

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

DISSERTAÇÃO DE MESTRADO

Estudos SAR e QSAR-2D de derivados de N-benzoil-2-hidroxibenzamidas ativos contra *Plasmodium falciparum*

Verlucia Amanda Machado de Freitas

João Pessoa – PB – Brasil
Dezembro / 2017

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

DISSERTAÇÃO DE MESTRADO

Estudos SAR e QSAR-2D de derivados de N-benzoil-2-hidroxibenzamidas ativos contra *Plasmodium falciparum*

Verlucia Amanda Machado de Freitas

Dissertação submetida ao programa de pós-graduação em química, da Universidade Federal da Paraíba, como requisito parcial à obtenção do título de Mestre em Química, área de concentração “Físico-Química”.

Orientador (a): Prof.^a. Dr.^a Karen Cacilda Weber

Bolsista (CAPES)

João Pessoa – PB – Brasil
Dezembro / 2017

Catálogo na publicação
Seção de Catalogação e Classificação

F866e Freitas, Verlucia Amanda Machado de..
Estudos SAR e QSAR-2D de derivados de N-benzoil-2-
hidroxibenzamidas ativos contra *Plasmodium falciparum* /
Verlucia Amanda Machado de Freitas. - João Pessoa, 2017.
84 f. : il.

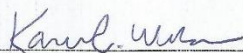
Orientadora: Karen Cacilda Weber.
Dissertação (Mestrado) - UFPB/CCEN

1. Química. 2. Quimiometria. 3. Estudos QSAR-2D.
4. *Plasmodium falciparum*. I. Título.

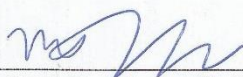
UFPB/BC

Estudos SAR e QSAR-2D de derivados de N-benzoil-2-hidroxibenzamidas ativos contra plasmodium falciparum.

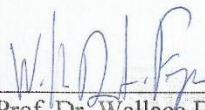
Dissertação de Mestrado apresentada pela aluna Verlúcia Amanda Machado de Freitas e aprovada pela banca examinadora em 14 de dezembro de 2017.



Profa. Dra. Karen Cacilda Weber
Orientadora/Presidente



Prof. Dr. Marcus Tullius Scotti
Examinador



Prof. Dr. Wallace Duarte Fragoso
Examinador

Levantarei os meus olhos para os montes, de onde vem o meu socorro.

O meu socorro vem do Senhor que fez o céu e a terra.

Não deixará vacilar o teu pé; aquele que te guarda não tosquenejará.

Eis que não tosquenejará nem dormirá o guarda de Israel.

O Senhor é quem te guarda; o Senhor é a tua sombra à tua direita.

O sol não te molestará de dia nem a lua de noite.

O Senhor te guardará de todo o mal; guardará a tua alma.

O Senhor guardará a tua entrada e a tua saída, desde agora e para sempre

Salmos 121: 1-8.

AGRADECIMENTOS

A Deus, que com sua infinita misericórdia tem me sustentado todos os dias.

Aos meus pais, José Silva de Freitas e Vera Lúcia M. de Freitas, que sonharam juntamente comigo para a conquista deste título.

Ao meu marido Thiago Borges, que me apoiou e sempre esteve ao meu lado.

Ao meu filho, um príncipe, que mesmo na fragilidade de ser apenas uma criança, entendeu todos os momentos de ausência da mamãe.

Às amigas Sayonara Araújo, Yanne Vasquez, Thayana Maria e Vanúbia Pontes, parceiras fundamentais neste processo.

Ao LQQC, ao LMMRQ e ao colega Railton Barbosa, que contribuiu pacientemente para meu aprendizado, buscando sempre ajudar.

Ao LAQA e aos colegas que pude fazer nas aulas de quimiometria.

Ao Programa de Pós-Graduação em Química e aos professores Otávio Santana, Mário Ugolino, Elizete Ventura e Wallace Duarte pela formação oferecida em suas aulas.

À prof^a. Dr^a. Karen Cacilda Weber, que me orientou e forneceu todo seu conhecimento, fundamental para construção deste trabalho, sendo paciente e compreensiva, sou verdadeiramente grata.

À CAPES, pelo financiamento da bolsa de pesquisa.

SUMÁRIO

1 INTRODUÇÃO	16
2 OBJETIVOS.....	18
2.1 Objetivo Geral.....	18
2.2 Objetivos Específicos.....	18
3 FUNDAMENTAÇÃO TEÓRICA	19
3.1 Planejamento Racional de Fármacos.....	19
3.2 Relação Quantitativa Estrutura-Atividade (QSAR).....	19
3.2.1 Conjunto de dados.....	22
3.2.2 Descritores Moleculares.....	23
3.2.3 Seleção de Variáveis.....	25
3.2.3.1 Peso de Fisher.....	25
3.2.3.2 Ordered Predictor Selection (OPS).....	26
3.2.3.3 Genetic Algorithm (GA).....	28
3.3 Quimiometria aplicada aos Estudos QSAR.....	29
3.3.1 Análise Hierárquica de Agrupamentos (HCA).....	29
3.3.2 Análise de Componentes Principais (PCA).....	31
3.3.3 <i>K-ésimo</i> Vizinho mais Próximo (KNN).....	32
3.3.4 Modelagem Independente por Analogia de Classes (SIMCA).....	33
3.3.5 Análise Discriminante pelos Mínimos Quadrados Parciais (PLS-DA).....	34
3.3.6 Técnicas de Calibração Multivariada.....	35
3.4 Validação dos modelos QSAR.....	37
4 METODOLOGIA.....	40
4.1 Otimização das geometrias e cálculo dos descritores.....	44
4.2 Análises Quimiométricas.....	45
5 RESULTADOS E DISCUSSÃO.....	50
5.1 Análise exploratória por HCA.....	50
5.2 Estudos da Relação entre Estrutura e Atividade (SAR).....	52
5.2.1 Seleção de variáveis por Peso de Fisher.....	53
5.2.2 HCA.....	54
5.2.3 PCA.....	55
5.2.4 Resultados do KNN.....	57
5.2.5 Resultados do SIMCA.....	58
5.2.6 Resultados do PLS-DA.....	62
5.3 Relação Quantitativa entre Estrutura e Atividade Biológica (QSAR).....	65
5.3.1 Resultados do modelo OPS-PLS para o conjunto com 31 amostras.....	69
5.3.1.1 Detecção de Outliers.....	69

5.3.1.2 Seleção de variáveis e construção do modelo PLS.....	70
5.3.1.3 Testes de leave-N-out, y-randomization e validação externa.....	72
5.3.1.4 Implicações para o planejamento de novas moléculas.....	75
6 CONCLUSÕES.....	78
7 PERSPECTIVAS.....	80
REFERÊNCIAS.....	81

LISTA DE FIGURAS

Figura 1: Estrutura do composto N-(4-etilbenzoil)-2-hidroxibenzamida (1a).....	17
Figura 2: Estrutura do composto 1r	17
Figura 3: Esquema ilustrativo dos passos na elaboração de um estudo QSAR e construção do modelo preditivo.....	22
Figura 4: Resumo da técnica OPS.....	28
Figura 5: Primeira e segunda PC em um espaço bidimensional.....	32
Figura 6: Exemplo do modelo SIMCA.....	34
Figura 7: Fluxograma da pesquisa.....	49
Figura 8: HCA com todos os descritores e o método de conexão completo.....	50
Figura 9: HCA com todos os descritores e o método de conexão simples.....	50
Figura 10: HCA com todos os descritores e o método de conexão da média.....	51
Figura 11: HCA com todos os descritores e o método de conexão centróide.....	51
Figura 12: Estrutura principal, estrutura 30 e estrutura 31.....	52
Figura 13: Dendrograma das 31 amostras com seleção de variáveis pelo Peso de Fisher.....	55
Figura 14: Representação gráfica dos escores PC1 <i>versus</i> PC2.....	56
Figura 15: Distâncias dos compostos para às classes obtidas para o conjunto de treinamento.....	60
Figura 16: Moléculas 1a, 12 e 16. O átomo de nitrogênio é representado na cor azul.....	61
Figura 17: Moléculas 1b (ativa) e 1i (inativa).....	62
Figura 18: Número de variáveis latentes por erro de classificação de CV.....	62
Figura 19: Importância das variáveis na projeção dos escores.....	63
Figura 20: Molécula 1h (inativa), 1m e 1n (ativas). Átomos de oxigênio representados em vermelho.....	64
Figura 21: Moléculas 1h (inativa), 1f e 1g (ativas). Átomos de flúor representados em azul ciano.....	64

Figura 22: Representação gráfica da distribuição dos valores de pIC_{50} nos conjuntos teste, treinamento e conjunto total para o conjunto de 37 e 31 amostras, respectivamente.....	66
Figura 23: Detecção de Outliers.....	70
Figura 24: Parâmetros utilizados para o OPS.....	70
Figura 25: Resultados da validação cruzada “leave-N-out”	73
Figura 26: Resultados do teste de Y-randomization. Os pontos azuis representam os modelos gerados com os valores de y aleatorizados e o ponto vermelho representa o modelo original.....	75
Figura 27: Valores pIC_{50} Preditos <i>versus</i> pIC_{50} Experimentais de pIC_{50} (triângulos vermelhos correspondem as amostras da validação externa).....	75
Figura 28: Contribuição de cada descritor para o vetor de regressão.....	76

LISTA DE TABELAS

Tabela 1: Estruturas químicas e respectivos valores de pIC ₅₀ e IC ₅₀	40
Tabela 2: Amostras ativas e inativas com valores de índice de seletividade (Stec <i>et al.</i> , 2012).....	44
Tabela 3: Resultados da classificação com as diferentes técnicas de seleção de variáveis e de classificação para conjunto treinamento e teste. Os valores em % referem-se ao cálculo da Taxa de Classificação Correta (equação 17).....	53
Tabela 4: Média, Variância e Peso de Fisher para as propriedades responsáveis pela separação das moléculas inativas e ativas.....	54
Tabela 5: <i>Loadings</i> das variáveis em cada PC's.....	57
Tabela 6: Sumário da classificação obtida com a técnica KNN para o conjunto de treinamento com as variáveis selecionadas pelo Peso de Fisher.....	58
Tabela 7: Número de componentes principais usados nos dois modelos SIMCA.....	58
Tabela 8: Resíduo entre as classes para os dois modelos SIMCA.....	59
Tabela 9: Distância entre as classes para os dois modelos SIMCA.....	59
Tabela 10: Matriz de confusão da classificação SIMCA.....	60
Tabela 11: Matriz de confusão da predição SIMCA.....	61
Tabela 12: Performance do PLS-DA calculada através do teorema de Bayes...65	
Tabela 13: Valores de coeficiente de determinação (r^2) para calibração e predição dos diferentes métodos utilizados.....	68
Tabela 14: Descritores selecionados pelo OPS para o melhor modelo.....	71
Tabela 15: Parâmetros estatísticos obtidos com a regressão PLS sobre os 12 descritores selecionados.....	72
Tabela 16: Valores experimentais e preditos de pIC ₅₀ para o conjunto de teste.....	74

LISTA DE SIGLAS E ABREVIATURAS

2D - Bidimensional

3D - Tridimensional

ADMET - Absorção, Distribuição, Metabolismo, Excreção e Toxicidade

AM1 – *Austin Model One*

APS - Algoritmo das Projeções Sucessivas.

DFT – *Density Functional Theory*

CV – Validação Cruzada

GA – *Genetic Algorithm*

HCA – *Hierarchical Cluster Analysis*

IC₅₀ - Concentração Inibitória de 50%

KNN – *K Nearest Neighbor*

KS – *Kennard-Stone*

LDA - *Linear Discriminant Analysis*

LOO - Deixe um de fora (do inglês *Leave-one-out*)

MLR – *Multiple Linear Regression*

NBO – *Natural Bond Orbital*

PC – *Principal Component*

PCA – *Principal Component Analysis*

PLS – *Partial Least Squares*

PLS-DA – *Partial Least Squares Discriminant Analysis*

PRESS – *Predictor Error Sum of Squares for Cross Validation*

q² – Coeficiente de determinação da validação cruzada

QSAR – *Quantitative Structure Activity Relationship*

r² - Coeficiente de determinação múltipla

r^2_{pred} - Coeficiente de determinação preditivo

SAR – *Structure-Activity Relationship*

SEV - *standard error of cross-validation* - erro padrão da validação cruzada leave-one-out

SIMCA – *Soft Independent Modeling of Class analysis*

VLs – Variáveis Latentes

RESUMO

Título: Estudos SAR e QSAR-2D de derivados de N-benzoil-2-hidroxibenzamidas ativos contra *Plasmodium falciparum*

Dentre as doenças parasitárias tropicais, a malária é uma das mais devastadoras. Um dos agentes causadores é o *Plasmodium falciparum*, que é responsável pelas mais altas taxas de mortalidade. Embora existam alguns medicamentos empregados no tratamento da malária, questões como a resistência dos parasitas e a baixa eficácia dos tratamentos torna o desenvolvimento de novos fármacos extremamente necessário. Existe um grande interesse comercial para prever a atividade biológica de novas moléculas e uma metodologia muito utilizada são os estudos da relação estrutura-atividade. Neste trabalho, utilizou-se um conjunto de 39 derivados de N-benzoil-2-hidroxibenzamidas e foram calculadas algumas propriedades eletrônicas utilizando o método DFT com o funcional M06-2X e o conjunto de bases 6-311+g(d,p), além de descritores físico-químicos e estruturais, com a utilização da plataforma on-line E-dragon 1.0. Análises de relação entre estrutura e atividade (SAR) e de relações quantitativas entre estrutura e atividade bidimensionais (QSAR-2D) foram realizadas com diferentes técnicas quimiométricas classificadas em três categorias: *reconhecimento de padrões não-supervisionado* (HCA e PCA), *reconhecimento de padrões supervisionado* (KNN, SIMCA e PLS-DA) e *calibração multivariada* (PLS e MLR) com seleção de variáveis para classificação pelos métodos de Peso de Fisher e GA (Algoritmo Genético), e seleção de variáveis para calibração usando OPS (Seleção de Preditores Ordenados) e GA. Nas análises HCA e PCA verificou-se agrupamentos característicos e a separação das amostras ativas e inativas. Nas técnicas de classificação, os modelos SIMCA e PLS-DA demonstraram confiabilidade e boa consistência interna com Taxa de Classificação Correta acima de 90% para o conjunto treinamento e teste. Os descritores selecionados sugerem que existem características estruturais que permitem separar os compostos ativos dos inativos no espaço químico definido. Para a calibração multivariada, o método de seleção de variáveis OPS em conjunto com o PLS conduziu a uma melhor proposta de modelo QSAR, que se mostrou estável, robusto e preditivo para a atividade antimalárica da classe de compostos estudados, com coeficientes de correlação $q^2 = 0,75$, $r^2 = 0,81$ e $r^2_{\text{pred}} = 0,89$. Para o melhor modelo, a atividade antimalárica está associada aos descritores do tipo 2D autocorrelations, Edge adjacency índices, Information Indices, descritores WHIM, RDF e 3D-Morse, e o descritor químico-quântico polarizabilidade. Os resultados apresentados demonstraram que o procedimento desenvolvido para o conjunto de N-benzoil-2-hidroxibenzamidas permitiu a obtenção de modelos confiáveis e preditivos, fornecendo subsídios para a síntese e avaliação biológica de novos compostos com características estruturais semelhantes às estudadas e potencialmente mais ativos contra o *P. falciparum*.

Palavras-Chave: SAR, QSAR-2D, *P. falciparum*, benzamidas.

ABSTRACT

Title: SAR and 2D-QSAR studies of N-benzoyl-2-hydroxybenzamide derivatives active against *Plasmodium falciparum*

Among parasitic diseases, malaria is one of the most devastating. One of the causative agents is *Plasmodium falciparum*, which is responsible for the highest mortality rates. Although there are some drugs employed in the treatment of malaria, issues such as parasitic resistance and low treatment efficacy make extremely necessary the design of new drugs. There is a huge commercial interest in biological activity prediction of new molecules and a largely utilized methodology are the structure-activity relationship studies. In this work, a set of 39 N-benzoyl-2-hydroxybenzamide derivatives were employed and some electronic properties were calculated using DFT method with the M06-2X functional and the 6-311+g(d,p) basis set, along with physical-chemical and structural descriptors, using the on-line platform E-dragon 1.0. Analyses of structure-activity relationships (SAR) and bi-dimensional quantitative structure-activity relationships (2D-QSAR) were performed with different chemometric techniques classified in three categories: unsupervised pattern recognition (HCA and PCA), supervised pattern recognition (KNN, SIMCA and PLS-DA) and multivariate calibration (PLS and MLR) with variable selection for classification using methods of Fisher weights and GA (Genetic Algorithm), and variable selection for calibration using OPS (*Ordered Predictor Selection*) and GA. In HCA and PCA, characteristic clusters and the separation of active and inactive samples were verified. In classification techniques, SIMCA and PLS-DA models have demonstrated reliability and good internal consistency with Correct Classification Rate above 90% for training and test set. The selected descriptors suggest that there are structural features which allow the separation of active and inactive compounds in the chemical space defined. For multivariate calibration, variable selection with OPS along with PLS has led to a better QSAR model proposal, which has demonstrated to be stable, robust and predictive for antimalarial activity of the class of compounds under study, with correlation coefficients $q^2 = 0,75$, $r^2 = 0,81$ and $r^2_{pred} = 0,89$. For the best model, antimalarial activity is associated to descriptors of 2D autocorrelation descriptors, Edge adjacency indices, Information Indices, WHIM, RDF and 3D-Morse descriptors, and the quantum chemical polarizability descriptor. The results indicate that the developed procedure for the set of N-benzoyl-2-hydroxybenzamides allow the achievement of reliable and predictive models, providing subsidies for synthesis and biological evaluation of new compounds with structural features similar to the ones studied here and potentially more active against *P. falciparum*.

Keywords: SAR, 2D-QSAR, *P. falciparum*, benzamides.

1 INTRODUÇÃO

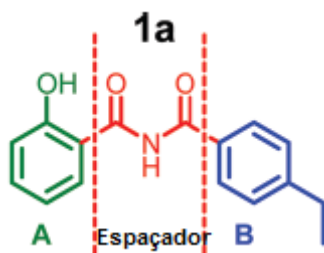
A malária, uma doença parasitária que tem como agentes etiológicos quatro espécies de protozoários do gênero *Plasmodium* (*P. vivax*, *P. ovale*, *P. malariae* e *P. falciparum*), é transmitida ao homem através da picada de fêmeas do inseto do gênero *Anopheles*. O *Plasmodium falciparum* é o grande responsável pela malária grave, caracterizada por distúrbios em diferentes órgãos e sistemas, com possibilidade de evolução ao óbito ^[1].

Segundo a Coordenação de Informações Estratégicas em Vigilância em Saúde – CIEVS, nos anos de 2015 e 2016 verificou-se um aumento no número de casos e aumento na proporção de casos infectados pelo *P. falciparum* em comparação aos infectados pelo *P. vivax* na Região das Américas, principalmente em países como Colômbia, Equador e Venezuela ^[1].

Durante muitos anos a cloroquina (4-aminoquinolina) foi utilizada como o principal fármaco no tratamento da malária, isso graças à sua eficácia, baixa toxicidade e o baixo custo. Contudo, a aparição da resistência à cloroquina levou ao desenvolvimento de outras drogas ^[2]. Atualmente, o tratamento de escolha para a malária são as chamadas terapias combinadas baseadas em artemisinina (do inglês, *artemisinin-based combination therapies*), que aliam este composto com um segundo fármaco antimalárico. No entanto, a resistência à artemisinina foi reportada em 2014 e já está disseminada no sudeste da Ásia, o que constitui um novo desafio e reforça a necessidade da busca por novas moléculas capazes de atuar como agentes antimaláricos ^[3].

Na litetaruta^[4] é descrita a descoberta de N-benzoil-2-hidroxibenzamidas como uma classe de compostos eficazes contra *T. gondii* *in vitro* e *in vivo*, e que a molécula N-(4-etilbenzoil)-2-hidroxibenzamida (**1a**) foi a mais eficaz (**Figura 1**). Stec e colaboradores^[5], reutilizaram as informações deste estudo e detalharam várias dessas sínteses, identificando agentes potentes e, em alguns casos, seletivos contra quatro parasitas protozoários: *Plasmodium falciparum*, *Leishmania donovani*, *Trypanosoma brucei* e *Trypanosoma cruzi*.

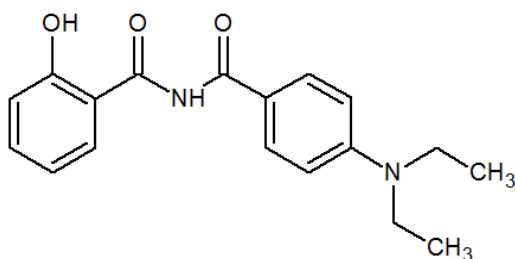
Figura 1: Estrutura do composto N-(4-etilbenzoil)-2-hidroxibenzamida (**1a**)



Fonte: Stec *et al.*, 2012 [5].

Em busca de compostos com atividade e propriedades ADMET (Absorção, Distribuição, Metabolismo, Excreção e Toxicidade) melhorada, os autores modificaram a estrutura **1a** em três locais: o anel fenil (A), anel de 4-etilfenil (B) e o espaçador de imida (**Figura 1**). Análises da relação estrutura-atividade levaram à identificação de compostos neste conjunto com excelente atividade antimalárica, a exemplo do composto **1r** (**Figura 2**), que apresentou a melhor atividade antimalárica, 21 vezes superior à do medicamento antimalárico padrão cloroquina contra o *P. falciparum* [5].

Figura 2: Estrutura do composto **1r**.



Fonte: Stec *et al.*, 2012 [5].

O presente trabalho surgiu como uma tentativa de obter modelos da relação quantitativa entre estrutura e atividade para a série congênere sintetizada por Stec e colaboradores (2012) [5]. Para este fim, foram empregadas diversas técnicas de reconhecimento de padrões e calibração multivariada em combinação com diferentes técnicas de seleção de variáveis. Os resultados obtidos possibilitaram a construção de um modelo robusto e preditivo, com poder de predição superior a 85%.

2 OBJETIVOS

2.1 Objetivo Geral

Construir modelos de relação entre estrutura e atividade (SAR) e modelos robustos e preditivos da relação quantitativa entre estrutura e atividade biológica (QSAR) de derivados de N-benzoil-2-hidroxibenzamidas, buscando evidenciar as propriedades relevantes à atividade biológica desta classe de compostos, com vistas ao desenvolvimento de novos agentes antimaláricos.

2.2 Objetivos Específicos

- Otimizar a geometria das moléculas que compõem o conjunto de dados selecionados para o estudo em nível DFT;
- Calcular propriedades eletrônicas, estruturais e topológicas para determinar quais as características que discriminam os compostos com maiores e menores atividades antimaláricas;
- Selecionar os descritores com maior poder de correlação com a atividade biológica;
- Obter modelos da relação entre estrutura e atividade (SAR) e da relação quantitativa entre estrutura e atividade (QSAR) a partir das estruturas modeladas;
- Validar os modelos obtidos utilizando as técnicas estatísticas apropriadas.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Planejamento Racional de Fármacos

As ferramentas de modelagem molecular desempenham um papel fundamental na busca por novas moléculas, sendo esta uma área de pesquisa que vem se desenvolvendo fortemente, o que pode ser observado pelo constante interesse da indústria farmacêutica em métodos da química medicinal para a otimização de compostos protótipos e na determinação estrutural de complexos ligante-receptor [6, 7, 8].

O processo de planejamento de compostos com atividade antimalárica baseia-se, principalmente, na investigação e comparação dos mecanismos bioquímicos associados à sobrevivência do parasito. Entender as características no sequenciamento do genoma de espécies *Plasmodium* e investigar as proteínas existentes pode proporcionar a descoberta de novos compostos [9].

A descoberta de um fármaco com estrutura completamente nova demanda de 10 a 15 anos de esforço de equipes multidisciplinares, envolvendo investimentos em torno de 500 milhões de dólares. Neste sentido e apesar da grande dificuldade de identificação e introdução de um fármaco novo no mercado, os métodos de busca de novos fármacos contam com forte contribuição dos estudos de QSAR [7, 10]. Uma das propostas mais interessantes desse estudo é a possibilidade de fornecer informações sobre a tendência de bioatividade para uma molécula ainda não sintetizada e/ou testada. Isto representa um importante passo durante o planejamento racional de novos fármacos por direcionar a pesquisa de uma série congênere, poupando custos e tempo neste processo [11].

3.2 Relação Quantitativa Estrutura-Atividade (QSAR)

A modelagem da relação quantitativa entre estrutura e atividade é uma das principais ferramentas computacionais empregadas em Química Medicinal [8, 11].

O primeiro passo na criação de um modelo QSAR é gerar um conjunto de compostos de treinamento com suas atividades experimentais. Idealmente, cada uma dessas atividades deve abranger a gama de possíveis valores para essa atividade. Se um número suficiente de atividades experimentais forem conhecidas, uma pequena porcentagem deles pode ser selecionada aleatoriamente para ser excluída do conjunto e ser usado como um conjunto de validação ^[12].

As estruturas químicas não contêm a informação relacionada à atividade biológica de maneira explícita. Esta informação deve ser extraída das estruturas na forma de descritores moleculares que acentuem diferentes propriedades químicas implícitas na estrutura da molécula. Tais propriedades, que vão desde parâmetros físico-químicos, eletrônicos ou mecânico-quânticos a características geométricas e topológicas das moléculas, podem ser então correlacionadas diretamente com a atividade ^[13]. Dentre os programas desenvolvidos para o cálculo de descritores moleculares diversos, um dos mais citados na literatura é o programa DRAGON ^[14].

Uma vez que os descritores foram calculados, é necessário selecionar quais devem ser incluídos no modelo QSAR e um coeficiente de correlação fornece uma medida quantitativa de quão bem cada descritor descreve a atividade. Assim, o descritor com maior coeficiente de correlação pode ser selecionado. O próximo descritor a ser selecionado deve ser um que se correlaciona bem com a atividade, mas não está fortemente correlacionado com o primeiro descritor selecionado (conforme descrito por um coeficiente de correlação cruzada). Este torna possível compensar deficiências no primeiro descritor sem redundância. Essas métricas fornecem uma base para selecionar descritores e, muitas vezes, sugerir várias opções alternativas ^[12].

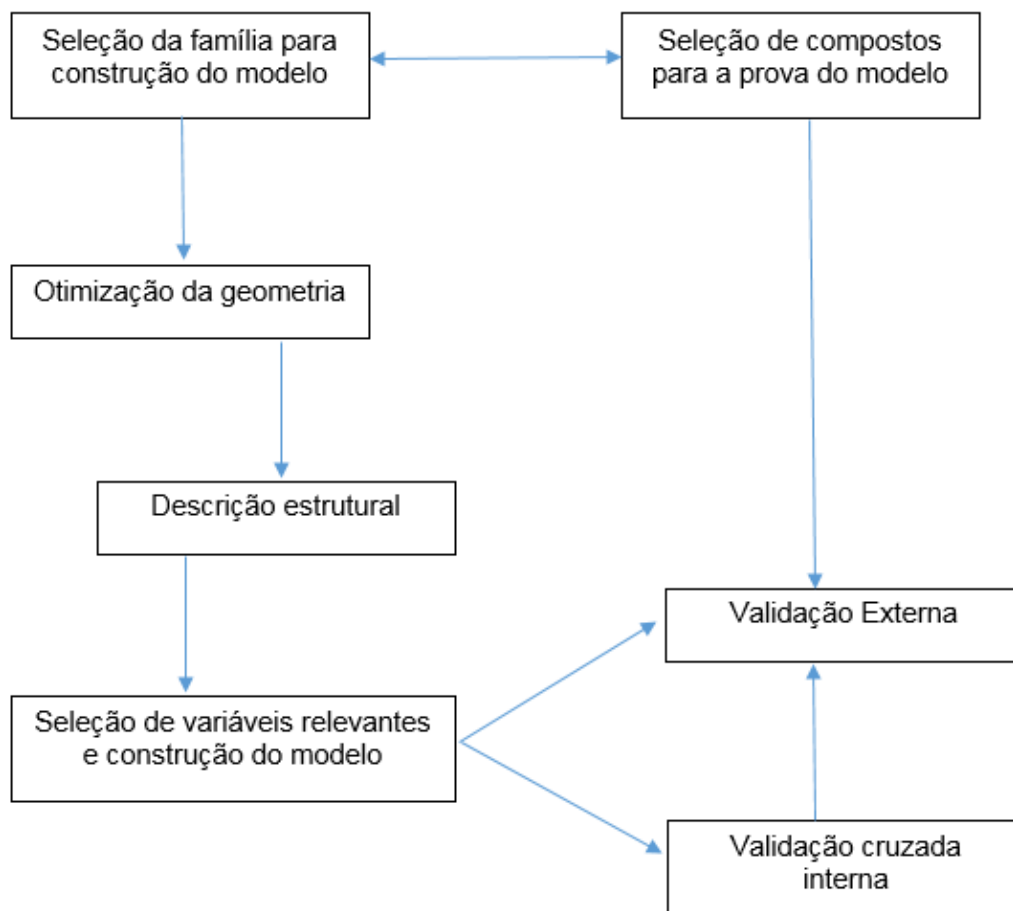
Após a obtenção de um determinado número de parâmetros moleculares, torna-se necessária a utilização de métodos que permitam a análise simultânea de todos os parâmetros obtidos, os métodos multivariados de análise são ferramentas de muita utilidade em estudos desse tipo ^[15].

Vários métodos quimiométricos ampliaram o arsenal de ferramentas que podem ser aplicadas a estudos QSAR. Entre eles, os mais amplamente utilizados são: i) os métodos de análise exploratória de dados como Análise de

Componentes Principais (PCA) e Análise Hierárquica de Agrupamentos (HCA); ii) métodos supervisionados de reconhecimento de padrões, como K-ésimo Vizinho mais Próximo (KNN) e Modelagem Independente por Analogia de Classes (SIMCA) e iii) métodos de Regressão Linear Múltipla (MLR), Regressão pelo Método das Componentes Principais (PCR) e Mínimos Quadrados Parciais (PLS) ^[16].

Para que uma equação de regressão seja promovida a modelo matemático é preciso muito mais do que simplesmente possuir elevado coeficiente de correlação. Para validar-se estatisticamente uma equação de regressão, é preciso executar diversos testes de validação, a análise dos resíduos e a validação cruzada. Bons resultados que eventualmente venham a ser obtidos em todos os testes de forma alguma garantem que a equação venha a ser útil para descrever a atividade biológica de um grupo de compostos. É preciso que a equação seja consistente com algum mecanismo de ação, em nível molecular, proposto para os compostos e que também sirva para fazer previsões sobre a atividade de compostos que não tenham sido incluídos no modelo ^[17]. A **Figura 3** apresenta o esquema dos passos essenciais para a elaboração do estudo QSAR ^[6,18].

Figura 3: Esquema ilustrativo dos passos na elaboração de um estudo QSAR e construção do modelo preditivo.



Fonte: Adaptado de Bajorath, 2004; Machado, 2008; Duarte, 2012 [6,18].

Ao longo dos anos, a execução dos estudos QSAR tem incorporado novas características e metodologias para modelar a relação entre a atividade biológica e a estrutura química de um composto. Os principais passos na elaboração de um estudo QSAR e construção do modelo preditivo serão comentados a seguir.

3.2.1 Conjunto de dados

A condição mais importante em um estudo QSAR é a disponibilidade de uma série congênere, incluindo apenas compostos com o mesmo mecanismo de ação. Todos os compostos de um conjunto devem ter o mesmo esqueleto principal, isto é, um composto original idêntico, com variação estrutural em apenas um ou várias posições [19].

Quanto aos dados biológicos, alguns tipos de dados de atividade biológica podem ser usados em QSAR, a exemplo de K_i (constante de inibição) e IC_{50} (concentração inibitória), desde que estejam na escala correta ^[19].

Uma vez que as constantes de equilíbrio ou de velocidade estão relacionadas aos valores de energia livre ΔG por relações como a da **equação 1**:

$$\Delta G = -2,303 RT \log K, \quad (1)$$

apenas constantes de equilíbrio (por exemplo, valores de K_i ou IC_{50} e não % de inibição a determinada concentração) e constantes de velocidade (como valores de $\log K$ e não % de absorção ou % de concentração) são apropriadas para estudos de QSAR, o que significa que todos os dados biológicos devem ser transformados de uma maneira apropriada antes de serem usados em análises quantitativas.

Como convenção, os negativos dos logaritmos, i.e., logaritmos dos recíprocos das concentrações molares (por exemplo, $\log 1/C$ ou pC) são usados para que se obtenham valores maiores para compostos mais ativos. Há ainda outra razão para o uso da escala logarítmica. Uma condição para aplicação das análises de regressão é uma distribuição normal do erro experimental na variável dependente. Para dados biológicos, isso é verdadeiro para escala logarítmica e não para a linear. Além disso, os dados de atividade biológica devem ser razoavelmente distribuídos sobre toda a faixa de valores, sem agrupamentos de dados ^[12-19].

3.2.2 Descritores Moleculares

Está bem fundamentado que muitas propriedades físico-químicas ou funções termodinâmicas de um composto dependem da sua estrutura química. O termo *estrutura* inclui os aspectos topológicos, eletrônicos e geométricos. Uma vez que muitas propriedades moleculares são medidas e os dados são registrados como valores numéricos, para a construção de modelos de QSPR/QSAR, a estrutura molecular também deve ser apresentada da mesma maneira. Portanto, são necessários dois conjuntos numéricos: um que

representa a propriedade/atividade e o outro que representa a estrutura molecular. Desta forma, pode-se estabelecer uma relação quantitativa entre os dois conjuntos de números através de uma equação algébrica [20].

Os descritores químicos estão no núcleo da modelagem QSAR e tantos tipos diferentes de descritores químicos são produtos que refletem vários níveis de estrutura química. Estes níveis variam de fórmula molecular (o chamado 1D), para o mais popular entre os químicos, que é a fórmula estrutural bidimensional (2D), para o tridimensional, dependente da conformação (3D), e até mesmo níveis mais elevados, levando em consideração a orientação mútua e temporal da dinâmica das moléculas (4D) [8].

Os vários descritores em uso podem ser classificados como sendo constitucionais, topológicos, geométricos, eletrostáticos ou químico-quânticos. Descritores constitucionais, como por exemplo, o peso molecular o número de heteroátomos, o número de anéis, o número de ligações duplas, etc., dão uma simples descrição do que está na molécula, e muitas vezes aparecem em uma equação QSAR quando a propriedade sendo prevista varia com o tamanho da molécula. Descritores topológicos são números que dão informações sobre o arranjo de ligações em uma molécula, a exemplo do índice de Weiner, índices de Randic, índices Kier e Hall e o índice de conectividade. Esses descritores podem quantificar algumas características, como, por exemplo, se a molécula é flexível ou rígida. Descritores geométricos são valores que descrevem a forma e o tamanho da molécula, alguns exemplos são momentos de inércia, o volume molecular e área da superfície molecular [12].

Descritores eletrostáticos são valores individuais que dão informações sobre a distribuição de carga molecular, tais como momento de dipolo e polarizabilidade. Descritores de química quântica dão informações sobre a estrutura eletrônica da molécula, estes incluem energias dos orbitais de fronteira (HOMO e LUMO), refratividade, potencial de ionização, afinidade eletrônica, entre outros [12].

Cálculos químico-quânticos são, então, uma grande fonte de descritores moleculares que podem, em princípio, expressar muitas propriedades geométricas e eletrônicas das moléculas e suas interações. Os métodos de química quântica podem ser aplicados em relações quantitativas estrutura-

atividade (QSAR) pela derivação direta dos descritores eletrônicos a partir da função de onda molecular [15].

A descrição teórica da matéria, bem como de muitos processos químicos, físicos e biológicos requer métodos precisos para a descrição das interações atômicas e de escala molecular. Embora existam muitas abordagens mecânico quânticas, nas últimas décadas a teoria do funcional da densidade (DFT) se estabeleceu como o método teórico de escolha para esta tarefa, sendo também utilizado em grandes partes da física, química e ciência dos materiais. O aumento do uso da DFT e sua aceitação na academia e na indústria tem sido amplamente discutido. O sucesso por trás desta teoria está no funcional de troca e correlação, que utiliza a densidade eletrônica para descrever intrinsecamente o efeito de muitos corpos com o formalismo de uma única partícula [20].

3.2.3 Seleção de Variáveis

A seleção de variáveis permite eliminar os termos que não são relevantes na modelagem. Isso gera um subconjunto com o melhor número de variáveis e que apresente maior sensibilidade ao alvo de interesse. Desta maneira, este procedimento minimiza ou até mesmo elimina características potenciais dos interferentes [21].

3.2.3.1 Peso de Fisher

Uma maneira de se fazer a redução de variáveis, consiste em decidir se um valor é mantido ou eliminado da matriz de dados. Pode-se também aplicar um critério mais suave, calculando o peso da variância ou o *peso de Fisher* para cada variável [22]. O Peso Fisher é uma técnica capaz de identificar as variáveis mais significativas para as análises e é dado pela **equação 2** [23].

$$W_{AB}(i) = \frac{(\bar{X}_i(A) - \bar{X}_i(B))^2}{S_i^2(A) + S_i^2(B)} \quad (2)$$

onde $\bar{X}_i(A)$ representa a média de todas as amostras na classe A, $\bar{X}_i(B)$ representa a média de todas as amostras na classe B, $S_i^2(A)$ a variância das

amostras da classe A e $S_t^2(B)$ a variância das amostras da classe B. Quanto maior o Peso Fisher, maior a capacidade da variável na separação. Neste trabalho, o conjunto de dados foi dividido em duas classes: Classe A, moléculas ativas e Classe B, moléculas inativas.

3.2.3.2 *Ordered Predictor Selection (OPS)*

A *seleção de preditores ordenados* (OPS) é um algoritmo desenvolvido para efetuar a seleção das variáveis e já foi usado com sucesso em estudos de QSAR/QSPR. A ideia básica desse algoritmo é atribuir importância a cada descritor com base em um vetor informativo. As colunas da matriz são rearranjadas de modo que os descritores mais importantes apareçam nas primeiras colunas. Em seguida, são realizadas sucessivas regressões PLS (**seção 3.3.6**) aumentando-se o número de descritores de modo a otimizar o modelo. O melhor modelo de regressão pode ser escolhido de acordo com alguns dos parâmetros citados na **seção 3.5**, o algoritmo OPS ^[24] está implementado no programa *QSAR modeling* ^[25].

Esse é o primeiro programa que implementa o método OPS de seleção de variáveis que recentemente foi desenvolvido. Incorpora os processos de validação cruzada *leave-N-out* e aleatorização de *y* (*y-randomization*), detalhados na **seção 3.5**, além de realizar a detecção de amostras anômalas conhecidas na literatura como *outliers*. A detecção destes compostos, frequentemente negligenciada em programas de QSAR, é implementada combinando os valores de influência (*leverage*) das amostras aos seus respectivos *resíduos de Student*. A influência indica exatamente o que o nome diz: a sua capacidade de influenciar na estimativa dos coeficientes de regressão, enquanto que o *resíduo de Student* é um resíduo (diferença entre o valor experimental da atividade biológica e o valor calculado pelo modelo de regressão) padronizado, obtido dividindo-se o resíduo por uma estimativa de seu próprio desvio padrão ^[25].

A detecção de amostras anômalas feita pelo programa *QSAR modeling* permite que o usuário escolha o número de variáveis latentes que serão usadas pelo modelo PLS e fornece como resultado uma tabela com os valores de

influência e do *resíduo de Student* para cada um dos compostos no conjunto de treinamento. Amostras que apresentam simultaneamente valores de influência e *resíduo de Student* acima dos limites supraindicados são atípicas e devem ser excluídas do conjunto de dados ^[25].

Na **Figura 4** é apresentado um resumo do método OPS. Em geral, a essência do método é obter um vetor (vetor informativo) que contém informações sobre a localização das melhores variáveis de resposta para a predição (Figura A). Este vetor informativo pode ser composto pelos coeficientes de correlação da variável dependente com cada variável independente (correlograma). Pode ser também composto pelos próprios coeficientes de regressão, partindo da premissa de que baixos coeficientes de regressão denotam pequena importância da variável para o modelo. Ou então pode ser composto por uma avaliação conjunta de diversos vetores informativos.

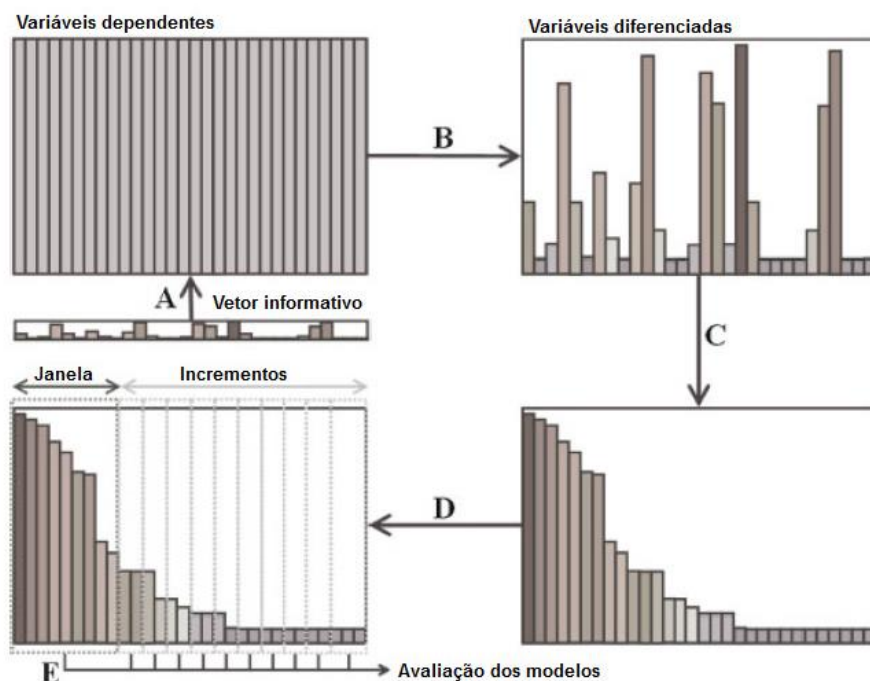
Neste trabalho, a metodologia desenvolvida baseia-se em vários vetores informativos e suas combinações. Essa combinação é obtida executando o produto do valor absoluto de cada elemento em um vetor vezes o elemento correspondente no outro vetor. Antes de fazer isso, os vetores são normalizados. No presente trabalho, pares desses vetores foram investigados^[24].

No segundo passo (Figura B), as colunas da matriz X são diferenciadas de acordo com os valores absolutos correspondentes dos elementos vetoriais informativos obtidos anteriormente no passo A. A matriz de variáveis independentes X tem suas colunas rearranjadas, de forma a posicionar-las em ordem decrescente do valor absoluto do vetor informativo obtido na etapa anterior (Figura C). Na quarta etapa (Figura D), modelos de regressão multivariada são construídos e avaliados usando uma estratégia de validação cruzada. Um subconjunto inicial de variáveis (janela) é selecionado para construir o primeiro modelo, em seguida, esta matriz é expandida pela adição de um número fixo de variáveis (incremento) e um novo modelo é construído e avaliado. Novos incrementos são adicionados até que todos os parâmetros de qualidade obtidos dos modelos são avaliados e armazenados para comparação futura.

Na última etapa (Figura E), os subconjuntos (janela e incrementos) são comparados usando os parâmetros calculados durante as validações e o modelo

com os melhores parâmetros de qualidade devem conter variáveis com capacidade de previsão e assim estas são as variáveis selecionadas [24,26].

Figura 4: Resumo da técnica OPS.



Fonte: Adaptado de Teófilo; Martins; Ferreira, 2008 por Alencar, 2012 [25,27].

3.2.3.3 Algoritmo Genético (GA)

Algoritmos genéticos (GA) são algoritmos de busca baseados nos mecanismos de seleção natural e genética. Eles combinam a sobrevivência entre os melhores indivíduos com uma forma estruturada de troca de informação genética entre dois indivíduos para formar uma estrutura heurística de busca [27].

O GA se assemelha ao processo evolutivo natural, pois sua estrutura permite que os descritores do sistema em estudo sejam codificados de forma binária, ficando com uma estrutura análoga aos dos cromossomos biológicos. Por meio de um gerador aleatório, uma população de cromossomos é gerada aleatoriamente, de forma a evitar influências tendenciosas externas. Em seguida, testa-se a aptidão do indivíduo, verificando, matematicamente, quais apresentam melhor resposta para o sistema. Nesta dissertação, em classificação, a aptidão usada para o GA foi considerada como o inverso do risco

de classificação incorreta G (**equação 3**) calculado usando os descritores codificados no cromossomo [27].

$$G = \frac{1}{K_V} \sum_{K=1}^{K_V} g_k \quad (3)$$

onde o risco de uma classificação incorreta do objeto x_k da $késima$ amostra de validação é definido como na **equação 4**:

$$g_k = \frac{r^2(x_k, \mu_{Ik})}{\min_{Ij \neq Ik} r^2(x_k, \mu_{Ij})} \quad (4)$$

Na equação anterior, o numerador é o quadrado da distância de Mahalanobis entre o objeto, x_k (com índice de classe Ik) e a média de sua classe (μ_{Ik}). O denominador da **equação 4** corresponde ao quadrado da distância de Mahalanobis entre o objeto μ_{Ik} e o centro da classe errada mais próxima. Idealmente, g_k deverá ser tão pequeno quanto possível, ou seja, o objeto x_k deverá estar perto do centro da sua verdadeira classe e distante dos centros das demais classes [27].

Os indivíduos que gerarem as melhores respostas, ou seja, os menores valores de G , irão ter seus parâmetros cruzados para geração de novas combinações, descendentes. Alguns indivíduos descendentes podem ainda receber características diferentes dos pais e sofrerem mutações. Gerações e mutações ocorrem até que a condição de avaliação seja alcançada [21].

E para a técnica de seleção de variáveis em calibração multivariada, uma matriz $\mathbf{X}$ de variáveis independentes e um vetor $\mathbf{y}$ de variável dependente, o algoritmo escolhe um subconjunto de variáveis em $\mathbf{X}$ e, usando algum método de regressão multivariada (aqui, os métodos PLS e MLR) e alguma estratégia de validação cruzada, é determinado o valor de RMSECV (raiz do erro quadrático médio de validação cruzada apresentada na **equação 5**) para o subconjunto de variáveis. O algoritmo genético usa este conceito iterativamente para localizar o subconjunto ou subconjuntos que fornece(m) o menor valor de RMSECV [28].

$$RMSECV_h = \sqrt{\frac{\sum_{i=1}^{I_{cv}} (y_i - \widehat{y}_i)^2}{I_{cv}}} \quad (5)$$

em que $\widehat{y}_i$ é o valor estimado para a amostra i , não incluída na construção do modelo, h é o número de componentes no modelo, I_{cv} é o número de amostras da validação interna.

3.3 Quimiometria aplicada aos Estudos QSAR

Quimiometria é o uso de estatística e matemática para analisar dados químicos [29]. Uma área de atuação bem-sucedida da quimiometria é a química medicinal e computacional, especialmente nos estudos de QSAR, que são úteis para o planejamento de novos fármacos[30]. Os métodos quimiométricos utilizados neste trabalho serão descritos nas próximas seções.

3.3.1 Análise Hierárquica de Agrupamentos (HCA)

HCA é uma técnica não-supervisionada que examina as distâncias entre pontos de todas as amostras e representa essa informação sob a forma de um esquema bidimensional chamado dendrograma [29]. O dendrograma apresenta os resultados na forma de uma árvore hierárquica e os comprimentos dos ramos representam o grau de similaridade entre os objetos[30].

Para proceder a análise hierárquica por agrupamento, a distância euclidiana é a medida mais comum e intuitiva de distância. A distância entre dois pontos no espaço bidimensional pode ser expressa de acordo com a **equação 6** [30].

$$d_{AB} = [\sum_{j=1}^J (x_{aj} - x_{bj})^2]^{1/2} \quad (6)$$

em que x_{aj} e x_{bj} são os valores numéricos da j -ésima coordenada da amostra A e de B , respectivamente.

Para calcular a similaridade (grandeza entre 0 e 1) entre duas amostras A e B utiliza-se a **equação 7**, onde d_{AB} é a distância A e B e $d_{máx}$ é a distância

máxima entre todas as amostras consideradas. As amostras são ditas similares se elas apresentarem valores de S_{AB} próximos da unidade ^[30].

$$S_{AB} = 1,0 - \frac{d_{AB}}{d_{máx}} \quad (7)$$

Outro parâmetro que deve ser levado em consideração é a escolha do critério com que os subagrupamentos serão ligados e para isso diversas técnicas de conexão são utilizadas, como por exemplo: (i) *método do vizinho mais próximo ou método simples*: nesse método de agrupamento, a distância entre dois grupos é dada pela menor de todas as distâncias entre os objetos dos dois grupos; (ii) *método do vizinho mais distante ou o método completo*: a distância entre os grupos é dada pela maior de todas as distâncias entre quaisquer dois objetos dos grupos em questão; (iii) *método da média*: o critério de agrupamento neste caso é a média ponderada entre os objetos de ambos os grupos; (iv) *método centróide*: utiliza a soma quadrática das distâncias para agrupar as amostras, utiliza o “centro” de cada grupo no espaço para representá-lo como um todo^[30].

3.3.2 Análise de Componentes Principais (PCA)

PCA é um método utilizado para projetar os dados multivariados em um espaço de dimensão menor reduzindo, assim, a dimensionalidade do espaço original do conjunto dos dados, sem que as relações entre amostras sejam afetadas. Utilizando essa metodologia, é possível descobrir, visualizar e interpretar as diferenças existentes entre as variáveis e examinar as relações que podem existir entre as amostras ^[30].

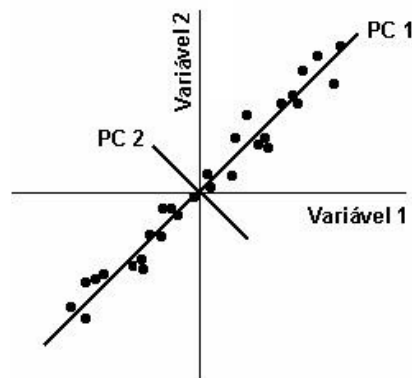
Do ponto de vista matemático, considera-se inicialmente uma matriz de dados $\mathbf{X}(\mathbf{m} \times \mathbf{n})$, sendo que $\mathbf{m}$ corresponde ao número de amostras e $\mathbf{n}$ o número de variáveis, que pode ser decomposta em 3 outras matrizes, $\mathbf{U}$, $\mathbf{S}$ e $\mathbf{V}$, conforme **equação 8**:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^t \quad (8)$$

As colunas de $\mathbf{U}$ e $\mathbf{V}$ são ortogonais. A matriz $\mathbf{V}$ é a matriz dos pesos, em que a primeira coluna contém os pesos de PC1 e assim por diante. O produto $\mathbf{U} \times \mathbf{S}$ corresponde à matriz $\mathbf{T}$ dos escores. Por fim, $\mathbf{S}$ é matriz diagonal, cujos elementos (valores singulares) contém informações sobre a quantidade de variância que cada componente principal descreve ^[16]. A matriz de escores é a projeção das amostras no novo sistema de coordenadas e os pesos são os coeficientes que determinam o peso de cada variável da matriz original em cada PC obtida ^[29].

Na prática, o processo de análise PCA parte da matriz dos descritores, a partir da qual as amostras podem ser projetadas em um espaço multidimensional. A primeira PC é traçada de forma a incorporar a maior variabilidade dos dados originais projetados, gerando um único eixo e um novo conjunto de coordenadas. A segunda PC é traçada perpendicularmente à primeira e assim sucessivamente. A representação gráfica de um conjunto de dados em um espaço bidimensional é ilustrada na **Figura 5**.

Figura 5: Primeira e segunda PC em um espaço bidimensional.



Fonte: Santos, 2009 ^[31].

A escolha do número de componentes principais utilizadas na descrição dos dados depende da particularidade de cada conjunto de dados e é feita levando-se em conta a porcentagem de variância descrita pelas PCs e a variância residual. Pode-se dizer que a escolha do número de PC deve permitir

a descrição do sistema com um número de fatores que descrevam a variação máxima, sem introduzir ruídos e informações desnecessárias ^[31].

3.3.3 K-ésimo Vizinho mais Próximo (KNN)

Nas técnicas de reconhecimento de padrões supervisionadas, o alvo é a construção de modelos de classificação em que a propriedade de interesse é categórica (discreta) ^[30].

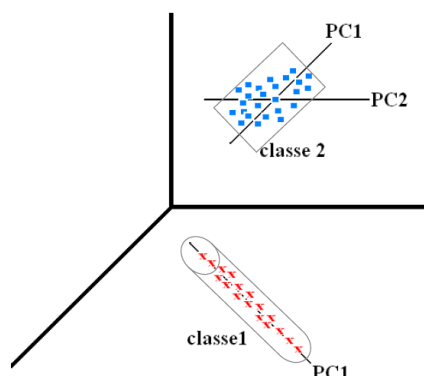
Uma vez definido o conjunto de treinamento, a próxima etapa é a construção propriamente dita do modelo de classificação KNN. Durante a construção do modelo, cada amostra do conjunto treinamento é excluída uma única vez e então classificada, usando-se, para isso, as amostras restantes. São calculadas as distâncias entre a amostra excluída e todas as outras amostras do conjunto de treinamento no espaço dimensional. Pode ser usada a distância euclidiana, que foi definida na **equação 6**. As distâncias de todas as amostras à amostra em questão são colocadas em ordem crescente para facilitar a identificação dos seus k vizinhos mais próximos. Essa amostra que havia sido excluída é, então, classificada de acordo com a maioria dos “votos” de seus vizinhos mais próximos ^[30].

Uma vez definido o valor de K , o modelo selecionado deve ser testado, e a classe prevista de amostras desconhecidas é atribuída como a classe das amostras mais próximas a ela no espaço multidimensional ^[29].

3.3.4 Modelagem Independente por Analogia de Classes (SIMCA)

Em contraste com KNN, que é baseado apenas em distâncias físicas entre as amostras, o método SIMCA constrói modelos de componentes principais, um para cada classe no conjunto de treinamento. O número de PCs ideais é determinado para cada classe e o modelo é completado pela definição de regiões de fronteira para cada modelo de PCA. Em outras palavras, é construída uma hipercaixa para cada classe. As componentes principais definem a orientação da caixa e os limites dentro do qual as amostras do conjunto de treinamento residem ^[17].

Figura 6: Exemplo do modelo SIMCA



Fonte: Sanches, 2009 [32].

Se um modelo não tiver poder de discriminação suficiente, a amostra pode ser atribuída a mais de uma classe, ou ainda, quando ela for um *outlier* em relação ao conjunto de treinamento, não pertencerá a nenhuma das classes [33].

A classificação de uma amostra externa àquelas empregadas na modelagem SIMCA é baseada nas diferentes projeções possíveis no espaço dos escores avaliando-se a possibilidade da mesma se encontrar dentro das fronteiras definidas para a respectiva classe em todas as projeções obtidas [34].

3.3.5 Análise Discriminante pelos Mínimos Quadrados Parciais (PLS-DA)

A Análise Discriminante pelos Mínimos Quadrados Parciais (PLS-DA) é um método de reconhecimento de padrões supervisionado baseado na regressão PLS para construção de modelos preditivos que buscam relacionar as respostas instrumentais ($\mathbf{X}$) com as propriedades de interesse, ou variáveis de resposta ($\mathbf{y}$), utilizando os rótulos das classes como um vetor dependente $\mathbf{y}$ [35]. As fontes relevantes de variância dos dados são modeladas pelo chamado Variáveis Latentes (VLs), é realizada com base em procedimentos de validação cruzada e obtida maximizando a covariância entre os escores da matriz e $\mathbf{y}$, assim se otimiza a decomposição na direção de previsão da propriedade $\mathbf{y}$ [36].

A principal diferença entre os dois está nas variáveis dependentes, uma vez que no PLS-DA elas representam valores qualitativos (classes) e não quantitativos. Em problemas de classificação binária, a variável $\mathbf{Y}$ pode ser arbitrariamente definida como 0 e 1, sendo 1 para amostras pertencentes à

classe, e 0 para amostras não pertencentes. O modelo então fornecerá uma resposta numérica, assim como nos métodos quantitativos, e este valor poderá não ser exatamente 1 ou 0. Então, um limite de decisão (*threshold*) precisa ser estabelecido [37]. A escolha mais simples seria adotar arbitrariamente o valor 0,5 como o limite entre as duas classes. No entanto, para definir o limite entre duas classes, utilizou-se um critério mais refinado, a teoria Bayesiana, onde o limiar é selecionado no ponto em que o número de falsos positivos e falsos negativos são minimizados [36,38].

Dois importantes parâmetros de avaliação: sensibilidade e especificidade são utilizados para avaliar os resultados do modelamento de classes. Sensibilidade é a capacidade do modelo em classificar corretamente as amostras da classe de interesse e especificidade é a habilidade do modelo em identificar corretamente as amostras que não são de interesse[36,38]. Os dois parâmetros são calculados a partir das **equações 9 e 10**:

$$SEN = \frac{TP}{TP+FN} \quad (9)$$

$$ESP = \frac{TN}{TN+FP} \quad (10)$$

onde:

TP é o número de verdadeiros positivos,

FN é o número de falsos negativos,

TN é o número de verdadeiros negativos e

FP é o número de falsos positivos.

3.3.6 Técnicas de Calibração Multivariada

Na etapa de desenvolvimento da QSAR o principal objetivo é encontrar a melhor expressão matemática, correlacionando a atividade com os parâmetros físico-químicos. São utilizadas técnicas estatísticas multivariadas e os modelos obtidos podem ser lineares ou não. Para se obter modelos lineares são utilizadas MLR e PLS [39].

Um dos primeiros modelos de calibração utilizados para análise de dados químicos é a Regressão Linear Múltipla, MLR. De uma maneira geral, a MLR é geralmente utilizada para ajustar um modelo de regressão linear onde a variável de resposta representada por **y** é uma combinação linear de **m** variáveis (descritores), com coeficientes **b**. O desvio entre **y** e o modelo é denominado resíduo (**e**) [39]. Assim, uma equação genérica de MLR pode ser representada como a **equação 11**:

$$y = x_1b_1 + x_2b_2 + \cdots + x_mb_m + e \quad (11)$$

A MLR apresenta a desvantagem de ser extremamente sensível à presença de colinearidade entre os descritores. Na aplicação do método de MLR a um conjunto de variáveis altamente correlacionadas, os coeficientes da regressão podem tornar-se instáveis e sem significado [17].

O método de PLS oferece vantagens sobre o método de MLR, pois a limitação para a razão entre o número de descritores e o número de compostos é reduzida, bem como não há sensibilidade às colinearidades entre as variáveis descritivas. Além disso, na técnica de MLR todas as variáveis selecionadas são incorporadas na matriz de correlação, sem qualquer parâmetro de distinção, quanto ao grau de relevância das mesmas. Assim, uma quantidade significativa de informação irrelevante é incorporada, o que não ocorre quando se usa o método PLS [17].

Da mesma maneira que o PCA, o método PLS encontra combinações lineares das variáveis independentes originais que contenham quantidades máximas de variação. Entretanto, no método PLS, a matriz de pesos é definida de forma que não apenas a variância seja maximizada, mas também o produto da variância pela correlação com Y seja otimizada. Assim, na forma matricial, a equação de regressão é descrita na **equação 12** [29].

$$Y = X\beta + F \quad (12)$$

Onde β é o vetor de regressão e F representa os erros na estimativa de Y .

O melhor modelo é escolhido com base na soma dos quadrados dos erros de predição (PRESS) obtidos no procedimento de validação cruzada do modelo. O número ótimo de componentes PLS é aquele que minimiza o valor PRESS (**equação 12**). A qualidade estatística dos modelos PLS pode ainda ser atestada pelos resíduos de predição e através dos parâmetros Coeficiente de Determinação do Modelo, r^2 (obtido na calibração para todo conjunto treinamento) e Coeficiente de Determinação da Validação Cruzada, q^2 , definidos nas **equações 14 e 15**, respectivamente.

3.4 Validação dos modelos QSAR

A validação de um modelo QSAR requer parâmetros que avaliem a qualidade preditiva e de robustez dos modelos construídos a partir do conjunto treinamento, os quais estão inseridos no contexto da *validação interna*. Não obstante, o modelo deve ser também avaliado utilizando amostras não inseridas na sua construção. Esta etapa, a qual utiliza o já definido conjunto teste, corresponde ao procedimento conhecido como *validação externa* [40,41,42,43].

Um dos principais métodos de validação interna de modelos QSAR é a *Validação Cruzada*, CV (do inglês, “cross-validation”), utilizando a técnica *leave-one-out*. Este procedimento consiste em retirar uma das amostras (compostos) do conjunto treinamento e construir um modelo de regressão para as amostras que sobram. Com este modelo, realiza-se a previsão do valor da variável dependente (atividade biológica) para a amostra excluída e compara-se com o seu valor verdadeiro. Isto acontece até que todas as amostras tenham sido submetidas ao teste. A partir dos valores de atividade previstos e experimentais podem-se calcular diversos parâmetros de qualidade. Dentre estes parâmetros está a Soma dos Quadrados dos Erros Residuais de Predição, *PRESS* (do inglês, “*Predicted Residual Error Sum of Squares*”), o qual pode ser utilizado na construção de parâmetros derivados. Sua forma é expressa pela **equação 13**.

$$PRESS = \sum (y_i^{pred} - y_i^{exp})^2 \quad (13)$$

Onde y_i^{pred} é a atividade predita para a amostra “i” com o modelo construído sem esta amostra e y_i^{exp} corresponde ao valor experimental (verdadeiro) da atividade.

A prática tem demonstrado que este processo sozinho não assegura a robustez do modelo na validação interna, o que é obtido pelo método *Leave-N-out*, LNO-CV, onde se eliminam **n** amostras por vez. Num conjunto com poucas amostras, recomenda-se fixar **n** entre 20% - 30% do total. A robustez do modelo é atestada se ocorrer diferença entre $q^2 - q^2_{LNO} < 0,1$.

Parâmetros Coeficiente de Determinação do Modelo, r^2 (obtido na calibração para todo conjunto treinamento) e Coeficiente de Determinação da Validação Cruzada, q^2_{loo} são dados pelas **equações 14 e 15**, respectivamente:

$$r^2 = 1 - \frac{\sum (y_i^{exp} - y_i^{cal})^2}{\sum (y_i^{exp} - y_i^{med})^2} \quad (14)$$

$$q^2_{loo} = 1 - \frac{\sum (y_i^{exp} - y_i^{pred})^2}{\sum (y_i^{exp} - y_i^{med})^2} \quad (15)$$

Onde y_i^{cal} corresponde ao valor da variável dependente calculado em um modelo gerado por todas as amostras treinamento, y_i^{pred} denota apenas os valores obtidos com *leave-one-out* e y_i^{med} corresponde à média dos valores de atividade experimentais.

Outro teste realizado é chamado *Y-randomization*, onde a variável dependente Y (atividade biológica) é randomicamente redistribuída no conjunto, mantendo inalterada a posição dos descritores. Esse tipo de validação avalia se o modelo foi obtido casualmente, ou seja, se existem outros modelos cujos valores dos coeficientes de determinação r^2 e a capacidade preditiva q^2_{loo} estão próximos aos valores do modelo original. Se este for o caso, pode-se afirmar que o modelo original é casual. Segundo Eriksson *et al.* [44], se forem atendidos os valores a $r^2 < 0,3$ e a $q^2 < 0,05$ o modelo de QSAR pode ser considerado robusto. Este estudo se utiliza de dez ciclos com 300 randomizações por ciclo para a validação, perfazendo 3000 randomizações [45].

A capacidade preditiva real dos modelos é analisada por validação externa utilizando um grupo teste que não foram incluídos no desenvolvimento dos modelos de QSAR. A capacidade preditiva na validação externa é expressa pelo valor de r^2 predito (r_{pred}^2), como calculado pela **equação 16**:

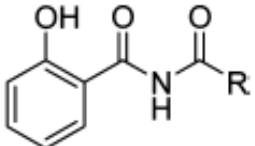
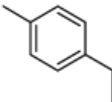
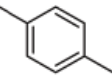
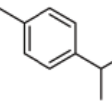
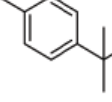
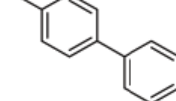
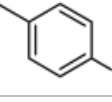
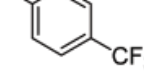
$$r_{pred}^2 = 1 - \frac{PRESS}{SS} \quad (16)$$

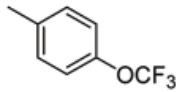
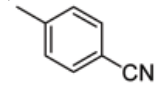
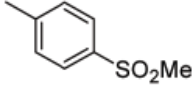
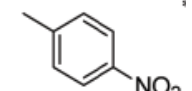
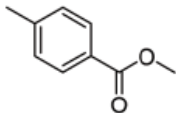
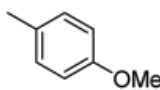
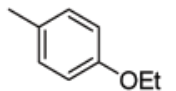
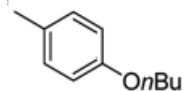
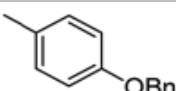
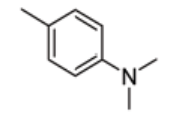
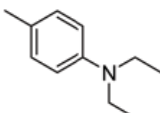
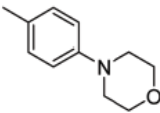
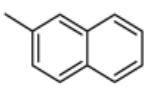
onde SS é a soma dos desvios quadrados entre o valor de atividade do conjunto de teste e o valor de atividade média do conjunto de treinamento e $PRESS$ é a soma dos quadrados do erro de predição proveniente da validação cruzada.

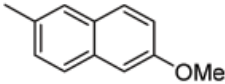
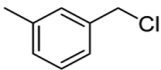
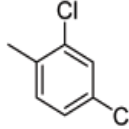
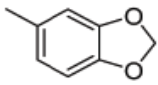
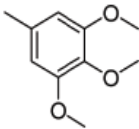
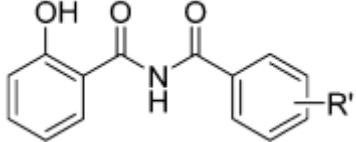
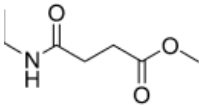
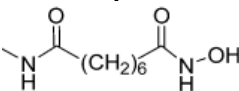
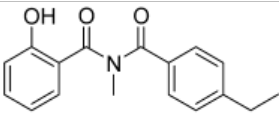
4 METODOLOGIA

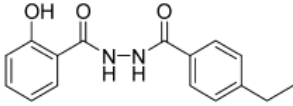
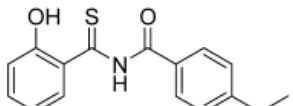
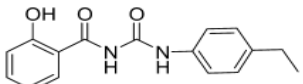
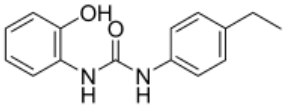
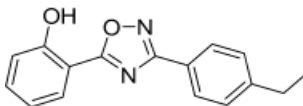
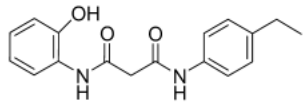
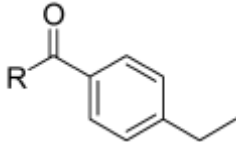
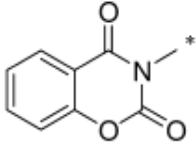
Para os estudos de modelagem, selecionou-se um conjunto de 39 derivados de N-benzoil-2-hidroxibenzamidas da literatura^[5] com valores de concentração inibitória de 50% (IC₅₀) contra o *Plasmodium falciparum* obtidos sob as mesmas condições experimentais. Os valores de IC₅₀ dos compostos foram transformados para a escala logarítmica pIC₅₀ (-log IC₅₀). Estes valores de atividade biológica do conjunto de dados variam de 4,42 a 8,30, uma faixa que se aproxima de quatro unidades logarítmicas, o que os torna adequados para as análises QSAR. As estruturas químicas dos compostos do conjunto de dados e seus respectivos valores de atividade biológica, expressa em pIC₅₀ e IC₅₀ (μg/mL), estão representados na **Tabela 1**.

Tabela 1: Estruturas químicas e respectivos valores de pIC₅₀ e IC₅₀.

				
Molécula	Radical - R	pIC ₅₀	IC ₅₀ (μg/mL)	Atividade
1a		5,25	5,6	ATIVA
1b		5,85	1,4	ATIVA
1c		5,82	1,5	ATIVA
1d		7,54	0,029	ATIVA
1e		5,41	3,9	ATIVA
1f		5,82	1,5	ATIVA
1g		5,42	3,8	ATIVA

Molécula	Radical - R	pIC ₅₀	IC ₅₀ (µg/mL)	Atividade
1h		4,66	22	INATIVA
1i		4,85	14	INATIVA
1j		4,42	38	INATIVA
1k		4,66	22	INATIVA
1l		5,54	2,9	ATIVA
1m		5,80	1,6	ATIVA
1n		6,00	1,0	ATIVA
1o		6,59	0,26	ATIVA
1p		5,39	4,1	ATIVA
1q		6,31	0,49	ATIVA
1r		8,30	0,005	ATIVA
1s		6,75	0,18	ATIVA
1t		6,66	0,22	ATIVA

1u		6,01	0,98	ATIVA
Radical - R	pIC ₅₀	IC ₅₀ (µg/mL)	Atividade	Atividade
1v		5,36	4,4	ATIVA
1w		4,85	14	INATIVA
1x		5,38	4,2	ATIVA
1y		6,11	0,78	ATIVA
1z		4,89	13	INATIVA
1aa		4,33	4,7	ATIVA
				
4a	<i>p</i> -NH ₂	5,41	3,9	ATIVA
4b	<i>o</i> -OH	5,96	1,1	ATIVA
5	<i>p</i> - 	4,68	21	INATIVA
8	<i>p</i> - 	5,46	3,5	ATIVA
Molécula	Estrutura	pIC ₅₀	IC ₅₀ (µg/mL)	Atividade
10		4,55	28	INATIVA

12		4,57	27	INATIVA
Molécula	Estrutura	pIC ₅₀	IC ₅₀ (µg/mL)	Atividade
13		5,38	4,2	ATIVA
16		4,59	26	INATIVA
19		5,69	2,1	ATIVA
23		4,57	27	INATIVA
28		4,92	12	INATIVA
				
21a		5,16	6,9	INATIVA

Para uma primeira análise de classificação das 39 amostras, adotou-se o seguinte critério: considerando que a molécula N-(4-etilbenzoyl)-2-hidroxibenzamida, molécula **1a** (**Figura 1**), possui moderada atividade antimalárica, com IC₅₀ igual a 5,6µg/mL e que a droga padrão, cloroquina, possui IC₅₀ igual a 0,11µg/mL, os análogos de N-(4-etilbenzoyl)-2-hidroxibenzamida

foram divididos em duas classes (ativos ou inativos): as amostras ativas foram consideradas as que possuem IC₅₀ menor que 5,6µg/mL e as inativas as que possuem IC₅₀ maior que 5,6µg/mL^[5].

Para a segunda análise, usou-se um conjunto de 33 amostras, a partir da seguinte verificação: algumas moléculas classificadas como ativas possuem baixos índices de seletividade (calculado como a razão entre o IC₅₀ para ratos mioblastos esqueléticos (L6) e o IC₅₀ para o respectivo protozoário parasita), ou seja, valores próximos às de moléculas inativas (**Tabela 2**), podendo interferir nos resultados e induzir a erros na classificação. As amostras com baixos índices de seletividade (**1g, 1v, 1x, 1aa, 13 e 19**) foram então retiradas do conjunto de dados.

Tabela 2: Amostras ativas e inativas com valores de índice de seletividade, SI* (Stec *et al.*, 2012).

Ativas	SI	Ativas	SI	Ativas	SI	Inativas	SI
1a	12	1o	50	1aa	3,4	1j	2,7
1b	33	1p	16	4a	10	1k	2,4
1c	12	1q	110	4b	36	1w	4,4
1d	80	1r	1640	5	>2,4	1z	0,18
1e	>26	1s	>272	8	>29	10	0,33
1f	25	1t	94	13	4,1	12	0,18
1g	3,4	1u	57	19	4,6	16	2,8
1l	34	1v	0,40	Inativas	SI	21a	3,7
1m	24	1x	56	1h	3,8	23	>3,7
1n	17	1y	26	1i	3,9	28	1,3

$$*SI = \frac{IC_{50} \text{ para L6}}{IC_{50} \text{ para P.falciparum}}$$

4.1 Otimização das geometrias e cálculo dos descritores

Inicialmente foi realizada a construção das estruturas de cada uma das moléculas em estudo utilizando o programa GaussView 5.0. Uma pré-otimização

com o método semiempírico AM1^[46] foi realizada, seguida de uma otimização mais rigorosa em nível DFT com o funcional M06-2X e o conjunto de base 6-311+g(d,p). Neste nível de cálculo foram obtidos diversos descritores químico-quânticos eletrônicos e estruturais, tais como energias dos orbitais de fronteira, momento dipolar, polarizabilidade e cargas NBO. Nesta etapa foi utilizado o programa Gaussian 09 ^[47].

Dentre os programas disponíveis para o cálculo de descritores físico-químicos e estruturais, o E-dragon 1.0 ^[48] foi escolhido por ser uma plataforma on-line de fácil acesso e de rápida obtenção dos resultados. O programa retorna uma matriz onde as linhas representam as amostras e as colunas representam os descritores, finalizando com uma matriz de 39 amostras *versus* 1388 descritores.

4.2 Análises Quimiométricas

Inicialmente foi realizada uma análise exploratória dos dados por HCA com a matriz completa contendo todos os descritores para observar o comportamento das amostras frente as variáveis. A medida utilizada para conceito de distância entre duas amostras foi a *distância euclidiana*, também utilizou-se diferentes técnicas de conexão dos agrupamento das amostras (**Figura 8 a 11**).

O pré-processamento utilizado neste conjunto de dados para os descritores foi o *autoescalamento*. Este procedimento consiste em subtrair de cada elemento de uma coluna da matriz de dados o valor médio da respectiva coluna e dividir o resultado pelo desvio-padrão dessa coluna. Este procedimento foi utilizado devido às diferentes ordens de grandeza de resposta encontrada para as variáveis consideradas. Assim, é possível garantir que as influências relativas de diferentes variáveis em todos os cálculos sejam independentes de suas unidades ^[49].

Na primeira análise, foram utilizadas todas as amostras selecionadas com base na análise exploratória do HCA e empregados os seguintes métodos quimiométricos, que podem ser classificados em três categorias: *reconhecimento de padrões não supervisionado* (HCA e PCA), *reconhecimento*

de padrões supervisionado (KNN, SIMCA e PLS-DA) e calibração multivariada (PLS e MLR). Na segunda análise, foram excluídas deste conjunto as amostras com baixo índice de seletividade e empregados os mesmos métodos quimiométricos da primeira análise.

O conjunto de dados foi dividido em dois conjuntos, *conjunto de treinamento* e *conjunto de teste*, usando o algoritmo Kennard-Stone (KS) ^[50], que foi aplicado separadamente para cada classe. As amostras de teste foram utilizadas apenas para a avaliação final dos dados e comparação dos modelos de classificação assinalando novos objetos.

Em seguida, foram usadas técnicas de reconhecimento de padrões supervisionadas a fim de obter modelos de classificação de amostras ativas e inativas do conjunto de dados. Para isso, o conjunto de dados foi dividido a priori em duas classes: **classe A** (amostras ativas, que possuem IC₅₀ menor do que 5,6µg/mL) e **classe B** (amostras inativas, que possuem IC₅₀ maior do que 5,6µg/mL). As técnicas de reconhecimento de padrões supervisionadas KNN, SIMCA e PLS-DA foram utilizadas com validação cruzada, e os modelos de classificação foram construídos com base nas variáveis selecionadas com as técnicas de Peso de Fisher e GA ^[51].

Para avaliação dos modelos, a taxa de classificação correta (TCC) das amostras foi utilizada para o conjunto de treinamento e teste (**equação 17**).

$$TCC (\%) = \frac{\text{Número de acertos de classificação}}{\text{número total de amostras}} \times 100\% \quad (17)$$

Onde o número de acertos de classificação significa o número de amostras alocadas corretamente em cada classe A e B.

Para a calibração multivariada, realizou-se um pré-tratamento dos dados a fim de excluir variáveis com baixa correlação com a atividade biológica. Assim, a matriz com os descritores químico-quânticos e os gerados pelo E-dragon foi reduzida eliminando-se os descritores que possuíam correlação com a atividade biológica menor do que 0,3. Além desses, foram eliminados os descritores que possuíam correlações entre si maiores do que 0,9, a fim de diminuir a colinearidade dos dados.

Em seguida, os descritores restantes foram selecionados para a construção de modelos QSAR utilizando os métodos de PLS e MLR. Primeiramente, utilizou-se as mesmas variáveis selecionadas pelos melhores modelos na classificação para gerar um modelo PLS. Posteriormente, outros métodos de seleção de variáveis para calibração foram utilizados para comparação: os métodos de seleção OPS ^[25] e GA ^[52].

Os modelos gerados com as diferentes combinações de técnica de seleção de variáveis e de regressão foram avaliados com base nos valores de PRESS (Soma dos Quadrados dos Erros de Predição), SEV (Erro Padrão de Validação), q^2 e r^2 provenientes de validação cruzada leave-one-out e r^2_{pred} da validação externa ^[40,41,42,43].

A técnica PLS foi utilizada neste trabalho com o intuito de modelar a propriedade experimental de interesse (pIC_{50}) com os descritores que representam as estruturas químicas dos compostos em estudo. O melhor modelo foi escolhido com base nos menores valores de PRESS e SEV obtidos no procedimento de validação cruzada. O número ótimo de componentes PLS é aquele que minimiza o valor de PRESS e SEV. Os valores dos coeficientes de determinação q^2 e r^2 são verificados para garantir que o modelo apresente boa qualidade. A técnica de MLR também foi realizada de forma comparativa ao modelo PLS, tendo em vista o pequeno número de descritores e ausência de correlação entre os mesmos.

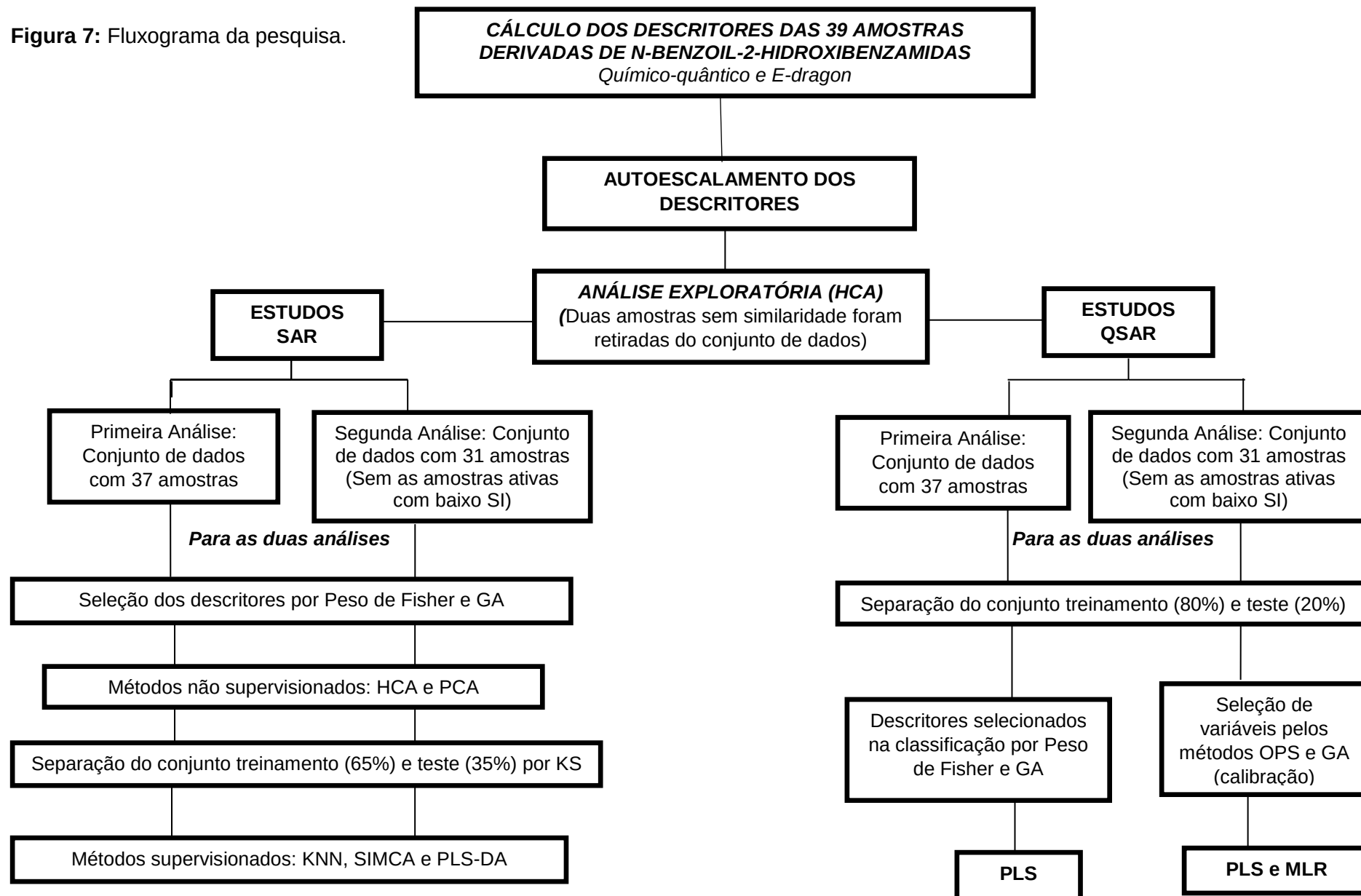
Uma vez escolhido o melhor modelo, validações internas adicionais foram realizadas para avaliar a robustez e a estabilidade dos modelos: os testes de leave-N-out e y-randomization.

Para garantir a aplicabilidade do modelo no planejamento de moléculas mais potentes, foi realizada uma validação externa, com o intuito de prever a atividade de moléculas não presentes no conjunto de calibração.

As análises quimiométricas foram realizadas usando os softwares Pirouette 3.11 ^[53] e o The Unscrambler 9.7 da CAMO Process AS. Os algoritmos de seleção de variáveis OPS, implementado no programa QSAR modeling, GA e as rotinas de classificação do PLS-DA foram executados em ambiente Matlab 7.10 (R2010a) (Mathworks, EUA), a seleção de variáveis pelo método de Peso

de Fisher no programa Microsoft Excel 2013®. Um fluxograma dos procedimentos realizados é apresentado na **Figura 7**.

Figura 7: Fluxograma da pesquisa.



5 RESULTADOS E DISCUSSÃO

5.1. Análise exploratória por HCA

A análise HCA realizada com todos os descritores e todas as amostras revelou o comportamento das amostras frente as variáveis. Utilizou-se distância euclidiana e diferentes técnicas de conexão dos agrupamentos das amostras (**Figuras 8, 9, 10 e 11**).

Figura 8: HCA com todos os descritores e o método de conexão completo.

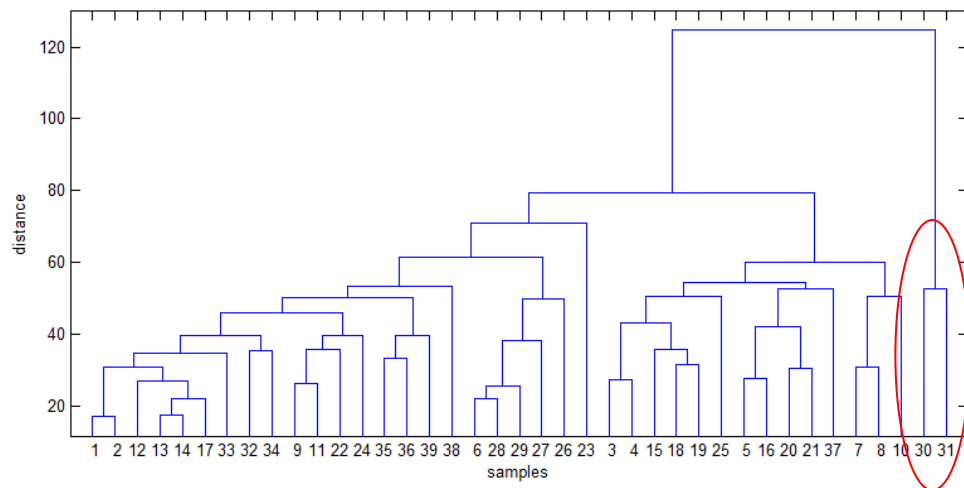


Figura 9: HCA com todos os descritores e o método de conexão simples.

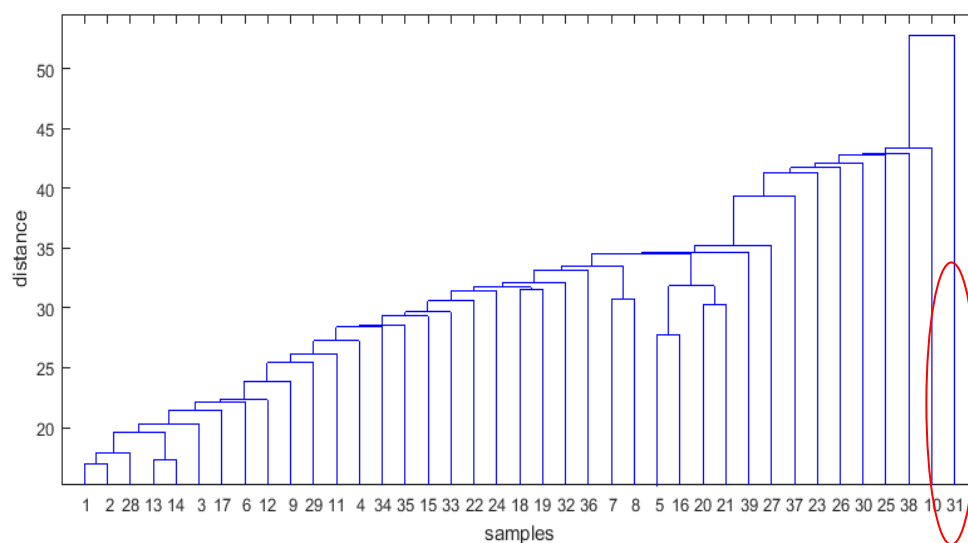


Figura 10: HCA com todos os descritores e o método de conexão da média.

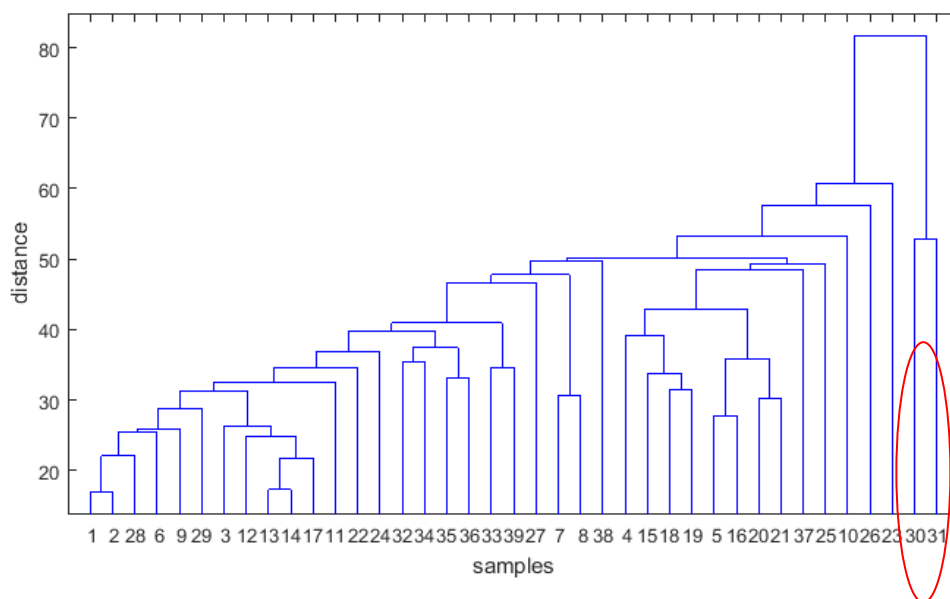
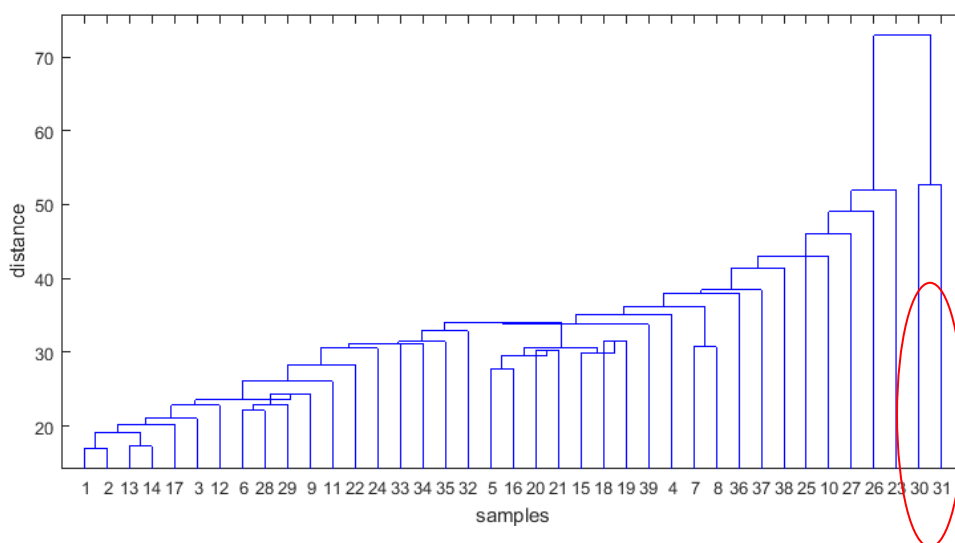


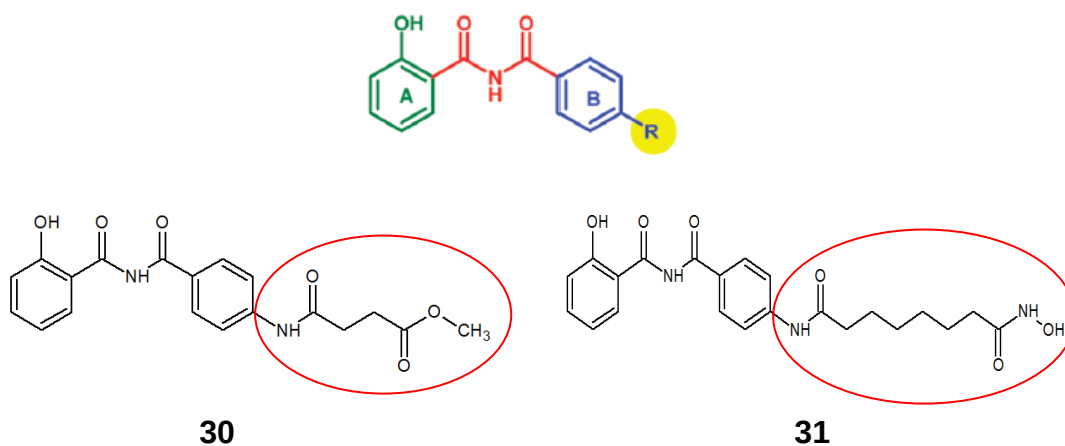
Figura 11: HCA com todos os descritores e o método de conexão centróide.



A análise geral dos agrupamentos evidencia que duas amostras **30** e **31**, (moléculas **5** e **8** – **Tabela 1**) aparecem isoladas e distantes das demais, para todos os métodos de conexão, exceto para o método simples, para o qual apenas a amostra **31** aparece isolada. A partir da verificação da estrutura de cada molécula na **Tabela 1** e analisando a **Figura 12**, nota-se que estas duas moléculas diferem significativamente do conjunto de dados quando se observa o radical na parte B da estrutura principal, que consiste de uma estrutura mais

longa quando comparada as das demais moléculas do conjunto de dados. Diante disto, as duas amostras foram retiradas do conjunto por possuírem baixa similaridade.

Figura 12: Estrutura principal, estrutura 30 e estrutura 31.



5.2 Estudos da Relação entre Estrutura e Atividade (SAR)

Nesta seção, são apresentados os resultados dos estudos SAR realizados por meio de duas análises. Como já foi mencionado anteriormente: na primeira análise, foram utilizadas as 37 amostras restantes após a exclusão das duas amostras com baixa similaridade e, na segunda análise, foram retiradas as amostras ativas, **1g**, **1v**, **1x**, **1aa**, **13** e **19**, por possuírem baixos índices de seletividade, restando 31 amostras. A **Tabela 3** apresenta os resultados da classificação utilizando as diferentes técnicas de seleção de variáveis e de classificação testadas, com o intuito de obter os descritores com maior poder de discriminação e que sejam capazes de classificar as amostras em ativas ou inativas. Anteriormente as análises supervisionadas, foram realizadas as análises não-supervisionadas, HCA e PCA, a fim de verificar a separação das amostras ativas e inativas a partir dos descritores selecionados pelos dois métodos: Peso de Fisher e GA.

Tabela 3: Resultados da classificação com as diferentes técnicas de seleção de variáveis e de classificação para conjunto treinamento e teste. Os valores em % referem-se ao cálculo da Taxa de Classificação Correta (**equação 17**).

Conjunto com 37 amostras				
Técnicas de seleção de variáveis				
Técnicas de Classificação	PESO DE FISHER		GA	
	Treinamento	Teste	Treinamento	Teste
KNN	80%	78,6%	84%	75%
SIMCA	78%	78,6%	96%	58,3%
PLS-DA	80%	78,6%	100%	91,7%
Conjunto com 31 amostras				
Técnicas de seleção de variáveis				
Técnicas de Classificação	PESO DE FISHER		GA	
	Treinamento	Teste	Treinamento	Teste
KNN	71,4%	100%	63,2%	66,7%
SIMCA	89,5%	91,7%	84,7%	90%
PLS-DA	90,5%	100%	100%	90%

Analisando a **Tabela 3**, a partir dos valores de TCC para o conjunto teste, pode-se concluir que, os melhores resultados foram para o conjunto com 31 amostras e seleção de variáveis por Peso de Fisher, para todos os métodos quimiométricos. Assim, serão detalhadas apenas para este conjunto de 31 amostras e as variáveis selecionadas pelo Peso de Fisher, as análises HCA e PCA, que antecederam as análises supervisionadas: KNN, SIMCA e PLS-DA, mostrando os detalhes dos modelos de classificação obtidos com estas técnicas.

5.2.1 Seleção de variáveis por Peso de Fisher

Após retirar as amostras com baixo índice de seletividade, as 31 moléculas restantes foram divididas em ativas e inativas de tal forma que as que possuem valores de IC₅₀ abaixo de 5,6µg/mL são classificadas como ativas (classe A), e as com valores acima foram classificadas como inativas (classe B). Encontra-se na **Tabela 1** os valores de IC₅₀ para cada composto e sua classificação ativo/inativo.

Os 1388 descritores passaram por uma seleção de acordo com o seguinte critério: descritores com peso de Fisher maior do que 0,64 foram selecionados e após este procedimento, foram testadas diferentes combinações destes descritores restando 13 variáveis. A escolha do valor 0,64 foi determinado a partir

da quantidade de descritores com maiores Pesos de Fisher. Com valores de Peso de Fisher abaixo de 0,60, selecionavam-se muitos descritores (mais de 40 descritores), enquanto valores acima de 0,70 selecionavam descritores que, sozinhos, não discriminavam as amostras em ativas e inativas. Com o valor de 0,64, foi possível obter a separação das amostras nas duas classes. Os descritores assim selecionados são mostrados na **Tabela 4**.

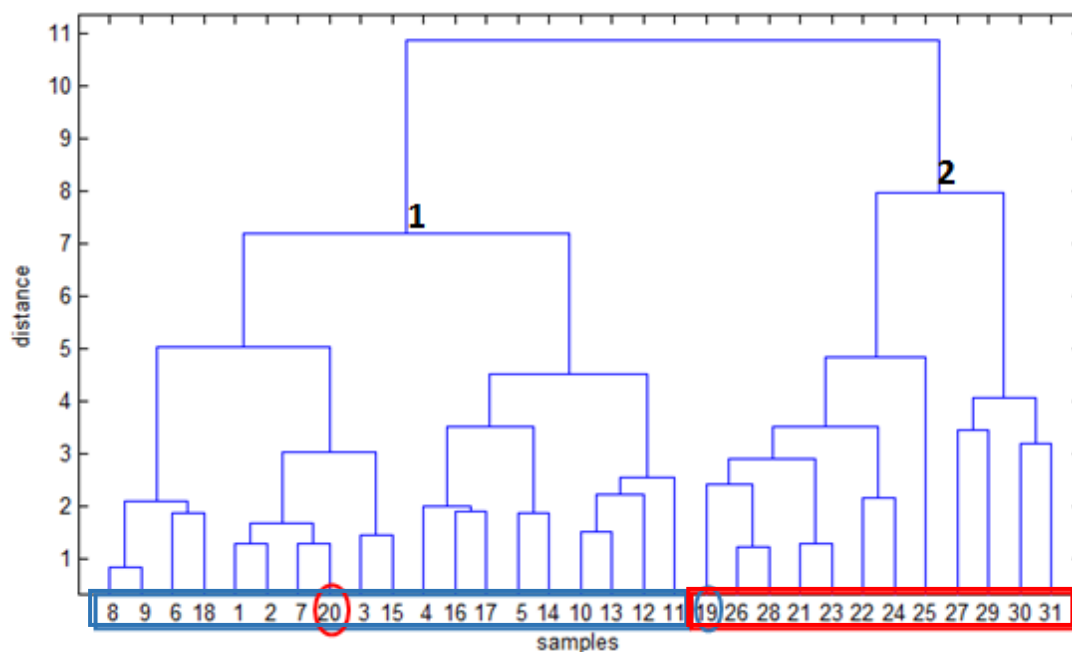
Tabela 4: Média, Variância e Peso de Fisher para as propriedades responsáveis pela separação das moléculas inativas e ativas.

Descritores	Média ativas	Média inativas	Variância ativas	Variância Inativas	Peso de Fisher
RDF085m	7,86	4,33	3,28	4,95	1,52
RDF050v	10,32	7,23	3,01	3,28	1,52
RDF085v	7,41	4,42	2,37	2,87	1,71
RDF050p	10,59	7,45	3,59	2,67	1,58
RDF085p	7,06	4,73	4,48	2,73	0,75
MATS6m	0,17	-0,07	0,03	0,04	0,84
MATS6v	0,17	-0,09	0,03	0,04	1,06
MATS6e	0,16	-0,05	0,03	0,05	0,64
MATS6p	0,18	-0,13	0,03	0,04	1,24
Mor18u	-1,72	-1,38	0,07	0,08	0,80
Mor18v	-1,00	-0,81	0,03	0,02	0,85
Mor18e	-1,71	-1,35	0,08	0,08	0,82
Mor18p	-1,05	-0,85	0,03	0,02	0,86

5.2.2 HCA

Após a seleção de variáveis com Peso de Fisher, observou-se a separação das moléculas em ativas e inativas a partir dos agrupamentos formados no dendrograma mostrado na **Figura 13**. A medida utilizada para conceito de distância entre duas amostras foi a *distância euclidiana* e a técnica de conexão dos agrupamentos das amostras foi o método completo.

Figura 13: Dendrograma das 31 amostras com seleção de variáveis pelo Peso de Fisher.

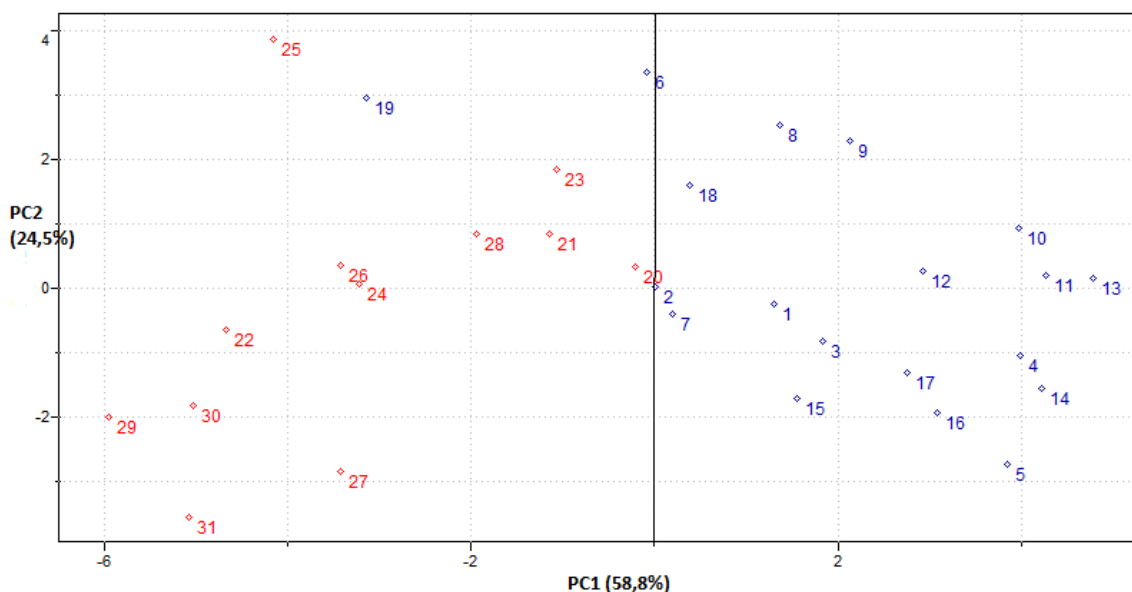


Ao analisar a **Figura 13** observam-se dois agrupamentos, o agrupamento 1, destacado em azul, são de moléculas ativas (classe A), e o agrupamento 2, destacado em vermelho, são de moléculas inativas (classe B). Verificou-se que as amostras **19** e **20** (moléculas **4b** e **1h**, respectivamente) não se agruparam corretamente, obtendo-se 93,5% de TCC. A formação dos agrupamentos demonstra que os descritores selecionados e adotados na classificação de atividades foram capazes de avaliar a similaridade entre os compostos por classes.

5.2.3 PCA

Para a análise de PCA foram utilizadas as 31 estruturas em estudo e as mesmas propriedades compondo 3 componentes principais (PCs), aplicando-se o procedimento de validação cruzada completa. A **Figura 14** mostra o resultado dos escores composto pela PC1 (58,8%), PC2 (24,5%) e PC3 (10,3%) em função das amostras, destacando os escores de PC1 x PC2, que juntas explicam a maior parte da variância.

Figura 14: Representação gráfica dos escores PC1 *versus* PC2.



A primeira componente explica 58,8% da variância dos dados e, considerando as três primeiras componentes, o resultado constitui 94,6% da variância total dos dados originais, fornecendo, assim, uma representação confiável dos mesmos. A análise PCA está em concordância com a análise de HCA: as amostras ativas estão destacadas em azul, e as inativas destacadas em vermelho. Esses resultados sugerem que esta análise tem uma boa representação e sobretudo, que essas diferenças discriminantes são modeladas pelos descritores selecionados.

A **Tabela 5** mostra os *loadings* de cada variável em PC1, PC2 e PC3. É importante observar a quantidade de variação descrita por uma PC ao interpretar os *loadings*: uma variável com um grande valor de *loading* indica que ela contribui significativamente para uma PC específica ^[29].

Nota-se que os descritores do tipo RDF descriptors possuem maior contribuição em PC1, os descritores do tipo 2D autocorrelations (MATS6m, MATS6v, MATS6e e MATS6p) possuem maior contribuição em PC2 e os descritores do tipo 3D-MoRSE descriptors possuem maior contribuição em PC3.

Tabela 5: *Loadings* das variáveis em cada PC's.

Propriedades	PC1	PC2	PC3
<i>RDF085m</i>	0,32	-0,03	0,28
<i>RDF050v</i>	0,31	0,02	0,32
<i>RDF085v</i>	0,32	-0,08	0,33
<i>RDF050p</i>	0,32	0,00	0,29
<i>RDF085p</i>	0,32	-0,09	0,33
<i>MATS6m</i>	0,21	0,42	-0,20
<i>MATS6v</i>	0,23	0,41	-0,17
<i>MATS6e</i>	0,21	0,42	-0,20
<i>MATS6p</i>	0,24	0,38	-0,14
<i>Mor18u</i>	-0,26	0,29	0,33
<i>Mor18v</i>	-0,28	0,27	0,30
<i>Mor18e</i>	-0,27	0,29	0,31
<i>Mor18p</i>	-0,28	0,27	0,31

5.2.4 Resultados do KNN

Para a construção dos modelos de classificação utilizando os métodos KNN, SIMCA e PLS-DA, foram utilizados os mesmos descritores selecionados por Peso de Fisher (ver Tabela 4). O conjunto de dados foi separado em conjuntos de treinamento, com 21 amostras (13 ativas e 8 inativas) e teste, com 10 amostras (6 ativas e 4 inativas) usando o algoritmo KS, que foi aplicado separadamente para cada classe.

A **Tabela 6** apresenta o sumário da classificação obtido com até sete vizinhos mais próximos. Todos os compostos do conjunto de treinamento foram classificados corretamente, o que mostra que as classes são bem distintas e que as variáveis selecionadas possuem boa habilidade para a discriminação entre as classes de compostos.

Tabela 6: Sumário da classificação obtida com a técnica KNN para o conjunto de treinamento com as variáveis selecionadas pelo Peso de Fisher.

Classe	Número de amostras	Número de compostos classificados incorretamente			
		K = 1	K = 3	K = 5	K = 7
A	13	2	2	2	3
B	8	4	4	4	4
Total	21	6	6	6	7
TCC (%)		71,4%	71,4%	71,4%	66,7%

A fim de avaliar o poder preditivo deste modelo (validação externa), ele foi empregado para classificar os compostos do conjunto de teste, não utilizado durante a construção do modelo de classificação. Seis amostras ativas e quatro inativas foram utilizadas no conjunto teste utilizando os mesmos critérios adotados para o conjunto de treinamento. Para K igual a 3, todas as amostras foram classificadas corretamente nas classes a que deveriam pertencer, de acordo com seus valores de IC_{50} (ativas, $IC_{50} < 5,6\mu\text{g/mL}$; inativas, $IC_{50} > 5,6\mu\text{g/mL}$).

5.2.5 Resultados do SIMCA

No método SIMCA foi utilizado o limite de confiança de 90% e cada classe foi auto escalada separadamente. O número de PCs necessário para descrever cada classe foi escolhido individualmente e mostrado na **Tabela 7** para os modelos construídos.

Tabela 7: Número de componentes principais usados nos dois modelos SIMCA.

	Número de PC	Variância
Classe 1	6	99,9%
Classe 2	5	99,7%

Para verificar o poder de discriminação dos modelos construídos foram utilizados os resíduos e as distâncias entre as classes os quais representam o quanto cada classe é compacta e o quão distante estão uma das outras, respectivamente. A **Tabelas 8** e **Tabela 9** mostram os resíduos e as distâncias entre as classes, respectivamente. Na **Tabela 8** pode-se verificar que as classes são compactas, ou seja, o resíduo de uma classe é menor que os resíduos da interclasse, para todos os modelos.

Tabela 8: Resíduo entre as classes para os dois modelos SIMCA.

	Classe 1	Classe 2
Classe 1	0,05	0,47
Classe 2	0,34	0,14

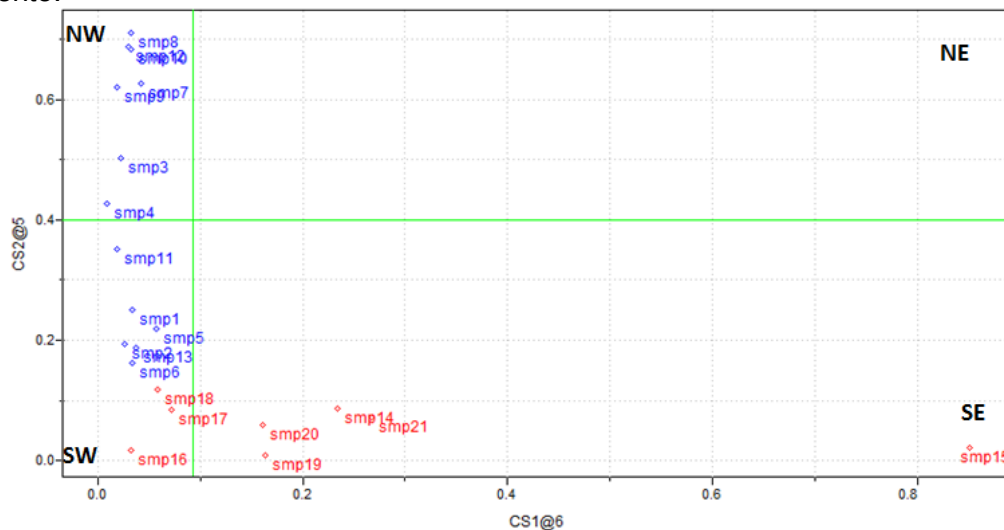
Analisando as distâncias entre as classes na **Tabela 9**, observa-se que elas estão suficientemente separadas ^[53], neste estudo a distância foi de 3,03, indicando um bom poder discriminante das variáveis.

Tabela 9: Distância entre as classes para os dois modelos SIMCA.

	Classe 1	Classe 2
Classe 1	0	3,03
Classe 2	3,03	0

Na **Figura 15** se analisa as distâncias das amostras às classes calculadas de acordo com os resíduos das amostras quando elas são ajustadas às classes. Este gráfico é dividido por duas linhas que representam valores críticos de variâncias residuais. Compostos posicionados no quadrante noroeste (NW) pertencem apenas à classe correspondente ao eixo-x, pois elas estão a distâncias pequenas o suficiente para serem consideradas membros desta classe. Da mesma forma, compostos no quadrante sudeste (SE) são membros apenas da classe do eixo-y. Compostos no quadrante sudoeste (SW) podem pertencer a ambas as classes, enquanto aquelas no quadrante nordeste (NE) pertencem a nenhuma das classes.

Figura 15: Distâncias dos compostos para às classes obtidas para o conjunto de treinamento.



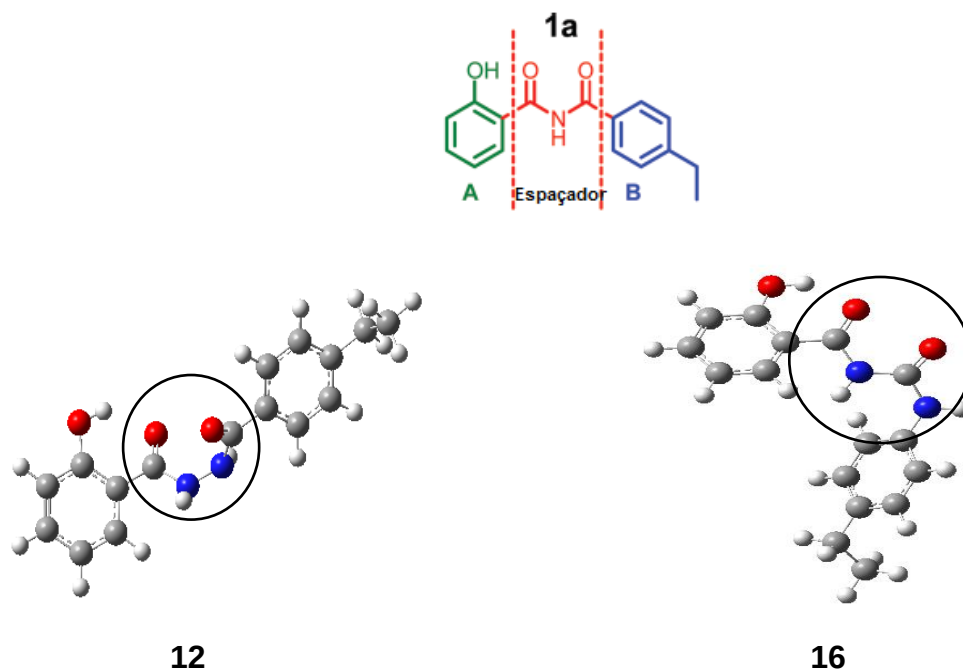
O melhor resultado obtido para o SIMCA atingiu uma TCC igual a 90,5% no conjunto de treinamento, **Tabela 10**, e 90% no conjunto de teste, **Tabela 11**. Para alcançar este resultado, o número ideal de componentes principais utilizado para cada classe foi: 6 PC's para a classe ativa e 5 PC's para a classe inativa.

Tabela 10: Matriz de confusão da classificação SIMCA.

Conjunto de Treinamento				
Classe verdadeira	Número de amostras	SIMCA		
		Ativos	Inativos	TCC (%)
Ativos	13	13	0	90,5%
Inativos	8	2	6	

As moléculas **12** e **16** inativas foram classificadas como ativas. Estas mesmas amostras também foram classificadas incorretamente em KNN. As estruturas destas duas amostras se assemelham a estrutura da amostra **1a** - ativa (**Figura 16**), mudando apenas a posição e quantidade de nitrogênio no espaçador de imida, e sem nenhuma alteração nos pontos A e B da estrutura, neste caso, os descritores selecionados não conseguiram discriminar estas diferenças.

Figura 16: Moléculas 1a, 12 e 16. O átomo de nitrogênio é representado na cor azul.

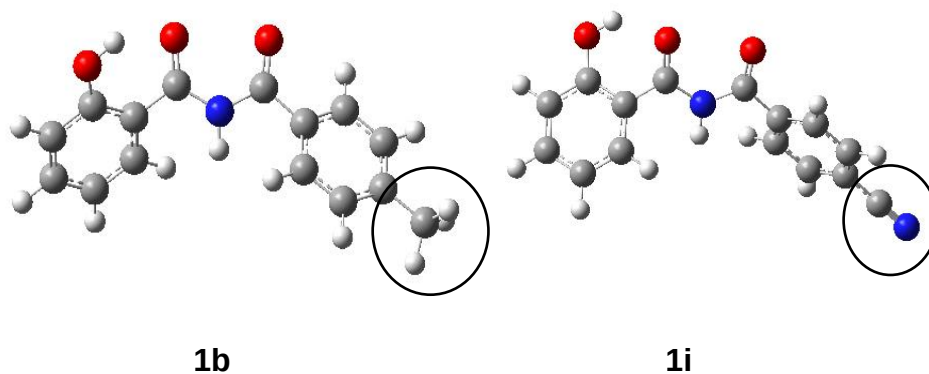


A fim de avaliar o poder preditivo deste modelo, ele foi empregado para classificar os compostos do conjunto teste, não utilizado durante a otimização do modelo (**Tabela 11**). Apenas uma amostra do conjunto teste não foi predita de forma correta: a estrutura inativa **1i** foi classificada como ativa. Na sua estrutura, há apenas uma modificação no ponto B (molécula **1a** – Figura 16) em relação a estrutura da molécula ativa **1b** (**Figura 17**), onde os átomos de hidrogênio da molécula **1b** foram substituídos por nitrogênio na molécula **1h**. Novamente para este modelo os descritores selecionados não conseguiram discriminar essas diferenças.

Tabela 11: Matriz de confusão da predição SIMCA.

Conjunto de Teste				
Classe verdadeira	Número de amostras	SIMCA		
		Ativos	Inativos	TCC (%)
Ativos	6	6	0	90%
Inativos	4	1	3	

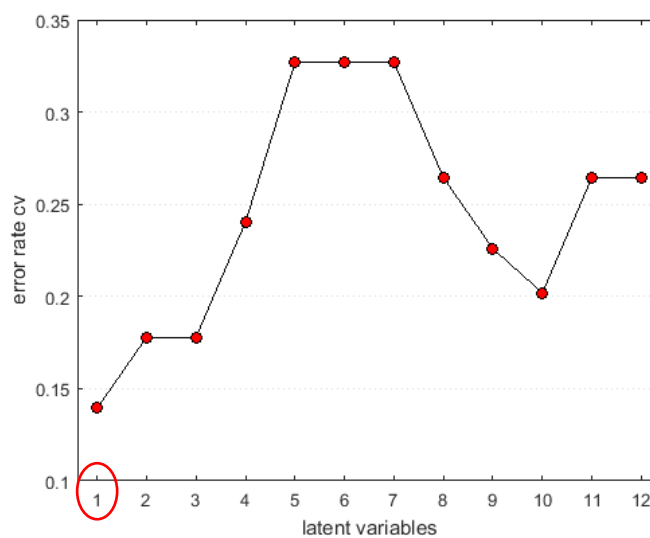
Figura 17: Moléculas 1b (ativa) e 1i (inativa).



5.2.6 Resultados do PLS-DA

Para selecionar o melhor número de VLs, foi feita uma validação cruzada usando o critério dos blocos contíguos (*contiguous blocks*). Foi selecionado o número de VLs que apresentou o menor erro de classificação de validação cruzada (CV) (**Figura 18**).

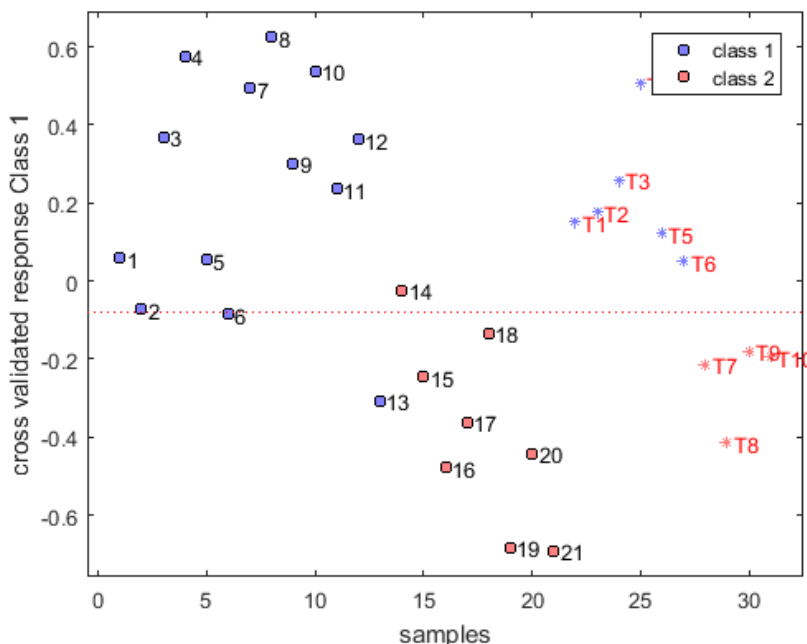
Figura 18: Número de variáveis latentes por erro de classificação de CV



Na **Figura 19** verifica-se os valores de classe estimada para o conjunto treinamento e teste. Amostras que se encontram acima do valor limite são classificadas como amostras ativas e as que se encontram abaixo são

consideradas amostras inativas. O valor limite é calculado pelo teorema de Bayes [38].

Figura 19: Importância das variáveis na projeção dos escores.



A partir da **Figura 19** pode-se fazer as seguintes observações: a amostra **13** (molécula **4b**), ativa, foi classificada como inativa, e a amostra **14** (molécula **1h**), inativa, foi classificada como ativa. Essas mesmas amostras também se agruparam incorretamente nas análises HCA e KNN. Para o conjunto teste, todas as amostras foram classificadas corretamente.

Pode-se fazer duas observações para a molécula **1h**, a primeira: sua estrutura possui oxigênio ligado diretamente ao anel no ponto B, não havendo nenhuma outra modificação no ponto A e nem no espaçador de imida (**Figura 1**), o que também pode ser observado nas moléculas **1m**, **1n**, **1o** e **1p**, que são consideradas ativas, na **Figura 20** são detalhadas as estruturas **1h**, **1m** e **1n** para comparação. A segunda: também na ramificação do ponto B, na molécula **1h** aparece átomos de flúor, característica observada nas moléculas **1f** e **1g**, ativas (**Figura 21**).

Figura 20: Molécula 1h (inativa), 1m e 1n (ativas). Átomos de oxigênio representados em vermelho.

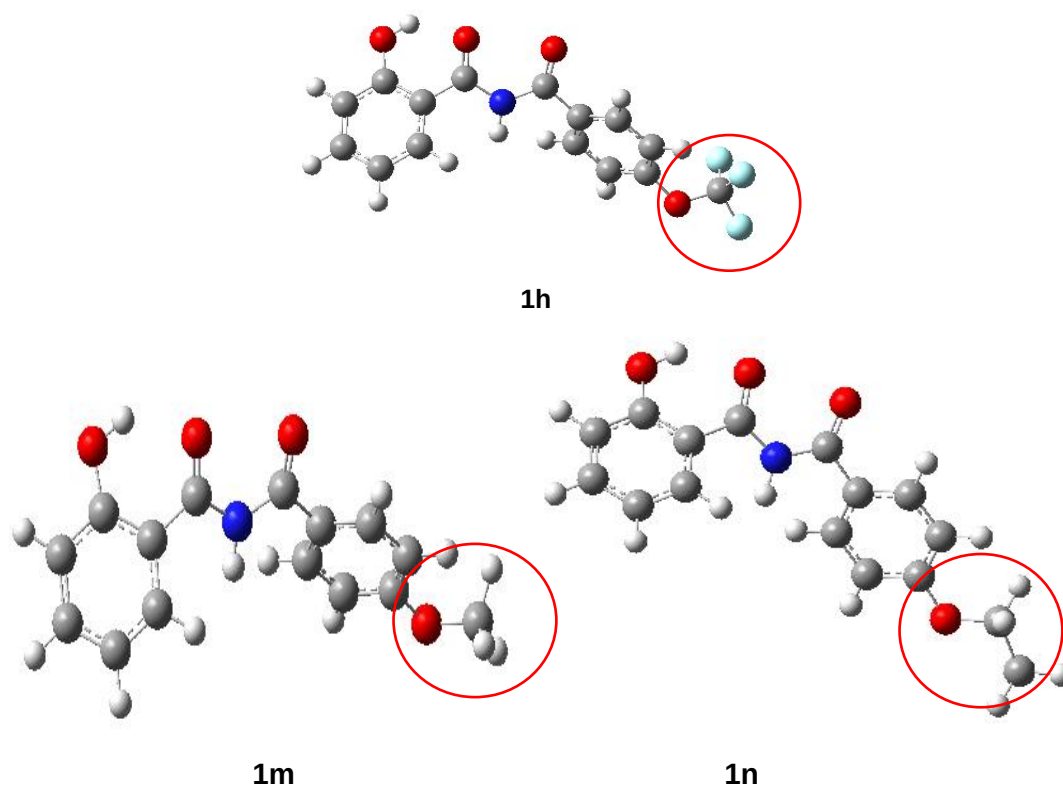


Figura 21: Moléculas 1h (inativa), 1f e 1g (ativas). Átomos de flúor representados em azul ciano.

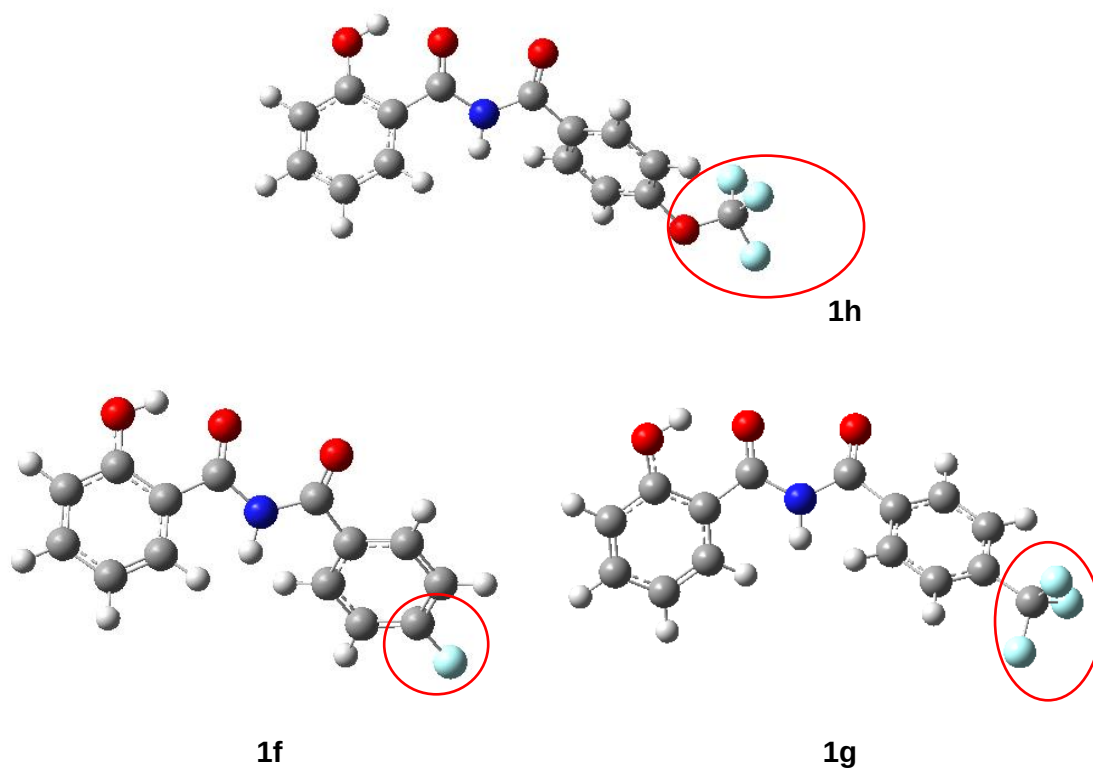


Tabela 12: Performance do PLS-DA calculada através do teorema de Bayes.

Parâmetros	Classe 1	Classe 2
VL	1	1
Sensibilidade (Treinamento)	0,85	0,88
Especificidade (Treinamento)	0,88	0,85
Sensibilidade (Teste)	1,0	1,0
Especificidade (Teste)	1,0	1,0
Limite	-0,08	0,08

O modelo mostra um nível excelente de sensibilidade (capacidade do modelo em classificar corretamente as amostras da classe de interesse) e especificidade (habilidade do modelo em identificar corretamente as amostras que não são de interesse).

Todos as análises dos modelos de classificação, KNN, SIMCA e PLS-DA, assim como as análises exploratórias de HCA e PCA, serviram para revelar a estrutura dos dados, em termos da similaridade das amostras no espaço definido por tais descritores e do poder discriminante dos descritores calculados. Assim, foi possível observar que os descritores selecionados por Peso de Fisher possuem o melhor poder de discriminação para classificar as amostras em ativas e inativas, em comparação com os selecionados por GA (**Tabela 3**). Posteriormente, estes mesmos descritores foram testados para a construção de modelos da relação quantitativa entre a estrutura e a atividade biológica. Os resultados dos estudos QSAR são apresentados nas próximas seções.

5.3 Relação Quantitativa entre Estrutura e Atividade Biológica (QSAR)

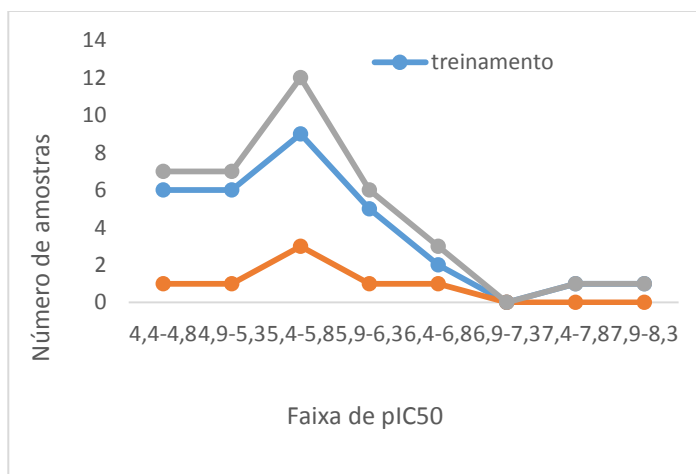
Os estudos QSAR também foram realizados com os dois conjuntos de dados: na primeira análise o conjunto de dados com as 37 amostras e na segunda análise, 31 amostras, retirando as amostras com baixo índice de seletividade.

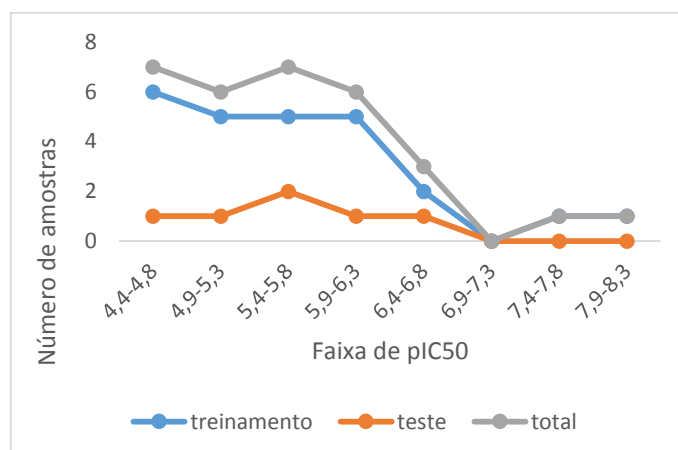
Os dados com 37 amostras foram divididos em conjunto de treinamento (30 compostos – 80% do conjunto total) e conjunto de teste (7 compostos - 20% do conjunto total). O conjunto de treinamento é formado pelos compostos **1a, 1b, 1c, 1d, 1e, 1f, 1g, 1h, 1j, 1k, 1l, 1q, 1r, 1s, 1t, 1u, 1v, 1w, 1y, 1z, 1aa, 4a, 4b, 10, 12, 13, 19, 21a, 23 e 28** e o conjunto de teste é formado pelos compostos **1i, 1m, 1n, 1o, 1p, 1x e 16**.

Os dados com 31 amostras foram divididos em conjunto de treinamento (25 compostos – 80% do conjunto total) e conjunto de teste (6 compostos - 20% do conjunto total). O conjunto de treinamento é formado pelos compostos **1a, 1b, 1d, 1e, 1f, 1h, 1j, 1k, 1l, 1m, 1q, 1r, 1s, 1t, 1u, 1w, 1y, 1z, 4a, 4b, 10, 12, 21a, 23, 28** e o conjunto teste é formado pelos compostos **1c, 1i, 1n, 1o, 1p e 16**.

Os conjuntos de treinamento e teste não foram separados pelo algoritmo KS, pois o mesmo selecionou apenas as amostras da mesma faixa de pIC_{50} para predição, em torno de 5,4 a 5,8, que possui o maior número de amostras com esses valores. A distribuição representativa do conjunto de treinamento em relação ao conjunto de teste é um fator importante na avaliação da consistência interna e externa dos modelos QSAR. Por essa razão, as amostras foram separadas de acordo com o número de moléculas para as faixas de pIC_{50} . Observa-se na **Figura 22** a distribuição para o conjunto com 37 amostras e com 31 amostras, respectivamente. Nota-se que a curva para cada conjunto (total, treinamento e teste) obedece a um mesmo perfil.

Figura 22: Representação gráfica da distribuição dos valores de pIC_{50} nos conjuntos teste, treinamento e conjunto total para o conjunto de 37 e 31 amostras, respectivamente.





A **Tabela 13** apresenta os valores de coeficiente de determinação (r^2) para calibração e predição de cada técnica quimiométrica utilizada e todas as técnicas de seleção de variáveis testadas, que serviram como parâmetro para a escolha do melhor modelo para calibração multivariada.

Tabela 13: Valores de coeficiente de determinação (r^2) para calibração e predição dos diferentes métodos utilizados.

Métodos quimiométricos	Conjunto contendo 37 amostras			
	Descritores selecionados na classificação			
	PESO DE FISHER		GA	
	R ² (calibração)	R ² (predição)	R ² (calibração)	R ² (predição)
PLS	0,66	0,71	0,79	0,69
Técnicas de seleção de variáveis para calibração				
Métodos quimiométricos	OPS		GA	
	R ² (calibração)	R ² (predição)	R ² (calibração)	R ² (predição)
PLS	0,77	0,80	0,68	0,76
MLR	0,81	0,45	0,83	0,23
Métodos quimiométricos	Conjunto contendo 31 amostras			
	Descritores selecionados na classificação			
	PESO DE FISHER		GA	
	R ² (calibração)	R ² (predição)	R ² (calibração)	R ² (predição)
PLS	0,55	0,29	0,74	< 0,0
Técnicas de seleção de variáveis para calibração				
Métodos quimiométricos	OPS		GA	
	R ² (calibração)	R ² (predição)	R ² (calibração)	R ² (predição)
PLS	0,81	0,89	0,31	0,12
MLR	0,91	0,42	0,94	0,42

Analisando a **Tabela 13**, verifica-se que os melhores resultados para calibração multivariada estão no modelo gerado pela seleção de variáveis OPS, seguido do método PLS para o conjunto com 31 amostras, quando foram excluídas do conjunto de dados as amostras com baixo índice de seletividade. Os descritores utilizados no estudo apresentado na **seção 5.2**, selecionados pela técnica de Peso de Fisher, não forneceram resultados bons para a construção de modelos quantitativos da relação estrutura-atividade para os compostos em estudo.

Verifica-se também que o OPS combinado com o PLS, para o conjunto com 37 amostras, gerou um modelo com bons coeficientes de determinação, mas na análise de resíduos (pIC_{50} experimental – pIC_{50} predito) de predição (validação externa), três amostras apresentaram resíduos relativamente altos com valores acima de 0,7. Nas seções seguintes serão apresentados os resultados para o modelo OPS-PLS para o conjunto de 31 amostras.

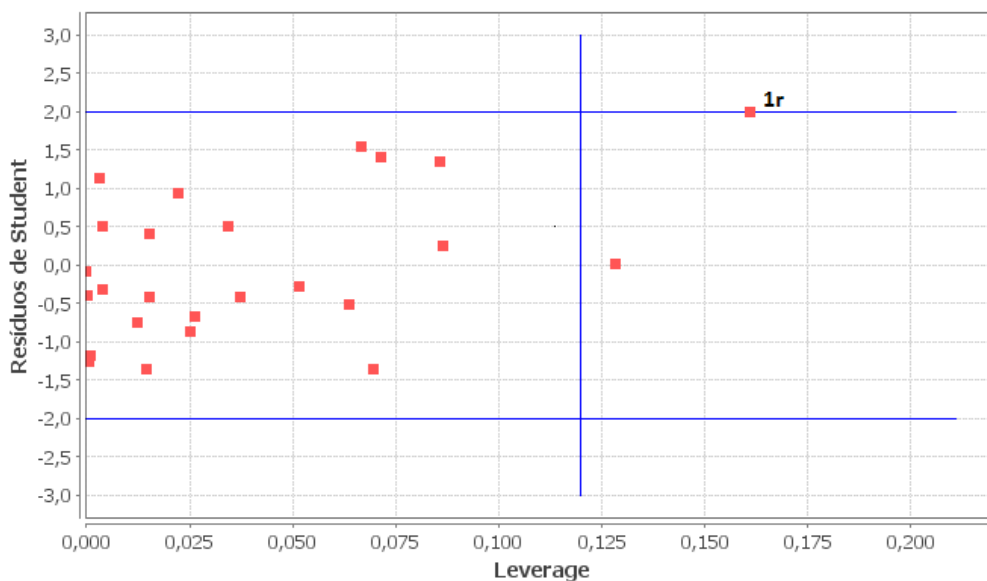
5.3.1 Resultados do modelo OPS-PLS para o conjunto com 31 amostras

5.3.1.1 Detecção de outliers

Em QSAR, *outliers* são compostos que apresentam atividade/propriedade inesperada ou não se ajustam ao modelo. Este tipo de amostra pode resultar de erros experimentais ou podem apontar falhas no modelo gerado. Por esse motivo, não podem ser removidos sem explicação, visto que sua remoção para melhorar os parâmetros estatísticos do modelo é considerada manipulação e essa melhoria pode não representar a real melhoria na habilidade de predição externa do modelo [25].

Para os modelos com 1 variável latente (este foi o número de VLS apontado como ótimo), pode-se notar que a amostra **1r** está próxima do limite superior de Resíduos de Student (2,0) e com alta influência (**Figura 23**). Esta amostra, no grupo de dados, é a que possui maior atividade biológica, o que explica sua alta influência, sendo assim, não foi considerada *outlier*.

Figura 23: Detecção de Outliers.



5.3.1.2 Seleção de variáveis e construção do modelo PLS

Os 304 descritores restantes, após terem sido eliminando os descritores que possuíam correlação com a atividade biológica menor do que 0,3, foram submetidos a uma seleção inicial utilizando o procedimento OPS, implementado no programa QSAR modeling ^[25]. Os seguintes parâmetros foram escolhidos na execução do OPS (**Figura 24**) o qual foi aplicado ao conjunto treinamento.

Figura 24: Parâmetros utilizados para o OPS.

OPS avançado

Número de variáveis latentes máximo para o OPS:

Número de variáveis latentes para o modelo:

Número de amostras a retirar na validação cruzada:

Janela: Incremento: Porcentagem de variáveis:

Número de modelos a guardar em cada passo do OPS:

Vetores

- ☐ Correlograma
- ☐ Correlograma e vetor de regressão
- ☐ Correlograma e produto
- ☒ Todos os vetores

Critério para classificar modelos

- ☐ RMSECV
- ☐ Rcv
- ☒ Q2
- ☐ Spres

O programa *QSAR modeling* gera modelos PLS construídos com os descritores selecionados e mediante validação interna *leave-one-out*. Este procedimento resultou na seleção dos 12 descritores apresentados na **Tabela 14**.

Tabela 14: Descritores selecionados pelo OPS para o melhor modelo.

DESCRITOR	TIPO	DEFINIÇÃO
Mor11p	3D-MoRSE descriptors	3D-MoRSE - signal 11 / weighted by atomic polarizabilities
Mor11v		3D-MoRSE - signal 11 / weighted by atomic van der Waals volumes
Mor30p		3D-MoRSE - signal 30 / weighted by atomic polarizabilities
RDF105u	RDF descriptors	Radial Distribution Function - 11.5 / unweighted
RDF115u		Radial Distribution Function - 11.5 / weighted by atomic Sanderson electronegativities
RDF105e		Radial Distribution Function - 10.5 / weighted by atomic Sanderson electronegativities
Polarizabilidade	Químico-quântico	Deformação molecular decorrente do movimento vibratório, resultando em uma variação dos momentos dipolares atômicos.
GATS4e	2D autocorrelations	Geary autocorrelation - lag 4 / weighted by atomic Sanderson electronegativities
BIC5	Information índices	bond information content (neighborhood symmetry of 5-order)
G2m	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic masses
E1u		1st component accessibility directional WHIM index / unweighted
EEig02r	Edge adjacency indices	Eigenvalue 02 from edge adj. matrix weighted by resonance integrals

Como mencionado anteriormente, o melhor modelo foi selecionado com base nos menores valores de PRESS e SEV. Os coeficientes de determinação obtidos foram $q^2 = 0,75$ e $r^2 = 0,81$ com o número ideal de variáveis latentes igual a 1. Os parâmetros estatísticos da regressão PLS para o melhor modelo são apresentados na **Tabela 15**.

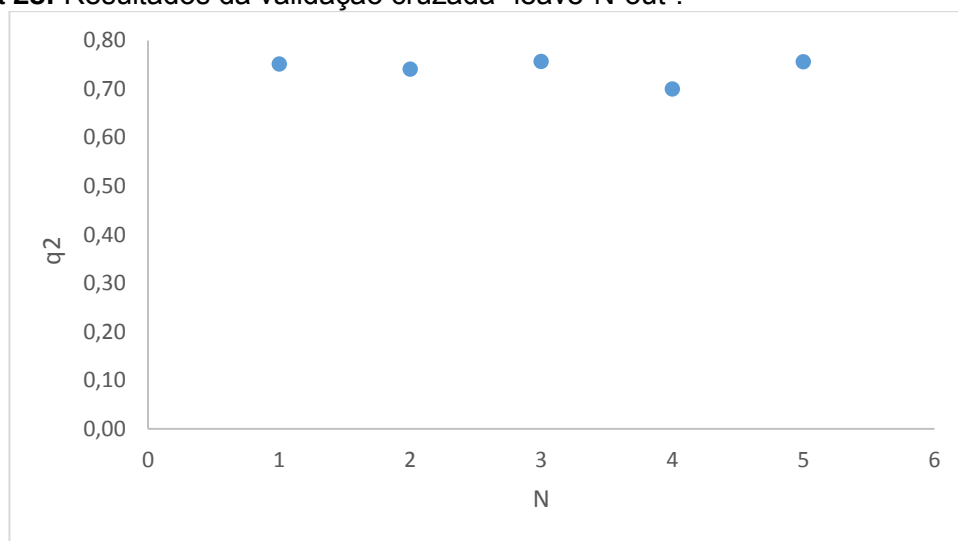
Tabela 15: Parâmetros estatísticos obtidos com a regressão PLS sobre os 12 descritores selecionados.

<i>Variáveis Latentes</i>	<i>SEV</i>	<i>PRESS</i>	<i>q²</i>	<i>r²</i>
1	0,48	5,71	0,75	0,81
2	0,58	8,31	0,65	0,82
3	0,66	10,76	0,57	0,82
4	0,70	12,25	0,52	0,83
5	0,73	13,46	0,49	0,83
6	0,74	13,66	0,48	0,83
7	0,78	15,06	0,46	0,84
8	0,81	16,41	0,41	0,85
9	0,81	16,47	0,42	0,87
10	0,77	14,94	0,45	0,87

5.3.1.3 Testes de leave-N-out, y-randomization e validação externa

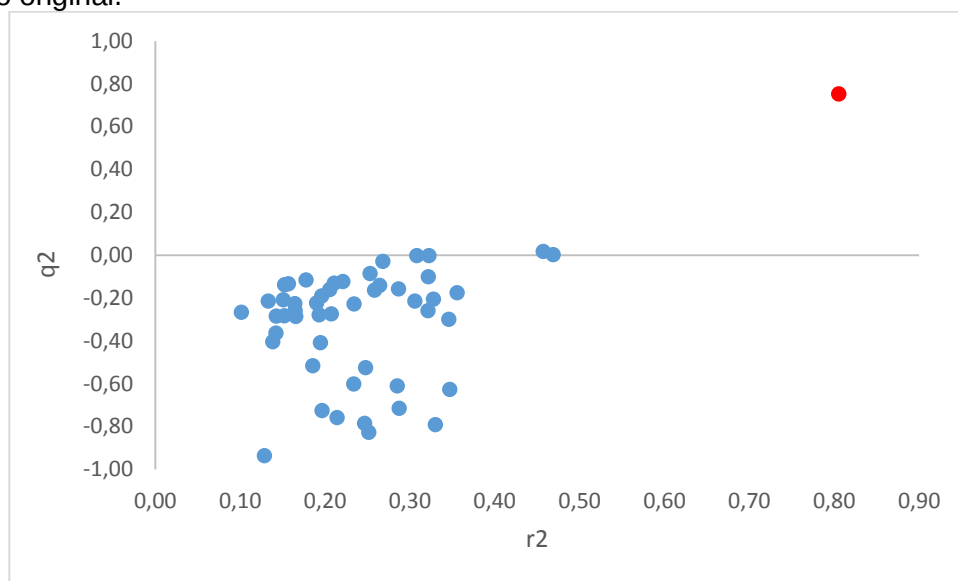
Uma vez escolhido o modelo, seguiu-se para a validação cruzada leave-N-out, realizada para avaliar a estabilidade do modelo PLS. No teste de leave-N-out, para cada N, os dados foram aleatorizados 20 vezes, ou seja, foram retirados 20 diferentes conjuntos de compostos, e os valores médios de q^2 foram tomados para a comparação com o q^2 obtido pela validação cruzada leave-one-out (LOO). Nota-se na **Figura 25** que um bom modelo de QSAR foi obtido pois o valor médio de q^2 para todos os valores de N está próximo ao q^2 da validação cruzada LOO do modelo original, com o desvio padrão para cada N não excedendo 0,1. Recomenda-se fixar N entre 20% - 30% do total ^[30], assim, para este modelo N foi igual a 5.

Figura 25: Resultados da validação cruzada “leave-N-out”.



Para o teste de validação *Y-randomization*, os valores da atividade biológica foram embaralhados aleatoriamente 50 vezes e os valores de r^2 e q^2 foram calculados. Como pode ser analisado na **Figura 26**, e conforme apresentado na fundamentação teórica (**seção 3.4**), valores de q^2 para modelos obtidos com a atividade biológica aleatorizada apresentam valores negativos, indicando que valores extremamente altos de resíduos de calibração e validação foram obtidos (**ver equações 13 e 14**). Embora alguns destes modelos tenham apresentado valores um pouco mais altos de r^2 , estes são acompanhados de valores negativos de q^2 , indicando que somente o modelo com os valores corretos de atividade biológica fornece parâmetros estatísticos apropriados (representado pelo ponto vermelho na **Figura 26**), o que não ocorre com números aleatórios. Assim, é possível afirmar que uma correlação verdadeira dos descritores com a variável de resposta existe no conjunto de dados.

Figura 26: Resultados do teste de Y-randomization. Os pontos azuis representam os modelos gerados com os valores de y aleatorizados e o ponto vermelho representa o modelo original.



Para a validação externa, na **Tabela 16** observa-se os valores da atividade biológica experimental, da atividade biológica prevista e resíduos, para cada composto. É possível verificar que os resíduos estão próximos de zero, o que indica um bom grau de ajuste para a equação e um bom poder preditivo do modelo gerado. O valor calculado para o parâmetro r^2_{pred} , calculado com a **equação 15**, foi igual a 0,89, indicando que o modelo obtido possui boa capacidade preditiva.

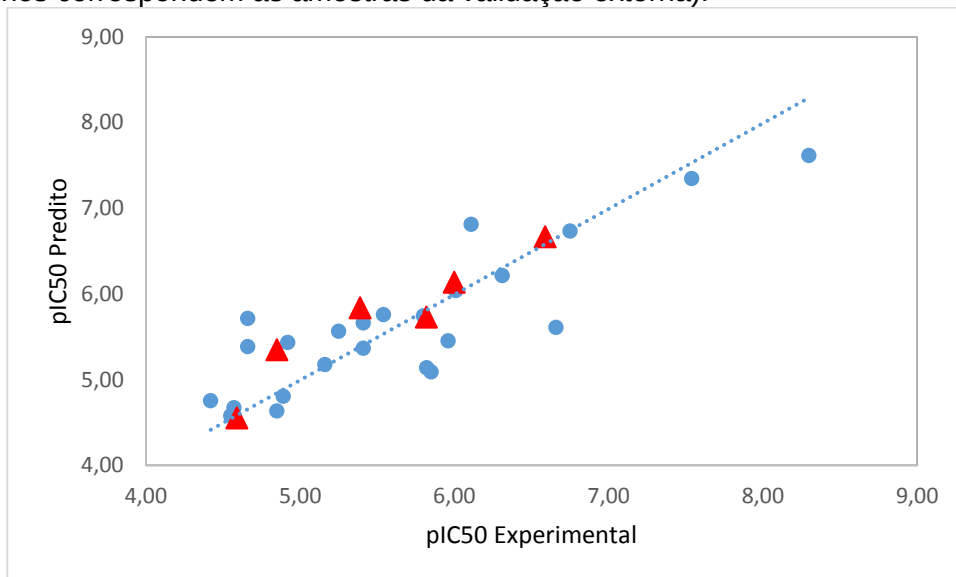
Tabela 16: Valores experimentais e preditos de pIC_{50} para o conjunto de teste.

Amostras	pIC_{50} Experimental	pIC_{50} Predito	Resíduo
1c	5,82	5,73	0,09
1i	4,85	5,35	-0,50
1n	6,00	6,14	-0,14
1º	6,59	6,67	-0,08
1p	5,39	5,84	-0,45
16	4,59	4,55	0,04

Os índices de confiabilidade são bons e há uma boa correlação entre os valores da atividade biológica calculada e experimental. É importante observar o alinhamento dos pontos em relação à reta ajustada, bem como a distribuição deles ao longo do intervalo de atividade estudado. Na **Figura 27** podemos

observar pIC_{50} Previsto vs pIC_{50} Experimental tanto para o conjunto teste (triângulos vermelhos) como para o conjunto treinamento (círculos azuis).

Figura 27: Valores pIC_{50} Preditos versus pIC_{50} Experimentais de pIC_{50} (triângulos vermelhos correspondem as amostras da validação externa).



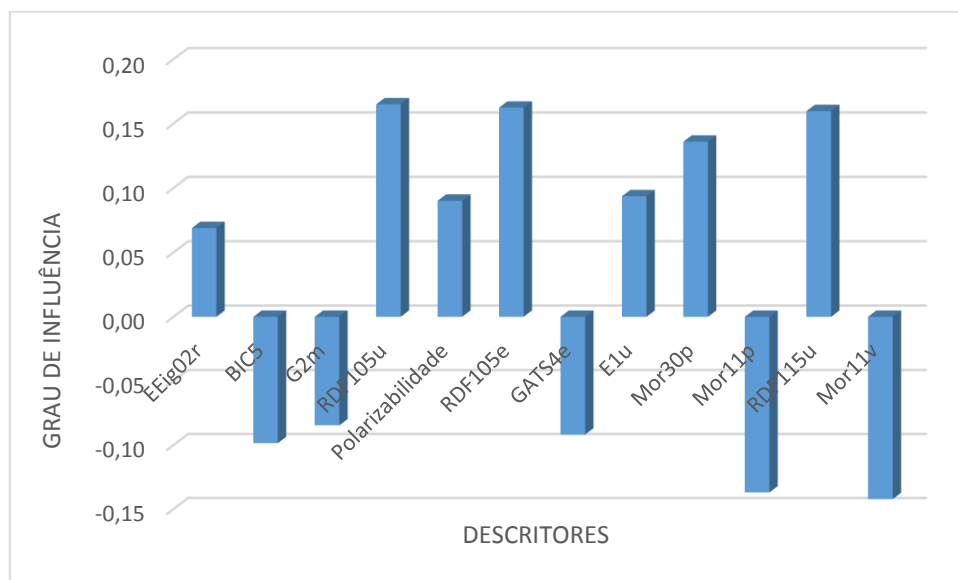
5.3.1.4 Implicações para o planejamento de novas moléculas

Uma vez que um modelo estável, robusto e com boa capacidade preditiva foi obtido, o vetor de regressão gerado pode ser usado para analisar os descritores moleculares empregados na construção do modelo e, assim, fornecer subsídios para a otimização da atividade biológica de moléculas análogas às estudadas neste trabalho.

As contribuições de cada descritor para o vetor de regressão (**equação 17**) podem ser observadas na **Figura 28**, onde verifica-se que as contribuições de maior peso no vetor de regressão são provenientes dos descritores do tipo RDF, baseados em uma função de distribuição radial, que pode ser interpretada como a distribuição da probabilidade de encontrar um átomo em um volume esférico de raio R [54].

$$pIC_{50} = 0,07(EEig02r) - 0,10(BIC5) - 0,08(G2m) + 0,17(RDF105u) + 0,09(Polariz.) + 0,16(RDF105e) - 0,09(GASTS4e) + 0,09(E1u) + 0,14(Mor30p) - (0,14Mor11p) + 0,16(RDF115u) - 0,14(Mor11v) \quad (17)$$

Figura 28: Contribuição de cada descritor para o vetor de regressão.



Os três descritores com contribuições mais altas são RDF105u, RDF105e e RDF115u. Estes três descritores apresentam contribuições positivas para o vetor de regressão, indicando que valores positivos altos dos mesmos são desejáveis para otimizar a atividade biológica das moléculas em estudo. Para o descritor RDF105e, quanto maior a eletronegatividade de átomos que estejam a uma distância aproximada de 10,5 Å, maior o valor deste descritor. Os descritores RDF105u (átomos que estejam a uma distância aproximadamente de 10,5 Å) e RDF115u (átomos que estejam a uma distância aproximadamente de 11,5 Å) são descritores estritamente de caráter estérico, para os quais não é utilizada nenhuma propriedade de ponderação (*“unweighted”*), assim, átomos volumosos, a estas distâncias, contribuem para o aumento no valor destes descritores e como resultado, o aumento na atividade biológica.

Além destes, outros descritores com contribuições relativamente altas para o vetor de regressão são os descritores 3D MoRSE, que são definidos predominantemente por pares de átomos de curta distância [55]. Para o descritor Mor30p, pode-se verificar que os pares de átomos que estão menores ou

próximos a 3Å são ponderados pela polarizabilidade e contribuem positivamente para a atividade biológica. Para os descritores Mor11p e Mor11v, pares de átomos que estão menores ou próximos a 1,1 Å, ponderados por polarizabilidade e volume de van der waals, contribuem negativamente para o aumento da atividade biológica.

Três outros descritores que contribuem positivamente para a atividade biológica são: (i) EEig02r, calculado a partir da teoria de grafos moleculares, que codificam informações sobre a conectividade dos átomos; (ii) polarizabilidade molecular, um descritor químico quântico que indica a suscetibilidade da molécula a um campo elétrico externo; e (iii) E1u, que pertence à classe de descritores WHIM (*Weighted Holistic Invariant Molecular*) que contém informação tridimensional ^[54].

Contribuindo negativamente para a atividade biológica aparecem ainda os descritores BIC5, que contém informação sobre ligações, G2m, que também é um descritor do tipo WHIM e GATS4e, um descritor do tipo autocorrelação 2D, que descreve como uma determinada propriedade é distribuída através da estrutura topológica molecular, neste caso, a eletronegatividade ^[54].

6 CONCLUSÕES

As diferentes técnicas utilizadas neste trabalho permitiram explorar alguns aspectos estruturais das moléculas estudadas, bem como relacionar alguns descritores à ação antimalárica. Com base nos resultados apresentados, podemos elencar algumas conclusões sobre a análise exploratória dos dados e os métodos de classificação para os melhores modelos.

Com as duas técnicas, HCA e PCA, apresentadas para o conjunto com 31 amostras e seleção de variáveis por Peso de Fisher, obteve-se resultados satisfatórios. Com o HCA foi possível verificar agrupamentos característicos, concluindo que há diferenças marcantes entre as amostras e que os descritores adotados na classificação de atividades foram capazes de avaliar a similaridade entre os compostos por classes. A PCA auxiliou na observação da separação das amostras e formação dos agrupamentos, sugerindo que existem diferenças físico-químicas e estruturais que permitem separar os compostos ativos dos inativos no espaço químico definido pelos descritores selecionados.

Com relação às técnicas de classificação, observou-se que os melhores modelos foram gerados por SIMCA e PLS-DA com os descritores selecionados por Peso de Fisher, demonstrando confiabilidade, boa consistência interna e capacidade de classificar as amostras em ativas ou inativas. No entanto, os descritores selecionados por esta abordagem não forneceram modelos QSAR adequados.

Para os resultados de calibração multivariada, a técnica OPS em conjunto com o PLS para 31 amostras, conduziu a uma proposta de modelo QSAR que se mostrou preditivo para a atividade antimalárica da classe de compostos estudadas, apresentando bons valores para os principais parâmetros estatísticos, a exemplo dos coeficientes de determinação, e um modelo estável e robusto, conforme indicado pelos testes de validação leave-N-out e Y-randomization.

Para o melhor modelo obtido, OPS-PLS (31 amostras), os coeficientes de correlação foram $q^2 = 0,75$ e $r^2 = 0,81$ com um coeficiente de determinação para a predição $r^2_{\text{pred}} = 0,89$ para o conjunto de teste. Esses resultados obtidos com o melhor modelo indicaram que a atividade biológica das moléculas em estudo

depende da combinação de múltiplos fatores, representados pelas propriedades estéricas bidimensionais (descritores EEig02r, BIC5 e GATS4e) e tridimensionais (descritores dos tipos RDF, MoRSE e WHIM), assim como por propriedades eletrônicas (polarizabilidade molecular).

Naturalmente, não é possível apenas olhar para o esqueleto principal dos compostos em estudo e propor modificações estruturais que otimizem a atividade biológica com base nos descritores selecionados, que são de difícil interpretação. Porém, os resultados aqui apresentados possuem considerável relevância, pois os testes de validação aplicados demonstraram que um modelo QSAR confiável e preditivo foi obtido. Este modelo pode ser utilizado por pesquisadores interessados na síntese e avaliação de novos compostos com características estruturais semelhantes às estudadas com vistas ao planejamento de moléculas potencialmente mais ativas contra o *P. falciparum*.

7 PERSPECTIVAS

O presente trabalho abre outras possibilidades de investigações a serem realizadas. Assim, como perspectivas deste trabalho, podemos elencar:

- A proposição e a triagem de outras estruturas congêneres de N-benzoil-2-hidroxibenzamidas, o cálculo dos descritores evidenciados neste trabalho e o uso destes nos modelos propostos para previsão de sua atividade antimalárica;
- A utilização de outros descritores, como 3D e/ou 4D, para obtenção de modelos com melhores parâmetros estatísticos;
- Utilizar os outros valores de atividade biológica, obtidos nas pesquisas de Stec e colaboradores^[5] para três parasitas protozoários e criar novos modelos preditivos da relação quantitativa entre estrutura e atividade biológica (QSAR) de derivados de N-benzoil-2-hidroxibenzamidas.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Coordenação de Informações Estratégicas em Vigilância em Saúde – CIEVS. Disponível em: <https://cievsrio.wordpress.com/malaria/>. **Malária**. Acessado em 13 de outubro de 2017.
- [2] Gurge, R. S. *A malária e a Resistência aos fármacos no seu tratamento*. Publicado em: <https://pt.linkedin.com/pulse/mal%C3%A1ria-e-resist%C3%A2ncia-aos-f%C3%A1rmacos-seu-tratamento-sim%C3%A3o-gurge>. Acessado em: 24 de novembro de 2017.
- [3] Corey, V. C., Lukens, A. K., Istvan, E. S., Lee, M. C. S., Franco, V., Magistrado, P., Coburn-Flynn, O., Sakata-Kato, T., Fuchs, O., Gnädig, N. F., Goldgof, G., Linares, M., Gomez-Lorenzo, M. G., Cristina De Cozar, Lafuente-Monasterio, M. Prats, S., Meister, S., Tanaseichuk, O., MelanieWree, Yingyao Zhou, Paul A. Willis, Gamo, F. G., Goldberg, D. E., Fidock, D. A., Wirth, D. F., Winzeler, E. A. *A broad analysis of resistance development in the malaria parasite*. **Nature Communications** | DOI: 10.1038/ncomms11901. Published 15 Jun 2016.
- [4] Fomovska, A.; Huang, Q.; El Bissati, K.; Mui, E. J.; Witola, W. H.; Cheng, G.; Zhou, Y.; Sommerville, C.; Roberts, C. W.; Bettis, S.; Prigge, S. T.; Afanador, G. A.; Hickman, M. R.; Lee, P. J.; Leed, S. E.; Auschwitz, J. M.; Marco Pieroni, M.; Stec, J.; Muench, S. P.; Rice, D. W.; Kozikowski, A. P.; McLeod, R. *Novel N-Benzoyl-2-Hydroxybenzamide Disrupts Unique Parasite Secretory Pathway*. **Antimicrobial Agents and Chemotherapy** p. 2666 –2682. May 2012 Volume 56 Number 5.
- [5] Stec, J.; Huang, Q.; Pieroni, M.; Kaiser, M.; Fomovska, A.; Mui, E.; Witola, W. H.; Bettis, S.; McLeod, R.; Brun, R.; Kozikowski, A. P. K. *Synthesis, Biological Evaluation, and Structure–Activity Relationships of N-Benzoyl-2-hydroxybenzamides as Agents Active against P. falciparum (K1 strain), Trypanosomes and Leishmania*. **J. Med. Chem.** 2012, 55, 3088–3100.
- [6] Bajorath, J. J. **Comput. Aided Mol. Des.** 26, 11-12, 2012.
- [7] Cramer, R. J **Comput Aided Mol Des.** 2012 Jan; 26(1): 35–38.
- [8] Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. *QSAR Modeling: Where Have You Been? Where Are You Going To?* **J. Med. Chem.** 2014, 57, 4977–5010.
- [9] Cator, L. J., Lynch, P. A., Andrew F. Read, A. F., Thomas, M. B. *Do malaria parasites manipulate mosquitoes?* Published in final edited form as: **Trends Parasitol.** 2012; 28(11): 466–470.

- [10] Tavares, L.C. QSAR: a abordagem de Hansch. **Química Nova**, Vol. 27, Nº. 4, 631-639, 2004.
- [11] Sliwoski, G., Kothiwale, S., Meiler, J., Lowe, E. W. Jr. *Computational Methods in Drug Discovery*. **Pharmacological Reviews** 66:334–395, January 2014.
- [12] Young, D.C. *COMPUTATIONAL DRUG DESIGN: A Guide for Computational and Medicinal Chemists*. **John Wiley & Sons**, Canada, 2009, 321 p.
- [13] Weber, K.C., *Modelagem molecular de compostos arilpiperazínicos e suas interações com o receptor 5-HT_{1A}*. 2008, 100f. Tese (Doutorado em Físico-Química) – Instituto de Química de São Carlos, Universidade de São Paulo, São Paulo.
- [14] Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - design and description, **Journal of Computer-Aided Molecular Design** 19, 453–463, 2005.
- [15] Arroio, A.; Honório, K.M.; Silva, A.B.F. *Propriedades químico-quânticas empregadas em estudos das relações estrutura-atividade*. **Química Nova**, Vol. 33, Nº. 3, 694-699, 2010.
- [16] Ferreira, M.M.C. *Multivariate QSAR*. **J. Braz. Chem. Soc.**, Vol. 13, Nº. 6, 2002.
- [17] Ferreira, M.M.C.; Montanari, C.A.; Gaudio, A.C. *Seleção de variáveis em QSAR*. **Química Nova**, Vol. 25, Nº. 3, 439-448, 2002.
- [18] Duarte, C.S. *Estudos de QSAR-2D em quinolina metanol com atividade antimalárica frente ao Plasmodium falciparum*. Dissertação (Mestrado) Universidade do Estado do Amazonas, 2012. 91f.
- [19] Kubunyi, H. *QSAR: Hansch Analysis and Related Approaches. Methods and Principles in Medicinal Chemistry*. **VCH Publishers**, New York, NY (USA), 1993, 241 p.
- [20] Cohen, A. J.; MORI-SÁNCHEZ, P.; YANG, W. *Challenges for density functional theory*. **Chemical reviews**, v. 112, n. 1, p. 289–320, 2012. Disponível em: <http://aip.scitation.org/doi/10.1063/1.4754130>.
- [21] Costa Filho, P. A.; Poppi, R. J. *Aplicação de algoritmos genéticos na seleção de variáveis em espectroscopia no infravermelho médio. Determinação simultânea de glicose, maltose e frutose*. **Química Nova**, Vol. 25, No. 1, 46-52, 2002.
- [22] Castro, L.S. *Estudo teórico da relação Estrutura–atividade de derivados análogos da cafeína contra o Câncer epitelial*. Dissertação apresentada ao

programa de pós-Graduação em gestão, pesquisa e desenvolvimento em tecnologia farmacêutica 84f, 2009.

[23] Santos, L.R. *Estudo químico quântico e quimiométrico dos derivados análogos da ciano pirazolina contra o diabetes tipo II*. Trabalho de conclusão do curso de licenciatura em química apresentado à coordenação de licenciatura em química do Instituto Federal de Educação, Ciência e Tecnologia de Goiás. Anápolis, dezembro, 2013.

[24] Teófilo, R. F.; Martins, J. P.; Ferreira, M. M. C. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. Published online in **Wiley InterScience**: 29 October 2008.

[25] Martins, J.P.A.; Ferreira, M.M.C. QSAR modeling: um novo pacote computacional open source para gerar e validar modelos QSAR. **Quim. Nova**, Vol. 36, No. 4, 554-560, 2013.

[26] Alencar, E.B.F. *Estudos da relação quantitativa estrutura-atividade (QSAR) de adutos de morita-baylis-hillman bioativos contra leishmania amazonenses*. Tese de doutorado. Universidade Federal da Paraíba, área de concentração Química Orgânica. Dezembro /2012, 155f.

[27] Mitchell, M. *An Introduction to Genetic Algorithms*. **MIT Press**, Massachusetts, London England, 1996.

[28] Teófilo, R. F. *Chemometric methods in the electrochemical studies of phenols on boron-doped diamond films*. Universidade Estadual de Campinas, Campinas, 2007.

[29] Beebe, K.R.; Pell, R. J. e Seasholtz; M. B. *Chemometrics: A practical Guide*; **Wiley**: New York, 1998.

[30] Ferreira, M. M. C. *Quimiometria – conceitos, métodos e aplicações*. **Editora da Unicamp**: São Paulo, 2015.

[31] Santos, B.C.D.B. *Desenvolvimento, aperfeiçoamento e validação de método cromatográfico para previsão da qualidade e propriedades físico-químicas do óleo diesel tipo B*. Dissertação de Mestrado. Universidade Estadual Paulista, área de concentração Química Orgânica. 2009. 188 f.

[32] Sanches, F.A.C. *Desenvolvimento de uma metodologia para o controle de qualidade de medicamento injetável sem violação da ampola, usando espectroscopia NIR e Técnicas Quimiométricas*. Dissertação de Mestrado. Química – Química Analítica, João Pessoa –PB, 2009, 88f.

[33] Gavaghana, C. L., Wilsonb, I. D., Nicholsona, J. K. *Physiological variation in metabolic phenotyping and functional genomic studies: use of orthogonal signal correction and PLS-DA*. **FEBS Letters** 530 (2002) 191-196.

- [34] Sabin, J.G.; Ferrão, M. F., Furtado, J. C. *Análise multivariada aplicada na identificação de fármacos antidepressivos. Parte II: Análise por componentes principais (PCA) e o método de classificação SIMCA*. **Brazilian Journal of Pharmaceutical Sciences** vol. 40, n. 3, jul./set., 2004.
- [35] Costa, G.B. *Classificação multivariada de biodiesel b100 e b5 usando imagens digitais*. Dissertação. Área de concentração: energias renováveis e biocombustíveis. Campina Grande- PB. Fevereiro de 2015. 80f.
- [36] Ballabio D, Consonni V. *Classification tools in chemistry. Part 1: Linear models. PLS-DA*. **Analytical Methods**, 5, 3790-3798, 2013.
- [37] Barker, M. & Rayens, W. *Partial least squares for discrimination*. **Journal of Chemometrics**. 17, 166–173 (2003).
- [38] N. F. Pérez, J. Ferré and R. Boqué. *Calculation of the reliability of classification in discriminant partial least-squares binary classification*. **Chemom. Intell. Lab. Syst.**, 2009, 95, 122.
- [39] Melo, E.B. *Estudos teóricos (modelagem molecular e QSAR) de inibidores de HIV-1 integrase*. Tese de doutorado. Unicamp - Universidade Estadual de Campinas. Instituto de Química - Departamento de Físico-Química, 2009, 246F.
- [40] Golbraikh, A. and A. Tropsha (2002a), "Beware of q² !". **Journal of Molecular Graphics and Modelling**, 20, 269-276.
- [41] Roy, K.; Mandal, A. S. *Predictive QSAR modeling of CCR5 antagonist piperidine derivatives using chemometric tools*. **Journal of Enzyme Inhibition and Medicinal Chemistry** ISSN: 1475. 2008.
- [42] Kiralj, R.; Ferreira, M. M. C..J. *Basic Validation Procedures for Regression Models in QSAR and QSPR Studies: Theory and Application*. **Braz. Chem. Soc.**, Vol. 20, No. 4, 770-787, 2009.
- [43] Tropsha, A. *Best Practices for QSAR Model Development, Validation, and Exploitation*. **Mol. Inf.** 2010, 29, 476 – 488.
- [44] Eriksson, L.; Jaworska, J.; Worth, P. A.; Cronin, M. T. D.; Macdowell, R. M.; Gramatica, P. *Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification – and Regression – Based QSARs*. **Environmental Health Perspectives**, 111, 1361-1375, 2003.
- [45] Cramer, B. *Estudos de QSAR-2D aplicados a diterpenóides clerodanos e dibenzoilidrazinas*. 316 f. Tese de Doutorado – Departamento de Físico-Química, Instituto de Química, Universidade Estadual de Campinas, Campinas, São Paulo, 2011.

- [46] Dewar, M. J. S.; Zoebisch, E.G.; Healy, E.F.; Stewart, J.J.P. **J. Am. Chem. Soc.** 107, 3902-3909, 1985.
- [47] Gaussian 09, Revision D.01, 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, J. A., Jr.; 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, J. M.; 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, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian, Inc., Wallingford CT, 2009. Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V., **J. Comput. Aid. Mol. Des.** 19, 453-463, 2005.
- [48] Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - design and description, **Journal of Computer-Aided Molecular Design** 19, 453–463, 2005.
- [49] Sharaf, M.A.; Illman, D.L.; Kowalski, B.R. *Chemometrics*. New York: **Wiley**, 1986. p. 193.
- [50] Kawakami, R. H. G. Seleção de amostras de treinamento, validação e teste empregando Kennard-Stone. ITA, 2004.
- [51] Kawakami, R. H. G. Seleção de Variáveis para Classificação baseada no Algoritmo das Projeções Sucessivas (SPA) empregando a built-in function qr do Matlab. ITA/IEES. 2008.
- [52] R. Leardi and A. Lupiáñez. “Genetic algorithms applied to feature selection in PLS regression: how and when to use them.” **Chemolab**, 41 (1998) 195-207.
- [53] Infometrix Inc. Pirouette 4.0. **Woodinville**, 2008.
- [54] Todeschini, R. and Lasagni, M. New molecular descriptors for 2D and 3D structures theory. **Journal of Chemometrics**, vol. 8, 263-272 (1994).
- [55] Devinyak, O., Havrylyukb, D., Lesyk, R. 3D-MoRSE descriptors explained. **Journal of Molecular Graphics and Modelling** 54 (2014) 194–203.