UMA NOVA AMOSTRAGEM DE DESCRITORES PARA PREDIÇÃO DE ATIVIDADE BIOLÓGICA

JOÃO VITOR SOARES TENÓRIO

UMA NOVA AMOSTRAGEM DE DESCRITORES PARA PREDIÇÃO DE ATIVIDADE BIOLÓGICA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: LOÏC PASCAL GILLES CERF

Belo Horizonte

Outubro de 2018

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Tenório, João Vitor Soares
T312n Uma nova amostragem de descritores para predição de atividade biológica / João Vitor Soares Tenório — Belo Horizonte, 2018.
xviii, 55 p.:il.; 29 cm.

Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Loïc Pascal Gilles Cerf

1. Computação – Teses. 2. Aprendizado do Computador. 3. Quimiometria. 4. Bioinformática. 5. QSAR (Bioquímica) I. Orientador. II. Título.

CDU 519.6*82(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Uma Nova Amostragem de Descritores para Predição de Atividade Biológica

JOÃO VITOR SOARES TENÓRIO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. LOIC PASCAL GILLES CERF - Orientador
Departamento Ciência da Computação - UFMG

PROF. JOÃO PAULO ATAIDE MARTINS
Departamento de Química - UFMG

PROFA. RAQUEL CARDOSO DE MELO MINARDI
Departamento de Ciência da Computação - UFMG

PROF. RENATO MARTINS ASSUNÇÃO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 26 de outubro de 2018.

Resumo

Os métodos de aprendizado de máquina estão sendo usados para resolver diversos problemas nas áreas de bioinformática e quimiometria. Um desses problemas é o planejamento de fármacos auxiliado por computador (CADD), que usa modelos preditivos para planejar e aprimorar compostos que possuem atividade biológica e podem ser usados como fármacos. Uma das técnicas de CADD mais usadas é o estudo das relações quantitativas estrutura-atividade (QSAR), que possibilita desenvolver um modelo preditivo que relaciona as propriedades dos compostos e suas atividades biológicas, normalmente esse modelo é uma regressão linear. O LQTA-QSAR é uma técnica 4D-QSAR, onde a amostragem dos descritores usados para o treinamento do modelo preditivo é feita inserindo os perfis de amostragem conformacional (PAC) dos compostos em uma grade 3D e calculando as interações entre o PAC e uma sonda (que pode ser um átomo, íon ou grupo funcional) nos pontos dessa grade. O problema dessa amostragem é que a sonda passa por pontos internos ao PAC, amostrando descritores com valores irreais. Para contornar esse problema, nessa dissertação foi proposta uma nova amostragem de descritores que usa ampliações da superfície definida pelo fecho convexo para construir camadas ao redor do PAC onde a sonda deve passar. Essa amostragem impede que a sonda passe por pontos internos ou próximos demais ao PAC. Para validar a proposta foram realizados diversos experimentos em conjuntos de compostos que podem ser usados como fármacos para tratamento de diversas doenças. Os resultados mostraram que a proposta conseguiu gerar modelos preditivos com maior precisão que o método original nos seis cenários avaliados. O maior aumento percentual obtido foi de 44%. Também foi proposto um fluxo de trabalho onde a regressão linear foi substituída por árvore de regressão, essa mudança possibilitou gerar modelos mais fáceis de serem interpretados. Também foram realizados os experimentos com essa modificação em seis cenários, onde em um caso a precisão foi superior aos modelos lineares e nos demais casos foi inferior, porém com precisão ainda satisfatória.

Palavras-chave: Aprendizado de Máquina, Quimiometria, QSAR.

Abstract

Machine learning methods are being used to solve different problems in the areas of bioinformatics and chemometrics. One such problem is computer-aided drug design (CADD), which uses predictive modeling to design and improve compounds that have biological activity and can be used as drugs. One of the techniques used CADD is the study of quantitative structure-activity relationships (QSAR), which allows to develop a predictive model that relates the properties of the compounds and their biological activities, this model is typically a linear regression. LQTA-QSAR is a 4D-QSAR technique, where the descriptors used for predictive model training are sampled by aligning the conformational ensemble profiles (CEP) of the compounds in a 3D grid and calculating the interaction between the CEP and a probe (it can be an atom, ion, or functional group) in each point of this grid. The problem with this sampling is that the probe crosses the CEP, when the probe falls into or close to an atom of the CEP, some descriptors presents unrealistic values. To overcome this problem, a new approach for sampling descriptors was proposed in this thesis, which uses surface expansions defined by the convex hull to construct layers around the CEP where the probe must pass. This sampling prevents the probe from passing through the points inside or too close the CEP. To validate the proposal, several experiments were carried out on sets of compounds that can be used as drugs for the treatment of several diseases. The results showed that the proposal was able to build predictive models with greater precision than the original method in the six scenarios evaluated. The highest percentage increase was 44%. We also proposed a workflow where linear regression was replaced by regression tree, which allows to build models easier to interpret. Experiments with this new workflow were also carried out in six scenarios, where in one case the precision was superior to the linear models and in the other cases it was lower, but still satisfactory.

Keywords: Machine Learning, Chemometrics, QSAR.

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Sumário

Capítulo 1

Introdução

Os métodos de aprendizado de máquina estão sendo amplamente utilizados para resolver diversos problemas relevantes no mundo real, em áreas como química, biologia, economia entre outras [Keilis-Borok et al., 2005; Dougherty, 2005]. Esses métodos são amplamente usados no planejamento de fármacos auxiliado por computador (CADD), onde os modelos preditivos são usados para planejar e aprimorar compostos que possuem atividade biológica e também podem ser usados como fármacos [Sant'Anna, 2002]. Nessa dissertação, a atividade biológica representa a dose necessária parar interferir em um efeito biológico.

O CADD é bastante usado quando desejamos criar um novo medicamento para tratar uma determinada doença, mas não é possível obter informações sobre os alvos terapêuticos envolvidos na função biológica do efeito que se deseja fazer intervenção para tratar essa doença [Barreiro, 2009].

Uma das abordagens de CADD mais usadas é o estudo das relações quantitativas estrutura-atividade (QSAR). Essa abordagem consiste no desenvolvimento de um modelo preditivo que relaciona as atividades biológicas de compostos previamente testados com suas propriedades estruturais, físico-químicas e conformacionais [Martins & Ferreira, 2013]. Nesses estudos o modelo preditivo normalmente usado é uma regressão linear.

Dentre as estratégias de QSAR, Hopfinger et al. [1997] introduziram o 4D-QSAR que incorpora, através de uma simulação de dinâmica molecular, a informação de flexibilidade conformacional e liberdade de alinhamento na quarta dimensão para o cálculo dos descritores que quantificam as propriedades dos compostos no espaço 3D. Esses descritores são usados para o treinamento do modelo preditivo. Essa representação do composto onde calculamos os descritores no 4D-QSAR é chamada de perfil de amostragem conformacional (PAC). A Figura 1.1 ilustra quatro exemplos de PACs gerados

por uma simulação de dinâmica molecular.

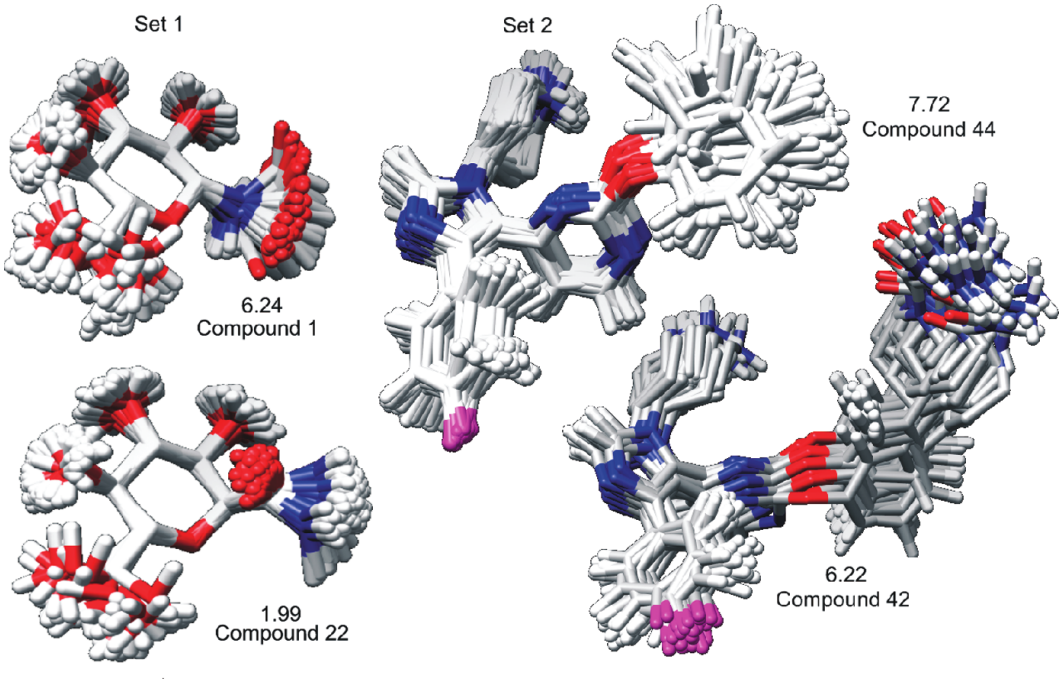


Figura 1.1: Quatro exemplos de PACs gerados por uma simulação de dinâmica molecular. Fonte: [Martins et al., 2009].

O LQTA-QSAR é uma técnica de 4D-QSAR, proposta por Martins et al. [2009], que faz uso dos descritores de Coulomb (C) e Lennard-Jones (LJ) para quantificar interações entre o PAC e uma sonda (que pode ser um átomo, íon ou grupo funcional) posicionada em algum lugar no espaço 3D. Esses descritores são usados posteriormente para o treinar um modelo de regressão linear. A amostragem desses descritores no espaço é feita da seguinte maneira: o PAC é inserido no centro de uma grade 3D, em seguida a sonda percorre cada ponto dessa grade calculando os descritores de C e LJ, entre o PAC e a sonda nesse ponto.

O problema da amostragem feita pelo LQTA-QSAR é que a sonda passa por pontos internos ao PAC, amostrando descritores com valores irreais que prejudicam todo o fluxo de trabalho para treinar o modelo preditivo. Os pontos internos ao PAC não podem ser associados como uma região do ligante interagindo com um receptor, que pode ser qualquer componente biológico que interage com o composto e provoca um efeito. Quando esses descritores são usados no modelo de regressão, além deles degradarem a precisão, eles fazem com o que modelo perca o sentido para interpretações químicas, pois não é possível mapear esses pontos em regiões reais do receptor.

Para contornar os problemas apresentados, o objetivo dessa dissertação é propor uma nova amostragem de descritores que considera a posição no espaço e formato

do PAC usando técnicas de geometria computacional. A proposta faz ampliações da superfície definida pelo fecho convexo para construir camadas ao redor do PAC, que representam os pontos onde a sonda deve percorrer para calcular os descritores de interação. Essa amostragem impede que a sonda passe por pontos internos ou próximos demais ao PAC.

A principal contribuição dessa proposta é melhorar a qualidade dos descritores amostrados e consequentemente aumentar a precisão dos modelos preditivos treinados com esses descritores. Além de possibilitar que os modelos preditivos façam sentido para interpretações química, descartando a possibilidade de que sejam mapeados pontos internos ao PAC.

Outro objetivo dessa dissertação é propor um novo fluxo de trabalho que substitui o modelo linear por uma árvore de regressão. A principal contribuição dessa proposta é gerar modelos mais fáceis de serem interpretados. Pois as árvores de regressão são modelos de fácil interpretação que conseguem encontrar padrões não lineares nos dados, que podem complementar a análise que é feita apenas com os modelos lineares.

Os próximos capítulos dessa dissertação estão organizados da seguinte maneira: O Capítulo 2 apresenta alguns conceitos básicos necessários para entendimento da dissertação, bem como os trabalhos relacionados em QSAR que serviram de base para o desenvolvimento dessa dissertação. O Capítulo 3 apresenta a nova amostragem de descritores proposta, bem como o fluxo de trabalho usando árvores de regressão. Já no Capítulo 4 são apresentados os experimentos realizados, bem como a discussão dos resultados. Finalmente, no Capítulo 5 é feita uma conclusão sobre a dissertação, bem como são apontados possíveis trabalhos futuros.

Capítulo 2

Embasamento Teórico

Esse Capítulo apresenta os conceitos básicos necessários para entendimento da dissertação, bem como os trabalhos relacionados que serviram como base teórica. Na Seção 2.1 é definida abordagem para predição de atividade biológica nesse trabalho, o estudo das relações quantitativas estrutura-atividade (QSAR). Já a Seção 2.2 apresenta um fluxo de trabalho para aprendizado dos modelos QSAR. Na Seção 2.3 é apresentada a etapa responsável pela geração de características, bem como o LQTA-QSAR [Martins et al., 2009], que foi a principal referência desse trabalho. Na Seção 2.4 são apresentados os conceitos básicos sobre redução de dimensionalidade, bem como o método proposto por Teófilo et al. [2009] que é usado nessa dissertação. Na Seção 2.5 são apresentados conceitos sobre análise de regressão, que é o modelo de predição usado nesse trabalho. E por fim, na Seção 2.6 são apresentados outros trabalhos relacionados em QSAR que usam diferentes métodos no fluxo de trabalho para construir os modelos.

2.1 Relações Quantitativas Estrutura-Atividade (QSAR)

Planejar racionalmente novos medicamentos demanda conhecimento sobre os mecanismos que influenciam no efeito biológico de um ligante interagindo com um receptor, onde se deseja fazer intervenção nesse efeito para tratar uma determinada doença. Dessa forma deve-se obter por meios de estudos elementares informações sobre os alvos terapêuticos envolvidos na função biológica. [Barreiro, 2009].

Quando não é possível obter essas informações, devem ser usadas outras estratégias para planejar compostos candidatos. Uma estratégia bastante usada é o planejamento de fármacos auxiliado por computador (CADD), que consiste em usar modelos

preditivos para planejar e aprimorar compostos que possuem atividade biológica e podem ser usados como fármacos [Sant'Anna, 2002].

Dentre as abordagens de CADD, uma das mais usadas é o estudo das relações quantitativas estrutura-atividade (QSAR). Essa abordagem consiste no desenvolvimento de um modelo preditivo que relaciona propriedades estruturais, físico-químicas e conformacionais de compostos previamente testados com suas atividades biológicas [Martins & Ferreira, 2013].

No treinamento dos modelos QSAR, as variáveis independentes são modeladas com os descritores obtidos em estudos de química teórica sobre perfis de amostragem conformacional (PAC) dos compostos. Esses descritores quantificam propriedades físico-químicas dos compostos. Já a variável dependente, ou seja, a que deseja-se prever, é modelada com alguma atividade biológica obtida experimentalmente. Uma das atividades mais usadas, inclusive nesse trabalho, é a IC_{50}, que representa a dose necessária para interferir em um efeito biológico em 50%. Como as atividades biológicas podem apresentar valores em escalas muito altas, é aplicada a transformação do logaritmo inverso, onde temos a atividade como pIC_{50} [Nantasenamat et al., 2009].

2.2 Fluxo de Trabalho para Aprendizado de Modelos QSAR

Podemos definir genericamente o fluxo de trabalho para aprendizado de modelos QSAR em três etapas básicas: geração de características, redução de dimensionalidade e modelo de predição.

A geração de características é a primeira etapa, onde recebemos os PACs dos compostos e devolvemos o conjunto de dados que tem os descritores como características.

A etapa posterior é a redução de dimensionalidade, que é responsável por reduzir a dimensão do conjunto de dados gerado na primeira etapa, de maneira que sejam preservadas as informações mais relevantes sobre os dados.

Por fim, a última etapa consiste no treinamento do modelo preditivo, que no nosso caso é uma regressão. Nessa etapa é possível extrair métricas que permitem fazer as avaliações quantitativas e interpretar qualitativamente o modelo que foi construído.

A Figura 2.1 ilustra esse fluxo de trabalho e permite ter uma visão de como é o processo de construção do modelo final.

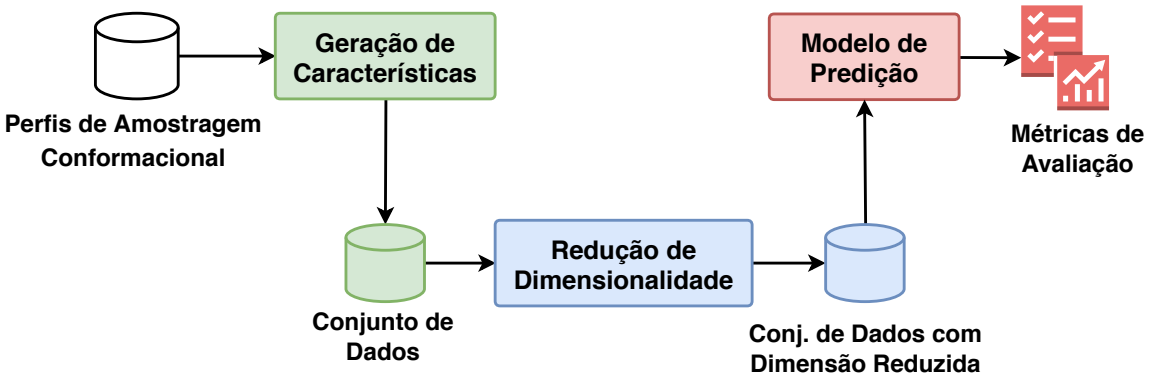


Figura 2.1: Fluxo de trabalho para predição de atividade biológica.

2.3 Geração de Características

A geração de características é a primeira e mais importante etapa na construção de modelos QSAR, pois é ela que define quais tipos de descritores devem ser amostrados. Existem vários métodos para fazer amostragem sobre os compostos, um dos mais conhecidos é o 3D-QSAR, que usa informações sobre o composto no espaço 3D, como as posições dos átomos e cargas, para calcular os descritores que quantificam as propriedades que deseja-se estudar [Andrade et al., 2010; Ferreira, 2002].

Uma abordagem que traz melhorias a esse processo é o 4D-QSAR, que incorpora uma simulação de dinâmica molecular no cálculo dos descritores sobre o composto no espaço 3D, ou seja, a quarta dimensão pode ser considerada como a informação de conformação do composto em um determinado tempo da simulação. A representação onde calculamos os descritores na abordagem 4D é chamada de perfil de amostragem conformacional (PAC).

A primeira proposta de 4D-QSAR foi apresentada por [Hopfinger et al., 1997] que incorpora a informação de flexibilidade conformacional e liberdade de alinhamento na quarta dimensão para o cálculo dos descritores. Desse modo, foram obtidos resultados bastante promissores que motivaram o desenvolvimento dessa abordagem.

2.3.1 LQTA-QSAR

Dentre as diversas abordagens 4D-QSAR, a presente dissertação foi desenvolvida usando como referência o método LQTA-QSAR, proposto por Martins et al. [2009]. Para gerar os PACs, o LQTA-QSAR faz uso das trajetórias da dinâmica molecular e informações espaciais das moléculas recuperadas do pacote GROMACS [Van Der Spoel et al., 2005]. Os PACs gerados na simulação contém a conformação do composto no

espaço tridimensional (x, y, z) para cada instante t da simulação, formando assim a quarta dimensão.

Para cada PAC, os descritores são gerados da seguinte maneira: Primeiro o PAC é inserido em uma grade 3D, que representa o receptor, com volume $a \times b \times c$ Å^3, onde o passo entre dois pontos da grade é definido por p Å (normalmente com $p = 1$). Em cada ponto dessa grade, são calculados dois descritores de interação entre a sonda e o PAC: Coulomb (C) e Lennard-Jones (LJ).

Os descritores de C e J são calculados de acordo com as Equações 2.1 e 2.2 respectivamente. Onde q_i é a carga da sonda i, q_j é a carga do j-ésimo átomo do PAC, ϵ_0 é a constante de permissividade do vácuo, r_{ij} é a distância entre a sonda i e o j-ésimo átomo do PAC, o conjunto $\{K_{ii}^{12}, K_{ii}^{6}, K_{jj}^{12}, K_{jj}^{6}\}$ são os parâmetros relacionados aos campos de força para a sonda e os átomo do PAC, por fim n é quantidade de instantes t do PAC.

$$C = \sum_{j} \frac{1}{n} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.1}$$

$$\begin{aligned} LJ &= \sum_{j} \frac{K_{ij}^{12}}{r_{ij}^{12}} - \frac{K_{ij}^{6}}{r_{ij}^{6}} \\ K_{ij}^{12} &= \sqrt{\frac{K_{ii}^{12} \cdot K_{jj}^{12}}{n}} \\ K_{ij}^{6} &= \sqrt{\frac{K_{ii}^{6} \cdot K_{jj}^{6}}{n}} \end{aligned} \tag{2.2}$$

Dessa forma, os descritores amostrados geram um conjunto de dados $X = \{\vec{x_1}, \ldots, \vec{x_m}\}$ que é usado para construir os modelos de regressão, onde cada $\vec{x} = (f_1, \ldots, f_n)$, sendo m igual ao número de PACs do conjunto e n igual a quantidade de descritores amostrados.

O conjunto de dados é usado para construir um modelo de predição de atividade biológica usando o Ordered Predictors Selection (OPS) [Teófilo et al., 2009] para redução de dimensionalidade, que será detalhado posteriormente na Seção 2.4 e *Partial Least Squares* (PLS) [Hastie et al., 2003] para treinamento do modelo de regressão, que será detalhado na Seção 2.5. Os experimentos mostraram que é possível treinar modelos com alta precisão para predição de atividade biológica usando os conjuntos de dados gerados com o LQTA-QSAR.

2.4 Redução de Dimensionalidade

Os problemas para mineração de dados estão frequentemente associados a alta dimensionalidade dos conjuntos de dados. Os modelos para predição de atividade biológica também precisam lidar com esse problema, sendo essa uma etapa bastante relevante para a construção do modelo.

Um dos grandes problemas enfrentados para mineração de dados em bases de alta dimensionalidade, além da complexidade de tempo e espaço, é a precisão dos modelos preditivos. Essas bases normalmente tem muitas informações redundantes ou irrelevantes, que comprometem o desempenho dos modelos treinados para solucionar o problema. Como exemplo, nos conjuntos de dados para predição de atividade biológica, onde podemos ter várias características que são calculadas sobre pontos do espaço que podem ser irrelevantes ou redundantes, consequentemente essas características não contribuem para prever a atividade biológica.

O efeito negativo da grande quantidade de características é conhecido como "maldição da dimensionalidade" e foi abordado inicialmente em estudos sobre teoria de controle dos processos [Bellman, 1957]. Também já foi mostrado que a precisão de um modelo preditivo melhora de acordo com o aumento da quantidade de características usadas até um tamanho ótimo. Se for usada uma quantidade de características maior que o tamanho ótimo, a precisão diminui [Hughes, 2006].

Existem dois tipos abordagens para se fazer redução de dimensionalidade: seleção ou extração de características. Embora sejam abordagens diferentes, normalmente as duas são utilizadas durante o processo para obter um modelo melhor. Primeiro é realizada a seleção e depois no resultado é executada a extração. Essas duas abordagens serão definidas nas subseções a seguir.

2.4.1 Seleção de Características

A seleção de características é uma das abordagens para redução de dimensionalidade mais utilizadas. Essa abordagem tem o objetivo de identificar e descartar as características mais irrelevantes para o problema [Blum & Langley, 1997].

O processo de seleção pode ser definido de maneira genérica da seguinte forma: Seja o conjunto de dados $X = \{\vec{x_1}, \ldots, \vec{x_m}\}$, onde cada $\vec{x} = (f_1, \ldots, f_n)$ sendo $F = \{f_1, \ldots, f_n\}$. O método de seleção busca por um subconjunto $F' \subset F$ tal que $|F'| < |F|$ e F' contenha os elementos mais representativos do conjunto F, ou seja, os que carregam maior quantidade de informação do conjunto original de características. O critério dessa seleção e como a representatividade do subconjunto é medida depende da abordagem

que está sendo utilizada.

Baseando-se na disponibilidade da variável que desejamos prever, os métodos de seleção podem ser divididos em duas abordagens: supervisionado e não-supervisionado. Quando é usada a variável que desejamos prever no processo de seleção o método é chamado de supervisionado, caso contrário é considerado não-supervisionado. [Liu & Yu, 2005].

Um dos tipos mais básicos entre os métodos para seleção é o *wrapper*. Esse tipo de método avalia as possíveis combinações de subconjuntos características usando alguma medida que quantifique a precisão do modelo preditivo treinado com essas características [Blum & Langley, 1997]. Exemplos de *wrappers* bastantes conhecidos são o *Foward Feature Selection* e o *Backward Feature Selection*, esses algoritmos utilizam uma busca heurística para encontrar o subconjunto de características que consegue obter maior precisão com um determinado modelo preditivo [Kohavi & John, 1997].

2.4.2 Extração de Características

A extração de características é outra abordagem bastante utilizada para redução de dimensionalidade. Esses métodos executam uma transformação no conjunto de dados, para obter novas características que podem ter capacidade de discriminação melhor do que um subconjunto dos dados originais [Sammut & Webb, 2011].

Uma das formas mais usadas para aplicar essa abordagem é definir a extração através de uma projeção linear dos dados em um espaço de dimensão menor. Esse processo pode ser definido genericamente da seguinte maneira: Seja $X = [x_1, \ldots, x_m]^T$ o conjunto de dados, onde cada $x_i = [f_1, \ldots, f_n]$ e $W = [w_1, \ldots, w_k]$ a base da transformação com cada vetor $w_i = [t_1, \ldots, t_n]^T$ onde $k < n$. O novo conjunto de dados X' com dimensão reduzida igual a k é dado por $X' = XW$. A base da transformação W é construída de acordo com o método que está sendo usado.

Um dos principais exemplos desse tipo de extração é a Análise de Componentes Principais (PCA), que faz a redução de dimensionalidade projetando o conjunto de dados nos vetores com direção de maior variância nos dados. Para isso, a base transformação W é definida com os k autovetores da matriz de covariância dos dados com maiores autovalores associados [Bishop, 2006; Theodoridis & Koutroumbas, 2008; Zaki & Jr, 2014].

2.4.3 Ordered Predictors Selection (OPS)

A redução de dimensionalidade é sem dúvidas uma das mais importantes no fluxo de trabalho para construção de modelos QSAR. Essa etapa tem impacto direto tanto na precisão dos modelos de regressão, quando na interpretação das caraterísticas selecionadas.

Os algoritmos genéticos podem ser usados para realizar seleção de características em modelos QSAR [Rogers & Hopfinger, 1994]. Esse tipo de abordagem gera uma população com diversos modelos de regressão que usam diferentes combinações das características originais. Depois é procurado o modelo que maximiza a precisão da predição. O problema desse método é que o tempo de execução é alto e nem sempre é possível encontrar um modelo de alta precisão em tempo hábil.

Para contornar os problemas encontrados ao usar técnicas como algoritmos genéticos, Teófilo et al. [2009] propuseram o Ordered Predictors Selection (OPS), um método para seleção e extração de características que tem a proposta de ser rápido e eficaz para problemas de regressão.

O algoritmo ordena as características de acordo com um vetor informativo (IV) que quantifica a relevância de cada característica. Em seguida, para selecionar a quantidade de características que serão selecionadas, bem como o número de variáveis latentes que serão extraídas, o algoritmo vai acrescentando cada característica em ordem a um modelo de regressão PLS. Por fim, são retornados os hiperparâmetros do modelo onde a regressão PLS obteve melhor desempenho em uma validação cruzada Leave-One-Out (LOO) de acordo com uma métrica de qualidade (normalmente o Q^2, que será detalhado no Capítulo 4).

Seja f_i uma característica qualquer, IV_i o valor correspondente a essa em IV e Y o vetor com as atividades biológicas. O OPS pode ser usado com três tipos de IV:

1. Correlação: $\text{IV}_i = |r(f_i, Y)|$, onde r é a correlação de Pearson.
2. Regressão: $\text{IV}_i = |\beta_i|$, onde β_i é o coeficiente que f_i recebe no treinamento de uma regressão PLS para prever Y.
3. Produto: $\text{IV}_i = |\beta_i| \cdot |r(f_i, Y)|$, o produto dos dois vetores anteriores.

O método recebe como entrada os seguintes hiperparâmetros: ρ que é o valor mínimo de correlação que uma f_i deve ter com Y, maxNV que é número máximo de características que devem ser testadas, maxNVL que é número máximo de variáveis latentes, IV como o vetor informativo usado e maxOVL que é usado apenas com os IV de regressão ou produto, ele representa o número máximo de variáveis latentes testadas

na regressão usada para ordenação, no caso do vetor de correlação ele é configurado como 1 e não tem impacto.

Um pseudocódigo do OPS é apresentado no Algoritmo 1, que recebe como entrada as características do conjunto de dados $F = \{\vec{f_1}, \ldots, \vec{f_n}\}$ o vetor Y com as atividades biológicas e os hiperparâmetros ρ, maxNV, maxNVL, maxOVL e IV. O primeiro passo é um pré-processamento onde são descartadas as características que possuem correlação de Pearson com o vetor de atividades biológicas menor que o limiar ρ, pois essas características podem ter capacidade de discriminação desprezível. Em seguida, a busca pelos hiperparâmetros do melhor modelo é executada de fato. O laço mais externo serve para quando são usados os IV de regressão ou produto, onde são testadas ordenações com diferentes números de variáveis latentes do PLS. Para o caso do IV de correlação esse laço é executado apenas uma vez, pois só existe uma ordenação possível.

O algoritmo retorna os seguintes hiperparâmetros do melhor modelo encontrado: NV como a quantidade de características selecionadas, NVL como número de variáveis latentes extraídas com o PLS e OVL, para os casos onde IV é de regressão ou produto, como o número de variáveis latentes usadas no PLS para ordenação.

Algoritmo 1 OPS

Entrada: $F = \{\vec{f_1}, \ldots, \vec{f_n}\}$, Y, ρ, maxNV, maxNVL, maxOVL, IV.
Saída: NVL, NVL, OVL e Q^2_{max}.

$F' \leftarrow \{f \in F | r(f, y) \leq \rho\}$
$Q^2_{\text{max}} \leftarrow -\infty$
OVL, NV, NVL $\leftarrow 0$
para $i \leftarrow 1, i \leftarrow i + 1$ **até** maxOVL **faça**
    Ordene F' usando IV com i variáveis latentes no PLS.
    para $j \leftarrow 1, j \leftarrow j + 1$ **até** maxNV **faça**
        para $k \leftarrow 1, k \leftarrow k + 1$ **até** maxNVL **faça**
            $T \leftarrow$ As f_j primeiras características de F'.
            Execute a validação cruzada usando T e k variáveis latentes no PLS.
            Obtenha o Q^2 da validação cruzada.
            se $Q^2_{\text{max}} \leq Q^2$ **então**
                OVL, NV, NVL $\leftarrow i, j, k$.
                $Q^2_{\text{max}} \leftarrow Q^2$.
            fim se
        fim para
    fim para
fim para
    devolve OVL, NV e NVL.

O OPS é simples de ser implementado e pode ser executado rapidamente, pois o número de iterações é definido pelo produto dos hiperparâmetros maxOVL, maxNV

e maxNVL. Esse método também possibilita que sejam criados modelos que não são impactados pelo problema de sobreajuste da regressão, pois a busca pelo melhor modelo é feita de forma supervisionada, maximizando uma medida de qualidade sobre uma validação cruzada.

Isso faz com que o OPS seja uma alternativa mais adequada para redução de dimensionalidade em modelos QSAR, pois os conjuntos de dados sobre QSAR normalmente possuem poucos exemplos, o que facilmente sobreajusta os modelos de regressão caso não sejam usadas técnicas para redução de dimensionalidade que levam em consideração métricas que ajudam a mitigar esse problema, como o Q^2.

2.5 Análise de Regressão

Podemos definir formalmente a análise de regressão como dado um conjunto $X = \{\vec{x_1}, \ldots, \vec{x_m}\}$, onde cada $\vec{x} = (f_1, \ldots, f_n)$ sendo $F = \{f_1, \ldots, f_n\}$, que podemos chamar também de variáveis independentes e $Y = \{y_1, \ldots, y_n\}$ com os valores que desejamos prever, que também podemos chamar de variável dependente. A regressão consiste no ajuste de uma função que tenta explicar a variação da variável dependente Y através das variações nas variáveis independentes em F [Jain, 1991].

A relação entre Y e as variáveis contidas em F pode ser ajustada tanto em uma função linear quanto em uma não linear, que pode ser quadrática, cúbica, exponencial, logarítmica ou qualquer outra função válida. No caso do problema desse trabalho, modelamos a relação entre os descritores de interação e atividade biológica que desejamos prever.

2.5.1 Regressão Linear

A regressão linear é amplamente usada e consegue atender vários problemas no mundo real. Podemos definir esse modelo para prever um valor y_i usando um exemplo x_i de acordo com a Equação 2.3, no caso onde temos n variáveis independentes em F.

$$y_i = \beta_0 + \beta_1 x_i(f_1) + \ldots + \beta_n x_i(f_n) + \varepsilon_i \tag{2.3}$$

Nesse caso, a função de regressão possui os seguintes parâmetros: β_0 como a constante de regressão, $\beta_1, \ldots, \beta_n$ como os coeficientes de regressão para cada valor de x_i no nível f_j e ε_i como o erro associado.

O ajuste dos parâmetros desse modelo de regressão são feitos pelo método dos mínimos quadrados, onde são buscados parâmetros que minimizam a soma dos erros

quadrados entre os valores reais da variável dependente e os previstos pela função que deseja-se ajustar [Jain, 1991].

2.5.2 Partial Least Squares (PLS)

Um modelo bastante usado em problemas para predição de atividade biológica é o *Partial Least Squares* (PLS) [Hastie et al., 2003], que usa redução de dimensionalidade para melhorar a precisão da regressão linear.

O PLS faz extração de características decompondo a matriz $X_{m\times n}$ formada pelo conjunto de dados em uma matriz $X'_{m\times k}$ dado o conhecimento da variável que desejamos prever $Y_{m\times 1}$, de maneira que conseguimos obter a maior covariância possível entre X e Y. As k variáveis do conjunto de dados com dimensão reduzida são chamadas de variáveis latentes (VL).

A decomposição é feita de acordo com a Equação 2.4,

$$\begin{aligned} X &= TP^T + E \\ Y &= UQ^T + F \end{aligned} \tag{2.4}$$

Onde T e U são matrizes $m \times k$, chamadas de scores, que contém os k vetores que formam as variáveis latentes. P é uma matriz $n \times k$ e Q é uma matriz $1 \times k$ que são chamados de *loadings*. Já o erro é representado pelas matrizes E $(m \times k)$ e F $(m \times 1)$. Nessa decomposição, deseja-se maximizar a relação linear entre T e U de maneira que os erros E e F sejam os menores possíveis.

Um dos métodos mais usados para fazer essa decomposição é o Nonlinear Iterative Partial Least Squares (NIPALS) [Wold, 2006; Schwartz et al., 2009]. Nesse método é construída a base da transformação $W = [w_1, \ldots, w_k]$ de acordo com a Equação 2.5,

$$[cov(t_i, u_i)]^2 = \max_{|w_i|=1} [cov(Xw_i, Y)]^2 \tag{2.5}$$

Onde em cada interação, t_i e u_i são respectivamente as i-ésimas colunas de T e U, de forma que $cov(t_i, u_i)$ é a covariância entre os vetores t_i e u_i. Logo após, X e Y são subtraídos pelas suas aproximações de posto baseadas em t_i e u_i. O processo é executado até que os k vetores da transformação sejam extraídos.

Após fazer a redução de dimensionalidade, podemos aprender os coeficientes da regressão linear no conjunto de dados X projetado em W.

2.6 Outros Trabalhos em Aprendizado de Modelos QSAR

Ao longo do tempo, foram desenvolvidos diversos trabalhos que exploram várias técnicas para treinamento de modelos QSAR, contribuindo para cada etapa do fluxo de trabalho: geração de características, redução de dimensionalidade e análise de regressão. A etapa para geração de características está diretamente relacionada a abordagem QSAR adotada para amostrar descritores. As demais etapas focam em técnicas de aprendizado de máquina para melhorar a qualidade dos modelos de predição construídos.

Nos parágrafos seguintes, serão pontuados alguns trabalhos da área que exploram diferentes técnicas para geração de caraterísticas, redução de dimensionalidade e análise de regressão no treinamento de modelos QSAR.

Modeling Steric and Electronic Effects in 3D- and 4D-QSAR. Polanski & Bak [2003] apresentaram um estudo sistemático que avalia o desempenho de modelos QSAR 3D e 4D na modelagem de efeitos estéricos e eletrônicos.

Foram aplicados os métodos 4D-QSAR de Hopfinger e Análise Comparativa de Campo Molecular (CoMFA) para gerar características que são usadas na predição de valores pK_a de ácidos benzoicos, que foram divididos em três subséries para simular diferentes níveis de controle estérico e eletrônico. Após a geração do conjunto de dados, um modelo de regressão linear foi treinado usando o PLS.

Ambos os métodos conseguiram construir modelos com desempenho comparáveis. Adicionalmente foi proposto o novo método 4D-QSAR baseado em redes neurais auto-organizáveis [Anzali et al., 1998] para geração de características, que teve resultados comparáveis com os demais métodos.

Self-organizing Neural Networks for Pharmacophore Mapping. Esse trabalho desenvolvido por Polanski [2003], mostrou que as redes neurais auto-organizáveis podem ser uma boa estratégia para mapeamento de fármacos, que é uma representação generalizada das características dos compostos. Esse tipo de rede permitem gerar representações difusas que podem descobrir aspectos de similaridade que facilmente podem passar despercebidos por uma análise humana. Para chegar a essa conclusão, foram conduzidos experimentos com dados sobre HIV-1 *integrase*.

Esse trabalho contribui para etapa de geração de características, porém uma desvantagem é a complexidade do treinamento de redes neurais auto-organizáveis, que demanda conhecimento prévio em otimização de hiperparâmetros. Esse método também não pode ser facilmente disponibilizado em um software com interface amigável

para o usuário, isso poderia transformar a análise em uma tarefa bastante custosa.

Self-organizing Neural Networks for Modeling Robust 3D and 4D QSAR. Já Polanski et al. [2004] apresentam contribuições em todas as etapas do fluxo de trabalho para predição de atividade biológica. São propostos modelos para geração de características QSAR 3D e 4D usando redes neurais auto-organizáveis. Já a redução de dimensionalidade é feita com o método *Uninformative Variable Elimination* (UVE), proposto por Centner et al. [1996], que remove características de acordo com os coeficientes obtidos em uma regressão usando o PLS. Para a predição, o modelo linear foi mantido.

Esse fluxo de trabalho foi usado na modelagem da atividade de uma série de inibidores para diidrofolato redutase. Foram conduzidos diversos experimentos quantitativos onde o modelo obteve bons resultados na predição, de acordo com a métrica Q^2, o que nos permite concluir que o método pode ser uma alternativa para prever a atividade desse tipo de inibidores avaliados.

Coding Molecules in 3D-QSARs - from a Point to Surface Sectors and Molecular Volumes. Gieleciak et al. [2005] apresentam estudos comparativos de modelos 3D-QSAR usando as abordagens Análise Comparativa de Campo Molecular (CoMFA) e Análise Comparativa de Superfície Molecular (CoMSA), que tem foco em analisar a superfície do composto. Os métodos foram usados para gerar características sobre uma série de inibidores: hipolipemiantes, antiagregantes plaquetários e antifúngicos.

Os estudos mostraram que o CoMSA permite fazer uma análise da atividade biológica com maior foco nas particularidades e formatos de um composto individualmente, diferente do CoMFA que é mais genérico. A redução de dimensionalidade também foi feita usando o UVE, onde o processo foi chamado de *Iterative Variable Elimination* (IVE). Por fim, a predição foi feita usando um modelo linear. Nos experimentos quantitativos foi mostrado que os métodos avaliados conseguem gerar descritores com alta capacidade de predição, avaliando pela métrica Q^2.

Iterative Variable Elimination Schemes for CoMSA. Gieleciak & Polanski [2007] questionam sobre a eficiência de métodos 3D-QSAR e que sua importância prática no projeto de novos fármacos ainda é um assunto controverso, tanto na predição da atividade biológica, quanto na identificação de regiões que são responsáveis por esses efeitos biológicos, que são influenciadas pelos métodos para seleção e extração de características que está sendo usado.

Para abordar esse problema, nesse trabalho foi usada uma série de ácidos benzoicos para modelar usando CoMSA a constante de Hammett (que é correlacionada com os valores pK_a), com o objetivo de avaliar a dependência entre precisão e seleção de

características nos modelos construídos com IVE e PLS. Ao modelar esse efeito químico, foi construída uma variante do IVE-PLS modificando o critério para descartar características usando uma validação cruzada. Os valores que quantificam a relevância das características foram usados para traçar os mapas de contorno que indicam uma função carboxílica, ou seja, a região que inclui o centro de reação de dissociação que determina os respectivos valores de pK_a.

A contribuição maior desse tipo de trabalho é na identificação das regiões do composto que influenciam na atividade biológica, que está diretamente relacionada com a redução de dimensionalidade.

SOM-4D-QSAR with Iterative Variable Elimination IVE-PLS. Bak & Polanski [2007] propuseram um método de 4D-QSAR baseado em redes neurais auto-organizáveis (SOM-4D-QSAR) com foco na sua capacidade para mapeamento de fármacos. Foi usado um processo estocástico para verificar a capacidade de precisão em uma grande população de modelos 4D-QSAR gerados, visando contribuir para etapa de geração de características.

Esse estudo foi conduzido em uma série de ácidos benzoicos, azo-compostos e esteroides. Os experimentos mostraram que o método 4D-QSAR proposto em conjunto com o IVE-PLS conseguem construir modelos para predição de atividade biológica com alta precisão para os azo-compostos e esteroides, já para os ácidos benzoicos o método se mostrou menos eficaz.

Grande parte dos trabalhos apresentados anteriormente usam redes neurais auto-organizáveis para construção dos modelos QSAR, a desvantagem desses métodos é que eles são complexos e de difícil interpretação. Já os trabalhos que usam 3D-QSAR tem várias limitações da própria abordagem. Dessa forma, o método LQTA-QSAR se mostrou um dos mais promissores para o desenvolvimento do modelos 4D-QSAR, pois ele é simples de ser aplicado e permite gerar modelos com boa precisão.

Capítulo 3

Métodos Propostos

Nesse Capítulo, serão apresentadas as principais contribuições desse trabalho. Na Seção 3.1 será apresentada uma nova abordagem para geração de características em modelos 4D-QSAR, que apresenta algumas melhorias em relação à abordagem original LQTA-QSAR [Martins et al., 2009]. Já na Seção 3.2 é apresentado um novo modelo para predição de atividade biológica usando Árvores de Regressão. Esse método proposto gera modelos que são fáceis de serem interpretados e podem ser uma alternativa aos modelos lineares.

3.1 Geração de Características usando Geometria Computacional

Um dos grandes problemas encontrados na amostragem feita pelo LQTA-QSAR, que foi discutido no Capítulo 2, é que o perfil de amostragem conformacional (PAC) do composto é inserido em uma grade 3D como representante do receptor, que pode ser qualquer componente biológico que interage com o composto e provoca um efeito. Dessa forma, a sonda (que pode ser um átomo, íon ou grupo funcional) vai passar por pontos internos ao PAC durante a amostragem (como na Figura 3.1). Isso gera descritores com valores irreais, pois pontos internos ao PAC não podem ser associados como uma região do ligante interagindo com o receptor. Os descritores gerados a partir de pontos internos ao PAC prejudicam todo o fluxo de trabalho para construir os modelos para prever a atividade biológica. Quando esses descritores são incluídos no treinamento do modelo de regressão, além deles degradarem a precisão, eles fazem com o que modelo perca o sentido para interpretações químicas, pois não é possível mapear esses pontos em regiões reais do receptor.

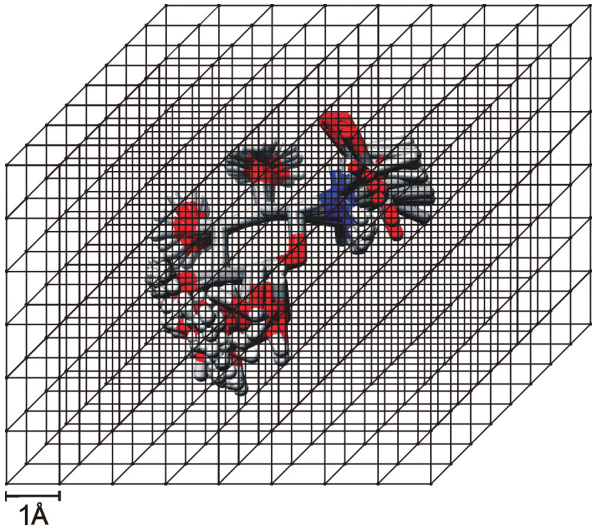


Figura 3.1: Ilustração do espaço de amostragem do LQTA-QSAR, onde cada ponto é amostrado com uma passo 1Å. Fonte: [Martins et al., 2009].

Para contornar os problemas encontrados, essa dissertação propõe uma nova amostragem que considera a posição no espaço e formato do PAC na definição do conjunto de pontos $Z \subseteq \mathbb{R}^3$, onde a sonda deve percorrer para amostrar os descritores. Técnicas da geometria computacional permitem estimar o formato dos compostos para que seja possível impedir a sonda de passar por pontos internos ou próximos demais ao PAC durante a amostragem. Para detectar o formato do composto, o método proposto calcula o fecho convexo dos pontos que representam os átomos do PAC. Depois disso, o conjunto de pontos Z, que nessa proposta representa o receptor, é definido a partir de camadas ao redor do PAC construídas através de ampliações da superfície definida pelo fecho convexo.

Formalmente, essas camadas são resultados de homotetias com centro igual ao centroide da superfície e com razões definidas de forma que a primeira camada esteja a uma distância miníma d_{min} da superfície e que o espaço entre as demais camadas seja uma constante Δ_r definida pelo analista, até que a última camada respeite uma distância máxima d_{max}. Um mesmo número de pontos são amostrados em cada camada da seguinte maneira: os pontos que intersetam com retas que passam pelo centroide com um ângulo (θ, ϕ) que varia por um passo constante Δ_α (em cada uma das duas dimensões), também escolhido pelo analista.

A apresentação dessa contribuição está organizada da seguinte maneira: A Subseção 3.1.1 define formalmente a abordagem proposta. Por fim, as Subseções 3.1.2 e 3.1.3 apresentam duas maneiras de implementar a proposta para definição do conjunto de pontos Z.

3.1.1 Definição do Método

Seja $M = \{\mu_1, \ldots, \mu_m\}$ o conjunto de PACs, onde cada μ está associado a um composto que possui atividade biológica e pode ser usado como fármaco. Cada PAC μ possui pontos tridimensionais $P = \{p_1, \ldots, p_k\}$, em que cada $p \subseteq \mathbb{R}^3$. Esses pontos representam as coordenadas no espaço de cada átomo pertencente ao PAC.

Para definirmos o conjunto de pontos Z, onde a sonda deve percorrer e calcular os descritores, calculamos o fecho convexo H no conjunto dos pontos P de um determinado PAC μ_i. No contexto de um conjunto finito de pontos 3D, H representa o menor volume convexo que inclui P. Ele é único e pode ser calculado através do algoritmo de Chan em tempo $\mathcal{O}(|P| \log h)$, onde h é o número de vértices, necessariamente em P, na superfície do fecho convexo [Chan, 1996; Berg et al., 2008]. Para ilustrar o problema, na Figura 3.2 temos um exemplo de fecho convexo para pontos em 2D, que é calculado da mesma forma em 3D. A partir de H, conseguimos definir uma superfície que envolve o PAC, dessa forma consideramos o formato do PAC na definição dos pontos onde a sonda deve percorrer.

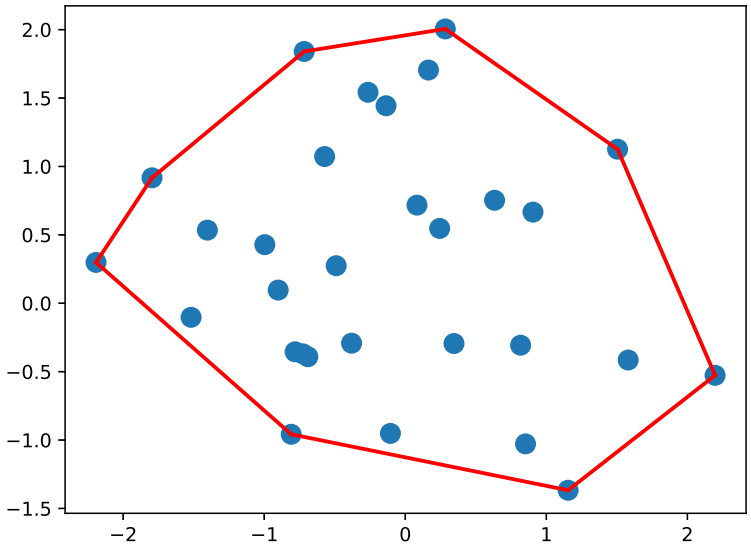


Figura 3.2: Exemplo de um fecho convexo calculado sobre pontos 2D.

Após calcularmos o fecho convexo H, definimos em coordenadas esféricas o conjunto de pontos Z, onde para todo $z \in Z$, $z = (\theta, \varphi, r)$ $|\theta \in \{k\Delta_\alpha \; | k \in \mathbb{N} \text{ e } k\Delta_\alpha \leq 179°\}$, $\varphi \in \{k\Delta_\alpha \; | k \in \mathbb{N} \text{ e } k\Delta_\alpha \leq 359°\}$ e $r \in \{d_{min} + k\Delta_r | k \in \mathbb{N} \text{ e } k\Delta_\alpha \leq d_{max}\}$. Dessa forma, conseguimos encontrar eficientemente os pontos nos quais a sonda deve passar, usando a interseção entre o fecho convexo H e uma reta que parte do seu cen-

tro na direção (θ, φ). Essa interseção pode ser definida usando qualquer algoritmo que resolva o problema *line clipping*, que permite encontrar o ponto de interseção entre uma reta e um poliedro convexo. O algoritmo de Cyrus-Beck resolve o problema em tempo linear pelo número de faces do poliedro [Cyrus & Beck, 1978]. A partir do ponto de interseção com H os pontos das demais camadas na direção (θ, φ) são encontrados incrementando o raio.

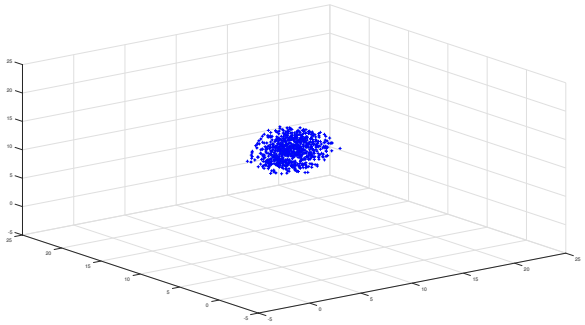
Nessa etapa, consideramos os seguintes hiperparâmetros: Δ_α como o passo de variação dos ângulos θ e φ, Δ_r como o passo de variação da distância radial em r. Também consideramos as constantes $d_{min} = 1,5$Å e $d_{max} = 8$Å, esses valores foram definidos em discussão com Martins et al. [2009], de maneira que não sejam amostrados pontos próximos ou longe demais do PAC.

A amostragem é feita da seguinte maneira, para cada passo Δ_α nos ângulos θ e φ, o primeiro valor de r é definido como o raio da interseção com o fecho convexo na direção (θ, φ) acrescido de uma distância inicial predefinida como d_{min}. Após a definição desse primeiro ponto, os demais pontos são definidos variando o raio r com incremento igual a Δ_r, até que $r \leq d_{max}$. Com essa nova abordagem, conseguimos definir o conjunto de pontos Z que considera o formato de cada composto e respeita as distâncias mínimas e máximas predefinidas, para assim obtermos um Z que nos permite calcular descritores que melhor representam as condições reais do receptor. Isso possibilita construir modelos para prever a atividade biológica com maior precisão.

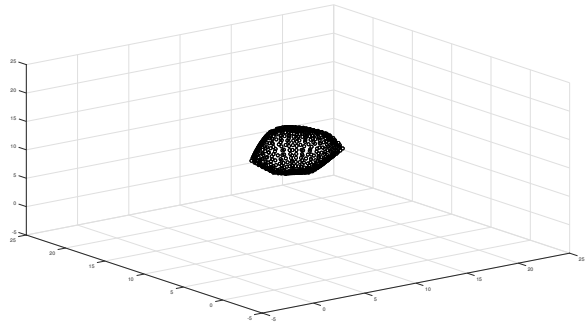
Essa primeira fase de amostragem, consiste apenas em definir os pontos onde devemos calcular os descritores. A Equação 3.1 mostra como obter a quantidade de pontos em Z dado os hiperparâmetros. Podemos observar através dessa Equação que o número de pontos amostrados é inversamente proporcional a $\Delta_r \Delta_\alpha^2$.

$$|Z| = \left\lfloor \frac{\Delta_r + d_{max} - d_{min}}{\Delta_r} \right\rfloor \cdot \left[\left\lfloor \frac{180}{\Delta_\alpha} \right\rfloor \cdot \left\lfloor \frac{360}{\Delta_\alpha} \right\rfloor - 2 \left(\left\lfloor \frac{180}{\Delta_\alpha} \right\rfloor - 1 \right) \right] \tag{3.1}$$

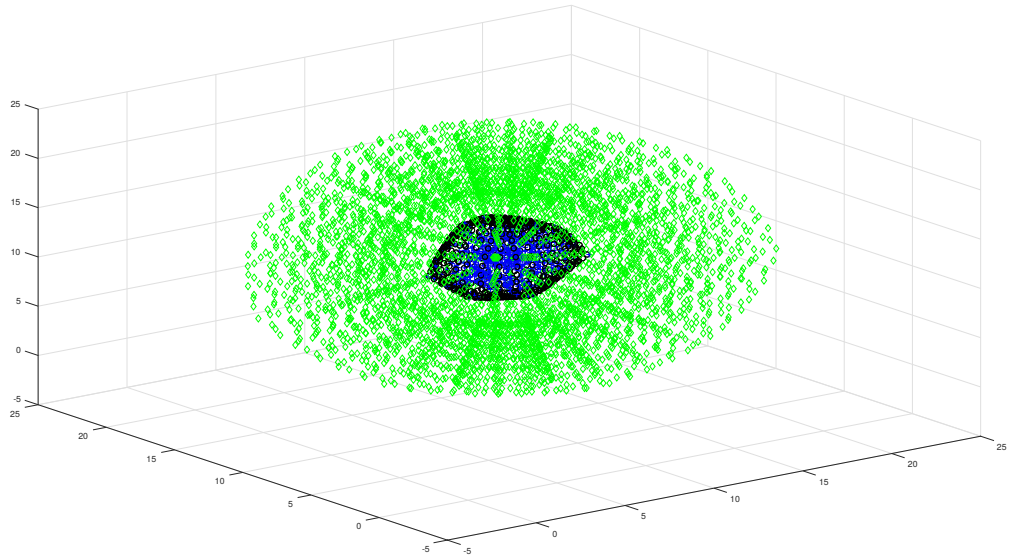
Uma ilustração da região construída pela nova amostragem é apresentada na Figura 3.3. A Subfigura 3.3(a) mostra os pontos que representam os átomos de um PAC. Já a Subfigura 3.3(b) mostra a superfície definida pelo fecho convexo. Por fim, na Subfigura 3.3(c) são mostrados os pontos que a sonda vai percorrer.



(a) Pontos do PAC.



(b) Região de pontos definida pelo fecho convexo.



(c) Pontos onde a sonda deve percorrer.

Figura 3.3: Ilustração do método proposto.

A segunda fase, onde ocorre de fato a geração dos descritores, é a mesma usada por Martins et al. [2009]: para cada ponto em Z, dois descritores de interação são calculados entre o PAC e sonda: Coulomb (C) e Lennard-Jones (LJ). Dessa forma, dado m PACs é gerado um conjunto de dados $X = \{\vec{x_1}, \ldots, \vec{x_m}\}$, onde cada $\vec{x} = (f_1, \ldots, f_n)$ e $n = 2|Z|$ que representa a dimensionalidade de X. O Algoritmo 2 mostra um pseudocódigo da amostragem definida nessa subseção, para um certo PAC μ e fecho convexo H dado os hiperparâmetros e constantes. Como o laço da linha 3 executa de 0° até 359°, o laço mais externo da linha 2 precisa ser executado apenas de 0° até 179°, pois após esse valor todos os pontos que cobrem o PAC já foram contemplados, os demais serão coordenadas repetidas.

Algoritmo 2 amostrarPorFechoConvexo

Entrada: μ, Δ_α, Δ_r, d_{min}, d_{max}, H.
Saída: $\vec{x}$

```
1: x⃗ ← []
2: para i ← 0°, i ← i + Δα até Δα ≤ 179° faça
3:     para j ← 0°, j ← j + Δα até Δα ≤ 359° faça
4:         [Θ, φ, r] ← intersecao(H,[i, j])
5:         k ← r + dmin
6:         enquanto k ≤ r + dmax faça
7:             C ← couloumb(m,[Θ, φ, k])
8:             LJ ← lennardJones(m,[Θ, φ, k])
9:             x⃗ ←+ [C, LJ]
10:            k ← k + Δr
11:        fim enquanto
12:    fim para
13: fim para
```

O método proposto foi desenvolvido para levar em consideração o formato de cada PAC específico, calculando um fecho convexo para cada. Porém também foi testado um método alternativo, onde calculamos um fecho convexo para ser usado em todos os PAC. As duas propostas correspondem a hipóteses extremas: a principal e mais provável é que cada composto possui seu próprio formato e o receptor se adapta à sua forma quando entra em contato. Já a alternativa é que todos os compostos possuem formato similar e o receptor é rígido e não se adapta ao seu formato. Comparar a precisão do modelo de regressão usado os descritores gerados nos dois casos, permite ver qual dessas duas hipóteses extremas pode ser melhor para prever a atividade biológica. A implementação da proposta original, chamada de Fecho Convexo Singular (FCS), é detalha na Subseção 3.1.2. Já o método alternativo, chamado de Fecho Convexo Compartilhado (FCC), é apresentado na Subseção 3.1.3.

3.1.2 Fecho Convexo Singular (FCS)

O método proposto, detalhado na subseção anterior, pode ser implementado através do pseudocódigo apresentado no Algoritmo 3. O método recebe como entrada os PACs $M = \{\mu_1, \ldots, \mu_m\}$ e os hiperparâmetros e constantes. Para cada PAC, esse processo consiste em: encontrar o fecho convexo sobre os pontos do PAC e calcular os descritores usando o Algoritmo 2. No final, os vetores de descritores gerados para cada PAC são retornados como o conjunto de dados $X = \{\vec{x_1}, \ldots, \vec{x_m}\}$.

Algoritmo 3 FCS

Entrada: $M = \{\mu_1, \ldots, \mu_m\}$, Δ_α, Δ_r, d_{min}, d_{max}.
Saída: $X = \{\vec{x_1}, \ldots, \vec{x_m}\}$

1: $X \leftarrow \emptyset$
2: **Para cada** $\mu \in M$ **faça**
3: $\quad H \leftarrow$ calcularFechoConvexo($\mu.P$)
4: $\quad \vec{x} \leftarrow$ amostrarPorFechoConvexo(μ, Δ_α, Δ_r, d_{min}, d_{max}, H)
5: $\quad X \leftarrow X \cup \vec{x}$
6: **fim para**

3.1.3 Fecho Convexo Compartilhado (FCC)

Adicionalmente, nesse trabalho também foi considerada uma implementação alternativa ao método proposto original, onde em vez de calcularmos um fecho convexo para cada PAC, calculamos apenas um sobre os pontos de todos os PACs do conjunto de dados. Essa modificação simplifica a proposta, porém ela só poderia gerar descritores próximos das condições reais, se for válida a hipótese alternativa que todos os PACs possuem o formato mais similar possível e que o receptor é rígido e não se adapta ao PAC, o que pode não ser fácil de se encontrar em dados reais.

A implementação da proposta alternativa é apresentada no Algoritmo 4. Os parâmetros de entrada são o mesmo da proposta original e o comportamento do método diferencia-se apenas pelo fato do fecho convexo H ser calculado em uma etapa anterior usando a união dos pontos de todos os PAC.

Algoritmo 4 FCC

Entrada: $M = \{\mu_1, \ldots, \mu_m\}$, Δ_α, Δ_r, d_{min}, d_{max}.
Saída: $X = \{\vec{x_1}, \ldots, \vec{x_m}\}$

1: $X \leftarrow \emptyset$
2: **Para cada** $\mu \in M$ **faça**
3: $\quad P' \leftarrow P' \cup \mu.P$
4: **fim para**
5: $H \leftarrow$ calcularFechoConvexo(P')
6: **Para cada** $\mu \in M$ **faça**
7: $\quad \vec{x} \leftarrow$ amostrarPorFechoConvexo(μ, Δ_α, Δ_r, d_{min}, d_{max}, H)
8: $\quad X \leftarrow X \cup \vec{x}$
9: **fim para**

3.2 Treinamento de Modelos QSAR usando Árvores de Regressão

Grande parte dos trabalhos encontrados na área de QSAR, incluindo [Martins et al., 2009], usam modelos de regressão linear para a predição da atividade biológica. A vantagem de usar esse tipo de regressão é que se forem usadas boas técnicas para geração, seleção e extração de características, é possível construir modelos simples e com alta precisão, pois o modelo final vai consistir em uma soma ponderada dos valores de cada característica. Regressões lineares são amplamente utilizadas em estudos sobre QSAR, pois a interpretação do modelo de predição é um ponto importante nessa área.

Apesar dos modelos lineares serem interpretáveis, eles não conseguem capturar se existe qualquer outro tipo de relação entre as características do modelo, o que é possível acontecer em problemas reais. Por outro lado, modelos não lineares podem ser mais complexos para interpretação, que é o caso do *Support Vector Regression* [Hastie et al., 2003].

Para contribuir com uma solução para o problema da interpretabilidade, nessa dissertação foi proposto um modelo para predição de atividade biológica usando árvores de regressão, pois esse é um caso de modelo preditivo não linear, que também é fácil de ser interpretado e consegue precisão satisfatória para vários tipos de problemas.

A árvore prevê o valor de atividade biológica após passar por uma sequência de decisões, onde cada decisão tomada depende da precedente. Falta esse aspecto hierárquico nos modelos lineares, pois eles atribuem pesos às características que são sempre usadas na decisão. Se existem razões diferentes que influenciam na atividade biológica em certos grupos de compostos, apenas um modelo linear não é bem adaptado para encontrar esses padrões. Já a árvore pode separar essas razões nos níveis mais próximos à raiz e tratar esse casos em subárvores diferentes.

O modelo adotado nesse trabalho para treinar uma árvore de regressão foi o CART [Breiman et al., 1984], pois esse algoritmo de treinamento já é consolidado no estado da arte. Existem métodos que treinam várias árvores de regressão e combinam o resultado de cada uma para conseguir maior precisão na predição, como o *Random Forest* [Breiman, 2001] e *XGBoost* [Chen & Guestrin, 2016]. Porém esses métodos não foram considerados, porque eles aumentam a precisão em detrimento da interpretabilidade, que é o principal objetivo dessa contribuição.

O restante do fluxo de trabalho usado nessa dissertação foi mantido, a redução de dimensionalidade continua sendo feita com o *Ordered Predictors Selection* (OPS) [Teófilo et al., 2009], pois esse método possibilita encontrar características que evitam

o sobreajuste do modelo. Uma ilustração do fluxo de trabalho proposto é apresentada na Figura 3.4.

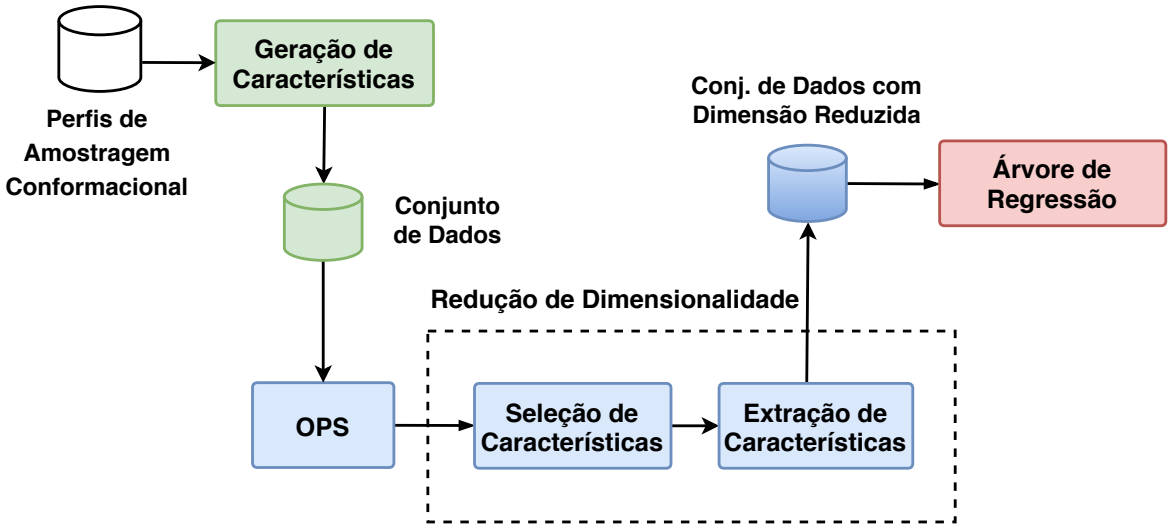


Figura 3.4: Fluxo de trabalho para o modelo QSAR proposto baseado em árvores de regressão.

Capítulo 4

Experimentos e Resultados

Esse Capítulo apresenta os experimentos realizados e uma discussão sobre os resultados obtidos nessa dissertação. A Seção 4.1 apresenta os conjuntos de dados usados para os experimentos. Na Seção 4.2 é definida a metodologia experimental, bem como as métricas de qualidade para avaliar os modelos. Já na Seção 4.3 é discutida a importância de cada hiperparâmetro considerado na amostragem proposta através de um experimento quantitativo. Nas Seções 4.4 e 4.5 são mostrados os experimentos comparativos entre a proposta da dissertação e o método original. A Seção 4.6 apresenta os resultados dos modelos baseados em Árvores de Regressão. Por fim, na Seção 4.7 é feita uma breve discussão dos resultados alcançados com o trabalho, bem como são apresentadas visualizações de dados que podem auxiliar na interpretação dos modelos obtidos.

4.1 Conjuntos de Dados

Os experimentos desse trabalho foram realizados em seis conjuntos de dados contendo PACs derivados de estudos QSAR para problemas relevantes na indústria farmacêutica, cada um desses conjuntos contém de 33 até 81 PACs, onde para cada um existe o valor de atividade biológica real. Esses conjuntos serão apresentados nos parágrafos a seguir.

prostate cancer. Esse conjunto consiste em 49 PACs derivados de benzilideno, oxazolidinediona e tiazolidinedionas com efeito inibitório sobre a enzima 17β-hidroxiesteroide dehidrogenase tipo 3, que são usados para o desenvolvimento de fármacos para o tratamento de câncer de próstata [Harada et al., 2012].

TP. Nesse conjunto temos 37 PACs derivados da uracila com efeito inibitório sobre a enzima timidina fosforilase, proteína frequentemente expressada em muitos tipos de câncer, como bexiga, colorretal, esofágica, pancreática e outros [Yano et al., 2004b]

[Yano et al., 2004a].

HIV. Esse conjunto é composto pelos 33 PACs apresentados nos estudos QSAR realizados por de Melo & Ferreira [2009] sobre derivados de 4,5-dihidroxi-pirimidina carboxamida que agem como inibidores da enzima HIV-1 integrase.

autism. Esse conjunto possui 38 PACs derivados de compostos que interagem com o receptor de serotonina 5-HT-2A. Esses compostos são efetivos na produção de antipsicóticos frequentemente usados no tratamento do autismo [Wilson et al., 2007; Ladduwahetty et al., 2006, 2010].

antimalarial* e *antimalarial+sulfur. Nesses dois conjuntos temos PACs derivados de ureia com atividade contra Plasmodium falciparum, responsável pela malária. O primeiro conjunto possui 48 PACs, enquanto o segundo tem 81. A diferença entre os dois conjuntos é que o segundo é acrescido de compostos que contém enxofre [Madapa et al., 2009].

A Tabela 4.1 apresenta informações gerais sobre os seis conjuntos de dados, como a quantidade de características geradas usando o LQTA-QSAR, que é o *baseline* desse trabalho, bem como a quantidade gerada para cada valor de Δ_α usados nos experimentos com os métodos propostos, que serão justificados nas seções posteriores.

Tabela 4.1: Detalhes dos conjunto de dados gerados com o método proposto e o baseline.

Dataset	PACs	Núm. Características						
		Baseline	$\Delta_\alpha = 2$	$\Delta_\alpha = 3$	$\Delta_\alpha = 4$	$\Delta_\alpha = 6$	$\Delta_\alpha = 8$	$\Delta_\alpha = 10$
antimalarial	48	67518	224308	99148	55468	24388	13272	8596
prostate cancer	49	73440	224308	99148	55468	24388	13272	8596
antimalarial+sulfur	81	124722	224308	99148	55468	24388	13272	8596
TP	37	27600	224308	99148	55468	24388	13272	8596
HIV	33	43680	224308	99148	55468	24388	13272	8596
autism	38	37700	224308	99148	55468	24388	13272	8596

4.2 Metodologia Experimental

Para avaliar a qualidade dos modelos propostos nesse trabalho, foi definida uma metodologia para a execução dos experimentos quantitativos. A Figura 4.1 ilustra uma visão geral do processo adotado. Esse fluxo de trabalho recebe como entrada um conjunto de dados, que pode ser gerado tanto pelos métodos propostos FCS ou FCC quanto pelo método original LQTA-QSAR.

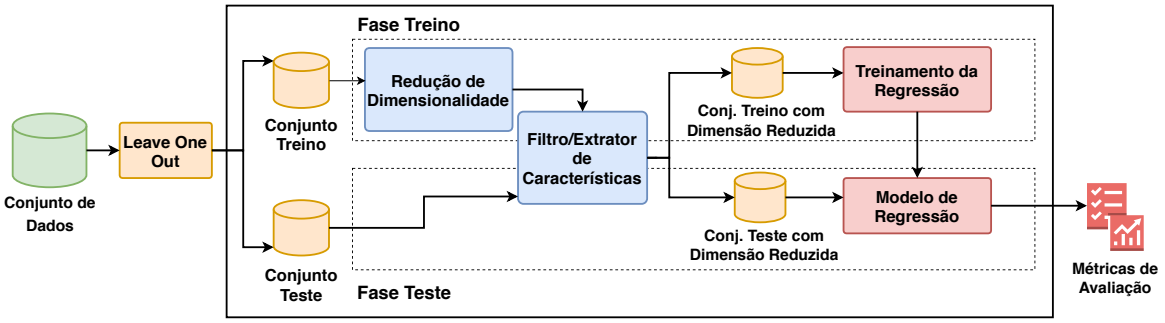


Figura 4.1: Processo de validação cruzada para avaliar modelos QSAR.

Foi usada a técnica de validação cruzada *Leave One Out* (LOO), assim como grande parte dos trabalhos dessa área, inclusive [Martins & Ferreira, 2013]. Em cada iteração, dos m compostos um é separado para a fase de teste e os demais $m-1$ são usados para a fase treino. O processo é executado m vezes, de maneira que cada composto seja avaliado uma vez como conjunto de teste.

A fase de treino consiste em executar um método para redução de dimensionalidade, que gera os filtros e extratores de características, e treinar um modelo de regressão com os dados de treino com dimensão reduzida. Já a fase de teste consiste em aplicar os filtros e extratores de características no composto que foi deixado de lado e prever sua atividade biológica usando o modelo de regressão treinado. No final do processo, para cada um dos m compostos, vamos ter a real atividade biológica do i-ésimo composto y_i e o valor previsto na validação cruzada $\hat{y}_i$. Isso possibilita calcular as métricas de avaliação que são usadas amplamente na literatura para quantificar a qualidade de modelos QSAR. As métricas usadas nesse trabalho recebem como entrada as atividades biológicas reais y e os valores previstos $\hat{y}$.

A principal métrica usada para avaliar e comparar modelos QSAR é o Q^2, que é uma definição pro R^2 quando se usa a validação cruzada (e não os mesmos objetos para treinar e testar). O Q^2 que é calculado de acordo com a Equação 4.1, onde além de y e $\hat{y}_i$, temos m como a quantidade de amostras do conjunto de dados e $\overline{y}$ como a média das atividades biológicas reais y. Roy & Roy [2008] sugere que bons modelos QSAR apresentam valores de Q^2 >0,5.

$$Q^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{m}[y_i - \hat{y}_i]^2}{\sum_{i=1}^{m}[y_i - \overline{y}]^2} \tag{4.1}$$

Além do Q^2, outras duas métricas são calculadas a partir do resultado de teste na validação cruzada: O erro médio quadrado (RMSECV), que é calculada conforme a Equação 4.2 e o coeficiente de correlação (CORRCV), apresentado na Equação 4.3.

$$\text{RMSECV}(y, \hat{y}) = \sqrt{\frac{\sum_{i=1}^{m}[y_i - \hat{y}_i]^2}{m}} \tag{4.2}$$

$$\text{CORRCV}(y, \hat{y}) = \frac{\sum_{i=1}^{m}[y_i - \hat{y}_i] \cdot [y_i - \hat{y}_i]}{\sqrt{\sum_{i=1}^{m}[y_i - \hat{y}_i]^2} \cdot \sqrt{\sum_{i=1}^{m}[y_i - \hat{y}_i]^2}} \tag{4.3}$$

Adicionalmente, também foram calculadas métricas sobre a fase de treinamento dos modelos, passando os valores previstos $\hat{y}$ na fase de treino, apesar de não serem usadas para avaliar e comparar modelos QSAR, elas podem agregar a análise. Considerando $\hat{y}_i$ como o valor previsto para o i-ésimo composto na fase de treino, temos as seguintes métricas para avaliar a qualidade do treinamento: o R^2, que corresponde ao Q^2, RMSE correspondendo ao RMSECV e por fim, CORR representando o CORRCV. Como no LOO cada composto participa do conjunto de treino em $m - 1$ iterações, essas métricas são calculadas na fase de treino em todas m interações e o valor final corresponde a média dos valores obtidos.

Também é necessário definir os hiperparâmetros do *O Ordered Predictors Selection* (OPS), que foi usado para redução de dimensionalidade. Para cada um dos seis conjuntos de dados, os hiperparâmetros do OPS foram definidos de acordo com a Tabela 4.2, considerando o tempo de execução e particularidades de cada conjunto. O parâmetro maxNV foi fixado em 250 para todos os casos, já ρ foi definido para cada conjunto observando a distribuição das correlações entre as características e a atividade biológica real. Foi selecionado o valor de corte onde o conjunto filtrado tivesse tamanho de no mínimo maxNV, o valor de 0,3 permitiu que essa condição fosse satisfeita para quase todos os conjuntos, nos demais 0,2 foi suficiente. Por fim, os parâmetros maxNVL e maxOVL foram definidos respectivamente como $m/5$ e $m/2$, onde m é o número de compostos do conjunto. Eles foram escolhidos como frações dessa dimensão porque o número de compostos é bem menor que o de características, isso permite extrair as variáveis latentes em tempos aceitáveis.

Tabela 4.2: Parâmetros utilizados no treinamento do OPS.

Parâmetros OPS	ρ	maxNV	maxNVL	maxOVL
antimalarial	0,3	250	9	24
prostate cancer	0,2	250	9	24
TP	0,3	250	7	18
antimalarial+sulfur	0,2	250	16	40
HIV	0,3	250	7	17
autism	0,3	250	8	19

Em todos os métodos propostos para geração de características, a sonda usada foi NH_3^+ assim como em [Martins et al., 2009]. Já para os método propostos, Δ_r foi fixado em 1 e Δ_α, devido a sua relevância, foi variado em cada experimento nos valores $\Delta_\alpha = \{2, 3, 4, 6, 8, 10\}$, para avaliarmos qual é o mais adequado para cada conjunto de dados específico. A Seção seguinte apresenta um experimento para justificar a forma como esses hiperparâmetros foram escolhidos.

4.3 Escolhendo os Hiperparâmetros para a Amostragem Proposta

Os métodos propostos foram construídos para amostrar os descritores com base em dois hiperparâmetros: Δ_α, que está relacionado ao passo nos ângulos e Δ_r que é o passo no raio r. Para avaliar a relevância desses hiperparâmetros, foi realizado um experimento com o objetivo de responder a seguinte questão: Em quais dimensões devemos amostrar um maior número de pontos em relação as demais?.

O primeiro passo foi executar o método proposto FCS no conjunto de PACs "HIV", para analisarmos a variação dos descritores C e LJ em cada uma das dimensões: θ, φ e r. Definimos $\Delta_\alpha = 6$ e $\Delta_r = 0.1166$, de maneira que sejam definidas 60 camadas ao redor do PAC com 60 coordenadas φ diferentes e 30 em θ, pelo fato de em θ ser necessário amostrar apenas de 0° até 179°. Dessa forma, essa escolha permite que seja feita uma comparação na qual em cada dimensão a quantidade de pontos seja o mais próximo possível das demais.

O segundo passo foi realizar um esboço dos valores de um tipo de interação em função de uma das três coordenadas, com as duas outras fixadas. O número de extremos locais na curva esboçada informa sobre a necessidade de um passo maior ou menor na enumeração da coordenada para a amostragem não esconder variações que podem ser relevantes. Esse processo é executado para todos os possíveis pares de valores das dimensões fixadas e no final é retornado a média do número extremos locais

encontrados para cada descritor na dimensão restante. Exemplos das curvas esboçadas em um PAC são mostrados na Figura 4.3. As Subfiguras da coluna esquerda são relacionadas aos descritores LJ. Por outro lado, as Subfiguras da coluna direta são referentes aos descritores C. A variável l_r foi usada para possibilitar o entendimento dos gráficos, ela representa o índice dos pontos gerados ao variar o raio r, por exemplo $l_r = 8$ representa a oitava camada entre as 60 que foram geradas nesse experimento.

Podemos observar qualitativamente através da Figura 4.3, que ao variar um ângulo existem mais variações nos valores dos descritores, em comparação com a distância radial. Isso também é corroborado quantitativamente nas médias do número de extremos locais apresentadas na Figura 4.2, onde cada barra representa a média calculada usando todos os possíveis pares das demais dimensões. Esses resultados mostram que ao usar valores baixos de Δ_α conseguimos capturar mais informação do que usando valores baixos de Δ_r. Isso implica que Δ_α pode ter maior impacto na quantidade de informação capturada em comparação com Δ_r. Também podemos concluir através dos resultados que amostrar uma quantidade maior de pontos nos ângulos, ou seja, com Δ_α baixo, pode ser mais relevante do que amostrar valores com Δ_r baixo.

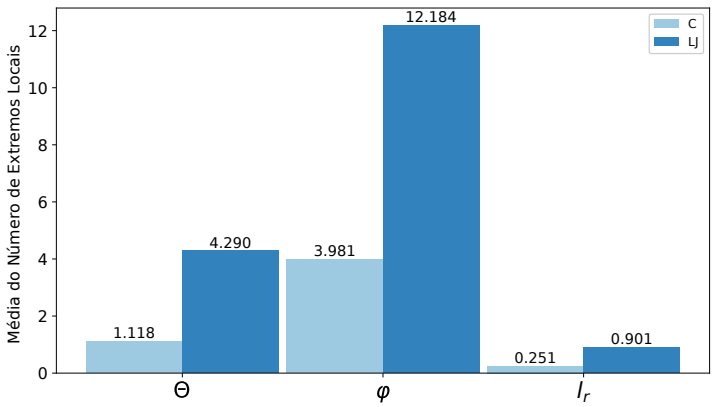
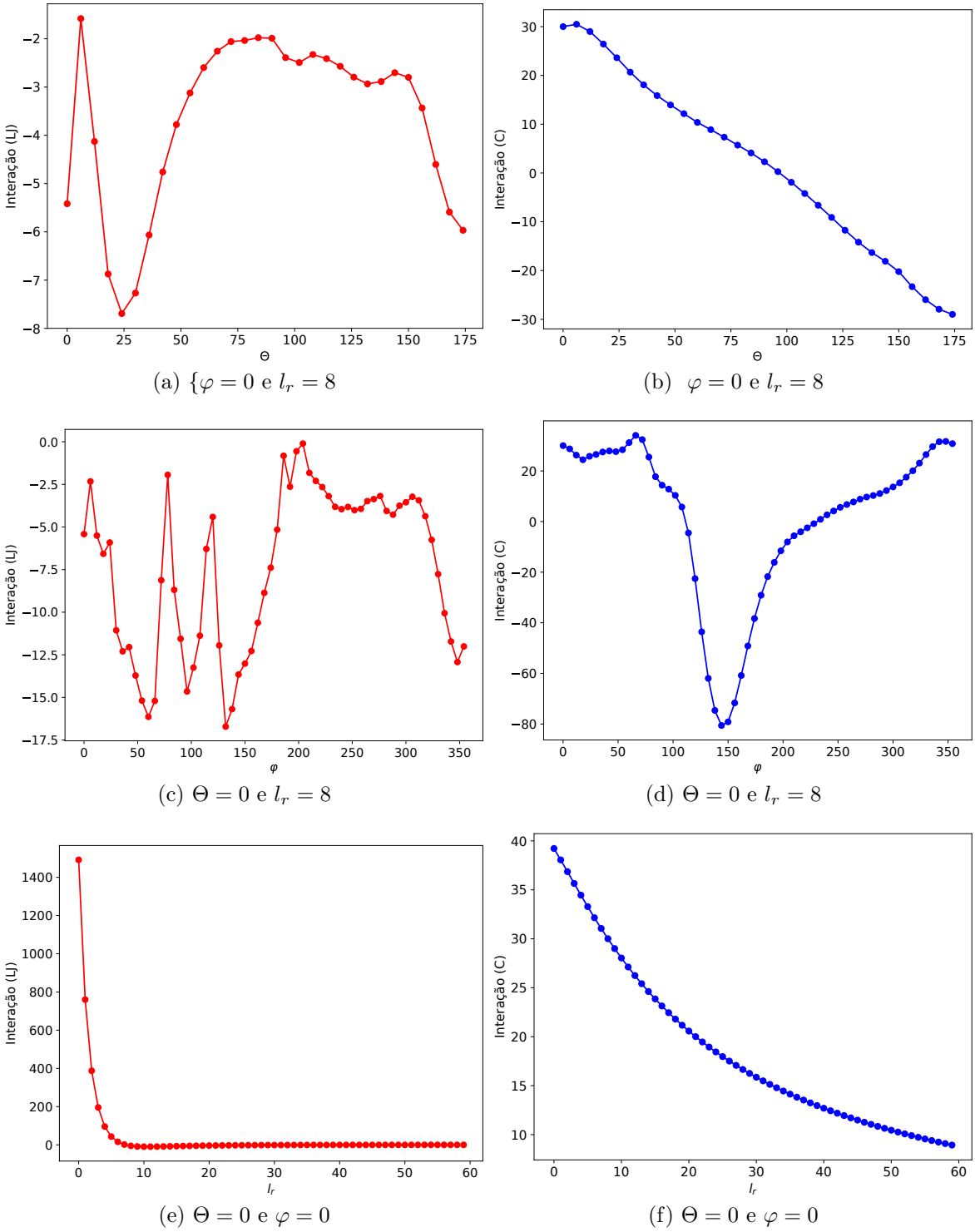


Figura 4.2: Média do número de extremos locais com os descritores de interação em cada dimensão.



(a) $\{\varphi = 0$ e $l_r = 8$ (b) $\varphi = 0$ e $l_r = 8$

(c) $\Theta = 0$ e $l_r = 8$ (d) $\Theta = 0$ e $l_r = 8$

(e) $\Theta = 0$ e $\varphi = 0$ (f) $\Theta = 0$ e $\varphi = 0$

Figura 4.3: Variação das interações ao longo de cada dimensão.

4.4 Comparativo entre FCS e FCC

Foram realizados experimentos comparativos, seguindo a metodologia descrita na Seção 4.2, para comparar as duas possibilidades de implementação da proposta: usar um fecho convexo para cada PAC (FCS), ou usar um fecho convexo para todos os PACs (FCC). Para cada conjunto de PACs, os dois métodos foram usados para gerar conjuntos de dados variando os hiperparâmetros Δ_α, conforme definido na metodologia experimental. Em seguida, para cada conjunto de dados, foi calculado o valor de Q^2 usando o processo de validação cruzada com o PLS.

Os resultados para cada um dos seis conjuntos de PACs são apresentados na Figura 4.4. O método FCS conseguiu construir modelos com valores de Q^2 muito maiores que o FCC em todos os conjuntos de dados avaliados. Pois em problemas reais dificilmente conseguimos validar a hipótese alternativa de que o receptor é rígido e que todos os PACs vão possuir o mesmo formato, ou bastante similar.

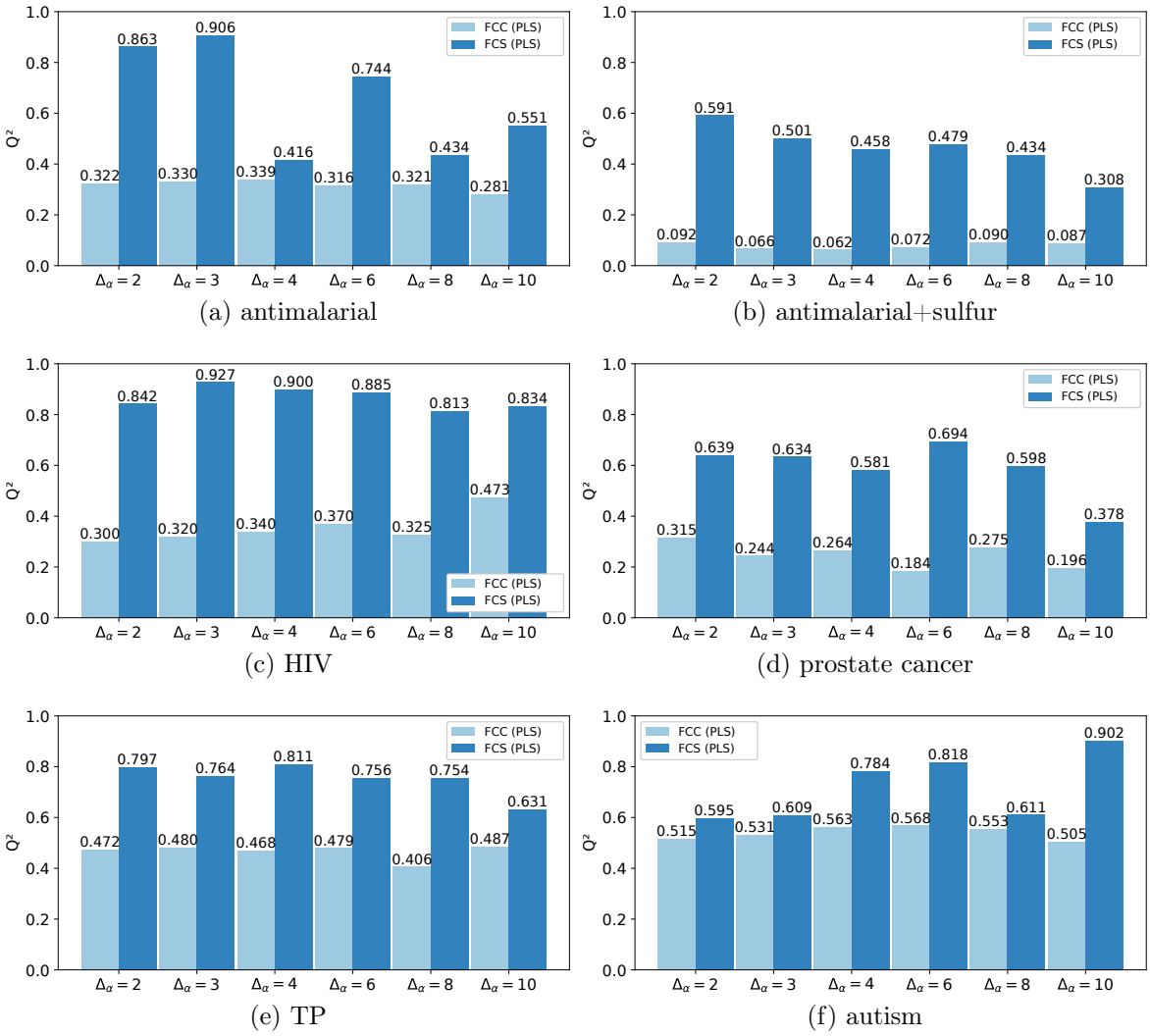
O único caso onde o FCC conseguiu modelos com $Q^2 > 0.5$ foi no conjunto autism. Embora esses resultados sejam inferiores aos obtidos com o FCS, isso nos fornece evidências que para casos específicos, a abordagem FCC pode produzir resultados válidos. Outro ponto que pode ser levado em consideração é que nos casos onde o FCC obtiver resultados razoáveis, isso pode nos fornecer evidências que a hipótese alternativa pode ser válida para o conjunto PACs estudado.

Como nesse experimento conseguimos corroborar quantitativamente a hipótese que de o método FCS é mais adequado para resolver o problema de forma genérica do que o FCC, nos demais experimentos, o FCS será adotado como proposta para geração de características.

4.5 Comparativo entre LQTA-QSAR e FCS

Anteriormente validamos que o método proposto FCS é mais adequado para resolver o problema. Agora será feito um experimento comparativo entre o FCS e método original LQTA-QSAR [Martins et al., 2009], usado como *baseline* dessa dissertação. Esses experimentos também seguiram a metodologia adotada: foram gerados conjuntos de dados com cada uma das abordagens para cada um dos conjuntos de PACs e em seguida obtivemos os valores de Q^2 usando o processo de validação cruzada com o PLS.

Os resultados obtidos nesse experimentos são ilustrados na Figura 4.5 para cada um dos conjuntos de PACs. Em todos os casos, o método proposto conseguiu obter valores de Q^2 superiores ao *baseline*, em pelo menos um dos valores de Δ_α testados. No conjunto HIV, o melhor modelo obtido com o método proposto (com $\Delta_\alpha = 3$), obteve



(a) antimalarial (b) antimalarial+sulfur

(c) HIV (d) prostate cancer

(e) TP (f) autism

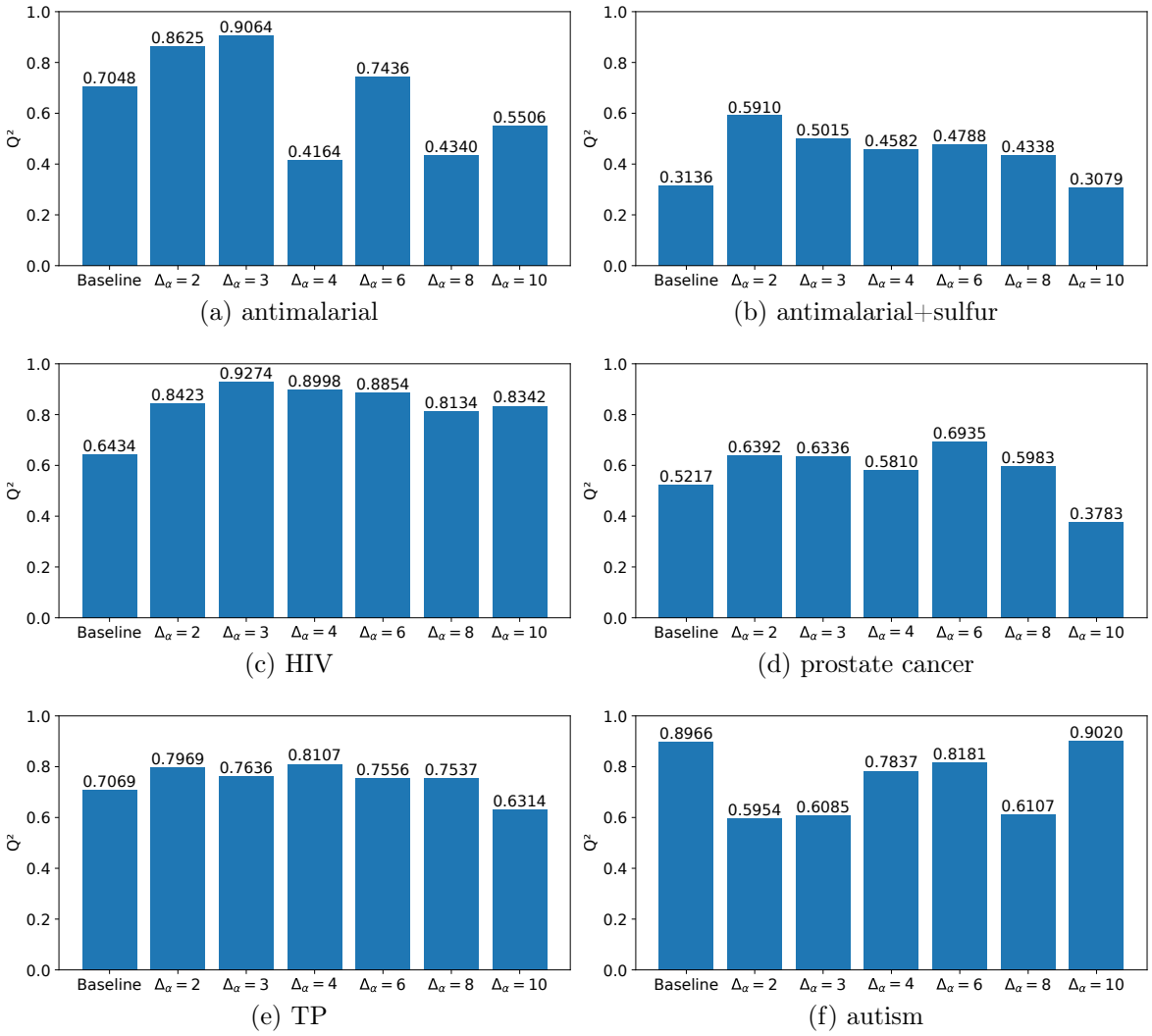
Figura 4.4: Experimentos comparativos entre FCS e FCC.

valor de Q^2 superior em 0,284. Esse melhoria é bastante significativa, considerando a complexidade do problema em questão.

Esses resultados nos fornecem evidências quantitativas de que as características geradas com o método proposto conseguem construir modelos QSAR com melhor qualidade que o método original LQTA-QSAR. Além de que ao contrário do LQTA-QSAR, a proposta considera o formato dos compostos, o que permite gerar modelos que fazem mais sentido para interpretação, como já foi discutido no Capítulo 3.

Além de avaliar os valores de Q^2 nesse experimento comparativo, os modelos com melhores resultados usando o método proposto também foram avaliados com as demais métricas adotadas nesse trabalho, com o objetivo de obter uma avaliação mais detalhada sobre esses resultados.

Na Tabela 4.3 temos para cada conjunto de PACs, os valores de Δ_α onde foram



(a) antimalarial (b) antimalarial+sulfur

(c) HIV (d) prostate cancer

(e) TP (f) autism

Figura 4.5: Experimentos comparativos entre método proposto e baseline.

obtidos os melhores resultados, bem como os hiperparâmetros de saída do OPS, que nos permite interpretar melhor os modelos que estão sendo construídos. Como exemplo para o conjunto antimalarial, o melhor modelo foi obtido com 97 características e a partir delas foram extraídas 6 variáveis latentes, que são a entrada para o treinamento do modelo de regressão.

Tabela 4.3: Hiperparâmetros dos melhores modelos encontrados com o método proposto e o OPS em cada conjunto de dados.

Dataset	Δ_α	NV	NVL	OVL	IV
antimalarial	3	97	6	7	2
prostate cancer	6	43	5	11	3
TP	4	113	7	16	2
antimalarial+sulfur	2	249	4	29	2
HIV	3	249	7	8	3
autism	10	138	8	12	2

Os resultados com as demais métricas de avaliação são apresentados na Tabela 4.4. É possível observar que em todos os casos, esses valores são coerentes com os resultados obtidos apenas observando o Q^2. Pois é esperado que as métricas relacionadas a fase de treino tenham melhores valores em relação as da fase de teste, porém é desejável que elas não tenham valores de $R^2 = 1$ ou RMSE $= 0$ no treino, porque isso pode ser um indicativo de que os modelos estão sendo impactados com o sobreajuste ao conjunto de treino. Como podemos observar, isso não ocorre nos modelos obtidos.

Tabela 4.4: Medidas de avaliação dos melhores modelos com o PLS em cada conjunto de dados.

Dataset	Q^2	R^2	RMSECV	RMSE	CORRCV	CORR
antimalarial	0,9064	0,9577	0,0422	0,0190	0,9526	0,9787
prostate cancer	0,6935	0,9157	0,1524	0,0419	0,8570	0,9569
TP	0,8107	0,9869	0,2652	0,0183	0,9013	0,9934
antimalarial+sulfur	0,5910	0,8107	0,1089	0,0504	0,7887	0,9004
HIV	0,9274	0,9980	0,0874	0,0024	0,9630	0,9990
autism	0,9020	0,9972	0,0683	0,0020	0,9517	0,9986

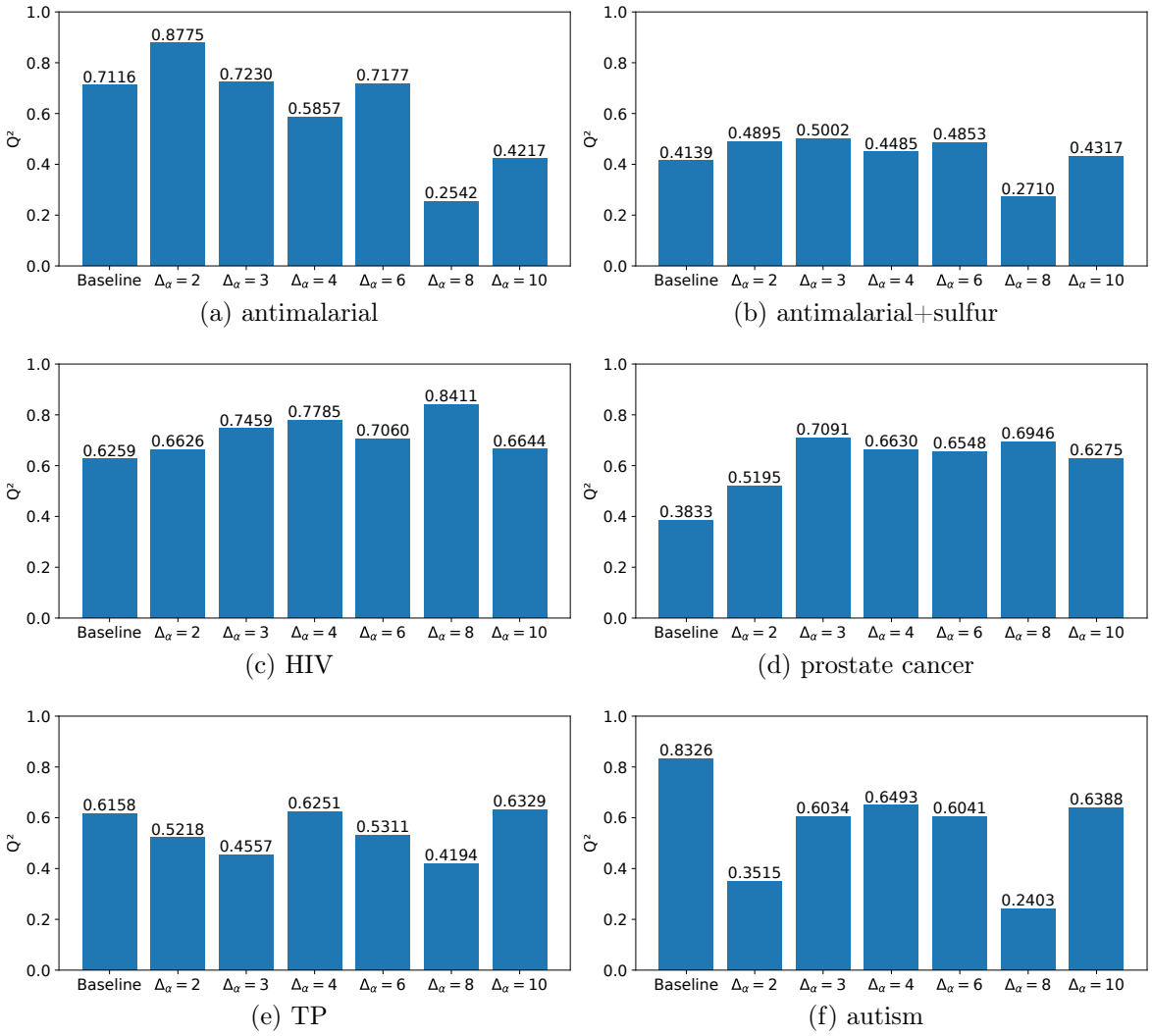
4.6 Experimentos usando Árvores de Regressão

Uma das propostas dessa dissertação foi avaliar um fluxo de trabalho alternativo que usa árvore de regressão como modelo preditivo. Nesse experimento, incluímos o treinamento da árvore de regressão na saída do OPS e avaliamos os valores de Q^2, seguindo a metodologia experimental adotada, para os dados gerados tanto com o LQTA-QSAR quanto com o FCS.

Para o treinamento da árvore de regressão, também foi necessário ajustar o parâmetro "minSamplesLeaf", que indica o número mínimo necessário de exemplos que devem alcançar um nó da árvore para que ele seja definido com uma folha da árvore.

Esse parâmetro foi variado de 1 até o total de exemplos dos conjuntos de dados e foi considerado o que obtivemos valor máximo de Q^2.

A Figura 4.6 mostra os resultados obtidos nesse experimento para todos os conjuntos de dados adotados nesse trabalho. Podemos observar que no geral os valores de Q^2 foram satisfatórios, pois atenderam o requisito de $Q^2 > 0.5$ e se aproximaram dos resultados obtidos usando o modelo linear. Os modelos construídos com o FCS também conseguiram valores de Q^2 superiores ao *baseline* em pelo menos um dos valores de Δ_α testados, com exceção apenas do conjunto autism, onde o modelo construído com o *baseline* superou os demais com o método proposto.

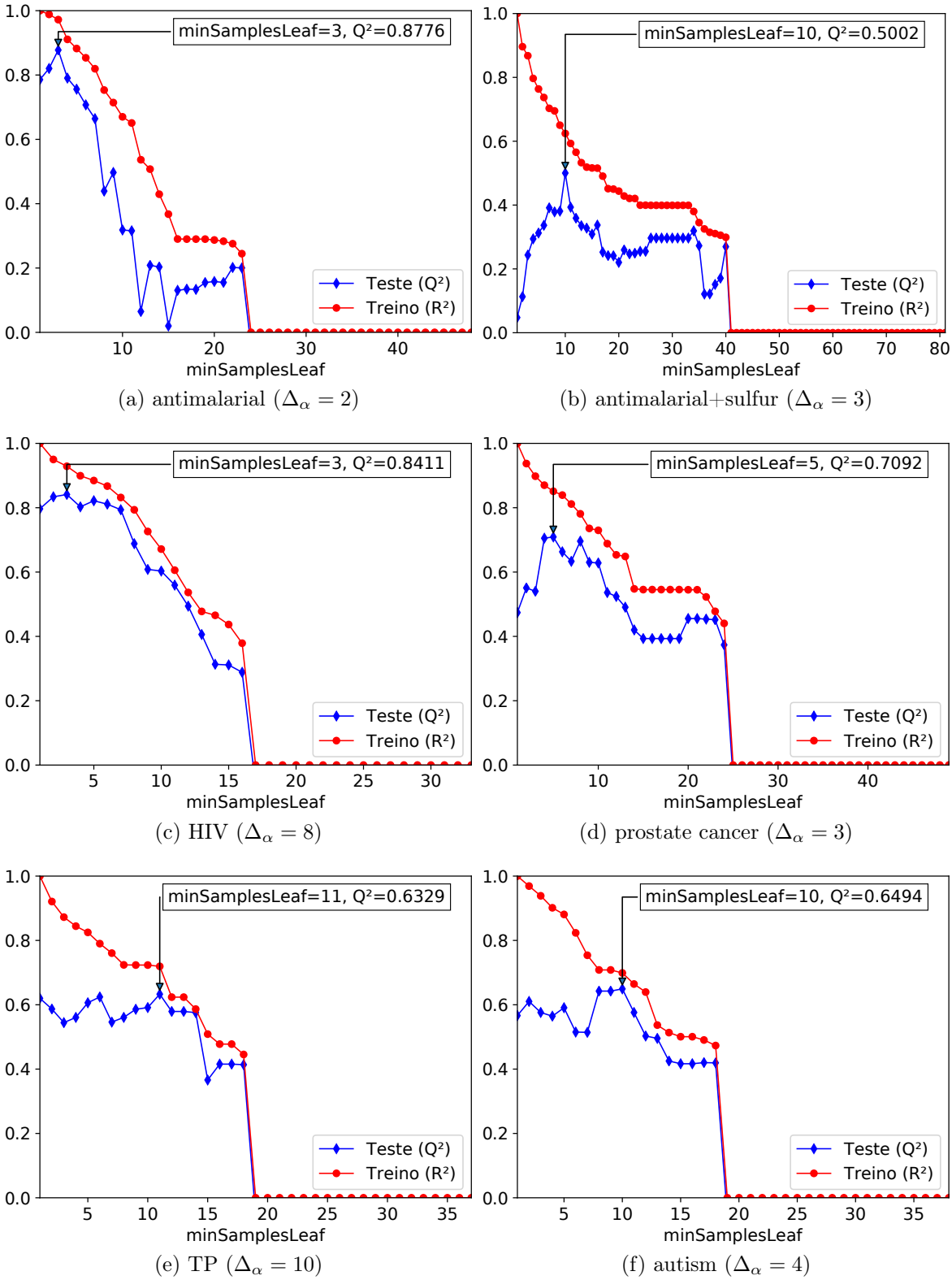


(a) antimalarial (b) antimalarial+sulfur

(c) HIV (d) prostate cancer

(e) TP (f) autism

Figura 4.6: Experimentos comparativos usando árvore de regressão.



(a) antimalarial ($\Delta_\alpha = 2$)

(b) antimalarial+sulfur ($\Delta_\alpha = 3$)

(c) HIV ($\Delta_\alpha = 8$)

(d) prostate cancer ($\Delta_\alpha = 3$)

(e) TP ($\Delta_\alpha = 10$)

(f) autism ($\Delta_\alpha = 4$)

Figura 4.7: Ajuste do hiperparâmetro "minSamplesLeaf" da árvore de regressão usando os modelos com a melhor configuração de Δ_α encontrada.

Para possibilitar uma interpretação melhor desses resultados, na Figura 4.7 são apresentados os valores de Q^2 (representando a fase de teste) e de R^2 (representando a fase de treino), para cada valor de "minSamplesLeaf" nos melhores modelos obtidos em cada conjunto de dados.

Podemos observar que quando "minSamplesLeaf" é testado para o valor 1, obtemos valores de $R^2 = 1$ com Q^2 baixos, a medida que vamos incrementado o valor de "minSamplesLeaf", os valores de Q^2 aumentam enquanto os de R^2 diminuem, até convergir em um valor próximo. Os valores de Q^2 e R^2 são aceitáveis até "minSamplesLeaf" assumir valores próximos a metade do conjunto de dados, maior que isso os modelos construídos possuem desempenho desprezível.

Com esses resultados, temos evidências quantitativas de que a árvore de regressão pode ser considerada na construção de modelos para predição de atividade. Sendo uma alternativa interessante ao modelo linear, que nos permite explorar novas possibilidades de interpretação dos modelos de preditivos.

Assim como nos demais experimentos comparativos, também avaliamos os melhores modelos obtidos usando árvore de regressão com as demais métricas. Os resultados são apresentados na Tabela 4.5. Podemos observar que assim como nos modelos obtidos com o PLS, os valores das demais métricas também são satisfatórios.

Tabela 4.5: Medidas de avaliação dos melhores modelos com a árvore de regressão em cada conjunto de dados.

Dataset	Q^2	R^2	RMSECV	RMSE	CORRCV	CORR
antimalarial	0,8775	0,9722	0,0553	0,0125	0,9399	0,9860
prostate cancer	0,7091	0,8511	0,1447	0,0740	0,8465	0,9225
TP	0,6329	0,7194	0,5145	0,3929	0,7983	0,8482
antimalarial+sulfur	0,5002	0,6243	0,1331	0,1000	0,7114	0,7901
HIV	0,8411	0,9288	0,1913	0,0851	0,9202	0,9637
autism	0,6493	0,6989	0,2444	0,2095	0,8065	0,8359

4.7 Interpretação e Visualização de Dados

Nos experimentos apresentados anteriormente, avaliamos o método proposto para geração de características com diversos conjuntos de dados e diferentes modelos de regressão, seguindo a metodologia de experimentação quantitativa adotada. Os resultados dos melhores modelos obtidos com o FCS usando tanto PLS quanto a árvore de regressão são mostrados na Figura 4.8, bem como os resultados obtidos método original LQTA-QSAR.

Podemos observar que os modelos construídos usando o FCS e PLS obtiveram maiores valores de Q^2 em cinco dos seis conjuntos de dados avaliados. O único conjunto onde o modelo com FCS e PLS não conseguiu obter maior Q^2 foi o prostate cancer, que nesse caso, o maior valor foi obtido usando o FCS com árvore de regressão.

Esses resultados nos mostram que ao usarmos o método proposto FCS em vez do LQTA-QSAR para gerar características, conseguimos construir modelos mais precisos, mesmo usando outras técnicas de regressão além do PLS. Isso nos possibilita realizar estudos QSAR de diferentes perspectivas.

Os modelos construídos usando o FCS e árvore de regressão conseguiram resultados superiores aos demais em apenas um conjunto de dados. Porém nos outros conjuntos de dados, os resultados foram próximos aos obtidos usando o PLS. A maior contribuição ao usar árvores de regressão é a facilidade de fazer interpretações mais amplas sobre a predição.

Considerando a facilidade de interpretação das árvores de regressão, além dos experimentos quantitativos, também foram feitas algumas visualizações de dados para possibilitar uma avaliação qualitativa mais detalhada dos modelos.

O primeiro passo para interpretar um modelo QSAR, é observar quais características foram selecionadas para construir o modelo de regressão e em quais posições do receptor elas se encontram. Para possibilitar essa análise, foi construída uma visualização usando o software Chimera [Pettersen et al., 2004], onde foram esboçados os átomos do PAC de uma molécula específica e as coordenadas dos descritores selecionados ao seu redor. Essa visualização permite identificar quais átomos estão próximos dos pontos selecionados e interpretar se esse modelo faz sentido biologicamente.

Dois exemplos dessa visualização são mostrados na Figura 4.9. Onde usamos os descritores selecionados no melhor modelo do conjunto de dados HIV para esboçar essa visualização. A Subfigura 4.9(a) ilustra essa visualização para ao PAC de maior atividade biológica do conjunto HIV. Por outro lado, a Subfigura 4.9(b) mostra a mesma visualização para o PAC de menor atividade biológica desse mesmo conjunto.

Outra particularidade que conseguimos observar através dessa visualização, é que os descritores selecionados pertencem a regiões específicas do receptor, formando alguns pequenos grupos. Isso permite que seja feita uma avaliação mais detalhada sobre a influência dessas regiões na atividade biológica em condições reais.

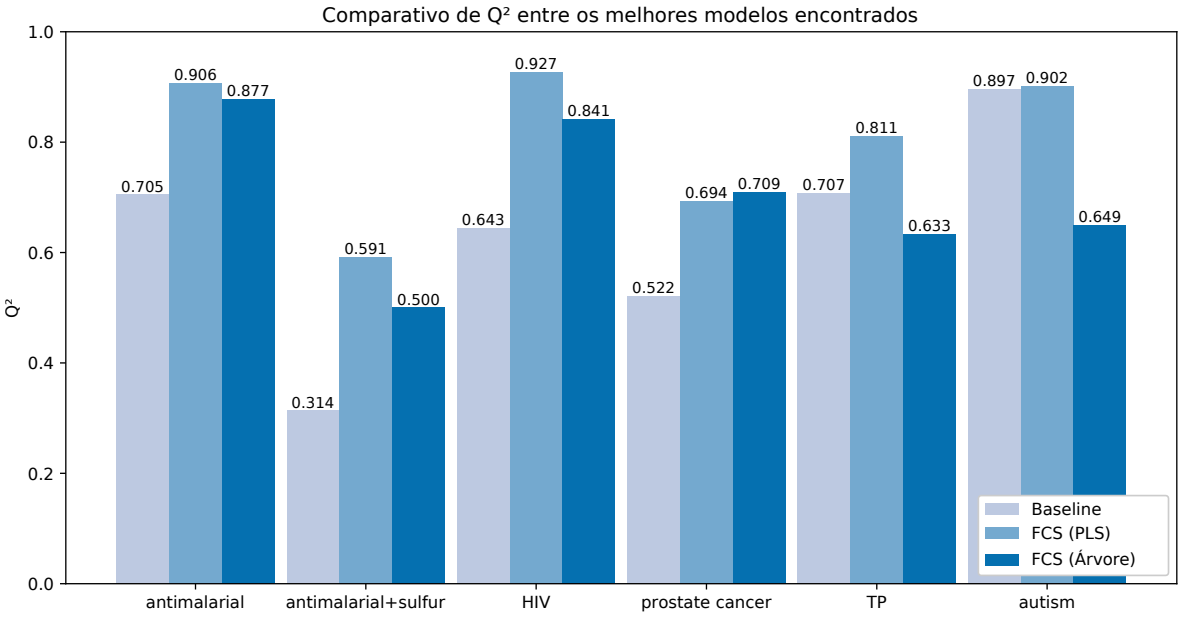
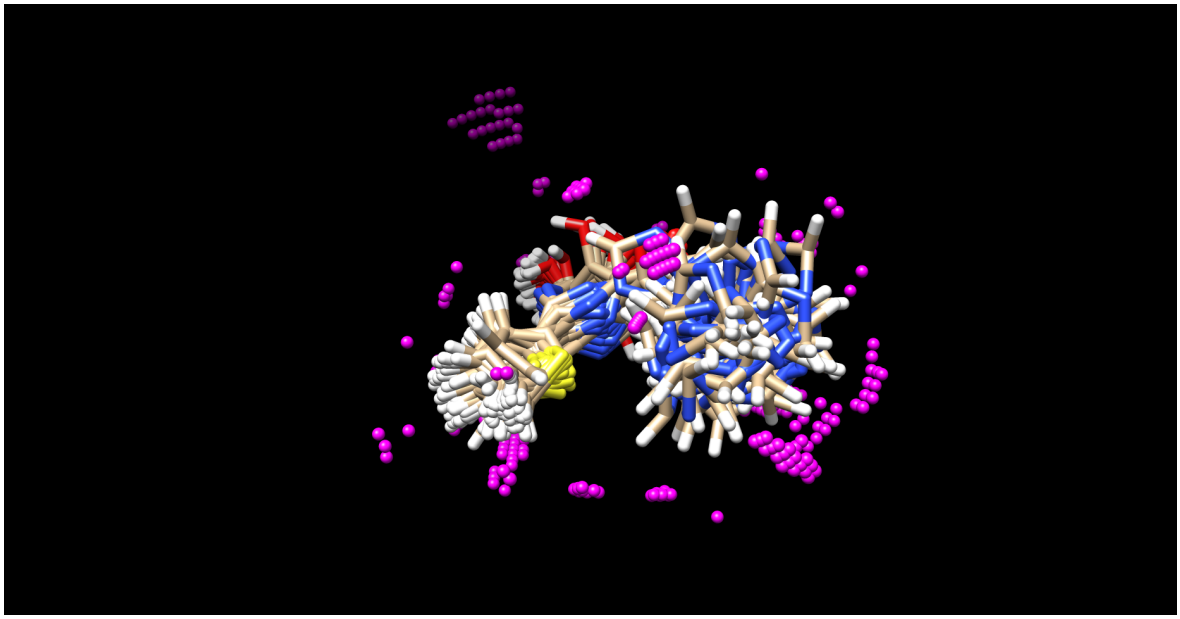
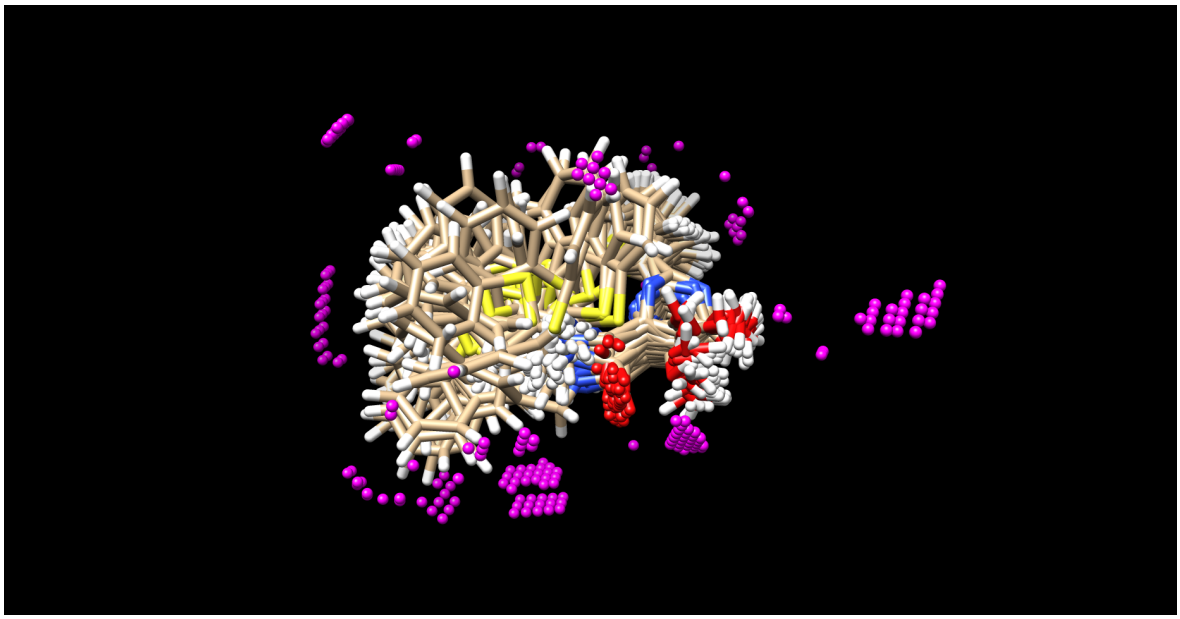


Figura 4.8: Comparativo de Q^2



(a) Molécula de menor atividade



(b) Molécula de maior atividade

Figura 4.9: Visualização dos descritores selecionados no melhor modelo da base HIV, usando as moléculas de menor e maior atividade biológica.

A próxima etapa da interpretação, é visualizar a árvore de regressão construída usando as variáveis latentes geradas com as características selecionadas. Para construir essa visualização, selecionamos os parâmetros do melhor modelo no conjunto HIV e treinamos uma árvore usando todos os PACs desse conjunto. A Figura 4.10 mostra a árvore treinada, para cada nó temos a regra de decisão, o número de exemplos que alcançam esse nó e a média das atividades biológicas desse exemplos. A intensidade da cor do nó está associada a média da atividade biológica nesse nó, quanto maior mais intensa será a cor. Através dessa visualização, conseguimos visualizar regras que não são possíveis de serem extraídas nos modelos lineares. Por exemplo, se desejamos interpretar o porquê de um valor ter sido previsto para um PAC específico, basta observar o caminho que foi percorrido na árvore a partir da raiz até a folha.

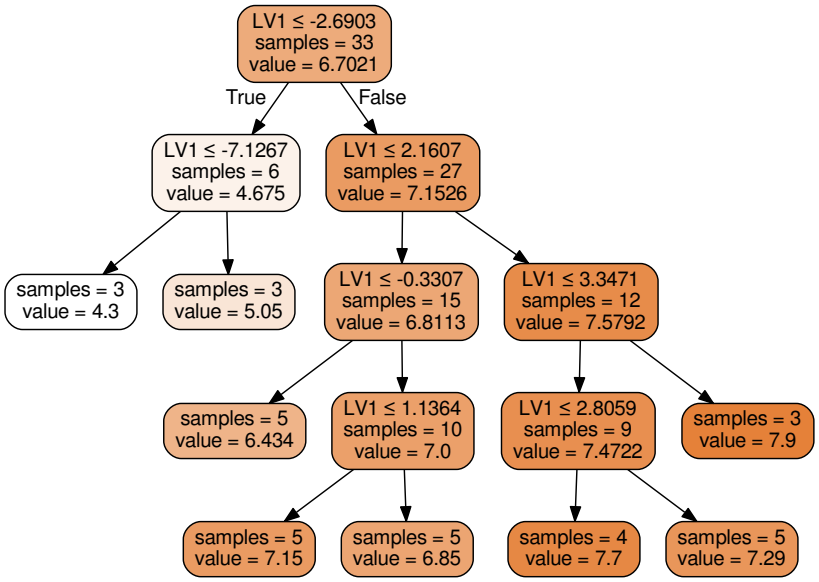


Figura 4.10: Árvore de regressão na base HIV.

Como o melhor modelo encontrado para o conjunto de dados HIV possui apenas uma variável latente, é possível visualização a função de regressão em um gráfico de dispersão. Onde cada PAC é esboçado com as seguintes coordenadas: no eixo das abcissas temos o valor da variável latente para esse PAC e no eixo das ordenadas temos a atividade biológica associada a esse PAC. Em seguida, a função de regressão entre a variável latente e atividade biológica é esboçada.

Com essa visualização podemos observar que a função da árvore de regressão consegue capturar padrões que não são possíveis de serem obtidos em modelos lineares. Esses padrões podem trazer informações relevantes sobre o conjunto de dados analisado e também podem ser usados como complemento de uma análise com o modelo linear.

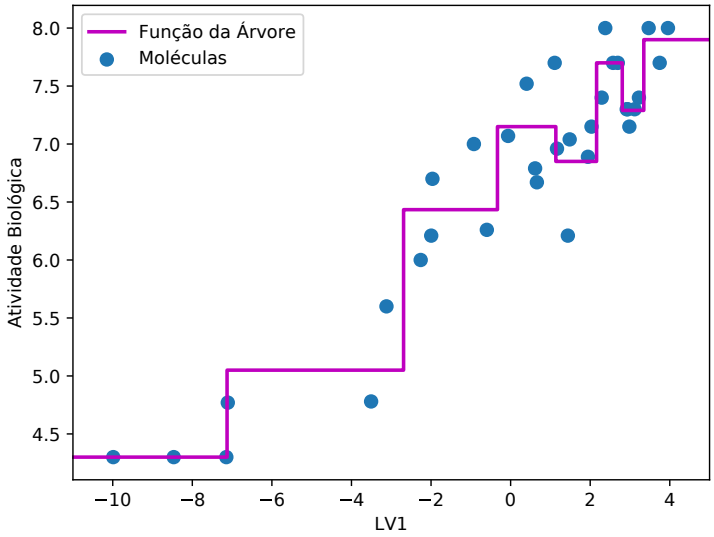


Figura 4.11: Visualização da função de regressão na base HIV.

Adicionalmente, também foi feita a mesma visualização para o conjunto prostate cancer, que foi onde o modelo usando árvore de regressão consegui ter desempenho superior aos demais comparados. A Figura 4.12 mostra a visualização gerada.

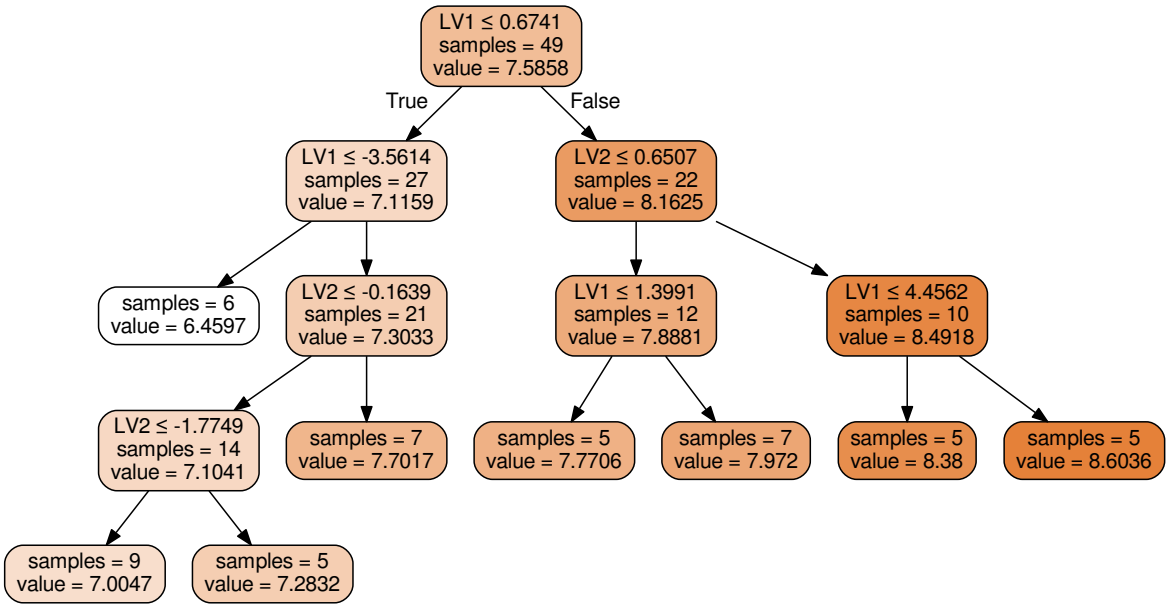


Figura 4.12: Árvore de regressão na base prostate cancer.

O melhor modelo obtido nesse conjunto de dados possui duas variáveis latentes, diferente do modelo anterior que possuía apenas uma. Nesse caso, foi construída uma visualização mais elaborada do que o gráfico de dispersão para visualizar a função de regressão, inspirada em um mapa de calor. A Figura 4.13 mostra a visualização gerada, onde no eixo das abcissas temos a primeira variável latente e no eixo das ordenadas

temos a segunda. As regiões formadas pela função de regressão são preenchidas com cor cuja a intensidade também é associada ao valor da atividade biológica que é ajustado naquela região.

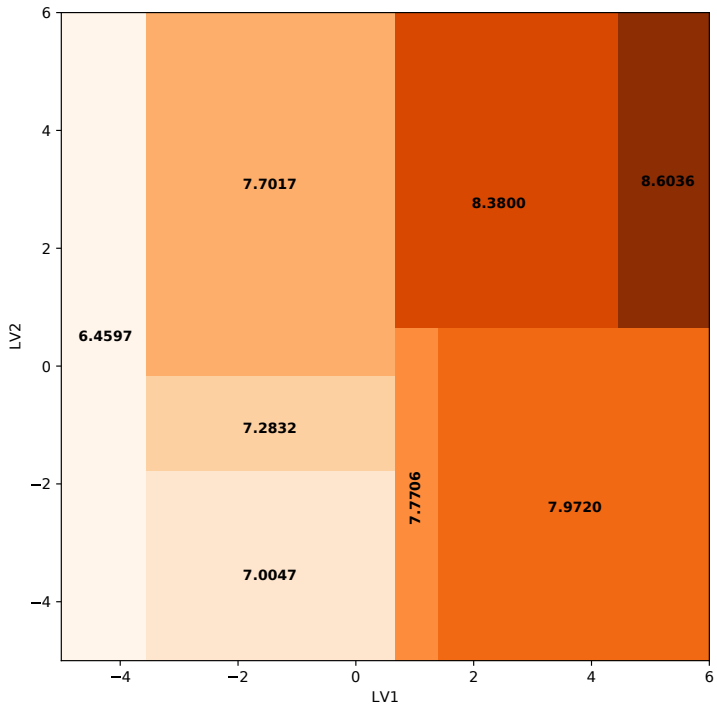


Figura 4.13: Visualização da árvore de regressão na base prostate cancer.

Essas visualizações apresentadas possibilitam novos tipos de interpretações que podem agregar valor ao estudos de QSAR, servindo como apoio às demais análises com modelos lineares já consolidadas na área.

Capítulo 5

Conclusão

Essa dissertação propôs uma nova amostragem de descritores para métodos 4D-QSAR que usa geometria computacional para considerar no processo a posição espacial e formato do perfil de amostragem conformacional (PAC). A proposta faz ampliações da superfície definida pelo fecho convexo para construir camadas ao redor do PAC, que representam os pontos onde a sonda (que pode ser um átomo, íon ou grupo funcional) deve percorrer para calcular os descritores de interação. Essa amostragem impede que a sonda passe por pontos internos ou próximos demais ao PAC.

A maior contribuição dessa proposta é que o conjunto de descritores amostrados possibilita construir modelos preditivos de maior precisão, em comparação com os descritores amostrados com o LQTA-QSAR [Martins et al., 2009]. A amostragem proposta também possibilita que os modelos preditivos façam sentido para interpretações químicas, pois é evitado que sejam mapeados pontos internos ao PAC.

Foram realizados experimentos usando seis conjuntos de compostos que podem ser usados como fármacos para diversas doenças. Nos seis cenários avaliados, o método proposto conseguiu gerar modelos mais precisos em relação ao LQTA-QSAR, de acordo com a métrica Q^2. O maior aumento percentual obtido foi de 44%.

Também foi avaliada uma implementação alternativa do método proposto, onde foi usado um mesmo fecho convexo para todos os PACs, que foi calculado considerando os átomos de todos os PACs do conjunto. Os experimentos mostraram que em todos os cenários essa implementação alternativa não conseguiu gerar modelos precisos. O que corrobora a hipótese principal que é melhor usar um fecho convexo para cada PAC, calculado usando apenas os átomos do mesmo PAC.

Outra contribuição dessa dissertação foi propor um novo fluxo de trabalho que substitui o modelo linear por uma árvore de regressão, pois esse tipo de modelo é fácil de ser interpretado e consegue encontrar padrões não-lineares nos dados, que podem

complementar a análise que é feita apenas com os modelos lineares.

Foram realizados experimentos usando a amostragem proposta com essa modificação nos seis cenários. De acordo com o Q^2, em um caso a precisão foi superior aos modelos lineares e nos demais casos foi inferior, porém com precisão ainda satisfatória. Adicionalmente realizamos esse experimento com o LQTA-QSAR usando árvores de regressão, os resultados também foram inferiores aos modelos construídos com a amostragem proposta, com exceção de um caso apenas.

Também foram apresentadas algumas visualizações de dados que possibilitam interpretar os modelos com árvores de regressão e podem ser incluídas no processo de análise por profissionais da área de química.

De acordo com os resultados obtidos durante o desenvolvimento dessa dissertação, podemos apontar as seguintes possibilidades de trabalhos futuros promissores:

- Desenvolver um método para estimar os hiperparâmetros da amostragem proposta de maneira totalmente automatizada.
- Estender a amostragem proposta para outras técnicas 4D-QSAR que usam outros descritores de interação, além dos de Coulomb e Lennard-Jones.
- Investigar estratégias para melhorar a precisão dos modelos que usam árvores de regressão, sem prejudicar a interpretabilidade.
- Propor visualizações de dados interativas para dar suporte a interpretação dos modelos por profissionais da área de química.

Referências Bibliográficas

Andrade, C. H.; Pasqualoto, K. F. M.; Ferreira, E. I. & Hopfinger, A. J. (2010). 4D-QSAR: Perspectives in Drug Design. *Molecules*, 15(5):3281--3294. ISSN 1420-3049.

Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Sadowski, J.; Teckentrup, A. & Wagener, M. (1998). The use of self organizing neural networks in drug design. *Perspectives in Drug Discovery and Design*, 9(11):273--299.

Bak, A. & Polanski, J. (2007). Modeling robust QSAR 3: SOM-4D-QSAR with iterative variable elimination IVE-PLS: Application to steroid, azo dye, and benzoic acid series. *Journal of Chemical Information and Modeling*, 47(4):1469--1480. ISSN 15499596.

Barreiro, E. J. (2009). A Quimica Medicinal e o Paradigma do Composto-Prototipo. *Revista Virtual de Química*, 1(1):26--34. ISSN 1984-6835.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edição.

Berg, M. d.; Cheong, O.; Kreveld, M. v. & Overmars, M. (2008). *Computational Geometry: Algorithms and Applications*. Springer-Verlag TELOS, Santa Clara, CA, USA, 3rd ed. edição. ISBN 3540779736, 9783540779735.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. ISBN 0387310738.

Blum, A. L. & Langley, P. (1997). Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence*, 97(1-2):245--271. ISSN 0004-3702.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5--32. ISSN 0885-6125.

Breiman, L.; Friedman, J. H.; Olshen, R. A. & Stone, C. J. (1984). *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA.

Centner, V.; Massart, D.-L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M. & Sterna, C. (1996). Elimination of Uninformative Variables for Multivariate Calibration. *Analytical Chemistry*, 68(21):3851--3858.

Chan, T. M. (1996). Optimal output-sensitive convex hull algorithms in two and three dimensions. *Discrete & Computational Geometry*, 16(4):361--368. ISSN 1432-0444.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Em *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785--794, New York, NY, USA. ACM.

Cyrus, M. & Beck, J. (1978). Generalized two- and three-dimensional clipping. *Computers & Graphics*, 3(1):23--28. ISSN 0097-8493 (print), 1873-7684 (electronic).

de Melo, E. B. & Ferreira, M. M. C. (2009). Multivariate QSAR study of 4,5-dihydroxypyrimidine carboxamides as HIV-1 integrase inhibitors. *European Journal of Medicinal Chemistry*, 44(9):3577--3583. ISSN 02235234.

Dougherty, E. R. (2005). The fundamental role of pattern recognition for gene-expression/microarray data in bioinformatics. *Pattern Recognition*, 38(12):2226 – 2228. ISSN 0031-3203.

Ferreira, M. M. C. (2002). Multivariate QSAR. Em *Journal of the Brazilian Chemical Society*. ISSN 01035053.

Gieleciak, R.; Magdziarz, T.; Bak, A. & Polanski, J. (2005). Modeling robust QSAR. 1. Coding molecules in 3D-QSAR - From a point to surface sectors and molecular volumes. *Journal of Chemical Information and Modeling*. ISSN 15499596.

Gieleciak, R. & Polanski, J. (2007). Modeling robust QSAR. 2. Iterative variable elimination schemes for CoMSA: Application for modeling benzoic acid pKa values. *Journal of Chemical Information and Modeling*. ISSN 15499596.

Harada, K.; Kubo, H.; Tanaka, A. & Nishioka, K. (2012). Identification of oxazolidinediones and thiazolidinediones as potent 17β-hydroxysteroid dehydrogenase type 3 inhibitors. *Bioorganic and Medicinal Chemistry Letters*, 22(1):504--507. ISSN 0960894X.

Hastie, T.; Tibshirani, R. & Friedman, J. (2003). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, corrected edição. ISBN 0387952845.

Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J. & Duraiswami, C. (1997). Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *Journal of the American Chemical Society.* ISSN 00027863.

Hughes, G. (2006). On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Trans. Inf. Theor.*, 14(1):55--63. ISSN 0018-9448.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling.* Wiley Professional Computing. Wiley. ISBN 978-0-471-50336-1.

Keilis-Borok, V.; Soloviev, A.; Allegre, C.; Sobolevskii, A. & Intriligator, M. (2005). Patterns of macroeconomic indicators preceding the unemployment rise in western europe and the {USA}. *Pattern Recognition*, 38(3):423 – 435. ISSN 0031-3203.

Kohavi, R. & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273--324. ISSN 0004-3702.

Ladduwahetty, T.; Boase, A. L.; Mitchinson, A.; Quin, C.; Patel, S.; Chapman, K. & MacLeod, A. M. (2006). A new class of selective, non-basic 5-HT2A receptor antagonists. *Bioorganic and Medicinal Chemistry Letters*, 16(12):3201--3204. ISSN 0960894X.

Ladduwahetty, T.; Gilligan, M.; Humphries, A.; Merchant, K. J.; Fish, R.; McAlister, G.; Ivarsson, M.; Dominguez, M.; O'Connor, D. & MacLeod, A. M. (2010). Non-basic ligands for aminergic GPCRs: The discovery and development diaryl sulfones as selective, orally bioavailable 5-HT2Areceptor antagonists for the treatment of sleep disorders. *Bioorganic and Medicinal Chemistry Letters*, 20(12):3708--3712. ISSN 0960894X.

Liu, H. & Yu, L. (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):491--502. ISSN 1041-4347.

Madapa, S.; Tusi, Z.; Sridhar, D.; Kumar, A.; Siddiqi, M. I.; Srivastava, K.; Rizvi, A.; Tripathi, R.; Puri, S. K.; Shiva Keshava, G. B.; Shukla, P. K. & Batra, S. (2009). Search for new pharmacophores for antimalarial activity. Part I: Synthesis

and antimalarial activity of new 2-methyl-6-ureido-4-quinolinamides. *Bioorganic and Medicinal Chemistry*, 17(1):203--221. ISSN 09680896.

Martins, J. P. A.; Barbosa, E. G.; Pasqualoto, K. F. M. & Ferreira, M. M. C. (2009). LQTA-QSAR: a new 4D-QSAR methodology. *Journal of Chemical Information and Modeling*, 49(6):1428--1436. ISSN 1549-9596 (Print).

Martins, J. P. a. & Ferreira, M. M. C. (2013). QSAR Modeling: a New Open Source Computational Package To Generate and Validate QSAR Models. *Química Nova*, 36(4):554--U250. ISSN 0100-4042.

Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Naenna, T. & Prachayasittikul, V. (2009). A practical overview of quantitative structure-activity relationship. *EXCLI Journal*, 8:74--88. ISSN 16112156.

Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C. & Ferrin, T. E. (2004). UCSF Chimera ? A Visualization System for Exploratory Research and Analysis. *Journal of Computational Chemistry*, 25(13):1605--12.

Polanski, J. (2003). Self-organizing Neural Networks for Pharmacophore Mapping. *Advanced Drug Delivery Reviews*, 55:1149--1162. ISSN 0169409X.

Polanski, J. & Bak, A. (2003). Modeling Steric and Electronic Effects in 3D- and 4D-QSAR Schemes: Predicting Benzoic pKa Values and Steroid CBG Binding Affinities. *Journal of Chemical Information and Computer Sciences*. ISSN 00952338.

Polanski, J.; Bak, A.; Gieleciak, R. & Magdziarz, T. (2004). Self-organizing Neural Networks for Modeling Robust 3D and 4D QSAR: Application to Dihydrofolate Reductase Inhibitors. *Molecules*, 9:1148--1159. ISSN 1420-3049.

Rogers, D. & Hopfinger, A. J. (1994). Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.*, 34:854--866.

Roy, P. & Roy, K. (2008). On Some Aspects of Variable Selection for Partial Least Squares Regression Models. *QSAR & Combinatorial Science*, 27(3):302--313. ISSN 1611020X.

Sammut, C. & Webb, G. I. (2011). *Encyclopedia of Machine Learning*. Springer Publishing Company, Incorporated, 1st edição. ISBN 0387307680, 9780387307688.

Sant'Anna, C. M. R. (2002). Glossário de termos usados no planejamento de fármacos (recomendações da IUPAC para 1997). *Química Nova*, 25:505 – 512. ISSN 0100-4042.

Schwartz, W. R.; Kembhavi, A.; Harwood, D. & Davis, L. S. (2009). Human detection using partial least squares analysis. Em *IEEE International Conference on Computer Vision*, pp. 24–31.

Teófilo, R. F.; Martins, J. P. A. & Ferreira, M. M. C. (2009). Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *Journal of Chemometrics*, 23(1):32--48. ISSN 08869383.

Theodoridis, S. & Koutroumbas, K. (2008). *Pattern Recognition, Fourth Edition*. Academic Press, 4th edição. ISBN 1597492728, 9781597492720.

Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E. & Berendsen, H. J. C. (2005). Gromacs: Fast, flexible, and free. *Journal of Computational Chemistry*, 26(16):1701--1718. ISSN 1096-987X.

Wilson, K. J.; van Niel, M. B.; Cooper, L.; Bloomfield, D.; O'Connor, D.; Fish, L. R. & MacLeod, A. M. (2007). 2,5-Disubstituted pyridines: The discovery of a novel series of 5-HT2Aligands. *Bioorganic and Medicinal Chemistry Letters*, 17(9):2643--2648. ISSN 0960894X.

Wold, H. (2006). *Partial Least Squares*. American Cancer Society.

Yano, S.; Kazuno, H.; Sato, T.; Suzuki, N.; Emura, T.; Wierzba, K.; Yamashita, J. I.; Tada, Y.; Yamada, Y.; Fukushima, M. & Asao, T. (2004a). Synthesis and evaluation of 6-methylene-bridged uracil derivatives. Part 2: Optimization of inhibitors of human thymidine phosphorylase and their selectivity with uridine phosphorylase. *Bioorganic and Medicinal Chemistry*, 12(13):3443--3450. ISSN 09680896.

Yano, S.; Kazuno, H.; Suzuki, N.; Emura, T.; Wierzba, K.; Yamashita, J. I.; Tada, Y.; Yamada, Y.; Fukushima, M. & Asao, T. (2004b). Synthesis and evaluation of 6-methylene-bridged uracil derivatives. Part 1: Discovery of novel orally active inhibitors of human thymidine phosphorylase. *Bioorganic and Medicinal Chemistry*, 12(13):3431--3441. ISSN 09680896.

Zaki, M. J. & Jr, W. M. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, New York, NY, USA. ISBN 0521766338, 9780521766333.