UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

Henrique Rodrigues Teles

Estratégias em modelagem molecular para a identificação de novos candidatos à fármacos para a leishmaniose

São Carlos

2022

Henrique Rodrigues Teles

Estratégias em modelagem molecular para a identificação de novos candidatos à fármacos para a leishmaniose

Dissertação apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Mestre em Ciências.

Área de concentração: Física Aplicada

Orientador: Prof. Dr. Adriano Defini Andricopulo

Versão original

São Carlos

2022

Teles, Henrique Rodrigues
Estratégias em modelagem molecular para a identificação de novos candidatos à fármaco para a leishmaniose / Henrique Rodrigues Teles; orientador Adriano Defini Andricopulo -- São Carlos, 2022.
98 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Física Aplicada) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2022.

1. Leishmania infantum. 2. Triagem virtual. 3. SBDD. 4. LBDD. 5. QSAR. I. Andricopulo, Adriano Defini, orient. II. Título.

FOLHA DE APROVAÇÃO

Henrique Rodrigues Teles

Dissertação apresentada ao Instituto de Física de São Carlos da Universidade de São Paulo para obtenção do título de Mestre em Ciências. Área de Concentração: Física Aplicada - Opção: Física Biomolecular

Aprovado(a) em: 23/06/2022

Comissão Julgadora

Dr(a). Adriano Defini Andricopulo

Instituição: (IFSC/USP)

Dr(a). Kathia Maria Honorio

Instituição: (EACH/USP)

Dr(a). Lílian Sibelle Campos Bernardes

Instituição: (UFSC/Florianópolis)

AGRADECIMENTOS

Agradeço à Universidade de São Paulo, ao Instituto de Física de São Carlos e ao Laboratório de Química Medicinal e Computacional por todo suporte e conhecimento passado, que possibilitou a realização desse trabalho. Agradeço também ao CNPq e à FAPESP pelo financiamento que permitiu e permite a produção científica nacional e em particular à CAPES pela Bolsa de Mestrado concedida durante parte deste trabalho.

Agradeço ao meu orientador, Prof. Dr. Adriano D. Andricopulo pela oportunidade, confiança e incentivo, o que tornou possível a conclusão deste trabalho e ampliação da minha experiência e visão acadêmico-científica.

Agradeço aos meus colegas de Laboratório, Alex Medeiros, Aldo Senna, David Palomino, Julia Medeiros, Matheus Souza, Renata Andricopulo e Simone Michelan, por todos cafés e risadas na copa, companhias no almoço e no traslado entre os campi. Agradeço em particular ao Leonardo Ferreira e à Marilia Valli pelas discussões, reuniões e proposições que foram fundamentais para essa dissertação.

Agradeço aos meus amigos da família Rep. Lasquera, Angelo, Paulão, Ronaldo e Etevaldo, que foram fundamentais pra me sentir acolhido em uma cidade nova. Agradeço pelas risadas, almoços de domingo, churrascos e também pelos estresses.

Agradeço aos meus amigos de Vitória, Bernard, Claudio, Matheus e Pedro, que, mesmo de longe e através de videochamadas, eram capazes de me impulsionar. Agradeço pelos momentos presenciais que foram fundamentais para recarregar o ânimo.

Agradeço aos meus amigos do balé, que quando viajava pra Vitória mesmo na correria me acolhiam nos ensaios e me proporcionavam momentos de muita alegria.

Agradeço à minha querida namorada Helena por todo carinho, paciência, ajuda, apoio, incentivo que mesmo à distância sempre esteve presente. Sem você a conclusão deste trabalho seria mais difícil. Obrigado por todos os minutos de ligação e pela parceria de todos esses anos.

Agradeço à minha família, que sem eles nada disso seria possível. Agradeço aos meus pais, Omar e Elisangela, pelo apoio, incentivo e educação que permitiram chegar onde eu queria. Agradeço à minha querida irmã, Maria Vitória, pelas conversas e momentos, você me inspira e me ensina em todos os momentos. Agradeço aos meus avós, meus tios e tias, em especial Ernani e Elaine, que estiveram presentes em toda minha trajetória, me dando muito carinho e amor durante toda minha vida, e servindo de exemplo e inspiração para mim.

A todos que não mencionei, mas que contribuíram de várias formas, meu muito

obrigado.

"Nada na vida deve ser temido, somente compreendido.
Agora é hora de compreender mais para temer menos."
Marie Curie

RESUMO

TELES, H. **Estratégias em modelagem molecular para a identificação de novos candidatos à fármacos para a leishmaniose**. 2022. 98p. Dissertação (Mestrado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2022.

A leishmaniose é uma doença tropical negligenciada (DTN) que acomete cerca de 1 milhão de pessoas por ano, sendo que 1 bilhão de pessoas vivem em áreas endêmicas. A leishmaniose visceral, causada pelas espécies *Leishmania donovani* e *Leishmania infantum*, é a forma mais grave da doença, apresentando altas taxas de mortalidade quando não tratada adequadamente. O tratamento disponível é complexo, tem longa duração e apresenta alta toxicidade. Portanto, o desenvolvimento de novos fármacos seguros e eficazes para a leishmaniose visceral é de grande urgência. Neste trabalho, estratégias de planejamento de fármacos baseado na estrutura do receptor (SBDD, do inglês *Structure-based drug design*) e na estrutura do ligante (LBDD, do inglês *Ligand-based drug design*) foram utilizadas. Modelos de relações quantitativas entre estrutura e atividade (QSAR, do inglês *Quantitative structure-actitivy relationships*) foram desenvolvidos para uma série de compostos bioativos frente a *L. infantum*. Os modelos gerados apresentaram alta consistência interna e capacidade de predição externa. Adicionalmente, estes modelos identificaram características importantes para a atividade biológica dos compostos e, portanto, são úteis para o planejamento de novas moléculas bioativas contra *L. infantum*. Na estratégia em SBDD, utilizou-se como alvo molecular a enzima metionil-tRNA sintetase de *L. infantum* (LiMetRS), essencial para a síntese de proteínas. A estrutura tridimensional da LiMetRS foi predita por modelagem por homologia, utilizando-se como *templates* a MetRS de *Trypanosoma brucei* e *L. major*. O modelo foi validado e, utilizando propriedades moleculares de inibidores da enzima, foram realizadas triagens virtuais no banco de dados ZINC, resultando em 5.340.480 compostos que foram selecionados para estudos de docagem molecular. Etapas adicionais de triagem virtual foram realizadas com um programa desenvolvido pelo autor para selecionar apenas os ligantes que interagem com aminoácidos específicos do sítio de interação da LiMetRS. Ao término do processo de triagem virtual, 10 moléculas foram selecionadas. Esse conjunto de moléculas interage com aminoácidos importantes para a estabilização de ligantes no sítio ativo da enzima, em uma região altamente conservada. A molécula melhor pontuada deste conjunto foi submetida a estudos de dinâmica molecular para avaliação da sua estabilidade e das interações intermoleculares no sítio de interação. As estratégias em LBDD e SBDD desenvolvidas neste estudo contribuem para uma melhor compreensão dos fenômenos de reconhecimento molecular entre LiMetRS e potenciais inibidores, e para o planejamento de novas moléculas bioativas para o tratamento da leishmaniose visceral.

Palavras-chave: *Leishmania infantum*. Triagem virtual. SBDD. LBDD. QSAR. Dinâmica molecular.

ABSTRACT

TELES, H. **Strategies in molecular modeling to identify new drug candidates for leishmaniasis**. 2022. 98p. Dissertação (Mestrado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2022.

Leishmaniasis is a neglected tropical disease (NTD) that affects 1 million people each year, with 1 billion people living in endemic areas. Visceral leishmaniasis, caused by *Leishmania donovani* and *Leishmania infantum*, is the most serious form of the disease, causing high mortality rates if not treated properly. The available treatment is complex and long, and presents high toxicity. Therefore, the development of novel, safe and effective drugs for visceral leishmaniasis is highly urgent. In this study, structure- and ligand-based drug design (SBDD and LBDD, respectively) strategies were applied. Quantitative structure-activity relationship (QSAR) models were developed for a series bioactive compounds against *L. infantum*. The generated models had high internal consistency and external predictive power. Additionally, these models identified important features for the biological activity of the compounds and, therefore, they are useful for the design of new anti-*L. infantum* molecules. In the SBDD strategy, the enzyme methionyl-tRNA synthetase of *L. infantum* (LiMetRS), which is essential for protein synthesis, was used as the molecular target. The three-dimensional structure of LiMetRS was predicted by homology modeling, using the MetRS from *Trypanosoma brucei* and *L. major* as templates. The model was validated and, using molecular properties of known inhibitors, the ZINC database was screened, which resulted in 5,340,480 compounds that were selected for molecular docking studies. Additional virtual screening steps were performed with a program developed by the author to select only ligands that interact with specific amino acids in the LiMetRS binding site. At the end of the virtual screening process, 10 molecules were selected. These molecules interact with important amino acids for the stabilization of ligands at the active site of the enzyme, in a highly conserved region. The best-scoring molecule of this set was subjected to molecular dynamics simulations to assess their stability and intermolecular interactions in the binding site. The LBDD and SBDD strategies developed in this study contribute to a better understanding of the molecular recognition between LiMetRS and putative inhibitors, and to the design of novel bioactive molecules for the treatment of visceral leishmaniasis.

Keywords: *Leishmania infantum*. Virtual screening. SBDD. LBDD. QSAR. Molecular dynamics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AP	Sítio Auxiliar da MetRS, do inglês, *Auxiliary Pocket*
C	Descritor de Coulomb
CoMFA	Análise Comparativa dos Campos Moleculares, do inglês *Comparative Molecular Field Analysis*
CP	do inglês, *Connective Peptide Domain*
DOPE	do inglês, *Discrete Optimized Protein Energy*
DTN	Doença Tropical Negligenciada
EMP	Sítio da metionina da MetRS, do inglês, *Enlarged Methionine Pocket*
GA	Algoritmo Genético
KPLS	do inglês, *Kernel Partial Least Squares Regression*
LBDD	Planejamento Baseado na Estrutura do Ligante, do inglês *Ligand-based drug design*
LiMetRS	*Leishmania infantum* Metionil-tRNA Sintetase
LJ	Descritor de Lennard-Jones
LmMetRS	*Leishmania major* Metionil-tRNA Sintetase
MCS	Máxima Subestrutura Comum
MDS	Escalonamento Multidimensional, do inglês, *Multidimensional Scaling*
MLR	Regressão Linear Múltipla, do inglês *Multiple Linear Regression*
OMS	Organização Mundial da Saúde
PAC	Perfil de Amostragem Conformacional
PCA	Análise de Componentes Principais, do inglês *Principal Component Analysis*
PCR	Regressão de Componentes Principais, do inglês *Principal Component Regression*
PLIP	do inglês *Protein-Ligand Interaction Profiler*

PLS	Regressão de Mínimos Quadrados Parciais, do inglês *Partial Least Squares Regression*
QSAR	Relações Quantitativas entre Estrutura e Atividade , do inglês *Quantitative structure–activity relationships*
QSPR	Relações Quantitativas entre Estrutura e Propriedade, do inglês *Quantitative Structure-Property Relationships*
RD	Receptor Dependente
RI	Receptor Independente
RMSD	Desvio Quadrático Médio, do inglês, *Root-Mean-Square Deviation*
RMSE	Raiz Quadrada do Erro-Médio, do inglês *Root Mean Squared Error*
SBDD	Planejamento Baseado na Estrutura do Receptor, do inglês *Structure-based drug design*
SD	Desvio Padrão, do inglês *Standard Deviation*
TbMetRS	*Trypanosoma brucei* Metionil-tRNA Sintetase

SUMÁRIO

1 INTRODUÇÃO

1.1 A leishmaniose

A leishmaniose é uma doença tropical negligenciada (DTN) causada por protozoários de mais de vinte espécies do gênero *Leishmania*. As DTNs compreendem um conjunto de doenças provocadas por agentes infecciosos ou parasitas e ocorrem em regiões tropicais e subtropicais[1]. Tais doenças possuem como característica comum o fato de atingirem populações vulneráveis em regiões e países pobres com condições precárias de saneamento e higiene. São consideradas negligenciadas por não receberem a mesma atenção em comparação à outras doenças, uma vez que o processo de planejamento e desenvolvimento de novos fármacos é caro e complexo, e as pessoas acometidas por estas doenças possuem baixo poder aquisitivo e não tem força política[2]. Sendo assim, não há interesse da indústria devido ao baixo retorno financeiro. Reflexos da desigualdade e incidência da leishmaniose também se mostram presentes no Brasil de modo que a região nordeste é a mais acometida pela doença[3]. Outro agravante das DTNs é o fato de que podem coexistir com outras doenças, como por exemplo a AIDS, dificultando o tratamento. A leishmaniose afeta cerca de 0,9 a 1,6 milhão de pessoas por ano em todo o mundo, causando de 20.000 a 30.000 mortes[4]. A Figura 1 representa o ciclo de vida da *Leishmania*. A transmissão do parasita ocorre através da picada da fêmea de mosquitos flebotomíneos, popularmente conhecidos como mosquitos-palha. O ciclo de transmissão inicia-se com a inoculação do parasita em fase promastigota, forma flagelada e extracelular, no hospedeiro. Os macrófagos, células do sistema imunológico, fagocitam os parasitas que passam da forma promastigota para amastigota, forma intracelular e não flagelada. Por meio de sucessivas mitoses o parasita, na forma amastigota, se multiplica e infecta novos macrófagos. A leishmaniose se manifesta em três principais formas: leishmaniose cutânea (LC); leishmaniose mucocutânea (LM) e leishmaniose visceral (LV). A LC é considerada a forma mais branda e a mais comum; manifesta-se através de erupções cutâneas. A LM produz lesões nas mucosas e cartilagens principalmente na face, boca, nariz e garganta. A LV é a forma mais agressiva com altos índices de letalidade em crianças desnutridas, portadoras da infecção pelo vírus da síndrome da imunodeficiência adquirida (AIDS, HIV) e em casos de não tratados adequadamente. A doença atinge órgãos viscerais como fígado, baço, linfonodos e medula óssea. É provocada pelas espécies *Leishmania infantum*, *L. chagasi* e *L. donovani*[5]. No Brasil, ocorre mais de 90% dos casos de VL em todo o continente americano. As infecções pela espécie *L. donovani*, em sua maior parte, ocorrem nas regiões da Ásia e da África, enquanto as infecções pela espécie *L. infantum* ocorrem nas regiões da América Latina e regiões de clima mediterrâneo[6].

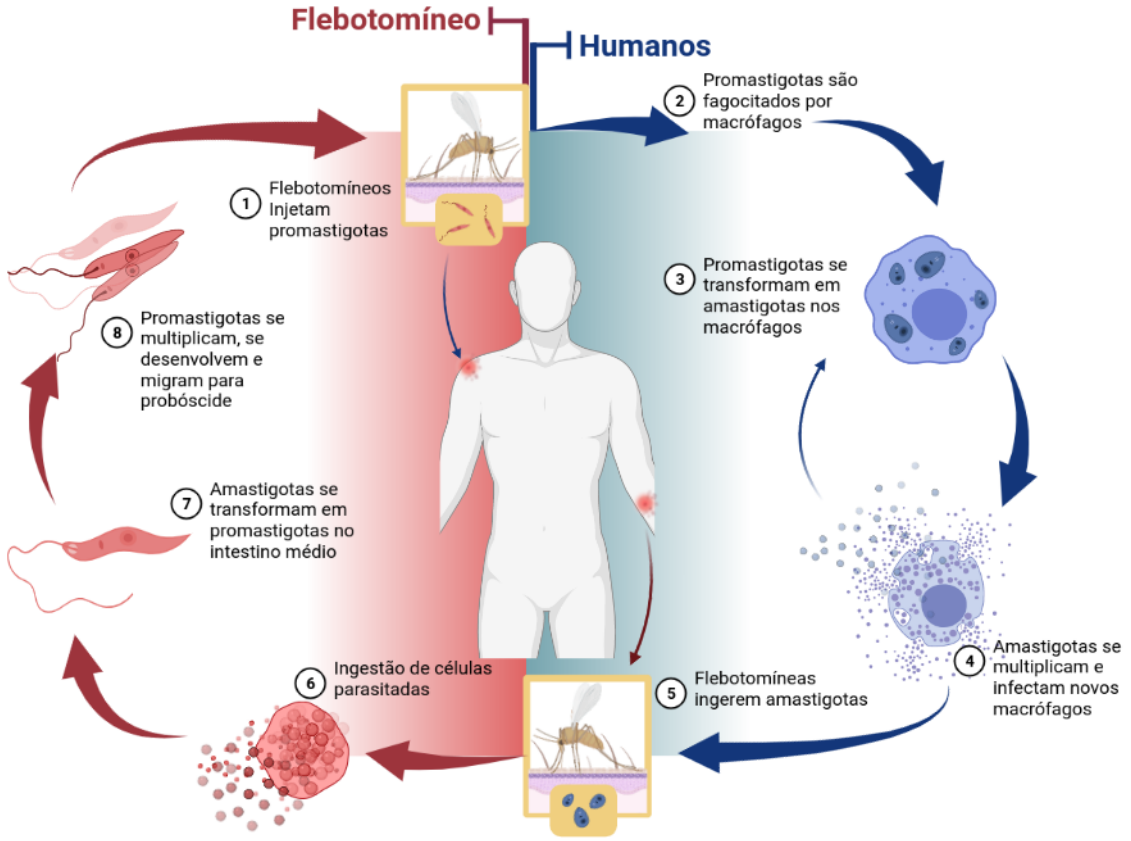


Figura 1: Ciclo de vida da *Leishmania*. Estágios do desenvolvimento no hospedeiro intermediário, flebotomíneo, e no hospedeiro definitivo.

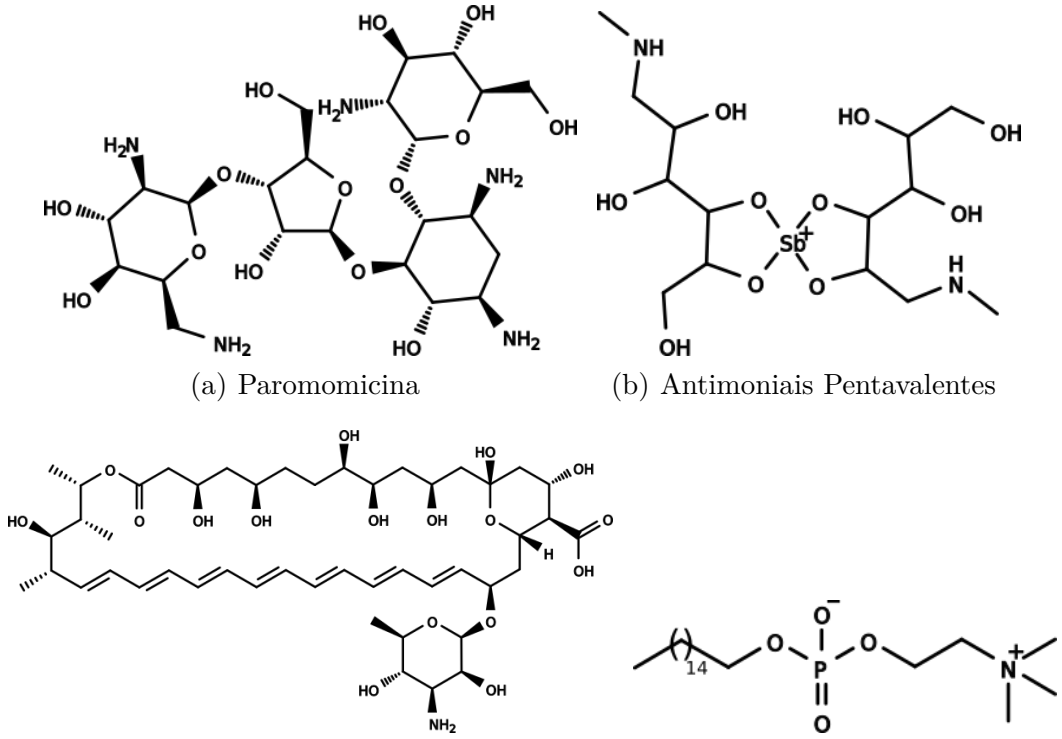
Fonte: Elaborada pelo autor.

O tratamento disponível para leishmaniose é bastante complexo devido a longa duração do tratamento, modo de administração e alta toxicidade. Os principais fármacos administrados são anfotericina B, antimoniais pentavalentes, miltefosina e paromomicina, administrados em doses diárias por via intramuscular ou intravenosa durante um período de 20 a 30 dias (Figura 2). Os efeitos adversos provocados pela alta toxicidade incluem fraqueza, náusea, vômito, diarréia, cólicas abdominais, hepatotoxicidade, cardiotoxicidade entre outros. Outro agravante refere-se à resistência adquirida pelo parasita, o que diminui a eficácia dos fármacos disponíveis[7–8]. Apesar de estes medicamentos serem muito utilizados, pouco se sabe sobre seus mecanismos de ação[9].

Diante deste cenário, é necessário o desenvolvimento de novos fármacos para a leishmaniose que apresentem menor toxicidade e possibilitem tratamentos mais seguros e menos invasivos, principalmente para a leishmaniose visceral devido à sua alta letalidade.

1.2 Desenvolvimento de fármacos

A pesquisa e desenvolvimento (PD) de novos fármacos é reconhecida pela interdisciplinaridade e multidisciplinaridade, incluindo conceitos de Ciências Biológicas, Químicas,



(a) Paromomicina (b) Antimoniais Pentavalentes

(c) Anfotericina B (d) Miltefosina

Figura 2: Estrutura química dos fármacos utilizados para o tratamento de leishmaniose.

Fonte: Elaborada pelo autor.

Farmacêuticas, Físicas, entre outras, além da utilização de métodos computacionais que possibilitam diminuir custos e tempo no planejamento de novos compostos bioativos. O processo de PD de um fármaco leva em média entre 12 e 15 anos, e é extremamente custoso, chegando na ordem de bilhão de dólares[10]. Neste processo, o fármaco deve passar por rigorosos testes que abrangem uma fase pré-clínica e fases clínicas[11].

Diversas estratégias podem ser empregadas na fase de planejamento pré-clínica. A estratégia na qual não há informação sobre o alvo molecular, entretanto, há conhecimento sobre os ligantes, é chamado de planejamento de fármacos baseado na estrutura do ligante (LBDD, do inglês, *Ligand-Based Drug Design*). Nesta estratégia, uma das principais ferramentas são os métodos estatísticos que são capazes de correlacionar a estrutura de uma molécula com uma propriedade de interesse (propriedade alvo), por exemplo atividade. Exemplos desta abordagem são os métodos de Relações Quantitativas entre Estrutura e Atividade (QSAR, do inglês *Quantitative Structure-Activity Relationship*), ou Relações Quantitativas entre Estrutura e Propriedade (QSPR, do inglês *Quantitative Structure-Property Relationships*)[12]. Esses métodos são capazes de, a partir de um conjunto de moléculas com suas respectivas propriedades-alvo, reconhecer características estruturais correlacionadas com a variação desta propriedade e, portanto, gerar conhecimento para propor novos conjuntos de moléculas otimizadas. Por outro lado, a estratégia conhecida

como planejamento de fármacos baseado na estrutura do receptor (SBDD, do inglês *Structure-Based Drug Design*) utiliza informações sobre a estrutura tridimensional do alvo macromolecular e suas interações com ligantes. Nessa estratégia, a partir da estrutura tridimensional do alvo molecular, é possível propor ligantes e predizer o seu modo de interação com o sítio[13]. Dentre as principais ferramentas de SBDD, a docagem molecular é uma das mais utilizadas e consiste em predizer a conformação e a energia de interação de uma molécula pequena ao interagir com uma macromolécula. Outro método é a dinâmica molecular, a partir da qual podemos avaliar a estabilidade de um ligante e das interações intermoleculares no sítio de interação do alvo molecular[13].

Uma vez que os compostos se mostram promissores em estudos de modelagem molecular, podem ser testados experimentalmente em testes *in vitro* e *in vivo*. Se o candidato a fármaco obtém sucesso em todas as etapas pré-clínicas, passa-se então para os estudos clínicos, que são divididos em 4 fases. A fase I é conduzida em um grupo pequeno de 20 a 100 indivíduos saudáveis a fim de se obter parâmetros iniciais como dose e segurança. A fase II é conduzida em grupos de 100 a 300 portadores da patologia. Nesta fase observa-se a eficácia do fármaco, ou seja, se o fármaco foi capaz de produzir o efeito desejado contra a doença, além de se avaliar os possíveis efeitos adversos. A fase III é a última antes da comercialização e utiliza grupos de 5 a 10 mil indivíduos para gerar dados estatísticos mais robustos e validar eficácia e segurança do candidato a fármaco. A fase IV consiste em um acompanhamento do fármaco após a comercialização em milhares de indivíduos. Dada a variabilidade deste conjunto de indivíduos, a fase IV é utilizada para se avaliar efeitos adversos que até então eram desconhecidos[14].

1.3 Planejamento de fármacos baseado na estrutura do ligante

1.3.1 Relações Quantitativas entre Estrutura e Atividade

As estratégias em QSAR são amplamente utilizadas em LBDD. O método baseia-se no princípio central de que a atividade biológica é resultante das características moleculares[15]. São gerados modelos matemáticos que correlacionam as estruturas de um conjunto de moléculas com a atividade biológica. Estes modelos, além de explicar a variação de atividade em função da estrutura, podem ser utilizados para predizer a atividade de novos compostos. A atividade biológica nos modelos de QSAR é denominada variável dependente (y), enquanto a estrutura é decodificada em descritores moleculares, ou variáveis independentes (x). Dessa forma, temos a expressão:

$$y = f(x) \tag{1.1}$$

Em que a função $f(x)$ é derivada através de métodos estatísticos de regressão.

O processo de construção de modelos QSAR inicia-se com a seleção do conjunto de dados. Quando se conhece o alvo molecular, os compostos devem atuar através do

mesmo mecanismo de ação[16–17]. A atividade biológica deve ter sido determinada usando-se o mesmo protocolo experimental. Outra etapa importante refere-se à escolha dos descritores adequados para o conjunto de dados. Descritores moleculares são representações quantitativas de uma molécula[18]. Os descritores podem ser unidimensionais (QSAR-1D), bidimensionais (QSAR-2D), tridimensionais (QSAR-3D) e quadridimensionais (QSAR-4D). Os descritores 1D consideram a fórmula química da molécula. Os descritores 2D consideram a estrutura bidimensional representada através de parâmetros topológicos. Os descritores 3D consideram a estrutura tridimensional, enquanto os descritores 4D consideram a dependência temporal. No caso dos métodos 4D, são geradas diversas conformações das moléculas obtidas por meio de simulações de dinâmica molecular.

Após a seleção dos descritores, o modelo é gerado a partir de métodos estatísticos de regressão que correlacionam a atividade biológica com os descritores. O objetivo dos métodos de regressão é encontrar a melhor equação que descreva a correlação entre os descritores e a resposta biológica, minimizando o erro entre os valores de atividade preditos e os valores experimentais. Um dos métodos fundamentais é o método de regressão linear múltipla (MLR, do inglês, *Multiple Linear Regression*). Neste método, os coeficientes de regressão são obtidos através da resolução direta da combinação linear entre atividade e descritores (Equação 1.2):

$$Y_i = a_0 + a_1 X_{i,1} + a_2 X_{i,2} + a_j X_{i,j} + ... + a_n X_{i,n} \tag{1.2}$$

Em que Y_i é atividade biológica de apenas uma molécula, $X_{i,j}$ são os descritores para a molécula i e a_j os coeficientes atribuídos a cada descritor. Para um conjunto de m moléculas, a determinação dos coeficientes torna-se um problema matricial dado por:

$$\mathbf{Xa} = \mathbf{Y} \tag{1.3}$$

Onde $\mathbf{X} = \mathbf{X_{m \times n}}$ é a matriz dos descritores, $\mathbf{a}$ o vetor dos coeficientes e $\mathbf{Y}$ o vetor da atividade biológica. A Equação 1.3 possui solução única quando $m > n$, ou seja, quando o número de linhas da matriz $\mathbf{X}$ for maior que o número de colunas (número de moléculas maior que o número de descritores). Os coeficientes de regressão são gerados pela resolução da Equação 1.3:

$$\mathbf{a} = (\mathbf{X^T X})^{-1} \mathbf{X^T Y_{obs}} \tag{1.4}$$

Outra restrição do método MLR refere-se à condição de existência da inversa da matriz $\mathbf{X^T X}$ para que a Equação 1.4 tenha solução. A matriz $\mathbf{X^T X}$ não possui inversa quando duas colunas são linearmente dependentes ou aproximadamente linearmente dependente, em outras palavras, se houver uma alta correlação entre duas variáveis. A

multicolinearidade pode gerar um inversa aproximada e consequentemente coeficientes de regressão imprecisos. Para contornar os problemas do método MLR, costuma- se usar métodos de projeção, que substituem os descritores originais por variáveis latentes que carregam grande parte das informações contidas nas condições iniciais. Em seguida, as regressões são feitas sobre as novas variáveis. A regressão de componentes principais (PCR, do inglês *Principal Component Regression*) é um método de regressão baseado na análise de componentes principais (PCA, do inglês *Principal Component Analysis*). A PCA utiliza uma transformação ortogonal para converter um conjunto de variáveis correlacionadas em um conjunto de variáveis linearmente não correlacionadas chamadas componentes principais. Por sua vez, na regressão de mínimos quadrados parciais (PLS, do inglês *Partial least Squares Regression*)[19] a substituição por variáveis latentes é feita tanto em X quanto em Y. Por fim, os modelos QSAR são submetidos à testes de validação, tanto interna quanto externa, que garantem sua robustez estatística e a capacidade preditiva.

Antes da criação dos modelos de QSAR, o conjunto de dados deve ser divido em dois subconjuntos. O conjunto utilizado para construção do modelo é denominado conjunto treinamento, enquanto aquele utilizado na validação externa do modelo, é o conjunto teste. Ambos os conjuntos são submetidos à testes de validação. Para o conjunto treinamento, o processo é denominado validação interna. A validação interna pode ser realizada por meio da validação cruzada. Este procedimento compreende dois processos. O primeiro verifica a capacidade do modelo de se ajustar aos dados, o que é representado pelo coeficiente de determinação R^2, calculado pela Equação 1.5. O outro verifica a capacidade de predição interna, que pode ser determinada por métodos como o *leave-one-out* e o *k-fold*. O *k-fold* divide o conjunto treinamento em K conjuntos. Utiliza-se os $K-1$ conjuntos para construção do modelo e o conjunto excluído é utilizado para a validação. O processo é repetido K vezes variando-se o conjunto utilizado para a validação e, por fim, os cálculos são combinados para a produção do resultado. O *leave-one-out* é um caso particular do *k-fold* quando K é igual ao número de moléculas do conjunto treinamento. Em ambas as metodologias, a qualidade de predição interna é avaliada através do coeficiente de correlação de validação cruzada (Q^2), calculado pela Equação 1.6. A Figura 3 apresenta o esquema do método *k-fold*.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - y_{ci})^2}{\sum_{i=1}^{N} (y_i - \langle \mathbf{y} \rangle)^2} \tag{1.5}$$

$$Q^2 = 1 - \frac{\sum_{i=1}^{K} (y_i - y_{vi})^2}{\sum_{i=1}^{K} (y_i - \langle \mathbf{y} \rangle)^2} \tag{1.6}$$

Nas Equações 1.5 e 1.6, N é o número de compostos presentes no conjunto treinamento, K representa o número de subconjuntos no método *k-fold*, y_i é o valor experimental da i-ésima amostra, y_{ci} é o valor da atividade calculado da i-ésima amostra, y_{vi} é o valor de

atividade calculado para a validação interna e $< y >$ é o valor médio de atividade para o conjunto de treinamento.

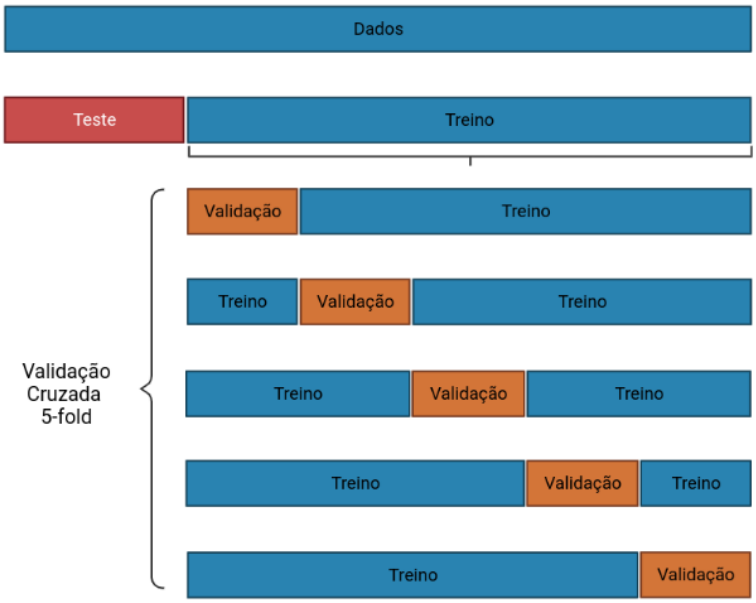


Figura 3: Esquema do método *k-fold* com k=5.

Fonte: Elaborada pelo autor.

Para o conjunto teste, o processo de validação é denominado validação externa e é representado pelo coeficiente de validação preditivo (R^2_{pred})(Equação 1.7).

$$R^2_{\text{pred}} = 1 - \frac{\sum_{i=1}^{M} (y_i - y_{pi})^2}{\sum_{i=1}^{M} (y_i - \langle \mathbf{y} \rangle)^2} \tag{1.7}$$

Na Equação 1.7, M é o número de compostos no conjunto teste e y_p é o valor de atividade predito para este conjunto. Os parâmetros estatísticos mais utilizados para se avaliar a qualidade dos modelos de QSAR são o Q^2 e R^2. A literatura reporta diversos valores mínimos para se obter modelos robustos de QSAR[15]. De maneira geral, o valor de Q^2 deve ser superior a 0,5 e R^2 superior a 0,6[20]. Além disso, modelos robustos não apresentam diferença superior a 0,3 entre os valores de R^2 e Q^2. Diferenças superiores, podem indicar sobreajuste dos dados.

1.3.2 Domínio de Aplicabilidade

O domínio de aplicabilidade (DA) ou domínio de aplicação é uma região teórica limitada pela aplicabilidade do modelo. Um dos objetivos do métodos QSAR é predizer a variável dependente para compostos ainda não testados. Neste contexto, o DA define o limite para o qual o modelo é capaz de realizar as predições de forma confiável. Existem diversos métodos para se avaliar o DA, dentre os quais o método *convex-hull*[15,21]. O

método *convex-hull* é um método geométrico em que o espaço de interpolação é definido pela menor área convexa contendo todo o conjunto treinamento. Esse método é indicado para modelos com dimensionalidade ≤ 3[22].

1.3.3 QSAR-3D e Análise Comparativa dos Campos Moleculares

A análise comparativa dos campos moleculares, (CoMFA, do inglês *Comparative Molecular Field Analysis*), é um dos métodos de QSAR-3D mais usados. O método é baseado nas propriedades estereoquímicas e eletrostáticas das moléculas do conjunto de dados. Cada molécula, em uma única conformação, é alinhada com as demais e posicionada em um retículo. A energia de interação entre um átomo de prova (sonda) e cada molécula é calculada em cada ponto do retículo, os quais constituirão os descritores moleculares (Figura 4(a)). As energias de interação empregadas são de Lennard-Jones e Coulomb, calculadas pelas Equações 1.8 e 1.9:

$$E_{\text{Coulomb}} = \sum_{i=1}^{N} \frac{Q_{\text{sonda}} Q_i}{4\pi\epsilon_0 r_{\text{sonda},i}} \tag{1.8}$$

$$E_{LJ} = \sum_{i=1}^{N} \left(\frac{a}{r_{sonda,i}^{12}} - \frac{c}{r_{sonda,i}^{6}} \right) \tag{1.9}$$

Em que N é o número de átomos da molécula, Q_{sonda} é a carga da sonda, Q_i é a carga do átomo i, $r_{\text{sonda},i}$ é a distância entre a sonda e o átomo i, a e c são parâmetros obtidos através do campo de força e ϵ_0 é constante de permissividade do vácuo. Nem todos os pontos são relevantes para explicar a atividade biológica e, por esta razão, o método aplica um filtro para selecionar variáveis não redundantes. Uma vez reduzido o número de variáveis, o método de PLS gera o modelo de QSAR. O método CoMFA é capaz de gerar mapas de contribuição, regiões do espaço em que forças de atração e repulsão estereoquímica e eletrostática indicam pontos de modificação estrutural para o planejamento de novos compostos. O alinhamento molecular pode ser feito de diversas maneiras tais como alinhamento por mínima estrutura comum, docagem molecular, e alinhamento por farmacóforos. Desta forma, o método CoMFA pode ser utilizado mesmo sem informação estrutural acerca do alvo molecular.

1.3.4 QSAR-4D

Apesar da variedade de métodos de alinhamento usados nas estratégias de QSAR-3D, a maioria destes métodos são independentes do receptor (RI-QSAR-3D). Por esta razão estes métodos podem apresentar dificuldades para identificar a conformação bioativa dos ligantes, um problema que pode ser minimizado pelo uso de métodos de QSAR-4D[23].

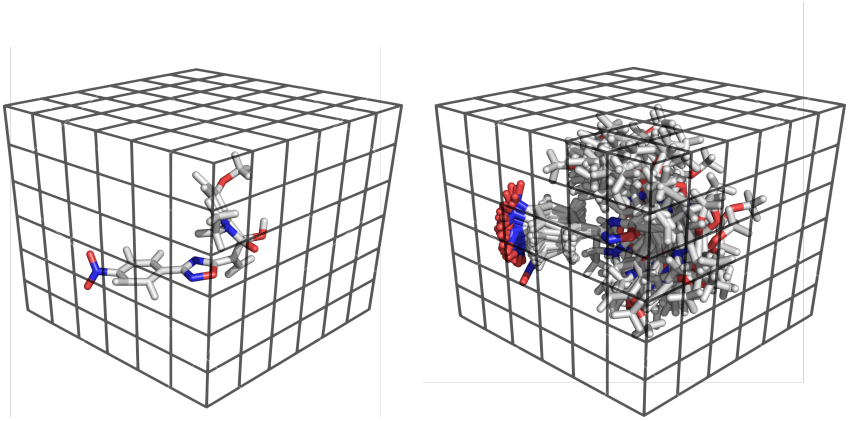
Um dos métodos de QSAR-4D, conhecido por LQTA-QSAR, utiliza a metodologia CoMFA, porém, com o acréscimo de flexibilidade conformacional[24]. Na metodologia LQTA-QSAR um perfil de amostragem conformacional (PAC) é utilizado. Os PACs são gerados

a partir de simulações de dinâmica molecular dos compostos em solvente e as trajetórias são alinhadas. Os PACs são gerados em uma grade tridimensional reticulada (Figura 4(b)) e cada ponto é percorrido por uma sonda. As interações eletrostáticas e estereoquímicas são calculadas pelas Equações 1.10 e 1.11.

$$E_{\text{Coulomb}} = \frac{1}{n}\sum_{i=1}^{N}\frac{Q_{\text{sonda}}Q_i}{4\pi\epsilon_0 r_{\text{sonda},i}} \tag{1.10}$$

$$E_{\text{LJ}} = \sum_{i=1}^{N}\Big[\frac{1}{r^{12}_{\text{sonda},i}}\sqrt{\frac{1}{n}\big(C_i^{(12)}C_{\text{sonda}}^{(12)}\big)} - \frac{1}{r^{6}_{\text{sonda},i}}\sqrt{\frac{1}{n}\big(C_i^{(6)}C_{\text{sonda}}^{(6)}\big)}\Big] \tag{1.11}$$

Nas Equações 1.10 e 1.11, os parâmetros C são obtidos através do campo de força utilizado. $r_{\text{sonda},i}$ é a distância entre a sonda e o átomo i, Q_{sonda} e Q_i são as cargas da sonda e átomo i, respectivamente, $\epsilon_0$constante de permissividade do vácuo, n o número de PAC e N número de átomos do ligante. É possível notar que a diferença entre o formalismo LQTA-QSAR e CoMFA é a presença de um fator de normalização $\frac{1}{n}$ como resultado da presença de várias cópias do mesmo composto nos PACs.



(a) Metodologia CoMFA. (b) Metodologia LQTA.

Figura 4: Diferença entre as metodologias CoMFA e LQTA.

Fonte: Elaborada pelo autor.

1.4 Planejamento de fármacos baseado na estrutura do receptor

1.4.1 Aminoacil-tRNA Sintetase

As aminoacil-tRNA sintetases (aaRS) são enzimas essenciais na etapa de tradução da síntese proteica. São responsáveis por anexar aminoácidos de maneira específica no RNA transportador (tRNA). A reação ocorre no citoplasma das células a partir da energia proveniente da hidrólise da adenosina trifosfato (ATP). Neste processo, ocorre o reconhecimento entre o aminoácido e sua aaRS correspondente, com a formação do

intermediário aminoacil-adenilato (aminoacil-AMP). Em seguida, a enzima reconhece o seu tRNA correspondente resultando na transferência do grupo aminoacil para a extremidade 3 do tRNA. As etapas da reação podem ser resumidas de acordo com as Equações 1.12 e 1.13 e a Figura 5.

$$\text{Aminoácido} + \text{ATP} + \text{aaRS} \rightarrow \text{Aminoacil-AMP} \cdot \text{aaRS} + \text{PP}_i \tag{1.12}$$

$$\text{Aminoacil-AMP} \cdot \text{aaRS} + \text{tRNA} \rightarrow \text{Aminoacil-tRNA} + \text{AMP} + \text{aaRS} \tag{1.13}$$

As aaRS são alvos moleculares validados para o desenvolvimento de fármacos para diversas doenças infecciosas [25–28]. Em um mesmo organismo existem cerca de 20 aaRS que podem ser explorados como alvos moleculares. A enzima metionil-tRNA sintetase (MetRS), por exemplo, tem sido explorada no planejamento de fármacos para a tripanossomíase africana, doença causada pelo protozoário *Trypanosoma brucei*. A diferença entre a estrutura primária de MetRS de mamíferos e tripanossomatídeos sugere que ligantes com boa seletividade podem ser desenvolvidos[29]. Outra característica relevante é a conservação estrutural entre diferentes parasitas[30], o que permite o planejamento de fármacos para diversas doenças parasitárias. Outro exemplo é o uso da MetRS de *L. dovani* (LdMetRS) como alvo molecular para o planejamento de fármacos para a leishmaniose visceral[30] e uso da MetRS de *L. major* (LmMetRS) para a leishmaniose cutânea[31].

1.4.2 Modelagem por Homologia

A modelagem por homologia, ou modelagem comparativa, possibilita a construção de estruturas 3D de proteínas através do alinhamento das estruturas primárias de proteínas

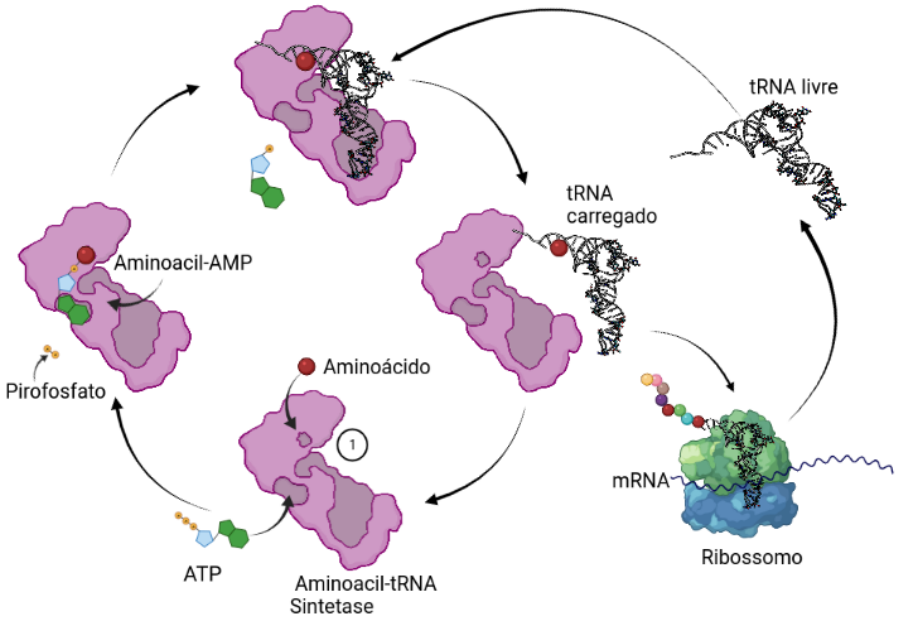


Figura 5: Esquematização das etapas de reação das aminoacil-tRNA sintetases (aaRS).

Fonte: Elaborada pelo autor.

com estrutura conhecida (*templates* ou moldes) e da estrutura alvo[32–33]. Portanto, é necessário que haja um mínimo de identidade entre as sequências de aminoácido da estrutura alvo e do *template*. Recomenda-se o alinhamento entre estruturas com identidade acima de 20% para a obtenção de modelos de qualidade. Além da identidade entre as estruturas primárias, fatores como a existência de ligantes e resolução da estrutura do *template* devem ser observados para a seleção das sequências. Uma vez selecionado o *template*, as sequências são alinhadas, o que possibilita a identificação de regiões da estrutura primária que são conservadas e aquelas que apresentam baixa identidade.

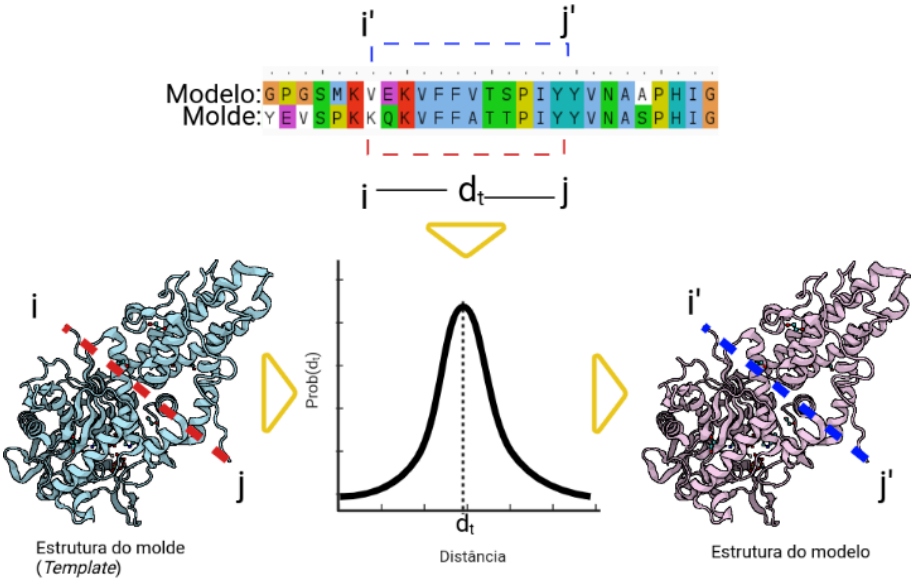


Figura 6: Metodologia de modelagem por satisfação de restrições espaciais.

Fonte: Adaptada de MEIER; SÖDING[34].

Existem diversos métodos para gerar a estrutura alvo, tais como[35] a modelagem por união de corpos rígidos, combinação de segmentos, e satisfação de restrições espaciais. O método utilizado neste trabalho foi a modelagem pela satisfação de restrições espaciais, o qual gera o modelo que satisfaz as restrições provenientes da estrutura *template*. Exemplos de restrição são o comprimento da ligação, ângulo de ligação, ângulos diédricos, entre outros. Esses valores são obtidos em bancos de dados, e a cada uma das restrições são atribuídas funções de densidade de probabilidade, definidas pela frequência com que ocorrem no banco de dados[36]. A função de densidade de probabilidade é aproximada a combinações de funções gaussianas e centrada de acordo com a restrição do *template*. Uma função objetivo é gerada agregando informações das restrições e dos termos que reforçam a conformação esperada. O melhor modelo é aquele que minimiza a função objetivo. A Figura 6 apresenta o método de restrições espaciais, em que as restrições entre o molde e o modelo são extraídas, e a função de densidade de probabilidade é construída com média centrada na distância do molde.

1.4.3 Docagem Molecular

A docagem molecular é um método de SBDD muito utilizado para a predição do modo de interação entre ligantes e macromoléculas. O processo envolve uma etapa de busca conformacional seguida por uma etapa de avaliação das conformações geradas. A literatura reporta diversos algoritmos de docagem molecular, dentre eles os algoritmos de correspondência, construção incremental, métodos de Monte Carlo, algoritmos genéticos e amostragem por dinâmica molecular

Os algoritmos genéticos são inspirados no processo de seleção natural. Os graus de liberdade translacionais, rotacionais e conformacionais (ângulos diedrais) do ligante são codificados nos genes. As informações codificadas em todos os genes formam o cromossomo. Processos de mutação e *crossover* ocorrem a cada ciclo gerando novas gerações de ligantes. As novas estruturas são avaliadas pela função de pontuação e aquelas que com maior pontuação são usadas como pontos de partida para a criação da próxima geração. As mutações alteram aleatoriamente os genes enquanto as operações de *crossover* cruzam genes de dois cromossomos. A Figura 7 apresenta o esquema geral do algoritmo genético.

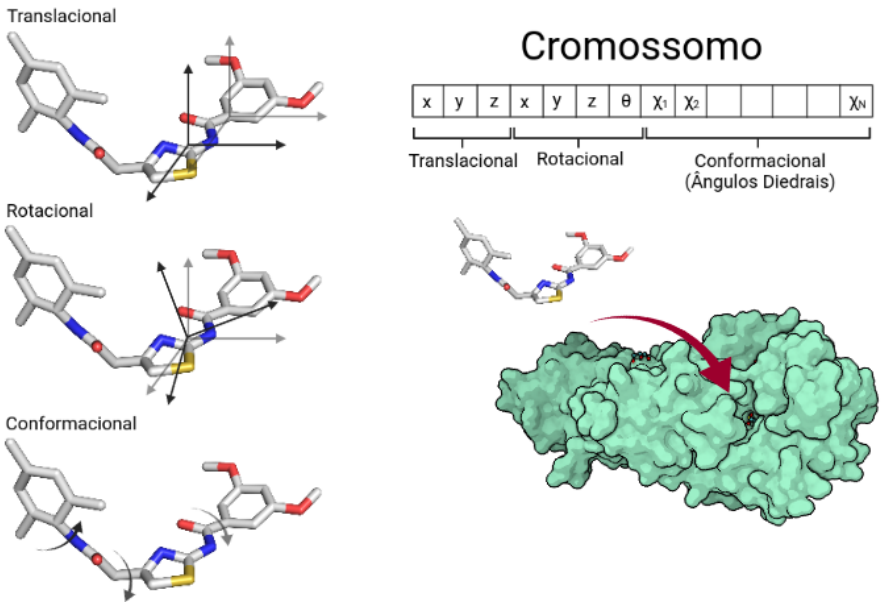


Figura 7: Algoritmo genético aplicado à docagem molecular.

Fonte: Adaptada de MORGON[13].

O objetivo das funções de pontuação é avaliar a afinidade entre o ligante e o alvo molecular e gerar uma classificação relativa entre os ligantes. As funções de pontuação podem ser divididas em três categorias: (i) baseadas em campo de força; (ii) empíricas, e (iii) baseadas no conhecimento. As funções de pontuação baseadas em campo de força calculam a energia de ligação como soma a das interações não ligadas (van de Waals e eletrostáticas). As funções de pontuação empíricas decompõem a energia total a partir da soma de interações de hidrogênio, interações iônicas, efeitos hidrofóbicos, entre outras interações.

Em seguida, coleções de complexos receptor-ligante cuja afinidade de ligação é conhecida são usadas para predizer os valores de ΔG das soluções de docagem. As funções baseadas em conhecimento usam dados estatísticos de banco de dados de complexos proteína-ligante para gerar potenciais de energia e gerar uma equação geral. Neste trabalho, foi usada a função baseada em campo de força GoldScore[37].

1.4.4 Dinâmica Molecular

A dinâmica molecular é um método de simulação computacional para a investigação de sistemas atômicos-moleculares. A realização de simulações de dinâmica molecular requer que se conheça os potenciais de interação entre as partículas do sistema e as equações que regem o seu movimento. Desta maneira, a evolução temporal da trajetória dos átomos é examinada a fim de relacioná-la com propriedades macroscópicas. Em sistemas de interesse biológico, a aproximação realizada com o uso de potenciais clássicos não acarreta grande perda de informação, pois as energias envolvidas em interações biomoleculares são estáveis nas temperaturas dos sistemas biológicos. Neste contexto, por exemplo, a ligação entre os átomos pode ser aproximada pelo uso de potenciais harmônicos.

Os átomos estão submetidos a um potencial total, derivado do campo de força, que pode ser dividido em potenciais ligados (comprimentos de ligação, ângulos de ligação e ângulos diedros) e potenciais não-ligados (interações de van der Waals e de Coulomb), de acordo com a Equação 1.14.

$$V = V_{ligado} + V_{não-ligado} \tag{1.14}$$

Os potenciais ligados podem ser representados de acordo com a Equação 1.15.

$$V_{ligado} = \sum \kappa_r (r - r_0)^2 + \sum \kappa_\theta (\theta - \theta_0)^2 + \sum \kappa_\chi (1 + cos(n\chi - \delta)) + \sum \kappa_\varphi \left(\varphi - \varphi_0\right)^2 \tag{1.15}$$

Na Equação 1.15, o primeiro termo refere-se ao potencial harmônico para as distâncias de ligação em relação à distância de equilíbrio (r_0), o segundo refere-se à variação do ângulo entre duas ligações em relação ao ângulo de equilíbrio (θ_0), e o terceiro termo representa a variação do ângulo diedro (χ), composto pelos parâmetros n(multiplicidade) e δ (fase). O quarto termo representa a variação do diedro impróprio em relação ao diedro de equilíbrio (φ_0). Os termos $\kappa_r, \kappa_\theta, \kappa_\chi, \kappa_\varphi$ são constantes de força para cada potencial.

As interações não-ligadas são representadas pela Equação 1.16.

$$V_{não-ligado} = \sum_{i,j} 4\epsilon_{ij} \Big[\frac{\sigma_{ij}}{r_{ij}^{12}} - \frac{\sigma_{ij}}{r_{ij}^{6}} \Big] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{1.16}$$

O primeiro termo correspondente ao potencial de Lennard-Jones, compostos pelos parâmetros ϵ_{ij}, relacionado ao poço de potencial, e δ_{ij} que é derivado da distância na qual o potencial inter-partícula é zero. No potencial de Lennard-Jones a razão $\frac{1}{r^{12}}$ descreve a repulsão e $\frac{1}{r^6}$ descreve a atração. O segundo termo corresponde ao potencial de Coulomb, no qual q_i e q_j são as cargas dos átomos e ϵ_0 a constante dielétrica.

O primeiro passo da dinâmica molecular é definir as condições iniciais do sistema, *i.e.*, a posição e a velocidade inicial de cada partícula. Diversos métodos podem ser utilizados nessa etapa inicial . Todos os átomos do sistema estão contidos em uma caixa onde condições de contorno periódicas são empregadas para se evitar efeitos de fronteira e permitir a aplicação do limite termodinâmico. Uma vez definidas as condições iniciais de cada átomo, é calculada a força resultante sobre cada partícula derivada do potencial dos outros átomos (Equação 1.17).

$$F_i(t) = -\nabla V = -\frac{\partial V(r_i)}{r_i} \tag{1.17}$$

Em que $F_i(t)$ é a resultante das forças agindo na partícula i no instante t, $V(r_i)$ é a energia potencial sistema em função das coordenadas das partículas e r_i a posição da partícula. Uma vez determinada a força aplicada pelos potenciais intermoleculares e intramoleculares, o próximo passo é determinar as posições em $t > t'$. Isso é feito através da solução da equação diferencial de segundo grau (Equação 1.18).

$$m_i \ddot{r}_i = F_i \tag{1.18}$$

Em que m_i é a massa da partícula i e $\ddot{r}_i$ é a derivada de segunda ordem da posição da partícula i em relação ao tempo. A Equação 1.18 pode ser resolvida por diversos métodos. Um dos métodos comumente utilizados é o *leap-frog*. Nesta metodologia, para encontrar as posições atômicas mais um incremento δt é utilizada a relação descrita na Equação 1.19.

$$r_i(t + \delta t) = r_i(t) + \delta t \dot{r}_i\Big(t + \frac{\delta t}{2}\Big) + \mathcal{O}(\delta^3) \tag{1.19}$$

Por fim, a trajetória de cada átomo é determinada, o que permite a análise de propriedades macroscópicas. Uma das análises, é o desvio quadrático médio (RMSD, do inglês *Root-Mean-Square Deviation*), que se refere à distância média entre os átomos. O cálculo do RMSD é importante em outras técnicas de modelagem molecular como a modelagem por homologia e a docagem molecular. Nos estudos de dinâmica molecular, o RMSD permite a quantificação da distância percorrida pelos átomos, através da Equação 1.20.

$$RMSD(t) = \left[\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{r}_i(t) - \mathbf{r}_i(0)\|^2\right]^{\frac{1}{2}} \tag{1.20}$$

Onde $\mathbf{r}_i(t)$ é o vetor posição do átomo i no tempo t, $\mathbf{r}_i(0)$ a posição inicial e N o número de átomos.

No processo de planejamento de fármacos, a dinâmica molecular desempenha um importante papel, uma vez que reconhecimento molecular receptor-ligante envolve ajustes conformacionais significativos[38]. Outras aplicações da dinâmica molecular no planejamento de fármaco incluem[39] a análise de moléculas de água na interação receptor-ligante, identificação de sítio alostéricos, cálculo de energia livre de ligação e de perturbação de energia livre, e análise de *cluster* de conformações aplicadas à docagem molecular e triagem de triagem virtual.

2 OBJETIVOS

O objetivo desta dissertação de mestrado compreende o desenvolvimento de estudos de modelagem molecular em LBDD e SBDD para a identificação de aspectos moleculares importantes para a atividade biológica de uma série de compostos frente a *L. infantum*, e para o reconhecimento intermolecular ligante-LiMetRS. Os objetivos específicos incluem:

1. Desenvolver modelos de QSAR 2D, 3D e 4D para uma série de compostos sintéticos ativos contra *L. infantum*.
2. Validar os modelos de QSAR e identificar características estruturais dos compostos que sejam relevantes para a atividade biológica.
3. Gerar modelos por homologia validados para o alvo molecular LiMetRS.
4. Realizar triagens virtuais para a identificação de novos ligantes para a LiMetRS.
5. Propor nova metodologia de filtragem de resultados de triagem virtual baseada nas interações ligante-receptor.
6. Avaliar a estabilidade dos ligantes identificados no sítio de interação da LiMetRS por meio de simulações de dinâmica molecular.

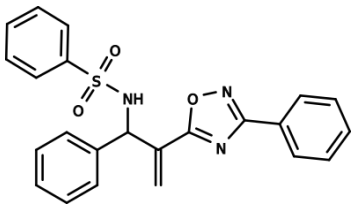
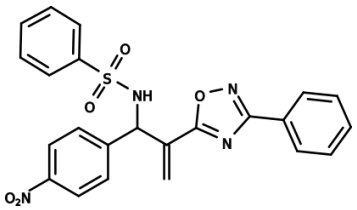
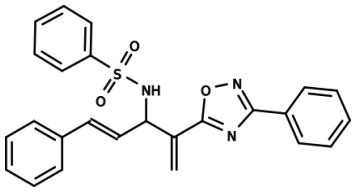
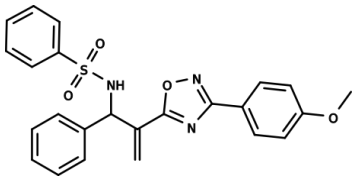
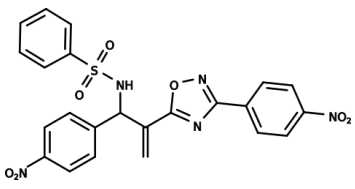
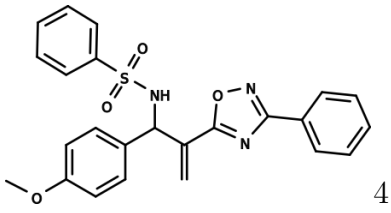
3 MATERIAIS E MÉTODOS

3.1 Estudos de LBDD

3.1.1 Conjunto de dados para os estudos de QSAR

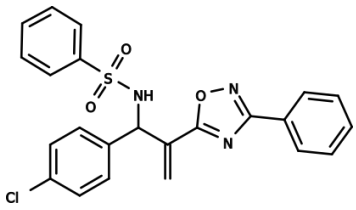
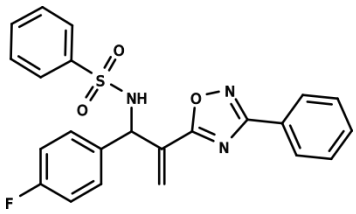
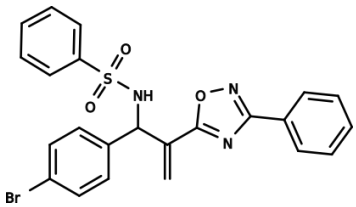
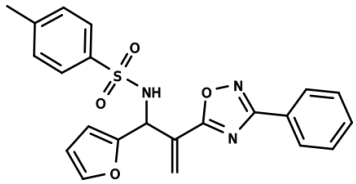
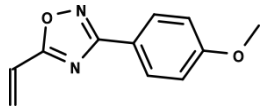
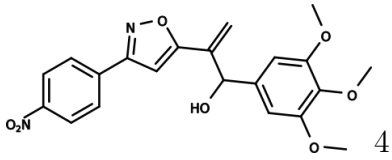
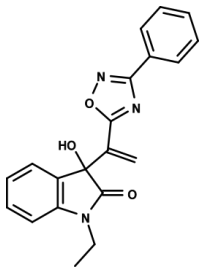
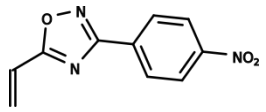
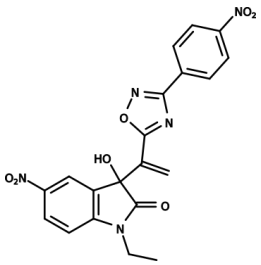
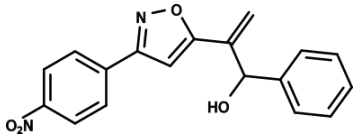
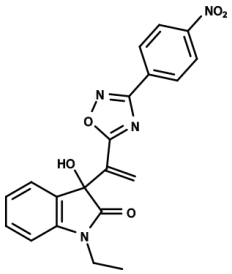
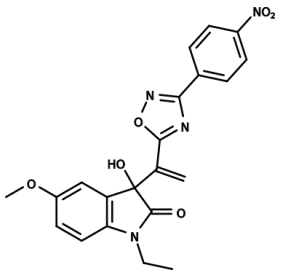
Uma série de 65 compostos ativos contra *L. infantum*, sintetizados pelo grupo do Professor Fernando Coelho (UNICAMP), e testados no Laboratório de Química Medicinal e Computacional (LQMC-IFSC-USP), foi usada para desenvolver os estudos de QSAR[40]. Os compostos foram avaliados em ensaios *in vitro* quanto a sua atividade frente a forma amastigota intracelular de *L. infantum*. A potência destes compostos foi expressa como a concentração necessária para inibir em 50% o crescimento do parasita (IC_{50}). O conjunto é majoritariamente composto por moléculas que possuem o grupo oxadiazol. A atividade contra o parasita foi determinada por meio da quantificação do número de amastigotas intracelulares em macrófagos da linhagem THP-1. Os valores de IC_{50} variam entre 2.38 µM e 52.59 µM, um intervalo de potência de cerca de 20 vezes, e foram convertidos em $pIC_{50}(-\log(IC_{50}))$ para escalonar corretamente os dados para a modelagem de QSAR. O conjunto de moléculas com suas estruturas e respectivos valores de pIC_{50} estão representados na Tabela 1.

Tabela 1: Conjunto de dados utilizado na modelagem QSAR.

Código	**Estrutura**	pIC_{50}	**Código**	**Estrutura**	pIC_{50}
1		5,033	**2**		5,115
3		5,084	**4**		4,592
5		5,081	**6**		4,976

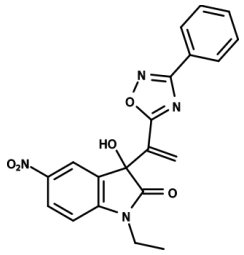
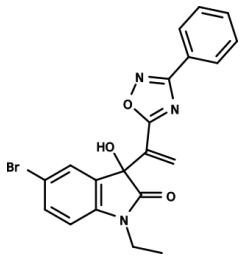
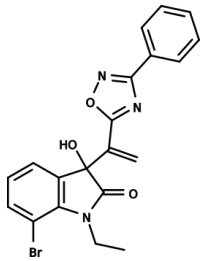
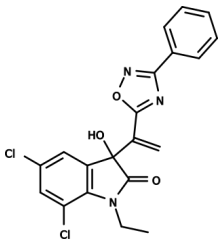
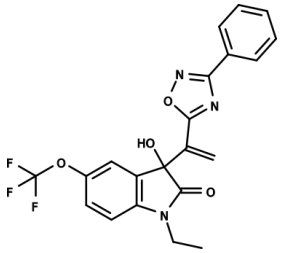
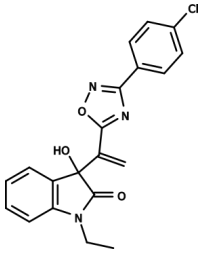
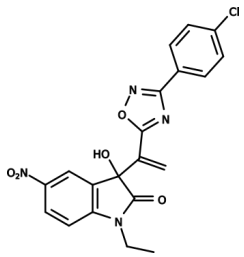
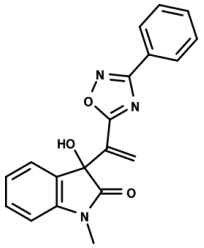
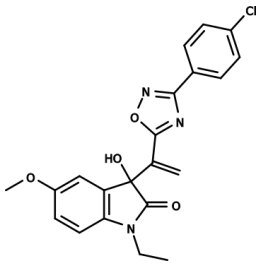
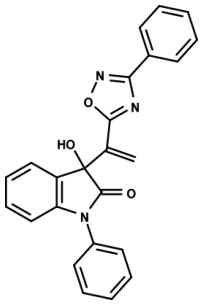
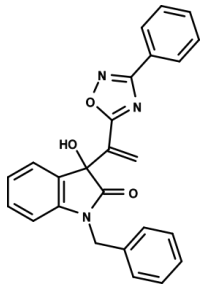
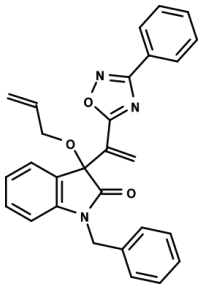
(continua)

(continuação)

7		5,096	**8**		4,932
9		5,135	**10**		4,981
11		4,598	**12**		4,279
13		5,138	**14**		4,913
15		5,133	**16**		4,426
17		5,478	**18**		4,922

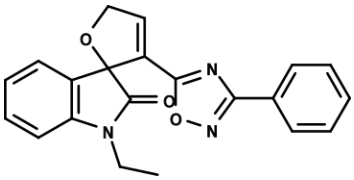
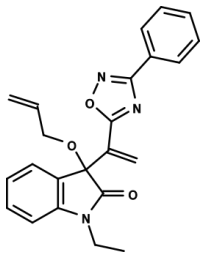
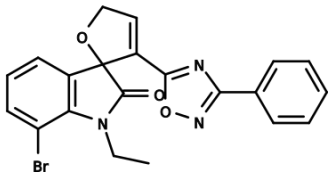
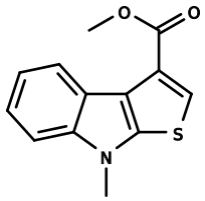
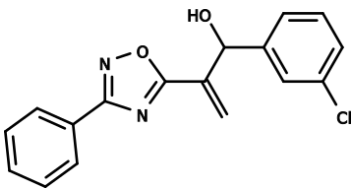
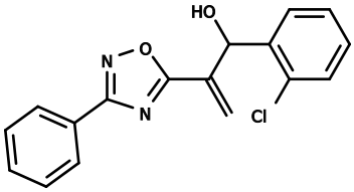
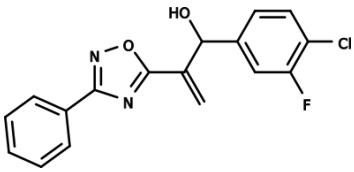
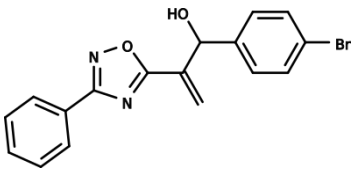
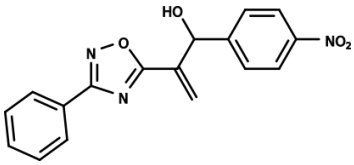
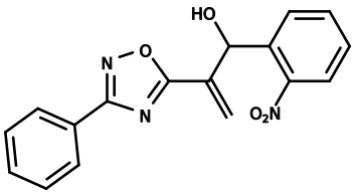
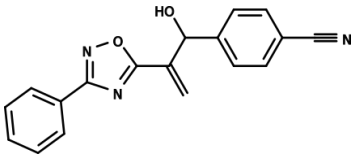
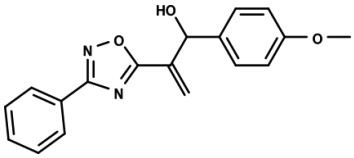
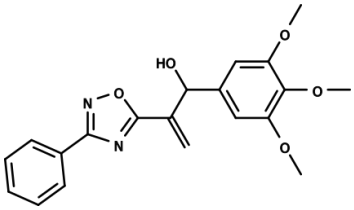
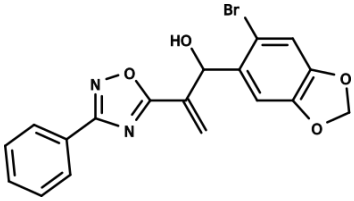
(continua)

(continuação)

19		5,29	
20		5,428	
21		4,984	
22		5,387	
23		4,955	
24		5,397	
25		5,188	
26		4,289	
27		5,293	
28		5,088	
29		5,248	
30		4,97	

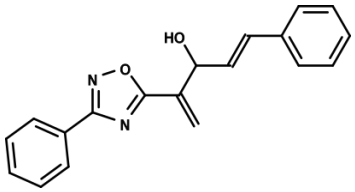
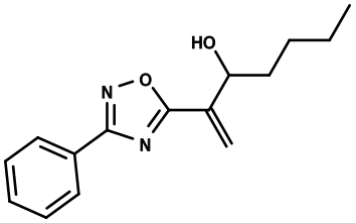
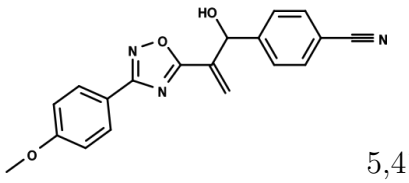
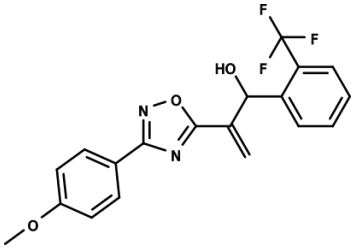
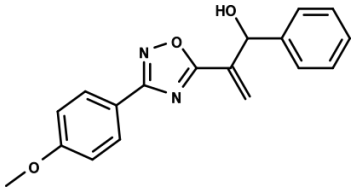
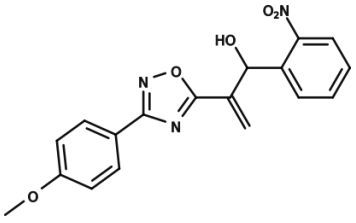
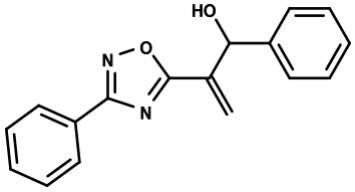
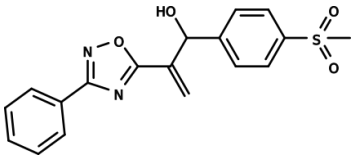
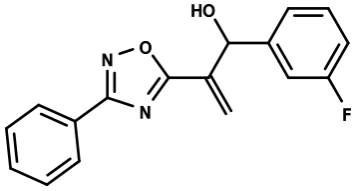
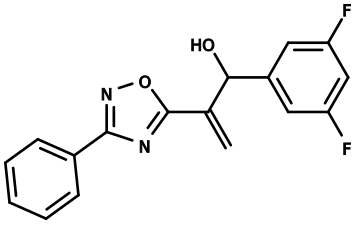
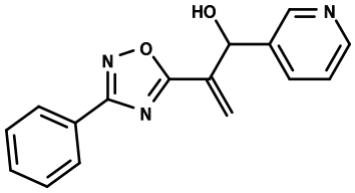
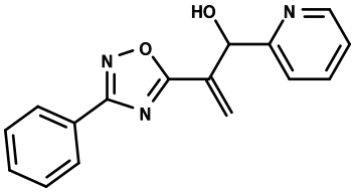
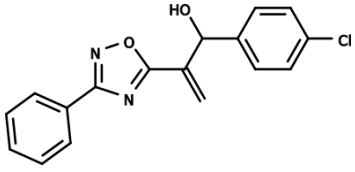
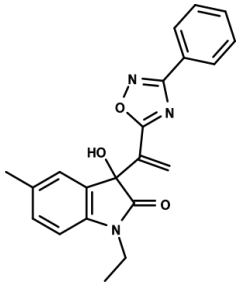
(continua)

(continuação)

31		4,931	
32		5,313
33		5,193
34		5,0783
35		5,11
36		4,755
37		4,723
38		4,358
39		4,985
40		4,988
41		4,663
42		4,744
43		4,92
44		5,049

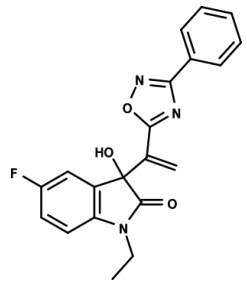
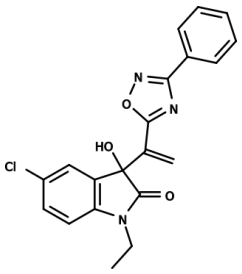
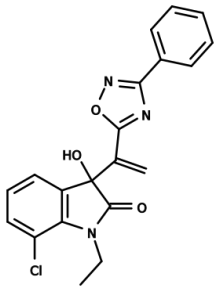
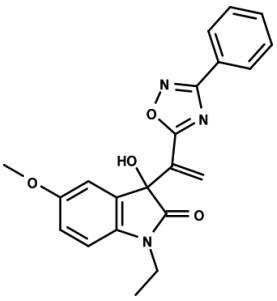
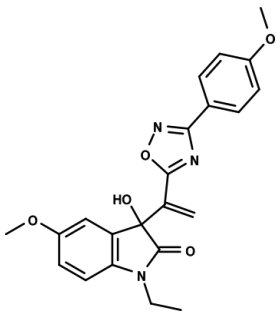
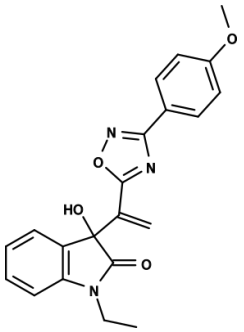
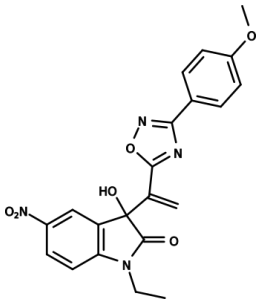
(continua)

(continuação)

45		4,687	**46** 	4,445
47		5,41	**48** 	5,068
49		4,94	**50** 	5,623
51		5,008	**52** 	5,072
53		5,137	**54** 	5,291
55		5,07	**56** 	4,747
57		5,16	**58** 	5,221

(continua)

(continuação)

59		5,455	**60**  5,602
61		5,314	**62**  5,137
63		4,658	**64**  5,545
65		4,587	

Fonte: Elaborada pelo autor.

3.1.2 QSAR-2D

O método AutoQSAR[41], disponível no pacote Maestro (Schrödinger)[42], foi utilizado para a geração dos modelos de QSAR 2D. Para a geração dos modelos de AutoQSAR, três variações de divisão entre série treinamento e série teste foram avaliadas. As proporções utilizadas entre o conjunto treinamento e teste foram: 70x30 (70% do conjunto total destinado ao conjunto treinamento e 30% ao conjunto teste), 75x25 e 80x20. O melhor

modelo foi selecionado com base em parâmetros estatísticos de validação interna (conjunto treinamento) como o coeficiente de regressão (R^2) e o desvio padrão (SD, do inglês *standard deviation*). Para o processo de validação externa, foram usados o coeficiente de regressão preditivo ($Q^2 = R^2_{pred}$) e a raiz quadrada média erro ($RMSE$) para os compostos do conjunto teste. Os melhores modelos foram recriados no módulo Canvas (Maestro, Schrödinger)[43] para a análise e visualização dos dados, com a manutenção do *fingerprint*, e conjuntos teste e treino previamente selecionados na etapa de criação e validação dos modelos.

3.1.3 *Clustering* Hierárquico

Para análise da similaridade estrutural 2D do conjunto de compostos, foram utilizados os parâmetros padrão do módulo de *clustering* hierárquico do programa Canvas. Os *fingerprints* gerados para cada composto foram submetidos ao cálculo de matriz de similaridade, a qual foi utilizada como entrada para a análise de agrupamento. O coeficiente de Tanimoto[44] foi utilizado como métrica de similaridade, e o método aglomerativo utilizado foi a média das distâncias. A seleção do número ótimo de grupos foi realizada com o índice Kelley[45]. Em seguida, selecionou-se dois grandes grupos, G_1 e G_2, com base na redução do número de grupos gerados pelo índice Kelley. Para os grupos G_1 e G_2, foram criados 10 modelos AutoQSAR, com os 7 *fingerprints* disponíveis e três combinações de série treinamento e teste. O processo gerou 210 modelos AutoQSAR. Foram selecionados os modelos que apresentaram os melhores parâmetros estatísticos de validação interna e externa para os grupos G_1 e G_2 simultaneamente. A Figura 8 representa a estratégia utilizada.

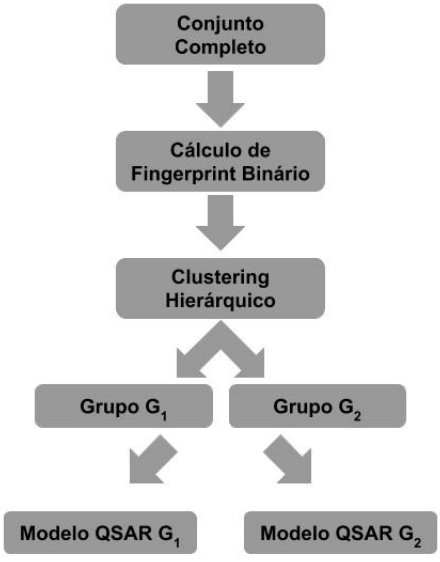


Figura 8: Esquema geral utilizado no método de *clustering*.

Fonte: Elaborada pelo autor.

3.1.4 Domínio de Aplicabilidade

Para a geração do Domínio de Aplicabilidade o *fingerprint* foi calculado e utilizado para a construção da matriz de distância pelo programa Canvas. Em seguida a matriz de distância foi reduzida em duas dimensões através do método de escalonamento multidimensional utilizando o nó de base do Knime[46]. Por fim, um código em Python foi escrito utilizando a biblioteca SciPy para gerar o polígono convexo que representa o domínio de aplicabilidade (**Apêndice A**).

3.1.5 Estrutura tridimensional e alinhamento molecular

As estruturas tridimensionais de mínima energia das moléculas utilizadas nos estudos de LBDD foram geradas pela teoria do funcional da densidade e o método híbrido B3LYP. Inicialmente, utilizou-se o programa Avogrado[47] para gerar as estruturas 3D e o *script* de entrada a ser utilizado pelo programa Gaussian[48]. Os parâmetros padrão do funcional B3LYP[49–50] com um conjunto base 6-31G(d,p) foram utilizados na otimização das estruturas pelo programa Gaussian. Três estratégias de alinhamento molecular foram empregadas neste trabalho: (i) alinhamento baseado nas estruturas de mínima energia, (ii) na máxima subestrutura comum e (iii) em grupos farmacofóricos. O alinhamento molecular foi realizado com o program SYBYL 2.1.1 (Certara). Na primeira estratégia, as moléculas do conjunto de dados foram alinhadas de forma rígida sobre os átomos dos anéis oxazol e oxadiazol. Na segunda, o alinhamento foi feito de forma flexível utilizando o composto mais potente (composto **50**) como referência. Para isso, utilizou-se a ferramenta Distill Rigid e os centros oxazóis e oxadiazóis como máxima subestrutura comum. Na terceira, utilizou-se a ferramenta GALAHAD[51] com uma população de 100 indivíduos e um número máximo de 50 gerações. Todas as propriedades farmacofóricas implementadas no GALAHAD foram utilizadas (grupos hidrofóbicos, doadores e aceptores de interações de hidrogênio, anéis de seis átomos, centros negativos e átomos de nitrogênio positivos). Como referência, utilizou-se a molécula mais potente do grupo (composto **50**).

3.1.6 QSAR-3D

Após o alinhamento o módulo CoMFA da plaforma SYBYL foi utilizado para o desenvolvimento dos modelos QSAR. Para todos os cálculos de QSAR-3D, utilizou-se as cargas GASTEIGER[52–53]. Para a geração dos modelos CoMFA, as moléculas foram posicionadas em um retículo tridimensional com espaçamento de 2Å entre as interseções e 4Å de extensão em todos os eixos além da superfície das moléculas. A energia de interação entre os átomos dos compostos e a sonda de carbono sp^3 com carga líquida +1 foi estimada com o campo de força Tripos e os potenciais de Lennard-Jones e Coulomb. O valor padrão de 30 kcal/mol foi utilizado como valor de corte para ambos os potenciais.

3.1.7 QSAR-4D

Para incorporar a variação conformacional dos ligantes nos modelos de QSAR, simulações de dinâmica molecular foram executadas. O programa GROMACS 4.6.5 [54–55] foi usado para gerar as trajetórias de dinâmica molecular. Em todas as simulações, as moléculas foram alocadas em um dodecaedro virtual, distantes 10Å das paredes, preenchido com moléculas de água explícitas TIP3. O campo de força ffG43a1[56–57] foi utilizado para gerar os potenciais. A pressão do sistema foi controlada pelo acoplamento Parrinello–Rahman[58] e a temperatura mantida constante pelo termostato de Berendsen[59]. Após a etapa de minimização, o volume do sistema foi balanceado usando um esquema de aquecimento dividido em etapas de 50K, 100K, 200K e 350K, com 20ps cada. O sistema foi então resfriado a 300K e simulado em um intervalo 500ps. A cada 1000 passos da simulação, um arquivo de trajetória foi salvo. As conformações obtidas de cada ligante foram organizadas no mesmo arquivo GRO e, para todos os ligantes, o PAC foi montado considerando as conformações registradas entre 50 e 500 ps. Esses conjuntos de conformações foram utilizados para a construção dos modelos de QSAR-4D. O alinhamento entre as moléculas foi realizado através da sobreposição das posições dos átomos dos centros oxazóis e oxadiazóis. A sonda virtual selecionada representa a unidade N-terminal NH_3^+. Os descritores com valores absolutos do coeficiente de correlação de Pearson ($|r|$) para o pIC_{50} menores que 0,2[24,60] e aqueles que apresentavam variância abaixo de 0,01 entre os compostos foram excluídos. Por fim, foi utilizado o corte de energia para descritores de Lennard-Jones:

$$LJ_{x,y,z} < 30 \text{ kcal/mol} \rightarrow LJ_{x,y,z} = LJ_{x,y,z}$$
$$LJ_{x,y,z} \geq 30 \text{ kcal/mol} \rightarrow LJ_{x,y,z} = 30 + \log\Big(LJ_{x,y,z} - 30\Big)$$

Os descritores remanescentes foram submetidos à seleção de variáveis pelo algoritmo PyQSAR[61]. Em seguida, 200 modelos foram criados variando aleatoriamente as moléculas destinadas ao conjunto teste e treinamento, mantendo a proporção 80x20. Foram utilizados os seguintes parâmetros para o algoritmo PyQSAR: *learning* = 500, *bank* = 10, *component* = 5. O melhor modelo foi selecionado a partir dos parâmetros estatísticos.

3.2 Estudos de SBDD

3.2.1 Obtenção da estrutura 3D por modelagem por homologia

A estrutura primária da LiMetRS foi obtida do banco de dados UniProt[62] (código A4HZ82) e a busca do *template* feita no banco de dados PDB[63] por meio da ferramenta *"Search by Sequences"*. A modelagem por homologia baseada na satisfação de restrições espaciais foi realizada com o programa MODELLER[64] versão 9.22, e com a estratégia de manutenção do ligante no sítio de ligação. O alinhamento foi realizado usando o comando *aling2d* no MODELLER. A extração e a satisfação das restrições de distância entre o ligante e a proteína foram obtidas automaticamente pelo MODELLER com base nas

interações entre o ligante e o *template*. A validação do modelo foi realizada através do gráfico de Ramachandran com o programa PROCHEK[65] que fornece as informações sobre a qualidade estereoquímica do modelo, a pontuação DOPE (do inglês *Discrete Optimized Protein Energy*), e o RMSD (do inglês *Root-Mean-Square Deviation*) entre o *template* e o modelo gerado com auxílio do programa Pymol versão 1.8.4.0[66].

3.2.2 Triagem virtual por docagem molecular

Na primeira etapa de triagem virtual, foram utilizados os seguintes filtros moleculares sobre o banco de dados ZINC15[67]: peso molecular de 300 a 450 Da e coeficiente de partição ($\log P$) de 3 a 5. Esses filtros foram definidos com base nas propriedades moleculares de inibidores da MetRS de *Trypanosoma brucei* (TbMetRS). Além destas propriedades, utilizou-se o subconjunto de compostos disponíveis para compra e não interferentes (PAINS, do inglês *Pan-Assay Interference Compounds*).

A coleção de compostos resultantes da primeira etapa de triagem virtual foi usada em uma segunda etapa de triagem baseada na estrutura do receptor. Para a docagem molecular dos compostos, foi utilizado o programa GOLD[37] versão 2020 e a função de pontuação GoldScore. O sítio de ligação foi definido como uma esfera de raio 10Å em torno do ligante. Inicialmente, o parâmetro de docagem do algoritmo genético foi mantido com o valor 1 e a eficiência da busca foi selecionada em 10% Os parâmetros de flexibilidade padrão foram mantidos para o ligante. Os 10.000 compostos de maior pontuação foram selecionados para uma próxima etapa de docagem. Nesta fase, o parâmetro de docagem foi aumentado para 10 e a eficiência da busca para 100%.

3.2.3 Análise Pós-Triagem Virtual

Um algoritmo para a análise das interações intermoleculares ligante-receptor foi desenvolvido neste trabalho. O algoritmo foi escrito em Python versão 2.7 e utiliza os softwares de código aberto Pymol 1.8.4.0 e PLIP v1.4.5[68] como ferramentas externas. O código está descrito no **Apêndice B**. A interface (API, do inglês, *Application Programming Interface*) do Pymol foi utilizada para importar as estruturas dos ligantes e do alvo molecular. Os complexos ligante-receptor foram exportados no formato .pdb e utilizados como entrada para o programa PLIP. Os 10.000 compostos resultantes das etapas de docagem molecular foram avaliados. Interações de hidrogênio com o resíduo Asp50 e interações π-*stacking* com a His52 ou Gly53 ou Tyr235 ou Val236, foram utilizadas como critérios de seleção[69].

3.2.4 Dinâmica Molecular

As moléculas selecionadas ao final da triagem virtual foram submetidas à simulações de dinâmica molecular em complexo com a LiMetRS, realizadas com o programa

GROMACS. A conformação obtida pela docagem molecular foi utilizada como conformação inicial nos estudos de dinâmica molecular. Foram geradas três simulações para a molécula escolhida. Foi utilizada uma caixa de solvente cúbica, preenchida com moléculas de água explícitas TIP3P, e com uma distância de 10Å entre a enzima e as paredes da caixa. Foi utilizado o campo de força CHARMM[70−71] e a topologia de ligante foi gerada com o programa CGenFF[72]. A minimização de energia foi realizada com um máximo de 50.000 passos. As fases de equilibração e contenção foram executadas com o *ensemble* isotérmico-isocórico (NVT) e o *ensemble* isotérmico-isobárico (NPT) com 500 ps e 1000 ps, respectivamente. O controle de temperatura foi mantido pelo termostato de redimensionamento de velocidade[73] em 300k e a pressão de referência em 1bar com o método de Berendsen[74]. As restrições de comprimento de ligação foram determinadas com o algoritmo LINCS[75]. O corte de van der Waals de curto alcance foi definido em 1,2nm e as interações eletrostáticas de longo alcance foram realizadas com esquema de malha de partículas de Ewald[76] (PME, do inglês *Particle Mesh Ewald*) com espaçamento de grade de 0,16nm. Simulações de 50 ns foram realizadas com a velocidade do equilíbrio NPT.

4 RESULTADOS E DISCUSSÃO

4.1 Estudos de LBDD

4.1.1 QSAR-2D

O conjunto de dados utilizado na construção dos modelos de QSAR demonstra uma boa distribuição da atividade biológica como representado na Figura 9. Desta forma, este conjunto de moléculas foi utilizado, inicialmente, na construção de modelos QSAR-2D. Estes estudos permitem identificar descritores bidimensionais relevantes para a atividade biológica. Dentre os descritores 2D, destacam-se os parâmetros topológicos, os fragmentos moleculares, e os parâmetros físico-químicos[77].

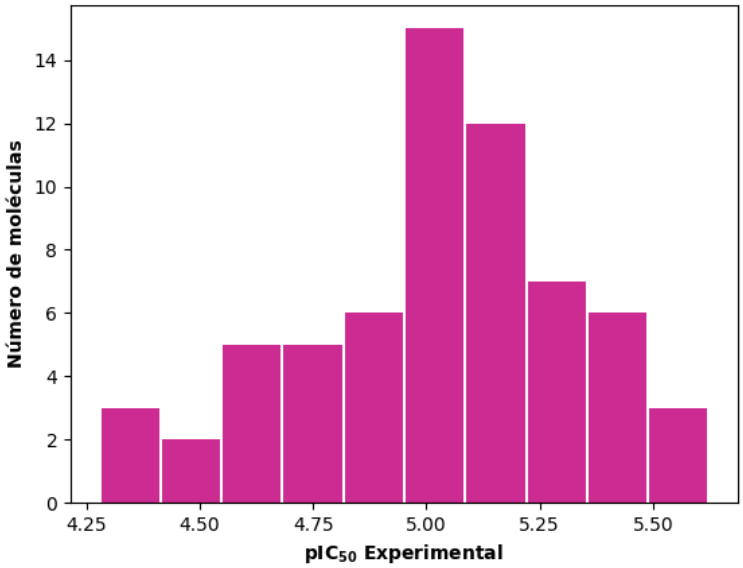


Figura 9: Histograma da distribuição dos valores de pIC_{50} do conjunto de dados utilizado na construção dos modelos de QSAR.

Fonte: Elaborada pelo autor.

Um primeiro modelo de QSAR-2D (**modelo 1**) foi construído utilizando o conjunto completo de 65 moléculas como entrada para o programa AutoQSAR. Este programa utiliza algoritmos de aprendizado de máquina para construção de modelos QSAR. Os seguintes descritores do tipo impressão digital (*fingerprints*) foram empregados para codificar as estruturas bidimensionais das moléculas: *dendritic*, *linear*, *atom pair*, *atom triplet*, *topological*, *MOLPRINT 2D*, e *radial*. A seleção dos conjuntos teste e treinamento foi feita através da ferramenta AutoQSAR. A proporção de compostos destinados ao conjunto teste e conjunto treinamento foi variada a fim de se obter a seleção capaz

de gerar os melhores modelos de QSAR-2D. O conjunto treinamento foi submetido a diversos métodos de regressão como regressão linear múltipla (MLR), regressão de mínimos quadrados parciais (PLS), regressão de componentes principais (PCR) e PLS baseado em kernel (KPLS). Cada modelo gerado foi avaliado quanto a sua robustez e capacidade preditiva através dos processos de validação interna e externa. Os resultados são exibidos na Tabela 2.

Tabela 2: Melhores modelos de QSAR-2D obtidos com o conjunto completo (**modelo 1**).

Conjunto Treino (%)	$\mathbf{R^2}$	**SD**	$\mathbf{Q^2}$	**RMSE**	**N**	**Fingerprint**
70	0,6256	0,1977	0,5258	0,2014	2	MOLPRINT 2D
75	0,5353	0,2158	0,5815	0,1938	1	Radial
80	0,6152	0,1988	0,5635	0,1932	2	Radial

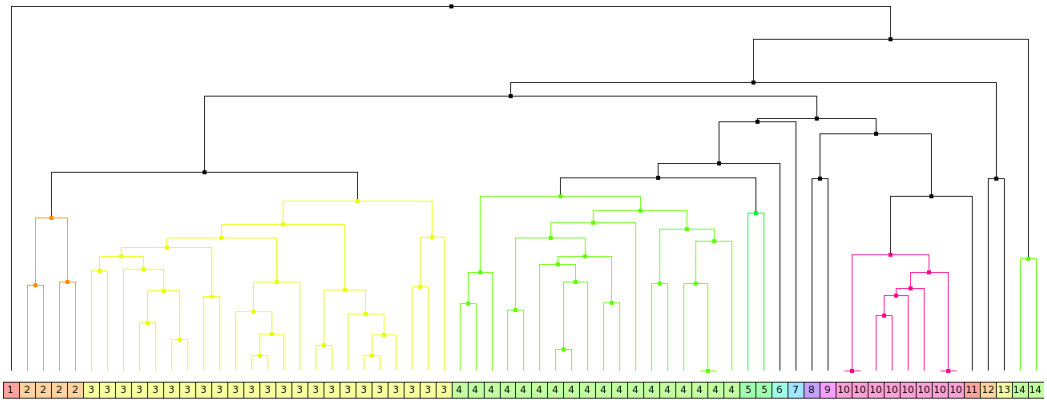
R^2: Coeficiente de determinação para o conjunto de treinamento; SD: desvio padrão; Q^2: coeficiente de correlação preditiva para o conjunto de teste (R^2_{pred}); $RMSE$: Raiz do erro quadrático médio para as previsões do conjunto de teste; N: número ideal de componentes

Fonte: Elaborada pelo autor.

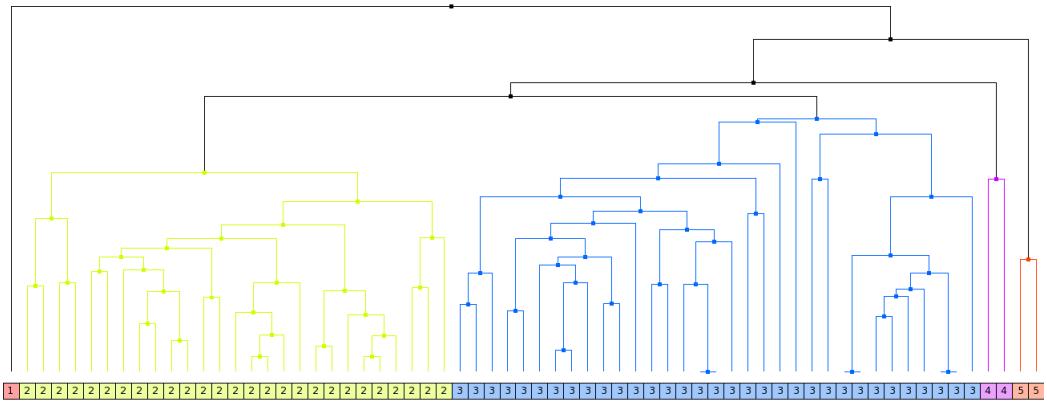
Nesta primeira etapa, o modelo que apresentou os melhores parâmetros estatísticos foi aquele gerado pelo *fingerprint* radial, método de regressão KPLS e razão 80x20 (52 moléculas no conjunto treinamento e 13 no conjunto teste). Este modelo apresentou os seguintes indicadores estatísticos: $R^2 = 0,6152$, SD= 0,1988, $Q^2 = 0,5635$, RMSE= 0,1932 e *score*=0,5996. Esses resultados apresentam parâmetros satisfatórios para construção e interpretação de modelos de QSAR. No entanto, com o intuito de se gerar modelos com maior capacidade preditiva, empregou-se a estratégia de *clustering* a fim de se otimizar parâmetros estatísticos. De maneira geral, os métodos de *clustering* visam agrupar dados de acordo com o seu nível de semelhança. Desta forma, esta técnica foi aplicada ao conjunto de dados para se investigar sua diversidade estrutural através do seu agrupamento em diferentes *clusters*.

Métodos de clustering tem sido implementados em outros trabalhos[78–79] , e tem sido eficientes em otimizar parâmetros estatísticos em modelos de QSAR. Através da análise de agrupamento e da utilização do índice Kelley, foi possível notar a presença de grupos de apenas uma molécula (*singletons*), o que indica a boa diversidade estrutural destes conjuntos de dados[78]. A Figura 10(a) apresenta o dendograma gerado com a utilização do índice Kelley e com o *fingerprint* MOLPRINT 2D para o conjunto de dados usado neste trabalho. Há 14 *clusters* sendo 8 *singletons*, o que demonstra a boa diversidade estrutural deste conjunto de moléculas. Por sua vez, a figura Figura 10(b) apresenta o dendograma gerado após a implementação da estratégia de *clustering*. O método de *clustering* resultou na divisão de dois grupos de moléculas, G_1 e G_2. Este agrupamento resultou da utilização do *fingerprint atom pair*, o qual levou aos melhores modelos de QSAR. As estruturas

representativas de cada grupo podem ser observadas na Figura 11.



(a)



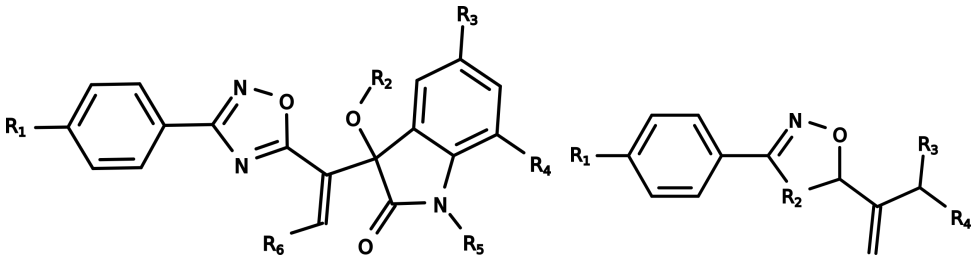
(b)

Figura 10: Dendogramas utilizados para a identificação dos grupos G_1 e G_2.(a) Dendograma gerado com a utilização do índice Kelley. (b) Dendograma gerado após a aplicação da estratégia de *clustering*.

Fonte: Elaborada pelo autor

Este agrupamento resultou na exclusão de 3 moléculas (**11**, **14** e **34**), totalizando um conjunto de 62 moléculas divididas em dois grupos. As moléculas excluídas podem ser consideradas *outliers* estruturais. A partir da geração de uma matriz de distância e do escalonamento multidimensional, é possível reduzir a dimensionalidade do conjunto de dados e representar a dispersão estrutural em um plano cartesiano (Figura 12). A caracterização das estruturas representativas foi feita através de uma busca automatizada pela máxima subestrutura comum e da análise dos grupos R.

É possível fazer algumas observações a partir do gráfico de dispersão na Figura 12. O conjunto de moléculas G_2 possui maior diversidade estrutural quando comparado ao conjunto G_1. Como pode ser observado na Figura 11, o grupo G_1 apresenta um padrão estrutural mais complexo e menor variação dos substituintes, ao passo que o grupo G_2 apresenta um padrão estrutural mais simples e maior variabilidade nos substituintes. Outra



(a) Groupo G_1 (b) Groupo G_2

Figura 11: Máxima subestrutura comum do Grupo G_1 e Grupo G_2.

Fonte: Elaborada pelo autor

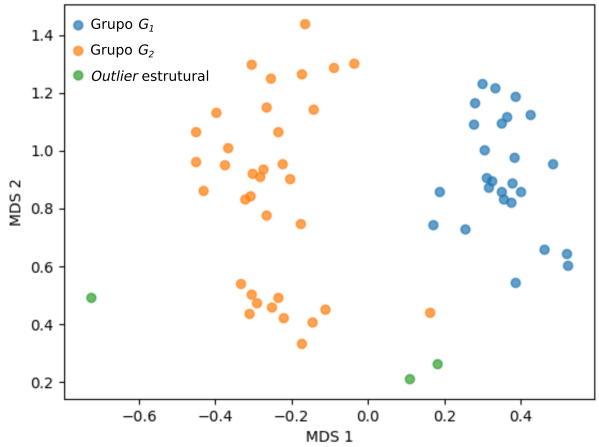


Figura 12: Gráfico de dispersão do conjunto de 65 moléculas derivado da matriz de distância e do escalonamento multidimensional.

Fonte: Elaborada pelo autor.

observação que pode ser feita a partir da análise de *cluster* é a presença do composto (**34**) com o núcleo indol, o qual é estruturalmente mais distante dos outros compostos. Diante disso, foram construídos modelos de QSAR-2D sem o composto **34** (**modelo 2**). A exclusão desse composto levou aos resultados apresentados na Tabela 3.

Tabela 3: Melhores modelos de QSAR-2D derivados com a exclusão da molécula **34** (**modelo 2**).

Conjunto Treino (%)	**R^2**	**SD**	**Q^2**	**RMSE**	**N**	**Fingerprint**
70	0,5378	0,2180	0,4937	0,2123	1	Radial
75	0,5997	0,2065	0,5284	0,1994	2	Dendritic
80	0,6304	0,2022	0,6107	0,1817	5	MOLPRINT 2D

Fonte: Elaborada pelo autor.

É possível constatar que a exclusão desse composto resultou em melhora discreta

dos indicadores. Para o modelo de QSAR com todos os compostos foram obtidos os valores de $R^2 = 0,6152$ e $Q^2 = R^2_{pred} = 0,5635$. O modelo construído após exclusão da molécula **34** resultou em parâmetros ligeiramente melhores ($R^2 = 0,6304$ e $Q^2 = R^2_{pred} = 0,6107$).

Dessa forma, adotou-se a estratégia de se gerar modelos específicos para os grupos G_1 (**modelo 3**) e G_2 (**modelo 4**). Cada grupo resultante da análise de agrupamento (o grupo G_1 com 27 compostos e o grupo G_2 com 35 compostos) foi submetido a construção de modelos de AutoQSAR de maneira independente. Os melhores modelos para os grupos G_1 e G_2 estão apresentados na Tabela 4 e na Tabela 5, respectivamente.

Tabela 4: Melhores modelos de AutoQSAR gerados para Grupo G_1 (**modelo 3**).

Conjunto Treino (%)	**R^2**	**SD**	**Q^2**	**RMSE**	**N**	**Fingerprint**
70	0,8982	0,1178	0,7132	0,1018	2	Radial
75	0,8012	0,1413	0,7022	0,1668	1	Radial
80	0,9069	0,1039	0,8201	0,0945	2	Radial

Fonte: Elaborada pelo autor.

O melhor modelo para o grupo G_1 foi aquele produzido pelo *fingerprint* radial e pela proporção 80x20 para a divisão das séries treinamento e teste. Este modelo gerou valores significativamente melhores de R^2 (0,9069) e $Q^2 = R^2_{pred}$ (0,8201).

Tabela 5: Melhores modelos gerados por AutoQSAR para Grupo G_2 (**modelo 4**).

Conjunto Treino (%)	**R^2**	**SD**	**Q^2**	**RMSE**	**N**	**Fingerprint**
70	0,6109	0,2050	0,4206	0,1829	2	MOLPRINT 2D
75	0,5693	0,2040	0,5351	0,1041	2	MOLPRINT 2D
80	0,8206	0,1377	0,8001	0,1081	3	Dendritic

Fonte: Elaborada pelo autor.

Para o grupo G_2, o modelo que produziu os melhores parâmetros estatísticos é aquele gerado pelo *fingerprint* dendritic e 28 moléculas no conjunto treinamento e 7 no conjunto teste (proporção 80x20). Foram obtidos valores de $R^2 = 0,8206$ e $Q^2 = R^2_{pred} = 0,8001$. Os valores experimentais e preditos para o conjunto de dados completo e para os grupos G_1 e G_2 estão listados na Tabela 6. A diferença entre os valores experimentais e preditos (resíduo) indica a boa capacidade preditiva dos modelos gerados. A representação gráfica dos valores de pIC_{50} preditos versus os valores experimentais para os grupos G_1 e G_2 é apresentada na Figura 13.

É interessante notar que apesar de ambos os grupos compartilharem uma estrutura comum (3-fenil-4,5-dihidro-1,2-oxazol ou 3-fenil-1,2,4-oxadiazol), a geração de um modelo

de QSAR único abrangendo os grupos G_1 e G_2 não resultou em modelos superiores. O grupo de 65 moléculas não apresentou os melhores resultados de R^2 e Q^2 em comparação com os modelos gerados após a análise de agrupamento. Dessa forma, é possível concluir que há um limite para que a diversidade estrutural apresentada por este conjunto de moléculas resulte em modelos de QSAR mais robustos. Entre os grupos G_1 e G_2, essa comparação também pode ser observada. O grupo G_2 apresenta maior diversidade estrutural, no entanto, apresenta parâmetros estatísticos ligeiramente menores quando comparado ao grupo G_1, que possui menor variabilidade estrutural. $R^2_{G_1} > R^2_{G_2}$, da mesma forma que $Q^2_{G_1} > Q^2_{G_2}$.

Tabela 6: Valores experimentais, preditos e residuais de pIC_{50} para os modelos de QSAR-2D para todo o conjunto de dados e para os grupos G_1 e G_2 definidos pela análise de *cluster*.

		modelo 2		**Cluster**		
Inibidor	$\mathbf{pIC_{50_{exp}}}$	$\mathbf{pIC_{50_{pred}}}$	**resíduo**	**Grupo**	$\mathbf{pIC_{50_{pred}}}$	**resíduo**
13	5,138	5,184	0,046	G_1	5,187	0,049
14	4,913	4,911	-0,002	–[1]	–[1]	–[1]
15	5,133	4,962	-0,171	G_1	5,061	-0,072
17	5,478	5,312	-0,165	G_1	5,433	-0,044
18	4,922	4,955	0,033	G_1	4,995	0,073
19	5,29	5,017	-0,273	G_1	5,173	-0,116
20	5,428	5,231	-0,197	G_1	5,385	-0,043
21	4,984	5,103	0,119	G_1	4,993	0,009
22	5,387	5,47	0,082	G_1	5,388	-0,001
23	4,955	4,904	-0,051	G_1	5,098	0,143
24	5,397	5,467	0,07	G_1	5,554	0,157
25	5,188	5,232	0,043	G_1	5,282	0,093
26	4,289	4,848	0,559	G_1	4,369	0,080
27	5,293	5,158	-0,135	G_1	5,345	0,052
28	5,088	4,848	-0,24	G_1	4,807	-0,281
29	5,248	4,848	-0,4	G_1	5,104	0,144
30	4,97	5,007	0,037	G_1	4,97	0,000
31	4,931	5,141	0,209	G_1	4,994	0,062
32	5,313	5,333	0,019	G_1	5,235	-0,079
33	5,193	5,062	-0,132	G_1	5,126	-0,067
58	5,221	5,333	0,112	G_1	5,245	0,024
59	5,455	5,543	0,088	G_1	5,493	0,038
60	5,602	5,53	-0,072	G_1	5,493	-0,109

(continua)

(continuação)

		modelo 2		Cluster		
Inibidor	**$pIC_{50_{exp}}$**	**$pIC_{50_{pred}}$**	**resíduo**	**Grupo**	**$pIC_{50_{pred}}$**	**resíduo**
61	5,314	5,185	-0,13	G_1	5,264	-0,050
62	5,137	4,902	-0,235	G_1	5,182	0,044
63	4,658	4,86	0,202	G_1	4,716	0,058
64	5,545	5,152	-0,393	G_1	5,47	-0,075
65	4,587	4,882	0,295	G_1	4,651	0,064
1	5,033	4,812	-0,221	G_2	4,856	-0,177
2	5,115	4,982	-0,133	G_2	5,146	0,031
3	5,084	5,094	0,01	G_2	5,12	0,036
4	4,592	4,785	0,193	G_2	4,841	0,249
5	5,081	4,982	-0,099	G_2	5,112	0,031
6	4,976	4,785	-0,192	G_2	4,907	-0,069
7	5,096	5,074	-0,022	G_2	5,269	0,173
8	4,932	4,883	-0,049	G_2	4,891	-0,041
9	5,135	4,857	-0,278	G_2	4,87	-0,265
10	4,981	4,992	0,011	G_2	4,926	-0,055
11	4,598	4,758	0,16	$-^1$	$-^1$	$-^1$
12	4,279	4,373	0,094	G_2	4,283	0,004
16	4,426	4,35	-0,076	G_2	4,34	-0,086
35	5,11	5,137	0,027	G_2	5,163	0,053
36	4,755	4,909	0,155	G_2	4,644	-0,110
37	4,723	4,83	0,107	G_2	4,595	-0,128
38	4,358	4,736	0,378	G_2	4,602	0,244
39	4,985	5,014	0,03	G_2	4,974	-0,011
40	4,988	5,373	0,385	G_2	5,186	0,198
41	4,663	4,874	0,212	G_2	4,868	0,205
42	4,744	4,925	0,181	G_2	4,728	-0,016
43	4,92	5,028	0,108	G_2	4,893	-0,027
44	5,049	5,05	0,001	G_2	5,092	0,043
45	4,687	4,753	0,067	G_2	4,828	0,141
46	4,445	4,608	0,164	G_2	4,456	0,011
47	5,41	4,933	-0,477	G_2	5,297	-0,113
48	5,068	4,916	-0,152	G_2	5,07	0,002
49	4,94	4,925	-0,015	G_2	4,978	0,038
50	5,623	5,444	-0,18	G_2	5,455	-0,168
51	5,008	4,874	-0,134	G_2	4,908	-0,100
52	5,072	5,094	0,022	G_2	4,871	-0,201

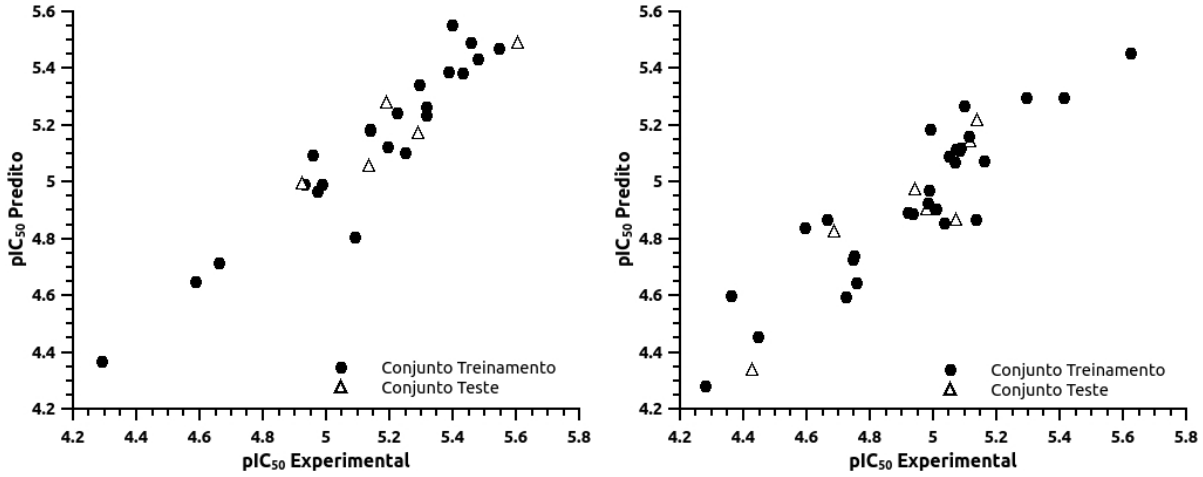
(continua)

(continuação)

		modelo 2		Cluster		
Inibidor	**$pIC_{50_{exp}}$**	**$pIC_{50_{pred}}$**	**resíduo**	**Grupo**	**$pIC_{50_{pred}}$**	**resíduo**
53	5,137	5,193	0,055	G_2	5,219	0,081
54	5,291	5,269	-0,022	G_2	5,299	0,008
55	5,07	4,875	-0,196	G_2	5,115	0,045
56	4,747	4,886	0,139	G_2	4,741	-0,007
57	5,16	5,12	-0,04	G_2	5,077	-0,083

[1] *outlier* estrutural

Fonte: Elaborada pelo autor.



(a) Grupo G_1 (b) Grupo G_2

Figura 13: Gráfico de pIC_{50} predito *versus* experimental para conjuntos de treinamento e teste do Grupo G_1 e Grupo G_2.

Fonte: Elaborada pelo autor.

Além da predição dos valores de pIC_{50}, os modelos de QSAR são capazes de fornecer informações sobre quais características estruturais contribuem de forma mais relevante para a variação da atividade biológica. Esses mapas de contribuição, gerados pelo módulo KPLS, são capazes de indicar regiões que contribuem positivamente ou negativamente para a propriedade biológica. As contribuições negativas são representadas pelas cores vermelho, e as contribuições positivas são representadas em verde (Figura 14).

Para o grupo G_1, a substituição por halogênios nos grupos R_1, R_3, R_4 e R_5 contribui positivamente para a atividade biológica. As moléculas **20** ($pIC_{50} = 5,428$), **59** ($pIC_{50} = 5,455$) e **60** ($pIC_{50} = 5,602$) possuem como única diferença, a substituição por bromo, flúor e cloro no grupo R_3, respectivamente. Neste caso, todas as substituições

apresentaram contribuições positivas, porém, o cloro é o substituinte que resulta na maior atividade (molécula **60**, $\text{pIC}_{50} = 5,602$). Em geral, substituições pelo grupo nitro foram desfavoráveis para as moléculas do grupo G_1. Para o grupo G_2, a ausência de substituintes na posição R_1 apresentou contribuição desfavorável. Para estas moléculas, o grupo fenil e, em especial, a substituição pelo grupo metoxi na posição R_1 contribuíram positivamente para a atividade biológica. Para moléculas com o núcleo oxazol, o grupo nitro em R_1, excepcionalmente, contribui positivamente. No grupo R_4, halogênios (exceto o bromo) e o grupo nitro ligados à fenila apresentam contribuição positiva. Anéis aromáticos com dois halogênios nas posições meta apresentam aumento da atividade em comparação aos análogos mono-susbstituídos, como é o caso dos compostos **53** ($\text{pIC}_{50} = 5,137$) e **54** ($\text{pIC}_{50} = 5,291$). Para o grupo R_3, o grupo hidroxila apresentou contribuição positiva enquanto a benzenosulfonamida apresentou contribuição negativa (composto **2**, $\text{pIC}_{50} = 5,115$).

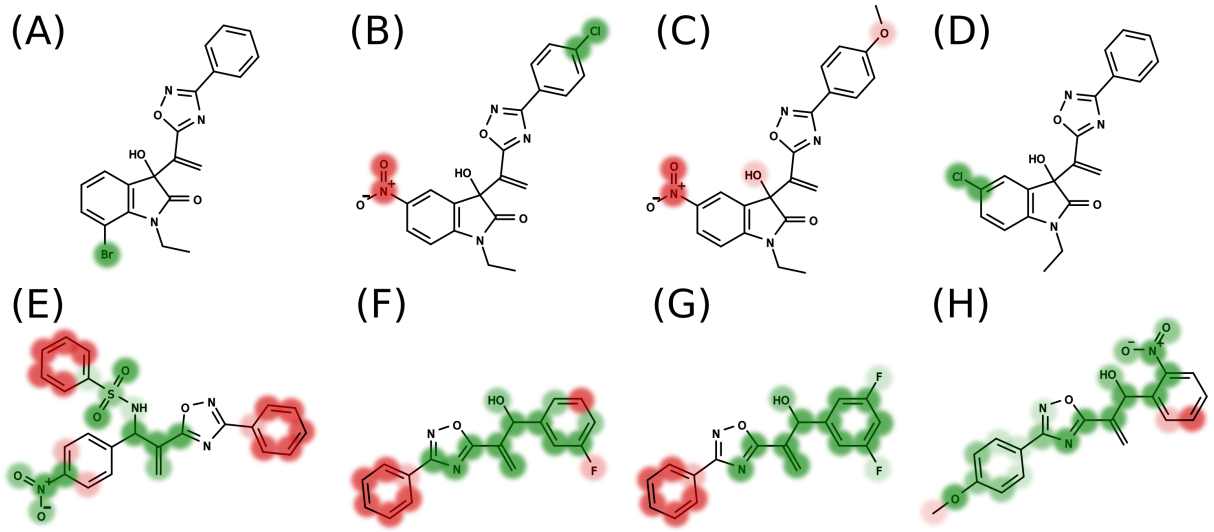
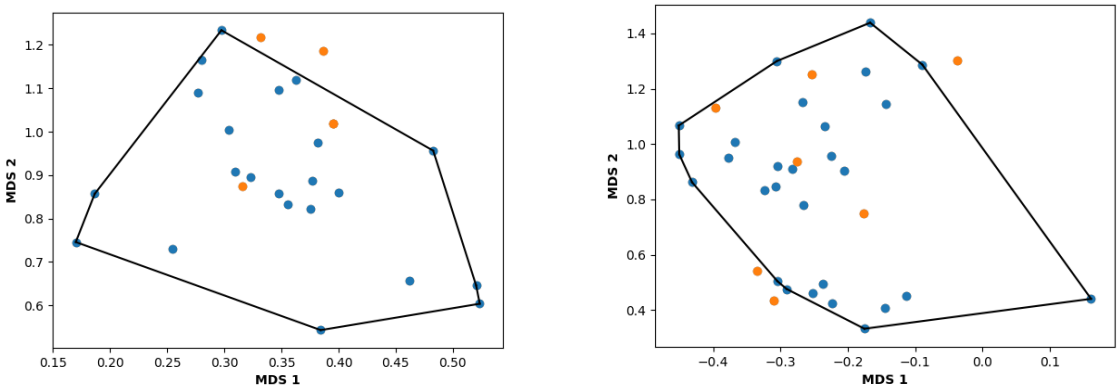


Figura 14: Mapas de contribuição gerados pelo método KPLS para as moléculas do grupo G_1 (A-D) e G_2 (E-H). Grupo G_1: (a) **21**; (b) **25**; (c) **65**; (d) **60**. Grupo G_2: (e) **2**; (f) **53**; (g) **54**; (h) **50**.

Fonte: Elaborada pelo autor.

Um requerimento essencial para os modelos de QSAR é a definição do domínio de aplicabilidade, o qual define o espaço estrutural para o qual as predições realizadas para o conjunto treinamento são válidas. Na Figura 15 os pontos azuis representam as moléculas utilizadas como conjunto treinamento e os pontos laranjas representam as moléculas utilizadas no conjunto teste. A linha preta limita a região onde o modelo é aplicável.



(a) Grupo G_1 (b) Grupo G_2

Figura 15: Domínio de Aplicabilidade definido pelo método *convex-hull*.

Fonte: Elaborada pelo autor.

4.1.2 QSAR-3D

Os métodos de QSAR-3D permitem a avaliação de como as conformações e descritores moleculares tridimensionais afetam a atividade biológica de uma série de compostos. Exemplos de descritores 3D são os campos moleculares eletrostáticos e estereoquímicos. Para gerar as conformações tridimensionais dos compostos foi utilizado o método DFT/B3LYP. Esta metodologia foi selecionada por ter demonstrado resultados superiores a métodos *ab initio* e semiempíricos quando geometrias e energias moleculares são consideradas[80]. Após a geração da estrutura tridimensional, as moléculas foram submetidas a 3 métodos diferentes de alinhamento molecular. A estratégia de *clustering* utilizada nos estudos de QSAR-2D também foi aplicada aos modelos CoMFA. Portanto, construiu-se um modelo de QSAR-3D com o conjunto completo (**modelo 5**) e para os grupos G_1 (**modelo 6**) e G_2 (**modelo 7**). A proporção entre o conjunto teste e treinamento foi mantida em 80x20, considerando que esta divisão resultou nos melhores parâmetros estatísticos (R^2 e Q^2) para os modelos de QSAR-2D. Para cada conjunto, 10 modelos foram criados variando-se os compostos nos conjuntos teste e treinamento. Para todos os modelos, parâmetros estatísticos insatisfatórios foram obtidos, o que pode ser atribuído a alinhamentos moleculares muito divergentes da conformação bioativa dos compostos. Os melhores modelos resultantes de cada uma das estratégias de alinhamento molecular são apresentados na Tabela 7.

Apesar dos modelos QSAR-3D apresentarem bons valores de R^2, os valores de Q^2 foram muito baixos. Diferenças maiores do que 0,2 – 0,3 unidades entre R^2 e Q^2 podem ser um indício de sobre-ajuste dos dados[81]. Os resultados obtidos indicam sobre-ajuste, ou seja, os modelos estatísticos se ajustam muito bem ao conjunto treinamento, mas se

mostram ineficazes para predizer o resultado de novas ocorrências.

Tabela 7: Melhores modelos de QSAR-3D.

Conjunto	Alinhamento	Proporção	R^2	Q^2	SD	Componentes
Completo	Mínima energia	80x20	0,686	0,335	0,191	3
Completo	MCS[1]	80x20	0,645	0,246	0,197	3
Completo	Farmacóforo	80x20	0,712	0,118	0,171	4
G_1	Mínima energia	80x20	0,794	0,175	0,156	3
G_1	MCS[1]	80x20	0,799	0,293	0,128	3
G_1	Farmacóforo	80x20	0,976	0,283	0,058	5
G_2	Mínima energia	80x20	0,833	0,038	0,123	3
G_2	MCS[1]	80x20	0,786	0,041	0,150	3
G_2	Farmacóforo	80x20	0,903	0,018	0,114	2

[1] MCS: Máxima Subestrutura Comum

Fonte: Elaborada pelo autor.

4.1.3 QSAR-4D

Além dos modelos de QSAR-2D e 3D, foram gerados modelos de QSAR-4D independentes do receptor (RI-QSAR-4D, do inglês *Receptor Independent* QSAR-4D). Este é um método reconhecido por superar alguns problemas das metodologias de QSAR-3D independentes do receptor. Assim como nos métodos 2D e 3D, as estratégias de *clustering* foram reproduzidas para se construir os modelos de QSAR-4D. Portanto, foram gerados modelos para o conjunto completo (exceto a molécula **34**, **modelo 8**) e para os grupos G_1 (**modelo 9**) e G_2 (**modelo 10**). Para cada um desses conjuntos, 200 modelos foram criados a partir de variações randômicas das moléculas presentes nos conjuntos treinamento e teste. Para o conjunto completo, os parâmetros estatísticos obtidos foram: $R^2 = 0,4353$, $R^2_{pred} = 0,4588$, o que indica resultados insatisfatórios.

Para o grupo G_1, a proporção entre série treinamento e teste foi mantida em 80x20 devido aos bons resultados obtidos com essa proporção nos modelos de QSAR-2D. O alinhamento molecular foi feito através dos átomos comuns dos centros oxazol e oxadiazol. Após o alinhamento, o arquivo contendo os perfis conformacionais de cada ligante foi submetido ao programa LQTAgridPy, resultando em 19.404 descritores para o grupo G_1. Após a exclusão dos descritores que possuíam variância inferior a 0,01 e correlação de Pearson inferior a 0,2, restaram 903 descritores. Esse conjunto de descritores foi submetidos à seleção de variáveis através do programa PyQSAR. O modelo de QSAR com as variáveis selecionadas obteve os seguintes parâmetros estatísticos: $R^2 = 0,8033$, $RMSE = 0,1313$, $Q^2_{5-fold} = 0,6600$, $RMSE_{cv} = 0,1716$, $R^2_{pred} = 0,6480$.

Estes resultados indicam a boa capacidade de correlação e predição do modelo. A diferença de 0,1432 entre R^2 e Q^2_{5-fold} indica que não houve sobre-ajuste. Além disso, o

índice R^2 ficou acima do limite mínimo que tem sido reportado na literatura[15,20] ($R^2 > 0,5$) e próximo do Q^2_{5-fold}, indicando que a capacidade de correlação interna e de predição externa e são equivalentes. A Equação 4.1 descreve o modelo de QSAR-4D para o grupo G_1.

$$\begin{aligned} \text{pIC}_{50} =& 5,1535 + 0,8409\Big[\text{15\_13\_6\_NH3+\_LJ}\Big] \\ & - 0,7075\Big[\text{16\_12\_5\_NH3+\_LJ}\Big] \\ & + 0,1484\Big[\text{16\_20\_10\_NH3+\_LJ}\Big] \\ & - 0,1210\Big[\text{21\_17\_13\_NH3+\_LJ}\Big] \\ & + 0,1913\Big[\text{22\_12\_12\_NH3+\_LJ}\Big] \end{aligned} \tag{4.1}$$

Os termos $\Big[\text{x\_y\_z\_NH3+\_LJ}\Big]$ e $\Big[\text{x\_y\_z\_NH3+\_C}\Big]$ representam respectivamente o descritor de Lennard-Jones e de Coulomb com a sonda posicionada nas coordenadas x, y, z do retículo.

A mesma estratégia utilzada no grupo G_1 foi reproduzida para o grupo G_2. Manteve-se a proporção 80x20, o que resultou em 28 moléculas para o conjunto treinamento e 5 para o conjunto teste. O programa LQTAgridPy gerou 21.252 descritores moleculares. O número de descritores foi inicialmente reduzido para 3.353 e posteriormente para XX descritores pelo programa PyQSAR. Os parâmetros estatísticos obtidos indicam a boa capacidade de correlação e de predição do modelo ($R^2 = 0,7005$, $RMSE = 0,1560$, $Q^2_{5-fold} = 0,6095$, $RMSE_{cv} = 0,1701$, $R^2_{pred} = 0,6581$). De maneira similar ao grupo G_1, a diferença entre R^2 e Q^2_{5-fold} foi de 0,0909, o que indica que não houve sobre-ajuste dos dados, e o valor de R^2 ficou acima de 0,5. A Equação 4.2 descreve o modelo de QSAR-4D para o grupo G_2.

$$\begin{aligned} \text{pIC}_{50} =& 4,9338 + 0,2170\Big[\text{16\_20\_11\_NH3+\_LJ}\Big] \\ & + 0,1303\Big[\text{17\_19\_15\_NH3+\_LJ}\Big] \\ & - 0,7328\Big[\text{17\_26\_15\_NH3+\_C}\Big] \\ & + 0,2770\Big[\text{18\_23\_14\_NH3+\_LJ}\Big] \\ & + 0,7227\Big[\text{19\_26\_20\_NH3+\_C}\Big] \end{aligned} \tag{4.2}$$

Os resultados dos modelos QSAR-4D estão resumidos na Tabela 8. É possível notar que a estratégia de *clustering* resultou em modelos de QSAR-4D significativamente mais robustos, assim como ocorreu para os modelos de QSAR-2D.

Tabela 8: Melhores modelos para metodologia QSAR-4D.

Dataset	$\mathbf{R^2}$	**RMSE**	$\mathbf{Q^2_{5-fold}}$	$\mathbf{RMSE_{cv}}$	$\mathbf{R^2_{pred}}$
Conjunto Completo (**modelo 8**)	0,4353	0,205	0,4035	0,3919	0,4588
G_1 (**modelo 9**)	0,8033	0,1313	0,6600	0,1716	0,6480
G_2 (**modelo 10**)	0,7005	0,1560	0,6095	0,1701	0,6581

Fonte: Elaborada pelo autor.

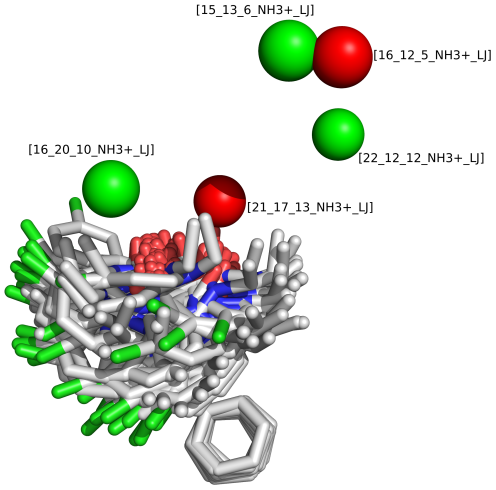
A Figura 16 apresenta os descritores moleculares em forma de mapas de contorno tridimensionais ao redor das moléculas alinhadas. Os mapas de contorno permitem a visualização das características moleculares que afetam mais significativamente a atividade biológica. As esferas verdes representam as contribuições estereoquímicas com coeficientes de regressão positivos e as vermelhas representam o mesmo tipo de contribuição, porém, com coeficientes de regressão negativos. As esferas azuis indicam descritores eletrostáticos com coeficientes negativos e as amarelas representam contribuições eletrostáticos com coeficientes positivos.

Os mapas de contorno indicam que a região molecular que mais contribui para a variação da atividade biológica é aquela acessível aos grupos nitrofenila, hidroxi-indol e trimetoxi. Por outro lado, regiões do espaço acessíveis ao grupo fenil-oxadiazol e fenil-oxazol não foram destacadas como regiões que afetam a variação da atividade. Este resultado é esperado visto que estes fragmentos moleculares são comuns a todos os compostos (exceto o outlier estrutural **34**).

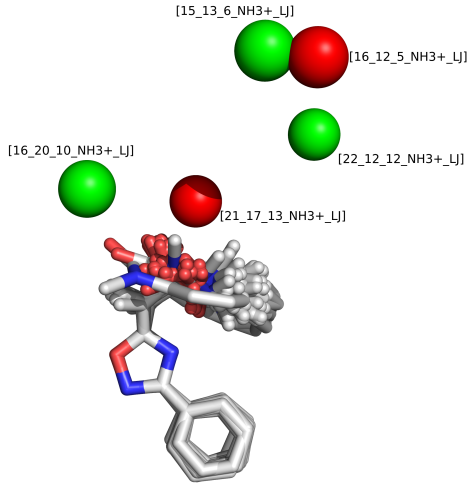
Para o grupo G_1, as contribuições estereoquímicas positivas [16_20_10_NH3+_LJ] e [22_12_12_NH3+_LJ] estão principalmente relacionadas aos halogênios nos grupos R_4 e R_5, que em virtude da flexibilidade conformacional das moléculas conseguem acessar essas regiões do espaço. A contribuição estereoquímica negativa [21_17_13_NH3+_ LJ] está relacionada ao grupo hidroxila e principalmente à substituintes volumosos em R_2. Um dos resultados dos modelos de QSAR-4D para o grupo G_1 é a correlação entre flexibilidade e atividade biológica. Moléculas com maior grau de flexibilidade apresentam atividade mais baixa, o que pode decorrer de maior dificuldade para formar interações estáveis com o alvo molecular[82].

Como no grupo G_1, a relação entre graus de liberdade e pIC_{50} também pode ser observada no grupo G_2. O descritor estereoquímico [18_23_14_NH3+_LJ] indica que substituintes volumosos no grupo R_4, representados principalmente no composto **3**, que possui um grupo estireno nesta posição, contribuem positivamente para a atividade biológica. A contribuição eletrostática positiva [19_26_20_NH3+_C] está associada à presença de halogênios na posição *para* do anel fenila no grupo R_4. No entanto, o descritor eletrostático [17_26_15_NH3+_C] indica que esses átomos podem contribuir para a

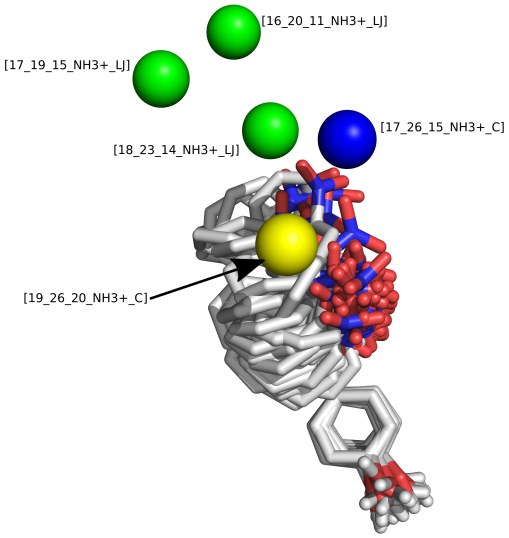
redução da atividade a depender da conformação adotada pela molécula.



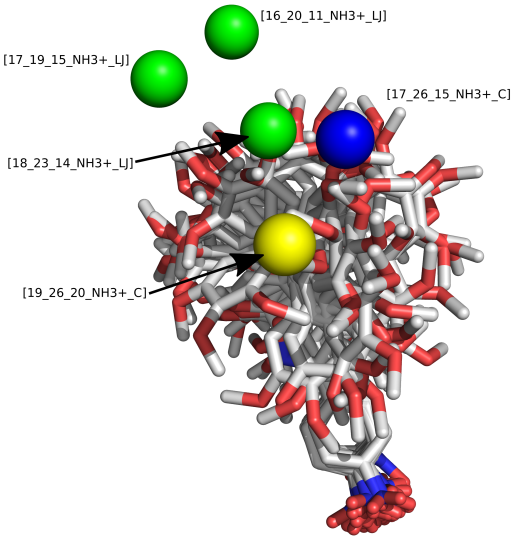
(a) Composto **60**.



(b) Composto **26**.



(c) Composto **50**.



(d) Composto **12**.

Figura 16: Mapas de contorno para os compostos mais (composto **60** e **50**) e menos potentes (composto **26** e **12**) dos grupos G_1 e G_2.

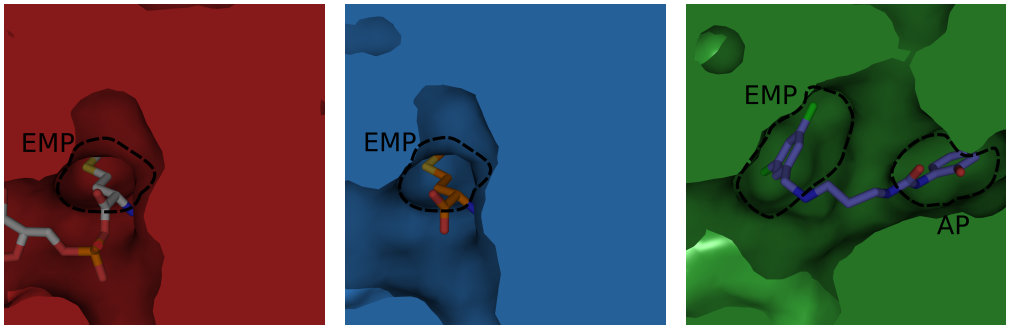
Fonte: Elaborada pelo autor

4.2 Estudos de SBDD

4.2.1 Obtenção da estrutura do alvo molecular por modelagem por homologia

Para a realização dos estudos de SBDD, a modelagem por homologia foi utilizada para a construção de um modelo tridimensional da LiMetRS. Esta enzima é altamente flexível e adota conformações substancialmente diferentes[83–84] quando ligada a diferentes moléculas[85] . O estado-M, em complexo com a metionina, é caracterizado pelo sítio da

metionina (EMP, do inglês *Enlarged Methionine Pocket*), fechamento do domínio de ligação do peptídeo (CP, do inglês, *Peptide Domain*) e ausência do sítio auxiliar (AP, do inglês *Auxiliary Pocket*). O estado-P, em complexo com o subproduto metionil-adenilato (MAMP, sigla do inglês para *Methionyl-Adenosine Monophosphate*), caracteriza-se por uma pequena região da EMP, domínio CP fechado e região AP inexistente. O estado-I, em complexo com o inibidor, caracteriza-se pela ampliação da região EMP, abertura do domínio CP e formação da região AP (Figura 17).



(a) Estado P (b) Estado M (c) Estado I

Figura 17: Sítios EMP e AP em diferentes estados.

Fonte: Elaborada pelo autor.

A estrutura primária da LiMetRS foi obtida no UniProt. Estruturas de MetRS com alta identidade com a enzima de *L. infantum* e alta resolução foram obtidas no PDB. As enzimas de *Trypanosoma brucei* (TbMetRS, código PDB 4MW0), e *L. major* (LmMetRS, código PDB 3KFL) foram selecionadas. A TbMetRS, em complexo com um inibidor derivado da ureia, apresenta a conformação I, enquanto a LmMetRS, em complexo com o MAMP, apresenta a conformação P. A identidade em relação à sequência de aminoácidos da LiMetRS, é de 68,6% para TbMetRS e 97,41% para LmMetRS (Figura 18). Os resíduos em um raio 10Å ao redor dos ligantes são altamente conservados em relação à LiMetRS, apresentando identidade de 84,33% para TbMetRS e 100% para LmMetRS.

Apesar da alta identidade entre as sequências de aminoácidos de LiMetRS e LmMetRS, a seleção da estrutura de TbMetRS no Estado-I como *template* para a modelagem foi importante para reproduzir este arranjo conformacional, no qual a atividade enzimática está inibida. Nesta conformação, o ligante interage com as regiões AP e EMP e impede a ligação da metionina. Este *template* também é importante para se modelar a região hidrofóbica[85] e as cadeias laterais no sítio AP que favorecem a formação de interações ligante-receptor.

Inicialmente, quatro estratégias diferentes foram adotadas para criar os modelos por homologia: (i) 4MW0 como *template*, sem restrição entre ligante e proteína; (ii) 4MW0

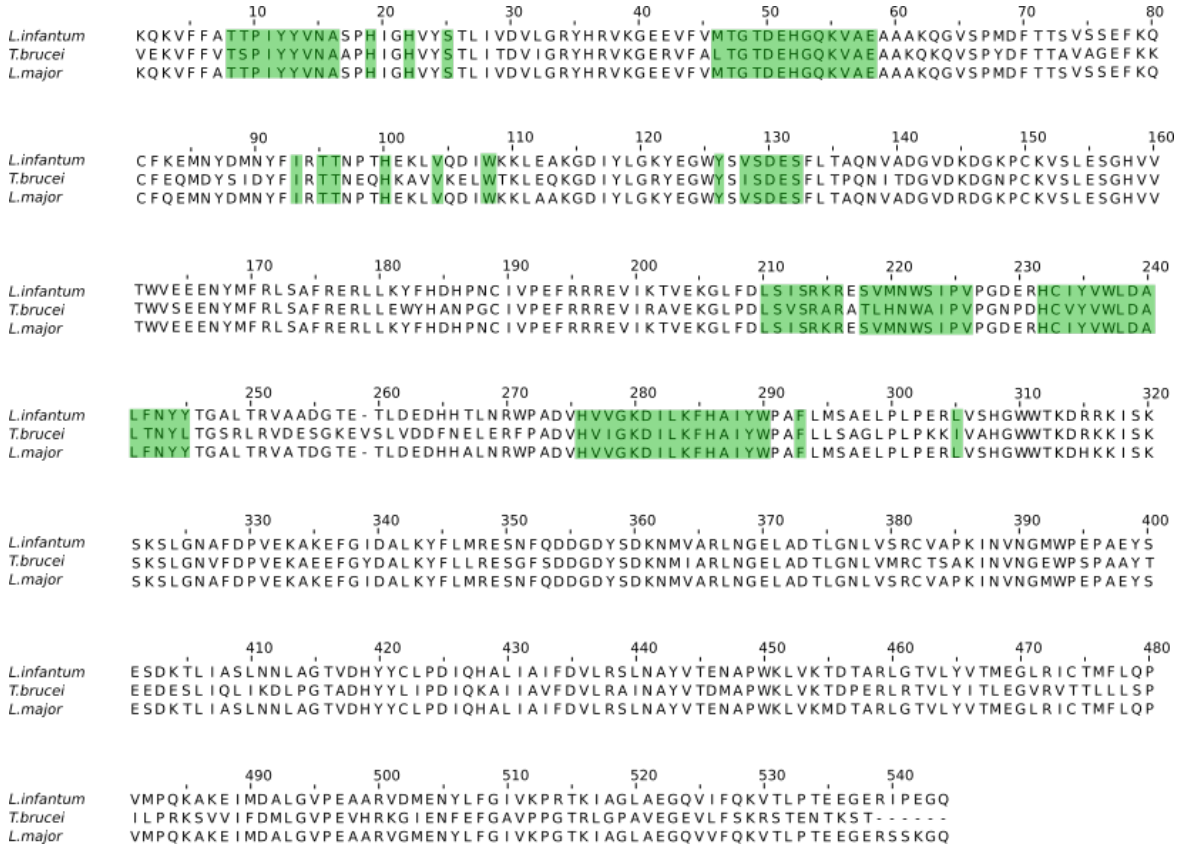


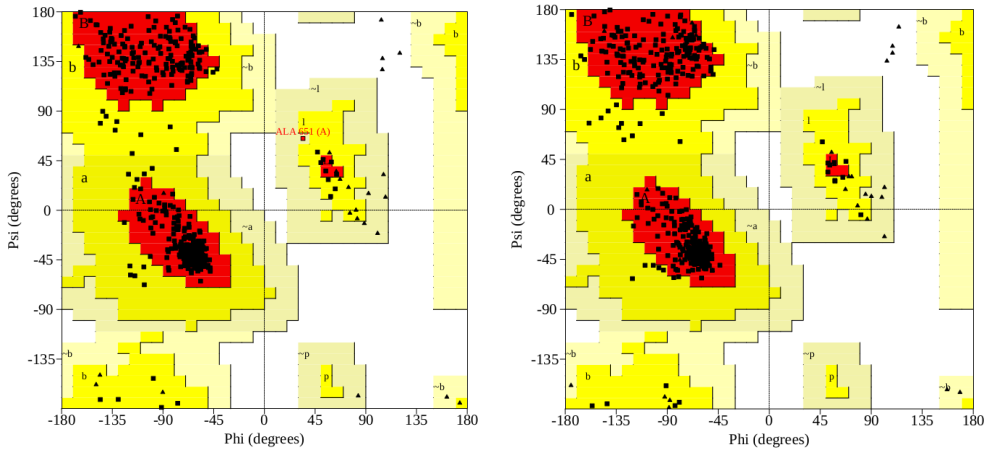
Figura 18: Alinhamento entre as estruturas primárias de LiMetRS, TbMetRS(4MW0) e LmMetRS(3KFL). Os aminoácidos destacados representam aqueles presentes em um raio de 10Å ao redor do ligante de 4WM0.

Fonte: Elaborada pelo autor.

como *template*, com restrição entre ligante e proteína; (iii) 3KFL como *template* (sem ligante) inicial, seguido por 4MW0, sem restrição entre ligante e proteína; (iv) 3KFL como *template* inicial (sem ligante), seguido por 4MW0, com restrição entre ligante e proteína. Para cada uma dessas estratégias, cinco modelos foram gerados e avaliados. Os modelos foram examinados através da pontuação DOPE gerado pelo MODELLER, análise do gráfico de Ramachandran e cálculo do RMSD entre *template* e modelo. O modelo escolhido deve carregar características do Estado-I, representado pelo *template* de 4MW0 de TbMetRS.

Além desses critérios de seleção, todos modelos gerados foram submetidos ao PROCHECK, para avaliação da qualidade estereoquímica através do gráfico de Ramachandran. Os gráficos de distribuição dos aminoácidos para os *templates* utilizados são apresentados na Figura 19. Para o *template* 3KFL, 93% dos resíduos estão na região mais favorável, 6,8% em regiões permitidas e 0,2% em regiões generosamente permitidas. Para o *template* 4MW0, 92,9% dos aminoácidos estão em regiões favoráveis e 7,1% em regiões permitidas. Essas regiões são derivadas dos ângulos diédricos, Ψ e Φ. Os limites que definem cada

região são obtidos a partir de distribuições de frequência desses ângulos, obtidos de bancos de dados[86]. Portanto, resíduos em regiões favoráveis apresentam ângulos Ψ e Φ próximos à média encontrada nesses bancos de dados. Os parâmetros de RMSD, pontuação DOPE e estereoquímica de todos os modelos gerados estão nas Tabelas 9 e 10.



(a) *Template* 3KFL (b) *Template* 4MW0

Figura 19: Gráfico de Ramachandran para os *templates* utilizados na geração do modelo de homologia para LiMetRS.

Fonte: Adaptada de LASKOWSKI *et al.*[65]

Tabela 9: RMSD e pontuação DOPE para os modelos gerados.

Método	**RMSD$_{4MW0}$**	**RMSD$_{3KFL}$**	**DOPE$_{4MW0}$**	**DOPE$_{3KFL}$**
4MW0 sem restrição	0,306	0,846	-66692,22656	
	0,163	0,892	-66774,42969	
	0,149	0,892	-66774,42969	
	0,193	0,936	-66590,46094	
	0,201	0,930	-66341,78125	
4MW0 com restrição	0,251	0,846	-66863,57031	
	0,182	0,863	-66819,41406	
	0,166	0,810	-67070,92188	
	0,116	0,812	-66518,95312	
	0,268	0,925	-66558,82812	
3KFL e 4WM0	0,872	0,538	-68811,00000	-71142,83594
sem restrição	0,855	0,561	-68675,94531	-71142,83594
	0,729	0,498	-68575,36719	-71142,83594
	0,439	0,802	-68195,55469	-71142,83594
	0,750	0,518	-68222,05469	-71142,83594

(continua)

(continuação)

Método	**$RMSD_{4MW0}$**	**$RMSD_{3KFL}$**	**$DOPE_{4MW0}$**	**$DOPE_{3KFL}$**
3KFL e 4WM0	0,827	0,573	-68706,04688	-71142,83594
com restrição	0,983	0,523	-68799,93750	-71142,83594
	1,071	0,595	-68850,16406	-71142,83594
	0,784	0,522	-68637,57031	-71142,83594
	0,802	0,554	-69107,23438	-71142,83594

Fonte: Elaborada pelo autor.

Tabela 10: Parâmetros obtidos pelo gráfico de Ramachandran para os modelos gerados.

Método	**Favorável(%)**	**Permitida(%)**	**Generosamente permitida(%)**	**Desfavorável(%)**
4MW0 sem restrição	93,1	6,2	0	0,6
	92,3	6,7	0,8	0,2
	92,7	6,9	0,2	0,2
	92,9	6,4	0,2	0,4
	93,3	5,8	0,2	0,6
4MW0 com restrição	93,3	6,2	0	0,4
	92,3	6,9	0,6	0,2
	93,3	6,0	0,2	0,4
	92,3	7,1	0,4	0,2
	93,6	6,0	0,2	0,2
3KFL e 4WM0	93,6	5,8	0,4	0,2
sem restrição	94	5,0	0,6	0,4
	93,1	6,0	0,4	0,4
	93,3	5,8	0,4	0,4
	93,6	5,4	0,6	0,4
3KFL e 4WM0	94,0	5,4	0,2	0,4
com restrição	94,0	5,4	0,2	0,4
	94,2	5,4	0,2	0,2
	93,8	5,6	0,4	0,2
	94,4	5,2	0	0,4

Fonte: Elaborada pelo autor.

O modelo escolhido utiliza como molde a estrutura de TbMetRS e restrição entre

ligante e proteína. Este modelo tem como parâmetros: RMSD entre modelo e *template* TbMetRS de 0,116, RMSD entre modelo e *template* LmMetRS de 0,812, e pontuação DOPE de -66518,95312. O modelo selecionado apresenta 92,3% de aminoácidos em regiões favoráveis, 7,1% em regiões permitidas, 0,4% em regiões generosamente permitidas e 0,2% em regiões desfavorável.

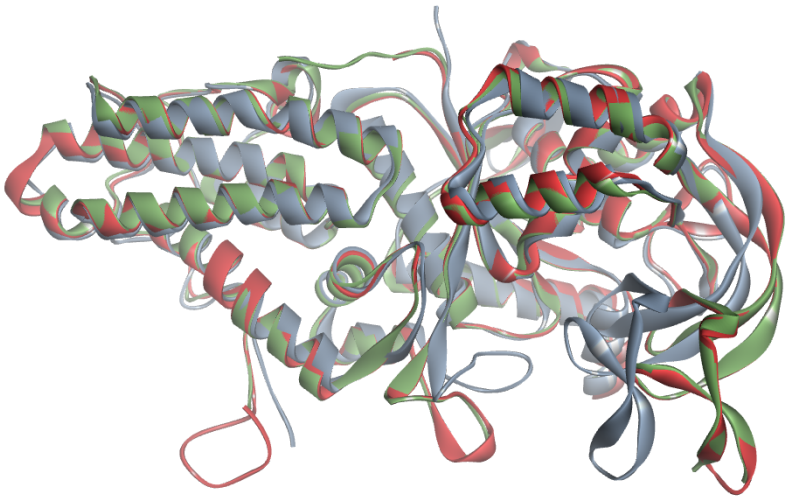


Figura 20: Alinhamento estrutural entre as enzimas de *T.brucei*(verde), *L.major*(azul), e o modelo gerado de *L.infantum* (vermelho).

Fonte: Elaborada pelo autor.

O modelo gerado reproduz características estruturais da enzima em complexo com inibidores, conforme observado na Figura 20. No sítio de interação, as posições dos aminoácidos são mantidas (Figura 21). Essas características indicam que o modelo gerado é adequado para ser usado nas etapas seguintes de triagem virtual, docagem molecular e dinâmica molecular.

Uma primeira avaliação para o planejamento de inibidores de LiMetRS é o reposicionamento nesta enzima, de inibidores de TbMetRS. Essa avaliação foi feita através da estabilidade do ligante (RMSD) e da verificação das interações ligante-receptor por meio de simulações de dinâmica molecular. As interações intermoleculares em TbMetRS poderiam ser mantidas em LiMetRS uma vez que os aminoácidos no sítio de ligação destas enzimas são conservados. Após a modelagem por homologia, a estrutura de LiMetRS possui em seu sítio o mesmo ligante do *template* TbMetRS, disposto nas mesmas posições atômicas. Para fins comparativos, duas dinâmicas foram realizadas. A primeira foi realizada com a TbMetRS e utilizada como controle, e a outra com a LiMetRS. A TbMetRS possui um resíduo modificado S-(dimetilarsênico)cisteína (CAS) substituindo a cisteína (Cys470). O resíduo foi alterado utilizando o programa Chimera e os resíduos completados através do Swiss PDB Viewer. É possível avaliar a estabilidade do ligante através da comparação do RMSD do ligante em complexo com o modelo LiMetRS e com o *template* TbMetRS.

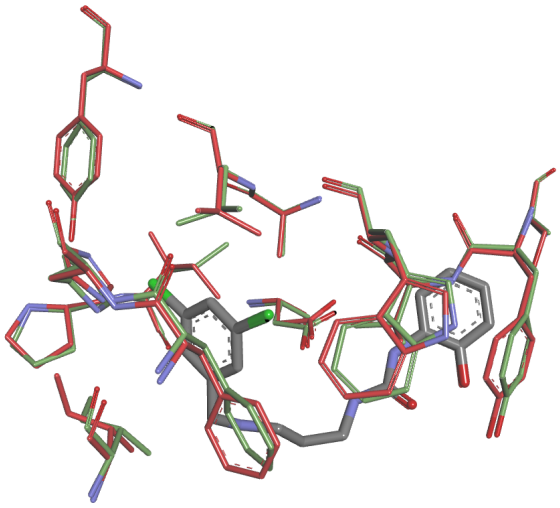


Figura 21: Sobreposição dos resíduos de aminoácidos do sítio ativo da TbMetRS(verde) e o modelo gerado de LiMetRS(vermelho). O inibidor, di-cloro, de TbMetRS está representado em cinza.

Fonte: Elaborada pelo autor.

A análise dos gráficos nas Figuras 22 à 25 permite concluir que o ligante não mantém as posições atômicas em LiMetRS, mas permanece estável em TbMetRS, o que está de acordo com resultados experimentais descritos na literatura[84]. Também houve dissociação das interações de hidrogênio com Asp50 e das interações π-*stacking* na região AP em LiMetRS. No entanto, essas interações permaneceram estáveis em TbMetRS. Esses achados ressaltam a necessidade de se planejar novos compostos que sejam específicos para LiMetRS.

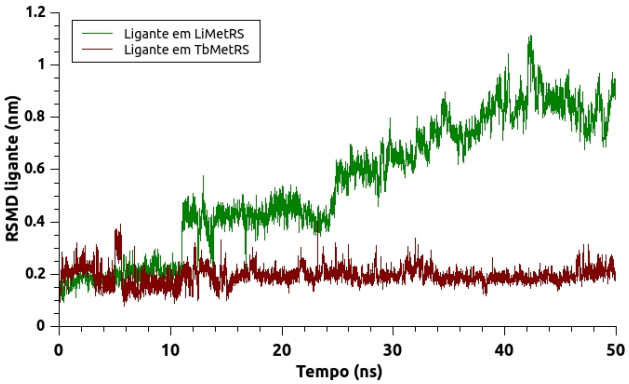


Figura 22: RMSD dos ligantes. Em verde RMSD do ligante complexado com TbMetRS e em vermelho o RMSD do ligante complexado com LiMetRS.

Fonte: Elaborada pelo autor.

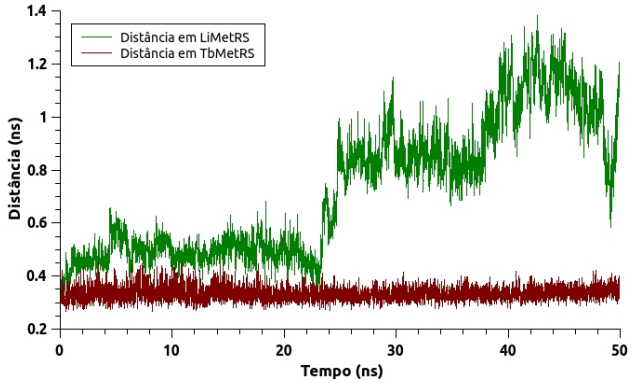


Figura 23: Comprimento da ligação de hidrogênio. Distância entre o NAR do ligante e OD1 do Asp50 (doador e o aceptor).

Fonte: Elaborada pelo autor.

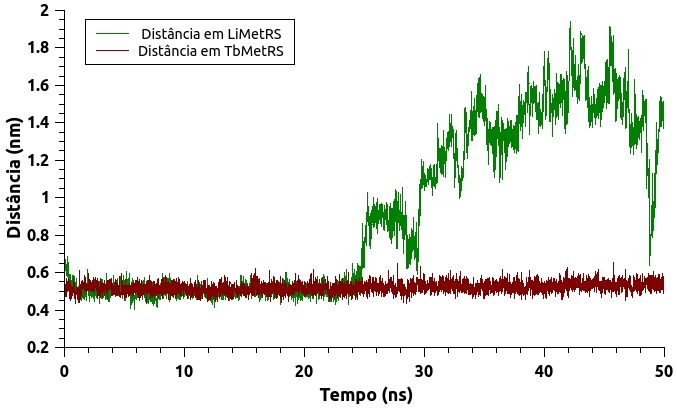


Figura 24: Distância entre o centro de massa do anel da Tyr235 e do ligante, envolvidos em uma interação π-*stacking*.

Fonte: Elaborada pelo autor.

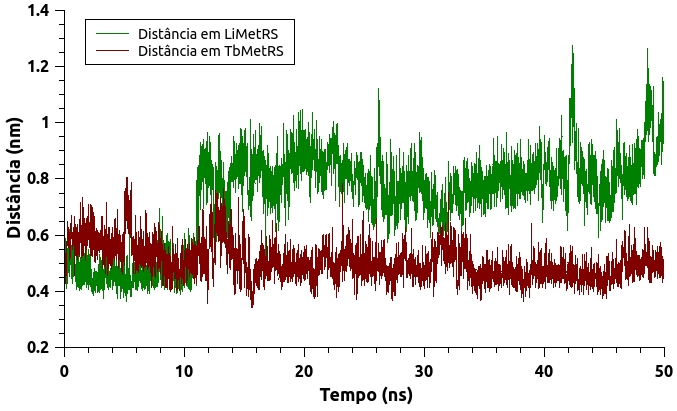


Figura 25: Distância entre o centro de massa do anel Di-Cloro e carbono CB da Ala240.

Fonte: Elaborada pelo autor.

4.2.2 Triagem Virtual

Uma busca na literatura por inibidores de TbMetRS foi feita para extrair um conjunto de propriedade químicas a serem utilizadas como filtros moleculares na triagem do banco de dados ZINC15. A pesquisa por dados reportados na literatura[29,87–92] resultou no uso dos seguintes critérios de busca: 3,00 $\leq$ LogP $\leq$ 5,00, 300 $\leq$ Peso Molecular $\leq$ 450, compostos em estoque, e filtro para PAINS. Esses critérios de filtragem resultaram em aproximadamente 5,3 milhões de compostos com características estruturais diversas, que foram usados nos estudos de docagem molecular com a LiMetRS. Após a primeira etapa de docagem molecular, selecionou-se os 10.000 compostos de maior pontuação. Este conjunto foi utilizado em nova etapa de docagem molecular, na qual parâmetros de eficiência e precisão mais robustos foram utilizados. Os resultados revelaram grande diversidade de modos de ligação ligante-receptor, no entanto, buscou-se identificar inibidores que ocupavam apenas as regiões EMP e AP.

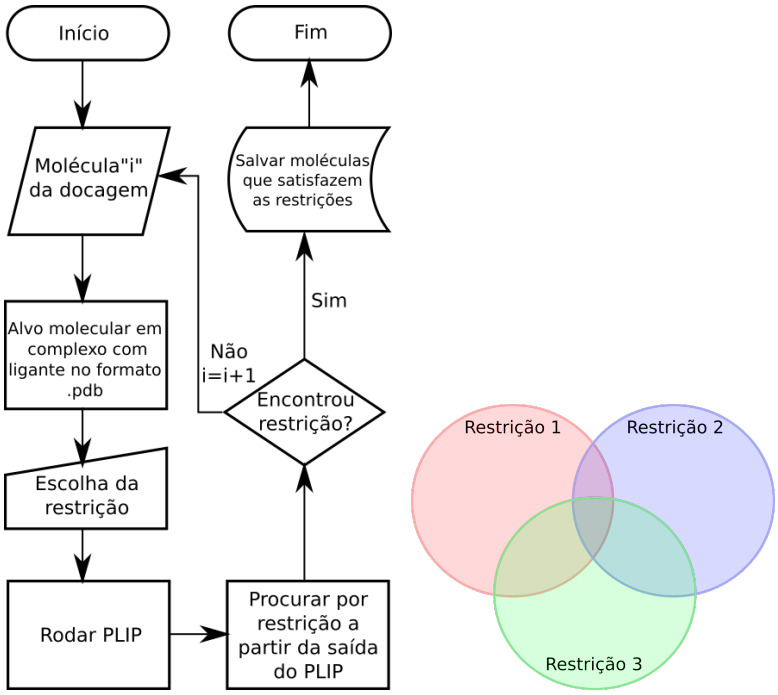
4.2.3 Análise das Interações Intermoleculares

Dado o grande volume de informações resultantes da docagem molecular, foi criado um algoritmo para selecionar moléculas que interagissem com aminoácidos específicos (Figura 26). Estudos anteriores utilizaram essa estratégia[69] para selecionar ligantes em triagens virtuais, em conjunto com simulações de dinâmica molecular para avaliação da estabilidade das interações de interesse. No entanto, a estratégia desenvolvida neste trabalho se destaca por poder filtrar por mais de um tipo de interação através dos operadores lógicos (e/ou).

Na estrutura de TbMetRS, os grupos benzil e fenil do ligante ocupam a região hidrofóbica EMP, que sofre um aumento quando em complexo com o ligante. Nesta região, os inibidores reportados mostram variabilidade estrutural, e sua interação com a enzima impede a ligação da metionina[84]. A porção aminoquinolona ou benzimidazol do ligante localiza-se na região AP. O caráter plano do inibidor no sítio AP é importante para sua estabilização por meio de múltiplas interações π-stacking[85]. Além dessas duas características, a grande maioria dos inibidores[29,87–92] interage com ácido aspártico por meio de uma ligação de hidrogênio que ocorre na região do *linker* do inibidor. Essa interação de hidrogênio é recorrente nos trabalhos anteriores, mesmo quando há alterações na região do *linker*. Este conjunto de características foi extrapolado para guiar estudos de identificação de inibidores da LmMetRS descritos na literatura[30].

A partir do resultado de docagem molecular, a melhor conformação dos 10.000 ligantes foi submetida ao programa desenvolvido para selecionar apenas os ligantes que formam interações de hidrogênio com o Asp50 e interações π-stacking com a His52 ou Gly53 ou Tyr235 ou Val236, localizados na região AP. A identificação das interações foi realizada pelo algoritmo PLIP, que utiliza um conjunto de quatro etapas para identificar

as interações relevantes. O programa utilizado neste trabalho é capaz de selecionar 5 interações (interações hidrofóbicas e de hidrogênio, π-stacking, interações cátion-π e ligações de halogênio). Uma restrição é dada por um conjunto de 3 parâmetros: o resíduo de aminoácido, o número do resíduo na estrutura primária, e o tipo de interação. Para um número de restrições maior do que um, o *script* é capaz de utilizar os operadores lógicos e/ou e gerar diferentes conjuntos. Neste caso, dado um número de restrições n, é possível escolher os elementos dos $2^n - 1$ subgrupos de maneiras diferentes. O esquema geral do programa é apresentado na Figura 26.



(a) Fluxograma do *script*. (b) Conjunto de restrições possíveis para $n = 3$.

Figura 26: Esquema geral do programa de pós-triagem virtual.

Fonte: Elaborada pelo autor.

O operador lógico utilizado identifica apenas os ligantes que formam interação hidrogênio com Asp e pelo menos uma interação do tipo *stacking* com os aminoácidos de interesse. Esses filtros selecionaram 222 ligantes. Como não foi imposta nenhuma obrigatoriedade de interação na região EMP, alguns compostos do conjunto não ocupavam essa região. Dessa forma, foi feita uma análise visual para selecionar os ligantes que ocupam a região EMP, o que resultou na seleção de 10 compostos (Tabela 11). Resultados em TbMetRS mostram que compostos que estão inseridos nessa região são mais potentes[85]. O fluxo de trabalho empregado para a proposição de novos ligantes para a LiMetRS é apresentado na Figura 27.

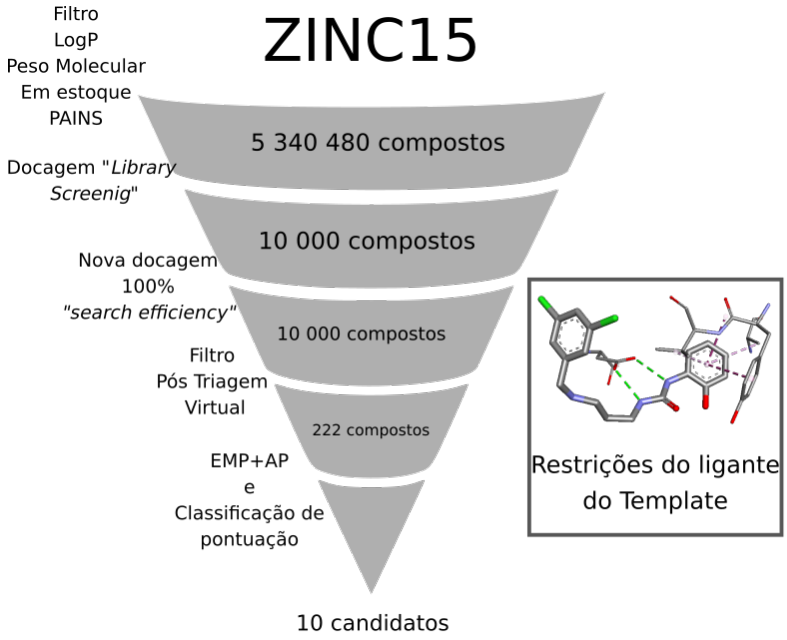
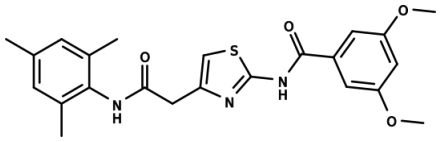
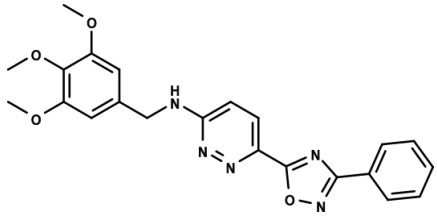
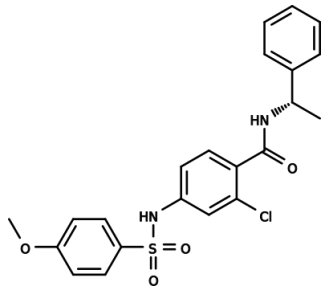
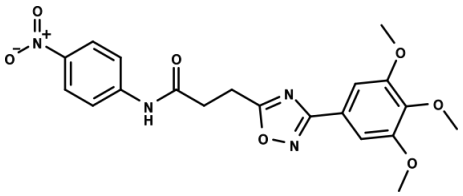
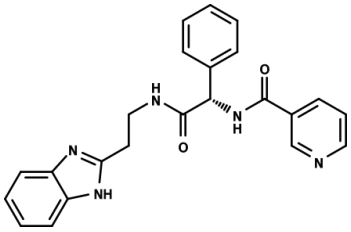
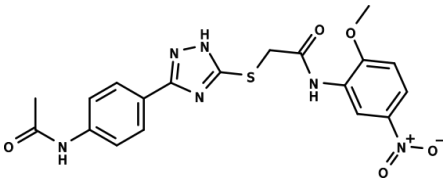
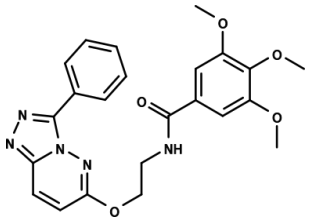


Figura 27: Representação esquemática da estratégia de Triagem Virtual empregada neste estudo.

Fonte: Elaborada pelo autor.

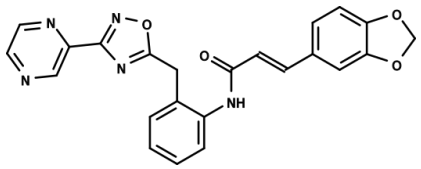
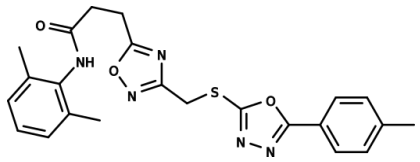
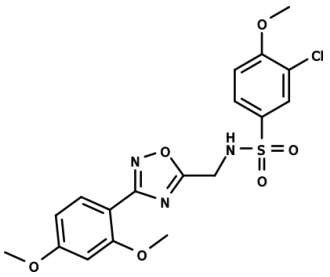
Uma observação importante após a docagem molecular, é o maior volume dos substituintes nos inibidores que interagem com a região AP da enzima em comparação aos inibidores de TbMetRS. Os resultados da docagem molecular também mostraram um deslocamento da posição dos anéis aromáticos localizados na região AP. Isso pode ser observado na Figura 28, na qual o melhor composto (ZINC9659242) é sobreposto ao ligante da enzima TbMetRS. Na Figura 28 a seta amarela indica um deslocamento do inibidor na região AP. A linha tracejada em vermelho representa o limite da região AP em TbMetRS e a linha tracejada em verde representa o limite da região AP em LiMetRS, indicando portanto um maior volume acessível ao ligante nessa região. A substituição de Leu456 em TbMetRS por Val219 em LiMetRS resulta em um maior volume da região AP, o que pode justificar a causa deste deslocamento. Além disso, a substituição de Ala460 em TbMetRs por Ser223 em LiMetRS podem implicar na formação de ligações de hidrogênio adicionais no sítio AP, o que justifica a presença do grupo dimetil-éter nos compostos selecionados. A Figura 29 ilustra o conjunto de compostos selecionados com suas conformações após a docagem com LiMetRS, evidenciando características importantes do modo de ligação que foram mantidas como a ocupação da região EMP, ocupação da região AP com anéis aromáticos e ligação de hidrogênio com Asp50.

Tabela 11: Compostos selecionados após a etapa de triagem virtual.

Classificação	**Código ZINC**	**Estrutura**
1	ZINC9659242	
2	ZINC64837843	
3	ZINC224021491	
4	ZINC9334822	
5	ZINC252484826	
6	ZINC408674884	
7	ZINC12443176	

(continua)

(continuação)

Classificação	Código ZINC	Estrutura
8	ZINC225605317	
9	ZINC21721631	
10	ZINC223767934	

Fonte: Elaborada pelo autor.

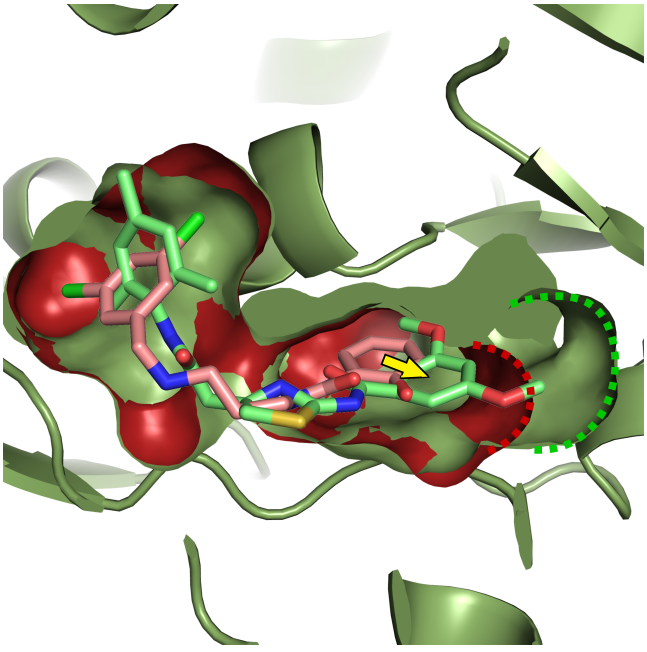


Figura 28: Composto ZINC223767934 complexado com LiMetRS em verde, comparado com o inibidor de TbMetRS(4MW0) em vermelho.

Fonte: Elaborada pelo autor.

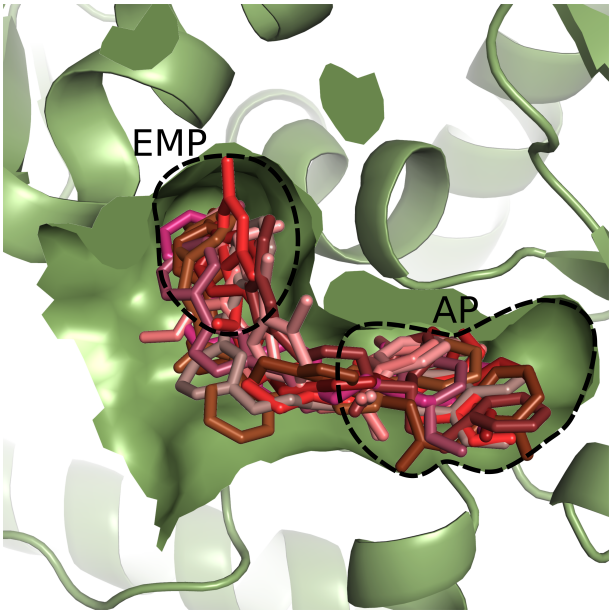


Figura 29: Compostos selecionados na docagem em complexo com a LiMetRS.

Fonte: Elaborada pelo autor.

4.2.4 Dinâmica Molecular

O composto que obteve a melhor pontuação (ZINC9659242) foi selecionado para estudos de dinâmica molecular. A Figura 30 representa o composto ZINC9659242 e suas interações na conformação inicial, definida a partir da docagem. Na Figura 30, a linha azul representa as ligações de hidrogênio, a linha tracejada verde representa a interação π-*stacking* e a linha tracejada cinza presenta as interações hidrofóbicas.

A estabilidade do ligante foi avaliada a partir de análises de RMSD do ligante e da manutenção das interações observadas na etapa de docagem molecular. Três simulações independentes foram feitas para reduzir flutuações estatísticas usando as coordenadas resultantes do processo de docagem como posição inicial do ligante. Valores de RMSD menores que 3,0Å(0,3 nm) foram estabelecidos como critério de estabilidade[93]. As interações de hidrogênio foram avaliadas tomando-se a distância média entre o doador e o aceptor definida em um intervalo de 0,27 a 0,33 nm. Os parâmetros da dinâmica são apresentados na Tabela 12. O inibidor se manteve no sítio de ligação (RMSD $\leq$ 0,3(nm)) em duas simulações. Além disso, as posições atômicas do ligante não variaram significativamente durante a simulação. A interação de hidrogênio com o Asp50 foi mantida durante todo o período em uma das simulações (Figura 30). Além das interações de hidrogênio, ligações π-*stacking* entre o grupo fenil-3,5-dimetoxi e os resíduos aromáticos na região AP foram avaliadas. Para a interação entre o ligante e a Tyr235, tomou-se como padrão uma distância de 0,56 nm. As distâncias de ligação entre o centro de massa do anel 3,5-dimetoxi e a cadeia lateral da Tyr235 permaneceram constantes durante as simulações MD1 e MD3, indicando a estabilidade do ligante na região AP do sítio (Figura 31).

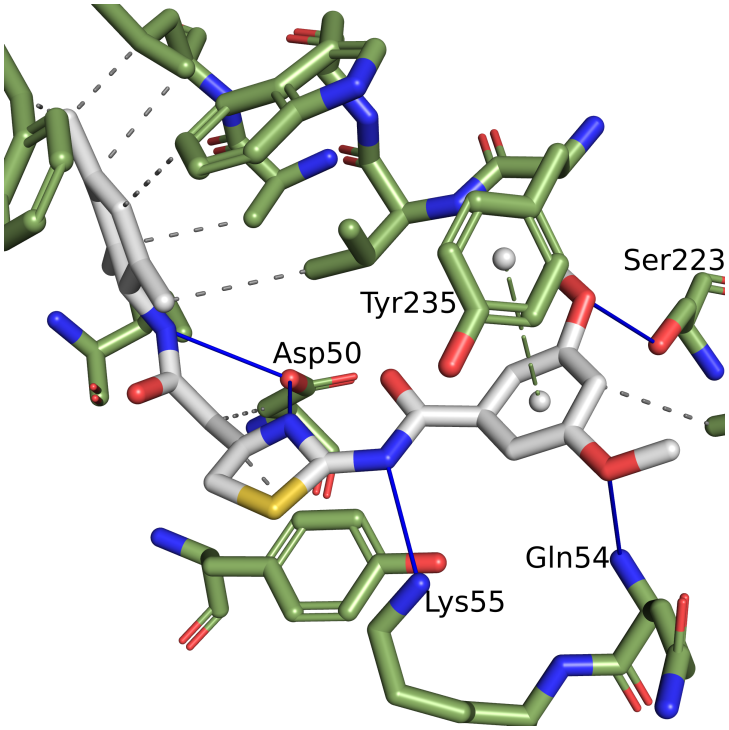


Figura 30: Interações entre ZINC965924 e LiMetRS.

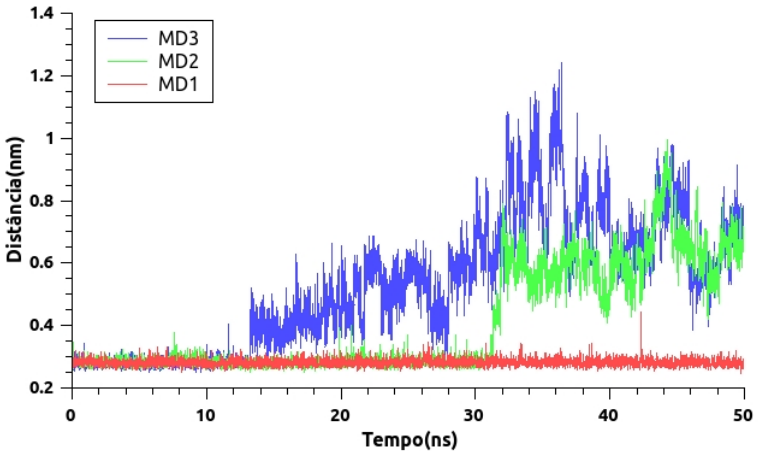
Fonte: Adaptada de SALENTIN *et al.*[68]

Tabela 12: Resultado da dinâmica molecular para o composto melhor pontuado.

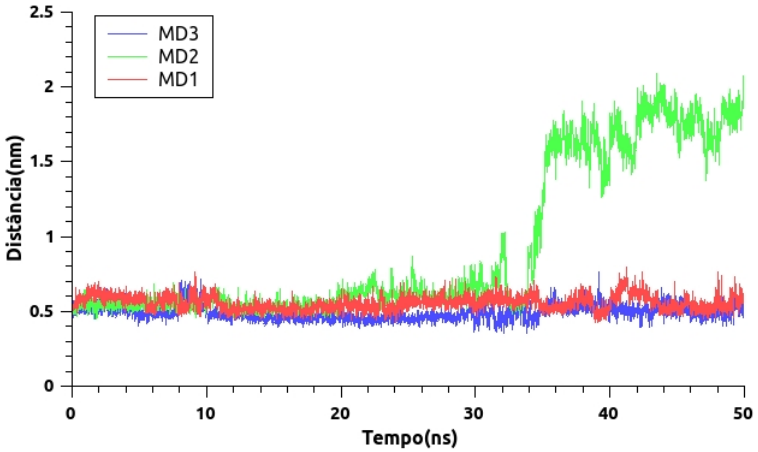
Composto	**Parâmetro**	**MD1**	**MD2**	**MD3**
ZINC9659242	Média RMSD (nm)	0,14 ± 0,01	0,5 ± 0,3	0,30 ± 0,05
	Ligação Asp50 (nm)	0,28 ± 0,01	0,4 ± 0,1	0,5 ± 0,2
	π-*stacking* Tyr235 (nm)	0,55 ± 0,05	0,9 ± 0,5	0,49 ± 0,04

Fonte: Elaborada pelo autor.

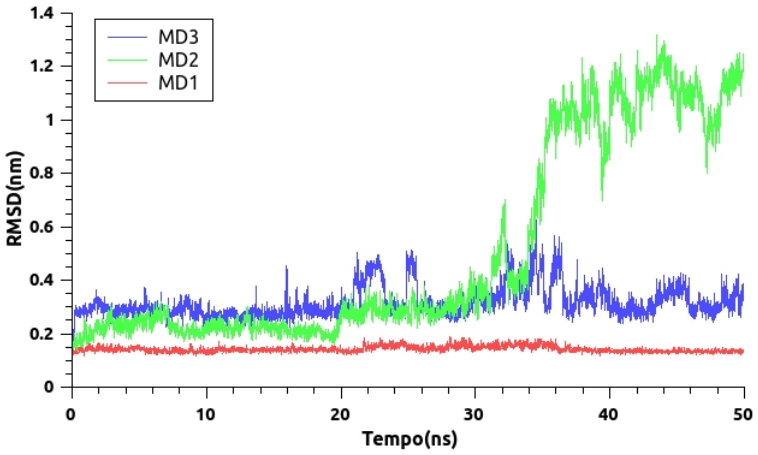
O processo de triagem baseada na estrutura do alvo LiMetRS identificou novas características não observadas em TbMetRS, como é o caso do aumento do volume molecular e presença de aceptor de hidrogênio do sítio AP. Considerando a inexistência de estudos para planejamento de fármacos tendo como alvo LiMetRS.



(a) Distância entre o átomo N2 do ligante e O2 do Asp50 do receptor (distância entre o doador e o aceptor).



(b) Distância entre o centro de massa do anel da Tyr235 e do ligante, envolvidos em uma interação π-*stacking*.



(c) RMSD do composto ZINC9659242.

Figura 31: RMSD do ligante e distâncias das interações intermoleculares entre o ligante e os resíduos do sítio ativo durante as simulações.

Fonte: Elaborada pelo autor.

5 CONCLUSÃO

Nessa dissertação de mestrado foram apresentados estudos de modelagem molecular em LBDD e SBDD para a leishmaniose visceral. Foram desenvolvidos modelos de QSAR com alta capacidade de correlação para o conjunto treinamento e predição para o conjunto teste. Análises de agrupamento, ainda pouco exploradas em técnicas de regressão, se mostraram úteis para o aumento da robustez estatística e capacidade preditiva dos modelos de QSAR-2D e QSAR-4D. No contexto da modelagem de QSAR, os métodos de agrupamento hierárquico são frequentemente utilizados para a definição dos conjuntos teste e treinamento, no entanto, poucas publicações exploram este tipo de análise para definir conjuntos independentes. Além de apresentarem boa capacidade de predição externa, os mapas de contribuição 2D e contorno 4D forneceram informações relevantes acerca das características estruturais e conformacionais mais significativamente correlacionadas com a variação da atividade biológica. Os mapas de contribuição 2D mostraram que a substituição por halogênios, e pelo grupo metoxi, contribuem positivamente para a atividade biológica, no entanto, grupos nitro, em geral, afetam negativamente a atividade. Os mapas de contorno 4D indicaram que as regiões acessíveis aos grupos nitrofenila, hidroxi-indol e trimetoxi são as que mais afetam a atividade leishmanicida.

Nos estudos em SBDD, uma estrutura tridimensional validada foi gerada para a LiMetRS. Esta enzima é essencial para a reprodução do parasita e, portanto, é um alvo molecular promissor para o planejamento de fármacos para a leishmaniose visceral. O modelo gerado reproduz a configuração conformacional do estado-I da enzima, um aspecto fundamental para estudos de modelagem envolvendo a descoberta de novos ligantes. A análise de proteínas homólogas revelou um padrão recorrente no modo de ligação de inibidores que envolve aminoácidos específicos. Foi desenvolvida uma estratégia envolvendo a filtragem de bancos de dados, docagem molecular, análise de interações intermoleculares e dinâmica molecular. O algoritmo desenvolvido neste trabalho permite a triagem em larga escala de moléculas e a análise automatizada de interações ligante-receptor. Essa ferramenta é uma alternativa de análise em larga escala que confere mais eficiência ao processo de seleção de ligantes em comparação ao método tradicional de inspeção visual. Ao término do processo de SBDD, dez ligantes foram selecionados para estudos posteriores. Os estudos revelaram características ainda não observadas para o estado-I da enzima homóloga TbMetRS, como a acomodação de grupos mais volumosos no sítio de interação e a presença de aceptores de hidrogênio do sítio AP. Os resultados indicaram características importantes de compostos com potente atividade leishmanicida e do modo de interação entre inibidores e a LiMetRS. Este conhecimento pode contribuir para o avanço das pesquisas em novas terapias para a leishmaniose visceral.

REFERÊNCIAS

1 MOLYNEUX, D. H.; SAVIOLI, L.; ENGELS, D. Neglected tropical diseases: progress towards addressing the chronic pandemic. **The Lancet**, Elsevier, v. 389, n. 10066, p. 312–325, 2017.

2 LIESE, B.; ROSENBERG, M.; SCHRATZ, A. Programmes, partnerships, and governance for elimination and control of neglected tropical diseases. **The Lancet**, Elsevier, v. 375, n. 9708, p. 67–76, 2010.

3 BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Manual de vigilância e controle da leishmaniose visceral**. Brasília: Ministério da Saúde, 2014. 120 p.

4 LEISHMANIASIS. PAHO/WHO | Pan American Health Organization. Disponível em: https://www.paho.org/en/topics/leishmaniasis. Acesso em: 10 nov. 2021.

5 SUNDAR, S.; RAI, M. Laboratory diagnosis of visceral leishmaniasis. **Clinical and Diagnostic Laboratory Immunology**, v. 9, n. 5, p. 951–958, 2002.

6 READY, P. D. Epidemiology of visceral leishmaniasis. **Clinical Epidemiology**, Dove Press, v. 6, p. 147, 2014.

7 GOURBAL, B. *et al.* Drug uptake and modulation of drug resistance in leishmania by an aquaglyceroporin. **Journal of Biological Chemistry**, Elsevier, v. 279, n. 30, p. 31010–31017, 2004.

8 POLONIO, T.; EFFERTH, T. Leishmaniasis: drug resistance and natural products. **International Journal of Molecular Medicine**, Spandidos Publications, v. 22, n. 3, p. 277–286, 2008.

9 FRÉZARD, F.; DEMICHELI, C.; RIBEIRO, R. R. Pentavalent antimonials: new perspectives for old drugs. **Molecules**, Molecular Diversity Preservation International, v. 14, n. 7, p. 2317–2336, 2009.

10 DIMASI, J. A.; GRABOWSKI, H. G.; HANSEN, R. W. Innovation in the pharmaceutical industry: new estimates of r&d costs. **Journal of Health Economics**, Elsevier, v. 47, p. 20–33, 2016.

11 LOMBARDINO, J. G.; LOWE, J. A. The role of the medicinal chemist in drug discovery—then and now. **Nature Reviews Drug Discovery**, Nature Publishing Group, v. 3, n. 10, p. 853–862, 2004.

12 BACILIERI, M.; MORO, S. Ligand-based drug design methodologies in drug discovery process: an overview. **Current Drug Discovery Technologies**, Bentham Science Publishers, v. 3, n. 3, p. 155–165, 2006.

13 MORGON, N. **Métodos de química teórica e modelagem molecular**. São Paulo: Editora Livraria da Física, 2007. ISBN 9788588325876.

14 MONTANARI, C. A. **Química medicinal**: métodos e fundamentos em planejamento de fármacos. São Paulo: Edusp, 2011.

15 ROY, K.; KAR, S.; DAS, R. **Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment**. [*S.l.*]: Elsevier Science, 2015. ISBN 9780128016336.

16 SCIOR, T. *et al.* How to recognize and workaround pitfalls in qsar studies: a critical review. **Current Medicinal Chemistry**, Bentham Science Publishers, v. 16, n. 32, p. 4297–4313, 2009.

17 GUIDO, R. V.; ANDRICOPULO, A. D.; OLIVA, G. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. **Estudos Avançados**, SciELO Brasil, v. 24, n. 70, p. 81–98, 2010.

18 PUZYN, T.; LESZCZYNSKI, J.; CRONIN, M. **Recent advances in QSAR studies**: methods and applications. Netherlands: Springer, 2010. (Challenges and Advances in Computational Chemistry and Physics). ISBN 9781402097836.

19 GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. **Analytica Chimica Acta**, Elsevier, v. 185, p. 1–17, 1986.

20 GOLBRAIKH, A.; TROPSHA, A. Beware of q2! **Journal of Molecular Graphics and Modelling**, Elsevier, v. 20, n. 4, p. 269–276, 2002.

21 JAWORSKA, J.; NIKOLOVA-JELIAZKOVA, N.; ALDENBERG, T. Qsar applicability domain estimation by projection of the training set in descriptor space: a review. **Alternatives to Laboratory Animals**, SAGE Publications, v. 33, n. 5, p. 445–459, 2005.

22 SAHIGARA, F. *et al.* Comparison of different approaches to define the applicability domain of qsar models. **Molecules**, Molecular Diversity Preservation International, v. 17, n. 5, p. 4791–4810, 2012.

23 HOPFINGER, A. *et al.* Construction of 3d-qsar models using the 4d-qsar analysis formalism. **Journal of the American Chemical Society**, ACS Publications, v. 119, n. 43, p. 10509–10524, 1997.

24 MARTINS, J. P. A. *et al.* Lqta-qsar: a new 4d-qsar methodology. **Journal of Chemical Information and Modeling**, ACS Publications, v. 49, n. 6, p. 1428–1436, 2009.

25 KELLY, P. *et al.* Targeting trna-synthetase interactions towards novel therapeutic discovery against eukaryotic pathogens. **PLoS Neglected Tropical Diseases**, Public Library of Science, v. 14, n. 2, p. e0007983, 2020.

26 SILVIAN, L. F.; WANG, J.; STEITZ, T. A. Insights into editing from an ile-trna synthetase structure with trnaile and mupirocin. **Science**, American Association for the Advancement of Science, v. 285, n. 5430, p. 1074–1077, 1999.

27 ZHOU, H. *et al.* Atp-directed capture of bioactive herbal-based medicine on human trna synthetase. **Nature**, Nature Publishing Group, v. 494, n. 7435, p. 121–124, 2013.

28 ROCK, F. L. *et al.* An antifungal agent inhibits an aminoacyl-trna synthetase by trapping trna in the editing site. **Science**, American Association for the Advancement of Science, v. 316, n. 5832, p. 1759–1761, 2007.

29 SHIBATA, S. *et al.* Selective inhibitors of methionyl-trna synthetase have potent activity against trypanosoma brucei infection in mice. **Antimicrobial Agents and Chemotherapy**, American Society for Microbiology, v. 55, n. 5, p. 1982–1989, 2011.

30 TORRIE, L. S. *et al.* Chemical validation of methionyl-trna synthetase as a druggable target in leishmania donovani. **ACS Infectious Diseases**, ACS Publications, v. 3, n. 10, p. 718–727, 2017.

31 TORRIE, L. S. *et al.* Discovery of an allosteric binding site in kinetoplastid methionyl-trna synthetase. **ACS Infectious Diseases**, ACS Publications, v. 6, n. 5, p. 1044–1057, 2020.

32 HUBBARD, T.; BLUNDELL, T. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. **Protein Engineering, Design and Selection**, Oxford University Press, v. 1, n. 3, p. 159–171, 1987.

33 ŠALI, A.; BLUNDELL, T. L. Comparative protein modelling by satisfaction of spatial restraints. **Journal of Molecular Biology**, Academic Press, v. 234, n. 3, p. 779–815, 1993.

34 MEIER, A.; SÖDING, J. Automatic prediction of protein 3d structures by probabilistic multi-template homology modeling. **PLoS Computational Biology**, Public Library of Science, v. 11, n. 10, p. e1004343, 2015.

35 SANTOS FILHO, O. A.; ALENCASTRO, R. B. d. Modelagem de proteínas por homologia. **Química Nova**, SciELO Brasil, v. 26, n. 2, p. 253–259, 2003.

36 ŠALI, A.; OVERINGTON, J. P. Derivation of rules for comparative protein modeling from a database of protein structure alignments. **Protein Science**, Wiley Online Library, v. 3, n. 9, p. 1582–1596, 1994.

37 JONES, G. *et al.* Development and validation of a genetic algorithm for flexible docking. **Journal of Molecular Biology**, Elsevier, v. 267, n. 3, p. 727–748, 1997.

38 DURRANT, J. D.; MCCAMMON, J. A. Molecular dynamics simulations and drug discovery. **BMC Biology**, BioMed Central, v. 9, n. 1, p. 1–9, 2011.

39 VIVO, M. D. *et al.* Role of molecular dynamics and related methods in drug discovery. **Journal of Medicinal Chemistry**, ACS Publications, v. 59, n. 9, p. 4035–4061, 2016.

40 FERNANDES, F. S. *et al.* Discovery of highly potent and selective antiparasitic new oxadiazole and hydroxy-oxindole small molecule hybrids. **European Journal of Medicinal Chemistry**, Elsevier, v. 201, p. 112418, 2020.

41 DIXON, S. L. *et al.* Autoqsar: an automated machine learning tool for best-practice quantitative structure–activity relationship modeling. **Future Medicinal Chemistry**, Future Science, v. 8, n. 15, p. 1825–1839, 2016.

42 MAESTRO. Schrödinger release 2016-3. 2016. Disponível em: https://www.schrodinger.com/citations. Acesso em: 25 mar. 2022.

43 CANVAS. Schrödinger release 2016-3. 2016. Disponível em: https://www.schrodinger.com/citations. Acesso em: 25 mar. 2022.

44 BAJUSZ, D.; RÁCZ, A.; HÉBERGER, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? **Journal of Cheminformatics**, BioMed Central, v. 7, n. 1, p. 1–13, 2015.

45 KELLEY, L. A.; GARDNER, S. P.; SUTCLIFFE, M. J. An automated approach for clustering an ensemble of nmr-derived protein structures into conformationally related subfamilies. **Protein Engineering, Design and Selection**, Oxford University Press, v. 9, n. 11, p. 1063–1065, 1996.

46 BERTHOLD, M. R. *et al.* Knime - the konstanz information miner: Version 2.0 and beyond. **SIGKDD Explorations Newsletter**, ACM, v. 11, n. 1, p. 26–31, nov. 2009. ISSN 1931-0145.

47 HANWELL, M. D. *et al.* Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. **Journal of Cheminformatics**, BioMed Central, v. 4, n. 1, p. 1–17, 2012.

48 FRISCH, M. J. *et al.* **Gaussian09 Revision E.01**. Gaussian Inc. Wallingford CT 2009.

49 BECK, A. D. Density-functional thermochemistry. iii. the role of exact exchange. **The Journal of Chemical Physics**, v. 98, n. 7, p. 5648–6, 1993.

50 LEE, C.; YANG, W.; PARR, R. G. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. **Physical Review B**, APS, v. 37, n. 2, p. 785, 1988.

51 RICHMOND, N. J. *et al.* Galahad: 1. pharmacophore identification by hypermolecular alignment of ligands in 3d. **Journal of Computer-Aided Molecular Design**, Springer, v. 20, n. 9, p. 567–587, 2006.

52 GASTEIGER, J.; MARSILI, M. A new model for calculating atomic charges in molecules. **Tetrahedron Letters**, Elsevier, v. 19, n. 34, p. 3181–3184, 1978.

53 GASTEIGER, J.; MARSILI, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. **Tetrahedron Letters**, Elsevier, v. 36, n. 22, p. 3219–3228, 1980.

54 SPOEL, D. V. D. *et al.* Gromacs: fast, flexible, and free. **Journal of Computational Chemistry**, Wiley Online Library, v. 26, n. 16, p. 1701–1718, 2005.

55 BERENDSEN, H. J.; SPOEL, D. van der; DRUNEN, R. van. Gromacs: a message-passing parallel molecular dynamics implementation. **Computer Physics Communications**, Elsevier, v. 91, n. 1-3, p. 43–56, 1995.

56 SCHULER, L. D.; DAURA, X.; GUNSTEREN, W. F. V. An improved gromos96 force field for aliphatic hydrocarbons in the condensed phase. **Journal of Computational Chemistry**, Wiley Online Library, v. 22, n. 11, p. 1205–1218, 2001.

57 CHANDRASEKHAR, I. *et al.* A consistent potential energy parameter set for lipids: dipalmitoylphosphatidylcholine as a benchmark of the gromos96 45a3 force field. **European Biophysics Journal**, Springer, v. 32, n. 1, p. 67–77, 2003.

58 PARRINELLO, M.; RAHMAN, A. Crystal structure and pair potentials: A molecular-dynamics study. **Physical Review Letters**, American Physical Society, v. 45, p. 1196–1199, Oct 1980.

59 BERENDSEN, H. J. *et al.* Molecular dynamics with coupling to an external bath. **The Journal of Chemical Physics**, American Institute of Physics, v. 81, n. 8, p. 3684–3690, 1984.

60 MELO, E. B. de; FERREIRA, M. M. Four-dimensional structure–activity relationship model to predict hiv-1 integrase strand transfer inhibition using lqta-qsar methodology. **Journal of Chemical Information and Modeling**, ACS Publications, v. 52, n. 7, p. 1722–1732, 2012.

61 KIM, S.; CHO, K.-H. Pyqsar: A fast qsar modeling platform using machine learning and jupyter notebook. **Bulletin of the Korean Chemical Society**, Wiley Online Library, v. 40, n. 1, p. 39–44, 2019.

62 CONSORTIUM, T. U. Uniprot: the universal protein knowledgebase in 2021. **Nucleic Acids Research**, Oxford University Press, v. 49, n. D1, p. D480–D489, 2021.

63 BERMAN, H. M. *et al.* The protein data bank. **Nucleic Acids Research**, Oxford University Press, v. 28, n. 1, p. 235–242, 2000.

64 WEBB, B.; SALI, A. Comparative protein structure modeling using modeller. **Current Protocols in Bioinformatics**, Wiley Online Library, v. 54, n. 1, p. 5–6, 2016.

65 LASKOWSKI, R. A. *et al.* Procheck: a program to check the stereochemical quality of protein structures. **Journal of Applied Crystallography**, International Union of Crystallography, v. 26, n. 2, p. 283–291, 1993.

66 SCHRODINGER, L. **The pymol molecular graphics system**. [*S.l.: s.n.*]: Version, 2015.

67 STERLING, T.; IRWIN, J. J. Zinc 15–ligand discovery for everyone. **Journal of Chemical Information and Modeling**, ACS Publications, v. 55, n. 11, p. 2324–2337, 2015.

68 SALENTIN, S. *et al.* Plip: fully automated protein–ligand interaction profiler. **Nucleic Acids Research**, Oxford University Press, v. 43, n. W1, p. W443–W447, 2015.

69 ZHAO, H. *et al.* Discovery of brd4 bromodomain inhibitors by fragment-based high-throughput docking. **Bioorganic & Medicinal Chemistry Letters**, Elsevier, v. 24, n. 11, p. 2493–2496, 2014.

70 BJELKMAR, P. *et al.* Implementation of the charmm force field in gromacs: analysis of protein stability effects from correction maps, virtual interaction sites, and water models. **Journal of Chemical Theory and Computation**, ACS Publications, v. 6, n. 2, p. 459–466, 2010.

71 VANOMMESLAEGHE, K. *et al.* Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. **Journal of Computational Chemistry**, Wiley Online Library, v. 31, n. 4, p. 671–690, 2010.

72 VANOMMESLAEGHE, K.; MACKERELL JUNIOR, A. D. Automation of the charmm general force field (cgenff) i: bond perception and atom typing. **Journal of Chemical Information and Modeling**, ACS Publications, v. 52, n. 12, p. 3144–3154, 2012.

73 BUSSI, G.; DONADIO, D.; PARRINELLO, M. Canonical sampling through velocity rescaling. **The Journal of Chemical Physics**, American Institute of Physics, v. 126, n. 1, p. 014101, 2007.

74 BERENDSEN, H. J. *et al.* Molecular dynamics with coupling to an external bath. **The Journal of Chemical Physics**, American Institute of Physics, v. 81, n. 8, p. 3684–3690, 1984.

75 HESS, B. *et al.* Lincs: a linear constraint solver for molecular simulations. **Journal of Computational Chemistry**, Wiley Online Library, v. 18, n. 12, p. 1463–1472, 1997.

76 DARDEN, T.; YORK, D.; PEDERSEN, L. Particle mesh ewald: An n log (n) method for ewald sums in large systems. **The Journal of Chemical Physics**, American Institute of Physics, v. 98, n. 12, p. 10089–10092, 1993.

77 KHAN, A. U. *et al.* Descriptors and their selection methods in qsar analysis: paradigm for drug design. **Drug Discovery Today**, Elsevier, v. 21, n. 8, p. 1291–1302, 2016.

78 SANTOS FILHO, O. A.; CHERKASOV, A. Using molecular docking, 3d-qsar, and cluster analysis for screening structurally diverse data sets of pharmacological interest. **Journal of Chemical Information and Modeling**, ACS Publications, v. 48, n. 10, p. 2054–2065, 2008.

79 KADAM, R. U.; ROY, N. Cluster analysis and two-dimensional quantitative structure–activity relationship (2d-qsar) of pseudomonas aeruginosa deacetylase lpxc inhibitors. **Bioorganic & Medicinal Chemistry Letters**, Elsevier, v. 16, n. 19, p. 5136–5143, 2006.

80 YAN, X.-F. *et al.* A comparison of semiempirical and first principle methods for establishing toxicological qsars of nitroaromatics. **Journal of Molecular Structure**: THEOCHEM, Elsevier, v. 764, n. 1-3, p. 141–148, 2006.

81 ERIKSSON, L. *et al.* Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based qsars. **Environmental Health Perspectives**, v. 111, n. 10, p. 1361–1375, 2003.

82 MELO, E.; FERREIRA, M. A 4d structure-activity relationship model to predict hiv-1 integrase strand transfer inhibition using the lqta-qsar methodology. **Journal of Chemical Information and Modeling**, v. 52, n. 7, p. 1722–1732, 2012.

83 INGVARSSON, H.; UNGE, T. Flexibility and communication within the structure of the mycobacterium smegmatis methionyl-trna synthetase. **The FEBS Journal**, Wiley Online Library, v. 277, n. 19, p. 3947–3962, 2010.

84 KOH, C. Y. *et al.* Structures of trypanosoma brucei methionyl-trna synthetase with urea-based inhibitors provide guidance for drug design against sleeping sickness. **PLoS Neglected Tropical Diseases**, Public Library of Science, v. 8, n. 4, p. e2775, 2014.

85 KOH, C. Y. *et al.* Distinct states of methionyl-trna synthetase indicate inhibitor binding by conformational selection. **Structure**, Elsevier, v. 20, n. 10, p. 1681–1691, 2012.

86 MORRIS, A. L. *et al.* Stereochemical quality of protein structure coordinates. **Proteins**: Structure, Function, and Bioinformatics, Wiley Online Library, v. 12, n. 4, p. 345–364, 1992.

87 SHIBATA, S. *et al.* Urea-based inhibitors of trypanosoma brucei methionyl-trna synthetase: selectivity and in vivo characterization. **Journal of Medicinal Chemistry**, ACS Publications, v. 55, n. 14, p. 6342–6351, 2012.

88 PEDRÓ-ROSA, L. *et al.* Identification of potent inhibitors of the trypanosoma brucei methionyl-trna synthetase via high-throughput orthogonal screening. **Journal of Biomolecular Screening**, SAGE Publications, v. 20, n. 1, p. 122–130, 2015.

89 HUANG, W. *et al.* Structure-guided design of novel trypanosoma brucei methionyl-trna synthetase inhibitors. **European Journal of Medicinal Chemistry**, Elsevier, v. 124, p. 1081–1092, 2016.

90 ZHANG, Z. *et al.* 5-fluoroimidazo [4, 5-b] pyridine is a privileged fragment that conveys bioavailability to potent trypanosomal methionyl-trna synthetase inhibitors. **ACS Infectious Diseases**, ACS Publications, v. 2, n. 6, p. 399–404, 2016.

91 DEVINE, W. G. *et al.* From cells to mice to target: characterization of neu-1053 (sb-443342) and its analogues for treatment of human african trypanosomiasis. **ACS Infectious Diseases**, ACS Publications, v. 3, n. 3, p. 225–236, 2017.

92 HUANG, W. *et al.* Optimization of a binding fragment targeting the “enlarged methionine pocket” leads to potent trypanosoma brucei methionyl-trna synthetase inhibitors. **Bioorganic & Medicinal Chemistry Letters**, Elsevier, v. 27, n. 12, p. 2702–2707, 2017.

93 PHATAK, S. S.; GATICA, E. A.; CAVASOTTO, C. N. Ligand-steered modeling and docking: A benchmarking study in class a g-protein-coupled receptors. **Journal of Chemical Information and Modeling**, ACS Publications, v. 50, n. 12, p. 2119–2128, 2010.

Apêndices

APÊNDICE A – CÓDIGO DO DOMÍNIO DE APLICABILIDADE MÉTODO *CONVEX-HULL*

Programa desenvolvido em Python2.7 pelo autor.

A.1 main.py

```
# -*- coding: utf-8 -*-
from scipy.spatial import ConvexHull, convex_hull_plot_2d
import matplotlib.pyplot as plt
import numpy as np
points_test=np.array([[0.3867697333114147,1.1870519017506165],
[0.33137697848302666,1.2171123610433634],
[0.39523632905235645,1.0192438530653167],
[0.39523632905235645,1.0192438530653167],
[0.3161594669040111,0.8753036004231199]]) #Array gerado pelo Knime contendo as coordenadas
apos MDS da matriz de distancia



points = np.array([[0.3227301137856712, 0.8954368077273701],
[0.36242312835496604,    1.1186140006860243],
[0.3766660691923303      ,0.8875384405989947],
[0.3999572139062636      ,0.8599390283610591],
[0.3751599800885271      ,0.8219404527491097],
[0.4828088899362922      ,0.9563208294519653],
[0.3816739840496447      ,0.9754124811928725],
[0.1867413353183405      ,0.857130692896813],
[0.34780934548124165,    1.0958544157491235],
[0.17012698102253684,    0.7459750562183596],
[0.2550952066300566      ,0.7301853717744301],
[0.3844153007753263      ,0.5432871627554956],
[0.5203676746817316      ,0.6460655918214488],
[0.46225262701229936,    0.6576019635098832],
[0.5235302523442278      ,0.6039942109290259],
[0.30989583075039095,    0.9071300211933051],
[0.3478236188341116      ,0.858772969103502],
[0.3552119745326679      ,0.8323197343153832],
[0.3040546287439916      ,1.0046054731161322],
[0.2801340581839572      ,1.1649284285857235],
[0.27693339984625626,    1.0906942400208903],
[0.2975193243240986,     1.233451867091124]])#Array gerado pelo Knime contendo as coordenadas

apos MDS da matriz de distancia

hull = ConvexHull(points)
plt.plot(points[:,0], points[:,1], 'o', label="Training")
plt.plot(points_test[:,0], points_test[:,1], 'o', label="Test")
for simplex in hull.simplices:
    plt.plot(points[simplex, 0], points[simplex, 1], 'k-')
plt.rcParams.update({'font.size':12})
```

```
plt.xlabel("MDS 1",fontweight='bold')
plt.ylabel("MDS 2",fontweight='bold')
legend = plt.legend(loc='upper center', shadow=True, fontsize='x-large')

plt.show()
```

APÊNDICE B – ANÁLISE AUTÔNOMA PARA IDENTIFICAÇÃO DE INTERAÇÕES COMO MÉTODO DE FILTRAGEM DE RESULTADOS DE *VIRTUAL SCREENING*

Programa desenvolvido em Python2.7 pelo autor. Neste código o usuário deve criar duas pastas no diretório escolhido, juntamente com os códigos em aqui apresentados. Uma das pastas deve ser nomeada como "target"onde deverá conter a estrutura do alvo no formato PDB. Outra pasta deve ser nomeada como "ligand"e deverá conter as moléculas resultantes da docagem. Os formatos podem ser "sdf"ou "mol2", porém devem ser alterados como indicado no cógido "lingand_complex_preparation.py"

B.1 main.py

```
from inputs import *
from ligand_complex_preparation import *
from run_plip import *
from output import *

def main():
    #definindo as variaveis de input
    array_input, array_input_index, array_index, restriction_number=inputs()
    #imprimindo as variaveis
    print_inputs(restriction_number,array_input_index)
    ligand_complex_preparation()#gerar PDB com alvo e ligante
    #armazenar conjunto de dados resultante das restricoes
    database_result=run_plip(restriction_number, array_index, array_input)
    #imprimir em arquivo de txt as moleculas pos filtro
    output(database_result,array_input_index,restriction_number)


if __name__ == "__main__":
    main()
```

B.2 inputs.py

```
# -*- coding: utf-8 -*-

def inputs():
    #definindo inputs do programa
    array_input=[]
    restriction_dictionary={} #dicionario para mudar o index
    array_index=[]
    array_input_index=[]
    #numero de ligacoes que deseja restringir
    try:
        restriction_number=int(raw_input("Escolha o numero de restricoes q desejafazer \n"))
    except ValueError:
        print ("Deve ser um numero")
    #definindo as restricoes dada o numero de restricoes
    for restriction in range(restriction_number):
```

```
        bond_restriction_options=['Hydrophobic Interactions','Hydrogen Bonds',
'Water Bridges','Salt Bridges','pi-Stacking', 'pi-Cation Interactions','Halogen Bonds',
'Metal Complexes']
        bond_restriction_index=int(input("Escolha uma ligacao:\n[0]=>Hydrophobic Interactions
\n[1]=>Hydrogen Bonds\n[2]=>Water Bridges\n[3]=>Salt Bridges\n[4]=>pi-Stacking
\n[5]=>pi-Cation Interactions\n[6]=>Halogen Bonds\n[7]=>Metal Complexes\n"))
        #index da restricao escolhida
        bond_restriction=bond_restriction_options[bond_restriction_index]
        residue=raw_input("Escolha um ligante no seguinte formato: ALA\n")
        if type(residue) is not str:
            print ("Deve ser um aminoacido")
        residue_number=int(raw_input("Escolha o numero ligante de acordo com sua enumeracao
correta\n"))
        array_input.append((bond_restriction, residue, residue_number))
        restriction_dictionary["Restr{0}".format(restriction)]=array_input[restriction]
    for new_index in restriction_dictionary:
        array_index.append(new_index)
    array_index=sorted(array_index)
    array_input_index=zip(array_index, array_input)
    return array_input, array_input_index, array_index, restriction_number


def print_inputs(restriction_number,array_input_index):
    print("-------------------------------------------------------------------------------")
    print ("{:<20} {:<15} {:<10} {:<10}".format("Restr number","Bond", "Aminoacid","Number"))
    print("-------------------------------------------------------------------------------")
    for index in range(restriction_number):
        restriction=array_input_index[index][0]
        array_input=array_input_index[index][1]
        #print data
        print ("{:<20} {:<15}".format(restriction,array_input))
```

B.3 ligand_complex_preparation.py

```
# -*- coding: utf-8 -*-
import pybel
import openbabel
import os
import sys
from pymol import cmd
import pymol
from os import path

def ligand_complex_preparation():
    directory=os.getcwd()
    result_directory="complex_PDB"
    path_result_directory=os.path.join(directory,result_directory)
    check_complex_directory=path.exists(path_result_directory)
    if check_complex_directory is True:
        pass
    else:
        #criar pasta para resultados
        os.mkdir(path_result_directory)
        print path_result_directory
```

```
        ligand_directory=directory+"/ligand/"
        target_directory=directory+"/target/"
        target=os.listdir(target_directory)[0]
        pymol.pymol_argv = ['pymol', '-qc'] + sys.argv[1:]
        pymol.finish_launching()
        cmd = pymol.cmd
        #rodar sobre todos ligantes
        for ligands in os.listdir(ligand_directory):
            ligands_ext=ligands.split(".",1)
            ligands_name=ligands_ext[0]
            cmd.reinitialize()
            #carregar alvo
            cmd.load(target_directory+target)
            #mol2 ou sdf, usuario deve alterar
            cmd.load (ligand_directory+ligands,"ligand" ,0,"sdf" ,1,-1,0)
            #salvar resultado em pasta criada
            cmd.save("%s/%s_complex.pdb" %(path_result_directory,ligands_name), "all")
        cmd.quit()
```

B.4 run_plip.py

```
# -*- coding: utf-8 -*-
import os
import pandas as pd
import re

def run_plip(restriction_number, array_index, array_input):
    directory=os.getcwd()
    directory_complex=directory+"/complex_PDB/"
    array_result=[]
    #criar um vetor nulo do com o numero de restricoes
    for restriction in range(restriction_number):
        array_result.append([])
    #rodar sobre todos arquivos complexados da pasta criada
    for filenames in os.listdir(directory_complex):
        #caminhho para plipcmd, deve ser alterada pelo usuario
        os.system('python2.7 /home/pliptool/plip/plipcmd.py -f %s%s -t '%(directory_complex,
        filenames))
        print("Running PLIP with %s as ligand" %filenames)
        #separar cada subunidadade do vetor, inputs_var=[[lig1,res1,res_nmbr1],...]
        for i in range(restriction_number):
            bond=array_input[i][0]
            residue=array_input[i][1]
            residue_number=array_input[i][2]
            with open("report.txt") as report:
                report_informartion=report.read()
               #procurar ligacao definida bond_finder=report_informartion.find("%s" %bond)
                if bond_finder !=-1:
                   #index da ligacao no arquivo report.txt
                   bond_report_index=report_informartion.index("%s" %bond)
                   end_bond_report_information=report_informartion.find("+\n\n\n",
                   bond_report_index)
                   #extrai intervalo do arquivo report contendo as informacoes da ligacao
                   between=(report_informartion[bond_report_index:end_bond_report_
```

```
                    information])
                    if re.search("%s\s*\W\s%s" %(residue_number, residue), between):
                        array_result[i].append(filenames)
                    else:
                        continue
                else:
                    continue
    #criar um dataframe com nome pvs contendo os ligantes que fazem restricao
    pvs=pd.DataFrame(array_result, index=array_index)
    pvs=pvs.T
    print pvs
    return pvs
```

B.5 output.py

```
from inputs import *

def output(pvs,array_input_index, restriction_number):
#    tabular=print_inputs(restriction_number,array_input_index)
    print ("&————————>␣AND\n|————————>␣OR")
    (teste, teste_raw)=((list(input("set(pvs.Restr0)␣&␣set(pvs.Restr1)\n"))),
    raw_input("Write␣again␣to␣confirm"))
    print teste_raw
    txt_result=open("resultPOSVS.txt", "w")
    txt_result.write("———Results———␣\n")
    #write table in txt file
    txt_result.write("\n———————————————————————\n")
    txt_result.write("{:<20}␣{:<15}␣{:<10}
␣␣␣␣{:<10}".format("Restr␣number","Bond", "Aminoacid","Number"))
    txt_result.write("\n———————————————————————\n")
    for index in range(restriction_number):
        restriction=array_input_index[index][0]
        array_input=array_input_index[index][1]
        #print data
        txt_result.write("{:<20}␣{:<15}".format(restriction,array_input))
    txt_result.write("\n\n\n\n————————————Restriction␣choosed:␣%s————————————\n" %teste_raw)
    txt_result.write("\n\n&————————>␣AND\n|————————>␣OR")
    for i in teste:
        txt_result.write("\n␣%s\n" %i)
```