

Universidade Estadual de Campinas
Instituto de Física “Gleb Wataghin”

TESE DE MESTRADO

**Estudo da Atividade Carcinogênica dos
Hidrocarbonetos Policíclicos
Aromáticos Através de Descritores
Quânticos**

Karla Souza Troche

Orientador: Prof. Dr. Douglas Galvão

Julho de 2003

**FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO IFGW - UNICAMP**

T741e	<p>Troche, Karla Souza Estudo da atividade carcinogênica dos hidrocarbonetos policíclicos aromáticos através de descritores quânticos / Karla Souza Troche. – Campinas, SP : [s.n.], 2003.</p> <p style="text-align: center;">Orientador: Douglas Soares Galvão. Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Física “Gleb Wataghin”.</p> <p style="text-align: center;">1. Métodos semi-empíricos. 2. Hidrocarbonetos policíclicos aromáticos. 3. Relação estrutura-atividade (Bioquímica). I. Galvão, Douglas Soares. II. Universidade Estadual de Campinas. Instituto de Física “Gleb Wataghin”. III. Título.</p>
-------	--

- **Título em inglês:** Study of structure-activity in polycyclic aromatic hydrocarbons
- **Palavras-chave em inglês (Keywords):**
 1. Semi-empirical methods
 2. Polycyclic aromatic hydrocarbons
 3. Structure-activity relationships (Biochemistry)
- **Área de concentração:** Física da matéria condensada
- **Titulação:** Mestre em física
- **Banca examinadora:**

Prof. Douglas Soares Galvão
Prof. Pedro Geraldo Pascutti
Prof. Yakov Veniaminovitch Kopelevitch
- **Data da defesa:** 29.07.2003



MEMBROS DA COMISSÃO JULGADORA DA TESE DE Mestrado de **KARLA SOUZA TROCHE – R.A. 010560** APRESENTADA E APROVADA AO INSTITUTO DE FÍSICA “GLEB WATAGHIN”, DA UNIVERSIDADE ESTADUAL DE CAMPINAS, EM 29/07/2003.

COMISSÃO JULGADORA:

Prof. Dr. Douglas Soares Galvão (Orientador da Candidata) – DFA/IFGW/UNICAMP

Prof. Dr. Pedro Geraldo Pascutti – IBCCF/UFRJ

Prof. Dr. Yakov Veniaminovitch Kopelevitch – DFA/IFGW/UNICAMP

AGRADECIMENTOS

Agradeço

Ao Professor Douglas Galvão pela oportunidade de orientação, pela paciência, pela amizade e pela confiança, flexibilidade e dedicação.

Ao meu muito e queridíssimo pai Dr. Álvaro del Portillo, que com muito carinho e dedicação cuidou dos muitos detalhes e dificuldades, destes dois anos de pós-graduação.

Ao pessoal do GSONM (Scheila, Fernando, Sergio, Giro, Vitor, Rossana) grupo excelente de trabalho e convívio, agradeço pelas ajudas e orientação no trabalho de pesquisa e disciplinas do mestrado

Aos professores Cabrera, Kleiman, Sakanaka e Luzzi.

Aos meus familiares: meus irmãos, Érika, Carlos, José e meus pais, Suely e Adolfo que sempre me apoiaram.

A minha sogra Olga pelo apoio e grande ajuda neste tempo.

Aos meus cunhados: Claudia, José, Sergio, Fabiana, Martin, Jorge pelo apoio em nossos estudos.

A secretaria do DFA: Lúcia, sempre sorridente e prestativa.

E de forma muito especial: “Na saúde e na enfermidade, nos momentos difíceis e alegres” ao meu marido e amigo Daniel pelo amor, pelo carinho, pela compreensão e pela paciência. Sem o seu apoio este Mestrado não se realizaria.

Resumo

Neste trabalho apresentamos estudos de estrutura-atividade realizados para 81 hidrocarbonetos policíclicos aromáticos (PAHs) tentando identificar compostos carcinogênicos.

Em particular, empregamos uma nova metodologia desenvolvida para tratar o problema da correlação estrutura geométrica com atividade biológica, denominada Metodologia de Índices Eletrônicos (MIE). Utilizamos três métodos semi-empíricos para estudar a dependência entre a qualidade dos resultados obtidos pela MIE e o Hamiltoniano utilizado.

Para validarmos os descritores utilizados pela MIE, investigamos a relação entre a atividade experimental dos PAHs e descritores teóricos através de cinco metodologias de reconhecimento de padrões: a Análise de Componentes Principais (PCA), Análise Hierárquica de Agrupamentos (HCA), K-ésimo vizinho mais próximo (KNN), Soft Independent Modeling of Class Analogies (SIMCA) e as Redes Neurais Artificiais (NN). Para estas investigações distintas pudemos correlacionar a atividade dos PAHs estudados com parâmetros teóricos, em sua maioria eletrônicos, onde os parâmetros utilizados na MIE foram selecionados pelos diferentes métodos. Este estudo valida estatisticamente a MIE como uma nova metodologia capaz de identificar compostos biologicamente ativos, e com um custo computacional menor que técnicas convencionais de Relação estrutura-atividade (SAR) e apresentando um desempenho, em geral superior.

Abstract

In this work we present the study structure-activity realized for 81 polycyclic aromatic hydrocarbons (PAHs) trying to identify carcinogenic compounds.

Particularly, we used a new methodology developed to deal with the problem of correlation between geometrical structure and biological activity, the Electronic Indices Methodology (EIM). We used three semi-empirical methods to analyze the dependence between the quality of results obtained through EIM and the Hamiltonian used.

In order to validate the descriptors used in EIM, we investigated the relationship between the experimental activity of PAHs and the theoretical descriptors through five methodologies of pattern recognition: Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), Kth Nearest Neighbor (KNN), Soft Independent Modeling of Class Analogies (SIMCA) and Neural Networks (NN). From these different investigations, we could correlate the activity of PAHs studied with theoretical parameters, almost all electronic, where the used parameters on EIM were selected with the different methods. These studies validate the statistical value of electronic parameters derived from EIM analysis and their ability to identify active compounds. The EIM out performed more standard Structure-Activity Relationship (SAR) methodologies.

Conteúdo

1 Introdução e Objetivos	1
Referências	4
2 Métodos de Cálculo	5
2.1. Introdução.....	5
2.2. Mecânica Quântica de Moléculas.....	5
2.2.1. Aproximação de Born-Oppenheimer	7
2.2.2. Aproximação de Hartree-Fock.....	8
2.2.3. O procedimento LCAO (Combinação Linear de Orbitais Atômicos).....	13
2.2.4. Métodos semi-empíricos.....	17
2.3. Quimiometria.....	19
2.3.1. Análise de componentes principais (PCA)	19
2.3.2. Análise hierárquica de agrupamentos (HCA)	22
2.3.3. K-ésimo vizinho mais próximo (KNN)	23
2.3.4. Soft Independent Modeling of Class Analogies (SIMCA)	24
2.4. Redes neurais.....	25
Referências	31
3 Propriedades físico-químicas	32
3.1. Introdução.....	32
3.2. Parâmetros hidrofóbicos	33
3.2.1. Coeficiente de partição	33

3.3. Parâmetros estereoquímicos	33
3.4. Parâmetros eletrônicos.....	33
3.4.1. Energias do HOMO e LUMO.....	34
3.4.2. Contribuição dos orbitais atômicos.....	34
3.4.3. Diferença de energia HOMO-LUMO	34
3.5. Metodologia dos índices eletrônicos	35
Referências	40
4 Estudo estrutura-atividade dos hidrocarbonetos policíclicos aromáticos	41
4.1. Introdução.....	41
4.4.1. Hidrocarbonetos carcinogênicos.....	42
Referências	50
5 Resultados e discussões	51
5.1. Otimização geométrica e análise conformacional	51
5.2. Metodologia dos índices eletrônicos	54
5.3. Análise de componentes principais	58
5.4. Análise hierárquica de agrupamentos	71
5.5. K-ésimo vizinho mais próximo	75
5.6. Soft Independent Modeling of Class Analogies	79
5.7. Redes neurais.....	84
5.8. Resumo dos resultados	87
Referências	88
6 Conclusões	89

Capítulo 1

Introdução e Objetivos

O descobrimento de compostos biologicamente ativos e responsáveis pela indução de câncer teve e tem um processo complexo e que envolve muitas disciplinas científicas.

Entre os agentes químicos conhecidos como indutores de câncer os Hidrocarbonetos Aromáticos Policíclicos (PAHs) desempenham um papel muito importante. Como classe os PAHs perdem somente para as micotoxinas em potência carcinogênica¹.

Na década de 30 o descobrimento das propriedades carcinogênicas dos benzo[a]pireno, dibenz[a,h]antraceno, e outros poliarenos, determinou pela primeira vez para a ciência biomédica a evidência de desordens causados não pela presença de microorganismos, mas sim pela presença de compostos químicos relativamente simples. E foi o início da tentativa de explicar a razão pela qual moléculas estruturalmente relacionadas apresentam tão grande variação de atividade carcinogênica, desde altamente ativas até completamente inertes.

Cook e colaboradores² foram os pioneiros nas investigações que relacionam estrutura-atividade. Eles identificaram uma relação entre atividade carcinogênica (tumores malignos em ratos) e alguns aspectos geométricos das moléculas. E a partir desses trabalhos outras investigações como as de Coulson, Schmidt and Svarthölm³

exploraram esses aspectos tentando correlacionar geometria e aspectos eletrônicos dos PAHs. Os trabalhos de A. Pullman e B. Pullman^{4,5}, propuseram a teoria das regiões K e L baseados em cálculos de química quântica, usando a teoria de Hückel simples expressando os cálculos em valores de índices críticos sobre regiões moleculares específicas, outras teorias utilizando análise estatística, redes neurais e métodos de inteligência artificial^{6,7}, foram testadas, mas nenhuma é totalmente consistente com os dados experimentais disponíveis. Algumas delas funcionam bem para um subconjunto dos dados experimentais, mas falham para outros, e vice-versa. Assim, uma teoria simples e confiável, que possa prever e diferenciar, pelo menos ao nível qualitativo, se determinada molécula da família de PAH será ou não cancerígena continua em desenvolvimento.

Em 1996 uma nova metodologia que relaciona estrutura molecular com atividade biológica (SAR, Structure Activity Relationship), denominada Metodologia dos Índices Eletrônicos (MIE)⁸, foi proposta para identificar a atividade carcinogênica dos PAHs, e mostrou ser muito eficiente para determinar se um PAH será cancerígeno ou não.

A metodologia dos Índices eletrônicos está baseada nos conceitos físicos de densidade de estados (local e total), que permite obter informações detalhadas sobre as contribuições de determinadas regiões geométricas para a reatividade química e, conseqüentemente sobre o comportamento bioquímico. Esse estudo permitiu estabelecer que a diferença entre o valor de energia do penúltimo e o valor da energia do último orbital molecular ocupado em conjunto com valores críticos de contribuição para a densidade local de estados sobre o anel que contém a maior ordem de ligação, podem ser usados de forma muito eficiente para determinar a atividade carcinogênica dos PAHs estudados.

Essa metodologia foi aplicada também a outras classes de compostos orgânicos, tais como antibióticos⁹, antitumorais¹⁰, hormônios anticoncepcionais¹¹, esteróides¹², inibidores de integrase do HIV-1¹³, sempre com uma taxa de acerto de 85-90% na identificação da Atividade/Inatividade dos compostos.

Nesse contexto, o presente trabalho realiza testes mais detalhados sobre a metodologia MIE, determinando se existe uma dependência explícita entre a

qualidade dos resultados obtidos e o hamiltoniano eletrônico utilizado. Nós utilizamos os Hamiltonianos semi-empíricos PM3 (Parametric Method 3)¹⁴, PM5 (Parametric Method 5), AM1 (Austin Method One)¹⁵ para realizar testes comparativos, e também investigar a possibilidade da existência de outras regiões moleculares que potencialmente possam ser correlacionadas com a atividade biológica, estendendo o conjunto de PAHs também para compostos metilados.

Para validarmos estatisticamente e compararmos a MIE com outras metodologias mais testadas e conhecidas na literatura na área de reconhecimento de padrões, nós realizamos o estudo comparativo dos resultados obtidos da MIE (utilizando os vários Hamiltonianos semi-empíricos) com os métodos de reconhecimento de padrões: Análise de Componentes Principais (PCA - Principal Component Analysis)¹⁶, Análise Hierárquica de Agrupamentos (HCA - Hierarchical Cluster Analysis)¹⁶, K-ésimo vizinho mais próximo (KNN- Kth Nearest Neighbor)¹⁶, Soft Independent Modeling of Class Analogies (SIMCA)¹⁶ e as Redes Neurais (NN - Neural Networks)¹⁷.

O uso combinado dessas metodologias, permitiu identificarmos os parâmetros eletrônicos que possam ser correlacionados com a atividade biológica, e estabelecer a confiabilidade ou não dos descritores utilizados na metodologia MIE.

No capítulo 2 descrevemos a metodologia utilizada na modelagem molecular de otimização geométrica dos PAHs com os diferentes métodos semi-empíricos, e também os métodos de reconhecimento de padrões, a seguir os parâmetros físico-químicos obtidos e utilizados para correlacionar estrutura-atividade são discutidos. No capítulo 4 apresentamos os compostos utilizados no trabalho de tese, para finalmente apresentar os resultados e discussões dos mesmos, que nos levaram a conclusões bastante interessantes.

Referencias

- ¹ R. G. Harvey and N. E. Geacintov, *Acc. Chem. Res.* **21**,66 (1988).
- ² C. A. Coulson, *Adv. Canc. Res.* **1**, 1 (1953) e referências citadas
- ³ W. Herndon, *Int. J. Quant. Chem.Quant. Biol. Symp.* **1**, 123 (1974).
- ⁴ A. Pullman, *Bull. Soc. Chem. Fr.* **21**, 595 (1954).
- ⁵ A. Pullman and B. Pullman, *Adv. Canc. Res.* **3**, 117 (1955).
- ⁶ L. V. Szentpály, *J. Amer. Chem. Soc.* **106**, 6021 (1984).
- ⁷ D. Villemin, D. Cherqaoui and A. Mesbah, *J.Chem. Inf. Comput. Sci.* **34**, 1288 (1994).
- ⁸ P. M. V. B. Barone, A. Camilo Jr., and D. S. Galvão, *Phys. Rev. Lett.* **77**, 1186 (1996).
- ⁹ L. L. E. Santo and D. S. Galvão, *J. Mol. Struct (THEOCHEM)* **464**, 273 (1999).
- ¹⁰ L. L. Mazzali and D. S. Galvão, *J. Chem. Phys.* B-submetido.
- ¹¹ R. S Braga, R. Vendrame and D. S. Galvão, *J. Chem. Inf. Comp. Sci.* **40**, 1377 (2000).
- ¹² R. Vendrame, R. S Braga and D. S. Galvão, *J. Chem. Inf. Comp. Sci*
- ¹³ M. Cyrillo and D. S. Galvão, *J.Mol. Struct. (THEOCHEM)* **464**, 267 (1999).
- ¹⁴ J. J. P. Stewart, *J. Comp. Chem.* **10**, 209 (1991); **10**, 221 (1991).
- ¹⁵ M. J. S. Dewar, E. G. Zoebisch, E. F. Healy eand J. J. P. Stewart, *J. Amer. Chem. Soc.* **107**, 3902 (1985).
- ¹⁶ K. R. Beebe, R. J. Pell, and M. B. Seasholtz, *Chemometrics: A Practical Guide* (Wiley, New York, 1998).
- ¹⁷ J. Zupan and J. Gasteiger, *Neural Networks for Chemists* (VCH, New York, 1993).

Capítulo 2

Métodos de Cálculo

2.1 Introdução

A modelagem molecular¹ consiste na visualização gráfica e na representação geométrica de uma molécula através de um conjunto de técnicas computacionais que aliam métodos de química teórica e dados experimentais. Com esses métodos teóricos pode-se avaliar e prever certas características moleculares (conformação de mínima energia e propriedades físico-químicas), que são de fundamental importância no entendimento das correlações estrutura-atividade. Isto permite o entendimento racional (baseado na estrutura) das características biológicas dos compostos.

Os três principais métodos teóricos computacionais utilizados para se calcular as propriedades moleculares classificam-se em empíricos (mecânica molecular), semi-empíricos e *ab initio*.

2.2 Mecânica Quântica de Moléculas

Podemos tratar o sistema molecular dentro do formalismo da mecânica quântica^{2,3} que dará a descrição da distribuição eletrônica detalhada, e assim obter, dos cálculos, propriedades associadas à estrutura eletrônica da molécula.

A descrição quântica dos estados eletrônicos estacionários de uma molécula é obtida através da Equação de Schödinger Independente do Tempo:

$$\hat{\mathbf{H}}\Psi(\vec{\mathbf{r}}, \vec{\mathbf{R}}) = \mathbf{E}\Psi(\vec{\mathbf{r}}, \vec{\mathbf{R}}) \quad (1)$$

Onde $\hat{\mathbf{H}}$ é o operador hamiltoniano do sistema de partículas; $\Psi(\mathbf{r}, \mathbf{R})$ é a função de onda de um estado molecular total, dependente das coordenadas eletrônicas e nucleares. Ela contém todas as informações sobre o estado de sistema que ela descreve, e \mathbf{E} é a energia deste estado.

Quando as soluções são geradas sem referencias de dados experimentais ou aproximações, o método é *ab initio* (do latim: do início), caso contrario, o modelo é semi-empírico.

O primeiro passo para se tratar o sistema molecular quanticamente é estabelecer uma Hamiltoniana para seu conjunto de núcleos e elétrons, em que senão todas, a maioria das interações existentes sejam consideradas.

O Hamiltoniano molecular, onde consideramos núcleos e elétrons como massas pontuais e desprezamos as interações relativísticas e interações spin-órbita é dado por:

$$\mathbf{H}_{\text{tot}} = \mathbf{T}_n + \mathbf{T}_e + \mathbf{V}_{ee} + \mathbf{V}_{ne} + \mathbf{V}_{nn} \quad (2)$$

$$\hat{\mathbf{H}} = -\sum_{A=1}^M \frac{1}{2\mathbf{M}_A} \nabla_A^2 - \sum_{i=1}^N \frac{1}{2} \nabla_i^2 + \sum_{i=1}^N \sum_{j \neq i}^N \frac{1}{|\vec{\mathbf{r}}_{ij}|} - \sum_{A=1}^M \sum_{i=1}^N \frac{\mathbf{Z}_A}{|\vec{\mathbf{r}}_{iA}|} + \sum_{A=1}^M \sum_{B \neq A}^M \frac{\mathbf{Z}_A \mathbf{Z}_B}{|\vec{\mathbf{R}}_{AB}|} \quad (3)$$

O primeiro termo do Hamiltoniano refere-se à energia cinética de cada núcleo, o segundo e o terceiro termos são: a energia cinética e de repulsão eletrostática dos elétrons, respectivamente. O próximo termo representa a energia de atração entre os núcleos e os elétrons, e o último termo refere-se à energia de repulsão entre os núcleos.

Na equação (3) \mathbf{A} , \mathbf{B} referem-se aos núcleos, $|\vec{\mathbf{R}}_{AB}|$ designa a distância entre os núcleos A e B, $|\vec{\mathbf{r}}_{iA}|$ designa a distância entre o núcleo A e o elétron i, \mathbf{Z}_A refere-se à carga do núcleo A; i, j referem-se aos elétrons, $|\vec{\mathbf{r}}_{ij}|$ designa a distância entre os elétrons i e j, ∇_i^2 é o operador gradiente relativo às coordenadas eletrônicas e ∇_A^2 é o operador gradiente relativo às coordenadas nucleares.

A aplicação desse operador Hamiltoniano ao sistema molecular possui solução exata somente para o caso diatômico, sendo que para o restante dos casos este é um problema matematicamente intratável.

Algumas simplificações devem ser impostas, com o intuito de tornar solúveis, mesmo aproximadamente, problemas que envolvem um número muito grande de variáveis.

2.2.1 Aproximação de Born-Oppenheimer

Uma das formas de simplificar a equação (3) é procurar separar o movimento eletrônico de o movimento nuclear, isso é feito mantendo-se os núcleos fixos durante cada ciclo do movimento eletrônico, o que é conhecido como **aproximação de Born-Oppenheimer**⁴. Pelo fato de a massa nuclear ser muito maior que a eletrônica, é uma aproximação razoável considerar que os núcleos mudem de posição tão lentamente que permitam aos elétrons se ajustarem muito rapidamente à nova configuração nuclear. Desse modo os núcleos poderiam ser tomados como fixos para uma análise de movimento eletrônico. Isto se reflete em desconsiderar os termos cinéticos dos núcleos no Hamiltoniano e considerar a repulsão entre os núcleos constante. Assim qualquer constante adicionada ao operador só é adicionada aos autovalores e as autofunções do operador não são afetadas. Os termos restantes descrevem o movimento de N elétrons no campo de M cargas pontuais. A **aproximação de Born-Oppenheimer** desacopla matematicamente o problema de calcular a energia de um sistema molecular em dois problemas separados, um eletrônico (4) e outro nuclear (6).

Podemos reescrever o Hamiltoniano total do sistema molecular para núcleos fixos, incluindo a constante de repulsão nuclear como sendo a soma de uma parte eletrônica (5) e outra nuclear (7) e a função de onda total do sistema, que é autofunção da equação de Schödinger, como sendo um produto das funções de onda nuclear e eletrônica (8).

$$\hat{H}_{el} \Psi_{el}(\{\vec{r}_i\}; \{\vec{R}_A\}) = E(\vec{R}_A) \Psi_{el}(\{\vec{r}_i\}; \{\vec{R}_A\}), \quad (4)$$

onde

$$\hat{H}_{el} = -\sum_{i=1}^I \frac{1}{2} \nabla_i^2 - \sum_{A=1}^M \sum_{i=1}^N \frac{Z_A}{|\vec{r}_{Ai}|} + \sum_{i=1}^N \sum_{j>i}^N \frac{1}{|\vec{r}_{ij}|} \quad (5)$$

para o movimento eletrônico, a função de onda que descreve o movimento dos elétrons é dependente explicitamente das coordenadas eletrônicas, mas parametricamente dependente das coordenadas nucleares.

E

$$\hat{H}_N \Psi_N(\{\vec{R}_A\}) = E \Psi_N(\{\vec{R}_A\}) \quad (6)$$

onde

$$\hat{H}_N = -\sum_{A=1}^M \frac{1}{2M_A} \nabla_A^2 + V(\vec{R}_A) \quad (7)$$

para o movimento nuclear onde a energia eletrônica e a energia de repulsão dos núcleos atuam juntas como $V(\vec{R}_A)$.

A correspondente aproximação da função de onda total da equação (1) é

$$\Phi(\vec{r}_i; \vec{R}_A) = \Psi_{el}(\vec{r}_i; \vec{R}_A) \Psi_N(\vec{R}_A) \quad (8)$$

Assim a, equação de Schödinger para o sistema teria a forma (9).

$$(\mathbf{H}_{el} + \mathbf{H}_N) \Psi_{el} \Psi_N = E \Psi_{el} \Psi_N \quad (9)$$

Onde a energia total da aproximação Born-Oppenheimer inclui as energias eletrônicas, de vibração, de rotação e translação da molécula.

Embora a **aproximação de Born-Oppenheimer** represente uma simplificação considerável, a dificuldade fundamental ainda persiste: o problema eletrônico, equação (5), envolve muitas partículas, e deve ser também simplificado para se tornar aproximadamente solúvel.

2.2.2 Aproximação de Hartree-Fock

Consideremos agora a resolução do problema eletrônico. A equação de Schrödinger eletrônica (5) não permite soluções exatas devido às interações eletrônicas.

Uma das aproximações que se faz é a chamada **Aproximação de Hartree-Fock**, onde a interação de cada elétron com os restantes N-1 elétrons (último termo da eq. (5)) foi substituída por Hartree pela interação com uma distribuição contínua de carga ρ_j , e o potencial correspondente é:

$$U_i^{\text{elec}} = \sum_{j \neq i}^N -e \int d^3r' \rho_j \frac{1}{|\mathbf{r}_{ij}|} \quad (10)$$

e para um elétron no estado ψ_i , a distribuição de carga é dada por:

$$\rho_i = -e |\phi_i|^2 \quad (11)$$

Podemos resolver a equação de Schrödinger para cada elétron independente para uma função de onda ϕ que se estende por toda molécula, e que descreva o elétron individualmente (12):

$$\hat{F}(r_i) \phi_i(r_i) = \varepsilon_i \phi_i(r_i) \quad (12)$$

Essa equação possui múltiplas soluções ϕ_i e ε_i . As funções ϕ_i são chamadas de orbitais moleculares (OM) para o elétron i e a energia ε_i de um elétron no orbital ϕ_i e é chamada de energia orbital. E a função de onda total, $\Psi(1, 2, \dots, 2n)$, seria dada pelo produto de todas as funções mono eletrônico:

$$\Psi(1, 2, \dots, 2n) = \phi_1(1) \phi_2(2) \dots \phi_{2n-1}(2n-1) \phi_{2n}(2n) \quad (13)$$

A função de onda escrita na forma da eq.13 é conhecida pelo nome de **produto de Hartree** e sugere que a função que descreve um elétron qualquer é completamente independente de todos os outros elétrons. Em outras palavras, a função de onda representada pela eq.13 corresponderia ao que se chama de **modelo de partículas independentes**.

O sistema de equações (12) são as **Equações de Hartree**. O método usado para resolvê-las deve usar um conjunto fisicamente razoável de funções $\phi_i(\mathbf{r})$, que são usadas na equação (5) no termo da interação eletrônica e obter a solução deste sistema linear de equações em uma nova aproximação. O processo é repetido, e repete-se o ciclo otimizando-se até que o valor do potencial obtido após a otimização sofra uma mudança suficientemente pequena se comparada com o ciclo anterior. Por isso o método é chamado de autoconsistente.

Podemos obter a expressão para a energia eletrônica total do sistema como:

$$E_{ele} = \sum_{i=1}^N \varepsilon_i - \sum_{j>i} \sum_{i=1}^N \iint e^2 \frac{|\psi_i|^2 |\psi_j|^2}{r_{ij}} d\mathbf{v}_i d\mathbf{v}_j \quad (14)$$

onde se faz necessário considerar que a repulsão intereletrônica é contada duas vezes.

Na descrição completa de um elétron se faz necessário especificar seu spin. Como dois elétrons não podem ocupar funções idênticas (princípio da exclusão de Pauli) uma das formas de se garantir isso é escrever a função de onda multieletrônica como um produto antissimetrizado de spin-orbitais moleculares. Como os elétrons possuem spin, além das coordenadas espaciais, a função de onda também deve possuir coordenada de spin, que, podem assumir duas configurações, *up* ou *down*. As funções assim construídas são chamadas Spin-Orbitais Moleculares (SOM), $\chi(\bar{\mathbf{x}})$, e são comumente formadas pela multiplicação de um orbital atômico ou molecular por uma função de spin:

$$\chi_{2i-1}(\bar{\mathbf{x}}) = \phi_i(\bar{\mathbf{r}})\alpha \quad \text{para spin } +1/2 \quad (15)$$

$$\chi_{2i}(\bar{\mathbf{x}}) = \phi_i(\bar{\mathbf{r}})\beta \quad \text{para spin } -1/2 \quad (16)$$

$i=1,2,\dots,k$ e $\bar{\mathbf{x}}=(\bar{\mathbf{r}},\sigma)$ inclui as coordenadas espaciais e de spin. Os spin-orbitais satisfazem à condição de ortonormalidade:

$$\langle \chi_i(\bar{\mathbf{x}}) | \chi_j(\bar{\mathbf{x}}) \rangle = \delta_{ij} \delta_{\alpha\beta} \quad (17)$$

As propriedades dos determinantes garantem que as funções de onda eletrônicas serão antisimetrizadas e, por tanto, que o Princípio de exclusão de Pauli será satisfeito.

Em um determinante de **Slater** cada coluna refere-se a um mesmo spin-orbital e cada linha a um mesmo elétron, ou vice-versa:

$$\Psi^{\text{HF}}(\vec{1}, \vec{2}, \dots, \vec{k}) = (\mathbf{N}!)^{-1/2} \begin{vmatrix} \chi_1(\vec{1}) & \chi_2(\vec{1}) & \dots & \chi_k(\vec{1}) \\ \chi_1(\vec{2}) & \chi_2(\vec{2}) & \dots & \chi_k(\vec{2}) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_k(\vec{k}) & \chi_k(\vec{k}) & \dots & \chi_k(\vec{k}) \end{vmatrix} \quad (18)$$

onde $\vec{1}, \vec{2}, \dots, \vec{k}$ são as coordenadas dos elétrons 1, 2, ..., k respectivamente e $(\mathbf{N}!)^{-1/2}$ o fator de normalização.

Obtendo a função de onda eletrônica apropriada que descreve um elétron com o spin orbital, o seguinte é considerar funções de onda para uma coleção de elétrons, isto é, N funções de onda eletrônicas.

$$\Psi_e(\mathbf{r}_1, \dots, \mathbf{r}_N) = \sum_{\mathbf{p}=1}^{\mathbf{N}!} (\pm 1)^{\mathbf{p}} \hat{\mathbf{p}} [\chi_m(\vec{1}), \chi_n(\vec{2}), \dots, \chi_k(\vec{N})] \quad (19)$$

onde \mathbf{p} é o operador de permutação, a soma é sobre todas as permutações dos estados $\psi_i(\mathbf{r}_i)$ e $\mathbf{p}(\pm)$ indica -1 , para uma permutação ímpar e $+1$, para uma permutação par.

A avaliação da energia total do sistema molecular ($E_{\text{HF}} = \langle \Psi | \hat{\mathbf{H}}_{\text{el}} | \Psi \rangle$) quando a função de onda é escrita na forma (19) nos leva à:

$$E_{\text{HF}} = 2 \sum_i^{\frac{N}{2}} \mathbf{H}_{ii} + \sum_i^{\frac{N}{2}} \sum_j^{\frac{N}{2}} (2\mathbf{J}_{ij} - \mathbf{K}_{ij}) + \mathbf{V}_{\text{NN}} \quad (20)$$

com

$$\mathbf{H}_{ii} = \langle \psi_i(\vec{1}) | \hat{\mathbf{h}}_i | \psi_i(\vec{1}) \rangle \quad (21)$$

$$\mathbf{J}_{ij} = \langle \psi_i(\vec{1})\psi_j(\vec{2}) | \frac{e^2}{|\vec{r}_{12}|} | \psi_i(\vec{1})\psi_j(\vec{2}) \rangle \quad (22)$$

$$\mathbf{K}_{ij} = \langle \psi_i(\vec{1})\psi_j(\vec{2}) | \frac{e^2}{|\vec{r}_{12}|} | \psi_j(\vec{1})\psi_i(\vec{2}) \rangle \quad (23)$$

O termo H_{ii} representa a energia de 1-elétron no orbital molecular φ_i sujeito a um campo produzido por núcleos fixos. As integrais do tipo J_{ij} são chamadas integrais de Coulomb, e podemos dizer que fisicamente elas representam a interação entre duas densidades de carga ($\psi_i \psi_i^*$ e $\psi_j \psi_j^*$). As integrais K_{ij} são chamadas integrais de troca por possuírem dois produtos de funções de onda que diferem pela troca dos índices eletrônicos $\vec{1}$ e $\vec{2}$. Matematicamente o que temos nesse caso é a interação de uma distribuição eletrônica com outro elétron na mesma distribuição. Fisicamente é difícil atribuir um significado a essas integrais, mas elas são resultado do princípio de antissimetria das funções de onda moleculares e refletem a energia de estabilização devida à correlação de elétrons com spins antiparalelos.

A determinação dos orbitais $\psi_i(\vec{r})$ é feita aplicando-se o princípio variacional à equação de energia (20), sujeita à condição de ortonormalidade da função de onda. A melhor função de onda ψ é aquela que minimiza a energia do estado fundamental, e o método variacional nos mostra que os OM ótimos devem ser autofunções da equação (12) de autovalores, onde $\hat{F}(\mathbf{r}_i)$ é o denominado operador de Fock, que é um operador da energia efetiva de um elétron e cuja expressão é dada por:

$$\hat{F} = \hat{h}(\vec{1}) + \sum_j^{n/2} (2J_j(\vec{1}) - K_j(\vec{1})) \quad (24)$$

onde $h_i J$

$$\hat{h}_i = -\frac{1}{2}\nabla_i^2 + \sum_{A=1}^M \frac{Z_A}{|\vec{r}_{Ai}|} \quad (25)$$

$$\hat{J}_j(\vec{1})f(\vec{1}) = f(\vec{1}) \int \psi_j^*(\vec{2}) \frac{1}{|\vec{r}_{12}|} \psi_j(\vec{2}) \quad (26)$$

$$\hat{K}_j(\vec{1})f(\vec{1}) = \psi_j(\vec{1}) \int \psi_j^*(\vec{1}) \frac{1}{|\vec{r}_{12}|} f(\vec{2}) \quad (27)$$

h_i é o operador Hamiltoniano de 1-elétron. $J_j(1)$ é o chamado operador de Coulomb, pois origina os termos de energia correspondentes às repulsões entre as densidades de cargas (26), e $K_j(1)$ é os chamados operadores de troca, que leva, depois de aplicado à função spin orbital molecular, às integrais de troca (27).

Note que se os spins associados aos orbitais ψ_i e ψ_j forem diferentes, os termos referentes ao operador $K_j(1)$ serão nulos. Isso acontece porque a integração sobre o espaço das coordenadas de spin do elétron 1 (ou 2) na equação (27) leva à integração sobre duas funções de spins ortogonais.

O problema original de encontrar as soluções da equação de Schrödinger independente do tempo (5) foi reduzido, a encontrar as soluções para as equações de autovalores (12), conhecidas com equações de Fock. A maior vantagem com relação ao problema anterior é o fato do operador de Fock ser um operador de 1-elétron, porém a equação de autovalores continua tendo um número infinito de soluções.

2.2.3 O Procedimento LCAO (Combinação Linear de Orbitais Atômicos)

Devemos resolver agora um conjunto de equações diferenciais de um elétron.

O próximo passo é expandir as funções OM em um conjunto finito de funções de base conhecidas. Essas funções de base são freqüentemente escolhidas como sendo as funções orbitais atômicas pelo fato dessa escolha facilitar a interpretação dos resultados, pois as propriedades moleculares estão diretamente relacionadas aos átomos constituintes da molécula. Esse procedimento é conhecido como combinação linear dos orbitais atômicos (LCAO -linear combination of atomic orbitals):

$$\psi_i = \sum_{s=1}^h c_{si} \varphi_s \quad (28)$$

Substituindo a expansão acima na equação (12), multiplicando por φ_s^* e integrando, obtemos as equações de Hartree-Fock-Roothaan (HFR) (29):

$$\sum C_{si} (\hat{\mathbf{F}}_{rs} - \varepsilon_i \mathbf{S}_{rs}) = 0, \mathbf{r} = 1, 2, \dots, \mathbf{h} \quad (29)$$

onde definimos:

$$\mathbf{F}_{rs} \equiv \langle \varphi_r | \hat{\mathbf{F}} | \varphi_s \rangle, \quad (30)$$

$$\mathbf{S}_{rs} \equiv \langle \varphi_r | \varphi_s \rangle. \quad (31)$$

\mathbf{F}_{rs} (30) é a matriz densidade da Hamiltoniana de 1-elétron com respeito aos orbitais atômicos φ_r e φ_s . A matriz \mathbf{S}_{rs} (31) é chamada matriz de superposição.

O problema agora passa a ser resolver esse conjunto de equações algébricas não-lineares. Essa não-linearidade vem do fato do operador de Fock depender dos coeficientes \mathbf{C}_{si} ($\mathbf{F}=\mathbf{F}(\mathbf{C})$), que não são conhecidos *a priori*.

Para as equações (29) tenham soluções diferentes da trivial, a seguinte condição deve ser satisfeita:

$$\det(\mathbf{F}_{rs} - \varepsilon_i \mathbf{S}_{rs}) = 0 \quad (32)$$

As raízes desta equação secular fornecem as energias dos orbitais moleculares ε_i .

Uma maneira de resolver o sistema de equações HFR (32) é através do método iterativo conhecido como método do campo autoconsistente (SCF). Esquemáticamente o procedimento adotado é o seguinte:

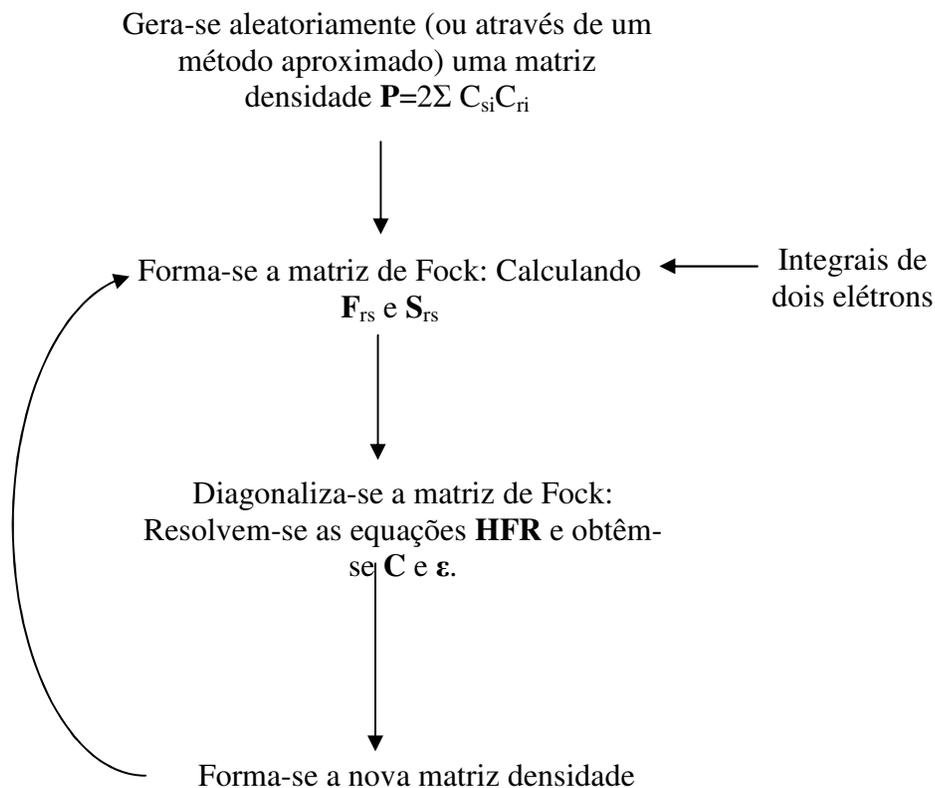


Fig.1: Esquema do procedimento iterativo da resolução do sistema de equações HFR.

Esse processo iterativo prossegue até que a energia ou a densidade de carga não sofra mudança dentro de uma precisão especificada entre duas iterações sucessivas.

Ao atingir essa convergência a solução das equações HFR fornecerá o conjunto de coeficientes de cada orbital molecular na base dos orbitais atômicos juntamente com a energia associada com cada um dos orbitais moleculares.

Podemos sumarizar o procedimento para se realizar um cálculo Hartree-Fock de um sistema molecular como:

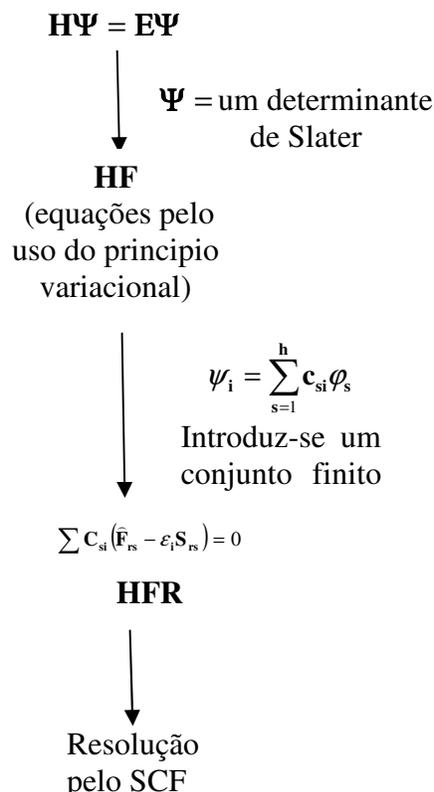


Fig.2: Esquema do cálculo Hartree-Fock de um sistema molecular.

Na resolução das equações de Roothaan (29) os termos dos elementos da matriz de Fock \mathbf{F}_{rs} são expressos em termos da função de base φ e na equação (26) e feita a substituição de f por φ e usando a expansão (28), multiplicando por φ^* e integrando sobre o espaço de coordenadas do elétron 1 nos temos as integrais de repulsão ($r s t u$). O maior obstáculo na resolução das equações de Hartree-Fock-Roothaan está no grande número de integrais de dois elétrons a serem calculadas. Essas integrais podem possuir um ($r^\alpha s^\alpha t^\alpha u^\alpha$), dois ($r^\alpha s^\alpha t^\alpha u^\beta$), três ($r^\alpha s^\alpha t^\beta u^\gamma$) ou quatro ($r^\alpha s^\alpha t^\gamma u^\delta$) centros, dependendo do número de centros atômicos (α, β, γ ou δ) nos quais estão localizadas as funções orbitais moleculares (r, s, t ou u). Os métodos *ab initio* (primeiros princípios) caracterizam-se pelo cálculo explícito de todas essas integrais, que são originadas da utilização de um conjunto finito de funções de base arbitrárias.

São cálculos extremamente custosos computacionalmente, o que restringe sua aplicação a sistemas moleculares pequenos. Outros métodos, que também adotam

uma formulação rigorosa da mecânica quântica, utilizam-se de algumas aproximações para tentar resolver essas integrais de dois elétrons. Os cálculos semi-empíricos podem utilizar hamiltonianos mais simples, desprezar algumas integrais decorrentes da formulação de Schrödinger ou ainda substituí-las por parâmetros obtidos de resultados experimentais, daí o nome semi-empírico, as aproximações variando de acordo com o método adotado⁵. As aproximações adotadas tornam esses métodos extremamente rápidos computacionalmente em comparação aos métodos *ab initio*, o que permite que sistemas maiores sejam investigados.

O conjunto das funções atômicas que descrevem os orbitais moleculares é chamado conjunto de base. Tanto nos métodos *ab initio* quanto nos métodos semi-empíricos é comum a utilização de um conjunto de base mínima. O conjunto de base mínima caracteriza-se pela utilização da combinação linear do menor número possível de funções atômicas que possibilite descrever os orbitais atômicos ocupados do átomo. A forma dessas funções de base pode variar, mas as funções de onda tipo Gaussianas (33), são as mais utilizadas nos cálculos *ab initio*, enquanto as funções de onda tipo Slater (34) são as mais empregadas nos cálculos semi-empíricos:

$$\psi(\mathbf{r}) = \exp(-\alpha r^2) \quad (33)$$

$$\psi(\mathbf{r}) = r^{n-1} \exp^{-\zeta r} Y_{lm}(\theta, \phi) \quad (34)$$

2.2.4 Métodos Semi-Empíricos

Os principais métodos semi-empíricos envolvem dois tipos de aproximação.

O primeiro passo na redução do problema computacional é considerar de forma explícita somente os elétrons de valência, os de caroço são considerados pela redução da carga nuclear ou introduzindo funções que modelem de forma combinada a repulsão do núcleo e do caroço eletrônico. Sé a consideração central dos métodos semi-empíricos, a aproximação Zero differential overlap (ZDO), na qual integrais de superposição que envolvem dois elétrons em átomos diferentes são desprezadas. O que traz as seguintes conseqüências:

- (1) A matriz de sobreposição S é reduzida a uma matriz unitária.

(2) Integrais de um elétron que envolvem três centros são consideradas zero.

(3) Todas as integrais de dois elétrons de três e quatro centros são desprezadas.

Para compensar estas aproximações, as integrais restantes são calculadas com o uso de parâmetros, e os seus valores são designados na base dos dados experimentais. Como exatamente estas integrais são desprezadas, e como a parametrização é feita, definem os vários métodos semi-empíricos.

O primeiro método a surgir, o **Método de Hückel** despreza completamente as integrais de 2-elétrons. A seguir veio o **Método de Hückel⁶ Estendido**, onde seu Hamiltoniano considera as interações eletrônicas e nucleares como uma média, podendo ser tratado como um modelo de elétrons independentes para os elétrons π da camada de valência. Não é um método autoconsistente, mas por sua formulação é útil no cálculo dos orbitais moleculares, não sendo utilizado para cálculos de otimização de geometria, posteriormente foram desenvolvidos os métodos: **CNDO** (*complet neglect of differential overlap*) sendo o método semi-empírico autoconsistente mais simples³. Utiliza como base um conjunto de orbitais atômicos do tipo Slater (38) e a aproximação ZDO. No **INDO** (*intermediate neglect of differential overlap*)⁷ os recobrimentos entre orbitais atômicos de mesmo átomo não são desprezados nas integrais de repulsão de 1 centro, mas ainda o são nas integrais de dois centros. Alguns métodos semi-empíricos utilizam parametrizações substituindo as integrais decorrentes dos cálculos por parâmetros, ao invés de desprezá-las. O **MINDO** (*modified intermediate neglect of differential overlap*)^{8,9} foi o primeiro destes métodos desenvolvidos. **NDDO** (*neglect of diatomic differential overlap*) onde apenas os recobrimentos entre orbitais atômicos em átomos diferentes são desprezados nos cálculos das integrais de Coulomb e de troca, e **MNDO** (*modified neglect of differential overlap*)^{10,11,12} que é parametrizado para a aproximação NDDO, e da mesma forma o AM1 (*Austin Method 1*)¹³ e o PM3 (*Parametric Method 3*)^{14,15}.

Os métodos AM1 e PM3 são os mais utilizados atualmente. O método PM3 é uma reparametrização do AM1, em que os parâmetros utilizados foram obtidos de um

número muito maior e representativo de dados experimentais. É vasta a literatura comparando a eficiência desses dois métodos^{16,17,18,19,20,21} porém não há consenso a respeito de qual deles reproduz melhor as geometrias e barreiras moleculares.

Uma Parametrização **INDO** foi preparada por Zerner e colaboradores (ZINDO/S) para a reprodução de espectro de absorção no UV-Visível associado as transições eletrônicas moleculares, utilizando a interação de configuração (CI) simples e duplas^{22,23}

Neste trabalho de tese utilizamos, de forma extensiva os métodos **AM1**, **PM3** e **PM5**. Sendo este último uma melhora sobre o **PM3**.

2.3 Quimiometria

O termo de Quimiometria foi introduzido em 1972 pelo Suíço Svate Wold e pelo Americano Bruce R Kowalski²⁴ cuja definição é:

A Quimiometria é uma disciplina da química que utiliza os métodos matemáticos e estatísticos, para desenhar ou selecionar os melhores procedimentos das medidas e experiências, e, para obter o máximo de informação química pela análise de dados químicos.

Uma das mais usadas aplicações dos métodos quimiométricos é na área das relações qualitativas e quantitativas da estrutura-atividade de compostos de interesse farmacológico.

2.3.1 Análise de Componentes Principais (PCA)

Análise de componentes Principais (PCA) é uma ferramenta muito útil que mapeia amostras através de scores e descritores individuais por loadings em um novo espaço vetorial definido pelas componentes principais (PC).

A idéia da PCA^{25,26} é aproximar a matriz **X** original, em um produto de duas matrizes menores –Score e Loading- de acordo com a seguinte transformação:

$$\mathbf{X} = \mathbf{TL}^T \quad (39)$$

$$\begin{array}{c} p \\ \boxed{\mathbf{X}} \\ n \end{array} = \begin{array}{c} d \\ \boxed{\phantom{\mathbf{L}^T}} \\ n \end{array} \begin{array}{c} p \\ \boxed{\mathbf{L}^T} \\ d \end{array} + \begin{array}{c} \boxed{\mathbf{E}} \end{array}$$

Fig 3: Sistema de decomposição da matriz X pelo método PCA.

Onde \mathbf{X} é a matriz original de dados. O arranjo dessa matriz vem dado da seguinte maneira:

As n linhas representam cada um dos compostos estudados no conjunto e as p colunas representam os descritores físico-químicos calculados para cada uma das moléculas; \mathbf{T} é a matriz dos scores ($n \times d$) e indica o número de componentes principais; \mathbf{L} é a matriz dos loadings ($n \times d$); e \mathbf{T} é a matriz transposta.

Em outras palavras, a projeção de \mathbf{X} sob um d -dimensional subespaço por meio da matriz de projeção \mathbf{L}^T obtém as coordenadas das moléculas no plano \mathbf{T} . As colunas em \mathbf{T} são os vetores scores e as linhas \mathbf{p} são chamadas de vetores Loading. Ambos os dois são ortogonais: $\mathbf{p}_i^T \mathbf{p}_j = 0$; $\mathbf{e} \cdot \mathbf{t}_i^T \mathbf{t}_j = 0$. para... $i \neq j$

As componentes principais são determinadas na base do critério de máxima variância residual no espaço n -dimensional. Cada subsequente componente principal descreve o máximo da variância, que não é modelada pela forma das componentes. De acordo com isso, a primeira componente principal é determinada olhando a direção da máxima variância residual dos dados. Depois de remover a primeira componente principal a segunda componente, completamente não correlacionada (ortogonal) com a primeira, contém a máxima variância restante do conjunto de dados

possível. O processo é repetido até que todas as PCs são geradas. Isto corresponde a um espaço n -dimensional (onde n é o número total de variáveis usadas).

Na interpretação geométrica, o conjunto de descritores de cada molécula define a sua posição geométrica no espaço n -dimensional, onde cada descritor corresponde a um eixo. As componentes principais, **PCs**, podem ser consideradas como as projeções da matriz original de dados **X**, em scores, **T**. Para isso a equação 1 pode ser convertida em:

$$\mathbf{T} = \mathbf{XL} \quad (40)$$

As novas coordenadas são combinações lineares das variáveis originais. Os elementos da primeira componente principal serão:

$$\begin{aligned} \mathbf{t}_{11} &= \mathbf{x}_{11}\mathbf{l}_{11} + \mathbf{x}_{12}\mathbf{l}_{21} + \dots + \mathbf{x}_{1p}\mathbf{l}_{p1} \\ \mathbf{t}_{21} &= \mathbf{x}_{21}\mathbf{l}_{11} + \mathbf{x}_{22}\mathbf{l}_{21} + \dots + \mathbf{x}_{2p}\mathbf{l}_{p1} \\ &\vdots \\ \mathbf{t}_{n1} &= \mathbf{x}_{n1}\mathbf{l}_{11} + \mathbf{x}_{n2}\mathbf{l}_{21} + \dots + \mathbf{x}_{np}\mathbf{l}_{p1} \end{aligned} \quad (41)$$

Os Loadings e os scores são calculados iterativamente usando um algoritmo baseado na diagonalização das Matrizes como é o caso do **SVD** (*decomposição de Valor Singular*); existem mais dois métodos usados, mas este é considerado um dos melhores.

A interpretação dos resultados da Análise de componentes principais é feita pela visualização dos Scores e Loadings. Normalmente os testes preliminares revelam que, duas ou três componentes principais são significativas, e as figuras das mesmas são suficientes.

O gráfico dos Scores apesar da projeção linear dos compostos mostra a parte principal da variância total dos dados, ele permite a identificação clarificando se elas são similares ou não. O gráfico dos Loadings decide a correlação e a importância das variáveis usadas. A correlação das variáveis vem dada pelo coseno do ângulo entre dois vetores-Loadings. Isto significa que um ângulo pequeno descreve uma alta correlação das variáveis e as variáveis sem correlação são ortogonais entre si. Se as variáveis estão altamente correlacionadas é suficiente usar uma delas. A relação do tamanho dos Loadings na componente principal é uma medida da importância da

variável para o modelo das Componentes Principais. Loadings na origem do sistema de coordenadas das componentes principais geralmente não são importantes.

A capacidade de discriminação das variáveis pode ser deduzida da direção dos Loadings. A proximidade dos compostos aos vetores Loadings, quando se sobrepõem os dos gráficos (Loadings e Scores), reflete a importância da variável para a construção do modelo PC.

2.3.2 Análise Hierárquica de Agrupamentos (HCA)

A base da segunda estratégia da análise com métodos não supervisionados é a análise de clusters (agrupamentos) onde é enfatizado o fato do agrupamento natural de amostras similares. Os resultados são apresentados na forma de *dendrogramas*, permitindo a visualização dos agrupamentos e a correlação entre as amostras.

Com este método os compostos (objetos) são agrupados de acordo com a similaridade das suas variáveis. Para decidir sobre a similaridade dos compostos (objetos) normalmente é aplicado um padrão de reconhecimento. Distâncias menores dentre compostos (objetos) indicam maior similaridade. De forma geral a distância aplicada entre os compostos é a chamada Minkowski ou de Métrica L_p :

$$d_{ij} = \left[\sum_{k=1}^K |x_{ik} - x_{jk}|^p \right]^{\frac{1}{p}} \quad (42)$$

Onde:

K é o número de variáveis e i, j são os índices dos compostos (objetos).

Na maioria dos casos d_{ij} é usado com o valor de $p=2$, sendo assim a distância Euclidiana usada. Junto com as medidas das distâncias também são feitas as de Similaridade, que é dada pela expressão:

$$S_{ij} = 1 - \frac{d_{ij}}{d_{ij}(\max)} \quad (43)$$

Onde $d_{ij}(\max)$ representa a distancia máxima dos compostos (objetos), encontrada nos dados. Compostos (objetos) completamente similares tem uma

similaridade $S_{ij}=1$. Para compostos que não apresentem similaridades S_{ij} ficará longe do valor de um.

Existem duas formas de realizar o agrupamento: *hierárquico* ou *não hierárquico* (os métodos explicados a seguir foram os utilizados no nosso trabalho), utilizamos a primeira: HCA aglomerativa, onde o processo de agrupamento começa com objetos individuais e os junta aos outros do grupo. De forma geral, mede-se a distância entre o cluster A-B que acaba de ser formado e outro C já formado, no nosso caso utilizamos a Conexão incremental²⁵:

$$d_{AB \Rightarrow C} = \sqrt{\frac{(n_A + n_C)d_{AC}^2}{n_A + n_B + n_C} + \frac{(n_B + n_C)d_{BC}^2}{n_A + n_B + n_C} - \frac{n_C d_{AB}^2}{n_A + n_B + n_C}} \quad (44)$$

Existem vários tipos de conexão, porém com a conexão incremental se obtém os melhores resultados, sendo desnecessária a utilização dos outros tipos de conexão.

2.3.3 K-ésimo vizinho mais próximo (KNN)

O KNN é um método simples de classificação de amostras de acordo ao conjunto de treinamento. No caso de se conhecerem os objetos que formam um determinado cluster, nós podemos utilizar os métodos de reconhecimento de padrões supervisionado.

O método do K-ésimo vizinho mais próximo foi introduzido por Fix e Hodges no ano 1951²⁷, ele é um método não paramétrico (não considera informação da distribuição da população) simples de classificação.

Ele Calcula a distância entre uma amostra desconhecida e todas as amostras do conjunto de treinamento distribuídas no espaço n -dimensional construído na análise de PCA. A distância mínima é selecionada e o objeto é atribuído a uma das classes. A atribuição a uma classe é feita por votos. Por comparação, um ou três K-ésimos vizinhos mais próximos de cada objeto são escolhidos para votar. Cada um dá um voto para sua classe. A classe que recebe mais votos (menor distância acumulada) ganha a amostra.

2.3.4 Soft Independent Modeling of Class Analogies (SIMCA)

Fora dos métodos de discriminação, a classificação de uma molécula (objeto) pode também ser feita pela descrição de classes individuais, o método SIMCA²⁸ encontra modelos de componentes principais para cada classe. Em termos de geometria, ele descreve um envelope ou uma “caixa da classe” ao redor da classe, e dessa maneira um composto desconhecido pode ser classificado de acordo com seu “fit” do modelo da classe particular.

O cálculo das componentes principais para cada classe é feito utilizando o mesmo algoritmo matemático do PCA.

Para cada classe q um modelo separado é construído e para somente um x :

$$\mathbf{x}_{ij}^q = \bar{\mathbf{x}}_j^q + \sum_{a=1}^{A_q} \mathbf{t}_{ia}^q \mathbf{l}_{ja}^q + \mathbf{e}_{ij}^q \quad (45)$$

onde, $\bar{\mathbf{x}}_j^q$ - média da variável j na classe q , A_q - Número de componentes principais importantes na classe q , \mathbf{t}_{ia}^q - score do objeto i na componente a na classe q , \mathbf{l}_{ja}^q - loading da variável j na componente principal a na classe q , \mathbf{e}_{ij}^q - erro residual do objeto i e a variável j .

Poder de Modelagem: A variância residual da variável j da classe q é utilizada para estimar o poder de modelagem de uma variável particular se for relacionada à chamada variância média:

Variância residual de variável j :

$$s_j^2(\mathbf{error}) = \sum_{i=1}^n \frac{\mathbf{e}_{ij}^2}{\mathbf{n} - A_q - 1} \quad (46)$$

Variância média de variável j :

$$s_j^2(\mathbf{x}) = \sum_{i=1}^n \frac{(\mathbf{x}_{ij} - \bar{\mathbf{x}})^2}{\mathbf{n} - 1} \quad (47)$$

A comparação das duas variáveis indica, a medida da razão do sinal de ruído destas variáveis. O poder de modelagem para a variável j , R_j é:

$$\mathbf{R}_j = 1 - \frac{s_j(\text{error})}{s_j(\mathbf{x})} \quad (48)$$

Se o poder de modelagem aproxima-se do valor de 1 significa que a variável será altamente relevante, porque a razão entre o valor residual para a variável é pequena comparada com o seu da variância média.

A classificação de um objeto desconhecido com dados do vetor \mathbf{x}_u (dimensão $1 \times p$) feita numa classe particular q é feita pela regressão do modelo dessa classe.

Multiplicando os dados do vetor pela matriz loading \mathbf{L} ($p \times A_q$) obtemos um novo vetor score \mathbf{t} ($1 \times p$). Com o vetor score os resíduos são calculados e usados na decisão de classificação:

$$\begin{aligned} \hat{\mathbf{t}} &= \mathbf{x}_u \mathbf{L} \\ \mathbf{e} &= \mathbf{x}_u - \hat{\mathbf{t}} \mathbf{L}^T \end{aligned} \quad (49)$$

A variância residual do objeto u :

$$s_u^2 = \sum_{j=1}^p \frac{e_{uj}^2}{p - A_q} \quad (50)$$

O objeto pertencerá à classe q , se o valor das variâncias s_u^2 e s_0^2 são similares na ordem de magnitude, caso contrario, o objeto não formará parte da classe q .

2.4 Redes Neurais

Aplicações das redes neurais artificiais são caracterizadas pela analogia com o neurônio biológico. A alta eficiência do sistema biológico é uma consequência das 10^4 interligações de não menos de 10 bilhões de células no cérebro.

Com base nessa idéia, e na tentativa de simular o neurônio biológico, foi construído o artificial. No neurônio biológico (fig. 4) as entradas dos pulsos elétricos, vêm através dos dendritos no processo de sinapse. A informação é transformada e transferida para o neurônio seguinte através do axônio.

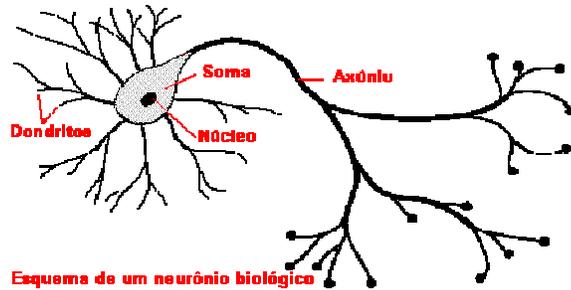


Fig.4: Esquema de um neurônio biológico.

Essa transferência de pulso é simulada no neurônio artificial pela multiplicação desse sinal de entrada, x , com o peso da sinapse, w , para obter o sinal de saída y (fig.5).

Considerando a operação de um neurônio j : Ele recebe de outros n neurônios os pulsos de entrada, x_i , interligados pelos pesos, w , das sinapses e passa o resultado depois de uma transformação como o sinal de saída y_i .

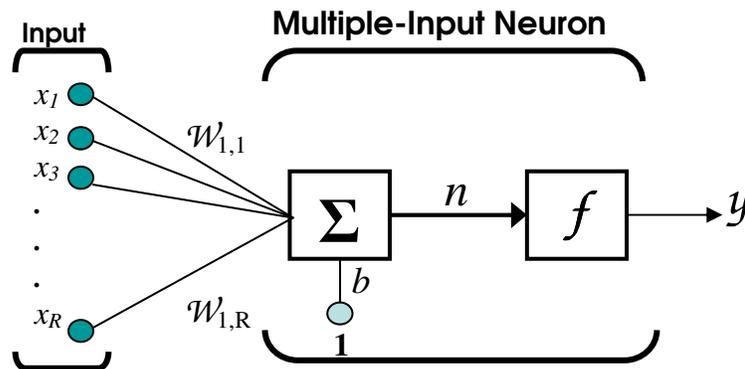


Fig.5: Esquema de um neurônio artificial.

Normalmente a junção dos pulsos de entrada é feita pela somatória $[\Sigma]$ (fig cinco). Depois de feita a junção a transformação é feita por uma função de transferência f , obtendo-se o pulso y de saída.

Ao estruturar uma arquitetura da rede neural existem diferentes possibilidades, nós podemos ter m camadas contendo n neurônios. No caso mais simples, a camada de entrada está interligada à segunda camada, que representa simultaneamente a camada de saída (rede de duas camadas). Entre as camadas de entrada e saída podem existir outras camadas que recebem o nome de camadas ocultas (*hidden layers*).

A rede neural mais simples é a do tipo perceptron. Ela foi introduzida por E. Rosenblatt (1950) com o objetivo de reconhecer padrões óticos, um modelo muito simples da retina dos olhos. O processo de aprendizado desta rede corresponde ao aprendizado supervisionado de acordo com as regras de aprendizado de associação, a rede perceptron pode ser comparada com uma rede de aprendizado linear, e como foi demonstrado por Minsky e Papert (1969), alguns problemas não podiam ser resolvidos usando a rede perceptron simples. A rede Neural perceptron tem sido aplicada com bastante sucesso nas investigações de SAR como uma ferramenta de classificação.

Atualmente são usadas perceptron de múltiplas camadas em conexão com o algoritmo de backpropagation. 90% das aplicações em química analítica são baseadas no algoritmo backpropagation.

Em nosso estudo nós utilizamos a rede neural do tipo perceptron com três camadas, de aprendizado supervisionado e que utiliza o algoritmo backpropagation (figura 6), implementados no pacote computacional PSDD (*perceptron simulator for drug design*), que foi elaborado para o desenvolvimento de drogas e reconhecimento de padrões de moléculas com atividade biológica²⁹.

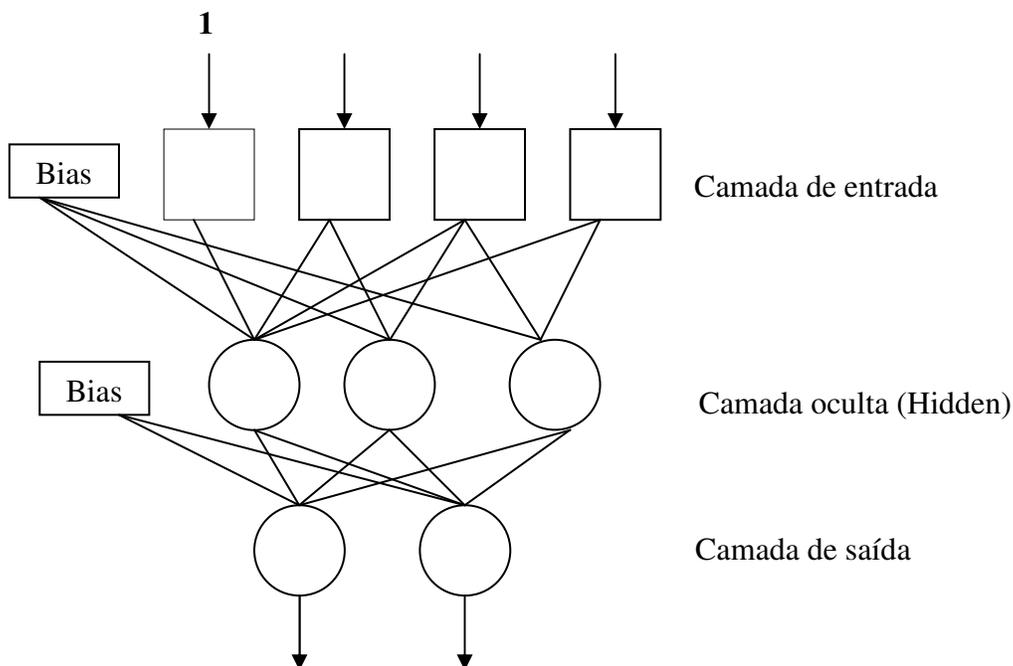


Fig.6: Esquema da rede neural perceptron de três camadas.

Cada círculo refere-se a um neurônio, exemplificado na figura 5. As forças das ligações entre os neurônios (linhas que interligam os círculos) são dadas pelo peso “w”. Os quadrados representam à primeira camada (de entrada) que no caso de QSAR (*quantitative structure activity relationship*) ou SAR (*structure activity relationship*) são os descritores associados à amostra.

Para esta arquitetura os dados de entrada são escalados para valores entre 0 e 1, pela equação:

$$\hat{x} = \frac{x_i - x_{\min} + 0.1}{x_{\max} + x_{\min} + 0.1} \quad (51)$$

onde x_{\min} e x_{\max} são os valores mínimos e máximos dos dados. Ao fazer o escalamento dos dados de entrada os valores mínimos serão de 0.1 e não de 0 para evitar a anulação de algum dado importante ao serem multiplicados pelo fator peso.

A função de ativação utilizada é do tipo sigmóide, os dados de saída de qualquer camada terão a forma dada pela equação:

$$O_j = \frac{1}{1 + e^{-\alpha y_i}} \equiv f(y_j) \quad (52)$$

onde

$$y_i = (\sum W_{ij}x_i) - \theta_j \quad (53)$$

x_i é o valor do neurônio na camada m-1; W_{ij} é o elemento da matriz peso da conexão entre os neurônios i e j; θ_j é o valor característico do neurônio j (é o parâmetro que expressa a não linearidade do neurônio). Os dados de entrada e de saída são representados como vetores. A dinâmica de inserção dos valores dos dados de entrada após percorrer as camadas das redes neurais, fornece o dado de saída e é denominado padrão de treinamento (t_{ij} , valor do neurônio da ultima camada). O treinamento é realizado de acordo com as equações abaixo:

$$W_{ij} = -d_j x_i \varepsilon \quad (54)$$

$$d_j^{(3)} = (O_j - t_j) f'(y_j) \quad (55)$$

$$d_j = (W_{ji} \cdot d_j') f'(y_j) \quad (56)$$

ε (parâmetro de ajuste da rede denominado EPSIRON) é o parâmetro que determina o acréscimo para a troca na correção entre os ciclos recursivos. Na equação 55 o índice sobrescrito indica que ela só será utilizada na ultima camada e a equação 56 é utilizada nas outras camadas. Depois de ocorrido o primeiro ciclo os novos valores de W'_{j1} e d'_1 de uma camada m passam a serem os antigos valores de W_{j1} e d_1 da camada m+1. A função $f''(y_j)$ (novo valor do neurônio j) é dada pela equação abaixo.

$$f''(y_i) = f(y_i)[1 - f(y_i)]\alpha \quad (57)$$

Aqui ambos os índices ε da equação 54 e α da equação 57 são ajustados independentemente da camada considerada. Depois de vários ciclos recursivos a minimização do erro, (equação 58), é feita até que exista uma pequena diferença entre a saída gerada pela rede com o padrão de saída desejado inicialmente (supervisionado).

$$E = \sum (O - t_j)^2 \quad (58)$$

Se a convergência é atingida a rede neural é chamada de treinada e tem a capacidade de classificar os inputs em M grupos. Assim a rede neural treinada é o ponto inicial para a classificação atividade biológica de novos compostos.

A figura seguinte descreve as características do algoritmo backpropagation e a correção feita para a minimização do erro.

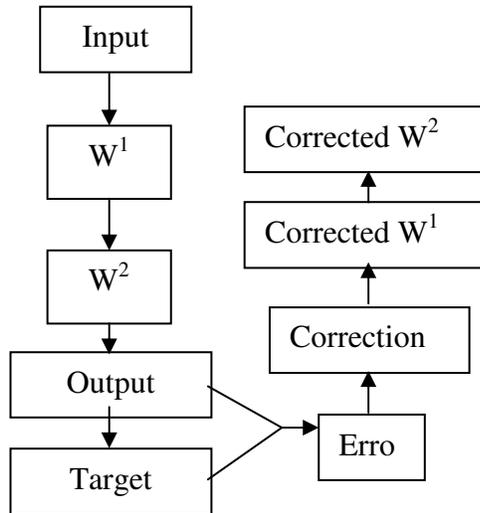


Fig.7: Algoritmo Backpropagation, correção feita por ciclos recursivos.

O uso das redes neurais na análise de SAR pode ser exemplificado no trabalho feito por Ayoma, Suzuki e Ichikawa ³⁰, eles utilizam um conjunto de 38 amostras e sete descritores. Separando 25 amostras para o conjunto de treinamento e 13 amostras para o conjunto de predição, ambos os grupos contém amostras do tipo endo e tipo exo. A saída para amostras do tipo exo é 1 0 0 1 e do tipo endo é 0 1 1 0. Os resultados para o grupo de treinamento são apresentados em alguns caso com valores absolutos de 1 ou 0 1, mas também existem valores próximos a 1 ou 0 como 0,99 ou 0,08 (respectivamente), o que indica um treinamento próximo aos dados experimentais. Depois de otimizar o grupo de treinamento foi feita a predição. O acerto foi estatisticamente muito bom 93% (12/13), o que mostra o poder das redes neurais.

Referencias

- ¹ N. C. Cohen, *Guidebook on Molecular Modeling in Drug Design* (Academic Press, San Diego, 1996).
- ² J. P. Lowe, *Quantum Chemistry* (Academic Press, San Diego, 1978).
- ³ J. A. Pople and D. L. Beveridge, *Approximate Molecular Orbital Theory* (McGraw-Hill, New York, 1970).
- ⁴ W. Kauzmann, *Quantum Chemistry* (Academic Press. Inc.-Publishers, New York, 1957).
- ⁵ J. Sadley, *Semi-Empirical methods of Quantum Chemistry* (Wiley and Sons.,1985); M.C. Zerner, *Rev Compt. Chem.* **2**, 313 (1991).
- ⁶ R. Hoffmann, *J Chem. Phys.* **39**, 1397 (1963).
- ⁷ G. A. Segal, *Semi-empirical Methods of Electronic Structure Theory* (Plenum, New York, 1977).
- ⁸ R. C. Bingham, M. J. S. Dewar *et al.*, *J. Am. Chem. Soc.* **97**, 1285 (1975).
- ⁹ M. J. S. Dewar *et al.*, *J. Am. Chem. Soc.* **97**, 1311 (1975).
- ¹⁰ M. J. S. Dewar *et al.*, *Science* **187**, 1037 (1975).
- ¹¹ M. J. S. Dewar and W. Theel, *J. Am. Chem. Soc.* **99**, 4899 (1977).
- ¹² M. J. S. Dewar and H. S. Rzepa, *J. Am. Chem. Soc.* **100**, 777 (1978).
- ¹³ M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.* **107**, 3902 (1985).
- ¹⁴ J. J. P. Stewart, *J. Comput. Chem.* **10**, 209 (1989).
- ¹⁵ J. J. P. Stewart, *J. Comput. Chem.* **11**, 543 (1990).
- ¹⁶ M. C. Zerner, *Reviews in Computational Chemistry II*, (Lipkowitz, KB, and Boyd, DB, Eds., VCH Publishers, 1991).
- ¹⁷ P. Scano and C. Thompson, *J. Comp. Chem.* **12**, 172 (1991).
- ¹⁸ Z. G. Zôos and D. S. Galvão, *J. Phys. Chem.* **98**, 1029 (1994), and references therein.
- ¹⁹ A. Koll, M. Rospenk, E. Jagodzinska and T. Dziembowska, *J. of Molecular Structure* **552**, 193 (2000).
- ²⁰ L. Gorb, A. Korkin, J. Leszczynski, A. Varnek, F. Mark and K. Schaffner, *J. Mol. Struct. (THEOCHEM)* **425**, 137 (1998).
- ²¹ M. A. Palafox and F. J. Melendez, *J. Mol. Struct. (THEOCHEM)* **459**, 239 (1999).
- ²² W. D. Edwards and M. C. Zerner, *Theor. Chem. Acta* **72**, 347 (1987).
- ²³ R. McWeeny, S. Wilson and G.H.F. Diercksen, *Methods in Computational Molecular Physics* (Plenum Press, New York, 1992).
- ²⁴ M. A. Sharaf, D. L. Illman and B. R. Kowalski, *Chemometrics Chemical Analysis Series Vol 82* (Wiley, New York, 1986).
- ²⁵ K. Varmuza, *Pattern recognition in chemistry* (Springer-Verlag, Berlin, 1980).
- ²⁶ M. Otto, *Chemometrics: Statistics and Computer Application in Analytical Chemistry* (Wiley, New York, 1999).
- ²⁷ D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte and L. Kaufman, *Chemometrics: a textbook 2* (Elsevier, 1998).
- ²⁸ J. H. Friedman, *J. Am. Stat. Assoc.* **84**, 165 (1989).
- ²⁹ H. Ichikawa, *PSDD: Perceptron-type Neural Network Simulator, QCPE 615* (Indiana, USA).
- ³⁰ T. Aoyama, Y. Suzuki and H. Ichikawa, *Chem. Pharm. Bull* **37** (9), 2558 (1989).

Capítulo 3

Propriedades Físico-químicas

3.1. Introdução

A resposta biológica que determinado composto induz em um organismo vivo depende de diferentes fatores que estão diretamente correlacionados às suas propriedades físico-químicas^{1,2}. Podemos agrupar esses fatores em três classes principais de propriedades físico-químicas: Hidrofóbica, estérea e eletrônica³. A hidrofobicidade está relacionada ao mecanismo de transporte do composto no meio biológico, que é governado principalmente pela partição entre fases lipídicas e aquosas. As propriedades estéreas estão relacionadas aos fatores tais como o tamanho e a forma do composto, o que pode ser determinante para o acoplamento do composto com seu receptor ou enzima e, por fim, temos as propriedades eletrônicas que estão relacionadas à susceptibilidade de determinada ligação ao ataque metabólico eletrofílico ou nucleofílico. Nos estudos de estrutura-atividade procura-se correlacionar parâmetros originados dessas propriedades à determinada resposta biológica de interesse.

3.2. Parâmetros Hidrofóbicos:

3.2.1 Coeficiente de Partição

Para tentar descrever o comportamento de um composto no meio celular criou-se um parâmetro chamado coeficiente de partição, P , que descreve a razão das concentrações do composto em dois meios líquidos imiscíveis em equilíbrio.

Os meios imiscíveis mais comumente utilizados são a água e o octanol. O octanol tem uma cabeça polar e uma cadeia alquila hidrofóbica que faz com que ele seja um bom modelo para a cadeia lipídica da célula. Assim a mistura octanol/água constitui um bom modelo para a barreira celular. Geralmente se utiliza o logaritmo do coeficiente de partição ($\log P$) como parâmetro hidrofóbico.

3.3. Parâmetros Estereoquímicos:

O volume e a forma das moléculas e dos grupos substituintes são características importantes a serem observadas no controle da atividade biológica.

Muitas vezes a droga e o receptor têm um formato complementar (conceito chave-fechadura), e qualquer modificação que altere a conformação da droga impedindo que haja um encaixe 'ótimo' com o receptor pode levar a um enfraquecimento das interações e uma conseqüente perda ou diminuição de atividade. Podemos citar como parâmetros relacionados às propriedades estéreas do composto o volume molecular, a área superficial e o raio de van der Waals.

3.4. Parâmetros Eletrônicos

A distribuição eletrônica sobre uma molécula e a facilidade com que essa distribuição pode ser modificada está diretamente relacionada às forças de interação intermoleculares. Essas interações exercem uma influência direta na resposta biológica que um composto é capaz de induzir. Com o objetivo de modelar essa

resposta biológica, vários parâmetros eletrônicos foram e vêm sendo propostos nos diversos estudos de estrutura-atividade. Existem os parâmetros clássicos, tais como o pKa, o momento de dipolo, a ligação de hidrogênio, a refratividade molar, etc., e os parâmetros obtidos dos cálculos de Química Quântica. Nos deteremos aos parâmetros químico-quânticos.

3.4.1 Energias do HOMO (*Highest Occupied Molecular Orbital*), e LUMO (*Lowest Unoccupied Molecular Orbital*)

A energia do **HOMO** (E_{HOMO}) representa a energia necessária para ‘arrancar’ um elétron da última camada de valência da molécula (potencial de ionização), portanto está relacionada à facilidade com que o elétron pode ser ‘doador’ pela molécula. Já a energia do **LUMO** (E_{LUMO}) está relacionada à facilidade com que a molécula pode ‘aceitar’ um elétron. Esses parâmetros geralmente estão correlacionados com a atividade biológica quando a interação dominante é a transferência de carga entre as moléculas.

Quando ocorre uma ‘aproximação’ dos orbitais moleculares de dois compostos, a interação responsável pelo maior ganho de energia na formação do novo composto é a dos orbitais ocupados de um dos compostos com os orbitais desocupados do outro. Nesse aspecto, as interações entre o **HOMO** e o **LUMO** são, comumente, as mais efetivas; isso porque quanto mais próximos em energia estiverem os orbitais interagentes maior será a separação dos níveis de energia do complexo formado pelos dois compostos.

3.4.2 Contribuição dos Orbitais Atômicos (C_r)

A C_r é definida como sendo a soma dos quadrados dos coeficientes C_{rj} do r -ésimo orbital atômico do j -ésimo orbital molecular.

3.4.3 Diferença de Energia HOMO LUMO

A diferença de energia entre o **HOMO** e o **LUMO** nos dá a energia necessária para promover o elétron do estado fundamental (HOMO) para o estado excitado

(LUMO), o que geralmente pode ser correlacionado às absorções óticas do composto.

3.5. Metodología dos Índices Electrônicos:

A MIE foi proposta inicialmente com o objetivo principal de distinguir de forma qualitativa o problema farmacológico de discriminação entre compostos ativos ou inativos dos PAHs com relação à atividade carcinogênica⁴. A MIE está baseada nos conceitos da densidade de estados local e total, e em valores críticos para a diferença de energia eletrônica entre os orbitais moleculares de fronteira.

A densidade de estados (**DOS**) é definida como o número de estados eletrônicos por unidade de energia. No caso molecular a formação dos estados (Orbitais Moleculares) é fruto da combinação dos átomos da molécula, com a ocupação máxima de dois elétrons em cada estado, obedecendo ao princípio de exclusão de Pauli.

Relacionado a este conceito da **DOS** está a densidade local de estados (**LDOS**) que é uma medida da contribuição dos átomos de uma região molecular específica para a formação dos estados acessíveis (HOMO, e seus vizinhos mais próximos ou LUMO e seus vizinhos mais próximos).

A **LDOS** é introduzida para descrever a distribuição espacial de estados específicos sobre determinadas regiões moleculares.

Isso nos permite investigar regiões moleculares específicas separadamente, observando sua importância na formação de um estado.

No caso dos PAHs, descobriu-se que a análise da densidade local de estados sobre o anel de maior ordem de ligação associada à diferença de energia entre os orbitais HOMO e HOMO-1 estava diretamente ligada à atividade carcinogênica desses compostos com essa metodologia foi possível avaliar a resposta biológica dos PAHs com um acerto maior a 83%.

Para hamiltonianos baseados na aproximação da combinação linear dos orbitais atômicos (**LCAO**) (que são os que utilizamos nos métodos semi-empíricos **AM1**,

PM3 e **PM5**), os cálculos da **LDOS** envolvem a contribuição de cada átomo para um nível eletrônico específico medida pelo coeficiente do orbital molecular elevado ao quadrado. Para cálculos envolvendo vários orbitais ou regiões moleculares a **LDOS** é obtida somando-se as contribuições dos orbitais atômicos selecionados:

$$\mathbf{LDOS}(E_i) = 2 \sum_{m=n_i}^{n_f} |c_{mi}|^2 \quad (1)$$

Onde, E_i a energia do orbital molecular, n_i e n_f são os sítios atômicos iniciais e finais respectivamente, C_{mi} é a contribuição do i -ésimo orbital atômico no m -ésimo sítio, O número 2 vem do princípio de exclusão de Pauli (máximo de 2 elétrons por nível eletrônico).

A equação 1 permite uma comparação direta da **DOS** e **LDOS** calculadas a partir de qualquer método **LCAO**.

A MIE utiliza um primeiro parâmetro de investigação, denominado η , ele está relacionado com as diferenças das contribuições **LDOS** de dois níveis eletrônicos, geralmente os níveis de fronteira HOMO (*Highest Occupied Molecular Orbital*), **HOMO-1** ou **LUMO** (*Lowest Unoccupied Molecular Orbital*), **LUMO+1**:

$$\eta = 2 \sum_{S=n_i}^{n_f} (|c_{snivel1}|^2 - |c_{snivel2}|^2) \quad (2)$$

A análise **LDOS** permite que, a identificação dos níveis moleculares relevantes, correlacionados com a atividade biológica, seja através de uma busca sistemática sobre várias regiões moleculares, ou de regiões importantes do ponto de vista químico ou biológico

O segundo parâmetro da MIE, denominado Δ , é definido como a diferença de energia entre os níveis moleculares da equação 2:

$$\Delta = E_{nível1} - E_{nível2} \quad (3)$$

Para os dois últimos orbitais moleculares ocupados o parâmetro Δ é dado por:

$$\Delta H = E_{HOMO} - E_{HOMO-1} \quad (4)$$

E para os dois primeiros desocupados:

$$HL = E_{LUMO+1} - E_{LUMO} \quad (5)$$

Exemplos típicos de **DOS** e **LDOS** estão indicados nas figuras 1 e 2. Os valores de η e Δ podem ser obtidos diretamente dos gráficos das figuras.

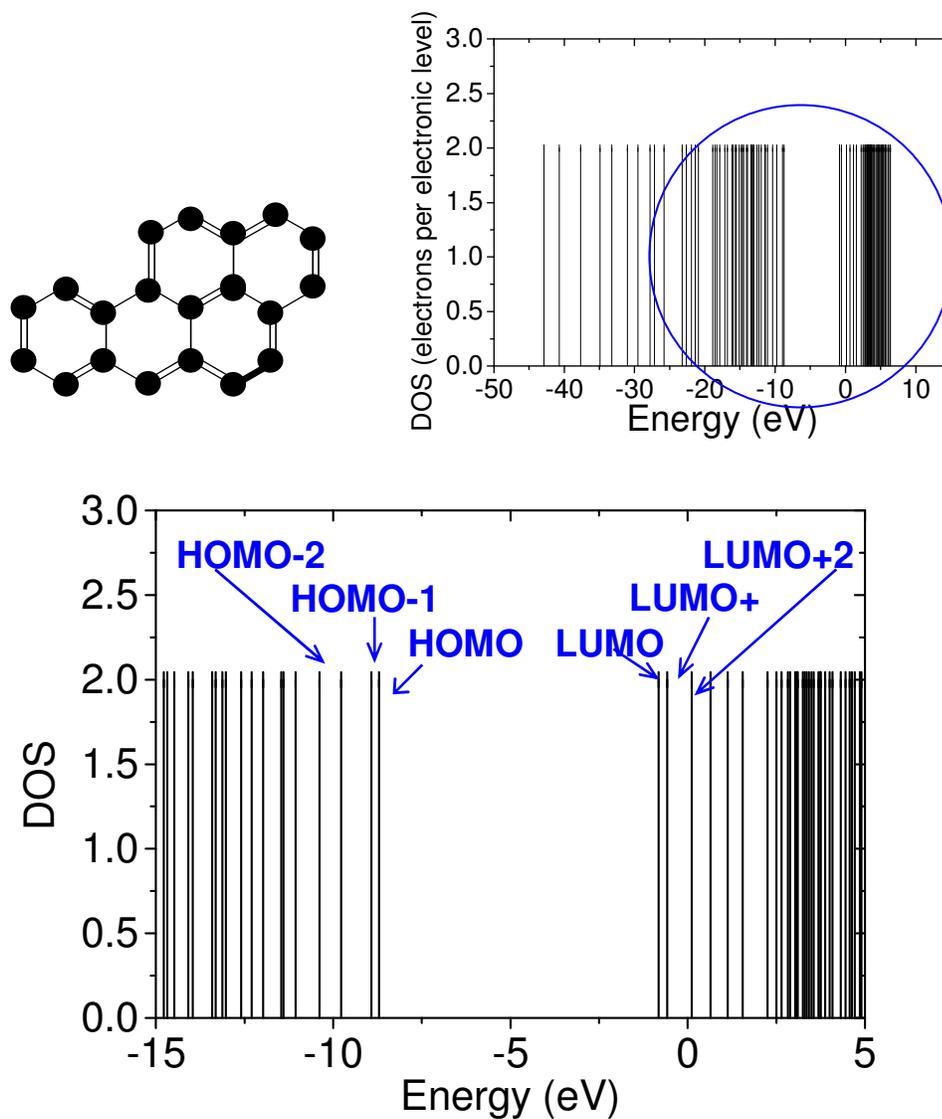


Fig1 : Gráfico da densidade total de estados DOS.

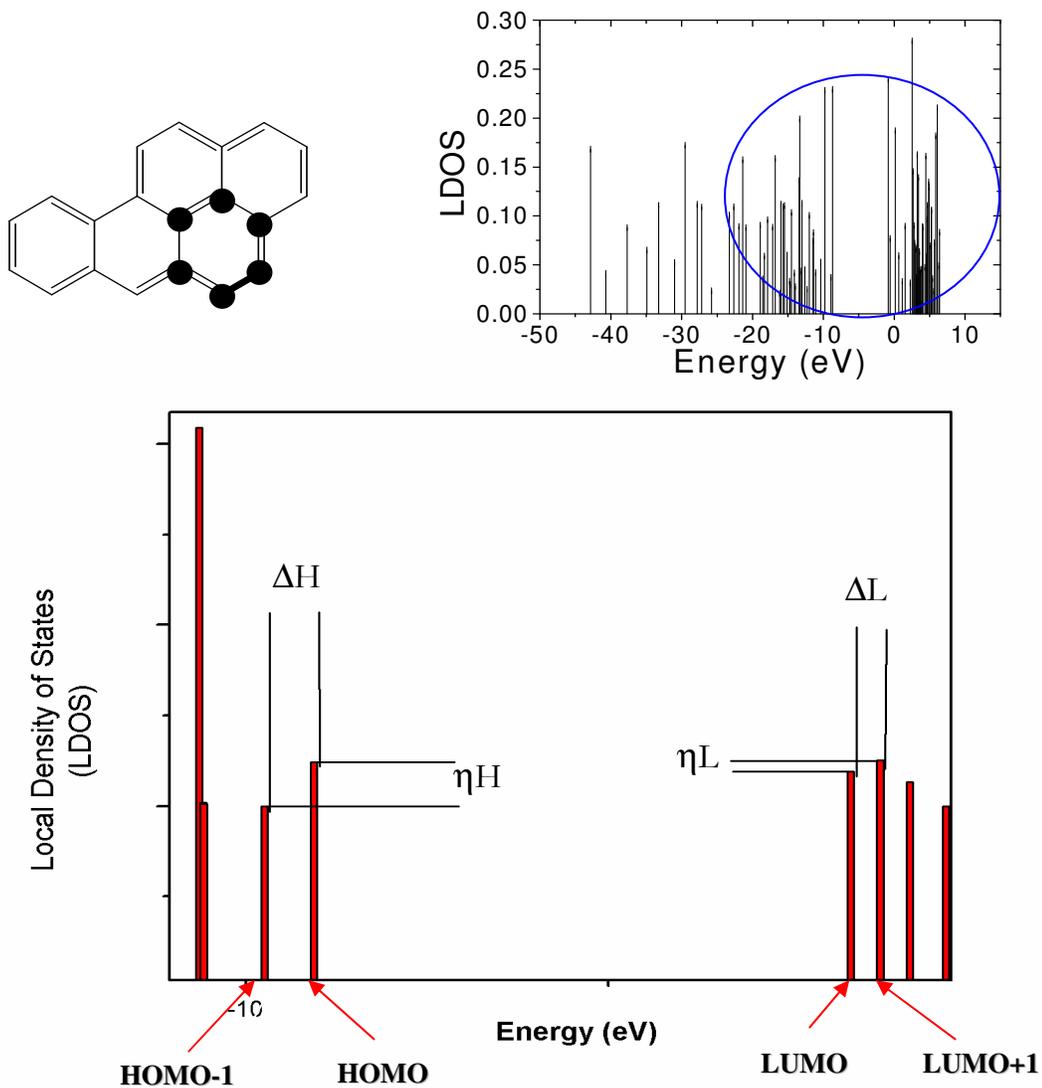


Fig 2 : Gráfico da densidade local de estados para uma região molecular específica.

Calculo dos parâmetros da MIE

Estes dois parâmetros nos permitem obter regras de seleção da atividade dos compostos. As regras estão relacionadas com valores críticos de Δ_c e η_c , que são obtidos como sendo valores limites para classificar e distinguir compostos ativos dos

inativos observando os dados experimentais da atividade biológica para cada classe de compostos.

Por exemplo uma regra típica é do tipo:

Regra: Se $\eta > \eta_c$ e $\Delta < \Delta_c$, a molécula será ativa; caso contrário será inativa.

As regras podem ser melhor sumarizadas com o emprego de tabelas Booleanas, como a mostrada abaixo.

Tabela 1 – Tabela Booleana para classificação da atividade dos compostos

(η)	(Δ)	Atividade Biológica (Predição)
+	+	Ativa
+	-	Inativa
-	+	Inativa
-	-	Inativa

Onde + e – indicam se a condição para o parâmetro especificado na regra é satisfeita (+) ou não (-).

Desta forma, a MIE utiliza apenas 2 parâmetros dentro de regras simplificadas para separar os compostos quanto a sua atividade biológica.

Com a utilização somente desses dois parâmetros foi possível classificar (ativos / inativos) com sucesso várias classes de compostos orgânicos, tais como antibióticos, antitumorais, hormônios anticoncepcionais, esteroides, inibidores de integrase do HIV-1 sempre com uma taxa de acerto da ordem de 80-90% .^{5,6,7,8,9,10}

Uma das motivações do presente estudo foi verificar e testar se na metodologia de índices eletrônicos, baseada nos parâmetros Δ e η , existe uma dependência explícita entre a qualidade dos resultados obtidos e o Hamiltoniano eletrônico utilizado.

Referencias

- ¹ N. C. Cohen, *Guidebook on Molecular Modeling in Drug Design* (Academic Press, San Diego, California, 1996).
- ² A. Korolkovas, *Essential of Molecular Pharmacology* (Wiley & Sons, New York, 1970).
- ³ W. Karcher and J. Devillers, *Practical Applications of Quantitative Structure-Activity Relationship (QSAR) in Environmental Chemistry and Toxicology* (Kluwer Academic Publishers, Holland, 1990).
- ⁴ P. M. V. B. Barone, A. Camilo Jr., D. S. Galvão, *Phys. Rev. Lett.* , **77**, 1186 (1996).
- ⁵ M. Cyrillo and D. S. Galvão, *J. Mol. Struct. (THEOCHEM)* **464** 267 (1999).
- ⁶ L. L. E. Santo and D. S. Galvão, *J. Mol. Struct. (THEOCHEM)*, **464** 273 (1999).
- ⁷ R. S. Braga, R. Vendrame and D. S. Galvão, *J. Chem Inf. Comput. Sci.* **40** (2000) 1377.
- ⁸ R. Vendrame, V.R. Coluci, R. S. Braga e D. S. Galvão - submetido
- ⁹ S. F. Braga e D. S. Galvão - submetido
- ¹⁰ S. F. Braga e D. S. Galvão - submetido.

Capítulo 4

Estudo Estrutura-Atividade dos Hidrocarbonetos Policíclicos Aromáticos

4.1. Introdução

Os hidrocarbonetos policíclicos aromáticos-PAHs, mais simplesmente conhecidos como poliarenos, constituem uma extraordinária classe em número e diversidade de compostos orgânicos. As maiores fontes de PAHs são o óleo cru, carvão e óleo de argila. A energia produzida por essas fontes é a principal fonte de energia dos países industrializados. Os PAHs representam uma grande classe de poluentes ambientais, presentes na exaustão dos automóveis, na fumaça de cigarros e inclusive como contaminantes da comida e água. Eles podem ser considerados um dos principais agentes cancerígenos, pelo fato de se encontrarem amplamente distribuídos no meio ambiente. O corpo humano interage com esses compostos via inalação, ingestão ou absorção cutânea. O trabalho de Lietch (1922)¹ feito no Research Institute of the Cancer Hospital em Londres, foi o estímulo para o começo, nesse período onde os hidrocarbonetos receberam substancial atenção, no estudo sobre estes compostos carcinogênicos. Atualmente com todos os estudos feitos, eles constituem uma das classes mais estudadas².

4.1.1 Hidrocarbonetos Carcinogênicos

Em 1961 Hieger³ trabalhou na indução experimental do câncer de pele feita em ratos, foi o primeiro sistema de teste essencial que, disponibilizou a identificação dos primeiros agentes químicos carcinogênicos. No método utilizado, os ratos eram pintados duas vezes por semana com um pincel de pelo de camelo empapado de uma solução, por exemplo, de benzopireno (0.75% em benzeno: parafina líquida, 9:1). O período latente (i.e., tempo de aparição da primeira papiloma), esteve entre 13 a 54 semanas.

Apesar de ser um método bastante laborioso e longo, ele permitiu um consenso para obter o grau da potência carcinogênica de vários hidrocarbonetos policíclicos aromáticos, que estavam disponíveis a partir dos anos 1930, em termos da aparição de tumores e média de período latente seguindo a aplicação de doses padronizadas. Este trabalho estimulou a idéia da relação estrutura-atividade dos PAHs^{4,5}.

Paralelamente, estudos teóricos foram propostos para conseguir a discriminação entre compostos ativos/inativos, poder também prever a potência carcinogênica dos compostos, e entender o mecanismo eletrônico da ativação metabólica. Pullman and Pullman⁶, a partir do estudo da mecânica quântica da estrutura eletrônica dos PAHs, considerando os métodos de orbitais moleculares, de ligação de Valencia e obtendo índices eletrônicos das moléculas isoladas e polarizadas, invariavelmente mostraram que a maioria dos PAHs contem duas regiões de particular importância para o seu comportamento químico (algumas só apresentam uma região). Essas regiões chamadas de **K** e **L** (Fig (1)), na aproximação da molécula isolada, contem a maior ordem de mobilidade e os carbonos que contem os maiores valores de valencia livre, respectivamente. E na teoria de localização são as regiões formadas pela ligação que

contem o menor B.L.E.¹ e os carbonos que tem o menor C.L.E.² e ao mesmo tempo o menor PLE³.

Com esse resultado eles chegaram as seguintes propostas fundamentais:

1. A atividade carcinogênica nos PAHs é determinada pela existência de uma região **K** ativa onde o índice complexo B.L.E. +C.L.E._{min} e igual o menor do que 3,31β.
2. Mas, se a molécula contem também uma região **L**, uma condição adicional e requerida, e é que esta região deve ser fracamente inativa e o índice complexo P.L.E. + C.L.E._{min} deve ser igual ou maior que 5,66β.

Esses foram os limites determinados para caracterizar a atividade ou inatividade dos PAHs. Os termos de regiões “Ativas” e “Inativas” são equivalentes a reativas ou não reativas, considerando que o passo essencial na carcinogênese, consiste na reação entre a molécula carcinogênica e o receptor celular.

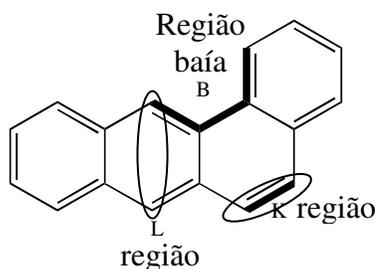


Fig 1: Regiões K, L, e bay (B) para os hidrocarbonetos policíclicos aromáticos.

Nos anos 70 e 80 foi determinado que uma considerável quantidade de PAHs carcinogênicos possuem uma região que foi chamada região baía (*bay-regions*), o que levou as teorias das regiões baía⁷.

¹ **B.L.E.:** A energia de localização de ligação é definida como a quantidade de energia necessária para perturbar a estrutura eletrônica de uma molécula conjugada, sendo a diferença de energia de ressonância entre a molécula inicial o fragmento conjugado que sobrou depois de ter eliminado a ligação em questão

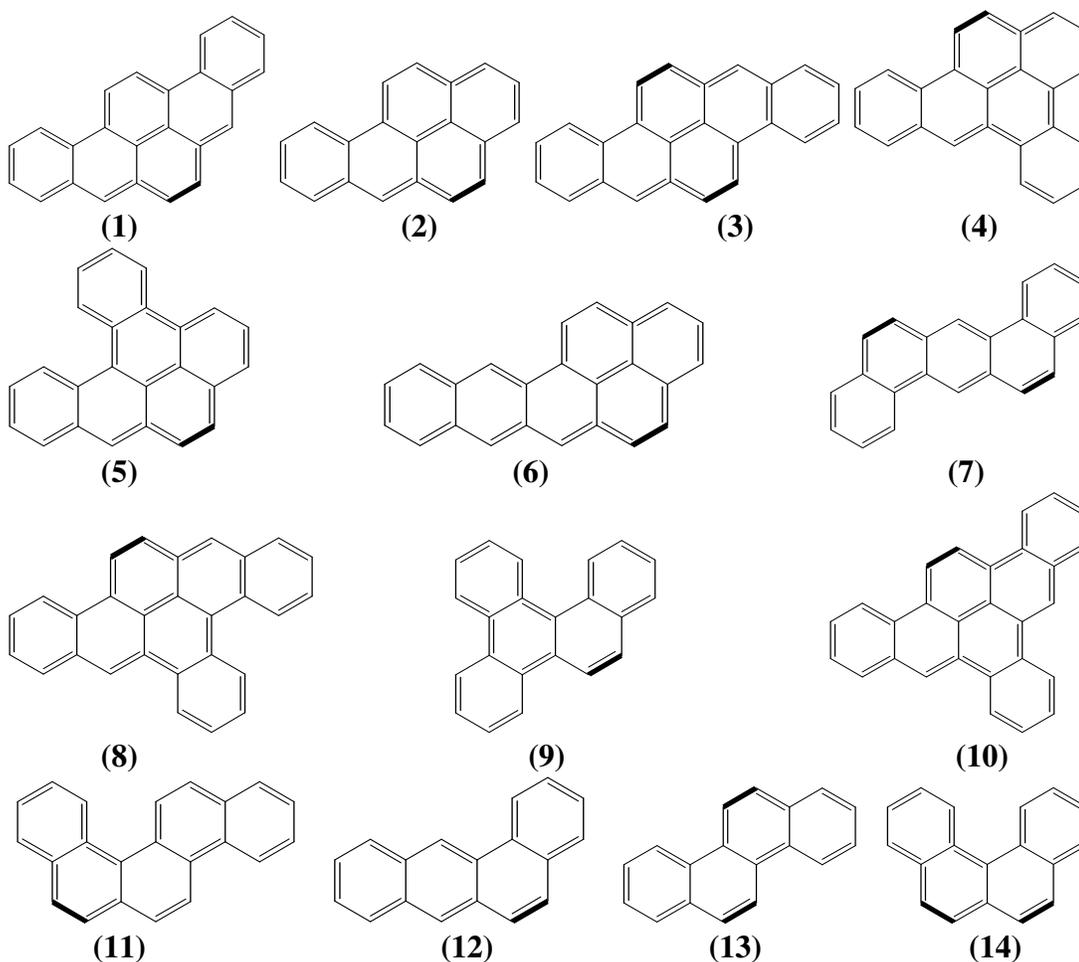
² **C.L.E.:** O nome da energia de localização do carbono tem sido dado a diferença da energia de ressonância entre o sistema inicial não perturbado e o hidrocarboneto polarizado, dessa maneira os eletros π estão localizados em um carbono e conseqüentemente não participam da conjugação.

³ **P.L.E.:** Esta quantidade de energia é a necessária para perturbar a estrutura eletrônica da molécula conjugada para conseguir localizar de forma simultânea dois elétrons nas posições *para* um do outro

Essas teorias baseadas em cálculos de química quântica, envolvendo energias de localização para as regiões K, L e B, e em índices eletrônicos e mais recentemente usando os métodos de análise estatística, redes neurais e inteligência artificial têm obtido um sucesso parcial. Alguns desses trabalhos descrevem bem a atividade de alguns compostos, no entanto falham para os outros.^{8,9,10,11}

As experiências e estudos sugerem que os PAHs requerem uma ativação metabólica onde o começo da carcinogênese propiciada pelo fato de que carcinogênicos químicos se ligam às macromoléculas celulares como o DNA, RNA e proteínas e que os carcinogênicos finais, de caráter electrofílico, reagem com os grupos nucleofílicos presentes nessas macromoléculas.

Para o nosso trabalho nós consideramos um grupo considerável de PAHs que são mostrados nas figuras 2 e 3.



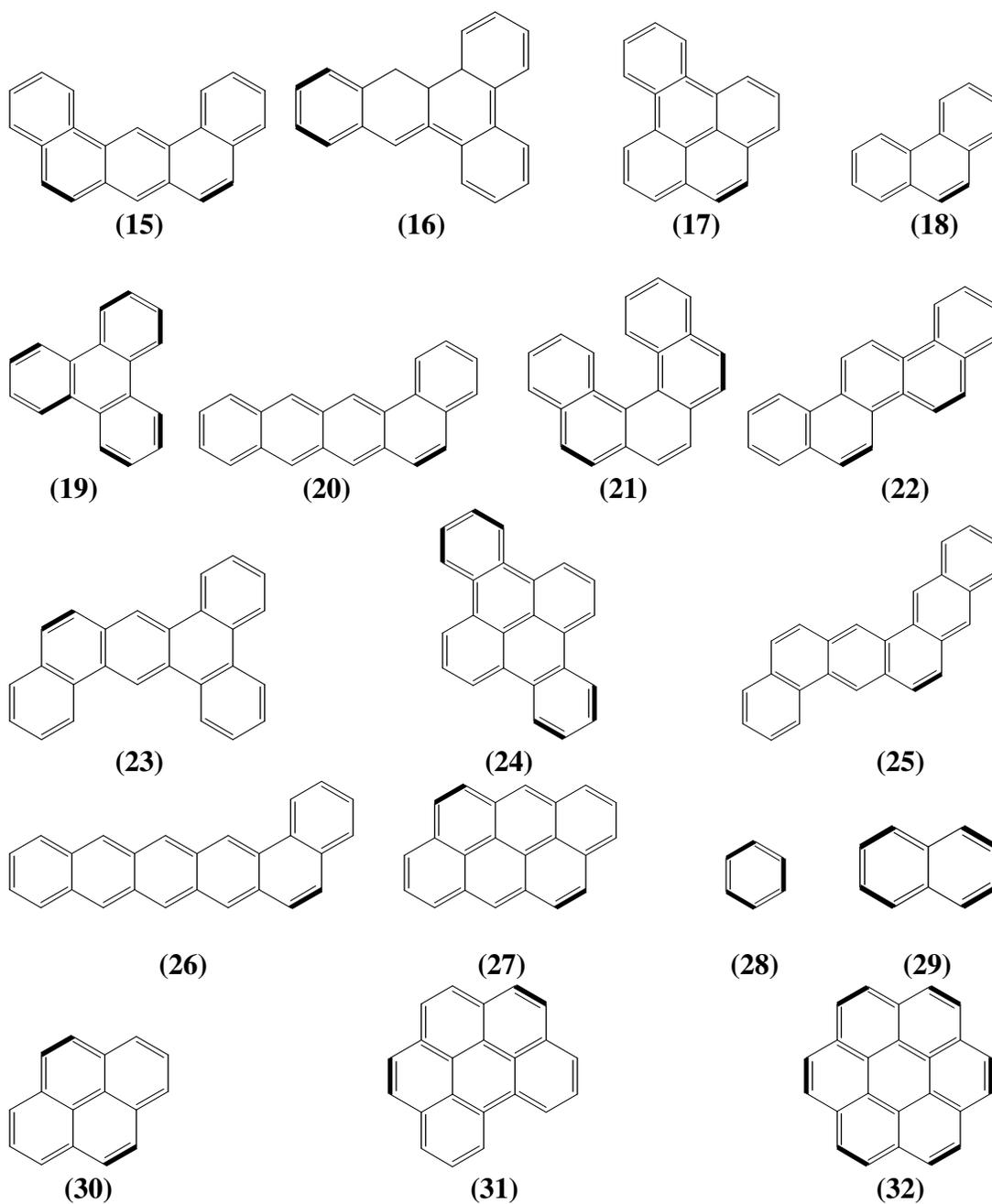


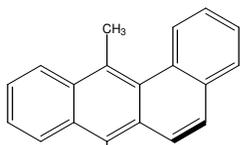
Fig 2: Estrutura Molecular dos 32 hidrocarbonetos policíclicos aromáticos não metilados. As ligações mais escuras indicam ligações com os valores da maior ordem de ligação.

Tabela 1: Atividade carcinogênica (AC) dos hidrocarbonetos policíclicos aromáticos não metilados correspondentes à figura 2, **A** e **I** referem-se a atividade carcinogênica (AC) observada experimentalmente, os compostos são ativos (**A**) e (**I**) inativos.

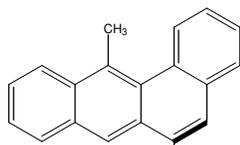
Molécula	AC	Molécula	AC
(1) dibenzo[3,4;9,10]pyrene	A	(17) benzo[1,2]pyrene	I
(2) benzo[3,4]pyrene	A	(18) phenantrene	I
(3) dibenzo[3,4;8,9]pyrene	A	(19) triphenylene	I
(4) dibenzo[3,4;6,7]pyrene	A	(20) benzo[1,2]naphthacene	I
(5) dibenzo[1,2;3,4]pyrene	A	(21) dibenzo[3,4;5,6]phenantrene	I
(6) naphtho[2,3;3,4]pyrene	A	(22) picene	I
(7) dibenzo[1,2;5,6]anthracene	A	(23) tribenzo [1,2;3,4;5,6]anthracene	I
(8) tribenzo[3,4;6,7;8,9]pyrene	A	(24) dibenzo[1,2;5,6]pyrene	I
(9) dibenzo[1,2;3,4]phenantrene	A	(25) phenanthra[2,3;1,2]anthracene	I
(10) tribenzo[3,4;6,7;8,9]pyrene	A	(26) benzo[1,2]pentacene	I
(11) dibenzo[1,2;5,6]phenantrene	I	(27) anthanthrene	I
(12) benzo[1,2]anthracene	I	(28) benzene	I
(13) chrysene	I	(29) naphthalene	I
(14) benzo[3,4]phenantrene	I	(30) pyrene	I
(15) dibenzo[1,2;7,8]anthracene	I	(31) benzo[ghi]preylene	I
(16) dibenzo[1,2;3,4]anthracene	I	(32) coronene	I

Tabela 2: Atividade carcinogênica (AC) dos hidrocarbonetos policíclicos aromáticos metilados correspondentes à figura 3, **A** e **I** referem-se a atividade carcinogênica (AC) observada experimentalmente, os compostos são ativos (**A**) e (**I**) inativos.

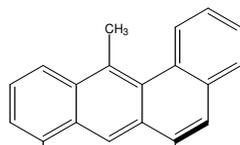
Molécula	AC	Molécula	AC
(33) 7,12-dimethylbenz[a]anthracene	A	(58) 3-methylbenzo[c]phenanthrene	I
(34) 6,12-dimethylbenz[a]anthracene	A	(59) 6-methylbenzo[c]phenanthrene	I
(35) 6,8,12-trimethylbenz[a]anthracene	A	(60) 6-methylbenz[c]anthracene	I
(36) 2-methylbenzo[a]pyrene	A	(61) 12-methylbenz[c]anthracene	I
(37) 4-methylbenzo[a]pyrene	A	(62) 6-methylanthanthrene	I
(38) 11-methylbenzo[a]pyrene	A	(63) 6,12-dimethylanthanthrene	I
(39) 12-methylbenzo[a]pyrene	A	(64) 1-methylbenzo[c]phenanthrene	I
(40) 1-methylbenzo[a]pyrene	A	(65) 2-methylbenzo[c]phenanthrene	I
(41) 4,5-dimethylbenzo[a]pyrene	A	(66) 10-methylbenzo[a]pyrene	I
(42) 3-methylbenzo[a]pyrene	A	(67) 6-methylchrysene	I
(43) 1,2-dimethylbenzo[a]pyrene	A	(68) 3-methylbenz[a]anthracene	I
(44) 2,3-dimethylbenzo[a]pyrene	A	(69) 1-methylbenz[a]anthracene	I
(45) 3,12-dimethylbenzo[a]pyrene	A	(70) 11-methylbenz[a]anthracene	I
(46) 1,3-dimethylbenzo[a]pyrene	A	(71) 9-methylbenz[a]anthracene	I
(47) 1, 4-dimethylbenzo[a]pyrene	A	(72) 2-methylbenz[a]anthracene	I
(48) 5-methylbenzo[a]phenanthrene	A	(73) 5-methylbenz[a]anthracene	I
(49) 5-methylchrysene	A	(74) 8-methylbenz[a]anthracene	I
(50) 6,8-dimethylbenz[a]anthracene	A	(75) 2-methylpyrene	I
(51) 7-methylbenz[a]anthracene	A	(76) 4-methylpyrene	I
(52) 5-methylbenzo[a]pyrene	A	(77) 1-methylpyrene	I
(53) 7-methylbenzo[a]pyrene	A	(78) 7,10-dimethylbenzo[a]pyrene	I
(54) 6-methylbenzo[a]pyrene	A	(79) 6,10-dimethylbenzo[a]pyrene	NA
(55) 1,6-dimethylbenzo[a]pyrene	A	(80) 8-methylbenzo[a]pyrene	NA
(56) 3,6-dimethylbenzo[a]pyrene	A	(81) 9-methylbenzo [a]pyrene	NA



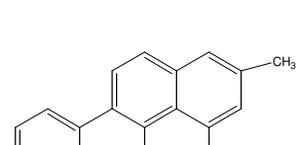
(33)



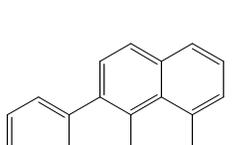
(34)



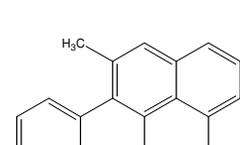
(35)



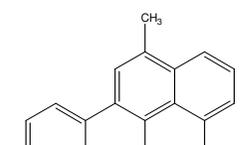
(36)



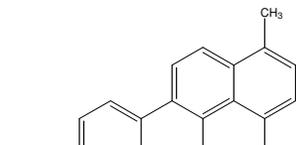
(37)



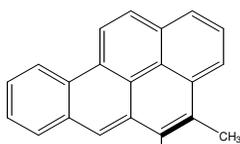
(38)



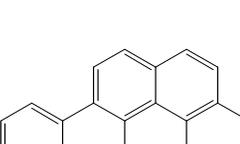
(39)



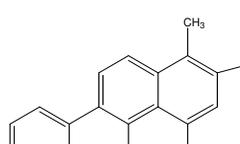
(40)



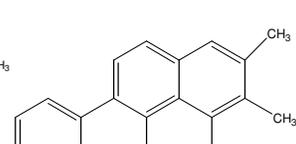
(41)



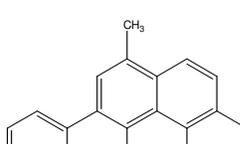
(42)



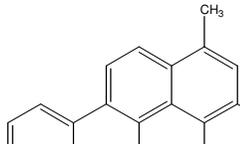
(43)



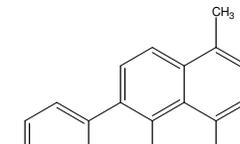
(44)



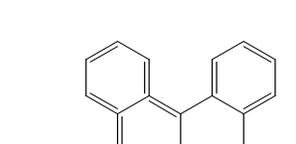
(45)



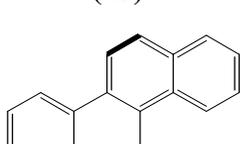
(46)



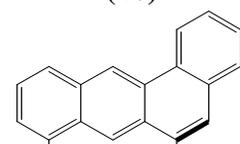
(47)



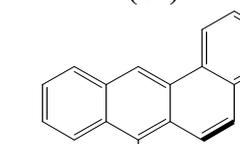
(48)



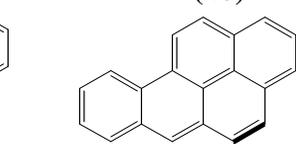
(49)



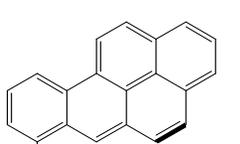
(50)



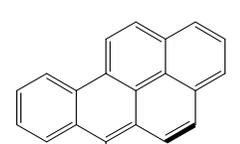
(51)



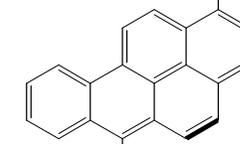
(52)



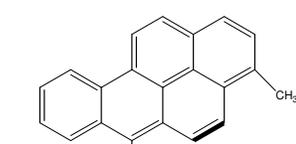
(53)



(54)



(55)



(56)

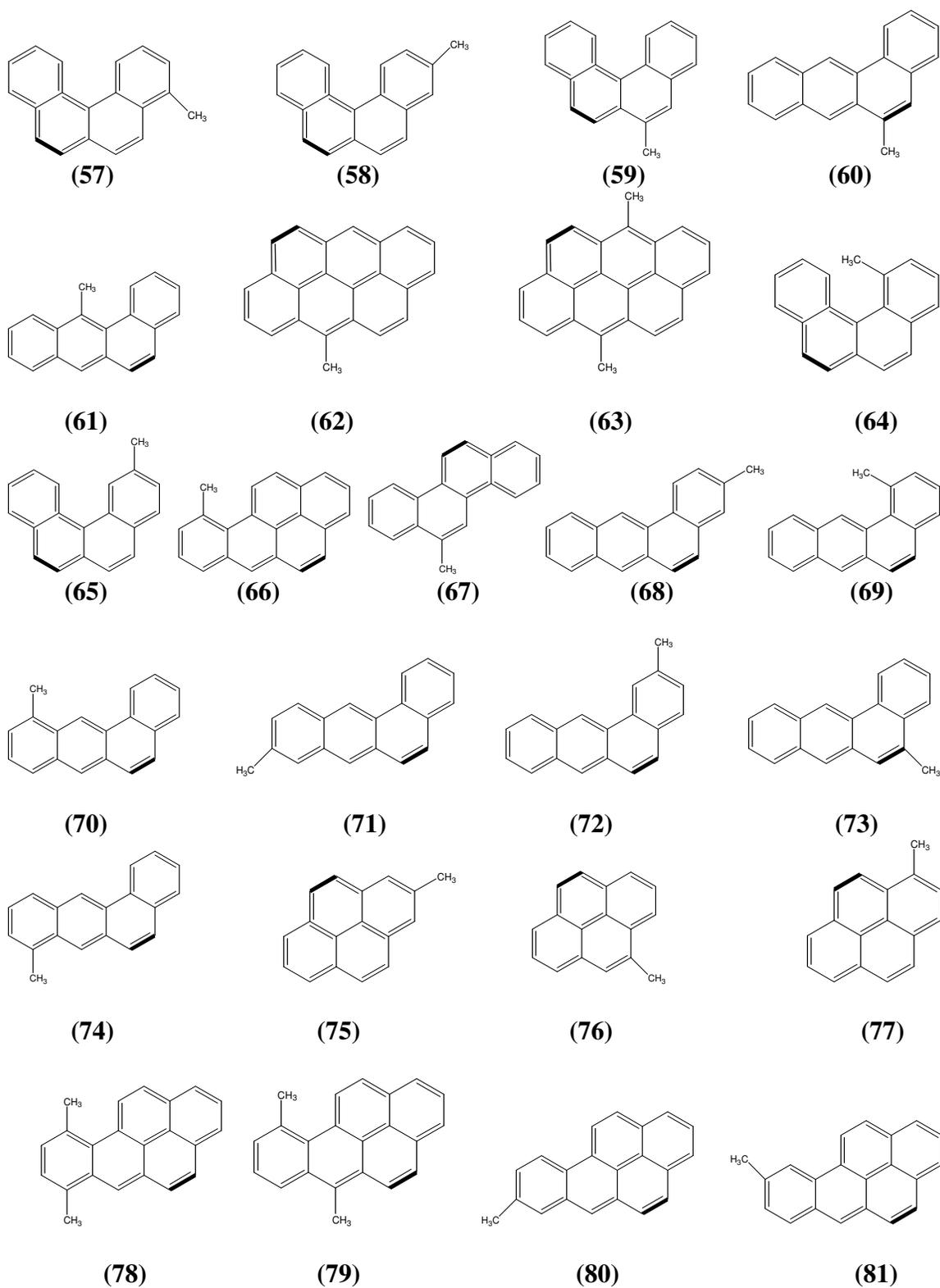


Fig 3: Estrutura Molecular dos 49 hidrocarbonetos policíclicos aromáticos metilados. As ligações mais escuras indicam ligações com os valores da maior ordem de ligação.

A classificação da atividade dos compostos que utilizamos está baseada no trabalho de Villemin *et al.*¹⁰, onde simplesmente se classificam os compostos em ativos (A) ou inativos (I). Os dados experimentais utilizados para definir a atividade estão baseados nos índices Iball⁸ e na escala proposta por Cavalieri *et al.*¹². Os índices Iball são definidos como a percentagem de câncer de pele ou de desenvolvimento de papiloma em ratos, em experimentos de pintura de pele de ratos com os materiais a serem estudados, dividida pelo período médio de latência em dias para os animais afetados.

A escala usada por Cavalieri¹³ para classificar a carcinogenicidade dos PAHs segue as definições: +++++ extremamente ativo; ++++ muito ativo; +++ ativo; ++ moderadamente ativo; + fracamente ativo; ± muito fracamente ativo e; - inativo. No nosso caso consideraremos que um composto é simplesmente denominado inativo se ele possuir a classificação ++, +, ± ou -, e ativo para as demais classificações^{13,14}.

Referencias

- ¹ A. Lietch, J. Brit. Med. **2**, 1104 (1922).
- ² M. Cooke and A. Dennis, *Polynuclear Aromatic Hydrocarbons: Chemistry, Characterization and Carcinogenesis, Ninth International Symposium* (1984).
- ³ I. Hieger, *Carcinogenesis* (Academic Press, London, 1961).
- ⁴ B. Pullman, **16**, 669 (1979).
- ⁵ G. Ronald and Harvey, *Polycyclic Aromatic Hydrocarbons* (Wiley-VCH- USA, 1997).
- ⁶ A. Pullman and B. Pullman, *Adv. Cancer Res.*, **3**, 117 (1955).
- ⁷ D. M. Jerina, *et al. Carcinogenesis: Fundamental Mechanisms and Environmental Effects*; B. Pullman, P. O. Ts'ao, H. Gelboin, *et al.*, Eds. D. Reidel Publishing Co. (Dordrecht, The Netherlands, 1980).
- ⁸ J. Gayoso and S. Kimri, *Int. J. Quantum Chem.*, **38**, 461 (1990); **38**, 487 (1990).
- ⁹ U. E. Nordén and W. Svante, *Acta Chem. Scand.* **B32**, 602 (1978).
- ¹⁰ D. Villemin, D. Cherqaoui and A. Mesbah, *J. Chem. Inf. Comput. Sci.* **34**, 1288 (1994).
- ¹¹ X. H. Song, M. Xiao and R. Q. Yu, *Comput. Chem.* **18**, 391 (1994).
- ¹² E. L. Cavalieri *et al.*, *Chem.-Biol. Interactions.* **47**, 87 (1983).
- ¹³ R. Vendrame, R. S. Braga, Y. Takahata and D. S. Galvão, *J. Mol. Struct. (THEOCHEM)*, **539**, 253 (2001).
- ¹⁴ P. M. V. B. Barone, A. Camilo Jr. and D. S. Galvão, *Phys. Rev. Lett.* **77**, 1186 (1996).

Capítulo 5

Resultados e Discussões

5.1 Otimização Geométrica e Análise Conformacional

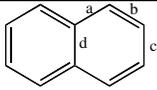
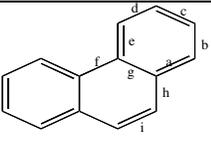
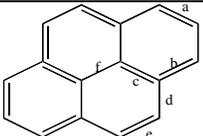
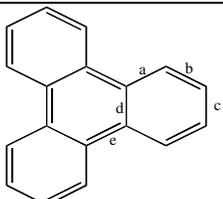
Como parte de nossa investigação na correlação da estrutura-atividade biológica dos compostos carcinogênicos, os cálculos para as propriedades eletrônicas, as quais são muito sensíveis à geometria dos compostos, foram feitos obtendo a estrutura geométrica de mínima energia, utilizando os três métodos semi-empíricos: AM1, PM3 e PM5 com o programa Chem2Pac que integra o pacote computacional MOPAC6 no ambiente Windows para os métodos PM3 (Parametric Method 3), AM1 (Austin Method One) e o programa Cache5 para o PM5 (Parametric Method 5).

No conjunto dos Hidrocarbonetos policíclicos Aromáticos (PAHs) estudados no trabalho, podemos separar dois grupos pelo fato de serem ou não metilados.

As trinta e duas primeiras moléculas são moléculas formadas pela união de vários anéis benzênicos aromáticos e em sua maioria são moléculas planares.

Foram feitas comparações das estruturas geométricas obtidas para cada um dos métodos, sobrepondo uma mesma estrutura otimizada de cada um dos métodos semi-empíricos utilizados, para comprovar a igualdade na otimização da estrutura. Os resultados foram comparados com dados de raios-X para algumas estruturas das quais se tem referencia experimental e que são a estrutura base fixa para os demais compostos.

Tabela 1: Comparação do comprimento de ligação (Å) mais relevante para alguns PAHs com referencia de dados experimentais de R-X^{1,2}

 R-X	AM1	PM3	PM5
a: 1.422 b: 1.371 c: 1.412 d: 1.420	a: 1.422 b: 1.368 c: 1.415 d: 1.421	a: 1.422 b: 1.368 c: 1.415 d: 1.421	a:1.422 b:1.360 c:1.409 d:1.421
 R-X	AM1	PM3	PM5
a: 1.423 b: 1.386 c: 1.394 d: 1.401 e: 1.409 f: 1.465 g: 1.420 h: 1.453 i: 1.350	a: 1.412 b:1.376 c: 1.405 d: 1.411 e: 1.411 f: 1.411 g: 1.406 h: 1.437 i: 1.354	a: 1.412 b:1.376 c: 1.405 d: 1.411 e: 1.411 f: 1.411 g: 1.406 h: 1.437 i: 1.354	a:1.437 b:1.369 c:1.402 d:1.370 e:1.411 f:1.411 g:1.406 h:1.437 i:1.343
 R-X	AM1	PM3	PM5
a: 1.395 b: 1.406 c: 1.425 d: 1.438 e: 1.367 f: 1.430	a: 1.398 b:1.398 c: 1.413 d: 1.441 e: 1.352 f: 1.413	a: 1.398 b:1.398 c: 1.413 d: 1.441 e: 1.352 f: 1.413	a:1.385 b:1.394 c:1.412 d:1.443 e:1.342 f:1.412
 R-X	AM1	PM3	PM5
a: 1.410	a: 1.407	a: 1.407	a:1.406

b: 1.381	b:1.380	b:1.380	b:1.374
c: 1.347	c: 1.398	c: 1.398	c:1.394
d: 1.413	d: 1.405	d: 1.405	d:1.404
e: 1.458	e: 1.454	e: 1.454	e:1.455

Nos compostos metilados pelo fato de apresentarem flexibilidade do ângulo diedral ϕ Fig(1) foi imprescindível uma análise detalhada dos seus vários conformémeros acessíveis na busca de sua geometria mais estável, de mínimo global (M.G.).

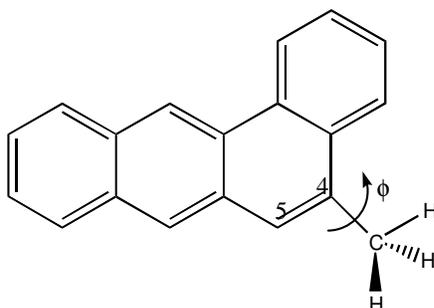


Fig 1: 5-metilbenz[a]antraceno, ângulo diedral ϕ formado pelos C4,C5,C e H.

Os diedrais foram girados simultaneamente em passos de 10 graus, varrendo toda a liberdade das ligações em um total aproximado de 400 conformações para o caso dos compostos com uma substituição, e em um total de 1032 conformações obtidas para cada uma das moléculas com duas substituições Fig (2).

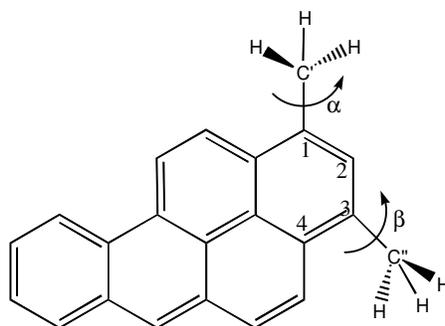


Fig 2 : 1,3-dimetilbenzo[a]pireno, ângulos diedrais α e β formados pelos átomos: C1,C2,C' e H; C3,C4, C'' e H respectivamente.

Depois de obter as geometrias otimizadas de mínima energia das 81 moléculas consideradas no vácuo nos empenhamos em obter as propriedades eletrônicas, físico-químicas e estereoquímicas associadas à cada composto para correlacioná-las com a atividade carcinogênica dos compostos.

5.2 Metodologia dos Índices Eletrônicos

Como já foi explicado, o objetivo da MIE é o de separar moléculas ativas das inativas com uma atividade biológica específica, utilizando dois parâmetros eletrônicos: Δ e η .

A LDOS foi analisada em detalhe para os orbitais de fronteira HOMO, HOMO-1, LUMO e LUMO+1.

Como no estudo da MIE a divisão das moléculas em regiões é essencial, observamos que para esta classe de compostos os valores críticos (Δ_c e η_c) e as regras de seleção que melhor classificam os PAHs com atividade carcinogênica ou não (de acordo com as informações experimentais) é o anel que contém o maior valor na ordem de ligação (bond order: b)^a nos orbitais de fronteira HOMO e HOMO-1.

Estes valores críticos são obtidos automaticamente em termos de relevância estatística. Observando o conjunto dos valores de Δ e η obtidos para os PAHs estudados, a busca consiste em encontrar os valores Δ e η que determinam a fronteira para poder classificar a molécula como ativa ou inativa que esteja em concordância com a classificação experimental que proporcione a maior percentagem de acerto.

A região, selecionada pelos parâmetros eletrônicos é considerada do ponto de vista biológico a região mais interessante e sensível aos ataques eletrofílicos.

Foi observado também que compostos como o caso das moléculas: 3, 7, 13, 14, 15, 19, 21, 22, 24, 27, 29, 30, 31, 32 e 63 que apresentam mais de um anel com o maior e

^a Ordem de ligação (Bond order): é um parâmetro muito útil para a discussão das características das ligações, porque está correlacionado com o comprimento e força da ligação. Um maior valor na ordem de ligação dentre átomos de um par de elementos, indica que o comprimento da ligação é o mais curto e a ligação mais forte. $b=1$, corresponde a uma ligação simples entre dois átomos e $b=0$ indica que não existe ligação.

mesmo valor da ordem de ligação, as somas da LDOS dos dois anéis, classificam de forma correta a atividade do composto. No caso de se considerar somente um anel no cálculo do parâmetro η_H o índice de acerto diminui.

Analisando os valores dos descritores frente à classificação experimental de atividade carcinogênica os valores críticos e regras de seleção de atividade para os três métodos semi-empíricos foram facilmente obtidos como:

Tabela 2: Valores críticos ΔH e η_H para os três métodos semi-empíricos, AM1, PM3 e PM5.

Valores Críticos (eV)	AM1	PM3	PM5
ΔH_c	0.605	0.31	0.304
$\eta_H c$	-0.107	0.020	0.020

Em resumo, as regras podem ser escritas de forma simplificada numa tabela Booleana de atividades.

Tabela 3: Regras de seleção para os três métodos semi-empíricos, AM1, PM3, PM5.

ΔH	η_H	Atividade
+	+	A
+	-	I
-	+	I
-	-	I

Onde os sinais + e - significam que os valores de ΔH e η_H para a molécula analisada são maiores/menores que os valores críticos.

Para o caso específico de η_H , na comparação entre os dois orbitais HOMO e HOMO-1, o sinal + indica que os átomos da região estudada contribuem mais fortemente para a formação do orbital H, no entanto o sinal - indica que a região contribui mais para a formação do H-1.

Aplicando as regras dos parâmetros da MIE para os 81 compostos com cada um dos métodos Semi-empíricos, temos:

Tabela 4: Valores de ΔH e ηH para os 81 PAHs calculados com os três métodos semi-empíricos AM1, PM3 e PM5, composto ativo (A), composto inativo(I), Det. refere-se a atividade calculada pela MIE e Exp. é a atividade determinada experimentalmente do composto. As células escuras mostram os erros de predição da MIE. Atividade experimental não determinada (ND).

Comp	PM3		Det	Exp	AM1		Det	PM5		Det
	ΔH	ηH			ΔH	ηH		ΔH	ηH	
1	0.69903	0.30022	A	A	0.72324	0.304058	A	0.6	0.26162	A
2	0.85952	0.072638	A	A	0.88588	0.076098	A	0.74	0.09535	A
3	1.07246	0.119455	A	A	1.10053	0.113276	A	0.936	0.120358	A
4	0.67604	-0.147808	I	A	0.69607	-0.145361	I	0.598	-0.19374	I
5	0.60036	0.095569	A	A	0.60916	0.093967	A	0.527	0.06238	A
6	0.91253	0.212004	A	A	0.93643	0.215458	A	1.217	0.182088	A
7	0.32574	-0.026368	I	A	0.34052	-0.007082	I	0.26	-0.089	I
8	0.82939	0.193138	A	A	0.8373	0.186002	A	0.732	0.145934	A
9	0.35864	0.426706	A	A	0.401	0.414875	I	0.344	0.41606	A
10	0.60697	-0.486259	I	A	0.63012	-0.470382	I	0.536	-0.49318	I
11	0.07021	-0.276654	I	I	0.05846	-0.327396	I	0.034	-0.16352	I
12	0.56334	-0.086206	I	I	0.57694	-0.077243	I	0.484	-0.08756	I
13	0.42044	-0.233936	I	I	0.43789	-0.231151	I	0.336	-0.2619	I
14	0.24337	-0.383451	I	I	0.26995	-0.39847	I	0.247	-0.37062	I
15	0.25358	-0.148093	I	I	0.25867	-0.133283	I	0.218	-0.1861	I
16	0.40234	-0.065036	I	I	0.41142	-0.050886	I	0.343	-0.14238	I
17	0.48955	-0.068675	I	I	0.50272	-0.058494	I	0.444	-0.13422	I
18	0.23629	0.139574	I	I	0.2496	0.142076	I	0.179	0.09407	I
19	0.00003	-0.000016	I	I	0.00001	0.000031	I	0	-0.000106	I
20	0.91791	-0.230368	I	I	0.93289	-0.224701	I	0.805	-0.18206	I
21	0.1636	0.221424	I	I	0.15205	0.291109	I	0.113	0.286674	I
22	0.20964	-0.023465	I	I	0.22961	-0.01194	I	0.139	-0.07298	I
23	0.15407	0.017977	I	I	0.1624	0.030169	I	0.122	-0.04518	I
24	0.30391	0.164707	I	I	0.3039	0.326337	I	0.301	0.360904	I
25	0.24792	0.249464	I	I	0.25901	0.252873	I	0.202	0.225696	I
26	1.08777	-0.337589	I	I	1.10968	-0.339497	I	1.022	-0.30328	I
27	1.17016	-0.125461	I	I	1.20095	-0.122282	I	1.013	-0.1387	I
28	0.00013	-0.000006	I	I	0.00001	-0.000017	I	0.001	-0.00002	I
29	0.59988	-0.000865	I	I	0.62974	-0.430583	I	0.489	-0.00036	I
30	0.79269	-0.11421	I	I	0.81908	-0.112682	I	0.683	0.014	I
31	0.5688	-0.065121	I	I	0.58641	-0.063165	I	0.528	-0.09226	I
32	0.00003	-0.000016	I	I	0.00002	0.000011	I	0.001	-0.00008	I
33	0.71396	-0.159295	I	A	0.71729	-0.137237	I	0.677	-0.13848	I
34	0.65034	-0.067575	I	A	0.65358	-0.056223	A	0.595	-0.06348	I
35	0.67751	-0.101557	I	A	0.68544	-0.099225	A	0.636	-0.11236	I
36	0.77534	0.057406	A	A	0.79955	0.061101	A	0.638	0.031216	A
37	0.8701	0.062059	A	A	0.89738	0.075397	A	0.753	0.04378	A
38	0.85402	0.083493	A	A	0.88015	0.086237	A	0.726	0.053918	A

39	0.90621	0.041755	A	A	0.92274	0.044434	A	0.793	0.021008	A
40	0.91331	0.073814	A	A	0.94074	0.076195	A	0.816	0.05136	A
41	0.90371	0.116118	A	A	0.91669	0.118361	A	0.777	0.075852	A
42	0.87046	0.131633	A	A	0.89524	0.135161	A	0.766	0.11902	A
43	0.83569	0.052922	A	A	0.86025	0.055612	A	0.721	0.028512	A
44	0.78944	0.111803	A	A	0.81463	0.119269	A	0.666	0.1025	A
45	0.89118	0.106809	A	A	0.92641	0.105148	A	0.808	0.08445	A
46	0.91262	0.137859	A	A	0.94766	0.133862	A	0.838	0.120702	A
47	0.92004	0.057966	A	A	0.94847	0.068729	A	0.825	0.03919	A
48	0.28361	0.012324	I	A	0.30837	0.002523	I	0.306	0.013036	I
49	0.41219	0.02637	A	A	0.44153	0.024299	I	0.323	0.02408	A
50	0.63564	-0.048187	I	A	0.63243	-0.044692	A	0.566	-0.05774	I
51	0.64851	-0.111528	I	A	0.66173	-0.106132	A	0.597	-0.09648	I
52	0.89389	0.116415	A	A	0.90971	0.115377	A	0.773	0.0814	A
53	0.87476	0.094849	A	A	0.89801	0.090587	A	0.768	0.057304	A
54	0.93233	0.084354	A	A	0.95356	0.094998	A	0.834	0.08032	A
55	0.99178	0.077086	A	A	1.00513	0.096401	A	0.905	0.085868	A
56	0.92848	0.153058	A	A	0.9568	0.155798	A	0.85	0.153202	A
57	0.26555	0.007012	I	I	0.29536	-0.002006	I	0.281	0.003386	I
58	0.26667	0.144428	I	I	0.29343	0.126572	I	0.278	0.153244	I
59	0.17502	0.084201	I	I	0.2047	0.097359	I	0.164	0.132056	I
60	0.59217	-0.0051	I	I	0.59784	-0.00828	I	0.521	0.001626	I
61	0.63625	-0.133979	I	I	0.64567	-0.12001	I	0.578	-0.13012	I
62	1.21791	-0.089117	I	I	1.24349	-0.093938	A	1.08	-0.10636	I
63	1.26611	-0.112456	I	I	1.28829	-0.108916	I	1.146	-0.05518	I
64	0.30293	0.126965	I	I	0.31079	0.121207	I	0.301	0.131616	I
65	0.18892	0.095274	I	I	0.21495	0.085885	I	0.181	0.11162	I
66	0.88051	0.080338	A	I	0.88826	0.092945	A	0.739	0.062368	A
67	0.50396	0.018856	I	I	0.51233	0.019396	I	0.432	0.013876	I
68	0.52193	-0.096977	I	I	0.53232	-0.088292	I	0.44	-0.10252	I
69	0.49077	0.006755	I	I	0.49203	0.020801	I	0.386	0.01687	I
70	0.61146	-0.155514	I	I	0.60974	-0.141239	I	0.53	-0.16974	I
71	0.52513	-0.115139	I	I	0.53736	-0.105367	I	0.443	-0.12464	I
72	0.59084	-0.029105	I	I	0.60402	-0.020494	I	0.527	-0.01948	I
73	0.5538	-0.078947	I	I	0.56738	-0.065358	I	0.471	-0.06682	I
74	0.60184	-0.133971	I	I	0.61843	-0.1244	I	0.539	-0.14616	I
75	0.66603	-0.227986	I	I	0.68747	-0.212063	I	0.531	-0.22944	I
76	0.81308	-0.244756	I	I	0.84138	-0.293039	I	0.715	-0.32784	I
77	0.8469	-0.172047	I	I	0.87375	-0.225268	I	0.764	-0.2384	I
78	0.87254	0.1147	A	I	0.89674	0.11139	A	0.764	0.07622	A
79	0.92156	0.10286	A	ND	0.95666	0.109755	A	0.83	0.186476	A
80	0.77212	0.110486	A	ND	0.79537	0.11449	A	0.638	0.1868	A
81	0.87365	0.042909	A	ND	0.90077	0.0464	A	0.762	0.024324	A

Acerto 85,9% (67/78)

Acerto 87,2% (68/78)

Acerto 85,9%

Na classificação da MIE os resultados da análise teórica e experimental discordam em apenas 11 compostos do conjunto de 78 moléculas no caso dos cálculos serem feitos com os métodos PM3 e o PM5. O que significa um acerto de 85,9% na predição da atividade biológica feita pela MIE, e de 10 compostos do conjunto de 78 moléculas no caso da MIE ser calculada com o método semi-empírico AM1 o que nos leva a um acerto de 87,2% na predição da atividade biológica. No caso das últimas três moléculas cuja atividade não foi determinada experimentalmente a MIE indica serem moléculas potencialmente carcinogênicas.

Observamos também que não existe uma dependência explícita entre a qualidade dos resultados obtidos e o Hamiltoniano eletrônico utilizado para a MIE.

5.3 Análise de Componentes Principais - PCA:

Com o objetivo de testar os índices eletrônicos da metodologia e sua dependência com o Hamiltoniano eletrônico utilizado, realizamos a análise de componentes principais.

Nos estudos de estrutura-atividade procura-se correlacionar parâmetros originados dessas propriedades à determinada resposta biológica de interesse. E em nosso caso investigaremos os parâmetros que possam correlacionar a atividade carcinogênica do conjunto de moléculas estudado.

Para a aplicação da técnica de PCA geramos inicialmente 22 parâmetros:

- E_{HOMO} : Valor da energia do orbital mais alto ocupado (eV),
- E_{LUMO} : Valor da energia do orbital mais baixo desocupado (eV),
- $E_{\text{HOMO}-1}$, (eV)
- $E_{\text{LUMO}+1}$, (eV),
- $\Delta H = E_{\text{HOMO}} - E_{\text{HOMO}-1}$, (eV),
- $\Delta L = E_{\text{LUMO}+1} - E_{\text{LUMO}}$, (eV),
- $HD = (\text{LUMO} - \text{HOMO})/2$: Dureza, (eV)
- $XE = -(\text{HOMO} + \text{LUMO})/2$: Eletronegatividade de Mulliken, (eV)

- CH e CH-1 contribuições do HOMO e HOMO-1 para a densidade local de estados sobre o anel da maior ordem de ligação, (eV),
- $\eta_H = (CH) - (CH-1)$, (eV),
- CL e CL+1 contribuições do LUMO e LUMO+1 para a densidade local de estados sobre o anel da maior ordem de ligação, (eV),
- $\eta_L = (CL+1) - (CL)$, (eV),
- Log P: Coeficiente de partição molecular octanol-agua.
- $Nat = (N-20)^3$, onde N é o número de átomos de carbono na molécula: Descritor empírico que correlaciona o tamanho da molécula com a sua atividade carcinogênica,
- Polarizabilidade, (u.a.),
- Massa, (u.a.),
- Volume, (\AA^3),
- Calor de formação, (kcal mol^{-1}),
- Refratividade,
- Energia de Hidratação (kcal mol^{-1}).

Depois de realizar a análise de componentes principais usando o pacote computacional Pirouette³, que contém a PCA e todos os métodos de quimiometria relacionados e usados nesse trabalho, observamos os seguintes resultados:

Antes de aplicar a Análise de Componentes Principais (PCA), cada uma das variáveis foi autoescalada para poderem ser comparadas entre si na mesma escala.

A figura.3. mostra a distribuição inicial dos compostos no espaço gerado pelas componentes principais PC1 e PC2 usando 22 descritores.

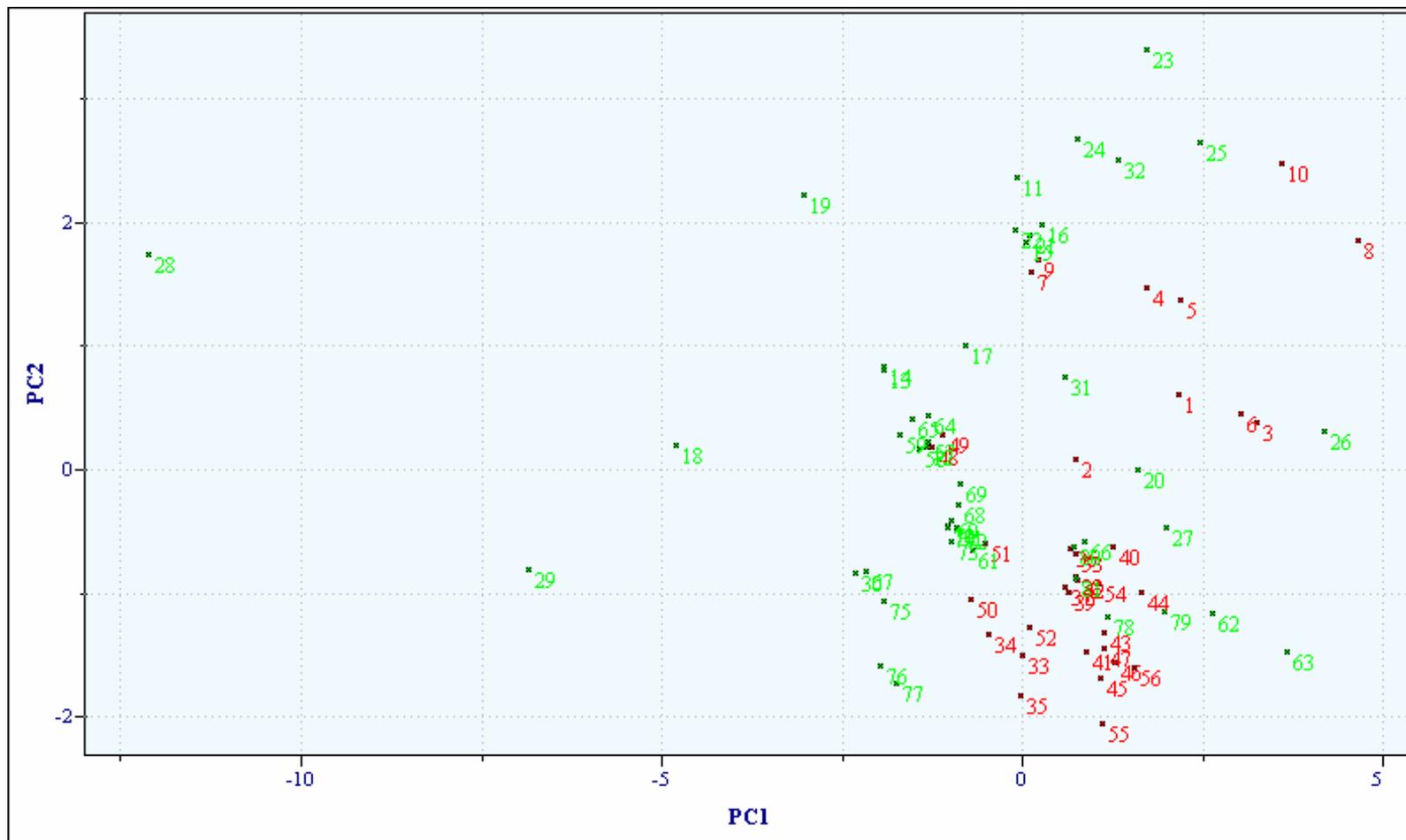


Fig 3: Scores plot das duas primeiras componentes principais dos 81 PAHs com todos os descritores obtidos pelo método semi-empirico AM1.
(Vermelhas) Ativas, (Verdes) inativas.

Depois de realizar uma seleção de variáveis i.e. eliminando as variáveis com alta correlação entre si, observando o gráfico dos Loadings (por ser suficiente a utilização de uma delas) e escolhendo os maiores valores dos loadings na componente principal que descrevam uma separação mais clara dos compostos ativos dos inativos no espaço das componentes principais conseguimos obter os seguintes resultados:

A fig.4 dos scores mostra que os PAHs estão distribuídos em duas diferentes regiões. O grupo dos PAHs ativos encontra-se ao lado esquerdo da figura dos Scores e os inativos ao lado direito. As moléculas que se encontram na região dos compostos ativos são as seguintes:

Tabela 5: Compostos distribuídos na região esquerda da figura dos Scores. Método AM1.

Composto na região dos Carcinogênicos	Atividade Carcinogênica (dado Experimental)
1	A
2	A
3	A
4	A
5	A
6	A
8	A
10	A
20	I
26	I
27	I
31	I
36	A
37	A
38	A
39	A
40	A
41	A
42	A
43	A
44	A
45	A
46	A
47	A
52	A
53	A
54	A
55	A
56	A
62	I
63	I
66	I
78	I
79	ND
80	ND
81	ND

As moléculas 20, 26, 27, 31, 62, 63, 66 e 78 estão distribuídas na região dos compostos ativos sendo compostos sem atividade carcinogênica.

As moléculas 7, 9, 33, 34, 35, 48, 49, 50 e 51 de atividade carcinogênica encontram-se na região dos compostos inativos.

Dos 78 compostos 61 foram corretamente classificados o que corresponde a um acerto de 78.2%.

Com a finalidade de investigar a atividade dos compostos 79, 80 e 81 não testados, a PCA mostrou que eles possuem as mesmas características, no espaço das PCs, dos compostos ativos.

As componentes principais PC1 e PC2 são dadas pelas equações, para o método AM1:

$$\mathbf{PC1} = -0.51484 \text{ HOMO} + 0.50603 \text{ LUMO} - 0.46265 \Delta\text{H} + 0.51462 \text{ HD} \quad (1)$$

$$\mathbf{PC2} = -0.13954 \text{ HOMO} + 0.38945 \text{ LUMO} + 0.87207 \Delta\text{H} + 0.26146 \text{ HD} \quad (2)$$

PC1 corresponde a 92.64% da variância e para a PC2 6.78% da variância tendo um total de 99.42%

E as componentes principais PC1 e PC2 dadas pelas equações, para o método PM3:

$$\mathbf{PC1} = -0.51417 \text{ HOMO} + 0.50550 \text{ LUMO} - 0.46462 \Delta\text{H} + 0.51404 \text{ HD} \quad (3)$$

$$\mathbf{PC2} = -0.13884 \text{ HOMO} + 0.39263 \text{ LUMO} + 0.87061 \Delta\text{H} + 0.26192 \text{ HD} \quad (4)$$

PC1 corresponde a 92.91% da variância e para a PC2 6.50% da variância tendo um total de 99.42% de informação acumulada nas duas primeiras componentes principais.

As equações 1 e 3 indicam que os descritores de maior importância e responsáveis pela separação dos dois grupos são o HOMO, LUMO, ΔH e HD, as equações 2 e 4 para a PC2 descrevem exatamente o mesmo, para os métodos AM1 e PM3 respectivamente.

A projeção das PC1 e PC2 (fig 4 e 5) conserva 99.42% da variância total dos dados originais e podemos esperar que a análise com estas duas componentes principais forneçam uma precisão razoável do espaço de ordem maior.

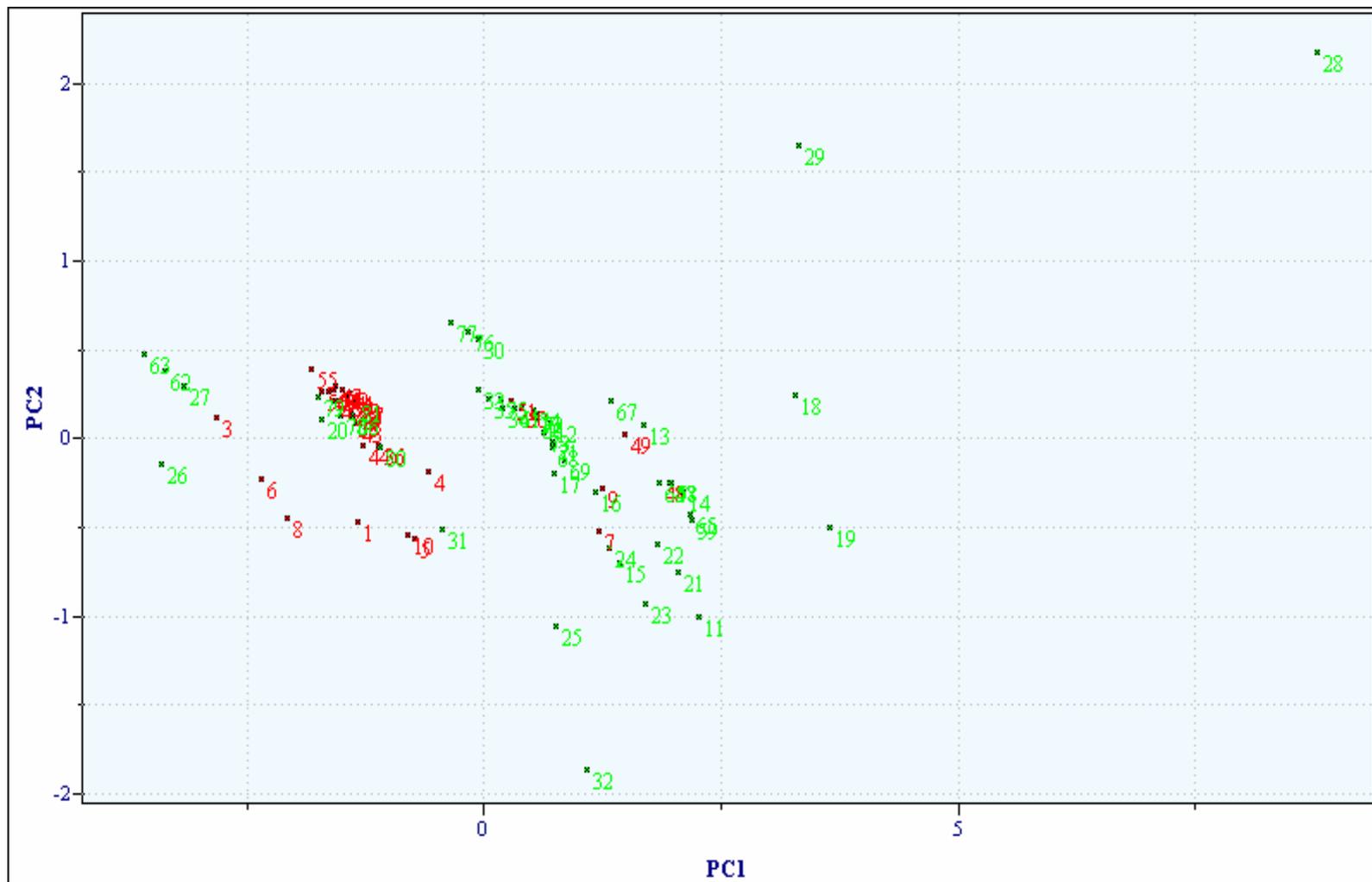


Fig 4: Scores plot das duas primeiras componentes Principais dos 81 PAHs com seleção de variáveis: HOMO, LUMO, ΔH , HD.calculadas com o método AM1
(Vermelhas) Ativas, (Verdes) inativas

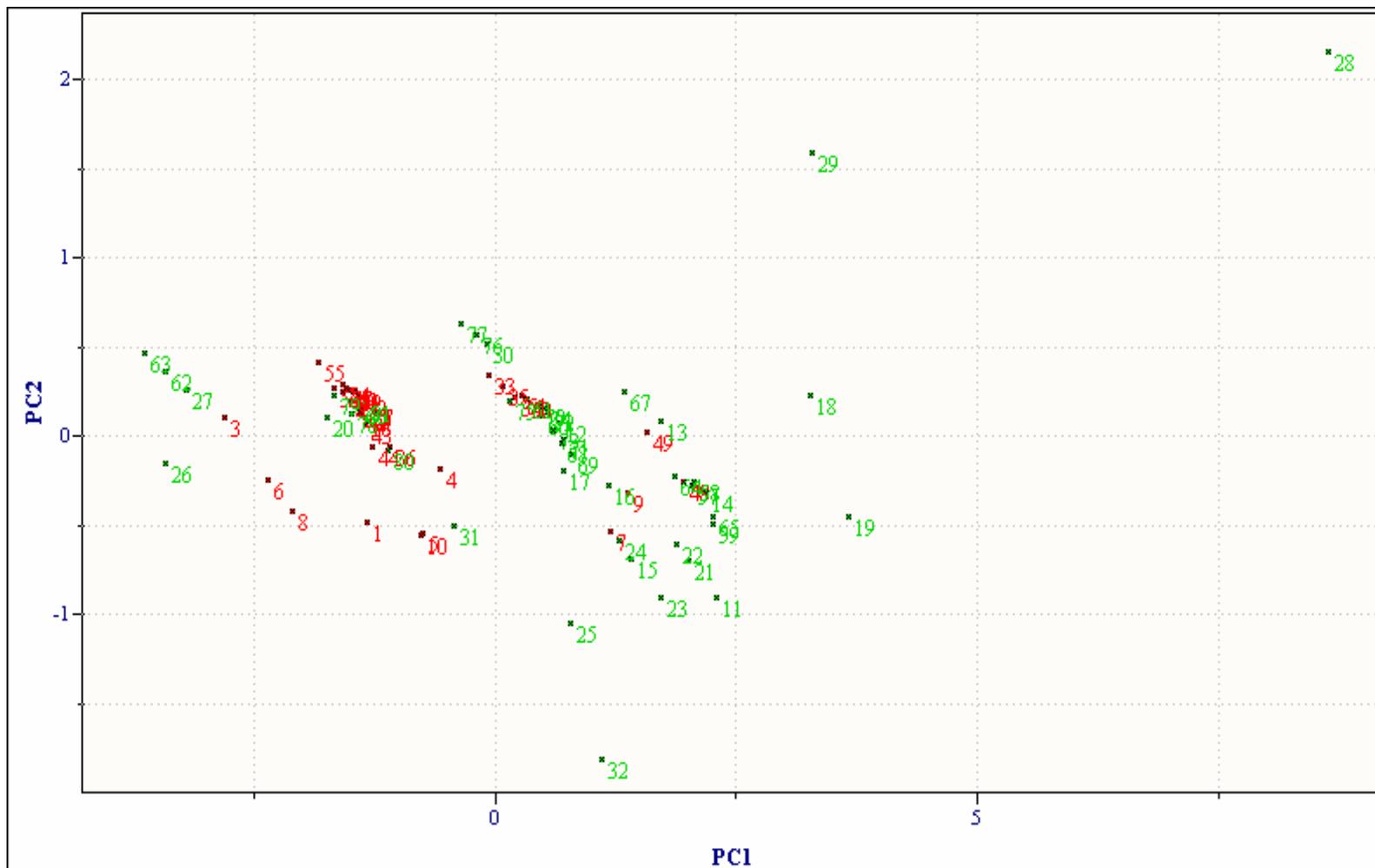


Fig 5: Scores plot das duas primeiras componentes Principais dos 81 PAHs com seleção de variáveis: HOMO, LUMO, ΔH , HD calculadas com o método PM3 (Vermelhas) Ativas, (Verdes) inativas.

Na fig.6 dos Loadings temos a referencia da distribuição dos descritores no espaço das componentes principais podemos observar que para os compostos que se encontram ao lado esquerdo (ativos) os parâmetros de maior importância para estes compostos são o HOMO e ΔH e no caso dos compostos inativos são os parâmetros LUMO e HD.

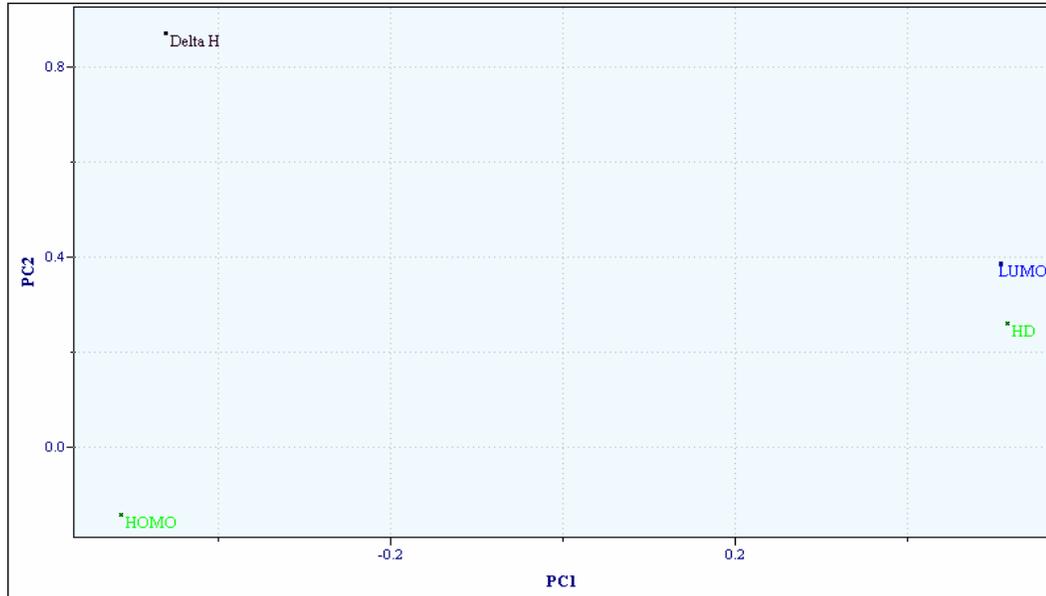


Fig 6: Loading Plot das duas primeiras componentes principais dos 81 PAHs com seleção de variáveis: HOMO, LUMO, ΔH e HD.

No resultado da primeira seleção feita pelo PCA observamos que um dos descritores da MIE foi escolhido como parâmetro importante na separação dos dois grupos, sendo também um descritor fundamental para se obter a separação visível dos dois grupos (ativos/inativos). No interesse de testar e observar a eficiência de utilizar os dois parâmetros da MIE, nos realizamos uma seleção de variáveis mantendo os dois parâmetros nas componentes principais, os resultados foram os seguintes:

A fig.7 dos Scores mostra uma nova distribuição dos compostos no espaço gerado pelas duas primeiras componentes principais.

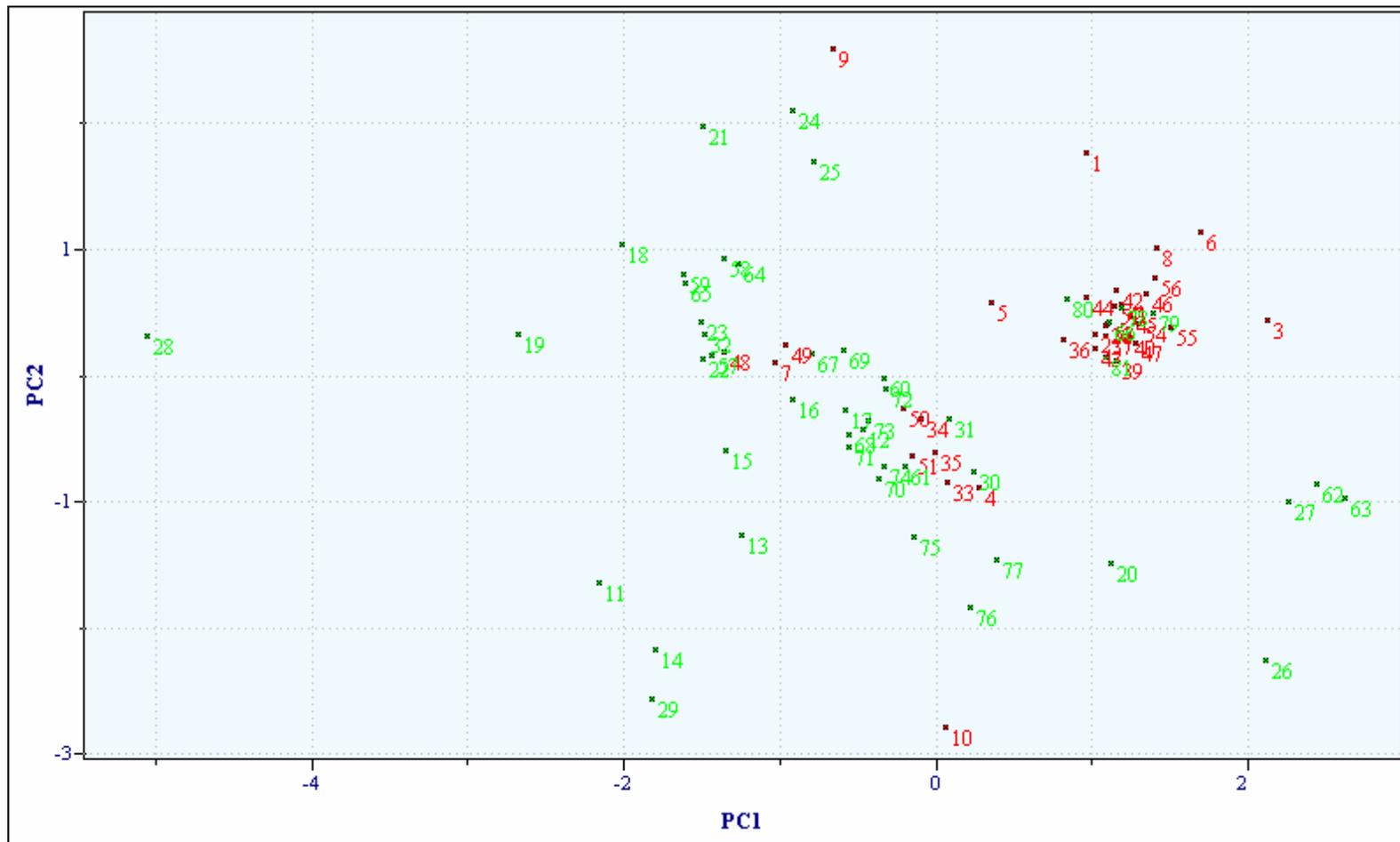


Fig.7: Scores plot das duas primeiras componentes Principais dos 81 PAHs com seleção de variáveis: ΔH , ηH e HD calculadas pelo método AM1.
(Vermelhas) Ativas, (Verdes) inativas

Do lado direito da figura encontram-se os compostos ativos (vermelho) e do lado esquerdo os inativos (verde). Na região dos compostos ativos temos os seguintes compostos inativos: 66 e 78.

Na região dos inativos temos as moléculas 4, 7, 9, 10, 33, 34, 35, 48, 49, 50, 51 incorretamente distribuídas no espaço das componentes principais quanto à sua atividade. Dos 78 compostos 65 foram corretamente classificados, um acerto de 83.3%.

Da mesma maneira os compostos 79, 80 e 81 encontram-se na região dos compostos ativos.

As componentes principais PC1 e PC2 para o método AM1 são dadas pelas equações:

$$\mathbf{PC1} = 0.69838 \Delta H + 0.10975 \eta H - 0.70726 HD \quad (5)$$

$$\mathbf{PC2} = -0.15796 \Delta H + 0.98744 \eta H - 0.00275 HD \quad (6)$$

A PC1 corresponde a 60.99% da variância e para a PC2 33.35% da variância, tendo um total de 94.34% de informação acumulada nas duas primeiras componentes principais o que fornece uma precisão razoável do espaço de ordem maior.

As componentes principais PC1 e PC2 para o método PM3 são dadas pelas equações:

$$\mathbf{PC1} = -0.70467 \Delta H + 0.05832 \eta H + 0.70713 HD \quad (7)$$

$$\mathbf{PC2} = 0.08337 \Delta H + 0.99652 \eta H - 0.00088 HD \quad (8)$$

A PC1 corresponde a 61.03% da variância e para a PC2 33.34% da variância, tendo um total de 94.37% de informação acumulada nas duas primeiras componentes principais.

As equações 5 e 7 indicam que os descritores de maior importância para a PC1 são ΔH e HD, enquanto que para a PC2 o descritor de maior importância é o ηH . Para os métodos semi-empíricos AM1 e PM3 respectivamente.

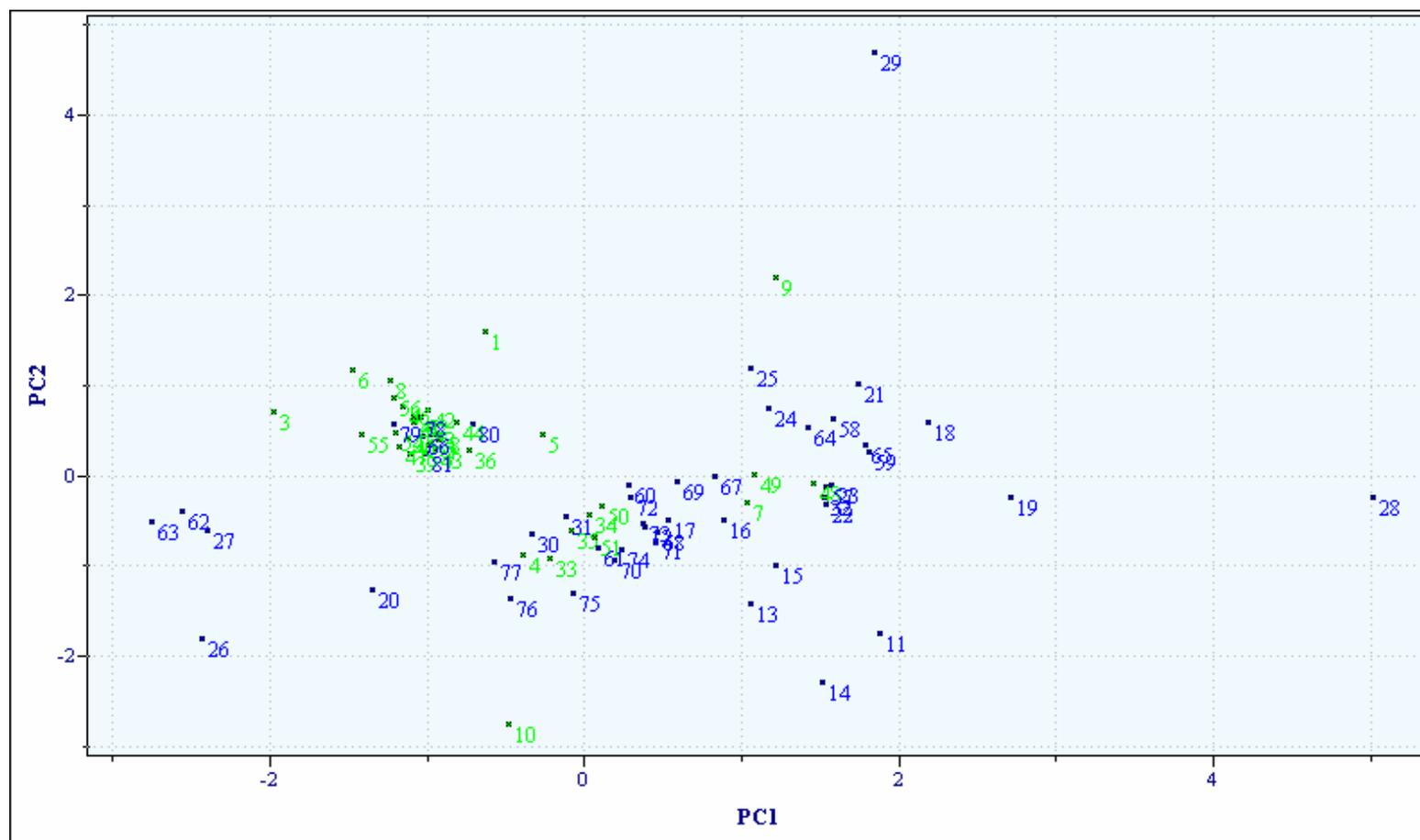


Fig 8: Scores plot das duas primeiras componentes Principais dos 81 PAHs com seleção de variáveis: ΔH , ηH e HD calculadas pelo método PM3. (Azuis) Ativas, (Verdes) inativas

O fato de obtermos esses resultados mostra que ambos os eixos são responsáveis pela separação na análise com PCA.

Na Fig. 9 dos Loadings podemos observar que na PC1 os três descritores encontram-se espalhados por toda a região da componente principal, e que na PC2 o descritor ηH é diferenciado no espaço com respeito aos outros dois parâmetros ΔH e HD.

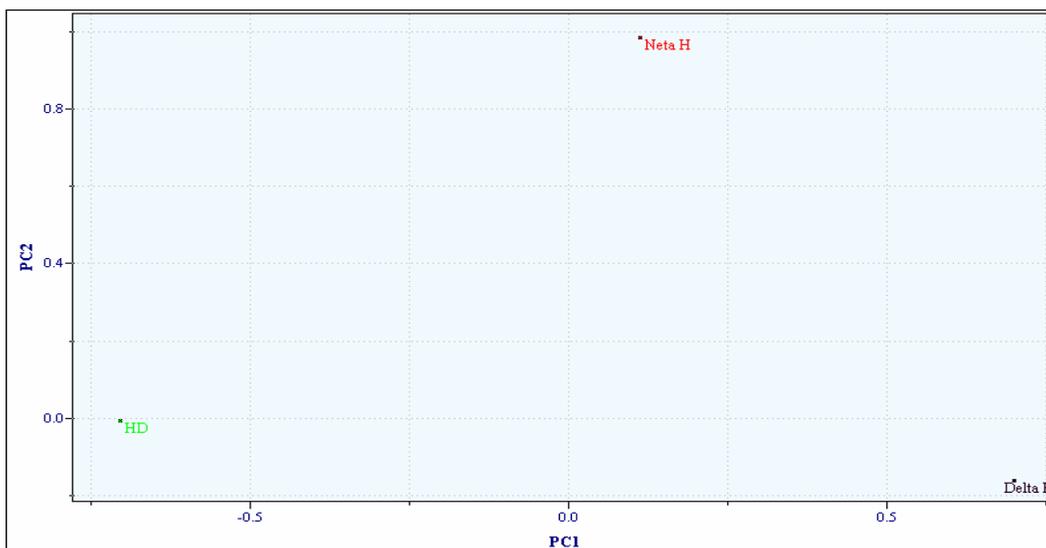


Fig 9: Loading Plot das duas primeiras componentes principais dos 81 PAHs com seleção de variáveis: ηH , ΔH e HD calculadas pelo método AM1 (Vermelhas) Ativas, (Verdes) Inativas.

No reconhecimento de padrões da análise de componentes principais verificamos que os descritores eletrônicos se fazem relevantes para conseguir a separação de dois grupos de moléculas (ativas/inativas), o acerto na classificação dos compostos como ativos e inativos (comparados com os dados experimentais) feita pela análise de componentes principais foi superior a 75%. A PCA identificou as moléculas 79,80 e 81 como potencialmente carcinogênicas.

Observamos também neste caso que não existe uma dependência explícita entre a qualidade dos resultados obtidos e o Hamiltoniano eletrônico utilizado para a PCA.

5.4 Análise Hierárquica de Agrupamentos - HCA:

Os resultados obtidos na análise hierárquica de agrupamentos HCA foram similares aos encontrados na análise com PCA.

Nos visualizamos os resultados em dendrogramas, onde as linhas verticais representam os compostos e as horizontais o valor da similaridade entre pares de compostos, a compostos e a grupos de compostos.

Utilizando as variáveis HOMO, LUMO, ΔH , HD, o agrupamento foi feito com o método incremental, o resultado mostra claramente duas classes com um valor do índice de similaridade (**S**) de zero, o que indica duas classes bem diferenciadas. Observando a aglomeração feita por este método Fig.(10 e 11) e, encontramos que dentro do agrupamento feito para os compostos onde a maior quantidade dos compostos são ativos, encontram-se as moléculas inativas: 20, 26, 27, 31, 62, 63, 66 e 78.

Na segunda classe o agrupamento foi feito com os compostos inativos onde as moléculas com atividade carcinogênica: 7, 9, 33, 34, 35, 48, 49, 50 e 51 foram incorretamente agrupadas.

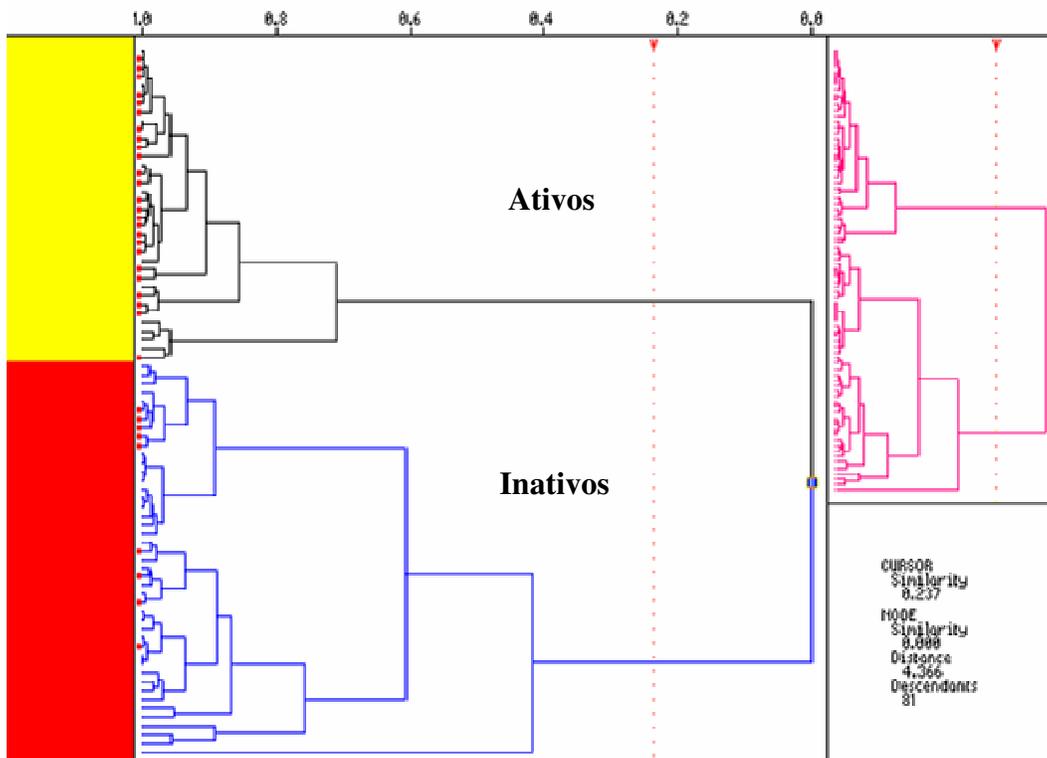


Fig 10: Dendrograma dos 81 PAHs com seleção de variáveis: HOMO, LUMO, ΔH e HD calculadas pelo método AM1 . As cores enfatizam as duas classes (ativos/inativos)

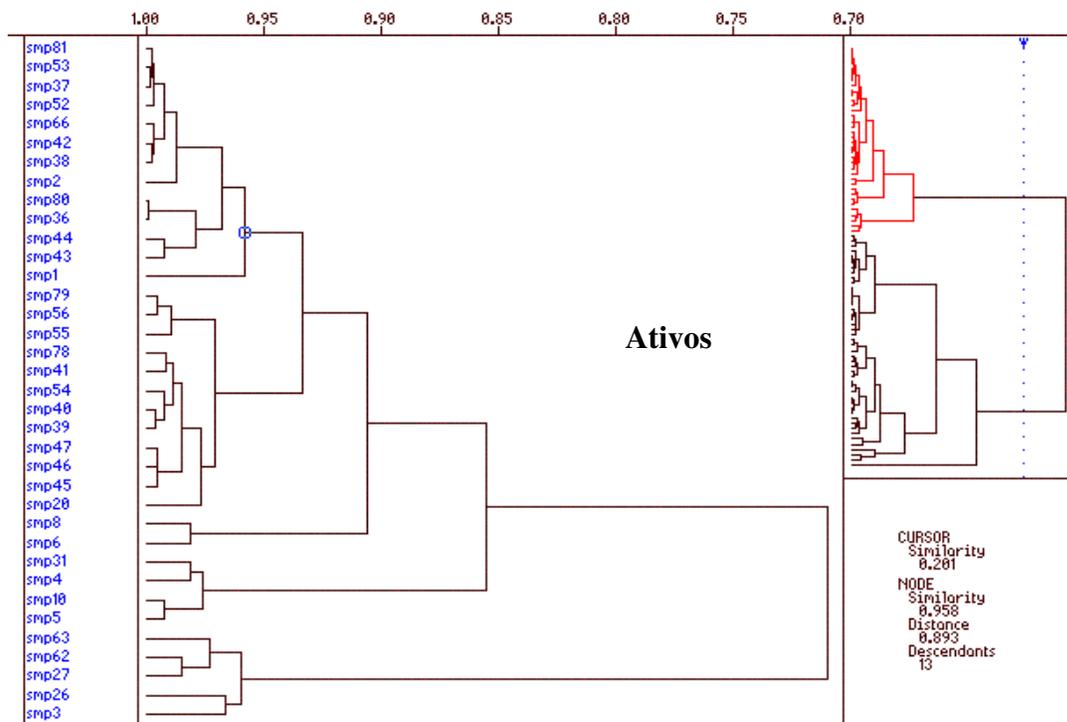


Fig 11: Dendrograma dos 81 PAHs com seleção de variáveis: HOMO, LUMO, ΔH e HD calculadas pelo método AM1 Mostrando os compostos de uma das classes (região vermelha de parte superior direita da figura) formadas pela análise de HCA.

Dos 78 compostos 61 foram corretamente classificados, o acerto foi de 78.2 %, com os compostos 79, 80 e 81 agrupados na classe dos compostos carcinogênicos.

Com as variáveis ηH , ΔH e HD , o agrupamento foi feito também com o método incremental, o resultado mostra claramente duas classes com um valor do índice de similaridade igual a zero, o que indica duas classes bem diferenciadas. Observando a aglomeração feita por este método encontramos que dentro do agrupamento feito para os compostos onde encontramos a maioria dos compostos ativos as moléculas inativas 26, 27, 62, 63, 66 e 78, foram incorretamente agrupadas

Na segunda classe o agrupamento foi feito com os compostos inativos onde as moléculas com atividade carcinogênica: 4, 7, 9, 10, 33, 34, 35, 48, 49, 50 e 51 foram incorretamente agrupadas.

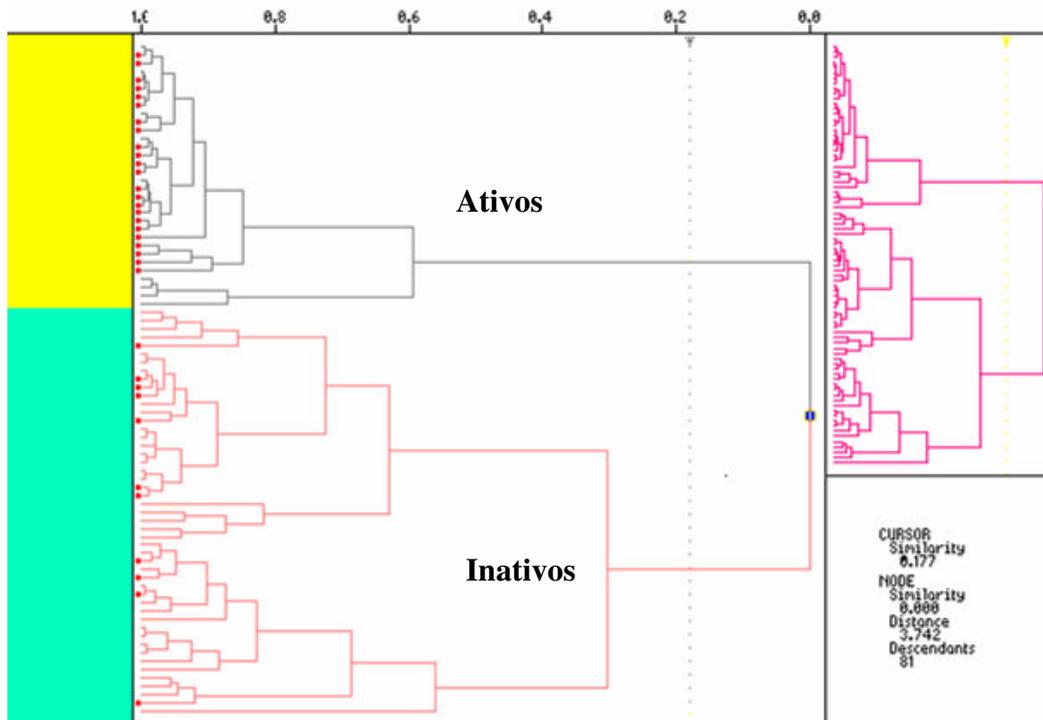


Fig 12: Dendrograma dos 81 PAHs com seleção de variáveis: η H, Δ H e HD calculadas pelo método AM1

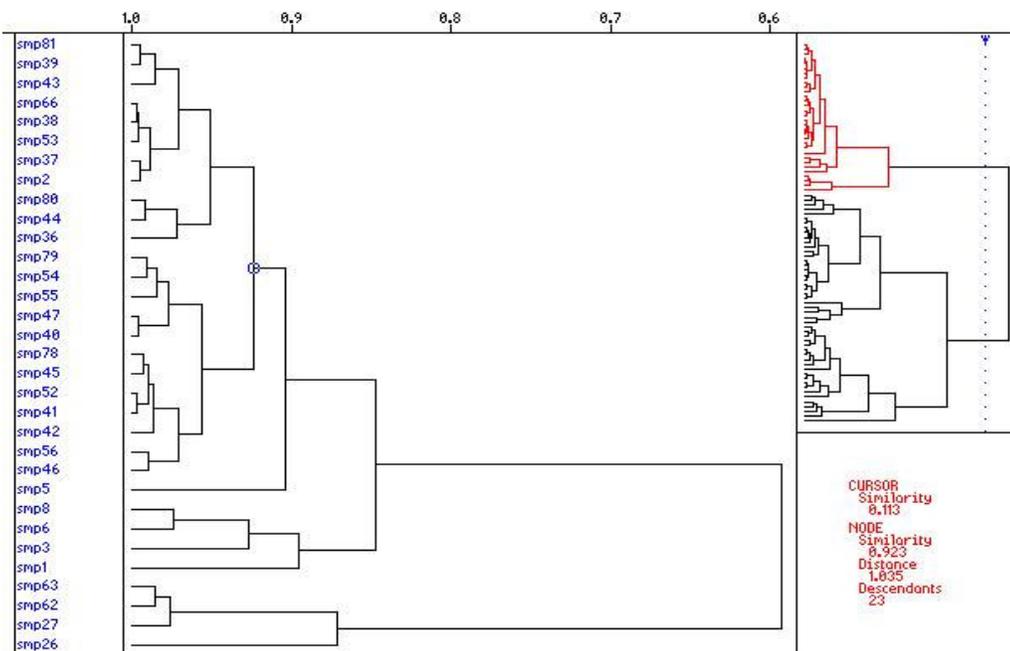


Fig 13: Dendrograma dos 81 PAHs com seleção de variáveis: η H, Δ H e HD calculadas pelo método AM1 Mostrando os compostos de uma das classes (região vermelha de parte superior direita da figura) formadas pela análise de HCA.

Dos 78 compostos 61 foram corretamente classificados, um acerto de 78.2%, com os compostos 79, 80 e 81 agrupados na classe dos compostos carcinogênicos.

5.5 K-ésimo vizinho mais próximo –KNN

Sendo um método de classificação, aplicamos o método KNN considerando a seleção das variáveis feita pela PCA e HCA, escolhemos 40 compostos aleatoriamente para formarem o grupo de treinamento como conjunto de treinamento (tabela.7) dos 78 PAHs com atividade carcinogênica conhecida para a predição da atividade de forma teórica de resto do conjunto. O método fez a classificação utilizando a distribuição no espaço gerado pelos padrões de reconhecimento escolhidos pelos dois anteriores métodos PCA e HCA.

No conjunto de treinamento foi calculada a distância entre um composto de atividade conhecida com o resto dos compostos do conjunto de treinamento. A distância mínima foi selecionada e o composto foi atribuído a uma das classes (Ativo/Inativo). A atribuição a uma classe foi feita por “votos” que por comparação de um ou três K-ésimos vizinhos mais próximos de cada classe são escolhidos para “votar”. Cada um dá um voto para sua classe. A classe que recebe mais votos (menor distância acumulada) ganha a amostra que será classificada como ativa ou inativa.

Obtivemos os seguintes resultados:

A escolha do número de vizinhos mais próximos para classificar cada um dos compostos foi de quatro para os parâmetros HOMO, LUMO ΔH e HD e de três para os parâmetros ηH , ΔH e HD, por serem os números que de forma estatística proporcionam os menores erros na previsões da atividade dos PAHs no conjunto de treinamento, como é indicado na tabela.6.

Tabela 6: Quantidade de erros na classificação da atividade dos PAHs do grupo de treinamento para os parâmetros escolhidos na PCA e HCA em função ao número de vizinhos mais próximos para realizar a classificação (Método semi-empírico AM1):

Votos: K-ésimo vizinho	Erros na previsão para os parâmetros HOMO, LUMO, ΔH e HD	Erros na previsão para os parâmetros ηH , ΔH e HD
1	12	8
2	12	7
3	13	5
4	10	6
5	11	7
6	11	6
7	10	8
8	10	5
9	11	7

Tabela 7: Resultados do treinamento com o método KNN para os dois conjuntos de descritores.

Conjunto de Treinamento	Número de Votos	Número de compostos	Número de compostos que foram corretamente classificados
Descritores: HOMO, LUMO, ΔH, HD			
Classe dos Inativos	4	22	17
Classe dos Ativos	4	18	11
Descritores: ΔH, ηH, HD			
Classe dos Inativos	3	22	18
Classe dos Ativos	3	18	16

Na seguinte tabela podemos identificar os compostos que foram incorretamente classificados no conjunto de treinamento.

Tabela 8: Compostos escolhidos para o conjunto de treinamento, previsão da atividade para os respectivos conjuntos de parâmetros.

Composto	Atividade	Previsão com	Previsão com
	Experimental	as variáveis HOMO, LUMO, ΔH , HD	as variáveis ηH , ΔH e HD
1	1	1	1
4	1	2	1
5	1	2	1
7	1	2	1
8	1	2	1
9	1	1	2
12	2	2	2
13	2	2	2
15	2	2	2
17	2	2	2
18	2	2	2
20	2	2	2
22	2	2	2
24	2	1	1
26	2	2	2
28	2	2	2
29	2	2	2
31	2	1	1
32	2	2	2
33	1	2	1
37	1	1	1
39	1	1	1
41	1	1	1
42	1	1	1
43	1	2	1
46	1	1	1
47	1	1	1
48	1	2	2
52	1	1	1
55	1	1	1
56	1	1	1
57	2	1	1
59	2	2	2
62	2	2	2
64	2	1	2
69	2	2	2
74	2	2	2
76	2	2	2
77	2	2	2
78	2	1	1

(1: Composto ativo), (2: composto Inativo).

Depois de obter nosso conjunto de treinamento e determinar o número de vizinhos que proporcionam o maior acerto na classificação, aplicamos o KNN ao resto do conjunto de PAHs (tabela:8), obtendo os seguintes resultados:

Tabela 9: Classificação dos PAHs a partir do conjunto de treinamento escolhido.

Composto	Atividade Experimental	Classificação HOMO, LUMO, ΔH , HD	Classificação ηH , ΔH e HD
2	1	1	1
3	1	2	1
6	1	1	1
10	1	1	2
11	2	2	2
14	2	2	2
16	2	1	2
19	2	2	2
21	2	2	2
23	2	2	2
25	2	1	2
27	2	2	2
30	2	2	2
34	1	1	2
35	1	1	2
36	1	1	1
38	1	1	1
40	1	1	1
44	1	1	1
45	1	1	1
49	1	2	2
50	1	2	2
51	1	2	2
53	1	1	1
54	1	1	1
58	2	2	2
60	2	2	2
61	2	2	2
63	2	2	2
65	2	2	2
66	2	1	1
67	2	2	2
68	2	2	2
70	2	2	2
71	2	2	2
72	2	2	2
73	2	2	2
75	2	2	2
79	ND	1	1
80	ND	1	1
81	ND	1	1

As células com sombra indicam os compostos que foram incorretamente classificados e onde, os compostos 79, 80 e 81 foram classificados como compostos potencialmente carcinogênicos. O acerto global (conjunto de treinamento + conjunto de predição) do KNN foi de 78,2% com as variáveis HOMO, LUMO, ΔH e HD e de 83.3% com as variáveis ηH , ΔH e HD.

5.6 Soft Independent Modeling of Class Analogies-SIMCA

A construção do modelo de classificação SIMCA utilizou um conjunto de treinamento de 40 compostos aleatoriamente escolhidos do conjunto de 78 PAHs estudados, este conjunto é o mesmo que foi utilizado pelo método KNN

A primeira análise feita com o método SIMCA foi realizada usando os quatro descritores HOMO, LUMO, ΔH e HD. Nos encontramos o número ótimo de PCs (Componentes Principais) para cada classe usando a validação cruzada “leave-one-out”. Determinando dessa maneira que o número de PCs foi de 3 e 2, para a classe dos ativos e inativos, respectivamente

A segunda análise com o método SIMCA considerou os descritores ΔH , ηH e HD, e depois de validar o nosso conjunto de treinamento usando a validação cruzada “leave-one-out” chegamos a determinar que duas PCs e o número ideal para cada uma das duas classes.

A tabela 9 mostra o resultado do treinamento usando as respectivas componentes principais para cada categoria y cada conjunto de descritores.

Tabela 9: Resultados do treinamento com o método SIMCA para os dois conjuntos de descritores.

Conjunto de Treinamento	Número de PCs	Número de compostos	Número de compostos que foram corretamente classificados
Descritores: HOMO, LUMO, ΔH, HD			
Classe dos Inativos	2	22	15
Classe dos Ativos	3	18	8
Descritores: ΔH, ηH, HD			
Classe dos Inativos	2	22	18
Classe dos Ativos	2	18	14

Na tabela 10, são apresentados os compostos escolhidos para o grupo de treinamento e os erros na classificação na validação do método

Tabela 9: Classificação dos PAHs a partir do conjunto de treinamento escolhido,

Composto	Atividade Experimental	Classificação	
		HOMO, LUMO, ΔH , HD	ηH , ΔH e HD
2	1	2	1
3	1	2	1
6	1	1	1
10	1	1	2
11	2	2	2
14	2	2	2
16	2	2	2
19	2	2	1
21	2	2	2
23	2	2	1
25	2	2	2
27	2	2	2
30	2	2	2
34	1	1	2
35	1	1	2
36	1	2	1
38	1	2	1
40	1	2	1
44	1	2	1
45	1	1	1
49	1	2	2
50	1	1	2
51	1	1	2
53	1	2	1
54	1	2	1
58	2	1	2
60	2	2	2
61	2	1	2
63	2	1	2
65	2	2	1
66	2	2	1
67	2	2	2
68	2	2	2
70	2	2	2
71	2	1	2
72	2	2	2
73	2	1	2
75	2	2	2
79	ND	1	1
80	ND	1	1
81	ND	2	1

Com o conjunto dos parâmetros: HOMO, LUMO, ΔH e HD o acerto foi de 24 moléculas para um conjunto de 38 compostos, o que significa uma porcentagem de acerto de 63.2%. Com o conjunto dos três parâmetros: ΔH , ηH e HD o acerto na

classificação dos 38 compostos foi de 28 moléculas o que significa um acerto percentual de 73,7%.

Com a finalidade de observar a classificação do conjunto dos 81 compostos nos aplicamos o método SIMCA com o mesmo grupo de treinamento nos 81 PAHs e a classificação que o método deu foi a seguinte:

Tabela 10: Classificação da atividade dos 81 PAHs feito pelo método SIMCA.

Composto	Atividade Experimental	Classificação HOMO, LUMO, ΔH , HD	Classificação ηH , ΔH e HD
1	1	2	1
2	1	2	1
3	1	2	1
4	1	2	2
5	1	2	1
6	1	1	1
7	1	1	2
8	1	1	1
9	1	2	0
10	1	1	0
11	2	2	1
12	2	2	2
13	2	2	1
14	2	2	1
15	2	2	1
16	2	2	2
17	2	2	1
18	2	2	2
19	2	2	2
20	2	2	2
21	2	2	2
22	2	1	2
23	2	2	2
24	2	1	2
25	2	2	2
26	2	1	2
27	2	0	2
28	2	2	0
29	2	2	0
30	2	2	2
31	2	1	2
32	2	2	2
33	1	1	2
34	1	1	2
35	1	1	2
36	1	2	1
37	1	2	1
38	1	2	1
39	1	2	1
40	1	2	1
41	1	1	1

42	1	2	1
43	1	2	1
44	1	2	1
45	1	1	1
46	1	1	1
47	1	1	1
48	1	1	2
49	1	2	2
50	1	1	2
51	1	1	2
52	1	2	1
53	1	2	1
54	1	2	1
55	1	1	1
56	1	2	1
57	2	1	2
58	2	1	2
59	2	2	2
60	2	2	2
61	2	1	2
62	2	2	2
63	2	1	2
64	2	1	2
65	2	2	2
66	2	2	1
67	2	2	2
68	2	2	2
69	2	1	2
70	2	2	2
71	2	1	2
72	2	2	2
73	2	1	2
74	2	2	2
75	2	2	2
76	2	2	2
77	2	2	2
78	2	2	1
79	ND	1	1
80	ND	2	1
81	ND	2	1

No caso das variáveis HOMO, LUMO, ΔH e HD a classificação das 78 moléculas com atividade conhecida teve um acerto de 60.2% (47/78). Com o conjunto dos três parâmetros: ΔH , ηH e HD dos 78 compostos só 13 foram incorretamente classificados o que significa um acerto percentual de 83,3%.

Com a utilização do método SIMCA também podemos observar o poder de modelagem e discriminação das variáveis na construção do modelo de predição, o que podemos observar é que os descritores da MIE são importantes (pelos valores

próximos de um), e de forma significativa o ΔH , que mostra ser uma das variáveis altamente relevantes na modelagem, e principalmente na discriminação do modelo feito pelo método SIMCA. A variável ηH possui um valor médio em ambos os casos.

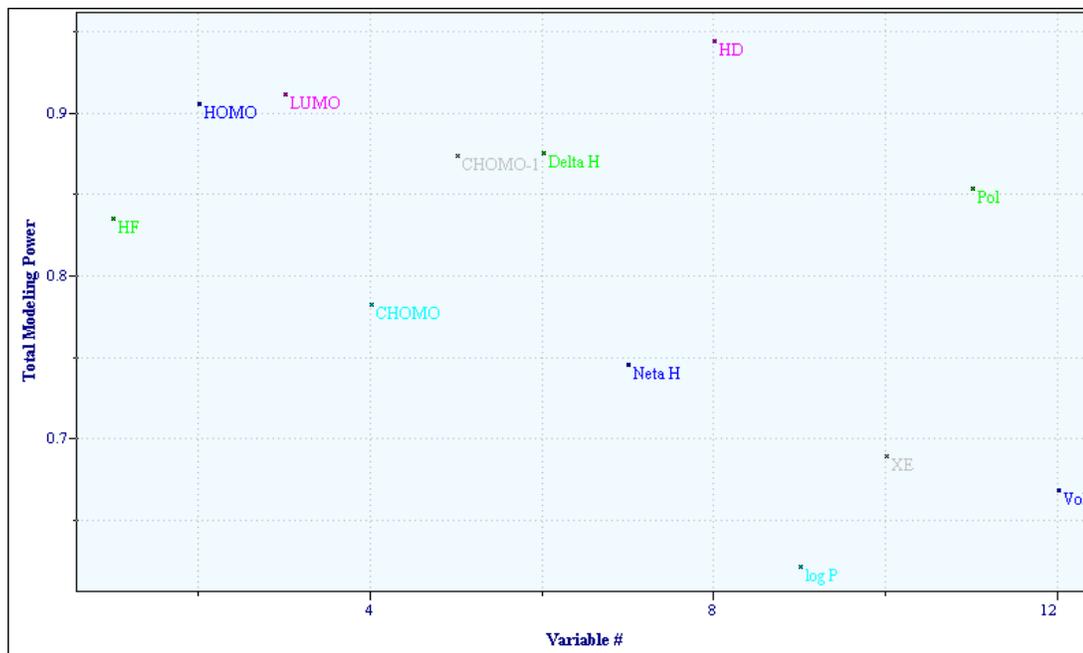


Fig 12: SIMCA poder de modelagem das variáveis

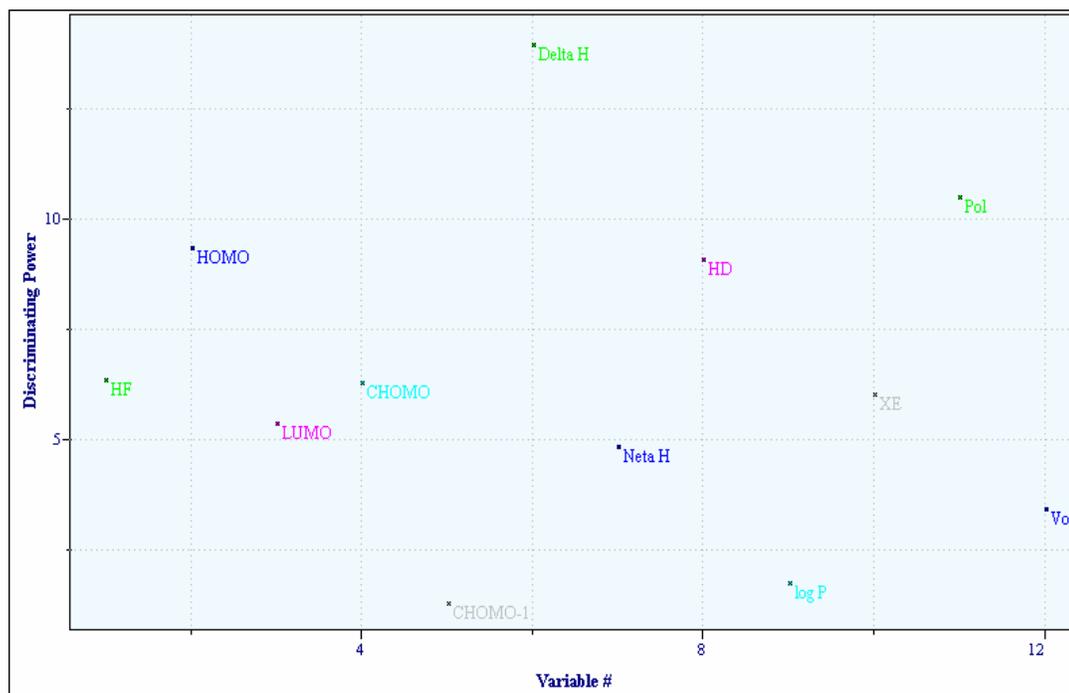


Fig 13: SIMCA poder de discriminação das variáveis.

Apesar do mencionado anteriormente, o que foi mostrado no desenvolvimento do trabalho é que o uso dos dois parâmetros de forma conjunta proporcionam um modelo superior na classificação dos PAHs o que não é observado no caso do modelo feito com as quatro variáveis HOMO, LUMO, ΔH e HD, que são escolhidas com o critério de valores perto de um, no poder de discriminação e modelagem.

5.7 Redes Neurais

A aplicação de redes neurais no estudo da relação estrutura/atividade pode nos levar a visualizar os padrões estabelecidos pelos dados experimentais de atividade biológica, esta parte consiste no treinamento da rede escolhendo os descritores, e também realizar a predição. Como nosso objetivo é o de observar se a rede escolhe os parâmetros ηH e ΔH como padrões no grupo de treinamento nos escolhemos aleatoriamente 40 compostos ativos e inativos e com este grupo nos escolhemos vários parâmetros físico-químicos para treinar a rede (determinar os parâmetros da rede) e conseguir a melhor percentagem de acerto frente à classificação experimental de atividade carcinogênica com os métodos semi-empíricos utilizados no trabalho.

Tabela 11: Conjunto de treinamento utilizado para o treinamento da rede

Conjunto de treinamento Compostos	Atividade Carcinogênica (dado experimental)
1	1
4	1
5	1
7	1
9	1
33	1
37	1
39	1
41	1
42	1

43	1
46	1
47	1
48	1
52	1
55	1
56	1
12	0
13	0
15	0
17	0
18	0
20	0
22	0
24	0
26	0
28	0
29	0
31	0
32	0
57	0
59	0
62	0
64	0
69	0
74	0
76	0
77	0
78	0

(1 = composto ativo, 0 = composto inativo)

O treinamento da rede consiste em variar os valores dos parâmetros de **ALPHA** e **HALPHA**, da não linearidade da função sigmoideal da segunda e terceira camada, de **HDNLAYER** que é o numero de neurônios da segunda camada, de **EPSIRON** e **HEPSIRON** que é o fator de troca dos pesos da segunda e terceira camada e a **bias**

que é o termo independente da função sigmoïdal da primeira camada, até obter os melhores resultados no treinamento e predição do conjunto de moléculas.

Os valores ótimos encontrados no treinamento da rede estão indicados na tabela 12, as duas últimas linhas da tabela contem os erros no treinamento da rede e na predição.

As colunas mostram os conjuntos dos descritores utilizados, testando sempre os da MIE.

Tabela 12: Parâmetros de treinamento da rede neural para os quatro grupos de descritores selecionados

Conjunto	HOMO, LUMO, ΔH, HD	ΔH, ηH, HD	ΔH, ηH, Log P	ΔH, ηH,
ALPHA	2.5	2.6	3.5	4.0
HALPHA	2.0	2.6	3.5	3.5
HDNLAYER	8	6	6	4
EPSIRON	0.4	0.2	0.35	0.2
HEPSIRON	0.4	0.2	0.35	0.2
BIAS	Não	Não	Não	Não
Número de iterações	30.000	30.000	30.000	30.000
Precisão	10^{-5}	10^{-5}	10^{-5}	10^{-5}
Erros no treinamento	1	2	2	3
Erros na predição	5	4	7	4

Os resultados mostram que as redes neurais reconhecem como padrão de classificação os descritores da MIE, a tabela 13 mostra que a classificação teve um acerto, frente aos dados experimentais superior ao 85% em todos os casos.

Tabela 13: Percentagem de acerto da rede neural para os diferentes conjuntos de descritores

Conjunto	HOMO, LUMO, ΔH , HD	ΔH , ηH , HD	ΔH , ηH , Log P	ΔH , ηH ,
% de acerto	92.3 %	92.3 %	88.5 %	91.0 %

5.8 Resumo dos Resultados

Para ter uma visão global dos resultados e testes feitos no trabalho de tese apresentamos em uma tabela a percentagem de acerto das

	AM1	PM3	PM5
MIE ($\eta H, \Delta H$)	85,9 %	87,2 %	85,9 %
PCA ($\eta H, \Delta H, HD$)	83,3 %	83,3 %	82,1 %
HCA ($\eta H, \Delta H, HD$)	78,2 %	78,2 %	78,2 %
SIMCA ($\eta H, \Delta H, HD$)	83,3 %	X	X
KNN ($\eta H, \Delta H, HD$)	83,3 %	X	X
RN ($\eta H, \Delta H, HD$)	92,3 %	92,3 %	X

Referencias

¹ M. Cooke and A. Dennis, *Polynuclear Aromatic Hydrocarbons: Chemistry, Characterization and Carcinogenesis, Ninth International Symposium* (1984).

² G. Ronald and Harvey, *Polycyclic Aromatic Hydrocarbons* (Wiley-VCH- USA, 1997).

³ *Pirouette Multivariate Data Analysis for IBM-PC Systems*, Version 2.0, Infometrix: Seattle, WA, 1996.

Capítulo 6

Conclusões

Neste trabalho apresentamos uma investigação teórica da atividade biológica de um conjunto de 78 hidrocarbonetos policíclicos aromáticos de atividade carcinogênica conhecida experimentalmente, e de três hidrocarbonetos policíclicos aromáticos não testados experimentalmente, e que os estudos feitos nos levam a sugerir que são potencialmente carcinogênicos.

Inicialmente obtivemos as geometrias de menor calor de formação e de mínimo global para as moléculas que apresentam diedrais flexíveis. Este ponto é fundamental porque as propriedades físico-químicas obtidas a partir da geometria obtida são utilizadas para avaliar e estudar as relações de estrutura atividade dos compostos estudados.

Nossa pesquisa mostra que é possível descrever a atividade biológica dos hidrocarbonetos policíclicos aromáticos utilizando somente os descritores quânticos da Metodologia dos Índices Eletrônicos (MIE).

Esta Metodologia foi introduzida recentemente por Barone durante a investigação da potencia carcinogênica de 26 hidrocarbonetos policíclicos aromáticos.

Esta metodologia baseada no conceito de densidade de estados (DOS), e da densidade local de estados (LDOS), permite investigar regiões moleculares específicas separadamente, observando sua importância na formação de um estado.

A MIE utiliza apenas dois descritores quânticos: Δ (é uma medida da diferença em energia entre dois estados de fronteira) e η (é uma diferença na LDOS de uma região molecular para a formação dos orbitais referentes aos estados observados em Δ). Estes dois parâmetros nos permitem obter regras de seleção da atividade dos compostos. As regras estão relacionadas com valores críticos de Δ e η . Estes valores por outro lado são obtidos como sendo valores limites entre compostos ativos e inativos do ponto de vista experimental.

Os testes mais detalhados realizados sobre a metodologia mostram com os PAHs, mostram a capacidade de seleção e classificação dos parâmetros Δ e η quando a LDOS é feita sobre o anel que contém a maior ordem de ligação da molécula nos estados HOMO e HOMO-1 o acerto na classificação foi de 84,6% .

Na determinação da dependência explícita entre a qualidade dos resultados obtidos e o hamiltoniano eletrônico utilizado com os métodos semi-empíricos PM3 (Parametric Method 3), PM5 (Parametric Method 5), AM1 (Austin Method One) para realizar os testes comparativos, observamos uma pequena variação nos valores críticos determinados pelos diferentes métodos, mas podemos indicar que não existe uma dependência explícita entre a qualidade dos resultados e o método semi-empírico utilizado.

No estudo da estrutura-atividade as metodologias utilizadas frequentemente como Análise de Componentes Principais (PCA - Principal Component Analysis), Análise Hierárquica de Agrupamentos (HCA - Hierarchical Cluster Analysis), K-ésimo vizinho mais próximo (KNN), Soft Independent Modeling of Class Analogies (SIMCA) e as Redes Neurais (RN - Neural Networks), permitem facilitar e economizar o demorado processo de otimização de atividade biológica, desenvolvimento de novas drogas e de predicação da atividade biológica de um composto. A diferença fundamental observada dentre as metodologias mencionadas e a MIE é a utilização de vários parâmetros contendo informações físico-químicas, estereoquímicas e eletrônicas com o objetivo de separar os compostos em grupos distintos (ativos e inativos) escolhendo os parâmetros que são responsáveis da separação observada.

Na análise de PCA, a seleção dentre um total de 22 descritores dos seguintes descritores eletrônicos HOMO, LUMO, ΔH , HD, com uma variância na primeira componente principal de 92,64%, permitiu a separação dos compostos ativos dos inativos com um acerto na discriminação de 78,2%. Usando os mesmos critérios na seleção das variáveis encontramos que a seleção dos descritores eletrônicos ηH , ΔH , HD, com uma variância na primeira componente principal de 60,99%, permite uma separação dos compostos ativos e inativos com um acerto na discriminação de 83,3%.

A análise HCA mostrou uma separação de duas classes com índice de similaridade zero para as moléculas distribuídas no espaço dimensional gerado pelas variáveis ηH , ΔH , HD, com um acerto na discriminação dos compostos ativos dos inativos de 78,2% e no caso das variáveis HOMO, LUMO, ΔH , HD, o acerto foi o mesmo.

Nos métodos de classificação SIMCA e KNN a classificação feita usando os descritores HOMO, LUMO, ΔH , HD mostrou um acerto de 60,2% e 78,2% respectivamente. No caso da classificação feita usando os descritores ηH , ΔH , HD, o acerto foi de 83,3% para ambos os casos.

Observando esses resultados podemos concluir que o uso combinado das metodologias, selecionam de forma natural parâmetros eletrônicos que permitiram a separação visual dos compostos estudados (ativos e inativos) e permitiu também identificarmos os parâmetros eletrônicos que possam ser correlacionados com a atividade biológica. É importante observar que os testes de relevância estatística, mostraram que Δ , η são importantes na discriminação e classificação dos PAHs. E que o índice de acerto é melhorado com o uso dos índices da MIE, o que nos permite estabelecer a confiabilidade desses descritores.

No resultado obtido com redes neurais do tipo *Perceptrom*, que tem como objetivo o reconhecimento de padrões, nos reforçamos o mencionado anteriormente porque a utilização dos parâmetros da MIE permitem que a rede consiga classificar os PAHs com um acerto de 92%.

Para finalizar nós podemos concluir que MIE, sendo uma metodologia que envolve apenas dois parâmetros, pode ser utilizada como uma ferramenta confiável na predição e seleção dos hidrocarbonetos policíclicos aromáticos.