Célio Fernando Lipinski

Estudo de Relações Quantitativas Estrutura-Atividade de chalconas
análogas à combretastatina A4

Dissertação apresentada ao Instituto de
Química de São Carlos como parte dos
requisitos para obtenção do título de Mestre
em Ciências.

Área de concentração: Físico-Química.

Orientador: Prof. Dr. Albérico Borges Ferreira
da Silva.

São Carlos
2014

Às duas mulheres que me aguardam, dia após dia.

AGRADECIMENTOS

Aos meus pais, Luiz e Terezinha Lipinski. Meu pai, que me mostrou que deformar as mãos de tanto trabalhar é também uma forma de demonstração de amor pelos filhos. Minha amada mãe, que criou seus sete filhos como sete mães não o fariam.

À moça que será minha esposa e mãe dos meus filhos, Bernadete Okopna. A única maneira que vejo de te agradecer por me esperar todos esses anos é permanecendo até o último dia da minha vida ao seu lado.

À toda a Lipinskada, família que tanto gosto. Não houve um único domingo aqui, que deixei de lembrar e imaginar vocês aí reunidos almoçando.

Ao professor Albérico que, como se não bastasse contribuir com todo o seu conhecimento científico, compartilha conosco sua experiência de homem vivido, ao lado do qual, cada dia que se passa é um novo aprendizado.

Ao pessoal da informática, por sempre nos atenderem com boa vontade e pelos ótimos trabalhos realizados.

A todos os colegas do grupo, em especial ao Gláucio (o 'Fino'), pelo incentivo inicial, e principalmente à dona Aline, sem a qual, este trabalho não teria o mesmo êxito.

Ao CNPq, pelo auxílio financeiro.

GURI

(César Passarinho)

Das roupas velhas do pai
queria que a mãe fizesse
uma mala de garupa, uma bombacha
e me desse.

Queria boinas, alpargatas
e um cachorro companheiro
Pra me ajudar a botar as vacas
no meu petiço sogueiro.

Levar comigo a tabuada
e o meu livro - queres ler?
Vou aprender a fazer conta
e algum bilhete escrever
Pra que a filha do sr. Bento saiba
que ela é o meu bem querer.
E se não for por escrito
eu não me animo a dizer

Quero gaita de oito baixos
pra ver o ronco que sai,
botas feitio do Alegrete,
esporas do Ibirocaí.
Lenço vermelho e guaiaca
compradas lá no Uruguai,
para que digam quando eu passe
- Saiu igualzito ao pai.

E se Deus não achar muito
tanta coisa que pedi,
não deixe que eu me separe
deste rancho onde nasci.
Nem me desperte tão cedo
do meu sonho de guri,
e de lambuja permita
que eu nunca saia daqui.

RESUMO

A combretastatina A4 é um promissor agente anticâncer. Na célula, inibe a polimerização dos microtúbulos, os quais são fundamentais nos processos de motilidade, manutenção estrutural e mitose. Essa inibição se dá a partir do sítio de interação da $\alpha\beta$ -tubulina bloqueando o fluxo do sangue que alimenta os tumores, o que resulta na morte dos mesmos. Com estrutura semelhante às combretastatinas, as chalconas constituem uma classe de compostos que atuam no mesmo sítio de interação na tubulina. Baseando-se nos trabalhos experimentais de Ducki e colaboradores, estudou-se a estrutura molecular de 87 chalconas análogas à combretastatina A4 por meio do método quântico DFT com o propósito de desenvolver modelos de Relações Quantitativas Estrutura-Atividade (QSAR) aplicados a tais antagonistas. A partir dos métodos dos Mínimos Quadrados Parciais (PLS) e de Redes Neurais Artificiais (ANN), foram gerados modelos que conduzem à elucidação da relação dos compostos estudados com suas respectivas atividades biológicas. Os descritores eletrônicos e moleculares selecionados apresentam alta concordância com as características das moléculas, havendo predominância de comportamento linear com a atividade biológica, podendo, eventualmente, apresentar comportamento não-linear, o que torna o modelo gerado altamente consistente.

Palavras-chave: Câncer, microtúbulos, chalconas, QSAR, PLS, ANN.

ABSTRACT

Combretastatin A4 is a promising anticancer agent. It inhibits the polymerization of microtubules in the cell, which are essential in the process of motility, structural maintenance and mitosis. This inhibition is given from the interaction site of $\alpha\beta$ -tubulin blocking the blood flow that feeds the tumor, what results in its death. The chalcones, sharing a similar structure of the combretastatin, are also a class of compounds that act in the same site of interaction in the tubulin. Based on the experimental work of Ducki and co-workers, we proposed a molecular structure study of 87 chalcones similar to combretastatin A4 using the DFT method in order to develop Quantitative Structure-Activity Relationships (QSAR) applied to the given antagonists. Through Partial Least Squares (PLS) and Artificial Neural Network (ANN) methods, some models has been generated to lead the understanding on the relationship between the compounds studied and their respective biological activities. The electronic and molecular descriptors selected have high correlation with the molecule features, being linear most of the time, although with eventual non-linear behavior, which makes the generated model highly consistent.

Keywords: Cancer, microtubules, chalcones, QSAR, PLS, ANN.

SUMÁRIO

1 INTRODUÇÃO	8
1.1 Câncer	8
1.2 Teoria do Funcional da Densidade (DFT).....	12
1.3 Relações Quantitativas Estrutura-Atividade (QSAR)	25
1.3.1 Algoritmo Genético	27
1.3.2 Mínimos Quadrados Parciais (PLS).....	29
1.3.3 Redes Neurais Artificiais.....	31
2 OBJETIVOS.....	34
3 METODOLOGIA	35
4 RESULTADOS.....	42
4.1 Seleção de descritores	42
4.2 Análises de PLS.....	43
4.3 Análises de ANN.....	60
4.4 Análise dos descritores selecionados	65
4.4.1 Orbital HOMO-3.....	65
4.4.2 Descritores RDF.....	67
4.4.3 Descritor SP02.....	69
4.4.4 Descritor RTv.....	71
4.4.5 Descritor UNIP.....	71
4.4.6 Polarizabilidade Isotrópica.....	72
5 CONCLUSÃO E PERSPECTIVAS	73
REFERÊNCIAS.....	74

1 INTRODUÇÃO

1.1 Câncer

Câncer ou neoplasia é a denominação de um conjunto de mais de 100 tipos de diferentes doenças que se caracterizam por um descontrole no processo de mitose no organismo, a partir do qual, a célula passa a se reproduzir indiscriminadamente, não respondendo mais aos mecanismos de controle fisiológicos, passando a se alimentar dos nutrientes que eram destinados às, até então, células sadias. Esse descontrole se deve a alguma alteração (de natureza desconhecida) que ocorre no DNA celular. Sua origem se dá por condições multifatoriais, que podem agir em conjunto ou em sequência para iniciar ou promover o câncer.^{1,2}

Segundo estimativas do projeto *GloboCan 2012*, realizado pela Organização Mundial da Saúde (OMS), houveram 14,1 milhões de casos novos de câncer, que resultaram em 8,2 milhões de mortes em todo o mundo, no ano de 2012. No Brasil, para o ano de 2014, foi estimada a ocorrência de aproximadamente 576.000 novos casos.³

A diferenciação entre os tipos de câncer se dá através do tipo de célula afetada. Como consequência disso, uma série de procedimentos e combinações de métodos são necessários para o tratamento dos diversos tipos de câncer, dependendo do seu tamanho, estágio e localização.

Atuando através dos mais diversos e complexos mecanismos de ação, uma grande quantidade de fármacos encontra-se disponível. Porém, fatores como elevada toxidez e resistência, causam grande limitação quanto ao uso desses medicamentos, o que motiva a pesquisa por novos fármacos com propriedades otimizadas, que possam representar novas alternativas terapêuticas.

O processo de divisão celular pode ser resumidamente descrito iniciando no estágio G1 (gap1), no qual são emitidos sinais para que o processo continue ou seja interrompido. A partir dessas manifestações por determinadas proteínas, a célula atesta que está apta para passar à etapa seguinte, a fase S (síntese), onde o DNA é duplicado. Como garantia de não haverem danos de duplicação, a célula passa pela fase G2 (gap2), onde é feita a análise de todo o material genético. Caso a cópia não seja exata, são liberadas proteínas supressoras que bloqueiam a passagem ao estágio final, a mitose, fase em que duas células idênticas são formadas.^{4,5,6}

Células defeituosas podem surgir do processo de divisão celular quando as etapas de verificação não ocorrem. E, se isso vir a acontecer sistematicamente, essas células passam a se reproduzir sem que o organismo necessite, resultando numa estrutura massiva composta de células aberrantes - o tumor. Por isso, se faz necessária a utilização de substâncias que interrompam tal processo de mitose.

A maior parte das substâncias identificadas com propriedades antimitóticas interage com os componentes dos microtúbulos, os quais, no processo de divisão celular, são responsáveis por diversos tipos de interações celulares, como absorção de alimentos, descarte de organelas e a separação dos cromossomos destinados à célula-filha durante o fuso mitótico. Os microtúbulos são formados pela polimerização da proteína tubulina. E essa é formada por dois heterodímeros: α -tubulina e β -tubulina.^{7,8}

Figura 1 – Estrutura dos microtúbulos.

Fonte: WILSON, L.; JORDAN, M. A., 2004, p. 252-265.

Se suprimida a dinâmica dos microtúbulos, a formação do eixo mitótico fica comprometida, desencadeando mecanismos que resultam em apoptose. Compostos que interferem nessa dinâmica estão entre os agentes anticâncer mais importantes na terapêutica. Os que interagem diretamente com a tubulina são denominados agentes antitubulina.⁹

As drogas que têm como alvo os microtúbulos despertam há muito a atenção da indústria farmacêutica. Essas drogas atuam em sítios específicos da tubulina. Os três mais importantes são: o dos taxóides, o dos alcalóides da vinca e o da colchicina.

Todos eles possuem a característica de atuarem interrompendo o ciclo de polimerização da tubulina.

Muitos compostos apresentam pouca eficácia em células que não estão se dividindo quando a droga alcança a seu sítio terapêutico. E, como consequência, muitas dessas células podem sobreviver e adquirir resistência à droga. Nesse caso, a interrupção do suprimento do sangue dos tumores é uma alternativa conveniente, já que, impedindo a polimerização dos microtúbulos, interrompe-se o fluxo de sangue da célula cancerosa, resultando na morte da mesma. O diferencial da colchicina em relação aos taxóides e os alcalóides da vinca, é o de agir também nessa vasculatura que irriga o tumor.¹⁰

Produto natural com atividade antiproliferativa, a colchicina foi o primeiro agente desestabilizador de microtúbulos a ser identificado. Apesar de sua importância histórica e de apresentar significativa propriedade citostática, a colchicina não está incluída entre os fármacos comercialmente disponíveis em razão da sua elevada neurotoxicidade, já que essa droga possui índice terapêutico baixo, o que a leva a ser administrada em concentrações próximas dos limites aceitáveis para o organismo.^{11,12}

Uma alternativa para os esforços intensivos direcionados à identificação de novos agentes antitubulina é selecionar conjuntos de novos compostos estruturalmente relacionados que possam atuar no mesmo sítio de interação da colchicina. Entre os mais estudados, encontram-se as combretastatinas e seus derivados.

As combretastatinas são pequenas moléculas orgânicas, cuja estrutura básica é constituída por dois anéis benzênicos polioxigenados, ligados por uma ponte de etileno.⁷ Trata-se de um grupo de agentes antimitóticos isolados a partir da casca da árvore sul-africana *Combretum caffrum*. Também podem ser encontradas na árvore *Combretum leprosum*, popularmente conhecida como Mufumbo, cuja ocorrência se dá na caatinga e pantanal brasileiro.¹³⁻¹⁵

Estudos etnofarmacológicos envolvendo tanto a *Combretum Leprosum* quanto a *Combretum Caffrum*, apontam diversas finalidades para o uso medicinal dessas plantas, entre elas: tratamento de tumores, contenção de hemorragias, cicatrizante, analgésico e antiinflamatório.¹⁶

Desde o início da década de 80, Pettit e colaboradores, isolam, sintetizam e avaliam diversos tipos de combretastatinas, verificando suas elevadas atividades antimitóticas, e consequentemente, suas aplicações no tratamento do câncer.¹⁷⁻²²

Nos últimos anos, um tipo de combretastatina vem sendo estudada e apontada com grande potencial anticâncer, a combretastatina A4 (CA4).²³⁻²⁷

Figura 2 – Estrutura da combretastatina A4.

Ligantes do sítio da colchicina, como as combretastatinas, são estruturalmente mais simples que os demais agentes antitubulina. Devido a essas propriedades químicas e biológicas, o estudo de ligantes do sítio da colchicina tem sido bastante recomendado, e um grande número de pequenas moléculas sintéticas ligantes desse sítio tem sido identificado. As chalconas que serão objeto de estudo neste trabalho, estão entre elas.

As chalconas inibem a polimerização dos microtúbulos a partir do mesmo sítio de interação que o alcalóide colchicina, a β -tubulina, conforme figura 3:

Figura 3 – Interação da colchicina com a β -tubulina.

Fonte: WILSON, L.; JORDAN, M. A., 2004, p. 252-265.

O termo chalcona caracteriza uma família de moléculas de cadeia aberta que contêm dois anéis aromáticos ligados por um fragmento enona de três carbonos, ou

seja, são cetonas α,β -insaturadas, em que um anel aromático está diretamente ligado à carbonila e o outro ao carbono β da função olefínica, conforme figura 4. As chalconas são precursoras na via biossintética dos flavonóides.²⁸

Figura 4 – Estrutura base das chalconas.

Muitas pesquisas utilizam as chalconas para o desenvolvimento de novos fármacos, pois estas apresentam uma extensa gama de atividades biológicas, como: antimalárica, antimitótica, anti-inflamatória, antiviral, bactericida, fungicida e anti-leishmania. De maneira análoga às combretastatinas, elas têm apresentado atividades biológicas bastante elevadas frente a processos antimitóticos.²⁸

1.2 Teoria do funcional da densidade (DFT)

Para o tratamento de sistemas atômicos, moleculares ou sólidos, pode ser aplicado o formalismo da mecânica quântica, que se baseia na construção de um operador hamiltoniano e na solução da equação de Schrödinger correspondente, a fim de serem obtidas as propriedades do sistema. Para sistemas de muitos corpos, se faz necessário um grande número de cálculos para sua resolução, onde o uso da simulação computacional é de fundamental importância. Para um sistema composto de N elétrons e M núcleos, a equação de Schrödinger pode ser escrita como:

$$\hat{H}\Psi(\vec{r}, \vec{R}, t) = i\hbar \frac{\partial \Psi}{\partial t} \quad (1)$$

onde $\hat{H}$ é o operador Hamiltoniano dado por:

$$\begin{aligned} \hat{H} = & -\frac{\hbar^2}{2m_e} \sum_i \nabla_i^2 - \frac{\hbar^2}{2} \sum_{\alpha} \frac{1}{M_{\alpha}} \nabla_{\alpha}^2 + \frac{e^2}{8\pi\epsilon_0} \sum_{\substack{i,j \\ i \neq j}} \frac{1}{|\vec{r}_i - \vec{r}_j|} + \\ & + \frac{e^2}{8\pi\epsilon_0} \sum_{\substack{\alpha,\beta \\ \alpha \neq \beta}} \frac{Z_{\alpha}Z_{\beta}}{|\vec{R}_{\alpha} - \vec{R}_{\beta}|} - \frac{e^2}{4\pi\epsilon_0} \sum_{i,\alpha} \frac{Z_{\alpha}}{|\vec{r}_i - \vec{R}_{\alpha}|} \end{aligned} \quad (2)$$

$$= \hat{T}_e + \hat{T}_N + \hat{V}_{ee} + \hat{V}_{NN} + \hat{V}_{eN} \quad (3)$$

sendo T_e e T_N as energias cinéticas dos elétrons e dos núcleos, V_{ee} , V_{NN} e V_{eN} as energias potenciais das interações elétron-elétron, núcleo-núcleo e elétron-núcleo, e Z_{α} e Z_{β} a carga de cada núcleo α e β . Na equação 1, Ψ é uma função de onda de todas as coordenadas eletrônicas e nucleares, denotados por r_i e R_{α} , respectivamente.

Podemos propor a separação de variáveis, já que o operador $\hat{H}$ não depende explicitamente do tempo:

$$\Psi(\vec{r}, t) = \Psi(\vec{r})\varphi(t) \quad (4)$$

Dessa forma, temos a separação entre a parte temporal e espacial, chegando à equação de Schrödinger independente do tempo:

$$\hat{H}\Psi(\vec{r}) = E\Psi(\vec{r}), \quad (5)$$

onde E representa os autovalores da energia do sistema.

Usando a aproximação de Born-Oppenheimer, desconsidera-se o termo referente à energia cinética dos núcleos (T_N). Consequentemente, o termo da energia de interação entre os núcleos (V_{NN}) torna-se constante, restando apenas o hamiltoniano eletrônico:

$$\hat{H}_{el} = \hat{T}_e + \hat{V}_{ee} + \hat{V}_{eN} \quad (6)$$

cujas solução é uma função de onda que depende das coordenadas eletrônicas e parametricamente das coordenadas nucleares, $\Psi_{ee} \{r, R\}$.

Com essa aproximação, a energia pode ser escrita como

$$E = E_{el} + V_{NN} \quad (7)$$

que representa a soma da energia total eletrônica com a energia de repulsão coulombiana dos núcleos. Como Ψ depende parametricamente das coordenadas nucleares, cada posição do núcleo corresponde a uma função de onda diferente.^{29,30}

As interações repulsivas entre os elétrons são a maior dificuldade encontrada para se resolver equação de Schrödinger para sólidos ou moléculas. A aproximação de Hartree foi uma das primeiras aproximações utilizadas para resolver esse problema. Ela consiste basicamente em considerar os elétrons como partículas independentes, movendo-se sob um potencial central efetivo, onde estão incluídas a atração nuclear e o efeito médio na repulsão dos outros elétrons. A função de onda Ψ proposta por Hartree para o sistema de N elétrons consiste em um produto de N funções de onda Ψ_i de um elétron.

$$\Psi(\vec{r}_1, \vec{r}_2, \vec{r}_3, \dots, \vec{r}_n) = \phi_1(\vec{r}_1)\phi_2(\vec{r}_2)\phi_3(\vec{r}_3) \dots \phi_n(\vec{r}_n) \quad (8)$$

Chega-se à equação de Hartree quando aplicado o princípio variacional, onde a energia total é minimizada com respeito a Ψ_n , mantendo-se o vínculo $\int d\vec{r}_i |\phi_i|^2 = 1$ (condição de normalização):

$$\left(-\frac{1}{2} \nabla_i^2 - \sum_{\alpha=1}^N \frac{Z_{\alpha}}{|\vec{r}_{i\alpha}|} + U(\vec{r}_i) \right) \phi_i(\vec{r}_i) = \varepsilon_i \phi_i(\vec{r}_i) \quad (9)$$

Nessa equação, $U(\vec{r}_i)$ é o potencial repulsivo médio dado por:

$$U(\vec{r}_i) = \sum_{j \neq i} \int \frac{d\vec{r}_j}{|\vec{r}_i - \vec{r}_j|} |\phi_j|^2 \quad (10)$$

chamado potencial de Hartree, V_H .

Como os elétrons são partículas indistinguíveis e obedecem ao princípio de exclusão de Pauli, devem ser descritos por funções antissimétricas,

$$\phi(\vec{x}_1, \dots, x_i, \vec{x}_j, \dots, \vec{x}_n) = -\phi(\vec{x}_i, \dots, \vec{x}_j, x_i, \dots, \vec{x}_n) \quad (11)$$

A falha apresentada na teoria de Hartree ao desconsiderar as coordenadas de spin foi corrigida por Fock, sendo incluída a antissimetria da função de onda através do determinante de Slater. Com essa modificação, essa teoria se tornou conhecida como Aproximação de Hartree-Fock.

Como visto, o método de Hartree-Fock está baseado na determinação da função de onda dependente de $3N$ variáveis (três variáveis espaciais para cada elétron) somadas às variáveis de spin. E sua resolução se torna bastante complexa para sistemas de muitas partículas, o que, na prática, gera um custo computacional bastante elevado. Com essa motivação, outros métodos e teorias que buscam maior eficiência passaram a ser desenvolvidos, uma delas é a Teoria do Funcional da Densidade (DFT – do inglês: *Density Functional Theory*).^{29,30}

A DFT é um método de estudo de sistemas interagentes, que tem como variável básica a densidade eletrônica. A densidade de carga eletrônica é tida como sua variável fundamental, que para uma partícula com $r_i = (x_i, \sigma_i)$, ou seja, de posição x_i e spin σ_i , num sistema de N partículas interagentes, é definida por:

$$\rho(r_i) = N \sum_{\sigma} \int dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_N |\Psi(r_1, \dots, r_{i-1}, r_i, r_{i+1}, \dots, r_N)|^2 \quad (12)$$

A integral atua sobre todas as coordenadas, exceto x_i . A soma é efetuada sobre todos os valores de spin.³¹

A densidade eletrônica é uma variável suficiente para determinar qualquer propriedade física do estado fundamental de um sistema eletrônico interagente, conforme é demonstrado nos teoremas a seguir. O primeiro método baseado na densidade eletrônica, e precursor do DFT, foi o método de Thomas-Fermi.

Modelo de Thomas-Fermi: foi o primeiro a calcular a energia eletrônica total de um sistema através da densidade. Esse modelo descreve elétrons em átomos interagentes movendo-se em um potencial externo $v(r)$, existindo uma relação implícita entre a densidade eletrônica $\rho(r)$ e o potencial $v(r)$, na posição r , por meio da equação:

$$\rho(r) = \frac{1}{3\pi^2} \left(\frac{2m}{\hbar} \right)^{\frac{3}{2}} (\mu - v_{eff}(r))^{\frac{3}{2}} \quad (13)$$

Sendo μ o potencial químico, e

$$v_{eff}(r) = v(r) + v_H(r) \quad (14)$$

$$v_H(r) = \int \frac{\rho(r')}{|r - r'|} dr' \quad (15)$$

onde $v_H(r)$ é o potencial de Hartree. A equação 13 é baseada na densidade de um gás de elétrons não-interagentes de densidade uniforme em um potencial efetivo $v_{eff}(r)$.

Segundo esse modelo, o cálculo da energia pode ser separado em duas partes. Uma delas consiste na soma das energias cinética e de interação (que só dependem do número de elétrons N), definida por $F[\rho]$:

$$F[\rho] = \langle \psi | \hat{T} + \hat{U} | \psi \rangle \quad (16)$$

A outra depende do potencial externo $v(r)$, e é definida por:

$$V[\rho] = \int v(r)\rho(r)dr \quad (17)$$

Dessa forma, energia total do sistema será:

$$E[\rho] = F[\rho] + V[\rho] = F[\rho] + \int v(r)\rho(r)dr \quad (18)$$

No entanto, esse modelo se mostrava insatisfatoriamente impreciso no resultado de cálculos da energia eletrônica. Além disso, apresentava algumas falhas teóricas, como por exemplo, não provar que havia um único potencial externo associado a uma única densidade eletrônica.^{30, 32}

Essas falhas foram resolvidas posteriormente com os teoremas Hohenberg-Kohn (HK), que garantiam tanto a unicidade da densidade para um dado potencial

externo, quanto a minimização de funcionais para obtenção da energia do estado fundamental.^{31,32}

Teoremas de Hohenberg-Kohn: Através de dois teoremas, propuseram a solução do impasse existente no modelo de Thomas-Fermi:

- Primeiro teorema de Hohenberg-Kohn:

“Dada a densidade eletrônica de um sistema $\rho(r)$, o potencial externo $v(r)$ correspondente é determinado univocamente.” (THOMAS, 1927).

Assim, pode-se determinar o potencial correspondente univocamente sendo conhecida a densidade eletrônica. O que é de suma importância, pois a validade desse teorema assegura que a densidade eletrônica é suficiente para caracterizar completamente o sistema, já que o funcional universal depende apenas do número de elétrons. A prova desse teorema foi proposta por Kohn e Sham, aqui válidas apenas para estados não-degenerados, pode ser enunciada como:

Seja $\rho(r)$ um estado não-degenerado de densidade eletrônica de um sistema de N elétrons em um potencial $v_1(r)$, correspondendo a um estado ϕ_1 de energia E_1 :

$$E_1 = \langle \phi_1 | \hat{H}_1 | \phi_1 \rangle = \int v_1(r) \rho(r) dr + \langle \phi_1 | \hat{T} + \hat{U} | \phi_1 \rangle \quad (19)$$

tomando $\hat{H}_1$ como o Hamiltoniano referente a v_1 , $\hat{T}$ e $\hat{U}$, com $\hat{T}$ e $\hat{U}$ sendo os operadores de energia cinética e de interação.

Por absurdo, suponha que existe um segundo potencial $v_2(r)$, que não é equivalente a $v_1(r)$, ou seja, $v_2(r) \neq v_1(r) + \text{cte}$ e $\phi_2 \neq e^{i\theta} \phi_1$. Dessa forma:

$$E_2 = \langle \phi_2 | \hat{H}_2 | \phi_2 \rangle = \int v_2(r) \rho(r) dr + \langle \phi_2 | \hat{T} + \hat{U} | \phi_2 \rangle \quad (20)$$

Desde que ϕ_1 é não degenerado, o princípio variacional de Rayleigh-Ritz é aplicável:

$$E_1 < \langle \phi_2 | \hat{H}_1 | \phi_2 \rangle = \int v_1(r) \rho(r) dr + \langle \phi_2 | \hat{T} + \hat{U} | \phi_2 \rangle \quad (21)$$

$$E_1 < E_2 + \int [v_1(r) - v_2(r)]\rho(r)dr \quad (22)$$

De maneira análoga:

$$E_2 < \langle \phi_1 | \hat{H}_2 | \phi_1 \rangle = \int v_2(r)\rho(r)dr + \langle \phi_1 | \hat{T} + \hat{U} | \phi_1 \rangle \quad (23)$$

$$E_2 < E_1 + \int [v_2(r) - v_1(r)]\rho(r)dr \quad (24)$$

Somando essas duas últimas equações, chega-se a:

$$E_1 + E_2 < E_1 + E_2 \quad (25)$$

O que é uma contradição, e mostra que a suposição de que para um mesmo $\rho(r)$, existe um segundo potencial $v_2(r)$ (tal que $v_2(r) \neq v_1(r) + \text{constante}$) é absurda, ou seja, $\rho(r)$ e $v(r)$ são determinados univocamente.³³

- Segundo teorema de Hohenberg-Kohn:

“Seja um sistema de densidade $\rho(r)$ no estado fundamental e $\tilde{\rho}(r)$ uma aproximação arbitrária de $\rho(r)$, tal que $\tilde{\rho}(r) \geq 0$ e $\int \tilde{\rho}(r)dr = N$. Seja $E[\tilde{\rho}]$ a energia correspondente a $\tilde{\rho}$ e $E[\rho]$ a energia correspondente a $\rho(r)$. Então:

$$E[\tilde{\rho}] \geq E[\rho] = E_0 \quad (26)$$

onde E_0 é a energia do estado fundamental do sistema.” (THOMAS, 1927)

Composta pelo funcional universal $F[\rho]$, que depende apenas do número de elétrons, e pelo potencial externo, que corresponde univocamente à densidade, a energia é um funcional da densidade que atinge o mínimo quando a densidade adotada corresponde a densidade eletrônica real do sistema. A prova pode ser apresentada conforme os argumentos a seguir.

Pode-se dizer que a densidade $\rho(r)$ está associada a um potencial $v(r)$ e a um hamiltoniano $\hat{H}$, assim como a densidade $\tilde{\rho}(r)$ arbitrária está associada um potencial $\tilde{v}(r)$ e um hamiltoniano $\tilde{H}$, conforme o 1º teorema HK. Da mesma forma, a densidade

$\rho(r)$ está para uma função de onda Φ , assim como $\tilde{\rho}(r)$ está associada a $\tilde{\Phi}$. Incluindo a densidade arbitrária $\tilde{\rho}(r)$ no hamiltoniano normal ($\hat{H}$):

$$E[\tilde{\rho}] = \langle \tilde{\Phi} | \hat{H} | \tilde{\Phi} \rangle = \int v(r) \tilde{\rho}(r) dr + \langle \tilde{\Phi} | \hat{T} + \hat{U} | \tilde{\Phi} \rangle \quad (27)$$

Conforme o princípio variacional de Rayleigh-Ritz, pode-se afirmar:

$$\int v(r) \tilde{\rho}(r) dr + \langle \tilde{\Phi} | \hat{T} + \hat{U} | \tilde{\Phi} \rangle \geq \int v(r) \rho(r) dr + \langle \Phi | \hat{T} + \hat{U} | \Phi \rangle, \quad (28)$$

$$E[\tilde{\rho}] \geq E[\rho] = E_0 \quad (29)$$

Segundo esse teorema, o estado de menor energia corresponde a densidade eletrônica $\rho(r)$, e qualquer aproximação $\tilde{\rho}(r)$ aplicada ao funcional de energia, provocará um resultado acima do esperado, ou seja $E[\tilde{\rho}] \geq E[\rho]$.

Como citado anteriormente, a prova dada por Hohenberg e Kohn era válida apenas para os estados não-degenerados. Dessa forma, persistia a dúvida quanto a validade desses teoremas para os estados degenerados.

Apesar de provar que a densidade eletrônica era suficiente para descrever um sistema, os teoremas Hohenberg-Kohn eram incapazes de determiná-la. Pouco depois, Kohn e Sham concluíram que seria possível estudar um sistema interagente a partir de um sistema não interagente de potencial convenientemente escolhido. Esse processo ficou conhecido como esquema Kohn-Sham (KS). Os teoremas de Hohenberg-Kohn, juntamente com o esquema de Kohn-Sham, formam os alicerces do DFT.³⁴

Como os teoremas de Hohenberg-Kohn não estabelecem um procedimento para o cálculo das energias, Kohn e Sham, propuseram um conjunto de equações, denominadas equações de Kohn-Sham, nas quais um sistema de partículas interagentes é substituído por um sistema de partículas não-interagente (independentes) submetidas a um potencial arbitrário, que reproduz as condições do sistema interagente.

O uso de um sistema não interagente para calcular a energia de um sistema interagente, como proposto no esquema Kohn Sham, implica no surgimento do termo

chamado de energia de troca e correlação, que, de maneira sucinta, reúne as interações entre as partículas. Ou seja, é nessa energia que surge a descrição do número de elétrons e o quão intensamente eles interagem entre si.

Considera-se, nessa mudança, o fato de que em ambos os sistemas, interagentes ou não, a densidade eletrônica é a mesma, onde toma-se proveito da simplicidade da solução de um sistema não-interagente.

O funcional da densidade para a energia eletrônica total contém basicamente três termos: a energia cinética ($T[\rho]$), a interação U e o potencial V :

$$E[\rho] = T[\rho] + V[\rho] + U[\rho] \quad (30)$$

As contribuições para a energia eletrônica podem ser reorganizadas quanto a cada tipo de energia.

A energia cinética pode ser dividida em duas formas: da partícula independente $T_s[\rho]$ e de correlação $T_c[\rho]$. O funcional da energia cinética de partícula independente, $T_s[\rho]$, é dado por:

$$T_s[\rho] = T_s(\phi_i(r)) = \frac{\hbar^2}{2m} \sum_i^N \int \phi_i^*(r) \nabla^2 \phi_i(r) d^3r \quad (31)$$

Isso representa uma pequena mudança na equação da energia cinética: $\langle \hat{T} \rangle = -\frac{\hbar^2}{2m} \sum_i^N \int \psi_i^*(r_i) \nabla^2 \psi_i(r_i) dx_i$, na qual a função de onda total ψ é trocada pela função de onda ϕ_i de cada partícula, já que é atribuída à a energia cinética de um sistema de partículas independentes. Essa substituição é feita por causa da complexidade dos cálculos envolvendo a função de onda total ψ , que agrega os efeitos de muitos corpos. Por outro lado, a energia cinética de correlação, que corresponde a diferença entre a energia cinética total e $T_s[\rho]$, tem a forma:

$$T_c[\rho] = T[\rho] - T_s[\rho] \quad (32)$$

A energia que corresponde a soma das interações elétron-elétron, ou seja, a energia de interação U , tem como parte principal a energia de Hartree, denotada por $U_H[\rho]$:

$$U_H[\rho] = \frac{1}{2} \int d^3r \int d^3r' \frac{\rho(r)\rho(r')}{|r - r'|} \quad (33)$$

sendo $\rho(r)$ a densidade eletrônica na posição r . Essa expressão fornece a energia de interação eletrostática clássica entre os elétrons, ou seja, a interação entre cada elétron e a densidade média de carga de todos os elétrons do sistema. A diferença, decorrente da substituição de U por U_H denotada por U_{xc} é dada por:

$$U_{xc}[\rho] = U[\rho] - U_H[\rho] \quad (34)$$

Conhecidos $T_s[\rho]$ e $U_H[\rho]$, pode-se reorganizar o funcional de energia da equação 30:

$$E[\rho] = T_s[\rho] + U_H[\rho] + V[\rho] + E_{xc}[\rho] \quad (35)$$

$E_{xc}[\rho]$ é a energia de troca e correlação que corresponde a diferença de energia decorrente da substituição de $T[\rho]$ por $T_s[\rho]$ e de $U[\rho]$ por $U_H[\rho]$. Assim:

$$E_{xc}[\rho] = (T[\rho] - T_s[\rho]) + (U[\rho] - U_H[\rho]) = T_c[\rho] + U_{xc}[\rho] \quad (36)$$

Dessa forma, a energia de troca e correlação reúne toda a diferença decorrente da substituição do sistema interagente por um sistema não-interagente. Em geral, escreve-se $E_{xc} = E_x + E_c$, onde E_x é a energia de troca e E_c é a energia de correlação.³⁵

Como a DFT se baseia na densidade eletrônica para descrever os sistemas eletrônicos e que, de acordo com o 2º teorema de HK, a densidade eletrônica real de um sistema no estado fundamental é a que minimiza o funcional de energia, é válida a relação:

$$\begin{aligned} \frac{\delta E[\rho]}{\delta \rho(r)} &= \frac{\delta T_s[\rho]}{\delta \rho(r)} + \frac{\delta V[\rho]}{\delta \rho(r)} + \frac{\delta U_H[\rho]}{\delta \rho(r)} + \frac{\delta E_{xc}[\rho]}{\delta \rho(r)} = 0 \\ &= \frac{\delta T_s[\rho]}{\delta \rho(r)} + v[\rho, r] + v_H[\rho, r] + v_{xc}[\rho, r] = 0 \end{aligned} \quad (37)$$

onde:

$v[\rho, r] = \delta V[\rho]/\delta \rho(r)$ é o potencial produzido pelos núcleos dos átomos do sistema, $v_H[\rho, r] = \delta U_H[\rho]/\delta \rho(r)$ é o potencial de Hartree e $v_{xc}[\rho, r] = \delta E_{xc}[\rho]/\delta \rho(r)$ é o potencial de troca e correlação.

Considerando um sistema não-interagente (composto por partículas independentes), de densidade eletrônica ρ_s , só há a energia cinética de cada partícula (T_s) e o potencial, não havendo o termo de interação (U). Então:

$$E_s[\rho_s] = T_s[\rho_s] + V[\rho_s] \quad (38)$$

Deve-se aplicar a minimização em relação a ρ_s (assim como foi feito na equação 37), de acordo com o segundo teorema de HK:

$$\begin{aligned} \frac{\delta E_s[\rho_s]}{\delta \rho_s(r)} &= \frac{\delta T_s[\rho_s]}{\delta \rho_s(r)} + \frac{\delta V_s[\rho_s]}{\delta \rho_s(r)} = 0 \\ &= \frac{\delta T_s[\rho_s]}{\delta \rho_s(r)} + v_s[\rho_s, r] = 0 \end{aligned} \quad (39)$$

Com $v_s[\rho_s, r] = \delta V[\rho_s]/\delta \rho_s(r)$ sendo o potencial arbitrário, escolhido de maneira conveniente, de forma a reproduzir, no esquema Kohn-Sham, o sistema interagente.

Kohn e Sham usaram o sistema não-interagente, de potencial $v_s(r)$ arbitrário, para determinar a densidade do estado fundamental do sistema interagente. Em ambos os sistemas, a densidade eletrônica é a mesma, ou seja, $\rho_s(r) = \rho(r)$. Comparando as equações 37 e 39, tem-se que:

$$v_s[\rho, r] = v(r) + v_H[\rho, r] + v_{xc}[\rho, r] \quad (40)$$

essa é a expressão que calcula o potencial a ser escolhido para que o sistema não-interagente reproduza o sistema interagente.³⁶

O sistema não-interagente pode ser resolvido facilmente, pois é um sistema de partículas independentes. Resolvendo a equação de Schrödinger correspondente para a partícula i:

$$\left[\frac{\hbar^2}{2m} \nabla^2 + v_s[\rho, r] \right] \phi_i(r) = \varepsilon_i \phi_i(r) \quad (41)$$

A densidade eletrônica $\rho(r)$ é determinada por:

$$\rho(r) = \rho_s(r) = \sum_i^N f_i |\phi_i(r)|^2 \quad (42)$$

onde $f_i = 0$ ou 1 é a ocupação do i -ésimo orbital.

Dessa forma, o esquema KS é exato (apesar de sua implementação só poder ser feita numericamente) e auto-consistente (já que o potencial v_s depende da densidade eletrônica). Porém, não há um funcional de troca e correlação exato, ou mesmo uma forma de determiná-lo exatamente.

Apesar do DFT ser um formalismo exato, se faz necessária uma aproximação adequada para o funcional da energia de troca e correlação de um sistema interagente, e a busca por um funcional universal cada vez mais acurado é motivo de grande interesse, cuja pesquisa se estende há mais de três décadas.³⁷

Várias linhas de pesquisa acerca do DFT voltam-se para construção de aproximações para a energia de troca e correlação. A primeira delas, proposta por Kohn e Sham, foi a LDA (do inglês: *Local Density Approximation*), na qual as propriedades de sistemas homogêneos são aplicadas localmente para sistemas quaisquer. A precisão é satisfatória em muitos casos, sendo muito empregada devido ao seu custo computacional ser relativamente baixo.

Aproximação da Densidade Local: A LDA foi proposta por Kohn e Sham, cujo procedimento consiste na aplicação local da energia de troca e correlação do gás de elétrons homogêneo. A energia de troca por unidade de volume é:

$$e_x^{hom}(\rho(r)) = -\frac{3q^2}{4} \left(\frac{3}{\pi} \right)^{\frac{1}{3}} \rho(r)^{\frac{4}{3}} \quad (43)$$

Tem-se a energia de troca total de um sistema homogêneo quando se soma em todo o espaço:

$$E_x[\rho] \approx E_x^{LDA}[\rho] = \int d^3r e_x^{hom}(\rho(r)) \quad (44)$$

A LDA para o funcional da energia de correlação $E_c[\rho]$ é determinada a partir da parametrização dos resultados de Monte Carlo quântico. É muito utilizada no estudo de sólidos metálicos, pois além de eficiente, possui um baixo custo computacional. No entanto, para átomos e moléculas, nos quais a distribuição eletrônica é altamente não-homogênea, a LDA não apresenta bons resultados.

Aproximação Generaliza do Gradiente da Densidade: As correções à LDA foram introduzidas na tentativa de corrigir os erros provenientes da não-homogeneidade da densidade eletrônica no espaço, principalmente nos sistemas moleculares, já que dependiam do gradiente da densidade eletrônica. O primeiro desses modelos foi o GEA (Gradient Expansion Approximation), no qual foram introduzidos termos como o gradiente da densidade (tais como $|\nabla\rho(r)|^2$, $\nabla^2\rho(r)$, etc). Um exemplo disso é o funcional de Weisäcker:

$$T_s^{vw} = \int d^3r \frac{|\nabla\rho(r)|^2}{8\rho(r)} \quad (45)$$

Porém, a introdução do gradiente, mesmo tornando os cálculos muito mais complexos, não implica em resultados mais refinados que do método LDA.³⁵

Posteriormente, o gradiente da densidade foi introduzido no funcional LDA, gerando a GEA (do inglês: *Gradient Expansion Approximation*), seguido de expansões mais elaboradas do gradiente da densidade eletrônica, que são os funcionais GGA (do inglês: *Generalized Gradient Approximation*), que apresentam a forma geral:

$$E_{xc}^{GGA}[\rho] = \int d^3r f(\rho(r), \nabla\rho(r)) \quad (46)$$

A diferença entre os GGA's é a função f escolhida. Os GGA's mais usados são o PBE (proposto por Perdew, Burke e Ernzerhof) e o BLYP (denotando a combinação do funcional de troca de Becke com o funcional de correlação de Lee, Yang e Parr).³⁶

O uso do DFT em suas mais variadas extensões tornou os cálculos de propriedades eletrônicas de compostos (por exemplo, materiais do estado sólido, polímeros e biomoléculas) muito mais rápidos e precisos, produzindo um grande impacto em pesquisas de diversas áreas da ciência.³⁶

Na função de base a ser utilizada em um determinado estudo, podem ser acrescentadas funções de polarização e funções difusas, as quais são de grande importância na descrição do comportamento eletrônico da molécula.

Funções de Polarização: As funções de polarização são um conjunto de funções de base de valência desdobrada que permitem mudar o tamanho dos orbitais, mas não sua forma. As funções de polarização removem essa limitação adicionando orbitais com momento angular além do que seria necessário para o estado fundamental para a descrição de cada átomo. Por exemplo, as funções de polarização adicionam funções d para átomos de carbono e funções f para metais de transição. Algumas adicionam funções p para os átomos de hidrogênio.

Por exemplo, a denominação 6-31G(d) indica que foram adicionadas funções d para átomos pesados nas funções de base do conjunto 6-31G. Esse conjunto também é conhecido com 6-31G*. Outro conjunto de funções bastante popular é o 6-31G(d,p) ou 6-31G**, que adiciona funções p para átomos de hidrogênio além das funções d para todos átomos pesados.³⁸

Funções difusas: Permitem que os orbitais ocupem uma região maior do espaço. Conjuntos de funções de base com funções difusas são importantes para sistema onde os elétrons estão relativamente longe do núcleo: moléculas com pares solitários, ânions e outros sistemas com significativas cargas negativas, sistemas em estados excitados, sistemas com baixos potenciais de ionização, etc.

Por exemplo, a função 6-31+G(d) é o conjunto 6-31G(d) acrescido de funções difusas em átomos pesados. A versão com duplo sinal de adição, 6-31++G(d), adiciona funções difusas também aos átomos de hidrogênio.³⁸

1.3 Relação Quantitativa Estrutura-Atividade (QSAR)

Obter Relações Quantitativas entre Estrutura e Atividade de um conjunto de compostos significa gerar modelos capazes de correlacionar as propriedades estruturais e físico-químicas dos mesmos com suas respectivas respostas biológicas de interesse. Isso possibilita olhar para os compostos em termos dessas propriedades estruturais, ao invés de considerar apenas certos grupos farmacofóricos.

Essas relações são amplamente utilizadas na indústria farmacêutica para compreender como as características estruturais de moléculas biologicamente ativas contribuem para sua atividade.³⁹

As QSARs são apresentadas como equações, na maioria das vezes equações lineares, também chamadas de modelos, que quantificam a atividade biológica como função das propriedades das moléculas. Uma vez que o modelo é desenvolvido, pode ser utilizado para priorizar em pesquisas posteriores mais aprofundadas, moléculas cujas atividades biológicas previstas pelo modelo foram as mais altas.⁴⁰

Vários aspectos devem ser considerados para que um modelo de QSAR seja desenvolvido com sucesso. A seleção dos descritores que representam as moléculas e as análises estatísticas devem ser feitas de maneira bastante criteriosa, a fim de que sejam selecionados parâmetros realmente relevantes para a atividade biológica em questão. Também se torna imprescindível a realização de testes de validação dos modelos, para assegurar a confiabilidade e funcionalidade dos mesmos.

Uma condição importante nesse estudo é a que as moléculas estudadas devem possuir um esqueleto em comum, no qual pode haver variações estruturais. Pois, baseando-se nisso, pode-se avaliar o mecanismo de ação de alguns compostos, e assumir que todos os outros apresentem o mesmo comportamento frente a um receptor biológico, mesmo que, eventualmente, o mecanismo não seja provado para todos os compostos.⁴¹

Tratando os descritores moleculares como variáveis, podemos extrair importantes informações relacionadas à atividade biológica dos compostos. De cada molécula otimizada, uma grande quantidade de informações é obtida. Os descritores moleculares podem ser divididos em diversas classes, como os constitucionais, os geométricos, os topológicos, os físico-químicos e os eletrônicos.

Constitucionais: referem-se à constituição básica de cada molécula, como número de átomos de determinado elemento químico, número de ligações, massa molecular, quantidade de anéis presentes na cadeia molecular, coeficiente de partição entre fases hidrofílica e hidrofóbica.

Geométricos: representam a disposição dos elementos na molécula, ângulos de ligação, ângulos de diedro, localização de fragmentos moleculares, entre outros.⁴²

Topológicos: descrevem a estrutura da molécula, a partir dos quais podem ser definidos vários índices quantificando a conectividade molecular, como, por exemplo, os índices de Wiener, de Randic, de Balaban, de Schultz e de Kier e Hall.⁴¹

Físico-Químicos: representam propriedades físico-químicas como, por exemplo, temperaturas de fusão e ebulição, solubilidade, coeficientes de partição entre fases orgânica e aquosa, densidade, constantes dielétricas e pK_a .⁴³

Eletrônicos: atribuídos à distribuição de cargas pela molécula, ocupação de orbitais moleculares como: HOMO (*Highest Occupied Molecular Orbital*), LUMO (*Lowest Unoccupied Molecular Orbital*), e consequentemente, orbitais efetivos para reações químicas FERMO (*Frontier Effective-for-Reaction Molecular Orbital*).⁴¹

Sendo obtida a energia dos estados neutro, oxidado e reduzido de cada composto, é possível calcular propriedades como afinidade eletrônica, potencial de ionização, dureza, índice eletrofílico, moleza e eletropositividade.⁴⁴⁻⁴⁶

Após a obtenção dos cálculos das geometrias das moléculas, será possível obter diversos outros descritores moleculares. E com esses dados, pode-se utilizar métodos para separar as variáveis que são relevantes das que não são relevantes para explicar a variação da atividade biológica.

A análise de correlação com os valores de IC₅₀ pode ser usada para uma seleção inicial das variáveis, selecionando as variáveis que mantenham alta correlação com as respectivas atividades. Numa etapa seguinte da seleção dos descritores, pode ser utilizado o método de Algoritmo Genético.⁴⁷

Para a geração do modelo, podem ser utilizadas técnicas de PLS, bem como Redes Neurais Artificiais. Essas técnicas podem apresentar resultados semelhantes ou distintos, dependendo do tipo de relação entre os descritores envolvidos.

Essas técnicas são de fundamental importância na construção de um modelo matemático, já que são elas as responsáveis pela seleção das variáveis e suas possíveis combinações. Seguem suas descrições:

1.3.1 Algoritmo Genético

Esse método se constitui através algoritmos de otimização global, baseados nos mecanismos de seleção natural e da genética. Empregam uma estratégia de busca paralela e estruturada, mas aleatória, que é voltada em direção ao reforço da busca de pontos de alta aptidão, ou seja, pontos nos quais a função a ser minimizada (ou maximizada) tem valores relativamente baixos (ou altos). Não são caminhadas não direcionadas, apesar de aleatórios, pois exploram históricos de informações para encontrar novos pontos de busca onde são esperados melhores desempenhos, o que é feito através de processos iterativos, onde cada iteração é chamada de geração.

Os princípios de seleção e reprodução são aplicados a uma população de candidatos que pode variar, dependendo da complexidade do problema e dos

recursos computacionais disponíveis. Assim se determina quais indivíduos conseguirão se reproduzir, gerando um número determinado de descendentes para a próxima geração, com uma probabilidade determinada pelo seu índice de aptidão.

O princípio básico dos operadores genéticos é transformar a população através de sucessivas gerações, estendendo a busca até chegar a um resultado satisfatório. Os operadores genéticos são necessários para que a população se diversifique e mantenha características de adaptação adquiridas pelas gerações anteriores.

Se faz necessário um conjunto de operações para que, dada uma população, se consiga gerar populações sucessivas que melhorem sua aptidão com o tempo. Esses operadores são: cruzamento (*crossover*) e mutação. Eles são utilizados para assegurar que a nova geração seja totalmente nova, mas possui, de alguma forma, características de seus pais, ou seja, a população se diversifica e mantém características de adaptação adquiridas pelas gerações anteriores.⁴⁸

Para assegurar que os melhores indivíduos não desapareçam da população pela manipulação dos operadores genéticos, eles podem ser automaticamente colocados na próxima geração, através da reprodução elitista. Esse ciclo é repetido um determinado número de vezes. Durante esse processo, os melhores indivíduos, assim como alguns dados estatísticos, podem ser coletados e armazenados para avaliação.

Alguns parâmetros influem no comportamento dos Algoritmos Genéticos. Convém analisar como se dão, para que se possa estabelecê-los conforme as necessidades do problema e dos recursos disponíveis.

O tamanho da população afeta o desempenho global e a eficiência dos Algoritmos Genéticos. Com uma população pequena o desempenho pode cair, pois deste modo a população fornece uma pequena cobertura do espaço de busca do problema. Uma grande população geralmente fornece uma cobertura representativa do domínio do problema, além de prevenir convergências prematuras para soluções locais ao invés de globais.

Quanto maior for a taxa de cruzamento, mais rapidamente novas estruturas serão introduzidas na população. Porém, se essa for muito alta, estruturas com boas aptidões poderão ser retiradas precipitadamente. Com um valor baixo, o algoritmo pode tornar-se muito lento.

Uma baixa taxa de mutação previne que uma dada posição fique estagnada em um valor, além de possibilitar que se chegue em qualquer ponto do espaço de busca. Com uma taxa muito alta a busca se torna essencialmente aleatória.

O intervalo de geração controla a porcentagem da população que será substituída durante a próxima geração. Com um valor alto, a maior parte da população será substituída, mas com valores muito altos pode ocorrer perda de estruturas de alta aptidão. Com um valor baixo, o algoritmo pode tornar-se muito lento. Contudo, sua utilização como método para separação de variáveis é motivada pela sua eficiência, principalmente quando o número de variáveis for elevado.^{48,49}

1.3.2 Mínimos Quadrados Parciais (PLS)

Esse método de regressão possibilita o desenvolvimento de uma expressão matemática correlacionando as variáveis, sendo que a essa expressão é capaz de modelar a atividade biológica dos compostos.

Nesse método, duas matrizes X e Y são relacionadas, onde X representa os descritores e Y a atividade biológica:

$$X = TP' + E = \sum t_n p'_n + E \quad (47)$$

$$Y = UQ' + F = \sum u_n q'_n + F \quad (48)$$

onde T é a matriz de escore da matriz X, e P é o seu peso. U é a matriz de escore da matriz Y, e Q é o seu peso. Os respectivos resíduos são representados por E e F. Os termos dos somatórios representam as decomposições em n variáveis.

Um bom modelo de PLS tem um bom valor de coeficiente de correlação, R, o qual determina a razão entre o valor mensurado da atividade biológica e o valor predito pelo modelo matemático gerado. Outro parâmetro é o R^2 , que corresponde à fração da variabilidade total que é explicada pelo modelo, portanto quanto mais próximo o valor de R^2 for de 1, maior a capacidade do modelo explicar as variações em torno da média.

$$R^2 = 1 - \frac{\sum (y_{iexp} - \hat{y}_{ical})^2}{\sum (y_{iexp} - y_{mean})^2} \quad (49)$$

onde:

y_{iexp} : valor experimental da atividade da i -ésima amostra;

$\hat{y}_{ical}$: previsão da atividade da i -ésima amostra utilizando o modelo com todas as amostras;

y_{mean} : valor médio experimental da atividade;

A razão entre a variabilidade explicada pelo modelo e a variabilidade não explicada pelo modelo é denominada equação de Teste F:

$$Teste\ F = \frac{\sum (y_{iexp} - y_{mean})^2 - \sum (y_{iexp} - \hat{y}_{ical})^2}{\sum (y_{iexp} - \hat{y}_{ical})^2} \frac{(I - 1)}{VL} - 1 \quad (50)$$

onde I é o número de amostras e VL é número de variáveis latentes.

Um bom modelo de regressão deve ter um Teste F com o maior valor possível, o que remete um modelo que consiga explicar suas variações.

O último parâmetro que analisa a consistência do modelo de regressão é o Q^2 , que mensura a capacidade de predição de novos compostos.

$$Q^2 = 1 - \frac{\sum (y_{iexp} - \hat{y}_{ival})^2}{\sum (y_{iexp} - y_{mean})^2} \quad (51)$$

onde $\hat{y}_{ival}$ representa a previsão da atividade sem a i -ésima amostra, utilizando o modelo sem a mesma.

O Q^2 é obtido através da validação cruzada. O mecanismo que faz essa validação é o método “*leave-one-out*”, que cria um modelo de regressão sem uma molécula do conjunto, e utilizando esse novo modelo para predizer a atividade biológica do composto retirado do conjunto. Então é calculado o desvio padrão entre o valor previsto e observado. Esse processo é repetido até que todas as moléculas do conjunto sejam retiradas e validadas.

Dentre os parâmetros ligados aos erros de predição, os três principais são: *PRESS*, *SEV* e *SEC*, os quais são apresentados a seguir.

- Soma dos quadrados dos erros de predição (*PRESS* - do inglês, *Prediction Residues Error Squares Sum*):

$$PRESS = \sum (y_{iexp} - \hat{y}_{ival})^2 \quad (52)$$

- Erro padrão de validação (*SEV* - do inglês, *Standard Error of Validation*):

$$SEV = \left[\frac{PRESS}{I} \right]^{1/2} \quad (53)$$

- Erro padrão de calibração (*SEC* - do inglês, *Standard Error of Calibration*):

$$SEC = \left[\frac{\sum (y_{iexp} - \hat{y}_{ical})^2}{I - VL - 1} \right]^{1/2} \quad (54)$$

Resumidamente, pode-se afirmar que um bom modelo de PLS deve apresentar valores de R^2 e Q^2 o mais próximo possível de 1, os maiores valores possíveis de Teste-F e os menores valores possíveis de *PRESS*, *SEV* e *SEC*.^{50,51,52}

1.3.3 Redes Neurais Artificiais (ANN)

Inspiradas na estrutura neural de organismos inteligentes que adquirem conhecimento através da experiência, o método de Redes Neurais Artificiais (ANN) é um método computacional de reconhecimento de padrões que apresentam um modelo matemático como produto final. Uma grande rede neural artificial pode ter centenas ou milhares de unidades de processamento. O cérebro de um mamífero pode ter muitos bilhões de neurônios.⁵³

A habilidade de aprender com seu ambiente para com isso melhorar seu desempenho é a propriedade mais importante das redes neurais. Isso é feito através de um processo iterativo de ajustes aplicado a seus pesos, o treinamento. O aprendizado ocorre quando a rede neural atinge uma solução generalizada para uma classe de problemas. São muitos os tipos de algoritmos de aprendizado específicos

para determinados modelos de redes neurais, estes algoritmos diferem entre si principalmente pelo modo como os pesos são modificados.

As etapas do funcionamento básico das redes neurais artificiais podem ser divididas em camadas, que são classificadas em três grupos: camadas de entrada, intermediárias e de saída. Na camada de entrada é onde os padrões são apresentados à rede, para, em seguida, nas camadas intermediárias, a maior parte do processamento ser realizada, através das conexões ponderadas. Essas podem ser consideradas como extratoras de características. Por fim, na camada de saída, o resultado final é concluído e apresentado. Um modelo que possibilita a utilização de camadas intermediárias é o do tipo Perceptron de Multicamadas, MLP (do inglês, *Multilayer Perceptron*).⁵⁴

Figura 5 – Representação do funcionamento das Redes Neurais Artificiais do tipo MLP.

Fonte: SEGATTO, E. C.; COURY, D. V., 2008, p. 93-106.⁵⁵

Apresentando inicialmente um padrão à rede, a mesma produz uma saída. Fazendo uso de um algoritmo de aprendizagem, serão ajustados os valores numéricos até se obter outra resposta, de maneira análoga à introduzida inicialmente. Esse algoritmo consiste basicamente em, uma vez dada as variáveis de entrada e a geração da resposta, calcular o erro dos valores obtidos em relação aos fornecidos à rede. O erro é então utilizado para estimar os erros das camadas intermediárias, a fim de que ele seja retropropagado até as conexões da camada de entrada – os pesos, os quais são ajustados para ser obtido um menor erro na camada de saída. Essa é a

formulação matemática do algoritmo *backpropagation*, que consiste basicamente em atualizar os pesos para minimizar o erro.

Um conjunto de exemplos com as respectivas respostas deve ser utilizado para o treinamento da rede. Isso fará com que a mesma aprenda as relações entre as variáveis fornecidas por experiência. As respostas geradas pelas unidades são calculadas através de uma função de ativação. Existem vários tipos de funções de ativação, as mais comuns são: *Hard Limiter*, *Threshold Logic* e *Sigmoid*.^{54,56}

2 OBJETIVOS

2.1 Objetivo Geral

Desenvolver modelos matemáticos capazes de prever a atividade biológica de uma série de 87 chalconas frente à $\alpha\beta$ -tubulina, a partir de descritores eletrônicos e moleculares.

2.2 Objetivos Específicos

- Construir e otimizar a estrutura molecular de 87 chalconas análogas à combretastatina A4.
- Obter descritores eletrônicos e moleculares dos compostos estudados.
- Selecionar os descritores que apresentam as maiores influências nas atividades biológicas dos compostos.
- Gerar um modelo matemático quantitativo capaz de prever a atividade biológica dos compostos utilizando os métodos PLS e ANN.
- Verificar as relações de linearidade / não-linearidade entre os descritores selecionados.
- Analisar o significado químico do modelo gerado.

3 METODOLOGIA

Diversos estudos sucederam os trabalhos de Petit, buscando isolar e sintetizar compostos similares à combretastatina, com o intuito de otimizar a atividade biológica dessa classe de compostos. Entre esses trabalhos, está o desenvolvido por Ducki e colaboradores¹³, de onde foram sintetizadas e testadas biologicamente. Desse estudo foram selecionadas as 87 chalconas que serão usadas neste trabalho.

A descrição das moléculas selecionadas com suas respectivas representações estruturais e valores de atividade biológica (IC₅₀) é apresentada a seguir.¹³

Tabela 1 – Estruturas utilizadas com seus respectivos ligantes e valores de atividade.

Estrutura base 2:

	R2	R3	R4	R5	R6	R2'	R3'	R4'	R5'	R6'	IC ₅₀ (μM)
2a	H	OMe	OMe	OMe	H	H	OH	OMe	H	H	0,0043
2b	H	OMe	OMe	OMe	H	H	NH ₂	OMe	H	H	0,01
2c	OMe	OMe	OMe	H	H	H	OH	OMe	H	H	0,01
2d	OMe	H	H	OMe	H	H	H	NMe ₂	H	H	0,04
2e	H	OMe	OMe	OMe	H	H	O-(CH ₂) ₂ -O		H	H	0,08
2f	H	OMe	OMe	OMe	H	OMe	H	OMe	H	OMe	0,20
2g	H	OMe	OMe	OMe	H	H	OMe	OMe	H	H	0,30
2h	H	OMe	OMe	OMe	H	H	O-CH ₂ -O		H	H	0,30
2i	H	OMe	OMe	OMe	H	H	F	OMe	H	H	0,30
2j	H	OMe	OMe	OMe	H	H	H	OMe	H	H	0,30

2k	H	OMe	OMe	H	H	OMe	H	OMe	H	OMe	0,30
2l	OMe	H	OMe	H	OMe	H	OMe	OMe	OMe	H	0,30
2m	OMe	OMe	OMe	H	H	H	H	OMe	H	H	0,30
2n	OMe	H	H	OMe	H	H	O-(CH ₂) ₂ -O		H	H	0,30
2o	OMe	OMe	OMe	H	H	H	F	OMe	H	H	0,35
2p	H	O-CH ₂ -O		H	H	OMe	H	OMe	OMe	OMe	1,1
2q	H	OMe	OMe	H	H	H	OMe	OMe	OMe	H	1,1
2r	H	OMe	OMe	OMe	H	H	OH	OH	H	H	1,2
2s	H	H	Me	H	H	H	OH	OMe	H	H	1,5
2t	H	OMe	OMe	OMe	H	H	OMe	OMe	OMe	H	1,5
2u	H	OMe	OMe	OMe	H	H	H	Cl	H	H	1,5
2v	H	OMe	OMe	OMe	H	H	H	F	H	H	1,5
2w	H	H	H	H	H	H	OMe	OMe	OMe	H	1,6
2x	H	H	F	H	H	H	OMe	OMe	OMe	H	1,6
2y	OMe	OMe	OMe	H	H	H	OH	OH	H	H	1,7
2z	H	OMe	OMe	H	H	H	OMe	OMe	H	H	2,2
2aa	OMe	H	OMe	H	H	H	H	OMe	H	H	2,7
2bb	H	O-CH ₂ -O		H	H	H	OMe	OMe	OMe	H	2,9
2cc	H	OMe	OH	H	H	H	OMe	OMe	OMe	H	3,5
2dd	H	OMe	OMe	OMe	H	Cl	H	Cl	H	H	3,8
2ee	H	H	H	H	H	H	H	H	H	H	3,8
2ff	f	H	OMe	H	H	OMe	H	H	OMe	H	4,0
2gg	H	OMe	OMe	OMe	H	H	NO ₂	OMe	H	H	4,0
2hh	OH	H	OMe	H	H	H	OMe	OMe	OMe	H	4,4
2ii	H	H	OMe	H	H	H	OMe	OMe	OMe	H	4,5
2jj	OMe	OMe	OMe	H	H	H	F	OMe	F	H	4,8

2kk	H	H	F	H	H	H	H	F	H	H	4,9
2ll	H	H	F	H	H	H	O-CH ₂ -O		H	H	5,1
2mm	H	O-(CH ₂) ₂ -O		H	H	H	OMe	OMe	OMe	H	5,6
2nn	H	OMe	OMe	OMe	H	H	F	OMe	F	H	6,2
2oo	H	O-CH ₂ -O		H	H	OMe	H	H	OMe	H	6,4
2pp	H	O-CH ₂ -O		H	H	H	OMe	OMe	H	H	7,1
2qq	OMe	H	OMe	H	OMe	OMe	H	H	OMe	H	7,2
2rr	H	O-(CH ₂) ₂ -O		H	H	H	O-CH ₂ -O		H	H	8,0
2ss	H	H	OH	H	H	H	OMe	OMe	OMe	H	8,9
2tt	OH	H	OMe	H	H	H	H	Cl	H	H	9,7
2uu	H	H	OMe	H	H	H	H	OMe	H	H	9,9
2vv	H	OMe	OMe	OMe	H	H	H	OBn	H	H	9,9
2ww	H	OMe	OMe	H	H	H	H	OMe	H	H	10
2xx	H	H	Me	H	H	H	H	Me	H	H	11
2yy	H	H	OMe	H	H	H	H	OMe	H	H	15
2zz	OH	H	OMe	H	H	H	OMe	OMe	H	H	16
2α	H	H	Me	H	H	H	H	OMe	H	H	20
2β	H	H	Me	H	H	H	OMe	OMe	OMe	H	24

Estrutura base 3:

	R	R2	R3	R4	R5	R6	R2'	R3'	R4'	R5'	R6'	IC ₅₀ (μM)
3a	Me	H	OMe	OMe	OMe	H	H	OH	OMe	H	H	0,00021
3b	Pr	H	OMe	OMe	OMe	H	H	OH	OMe	H	H	0,0022
3c	Me	H	OMe	OMe	OMe	H	H	F	OMe	H	H	0,0020
3e	Me	OMe	H	H	OMe	H	H	H	NMe ₂	H	H	0,012
3g	Me	H	OMe	OMe	OMe	H	H	F	OMe	F	H	0,050
3h	Me	H	OMe	OMe	OMe	H	H	NO ₂	OMe	H	H	0,060
3i	Me	H	H	OMe	H	H	H	H	OMe	H	H	1,9
3j	Me	H	H	OMe	H	H	OMe	H	H	OMe	H	2,3
3k	Me	H	OMe	OMe	OMe	H	H	OMe	OMe	OMe	H	2,5

Estrutura base 4:

	R	R2	R3	R4	R5	R6	R2'	R3'	R4'	R5'	R6'	IC ₅₀ (μM)
4a	Me	H	OMe	OMe	OMe	H	H	OH	OMe	H	H	0,0015
4b	Et	H	OMe	OMe	OMe	H	H	OH	OMe	H	H	0,0026
4c	Me	H	OMe	OMe	OMe	H	H	F	OMe	H	H	0,0037
4d	Et	H	OMe	OMe	OMe	H	H	F	OMe	H	H	0,011
4e	Pr	H	OMe	OMe	OMe	H	H	F	OMe	H	H	0,020
4f	Et	H	OMe	OMe	OMe	H	H	F	OMe	F	H	0,22

4g	Et	H	OMe	OMe	OMe	H	H	F	OMe	F	H	0,23
4h	Me	H	OMe	OMe	OMe	H	H	F	OMe	F	H	0,36

Estrutura base 5:

	R2	R3	R4	R5	R6	R2'	R3'	R4'	R5'	R6'	IC ₅₀ (μM)
5a	H	OMe	OMe	OMe	H	H	H	Me	H	H	3,4
5b	H	OMe	OMe	OMe	H	H	OMe	OMe	OMe	H	3,7
5c	H	OMe	OMe	OMe	H	H	H	OMe	H	H	4,9
5d	H	H	Me	H	H	H	H	Me	H	H	12
5e	H	H	Me	H	H	H	H	OMe	H	H	15

Estruturas base 6 e 7, respectivamente:

	R	R2	R3	R4	R5	R6	R2'	R3'	R4'	R5'	R6'	IC ₅₀ (μM)
6a	Me	H	OMe	OMe	OMe	H	H	OH	OMe	H	H	0,09
6b	H	H	OMe	OMe	OMe	H	H	OH	OMe	H	H	0,5

6c	H	H	OMe	OMe	OMe	H	H	H	OMe	H	H	4,0
7a	Me	H	OMe	OMe	OMe	H	H	OH	OMe	H	H	0,6
7b	H	H	OMe	OMe	OMe	H	H	OH	OMe	H	H	1,0
7c	H	H	OMe	OMe	OMe	H	H	H	OMe	H	H	4,0

Estruturas base 8 e 9, respectivamente:

	R	R2	R3	R4	R5	R6	R2'	R3'	R4'	R5'	R6'	IC ₅₀ (μM)
8a	Me	H	OMe	OMe	OMe	H	H	OH	OMe	H	H	0,03
8b	H	H	OMe	OMe	OMe	H	H	OH	OMe	H	H	1,0
8c	H	H	OMe	OMe	OMe	H	H	H	OMe	H	H	10
8d	H	H	OMe	OMe	OMe	H	H	H	OMe	H	H	12
9a	H	H	OMe	OMe	OMe	H	H	H	OMe	H	H	0,5

Dessas moléculas, 80% foram selecionadas para constituir o conjunto de treinamento, que foi utilizado na construção do modelo, e 20% constituíram o conjunto de teste, responsável pela validação externa do modelo.

Para fazer os cálculos quânticos desse conjunto de moléculas foram utilizados computadores Core i7 do Grupo de Química Quântica do Instituto de Química de São Carlos, nos quais está instalado o software Gaussian 09⁵⁷. O método de cálculo DFT com o funcional híbrido B3LYP⁵⁸ e função de base 6-31g++(d,p)⁵⁹ foi utilizado a fim de otimizar a geometria dos compostos e calcular as frequências vibracionais. Também foram realizados cálculos com as moléculas em seus estados reduzido e oxidado,

com o intuito de estimar suas respectivas capacidades de doação e recepção de elétrons.

A partir desses cálculos, puderam ser obtidos os descritores eletrônicos. Após a otimização, as moléculas foram submetidas ao software Dragon⁶⁰, de onde foram obtidos os descritores moleculares.

Com o intuito de dar maior ênfase aos descritores com alta relevância, uma análise prévia do coeficiente de correlação de cada um dos descritores com a atividade biológica foi realizada, onde foram excluídos os descritores que apresentavam uma correlação menor que 0,3.

Uma seleção mais acurada das variáveis foi realizada através do método de Algoritmo Genético com o uso do software BuildQSAR⁶¹. Levou-se em consideração o número de gerações, descritores por modelo e alta correlação de variáveis. A escolha dos melhores modelos gerados foi baseada nos valores de R e de Q^2 .

Fazendo uso dos descritores mais significativos, cujos valores foram autoescalados, as análises de PLS foram realizadas através do software Pirouette 3.10⁶². O conjunto com todos os 87 compostos e também o conjunto de treinamento (70 compostos) foram estudados. Parâmetros estatísticos como Q^2 , R^2 , PRESS, SEC, SEV nortearam a escolha do melhor modelo. Também, processos de validação do modelo foram realizados, como o gráfico de resíduos de Student *versus* alavancagem (detecção de outliers), testes de leave-N-out (robustez) e randomização da atividade biológica (geração ao acaso). Todos com o uso do software QSAR Modeling⁶³.

Com as mesmas variáveis, outro modelo foi gerado a partir do método de Redes Neurais Artificiais, cujo treinamento da rede se deu através do software Matlab⁶⁴. O algoritmo utilizado foi o de retropropagação e as funções de transferência foram sigmoidais. Baseados na normalização de valores máximos e mínimos, os dados foram escalados em uma faixa de 0 a 1.

4 RESULTADOS

4.1 Seleção dos descritores

Após serem concluídos os cálculos de otimização de energia e frequências vibracionais de todos os compostos, puderam ser gerados 21 descritores eletrônicos e 748 moleculares, o que resultou num total de 769 descritores. Dentre os descritores eletrônicos gerados, podemos citar a energia da molécula, energia dos orbitais HOMO e LUMO, polarizabilidade isotrópica, energia do ponto zero, momento de dipolo, afinidade eletrônica, dureza, moleza, índice eletrofílico e eletropositividade.

A fim de selecionar os que melhor descrevem a atividade dos compostos estudados, foram excluídos, inicialmente, os que apresentavam uma correlação menor que 0,3 com as respectivas atividades biológicas. Ao final dessa etapa, restaram 386 descritores, os quais foram submetidos ao método de Algoritmo Genético.

Variando o número de gerações e eliminando variáveis com altas correlações entre si, chegou-se aos melhores modelos de Algoritmo Genético com 9.500 gerações e correlação menor que 0,9, com um número ideal de 7 descritores.

Dos 250 modelos gerados, um deles se destacou apresentando os descritores e parâmetros contidos na tabela 2, cujos parâmetros analisados foram o coeficiente de correlação (R), o desvio padrão (s), o teste F (F) e o coeficiente de correlação da validação cruzada (Q^2):

Tabela 2 – Valores de R, s, F e Q^2 do modelo gerado com o método de algoritmo genético.

Descritores selecionados	R	s	F	Q^2
E_{HOMO-3} , Polarizabilidade isotrópica, RDF045e, RDF155v, RDF035m, UNIP, SP02	0,87	0,591	35,017	0,71

Dentre todos os modelos gerados por esse método, foram selecionados os que apresentaram os melhores valores de R e de Q^2 . Também, foram contabilizadas as variáveis que apresentaram maior ocorrência nos mesmos. Dessas, foram selecionadas 20, no intuito de, a partir de análises de PLS, construir um modelo que não tenha sido gerado pelo algoritmo genético, ou porventura, validar os oriundos do mesmo.⁴³

4.2 Análises de PLS

Os testes realizados com o método de Algoritmo Genético auxiliaram nas análises de PLS fornecendo bons modelos e também uma série de variáveis com forte indicativo de influência na atividade biológica. Os melhores modelos foram analisados e também outros novos foram construídos a partir das variáveis com maior notabilidade.

Ao ser realizada a análise de PLS, percebeu-se que os modelos obtidos a partir do método de Algoritmo Genético apresentaram altos valores de variância e de percentual, como pode ser observado na tabela 3, exemplificados pelo melhor modelo. Isso motivou a busca por melhores modelos a partir das análises de PLS, já que altos valores desses parâmetros podem acarretar, em validações futuras, em predições não satisfatórias da atividade biológica, como também influenciar na estabilidade e robustez do modelo.

Tabela 3 – Parâmetros obtidos nas análises PLS para o modelo gerado pelo método de AG.

Variáveis latentes	6
Variância	85,756
Percentual	14,245
Acumulativo (%)	96,728
SEV	0,616
PRESS	33,042
SEC	0,587
Q²	0,709
R²	0,756

As análises de algoritmo genético admitem um número máximo de 7 variáveis. Já nas análises de PLS, com 7 variáveis, os valores dos parâmetros obtidos não foram satisfatórios, da mesma forma que com 9 variáveis. Assim, verificou-se que o número de 8 variáveis era o número ideal.

Isso mostra que, apesar do método de algoritmo genético ser capaz prever bons modelos com valores de R de Q² satisfatórios, alguns parâmetros estatísticos podem passar despercebidos. E essas escolhas podem ser decisivas na geração de um bom modelo.

As análises de PLS indicaram quatro modelos com alta capacidade de predição da atividade biológica. O primeiro deles, o modelo A, é constituído pelas 87 moléculas estudadas neste trabalho. Os outros três modelos, B1, B2 e B3, são formados por um conjunto de treinamento com 70 moléculas e um conjunto de teste com 17 moléculas, cada.

O modelo A, com todos os compostos no conjunto de treinamento, apresentou os seguintes valores:

Tabela 4 – Parâmetros obtidos pelo método de PLS para o modelo A.								
VL	1	2	3	4	5	6	7	8
Variância	346,285	49,780	37,767	51,186	77,299	79,850	25,960	19,874
Percentual	50,332	7,235	5,489	7,440	11,235	11,606	3,773	2,889
Acumulativo (%)	50,332	57,568	63,057	70,497	81,732	93,338	97,111	100,000
SEV	0,907	0,839	0,686	0,664	0,603	0,604	0,603	0,603
PRESS	71,541	61,205	40,924	38,389	31,637	31,764	31,635	31,615
Q²	0,370	0,469	0,643	0,663	0,721	0,720	0,722	0,722
SEC	0,876	0,727	0,624	0,589	0,575	0,568	0,571	0,575
R²	0,423	0,608	0,714	0,750	0,764	0,771	0,772	0,772

Como pode ser observado na tabela acima, o melhor modelo é obtido com 7 variáveis latentes, descrevendo 97% da informação original. À medida que aumentam o número de variáveis latentes, diminuem os erros de predição e validação e aumentam os valores de Q^2 e R^2 . Conforme recomenda a literatura, bons valores de Q^2 e R^2 são valores acima de 0,5 e 0,6, respectivamente. Nota-se, nesse caso, a considerável diminuição nos valores da variância e do percentual, bem como melhoras nos demais parâmetros. Isso mostra que o acréscimo de um descritor em relação ao modelo obtido por Algoritmo Genético contribui significativamente para a construção de um melhor modelo.⁴³

A seguir encontram-se os respectivos pesos de cada um dos descritores na predição da atividade biológica, onde os descritores que apresentaram maior contribuição para o modelo foram: RDF045e, RTv, SP02 e E_{HOMO-3} :

Tabela 5 – Coeficientes dos descritores constituintes do modelo A.

E_{HOMO-3}	0.341758
UNIP	0.058915
Polarizabilidade isotrópica (PI)	-0.188553
SP02	0.446925
RDF035m	0.043860
RDF155u	-0.208197
RTv	0.452270
RDF045e	-0.628187

A tabela 6 apresenta o valor dos descritores para cada um dos compostos:

Tabela 6 – Respective valores dos descritores para cada um dos compostos.

Composto	E_{HOMO-3}	UNIP	PI	SP02	RDF035m	RDF155u	RTv	RDF045e
2a	-6,207	96	263,480	6	4	6	6	18
2aa	-6,582	79	240,490	5	2	0	6	12
2b	-6,445	96	269,920	6	4	6	6	18
2bb	-6,561	97	260,010	6	4	0	6	19
2c	-6,332	94	261,430	6	4	3	6	22
2cc	-6,499	97	261,720	6	3	0	5	22
2d	-6,013	85	266,390	5	2	0	6	16
2dd	-7,018	87	294,650	5	4	0	7	15
2e	-6,494	103	271,170	6	4	1	7	17
2ee	-6,889	46	174,410	5	0	0	6	4
2f	-6,495	112	293,870	6	5	6	6	20
2ff	-6,646	83	235,530	5	2	0	6	14
2g	-6,477	103	276,470	6	4	6	6	19
2gg	-6,518	108	267,910	6	6	2	6	22
2h	-6,504	95	258,220	6	6	0	6	15
2hh	-6,389	97	258,240	6	3	0	6	20
2i	-6,565	96	259,940	6	4	0	6	17
2ii	-6,609	92	259,000	6	2	1	5	20
2j	-6,402	90	262,040	6	4	1	6	19
2jj	-6,518	100	259,100	6	5	2	6	18
2k	-6,169	101	278,830	6	2	3	6	17

2kk	-6,862	58	177,860	5	2	0	6	9
2l	-6,316	112	282,270	6	5	0	7	21
2ll	-6,701	69	202,930	5	4	0	5	10
2m	-6,311	88	257,830	6	4	3	6	20
2mm	-6,513	105	272,380	6	3	6	6	19
2n	-6,345	92	251,760	6	2	0	6	14
2nn	-6,714	102	261,090	6	5	3	6	14
2o	-6,392	94	257,840	6	4	1	6	20
2oo	-6,407	83	238,730	5	3	1	5	12
2p	-6,266	101	276,460	6	4	0	6	20
2pp	-6,570	87	245,270	5	4	2	6	14
2q	-6,480	105	276,340	6	3	0	6	23
2qq	-6,232	95	264,280	5	3	0	6	17
2r	-6,472	88	247,030	5	4	0	6	17
2rr	-6,557	86	240,610	5	3	0	6	12
2s	-6,569	70	222,390	5	1	0	5	12
2ss	-6,605	83	242,020	5	2	0	6	18
2t	-6,298	116	291,110	6	5	1	6	24
2tt	-6,673	67	217,660	5	1	0	6	9
2u	-6,620	81	249,950	5	3	0	6	13
2uu	-6,722	72	226,280	5	1	1	5	11
2v	-6,605	66	194,990	5	2	0	6	10
2vv	-6,314	149	325,700	6	5	9	7	21
2w	-6,994	75	264,890	5	2	0	6	17
2ww	-6,530	81	244,920	5	1	4	5	15
2x	-6,682	83	235,460	5	3	0	6	19
2xx	-6,616	58	208,440	5	0	0	5	7
2y	-6,309	86	245,070	5	3	0	6	20
2yy	-6,722	72	226,280	5	1	1	5	11
2z	-6,521	94	261,560	6	2	4	6	18
2zz	-6,381	87	243,540	5	2	0	6	15
2α	-6,679	65	218,700	5	1	0	5	9
2β	-6,534	83	248,990	5	2	0	6	17
3a	-6,459	98	268,150	6	5	6	7	27
3b	-6,384	105	279,620	6	7	4	8	37
3c	-6,539	98	264,980	6	5	2	7	27
3e	-5,838	87	273,740	5	2	3	7	23

3g	-6,655	104	263,790	6	6	1	7	22
3h	-6,711	112	274,190	6	7	3	7	27
3i	-6,651	73	230,590	5	2	1	6	17
3j	-6,449	80	234,200	5	2	0	7	21
3k	-6,476	117	292,920	6	6	1	7	29
4a	-6,473	101	276,840	6	7	7	7	25
4b	-6,489	105	287,230	6	7	7	7	32
4c	-6,567	101	272,960	6	7	1	7	22
4d	-6,584	105	283,010	6	7	2	7	28
4e	-6,587	109	294,100	6	7	1	7	27
4f	-6,700	114	294,860	6	8	2	7	25
4g	-6,691	110	282,030	6	9	2	7	26
4h	-6,669	107	274,200	6	7	0	6	22
5a	-6,835	84	217,420	5	5	0	8	21
5b	-6,691	117	266,290	6	6	0	7	22
5c	-6,666	92	231,180	6	4	1	7	19
5d	-6,873	59	188,060	5	0	0	7	9
5e	-6,644	75	197,860	5	4	0	7	10
6a	-6,514	98	250,930	5	7	0	7	31
6b	-6,660	96	236,080	5	7	0	7	24
6c	-6,642	90	238,000	5	6	0	7	21
7a	-6,457	98	237,560	5	7	0	7	36
7b	-6,438	96	235,460	5	5	0	7	27
7c	-6,507	90	222,390	5	4	0	7	24
8a	-6,585	98	242,700	5	8	0	7	30
8b	-6,563	96	230,180	6	5	0	7	22
8c	-6,621	90	225,950	5	4	0	7	19
8d	-6,749	92	225,700	6	3	0	7	19
9a	-6,404	88	221,050	5	5	0	7	22

Amostras com algum tipo de anomalia em relação ao modelo puderam ser identificadas a partir do gráfico de alavancagem *versus* resíduos de Student, onde nota-se que as moléculas 3e, 2dd e 2vv apresentaram alto valor de alavancagem e limites aceitáveis de resíduos de Student, como também as 2m, 3a e 3c têm altos resíduos de Student, porém aceitáveis valores de alavancagem. Conforme demonstra

a figura 6, não foram detectados *outliers*, dessa forma, esse modelo pôde ser gerado com a participação de todos os compostos.

Figura 6 – Detecção de *outliers* do modelo A.

A fim de verificar se tal modelo é robusto e suas correlações não foram geradas ao acaso, foram realizados testes de *leave-N-out* e randomização. Para os testes de *leave-N-out*, foram excluídos até 20 compostos, com 50 repetições para cada valor de N. Como pode-se observar na figura 7, os valores de Q^2 mantiveram-se maiores que 0,7 e os desvios apresentam-se um tanto quanto menores que 0,1, mostrando que o modelo é robusto.⁴³

Figura 7 – Teste de *leave-N-out* aplicado ao modelo A.

Os testes de randomização da atividade biológica mostraram que não foram geradas correlações ao acaso, já que os valores obtidos para Q^2 e R^2 foram menores que 0,4, conforme recomenda a literatura.⁴³

Figura 8 – Teste de randomização do modelo A.

Na figura 9, está apresentado o gráfico da atividade predita (y predito) *versus* a atividade mensurada (y mensurada). Como podemos ver, com exceção de alguns poucos desvios, a grande maioria dos valores está em conformidade com a reta.

Figura 9 – Predição da atividade do modelo A.

Com base nos dados apresentados, pode-se afirmar que o modelo A é capaz de prever satisfatoriamente a atividade biológica dessa classe de compostos.

Após a geração do modelo com todos os compostos, alguns compostos (aproximadamente 20%) foram selecionados para constituírem o conjunto de teste. Os demais constituíram o conjunto de treinamento. Ao final, três conjuntos de treinamento e teste apresentaram os melhores modelos.

Os modelos B1, B2 e B3, constituídos por 70 compostos no conjunto de treinamento e 17 no conjunto de teste, foram gerados com os mesmos descritores do modelo A, porém com pesos diferentes. Os pesos dos três respectivos modelos estão apresentados na tabela 7:

Tabela 7 - Coeficientes dos descritores constituintes dos modelos B1, B2 e B3.

	B1	B2	B3
E_{HOMO-3}	0.156686	0.200289	-0.041363
UNIP	-0.011797	-0.035156	-0.017041
PI	-0.252282	0.050814	0.448960
SP02	0.507394	0.561700	0.066983
RDF035m	-0.243477	-0.439594	-0.592550
RDF155u	-0.117437	-0.405725	-0.109842
RTv	0.640743	0.468413	0.612678
RDF045e	-0.413027	-0.251581	-0.231066

Nas tabelas 8, 9 e 10 estão apresentados valores dos parâmetros obtidos para cada um desses modelos. Nota-se que os valores ideais são alcançados quando o modelo é apresentado com 7 variáveis latentes.

Tabela 8 – Parâmetros obtidos pelo método de PLS para o modelo B1.

VL	1	2	3	4	5	6	7	8
Variância	262,221	51,461	47,938	73,541	49,982	30,490	21,514	14,853
Percentual	47,504	9,323	8,684	13,323	9,055	5,524	3,897	2,691
Acumulativo (%)	47,504	56,826	65,511	78,834	87,888	93,412	97,309	100,000
SEV	0,958	0,885	0,804	0,696	0,664	0,634	0,634	0,634
PRESS	64,196	54,818	45,282	33,878	30,878	28,138	28,127	28,139
Q²	0,351	0,457	0,555	0,657	0,687	0,715	0,715	0,715
SEC	0,911	0,749	0,658	0,624	0,602	0,591	0,596	0,601
R²	0,424	0,616	0,709	0,741	0,763	0,775	0,776	0,776

Tabela 9 – Parâmetros obtidos pelo método de PLS para o modelo B2.

VL	1	2	3	4	5	6	7	8
Variância	292,418	34,494	43,363	68,429	35,421	39,842	23,694	14,339
Percentual	52,974	6,249	7,856	12,397	6,417	7,218	4,292	2,598
Acumulativo (%)	52,974	59,223	67,079	79,475	85,892	93,110	97,402	100,000
SEV	0,935	0,883	0,703	0,648	0,628	0,602	0,596	0,596
PRESS	61,208	54,568	34,557	29,356	27,595	25,404	24,896	24,881
Q²	0,358	0,441	0,638	0,692	0,710	0,734	0,739	0,739
SEC	0,895	0,711	0,631	0,600	0,572	0,558	0,560	0,565
R²	0,424	0,643	0,722	0,753	0,779	0,793	0,794	0,794

Tabela 10 – Parâmetros obtidos pelo método de PLS para o modelo B3.

VL	1	2	3	4	5	6	7	8
Variância	278,168	47,503	20,950	57,536	71,050	44,469	19,508	12,817
Percentual	50,393	86,056	37,952	10,423	12,871	8,056	3,534	2,322
Acumulativo(%)	50,393	58,998	62,794	73,217	86,088	94,144	97,678	100,000
SEV	0,869	0,802	0,723	0,731	0,671	0,650	0,649	0,649
PRESS	52,901	45,023	36,630	37,429	31,521	29,558	29,522	29,462
Q²	0,426	0,514	0,606	0,598	0,661	0,681	0,681	0,682
SEC	0,843	0,726	0,661	0,626	0,608	0,597	0,601	0,606
R²	0,474	0,615	0,687	0,723	0,743	0,756	0,756	0,756

Da mesma forma que no modelo A, em nenhum desses modelos foram detectados *outliers*, como confirmam as figuras 10, 11 e 12:

Figura 10 – Detecção de *outliers* do modelo B1.

Figura 11 – Detecção de *outliers* do modelo B2.

Figura 12 – Detecção de *outliers* do modelo B3.

Os conjuntos de teste selecionados foram utilizados para as validações externas dos modelos, cujos respectivos gráficos de atividade predita *versus* atividade calculada encontram-se nas figuras 13, 14 e 15:

Figura 13 – Previsão da atividade do modelo B1.

Figura 14 – Previsão da atividade do modelo B2.

Figura 15 – Previsão da atividade do modelo B3.

O que também pode ser verificado através do cálculo dos respectivos resíduos, que como pode-se notar nas tabelas 11, 12 e 13, encontram-se dentro do admitido, já que apresentam valores menores que 1:

Tabela 11 – Valores dos resíduos do modelo B1.

	pIC₅₀ experimental	pIC₅₀ predito	Resíduo
2c	8,000	7,273	0,727
2dd	5,420	6,384	-0,964
2ee	5,420	5,290	0,131
2ff	5,398	4,737	0,661
2kk	5,310	5,033	0,277
2l	6,523	6,241	0,282
2mm	5,252	6,166	-0,914
2tt	5,013	5,285	-0,271
2u	5,824	5,627	0,197
2v	5,824	5,212	0,612
2ww	5,000	5,377	-0,377
2y	5,770	5,906	-0,137
2z	5,658	6,316	-0,659
3j	5,638	6,183	-0,545
4d	7,959	7,749	0,209
5d	4,921	5,080	-0,160
6b	6,301	5,790	0,511

Tabela 12 – Valores dos resíduos do modelo B2.

	pIC₅₀ experimental	pIC₅₀ predito	Resíduo
2aa	5,569	5,279	0,290
2b	8,000	7,337	0,663
2dd	5,420	6,304	-0,884
2gg	5,398	6,183	-0,785
2jj	5,319	6,250	-0,931
2q	5,959	5,503	0,455
2v	5,824	5,276	0,548
2zz	4,796	5,058	-0,262
3g	7,301	6,506	0,795
4b	8,585	9,036	-0,451
4f	6,658	7,188	-0,530
4h	6,444	6,214	0,230
5c	5,310	6,027	-0,718

6c	5,398	6,182	-0,784
7a	6,222	6,729	-0,508
8a	7,523	6,692	0,831
9a	6,301	6,101	0,200

Tabela 13 – Valores dos resíduos do modelo B3.

	pIC₅₀ experimental	pIC₅₀ predito	Resíduo
2g	6,523	6,749	-0,226
2p	5,959	6,634	-0,676
2rr	5,097	4,879	0,217
2u	5,824	5,534	0,290
2v	5,824	5,467	0,356
2vv	5,004	4,643	0,361
2xx	4,959	5,321	-0,363
3e	7,921	8,456	-0,536
3j	5,638	6,275	-0,637
4b	8,585	9,022	-0,437
4e	7,699	7,359	0,340
5b	5,432	4,874	0,558
5c	5,310	6,199	-0,889
5e	4,824	5,606	-0,782
6a	7,046	6,668	0,378
8b	6,000	6,319	-0,319
8d	4,921	5,397	-0,477

A robustez desses modelos também foi testada através dos testes de *leave-N-out*, apresentados nas figuras 16, 17 e 18. Nesses casos, com até 16 amostras retiradas e 35 repetições:

Figura 16 – Teste *leave-N-out* do modelo B1. Figura 17 – Teste *leave-N-out* do modelo B2.

Figura 18 – Teste de *leave-N-out* do modelo B3.

os quais também apresentaram valores seguramente dentro do admitido.

Os testes de randomização das atividades foram aplicados nesses modelos, a fim de garantir que nenhum deles foi gerado ao acaso⁴³:

Figura 19 – Teste de randomização do modelo B1.

Figura 20 – Teste de randomização aplicado ao modelo B2.

Figura 21 – Teste de randomização aplicado ao modelo B3.

Assim, através das validações aqui apresentadas, bem como os parâmetros estatísticos, a escolha desses três modelos é justificada.

Dentre os modelos gerados por esse método, os modelos B1 e B3 foram os que apresentaram as melhores capacidades preditivas, apresentando valores médios de resíduos relativos de 7,8% e 8%, respectivamente. Para o modelo B2, esse valor foi de 9,6%.

4.3 Análises de ANN

Três modelos foram apresentados como os melhores gerados pela análise de ANN. Esses modelos, denominados C1, C2 e C3, apresentam o mesmo conjunto de treinamento e de teste que os respectivos modelos B1, B2 e B3 gerados nas análises de PLS. Dessa forma, torna-se passível a investigação da existência de relação não-linear entre as variáveis e a atividade biológica, já que a análise de ANN está baseada em métodos não-lineares.

Foram utilizados 8 neurônios na camada de entrada. E, a partir dos testes realizados, chegou-se a um número ideal de neurônios para cada um dos modelos obtidos. Para o modelo C1, o melhor número de neurônios foi 10. Para C2 e C3 foram 12. A função que apresentou o melhor desempenho foi a função sigmoideal – tansig (simétrica). As redes de alimentação utilizadas no treinamento foram do tipo MLP-ANN.

Os valores de R^2 e Q^2 para os três modelos estão apresentados a seguir:

Tabela 14 – Valores de Q^2 , R^2 dos conjuntos de treinam. e teste dos modelos C1, C2 e C3.

	Q^2	R^2 treinam.	R^2 teste
C1	0,557	0,827	0,753
C2	0,842	0,799	0,482
C3	0,761	0,803	0,663

Como pode-se observar, para o modelo C2, os valores de R^2 e Q^2 obtidos não foram satisfatórios, onde apenas os modelos C1 e C3 apresentaram valores aceitáveis para os mesmos.

Nas figuras 22, 23 e 24, são apresentados os gráficos dos valores da atividade preditos *versus* os valores experimentais:

Figura 22 – Predição da atividade do modelo C1.

Figura 23 – Predição da atividade do modelo C2.

Figura 24 – Predição da atividade do modelo C3.

Nas tabelas de resíduos 15, 16 e 17, pode ser observada a capacidade de predição de cada um dos modelos, onde nota-se um considerável aumento nos valores dos resíduos:

Tabela 15 – Valores dos resíduos do modelo C1.

	pIC ₅₀ experimental	pIC ₅₀ predito	Resíduo
2c	8,000	9,517	1,517
2dd	5,420	4,043	-1,377
2ee	5,420	5,905	0,485
2ff	5,398	5,674	0,276
2kk	5,310	5,728	0,418
2l	6,523	6,260	0,968
2mm	5,252	4,028	-1,224
2tt	5,013	4,893	-0,120
2u	5,824	4,911	-0,913
2v	5,824	6,820	0,996
2ww	5,000	4,408	-0,592
2y	5,770	5,780	0,011
2z	5,658	4,975	-0,682
3j	5,638	5,144	-0,495
4d	7,959	8,455	0,497
5d	4,921	4,936	0,015
6b	6,301	6,445	0,144

Tabela 16 – Valores dos resíduos do modelo C2.

	pIC₅₀ experimental	pIC₅₀ predito	Resíduo
2aa	5,569	5,659	0,091
2b	8,000	9,545	1,545
2dd	5,420	3,561	-1,859
2gg	5,398	4,764	-0,634
2jj	5,319	4,550	-0,768
2q	5,959	6,629	0,670
2v	5,824	5,923	0,099
2zz	4,796	4,095	-0,701
3g	7,301	8,317	1,017
4b	8,585	7,570	-1,015
4f	6,658	5,809	-0,848
4h	6,444	7,486	1,043
5c	5,310	5,123	-0,186
6c	5,398	3,993	-1,405
7a	6,222	5,402	-0,820
8a	7,523	8,696	1,174
9a	6,301	6,116	-0,185

Tabela 17 – Valores dos resíduos do modelo C3.

	pIC₅₀ experimental	pIC₅₀ predito	Resíduo
2g	6,523	6,571	0,049
2p	5,959	5,617	-0,342
2rr	5,097	4,813	-0,284
2u	5,824	6,534	0,710
2v	5,824	5,500	-0,324
2vv	5,004	3,159	-1,846
2xx	4,959	4,568	-0,391
3e	7,921	8,066	0,145
3j	5,638	5,202	-0,436
4b	8,585	7,861	-0,724
4e	7,699	9,430	1,731
5b	5,432	4,555	-0,876
5c	5,310	4,855	-0,454

5e	4,824	3,721	-1,103
6a	7,046	6,917	-0,129
8b	6,000	6,446	0,446
8d	4,921	4,649	-0,272

Dessa forma, as análises realizadas através dos métodos de PLS e de ANN podem ser comparadas. Na tabela 18, é feita uma comparação dos resíduos de cada um dos modelos PLS e ANN:

Tabela 18 – Valores dos resíduos e valores médios dos resíduos dos conjuntos de teste dos modelos analisados pelos métodos de PLS e ANN.

Modelo 1 (B1/C1)			Modelo 2 (B2/C2)			Modelo 3 (B3/C3)		
	PLS	ANN		PLS	ANN		PLS	ANN
2c	0,727	1,517	2aa	0,290	0,091	2g	-0,226	0,049
2dd	-0,964	-1,377	2b	0,663	1,545	2p	-0,676	-0,342
2ee	0,131	0,485	2dd	-0,884	-1,859	2rr	0,217	-0,284
2ff	0,661	0,276	2gg	-0,785	-0,634	2u	0,290	0,710
2kk	0,277	0,418	2jj	-0,931	-0,768	2v	0,356	-0,324
2l	0,282	0,968	2q	0,455	0,670	2vv	0,361	-1,846
2mm	-0,914	-1,224	2v	0,548	0,099	2xx	-0,363	-0,391
2tt	-0,271	-0,120	2zz	-0,262	-0,701	3e	-0,536	0,145
2u	0,197	-0,913	3g	0,795	1,017	3j	-0,637	-0,436
2v	0,612	0,996	4b	-0,451	-1,015	4b	-0,437	-0,724
2ww	-0,377	-0,592	4f	-0,530	-0,848	4e	0,340	1,731
2y	-0,137	0,011	4h	0,230	1,043	5b	0,558	-0,876
2z	-0,659	-0,682	5c	-0,718	-0,186	5c	-0,889	-0,454
3j	-0,545	-0,495	6c	-0,784	-1,405	5e	-0,782	-1,103
4d	0,209	0,497	7a	-0,508	-0,820	6a	0,378	-0,129
5d	-0,160	0,015	8a	0,831	1,174	8b	-0,319	0,446
6b	0,511	0,144	9a	0,200	-0,185	8d	-0,477	-0,272
Média: 0,449 0,631			0,580 0,827			0,461 0,604		

Observando a tabela, nota-se que os modelos gerados a partir do método de PLS são consideravelmente superiores aos gerados pelas ANN. Isso mostra que os descritores selecionados descrevem mais fielmente a atividade biológica quando

sujeitos a relações lineares. Apesar dos resultados das análises de ANN terem sido inferiores aos do PLS, eles mostram que essas variáveis podem ser, de alguma forma, mesmo que menos precisas, relacionadas de maneira não-linear, o que mostra que a relação entre os descritores selecionados é altamente consistente.

4.4 Análise dos descritores selecionados

4.4.1 Orbital HOMO-3:

Nas figuras 25-29 estão apresentadas as respectivas ilustrações dos orbitais HOMO-3 das moléculas estudadas com os maiores valores de atividade biológica:

Figura 25 – Representação do orbital HOMO-3 do composto 2a (IC_{50} : 0,0043 μ M).

Figura 26 – Representação do orbital HOMO-3 do composto 3a (IC_{50} : 0,00021 μ M).

Figura 27 – Representação do orbital HOMO-3 do composto 3b (IC_{50} : 0,0022 μ M).

Figura 28 – Representação do orbital HOMO-3 do composto 4a (IC_{50} : 0,0015 μ M).

Figura 29 – Representação do orbital HOMO-3 do composto 4b (IC_{50} : 0,0026 μ M).

Para as moléculas com os menores valores de atividade, temos as representações nas figuras 30 e 31:

Figura 30 – Representação do orbital HOMO-3 do composto 2 α (IC₅₀: 20 μ M).

Figura 31 – Representação do orbital HOMO-3 do composto 2 β (IC₅₀: 24 μ M).

Nitidamente, a presença do orbital HOMO-3 em determinadas regiões das moléculas influencia nas suas respectivas capacidades de interação com a tubulina, tendo em vista o alto contraste da representação entre as moléculas mais ativas e menos ativas, concordando com o procedimento estatístico anteriormente realizado.

4.4.2 Descritores RDF

A função de distribuição radial (RDF) de um conjunto de N átomos pode ser interpretado como a distribuição de probabilidade de encontrar um átomo em um volume esférico de raio r . A equação abaixo representa a função de distribuição radial:

$$g(r) = f \sum_i^{N-1} \sum_{j>i}^N A_i A_j e^{-B(r-r_{ij})^2} \quad (55)$$

onde:

f é um fator de escalonamento;

N é o número de átomos;

A: propriedades características atômicas dos átomos i e j (A função RDF pode ser usada em diferentes tarefas para servir aos requerimentos das informações a serem representadas).

O termo exponencial contém a distância r_{ij} entre os átomos i e j e o parâmetro de suavização B, que define a distribuição da probabilidade das distâncias individuais.

As propriedades atômicas A_i e A_j usadas na equação, permitem a discriminação dos átomos de uma molécula para quase todas as propriedades que podem ser atribuídas a um átomo.

Os descritores da Função de Distribuição Radial (RDF) são baseados na distribuição das distâncias dos compostos. Esses descritores 3D sugerem a ocorrência de alguma dependência linear entre a atividade dos compostos e a distribuição molecular 3D de certos átomos para um dado valor de raio a partir do centro geométrico da molécula.^{65,66}

Baseando-se nessas informações, podemos discorrer sobre a influência dos descritores RDF obtidos:

RDF045e: Distribuição dos átomos num raio de 4,5Å, baseada na eletronegatividade de Sanderson. Esse descritor destaca a importância do ambiente eletrostático que rodeia o centro da molécula nesse valor de raio.

RDF155u: Distribuição de quaisquer tipos de átomos num raio de 15,5Å. Mostra que para raios dessa magnitude, a presença de um grande número de átomos é fundamental para serem obtidos altos valores de atividade biológica dos compostos.

RDF035m: Distribuição dos átomos num raio de 3,5Å, baseada nas massas atômicas. Destaca a importância dos átomos com maior massa próximos ao centro geométrico da molécula.

4.4.3 Descritores SP02 (*Shape Profile* nº 02)

A forma molecular é algo bastante subjetivo quando abordado simplesmente por aspectos gráficos ou visuais. Para um estudo mais preciso, devem ser levados em consideração fatores numéricos nesse estudo.

Uma representação mais precisa da forma molecular deve estar baseada em detalhes da geometria molecular. Tais informações podem ser fornecidas por uma lista de todas as distâncias interatômicas, que podem ser convenientemente representadas com uma matriz de distâncias. Referimo-nos a elas como caracterização da forma ou expansão da forma, em analogia com as expansões de funções potenciais. O procedimento fornece uma caracterização para qualquer forma, para que dessa maneira uma função simples possa ser atribuída. A significância desse resultado é que sucedemos na representação uma forma em uma função simples bem definida. Essas funções são denominadas perfis de forma molecular. Aqui está o esboço do método que gera a representação. Inicia-se com a matriz de distância de uma estrutura:

$$D = \begin{pmatrix} d_{1,1} & d_{1,2} & d_{1,3} & d_{1,4} & \dots \\ d_{2,1} & d_{2,2} & d_{2,3} & d_{2,4} & \dots \\ d_{3,1} & d_{3,2} & d_{3,3} & d_{3,4} & \dots \\ d_{4,1} & d_{4,2} & d_{4,3} & d_{4,4} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix} \quad (56)$$

Da qual é extraída a soma das linhas ou colunas. A partir da qual será obtida a média da soma das linhas: $(R = R_1 + R_2 + R_3 + \dots) / N$, onde N é o número de átomos na estrutura. O segundo passo é considerar a matriz:

$$D^2 = \begin{pmatrix} d_{1,1}^2 & d_{1,2}^2 & d_{1,3}^2 & d_{1,4}^2 & \dots \\ d_{2,1}^2 & d_{2,2}^2 & d_{2,3}^2 & d_{2,4}^2 & \dots \\ d_{3,1}^2 & d_{3,2}^2 & d_{3,3}^2 & d_{3,4}^2 & \dots \\ d_{4,1}^2 & d_{4,2}^2 & d_{4,3}^2 & d_{4,4}^2 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix} \quad (57)$$

onde d_{ij}^2 é o quadrado do elemento d_{ij} . Novamente consideramos a soma das linhas e construímos a média da soma das linhas: $R^2 = (R_1^2 + R_2^2 + R_3^2 + \dots) / N$. O processo

se estende para matrizes construídas a partir dos elementos obtidos das maiores energias de d_{ij} .

Dessa forma, obtemos a sequência: $R, R^2, R^3, R^4, R^5, R^6, \dots$. Com o intuito de reduzir o papel de qualquer aumento de energia, normalizamos a média das somas das linhas pelo fatorial do expoente usado. Então, obtemos a sequência: $R^1, R^2/2!, R^3/3!, \dots$, o que pode ser alternativamente representado como uma série de potência:

$$S = N + R^1x + R^2/2!x^2 + R^3/3!x^3 + \dots \quad (58)$$

ou então pela sequência:

$$S = S^1, S^2, S^3, S^4, S^5, S^6, \dots \quad (59)$$

Podemos formalmente introduzir a constante N , que indica o tamanho do sistema, como termo principal da sequência: $S = S^0, S^1, S^2, S^3, \dots$. O termo principal S^0 pode ser visto como derivado da matriz d_{ij}^0 .

Os descritores da forma $S^1, S^2, S^3, S^4, S^5, S^6$, ou a sequência S , representam o perfil molecular. Por causa da presença dos fatoriais nos denominadores a sequência irá sempre convergir.^{67,68}

Baseados nos parâmetros de classificação e normalização propostos por Randic, os perfis das mais variadas estruturas podem ser resumidos pelas 12 formas apresentadas na figura 32. Dos perfis sugeridos por Randic, o descritor SP02 (*shape profile* nº 02) tem a forma II, onde a forma I remete a moléculas lineares:

Figura 32 – Perfis de formas propostos por Randic.

Fonte: RANDIC, M., 1995, p. 373-382.

Como pode ser visualizado, essa forma genérica apresenta grande semelhança com as estruturas dos compostos estudados. E as análises estatísticas apontaram sua influência na atividade biológica dos compostos aqui estudados.

4.4.4 Descritor RTv

Seja M a matriz molecular constituída por a linhas correspondentes aos átomos da molécula e três colunas correspondentes às coordenadas cartesianas x, y e z de cada átomo, com respeito ao centro geométrico da molécula. A matriz de influência molecular (MIM), denotada por H, é calculada a partir da matriz M como:

$$H = M \cdot (M^T \cdot M)^{-1} \cdot M^T \quad (60)$$

onde T se refere à matriz transposta.

Esse descritor combina a informação provinda da matriz de influência molecular H com a da matriz de geometria G, cujos elementos r_{ij} são distâncias geométricas 3D entre cada par de átomos i e j.

São quantidades matemáticas calculadas a partir de duas novas representações moleculares dependentes da geometria molecular: a matriz de influência molecular e a matriz de influência/distância.

Esses descritores carregam consigo características do tipo: (a) eles contêm informação 3D, (b) suas definições funcionais estão baseadas em informações e formulas bem conhecidas, (c) podem ser facilmente calculados, (d) mostram bom poder de predição em modelagens de propriedades físico-químicas, (e) comprimento uniforme adequado para análises de similaridade/diversidade.^{69,70}

Dessa forma, nota-se que as coordenadas dos átomos e suas distâncias geométricas, que são as características nas quais está embasado o descritor RTv, apresentam considerável influência na atividade biológica dos compostos estudados.

4.4.5 Descritor UNIP

É um descritor obtido a partir de um grafo molecular 2D. O grafo é um conceito topológico cuja representação é constituída de vértices representada pelos átomos, e linhas, representadas pelas ligações. Para simplificar sua manipulação, representam-

se os grafos omitindo-se os átomos de hidrogênio e suas ligações, apresentando apenas o esqueleto molecular.

Muitas propriedades de grafos são determinadas pela menor distância entre seus vértices. A característica excêntrica, que está baseada na excentricidade do vértice, é a característica fundamental do descritor UNIP (unipolaridade), que é o menor comprimento do grafo molecular. Ou seja, está relacionado com a compactação da molécula. Portanto, quanto maior a unipolaridade, mais compacta e mais reativa será a molécula.⁷¹

4.4.6 Polarizabilidade isotrópica

A polarizabilidade (α) é definida como o momento de dipolo induzido por unidade da intensidade do campo elétrico conforme a equação:

$$\alpha = \frac{\mu_{ind}}{E} \quad (61)$$

É uma grandeza que determina a facilidade com que uma densidade de carga eletrônica de um átomo (molécula) pode ser distorcida, em função do efeito do dipolo de um campo elétrico externo que promove uma maior deslocalização.⁷²

A polarizabilidade isotrópica sugere que essa distorção ocorra da maneira mais uniforme possível.

Assim, pode-se perceber que as moléculas que tendem a ser eletronicamente mais equilibradas na sua estrutura, apresentam maiores valores de polarizabilidade isotrópica (acima de 250 em geral), o que torna esses compostos mais ativos frente à inibição da tubulina.

5 CONCLUSÃO E PERSPECTIVAS

Estruturas químicas e atividades biológicas foram aqui relacionadas conforme proposto no início deste trabalho. Relações essas que se mostraram capazes de prever satisfatoriamente a atividade de novos compostos pertencentes à classe analisada.

A escolha do método quântico adequado, bem como o funcional e a função de base, foram fundamentais para a correta obtenção dos descritores eletrônicos, que se mostraram de grande influência na predição da atividade biológica desses compostos. Mesmo não participando diretamente da geração do modelo, o método de Algoritmo Genético foi essencial na seleção dos melhores descritores, que posteriormente foram corretamente relacionados nas análises de PLS.

O modelo principal (modelo A) foi gerado com um número ideal de oito variáveis, apresentando os melhores resultados para todos os parâmetros estatísticos analisados. Para sua validação interna, as análises realizadas mostraram que o modelo é estável, robusto e que não foram geradas correlações ao acaso. Para sua validação externa, foram formados três conjuntos de treinamento e três conjuntos de teste, os quais foram submetidos às análises de PLS e de ANN.

Dessas análises, as que apresentaram os melhores resultados foram as análises de PLS, o que denuncia uma predominância de comportamento linear entre os descritores e a atividade biológica. Entretanto, as análises de ANN também apresentaram alguns resultados satisfatórios, indicando que alguma reação não-linear também é possível entre os mesmos.

Além dos parâmetros estatísticos e das validações interna e externa, o significado químico dos descritores obtidos é de extrema importância num estudo de QSAR. Neste estudo, os descritores selecionados para representar a estrutura das moléculas em função das suas respectivas atividades apresentaram alta concordância com o perfil eletrônico e estrutural dos compostos, aprovando assim todas as análises realizadas anteriormente.

Estando devidamente indicados os descritores que influenciam na atividade biológica desses compostos, se tornam convenientes posteriores estudos de *docking* e de dinâmica molecular, a fim de que esta contribuição possa, de alguma forma, auxiliar no planejamento de compostos para o tratamento do câncer.

REFERÊNCIAS

- 1 LODISH, H. et al. **Molecular Cell Biology**, 5 ed., New York: W. H. Freeman and Company, 2000.

- 2 KERSSEMAKERS, J. W.; MUNTEANU, E. L.; LAAN, L.; NOETZEL, T. L.; JANSON, M. E.; DOGTEROM, M. Assembly dynamics of microtubules at molecular resolution. **Nature**, v. 442, n. 7103, p. 709-712, 2006.

- 3 INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA (INCA). **Estimativa 2014: Incidência de Câncer no Brasil**. Rio de Janeiro, 2014. 124p.

- 4 LAZO, J. S.; TAKAMURA, K.; VOGT, A.; JUNG, J. K.; RODRIGUEZ, S.; BALACHANDRAN, R.; DAY, B. W.; WIPF, P. Antimitotic actions of a novel analog of the fungal metabolite palmarumycin CP1. **The Journal of Pharmacology and Experimental Therapeutics**, v. 296, n. 2, p. 364-371, 2001.

- 5 JORDAN, A.; HADFIELD, J. A.; LAWRENCE, N. J.; MCGOWN, A. T. Tubulin as a target for anticancer drugs: agents which interact with the mitotic spindle. **Medicinal Research Reviews**, v. 18, n. 4, p. 259-296, 1998.

- 6 FRENSE, D. Taxanes: perspectives for biotechnological production. **Applied Microbiology and Biotechnology**, v. 73, n. 6, p. 1233-1240, 2007.

- 7 WILSON, L.; JORDAN, M. A. Microtubules as a target for anticancer drugs. **Nature**, v. 4, p.252-265, 2004.

- 8 BERGSTRALH, D. T.; TING, J. P. Microtubule stabilizing agents: their molecular signaling consequences and the potential enhancement drug combination. **Cancer Treatment Reviews**, v. 32, n. 3, p. 166-179, 2006.

- 9 GUPTA, S.; BHATTACHARYYA, B. Antimicrotubular drugs binding to vinca domain of tubulin. **Molecular and Cellular Biochemistry**, v. 253, n. 1-2, p. 41-47, 2003.

- 10 GIGANT, B.; WANG, C.; RAVELLI, R. B.; ROUSSI, F.; STEINMETZ, M. O.; CURMI, P. A.; SOBEL, A.; KNOSSOW, M. Structural basis for the regulation of tubulin by vinblastine. **Nature**, v. 435, n. 7041, p. 519-522, 2005.

11 BISWAS, B. B.; SEN, K.; GHOSH CHOUDHURY, G.; BHATTACHARYYA, B. Molecular biology of tubulin: its interaction with drugs and genomic organization. **Journal of Biosciences**, v. 6, n. 4, p. 341-457, 1984.

12 BAI, R.; COVELL, D. G.; PEI, X. F.; EWELL, J. B.; NGUYEN, N. Y.; BROSSI, A.; HAMEL, E. Mapping the binding site of colchicinoids on beta-tubulin. 2-chloroethyl-2-demethylthiocolchicine covalently reacts predominantly with cysteine 239 and secondarily with cysteine 354. **The journal of Biological Chemistry**, v. 275, n. 51, p. 40443-40452, 2000.

13 DUCKI, S.; RENNISON, D.; WOO, M.; KENDALL, A.; CHABERT, J. F. D.; McGOWN, A. T.; LAWRENCE, N. Combretastatin-like chalcones as inhibitors of microtubule polymerization. Part 1: Syntesis and biological evaluation of antivasular activity. **Bioorganic & Medicinal Chemistry**, v. 17, p. 7698-7710, 2009.

14 OLIVEIRA, F. C. S.; BARROS, R .F. M; MOITA NETO, J. M. Plantas medicinais utilizadas em comunidades rurais de Oeiras, semiárido piauiense. **Revista Brasileira de Plantas Mediciniais**, v. 12, p.282-301, 2010.

15 LOPES, L. S.; MARQUES, R. B.; PEREIRA, S. S.; AYRES, M. C.; CHAVES, M. H.; CAVALHEIRO, A. J.; VIEIRA JR, G. M.; ALMEIDA, F. R. C. Antinociceptive effect on mice of the hydroalcoholic fraction and epicatechin obtained from *Combretum leprosum* Mart & Eich. **Brazilian Journal of Medicinal and Biological Research**. Ribeirão Preto, v. 43, p. 1184-1192, 2010.

16 SILVA, J. M. C.; TABARELLI, M.; FONSECA, M. T.; LINS, L. V. **Biodiversidade da Caatinga: áreas prioritárias para a conservação**. Brasília: Ministério do Meio Ambiente, 2003. 382p.

17 PETTIT, G. R.; SINGH, S. B.; HAMEL, E.; LIN, C. M.; ALBERTS, D. S.; KENDALL, D. G. Isolation and structure of the strong cell growth and tubulin inhibitor combretastatin A-4. **Experientia**, v. 45, 1989.

18 PETTIT, G. R.; SINGH, S. B.; BOYD, M. R.; HAMEL, E.; PETTIT, R. K.; SCHMIDT, J. M.; HOGAN, F. Antineoplastic Agents. 291. Isolation and Synthesis of Combretastatins A-4, A-5 and A-6. **Journal of Medicinal Chemistry**, v. 38, p. 1666-1672, 1995.

19 PETTIT, G. R.; LIN, C. M.; SINGH, S. B.; CHU, P. S.; DEMPCY, R. O.; SCHMIDT, J. M.; HAMEL, E. Interactions of Tubulin with Potent Natural and Synthetic Analogs of the Antimitotic Agent Combretastatin: a Structure-Activity Study. **Molecular Pharmacology**, v. 34, p. 200-208, 1988.

20 PETTIT, G. R.; SINGH, S. B.; SCHMIDT, J. M. Isolation, Structure, Synthesis and Antimitotic Properties of Combretastatins B-3 and B-4 from *Combretum Caffrum*. **Journal of Natural Products**, v. 51, p. 517-527, 1988.

21 PETTIT, G. R.; CRAGG, G. M.; HERALD, D. L.; SCHMIDT, J.M.; LOHAVANUAYA, P. Isolation and structure of combretastatin. **Canadian Journal of Chemistry**, v. 60, 1982.

22 PETTIT, G. R.; SINGH, S. B. Isolation, structure, and synthesis of combretastatin A-2, A-3 and B-2. **Canadian Journal of Chemistry**, v. 65, 2390, 1987.

23 NAM, N. H. Combretastatin A-4 Analogues as Antimitotic Antitumor Agents. **Current Medicinal Chemistry**, vol. 10, p. 1697-1722, 2003.

24 DORR, R. T.; DVORAKOVA, K.; SNEAD, K.; ALBERTS, D. S.; SALMON, S. E.; PETTIT, G. R. Antitumor activity of combretastatin-A4 phosphate, a natural product tubulin inhibitor. **Investigational New Drugs**, vol. 14, p. 131-137, 1996.

25 OHSUMI, K.; NAKAGAWA, R.; FUKUDA, Y.; HATANAKA, T.; MORINAGA, Y.; NIHEI, Y.; OHISHI, K.; SUGA, Y.; AKIYAMA, Y.; TSUJI, T. Novel Combretastatin Analogues Effective against Murine Solid Tumors: Design and Structure – Activity Relationships. **Journal of Medicinal Chemistry**, vol. 41, p. 3022-3032, 1998.

26 WANG, L.; WOODS, K. W.; LI, Q.; BARR, K. J.; McCROSKEY, R. W.; HANNICK, S. M.; GHERKE, L.; CREDO, R. B.; HUI, Y. H.; MARSH, K.; WARNER, R.; LEE, J. Y.; MOZNG, N. Z.; FROST, D.; ROSENBERG, S. H.; SHAM, H. L. Potent, Orally Active Heterocycle-Based Combretastatin A-4 Analogues: Synthesis, Structure – Activity Relationship, Pharmacokinetics, and In Vivo Antitumor Activity Evaluation. **Journal of Medicinal Chemistry**, vol. 45, p. 1697-1711, 2002.

27 McGOWN, A. T.; FOX, B. W. Differential cytotoxicity of Combretastatins A1 and A4 in two daunorubicin-resistant P388 cell lines. **Cancer Chemotherapy and Pharmacology**, vol. 26, p.79-81, 1990.

28 RAHMAN, M. A. Chalcone: A Valuable Insight into the Recent Advances and Potential Pharmacological Activities. **Chemical Sciences Journal**, CSJ-29, 2011.

29 SZABO, A.; OSTLUND, N. S.; **Modern Quantum Chemistry: introduction to advanced electronic structure theory**. Dover Publications, 1996.

- 30 LEVINE, I.N **Química Cúantica**. 5 ed. Madrid: Pearson Educación, 2001, 736p.
- 31 HOHENBERG, P.; KOHN, W.; **Physical Review**, v. 136, B864, 1964.
- 32 FITTS, D. D. **Principles of quantum mechanics: as Applied to Chemistry and Chemical Physics**. Cambridge University Press, 2002.
- 33 THOMAS, L. H. ; The calculation of atomic fields. **Proc. Cambridge Phil. Soc.**, v. 23, 542-548, 1927.
- 34 PARR, R. G.; YANG, W. **Density-Functional Theory of atoms and molecules**. New York: Oxford, 1989. 333p.
- 35 BURKE, K. **The ABC of DFT**. 2007. Disponível em: <http://dft.uci.edu/doc/g1.pdf>, acesso em 05 de Out. 2014.
- 36 KOHN, W.; SHAM, L. J.; **Physical Review**, vol. 140, 4A, 1965.
- 37 KOCH, W.; HOLTHAUSEN, M. **A Chemist's Guide to Density Functional Theory**. 2 ed. Weinheim: Wiley, 2001.
- 38 FORESMAN, J. B.; FRISCH, A. **Exploring Chemistry with Electronic Structure Methods**. Pittsburg: Gaussian Inc., 1998.
- 39 KUBINNYI, H. **QSAR: Hansch analysis and related approaches**. New York: VHC, 1993. 240p.
- 40 CRAMER, C. J. **Essentials of Computacional Chemistry**. 2 ed. Chichester: John Wiley & Sons, 2004.
- 41 ABRAHAM, D. J. **Burguer's medicinal chemistry and drug discovery**. 6 ed. Hoboken: John Wiley & Sons, 2003.
- 42 LABUTE. P. A widely aplicable set of descriptors. **Journal of Molecular Graphics and Modelling**. vol. 18, p.464-477, 2000.

43 FERREIRA, M. M. C.; KIRALJ, R. **Métodos Quimiométricos em Relações Quantitativas Estrutura-Atividade (QSAR)**. In: MONTANARI, C. A. Química Medicinal: Métodos e Fundamentos em Planejamento de Fármacos. São Paulo: Edusp, 2011. 732p.

44 PEARSON, R. G. Absolute electronegativity and hardness correlated with molecular orbital theory. **Proceedings of the National Academy of Sciences of the United States of America**, n. 22, v. 83, 1986.

45 REIS, M.; LAMEIRA, J. A theoretical study of phenolic compounds with antioxidant properties. **European Journal of Medicinal Chemistry**, v. 42, 2007.

46 LAMEIRA, J.; ALVES, C.N. A density functional study of flavonoid compounds with anti-HIV activity. **European Journal of Medicinal Chemistry**, v. 41, 2006.

47 OLIVEIRA, D. B.; GAUDIO, A. C. BuildQSAR: A New Computer Program for QSAR Analysis. **Quantitative Structure-Activity Relationships**, v. 19, p. 599-601, 2000.

48 POPPI, R. J.; COSTA FILHO, P. A. Algoritmo genético em química. **Química Nova**, v. 22, n. 3, p. 405-411, 1999.

49 FERREIRA, M. M. C.; MONTANARI, C. A.; GAUDIO, A. C. Seleção de variáveis em QSAR. **Química Nova**, v. 25, n. 3, p. 439-448, 2002.

50 BEEBE, K. R.; PEEL, R. J.; SEASHOLTZ, M.B. **Chemometrics: a practical guide**. New York: Jon Wiley, 1998. 348p.

51 KIRALJ, R.; FERREIRA, M. M. C. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. **Journal of the Brazilian Chemical Society**, v. 20, n. 4, p. 770-787, 2009.

52 GAUDIO, A.C.; ZADONA de, E., Preposição, validação e análise dos modelos que correlacionam estrutura química e atividade biológica. **Química Nova**, n. 5, v. 24, 2001.

53 DOROFKI, M.; ELSHAFIE, A.; JAAFAR, O.; KARIM, O. A.; Comparision of artificial neural network transfer functions abilities to simulate extreme runoff data. **IPCBEE**, v. 33, p. 39-44, 2012.

54 PASTI, R.; DE CASTRO, L. N.; An Immune and a Gradient-Based Method to Train Multi-Layer Perceptron Neural Networks. In: **International Joint Conference on Neural Networks**. Vancouver. Proceedings do World Congress on Computational Intelligence. p. 4077-4084, 2006.

55 SEGATTO, E. C.; COURY, D. V.; Redes neurais aplicadas a relés diferenciais para transformadores de potência. **Controle & Automação**, v. 19, n. 1, 2008.

56 CASTRO, L. N.; VON ZUBEN, F. J.; Optimized training techniques for feedforward neural networks. **Technical Report**, DCA-RT 03/98. Campinas, 1998.

57 GAUSSIAN. FRISCH, M. J., TRUCKS, G. W., SCHLEGEL, H. B., SCUSERIA, G. E., ROBB, M. A., CHEESEMAN, J. R., SCALMANI, G., BARONE, V., MENNUCCI, B., PETERSSON, G. A., NAKATSUJI, H., CARICATO, M., LI, X. HRATCHIAN, H.P., IZMAYLOV, A. F., BLOINO, J., ZHENG, G., SONNENBERG, J. L.; HADA, M., EHARA, M., TOYOTA, K., FUKUDA, R., HASEGAWA, J., ISHIDA, M., NAKAJIMA, T., HONDA, Y., KITAO, O., NAKAI, H., VREVEN, T., MONTGOMERY, JR., J. A., PERALTA, J. E., OGLIARO, F., BEARPARK, M., HEYD, J. J., BROTHERS, E., KUDIN, K. N., STAROVEROV, V. N., KOBAYASHI, R., NORMAND, J., RAGHAVACHARI, K., RENDELL, A., BURANT, J. C., IYENGAR, S.S., TOMASI, J., COSSI, M., REGA, N., MILLAM, N. J., KLENE, M.; KNOX, J. E., CROSS, J. B., BAKKEN, V., ADAMO, C., JARAMILLO, J., GOMPERTS, R., STRATMANN, R. E., YAZYEV, O., AUSTIN, A. J., CAMMI, R., POMELLI, C., OCHTERSKI, J. W., MARTIN, R. L., MOROKUMA, K., ZAKRZEWSKI, V. G., VOTH, G. A., SALVADOR, P., DANNENBERG, J. J., DAPPRICH, S., DANIELS, A.D., FARKAS, O., FORESMAN, J. B., ORTIZ, J. V., CIOŚŁOWSKI, J., FOX, D. J. Gaussian, Inc., Wallingford CT, 2009.

58 LEE, C.; YANG, W.; PARR, R. G.; **Physical Review B**, n. 37, v. 785, 1988.

59 POOPLE, J. A.; SEEGER, R.; BINKLEY, J. S.; KRISHNAN, R. Self-Consistent Molecular Orbital Methods. XX. A Basis Set for Correlated Wavefunctions. **Journal of Chemical Physics**, v. 72, 650, 1980.

60 TODESCHINI, R.; CONSONNI, V.; PAVAM, M. **Dragon 2.1**. Milano, 2002.

61 OLIVEIRA, D. B.; GAUDIO, A. C. BuildQSAR : A new computer program for QSAR analysis. **Quantitative Structure-Activity Relationships**, v. 19, n. 6, p. 599–601, 2001.

62 INFOMETRIX INC. **Pirouette 3.11**. Woodinville, 2002.

63 MARTINS, J. P. A.; FERREIRA, M. M. C. QSAR modeling: um novo pacote computacional open source para gerar e validar modelos QSAR. **Química Nova**, v. 36, n. 4, p. 554–560, 2013.

64 MATHWORKS. **Matlab 7.12**. 2011.

65 HEMMER, M. C.; STEINHAUER, V.; GASTEIGER, J.; Deriving the 3D structure of organic molecules from their infrared spectra. **Journal of Vibrational Spectroscopy**, v. 19, p. 151-164, 1999.

66 ABREU, R. M. V.; FERREIRA, I. C. F. R.; QUEIROZ, M. J. R. P. QSAR model for predicting radical scavenging activity of di(hetero)arylamines derivatives of benzo[b]thiophenes. **European Journal of Medicinal Chemistry**, v. 44, p. 1952-1958, 2009.

67 RANDIC, M. Molecular shape profiles. **J. Chem. Inf. Comput. Sci.**, v. 35, p. 373-382, 1995.

68 RANDIC, M.; BASAK, S. C. Optimal molecular descriptors based on weighted path numbers. **J. Chem. Inf. Comput. Sci.**, v. 39, p. 261-266, 1999.

69 CONSONNI, V.; TODESCHINI, R.; PAVAN, M.; Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. **J. Chem. Inf. Comput. Sci.**, v. 42, p. 682-692, 2002.

70 CONSONNI, V.; TODESCHINI, R.; PAVAN, M.; Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. **J. Chem. Inf. Comput. Sci.**, v. 42, p. 693-705, 2002.

71 SKOROBOGATOV, V. A.; DOBRYNIN, A. A. Metric analysis of graphs. **Match**, n. 23, p. 105-151, 1988.

72 HILL, N. E. The dielectric relaxation of polar molecules in solution. **Nature**, v. 171, n. 4358, p. 836-837, 1953.