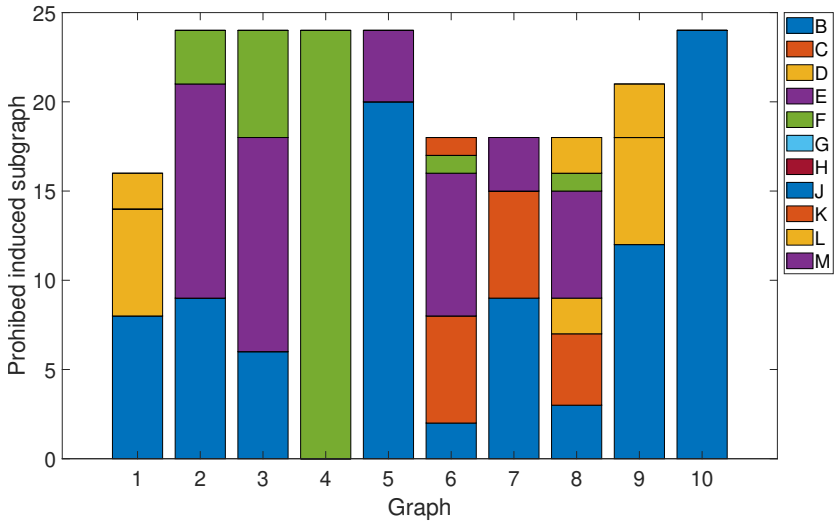

Figure 6: Frequency with which some of the prohibited induced subgraphs appear in the connected graphs with 8 nodes which display the minimum values of the index $H_{-1,1}(G)$. The induced subgraphs are given in Fig. 3 and graphs are shown in Fig. 5. The prohibed induced subgraphs not depicted in the figure do not appear in the graphs considered.

□

6. Subgraph contributions to $H_{-1,1}(G)$

We can expand the index $H_{-1,1}(G)$ as a Taylor series to obtain

$$\begin{aligned} e^{-(A^2-I)^2} &= e^{-I}e^{2A^2}e^{-A^4} \\ &= \frac{1}{e}Ie^{-A^4}e^{2A^2} \\ &= \frac{1}{e}\left(\sum_{j=0}^{\infty}\frac{(-1)^j}{j!}A^{4j}\right)\left(\sum_{i=0}^{\infty}\frac{(2)^i}{i!}A^{2i}\right). \end{aligned} \tag{59}$$

Now, using Cauchy product of two infinite series we arrive at

$$e^{-(A^2-I)^2} = \frac{1}{e}\left(\sum_{k=0}^{\infty}a_kA^{2k}\right),$$

where $a_k = \sum_{4m+2n=2k}(-1)^m\frac{2^n}{m!n!}$, and m, n are non negative integers such that $4m+2n=2k$. For example,

$$e^{-(A^2-I)^2} \approx \frac{1}{e}\left(\sum_{k=0}^{11} a_k A^{2k}\right)$$
$$= \frac{1}{e}\left(I + 2A^2 + A^4 - \frac{2}{3}A^6 - \frac{5}{6}A^8 - \frac{1}{15}A^{10} + \frac{23}{90}A^{12} + \frac{29}{315}A^{14} - \frac{103}{2520}A^{16} - \frac{4}{35}A^{18} - \frac{1}{15}A^{20} - \frac{2}{63}A\right.$$

So, the trace of $e^{-(A^2-I)^2}$ can be expressed as

$$H_{-1,1}(G) = \frac{1}{e}\left(\sum_{k=0}^{\infty} a_k TrA^{2k}\right)$$
$$= \frac{1}{e}\left(TrI + 2TrA^2 + TrA^4 - \frac{2}{3}TrA^6 - \frac{5}{6}TrA^8 - \frac{1}{15}TrA^{10} + \frac{23}{90}TrA^{12} + \ldots\right). \tag{60}$$

We know that:

Lemma 14. *The number of walks of length k between the nodes p and q of a graph is given by $\left(A^k\right)_{pq}$.*

Consequently, TrA^k counts the number of closed walks of length k in the graph. Every close walk encloses a given subgraph. For instance, a closed walk of length two encloses an edge, therefore the TrA^2 counts twice the number of edges in the graph. Thus, we can related every term TrA^k with a weighted sum of subgraphs. However, due to the presence of positive and negative signs in Eq. (60) the convergence of the Taylor series is extremely slow, which make necessary a long list of terms to compute $H_{-1,1}(G)$ this formula. Therefore, the main importance of this power-series expansion resdes in its use for structural interpretations as the researcher can identify those terms with a positive or a negative contribution to the index as well as the contribution of a given specific subgraph.

7. Application. Carcinogenicity of polycyclic aromatic hydrocarbons

Polycyclic aromatic hydrocarbons (PAHs) are compounds formed by carbon in fused hexagonal shapes and hydrogen, for which the eigenvalues ± 1 play an important role [17]. The excessive exposure to PAHs may result in cancer in humans. A typical route of exposition is through the consumption of charcoal broiled foods [18]. The general mechanism by which PAHs produce cancer is by their metabolic activation which leads to the formation of the active carcinogens like diol-epoxides, radical cations, and o-quinones (see first line in Fig. 7) [18]. These metabolites then react with DNA forming DNA adducts which results in DNA mutations, alteration of gene expression profiles, and tumorigenesis (see second line in Fig. 7). The metabolic activation of PAHs depends on the

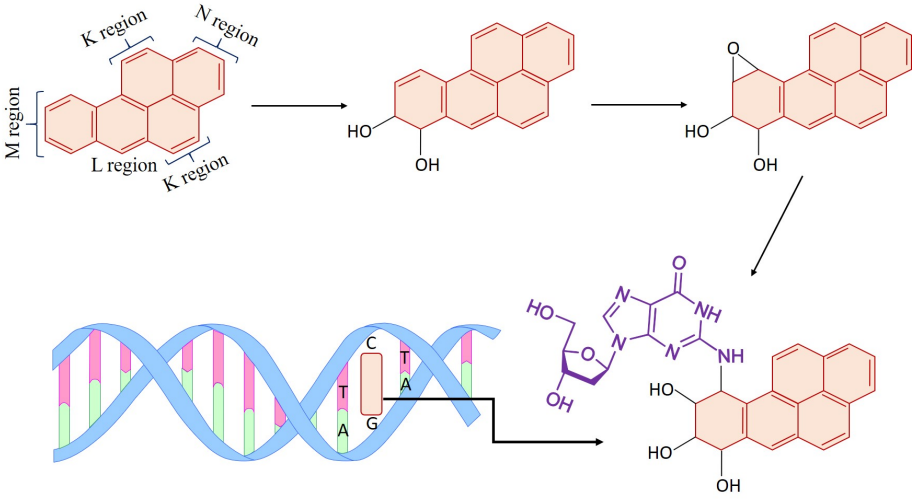

Figure 7: Schemetaic illustration of the metabolic activation of a PAH (first line) and the reaction of the reactive metabolite with a DNA base producing alterations in DNA (second line).

chemical reactivity of these compounds, and their electron donation/acceptance capacities, which are mainly determined by their HOMO and LUMO. Then, it is not strange to find reports on the use of these frontier molecular orbitals or electronic parameters like superdelocalizability in explaining the carcinogenic power of PAHs [19].

However, because chemical reactions occur at some specific atoms in a molecule different atomic regions may have distinct contributions to the carcinogenicity of PAHs. This has been widely recognized in the literature where four main atomic regions have been identified with different contributions to the carcinogenic activity of PAHs. These regions are known as K, L, M and N, which are illustrated in Fig. 7 (see for instance [19, 20]). The regions K and L where proposed by Pullman and Pullman [21, 22, 23] and have proved to be predictive for the carcinogenicity of a large number of PAHs [21]. Due to more recent findings the other two regions, M and N, where proposed and studied in quantitative structure-carcinogenicity activity of PAHs for instance by Vijayalakshmi and Suresh [20].

Here we use the series of 28 PAHs for which the carcinogenic power has been reported and studied by Vijayalakshmi and Suresh [20]. The list of PHAs and their carcinogenic activity (CA) is reported in Table 1. We consider here the $H_{-1,1}$ index split as follows:

$$H_{-1,1} = H_{-1,1}(K) + H_{-1,1}(L) + H_{-1,1}(M) + H_{-1,1}(N) + H_{-1,1}(F), \quad (61)$$

where $H_{-1,1}(K)$ is the sum of the contributions of the atoms in the region K

to the global $H_{-1,1}$ index, and the term F is used for the atoms in the frame of the PAHs, i.e., those not in any of the four mentioned regions. We recall that the contribution of an atom p to the $H_{-1,1}$ index is:

$$H_{-1,1}(p) = \sum_{j=1}^{n} \psi_j^2(p) \exp\left(-\left(\lambda_j^2 - 1\right)^2\right). \tag{62}$$

We use here the average of the atomic contributions for each region $\bar{H}_{-1,1}$ whose values for the 28 PAHs analyzed are given in Table 1. We grouped the carcinogenic activity (CA) of these 28 PAHs into two categories, which correspond to the class I which groups inactive and the class A, which groups PAHs with CA ranging from + to ++++ (see Table 1). The main reason is that a classification based on the strength of the carcinogenic activity is impossible as some of the classes contain only one member, e.g., CA +++.

We now focus in classification techniques that allow to split the set of PAHs into the two groups devised here on the basis of the regional $\bar{H}_{-1,1}$ indices. For that purpose we explore the use of discriminant analysis, classification trees, support vector machine, and K-nearest neighbors (KNN) techniques, all implemented in the "classification learner" toolbox of Matlab R2018b. In all cases we observe that the use of the $\bar{H}_{-1,1}$ indices for the K and L regions are enough for the classification of these compounds, with no improvement by adding information about M and N regions. Both, linear discriminant analysis (LDA) and support vector machine (SVM) classify correctly 85.7% of PAHs in the two classes. From the carcinogenic compounds these methods classify correctly 87.5% of PAHs and 83.3% of inactive ones. In Fig. 8 (a and b) we illustrate the results for the LDA. The classification tree improves the previous results and classifies correctly 93.75% of carcinogenic PAHs (see Fig. 8 (c and d). The best results are obtained by using KNN which classifies correctly 100% of compounds in the two classes as illustrated in Fig. 8 (e and f). Our analysis has no exception like in the case of Vijayalakshmi et al. [20] for which 5 PAHs were excluded from the analysis as outliers (phenanthrene, chrysene, triphenylene, naphthalene and coronene).

These results coincide qualitatively with those published long time ago by Pullman and Pullman [21, 22, 23] which shown that the K and L regions are enough to classify correctly PAHs into carcinogenic/inactive classes. As can be observed in Fig. 8 (a, c, and e) carcinogenic compounds are those having large values of $\bar{H}_{-1,1}(K)$ as well as of $\bar{H}_{-1,1}(L)$ (red regions in the mentioned plots). This indicates that carcinogenic PAHs have large contributions of the HOMO/LUMO eigenvalues and those close to them, which parallel the idea of compounds of high reactivity. In fact, the three only PAHs having the strongest carcinogenicity, i.e., "++++", are the ones having the largest values of $\bar{H}_{-1,1}(K)$: dibenzo[a,i]pyrene (0.5474); dibenzo[a,h]pyrene (0.5410); benzo[a]pyrene (0.5299). However, in general having low values of either $\bar{H}_{-1,1}(K)$ or $\bar{H}_{-1,1}(L)$ result in inactive compounds although the other index display large values. This result possibly indicates that a combined intervention of both regions, K and L, are important for the diverse processes giving

No.	compound	$\bar{H}_{-1,1}$				CA	Class
		K	L	M	N		
1	dibenzo[a,i]pyrene	0.5474	0.3511	0.5243	0	++++	A
2	dibenzo[a,h]pyrene	0.541	0.3582	0.5229	0	++++	A
3	benzo[a]pyrene	0.5299	0.3531	0.5266	0.4902	++++	A
4	tribenzo[a,e,i]pyrene	0.5244	0.3806	0.5233	0	++	A
5	dibenzo[a,e]pyrene	0.5024	0.379	0.5222	0.4903	+++	A
6	naphtho[2,3,a]pyrene	0.4899	0.393	0.5088	0.5066	++	A
7	benzo[g,h,i]perylene	0.4799	0	0	0.488	++	A
8	dibenzo[a,h]anthracene	0.4234	0.4388	0.5088	0	++	A
9	dibenzo[a,j]anthracene	0.4192	0.4363	0.5087	0	++	A
10	dibenzo[a,c]anthracene	0	0.4229	0.5192	0	++	A
11	peropyrene	0.5372	0	0	0.4832	+	A
12	benzo[a]naphtho[3,4,a]pyrene	0.5007	0.4011	0.5138	0.4861	+	A
13	benzo[a]naphtho[2,1,8,h,i,j]naphthacene	0.4914	0.417	0.5081	0.485	+	A
14	benzo[a]anthracene	0.4255	0.3947	0.5106	0	+	A
15	tribenzo[a,c,h]naphthacene	0.4225	0.4316	0.5207	0	+	A
16	dibenzo[a,c]naphthacene	0	0.4157	0.5214	0	+	A
17	pyrene	0.5183	0	0	0.4783	-	I
18	coronene	0.4934	0	0	0	-	I
19	anthanthrene	0.4833	0.4123	0	0.495	-	I
20	benzo[e]pyrene	0.48	0	0.5161	0.48	-	I
21	chrysene	0.4301	0	0.5091	0	-	I
22	phenanthrene	0.3975	0	0.5072	0	-	I
23	triphenylene	0	0	0.5258	0	-	I
24	dibenzo[e,l]pyrene	0	0	0.5224	0.4889	-	I
25	tetracene	0	0.3881	0.515	0	-	I
26	anthracene	0	0.3525	0.5114	0	-	I
27	naphthalene	0	0	0.4723	0	-	I
28	perylene	0	0	0	0.5008	-	I

Table 1: Names of the PAHs studied here, their carcinogenic action (CA), the values of $\bar{H}_{-1,1}$ for the four atomic regions of PAHs.

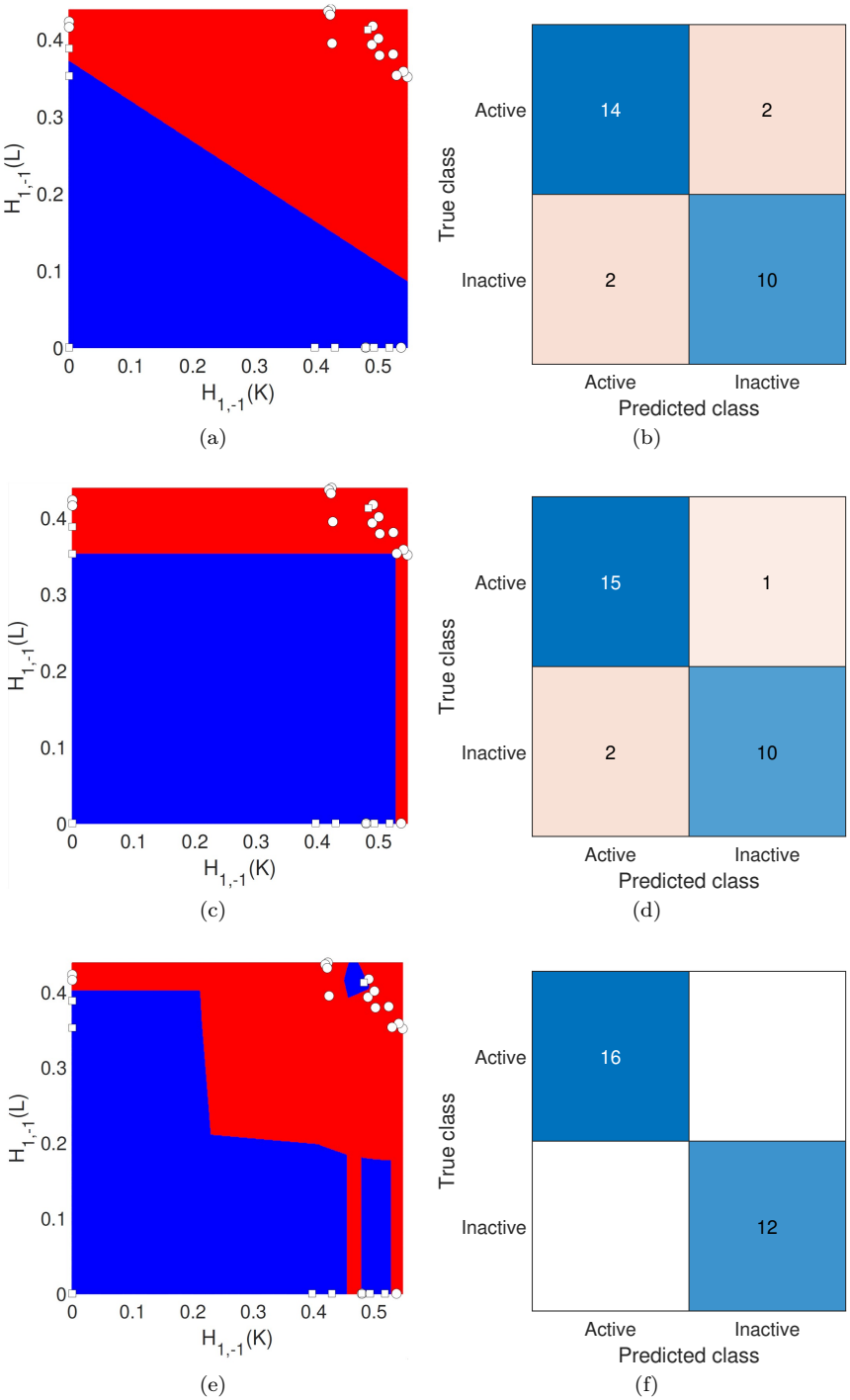

Figure 8: Illustration of the classification plots for carcinogenic PAHs (white circles) and inactive ones (white squares) using LDA (a), classification tree (c) and KNN (e). Confusion charts for the results obtained with the three classification methods used: LDA (b), classification tree (d) and KNN (f).

rise to the carcinogenicity of these compounds.

8. Conclusions

We have defined here a generalization of Gaussian function of the adjacency matrix of a graph to account for the spectral folding at two eigenvalues. This double Gaussianization of the graph spectra allows to generate indices that give more importance to a couple of reference eigenvalues in the graph, instead of all previous matrix functions which give the highest contribution to one eigenvalue, typically the spectral radius. The current approach is general enough as to focus on any pair of eigenvalues of the graph, but here we have concentrated on the pair ± 1. The main motivation for this selection has been the role played by these two eigenvalues in the spectra of molecular graphs, where the HOMO and LUMO eigenvalues seem to be bounded by these two numbers. The importance of the HOMO and LUMO in organic molecules has been widely documented and the current work is a contribution to the search of indices that pay more importance to these eigenvalues and those close to them. As we have shown here the indices derived from double Gaussianization of the graph spectra describe very well the carcinogenicity of PAHs, opening new avenues for the analysis of quantitative structure-property/activity relations in molecular sciences.

References

[1] E. Estrada, J. A. Rodriguez-Velazquez, Subgraph centrality in complex networks, Physical Review E 71 (5) (2005) 056103.

[2] E. Estrada, D. J. Higham, Network properties revealed through matrix functions, SIAM review 52 (4) (2010) 696–714.

[3] E. Estrada, Generalized walks-based centrality measures for complex biological networks, Journal of Theoretical Biology 263 (4) (2010) 556–565.

[4] E. Estrada, M. Benzi, What is the meaning of the graph energy after all?, Discrete Applied Mathematics 230 (2017) 71–77.

[5] E. Estrada, A. A. Alhomaidhi, F. Al-Thukair, Exploring the "middle earth" of network spectra via a gaussian matrix function, Chaos: An Interdisciplinary Journal of Nonlinear Science 27 (2) (2017) 023109.

[6] E. Estrada, G. Silver, Accounting for the role of long walks on networks via a new matrix function, Journal of Mathematical Analysis and Applications 449 (2) (2017) 1581–1600.

[7] E. Estrada, The electron density function of the huckel (tight-binding) model, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 474 (2210) (2018) 20170721.

[8] A. Alhomaidhi, F. Al-Thukair, E. Estrada, Gaussianization of the spectra of graphs and networks. theory and applications, Journal of Mathematical Analysis and Applications 470 (2) (2019) 876–897.

[9] E. Estrada, Back to the Origins. Using Matrix Functions of Huckel Hamiltonian for Quantum Interference, Apple Academic Press: Oakville, ON, 2018.

[10] M. Benzi, P. Boito, Matrix functions in network analysis, GAMM Mitteilungen.

[11] E. Heilbronner, H. Bock, HMO model and its application, Wiley, 1976.

[12] C. A. Coulson, B. O'Leary, R. B. Mallion, Hückel theory for organic chemists, Academic Pr, 1978.

[13] A. Streitwieser, Molecular orbital theory for organic chemists, in: Pioneers of Quantum Chemistry, ACS Publications, 2013, pp. 275–300.

[14] P. W. Fowler, T. Pisanski, Homo-lumo maps for chemical graphs, MATCH Commun. Math. Comput. Chem 64 (2) (2010) 373–390.

[15] S. M. Cioabă, W. H. Haemers, J. R. Vermette, W. Wong, The graphs with all but two eigenvalues equal to ± 1, Journal of Algebraic Combinatorics 41 (3) (2015) 887–897.

[16] I. Gutman, A. Graovac, Estrada index of cycles and paths, Chemical physics letters 436 (1-3) (2007) 294–296.

[17] Y.-S. Jiang, G.-Y. Chen, On subspectral problem–benzenoid hydrocarbons with common eigenvalues± 1, Theoretica chimica acta 76 (6) (1990) 437–450.

[18] B. Moorthy, C. Chu, D. J. Carlin, Polycyclic aromatic hydrocarbons: from metabolism to lung cancer, Toxicological Sciences 145 (1) (2015) 5–15.

[19] I. A. Smith, G. D. Berger, P. G. Seybold, M. Serve, Relationships between carcinogenicity and theoretical reactivity indices in polycyclic aromatic hydrocarbons, Cancer research 38 (9) (1978) 2968–2977.

[20] K. P. Vijayalakshmi, C. H. Suresh, Theoretical studies on the carcinogenicity of polycyclic aromatic hydrocarbons, Journal of computational chemistry 29 (11) (2008) 1808–1817.

[21] A. Pullman, B. Pullman, Electronic structure and carcinogenic activity of aromatic molecules new developments, in: Advances in Cancer Research, Vol. 3, Elsevier, 1955, pp. 117–169.

[22] A. Pullman, B. PULLMAN, Quantum biochemistry, in: Comprehensive Biochemistry, Vol. 22, Elsevier, 1967, pp. 1–60.

[23] B. Pullman, Recent developments on the mechanism of chemical carcinogenesis by aromatic hydrocarbons, International Journal of Quantum Chemistry 16 (3) (1979) 669–689.