PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TRIAGEM VIRTUAL EM BANCO DE DADOS DE LIGANTES CONSIDERANDO PROPRIEDADES FÍSICO-QUÍMICAS DE UM MODELO DE RECEPTOR TOTALMENTE FLEXÍVEL

CHRISTIAN VAHL QUEVEDO

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz
Co-Orientador: Prof. Dr. Osmar Norberto de Souza

Porto Alegre
2016

Ficha Catalográfica

Q5 t Quevedo, Christian Vahl

Triagem Virtual em Banco de Dados de Ligantes Considerando Propriedades Físico-químicas de um Modelo de Receptor Totalmente Flexível / Christian Vahl Quevedo . – 2016.

161 f.

Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz.

Co-orientador: Prof. Dr. Osmar Norberto de Souza.

1. Planejamento Racional de Fármacos. 2. triagem virtual baseada em estrutura. 3. modelo de receptor totalmente flexível. 4. docagem molecular. 5. seleção de ligantes. I. Ruiz, Duncan Dubugras Alcoba. II. Norberto de Souza, Osmar. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).

TERMO DE APRESENTAÇÃO DE TESE DE DOUTORADO

Tese intitulada "Triagem Virtual em Banco de Dados de Ligantes Considerando Propriedades Físico-químicas de um Modelo de Receptor Totalmente Flexível" apresentada por Christian Vahl Quevedo como parte dos requisitos para obtenção do grau de Doutor em Ciência da Computação, aprovada em 21 de janeiro de 2016 pela Comissão Examinadora:

Prof. Dr. Duncan Dubugras Alcoba Ruiz – Orientador — PPGCC/PUCRS

Prof. Dr. Osmar Norberto de Souza – Coorientador — PPGCC/PUCRS

Prof. Dr. Márcio Porto Basgalupp – — PPGCC/UNIFESP

Prof. Dr. Felipe Rech Meneguzzi - — PPGCC/PUCRS

Prof. Dr. Laurent Emmanuel Dardenne - — PPGMC/LNCC

Homologada em 17/01/2017, conforme Ata No. 001 pela Comissão Coordenadora.

Prof. Dr. Luiz Gustavo Leão Fernandes
Coordenador.

Campus Central
Av. Ipiranga, 6681 – P. 32 – sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br

AGRADECIMENTOS

O desenvolvimento deste trabalho foi o maior desafio da minha carreira. Muitas pessoas colaboraram para que este objetivo se tornasse realidade. É com extrema satisfação que dedico este espaço para expressar o meu mais profundo agradecimento a estas pessoas.

Começo agradecendo a minha família, pois sempre foram uma base sólida nos momentos difíceis. Agradeço primeiramente aos meus pais, que abriram mão de muita coisa para me proporcionar a chance de estudar em um local adequado. Agradeço também a minha esposa Andréia pela cumplicidade e apoio neste tempo de ausência. As dificuldades que soubemos enfrentar juntos serão o alicerce da nossa própria família.

Também preciso agradecer especialmente a três famílias que foram essenciais ao longo da minha jornada: Nádia e Cláudio, Marisol e Fernando e meus padrinhos Ivone e Leci. Sem o apoio delas, certamente eu não teria conseguido. Especialmente meus padrinhos, que cuidaram de mim por mais de 8 anos como um filho.

Quando terminei a minha graduação tinha duas oportunidades de afazer o mestrado: UFRGS e PUCRS. Resolvi seguir os conselhos do Rodrigo e optei pela PUCRS. Sempre fiquei com o pensamento se realmente tinha feito a escolha certa, pois na UFRGS havia uma segurança maior com relação a bolsa.

Agradeço ao Prof. Duncan e ao Prof. Osmar pela oportunidade de crescimento profissional e pelas pacienciosas e esclarecedoras orientações.

Por fim, agradeço a PUCRS por proporcionar este ambiente enriquecedor. Poucos são os pesquisadores neste país que tem a sua disposição uma infraestrutura de pesquisa como a da PUCRS. Agradeço imensamente a CAPES, a FAPERGS e ao Projeto Internacional Marie Curie pelo financiamento dos meus estudos.

TRIAGEM VIRTUAL EM BANCO DE DADOS DE LIGANTES CONSIDERANDO PROPRIEDADES FÍSICO-QUÍMICAS DE UM MODELO DE RECEPTOR TOTALMENTE FLEXÍVEL

RESUMO

Modelos farmacofóricos têm sido amplamente utilizados no processo de triagem virtual de ligantes, permitindo selecionar ligantes que contenham as propriedades físico-químicas essenciais em um arrranjo espacial adequado. Essas propriedades são obtidas a partir da avaliação das interações similares identificadas de complexos receptor-ligante conhecidos. Atualmente, esses modelos farmacofóricos baseados em ligantes são dependentes das características físico-químicas presentes nos complexos receptor-ligante conhecidos. Desta forma, o modelo farmacofórico gerado pode negligenciar as proteínas que não possuem ligantes complexados conhecidos e cujas propriedades físico-químicas não estabelecem interação nos complexos avaliados. Ou seja, regiões dentro da cavidade que não interagem com o conjunto de ligantes geradores do modelo farmacofórico e que podem permitir a interação de ligantes estruturalmente diferentes não estão incluídas nessa busca seletiva. Além disso, diversos autores têm mostrado que não considerar a flexibilidade da proteína no processo de seleção de candidatos a fármacos acaba limitando significativamente a precisão dos resultados. Assim, esta tese apresenta um novo método para realizar uma triagem virtual de ligantes baseada na avaliação das propriedades físico-químicas 3D da cavidade de ligação do substrato, e sem a presença de ligantes complexados, de estruturas representativas de um modelo de Receptor Totalmente Flexível (FFR). O resultado desse método permite identificar modelos farmacofóricos 3D de regiões flexíveis que podem não ser obtidos de modelos desenvolvidos apenas a partir de estruturas cristalinas de complexos receptor-ligante. Uma lista de hipóteses farmacofóricas é proposta para selecionar um conjunto de ligantes do banco de dados ZINC. Testes da eficácia desse método foram baseados em experimentos de *cross docking* com um modelo de FFR de 19,5 ns da enzima InhA de *Mycobacterium tuberculosis*. Os experimentos de docagem molecular com o conjunto de ligantes selecionado mostraram que 95,0% desse conjunto obtiveram valores negativos de FEB, sendo 20,6% desses valores melhores que os valores de FEB obtidos com experimentos de docagem com a estrutura cristalina que gerou o modelo avaliado. Esses resultados promissores comprovam que o método desenvolvido tem condições de ser uma importante ferramenta de apoio aos pesquisadores na busca por novos candidatos a fármacos, acelerando o processo de seleção dos possíveis candidatos a serem testados com modelos FFR de moléculas alvo. O método apresentado também fornece uma ótima forma de avaliar o modelo FFR empregado, possibilitando ao especialista de domínio identificar se as regiões obtidas são realmente acessíveis na proteína investigada.

Palavras-Chave: planejamento racional de fármacos, triagem virtual baseada em estrutura, modelo de receptor totalmente flexível, docagem molecular, seleção de ligantes.

VIRTUAL SCREENING IN LIGAND DATABASES CONSIDERING PHISICAL-CHEMICAL PROPERTIES OF A FULLY-FLEXIBLE RECEPTOR MODEL

ABSTRACT

Pharmacophore models have been widely used in the virtual screening, allowing to select ligands that containing the spatial arrangement of essential physico-chemical properties. These properties are obtained from the evaluation of similar interactions identified in known receptor-ligand complexes. Currently, these pharmacophore models based on ligands are dependent on the physicochemical characteristics present in the known receptor-ligand complex. Thus, the pharmacophore model generated can overlook the proteins that have no known ligands complexed and whose physical and chemical properties do not establish interaction in the evaluated complex. That is, regions in the cavity that do not interact with ligands that generate the pharmacophore model and that may allow the interaction of structurally diverse ligands are not included in the selective search. Furthermore, several authors have shown that not taking the protein's flexibility into account during the selection of drug candidates limits the result's accuracy significantly. Thus, this thesis presents a new method for performing a virtual screening of ligands based on the evaluation of the 3D physico-chemical properties of the substrate binding pocket, and without the presence of complexed ligands, of representative structures of a Fully-Flexible Receptor (FFR) model. This method allows identifying 3D pharmacophoric models of flexible regions, which cannot be obtained from 3D pharmacophore models developed only from crystal structures of the ligand-receptor complex. A list of pharmacophoric hypothesis is proposed to select a set of ligands ZINC DB. Tests of this method's efficacy were based on cross-docking experiments with the FFR model of 19.5 ns of the InhA enzyme from Mycobacterium tuberculosis. Molecular docking experiments with selected ligands showed that 95.0% of this group were negative values FEB, with 20.6% of these values that the best values obtained with FEB docking experiments with the crystalline structure that generated the rated model. These promising results show that the developed method may be an important support tool for researchers in the search for new drug candidates, accelerating the selection of possible candidates to be tested with FFR models of target molecules. The method presented also provides a great way to evaluate FFR models, enabling the domain expert to identify whether the obtained regions are really accessible in the investigated protein.

Keywords: rational drug design, structure-based virtual screening, fully-flexible receptor model, molecular docking, ligand selection.

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Abreviaturas

AMBER – (do inglês *Assisted Model Building with Energy Refinement)*

BD – Banco de Dados

CSV – Valores separados por vírgula
(do inglês *Comma-separated values*)

DM – Dinâmica Molecular

ETH – Etionamida
(do inglês *Ethionamide)*

FEB – Energia Livre de Ligação
(do inglês *Free Energy of Binding)*

FFR – Receptor Totalmente Flexível
(do inglês *Fully-Flexible Receptor)*

FLAP – *Fingerprints* de Ligantes e/ou de Proteínas
(do inglês *Fingerprints for Ligands And Proteins*)

INH – Isoniazida
(do inglês *Isoniazid)*

InhA – Enzima 2-trans-enoil ACP(CoA) Redutase de *Mycobacterium tuberculosis*

LBDD – Planejamento de Fármacos Baseado na Estrutura do Ligante
(do inglês *Ligand-Based Drug Design*)

LSH – *Hash* Sensível à Localidade
(do inglês *Locality Sensitive Hashing*)

MDR-TB – Tuberculose Multirresistente
(do inglês *MultiDrug-Resistant TuBerculosis*)

Mtb – *Mycobacterium tuberculosis*

MIF – Campos de Interação Molecular
(do *Molecular Interaction Field*)

NADH – Nicotinamida Adenina Dinucleotídeo
(do inglês *Nicotinamide adenine dinucleotide)*

P-SaMI – Padrão Múltiplas Instâncias AutoAdaptáveis
(do inglês *Self-adaptive Multiple Instances)*

PDB – Banco de Dados de Proteínas
(do inglês *Protein Data Bank)*

PIF – Isoniazida Pentacianoferrato
(do inglês *Pentacyano (isoniazid) ferrate II)*

RDD – Planejamento Racional de Fármacos
(do inglês *Rational Drug Design*)

RFFR – Receptor Totalmente Flexível Reduzido
(do inglês *Reduced Fully-Flexible Receptor)*

RMSD – Raiz Quadrada do Desvio Médio Quadrático
(do inglês *Root Mean Square Deviation)*

SBDD – Planejamento de Fármacos Baseado na Estrutura do Receptor
(do inglês *Structure-Based Drug Design*)

SDQ – Soma das Diferenças entre os Quartis

TCL – Triclosano
(do inglês *Triclosan)*

TDR-TB – Tuberculose Totalmente Resistente
(do inglês *Totally Drug-Resistant TuBerculosis*)

THT – *Trans-2-Hexadecenoyl-(N-Acetyl-Cysteamine)-Thioester*

UPGMA – Método de agrupamento baseado na média aritmética não ponderada
(do inglês *Unweighted Pair Group Method with Arithmetic mean*)

XDR-TB – Tuberculose Extensivamente Resistente
(do inglês *eXtensively Drug-Resistant TuBerculosis*)

wFReDoW – (do inglês *web Flexible Receptor Docking Workflow)*

WPGMA – Método de agrupamento baseado na média aritmética ponderada
(do inglês *Weighted Pair Group Method with Arithmetic mean*)

Sumário

1. Introdução

Nesta última década, o uso de técnicas de triagem virtual baseadas em estrutura para o descobrimento de novos fármacos tem alcançado resultados promissores. Assim, essas técnicas vêm sendo amplamente utilizadas, embora ainda se encontrem em um estado de aprimoramento [Yur14]. Diversos métodos têm sido propostos com o objetivo de reduzir custos e o tempo necessário para o desenvolvimento de novos fármacos. No entanto, o processo necessário até a aprovação de uma nova droga continua demorado, podendo durar de 10 a 15 anos [Cas07, PMD$^+$10], com um custo de, aproximadamente, 1,8 bilhão de dólares [PMD$^+$10]. Esses fatores motivam o desenvolvimento de soluções alternativas para otimizar os métodos existentes.

Atualmente, existem diversos fatores que podem influenciar a qualidade dos resultados do conjunto de candidatos à solução de uma proteína alvo como: a simulação da flexibilidade da proteína, a quantidade de pequenas moléculas a serem testadas, a correta predição da estrutura, a função de pontuação para calcular a interação entre o complexo receptor-ligante e a busca exaustiva para encontrar a melhor orientação do ligante dentro do sítio ativo [Mac06, SBBW12]. Dessa forma, os métodos desenvolvidos necessitam encontrar um equilíbrio entre o ganho proporcionado pela aplicação e o tempo despendido.

Apesar de a redução do tempo ser um dos objetivos, limitar a flexibilidade da proteína no processo de seleção de candidatos a fármacos pode reduzir significativamente a precisão dos métodos de docagem molecular e, por consequência, restringir o conjunto de candidatos à solução [BR02, TA08]. Uma forma de contornar esse problema é a utilização da simulação da flexibilidade da proteína por Dinâmica Molecular (DM). A simulação por DM é uma abordagem que fornece informações sobre o comportamento dinâmico da estrutura e da função de macromoléculas biológicas em função do tempo [ABG06, KM02]. Uma típica simulação por DM pode gerar acima de 10^4 conformações para explorar o espaço conformacional de uma proteína. A simulação captura os movimentos individuais das partículas em função do tempo [KM02]. Esse conjunto de conformações, derivado de uma trajetória de DM, é denominado modelo de Receptor Totalmente Flexível (FFR - do inglês *Fully-Flexible Receptor*) [MWRNdS11]. Esse modelo visa representar um conjunto de microestados capaz de contemplar a flexibilidade total de uma proteína.

Embora essa abordagem possibilite maiores oportunidades para a descoberta de candidatos a fármaco, ela também torna o processo de avaliação mais demorado [ABG06]. Normalmente, esses modelos FFR são utilizados por programas de docagem molecular para analisar a interação de cada pequena molécula armazenada nos Bancos de Dados (BD) de ligantes. Em cada avaliação do complexo receptor-ligante, o programa de docagem molecular simula parte da flexibilidade do ligante buscando encontrar a melhor conformação e orientação dentro da cavidade do receptor. Nesse tipo de avaliação, o programa

de docagem molecular simula a flexibilidade do ligante enquanto o modelo FFR simula a flexibilidade do receptor. O conjunto de ligantes a serem testados podem ser obtidos de BD de ligantes.

Atualmente, existem diversos BD de ligantes de acesso público disponibilizando acima de 10^5 estruturas. Um exemplo desses BD é o ZINC [ISM+12], cujo número de estruturas comercialmente disponibilizadas é superior a 20 milhões. Portanto, o alto custo computacional na utilização de simulações de DM para identificar possíveis ligantes candidatos através de experimentos de docagem molecular com os ligantes dos BD pode tornar a tarefa inviável. Por essa razão, abordagens novas e promissoras para reduzir a dimensionalidade de experimentos de *cross-docking* sem perda de informações devem ser investigadas [AL10].

1.1 Caracterização do problema

Um dos principais desafios da triagem virtual é conseguir manipular a grande quantidade de estruturas disponibilizadas de forma a permitir a seleção de candidatos à solução em um período aceitável de tempo [Yur14]. Esse problema pode ser separado em duas categorias: uma visando reduzir a quantidade de conformações do modelo FFR sem impactar drasticamente na perda de movimentos flexíveis importantes e, outra, selecionando de forma eficaz um conjunto de ligantes promissores de acordo com as propriedades do modelo FFR, sem a obrigatoriedade da existência de ligantes já cristalizados na cavidade de ligação do substrato.

1.1.1 Seleção de estruturas representativas de um modelo FFR

A exploração da flexibilidade de uma proteína possibilita a análise de diferentes funções biológicas [MSWN02]. Naturalmente, essa elevada quantidade de conformações de um receptor também proporciona um aumento na quantidade de estruturas similares. A técnica de agrupamento é bastante utilizada para reduzir o número de estruturas do receptor a serem avaliadas, gerando um conjunto de estruturas representativas. Diferentes metodologias são utilizadas para agrupar estruturas de proteínas. Essas metodologias podem variar o algoritmo de agrupamento, o número de grupos e a métrica de similaridade a ser usada.

Quando as estruturas 3D das proteínas são conhecidas, as medidas de similaridade estrutural são frequentemente escolhidas como comparativo para avaliar o grau de semelhança entre duas proteínas, devido à precisão dos resultados gerados [SP04]. Essas medidas são tipicamente baseadas no pareamento das coordenadas atômicas de diferentes conformações.

Em anos recentes, diversos trabalhos foram desenvolvidos baseando-se somente na informação estrutural fornecida pela raiz quadrada dos desvios médios quadráticos (RMSD - do inglês *Root Mean Square Deviation*) para definir o conjunto de estruturas representativas [PMF$^+$09, FPSB11, ABE$^+$13, DABR$^+$13]. No entanto, quando existe o interesse em uma região específica da proteína, por exemplo, a cavidade de ligação do substrato, a utilização dessa métrica não se mostra adequada [CLP11]. Uma análise qualitativa evidencia que o RMSD possui limitações geométricas que o tornam incapaz de distinguir conformações com regiões flexíveis.

Para ilustrar essa questão, a Figura 1.1 apresenta uma comparação entre duas proteínas obtidas de um modelo FFR de 19,5 ns da enzima InhA de *Mycobacterium tuberculosis* (*Mtb*). Essas estruturas foram selecionadas por conter o mesmo valor de RMSD, seja pela estrutura do Cα ou pela estrutura do *backbone* (N, Cα, C, and O). Analisando a Figura 1.1 é possível notar que ambas estruturas apresentam sítios de ligação diferentes. No entanto, essas estruturas são consideradas idênticas conforme o valor apresentado pela métrica do RMSD. Assim, um processo típico de agrupamento considerando somente os valores do RMSD certamente selecionaria essas estruturas dentro da mesma partição.

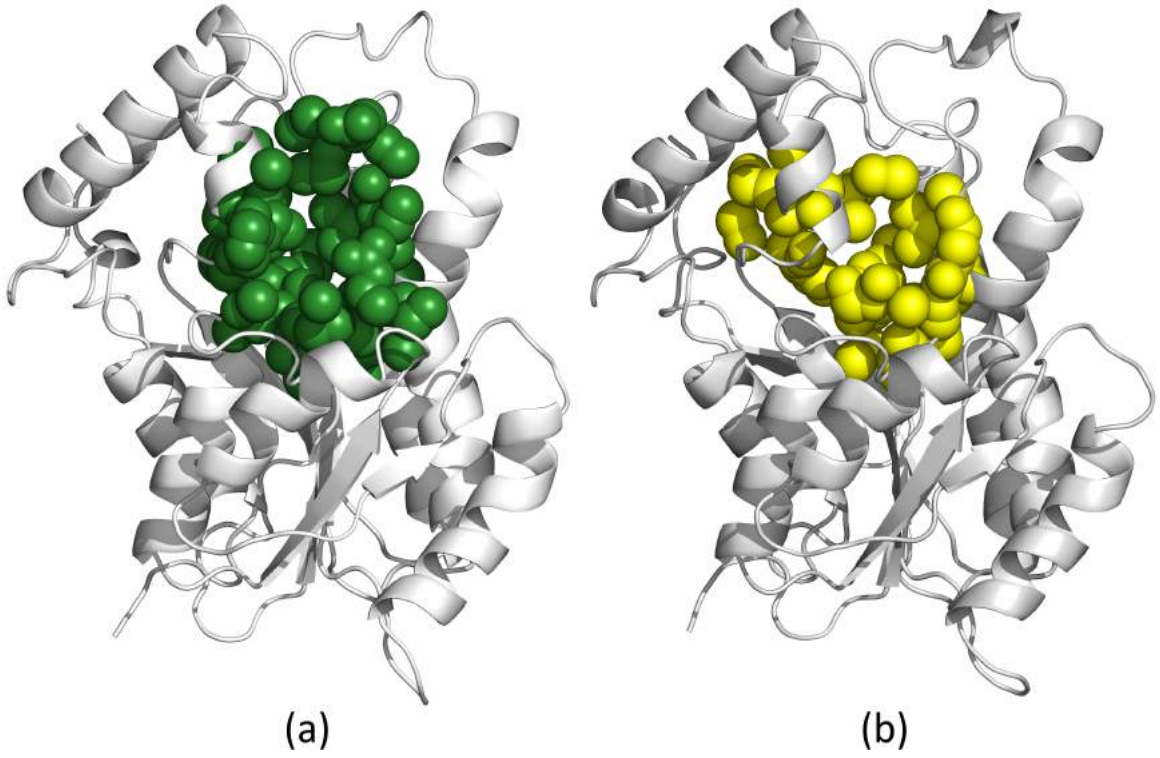


Figura 1.1 – Representação de conformações da InhA contendo o mesmo valor de RMSD (1,14 Å para o Cα e o *backbone* (N, Cα, C e O)) e que apresentam diferentes regiões da cavidade de ligação do substrato. Ambas as estruturas foram obtidas de um modelo FFR de 19,5 ns da enzima InhA de *Mycobacterium tuberculosis* [GCNdS07]. A estrutura da InhA está representada no diagrama de fitas na cor cinza e a cavidade de ligação do substrato está representada por esferas de van der Waals. (a) Em verde, a cavidade de ligação do substrato da conformação 1.903 ps com um volume 515,9 Å^3 (b) Em amarelo, a cavidade de ligação do substrato conformação 1.970 ps com um volume 1.242,0 Å^3.

A Figura 1.1 exemplifica claramente um dos problemas que ocorrem em técnicas que se baseiam na seleção de estruturas representativas considerando apenas o RMSD como métrica. Esse problema ocorre devido às restrições geométricas da fórmula do RMSD, que utiliza como base as distâncias pareadas dos átomos da proteína. A utilização somente das distâncias torna possível que diferentes arranjos espaciais possuam o mesmo valor de RMSD. Ou seja, uma simulação de DM pode conter diversas conformações com o mesmo valor de RMSD e essas estruturas projetarem sítios de ligação distintos. Além disso, outro problema constatado é que essa métrica pode apresentar diferentes resultados de acordo com o tipo de alinhamento aplicado nas estruturas comparadas. Isso pode ocorrer devido às mudanças do posicionamento atômico, influenciando diretamente nos valores das distâncias calculadas pela fórmula do RMSD.

1.1.2 Triagem virtual baseada na estrutura do receptor considerando um conjunto de estruturas representativas de um modelo FFR

Nas últimas duas décadas, diferentes abordagens têm sido desenvolvidas buscando encontrar métodos *in silico* eficazes para selecionar bons candidatos a fármacos para testes *in vitro* de forma rápida. Uma solução, complementar à apresentada na Seção 1.1.1, para reduzir esse problema é a realização de uma completa varredura em BD de ligantes para eliminar estruturas não promissoras conforme as características da proteína antes de aplicar os experimentos de docagem molecular [Was08]. Modelos farmacofóricos 3D têm sido utilizados para realizar a redução da dimensionalidade, selecionando ligantes que contenham o arranjo espacial de propriedades físico-químicas essenciais [LB12]. Essas propriedades essenciais são definidas a partir da avaliação das interações similares identificadas de complexos receptor-ligante conhecidos [LGLT10]. Dessa forma, os modelos farmacofóricos baseados em ligantes são dependentes das características físico-químicas presentes nos complexos receptor-ligante conhecidos. Esses modelos podem negligenciar ligantes promissores em proteínas que possuem regiões que não estabelecem interações com os complexos avaliados ou simplesmente não possuem ligantes complexados conhecidos [HL13]. Ou seja, as regiões dentro da cavidade, que não interagem com o conjunto de ligantes geradores do modelo farmacofórico e que podem permitir a interação de ligantes estruturalmente diferentes, não são contempladas nessa busca seletiva.

Pesquisas recentes têm explorado as propriedades da cavidade de ligação do substrato da proteína para gerar modelos farmacofóricos baseados na proteína, sem considerar as informações de ligantes complexados [TCM+08, HL12, HL13]. Tintori e colaboradores [TCM+08] apresentaram um estudo combinando diferentes abordagens, gerando um protocolo de múltiplos passos para auxiliar na geração e na aplicabilidade de modelos farmacofóricos 3D. O primeiro passo calcula o Campo de Interação Molecular (MIF - do inglês *Molecular Interaction Fields*) de cada estrutura utilizando diferentes sondas químicas. De-

pois, para reduzir o escopo, aplicam um filtro para identificar as regiões que apresentam as menores interações de energia. O segundo passo converte os pontos gerados em características farmacofóricas para que, no terceiro passo, essas características sejam refinadas. A limitação desse trabalho é que o método proposto não permite a exploração de muitas estruturas devido ao tempo necessário para processar as informações da proteína. O estudo de caso apresentado considerou um modelo flexível contendo apenas 6 conformações.

Hu e Lill [HL12] descreveram um método para reduzir a quantidade de características farmacofóricas utilizando as informações da energia liberada por moléculas de água dentro da cavidade de ligação do substrato. O estudo de caso foi realizado com uma DM contendo 1.000 conformações. Após aplicar técnicas de agrupamento, a construção do modelo farmacofórico foi aplicada em apenas 3 estruturas. Posteriormente, esses mesmos autores descreveram um estudo comparando a acurácia de modelos farmacofóricos baseados somente na estrutura da proteína com modelos farmacofóricos gerados a partir dos contatos de um conjunto de 190 complexos receptor-ligante determinados experimentalmente [HL13]. Na avaliação deles, um processo de otimização foi aplicado devido ao tamanho da cavidade e, mesmo após a redução da complexidade, os modelos farmacofóricos baseados somente na estrutura da proteína foram capazes de identificar acima de 95% dos contatos existentes entre os complexos receptor-ligante experimentalmente conhecidos. Além disso, foram identificados sítios de ligação não reconhecidos pelos modelos farmacofóricos baseados nos contatos entre os complexos receptor-ligante, corroborando a importância de avaliações mais abrangentes da cavidade alvo. A principal limitação desse trabalho está no fato de o método utilizar apenas dois tipos de farmacóforos dos tradicionais 5 e 6 normalmente utilizados na literatura.

Diversos casos na literatura têm confirmado que estratégias que consideram sistemas flexíveis obtêm melhores resultados na identificação de novos candidatos à fármacos [ABG06, TA08]. Trabalhos recentes têm utilizado trajetórias de DM para definir um conjunto de hipóteses farmacofóricas. No entanto, todos os trabalhos encontrados até o momento aplicam reduções de dimensão bastante expressivas, reduzindo as simulações a conjuntos com não mais que 10 estruturas. Essa quantidade não representa 0,1% das simulações de flexibilidade obtidas atualmente. Portanto, os métodos descritos nesta seção não se mostram adequados devido aos cortes expressivos da flexibilidade da proteína.

1.2 Objetivos e principais contribuições

Apesar dos esforços de diversos pesquisadores, grande parte dos modelos farmacofóricos baseados somente na estrutura da proteína tem utilizado apenas as informações de estruturas experimentais, reduzindo as possibilidades de uma seleção de ligantes alternativos aos já conhecidos por não considerar a flexibilidade da proteína alvo. Isso significa que as avaliações realizadas somente com estruturas cristalinas e/ou poucas estruturas de

um modelo FFR não contemplam o conjunto completo de soluções possíveis. Assim, evidenciar regiões alternativas às encontradas pelos métodos que utilizam estruturas “rígidas” constitui uma estratégia muito promissora.

Diferentemente das pesquisas abordadas na seção 1.1.2, o objetivo desta tese é o de apresentar um novo método para identificar um conjunto de hipóteses farmacofóricas alternativas de um modelo totalmente flexível, possibilitando a exploração das regiões acessíveis não contempladas por estruturas cristalinas depositadas no sítio do PDB. Essas hipóteses, obtidas da avaliação das propriedades físico-químicas 3D da cavidade de ligação do substrato de um modelo FFR sem a informação de complexos receptor-ligante experimentalmente conhecidos, são então empregadas na triagem virtual de estruturas 3D de ligantes.

Como contribuição desta pesquisa, espera-se que o método desenvolvido seja uma importante ferramenta de apoio aos pesquisadores, auxiliando na busca de novos candidatos a fármacos que possuam algumas características físico-químicas diferente dos ligantes já cristalizados e que sejam propícias ao encaixe no receptor. Assim, este método deve contribuir na aceleração do processo de seleção dos possíveis candidatos a serem testados com modelos FFR de moléculas alvo. As propriedades elencadas também fornecem uma ótima forma para avaliar o modelo FFR gerado, possibilitando ao especialista de domínio identificar se as regiões obtidas são realmente acessíveis na proteína investigada.

1.2.1 Objetivos específicos

- Realizar uma análise detalhada das estruturas cristalinas da enzima InhA de *Mycobacterium tuberculosis (Mtb)* disponibilizadas no sítio *Protein Data Bank* (PDB). Essa enzima foi usada como o estudo de caso desta tese.

- Identificar a cavidade de ligação do substrato contemplando as variações estruturais de cada conformação de um modelo FFR, armazenando as suas informações descritivas, tais como o volume e os átomos que compõem a superfície da área acessível ao solvente. Neste trabalho, um modelo FFR de 19,5 ns da enzima InhA de *Mtb* é usado como estudo de caso.

- Desenvolver e aplicar técnicas de agrupamento às conformações do modelo FFR considerando o valor do RMSD em conjunto com outras propriedades da cavidade de ligação do substrato. Esse estudo busca identificar se outras propriedades como o volume e a presença de resíduos importantes na fronteira da cavidade de ligação do substrato são suficientes para propiciar partições mais homogêneas que as partições formadas a partir de métricas como o RMSD. Nesta tese, três estudos de agrupamento são apresentados.

- Construir um amplo conjunto de hipóteses farmacofóricas 3D baseado nas características físico-químicas da estrutura 3D do receptor InhA de *Mtb* a partir do conjunto de estruturas similares identificadas pelo agrupamento das propriedades da cavidade. As regiões não contempladas por estruturas cristalinas são destacadas com o raio de interação que altera conforme a sua frequência em cada grupo. As propriedades comuns às estruturas cristalinas recebem um raio reduzido (0,2 Å), conforme pode ser visto na seção 6.4.2.

- Executar experimentos de docagem molecular, com o programa AutoDock 4 [MHL+09], com os 20 ligantes selecionados pelas hipóteses farmacofóricas 3D. A seleção de compostos capazes de propiciar bons resultados da estimativa da energia livre de ligação (FEB - do inglês *Free Energy of Binding*) com o modelo FFR e não contemplados pelas estruturas cristalinas demonstram a relevância desta tese.

1.3 Organização deste trabalho

Esta tese está organizada em 8 capítulos:

- Capítulo 2 - Fundamentação teórica: Esse capítulo apresenta conceitos fundamentais necessários a um melhor entendimento desta tese, destacando as principais características dos métodos de triagem virtual baseada em estrutura e sobre as técnicas de agrupamento utilizadas na redução de modelos FFR.

- Capítulo 3 - Proteína investigada- enzima InhA de *Mycobacterium tuberculosis (Mtb)*: Esse capítulo descreve a motivação social e as principais propriedades da enzima InhA de *Mtb*, definida como o estudo de caso desta tese. É apresentado um conjunto de estudos que descrevem importantes características físico-químicas das interações entre 20 pequenas moléculas candidatas a fármaco e a proteína InhA de *Mtb*.

- Capítulo 4 - Experimentos de docagem molecular com as estruturas cristalinas da InhA de *Mtb* presentes no *Protein Data Bank* e com o modelo FFR: Esse capítulo fornece detalhes de um novo conjunto de experimentos de docagem molecular realizados entre essas pequenas moléculas cristalizadas com as proteínas da InhA e o modelo FFR de 19,5 ns. Essas avaliações buscam um melhor entendimento da flexibilidade avaliando como são as interações *in silico* entre complexo receptor-ligante de estruturas bem conhecidas *in vivo*.

- Capítulo 5 - Seleção de estruturas representativas: Aplicação de técnicas de agrupamento considerando o valor do RMSD em conjunto com outras propriedades da cavidade de ligação do substrato. Esse capítulo descreve três métodos para particionar as conformações de DM baseado em um conjunto de propriedades da cavidade

de ligação do substrato. Cada método descrito apresenta uma avaliação das partições encontradas e o quanto reduziram a dimensionalidade do modelo FFR da enzima selecionada como o estudo de caso. As partições formadas a partir das propriedades físico-químicas da cavidade alvo utilizando farmacóforos de 3 pontos apresentaram a seleção de estruturas representativas mais específicas, possibilitando agrupamentos mais adequados em uma menor quantidade de tempo.

- Capítulo 6 - Algoritmo para a determinação de modelos farmacofóricos 3D baseado nas características físico-químicas da estrutura 3D de regiões inacessíveis por estruturas cristalinas do receptor InhA de *Mtb*: Nesse capítulo, as propriedades farmacofóricas 3D da cavidade de ligação do substrato das estruturas representativas das partições formadas no Capítulo 5 são avaliadas. Cada partição possui um peso determinado pela quantidade de conformações agrupadas. Cada propriedade farmacofórica já existente na estrutura cristalina é expressa com pequenas esferas de *van der Waals*, enquanto que propriedades farmacofóricas de regiões inacessíveis por estruturas cristalinas possuem raios de acordo com a frequência de cada propriedade. Essa marcação evidencia para o pesquisador quais propriedades farmacofóricas 3D encontradas são devidas à consideração da flexibilidade do modelo FFR.

- Capítulo 7 - Trabalhos relacionados: Nesse capítulo, um conjunto de trabalhos já publicados e relevantes para esta pesquisa são apresentados, relacionando suas contribuições com o conteúdo descrito desta tese.

- Capítulo 8 - Considerações finais: Esse capítulo apresenta as considerações finais desta pesquisa, citando as principais contribuições e limitações. Também apresenta um conjunto de oportunidades não solucionadas por este trabalho. Por fim, descreve as principais publicações deste trabalho.

2. Referencial teórico

Este Capítulo descreve conceitos fundamentais para o melhor entendimento desta tese. A primeira seção descreve as principais etapas do Planejamento Racional de Fármacos. A segunda seção apresenta o método da docagem molecular, cujo objetivo é identificar a orientação e a estrutura conformacional que apresenta a melhor interação com a cavidade de ligação do receptor. A terceira seção caracteriza o método utilizado neste trabalho para se considerar a flexibilidade da proteína, elencando suas vantagens e desvantagens. A quarta seção aborda os principais métodos de agrupamento de conformações e descreve importantes medidas de validação de partições de métodos não supervisionados. A quinta seção descreve os principais BD de pequenas moléculas públicos, elencando as suas propriedades. A sexta seção apresenta os recentes avanços dos métodos de Triagem Virtual baseados na estrutura. Por fim, a última seção apresenta as considerações finais deste Capítulo.

2.1 Planejamento racional de fármacos

No início do século XX, a busca por novos fármacos empregava metodologias baseadas no uso de testes *in vitro* de maneira aleatória, resultando em um processo caro e lento. Evidentemente, esse tipo de pesquisa não apresentava uma relação de custo-benefício adequada aos interesses das empresas farmacêuticas [PMD+10]. Com o avanço da ciência, passou-se a investir em procedimentos mais lógicos, método conhecido como Planejamento Racional de Fármacos (RDD - do inglês *Rational Drug Design*) [Kun92].

O RDD é definido como um estudo que trata do reconhecimento de moléculas capazes de ter afinidade com determinados receptores e baseia-se na interação molecular entre ligantes e suas proteínas-alvo [LR96]. Esses métodos *in silico* podem alterar as propriedades conformacionais das proteínas e ligantes de modo a ampliar a busca pelas melhores interações, antes da aplicação de testes *in vitro*, o que torna essa busca mais eficaz. O RDD consiste em 4 etapas [Kun92]:

- Etapa 1: Identificar a doença a ser tratada e isolar um alvo específico, determinando essa proteína como o receptor a ser tratado. Um especialista de domínio realiza uma análise da estrutura 3D desse receptor, determinando a cavidade alvo cuja ocorrência de interação com ligantes se deseja investigar.

- Etapa 2: Selecionar um conjunto de ligantes que apresentem interações favoráveis do complexo receptor-ligante considerando a cavidade alvo identificada na etapa 1. As diferentes orientações que determinado ligante pode assumir dentro da cavidade

alvo do receptor podem ser simuladas por programas de docagem molecular, que quantificam a interação do complexo receptor-ligante com um escore.

- Etapa 3: Sintetizar experimentalmente e avaliar se o conjunto de ligantes identificados como promissores na etapa 2 obtêm bons resultados.
- Etapa 4: Avaliar os resultados experimentais, definindo se o medicamento deve ser produzido ou se o processo deve retornar à etapa 1, de forma iterativa, com pequenas modificações das características na busca dos ligantes.

De acordo com Broughton [Bro00], o maior desafio na área de desenvolvimento de fármacos está relacionado com a predição da estrutura e a energia envolvida na interação entre os ligantes e a proteína alvo. A seção 2.2 apresenta detalhes sobre o método desenvolvido para avaliar a interação entre o complexo receptor-ligante.

2.2 Docagem molecular

A docagem molecular é o método de simulação computacional que avalia a afinidade entre o receptor e o ligante. Essa afinidade é mensurada por um escore, onde valores mais baixos significam melhores interações [LR96, Jia08]. Nos últimos anos, esse método de avaliação *in silico* tem sido normalmente empregado no RDD para predizer o melhor modo de interação, explorando as interações específicas que podem ser formadas e estimar qualitativamente a afinidade de interação do ligante [WCG11].

Atualmente existem mais de 30 programas de docagem molecular disponíveis, podendo estes variarem quanto ao algoritmo de busca e quanto a função de escore. O algoritmo de busca é utilizado para amostrar todos os graus de liberdade do ligante, gerando as diferentes orientações possíveis considerando as translações e rotações dentro da cavidade alvo. A função de pontuação avalia as poses geradas pelo algoritmo de busca e as classifica de acordo com as interações com o receptor, calculando a FEB [Rog11].

Nesta tese, o AutoDock 4.2.5 [MHL$^+$09] foi escolhido como o programa de docagem molecular a ser utilizado. Esse programa é muito bem estabelecido pela comunidade científica e também disponibiliza licenças gratuitas. Os seus resultados são ordenados pelo valor da FEB mais negativa. No caso da existência de poses com a mesma energia livre, o valor de RMSD, calculado a partir da estrutura de referência, é utilizado. Da mesma forma que a FEB, quanto menor o valor do RMSD, melhor será o resultado, sendo o menor valor possível igual a 0,0 Å. A Figura 2.1 mostra um exemplo da melhor posição encontrada pela docagem molecular feita entre uma conformação da InhA de *Mtb* e a coenzima NADH.

Normalmente, as aplicações de docagem molecular utilizando modelos FFR consideram a estrutura do receptor como rígida, enquanto a molécula do ligante pode ter uma flexibilidade parcial, variando os ângulos de torção. Assim, o ligante a ser testado assume

diferentes conformações estruturais em diversas posições dentro do sítio ativo da molécula receptora. Desta maneira, o tempo computacional necessário para executar esses experimentos é bastante custoso, consumindo em média 1 minuto para cada avaliação (o tempo pode variar de acordo com os parâmetros do programa, do *hardware* e do ligante). A quantidade de ligantes a serem avaliados depende do BD de pequenas moléculas selecionado e dos filtros de seleção de estruturas que são aplicados. A próxima seção apresenta detalhes sobre o método mais tradicional de simulação de parte da flexibilidade da proteína.

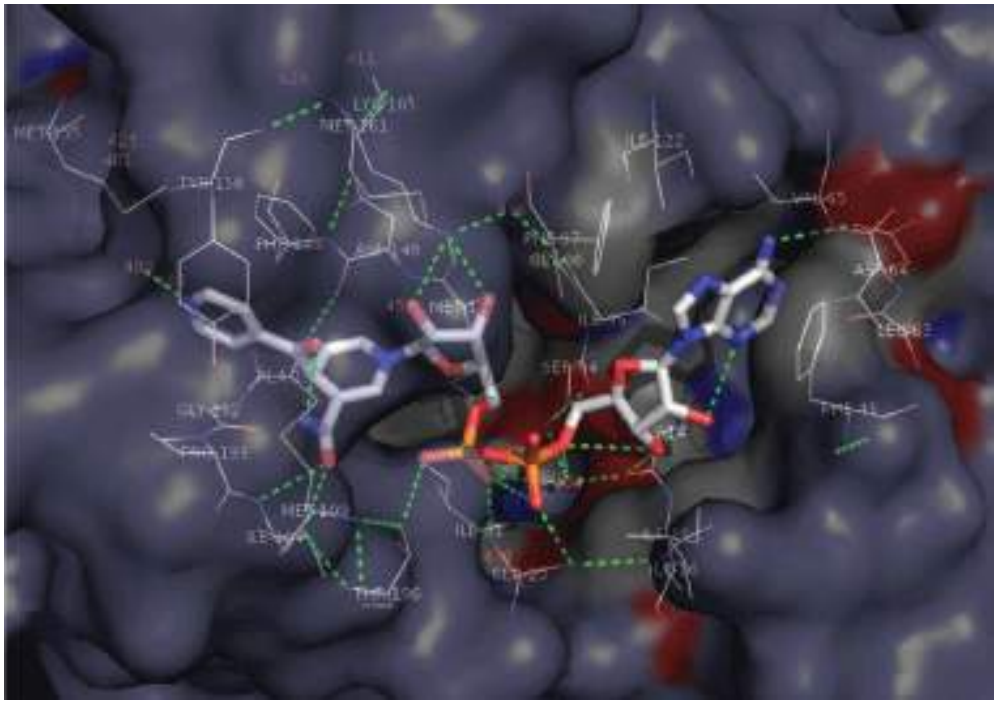


Figura 2.1 – Visualização da melhor orientação encontrada pelo programa de docagem molecular entre uma conformação da enzima InhA e o aduto INH-NAD. A superfície molecular da enzima está representada em roxo e os resíduos que interagem com o aduto estão representados em palitos. O aduto está representado em palitos e os átomos são coloridos pelo tipo de átomo (carbono em branco, nitrogênio em azul, oxigênio em vermelho e fósforo em laranja). As linhas tracejadas em verde demonstram a ocorrência de interação entre os resíduos da InhA com os átomos do aduto. Adaptada de [APZF08].

2.3 Consideração da flexibilidade da proteína utilizando múltiplas conformações

Diversos autores têm comprovado que não considerar a flexibilidade da proteína no processo de seleção de candidatos a fármacos acaba limitando significativamente a precisão dos métodos de docagem molecular [BR02, ABG06, TA08]. Assim, a escolha dos métodos utilizados para considerar a flexibilidade tornou-se um fator essencial para aumentar a probabilidade de sucesso, tendo sido amplamente estudada nos últimos anos [TA08, MRB+15, BRM15].

Atualmente é possível encontrar uma grande quantidade de métodos capazes de simular parte da flexibilidade da proteína [TA08, MRB+15] (importantes revisões sobre métodos que consideram da flexibilidade da proteína podem ser encontrados em [CSS09, YAR11]). Dentre os métodos existentes, a simulação por Dinâmica Molecular (DM) é o método capaz de gerar múltiplas conformações considerando todos os graus de liberdade existentes em uma proteína, tornando possível a representação de um modelo de Receptor Totalmente Flexível (FFR) [MWRNdS11]. A Figura 2.2 apresenta um conjunto de conformações capturadas de uma simulação por DM da enzima InhA em diferentes momentos de tempo.

Segundo Alonso et al. [ABG06], modelos FFR com elevadas quantidades de conformações possuem probabilidades maiores de simular movimentos internos e trocas conformacionais distintas do receptor que métodos com poucas conformações. Essas mudanças conformacionais aumentam a possibilidade de a cavidade de ligação do substrato ser formada por diferentes resíduos ao longo do modelo FFR, possibilitando à cavidade de ligação do substrato a mudança da sua atividade biológica [MSWN02]. Assim, modelos FFR de uma proteína podem simular diferentes funções biológicas.

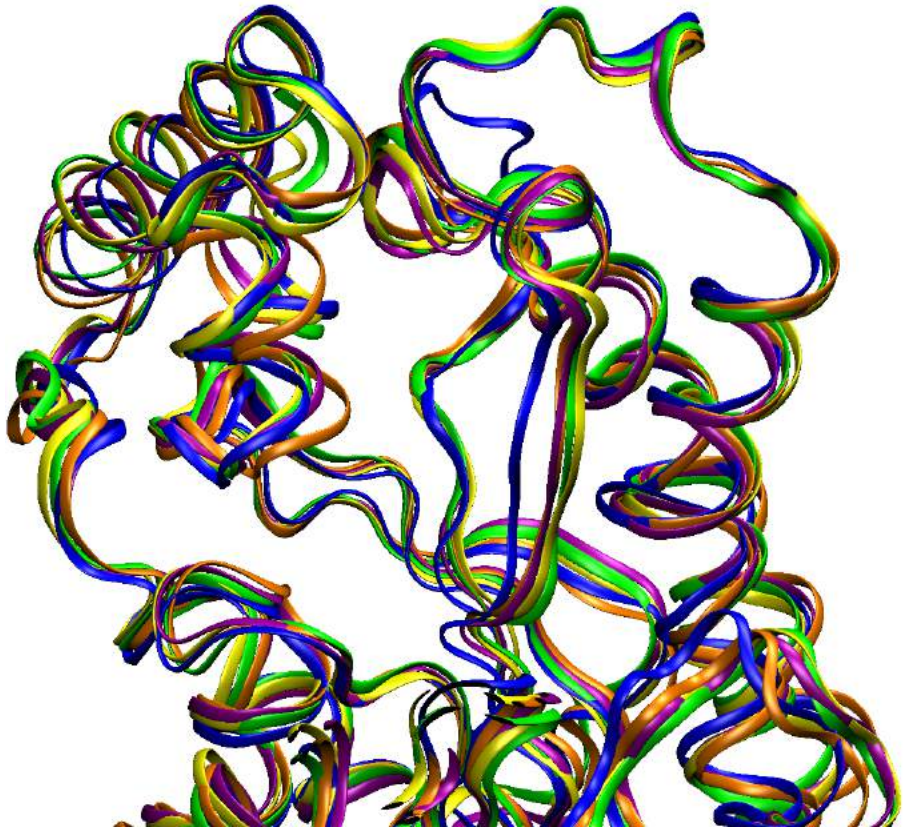


Figura 2.2 – Conjunto de conformações da enzima InhA de *Mtb* capturadas em diferentes momentos de um modelo FFR para demonstrar a flexibilidade da enzima. As proteínas estão representadas em fitas e as cores representam o momento da captura de cada conformação ao longo da DM (azul em 1 ns, laranja em 5 ns, magenta em 10 ns, verde em 15 ns e amarelo em 20 ns. Adaptada de [MWRNdS11].

Embora essa abordagem possibilite maiores oportunidades para a descoberta de potenciais candidatos a fármaco, ela também implica em uma considerável elevação na complexidade computacional para realizar as simulações de docagem molecular entre os ligantes flexíveis e o modelo FFR. A seção 2.4 descreve detalhes sobre os algoritmos de agrupamento utilizados no processo de redução de dimensionalidade de modelos FFR.

2.4 Agrupamento de conformações de modelos FFR

As técnicas de agrupamento são métodos não supervisionados que buscam identificar padrões em conjunto de dados, formando partições com o conjunto de objetos que são mais homogêneos entre si [JD88, TSK06]. A partir desse conjunto de dados, a análise por agrupamento se torna a mais adequada técnica de inteligência computacional para particionar conformações de modelos FFR [BD92]. Assim, todas as conformações de um modelo FFR são atribuídas a uma partição usando uma medida de similaridade ou dissimilaridade.

O agrupamento das conformações contribui especialmente quando existe a necessidade da execução de experimentos de docagem molecular exaustivos, uma vez que, após o agrupamento, as estruturas estarão separadas em conjuntos de receptores similares. As conformações dentro da mesma partição são, segundo os critérios de um algoritmo de agrupamento sobre os dados de entrada, semelhantes umas às outras e dessemelhantes das conformações de outras partições [HW79]. Desta forma, se uma conformação do receptor pertence a uma partição que interage favoravelmente com um ligante específico, pode-se assumir que outras conformações dentro da mesma partição também possuam boas interações com esse ligante. Por outro lado, as conformações que pertencem a conjuntos considerados como pouco promissores são, consequentemente, descartadas a fim de se reduzir o número de experimentos de docagem molecular.

Diversos trabalhos que aplicam algoritmos de agrupamento têm sido desenvolvidos com o objetivo de reduzir a complexidade no tratamento de receptores flexíveis. Os mesmos abordam diferentes métodos, incluindo: (1) a seleção de estruturas representativas com base somente no RMSD; (2) a seleção de estruturas com base em um conjunto de propriedades da molécula; (3) variações na fórmula do RMSD; (4) novos métodos de alinhamento estrutural; e (5) utilização dos farmacóforos mais frequentes do modelo FFR.

O valor do RMSD obtido pela comparação dos átomos pareados de duas proteínas é a mais tradicional e popular medida de similaridade utilizada para particionar modelos FFR. Por exemplo, Lyman *et al.* [LZ06] gerou um conjunto de estruturas representativas ao aplicar um raio de corte do RMSD para identificar as partições de modelos flexíveis de met-encefalina, um neurotransmissor pentapeptídico. Shao e colaboradores [STTC07] compararam os conjuntos de partições de um modelo flexível, gerados por 11 algoritmos de agrupamento e avaliaram esses resultados com duas medidas de validação para determinar a melhor partição.

Esses trabalhos apresentaram importantes contribuições nesta área. No entanto, o agrupamento de conformações de modelos FFR ainda é um problema não resolvido. Em [RFS10], Rajan e colaboradores elencaram um conjunto de restrições frequentemente encontradas nos agrupamentos de modelo FFR:

- Instabilidade: as partições geradas são muitas vezes instáveis a mudanças nos parâmetros de corte e a ruídos.
- Relação intergrupos: não existe qualquer informação sobre as relações entre as partições formadas.
- Inspeção visual: o conjunto de partições geradas são geralmente validados por inspeção visual das estruturas atribuídas como representativas. No entanto, modelos flexíveis possuem geralmente centenas de milhares de conformações, tornando muito complexa a avaliação individual de cada conformação pelo especialista de domínio.
- Incerteza: existe uma grande dificuldade na análise dos movimentos de regiões flexíveis, sendo complexa a análise envolvendo um conjunto de coordenadas coletivas.
- Validação: um dos grandes desafios dos algoritmos não supervisionados é a avaliação dos resultados. Diversas medidas de validação foram desenvolvidas; no entanto é comum cada medida indicar diferentes números de partições como o mais adequado para um mesmo conjunto de dados.

Nesta tese, o Capítulo 5 apresenta três métodos utilizados no processo de redução de dimensionalidade de modelos FFR. Como foi descrito anteriormente, conformações pertencentes a uma partição devem apresentar valores de interação similares do complexo receptor-ligante. Para avaliar os agrupamentos formados são avaliados diversos experimentos docagem molecular de complexos receptor-ligante. A seção 2.5 apresenta um estudo sobre os principais BD de pequenas moléculas públicos disponibilizados na literatura [GRFS15].

2.5 Bancos de dados de pequenas moléculas

Os Banco de Dados (BD) de ligantes são repositórios capazes de armazenar as informações de pequenas moléculas. O surgimento desses BD ocorreu devido à necessidade de os pesquisadores terem acesso a dados biológicos de forma ágil. Desta forma, classifica-se como um BD ideal aquele que disponibiliza o maior número possível de informações, provê um acesso fácil às suas informações, fornece respostas rápidas às requisições e disponibiliza os dados em formatos que são acessíveis por um grande número de sistemas de computação [IS05].

Atualmente, existe uma grande quantidade desses BD disponíveis na comunidade científica e esse número está crescendo consideravelmente nos últimos anos, conforme pode ser observado nos volumes anuais divulgados pelo *Nucleic Acids Research* [GRFS15]. Embora exista uma grande quantidade desses BD, todos possuem determinadas particularidades, tornando necessária uma análise criteriosa a fim de compreender quais são as propriedades armazenadas e a forma como as informações de cada ligante são geradas.

Um dos estudos iniciais desta tese [WQM+12] abordou as principais particularidades envolvendo um conjunto com os principais BD de pequenas moléculas públicos, como o ChemBank [SGH+08], ChemDB [CLS+07], ZINC [ISM+12], NCI Database [IVB+02] e o PubChem [ABIC04]. Esse estudo elencou os aspectos positivos e negativos de cada BD relacionados com as necessidades desta pesquisa. Esse trabalho também identificou discrepâncias nos parâmetros calculados devido às diferentes ferramentas utilizadas por cada BD. Assim, uma determinada busca por propriedades pode resultar na obtenção de diferentes conjuntos de ligantes. Um resumo dessa avaliação está descrito na Tabela 2.1. O BD PubChem disponibiliza a maior quantidade de ligantes e também é o único a informar o volume de cada molécula. Contudo, esse BD não fornece as informações das cargas atômicas parciais dos ligantes e, assim, seus dados não estão prontos para execução em programas de docagem molecular. O BD ZINC, por sua vez, é o segundo maior em quantidade de ligantes disponíveis e as moléculas armazenadas nesse BD estão prontas para execução em programas de docagem molecular [ISM+12]. As características disponibilizadas pelo BD ZINC são preponderantes para esta tese.

Tabela 2.1 – Descrição das principais características existentes em alguns dos Banco de Dados de ligantes pesquisados [WQM+12].

	ChemBank	ChemDB	NCI	PubChem	ZINC
Baixar:					
- subconjuntos	x	x	x	x	x
- todos os dados	x	x	x	x	x
Pesquisar por:					
- nome	x	x	x	x	x
- código SMILE	x	x	x	x	x
- estrutura exata	x		x		x
- subestrutura	x	x	x	x	x
- descritores moleculares	x	x	x	x	x
- similaridade	x	x	x	x	x
Dados:					
- prontos para docagem					x
- volume				x	
- quantidade	4,5 M	5 M	0,26M	27 M	20 M
- formatos	sdf	mol, mol2, sdf, PDB	mol2, sdf	ASN.1, XML sdf	mol2, sdf, flexibase

M = 10^6

2.6 Triagem virtual

Na literatura, existe uma grande variedade de abordagens buscando selecionar um conjunto de ligantes que apresentem interações favoráveis do complexo receptor-ligante. Essas abordagens podem ser classificadas basicamente em duas vertentes, o Planejamento de Fármacos Baseado na Estrutura do Ligante (LBDD - do inglês *Ligand-Based Drug Design*) e o Planejamento de Fármacos Baseado na Estrutura do Receptor (SBDD - do inglês *Structure-Based Drug Design*).

O LBDD captura as características de ligantes que são conhecidos por apresentar interações favoráveis com o receptor alvo. Esse conjunto de características é usado para identificar ligantes similares em BD de pequenas moléculas. No entanto, essa abordagem costuma ser utilizada somente quando não se tem o conhecimento da estrutura do receptor, em virtude das suas diversas limitações [MKT02].

O SBDD utiliza as informações da estrutura do receptor para avaliar e selecionar os compostos que melhor interagem com a cavidade alvo [Lyn02]. Existem muitas abordagens baseadas no SBDD visando reduzir o número de compostos a serem testados com experimentos de docagem molecular. No entanto, a alta capacidade computacional demandada para avaliar a flexibilidade no SBDD tem sido um fator limitante comum a todas as abordagens [Yan10].

Ambas estratégias possibilitam a extração de um conjunto de características estruturais e físico-químicas do ligante ou do receptor. Essas características propiciam a elaboração de um conjunto de regras para averiguar a afinidade da formação de um complexo receptor-ligante, procedimento conhecido na literatura como a identificação de farmacóforos. No entanto, o SBDD fornece uma quantidade maior de propriedades da cavidade de ligação que o LBDD. Assim, modelos farmacofóricos gerados a partir das características do SBDD tendem a ser mais precisos que modelos gerados com base no LBDD, devido a complementariedade da cavidade alvo. Desta forma, modelos farmacofóricos devem ser, preferencialmente, baseados na estrutura do receptor. Maiores detalhes sobre essa abordagem são descritos na próxima subseção.

2.6.1 Triagem virtual baseada em farmacóforos 3D

Quando se tem o conhecimento da estrutura 3D da proteína ou do ligante, é possível definir as características estruturais mínimas que os ligantes devem possuir a fim de se ligar com a proteína alvo [Gun00], sendo esta uma forma eficiente para a identificação de bons candidatos a solução. Esse conjunto de características é conhecido como modelo farmacofórico. Outra definição, muito frequente na literatura, define um modelo farmacofórico como um conjunto de características estéricas e eletrostáticas necessárias para garantir interações supramoleculares ótimas com um alvo biológico específico e para acionar ou impe-

dir a sua resposta biológica [WGLM98]. Modelos farmacofóricos que apresentem uma alta quantidade de características físico-químicas são, normalmente, decompostos em combinações de hipóteses farmacofóricas. Assim, a triagem virtual em BD de pequenas moléculas é feita com o objetivo de identificar moléculas que possuam, em sua estrutura, um arranjo 3D das propriedades físico-químicas de acordo com a hipótese farmacofórica [HNW$^+$97].

Segundo Seidel et al. [SIBW11], as técnicas baseadas em hipóteses farmacofóricas considerando estruturas 3D se tornaram uma das formas mais rápidas e eficientes de se selecionar compostos. A Figura 2.3 apresenta um fluxograma descrevendo a triagem virtual com farmacóforos para buscar novos compostos candidatos. Conforme pode ser visto na Figura 2.3-a, um modelo farmacofórico pode ser obtido a partir das informações da proteína (SBDD) e/ou do ligante (LBDD). Assim como foi descrito no início desta seção, modelos farmacofóricos baseados no receptor tendem a ser mais precisos que modelos gerados com base nas estruturas do ligante.

A Figura 2.3-b mostra a etapa de preparação do conjunto de ligantes a serem pesquisados com a hipótese farmacofórica. Primeiramente, seleciona-se um conjunto de ligantes (um ou mais BD de pequenas moléculas). Esses BD normalmente armazenam apenas as estruturas canônicas, ou seja, não disponibilizam as variações conformacionais dos ligantes. Desta forma, é necessário gerar um conjuntos de possíveis confôrmeros para cada ligante para simular parte da sua flexibilidade. Não existe um número exato de confôrmeros a serem gerados, devendo esse valor ser um equilíbrio entre a flexibilidade do ligante e o tempo necessário para o método avaliar cada confôrmero. A Figura 2.3-c descreve o resultado da avaliação de cada ligante (e seus confôrmeros) com as hipóteses do modelo farmacofórico. Essas avaliações, na maioria das vezes, são ponderadas pela quantidade de acertos dos seus famacóforos. O valor obtido por essa métrica serve para ordenar o conjunto de ligantes aprovados, formando uma lista ordenada (Figura 2.3-d). A próxima subseção descreve alguns trabalhos que geraram modelos farmacofóricos 3D a partir de um conjunto de conformações.

2.6.2 Triagem virtual baseada em farmacóforos 3D formados a partir de conjuntos de conformações

Modelos farmacofóricos 3D baseados na estrutura do receptor têm sido amplamente utilizados para realizar a redução da dimensionalidade, selecionando ligantes que contenham o arranjo espacial de propriedades físico-químicas essenciais. Contudo, limitar a flexibilidade da proteína no processo de seleção de candidatos a fármacos pode reduzir significativamente a precisão dos métodos de docagem molecular e, por consequência, restringir o conjunto de candidatos à solução [BR02, TA08].

Algumas abordagens têm adotado métodos simplistas de aumentar o raio do ponto farmacofórico de interação para considerar a flexibilidade das proteínas sem realizar um

estudo específico da região. Aumentar o raio da propriedade sem uma análise detalhada da flexibilidade da proteína pode incorrer em arranjos conformacionais não factíveis para essa proteína sendo, assim, uma forma inadequada de considerar a sua flexibilidade.

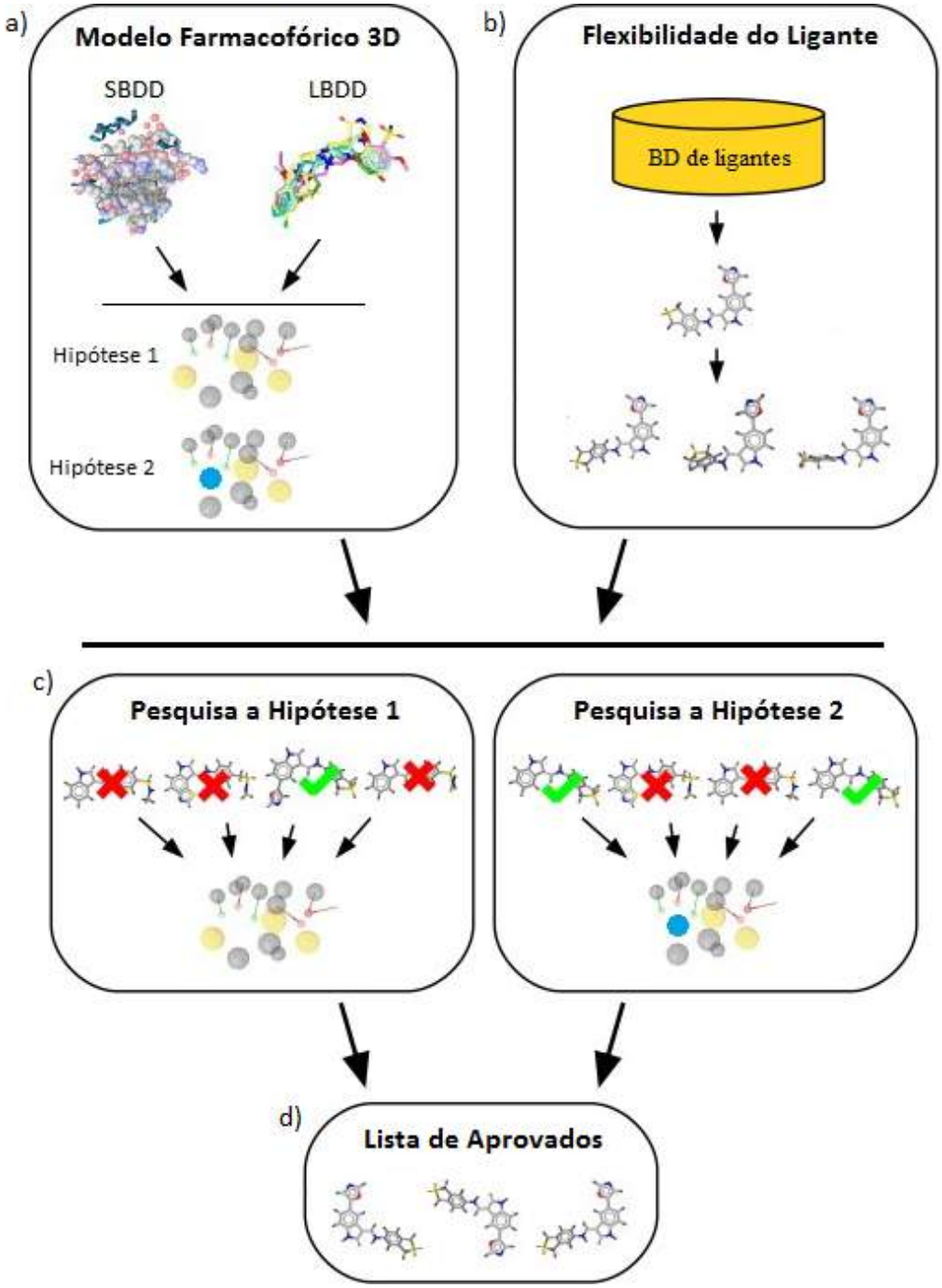


Figura 2.3 – Fluxograma da triagem virtual baseada em hipóteses farmacofóricas para selecionar novos compostos candidatos a fármaco. (a) Identificação das hipóteses farmacofóricas a partir das informações de um conjunto de ligantes e/ou da cavidade de ligação do receptor. (b) Simulação da flexibilidade de cada estrutura disponibilizada em BD de pequenas moléculas. (c) Avaliação das pequenas moléculas para selecionar as estruturas 3D de acordo com a hipótese farmacofórica. O símbolo v em verde significa que o ligante satisfaz a hipótese farmacofórica, devendo este ser selecionado. (d) Lista dos ligantes selecionados ordenados pela afinidade obtida com a hipótese farmacofórica. Adaptada de [Yan10].

Trabalhos como o Carlson [CM00] e Dror [DSPNW04] iniciaram a investigação de sistemas flexíveis no desenvolvimento de modelos farmacofóricos 3D. Dror et al. [DSPNW04] propôs quatro etapas fundamentais para o desenvolvimento de modelos farmacofóricos baseados na estrutura de receptores flexíveis. Essas etapas são:

- Etapa 1: Avaliar a qualidade da estrutura da proteína alvo e das estruturas do modelo flexível a ser utilizado na geração dos modelos farmacofóricos 3D.
- Etapa 2: Selecionar um conjunto de estruturas representativas do modelo FFR da proteína alvo.
- Etapa 3: Identificar as propriedades farmacofóricas complementares da cavidade de ligação do substrato.
- Etapa 4: Analisar o conjunto de ligantes selecionados pelo modelo farmacofórico.

Atualmente, métodos alternativos têm sido desenvolvidos visando considerar os movimentos flexíveis da cavidade de ligação do substrato na geração modelos farmacofóricos 3D [DLS+05, DSNB06, HL13]. Deng e colaboradores [DLS+05, DSNB06] desenvolveram dois relevantes trabalhos com a utilização de estruturas representativas de modelos FFR. Em [DLS+05], Deng baseou a geração das hipóteses farmacofóricas na utilização de 10 estruturas selecionadas pela comparação dos valores do RMSD. Uma limitação importante desse trabalho é justamente a seleção desse conjunto de estruturas representativas ser baseado somente nos valores do RMSD, visto essa selação torna possível a formação de partições contendo conjuntos de estruturas com cavidades de ligação do substrato distintas entre si. O principal problema dessa abordagem está na seleção das estruturas representativas a partir de conjuntos particionados pelas informações do RMSD.

Desta forma, as limitações encontradas nesses modelos farmacofóricos 3D comprovam que ainda hoje não existe um método baseado somente nas propriedades físico-químicas da cavidade de ligação do substrato considerando um modelo FFR, de forma a efetivamente reduzir esse número elevado de ligantes a um valor gerenciável.

2.7 Considerações finais

Este capítulo apresentou conceitos fundamentais sobre o RDD, descrevendo suas etapas e as principais abordagens utilizadas nesta tese. Os diferentes conceitos apresentados sobre o processo de RDD empregados para considerar a flexibilidade das proteínas, juntamente com a avaliação dos BD de ligantes contendo elevadas quantidades de moléculas, corroboram a necessidade de novas soluções que atendam às duas primeiras etapas definidas por Kuntz [Kun92]. Especificamente, o escopo deste trabalho está inserido na etapa 2, cujo objetivo é identificar conjuntos promissores de ligantes candidatos a fármaco.

Este capítulo também descreveu a necessidade de considerar a flexibilidade da proteína no processo de seleção de ligantes candidatos a fármacos para não limitar a precisão dos métodos de docagem molecular. Assim, um modelo FFR da enzima InhA de *Mtb* foi adotado como o estudo de caso neste trabalho. No entanto, a utilização da simulação da flexibilidade da proteína alvo por DM contendo milhares de conformações implica em um aumento no esforço computacional necessário. Desta forma, modelos farmacofóricos 3D avaliando conjuntos de estruturas cristalinas têm sido amplamente utilizados no processo de triagem virtual de ligantes para reduzir a dimensionalidade dessa tarefa. Esses modelos selecionam apenas os ligantes que possuam o arranjo espacial de propriedades físico-químicas essenciais obtidas da avaliação das interações similares identificadas de complexos receptor-ligante conhecidos [LGLT10]. No entanto, os modelos desenvolvidos ao não considerarem a flexibilidade das proteínas na geração dos modelos farmacofóricos 3D acabam limitando o potencial de filtrar estruturas mais acuradas. O próximo capítulo apresenta uma avaliação da estrutura da proteína alvo e das estruturas do modelo flexível a serem utilizados na geração dos modelos farmacofóricos 3D.

3. Proteína investigada: Enzima InhA de *Mycobacterium tuberculosis*

Este Capítulo descreve detalhes da proteína alvo investigada como estudo de caso nesta tese: a enzima InhA de *Mycobacterium tuberculosis* (*Mtb*). Essa proteína representa um alvo interessante para o desenvolvimento de novos fármacos anti-tuberculose. A primeira seção deste Capítulo apresenta a motivação social para a definição da proteína alvo a ser analisada. Após, uma descrição da cavidade de ligação do substrato é feita, caracterizando regiões relevantes dessa cavidade. A terceira seção mostra detalhes do conjunto de estruturas cristalinas da InhA de *Mtb* catalogadas no sítio do *Protein Data Bank* (PDB) [BWF+00]. Por fim, um modelo de Receptor Totalmente Flexível (FFR) de 20 ns é avaliado.

3.1 Motivação social

A tuberculose é considerada uma doença negligenciada por empresas farmacêuticas que não investem no desenvolvimento de novos medicamentos devido à falta de retorno financeiro. Segundo a Organização Mundial de Saúde, foram estimados entre 8,6 e 9,4 milhões de novos casos de tuberculose em 2013 no mundo [Tub14]. Essa organização também estima que um terço da população está infectada com o bacilo; e aproximadamente 10% dos que possuem o bacilo devem desenvolver a doença ao longo da sua vida.

Embora o número de mortes em decorrência da tuberculose esteja declinando gradativamente em valores mundiais, países em desenvolvimento ainda permanecem com taxas preocupantes. Uma das principais causas apontadas para a ocorrência desse índice é o crescimento do número de casos de pacientes com tuberculose que acabam adquirindo resistência a isoniazida e a rifampicina, que são os principais fármacos utilizados no combate dessa doença [SBSNdS05, JTJ+10]. Esses casos que apresentam resistência aos medicamentos utilizados no tratamento da tuberculose são categorizados de três formas [fDCC06, VMF+09]:

- Tuberculose Multirresistente (MDR-TB - do inglês *MultiDrug-Resistant TuBerculosis*): ocorre quando o bacilo da tuberculose é resistente aos dois fármacos mais potentes no tratamento contra a tuberculose (isoniazida e rifampicina).

- Tuberculose Extensivamente Resistente (XDR-TB - do inglês *eXtensively Drug-Resistant TuBerculosis*): ocorre quando, além de ser MDR-TB, o bacilo da tuberculose também é resistente a qualquer fluoroquinolonas e, pelo menos, a um dos três fármacos de segunda linha (capreomicina, canamicina, e amicacina).

- Tuberculose Totalmente Resistente (TDR-TB - do inglês *Totally Drug-Resistant TuBerculosis*): ocorre quando o bacilo da tuberculose é resistente aos fármacos de primeira e segunda linha do tratamento contra a tuberculose.

O surgimento de casos resistentes aos fármacos existentes evidencia a necessidade de novas pesquisas para identificar possíveis novos candidatos a fármacos para a tuberculose. Neste contexto, a enzima InhA ou 2-trans-enoil-ACP(COA) redutase tem sido amplamente utilizada como uma estrutura alvo, devido ao seu importante papel no metabolismo do *Mtb*. A próxima Seção apresenta detalhes dessa enzima, bem como da sua cavidade de ligação do substrato.

3.2 A enzima InhA de *Mycobacterium tuberculosis*

A InhA é a enzima responsável por estender as cadeias dos ácidos graxos, contribuindo na biossíntese dos ácidos micólicos. Esse processo contribui na formação da parede celular do *Mtb*, que protege o bacilo de danos químicos, impedindo a atividade de diversos fármacos. Assim, estratégias capazes de inibir a ação da enzima InhA causariam a interrupção do processo de formação da parede celular, ocasionando posteriormente a morte do bacilo. Devido à função farmacológica desempenhada por essa enzima, esta tese define a InhA de *Mtb* como o modelo de receptor a ser utilizado.

Essa enzima é composta por 268 resíduos de aminoácidos, enovelando uma estrutura formada por 7 fitas e 8 hélices. A primeira estrutura 3D da InhA, cristalizada junto com a coenzima NADH, foi disponibilizada no sítio do PDB por Dessen e colaboradores [DQB+95]. Essa estrutura foi cristalizada contendo 4 subunidades, sendo que todas apresentam a cavidade de ligação do substrato independente. Assim, optou-se por utilizar apenas uma subunidade neste trabalho. A próxima subseção descreve a cavidade de ligação do substrato dessa enzima.

3.2.1 Cavidade de ligação do substrato

A cavidade de ligação do substrato da enzima InhA localiza-se entre duas hélices transversais sustentadas por alças da proteína e, mais precisamente, acima do anel da nicotinamida da coenzima NADH [DQB+95]. A Figura 3.1 mostra, em azul, o volume ocupado pela coenzima NADH no sítio ativo da enzima InhA (PDB ID: 1BVR). A região em verde representa o volume do ligante THT (análogo do substrato). Essa região, ocupada pelo análogo do substrato, é denominada cavidade de ligação do substrato. Em destaque na Figura 3.1, há a representação da superfície molecular e em palitos da coenzima NADH com a estrutura do análogo do substrato.

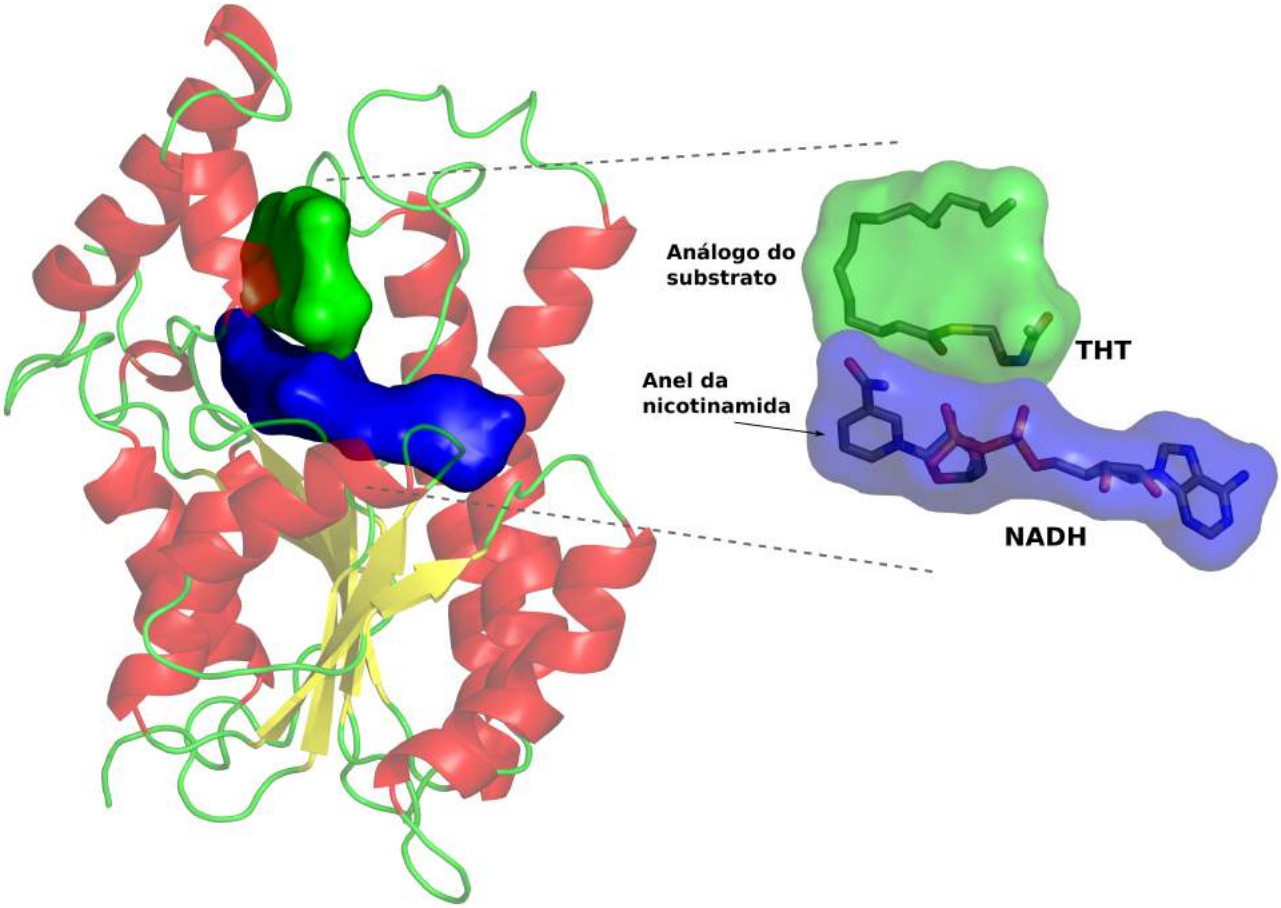


Figura 3.1 – Cofator NADH e um ligante análogo à cavidade do substrato extraídos da estrutura cristalográfica da 1BVR. Em verde está a região do ligante análogo à cavidade do substrato e, em azul, a região ocupada pela coenzima NADH. Nota-se que a região do substrato está situada logo acima do anel da nicotinamida do NADH.

Segundo [QDS+96], a enzima InhA depende da presença da coenzima NADH em seu sítio ativo para poder interagir com o ligante INH e, assim, inibir a ação da bactéria. Nos casos em que a coenzima não estava presente no sítio ativo da InhA, o ligante não conseguiu inibir a atividade da bactéria. Esse estudo ressalta a importância de não selecionar novos candidatos a fármaco que possam competir com a coenzima NADH pela sua região.

Alguns dos estudos anteriores realizados em nosso laboratório não consideraram a coenzima NADH como parte da estrutura da proteína [De 12, MSRNdS07, MSR+08, WMNdSR09]. O objetivo desses estudos era o de entender como as pequenas moléculas se encaixariam no sítio ativo da InhA quando a coenzima não estivesse presente. Grande parte dos resultados encontrados mostraram ligantes posicionados em regiões necessárias para a coenzima, fato que inviabilizaria o seu posterior encaixe.

A fim de evitar encontrar candidatos que possam concorrer com a posição da coenzima, os estudos elaborados nesta tese sempre consideram a estrutura da coenzima dentro do sítio ativo da enzima InhA. A composição da estrutura do receptor depende apenas do tipo de composto a ser avaliado:

- Ligantes: a coenzima NADH é acoplada na estrutura da enzima InhA.
- Adutos: são formados por ligantes que estão conectados com a coenzima NADH, assim a coenzima NADH não deve integrar a estrutura da enzima InhA.

Após caracterizar a estrutura da InhA, um estudo das estruturas cristalinas da proteína InhA de *Mtb* do sítio *Protein Data Bank* é apresentado na próxima Seção. Esse estudo serve como base para a compreensão das avaliações a serem realizadas considerando o modelo FFR.

3.3 Estruturas cristalinas da enzima InhA armazenadas no *Protein Data Bank*.

O *Protein Data Bank* (PDB) é um BD de estruturas experimentais de macromoléculas biológicas criado com a função de armazenar, organizar e distribuir estruturas 3D de proteínas. As estruturas disponibilizadas nesse sítio possuem um alto índice de confiabilidade, uma vez que essas estruturas são determinadas experimentalmente através da cristalografia por difração de raio X, ou ressonância magnética nuclear ou microscopia eletrônica [BWF$^+$00]. Atualmente existe um conjunto de 52 estruturas da InhA de *Mycobacterium tuberculosis* armazenadas neste BD. No entanto, este estudo descarta as estruturas mutantes ou que não contenham a coenzima NADH, resultando em um conjunto de 34 estruturas.

A Tabela 3.1 descreve as 34 estruturas selecionadas e algumas características de cada proteína. Observando essa tabela, nota-se que algumas proteínas foram cristalizadas com ligantes, adutos ou sem nenhum composto na cavidade do substrato. Esses diferentes tipos de compostos docados podem destacar importantes variações estruturais dos resíduos que compõem a cavidade do substrato da enzima. Então, para estudar tais propriedades, o conjunto de proteínas necessita do alinhamento estrutural de todas as proteínas evidenciadas na Tabela 3.1. Esse processo de alinhamento começa com a seleção de uma cadeia da proteína que contenha o maior número de resíduos (nem todas as cadeias possuem 268 resíduos) com as coordenadas 3D de cada átomo dispostas em arquivos no formato PDB de cada estrutura. As próximas subseções descrevem os métodos adotados para calcular as propriedades do RMSD e do volume descritos na Tabela 3.1.

3.3.1 Alinhamento das estruturas cristalinas da proteína InhA.

Na literatura existem diversos programas para alinhar proteínas com base em sua estrutura. Contudo, uma considerável quantidade não permite extrair todas as coordenadas 3D da molécula resultante do alinhamento. Dentre os programas existentes, o VMD [HDS96] e o PyMol [DeL02] permitem a extração das moléculas alinhadas. Porém o

conjunto de proteínas deve ser equiparável, ou seja, o conjunto de átomos que serve como base para o alinhamento deve existir em todas as estruturas. Assim, as estruturas da Tabela 3.1 foram submetidas para a avaliação no programa VMD para realizar o alinhamento estrutural dessas moléculas e foram constatados alguns problemas como a duplicidade de resíduos e/ou a falta de alguns átomos/resíduos.

Tabela 3.1 – Estruturas cristalinas da InhA de *Mtb* disponibilizadas no sítio do PDB.

PDB ID	**Composto**	**RMSD(Å)**	**Volume da cavidade (Å³)**	**Referência**
1ENY	NADH	0,0	598,2	Dessen et al. [DQB+95]
3OEW	NADH	0,8	1.516,3	Molle et al. [MGV+10]
1BVR	NADH + THT	2,4	734,9	Rozwarski et al. [RVS+99]
1P44	NADH + GEQ	1,2	1.691,9	Kuo et al. [KMA+03]
1P45	NADH + TCL	1,1	970,4	Kuo et al. [KMA+03]
2B35	NADH + TCL	1,0	1.080,3	Sullivan et al. [STB+06]
2B36	NADH + 5PP	1,2	946,6	Sullivan et al. [STB+06]
2B37	NADH + 8PS	1,0	445,1	Sullivan et al. [STB+06]
2H7I	NADH + 566	0,9	1.451,0	He et al. [HASOdM06]
2H7L	NADH + 665	0,9	1.674,4	He et al. [HASOdM06]
2H7M	NADH + 641	0,9	1.482,2	He et al. [HASOdM06]
2H7N	NADH + 744	0,9	781,1	He et al. [HASOdM06]
2H7P	NADH + 468	0,9	850,9	He et al. [HASOdM06]
3FNE	NADH + 8PC	1,3	1.503,0	Freundlich et al. [FWV+09]
3FNF	NADH + JPM	1,4	756,6	Freundlich et al. [FWV+09]
3FNG	NADH + JPL	1,2	1.144,0	Freundlich et al. [FWV+09]
3FNH	NADH + JPJ	1,3	693,8	Freundlich et al. [FWV+09]
2NSD	NADH + 4PI	1,2	1.468,6	He et al. [HAOdM07]
2X22	NADH + TCU	1,7	1.225,1	Luckner et al. [LLaE+10]
2X23	NADH + TCU	1,5	601,6	Luckner et al. [LLaE+10]
2H9I	Aduto ETH-NAD	1,0	1.681,1	Wang et al. [WLG+07]
1ZID	Aduto INH-NAD	1,0	1.714,2	Rozwarski et al. [RGB+98]
2IDZ	Aduto INH-NAD	1,0	928,5	Dias et al. [DVP+07]
2PR2	Aduto INH-NAD	0,7	836,9	Argyrou et al. [AVB07]
2NTJ	Aduto PTH-NAD	0,8	1.752,9	Wang et al. [WLG+07]
4BQR	NADH + IBH	1,0	484,4	Shirude et al. [SMN+13]
4BQP	VMY/Na	1,0	664,2	Shirude et al. [SMN+13]
4OYR	1US	1.5	782,5	Li et al. [LLP+14]
4OXY	1TN	1,6	949,8	Li et al. [LLP+14]
4OXN	1S5/2NV/Cl/EPE	1,3	2.032,8	Li et al. [LLP+14]
4OXK	1S5/2NV	1,4	794,9	Li et al. [LLP+14]
4OHU	2TK	1,5	659,7	Li et al. [LLP+14]
4COD	KV1	1,1	696,7	Encinas et al. [EON+14]
4OIM	ACT/JUS	1,2	701,3	Pan et al. [PKB+14]

*O cálculo do valor do RMSD comparou todos os átomos existentes na enzima 1ENY com seus correspondentes em outras enzimas. **O volume da cavidade do substrato foi calculado considerando os átomos da coenzima NADH fazendo parte da enzima.

A duplicidade dos resíduos foi resolvida mantendo-se apenas uma das predições da proteína. Os principais problemas encontrados em cada proteína sobre a ausência de átomos ou resíduos, ocorridas em 6 moléculas, estão descritos abaixo:

- Ausência de uma sequência de resíduos:
  - 2B35: lacuna entre os resíduos LEU197 e a GLI212;
  - 2B37: lacuna entre os resíduos LEU197 e a LEU217;
  - 4OXN: lacuna entre os resíduos ALA201 e o GLU209.

- Alguns resíduos não apresentam o átomo Cα:
  - 2X22: não possui Cα nos resíduos: ARG43, LEU61, MET130 e SER152;
  - 2X23: não possui Cα nos resíduos: MET199 e SER152;
  - 3OEW: não possui Cα nos resíduos: ARG153 e GLU219.

Uma consulta a um especialista de domínio foi feita para obter informações sobre os problemas dessas estruturas armazenadas no sítio do PDB. Esse especialista constatou que a ausência de alguns resíduos bem como a duplicidade de resíduos pode ocorrer quando trata-se de uma região muito flexível. Assim, há uma grande complexidade em determinar exatamente a posição correta desses átomos. Por exemplo, como ocorreu nas estruturas 2B35, 2B37 e 4OXN, onde o conjunto de resíduos ausentes pertencem exatamente à alça de ligação da InhA.

A solução encontrada para alinhar essas estruturas foi criar diferentes composições do arquivo da 1ENY, proteína que é a estrutura base definida para o alinhamento das demais moléculas. Após as modificações citadas, tornou-se possível obter todas as estruturas do sítio do PDB em um mesmo arranjo espacial. A Figura 3.2 mostra o alinhamento estrutural das estruturas cristalográficas em diferentes cores para realçar as diferentes conformações que cada molécula pode assumir. O cálculo do valor do RMSD foi realizado comparando todos os átomos existentes na enzima 1ENY com seus correspondentes nas outras enzimas presentes na Tabela 3.1.

3.3.2 Volume da cavidade de ligação do substrato das estruturas cristalinas da InhA.

A avaliação do volume da cavidade de ligação do substrato de cada estrutura cristalina catalogada na Tabela 3.1 foi realizada pelo programa CASTp. As estruturas complexadas com adutos (2H9I, 1ZID, 2IDZ, 2PR2 e 2NTJ) foram editadas de maneira que todas as estruturas presentes na Tabela 3.1 fossem compostas apenas pelo complexo InhA-NADH. Os valores obtidos pelo programa foram calculados utilizando uma sonda química de 1,4 Å, valor padrão definido por esse programa por ser o raio de uma molécula de água.

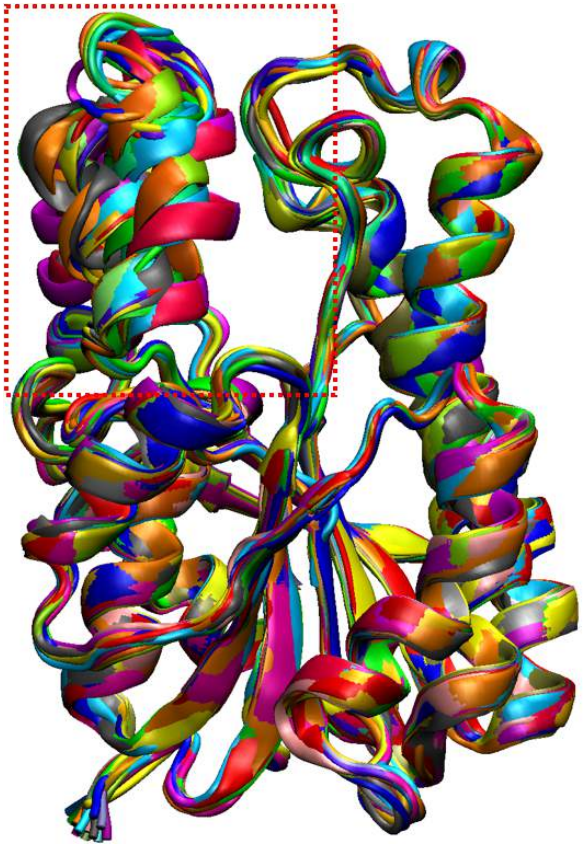


Figura 3.2 – Estruturas cristalinas da InhA de *Mycobacterium tuberculosis*, armazenadas no sítio do PDB, alinhadas pelo programa VMD [HDS96]. As estruturas estão em diferentes cores na representação de fitas, ressaltando as similaridades e dissimilaridades estruturais entre as enzimas. Em destaque, o quadro vermelho mostra a alça da esquerda, região com a maior dissimilaridade entre as proteínas.

As estruturas cristalinas analisadas apresentaram uma alta variabilidade no volume da cavidade de ligação do substrato (ver Tabela 3.1), evidenciando a necessidade de se considerar estruturas flexíveis da enzima InhA. Essas diferenças existentes nas características da superfície possibilitam a exploração de regiões funcionais diferentes na proteína. A próxima seção apresenta detalhes da geração do modelo FFR da proteína InhA de *Mtb*.

3.4 Modelo FFR da enzima InhA de *Mycobacterium tuberculosis*

A evolução computacional e o aprimoramento das técnicas de simulações de DM possibilitam a obtenção de simulações por DM capazes de reproduzirem uma flexibilidade maior com um aumento no número de estados gerados. Além disso, as diferentes cavidades de ligação do substrato identificadas na seção 3.3.1 evidenciaram a necessidade de se trabalhar um modelo de Receptor Totalmente Flexível para a enzima de InhA-NADH. A trajetória adotada como estudo de caso nesta tese foi gerada a partir da estrutura cristalina da enzima Enoyl-Reductase ou complexo InhA-NADH de *Mycobacterium tuberculosis* (PDB ID: 1ENY) [DQB$^+$95] contendo moléculas de água, conforme descrito em [GCNdS07].

O pacote do programa AMBER 9.0 [PCC+95] foi utilizado no desenvolvimento dessa trajetória, considerando o campo de força ff99SB [HAO+06]. Os parâmetros da macromolécula foram obtidos a partir da biblioteca parm99.dat [PCC+95], com todos os átomos explicitamente considerados. As cargas atômicas parciais da coenzima NADH foram geradas por cálculos *ab initio* com RHF/6-31G* e ajustados com o programa RESP [BCCK93]. Segundo [CCB+95], essas cargas são totalmente compatíveis com o campo de força do AMBER. Todos os átomos de hidrogênio, íons e moléculas de água foram submetidos a 100 passos de minimização de energia para remover os contatos muito próximos das forças de *van der Waals*. A pressão da simulação foi mantida em 1 atm e, para evitar a perturbação ao sistema, a temperatura foi aumentada gradualmente de 10 K para 298 K dividas em 6 etapas [GCNdS07]. Para essas etapas da temperatura, as velocidades foram novamente calculadas de acordo com a distribuição Maxwell-Boltzmann e equilibradas por 200 ps.

Nessa trajetória, as conformações foram capturadas em intervalos de 1 ps, resultando em um conjunto de microestados de 20.000 ps. Após a geração desse conjunto, é necessário avaliar as mudanças que ocorrem com a estrutura da proteína durante a simulação. Essa avaliação é frequentemente calculada com base no valor do RMSD entre a primeira estrutura da simulação (ou em relação a estrutura cristalina que originou a simulação), que é a referência para o cálculo da variação da posição de cada átomo, com as demais conformações. Os átomos utilizados para calcular o valor do RMSD variam, podendo ser considerados todos os átomos da proteína, ou somente os átomos do *backbone* (N, Cα, C e O), ou apenas os Cα. Nessa fórmula, quanto mais próximo de zero for o valor da dissimilaridade, maior será a sobreposição (similaridade) das estruturas comparadas.

Em simulações de DM, o sistema necessita de um tempo para estabilizar todos os parâmetros, dividindo a trajetória em duas partes. Na primeira parte, também chamada de fase de equilibração, as conformações apresentam variações estruturais significativas, não sendo consideradas como estruturas adequadas devido à instabilidade da simulação. Na segunda parte, também conhecida como fase de produção, as estruturas podem ser efetivamente estudadas uma vez que o sistema se mantém em um estado de não-equilíbrio constante [vGB90]. A Figura 3.3 apresenta a avaliação feita com essa simulação considerando os valores dos átomos do *backbone* (N, Cα, C e O) da enzima InhA. A linha vermelha pontilhada delimita a separação entre as fases de equilibração e produção (aos 500 ps). A linha verde representa a média da variação dos valores do RMSD do estado fase de produção. Esse valor é estabilizado entre 1,00 Å e 1,80 Å, sendo o valor da sua média em torno de 1,38 ± 0,11 Å. Assim, devido à instabilidade inicial, as primeiras 500 conformações dessa simulação são descartadas. O conjunto restante dessa simulação forma o modelo de FFR [MWRNdS11]) da enzima de InhA-NADH de *Mtb* com 19,5 ns.

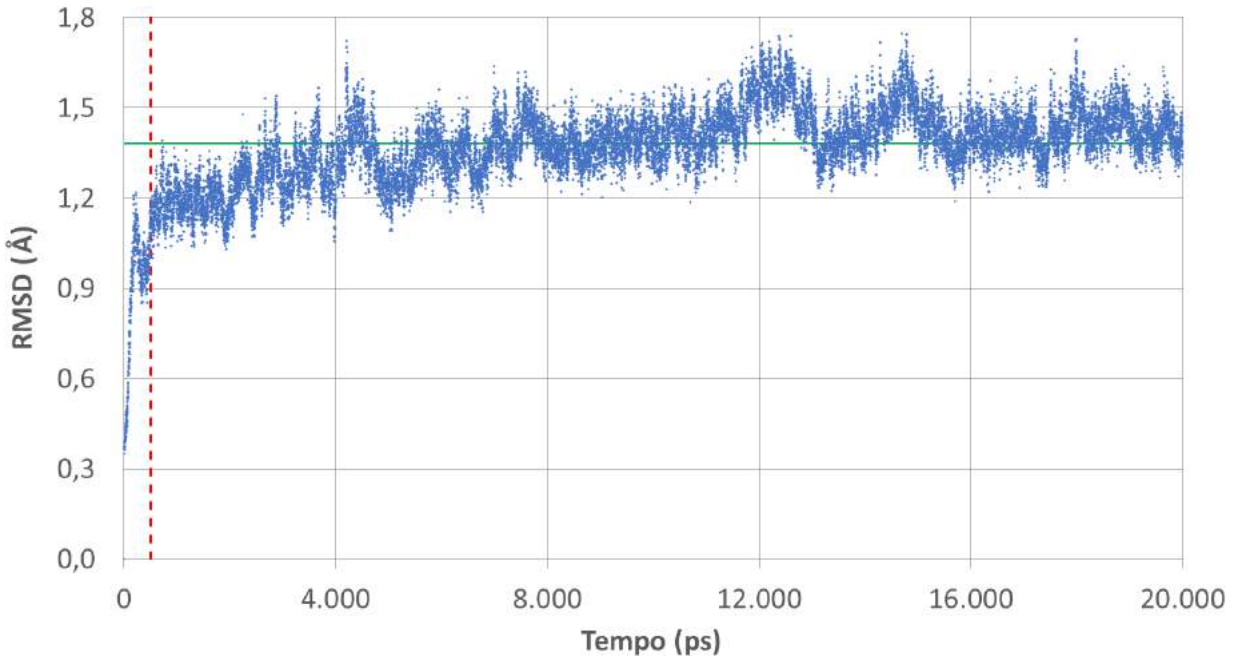


Figura 3.3 – Avaliação da variação do posicionamento dos átomos do *backbone* (N, Cα, C e O) entre a primeira conformação e o modelo FFR. A linha vermelha pontilhada delimita, aos 500 ps, a separação entre as fases de equilibração e produção. A linha verde representa a média da variação dos valores do RMSD do estado da fase de produção. Este valor é estabilizado entre 1,00 Å e 1,80 Å, sendo o valor da sua média em torno de 1,38 ± 0,11 Å. Valores gerados pelo programa Amber 12 [CDCI$^+$12].

3.5 Considerações finais

Este Capítulo apresentou detalhes sobre a enzima InhA de *Mtb*, descrevendo, primeiramente, a motivação social da escolha dessa enzima como o estudo de caso desta tese. Após, uma avaliação realizada com as estruturas cristalinas disponibilizadas no PDB permitiu uma série de análises utilizadas como parâmetro para a avaliação do modelo flexível gerado computacionalmente. As considerações sobre as avaliações das estruturas cristalinas e do modelo FFR são descritas nas próximas duas seções.

3.5.1 Estruturas cristalinas da enzima InhA

As estruturas cristalinas são frequentemente utilizadas como parâmetro para avaliar as interações entre os complexos receptor-ligante. Na seção 3.3, as estruturas cristalinas da enzima InhA de *Mtb* disponibilizadas no PDB foram listadas. O alinhamento realizado com as estruturas cristalinas possibilita uma análise estrutural da cavidade de ligação do substrato conforme o tipo de ligante/aduto cristalizado com a estrutura. Embora a Figura 3.2 tenha apresentado, visualmente, um alto grau de similaridade entre as estruturas, existem diferenças estruturais que somente podem ser visualizadas com outros tipos de representação. Para ressaltar essas diferenças, um novo tipo de representação é ado-

tado, focando apenas na cavidade do substrato. Essa mudança torna mais compreensível a avaliação estrutural, pois, devido ao grande número de estruturas, a representação com todos os resíduos diminui a qualidade da visualização.

A Figura 3.4 mostra alguns resíduos das proteínas da InhA na representação em palito. Esses resíduos determinam parte da cavidade do substrato, conforme a identificação do programa CASTp [BNL03]. As cores correspondem ao tipo de estrutura cristalizada com a enzima: em roxo as estruturas contendo somente a coenzima NADH, em laranja as proteínas contendo um aduto e em amarelo as estruturas cristalinas que possuem a coenzima NADH mais algum ligante. Essa figura evidencia uma importante situação observada em dois resíduos que pertencem à cavidade do substrato dessa enzima. Os resíduos TYR158 e PHE149 apresentam uma forma estrutural padrão, dependendo do tipo de composto que interage com a cavidade de ligação do substrato. Essas características estruturais definem o tipo de composto que possui a maior probabilidade de afinidade com a estrutura cristalina.

Um conjunto de experimentos de docagem molecular foi realizado entre os tipos de compostos e de proteínas. Os resultados mostraram que há tanto perdas da FEB, quanto alterações da correta predição da forma estrutural do composto. Esses testes demonstraram, também, haver diferenças na afinidade do complexo receptor-ligante conforme o tipo de estrutura cristalina. Assim, esses resultados auxiliam na elaboração de um método para a avaliação de modelos FFR, visto que esses modelos são criados a partir de uma estrutura cristalina. Desta forma, os modelos FFR, gerados a partir de um tipo específico e que sejam capazes de reproduzir a afinidade e a mesma forma estrutural dos compostos de outros tipos de estruturas, poderiam ser definidos como modelos de boa qualidade.

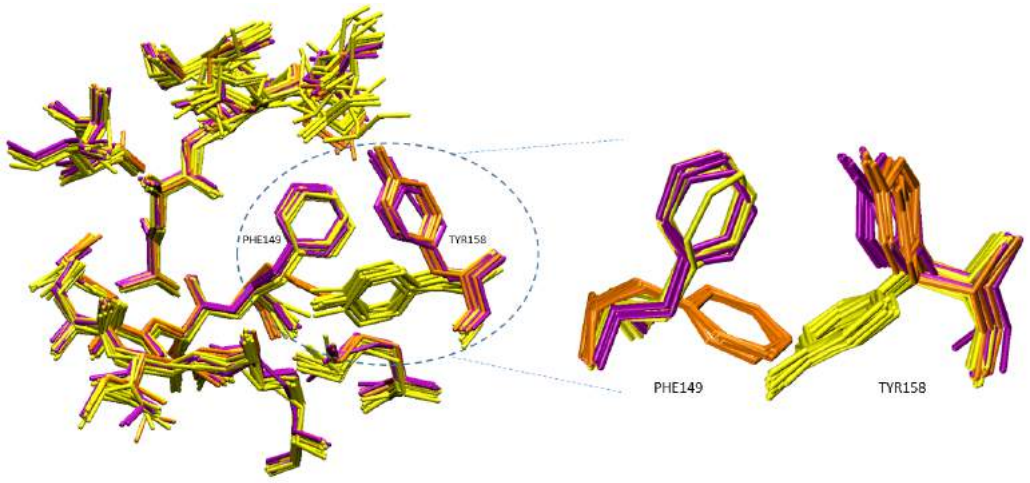


Figura 3.4 – Alguns resíduos delimitadores da cavidade do substrato apresentados no formato palito. As estruturas cristalinas foram divididas em cores que correspondem aos 3 tipos de estruturas encontradas: Em roxo estão as enzimas que possuem somente a coenzima NADH. Em laranja estão as estruturas cristalinas com adutos e, em amarelo, as proteínas que possuem a coenzima com algum ligante cristalizado. Em destaque, os resíduos TYR158 e PHE149 que apresentam similaridades na variação conforme o tipo de composto complexado.

Além disso, o alinhamento das estruturas cristalinas, apresentado na seção 3.3.1, possibilita aos pesquisadores uma forma rápida e acurada de verificar os resultados de experimentos de docagem molecular com esses compostos. Os receptores que não possuem um composto conhecido necessitam de uma análise detalhada de um especialista de domínio para identificar o posicionamento mais adequado. Desta forma, utilizar o posicionamento dos ligantes de estruturas cristalinas como referência fornece uma maior confiabilidade sobre a correta predição da estrutura encontrada pelo algoritmo de docagem molecular.

3.5.2 Modelo FFR da enzima InhA

Este capítulo descreveu um modelo FFR contendo, inicialmente, 20.000 conformações. No entanto, as avaliações do RMSD mostraram uma grande instabilidade no início da trajetória, sendo as primeiras 500 conformações descartadas. Esse modelo FFR não foi o primeiro a ser gerado em nosso grupo. Nos últimos anos, importantes progressos foram realizados em nossas pesquisas considerando as avaliações obtidas a partir das interações entre o modelo FFR de 3.100 ps e 4 pequenas moléculas (NADH, TCL, PIF e ETH) [De 12, WMNdSR09]. No entanto, no trabalho apresentado em De Paris [De 12] não houve uma investigação aprofundada sobre a qualidade das docagens (sua contribuição foi direcionada para a redução do tempo necessário em simulações de docagens moleculares), e em [MSRNdS07, MSR$^+$08, WMNdSR09] os estudos foram feitos na cavidade alvo da enzima InhA sem a presença da coenzima NADH. Essa coenzima é um composto orgânico encontrado nas células de todos os seres vivos e possui fundamental importância nas reações metabólicas, tendo um papel preponderante na produção de energia para a célula. Assim, encontrar candidatos a fármaco que atuem como concorrentes à mesma posição adotada por essa coenzima, pode impossibilitar a docagem da coenzima na InhA. Caso essa coenzima não forme o complexo na cavidade da InhA, isso pode ocasionar reações biológicas indesejáveis no paciente [PdSR$^+$13, QDS$^+$96]. Então o objetivo é buscar novos inibidores que sejam seletivos para a enzima InhA-NADH de *Mtb* com a coenzima NADH fazendo parte dessa proteína, restringindo o sítio de ligação de modo que fique livre apenas a cavidade de ligação do substrato.

Após a geração desse modelo FFR, é apropriado que essa nova trajetória seja submetida a diversos experimentos *in silico* baseados em docagem molecular a fim de validar as estruturas obtidas. Desta forma, espera-se que modelos FFR, gerados a partir de uma estrutura cristalina, sejam capazes de reproduzir a afinidade e a mesma forma estrutural dos compostos de outros tipos de estruturas cristalinas (Figura 3.4). Contudo, uma validação completa desse modelo FFR somente poderia ser feita utilizando a completa reprodução de experimentos *in silico* com todos os ligantes, obtidos das estruturas cristalinas, conhecidos atualmente na comunidade científica. Esse conjunto de experimentos validaria com um alto grau de confiabilidade o modelo FFR gerado.

Nesse sentido, o próximo Capítulo apresenta 3 estudos com a enzima InhA de *Mycobacterium tuberculosis*. O primeiro realiza uma avaliação das estruturas cristalinas da InhA depositadas no PDB [BWF^{+}00] com o conjunto dos seus 20 respectivos ligantes. O segundo estudo avalia se esse mesmo conjunto de ligantes consegue obter boas interações com a estrutura cristalina 1ENY, que gerou o modelo FFR. Por fim, o terceiro estudo avalia os experimentos de docagem molecular realizados entre o modelo FFR da enzima InhA de *Mtb* de 19,5 ns e o conjunto de ligantes/adutos das estruturas cristalinas.

4. Avaliação da qualidade das estruturas do modelo FFR da InhA de *Mtb* de 19,5 ns

Neste Capítulo são apresentados detalhes da avaliação de 3 conjuntos de experimentos de docagem molecular. Esses conjuntos de experimentos buscam identificar padrões de interações *in silico* do complexo receptor-ligante de estruturas bem conhecidas *in vivo*. Desta forma, as análises consideram uma seleção de estruturas cristalinas de proteínas com ligantes presentes no sítio do *Protein Data Bank* (sítio web que disponibiliza estruturas experimentais de proteínas) [BWF+00]. Essas proteínas cristalinas possuem grande respaldo na comunidade científica. No entanto, existe uma limitada quantidade de estruturas da InhA em virtude da complexidade e dos custos envolvidos no processo da cristalografia. Devido a essas limitações, pesquisadores têm investido na utilização de modelos capazes de incorporarem a flexibilidade da proteína, possibilitando uma variedade de estruturas que não são contempladas cristalograficamente. Contudo, estudos precisam aferir, também, a qualidade dos modelos criados, porque o aumento da quantidade dos graus de liberdade pode ocasionar a geração de estruturas falso positivas. Assim, as informações resultantes das avaliações de estruturas cristalinas servem como parâmetros para analisar os resultados de docagem molecular realizados com o modelo FFR de 19,5 ns.

A próxima seção descreve o protocolo utilizado para a preparação dos arquivos a serem submetidos a execução dos experimentos de docagem molecular. Em seguida, os experimentos de *redocking* e *cross docking* com os compostos obtidos das estruturas experimentais existentes da InhA de *Mtb*, descritas na Tabela 3.1, são descritos em detalhes.

4.1 Protocolo de docagem molecular

Esta seção apresenta os passos adotados para a preparação dos arquivos submetidos aos experimentos de docagem molecular. O protocolo apresentado nesta seção segue as orientações definidas em [HM08, MHO08]. A subseção 4.1.1 mostra as etapas para a preparação do receptor e do ligantes. Após, as subseções 4.1.2 e 4.1.3 descrevem os arquivos de parâmetros gerados para os programas AutoGrid e AutoDock, respectivamente.

4.1.1 Preparação da proteína e do ligante

A preparação dos arquivos da proteína e do ligante seguiu o protocolo definido por [HM08, MHO08], utilizando como interface gráfica o AutoDockTools 1.5.6 [MHL+09]. Todas as moléculas foram editadas, seguindo os seguintes passos:

- Passo 1: Analisar a estrutura, verificando a correta composição estrutural das interações entre os átomos.
- Passo 2: Remover as moléculas de água, uma vez que os experimentos de docagem molecular utilizados nesta tese não consideram a presença deste solvente.
- Passo 3: Adicionar todos os Hidrogênios da proteína, renumerando os átomos para manter a sequência ordenada ao final da preparação.
- Passo 4: Atribuir as cargas parciais de cada tipo de átomo. Nesta tese foi utilizado o método desenvolvido por Gasteiger para a atribuição das cargas parciais. Nesse método, o valor total da carga da estrutura deve ser próximo de um número inteiro [GM80].
- Passo 5: Unir os hidrogênios não-polares com o átomo pesado mais próximo, somando as cargas de Gasteiger dos átomos unidos.
- Passo 6: Assinalar os tipos de átomos “AD4.0”, convertendo os códigos de átomos que o AutoDock não identifica.
- Passo 7: Salvar a estrutura no formato “pdbqt”, mantendo apenas o rótulo “ATOM” neste arquivo final.

Neste ponto, o arquivo do receptor está pronto para a preparação dos arquivos de parâmetros do AutoGrid e do AutoDock. No entanto, existe mais uma sequência de passos para terminar a edição do ligante, considerando, ou não, a sua flexibilidade:

- Passo 8: Abrir a estrutura do ligante no formato “pdbqt” salva no Passo 7.
- Passo 9: Selecionar as torções mais adequadas identificadas pelo programa ou definir o número 0 de torção para tratar o ligante rígido.
- Passo 10: Salvar novamente a estrutura no formato “pdbqt”.

Esses passos para a preparação das moléculas foram aplicados no modelo FFR de 19,5 ns e nos ligantes e, também em seus respectivos receptores, detalhados nas Figuras 4.1 e 4.2. Conforme as diferenças estruturais descritas na seção 3.2.1, a preparação dos arquivos para os experimentos de docagem molecular são feitas de formas distintas, de acordo com o tipo de composto a ser testado:

- Ligantes: é necessário avaliar a interação molecular da pequena molécula com a enzima InhA contendo a coenzima NADH já acoplada em sua estrutura.
- Adutos: são formados por ligantes conectados com a coenzima NADH. Os experimentos de docagem molecular com adutos são realizados apenas com a enzima InhA.

As Figuras 4.1 e 4.2 descrevem as estruturas 3D dos ligantes e adutos extraídos das estruturas cristalinas descritas na Tabela 3.1 disponíveis à época em que esses experimentos foram avaliados. Além da forma estrutural, estão incluídas as informações sobre as ligações rotacionáveis de cada composto. Portanto, os cálculos de docagem molecular nesta etapa consideram o alvo molecular rígido e o ligante na forma flexível.

Ao final desse conjunto de passos, ambos, receptores e ligantes, estão prontos para serem submetidos ao programa AutoGrid e, posteriormente, ao programa AutoDock. As próximas subseções descrevem os parâmetros que definem o tamanho e o posicionamento da região que delimita o espaço de busca e, também, os parâmetros de entrada do algoritmo de docagem molecular.

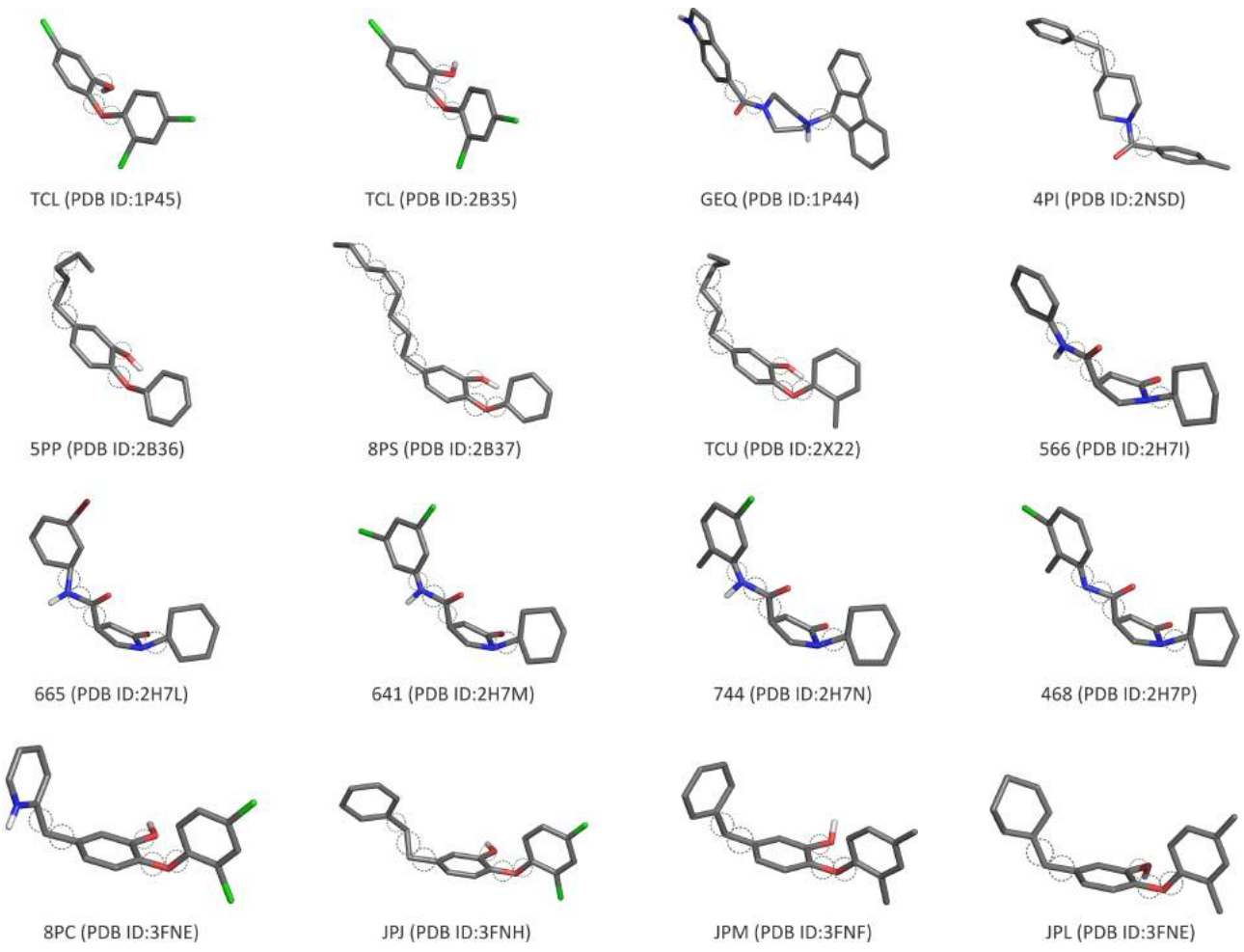


Figura 4.1 – Estrutura 3D dos ligantes usados nos experimentos de docagem molecular. Cada ligante está colorido de acordo com o tipo de átomo (Carbono: cinza, Nitrogênio: azul, Oxigênio: vermelho, Cloro: verde e Hidrogênio: branco), identificado com o seu nome e a estrutura cristalina de origem (PDB ID). Os círculos pontilhados identificam as ligações rotacionáveis que foram selecionadas pelo programa AutoDockTools 1.5.6 [MHL+09].

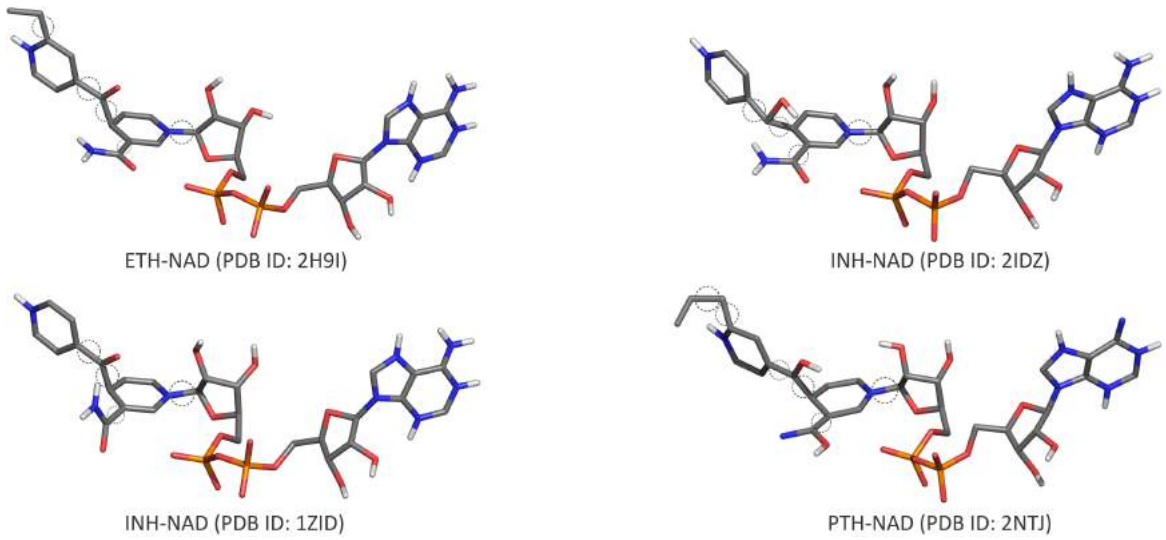


Figura 4.2 – Estrutura 3D dos adutos utilizados nos experimentos de docagem molecular. Cada aduto está colorido de acordo com o tipo de átomo (Carbono: cinza, Nitrogênio: azul, Oxigênio: vermelho, Cloro: verde, Fósforo: laranja e Hidrogênio: branco). As moléculas estão identificadas com o seu nome e os nomes das estruturas cristalinas de origem (PDB ID). Os círculos pontilhados identificam as ligações rotacionáveis que foram selecionadas pelo programa AutoDockTools 1.5.6 [MHL+09].

4.1.2 Preparação do arquivo de parâmetros do AutoGrid

O AutoGrid é um programa que calcula um conjunto de malhas para medir a afinidade entre os átomos do ligante e os átomos da cavidade alvo do receptor. Essas malhas definem a energia potencial dos átomos do ligante com todos os átomos do receptor com uma distância de até 8 Å [MHL+09]. O tamanho dessa malha é definido pelas variáveis n_x, n_y e n_z, formando uma caixa 3D (conforme pode ser visto na Figura 4.3). Antes de calcular os valores da energia potencial, o centro dessa caixa 3D deve ser posicionado na região onde espera-se que ocorra o posicionamento da pequena molécula na macromolécula. Assim, essa caixa serve para restringir as regiões acessíveis ao ligante. A Figura 4.3 mostra um exemplo de uma caixa 3D.

Neste trabalho, o posicionamento da caixa 3D é definido com as mesmas coordenadas do centro geométrico da pequena molécula de referência. O tamanho da caixa é definido conforme o tipo da pequena molécula a ser avaliada, podendo ser um ligante ou um aduto. Após a centralização das caixas 3D projetadas para as estruturas cristalinas e para a primeira conformação do modelo FFR, elas são avaliadas manualmente por inspeção visual para validar se os limites da caixa englobam ao mesmo tempo a estrutura da pequena molécula e a área acessível ao solvente da cavidade de ligação do substrato.

As malhas foram geradas conforme a composição dos átomos do complexo receptor-ligante. Além das malhas de afinidade dos átomos, outras duas malhas são geradas para catalogar o potencial eletrostático e o potencial de dessolvatação. Um fragmento desse arquivo de parâmetros do programa AutoGrid está na Figura 4.4.

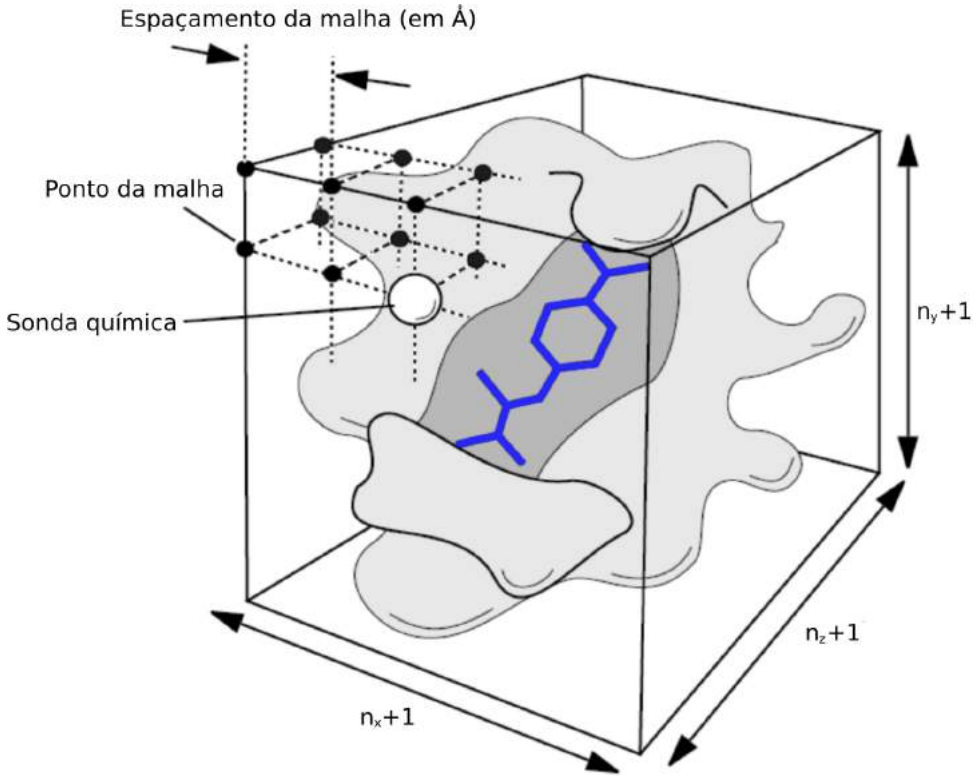


Figura 4.3 – Exemplo da malha de afinidade entre os átomos do ligante e os átomos da cavidade alvo do receptor gerada pelo programa AutoGrid. A sonda química passa pela malha 3D delimitada pelos parâmetros n_x, n_y e n_z, registrando a energia potencial em cada ponto desta malha. A superfície da proteína está representada em cinza claro e em cinza escuro está a região da cavidade de ligação. O ligante está na representação de palitos em azul escuro. Figura adaptada de [MGH+01].

```
npts 50 50 50                  # num.grid points in xyz
gridfld receptor.maps.fld      # grid_data_file
spacing 0.375                  # spacing(A)
receptor_types __              # receptor atom types
ligand_types __                # ligand atom types
receptor receptor.pdbqt        # macromolecule name
gridcenter  x   y   z          # xyz-coordinates or auto
smooth 0.5                     # store minimum energy w/in rad(A)
map receptor.__.map            # atom-specific affinity map
elecmap receptor.e.map         # electrostatic potential map
dsolvmap receptor.d.map        # desolvation potential map
dielectric -0.1465             # <0, AD4 distance-dep.diel;>0, constant
```

Figura 4.4 – Fragmento do arquivo de parâmetros do programa AutoGrid contendo informações dos átomos do ligante e do receptor. Este arquivo também determina o tamanho e o posicionamento da caixa utilizadas pelo programa AutoGrid. O tamanho da caixa 3D (parâmetros n_x, n_y e n_z) é ajustado conforme o tipo da pequena molécula, sendo a 50x50x50 para a docagem com ligantes e 60x50x50 para a docagem com adutos. Os tipos de átomos são definidos conforme a composição dos átomos do complexo receptor-ligante.

4.1.3 Preparação do arquivo de parâmetros do Autodock4

Assim como o programa AutoGrid, o AutoDock também necessita de um arquivo de parâmetros para determinar as principais propriedades dos arquivos do receptor e do ligante a serem avaliados (arquivos gerados na seção 4.1.1). As conformações a serem testadas em experimentos de docagem molecular são definidas como estruturas rígidas (a flexibilidade do receptor é considerada somente quando existe a submissão de diversas conformações da mesma proteína, como no modelo FFR). Os ligantes são definidos como estruturas flexíveis, conforme as ligações rotacionáveis definidas nas Figuras 4.1 e 4.2.

Nesta tese foi utilizado o algoritmo genético Lamarckiano como método de busca global. A população inicial começou com 150 indivíduos, contendo um número máximo de 27.000 gerações, ou 3.000.000 de avaliações de energia, mantendo sempre o melhor indivíduo a cada geração. Os operadores de mutação e *crossover* foram aplicados com valores de 0,02 e 0,80, respectivamente. A busca local é realizada pelo algoritmo Solis e Wets [SW81]. Os ligantes são definidos como estruturas flexíveis e desta forma, o número de execuções a serem feitas aumentou de 10 para 25 para explorar uma quantidade maior de resultados de docagem. Um fragmento desse arquivo contendo os parâmetros de entrada do AutoDock é mostrado na Figura 4.5.

4.2 Experimentos de Docagem Molecular

Os experimentos de docagem molecular foram realizados com o programa Autodock4 *release* 4.2.5.1 [MHL+09], utilizando uma função de pontuação empírica para estimar a energia livre de ligação da interação. Esses experimentos podem ser separados em duas categorias quanto ao tipo de complexos receptor-ligante avaliados [MGH+98]:

- *Redocking*: é o processo cujo objetivo é o de reencontrar, a partir de uma simulação computacional com experimentos de docagem molecular, a posição original da pequena molécula dentro da própria estrutura do receptor que ela foi retirada.
- *Cross docking*: é o processo de se avaliar um complexo receptor-ligante, onde o ligante a ser testado não tenha sido obtido da própria estrutura do receptor o qual esteja sendo avaliado.

O processo de *redocking* é frequentemente realizado para verificar se os parâmetros de docagem especificados no arquivo de entrada para o método de docagem são corretos e capazes de recuperar a interação e a estrutura de um complexo conhecido. Já o processo de *cross docking* é interessante para estudar a especificidade entre os receptores e os ligantes. Quando existem vários receptores da mesma proteína, com diferentes conformações, a avaliação pode fornecer valiosas informações sobre os efeitos do encaixe induzido após ligação.

```
## Parâmetros Gerais
autodock_parameter_version 4.2      # used by autodock to set parameter
outlev 1                            # diagnostic output level
seed 71277 142554                   # seeds for random generator
ligand_types  __                    # atoms types in ligand
fld receptor.maps.fld               # grid_data_file
map receptor.__.map                 # atom-specific affinity map
elecmap receptor.e.map              # electrostatics map
desolvmap receptor.d.map            # desolvation map
move ligand_ini.pdbqt               # small molecule
about  coord_x  coord_y  coord_z    # small molecule center
rmsref ligand_ref.pdbqt             # reference small molecule
torsdof  X                          # torsional degrees of freedom

## Parâmetros do estado inicial da busca
tran0 random                        # initial coordinates/A or random
quaternion0 random                  # initial orientation
dihe0 random                        # initial dihedrals (relative) or random

## Parâmetros do Algoritmo Lamarckiano
ga_pop_size 150                     # nº of individuals in population
ga_num_evals 3000000                # max nº of energy evaluations
ga_num_generations 27000            # max nº of generations
ga_mutation_rate  0.02              # rate of gene mutation
ga_crossover_rate 0.80              # rate of crossover
set_ga                              # set the above parameters for GA or LGA
ga_run 25                           # do this many hybrid GA-LS runs
ga_elitism 1                        # nº of top individuals to survive to
                                      next generation

## Parâmetros da análise dos agrupamentos
rmstol 2.0                          # cluster_tolerance/A
analysis                            # perform a ranked cluster analysis
```

Figura 4.5 – Fragmento do arquivo contendo os parâmetros utilizados nos experimentos de docagem molecular. O valor da variável *torsdof* varia conforme o ligante/aduto a ser testado de acordo com o número de torções da pequena molécula. O algoritmo genético tem uma população inicial de 150 indivíduos, sendo cada cálculo efetuado por até 27.000 gerações ou 3 milhões de avaliações da energia. Devido à flexibilidade da molécula, o número de execuções escolhido é de 25.

Um dos principais objetivos deste Capítulo é a avaliação da flexibilidade do modelo FFR de InhA de *Mtb* de 19,5 ns, descrevendo importantes características físico-químicas das interações deste modelo com os 20 ligantes/adutos candidatos a fármaco da proteína InhA presentes no sítio do *PDB* e descritos na subseção 4.1.1. Contudo, antes de realizar os experimentos de *cross docking* com o modelo FFR, são necessárias duas avaliações. A primeira trata-se de uma avaliação de *redocking* com os ligantes/adutos das estruturas da InhA de *Mtb* disponíveis no PDB para validar os parâmetros adotados para a definição das malhas e da docagem molecular. A segunda avalia a flexibilidade do modelo FFR com um estudo da proteína 1ENY, cuja estrutura cristalina foi a base para a geração do modelo FFR de 19,5 ns. Desta forma, docagens moleculares com o modelo FFR, reproduzindo os sítios de ligação das estruturas cristalinas não acessíveis à 1ENY, evidenciariam a importância de modelo flexíveis. Assim, as próximas Subseções descrevem os 3 experimentos de docagem molecular realizados nesta tese para avaliar as estruturas da enzima InhA de *Mtb*.

4.2.1 Experimentos de *redocking* com as estruturas cristalinas da proteína InhA.

A execução desse conjunto de experimentos de docagem molecular é baseada no protocolo de preparação dos arquivos de parâmetros, descritos nas seções 4.1.2 e 4.1.3, considerando o conjunto de pequenas moléculas das Figuras 4.1 e 4.2 testados com suas respectivas proteínas. A quantidade de experimentos necessários é relativamente baixa, visto que os testes de docagem molecular avaliam cada ligante/aduto com a sua respectiva estrutura cristalina. Assim, as avaliações de *redocking* não representam uma tarefa onerosa.

Após a execução dos experimentos de docagem molecular, o melhor encaixe indicado pelo programa AutoDock para cada interação receptor-ligante é identificado pelo menor valor da FEB e, depois, pelo menor valor do RMSD com relação ao ligante de referência. Esses resultados estão descritos na coluna *redocking* da Tabela 4.1. A posição resultante da predição com a menor energia fornecida pelo AutoDock é definida como *"adequadamente docada"* quando o valor do RMSD resultar em um valor menor ou igual a 2.0 Å a partir da posição da estrutura de referência [GHK00, WLW03, VGK05].

Na Tabela 4.1, pode-se observar que os experimentos de *redocking* apresentaram um alto grau de similaridade entre as poses preditas pelo programa de docagem e a pose real obtida da estrutura cristalina. Apenas 20% dos experimentos executados apresentam o valor do RMSD com 1,0 Å ou mais da estrutura experimental, corroborando a alta precisão dos parâmetros de docagem estipulados. Com base nesses resultados, pode-se afirmar que tanto os valores dos parâmetros dos programas de docagem (definições do algoritmo de busca e da função de escore) quanto os valores calculados das cargas parciais mostraram-se adequados.

4.2.2 Experimentos de *cross docking* entre os ligantes das estruturas cristalinas da proteína InhA com o receptor 1ENY.

O principal objetivo desse conjunto de experimentos é identificar se os ligantes cristalizados com as estruturas da InhA conseguem ser gerados com uma estrutura conformacional similar, quando estes ligantes são testados com uma estrutura que tenha sido cristalizada apenas com a coenzima dentro do sítio ativo. A estrutura cristalina 1ENY é um exemplo desse tipo de estrutura, sendo esta também a estrutura base para a geração do modelo FFR de 19,5 ns.

A execução desse conjunto de experimentos segue o protocolo de preparação dos arquivos de parâmetros descritos nas seções 4.1.2 e 4.1.3, considerando o conjunto de pequenas moléculas sendo testadas com a estrutura 1ENY. Assim como na etapa de *redocking*, a quantidade de experimentos necessários também é baixa, visto que os testes consideram os ligantes/adutos das estruturas cristalinas com apenas uma proteína. Assim, essas avaliações também não representam uma tarefa onerosa.

Na Tabela 4.1, a coluna *cross docking* com a 1ENY mostra os valores da FEB e do RMSD das docagens moleculares realizadas com o conjunto de dados desta seção. Esses resultados demonstram que 70% dos experimentos de *cross docking* com a estrutura da 1ENY não conseguiram reproduzir a estrutura conformacional indicada pela estrutura de referência, ou seja, apenas 30% dos resultados apresentaram valores do RMSD menores ou iguais a 2.0 Å, sendo considerados empiricamente como estruturas dissimilares [GHK00, WLW03, VGK05]. A análise dos valores de FEB desse conjunto de experimentos também apresentou valores maiores que os encontrados nos experimentos de *redocking* em 95% dos casos.

4.2.3 Experimentos de *cross docking* entre os ligantes das estruturas cristalinas da proteína InhA com o modelo FFR de 19,5 ns.

Os experimentos de *cross docking* descritos nesta subseção são aplicados para avaliar o modelo FFR da InhA de 19,5 ns com o mesmo conjunto de ligantes e adutos utilizados nos experimentos anteriores. A Figura 4.6 esboça a forma de avaliação dos experimentos de *cross docking*, mostrando os experimentos de docagem molecular de cada conformação do modelo FFR com cada pequena molécula a ser avaliada.

Nos experimentos anteriores, os arquivos de configuração da caixa e da docagem molecular haviam sido criados manualmente. No entanto, a preparação do modelo FFR de 19,5 ns requer a preparação de 60.000 arquivos para cada experimento. O *middleware FReMI*, desenvolvido por [DPFdSR11], foi utilizado para criar os arquivos de preparação de forma automática, utilizando um arquivo modelo. Os arquivos modelo variam de acordo com as características de cada ligante/aduto a ser testado.

Tabela 4.1 – Resultado dos experimentos de *redocking* das estruturas cristalinas e *cross docking* dos ligantes das estruturas cristalinas com a estrutura 1ENY e com o modelo FFR. A comparação desses experimentos evidencia os melhores valores da FEB obtidos nos experimentos de *redocking*. O experimento de *cross docking* com o modelo FFR, embora não tenha apresentado os valores da FEB similares aos experimentos de *redocking*, resultou em um ganho tanto nos valores da FEB quanto do RMSD quando comparados aos experimentos de *cross docking* com a estrutura 1ENY. Em verde estão os valores de RMSD até 2.0 Å e, em vermelho os valores acima deste limiar.

PDB ID	Composto	*Redocking*		*Cross docking* com a 1ENY		*Cross docking* com o modelo FFR 19,5 ns	
		FEB	RMSD	FEB	RMSD	FEB	RMSD
1P44	GEQ	-11,6	0,5	-9,5	3,9	-10,4	2,8
1P45	TCL	-8,7	0,7	-7,1	2,1	-7,0	0,4
2B35	TCL	-6,8	0,5	-7,2	2,0	-8,1	1,5
2B36	5PP	-7,8	1,0	-7,7	3,8	-8,1	1,4
2B37	8PS	-7,3	0,9	-6,9	2,9	-7,3	2,9
2H7I	566	-8,7	1,2	-7,0	1,9	-7,2	1,7
2H7L	665	-10,5	0,9	-8,1	3,4	-8,1	2,3
2H7M	641	-9,8	0,8	-8,6	3,3	-8,1	2,4
2H7N	744	-10,1	0,9	-8,3	4,3	-8,3	2,7
2H7P	468	-9,5	0,7	-8,4	2,9	-8,6	2,5
3FNE	8PC	-9,9	1,1	-8,2	4,6	-8,2	1,8
3FNF	JPM	-10,0	0,4	-7,7	2,4	-6,6	1,9
3FNG	JPL	-10,4	0,5	-8,2	2,1	-9,7	1,6
3FNH	JPJ	-10,2	1,3	-8,3	4,2	-9,3	1,8
2NSD	4PI	-11,0	0,8	-7,8	4,4	-9,0	2,0
2X22	TCU	-10,0	0,5	-6,5	3,0	-8,3	1,4
2H9I	ETH-NAD	-11,7	0,8	-8,8	1,9	-11,5	1,2
1ZID	INH-NAD	-10,3	0,8	-6,9	2,0	-7,0	1,6
2IDZ	INH-NAD	-11,4	0,7	-8,3	2,8	-8,9	1,3
2NTJ	PTH-NAD	-12,9	0,8	-10,3	1,9	-11,1	1,8
Média:		-9,9	0,8	-8,0	3,0	-8,5	1,9
Desvio padrão:		1,5	0,2	0,9	0,9	1,3	0,6

Devido ao tempo computacional necessário para executar esses experimentos, o *middleware* FReMI [DPFdSR11] foi utilizado para gerenciar e paralelizar as 19.500 experimentos de docagem molecular. Essas execuções foram realizadas em 3 computadores com processador Quad-core i7 2600 3.4 GHz, com 12 GB de RAM, SO Linux Ubuntu 13.04. Embora os testes de cada conformação do modelo FFR tenham sido paralelizados e executados em máquinas de alto desempenho, alguns experimentos necessitaram acima de 4 semanas para serem finalizados. No total, aproximadamente 4 meses foram necessários para a execução completa.

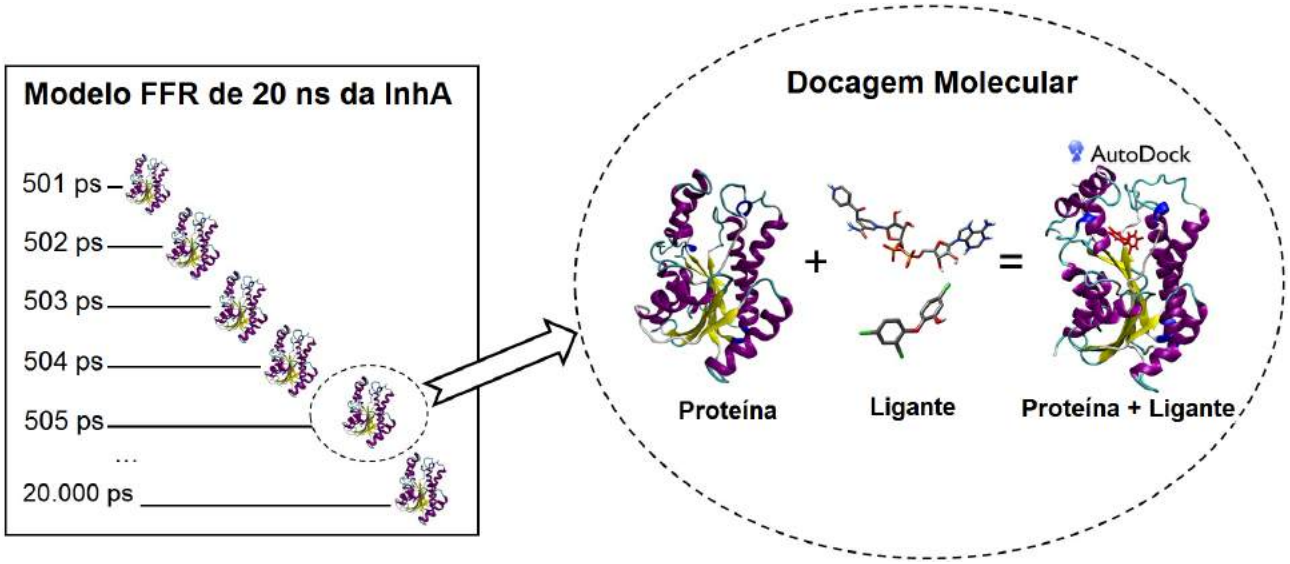


Figura 4.6 – Representação do processo da docagem molecular entre o modelo FFR e pequenas moléculas. As estruturas dentro do retângulo representam o modelo FFR de 20.000 ps definido como o estudo de caso deste trabalho [GCNdS07]. A região em destaque mostra os experimentos de docagem molecular realizados entre cada conformação com diferentes ligantes. Adaptado de [QDPRNdS14].

Na Tabela 4.1, a coluna *cross docking* com o modelo FFR de 19,5 ns apresenta os resultados dos experimentos de docagem molecular descritos nesta subseção. Um comparativo mostra que, de modo geral, os resultados desta seção são melhores que os apresentados no experimento de *cross docking* com a estrutura da 1ENY. Essa diferença fica mais evidente na comparação entre as médias e os desvios padrão da FEB e do RMSD. Existem particularidades nas medidas de FEB e RMSD que reforçam a necessidade de uma interpretação diferente entre os valores do desvio padrão da FEB e do RMSD[1].

As avaliações das médias da FEB e do RMSD mostram que os experimentos de docagem molecular descritos nesta subseção foram mais adequados que os experimentos de *cross docking* com a estrutura da 1ENY. Além disso, os valores do desvio padrão do RMSD apontaram que as estruturas resultantes da docagem molecular são estruturalmente mais similares à estrutura de referência. As Figuras 4.7 e 4.8 apresentam as comparações dos valores da FEB e do RMSD entre os experimentos de *redocking* das estruturas cristalinas e *cross docking* dos ligantes das estruturas cristalinas com a estrutura 1ENY e com o modelo FFR de 19,5 ns [GCNdS07], respectivamente.

Esses experimentos também apontam que a diferença entre os valores encontrados nos experimentos de *cross docking* da estrutura cristalina 1ENY e do modelo FFR estão diretamente relacionados à flexibilidade da proteína. Nos adutos, a parte da coenzima NADH se encaixa de maneira adequada à sua cavidade de ligação. No entanto,

[1] A medida da FEB pode apresentar bons resultados com valores muitos diferentes devido as estruturas envolvidas na interação do complexo receptor-ligante. No entanto, independente da estrutura avaliada, valores próximos de 0.0 Å representam bons resultados para o RMSD.

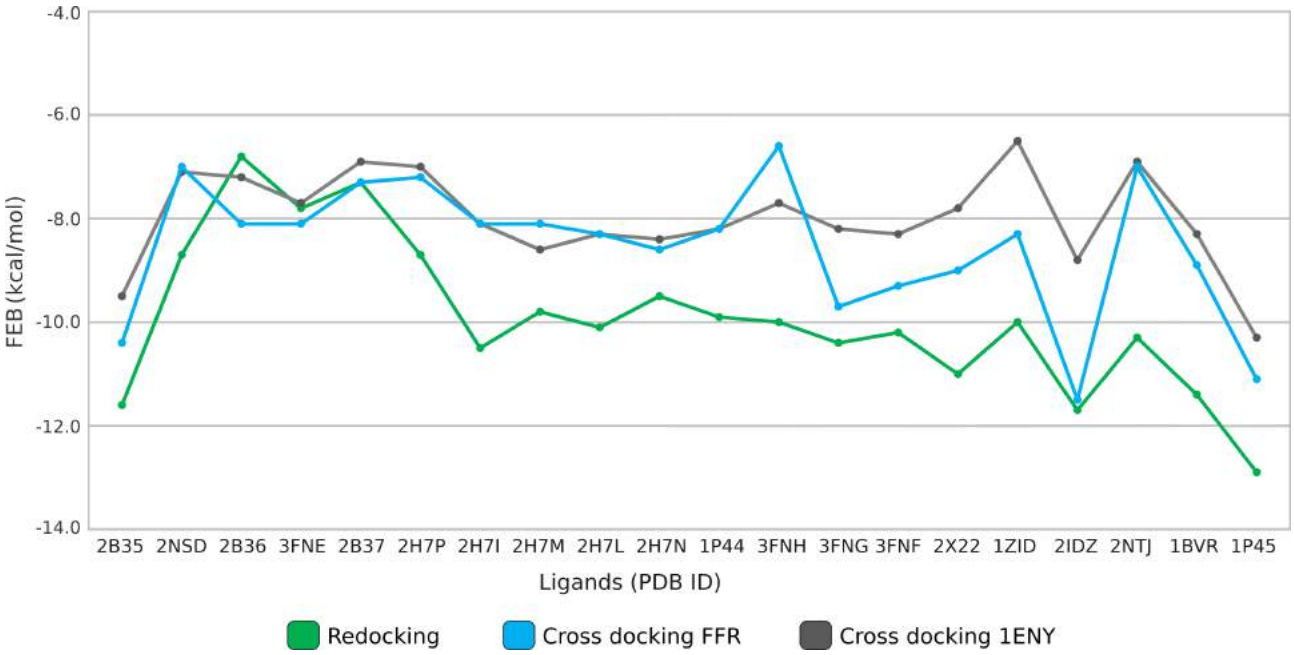


Figura 4.7 – Comparação dos valores da FEB entre os experimentos de *redocking* das estruturas cristalinas e *cross docking* dos ligantes das estruturas cristalinas com a estrutura 1ENY e com o modelo FFR. A comparação desses experimentos evidencia os melhores valores da FEB obtidos nos experimentos de *redocking* (em verde). Os resultados obtidos nos experimentos de *cross docking* com o modelo FFR apresentaram, em geral, valores mais adequados que os experimentos de *cross docking* com a estrutura 1ENY.

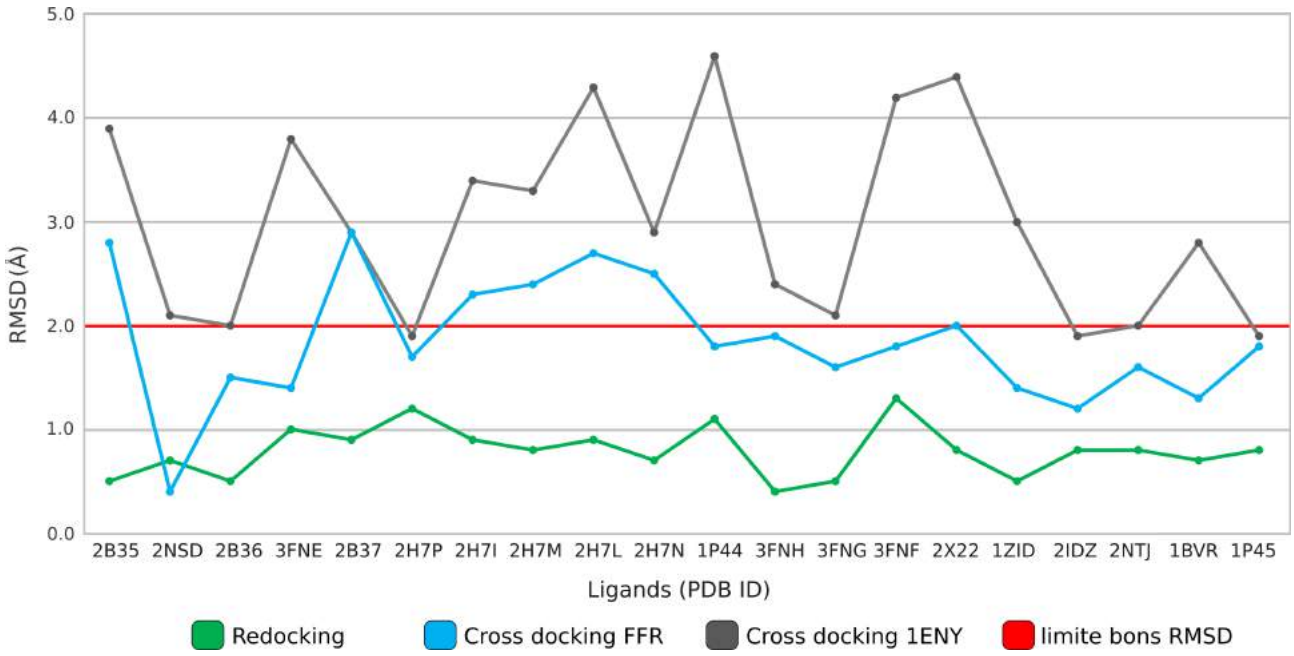


Figura 4.8 – Comparação dos valores do RMSD entre os experimentos de *redocking* das estruturas cristalinas e *cross docking* dos ligantes das estruturas cristalinas com a estrutura 1ENY e com o modelo FFR. Os experimentos de *redocking* apresentaram resultados muitos próximos das conformações de referência. Os experimentos de *cross docking* com o modelo FFR também mostraram bons resultados (em azul), obtendo, em geral, valores de RMSD menores ou iguais a 2.0 Å a partir da posição da estrutura de referência (em vermelho).

a formação de fendas na cavidade de ligação do substrato para o encaixe adequado da porção do ligante tornou-se possível somente no modelo FFR. Desta forma, os melhores resultados tanto de RMSD quanto os valores da FEB são obtidos nos experimentos com o modelo FFR. Cabe ressaltar que o modelo FFR foi gerado a partir de uma simulação DM do complexo enzima InhA-NADH. Os resultados obtidos com os ligantes não apresentaram sensíveis diferenças com relação aos valores da FEB, como já era esperado, devido às interações ocorrerem no mesmo sítio ligação. Por outro lado, os valores do RMSD foram consideravelmente menores que nos experimentos de *cross docking* com a 1ENY, claramente mostrando que o modelo FFR da InhA é capaz de capturar suas conformações de acordo com uma estrutura cristalina experimental.

4.3 Considerações Finais

Trabalhos prévios realizados em nosso grupo de pesquisa consideravam experimentos com até 4 ligantes em um modelo FFR de 3.100 conformações para aplicações de técnicas de mineração de dados [MSRNdS07, MSR$^+$08, WMNdSR09, De 12]. Este Capítulo descreveu a avaliação de um conjunto de 20 experimentos de docagem molecular com um modelo FFR de 19.500 conformações. Esse novo conjunto de dados permite validações com uma maior confiabilidade envolvendo os estudos de mineração de dados, visto que esses dados são um estudo completo das estruturas cristalinas reconhecidas pela comunidade científica.

Uma análise com um conjunto de experimentos de *redocking* e *cross docking* foi realizada, buscando aferir a qualidade dos experimentos *in silico* estudados. Os valores do RMSD da Tabela 4.1 mostraram que os experimentos de *cross docking* com o modelo FFR reproduziram adequadamente o modo de ligação existente nas estruturas cristalinas em 70%, enquanto os experimentos de *cross docking* com a 1ENY contemplaram apenas 30% dos resultados. Ao final, essa validação do modelo FFR de 19,5 ns, gerado a partir da estrutura cristalina 1ENY, foi capaz de reproduzir parte da afinidade e da forma estrutural dos compostos de outros tipos de estruturas, validando o modelo FFR gerado, ou seja, os experimentos com o modelo FFR mostraram que a partir de uma estrutura cristalina contendo somente a coenzima NADH, é possível gerar as conformações dos ligantes/adutos de outras enzimas cristalinas da InhA.

No próximo capítulo é descrito o desenvolvimento de métodos para selecionar um conjunto de estruturas representativas do modelo FFR utilizando técnicas de mineração de dados. Diferentes técnicas de agrupamento são relacionadas com base nas propriedades da cavidade de ligação do substrato. Após a identificação das partições pelo algoritmo, o conjunto de experimentos descritos neste capítulo é utilizados para verificar se os resultados da FEB de cada partição apresentam valores similares.

5. Seleção de conjuntos de conformações similares do modelo FFR de 19,5 ns baseado nas propriedades estruturais da cavidade de ligação do substrato da enzima InhA de *Mtb*

Algoritmos de agrupamento têm sido amplamente utilizados para reduzir a dimensionalidade de trajetórias de modelos FFR [STTC07, TvG94]. A maioria desses estudos investigam diferentes algoritmos de agrupamento usando os valores de RMSD de pares de átomos entre estruturas. No entanto, essa métrica pode não ser o indicador mais apropriado para particionar as conformações quando há uma cavidade de ligação do substrato conhecida, uma vez que os valores de RMSD podem ser influenciados pelas mudanças que ocorrem em partes externas da cavidade de ligação do substrato do receptor. Assim, é possível ocorrer cavidades de ligação do substrato com diferentes sítios de ligação e com o mesmo valor de RMSD. Por essa razão, abordagens novas e promissoras para reduzir modelos FFR, sem perda de informação estrutural crítica, devem ser investigadas [AL10].

Uma forma promissora de se agrupar as conformações de um modelo FFR é considerar apenas as estruturas que têm influência na cavidade de ligação do substrato. Assim, há uma redução do efeito das diferenças estruturais que ocorrem longe da cavidade alvo e que, muitas vezes, não influenciam na interação entre o complexo receptor-ligante. Para solucionar esse problema, novos métodos são propostos para agrupar as conformações baseados nas características da cavidade de ligação do substrato de cada receptor [QDPRNdS14, DPQRdNdS15]. Este capítulo descreve três métodos para identificar conjuntos de conformações similares no modelo FFR de 19,5 ns da enzima InhA de *Mtb*. Os dois primeiros métodos descritos neste capítulo são apresentados em maiores detalhes em dois artigos já publicados, sendo um artigo publicado na revista *Expert Systems With Aplications* [QDPRNdS14] e o outro artigo publicado na revista *PloS one* [DPQRdNdS15].

A seção 5.1 apresenta um método de agrupamento de conformações baseado na análise de 4 propriedades da cavidade de ligação do substrato. Esse agrupamento foi o primeiro estudo a buscar por grupos bem separados e compactos em relação aos grupos formados somente a partir dos valores do RMSD. A seção 5.2 define o segundo método de agrupamento de conformações baseado na análise de 12 propriedades da cavidade de ligação do substrato. Ao final, esse método selecionou 192 estruturas representativas comparando-as com outros dois conjuntos de estruturas representativas geradas a partir dos valores de RMSD da estrutura inteira e da cavidade de ligação do substrato. A seção 5.3 apresenta um método de seleção de estruturas similares baseado em vetores de características físico-químicas da cavidade de ligação do substrato. Essas seções descrevem as características estruturais extraídas do modelo FFR consideradas para formar o arquivo de entrada do algoritmo de agrupamento, os métodos estatísticos utilizados para encontrar o número de grupos ideal e, por fim, uma avaliação de cada agrupamento gerado.

5.1 Agrupamento baseado na análise de 4 propriedades da cavidade de ligação do substrato.

O agrupamento descrito nesta seção apresenta um método de agrupamento de conformações baseado na análise de 4 propriedades da cavidade de ligação do substrato. Neste primeiro estudo, a intenção é identificar conjuntos de conformações similares, avaliando se os grupos gerados são mais separados e mais compactos em relação aos grupos formados somente a partir dos valores do RMSD.

5.1.1 Propriedades estruturais da cavidade de ligação do substrato.

As propriedades estruturais da cavidade de ligação do substrato de cada conformação foram extraídas das 19.500 conformações que compõem o modelo FFR. Esse conjunto de dados contém 4 atributos das propriedades estruturais da cavidade de ligação do substrato, sendo 1 atributo gerado pela ferramenta *ptraj* [CDCI$^+$12] e os outros 3 atributos capturados pelo programa CASTp [BNL03]. Os valores do RMSD foram calculados comparando a primeira conformação com as estruturas do modelo FFR, usando o módulo *ptraj* pertencente ao pacote AMBER 12 [CDCI$^+$12]. Os outros atributos foram obtidos utilizando o programa CASTp. Esse programa foi usado para detectar a área de superfície acessível ao solvente, o volume da cavidade de ligação do substrato e identifica os átomos que determinam essa cavidade [BNL03]. As características estruturais extraídas de cada cavidade de ligação do substrato foram:

1. a área de superfície acessível na cavidade de ligação do substrato (em Å^2);
2. o volume da cavidade de ligação do substrato (em Å^3); e
3. a quantidade de átomos pesados da enzima 1BVR [RVS$^+$99] presentes na cavidade de ligação do substrato para cada conformação.

O CASTp fornece uma lista de todas as cavidades existentes na proteína, sendo necessário desenvolver um método externo para identificar a cavidade de ligação do substrato correta em cada conformação do modelo FFR. Então, para identificar as cavidades corretas, foi desenvolvido um método heurístico [DPQR$^+$15] com base no número de átomos no interior da cavidade de ligação ao substrato da enzima 1BVR [RVS$^+$99]. O análogo de substrato, ligante situado dentro da estrutura cristalográfica 1BVR, permitiu identificar a cavidade de ligação do substrato e os átomos que a constituem. A Figura 5.1 mostra a cavidade de ligação do substrato da enzima 1BVR identificada pelo programa CASTp.

Um programa foi desenvolvido para submeter os arquivos das 19.500 conformações do modelo FFR ao sítio web do CASTp e identificar a cavidade de ligação do substrato

em cada conformação. Após essa identificação, o conjunto de dados com as propriedades da cavidade de ligação do substrato está pronto para ser submetido ao algoritmo de agrupamento. Uma análise da representação dos dados a serem agrupados mostrou características adequadas para a utilização do algoritmo *k*-means. A quantidade limitada de grupos foi definida com base nos experimentos anteriores realizados com o padrão de dados chamado Padrão Múltiplas Instâncias Autoadaptáveis (P-SaMI - do inglês *Self-adaptive Multiple Instances*) [Hüb10, HRFNdS15] que necessita, como dados de entrada, um conjunto de agrupamentos do modelo FFR variando de 2 até 15 grupos.

No entanto, o *k*-means é um algoritmo de agrupamento que necessita previamente da quantidade de grupos a serem formados como parâmetro [HW79, Llo82]. Desta forma, é necessária a aplicação de métodos estatísticos para determinar o número de grupos mais adequado para o conjunto de dados a ser avaliado. A próxima seção descreve três métodos estatísticos, bastante sedimentados na literatura, que foram utilizados neste estudo para avaliar esse conjunto de dados variando o valor de *k* entre 2 e 15.

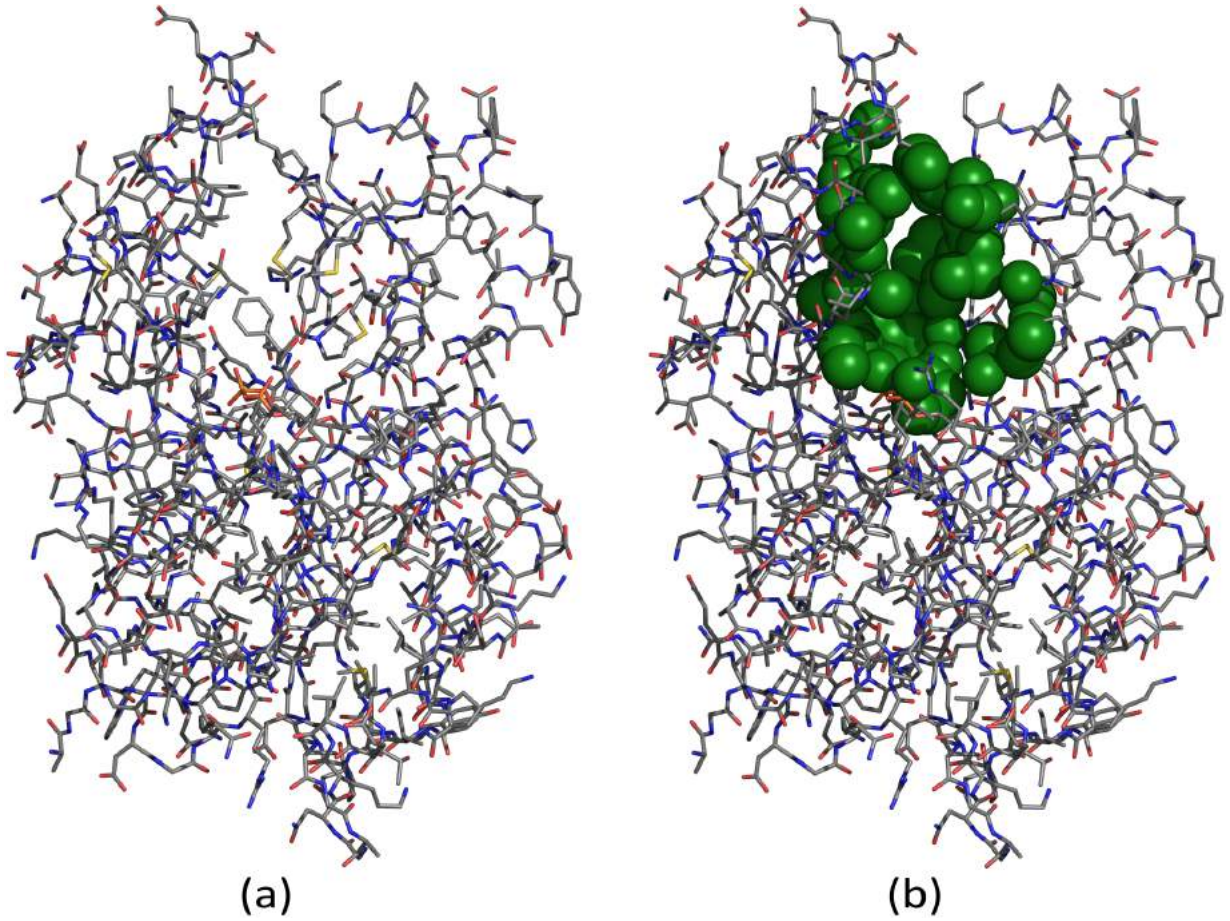


Figura 5.1 – Cavidade de ligação do substrato da enzima InhA de *Mycobacterium tuberculosis* (PDB ID: 1BVR) identificada pelo programa CASTp. Proteína 1BVR usando a representação de palitos colorida pelo tipo de átomo (carbono em cinza, nitrogênio em azul, oxigênio em vermelho, enxofre em amarelo e o fósforo em laranja). (a) Cadeia A da estrutura cristalina 1BVR que foi submetida ao programa CASTp. (b) Em verde, a cavidade de ligação do substrato da enzima 1BVR representada pelas esferas de van der Waals.

5.1.2 Medidas de validação de agrupamento para estimar o número de grupos.

As medidas de validação usadas para avaliar a qualidade dos grupos geradas neste estudo são o índice Davies-Bouldin (DB) [DB79], índice Dunn [Dun73] e o *gap* estatístico [TWH01]. Essas medidas têm se mostrado importantes estratégias para avaliar a qualidade de agrupamentos, especialmente quando elas são utilizadas em conjunto com uma inspeção visual dos grupos gerados [HBV02, STTC07]. O índice DB (Equação 5.1) representa a similaridade média entre cada um dos grupos do conjunto de dados com o grupo mais semelhante correspondente. Esse índice é a razão entre a soma da dispersão intraclasses e a separação interclasses, definido como:

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq i}\left\{D_{i,j}\right\}, D_{i,j} = \frac{\bar{d}_i + \bar{d}_j}{d_{i,j}} \tag{5.1}$$

onde *k* é o número de grupos, $D_{i,j}$ é a distância da dispersão entre os grupos i^{th} e j^{th}, $\bar{d}_i$ e $\bar{d}_j$ são as distâncias médias entre cada ponto do grupo com o centroide do agrupamento e $d_{i,j}$ é a distância entre o centroide do grupo i^{th} e j^{th}. O objetivo é minimizar a dispersão e maximizar a separação entre os grupos, assim, o número de grupos *k* que minimiza a Equação 5.1 é definido como o valor mais adequado para este conjunto de dados.

Da mesma forma que o índice DB, o índice Dunn também identifica a melhor partição baseada em aspectos geométricos sobre as maiores distâncias entre os grupos e compactação dentro do grupo. As partições que apresentam seus dados bem compactados e separados recebem valores elevados de Dunn, tal como indicado na seguinte equação:

$$D_n = \min_{1\leqslant i\leqslant n}\left\{\min_{1\leqslant j\leqslant n, j\neq i}\left\{\frac{\delta(C_i, C_j)}{\max\limits_{1\leqslant k\leqslant n} diam(C_k)}\right\}\right\} \tag{5.2}$$

onde $\delta(C_i, C_j)$ é a diferença entre os grupos C_i e C_j, obtido pelo cálculo do conjunto das distâncias intergrupos entre os grupos C_i e C_j. $diam(C_k)$ é o diâmetro máximo entre os pares de elementos do grupo k^{th}. Desta forma, se o conjunto de dados contém grupos compactos e bem separados, a separação entre os grupos deve apresentar valores altos enquanto que o diâmetro dos grupos deve ser pequeno. Assim, valores que maximizem o índice Dunn indicam grupos compactos e bem separados.

A estatística *gap* [TWH01] é baseada na comparação da soma do quadrado das distâncias dentro do grupo de uma dada partição com um grupo obtido a partir de dados aleatórios. Essa estatística é um procedimento para estimar o número de grupos para um conjunto de dados, a qual compara as alterações na dispersão dentro de cada grupo com o valor esperado sob uma distribuição nula adequada utilizada como referência:

$$Gap_n(k) = E_n^* \left\{ \log(W_k) \right\} - \log(W_k) \tag{5.3}$$

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

onde E_n^* representa o valor esperado para uma amostra de tamanho n, a partir da distribuição de referência, *n* é o tamanho da amostra, *k* é a quantidade de grupos sendo avaliados, W_k é a dispersão agrupada dentro do grupo, n_r é o número de objetos de dados no grupo *r* e D_r é a soma das distâncias entre pares para todos os objetos no grupo *r*. A estatística *gap* é calculada para grupos com diferentes valores de *k* e a estimativa do número de grupos é o valor de *k* para o qual a diferença entre $\log(W_k)$ e a curva de referência é maior.

Considerando que os índices DB e Dunn têm como objetivo identificar grupos compactos e bem separados, a estatística *gap* tende a estimar o número adequado de grupos com base na dispersão desses grupos. Portanto, uma partição ideal deve fornecer altos valores para o índice Dunn e a estatística *gap* e um baixo valor para o índice DB.

Após definir os índices de validação a serem mensurados, o procedimento para identificar a partição ideal de acordo com os índices de validade do agrupamento é dividido em três passos. Primeiro, o conjunto de dados de entrada é criado para o algoritmo de agrupamento. Então, o algoritmo *k*-means é executado com o valor de k variando de 2 até 15 grupos. Finalmente, se identifica o agrupamento mais adequado usando os índices DB [DB79], Dunn [HBV02] e a estatística *gap* [TWH01].

Conforme descrito na Seção 5.1.1, foram extraídas as propriedades estruturais da cavidade de ligação do substrato de cada conformação do receptor que compõe o modelo FFR. A área, o volume, o RMSD e a pontuação de átomos pesados para os 19.500 conformações foram colocados em um arquivo CSV. Esse conjunto de dados compreende atributos com diferentes unidades e escalas. Desta forma, todos os valores foram normalizados em uma escala dentro do intervalo [0,1] antes de executar o algoritmo *k*-means. O arquivo CSV com os dados normalizados foi submetido ao algoritmo *k*-means com *k* variando de 2 a 15 de grupos.

A fim de avaliar a qualidade da partição formada pelo algoritmo *k*-means e identificar qual a melhor solução, o conjunto de índices DB, Dunn e estatística *gap* são calculados para as partições com números distintos de grupos. Partições que fornecem um valor baixo para o índice DB e valores altos para a estatística *gap* e índice Dunn indicam ser o melhor agrupamento. As Figura 5.2-a e Figura 5.2-b mostram que o *k*-means gera uma partição que mostra um valor máximo para a estatística *gap* e índice Dunn, quando o agrupamento de dados possui 10 grupos. No entanto, o índice DB (Figura 5.2-c) mostra uma pequena preferência por 11 grupos (ao invés de 10).

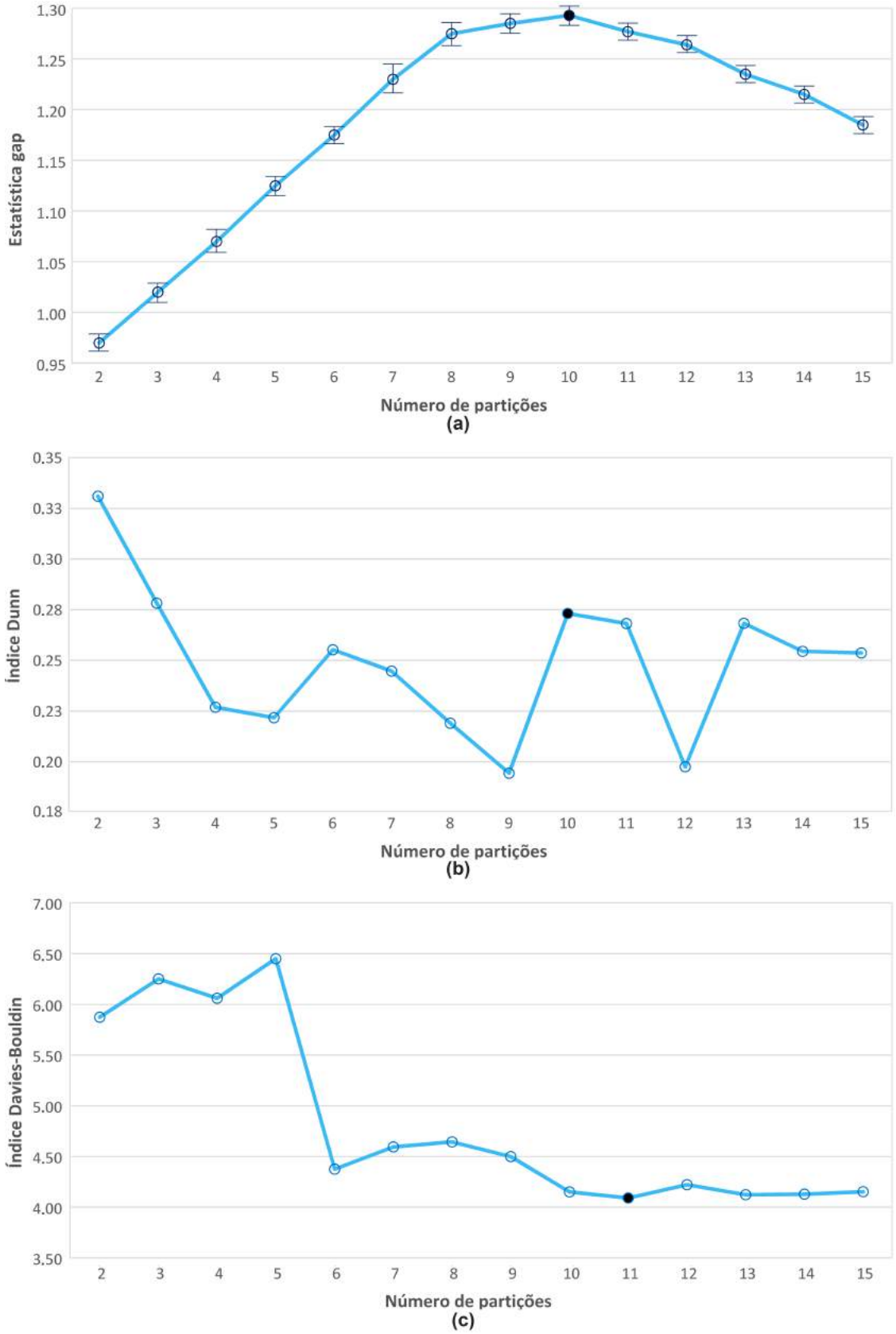


Figura 5.2 – Gráficos das avaliações dos índices DB, Dunn e estatística *gap* identificando o valor de *k* mais adequado considerando o modelo FFR. (a) Estatística *gap*. (b) Índice Dunn. (c) Índice DB. Os círculos pretos identificam o melhor número de grupos para cada índice. A estatística *gap* foi utilizada como critério de desempate para selecionar entre *k=10* e *k=11*, como sugerido pelos índices Dunn e DB, respectivamente.

O índice de Dunn também indica que a partição formada por dois grupos é uma boa solução. No entanto, esta mesma partição é mal avaliada pelo índice DB e pela estatística *gap*. Assim, a estatística *gap* foi utilizada como critério decisivo para as duas partições sugeridas como ideal pelo índice DB, ou seja, *k* com 10 ou 11 grupos. Seguindo esta estratégia, a partição com 10 grupos é selecionada considerando que os valores do índice Dunn e da estatística *gap* são maiores do que a partição com 11 grupos.

Para ilustrar a partição ideal do *k*-means, a Figura 5.3 mostra o grupo atribuído a cada conformação do modelo FFR com base em suas características estruturais. A variabilidade do resultado do agrupamento é fortemente influenciado pelas mudanças estruturais na cavidade de ligação de substrato para determinar a similaridade das configurações moleculares diferentes. Diferentemente dos tradicionais resultados de agrupamento baseados no RMSD, a Figura 5.3 apresenta uma distribuição bastante heterogênea dos grupos em relação aos valores de RMSD durante as diferentes escalas de tempo da trajetória [STTC07, PCN11].

Diversos testes foram realizados no intuito de descobrir as relações entre as distâncias do RMSD pareadas e os valores FEB das conformações desse estudo. No entanto, nenhuma relação satisfatória foi perceptível entre eles, provavelmente devido ao fato de o RMSD abstrair todas as características importantes que influenciam significativamente o valor de FEB. A distribuição na Figura 5.3 descreve a identificação de cavidades semelhantes em diferentes escalas de tempo ao longo do modelo FFR. Na próxima seção será apresentado um estudo sobre a relação entre do agrupamento encontrado e os valores de FEB obtidos a partir de experimentos de docagem molecular.

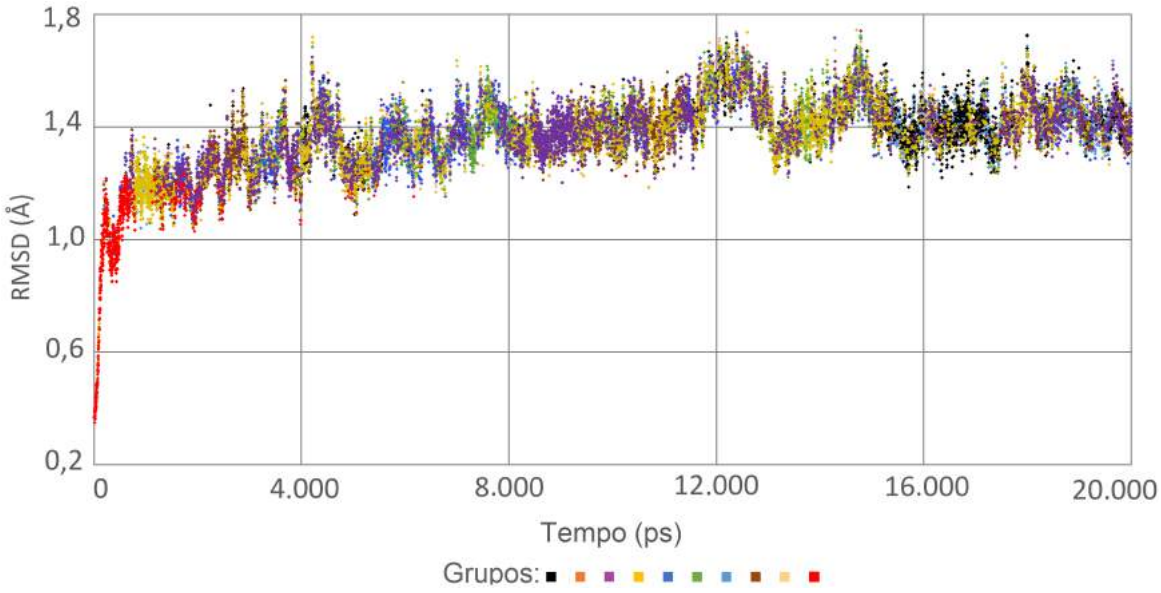


Figura 5.3 – Gráfico de dispersão mostrando a distribuição dos grupos atribuídos para cada conformação do modelo FFR da enzima InhA de 19,5 ns. Cada ponto representa um valor de RMSD em função do tempo correspondente a geração da conformação na trajetória e a cor identifica a qual grupo pertence.

5.1.3 Análise do agrupamento gerado.

A análise dos 10 grupos gerados na seção anterior é realizada a partir dos resultados das execuções de docagem molecular entre o modelo FFR e o conjunto de 20 ligantes descritos na seção 4.1.1. Os valores da FEB e do RMSD foram extraídos das execuções de docagem molecular de todas as conformações do modelo FFR. As análises realizadas consideram os valores da mediana da FEB e do RMSD de cada conjunto de conformações do agrupamento gerado[1].

A Figura 5.4 mostra a variação dos valores das medianas obtidos para cada grupo considerando os valores da FEB. A linha vermelha identifica o grupo que apresentou os melhores valores da mediana da FEB em todos os ligantes testados. Esse resultado permite concluir que o agrupamento foi capaz de detectar um padrão de similaridade para representar os melhores resultados de docagem molecular dos 20 experimentos testados. A distribuição dos agrupamentos é outro fator representado na Figura 5.4. Em grande parte dos experimentos pode-se verificar uma separação significativa entre o primeiro e o segundo melhores resultados da mediana da FEB. Esse comportamento apenas não ocorreu com os ligantes JPM e TCL (1P45) (cerca de 10% dos ligantes avaliados).

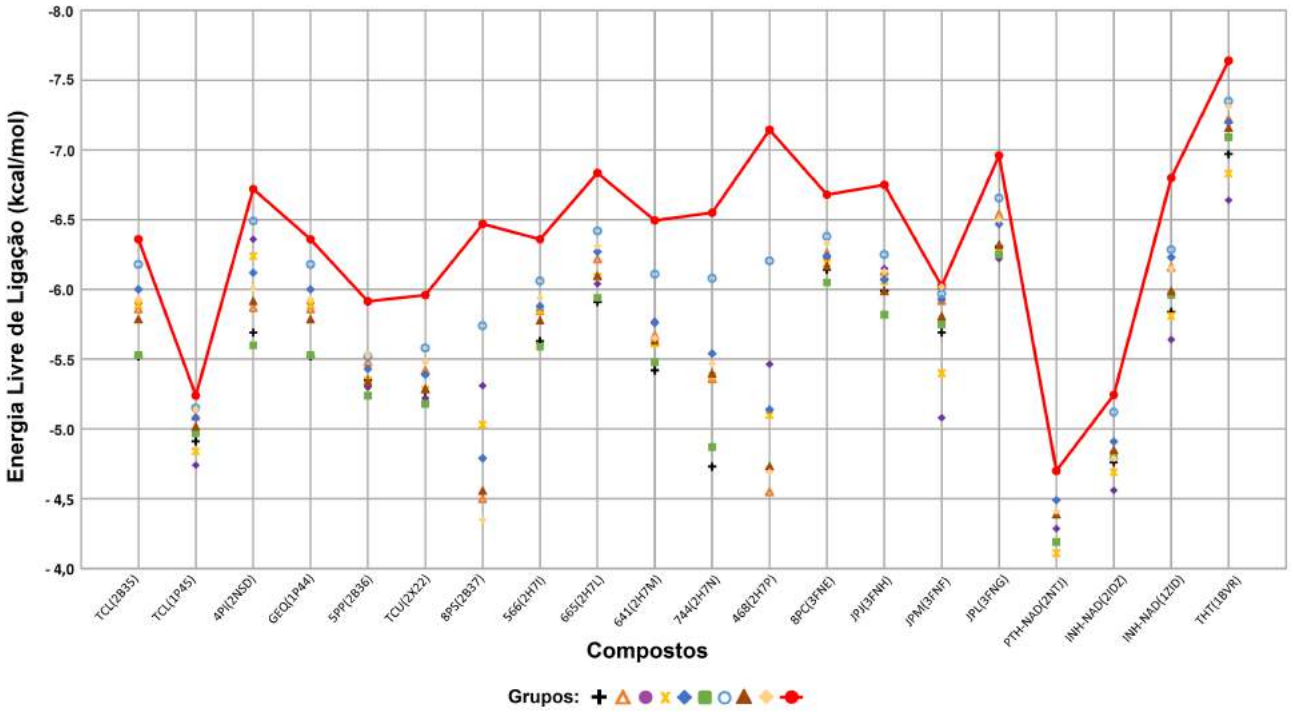


Figura 5.4 – Avaliação dos experimentos de docagem molecular entre o conjunto de 20 ligantes e o modelo FFR de 19,5 ns mostrando os valores das medianas da FEB de cada grupo. O círculo vermelho representa o grupo com os melhores valores para cada ligante. A linha vermelha interliga os resultados do grupo de círculo vermelho, destacando que este grupo obteve o maior valor da mediana da FEB em todos os experimentos.

[1] A mediana é uma medida resistente a valores *outliers* que pode ocorrer nas análises de FEB e do RMSD de casos de docagem molecular, principalmente em casos onde há uma colisão da estrutura a ser docada.

A análise realizada com a medida do RMSD segue o mesmo critério da avaliação da FEB, sendo calculada a mediana dos valores do RMSD com relação a estrutura de referência para cada grupo. A Figura 5.5 mostra a variação dos valores obtidos na análise do RMSD para cada grupo[2]. A linha vermelha identifica o grupo que apresentou as menores distâncias com relação as respectivas estruturas de referência. A análise dos valores do RMSD mostra que o grupo com as menores distâncias de RMSD para cada estrutura de referência em 90% dos casos é o mesmo conjunto que obteve os maiores valores de FEB apresentados na Figura 5.4. Nota-se, também, a existência desta mesma correlação dos melhores valores de FEB x menor distância do RSMD nos demais grupos analisados.

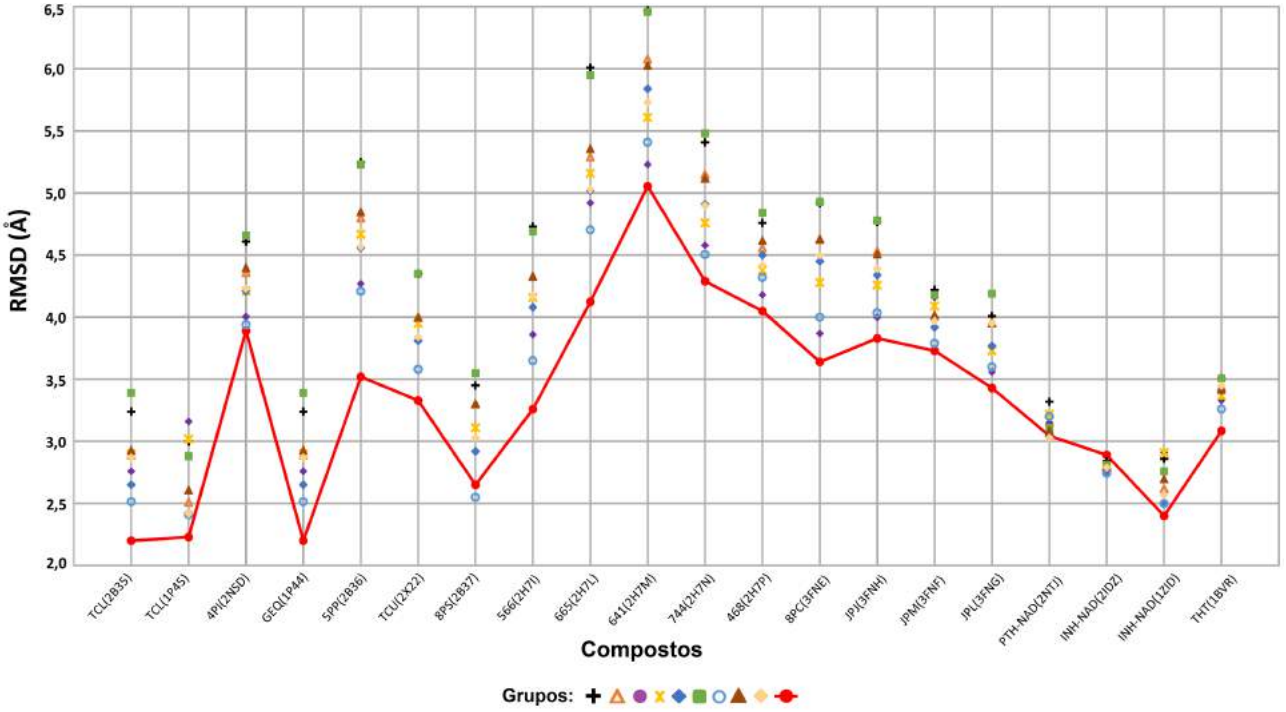


Figura 5.5 – Avaliação dos experimentos de docagem molecular entre o conjunto de 20 ligantes e o modelo FFR de 19,5 ns mostrando os valores das medianas do RMSD de cada grupo. O círculo vermelho representa o grupo com os menores valores da mediana do RMSD para cada experimento. A linha vermelha interliga os resultados do grupo de círculo vermelho, destacando que este grupo obteve a menor distância da estrutura de referência em 90% dos casos.

Apenas duas estruturas do conjunto de experimentos apresentaram resultados diferentes dos demais resultados, o aduto INH-NADH e o ligante 8PS. Esse aduto foi obtido da estrutura cristalina 2IDZ e é um dos fármacos de primeira ordem. Um ponto importante refere-se aos resultados obtidos da docagem molecular de um aduto similar da proteína 1ZID. Os resultados da 1ZID apresentam um comportamento idêntico aos demais valores encontrados nos demais experimentos. O ligante 8PS tem uma posição da estrutura de referência bastante crítica, visto que nossas análises mostraram que a Tyrosina 158 ocupa o espaço ideal de ancoragem deste ligante flexível.

[2]Os marcadores idênticos utilizados na Figura 5.4 e na Figura 5.5 representam o mesmo grupo.

De modo geral, o resultado destas análises mostra que o agrupamento foi capaz de detectar um padrão de similaridade para representar os melhores resultados de docagem molecular dos 20 experimentos testados. A distribuição dos agrupamentos é outro fator representado na Figura 5.5. Em grande parte dos experimentos pode-se verificar uma separação significativa entre o primeiro e o segundo melhores resultados. Esse comportamento apenas não ocorreu na avaliação do aduto da estrutura 2IDZ.

Uma análise inicial desses resultados aponta para a possibilidade de reduzir consideravelmente o tempo necessário para realizar os experimentos de *cross docking* com o modelo FFR. No entanto, não é sensato considerar que um único grupo irá representar o comportamento da flexibilidade existente em todo o modelo FFR. Uma estratégia opcional para aplicar a triagem virtual de BD de ligantes pode ser a seleção de uma amostragem de conformações de cada grupo ou aumentar a quantidade de grupos para representar o modelo FFR. Assim, as mudanças de comportamento dos grupos para cada composto podem ser detectadas com maior facilidade, bem como as análises das conformações do modelo FFR podem ser feitas de forma mais precisa.

Embora o estudo apresentado nesta seção tenha mostrado bons resultados, a seleção de uma quantidade baixa de estruturas representativas implica diretamente na perda de informações do modelo flexível. Além disso, o atributo que contém a quantidade de átomos pesados da enzima 1BVR no conjunto de entrada pode ser melhor detalhado para possibilitar uma avaliação mais precisa da cavidade de ligação do substrato. A próxima seção apresenta um conjunto de dados com informações mais específicas da cavidade de ligação do substrato e uma avaliação mais abrangente do número ideal de conformações representativas do modelo FFR.

5.2 Agrupamento baseado na análise de 12 propriedades da cavidade de ligação do substrato.

Esta seção apresenta a seleção de um conjunto de estruturas representativas do modelo FFR baseado na análise de 12 propriedades da cavidade de ligação do substrato de um modelo FFR. Essa limitação da quantidade de estruturas é imposta devido ao tempo de pré-processamento e avaliação necessários para agrupar e selecionar as conformações representativas. Não existe um valor exato da quantidade de grupos existentes em um modelo FFR, devendo haver uma análise considerando o equilíbrio entre a redução do tempo necessário ao selecionar estruturas representativas e a perda relacionada a redução da quantidade de conformações a serem avaliadas.

As medidas de validação de agrupamento para estimar o número de grupos (apresentadas na seção 5.1.2) apresentaram problemas relacionados ao uso e memória nas avaliações considerando mais 150 grupos de um conjunto de dados composto por 12 atributos e 19.500 instâncias. Uma forma de comparar a qualidade do agrupamento e também

identificar o número de grupos mais adequado para esse conjunto de dados é aplicar uma medida de avaliação dos resultados de docagem molecular. Diversos algoritmos de agrupamento foram utilizados para avaliar o conjunto de dados proposto nesta seção comparando-o com mais outros dois conjuntos de dados formados pelo RMSD da estrutura inteira e o outro pelo RMSD somente da cavidade de ligação do substrato. A próxima seção descreve a composição de cada um desses conjuntos de dados.

5.2.1 Propriedades estruturais da cavidade de ligação do substrato.

Esta subseção descreve o conjunto de dados baseado nas propriedades da cavidade de ligação do substrato e outros dois conjuntos de dados baseados somente na avaliação do RMSD. Um obtendo seus valores a partir do cálculo de toda a estrutura e o outro considerando apenas os resíduos da cavidade de ligação do substrato:

1. RMSD da proteína: possui a distância do RMSD de cada átomo pareado entre a primeira conformação e com cada conformação do modelo FFR considerando todos os resíduos da estrutura, da mesma forma como aplicado por [STTC07, LZ06, ZS04].

2. RMSD da cavidade de ligação do substrato: possui a distância do RMSD de cada átomo pareado entre a primeira conformação e com cada conformação do modelo FFR considerando apenas os resíduos que delimitam a cavidade de ligação ao substrato da enzima InhA em complexo com o NADH. Exemplos de aplicação dessa medida de similaridade estão em [LAB$^+$08, LWWZ08].

3. Atributos da Cavidade: possui um conjunto de características extraídas a partir da cavidade de ligação de substrato do modelo FFR. Esse é o conjunto de dados proposto e uma explicação mais detalhada é dada nesta seção.

Os dois primeiros conjuntos de dados foram gerados a partir de medidas típicas de similaridade utilizadas atualmente para definir o conjunto de estruturas representativas [PMF$^+$09, FPSB11, ABE$^+$13, DABR$^+$13] com a ferramenta *ptraj* [CDCI$^+$12]. O terceiro conjunto de dados a ser avaliado nesta subseção possui características similares ao conjunto descrito na Seção 5.1.1. As diferenças envolvem particularmente dois atributos, a área acessível ao solvente e a quantidade de átomos pesados da enzima 1BVR que delimitam a cavidade de ligação do substrato. O atributo mensurando a área acessível ao solvente é altamente correlacionado com o atributo do volume da cavidade de ligação do substrato, sendo então o atributo referente a área acessível ao solvente descartado a fim de evitar informações redundantes. O atributo da quantidade de átomos pesados da enzima 1BVR que delimitam a cavidade de ligação do substrato suprime muitas informações importantes ao não discriminar quais são os resíduos que definem as propriedades físico-químicas da cavidade avaliada.

Esses fatores permitem caracterizar mais especificamente os diferentes comportamentos encontrados no sítio de ligação ao longo do modelo FFR, permitindo a identificação de um conjunto de estruturas representativas capazes de cobrir os movimentos localizados das proteínas para melhorar o encaixe dos ligantes durante as execuções de docagem molecular. Assim, um novo conjunto de dados com 12 atributos das propriedades estruturais da cavidade de ligação do substrato foram extraídos pelo programa CASTp [BNL03] e pela ferramenta *ptraj* [CDCI+12]:

1. o valor do RMSD da cavidade de ligação do substrato da primeira estrutura do modelo FFR comparado com cada conformação do modelo FFR (em Å).

2. o volume da cavidade de ligação do substrato (em $Å^3$); e

3. dez atributos avaliando cada resíduo (GLY96, PHE97, MET98, MET103, PHE149, TYR158, MET161, LYS165, MET199 e coenzima NAD) da enzima 1BVR [RVS+99] presentes na cavidade de ligação do substrato para cada conformação. Cada atributo possui a quantidade de átomos pesados e a quantidade máxima de átomos pesados depende do resíduo avaliado.

Esse conjunto de dados é obtido de forma similar à descrita na seção 5.1.1. A Figura 5.6 mostra a cavidade de substrato da estrutura 1BVR identificada pelo programa CASTp, juntamente com os resíduos que delimitam a cavidade de ligação do substrato. O CASTp informa quais são os átomos pesados de cada resíduo que estão presentes na cavidade alvo de cada conformação do modelo FFR. A lista de resíduos identificadas pelo CASTp para o modelo FFR de 19,5 ns é composta pelos seguintes resíduos: GLY96, PHE97, MET98, MET103, PHE149, TYR158, MET161, LYS165, MET199 e mais a coenzima NADH.

Uma caracterização da estrutura do arquivo de entrada dessa avaliação é apresentado na Tabela 5.1. Nessa tabela também é possível verificar diferentes conformações contendo valores de RMSD idênticos, enquanto a diferença das conformações somente é identificada pela diferença do volume e dos átomos pesados dos resíduos. Por exemplo, as duas primeiras e as duas últimas estruturas contêm valores idênticos de RMSD. No entanto, há uma diferença significativa entre o valor do volume e do número de átomos pesados para cada um dos resíduos. Os valores da Tabela 5.1 estão em diferentes escalas, sendo necessária uma etapa de pré-processamento para adequar esse conjunto de dados. Um arquivo CSV é criado com esses dados normalizados em um intervalo de [0,1], visando preservar as distâncias entre os objetos considerando o mesmo atributo e tornando equivalente as grandezas entre atributos distintos.

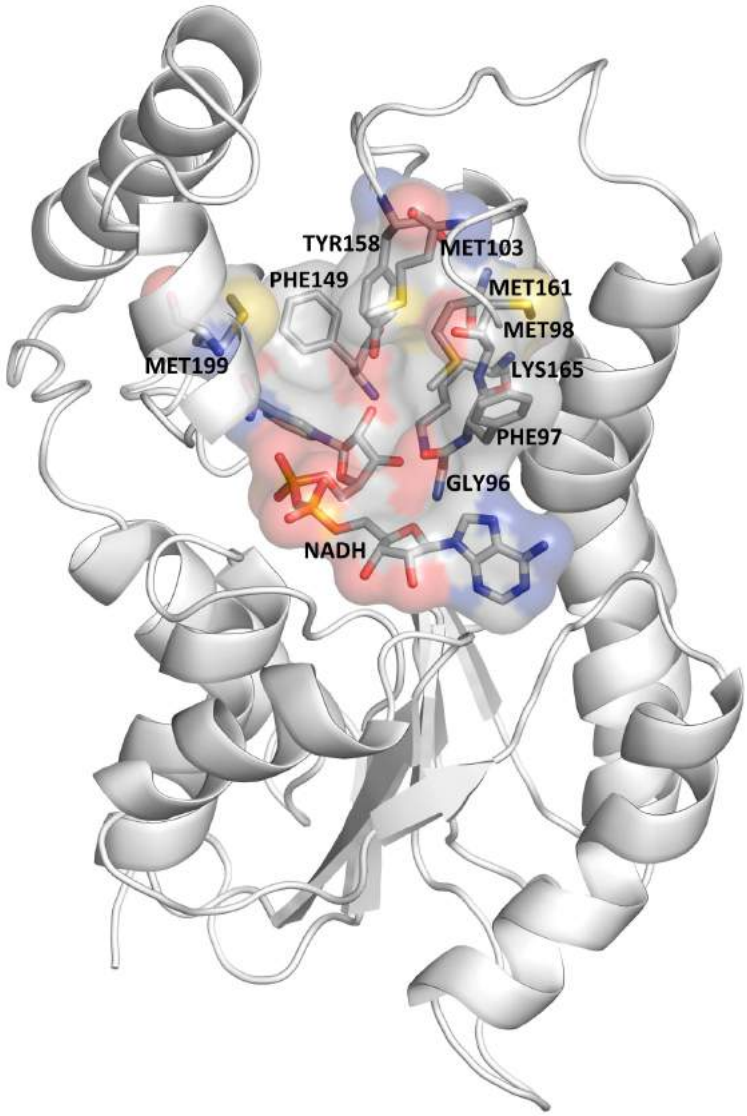


Figura 5.6 – Cavidade de ligação do substrato da enzima InhA de *Mtb* (PDB ID: 1BVR) identificada pelo programa CASTp. Proteína 1BVR usando a representação de fitas na cor branca. Os resíduos da cavidade de ligação do substrato são representados em palitos e em superfície molecular coloridos pelo tipo de átomo (carbono em cinza, nitrogênio em azul, oxigênio em vermelho, enxofre em amarelo e o fósforo em laranja). Imagem gerada pelo programa PyMol [DeL02].

5.2.2 Medidas de validação de agrupamento

O conjunto de dados proposto, que utiliza as propriedades da cavidade de ligação ao substrato do modelo FFR, e os outros dois conjuntos de dados baseados nos valores do RMSD foram submetidos a seis algoritmos de agrupamentos. Para cada conjunto de dados, os algoritmos particionais *k*-means e *k*-medoid e os algoritmos hierárquicos aglomerativos Complete linkage, UPGMA, WPGMA e Ward's foram aplicados para gerar grupos de conformações do modelo FFR. A qualidade desses grupos foi avaliada com base no cálculo dos valores do primeiro, segundo e terceiro quartis da estrutura representativa de cada grupo, sendo que o escopo do número de grupos varia de 10 até 200.

Tabela 5.1 – Fragmento do conjunto de dados contendo as informações detalhadas da cavidade de ligação do substrato usados para o agrupamento do modelo FFR. A primeira linha descreve cada atributo do conjunto de dados com as informações da cavidade de ligação de substrato. O número entre parênteses abaixo de cada resíduo indica o número máximo de átomos pesados que o resíduo pode conter.

RMSD (Å)	Volume (Å^3)	G96 (4)	F97 (11)	M98 (8)	M103 (8)	F149 (11)	Y158 (12)	M161 (8)	K165 (9)	M199 (8)	NAD (9)
0,37	607,0	2	4	3	2	6	4	3	2	6	6
0,37	795,1	2	4	3	2	5	3	2	0	6	6
...	...	...	...	...	...	...	...	...	...	...	...
1,38	929,8	2	6	4	3	2	5	3	3	1	5
1,38	516,9	3	4	3	2	2	4	3	2	2	6

*Código de 1 letra: G-Glicina, F-Fenilalanina, M-Metionina, Y-Tirosina e K-Lisina

Os quartis são medidas estatísticas de tendência central robustas para avaliar a dispersão dos objetos de um conjunto de dados. Outra característica importante é que essas medidas também são resistentes a *outliers*. À medida que se procura identificar os grupos que contenham conformações do modelo FFR com alta afinidade no seu modo de ligação, a investigação avalia o desempenho dos grupos gerados pelos algoritmos de agrupamento considerando a interação dos ligantes/adutos candidatos a solução com a enzima InhA. Assim, a avaliação desses agrupamentos utilizou as informações dos experimentos de docagem molecular com os 20 compostos testados experimentalmente (Figuras 4.1 e 4.2). Esses ligantes/adutos possibilitam uma avaliação diversificada das interações da cavidade de ligação do substrato.

O conjunto de conformações representativas do modelo FFR foi selecionado com base nos valores da FEB preditas de cada grupo gerado pelos algoritmos de agrupamento. A dispersão é quantizada no primeiro, segundo e terceiro quartil dos valores da FEB. Esses valores são calculados para comparar o nível de convergência entre os grupos resultantes dos três conjuntos de dados e o modelo FFR. Primeiramente, uma busca é feita para identificar a estrutura representativa de cada grupo. Essa estrutura é definida como a conformação que está mais próxima do centroide de cada grupo, também conhecido na literatura como medoide. Assim, são identificados todos os medoides dos agrupamentos a serem avaliados. Após é realizado o cálculo dos valores dos três quartis dos conjuntos de medoides de cada agrupamento como segue:

$$SomaQ_q = \frac{1}{N}\sum_{i=1}^{N} x_{iq} \tag{5.4}$$

onde *N* é a quantidade de ligantes/adutos avaliados, *q* identifica o primeiro, segundo e terceiro quartis que foram calculados com base nos valores estimados da FEB obtidos a partir de experimentos de docagem molecular.

Depois de calcular os quartis de cada composto, é possível avaliar a Soma das Diferenças entre os Quartis (SDQ), a fim de identificar os conjuntos de medoides com a dispersão mais similar ao modelo FFR. Assim, a SDQ é calculada da seguinte forma:

$$SDQ = \sum_{i \in (1,2,3)} |(\overline{x_{qi}} - \overline{y_{qi}})| \tag{5.5}$$

onde *i* indica o quartil, x_{qi} é o valor do $SomaQ_i$ e y_{qi} o valor do quartil *i* dos valores de FEB do modelo FFR. A Equação 5.5 calcula a dissimilaridade entre os valores dos quartis dos agrupamentos e os valores do modelo FFR. Desta forma, baixos valores da Equação 5.5 indicam que o conjunto de conformações representativas do modelo FFR possui uma alta similaridade com os resultados obtidos com a execução exaustiva do modelo FFR.

5.2.3 Análise dos agrupamentos gerados.

Os agrupamentos gerados para cada um dos três conjuntos de dados foram analisados e comparados, considerando o nível de cobertura alcançada por eles em termos de dispersão e representatividade do modelo FFR. Para cada conjunto, o número de grupos gerados variaram de 10 até 200 grupos analisados pelos valores da SDQ (Equação 5.5). As Figuras 5.7 e 5.8 apresentam os desempenhos comparativos entre os conjuntos de dados avaliados pela SDQ nos agrupamentos gerados pelos métodos particionais e hierárquicos, respectivamente.

A Figura 5.7 mostra valores com grandes oscilações no valor da SDQ obtidos para os algoritmos de agrupamento para cada conjunto de dados. Apesar dos algoritmos *k*-means e *k*-medoid atingirem algumas vezes valores baixos para essa medida, os valores estatísticos calculados apresentaram uma grande quantidade de valores distantes daqueles obtidos para o modelo FFR. Por exemplo, as Figuras 5.7-(a) e 5.7-(b) mostram que os conjuntos de dados do Atributos da Cavidade e do RMSD da Cavidade possuem valores mais adequados que o conjunto de dados do RMSD da Proteína.

A Tabela 5.2 apresenta as avaliações estatísticas dos agrupamentos com os menores valores da medida SDQ para cada algoritmo. Essa tabela também mostra a diferença da variância e da média entre os melhores agrupamentos dos métodos particionais e os valores do modelo FFR, resultando na redução da representatividade do modelo FFR. Essa variação pode ser ocasionada pelas características do algoritmo de agrupamento, que neste caso específico, favorece dispersões com formas esféricas [HKP11]. Uma consequência prática dessa característica está representada na Figura 5.7. Essas linhas mostram uma grande oscilação nos valores, mostrando também pequenas diferenças nos valores SDQ entre os dados em relação ao número de grupos.

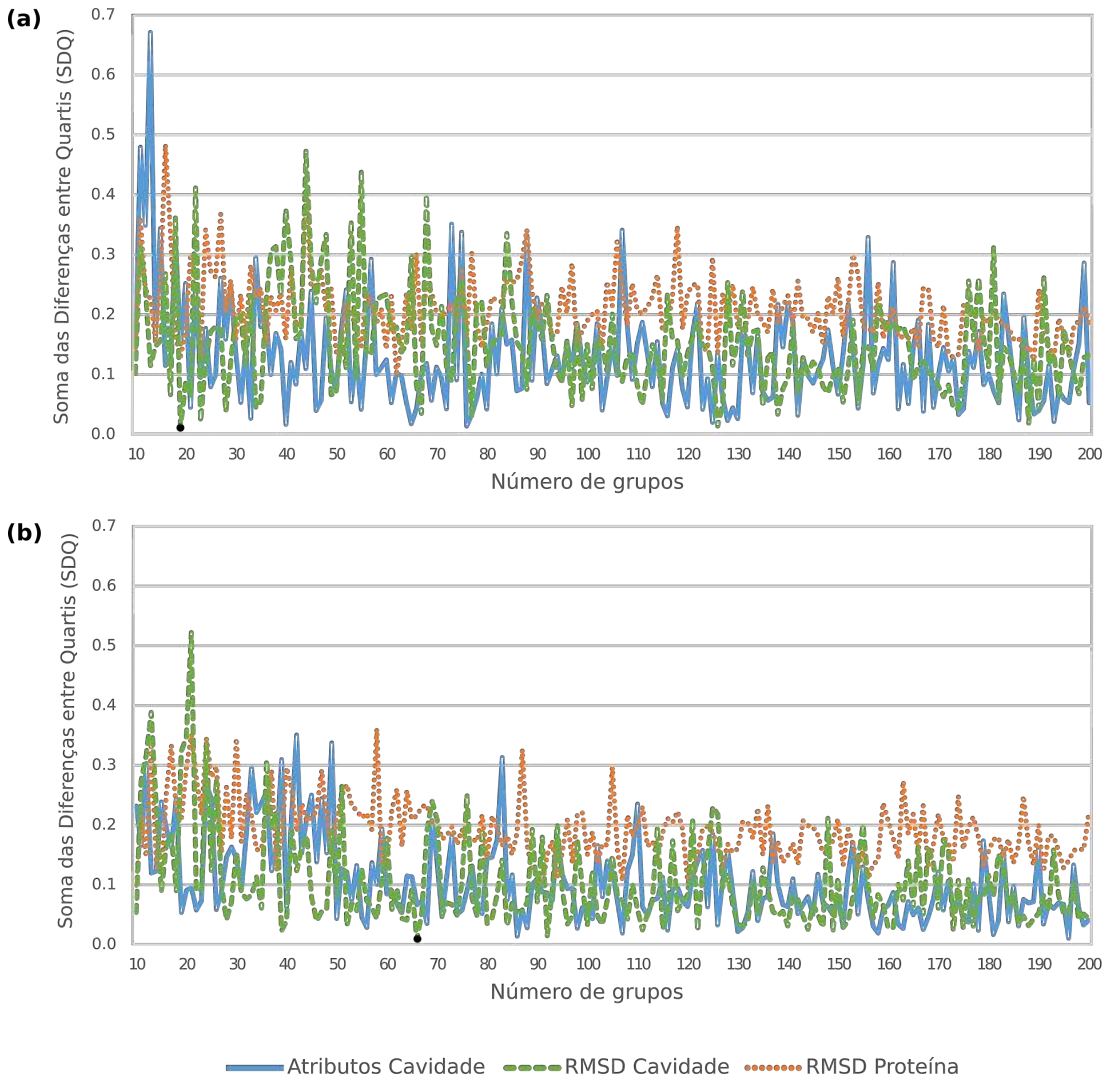


Figura 5.7 – Comparação do desempenho dos agrupamentos gerados pelos algoritmos particionados considerando os três conjuntos de dados em estudo. Os gráficos (a) e (b) mostram as variações dos valores SDQ em função do número de grupos para os algoritmos *k*-means e *k*-medoid, respectivamente. Os dois pontos pretos identificam as melhores soluções de particionamento para cada algoritmo.

Tabela 5.2 – Avaliação estatística dos melhores grupos (corresponde ao menor valor SDQ) de cada algoritmo de agrupamento. A terceira coluna indica a quantidade de grupos utilizados nas avaliações estatísticas. A média, o desvio padrão e a variância foram calculados para cada conjunto de grupos com base nos valores preditos da FEB. A primeira linha indica os valores estatísticos da avaliação de todo o modelo FFR.

Algoritmo	Conjunto de dados	Partições	SQD	Média	Desvio P.	Variância
-	modelo FFR	20	0,00	-6,58	-0,72	-0,57
k-means	RMSD Proteína	19	0,01	-6,61	-0,70	-0,54
k-medoid	RMSD Proteína	66	0,01	-6,63	-0,70	-0,55
UPGMA	Atributos Cavidade	133	0,04	-6,58	-0,72	-0,56
WPGMA	Atributos Cavidade	84	0,03	-6,59	-0,73	-0,58
Complete	Atributos Cavidade	48	0,01	-6,59	-0,69	-0,51
Ward's	Atributos Cavidade	95	0,01	-6,60	-0,68	-0,51

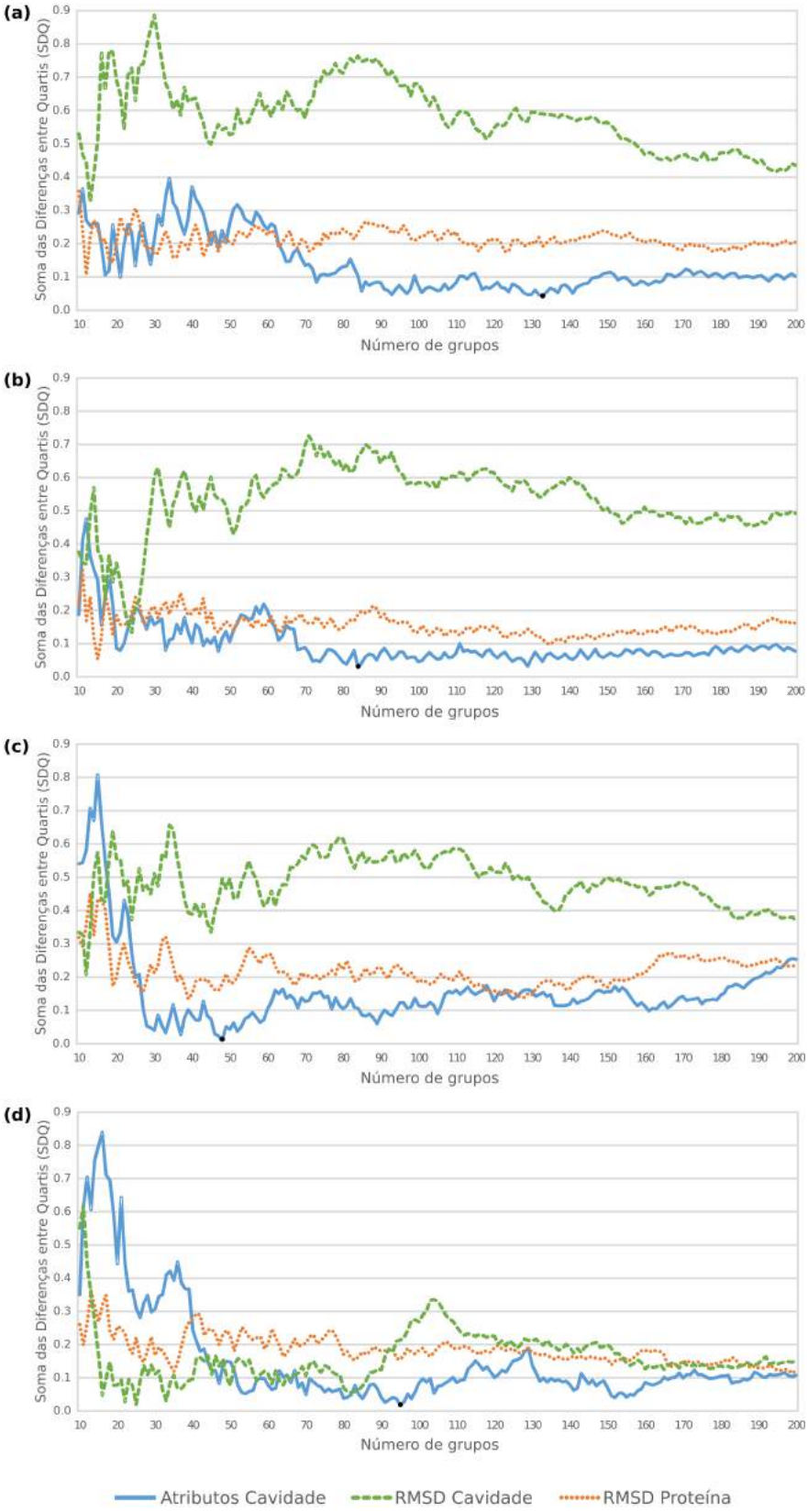


Figura 5.8 – Comparação do desempenho dos agrupamentos gerados pelos algoritmos hierárquicos considerando os três conjuntos de dados em estudo. Os gráficos (a), (b), (c) e (d) mostram as variações dos valores SDQ em função do número de grupos para os algoritmos UPGMA, WPGMA, Complete e Ward's, respectivamente. Os pontos pretos identificam as melhores soluções de particionamento para cada algoritmo.

Ao contrário dos métodos particionais, os algoritmos hierárquicos apresentaram melhores agrupamentos para todos os conjuntos de dados, segundo a SQD (Equação 5.5). Além disso, os valores obtidos pelos algoritmos hierárquicos também se mostraram mais estáveis que os algoritmos *k*-means e *k*-medoid. Assim como nos algoritmos particionais, os agrupamentos hierárquicos também apresentam resultados ruins quando um número pequeno de grupos é selecionado. Esse comportamento fica evidenciado na região inicial dos gráficos da Figura 5.8. Conforme esperado, os agrupamentos formados a partir do conjunto de dados dos Atributos da Cavidade determinam com uma maior precisão as alterações fundamentais que ocorrem na cavidade de ligação de substrato das conformações do modelo FFR em estudo. A Figura 5.8 representa esse conjunto de dados com uma linha azul clara, mostrando que este conjunto possui a maior parte dos menores valores da SDQ.

Embora a Figura 5.8 indique que os algoritmos UPGMA, WPGMA e Complete sejam bastante adequados para agrupar os conjuntos dos Atributos da Cavidade e do RMSD da Cavidade, isso não significa que ambos os conjuntos de dados possuem a melhor representatividade do modelo FFR, visto que os menores valores da SDQ são obtidos somente pelo conjunto de dados dos Atributos da Cavidade. Desta forma, os pontos pretos, que definem os melhores agrupamentos em ambas as Figuras 5.7 e 5.8, fazem parte da linha azul e estão distantes da linha que representa o conjunto de dados do RMSD da Cavidade. Analisando as Figuras 5.7 e 5.8 em conjunto com a Tabela 5.2, é possível constatar que os métodos de agrupamento hierárquico considerando o conjunto de dados dos Atributos da Cavidade superam os outros algoritmos de agrupamento e os outros conjuntos de dados, mostrando valores estatísticos semelhantes ao modelo FFR e também os menores valores da SDQ. Dentre os algoritmos hierárquicos, tanto o algoritmo Ward quanto o algoritmo Complete apresentaram uma boa representatividade do modelo FFR.

O algoritmo Ward é o que apresenta a média dos valores mais baixos da SDQ para os três conjuntos. Particularmente, os agrupamentos com 16, 22 e 25 grupos apresentam os melhores resultados com o conjunto de dados do RMSD da Proteína. Os valores da média e da variância dos medoids desses grupos são 6,62 e 0,51 para 16 grupos, 6,62 e 0,52 para 22 grupos, e por fim, 6,61 e 0,53 para 25 grupos. Semelhante aos métodos de particionamento, esses valores são bastante diferentes dos valores encontrados para o modelo FFR. Assim, esses agrupamentos não são capazes de selecionar um conjunto de conformações representativas. Além disso, o método de Ward utiliza todos os objetos do grupo no cálculo da medida de distância, favorecendo dispersões esféricas. Assim, da mesma forma que o algoritmo *k*-means, o Ward também é capaz de selecionar dispersões similares, mas incapaz de atingir a medida exata de tendência central para todo o modelo FFR. Essas avaliações indicam que o algoritmo Ward não contempla características essenciais para a seleção das estruturas representativas.

O algoritmo Complete tem a média e desvio padrão mais próximos dos valores obtidos pelo modelo FFR em relação à melhor solução indicada pelo algoritmo Ward. Dife-

rentemente do algoritmo Ward, este algoritmo determina a distância entre dois grupos de acordo com o par de objetos mais distantes intergrupos, resultando na seleção de grupos não tão alongados. Outro fator diferencial é que esse algoritmo identificou um agrupamento com 48 grupos com valores mais próximos do modelo FFR em estudo.

A Figura 5.9 apresenta um diagrama de caixa para comparar a distribuição de dados da solução ideal obtida das estruturas representativas do conjunto de dados com a distribuição do modelo FFR. Esse diagrama evidencia as diferenças entre os conjuntos de estruturas representativas geradas para os conjuntos de dados dos Atributos da Cavidade, do RMSD da Cavidade, do RMSD da Proteína e do modelo FFR. Os bigodes do diagrama são calculados com base na soma (assintótica superior) e na subtração (assintótica inferior) de 50% da diferença entre o 1º e o 3º quartis [DS13].

A Figura 5.9 mostra também a cobertura dos valores de docagem molecular, onde o conjunto dos Atributos da Cavidade mostrou-se muito semelhante à cobertura atingida pelo modelo FFR em comparação aos outros conjuntos. Uma análise por inspeção visual mostrou que as 48 conformações representativas do conjunto de dados dos Atributos da Cavidade possuem alterações significativas no interior da cavidade de ligação do substrato. Na avaliação da seleção de estruturas representativas, um conjunto de conformações com cavidades de ligação distintas entre si evidencia bons agrupamentos.

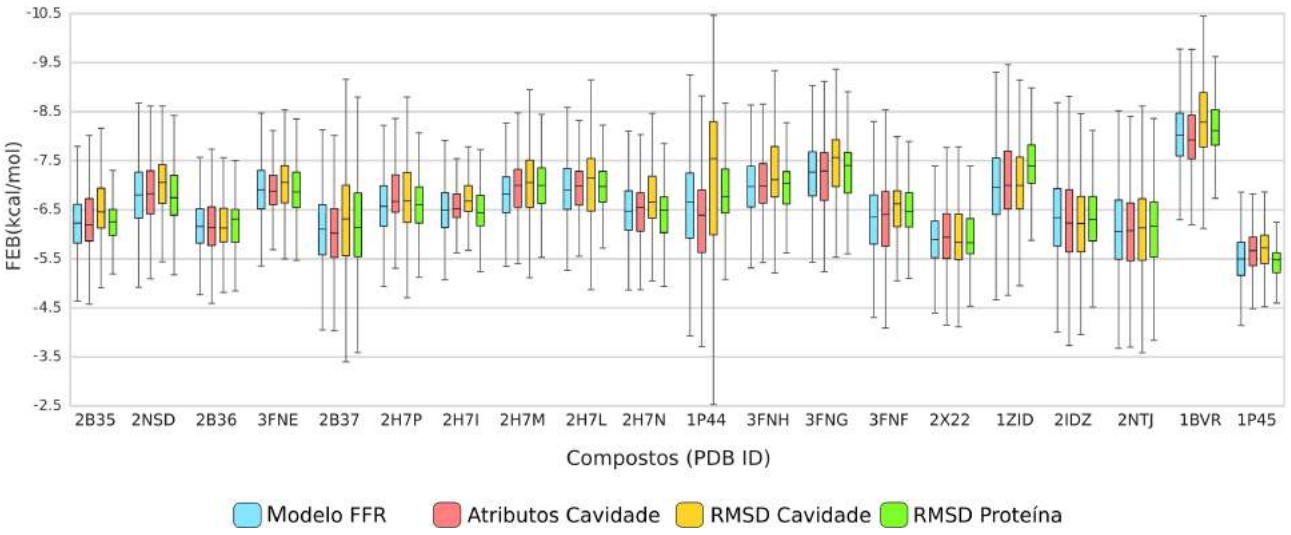


Figura 5.9 – Diagrama de caixa da dispersão da média dos valores de FEB do conjunto de 48 conformações representativas do modelo FFR categorizados pelos conjuntos de dados e compostos. Os conjuntos de dados são separados por cores diferentes, onde o azul claro, salmão, amarelo e verde representam os resultados dos conjuntos de dados do modelo FFR, Atributos da Cavidade, RMSD da Cavidade e RMSD da Proteína, respectivamente.

De acordo com os resultados apresentados, o conjunto de estruturas representativas do modelo FFR é capaz de representar os principais sítios de ligação da cavidade alvo em 75% dos ligantes testados considerando somente uma análise da Energia Livre de Ligação (FEB). Contudo, o nível da convergência desse conjunto representativo pode ser alterado dependendo dos compostos analisados. Por exemplo, a Figura 5.9 mostra uma

baixa convergência alcançada pelos ligantes das proteínas (PDB ID: 3FNE, 2H7P, 2H7I, 2H7L e 1P44) para o modelo FFR. Além disso, identificou-se que três dos 25% ligantes restantes contêm a mediana próxima do modelo FFR (PDB ID: 3FNE, 2H7I e 2H7L), onde um representa os melhores valores FEB (PDB ID: 2H7P) e o outro se sobrepõe ao pior valor de FEB (PDB ID: 1P44).

Até este momento, as análises deste agrupamento foram baseadas exclusivamente na avaliação da Energia Livre de Ligação (FEB). No entanto, assim como foi feito nos experimentos anteriores, é importante analisar a pose final predita pelo algoritmo de docagem molecular. Desta forma, um estudo comparativo foi realizado para analisar a variância dos valores do RMSD dos agrupamentos gerados com 48 grupos. Os 3 conjuntos de dados foram analisados: o conjunto baseado nos Atributos da Cavidade, o conjunto baseado no RMSD da Cavidade e o conjunto obtido a partir do RMSD da Proteína.

A avaliação do RMSD baseou-se na análise da variância dos valores de RMSD de cada estrutura para a média dos valores do RMSD existente em cada grupo[3]. A Figura 5.10 mostra o resultado deste estudo comparativo do RMSD dos conjuntos com os 48 grupos. O gráfico em barras representa a variância calculada para cada proteína. A representação em linhas apresenta a análise da variança média acumulada em cada experimento. Ao final desta avaliação, os valores resultantes da variança acumulada para os conjuntos dos Atributos da Cavidade, do RMSD da Cavidade e do RMSD da Proteína foram de 0,46 Å, 0,50 Å e 0,51 Å, respectivamente. Nesta figura, nota-se que os adutos possuem uma taxa de variância muito baixa com relação aos ligantes. Isto ocorre devido ao fato de haver uma ligação covalente entre o ligante e o anel da nicotinamida do NADH.

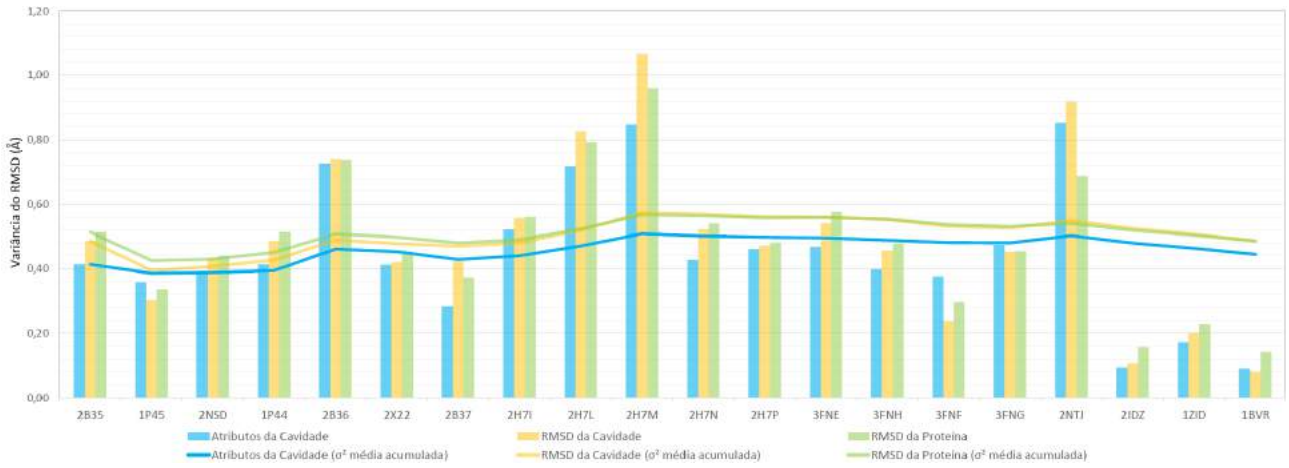


Figura 5.10 – Análise da variância dos valores do RMSD considerando 3 conjuntos de dados com 48 grupos. Os conjuntos de dados são separados por cores, onde o azul claro, amarelo e verde representam os resultados dos conjuntos de dados dos Atributos da Cavidade, RMSD da Cavidade e RMSD da Proteína, respectivamente. O gráfico em barras representa os valores da variância para cada proteína e o gráfico de linhas apresenta os valores médios acumulativos de cada conjunto.

[3]Nos casos envolvendo grupos únicos, é somado 1 para não zerar os pesos.

O conjunto de 48 estruturas representativas do modelo FFR apresentou uma considerável aproximação do modelo FFR para 75% dos ligantes obtidos das estruturas cristalinas do *Protein Data Bank*. Isso mostra a relevância de considerar propriedades mais específicas da cavidade de ligação de substrato além dos valores do RMSD. A avaliação do conjunto de Atributos da Cavidade pelo algoritmo hierárquico Complete mostrou-se capaz de reduzir a dimensão do modelo FFR a um tamanho gerenciável, mantendo as características mais relevantes para encontrar novos inibidores para a proteína em estudo.

Embora o estudo apresentado nesta seção tenha mostrado resultados interessantes, esperava-se obter conjuntos de estruturas representativas apresentando dispersões obtidas a partir de grupos bem separados e coesos. As diferenças entre os agrupamentos na Tabela 5.2 não apresentaram grandes variações. Essas variações podem ser melhor evidenciadas com uma preparação mais adequada do conjunto de dados fornecido. Por exemplo, a medida do RMSD que é utilizada nos três conjuntos de dados avaliados depende do tipo de alinhamento da estrutura. Ou seja, dependendo dos átomos considerados no alinhamento, o valor do RMSD pode variar. Outro fator relevante é melhorar as características do conjunto Atributos da Cavidade para definir mais propriedades da cavidade alvo. A próxima seção apresenta um novo agrupamento realizado com as propriedades físico-químicas da cavidade de ligação do substrato utilizando as triangularizações entre as propriedades farmacofóricas do receptor.

5.3 Agrupamento com vetores de propriedades farmacofóricas.

Recentemente, diversos métodos têm sido propostos para comparar as propriedades estruturais da cavidade de ligação do substrato buscando encontrar proteínas similares. Grande parte desses métodos avaliam as propriedades do conjunto de aminoácidos que compõem essa cavidade alvo e que possuem uma determinada distância dos átomos de ligantes cristalizados com a estrutura. Cada resíduo selecionado pode ser decomposto em um ou mais pontos farmacofóricos de interação molecular, dependendo especificamente dos átomos que compõem o resíduo avaliado [Cat00, SKK02, WR10]. Esse conjunto de pontos é, então, codificado em um vetor, onde cada posição representa a combinação de propriedades farmacofóricas. Basicamente, existem três formas de representar a ocorrência ou não das subestruturas de moléculas [SMdG07]:

- binária: cada posição do vetor descreve a presença ou a ausência da subestrutura.
- frequência: cada posição do vetor armazena a quantidade de ocorrências da subestrutura.
- frequência ponderada: cada posição do vetor armazena a frequência ponderada por pesos que variam conforme a representatividade da subestrutura considerando todas as subestruturas da molécula.

A utilização de vetores binários naturalmente implica na perda de parte das informações da molécula. Contudo essa representação tem sido amplamente utilizada devido à redução da complexidade computacional na busca por estruturas similares. A próxima subseção descreve detalhes sobre a composição de cada posição do vetor de propriedades farmacofóricas adotado para tratar o modelo FFR de 19,5 ns.

5.3.1 Composição do vetor de propriedades farmacofóricas

O conjunto de pontos farmacofóricos pode ser representado pela combinação de n pontos, onde n é geralmente 3, 4 ou 5 pontos farmacóforos. A definição desse valor de n impacta diretamente na complexidade da avaliação do espaço farmacofórico, sendo este definido como todas as combinações das propriedades farmacofóricas considerando todos os possíveis valores discretizados de cada distância entre os pontos.

As avaliações considerando 4 ou 5 pontos farmacofóricos tem apresentado importantes resultados no mapeamento do espaço do ligante na cavidade alvo como, por exemplo, o trabalho apresentado em [BCS+07]. Nesse trabalho, Baroni e colaboradores desenvolveram o FLAP, uma ferramenta para a seleção de ligantes com base nos farmacóforos da cavidade de ligação de um conjunto de ligantes cristalizados na proteína. Nesse trabalho, a análise foi limitada a um pequeno conjunto de proteínas devido ao custo computacional necessário para a avaliação de todas as combinações das quádruplas e também devido ao método de seleção dos pontos farmacofóricos feitos por diferentes sondas atômicas. Desta forma, avaliar extensos conjuntos de conformações, como por exemplo modelos FFR, se torna uma tarefa extremamente onerosa.

Segundo Weill et. al [WR10], composições considerando 3 pontos farmacofóricos têm sido amplamente utilizadas por apresentarem uma melhor relação entre o conteúdo informacional e complexidade necessária. Assim, nesta tese optou-se pela utilização da representação avaliando 3 pontos farmacofóricos, definindo um conjunto de 3 propriedades e 3 distâncias euclidianas entre esses pontos. Essas distâncias são discretizadas em intervalos para estabelecer um conjunto finito de objetos. A Figura 5.11 apresenta um exemplo dessa representação de 3 pontos que caracteriza uma posição do vetor de características de cada conformação do modelo FFR.

O espaço farmacofórico de um conjunto de 3 pontos depende dos tipos de farmacóforos a serem considerados e da quantidade de intervalos cujas distâncias foram discretizadas. O cálculo para se descobrir todas as combinações possíveis depende de duas etapas:

- 1º etapa: Trata do número de combinações da quantidade de tipos de farmacóforos para 3 posições. Para ilustrar esse cálculo é apresentado o mesmo exemplo descrito em [SMdG07], cuja avaliação considerou 4 tipos de propriedades: AAA, AAB, AAC,

AAD, ABB, ABC, ABD, ACC, ACD, ADD, BBB, BBC, BBD, BCC, BCD, BDD, CCC, CCD, CDD e DDD. Neste caso, o total de combinações possíveis de um conjunto de 3 famacóforos pode ser obtido pela fórmula da combinação com repetição.

- 2º etapa: Calcula as combinações dos intervalos das distâncias. O total de combinações desta etapa é obtido pela permutação da quantidade de intervalos existentes.

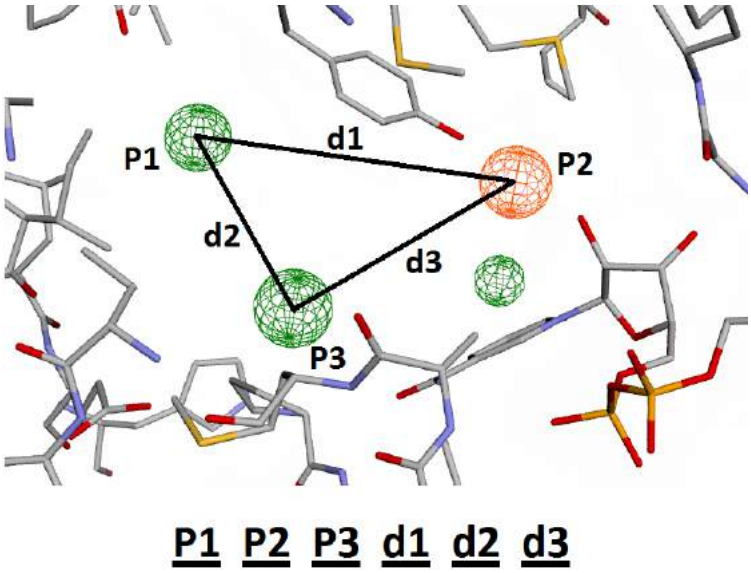


Figura 5.11 – Representação de um conjunto de 3 pontos farmacofóricos das propriedades da cavidade do receptor extraídas diretamente dos resíduos do receptor. Cada posição do vetor de características é composto pela informação de 3 famacóforos e suas respectivas distâncias entre os pontos avaliados. Neste exemplo, os pontos P1 e P3 denotam características hidrofóbicas (Hi), enquanto que o ponto P2 representa uma região aceitadora de Hidrogênio (Ac). Assim, a composição da posição do vetor seria formada por Hi Ac Hi, seguidos por d1 d2 e d3 representando as respectivas distâncias deste triângulo.

Recentes trabalhos foram analisados para definir as características necessárias para aplicar um agrupamento considerando o modelo FFR de 19,5 ns. As próximas subseções apresentam uma comparação dos atributos utilizados em outros trabalhos, tais como as propriedades consideradas e a quantidade de intervalos de distâncias avaliados.

5.3.2 Propriedades farmacofóricas avaliadas na cavidade de ligação do substrato.

Farmacóforos são frequentemente utilizados em etapas da triagem virtual de compostos [SKK02, BCS+07, SMdG07, WR10, ITS+12]. Embora diversos trabalhos tenham relatado relevantes contribuições na utilização desse método, é importante ressaltar que a simplicidade da representação das características farmacofóricas das moléculas implica na abstração das reais interações entre o complexo receptor-ligante. Portanto, existe uma margem de erro bastante considerável na utilização desse método.

As propriedades farmacofóricas dos átomos estão relacionadas com a sua natureza e o ambiente químico a sua volta, atribuindo a elas um conjunto de categorias pré-definidas que estão associadas a um comportamento de interação específico [LGLT10]. Assim, um farmacóforo está baseado em uma classificação físico-química simplista de átomos ou grupos funcionais que são classificados em hidrofóbicos (Hi), doadores de Hidrogênio (Do), aceitadores de Hidrogênio (Ac), anéis aromáticos (Ar), íons negativos (-) e íons positivos (+). Alguns autores consideram os átomos ou grupos funcionais que possuem átomos doadores e aceitadores de Hidrogênio (AD) como uma propriedade diferenciada. A Tabela 5.3 apresenta a descrição das propriedades consideradas por alguns dos principais trabalhos nesta área e também descreve as propriedades consideradas nesta tese.

Schmitt e colaboradores [SKK02] apresentaram uma importante contribuição nesta área, sendo um dos primeiros trabalhos a definir um conjunto de pseudocentros a partir de propriedades farmacofóricas. Contudo, seu trabalho não considerou a carga de íons, descartando assim duas relevantes propriedades. Weill e Rognam [WR10] descreveram um método capaz de identificar as conformações similares com base no alinhamento dos farmacóforos, chamado Fuzcav. Esse método utiliza um vetor de características para armazenar os triângulos formados por propriedades farmacofóricas que interagem com ligantes cristalizados na estrutura e, por consequência, não contemplando estruturas obtidas de modelos FFR ou que não possuam ligantes catalogados.

Tabela 5.3 – Propriedades famacofóricas atribuidas a cada resíduo em diferentes trabalhos.

Resíduos	Schmitt *et al.*, 2002	Weill & Rognan, 2010	Esta tese
ALA	Hi	Hi	Hi
ARG	Hi, Do	Hi, Do, +	Hi, Do, +
ASN	AD	Do, Ac	AD
ASP	Ac	Ac, -	Ac, -
CYS	Hi	Hi	Hi
GLN	Do	Do	Do
GLU	Ac	Ac, -	Ac, -
HIS	AD, Ar	Do, Ac, Ar	AD, Ar
HID	AD, Ar	Do, Ac, Ar	AD, Ar
HIE	AD, Ar	Do, Ac, Ar	AD, Ar
ILE	Hi	Hi	Hi
LEU	Hi	Hi	Hi
LYS	Hi, Do	Hi, Do, +	Hi, Do, +
MET	Hi	Hi	Hi
PHE	Hi, Ar	Hi, Ar	Hi, Ar
PRO	Hi	Hi	Hi
SER	AD	Do, Ac	AD
THR	Hi, AD	Hi, Do, Ac	Hi, AD
TRP	Do, Ar	Do, Ar	Do, Ar
TYR	AD, Ar	Do, Ac, Ar	AD, Ar
VAL	Hi	Hi	Hi

Outro fator limitante é que esse método não considera os resíduos que possuem tanto características aceitadoras quanto doadoras de Hidrogênio como uma propriedade exclusiva. Assim, o método Fuzcav trata duas variáveis dependentes (Ac e Do) de forma independente, aumentando a possibilidade da ocorrência de falsos positivos. Nesta tese, as propriedades avaliadas consideram tanto os íons positivos e negativos (descartados por [SKK02]), quanto adicionam a propriedade de átomos doadores e aceitadores de Hidrogênio (AD) (não avaliadas por [WR10]).

5.3.3 Discretização das distâncias entre os pontos farmacofóricos.

Para definir um conjunto finito de probabilidades, as distâncias entre os pontos farmacofóricos são discretizadas em valores que variam, normalmente, entre 5 e 10 categorias [DSPNW04, RTPRM05, WR10, ITS$^+$12]. Essa quantidade de categorias não pode ser muito elevada devido ao fato desse valor impactar diretamente na quantidade de elementos do vetor de características. A escolha dessa quantidade de categorias a serem criadas depende de uma análise do tamanho da cavidade de ligação e da proteína a ser investigada. Desta forma, uma avaliação amostral do conjunto de distâncias entre os pontos farmacofóricos do modelo FFR de 19,5 ns foi realizada.

A Figura 5.12 mostra o histograma resultante da avaliação do modelo FFR. Esse histograma é gerado a partir da combinação de todos os pontos farmacofóricos possíveis no conjunto de 19.500 conformações do modelo FFR. A distância utilizada como limiar para determinar o número de categorias e a avaliação de grandes volumes da cavidade de ligação do substrato considerando a flexibilidade de longos modelos flexíveis são dois fatores que, naturalmente, propiciam altos índices de frequência. A partir das informações geradas por esse histograma é possível determinar o número de categorias mais adequado para discretizar as distâncias entre os pontos farmacofóricos.

Na literatura existem trabalhos que diferem na maneira de discretizar o conjunto de distâncias entre os pontos farmacofóricos, considerando um conjunto de distâncias fixas ou incrementais para determinar os limites de cada categoria. Recentemente, os trabalhos apresentados por [ITS$^+$12, WR10] têm utilizado distancias fixas de 2.2 Å e 2.4 Å para discretizar o conjunto de distâncias entre os pontos farmacofóricos, respectivamente. Da mesma forma, nesta tese é adotado este parâmetro para determinar o intervalo das distâncias.

A Tabela 5.4 mostra os limites definidos de cada categoria dos trabalhos encontrados na literatura [DSPNW04, RTPRM05, WR10, ITS$^+$12]. O histograma apresentado na Figura 5.12 mostra uma quantidade elevada de distâncias que estão acima do maior limite dos trabalhos descritos na Tabela 5.4 ($\geq$ 18 Å em [DSPNW04]). Assim, a quantidade de categorias foi aumentada, visando contemplar adequadamente as distâncias acima de 18 Å. O intervalo entre as categorias foi mantido fixo, utilizando uma distância muito similiar a distância utilizada por Weill e Rognam [WR10].

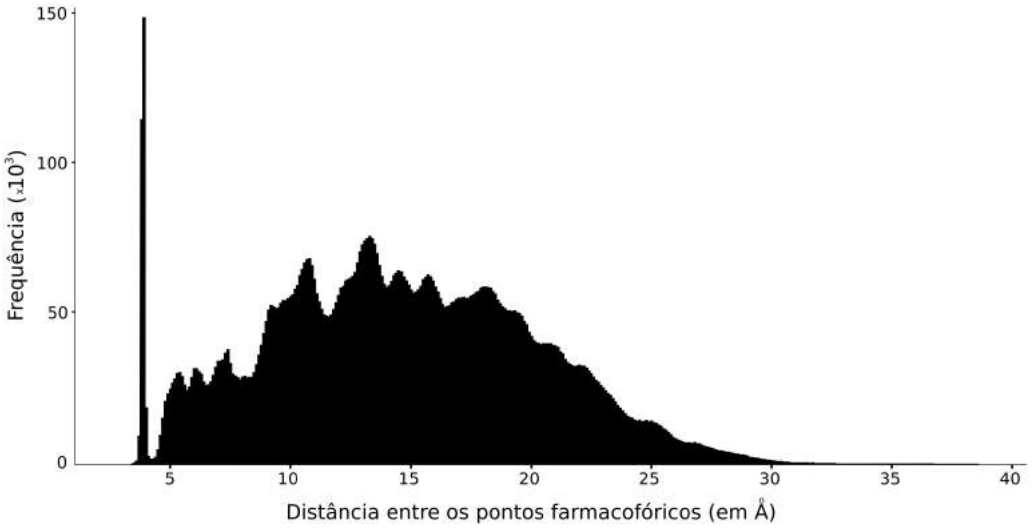


Figura 5.12 – Histograma mostrando a frequência das distâncias entre as propriedades farmacofóricas contidas na cavidade de ligação durante o modelo FFR de 19,5 ns. Imagem gerada utilizando um limiar de 0.1 Å.

Tabela 5.4 – Intervalos das distâncias entre as propriedades famacofóricas atribuídas a cada resíduo em diferentes trabalhos.

Intervalos	Dror_2004	Rodriguez_2005	Weill_2010	Ito_2011	Esta tese
1	< 2,5	< 3,0	< 4,8	< 4,8	< 4,8
2	< 4,0	< 4,5	< 7,2	< 7,0	< 7,2
3	< 6,0	< 6,0	< 9,5	< 9,2	< 9,6
4	< 9,0	< 7,5	< 11,9	< 11,4	< 12,0
5	< 13,0	< 9,0	< 14,3	< 13,6	< 14,4
6	< 18,0	< 10,5			< 16,8
7	$\geq$ 18,0	< 12,0			< 19,2
8		< 13,5			< 21,6
9		< 15,0			< 24,0
10		< 16,5			< 26,4
11					< 28,8
12					< 31,2
13					< 33,6
14					< 36,0
15					< 38,4
16					$\geq$ 38,4

Desta forma, os triângulos são categorizados com base em 7 propriedades famacofóricas e as distâncias entre estes pontos são discretizadas em 16 categorias. Assim, um vetor com o número total de combinações possíveis desta configuração é gerado, resultando em um vetor com CR_7^3 x 16^3 (344.064) posições. Devido à grande quantidade de dados manipulados, optou-se pela utilização de uma representação binária ao invés de contabilizar a frequência da quantidade de triângulos em cada cavidade de ligação. Uma equação foi elaborada para identificar o índice do vetor com base nas propriedades e as distâncias de um triângulo. Cada conformação é avaliada e descrita em um vetor binário de

344.064 posições, marcando 1 quando ocorrer o triângulo. Ao final da avaliação do modelo FFR, um vetor binário de mesmo tamanho identifica a ocorrência de cada posição. Esse vetor resultante é chamado de vetor de propriedades do modelo FFR. A Figura 5.13 apresenta uma representação visual da geração desse vetor de propriedades do modelo FFR, descrito neste parágrafo.

O vetor de propriedades do modelo FFR foi gerado para reduzir o custo de manipulação, uma vez que a avaliação total do modelo FFR necessitaria de uma matriz com a quantidade de conformações do modelo FFR como linhas e a quantidade de triângulos possíveis como colunas. A geração desse vetor de propriedades do modelo FFR possibilitou a redução de 344.064 posições para 60.654 triângulos distintos existentes, resultando na redução de aproximadamente 82,3% da memória necessária para analisar todo o modelo.

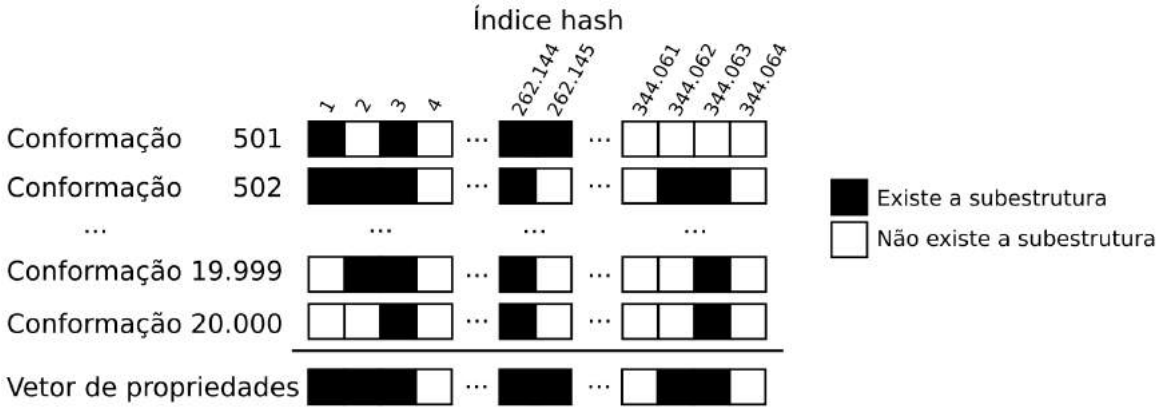


Figura 5.13 – Representação da geração do vetor de propriedades do modelo FFR, criado a partir da avaliação das subestruturas que ocorrem nas conformações do modelo FFR. Cada valor do índice desse vetor corresponde à codificação de um triângulo com 3 propriedades farmacofóricas e 3 distâncias entre esses pontos.

5.3.4 Agrupamento com base no vetor de propriedades de cada conformação

Após a identificação dos triângulos de propriedades farmacofóricas presentes em cada estrutura, esta etapa busca identificar e agrupar as estruturas com maior similaridade. Recentemente, trabalhos na área têm comparado diferentes coeficientes de similaridade existentes e reafirmado o coeficiente de Tanimoto como uma medida bem estabelecida para a avaliação da similaridade entre duas conformações [TCX+12, BAK15]. O coeficiente de Tanimoto é definido pela equação

$$Tanimoto(A, B) = \frac{N_{AB}}{N_A + N_B - N_{AB}} \tag{5.6}$$

onde N_A é o número de triângulos existentes na conformação A, N_B é o número de triângulos existentes na conformação B e N_{AB} é o número de triângulos existentes tanto na

conformação A como na conformação B. Os valores desse coeficiente variam em uma escala entre 0 e 1, sendo 1 para moléculas totalmente idênticas e 0 para moléculas sem nenhum triângulo em comum.

Uma matriz de similaridade é gerada utilizando o coeficiente de Tanimoto na comparação pareada entre as estruturas do modelo FFR. O resultado de cada comparação foi truncado após 6 casas decimais devido a quantidade de dados a serem armazenados. A aplicação desse truncamento, reduziu o tamanho do arquivo da matriz de similaridade para 3,3 GB, possibilitando a manipulação do arquivo em um tempo gerenciável. A Figura 5.14 mostra um fragmento do arquivo que armazena a matriz de similaridade resultante dessa avaliação. Desta forma, o número de linhas dessa matriz corresponde a mesma quantidade de conformações do modelo FFR. A quantidade de colunas corresponde ao número de conformações do modelo FFR mais uma coluna para descrever qual é a conformação.

501	1.000000	0.689888	0.629764	0.646742	0.639640		0.213011	0.457426	0.458176	0.543547	0.623323
502	0.689888	1.000000	0.594189	0.600391	0.558115		0.182153	0.399048	0.402708	0.543224	0.580233
503	0.629764	0.594189	1.000000	0.463873	0.621129	...	0.238472	0.505445	0.470397	0.504664	0.511679
504	0.646742	0.600391	0.463873	1.000000	0.570358		0.147564	0.332988	0.348590	0.479257	0.514603
...			...						...		
19997	0.320677	0.277778	0.357079	0.231368	0.314861		0.307457	1.000000	0.504039	0.302819	0.326087
19998	0.328690	0.286501	0.342643	0.246554	0.323457	...	0.302433	0.521739	1.000000	0.294613	0.340193
19999	0.404329	0.392812	0.386549	0.341463	0.371810		0.157813	0.358453	0.337145	1.000000	0.474511
20000	0.444444	0.405291	0.381339	0.354506	0.400641		0.182536	0.372194	0.375984	0.451210	1.000000

Figura 5.14 – Fragmento do arquivo que armazena a matriz de similaridade gerada pela comparação pareada entre os vetores de propriedades das conformações do modelo FFR utilizando o coeficiente de Tanimoto. A diagonal principal apresenta o valor máximo devido a comparação com a própria estrutura.

Ao final desta etapa, cada conformação possui um conjunto de colunas, onde cada valor define o grau de similaridade com relação a outra estrutura do modelo FFR. Assim, essa matriz de similaridade, gerada a partir do coeficiente de Tanimoto, permite o particionamento de conjuntos contendo conformações semelhantes do modelo FFR, com base na representatividade dos triângulos formados pelos pontos farmacofóricos.

Neste tipo de avaliação, o *single linkage* é um algoritmo muito apropriado devido ao seu método de particionamento permitir a identificação de conjuntos encadeados. A matriz de similaridade necessitaria ser invertida, de modo a obtermos a matriz de dissimilaridade utilizada pelo algoritmo *single linkage*: $(1 - x)$, onde x seria cada valor da matriz de similaridade. No entanto, mesmo com a matriz de dissimilaridade já calculada, a complexidade desse algoritmo é alta ($O(n^2)$). Isso ocorre devido à quantidade de repetições da etapa que examina todos os pontos dissimilares para gerar um grupo com os pontos mais próximos.

Outra possibilidade de particionar esse conjunto de conformações de forma mais ágil é utilizando o algoritmo *k*-means [Llo82]. No entanto, existe uma série de limitações na utilização desse algoritmo considerando esse conjunto de dados: (1) a possível existência de grupos pequenos próximos de grupos grandes; (2) a existência de *outliers*; e (3) a quantidade de memória necessária para a aplicação desse algoritmo. Recentemente, trabalhos na área têm enfatizado o problema do algoritmo *k*-means ao tratar conjuntos de dados muito grandes de serem armazenados na memória principal [SWM11, BBH+15]. Assim, um algoritmo mais robusto para identificar conjuntos de conformações similares do modelo FFR foi pesquisado.

Na literatura existe um algoritmo capaz de identificar estruturas "vizinhas" de acordo com a distância dos atributos. Esse algoritmo, chamado SketchSort [TUST10], favorece a formação de grupos similares, possibilitando a construção dos grafos de vizinhança das conformações. A Figura 5.15 descreve o funcionamento do método de ordenamento múltiplo, cuja função é uma das etapas do algoritmo SketchSort.

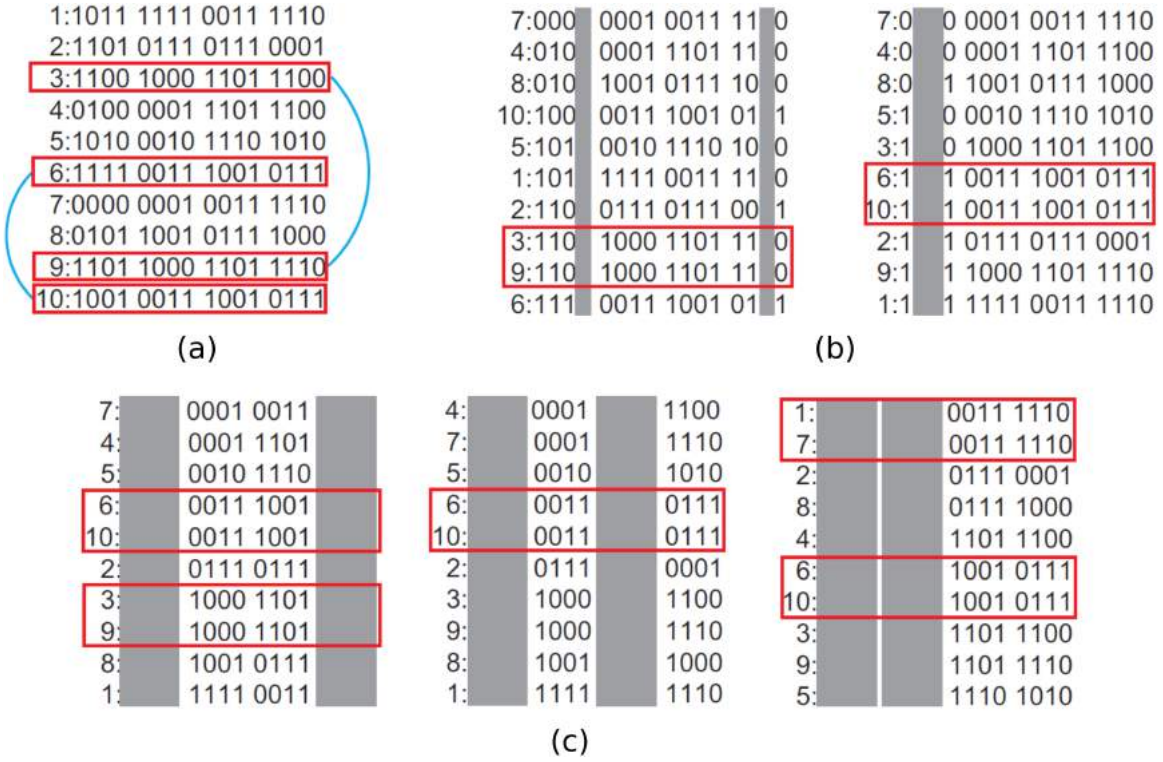


Figura 5.15 – Descrição do funcionamento do método de ordenamento múltiplo do algoritmo SketchSort. As distâncias de Tanimoto entre as conformações são convertidas em palavras binárias ordenadas. (a) As palavras similares podem ser detectadas com a aplicação de uma máscara de caracteres em todas as posições. Em destaque é apresentado um exemplo de pares de palavras similares com a distância de Hamming de até 2 unidades: (3,9) e (6,10). Exemplo da máscara de caracteres capaz de identificar os pares relatados no item (a). (c) As palavras binárias são divididas em blocos de caracteres, aos quais é aplicada uma máscara alternando os blocos de caracteres em todas as posições, assim como o funcionamento apresentado no item (b). Os blocos restantes são concatenados e as palavras resultantes idênticas são consideradas como pontos próximos. Adaptado de [UST04].

O primeiro passo do SketchSort trata da conversão dos valores reais, resultantes da comparação entre duas estruturas, em palavras binárias ordenadas[4]. Para realizar essa transformação, esse algoritmo utiliza uma função *Hash* Sensível à Localidade (LSH - do inglês *Locality Sensitive Hashing*) [IM98], cuja implementação utiliza o método da projeção aleatória. Nesse método, dada uma coleção de vetores R^d, um vetor aleatório é escolhido a partir de uma distribuição Gaussiana d-dimensional, em que a média é igual a 0 e o desvio padrão igual a 1. O vetor aleatório selecionado a partir dessa distribuição estável, garante que as medidas originais da vizinhança sejam mantidas após a conversão em palavras binárias [ITS+12].

Após a criação do conjunto de palavras binárias, o segundo passo consiste na ordenação dessas palavras pelo algoritmo radix [Dav92]. Esse algoritmo de ordenação de palavras é muito eficiente, cujo seu pior caso possui uma complexidade de $O(nk)$, onde *n* é o número de elementos e *k* é o tamanho da palavra.

O terceiro passo do algoritmo divide esse conjunto de palavras em blocos de caracteres de mesmo tamanho. Uma máscara é aplicada alternando os blocos de caracteres em todas as posições, conforme está exemplificado na Figura 5.15-(c). Os blocos restantes são concatenados e as palavras resultantes são comparadas pelo cálculo da distância de Hamming. Quando duas estruturas apresentam a distância igual a 0, o algoritmo define que duas estruturas são estruturas vizinhas.

No final da execução, o algoritmo SketchSort salva uma lista contendo os pares de conformações vizinhas. A partir dessa lista é possível elaborar os conjuntos de estruturas similares com a construção de um grafo não dirigido. A próxima seção apresenta a análise da aplicação desse algoritmo SketchSort com a matriz de similaridade do modelo FFR.

5.3.5 Análise dos conjuntos de estruturas similares identificadas utilizando o algoritmo SketchSort.

Ao utilizar um sistema de projeções, o algoritmo SketchSort não determina as estruturas vizinhas de todo o conjunto de conformações do modelo FFR. Ao total, o algoritmo avaliou 8.117 conformações, segmentando essa quantidade de conformações em 1.086 grupos. Contudo, diversos grupos apresentaram uma quantidade relativamente baixa de conformações, considerando a quantidade de conformações do modelo FFR avaliado. Assim, como forma de restringir a quantidade de estruturas a serem avaliadas posteriormente, foram selecionados apenas os grupos com uma representatividade acima de 0,5% das conformações identificadas pelo algoritmo SketchSort[5]. Esse critério selecionou 25 grupos e a quantidade de conformações de cada grupo está descrita na Tabela 5.5.

[4]No estudo de caso avaliado nesta tese, os valores reais são os resultados da comparação pareada entre as estruturas do modelo FFR feitas pelo coeficiente de Tanimoto.

[5]A porcentagem é calculada com base no número de estruturas amostradas pelo algoritmo SketchSort.

Tabela 5.5 – Grupos identificados pela avaliação do algoritmo SketchSort considerando o modelo FFR. Os 25 grupos com maior quantidade de conformações (> 0,5%) são selecionados para definir as estruturas representativas do modelo FFR.

Conformações no grupo	Quantidade de grupos	Representabilidade
570	1	6,78%
452	1	5,38%
429	1	5,10%
384	1	4,57%
256	1	3,05%
183	1	2,18%
176	1	2,09%
149	1	1,77%
143	1	1,70%
136	1	1,62%
117	1	1,39%
113	1	1,34%
97	1	1,15%
93	1	1,11%
90	1	1,07%
75	1	0,89%
60	1	0,71%
58	1	0,69%
52	1	0,62%
51	1	0,61%
50	1	0,59%
49	3	0,58%
45	1	0,54%
< 45	1.061	53,28%

Os grupos com representatividade acima de 0,5% foram selecionadas para definir o conjunto de estruturas representativas do modelo FFR[6]. Nesta tese, esse limiar foi definido por esse conjunto contemplar 46,72% do total de conformações selecionadas pelo algoritmo SketchSort e também para limitar o número de estruturas a serem analisadas. Conforme mostrado na Tabela 5.5, esses 25 grupos possuem uma porcentagem aproximada aos demais 1.061 grupos, cuja soma é de 53,28%. A partir da seleção desses grupos, a próxima etapa trata da identificação das estruturas representativas de cada grupo. Nesse sentido, o arquivo contendo os pares de conformações vizinhas gerado pelo SketchSort é essencial. Com base nesse arquivo, a estrutura presente em um maior número de arestas é definida como a estrutura representativa do grupo. As Figuras 5.16 e 5.17 mostram as diferentes estruturas conformacionais da cavidade de ligação do substrato obtidas das 25 estruturas representativas dos grupos gerados pelo SketchSort. A Figura 5.16 descreve o conjunto de 25 estruturas representativas ordenadas pelo volume. Uma análise do volume da cavidade de ligação do substrato mostra os diferentes arranjos 3D das estruturas representativas.

[6]Esse limiar pode variar de acordo com a quantidade de estruturas que o pesquisador desejar investigar.

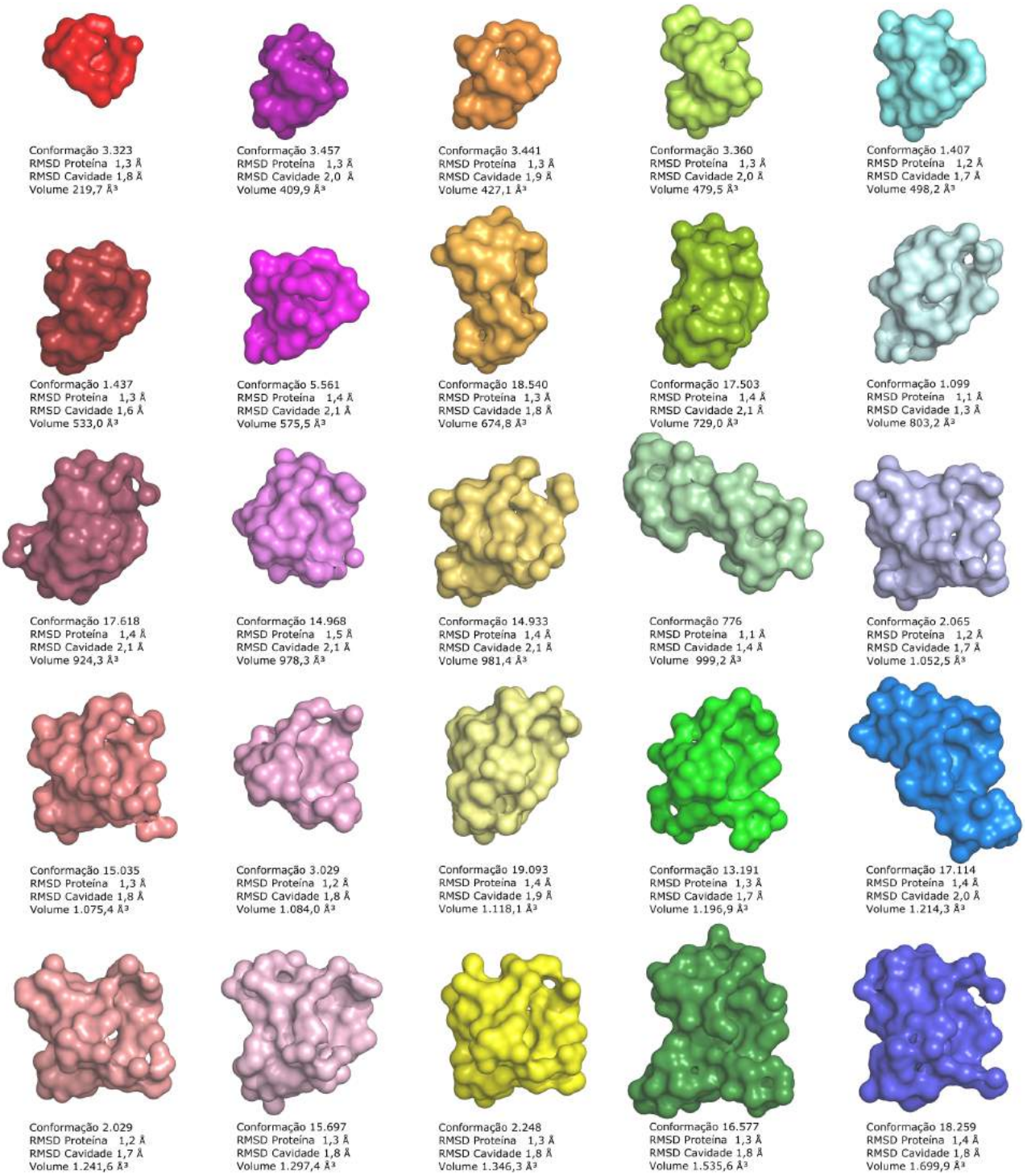


Figura 5.16 – Representação volumétrica das 25 estruturas representativas dos grupos identificados pelo algoritmo SketchSort. A análise por inspeção visual realça as diferenças conformacionais da cavidade de ligação do substrato existentes no modelo FFR. As conformações, que estão ordenadas pelo volume, descrevem outras informações comparativas como o valor do RMSD da proteína, o valor do RMSD dos átomos da cavidade de ligação do substrato e o volume identificado pelo CASTp [BNL03]. Um exemplo é a conformação 3.323 (vermelha) com o menor volume da cavidade, cujo valor do RMSD da proteína e o RMSD da cavidade resultaram em 1,3 Å e 1,8 Å, respectivamente. Essa conformação apresenta os valores de RMSD muito similares às 5 conformações com os maiores volumes encontrados, ressaltando o problema da métrica do RMSD.

A Figura 5.17 apresenta uma comparação entre as cavidades de ligação do substrato das estruturas representativas que apresentaram as maiores variações estruturais da cavidade alvo. Essa variação estrutural evidencia a flexibilidade da cavidade do substrato do modelo FFR dessa proteína. Além disso, esse conjunto possibilita a exploração de variados arranjos 3D da cavidade de ligação do substrato.

Diferentemente das estruturas representativas encontradas nas seções 5.1 e 5.2, o método utilizando um conjunto de triângulos de propriedades farmcofóricas com a seleção de estruturas vizinhas pelo SketchSort obteve uma seleção de estruturas representativas com cavidades do sítio de ligação bastante distintas. Um importante diferencial na utilização do algoritmo SketchSort foi o tratamento de uma quantidade densa de dados solucionada em um curto período de tempo (devido a sua complexidade ser linear).

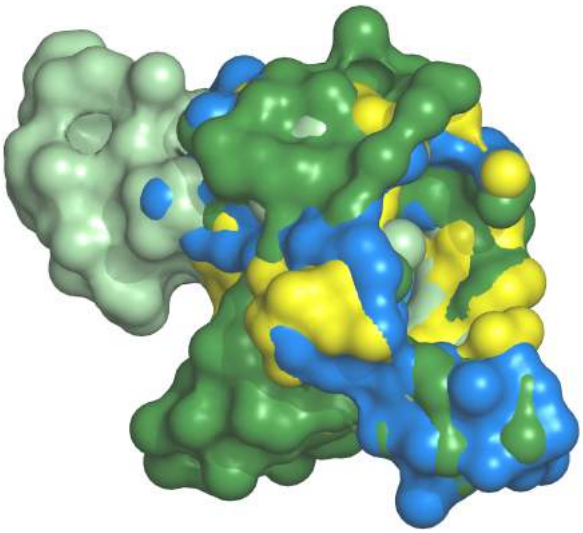


Figura 5.17 – Representação volumétrica dos átomos que delimitam a cavidade de ligação do substrato da enzima InhA de Mtb das conformações 776 (verde claro), 2.248 (amarelo), 16.577 (verde) e 17.114 (azul) do modelo FFR. Esta figura destaca as diferentes conformações representativas selecionadas pelo algoritmo SketchSort.

Uma avaliação da variância do RMSD foi realizada com as estruturas que estão conectadas no grafo resultante do algoritmo SketchSort. Esta avaliação comparou a variância dos valores de RMSD dos 25 conjuntos identificados nesta seção 5.3 e os comparou com os resultados da abordagem apresentada na seção 5.2. A Figura 5.18 apresenta o resultado da comparação destas abordagens. O gráfico em barras representa a variância calculada para cada proteína e as linhas mostram a análise da variança média acumulada em cada experimento. A final desta avaliação, os valores resultantes da variança acumulada para os conjuntos formados pelos Atributos da Cavidade, do RMSD da Cavidade, do RMSD da Proteína e as Propriedades Farmacofóricas foram de 0,46 Å, 0,50 Å, 0,51 Å e 0,40 Å, respectivamente. Desta forma, estes resultados corroboram os valores prévios identificados na análise apresentada na Figura 5.16.

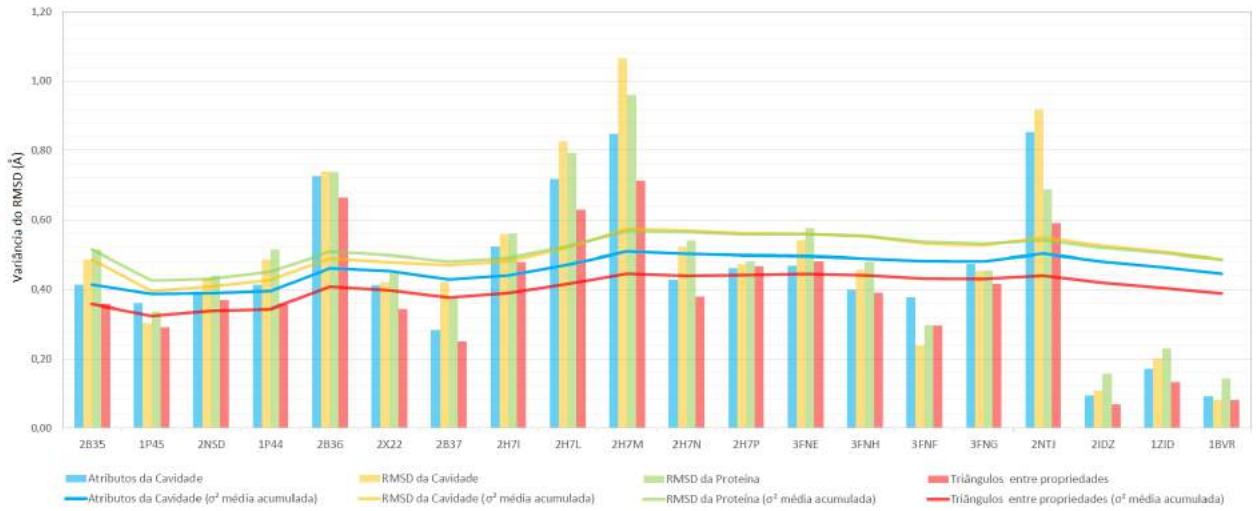


Figura 5.18 – Análise da variância dos valores do RMSD considerando os agrupamentos gerados a partir dos conjuntos de dados das propriedades farmacofóricas, Atributos da Cavidade, RMSD da Cavidade e o RMSD da Proteína. Os conjuntos de dados são separados por cores, onde o azul claro, amarelo, verde e salmão representam os resultados dos conjuntos de dados dos Atributos da Cavidade, RMSD da Cavidade, RMSD da Proteína e Propriedades Farmacofóricas, respectivamente. O gráfico em barras representa os valores da variância para cada proteína e o gráfico de linhas apresenta os valores médios acumulativos de cada conjunto. O agrupamento formado a partir das informações das propriedades farmacofóricas apresenta a menor variância em 95% dos casos analisados.

O conjunto formado pela pelas propriedades farmacofóricas apresentou uma variabilidade menor por conseguir formar conjuntos compostos por cavidades do sítio ativo muito similares. Assim como na análise da seção 5.2, os adutos apresentaram uma variância baixa. Isso ocorre pelo fato da coenzima NADH não apresentar uma grande alteração, visto que o local de encaixe é propício justamente para esta coenzima. Além disso, há uma ligação covalente entre o ligante e o anel da nicotinamida do NADH para formar o aduto. Assim, podemos considerar que uma parte do ligante fica “fixa” na posição de referência.

5.4 Considerações finais.

Segundo [TA08], em experimentos de triagem virtual é melhor utilizar um conjunto pequeno e cuidadosamente selecionado de conformações ao invés de avaliar todas as conformações de um modelo FFR. Nesse sentido, a utilização de algoritmos de agrupamento para particionar conformações similares tem sido a técnica de mineração de dados mais adequada para filtrar as informações estruturais de propriedades de uma trajetória de modelos FFR [STTC07, TvG94]. No entanto, a aplicação de técnicas de agrupamento baseadas apenas nos valores de RMSD não se mostra apropriada para separar conformações quando existe uma cavidade de ligação do substrato conhecida. Isso ocorre devido aos valores do RMSD serem obtidos a partir de medidas absolutas da distância euclidiana, não avaliando características fundamentais da cavidade de ligação do substrato do receptor.

Antes das avaliações descritas neste capítulo, os agrupamentos gerados previamente em nossas pesquisas já evitavam utilizar somente as informações do RMSD da proteína, considerando como conjunto de dados as interações entre uma coenzima e o modelo FFR. Contudo, esse tipo de avaliação tornava o agrupamento específico para um ligante, ocasionando a necessidade de análises manuais frequentes. Dessa forma, essa etapa da geração dos agrupamentos era um grande gargalo que inviabilizava a automatização do processo de testes sequenciais de conjuntos de ligantes utilizando o ambiente wFReDoW [DPFNdSR13].

Assim, neste capítulo foram apresentados três estudos sobre novos métodos para selecionar grupos de conformações similares a partir de modelos FFR com base na utilização de propriedades essenciais da cavidade de ligação do substrato do receptor. As seções a seguir apresentam um resumo de cada estudo.

5.4.1 Agrupamento baseado na análise de 4 propriedades da cavidade de ligação do substrato.

Este foi o primeiro conjunto de experimentos utilizando somente a análise da cavidade de ligação do substrato por alguns critérios de similaridade, sem utilizar quaisquer resultados de docagem molecular. Abaixo estão descritas as principais características deste agrupamento:

- conjunto de dados: 4 atributos, sendo eles: o RMSD da proteína, a área acessível ao solvente, o volume da cavidade do substrato e a quantidade de átomos pesados da enzima 1BVR presentes na cavidade de ligação do substrato da conformação.
- índice para estimar o melhor número de grupos: índice Davies-Bouldin, índice Dunn e o *gap* estatístico.
- algoritmo: *k*-means
- quantidade de grupos: 10
- principais vantagens: o conjunto de entrada apresenta detalhes que complementam as informações dos valores do RMSD. Os agrupamentos são obtidos de forma rápida devido ao algoritmo de agrupamento utilizado.
- principais desvantagens: a baixa quantidade de agrupamentos identificados. Além disso, também apresentou pouca precisão nos grupos particionados devido ao conjunto de dados possuir poucos atributos. As estruturas representativas de cada grupo não apresentaram relevantes variações conformacionais.

Esse novo agrupamento tornou possível realizar o processo de testes sequenciais de conjuntos de ligantes utilizando o ambiente wFReDoW(*web Flexible Receptor Docking*

Workflow) de forma automatizada. As particularidades desse agrupamento solucionam o problema da geração automática dos grupos de entrada do P-SaMI [Hüb10, HRFNdS15], o qual procura selecionar grupos de conformações que sejam favoráveis entre as interações do modelo FFR com o ligante. Os resultados obtidos com este agrupamento foram publicados na revista *Expert Systems With Aplications* [QDPRNdS14].

Embora este agrupamento tenha apresentado bons resultados com o P-SaMI, as estruturas representativas de cada grupo não apresentaram relevantes variações conformacionais. Assim, este agrupamento não fornece uma variabilidade estrutural que pretende-se utilizar nesta tese.

5.4.2 Agrupamento baseado na análise de 12 propriedades da cavidade de ligação do substrato.

Uma das limitações mais importantes identificadas no agrupamento descrito na Seção 5.1 se refere à subjetividade ainda existente no conjunto de dados avaliado. A solução encontrada foi descrever de forma mais detalhada o atributo que armazenava a quantidade de átomos pesados da enzima 1BVR presentes na cavidade de ligação do substrato da conformação. Abaixo estão descritas as principais características deste agrupamento:

- conjunto de dados: 12 atributos, sendo eles: o RMSD da cavidade de ligação do substrato, o volume da cavidade de ligação do substrato e os outros 10 atributos descrevem a quantidade de átomos de cada um dos 10 resíduos da enzima 1BVR presentes na cavidade de ligação do substrato da conformação.
- índice para estimar o melhor número de grupos: utilizou-se a soma das diferenças dos três quartis entre os agrupamentos identificados e os valores do modelo FFR.
- algoritmos: algoritmos particionais (*k*-means e *k*-medoid) e algoritmos hierárquicos aglomerativos (*Complete linkage*, UPGMA, WPGMA e Ward's).
- quantidade de grupos: 48
- principais vantagens: o novo conjunto de dados permite caracterizar mais detalhadamente os diferentes comportamentos encontrados no sítio de ligação ao longo do modelo FFR, possibilitando a identificação de estruturas representativas capazes de garantir resultados aproximados com o modelo FFR tanto da FEB quanto do RMSD. Outra vantagem foi a descrição detalhada de diversos algoritmos de agrupamento, validando os resultados com informações dos valores da FEB e do RMSD.
- principais desvantagens: definir a quantidade ideal de grupos utilizando diferentes conjuntos de dados e diferentes algoritmos. As estruturas representativas de cada grupo também não possuem relevantes variações conformacionais entre os grupos.

A avaliação das 48 estruturas representativas identificadas por este agrupamento apresentou uma considerável aproximação com o modelo FFR, atingindo 75% dos ligantes obtidos das estruturas cristalinas do PDB. Isso destacou a relevância de considerar um conjunto maior de propriedades da cavidade de ligação de substrato (além dos valores do RMSD). A avaliação do conjunto de Atributos da Cavidade pelo algoritmo hierárquico *Complete linkage* mostrou-se capaz de reduzir a dimensão do modelo FFR a um tamanho gerenciável, mantendo as características mais relevantes para encontrar novos inibidores para a proteína em estudo. Esses resultados foram publicados na revista *PloS one*.

Embora este agrupamento tenha apresentado resultados interessantes, esperava-se obter agrupamentos de estruturas representativas apresentando dispersões com grupos bem separados e coesos. Além disso, as conformações representativas de cada grupo também não apresentaram relevantes variações conformacionais, fato também constatado no agrupamento descrito na Seção 5.1.

5.4.3 Agrupamento com vetores de propriedades farmacofóricas.

As estruturas representativas selecionadas nas seções 5.1 e 5.2 não apresentaram variações estruturais significativas entre os diferentes grupos. Após essa avaliação, foram pesquisadas formas mais adequadas de se identificar estruturas similares foram pesquisados. Trabalhos recentes na literatura têm apontado para os problemas relacionados com a utilização de informações como o RMSD, medida que pode variar conforme o alinhamento dos átomos considerados [CLP11, L$^+$13]. A fim de evitar o viés ocasionado pela medida do RMSD, a Seção 5.3 mostrou um novo método que avalia as propriedades físico-químicas da cavidade de ligação do substrato utilizando triangulações entre as propriedades farmacofóricas do receptor. Abaixo estão descritas as principais características deste agrupamento:

- conjunto de dados: vetor de propriedades farmacofóricas com 60.654 posições. Cada posição corresponde a um triângulo canônico formado a partir das propriedades farmacofóricas da cavidade de ligação do substrato do receptor.

- algoritmo(s): SkecthSort.

- quantidade de grupos: 1.086. no entanto, o número de estruturas representativas foi limitado em 25 por questões computacionais e devido a esse conjunto de grupos conter $\cong$ 46% das estruturas similares identificadas pelo algoritmo.

- principais vantagens: o conjunto de dados armazena uma grande quantidade de informações detalhadas da cavidade de ligação do substrato, possibilitando uma avaliação fidedigna às características de cada conformação do modelo FFR. O conjunto

de estruturas representativas elencadas possui uma grande variabilidade estrutural (conforme ilustrado na Figura 5.16).

- principais desvantagens: a etapa de caracterização da estrutura até a elaboração da matriz de similaridade pelo cálculo do coeficiente de Tanimoto é demorada. esse tempo é compensado pela aplicação do algoritmo SkecthSort. Contudo, esse algoritmo não identifica a similaridade de todas as conformações do modelo FFR.

Em comparação com outros estudos de agrupamento de trajetórias de modelos FFR, o método proposto na seção 5.3 utilizando um conjunto de triângulos de propriedades farmcofóricas com a seleção de estruturas vizinhas pelo SketchSort apresentou vantagens essenciais na seleção de estruturas representativas, identificando cavidades do sítio de ligação bastante distintas.

Embora o método apresentado aqui tenha selecionado um conjunto contendo 25 estruturas, outros valores podem ser utilizados, dependendo apenas da quantidade de estruturas que o pesquisador está disposto a avaliar. Cabe ressaltar que é preciso avaliar constantemente o equilíbrio do custo versus benefício do conjunto de dados e dos métodos a serem utilizados.

O próximo capítulo descreve o desenvolvimento do método para determinar conjuntos de hipóteses farmacofóricas 3D baseadas nas características físico-químicas da estrutura 3D de regiões inacessíveis por estruturas cristalinas do receptor InhA de *Mtb*. O conjunto de propriedades farmacofóricas são extraídas a partir das estruturas representativas dos grupos formados na seção 5.3.

6. Triagem virtual em BD de ligantes considerando propriedades físico-químicas das estruturas representativas do modelo FFR

No capítulo anterior um método para particionar conjuntos de estruturas com cavidades de ligação do substrato similares foi descrito. Esse método identificou 25 conjuntos de estruturas similares mais representativas. Após definir o conjunto amostrado do modelo FFR, este capítulo descreve a etapa final do método visando reduzir o conjunto de ligantes a serem avaliados nos experimentos de docagem molecular. A qualidade dos conjuntos de ligantes a serem selecionados depende diretamente da especificidade do arranjo espacial de propriedades físico-químicas essenciais para a formação das interações de complexos receptor-ligante utilizadas no processo de triagem virtual.

Segundo Hu [HL13], grande parte dos modelos farmacofóricos baseados em estrutura são dependentes das características físico-químicas presentes nos complexos receptor-ligante conhecidos. Desta forma, as regiões dentro da cavidade que não interagem com o conjunto de ligantes formadores do modelo farmacofórico e que podem permitir a interação de ligantes estruturalmente diferentes, consequentemente, não estarão contempladas nesta busca seletiva. Assim, abordagens alternativas têm sido desenvolvidas utilizando as propriedades da cavidade de ligação do substrato para gerar modelos farmacofóricos baseados na proteína sem considerar as informações de ligantes complexados [KBK$^+$01, TCM$^+$08, CC11, HL12]. No entanto, importantes limitações têm sido encontradas nessas abordagens, reforçando a necessidade de novas soluções capazes de endereçar o problema da flexibilidade de forma efetiva.

Assim, o método apresentado nesta tese busca selecionar um conjunto de ligantes contendo características físico-químicas 3D favoráveis ao encaixe na cavidade de ligação do substrato de um modelo FFR. As estruturas selecionadas irão compor uma lista dos ligantes mais promissores a se tornarem candidatos a fármacos para a enzima avaliada. Inicialmente, as etapas da metodologia são apresentadas. Após, para testar a eficácia da função desenvolvida serão utilizados experimentos com a enzima InhA de *Mycobacterium tuberculosis*.

6.1 Metodologia

Segundo Dror [DSPNW04], existem quatro etapas fundamentais para o desenvolvimento de modelos farmacofóricos 3D baseados na estrutura de receptores flexíveis. Essas etapas consideram: (1) a qualidade das estruturas a serem utilizadas, (2) a amostragem de conformações desse conjunto, (3) a identificação das propriedades farmacofóricas complementares da cavidade de ligação do substrato e, por fim, (4) a análise do conjunto de ligantes selecionados.

Algumas das etapas definidas por Dror [DSPNW04] foram readequadas no intuito de melhorar o entendimento deste trabalho. Assim, a metodologia adotada está dividida nas seguintes etapas:

- Etapa 1: Analisar a proteína alvo e avaliar a qualidade das estruturas do modelo flexível a ser utilizado na geração dos modelos farmacofóricos 3D.
- Etapa 2: Selecionar os conjuntos de conformações similares do modelo FFR da proteína alvo.
- Etapa 3: Gerar os modelos famacofóricos 3D com base nos grupos de estruturas similares, destacando as regiões que não se sobrepõem a estrutura cristalina.
- Etapa 4: Analisar o conjunto de ligantes selecionados pelo modelo farmacofórico.

6.2 Analisar a proteína alvo e avaliar a qualidade das estruturas do modelo flexível a ser utilizado na geração dos modelos farmacofóricos 3D

A avaliação da qualidade das estruturas a serem estudadas é fundamental para possibilitar, ao final do processo, a geração de modelos farmacofóricos acurados. A qualidade da avaliação dessas estruturas depende, primeiramente, da qualidade da estrutura que será utilizada como molde para a geração do modelo FFR. A qualidade das estruturas cristalinas depende principalmente da resolução e do fator-R, devendo serem consideradas nos estudos somente estruturas que possuam boas resoluções e com um fator-R satisfatório [Las95, Las03]. A avaliação dessa etapa foi realizada previamente durante a elaboração desta tese no capítulo 3 e no capítulo 4.

6.3 Selecionar os conjuntos de conformações similares do modelo FFR.

Algoritmos de agrupamento têm sido amplamente utilizados para reduzir a dimensionalidade de trajetórias de modelos FFR [STTC07, TvG94]. No entanto, trabalhos recentes na literatura têm apontado os problemas relacionados com a utilização de informações como o RMSD, medida que também tende a variar conforme o alinhamento dos átomos considerados [CLP11, L$^+$13].

A fim de evitar o viés ocasionado na geração de agrupamentos baseados somente na medida do RMSD, o capítulo 5 descreveu três métodos distintos para selecionar os conjuntos de conformações similares do modelo FFR da proteína alvo considerando as propriedades físico-químicas da cavidade de ligação do substrato. O método descrito na seção 5.3, utilizando triangularizações entre as propriedades farmacofóricas do receptor, apresentou os melhores resultados, identificando estruturas representativas de cada grupo com arranjos espaciais 3D bastante distintos.

6.4 Gerar os modelos famacofóricos 3D com base nos grupos de estruturas similares, destacando as regiões que não se sobrepõem a estrutura cristalina.

As informações 3D da cavidade de ligação do substrato contendo as propriedades físico-químicas têm possibilitado a descoberta de proteínas que possuem sítios funcionais similares de forma eficiente [SKK02, SPNW04]. A Figura 6.1 mostra um exemplo da pesquisa desenvolvida por Shulman-Peleg [SPNW04], cujo objetivo era identificar sítios funcionais similares entre as estruturas disponibilizadas no sítio do PDB. Nesse trabalho, Shulman-Peleg e colaboradores avaliaram a influência das propriedades farmacofóricas, definindo um conjunto de pseudocentros na coordenada atômica do átomo de influência localizado na cavidade da proteína. Esses pseudocentros representam farmacóforos 3D que definem propriedades físico-químicas que podem interagir com o ligante [LP03].

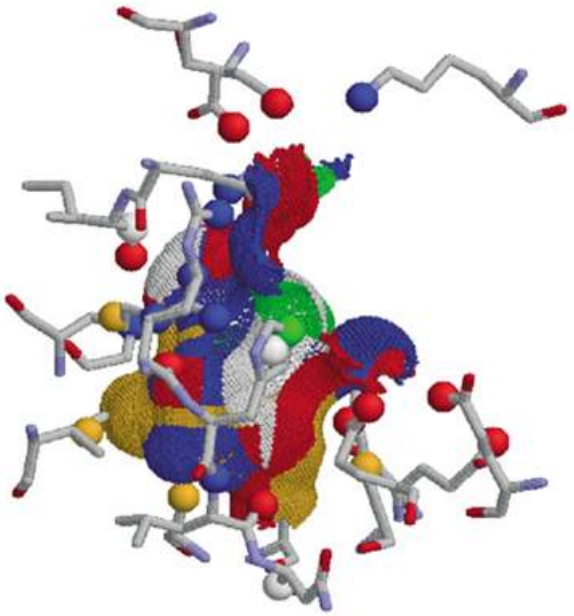


Figura 6.1 – Representação das propriedades físico-químicas de uma cavidade identificadas pelo conjunto de pseudocentros da proteína. A superfície acessível aos pseudocentros são coloridas de acordo com as propriedades físico-químicas do pseudocentro em contato (Azul: Do; Vermelho: Ac; Verde: Ac e Do; Laranja: Hi; Branco: Ar).

Deste modo, os pseudocentros podem ser definidos como a representação hipotética das propriedades farmacofóricas dos átomos ou grupos funcionais do receptor na cavidade de ligação do substrato [SKK02]. Desta forma, os pseudocentros estão baseados em uma classificação físico-química simplista de átomos ou grupos funcionais que são classificados em hidrofóbicos (Hi), doadores de Hidrogênio (Do), aceitadores de Hidrogênio (Ac), anéis aromáticos (Ar), íons negativos (-) e íons positivos (+). A Tabela 5.3 apresenta a relação dos resíduos e as suas respectivas propriedades farmacofóricas consideradas[1].

[1]Neste capítulo a propriedade do átomo aceitador e doador de Hidrogênio (AD) é considerada como 2 pontos independentes de mesma coordenada, permitindo que o especialista escolha as propriedades que deseja seleção.

6.4.1 Identificação das propriedades farmacofóricas complementares do receptor.

As propriedades farmacofóricas avaliadas pelos pseudocentros descreveram até o momento as características relacionadas ao receptor. Contudo, o objetivo desta tese é desenvolver um filtro para selecionar conjuntos de ligantes a partir da avaliação das propriedades da cavidade de ligação do receptor. Para definir o arranjo 3D das propriedades farmacofóricas da cavidade e selecionar conjuntos de ligantes é preciso obter a região complementar das propriedades do receptor. Assim, é possível gerar a complementariedade dos pseudocentros com base nas propriedades e distâncias definidas na Figura 6.2. O conjunto de distâncias adotado seguiu a avaliação de estruturas experimentais definidas por Liljefors [LP03].

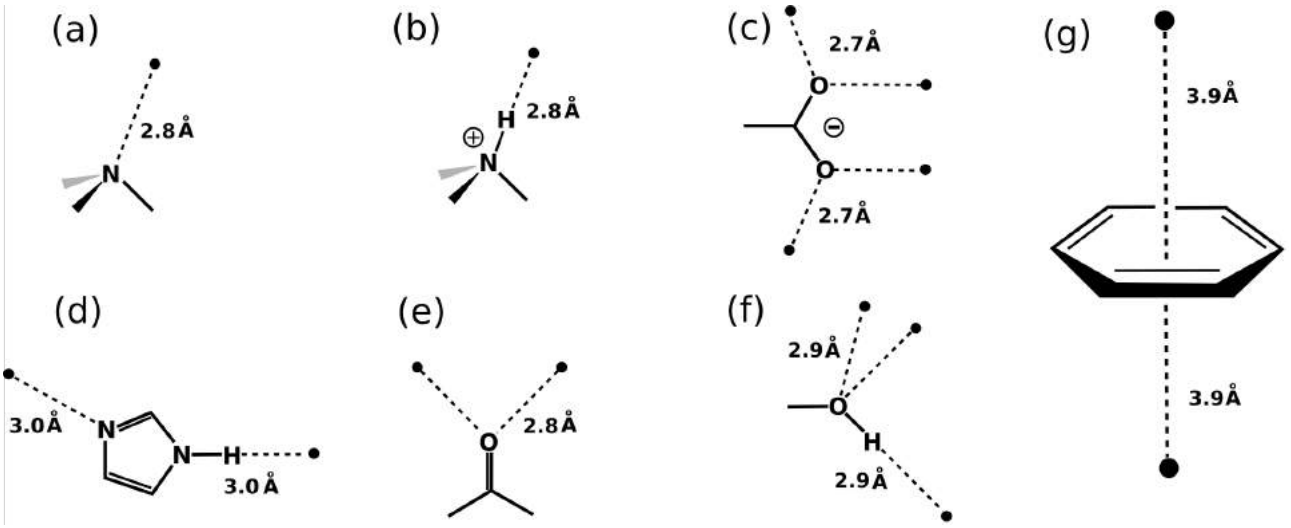


Figura 6.2 – Definição das distâncias das projeções do pseudocentro para cada átomo/grupo funcional da conformação. Os fragmentos pertencentes aos resíduos são representados pelos átomos Nitrogênio (N), Oxigênio (O) e Carbono (C). A região de contato do ligante corresponde aos círculos pretos. **(a)** Grupo amina (Ac) pode interagir com Do. **(b)** Grupo amina (Do) pode interagir com Ac. **(c)** Grupo carboxilato (-) pode interagir com um íon +. **(d)** HIS (Do,Ac) pode interagir com Do e Ac. **(e)** Grupo cetona (Ac) pode interagir com Do. **(f)** Grupo hidroxila (Do,Ac) pode interagir com Do e Ac. **(g)** Anel aromático (Ar) pode interagir com Ar em ambos os lados. Adaptada de [LP03].

A partir da identificação dessas propriedades já seria possível projetar todo o conjunto de propriedades da cavidade de ligação do substrato conforme o conjunto de resíduos que delimitam essa cavidade alvo. No entanto, avaliar muitas propriedades farmacofóricas sem haver uma prioridade (por exemplo definir interações essenciais) acaba tornando a busca redundante. Normalmente, modelos farmacofóricos que consideram acima de 7 propriedades químicas também acabam incorrendo no mesmo problema, obtendo como resultado um conjunto expressivo de estruturas a serem docadas [Yan10]. Desta forma, existem muitos métodos desenvolvidos para reduzir a dimensionalidade desse problema [LB12].

6.4.2 Projetar as propriedades farmacofóricas das estruturas representativas do modelo FFR, destacando as propriedades não acessíveis da estrutura cristalina modelo.

Nesta tese, o receptor investigado é um modelo de Receptor Totalmente Flexível. O principal motivo de se utilizar um modelo flexível é o de se conseguir explorar as regiões não contempladas pelos conjuntos de estruturas cristalinas. Assim, a simulação de parte da flexibilidade do receptor pode possibilitar o acesso de pequenas moléculas candidatas à fármaco às regiões "inéditas" proporcionadas pelo modelo FFR. O método elaborado avalia o modelo flexível e o compara com uma estrutura cristalina para encontrar os modelos farmacofóricos 3D que possam selecionar um conjunto de estruturas diferentes de um conjunto obtido a partir da avaliação realizada somente de uma estrutura cristalina. Desta forma, para cada propriedade farmacofórica já existente na estrutura cristalina, a propriedade é projetada na cavidade alvo com o raio da propriedade farmacofórica reduzido (0,2 Å). As propriedades farmacofóricas de regiões inacessíveis por estruturas cristalinas possuem raios de 1,0 Å. A Figura 6.3 mostra um exemplo de um modelo famacofórico 3D gerado a partir do método apresentado nesse capítulo avaliando a estrutura 776 ps do modelo FFR de InhA em comparação com a estrutura 2B37 [STB$^+$06].

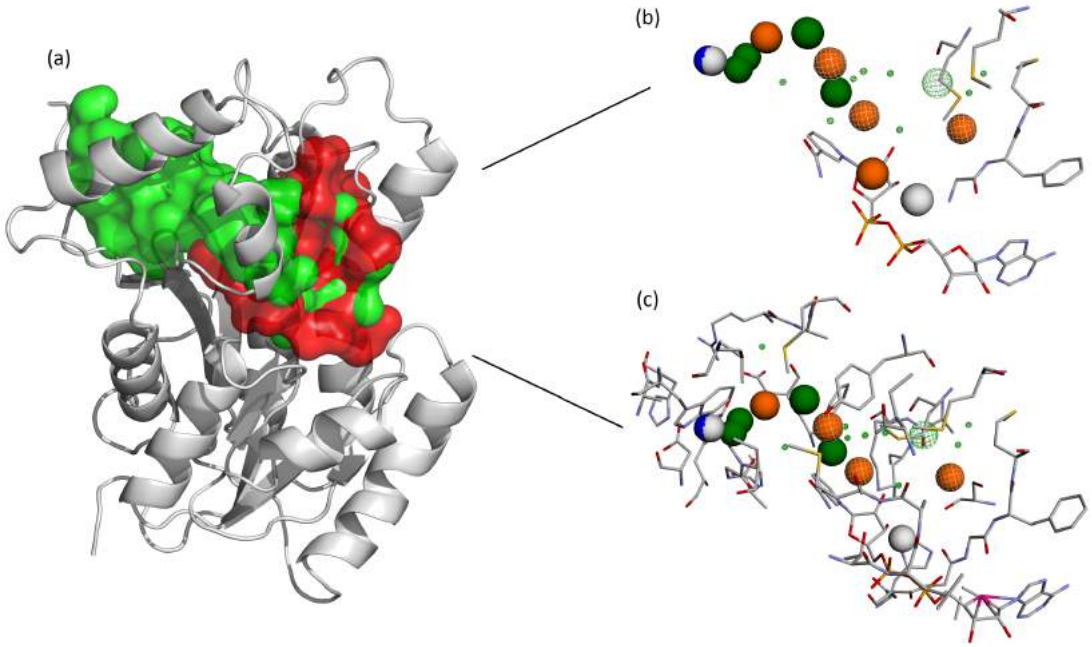


Figura 6.3 – Representação de um modelo famacofórico 3D obtido a partir da avaliação da função heurística da estrutura 776 ps do modelo FFR de InhA utilizando como estrutura de comparação a estrutura 2B37 (Amarelo: Do; Branco: Ac; Verde: Hi; Azul: íon positivo). (a) Diferença entre os volumes da cavidade de ligação do substrato da estrutura 776 ps do modelo FFR de InhA (verde) e da estrutura cristalina 2B37 (vermelho). (b) Projeção do modelo farmacofórico com os átomos que delimitam a cavidade de ligação do substrato da estrutura cristalina 2B37. Essa imagem ressalta ao especialista de domínio a região flexível que o modelo está avaliando. (c) Projeção do modelo farmacofórico com os átomos que delimitam a cavidade de ligação do substrato da conformação 776 do modelo FFR.

Esse tipo de projeção diferenciada das propriedades evidencia para o pesquisador quais propriedades farmacofóricas 3D encontradas são devidas à consideração da flexibilidade do modelo FFR. Na Figura 6.3-a é possível aferir a diferença entre os volumes da cavidade de ligação do substrato da estrutura 776 ps do modelo FFR de InhA (verde) e da estrutura cristalina 2B37 (vermelho). O volume da cavidade alvo da estrutura cristalina é menor que o volume identificado na estrutura do modelo FFR. A Figura 6.3-b e a Figura 6.3-c apresentam a projeção do modelo farmacofórico com os átomos que delimitam a cavidade de ligação do substrato da estrutura cristalina 2B37 e da conformação 776 do modelo FFR, respectivamente. Em especial, a Figura 6.3-b ressalta ao pesquisador os pontos farmacofóricos que não estão acessíveis na avaliação da estrutura cristalina. Assim, uma das principais contribuições desta tese é permitir ao especialista de domínio que ele mesmo manipule o modelo farmacofórico 3D, selecionando as hipóteses que sejam mais adequadas.

No processo de triagem virtual em BD de ligantes é natural ocorrer consultas que retornem uma grande quantidade de compostos. Uma forma de restringir efetivamente esse espaço de busca é inserir as informações das regiões que os ligantes a serem selecionados não podem ocupar. Esse volume ocupado pelo receptor é conhecido como volume essencial ocupado.

6.4.3 Identificação do volume essencial do receptor.

As avaliações de triagem virtual baseadas em modelos farmacofóricos que consideram na avaliação os volumes inacessíveis da cavidade têm apresentado resultados com maiores índices de acerto [SMdG07]. Desta forma, o método desenvolvido calcula o volume essencial do receptor e gera um arquivo independente contendo somente a informação dessas regiões inacessíveis. Esse arquivo serve para complementar o arquivo principal que armazena as informações farmacofóricas da cavidade alvo.

A Figura 6.4 mostra a projeção do mesmo modelo farmacofórico 3D da conformação 776 ps do modelo FFR. A diferença entre a Figura 6.4-a e a Figura 6.4-b está na avaliação ou não do volume essencial do receptor na consulta a ser utilizada no processo de triagem virtual. A projeção dessas esferas deve limitar sensivelmente a quantidade de ligantes a serem avaliados.

6.4.4 Editar as hipóteses famacofóricas 3D e aplicar o filtro para selecionar o conjunto de ligantes do Banco de Dados ZINC.

A função heurística desenvolvida nesta tese avalia o conjunto das 25 estruturas representativas usando como comparação as estruturas cristalinas descritas na Tabela 3.1. O resultado da avaliação de cada conformação representativa é armazenado em um arquivo

contendo as propriedades farmacofóricas. Esse arquivo da avaliação da conformação representativa serve como o arquivo de entrada da ferramenta ZINCPharmer [KC12]. Essa ferramenta permite a construção e o refinamento de modelos farmacofóricos gerados a partir da avaliação dos ligantes conhecidos, pesquisando por novos ligantes que contemplem a restrição farmacofórica extraída do ligante conhecido. Segundo Koes & Camacho [KC12], as principais características dessa ferramenta são:

- Editar: Existe uma interface gráfica que permite a edição e o refinamento dos modelos farmacofóricos a serem pesquisados. Os ligantes resultantes da pesquisa no BD podem ser visualizados para aferir o quanto as propriedades do ligante conseguiram sobrepor a hipótese farmacofórica.

- Banco de Dados de ligantes: O ZINCPharmer utiliza um BD modelado para o tipo de necessidade dessa ferramenta. Esse BD é atualizado mensalmente pelo conjunto de ligantes capturados a partir do BD ZINC. Cada molécula do ZINC é projetada tridimensionalmente pelo software Omega e os 10 melhores confôrmeros são armazenados, como uma forma de avaliar a flexibilidade do ligante.

- Busca por ligantes com base na avaliação dos farmacóforos: o ZINCPharmer é baseado no método de pesquisa por farmacóforos desenvolvido pelo Pharmer [KC11], cujo processamento é capaz de pesquisar quase 2 milhões de estruturas em menos de um minuto.

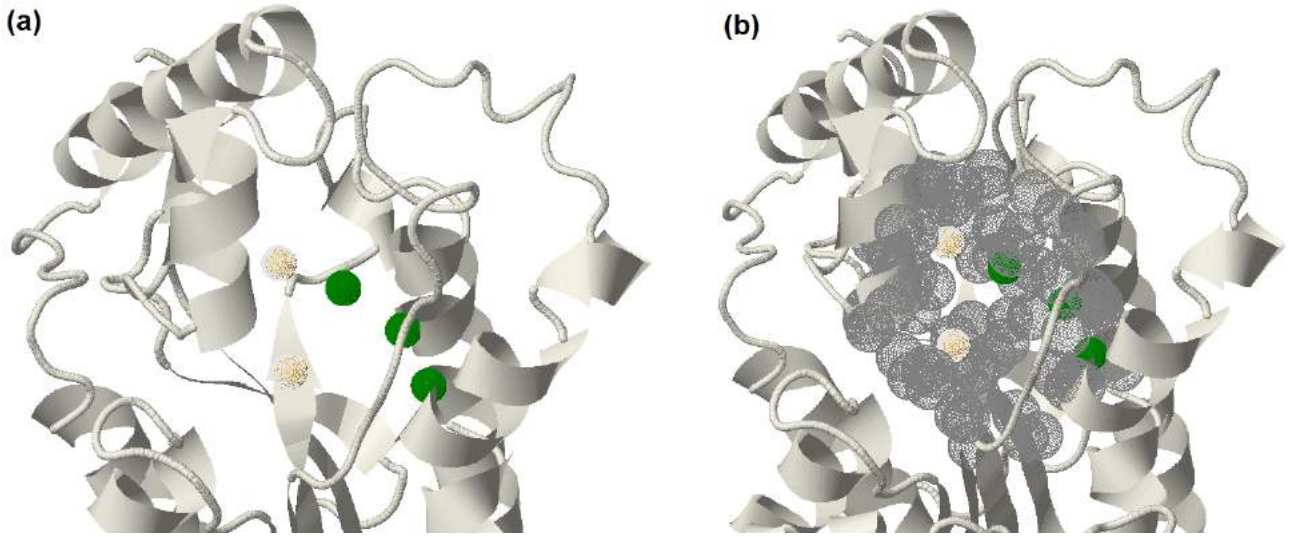


Figura 6.4 – Representação do volume ocupado pelo receptor na avaliação do modelo farmacofórico 3D gerado para a conformação 776 ps da enzima de InhA de *Mtb*. (a) Representação em fitas da estrutura 776 do modelo FFR e 7 propriedades farmacofóricas projetadas dentro da cavidade de ligação do substrato (Amarelo: Do; Branco: Ac; Verde: Hi). (b) Projeção do volume essencial do receptor considerando as mesmas propriedades farmacofóricas definidas no item a.

A Figura 6.5 apresenta a interface gráfica do ZINCPharmer [KC12] que pode ser acessada diretamente (http://zincpharmer.csb.pitt.edu/pharmer.html), sem a necessidade de precisar submeter uma estrutura cristalina antes de chegar nesta área. Assim, as propriedades farmacofóricas geradas pela função heurística envolvidas no modelo são projetadas nessa ferramenta. A partir da projeção dos pontos farmacofóricos do arquivo gerado pela função heurística, o especialista faz uma avaliação do modelo gerado, podendo ajustar o posicionamento, configurar o tamanho do raio da propriedade definida e também inserir novas propriedades. Por fim, o especialista pode abrir o arquivo que restringe a possibilidade de ligantes candidatos sobreporem a região do volume essencial do receptor. Contudo, essa ferramenta limita a abertura de arquivos contendo uma quantidade superior a 25 pontos farmacofóricos. Nos casos quando as quantidades são superiores a 25 pontos, a função heurística irá particionar os arquivos contendo as propriedades.

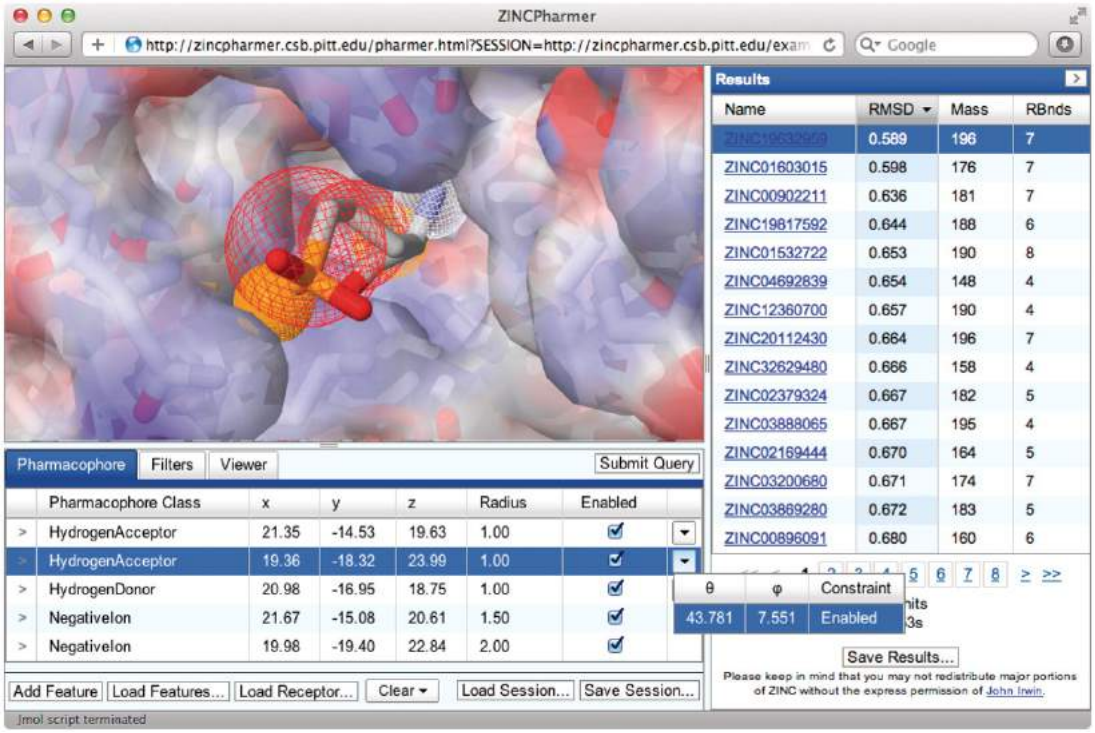


Figura 6.5 – Interface gráfica da ferramenta ZINCPharmer. O menu principal está disposto na base da ferramenta. Na parte da direita estão dispostos os ligantes selecionados conforme as propriedades farmacofóricas definidas [KC12].

6.4.5 Ordenar o conjunto de ligantes selecionados pela probabilidade de se tornarem candidatos a fármacos para a enzima avaliada.

A quantidade de ligantes selecionada pelo ZINCPharmer pode ser elevada, mesmo considerando poucas estruturas da proteína. Desta forma, realizar os experimentos de docagem molecular como todos os ligantes selecionados pode se tornar uma tarefa onerosa. No entanto, se o objetivo é selecionar estruturas para executar experimentos de docagem

molecular com o modelo FFR, uma lista com os ligantes com maior probabilidade de interagir com as estruturas desse modelo FFR pode contribuir na aceleração dessa avaliação. Os principais fatores que podem influenciar esse conjunto são:

- Volume da cavidade de ligação do substrato: As cavidades de ligação do substrato que possuem grandes volumes têm uma probabilidade maior de ancorar pequenas moléculas. Isso porque, elas permitem uma exploração maior da flexibilidade do ligante a ser testado e também aumentam a quantidade de regiões que podem interagir com o ligante.

- Representatividade da estrutura no modelo FFR: Cada uma das 25 estruturas selecionadas representam diferentes quantidades de conformações ocorridas na trajetória do modelo de Receptor Totalmente Flexível. Uma ponderação deve ser avaliada de modo que as estruturas representativas de conjuntos com muitas conformações tenham uma prioridade maior.

A ferramenta do ZINCPharmer fornece um valor do RMSD resultante da análise do erro de posicionamento das propriedades existentes nos ligantes selecionados com o modelo farmacofórico. No entanto, a projeção do modelo farmacofórico possui um limiar de incerteza na determinação dos pontos devido a abstração aplicada. Então, para se aplicar uma ordenação baseada nesta análise de erro seria necessário, primeiramente, reduzir o grau de incerteza da projeção do modelo. Assim, a ordenação dos ligantes a serem avaliados é baseada apenas no volume da cavidade de ligação do substrato e na representatividade da estrutura no modelo FFR. A próxima seção descreve a avaliação do método apresentado neste capítulo.

6.5 Avaliação do método desenvolvido

Para avaliar o método descrito, as estruturas representativas dos conjuntos de estruturas similares foram comparadas com a estrutura cristalina 1ENY [DQB+95] (Figura 6.6). Essa estrutura cristalina foi escolhida por ser a estrutura molde que originou o modelo FFR de 19,5 ns. Desta forma, essa avaliação simula um processo ideal: (1) uma estrutura cristalina somente com a coenzima NADH é escolhida; (2) um modelo FFR é gerado para simular a flexibilidade dessa proteína; e (3) o método descrito nessa tese é aplicado para evidenciar as regiões flexíveis do modelo FFR que a estrutura cristalina não contempla. O item 3 se torna fundamental a medida que os modelos flexíveis são cada vez mais extensos.

Nessa avaliação, primeiramente é caracterizada a hipótese de trabalho, sendo após apresentado os conjuntos de experimentos de docagem molecular utilizados nessa avaliação.

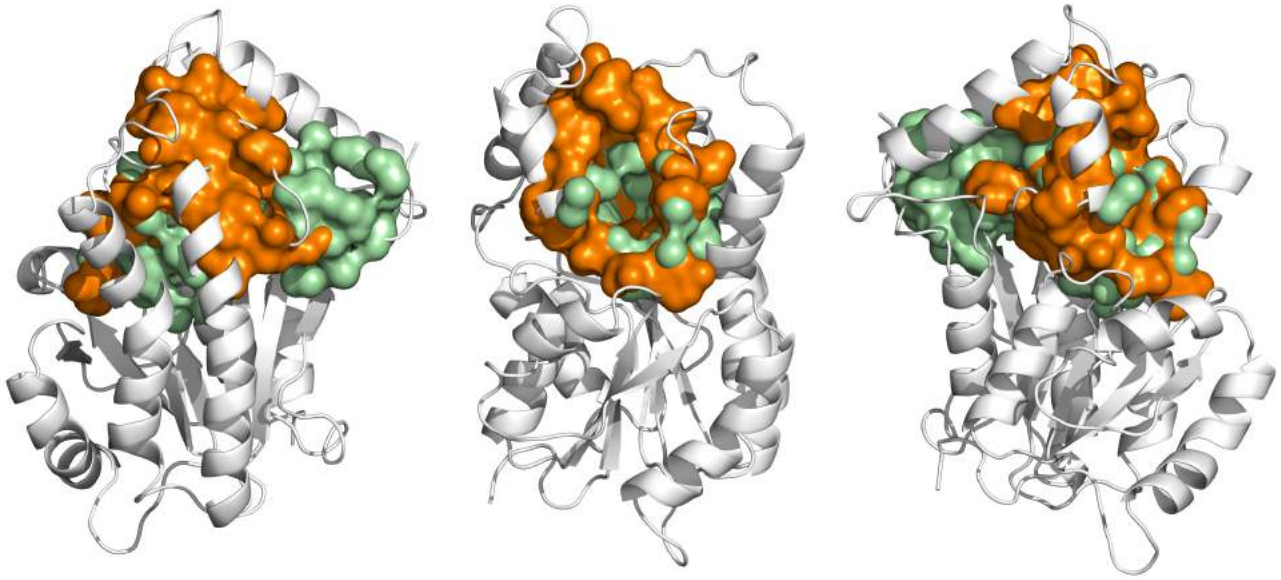


Figura 6.6 – Representação volumétrica da cavidade de ligação do substrato da estrutura 776 ps do modelo FFR de InhA (verde claro) e a estrutura cristalina 1ENY [DQB+95] (laranja) em 3 diferentes poses. A visualização frontal da cavidade (imagem do meio) mostra uma redução de altura do volume acessível da cavidade de ligação do substrato da estrutura 776 em relação à estrutura cristalina.

6.5.1 Caracterização formal da hipótese

Nesta tese, o método proposto tem como objetivo explorar o máximo da flexibilidade proporcionada pelo modelo FFR com a utilização de um conjunto de estruturas representativas. Desta forma, a comparação deste método é feita com a estrutura cristalina que originou o modelo FFR, elencando inicialmente a seguinte hipótese:

1. **Hipótese Nula**: H_0: Os ligantes selecionados apresentam resultados de energia livre de ligação (FEB) aproximados na avaliação dos experimentos de docagem molecular considerando a estrutura cristalina e as 25 estruturas representativas do modelo FFR de 19,5 ns. Neste caso, podemos inferir que os desempenhos obtidos são equivalentes ($FEB_{cristalina} \cong FEB_{representativas}$).

2. **Hipótese Alternativa**, H_1: Existe pelo menos um resultado na avaliação dos experimentos de docagem molecular que apresenta o valor de FEB com diferença significativa da estrutura cristalina, sendo adotado o limiar de até 1 kcal/mol entre as moléculas [HMOG07]. Esse valor ocorre devido ao acesso de alguma região flexível.

Caso a Hipótese Nula seja refutada em favor da Hipótese Alternativa, é realizada uma validação por inspeção visual para analisar as regiões de contato referentes a melhor orientação definida pelo experimento de docagem molecular. A próxima subseção descreve os experimentos de avaliação do conjunto de ligantes selecionados a partir da análise das 25 estruturas representativas.

6.5.2 Experimentos de avaliação do conjunto de ligantes selecionados no ZINCPharmer

As 25 estruturas representativas do modelo FFR foram comparadas com as estrutura cristalina 1ENY, resultando em um conjunto de 25 modelos farmacofóricos 3D. Esse conjunto foi avaliado na ferramenta ZINCPharmer. Os pontos farmacofóricos definidos como a hipótese de busca no BD foram selecionados de modo a possibilitar a identificação de conjuntos de ligantes promissores para realizar os experimentos de docagem molecular. A Figura 6.7 apresenta 4 modelos farmacofóricos do conjunto total (os modelos farmacofóricos 3D gerados estão descritos no Apêndice A).

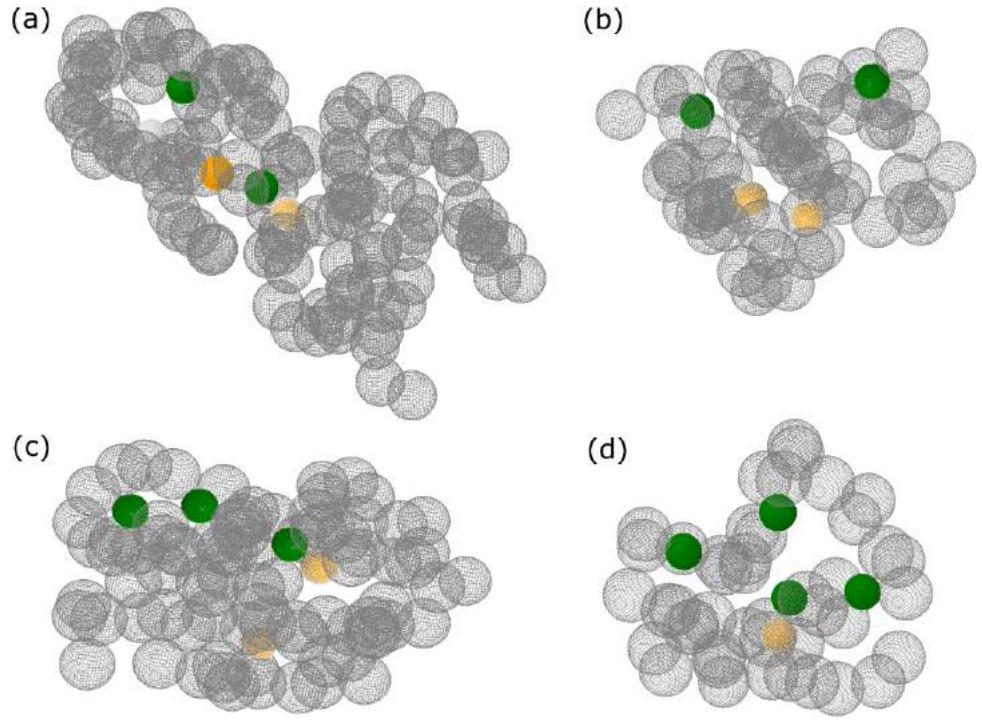


Figura 6.7 – Modelos farmacofóricos 3D de 4 estruturas representativas utilizadas no processo de triagem virtual dos ligantes na ferramenta ZINCPharmer (Verde: Hi; Branco: Ac; Amarelo: Do e Cinza: volume essencial do receptor). Os modelos farmacofóricos (a), (b), (c) e (d) correspondem as estruturas 776, 1.099, 1.407 e 3.441 do modelo FFR respectivamente e foram resultados da comparação com a estrutura cristalina 1ENY [DQB+95].

Um resumo da seleção realizada no ZINCPharmer está descrito na Tabela 6.1. Essa tabela também apresenta detalhes das estruturas representativas do modelo FFR, tais como o volume da cavidade alvo, o percentil da representabilidade da partição no modelo FFR e a quantidade de ligantes selecionados pela ferramenta ZINCPharmer. Neste exemplo, não houve a edição das propriedades projetadas pelo método desenvolvido nesta tese, havendo apenas o descarte de algumas propriedades quando a busca se tornava muito específica. Hipóteses farmacofóricas muito permissivas foram descartadas (o limite superior estabelecido foi de $\sqrt{n}$, sendo n a quantidade de conformações do modelo FFR). Esses critérios têm como objetivo testar as hipóteses definidas pelo método desenvolvido.

Tabela 6.1 – Características das 25 estruturas representativas utilizadas no processo de triagem virtual dos ligantes, descrevendo o volume e representabilidade dessas estruturas. A última coluna descreve a quantidade de ligantes selecionados após a aplicação da hipótese farmacofórica na ferramenta ZINCPharmer.

Conformação	Volume da cavidade	Representabilidade	Ligantes selecionados
776	999,2	5,3%	109
1.099	803,2	0,5%	7
1.407	498,2	5,3%	17
1.437	533,0	3,0%	7
2.029	1.241,6	1,0%	63
2.065	1.052,5	1,6%	18
2.248	1.346,3	2,0%	22
3.029	1.084,0	0,5%	115
3.323	219,7	5,1%	4
3.360	479,5	1,7%	50
3.441	427,1	6,7%	53
3.457	409,9	0,8%	8
5.561	575,5	1,7%	98
13.191	1.196,9	0,6%	122
14.933	981,4	2,1%	16
14.968	978,3	1,1%	2
15.035	1.075,4	0,5%	3
15.697	1.297,4	1,3%	18
16.577	1.535,6	0,6%	129
17.114	1.214,3	4,5%	24
17.503	729,9	1,3%	62
17.618	924,3	0,6%	8
18.259	1.699,9	0,5%	25
18.540	674,8	0,7%	1
19.093	1.118,1	1,1%	2
		Total:	957

No total, nesta avaliação foram selecionados 983 ligantes para os experimentos de docagem molecular, sendo apenas 26 ligantes selecionados em mais de um modelo farmacofórico. Removendo os ligantes redundantes, os experimentos de docagem molecular avaliaram um conjunto de 957 ligantes. A composição desse conjunto de ligantes selecionados provém de características distintas identificadas nas estruturas representativas. Assim, a quantidade de compostos selecionados por cada modelo farmacofórico pode variar conforme as propriedades avaliadas. A Tabela 6.1 apresenta a quantidade de ligantes selecionada por cada estrutura representativa.

Os experimentos de docagem molecular foram realizados para avaliar a qualidade da interação do conjunto de 957 ligantes selecionados. Essa avaliação analisa a interação com a estrutura cristalina da 1ENY e com as 25 estruturas representativas. A preparação dos arquivos da proteína e do ligante seguiu o protocolo definido por [HMF12], utilizando os algoritmos de preparação automática fornecidos pelo pacote do AutoDockTools [MHL+09].

Esses experimentos foram baseados no algoritmo genético Lamarckiano contendo um número máximo de 27.000 gerações ou 1.500.000 de avaliações de energia, mantendo sempre o melhor indivíduo a cada geração. Devido a preparação automática, as ligações rotacionáveis identificadas são definidas como ativas e o número máximo de execuções para 20. Houve uma redução no número de gerações e no números de execuções quando comparados com a preparação dos arquivos de docagem molecular descritos na seção 4.1.3, sendo anteriormente considerados 3.000.000 de avaliações de energia e 25 execuções. Essa alteração foi necessária devido a quantidade de ligantes a serem testados e a avaliação de testes iniciais que apontaram valores de FEB similares mesmo com a redução de 50% do número de gerações e 25% no número de execuções.

6.5.3 Avaliação dos experimentos de docagem molecular

A primeira avaliação dos experimentos de docagem molecular contabilizou a quantidade de ligantes que obtiveram resultados de FEB negativos. Quanto mais negativos forem os resultados da interação do complexo, mais forte será a interação do ligante com o receptor [Dun95]. A Tabela 6.2 apresenta um resumo descrevendo os valores obtidos considerando a estrutura cristalina e as estruturas representativas do modelo FFR. Esses resultados também mostram que o conjunto de ligantes selecionados pelo método apresentado nesta tese obteve acima de 95% de resultados favoráveis que geraram FEB negativas (tanto com a estrutura cristalina quanto com as estruturas representativas).

Tabela 6.2 – Resultado da avaliação dos experimentos de docagem molecular da estrutura cristalina e das estruturas representativas com os 957 ligantes selecionados do BD ZINC.

	Negativos	Positivos
Representativas	908 (95,0%)	48 (5,0%)
Cristalinas	924 (96,7%)	32 (3,3%)

Huey e colaboradores [HMOG07] realizaram um conjunto de experimentos avaliando complexos receptor-ligante, cujo objetivo era o de definir a precisão da FEB do programa AutoDock. Os resultados encontrados apresentaram uma predição correta de até 2.0 Å da pose da estrutura cristalina com uma variação de até 1 kcal/mol entre as moléculas com mais de 80% dos casos. Assim, a avaliação apresentada na c foi separada em 4 categorias:

- Vitória da estrutura representativa: são as estruturas que apresentaram uma diferença significativa favorável considerando o resultado da estrutura cristalina.
- Empate considerando a estrutura representativa: contabiliza os resultados que obtiveram melhores valores de FEB que a estrutura cristalina, no entanto não se mostrou superior a margem definida por Huey [HMOG07].

- Empate considerando a estrutura cristalina: o cálculo é similar ao empate das estruturas representativas, apenas considerando a avaliação dos valores de FEB da estrutura cristalina que não se mostraram superior a margem de 1 kcal/mol.
- Vitória da estrutura cristalina: são as estruturas que apresentaram uma diferença significativa favorável considerando o resultado das estruturas representativas.

Tabela 6.3 – Análise dos resultados dos experimentos de docagem molecular das 25 estruturas representativas e a estrutura cristalina com os 957 ligantes selecionados do BD ZINC, considerando a variação de até 1 kcal/mol entre os valores de FEB das moléculas.

Vitória	Empate		Vitória
Representativa	Representativa	Cristalina	Cristalina
< 1.0	1.0 > & < 0.0	1.0 > & < 0.0	< 1.0
197 (**20,6%**)	418 (43,7%)	255 (26,6%)	87 (9,1%)

A Tabela 6.3 mostra que mais de 64% dos melhores resultados de FEB foram obtidos com o conjunto de estruturas representativas, sendo aproximadamente 20% com diferença acima da margem definida pelo programa de docagem molecular utilizado. Esses resultados elencam um conjunto de pequenas moléculas que resultaram em bons valores de FEB e com diferenças significativas com relação a estrutura cristalina utilizada como molde do modelo FFR. Desta forma, a hipótese nula é rejeitada em favor da hipótese alternativa, cujo resultado obtido apresentou um valor de FEB com diferença significativa da estrutura cristalina. A seguir é realizada uma avaliação para analisar o conjunto de ligantes que apresentaram as maiores diferenças entre os resultados obtidos. Primeiramente, a avaliação considera três dos quatro ligantes que apresentaram as maiores vantagens com relação a estrutura cristalina[2].

A Figura 6.8 ilustra os 3 modelos farmacofórios utilizados para a seleção do conjunto de ligantes a serem testados. Esses pontos foram editados na ferramenta ZINCPharmer a partir da geração do conjunto de pontos farmacofóricos identificados nas conformações 15.035 (Figura 6.8-a), 13,191 (Figura 6.8-b) e 17,503 (Figura 6.8-c). Para cada um desses modelos foi selecionado um ligante como exemplo. Os ligantes ZINC31167913, ZINC75714056 e ZINC56919632 correspondem aos modelos gerados a partir das conformações 15.035, 13,191 e 17,503 respectivamente.

[2]O ligante que apresentou a maior diferença dos valores de FEB foi omitido desta tese devido ao fato deste ligante encontrar-se em fase de análise experimental. A avaliação deste ligante com a estrutura cristalina não apresentou bons resultados, resultando em execuções com o valor positivo de FEB. Segundo uma pesquisa realizada no DrugBank [KLJ+11], essa pequena molécula identificada pelo método apresentado nesta tese é um conhecido ligante inibidor, cujo antibiótico já está aprovado. No entanto, nenhuma relação desse fármaco foi encontrada com a enzima InhA de *Mtb*. Desta forma, o nome desse ligante será preservado até o término da investigação experimental que está sendo realizada.

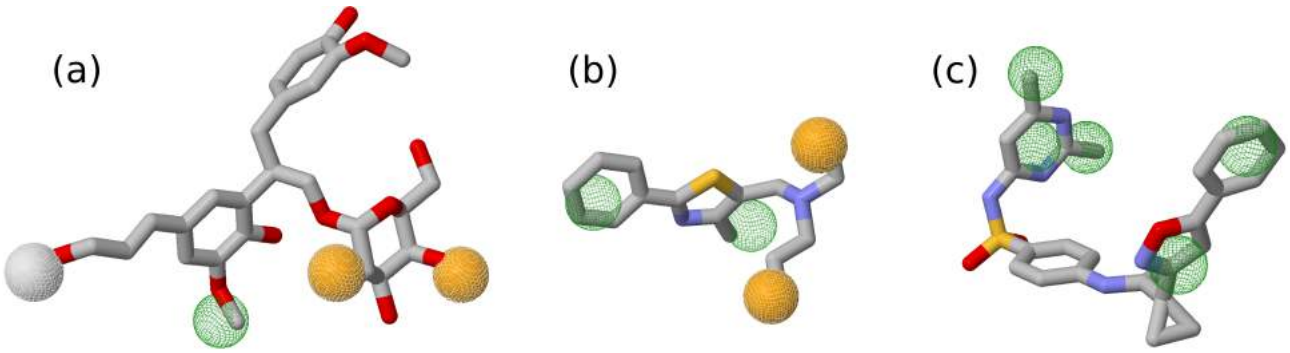


Figura 6.8 – Representação dos modelos farmacofóricos 3D gerados a partir das estruturas 15.035, 13.191 e 17.503 do modelo FFR em esferas de van der Waals (Amarelo: Do; Branco: Ac; e Verde: Hi). Os ligantes selecionados encaixam-se adequadamente nas propriedades farmacofóricas das estruturas representativas. Os ligantes de cada modelo farmacofórico são representados em palitos coloridos conforme o tipo de átomo (carbono em cinza, nitrogênio em azul, oxigênio em vermelho e enxofre em amarelo). (a) Modelo extraído da conformação 15.035 e selecionando o ligante ZINC31167913. (b) Modelo extraído da conformação 13.191 e selecionando o ligante ZINC75714056. (c) Modelo extraído da conformação 17.503 e selecionando o ligante ZINC56919632.

Os 3 modelos farmacofóricos submetidos à avaliação do ZINCPharmer continham a descrição das esferas que caracterizam a área essencial do receptor dos respectivos átomos que delimitam a cavidade de ligação do substrato. No entanto, esses conjuntos de esferas foram omitidos da Figura 6.8 para permitir uma visualização clara dos pontos farmacofóricos extraídos da cavidade de ligação do substrato e a sobreposição das propriedades dos ligantes que foram alinhadas com a hipótese farmacofóricas.

Os valores resultantes dos experimentos de docagem molecular desses ligantes foram disponibilizados na Tabela 6.4. Essa tabela apresenta os resultados das melhores interações identificadas no conjunto de estruturas representativas e com a estrutura cristalina 1ENY para cada ligante avaliado nesta seção. A maior diferença encontrada foi de 3,4 kcal/mol entre a docagem molecular do ligantes ZINC31167913 com a estrutura cristalina e com a estrutura representativa 18.540. A Figura 6.9 mostra as poses de melhor energia encontradas pelo programa de docagem molecular para cada um dos 3 ligantes analisados na Tabela 6.4. Essa figura destaca as posições que apresentaram os melhores valores de FEB devido ao acesso das regiões flexíveis disponíveis nas estruturas representativas.

Tabela 6.4 – Comparação dos resultados dos experimentos de docagem molecular realizados com os ligantes ZINC31167913, ZINC75714056 e ZINC56919632 com as estruturas cristalinas e as respectivas estruturas representativas que resultaram na melhor interação.

Ligantes	Representativas (kcal/mol)	Cristalinas (kcal/mol)	Diferença (kcal/mol)
ZINC31167913	-8,8	-5,4	3,4
ZINC75714056	-8,7	-6,5	2,2
ZINC56919632	-10,4	-7,7	2,7

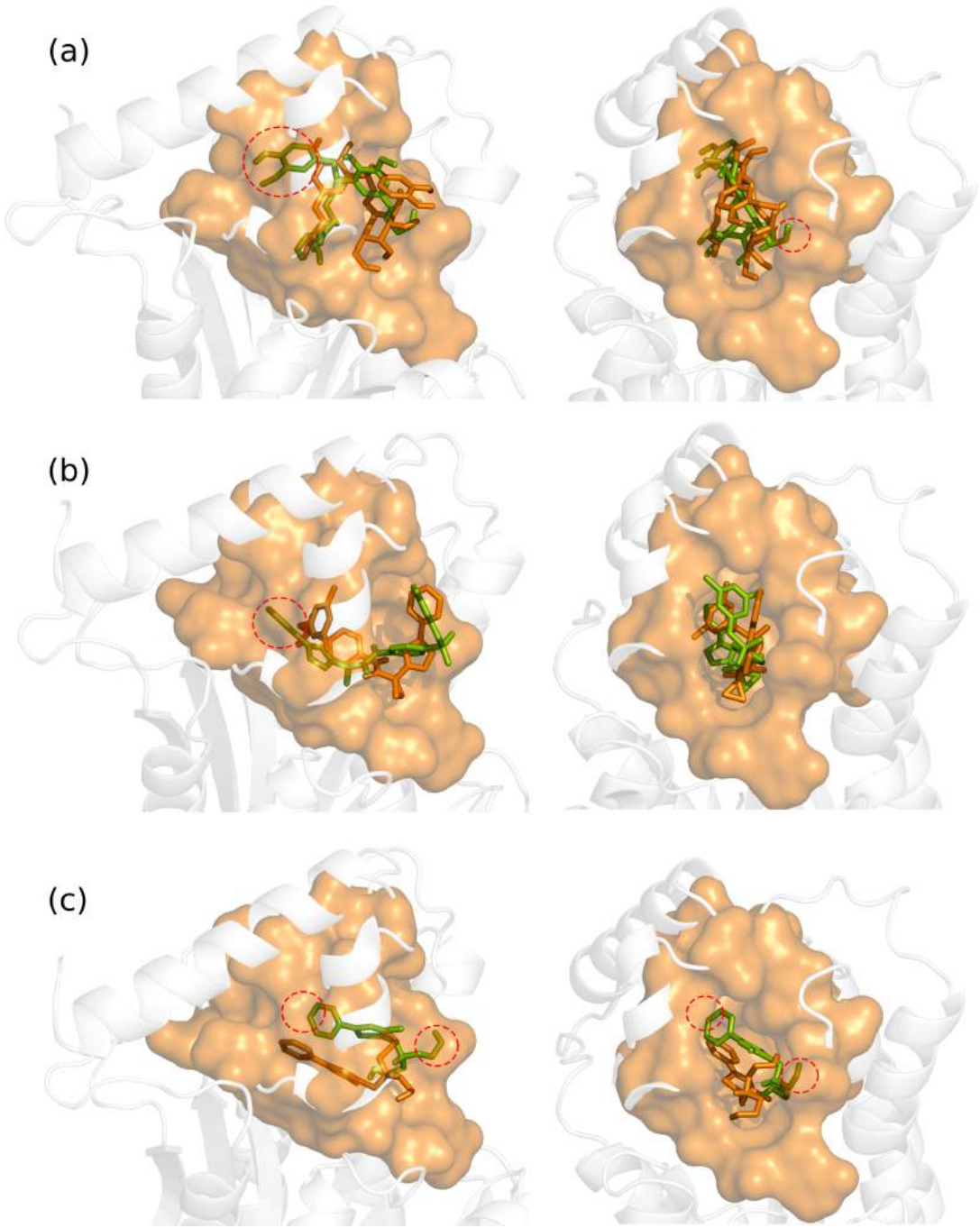


Figura 6.9 – Comparação da pose final de cada experimento de docagem molecular entre os 3 ligantes analisados e o conjunto de estruturas representativas do modelo FFR e a estrutura cristalina 1ENY. A estrutura cristalina 1ENY está na representação em fitas e os átomos que delimitam a cavidade de ligação do substrato desta enzima estão na representação volumétrica (ocre). Na representação de palitos estão os ligantes com melhores poses definidas pelo algoritmo de docagem molecular nas avaliações com as respectivas estruturas representativas (verde) e com estrutura cristalina 1ENY (laranja). Os círculos pontilhados (vermelho) destacam as regiões acessíveis devido a flexibilidade do modelo FFR que resultaram em melhores valores de FEB que os resultados obtidos pela estrutura cristalina. (a) Avaliação do ligante ZINC31167913 com a estrutura cristalina 1ENY e a conformação 18.259. (b) Avaliação do ligante ZINC75714056 com a estrutura cristalina 1ENY e a conformação 17.618. (c) Avaliação do ligante ZINC56919632 com a estrutura cristalina 1ENY e a conformação 2.029.

Além dos experimentos de docagem molecular, também foram realizadas as avaliações desses três ligantes utilizando o programa LigPlot+ [LS11] (Figura 6.10, Figura 6.12 e a Figura 6.11). Essas avaliações mostram a comparação das poses finais dos experimentos de docagem molecular entre a estrutura cristalina e a estrutura representativa.

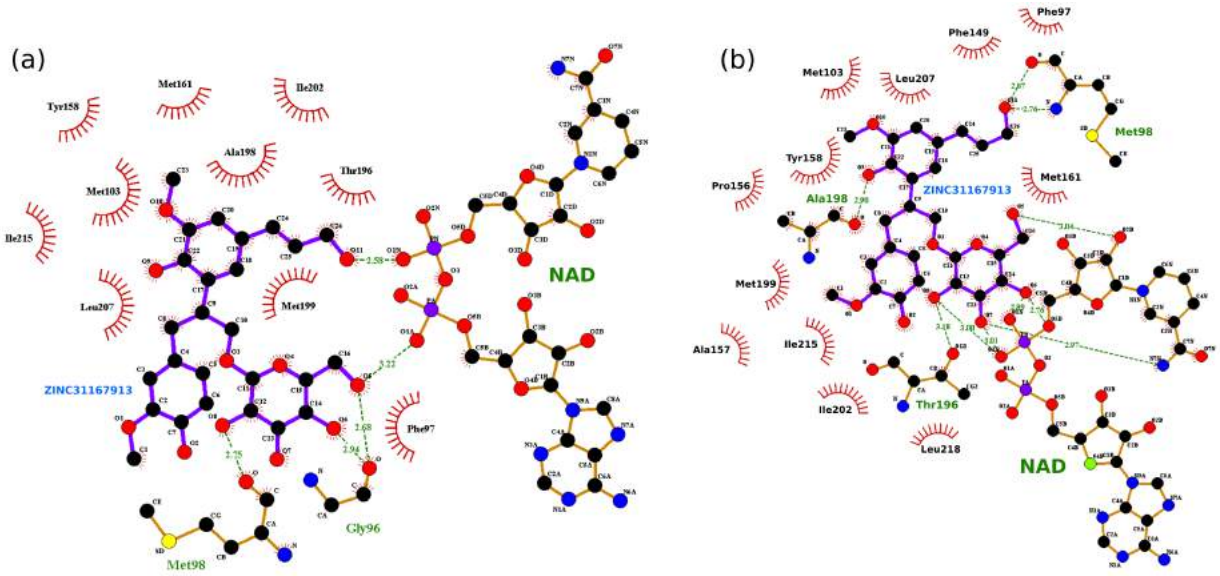


Figura 6.10 – Representação 2D mostrando a interação entre o complexo formado pelo ligante ZINC31167913 e a conformação 18.259 do modelo FFR gerada automaticamente pelo programa LigPlot+ [LS11].

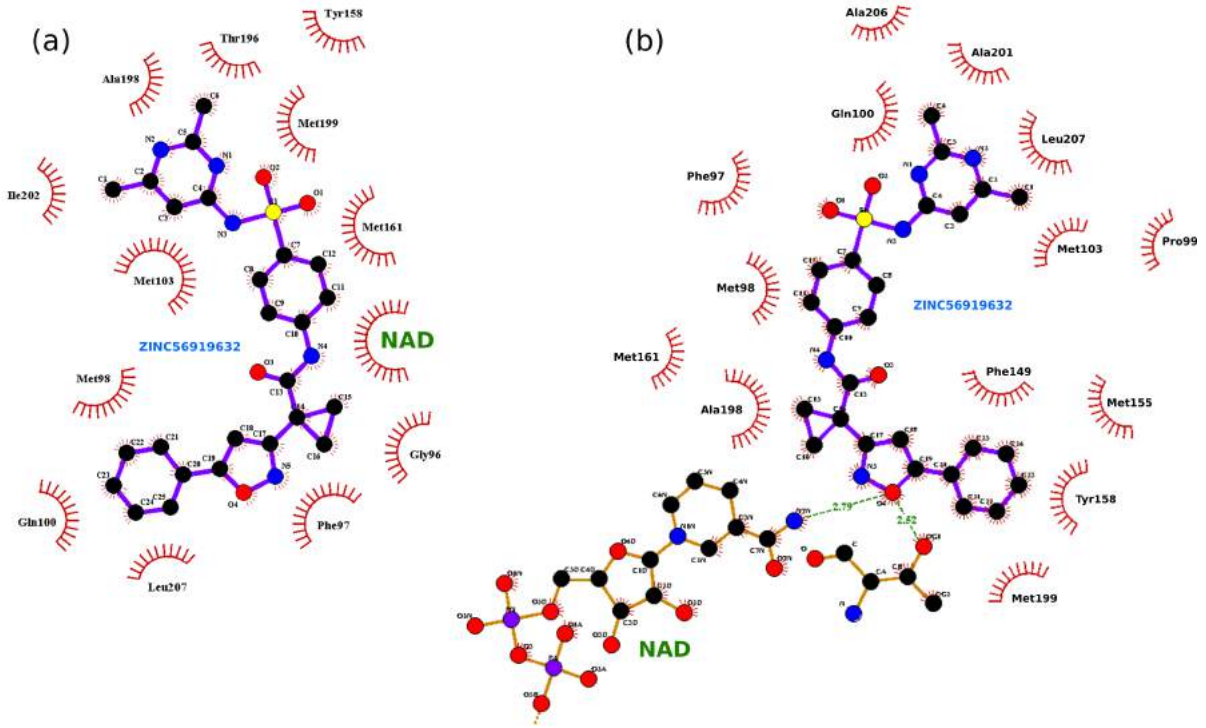


Figura 6.11 – Representação 2D mostrando a interação entre o complexo formado pelo ligante ZINC56919632 e a conformação 2.029 do modelo FFR gerada automaticamente pelo programa LigPlot+ [LS11].

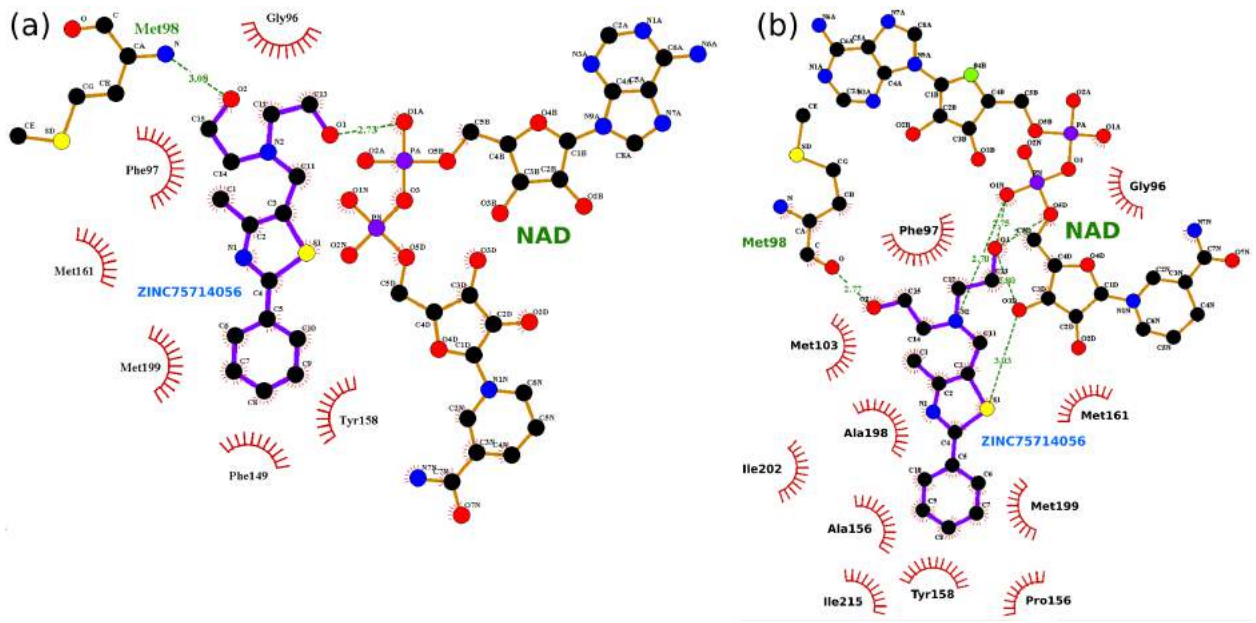


Figura 6.12 – Representação 2D mostrando a interação entre o complexo formado pelo ligante ZINC75714056 e a conformação 17.618 do modelo FFR gerada automaticamente pelo programa LigPlot+ [LS11].

As avaliações realizadas com o programa Ligplot+ mostraram que os ligantes ZINC56919632 e ZINC31167913 possuem duas interações por pontes de hidrogênio com o anel da nicotinamida da coenzima NAD. Essas importantes interações foram identificadas somente nos experimentos de docagem molecular com as estruturas representativas. Além disso, nesses dois casos, as interações desses ligantes correspondem exatamente aos resíduos que projetaram essas propriedades na cavidade de ligação do substrato.

6.6 Considerações finais.

Os resultados encontrados reforçam a importância da análise da flexibilidade das proteínas. O método apresentado nesta tese identificou um conjunto de ligantes que se mostraram mais favoráveis à interagir com pelo menos uma das estruturas representativas do modelo FFR do que com a estrutura cristalina utilizada como molde do modelo FFR avaliado. Também foram identificados casos onde a estrutura cristalina continuou sendo a melhor opção para os ligantes selecionados. Basicamente, isso pode haver ocorrido devido:

- Limitação das 25 estruturas representativas: O conjunto de estruturas representativas é uma seleção resumida de um modelo FFR com 19.500 conformações. Ou seja, por mais que a redução seja bastante distribuída, nada pode substituir totalmente o modelo FFR. Uma avaliação interessante seria a execução dos experimentos de docagem molecular de forma exaustiva com o modelo FFR ou com a utilização do P-SaMI. Esse experimento mostraria ao pesquisador se o conjunto de estruturas atende aproximadamente os resultados atingidos com as estruturas cristalinas.

- Modelo FFR limitado: Um modelo de receptor totalmente flexível deveria ser capaz de simular a estrutura cristalinas que o gerou em algum instante de tempo ao longo da trajetória. O fechamento das alças da proteína e/ou a projeção de parte da estrutura da coenzima para o interior da cavidade alvo causam uma grande restrição no encaixe de pequenas moléculas, conforme pode ser visto na Figura 6.6 e 6.13.

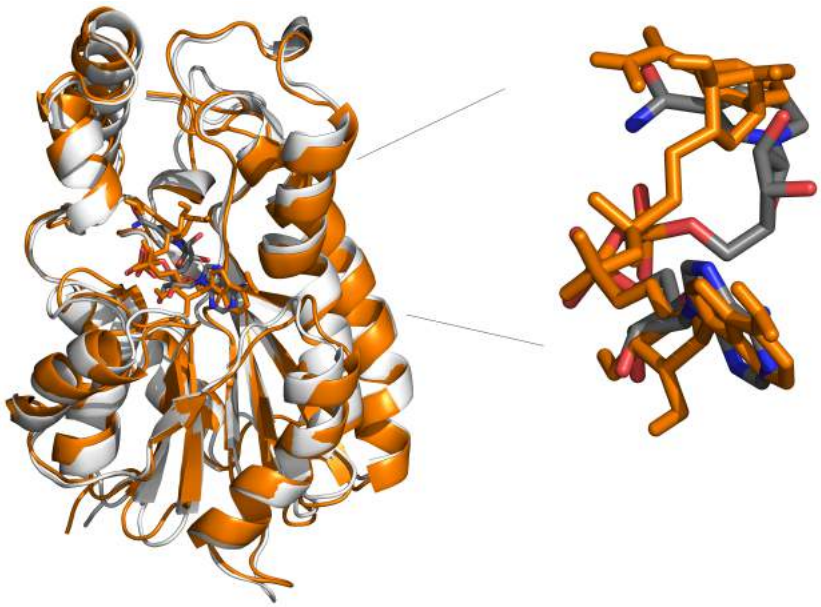


Figura 6.13 – Representação em fitas da enzima InhA da estrutura cristalina 1ENY (branco) e da estrutura representativa 16.577 (ocre) mostrando a diferença estrutural ocasionada pela flexibilidade. As coenzimas NADH estão dispostas na representação de palitos, sendo atribuída a coenzima da estrutura representativa 16.577 a cor ocre e a estrutura cristalina 1ENY na cor conforme o tipo de átomo (carbono em cinza, nitrogênio em azul, oxigênio em vermelho e fósforo em ocre) A projeção da coenzima para o interior da cavidade de ligação do substrato na conformação 16.577 tende a restringir o acesso dos ligantes.

Embora existam muitos pontos a serem aprimorados, o método desenvolvido mostrou resultados promissores ao elencar o conjunto de hipóteses farmacofóricas complementares a avaliação do conjunto de estruturas cristalinas. Os experimentos comprovam a possibilidade de definir um conjunto de propriedades da cavidade de ligação do substrato sem haver a restrição da necessidade de ligantes complexados com a molécula alvo. Naturalmente, a seleção de *algumas* dessas estruturas poderia ser realizada com um criteriosa etapa de geração de confôrmeros de cada ligante (por exemplo, utilizando programas como o CORINA ou OMEGA). Ainda assim, haveria uma grande limitação ao restringir os esforços apenas na geração de grandes repositórios de confôrmeros, visto que os conjuntos de estruturas cristalinas disponíveis de cada proteína não são suficientemente grandes para representar uma grande parte dos movimentos flexíveis possíveis dessas proteínas. Desta forma, os modelos FFR são uma oportunidade para simular a total flexibilidade da proteína e, em conjunto, o método elaborado nesta tese pode explorar as regiões flexíveis que os métodos baseados somente em estrutura cristalina não alcançam.

7. Trabalhos Relacionados

Existem diversas abordagens na literatura baseadas na investigação das propriedades farmacofóricas da estrutura do receptor. De maneira geral, essas abordagens podem ser divididas em pesquisas baseadas apenas nas informações das estruturas cristalinas e outras que investigam modelos FFR buscando incrementar seus modelos de forma a considerar a flexibilidade das proteínas. As próximas seções descrevem um conjunto de trabalhos que estão relacionados à esta pesquisa.

7.1 Abordagens considerando somente estruturas rígidas para a geração de modelos farmacofóricos 3D

Modelos farmacofóricos 3D avaliando conjuntos de estruturas cristalinas têm sido amplamente utilizados no processo de triagem virtual de ligantes, selecionando ligantes que contenham o arranjo espacial de propriedades físico-químicas essenciais obtidas da avaliação das interações similares identificadas de complexos receptor-ligante conhecidos [LGLT10]. Barillari e colaboradores [BMR08] desenvolveram um método para a triagem de ligantes baseado em um critério de relevância das cavidades com maior probabilidade de obter boas interações. As estruturas mais relevantes são identificadas a partir da utilização de técnicas de aprendizado de máquina treinadas com um conjunto de estruturas complexadas com ligantes conhecidos. Contudo, os modelos produzidos por essa abordagem são limitados por não avaliarem as regiões flexíveis da proteína e por considerarem somente propriedades físico-químicas que já possuam ligantes complexados. Abordagens que geram modelos farmacofóricos baseados em ligantes são dependentes das características físico-químicas presentes nos complexos receptor-ligante conhecidos. Desta forma, as proteínas que não possuem ligantes complexados conhecidos e as propriedades físico-químicas que não estabelecem interação nos complexos avaliados podem ser negligenciadas pelo modelo farmacofórico gerado.

Outro relevante trabalho nesta área foi apresentado por Tintori [TCM+08], cujo trabalho descreve um protocolo para a geração de hipóteses farmacofóricas 3D a partir de uma estrutura cristalina *apo*. Assim como no trabalho desenvolvido por Barillari [BMR08], Tintori também utiliza uma criteriosa avaliação dos campos de interação molecular da proteína. Esses campos de interação molecular são geralmente utilizados para caracterizar as moléculas de acordo com os seus sítios de interação favoráveis, permitindo uma projeção sobre como os ligantes podem interagir com a cavidade alvo [ACC+13]. Normalmente, essa é a primeira etapa na derivação de modelos farmacofóricos baseados unicamente na estrutura da proteína [HL12, CC11]. No cálculo desses campos de interação molecular, uma malha 3D é projetada sobre a cavidade de ligação do substrato e as energias de interação

de cada ponto da malha são calculadas entre a proteína e diversas forças moleculares, cada uma com diferentes propriedades físico-químicas. No entanto, avaliações muito elaboradas necessitam de muito tempo computacional para serem aplicadas, fato que limita a consideração da flexibilidade das estruturas a serem analisadas.

7.2 Abordagens considerando a avaliação da flexibilidade a partir de modelos FFR para a geração de modelos farmacofóricos 3D

Segundo Teodoro [TK03] e Alonso [ABG06], as simulações por modelos FFR possibilitam representar o comportamento natural do receptor em ambientes flexíveis, aumentando a quantidade de conformações e, desta forma, elevando a probabilidade de encontrar complexos receptor-ligante com melhores interações. Assim, métodos alternativos têm sido desenvolvidos visando considerar os movimentos flexíveis da cavidade de ligação do substrato para gerar modelos farmacofóricos 3D [DLS+05, DSNB06, HL13]. Deng [DLS+05, DSNB06] apresentou dois trabalhos visando incorporar parte da flexibilidade da proteína da HIV-1 *integrase* (trajetória com 1 ns). Em [DLS+05], Deng baseou a geração das hipóteses farmacofóricas na utilização de 10 estruturas selecionadas pela comparação dos valores do RMSD. A seleção desse conjunto de estruturas representativas, ao ser baseada somente nos valores do RMSD, torna-se uma limitação importante desse trabalho, visto que torna possível a formação de partições contendo conjuntos de estruturas com cavidades de ligação do subtrato distintas entre si. O principal problema dessa abordagem está na seleção das estruturas representativas a partir de conjuntos particionados pelas informações do RMSD. Os dois principais problemas desse tipo de abordagem são:

- ao avaliar somente as propriedades das estruturas representativas: selecionar estruturas desse tipo de agrupamento seria o mesmo que selecionar cavidades aleatórias, visto que conformações com diferentes valores de RMSD podem resultar em cavidades de ligação muito similares (havendo a alteração estrutural em uma região longe da cavidade alvo). Assim, não existe garantia que este tipo de abordagem realmente consiga explorar a flexibilidade do modelo FFR.

- ao avaliar as principais posições dos átomos das cavidades de cada agrupamento: esse tipo de abordagem avalia os resíduos mais frequentes da cavidade de ligação, contudo o agrupamento de cavidades distintas em uma mesma partição aumenta a probabilidade da geração de resultados falsos positivos. Isso pode ocorrer devido a combinação de conjuntos de resíduos que somente ocorrem em conformações distintas. Desta forma, a hipótese gerada possibilitaria a geração de um arranjo que não é aceito por nenhuma conformação da partição.

No trabalho apresentado em [DSNB06], Deng alterou sua estratégia para seleção das estruturas representativas, definindo o conjunto de estruturas representativas a partir de tempos equidistantes na trajetória. Embora o método de seleção tenha sido alterado, o problema de não considerar boa parte da flexibilidade persistiu. Além disso, ambos os métodos propostos visando considerar a flexibilidade utilizando informações de modelos FFR de 1 ns têm optado por simplificações na seleção de 5 até 10 estruturas representativas para a geração das hipóteses farmacofóricas, fato que restringe o espaço de busca da flexibilidade da proteína.

7.3 Trabalhos desenvolvidos nos grupos de pesquisa GPIN e LABIO

Nosso grupo de pesquisa tem apresentado importantes contribuições com a finalidade de reduzir o custo computacional envolvido no teste de novos candidatos a fármaco com o modelo FFR, sem perder informações críticas. Estudos apresentados previamente em [MWRNdS11, Hüb10, DPFNdSR13] têm descrito aplicações focadas na redução da dimensão dos modelos FFR, denominados como modelos de Receptor Totalmente Flexível Reduzido (RFFR) [DPFNdSR13]. Outros trabalhos como [Que11] e [Pau11] caracterizaram abordagens para selecionar conjuntos de ligantes específicos para um receptor alvo considerando as características da cavidade de ligação do substrato.

Em [Pau11], um conjunto de novos candidatos a inibidores são identificados após a seleção de ligantes contendo propriedades físico-químicas observadas a partir de uma avaliação de mapas farmacofóricos das interações do complexo receptor-ligante do receptor alvo. Esse relevante trabalho resultou no registro de novas patentes de alguns dos inibidores encontrados. Importantes descobertas têm sido feitas com a utilização de mapas farmacofóricos na seleção de ligantes na última década [CKW$^+$09, Pau11, NQT$^+$14].

Embora essas abordagens realizem reduções na quantidade de ligantes a serem testados com base em uma avaliação das propriedades físico-químicas da cavidade de ligação do substrato, os critérios identificados pelos mapas farmacofóricos são pesquisados nos BD de ligantes, desconsiderando a disposição 3D onde cada propriedade foi catalogada. O fator positivo dessa abordagem é o tempo necessário de retorno das consultas que ocorre de modo eficiente. No entanto, muitos ligantes selecionados não apresentam bons resultados de FEB devido à disposição 3D das características identificadas.

Quevedo [Que11] descreveu um filtro baseado nas propriedades geométricas da cavidade de ligação do substrato para determinar se os ligantes testados continham possibilidade de estabelecer interação com o receptor. Embora os resultados apresentados por essa heurística tenham sido promissores, existem limitações importantes quanto à determinação das coordenadas adequadas do centro das esferas concêntricas e com relação ao grande número de estruturas que poderiam ser descartadas. Outra limitação importante deve-se ao fato dessa função desenvolvida ser baseada somente na possibilidade de haver

o encaixe geométrico do complexo fármaco-receptor. Assim, muitos resultados aprovados pela função heurística acabaram não gerando bons resultados de FEB. Desta forma, há uma série de melhorias que podem tornar o filtro desenvolvido em [Que11] mais eficaz.

7.4 Considerações finais

Modelos farmacofóricos 3D avaliando conjuntos de estruturas cristalinas têm sido amplamente utilizados no processo de triagem virtual de ligantes, selecionando ligantes que contenham o arranjo espacial de propriedades físico-químicas essenciais obtidas da avaliação das interações similares identificadas de complexos receptor-ligante conhecidos [LGLT10]. No entanto, os modelos desenvolvidos não têm considerado a flexibilidade das proteínas na geração dos modelos farmacofóricos 3D. Outra restrição encontrada é a dependência de ligantes conhecidos para a identificação de resíduos alvo.

Atualmente, métodos alternativos têm sido desenvolvidos visando considerar os movimentos flexíveis da cavidade de ligação do substrato na geração modelos farmacofóricos 3D [DLS$^+$05, DSNB06, HL13]. No entanto, importantes limitações foram encontradas nesses poucos modelos farmacofóricos 3D analisando modelos FFR disponibilizados na literatura. Assim, ainda hoje não existe um método baseado somente nas propriedades físico-químicas da cavidade de ligação do substrato considerando um modelo FFR, de forma a efetivamente reduzir esse número elevado de ligantes a um valor gerenciável.

8. Considerações finais

A incorporação da flexibilidade em receptores de proteínas se tornou um fator diferencial no processo do Planejamento Racional de Fármacos (RDD). No entanto, as abordagens baseadas na execução de experimentos de docagem molecular exaustivas, considerando modelos de Receptor Totalmente Flexíveis (FFR) e a quantidade de ligantes disponíveis em BD, não são adequadas devido ao tempo computacional necessário. Desta forma, a aplicação de simplificações, tanto na avaliação do modelo FFR quanto na seleção de pequenas moléculas, faz-se necessária para tornar este processo exequível. Uma solução para possibilitar uma varredura completa nos BDs de ligantes é o desenvolvimento de métodos capazes de selecionar o conjunto de ligantes mais promissores conforme as características canônicas do modelo FFR antes de aplicar os experimentos de docagem molecular.

Nesse sentido, modelos farmacofóricos 3D avaliando receptores flexíveis têm realizado simplificações na quantidade de conformações a serem utilizadas com base nos valores do RMSD. Os modelos gerados a partir dessas análises não exploram grande parte das regiões flexíveis geradas pelo modelo FFR, visto que utilizar diferentes valores de RMSD não garante que a estrutura da cavidade de ligação do substrato seja diferente. Além disso, grande parte desses trabalhos avaliam as características físico-químicas presentes nos complexos receptor-ligante conhecidos. Desta forma, as regiões dentro da cavidade que não interagem com o conjunto de ligantes formadores do modelo farmacofórico 3D e, que podem permitir a interação de ligantes estruturalmente diferentes, consequentemente, não estarão contemplados nesta busca seletiva.

O presente trabalho relata um método para realizar a triagem virtual em BD de ligantes a partir da avaliação das propriedades físico-químicas 3D da cavidade de ligação do substrato de um modelo FFR, visando contribuir para a descoberta de novas moléculas candidatas à farmaco. Uma avaliação considerando a representatividade das partições das estruturas similares do modelo FFR em conjunto com o valor resultante do alinhamento das propriedades do ligantes com o seu modelo farmacofórico foram utilizados para gerar uma lista ordenada das moléculas mais promissoras a se tornarem candidatas a fármacos para a enzima avaliada.

Os capítulos iniciais desta tese descreveram conceitos fundamentais das áreas de pesquisa necessárias para o desenvolvimento das atividades que foram apresentadas nesta tese, destacando características do RDD tais como a triagem virtual baseada em estrutura, a simulação de parte da flexibilidade pela Dinâmica Molecular e características dos principais BD de ligantes disponibilizados na literatura. Essa avaliação descrevendo os BD de ligantes resultou na publicação de um artigo em uma conferência internacional em 2012, o *Brazilian Symposium on Bioinformatics*.

O capítulo 3 e o capítulo 4 apresentaram avaliações da enzima adotada como estudo de caso nesta tese, a enzima InhA de *Mycobacterium tuberculosis* (*Mtb*). O capítulo 3 caracterizou importantes conceitos e a motivação social do estudo sobre a enzima InhA de *Mtb*. Também descreveu os parâmetros utilizados na geração do modelo FFR de 19.500 ps da InhA de *Mtb*. O capítulo 4 reportou dois estudos baseados em técnicas *redocking* e *cross docking* para descrever importantes características físico-químicas das interações entre 20 ligantes candidatos a fármaco e a proteína InhA. Os experimentos de *redocking* foram utilizados para calibrar os parâmetros de programas de docagem molecular e, segundo [WLW03, VGK05], experimentos contendo bons resultados são aqueles cujo o valor de RMSD são inferiores a 2.0 Å.

Com base nesses resultados, tanto os valores dos parâmetros dos programas de docagem (definições do algoritmo de busca e da função de escore) quanto os valores calculados das cargas parciais estão adequados. Além disso, estes experimentos também forneceram um conjunto de ligantes de referência para serem utilizados por pesquisadores em avaliações de experimentos de *cross docking*. Esses experimentos de *cross docking* foram aplicados para avaliar o modelo da proteína InhA de 19,5 ns. Os experimentos com o modelo FFR mostraram que a partir de uma estrutura cristalina contendo somente a coenzima NADH, foi possível gerar as conformações dos ligantes/adutos de outras enzimas cristalinas da InhA. Resultados não alcançados pelo programa de docagem molecular nos experimentos de *cross docking* entre os mesmos ligantes e a estrutura cristalina 1ENY. Os resultados dos experimentos de docagem molecular realizados nesse capítulo garantem a confiabilidade do modelo FFR gerado e também ressaltam a importância de considerar a flexibilidade das proteínas no RDD.

O capítulo 5 apresentou três métodos para identificar partições de conformações similares de modelos FFR baseado nas propriedades geométricas da cavidade de ligação do substrato. As análises desse capítulo mostram uma série de agrupamentos realizados visando reduzir a dimensionalidade do modelo FFR. O primeiro método agregou atributos como a avaliação do volume e a quantidade de relevantes átomos presentes na cavidade de ligação do subtrato ao longo do modelo FFR. O agrupamento gerado por esse método foi utilizado como o conjunto de dados de entrada para o ambiente web baseado em computação em núvem desenvolvido em nosso grupo chamado de wFReDoW [DPFNdSR13]. Esse agrupamento possibilitou que a automatização do processo de testes sequenciais de conjuntos de ligantes utilizando o ambiente wFReDoW. Assim, este novo agrupamento solucionou o problema da geração automática dos grupos de entrada do P-SaMI [HRFNdS15], o qual procura selecionar grupos de conformações que sejam favoráveis entre as interações do modelo FFR com o ligante. Estes resultados foram publicados na *Expert Systems With Applications* [QDPRNdS14]. No entanto, uma das limitações mais importantes identificadas nesse agrupamento refere-se a subjetividade ainda existente no conjunto de dados avaliado.

O segundo método expandiu atributo que descrevia a quantidade de átomos pesados da enzima 1BVR presentes na cavidade de ligação do substrato da conformação. Esse agrupamento selecionou 192 estruturas representativas comparando com as estruturas representativas selecionadas por conjuntos gerados a partir os valores de RMSD da estrutura inteira e somente da cavidade de ligação do substrato. Os resultados mostraram-se mais adequados com o conjunto estruturas representativas agrupadas pelo método contendo as propriedades da cavidade de ligação do substrato. Estes resultados foram publicados em uma importante revista da área da Ciência da Computação, a *PloS one* [DPQRdNdS15]. Embora as avaliações da energia tenham se mostrado bastante adequadas, as estruturas representativas selecionadas não expressaram a variabilidade estrutural esperada. Isso também foi evidenciado na avaliação que identificava as características médias dos atributos de cada conjunto e comparava o valor da energia de ligação desta conformação. Assim, as estruturas representativas selecionadas pelos dois primeiros métodos desenvolvidos não apresentaram variações estruturais significativas entre as diferentes partições.

A fim de evitar o viés ocasionado pela medida do RMSD, o terceiro método avaliou as propriedades físico-químicas da cavidade de ligação do substrato utilizando triangularizações entre as propriedades farmacofóricas do receptor. Esse método consegue caracterizar o conjunto de propriedades envolvidos na cavidade de ligação do substrato sem depender do alinhamento das estruturas, que é uma restrição dos agrupamentos baseados no RMSD. Em comparação com outros estudos de agrupamento de trajetórias de modelos FFR, o terceiro método desenvolvido utilizando um conjunto de triângulos de propriedades farmcofóricas com a seleção de estruturas vizinhas pelo SketchSort apresentou vantagens essenciais na seleção de estruturas representativas, identificando cavidades do sítio de ligação bastante distintas. Embora esse método demande uma etapa de pré-processamento mais elaborada, os resultados alcançados foram significativos quando comparados com os resultados encontrados com os outros dois métodos. A evolução dos recursos computacionais deverá possibilitar a geração de simulações por DM mais extensas que as trajetórias produzidas atualmente. Assim, métodos capazes de reconhecer a ocorrência da repetições do arranjo 3D dos átomos da cavidade de ligação do substrato podem ressaltar a existência da formação de ciclos dentro da trajetória gerada.

O capítulo 6 apresentou um método para a determinação de modelos farmacofóricos 3D baseado nas características físico-químicas da estrutura 3D de regiões inacessíveis por estruturas cristalinas do receptor InhA de *Mtb*. Essa etapa é baseada nas propriedades farmacofóricas 3D da cavidade de ligação do substrato das estruturas representativas das partições formadas no capítulo 5. Cada partição possui um peso determinado pela quantidade de conformações agrupadas. As propriedades farmacofóricas existentes na estrutura cristalina são expressas como pequenas esferas de *van der Waals*, enquanto que propriedades farmacofóricas de regiões inacessíveis por estruturas cristalinas possuem raios de acordo com a frequência de cada propriedade. Essa projeção indica ao pesquisador quais

arranjos de propriedades farmacofóricas 3D correspondem a avaliação da flexibilidade do modelo FFR. Cabe ao pesquisador definir qual seria o arranjo mais adequado para se realizar a pesquisa no BD de ligantes. O sucesso dessa etapa está diretamente relacionada a qualidade do conjunto de estruturas similares particionadas. Isso porque as estruturas representativas definem os pontos farmacofóricos a serem utilizados na pesquisa no BD; mas a flexibilidade do conjuntos é considerada com a avaliação da variabilidade desses pontos conforme o posicionamento dos respectivos resíduos nas demais conformações da partição.

Embora os resultados apresentados ainda necessitem de aprimoramento, a ideia defendida nesta tese aponta uma nova e importante direção no processo de triagem virtual de ligantes. Além disso, a geração de modelos de Receptores Totalmente Flexíveis ainda encontra-se em fase de aprimoramento. Assim, o fato de a seleção dos ligantes não apresentar uma variabilidade de estruturas com relação a uma seleção realizada a partir de estruturas cristalinas, pode ser atribuído basicamente aos seguintes pontos:

- Posição dos pontos farmacofóricos: A projeção dos pontos farmacofóricos apresentados nesta tese possuem um alto índice de abstração para tornar o processo ágil. Consequentemente, a inserção de incerteza pode ocasionar a geração de resultados falsos positivos.

- Qualidade do modelo FFR: A geração de modelos flexíveis sem qualidade pode ocasionar a formação de microeestados não reproduzíveis pela estrutura em ambiente natural. A qualidade do modelo FFR também pode ser afetada pelo conjunto limitado de estruturas produzidas, fato que implica diretamente na captura de todos os movimentos de um modelo de Receptor Totalmente Flexível.

- Tamanho da cavidade: O tamanho da cavidade alvo influencia diretamente na quantidade de resíduos que compõem a área acessível ao solvente dessa cavidade. Consequentemente, uma quantidade maior de resíduos compondo a cavidade alvo tende a inserir uma quantidade maior de pontos farmacofóricos na cavidade alvo.

Por fim, se os ligantes apresentarem boas interações na avaliação dos experimentos de docagem molecular e os experimentos *in vitro* não apresentarem resultados efetivos, pode indicar que a flexibilidade da proteína não foi simulada de forma correta.

Como contribuição desta pesquisa, espera-se que o método desenvolvido se torne uma importante ferramenta de apoio aos pesquisadores, auxiliando a pesquisa na busca de novos candidatos à fármacos selecionando, os ligantes que possuem as características físico-químicas propícias para encaixar no receptor. Desta maneira, espera-se contribuir na aceleração do processo de seleção dos possíveis candidatos a serem testados com modelos FFR de moléculas alvo. A próxima seção descreve algumas das principais limitações desta tese.

8.1 Limitações

Naturalmente, os métodos desenvolvidos nesta tese podem ser evoluídos. Algumas limitações foram adotadas conscientemente, devido a necessidade da redução da dimensionalidade do conjunto de propriedades a serem avaliadas. Abaixo estão descritas as principais limitações:

- Utilizar um programa *online* para identificar as cavidades de ligação do substrato do modelo FFR: O CASTp utiliza um algoritmo baseado no método *alpha-shape* desenvolvido por Edelsbrunner and Mücke [EM94] e utiliza tanto o modelo da superfície acessível ao solvente desenvolvido por Richard [Ric77], quanto o modelo da superfície molecular desenvolvido por Connolly [Con83]. O CASTp foi escolhido como o programa à ser utilizado nesta tese devido a experiência de uso dessa ferramenta e pela metodologia abordada. Esse programa fornece bons resultados, mas torna-se um gargalo no momento em que não disponibiliza uma versão local.

- Vetor binário para contabilizar a existência da subestrutura na cavidade de ligação do substrato: A utilização de um vetor que avalia apenas a presença da subestrutura foi definida por questões de limitação computacional. Naturalmente, contabilizar a frequência da ocorrência dos triângulos em cada cavidade de ligação possibilitaria uma especificidade mais adequada da comparação da cavidade alvo.

- Discretização mais suave: As distâncias entre as propriedades foram discretizadas em intervalos de 2,2 Å para limitar a quantidade em 16 categorias. A avaliação de intervalos menores tornaria as subestruturas mais específicas, possibilitando caracterizar as regiões de uma forma mais específica.

- Necessidade de mais tempo para formar os agrupamentos que o método baseado no RMSD: A avaliação e a geração de partições formadas a partir do valor de RMSD pode ser realizada rapidamente, no entanto uma avaliação com maior especificidade necessita de um tempo superior. Contudo, um aprimoramento do algoritmo de avaliação pode reduzir o tempo necessário atualmente.

- Definição do ponto farmacofórico: Atribuir a posição do ponto farmacofórico a partir do resíduo foi uma etapa de grande abstração para a redução da dimensionalidade. No entanto, a possibilidade de utilizar os Campos de Interação Molecular (MIF - do inglês *Molecular Interaction Fields*) sem aplicar técnicas de redução de dimensionalidade tornaria o processo consideravelmente mais demorado e definir uma técnica ideal para a redução dessa dimensionalidade necessitaria de uma análise bastante detalhada.

- Parametrização do tamanho do raio do ponto farmacofórico: Com a seleção de grupos similares, a projeção do raio pode ser baseada na avaliação da variabilidade da posição considerando o conjunto de conformações do grupo.

8.2 Trabalhos futuros

A avaliação de modelos farmacofóricos 3D considerando a flexibilidade proporcionada pelo modelo flexível insere uma quantidade elevada de fatores que torna necessária a aplicação de métodos para reduzir esse espaço de busca. Desta forma, os métodos desenvolvidos devem analisar a qualidade da redução do conjunto a ser avaliado considerando o índice de confiança dos resultados obtidos.

Dentre as principais evoluções a serem desenvolvidas (com base nas limitações descritas na seção 8.1), a principal questão envolve a definição do ponto farmacofórico. A identificação mais precisa desse ponto pode impactar diretamente na qualidade. Um trabalho relevante nesta área aplicando uma heurística para transformar as informações de um MIF em um conjunto de pontos relevantes foi apresentado por Hu & Lill [HL13].

Outros fatores relevantes a serem aprimorados estão relacionados a utilização de outros programas de docagem molecular, a avaliação de modelos FFR mais extensos e as melhorias de implementação possíveis no código já elaborado.

8.3 Publicações

Durante o tempo de doutoramento, os seguintes artigos foram publicados:

- Revista A1: *A strategic solution to optimize molecular docking simulations using Fully-Flexible Receptor Models*. **Expert Systems With Applications.** 2014
- Revista A1: *An effective approach for clustering InhA molecular dynamics trajectory using substrate-binding cavity features*. **PloS one.** 2015
- Revista: *Clustering Molecular Dynamics Trajectories for Optimizing Docking Experiments*. **Computational Intelligence and Neuroscience.** 2015
- Conferência A1: *Medoid-based Data Clustering with Estimation of Distribution Algorithms*. **Symposium on Applied Computing.** 2016
- Conferência A2: *Clustering Molecular Dynamics Trajectories with a Univariate Estimation of Distribution Algorithm*. **Evolutionary Computation.** 2015
- Conferência B4: *A comparative analysis of public ligand databases based on molecular descriptors*. **Advances in Bioinformatics and Computational Biology.** 2012

Referências Bibliográficas

[ABE+13] Ahlstrom, L. S.; Baker, J. L.; Ehrlich, K.; Campbell, Z. T.; Patel, S.; Vorontsov, I. I.; Tama, F.; Miyashita, O. “Network visualization of conformational sampling during molecular dynamics simulation”, *Journal of Molecular Graphics and Modelling*, vol. 46, 2013, pp. 140–149.

[ABG06] Alonso, H.; Bliznyuk, A. A.; Gready, J. E. “Combining docking and molecular dynamic simulations in drug design”, *Medicinal Research Reviews*, vol. 26–5, 2006, pp. 531–568.

[ABIC04] Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. “NIH molecular libraries initiative”, *Science*, vol. 306–5699, 2004, pp. 1138–1139.

[ACC+13] Artese, A.; Cross, S.; Costa, G.; Distinto, S.; Parrotta, L.; Alcaro, S.; Ortuso, F.; Cruciani, G. “Molecular interaction fields in drug discovery: recent advances and future perspectives”, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 3–6, 2013, pp. 594–613.

[AL10] Amaro, R. E.; Li, W. W. “Emerging methods for ensemble-based virtual screening”, *Current Topics in Medicinal Chemistry*, vol. 10–1, 2010, pp. 2–13.

[APZF08] Andrade, C. H.; Pasqualoto, K. F. M.; Zaim, M. H.; Ferreira, E. I. “Rational approach in the new antituberculosis agent design: inhibitors of InhA, the enoyl-ACP reductase from *mycobacterium tuberculosis*”, *Revista Brasileira de Ciências Farmacêuticas*, vol. 44–2, 2008, pp. 167–179.

[AVB07] Argyrou, A.; Vetting, M. W.; Blanchard, J. S. “New insight into the mechanism of action of and resistance to isoniazid: interaction of *mycobacterium tuberculosis* enoyl-ACP reductase with INH-NADP”, *Journal of the American Chemical Society*, vol. 129–31, 2007, pp. 9582–9583.

[BAK15] Bajusz, D.; Anita, R.; Károly, H. “Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations?”, *Journal of Cheminformatics*, 2015, pp. 7–20.

[BBH+15] Bender, M. A.; Berry, J.; Hammond, S. D.; Moore, B.; Phillips, C. A. “k-means clustering on two-level memory systems”, *International Symposium on Memory Systems*, 2015, pp. 197–205.

[BCCK93] Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. “A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model”, *The Journal of Physical Chemistry*, vol. 97–40, 1993, pp. 10269–10280.

[BCS+07] Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. “A common reference framework for analyzing/comparing proteins and ligands. fingerprints for ligands and proteins (flap): theory and application”, *Journal of Chemical Information and Modeling*, vol. 47–2, 2007, pp. 279–294.

[BD92] Barnard, J. M.; Downs, G. M. “Clustering of chemical structures on the basis of two-dimensional similarity measures”, *Journal of Chemical Information and Computer Sciences*, vol. 32–6, 1992, pp. 644–649.

[BMR08] Barillari, C.; Marcou, G.; Rognan, D. “Hot-spots-guided receptor-based pharmacophores (hs-pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores”, *Journal of Chemical Information and Modeling*, vol. 48–7, 2008, pp. 1396–1410.

[BNL03] Binkowski, T. A.; Naghibzadeh, S.; Liang, J. “CASTp: computed atlas of surface topography of proteins”, *Nucleic Acids Research*, vol. 31–13, 2003, pp. 3352–3355.

[BR02] Bursavich, M. G.; Rich, D. H. “Designing non-peptide peptidomimetics in the 21st century: inhibitors targeting conformational ensembles”, *Journal of Medicinal Chemistry*, vol. 45–3, 2002, pp. 541–558.

[BRM15] Buonfiglio, R.; Recanatini, M.; Masetti, M. “Protein flexibility in drug discovery: From theory to computation”, *ChemMedChem*, 2015.

[Bro00] Broughton, H. B. “A method for including protein flexibility in protein-ligand docking: improving tools for database mining and virtual screening”, *Journal of Molecular Graphics and Modelling*, vol. 18–3, 2000, pp. 247–257.

[BWF+00] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. “The Protein Data Bank”, *Nucleic Acids Research*, vol. 28–1, 2000, pp. 235–242.

[Cas07] Caskey, C. T. “The drug development crisis: efficiency and safety”, *Annual Review of Medicine*, vol. 58, 2007, pp. 1–16.

[Cat00] Cato, S. “Exploring pharmacophores with Chem-X”, *Pharmacophore Perception, Development, and Use in Drug Design*, 2000, pp. 107–125.

[CC11] Cross, S.; Cruciani, G. “Grid-derived structure-based 3D pharmacophores and their performance compared to docking”, *Drug Discovery Today: Technologies*, vol. 7–4, 2011, pp. e213–e219.

[CCB+95] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. “A second generation force field for the simulation of proteins, nucleic acids, and organic molecules”, *Journal of the American Chemical Society*, vol. 117–19, 1995, pp. 5179–5197.

[CDCI+12] Case, D.; Darden, T.; Cheatham III, T.; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Walker, R.; Zhang, W.; Merz, K.; et al.. “AMBER 12”, San Francisco, CA, 2012.

[CKW+09] Chiang, Y.-K.; Kuo, C.-C.; Wu, Y.-S.; Chen, C.-T.; Coumar, M. S.; Wu, J.-S.; Hsieh, H.-P.; Chang, C.-Y.; Jseng, H.-Y.; Wu, M.-H.; et al.. “Generation of ligand-based pharmacophore model and virtual screening for identification of novel tubulin inhibitors with potent anticancer activity”, *Journal of Medicinal Chemistry*, vol. 52–14, 2009, pp. 4221–4233.

[CLP11] Cossio, P.; Laio, A.; Pietrucci, F. “Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory?”, *Physical Chemistry Chemical Physics*, vol. 13–22, 2011, pp. 10421–10425.

[CLS+07] Chen, J. H.; Linstead, E.; Swamidass, S. J.; Wang, D.; Baldi, P. “ChemDB update - full-text search and virtual chemical space”, *Bioinformatics*, vol. 23–17, 2007, pp. 2348–2351.

[CM00] Carlson, H. A.; McCammon, J. A. “Accommodating protein flexibility in computational drug design”, *Molecular Pharmacology*, vol. 57–2, 2000, pp. 213–218.

[Con83] Connolly, M. L. “Analytical molecular surface calculation”, *Journal of Applied Crystallography*, vol. 16–5, 1983, pp. 548–558.

[CSS09] Chandrika, B.-R.; Subramanian, J.; Sharma, S. D. “Managing protein flexibility in docking and its applications”, *Drug Discovery Today*, vol. 14–7, 2009, pp. 394–400.

[DABR+13] De Almeida, H.; Bastos, I. M.; Ribeiro, B. M.; Maigret, B.; Santana, J. M. “New binding site conformations of the dengue virus NS3 protease

accessed by molecular dynamics simulation", *PloS one*, vol. 8–8, 2013, pp. e72402.

[Dav92] Davis, I. J. "A fast radix sort", *The Computer Journal*, vol. 35–6, 1992, pp. 636–642.

[DB79] Davies, D. L.; Bouldin, D. W. "A cluster separation measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, 1979, pp. 224–227.

[De 12] De Paris, R. "Fremi - a middleware to handle molecular docking simulations of fully-flexible receptor modes in hpc environments", Dissertação de Mestrado, Faculdade de Informática – PUCRS, Porto Alegre, RS, Brasil, 2012, 74p.

[DeL02] DeLano, W. L. "Pymol. the pymol molecular graphics system", San Carlos, CA, 2002.

[DLS+05] Deng, J.; Lee, K. W.; Sanchez, T.; Cui, M.; Neamati, N.; Briggs, J. M. "Dynamic receptor-based pharmacophore model development and its application in designing novel hiv-1 integrase inhibitors", *Journal of Medicinal Chemistry*, vol. 48–5, 2005, pp. 1496–1505.

[DPFdSR11] De Paris, R.; Frantz, F. A.; de Souza, O. N.; Ruiz, D. D. "A conceptual many tasks computing architecture to execute molecular docking simulations of a fully-flexible receptor model", *Brazilian Symposium on Bioinformatics*, 2011, pp. 75–78.

[DPFNdSR13] De Paris, R.; Frantz, F. A.; Norberto de Souza, O.; Ruiz, D. D. "wFReDoW: a cloud-based web environment to handle molecular docking simulations of a fully flexible receptor model", *BioMed Research International*, vol. 2013, 2013, pp. 1–12.

[DPQR+15] De Paris, R.; Quevedo, C. V.; Ruiz, D. D.; Norberto de Souza, O.; Barros, R. C. "Clustering molecular dynamics trajectories for optimizing docking experiments", *Computational Intelligence and Neuroscience*, vol. 2015, 2015.

[DPQRdNdS15] De Paris, R.; Quevedo, C. V.; Ruiz, D. D.; de Norberto de Souza, O. "An effective approach for clustering inha molecular dynamics trajectory using substrate-binding cavity features", *PloS one*, vol. 10–7, 2015, pp. e0133172.

[DQB+95] Dessen, A.; Quemard, A.; Blanchard, J. S.; Jacobs Jr, W. R.; Sacchettini, J. C. “Crystal structure and function of the isoniazid target of *mycobacterium tuberculosis*”, *Science*, vol. 267–5204, 1995, pp. 1638–1641.

[DS13] Doane, D. P.; Seward, L. E. “Applied statistics in business and economics”. USA: Irwin, 2013.

[DSNB06] Deng, J.; Sanchez, T.; Neamati, N.; Briggs, J. M. “Dynamic pharmacophore model optimization: identification of novel hiv-1 integrase inhibitors”, *Journal of Medicinal Chemistry*, vol. 49–5, 2006, pp. 1684–1692.

[DSPNW04] Dror, O.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. “Predicting molecular interactions in silico: I. a guide to pharmacophore identification and its applications to drug design”, *Current Medicinal Chemistry*, vol. 11–1, 2004, pp. 71–90.

[Dun73] Dunn, J. C. “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters”, *Journal of Cybernetics*, vol. 3–3, 1973, pp. 32–57.

[Dun95] Dunitz, J. D. “Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions”, *Chemistry & Biology*, vol. 2–11, 1995, pp. 709–712.

[DVP+07] Dias, M. V. B.; Vasconcelos, I. B.; Prado, A. M. X.; Fadel, V.; Basso, L. A.; et al.. “Crystallographic studies on the binding of isonicotinyl-nad adduct to wild-type and isoniazid resistant 2-*trans*-enoyl-ACP (CoA) reductase from *mycobacterium tuberculosis*”, *Journal of Structural Biology*, vol. 159–3, 2007, pp. 369–380.

[EM94] Edelsbrunner, H.; Mücke, E. P. “Three-dimensional alpha shapes”, *ACM Transactions on Graphics*, vol. 13–1, 1994, pp. 43–72.

[EON+14] Encinas, L.; O’Keefe, H.; Neu, M.; Remuiñán, M. J.; Patel, A. M.; Guardia, A.; Davie, C. P.; Pérez-Macías, N.; Yang, H.; Convery, M. A.; et al.. “Encoded library technology as a source of hits for the discovery and lead optimization of a potent and selective class of bactericidal direct inhibitors of *mycobacterium tuberculosis* InhA”, *Journal of Medicinal Chemistry*, vol. 57–4, 2014, pp. 1276–1288.

[fDCC06] Centers for Disease Control and Prevention (CDC). “Emergence of mycobacterium tuberculosis with extensive resistance to second-line

drugs–worldwide (2000-2004)", Relatório Técnico, Morbidity and Mortality Weekly Report, 2006, 301p.

[FPSB11] Fraccalvieri, D.; Pandini, A.; Stella, F.; Bonati, L. "Conformational and functional analysis of molecular dynamics trajectories by self-organising maps", *BMC Bioinformatics*, vol. 12–1, 2011, pp. 158.

[FWV+09] Freundlich, J. S.; Wang, F.; Vilchèze, C.; Gulten, G.; Langley, R.; Schiehser, G. A.; Jacobus, D. P.; Jacobs, W. R.; Sacchettini, J. C. "Triclosan derivatives: Towards potent inhibitors of drug-sensitive and drug-resistant *mycobacterium tuberculosis*", *Journal ChemMedChem*, vol. 4–2, 2009, pp. 241–248.

[GCNdS07] Gargano, F.; Costa, A. L.; Norberto de Souza, O. "Effect of temperature on enzyme structure and function: a molecular dynamics simulation study", *Annals of the 3rd International Conference of the Brazilian Association for Bioinformatics and Computational Biology, São Paulo, Brazil*, 2007.

[GHK00] Gohlke, H.; Hendlich, M.; Klebe, G. "Knowledge-based scoring function to predict protein-ligand interactions", *Journal of Molecular Biology*, vol. 295–2, 2000, pp. 337–356.

[GM80] Gasteiger, J.; Marsili, M. "Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges", *Tetrahedron*, vol. 36–22, 1980, pp. 3219–3228.

[GRFS15] Galperin, M. Y.; Rigden, D. J.; Fernández-Suárez, X. M. "The 2015 nucleic acids research database issue and molecular biology database collection", *Nucleic Acids Research*, vol. 43–D1, 2015, pp. D1–D5.

[Gun00] Guner, O. F. "Pharmacophore perception, development, and use in drug design". La Jolla, CA, 2000, vol. 29, 560p.

[HAO+06] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. "Comparison of multiple amber force fields and development of improved protein backbone parameters", *Proteins: Structure, Function, and Bioinformatics*, vol. 65–3, 2006, pp. 712–725.

[HAOdM07] He, X.; Alian, A.; Ortiz de Montellano, P. R. "Inhibition of the *mycobacterium tuberculosis* enoyl acyl carrier protein reductase InhA by arylamides", *Bioorganic & Medicinal Chemistry*, vol. 15–21, 2007, pp. 6649–6658.

[HASOdM06] He, X.; Alian, A.; Stroud, R.; Ortiz de Montellano, P. R. "Pyrrolidine carboxamides as a novel class of inhibitors of enoyl acyl carrier protein

reductase from *mycobacterium tuberculosis*", *Journal of Medicinal Chemistry*, vol. 49–21, 2006, pp. 6308–6323.

[HBV02] Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. "Clustering validity checking methods: part II", *ACM Sigmod Record*, vol. 31–3, 2002, pp. 19–27.

[HDS96] Humphrey, W.; Dalke, A.; Schulten, K. "VMD: visual molecular dynamics", *Journal of Molecular Graphics*, vol. 14–1, 1996, pp. 33–38.

[HKP11] Han, J.; Kamber, M.; Pei, J. "Data mining: concepts and techniques". Elsevier, 2011, 744p.

[HL12] Hu, B.; Lill, M. A. "Protein pharmacophore selection using hydration-site analysis", *Journal of Chemical Information and Modeling*, vol. 52–4, 2012, pp. 1046–1060.

[HL13] Hu, B.; Lill, M. A. "Exploring the potential of protein-based pharmacophore models in ligand pose prediction and ranking", *Journal of Chemical Information and Modeling*, vol. 53–5, 2013, pp. 1179–1190.

[HM08] Huey, R.; Morris, G. M. "Using autodock 4 with autodocktools: A tutorial", *The Scripps Research Institute, USA*, 2008, pp. 1–56.

[HMF12] Huey, R.; Morris, G. M.; Forli, S. "Using autodock 4 and autodock vina with autodocktools: A tutorial", *The Scripps Research Institute, USA*, 2012, pp. 1–32.

[HMOG07] Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. "A semiempirical free energy force field with charge-based desolvation", *Journal of Computational Chemistry*, vol. 28–6, 2007, pp. 1145–1152.

[HNW+97] Hong, H.; Neamati, N.; Wang, S.; Nicklaus, M. C.; Mazumder, A.; Zhao, H.; Burke, T. R.; Pommier, Y.; Milne, G. W. "Discovery of hiv-1 integrase inhibitors by pharmacophore searching", *Journal of Medicinal Chemistry*, vol. 40–6, 1997, pp. 930–936.

[HRFNdS15] Hübler, P.; Ruiz, D.; Ferreira, J. E.; Norberto de Souza, O. "P-SaMI: a data-flow pattern to perform massively-parallel molecular docking experiments using a fully-flexible receptor model", *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, 2015, pp. 54–57.

[Hüb10] Hübler, P. "P-MIA: Padrão múltiplas instâncias autoadaptáveis - um padrão de dados para workflows científicos", Tese de Doutorado, Faculdade de Informática – PUCRS, Porto Alegre, RS, Brasil, 2010, 179p.

[HW79] Hartigan, J. A.; Wong, M. A. “A k-means clustering algorithm”, *Applied Statistics*, vol. 28, 1979, pp. 100–108.

[IM98] Indyk, P.; Motwani, R. “Approximate nearest neighbors: towards removing the curse of dimensionality”. In: Proceedings of the thirtieth annual ACM symposium on Theory of computing, 1998, pp. 604–613.

[IS05] Irwin, J. J.; Shoichet, B. K. “ZINC - A free database of commercially available compounds for virtual screening”, *Journal of chemical information and modeling*, vol. 45–1, 2005, pp. 177–182.

[ISM+12] Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. “ZINC: a free tool to discover chemistry for biology”, *Journal of chemical information and modeling*, vol. 52–7, 2012, pp. 1757–1768.

[ITS+12] Ito, J.-I.; Tabei, Y.; Shimizu, K.; Tomii, K.; Tsuda, K. “Pdb-scale analysis of known and putative ligand-binding sites with structural sketches”, *Proteins: Structure, Function, and Bioinformatics*, vol. 80–3, 2012, pp. 747–763.

[IVB+02] Ihlenfeldt, W.-D.; Voigt, J. H.; Bienfait, B.; Oellien, F.; Nicklaus, M. C. “Enhanced CACTVS browser of the Open NCI Database”, *Journal of Chemical Information and Computer Sciences*, vol. 42–1, 2002, pp. 46–57.

[JD88] Jain, A. K.; Dubes, R. C. “Algorithms for clustering data”. Prentice-Hall, Inc., 1988, 334p.

[Jia08] Jiang, Z. “Computational analysis of the interaction between ligand-receptor pairs”, *Current Pharmaceutical Design*, vol. 14–6, 2008, pp. 588–592.

[JTJ+10] Jacobson, K. R.; Tierney, D. B.; Jeon, C. Y.; Mitnick, C. D.; Murray, M. B. “Treatment outcomes among patients with extensively drug-resistant tuberculosis: systematic review and meta-analysis”, *Clinical Infectious Diseases*, vol. 51–1, 2010, pp. 6–14.

[KBK+01] Kirchhoff, P. D.; Brown, R.; Kahn, S.; Waldman, M.; Venkatachalam, C. “Application of structure-based focusing to the estrogen receptor”, *Journal of Computational Chemistry*, vol. 22–10, 2001, pp. 993–1003.

[KC11] Koes, D. R.; Camacho, C. J. “Pharmer: efficient and exact pharmacophore search”, *Journal of Chemical Information and Modeling*, vol. 51–6, 2011, pp. 1307–1314.

[KC12] Koes, D. R.; Camacho, C. J. “Zincpharmer: pharmacophore search of the zinc database”, *Nucleic Acids Research*, vol. 40, 2012, pp. W409–W414.

[KLJ+11] Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; et al.. “DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs”, *Nucleic Acids Research*, vol. 39, 2011, pp. D1035–D1041.

[KM02] Karplus, M.; McCammon, J. A. “Molecular dynamics simulations of biomolecules”, *Nature Structural & Molecular Biology*, vol. 9–9, 2002, pp. 646–652.

[KMA+03] Kuo, M. R.; Morbidoni, H. R.; Alland, D.; Sneddon, S. F.; Gourlie, B. B.; Staveski, M. M.; Leonard, M.; Gregory, J. S.; Janjigian, A. D.; Yee, C.; et al.. “Targeting tuberculosis and malaria through inhibition of enoyl reductase compound activity and structural data”, *Journal of Biological Chemistry*, vol. 278–23, 2003, pp. 20851–20859.

[Kun92] Kuntz, I. D. “Structure-Based Strategies for Drug Design and Discovery”, *Science*, vol. 257–5073, 1992, pp. 1078–1082.

[L+13] Li, S. C.; et al.. “The difficulty of protein structure alignment under the rmsd.”, *Algorithms for Molecular Biology*, vol. 8–1, 2013.

[LAB+08] Landon, M. R.; Amaro, R. E.; Baron, R.; Ngan, C. H.; Ozonoff, D.; Andrew McCammon, J.; Vajda, S. “Novel druggable hot spots in avian influenza neuraminidase h5n1 revealed by computational solvent mapping of a reduced and representative receptor ensemble”, *Chemical Biology & Drug Design*, vol. 71–2, 2008, pp. 106–116.

[Las95] Laskowski, R. A. “SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions”, *Journal of Molecular Graphics*, vol. 13–5, 1995, pp. 323–330.

[Las03] Laskowski, R. A. “Structural quality assurance”, *Structural Bioinformatics*, vol. 44, 2003, pp. 273–303.

[LB12] Luque, F. J.; Barril, X. “Pharmacophore Models in Drug Design”. Royal Society of Chemistry, 2012, 418p.

[LGLT10] Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. “Three-dimensional pharmacophore methods in drug discovery”, *Journal of Medicinal Chemistry*, vol. 53–2, 2010, pp. 539–558.

[LLaE+10] Luckner, S. R.; Liu, N.; am Ende, C. W.; Tonge, P. J.; Kisker, C. “A slow, tight binding inhibitor of InhA, the enoyl-acyl carrier protein reductase from *mycobacterium tuberculosis*”, *Journal of Biological Chemistry*, vol. 285–19, 2010, pp. 14330–14337.

[Llo82] Lloyd, S. P. “Least squares quantization in PCM”, *IEEE Transactions on Information Theory*, vol. 28–2, 1982, pp. 129–137.

[LLP+14] Li, H.-J.; Lai, C.-T.; Pan, P.; Yu, W.; Liu, N.; Bommineni, G. R.; Garcia-Diaz, M.; Simmerling, C.; Tonge, P. J. “A structural and energetic model for the slow-onset inhibition of the *mycobacterium tuberculosis* enoyl-ACP reductase InhA”, *ACS Chemical Biology*, vol. 9–4, 2014, pp. 986–993.

[LP03] Liljefors, T.; Pettersson, I. “Computer-aided development and use of three-dimensional pharmacophore models”, *Textbook of Drug Design and Discovery*, 2003, pp. 86–115.

[LR96] Lengauer, T.; Rarey, M. “Computational methods for biomolecular docking”, *Current Opinion in Structural Biology*, vol. 6–3, 1996, pp. 402–406.

[LS11] Laskowski, R. A.; Swindells, M. B. “Ligplot+: multiple ligand–protein interaction diagrams for drug discovery”, *Journal of chemical information and modeling*, vol. 51–10, 2011, pp. 2778–2786.

[LWWZ08] Liu, Z.-P.; Wu, L.-Y.; Wang, Y.; Zhang, X.-S. “Protein cavity clustering based on community structure of pocket similarity network”, *International Journal of Bioinformatics Research and Applications*, vol. 4–4, 2008, pp. 445–460.

[Lyn02] Lyne, P. D. “Structure-based virtual screening: an overview”, *Drug Discovery Today*, vol. 7–20, 2002, pp. 1047–1055.

[LZ06] Lyman, E.; Zuckerman, D. M. “Ensemble-based convergence analysis of biomolecular trajectories”, *Biophysical Journal*, vol. 91–1, 2006, pp. 164–172.

[Mac06] Macarron, R. “Critical review of the role of hts in drug discovery”, *Drug Discovery Today*, vol. 11–7, 2006, pp. 277–279.

[MGH+98] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J.; et al.. “Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function”, *Journal of Computational Chemistry*, vol. 19–14, 1998, pp. 1639–1662.

[MGH+01] Morris, G. M.; Goodsell, D. S.; Huey, R.; Hart, W. E.; Halliday, S.; Belew, R.; Olson, A. J. “Autodock”, *Automated Docking of Flexible Ligands to Receptor-User Guide*, 2001, pp. 1–66.

[MGV+10] Molle, V.; Gulten, G.; Vilchèze, C.; Veyron-Churlet, R.; Zanella-Cléon, I.; Sacchettini, J. C.; Jacobs Jr, W. R.; Kremer, L. “Phosphorylation of InhA inhibits mycolic acid biosynthesis and growth of *mycobacterium tuberculosis*”, *Molecular Microbiology*, vol. 78–6, 2010, pp. 1591–1605.

[MHL+09] Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. “AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility”, *Journal of Computational Chemistry*, vol. 30–16, 2009, pp. 2785–2791.

[MHO08] Morris, G. M.; Huey, R.; Olson, A. J. “Using autodock for ligand-receptor docking”, *Current Protocols in Bioinformatics*, 2008, pp. 8–14.

[MKT02] Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. “Do structurally similar molecules have similar biological activity?”, *Journal of Medicinal Chemistry*, vol. 45–19, 2002, pp. 4350–4358.

[MRB+15] Mortier, J.; Rakers, C.; Bermudez, M.; Murgueitio, M. S.; Riniker, S.; Wolber, G. “The impact of molecular dynamics on drug design: applications for the characterization of ligand–macromolecule complexes”, *Drug Discovery Today*, 2015.

[MSR+08] Machado, K. S.; Schroeder, E. K.; Ruiz, D. D.; Wink, A.; Norberto de Souza, O. “Extracting information from flexible receptor-flexible ligand docking experiments”. In: *Advances in Bioinformatics and Computational Biology*, Springer, 2008, pp. 104–114.

[MSRNdS07] Machado, K. S.; Schroeder, E. K.; Ruiz, D. D.; Norberto de Souza, O. “Automating molecular docking with explicit receptor flexibility using scientific workflows”. In: *Advances in Bioinformatics and Computational Biology*, Springer, 2007, pp. 1–11.

[MSWN02] Ma, B.; Shatsky, M.; Wolfson, H. J.; Nussinov, R. “Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations”, *Protein Science*, vol. 11–2, 2002, pp. 184–197.

[MWRNdS11] Machado, K. S.; Winck, A. T.; Ruiz, D. D.; Norberto de Souza, O. “Mining flexible-receptor molecular docking data”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1–6, 2011, pp. 532–541.

[NQT+14] Niu, M.-m.; Qin, J.-y.; Tian, C.-p.; Yan, X.-f.; Dong, F.-g.; Cheng, Z.-q.; Fida, G.; Yang, M.; Chen, H.; Gu, Y.-q. “Tubulin inhibitors: pharmacophore modeling, virtual screening and molecular docking”, *Acta Pharmacologica Sinica*, vol. 35–7, 2014, pp. 967–979.

[Pau11] Pauli, I. “Estudos *in silico* da interação da enzima InhA de *mycobacterium tuberculosis* com pequenas moléculas do tipo fármaco”, Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, 2011.

[PCC+95] Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham III, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. “AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules”, *Computer Physics Communications*, vol. 91–1, 1995, pp. 1–41.

[PCN11] Phillips, J. L.; Colvin, M. E.; Newsam, S. “Validating clustering of molecular dynamics simulations using polymer models”, *BMC Bioinformatics*, vol. 12–1, 2011, pp. 445–468.

[PdSR+13] Pauli, I.; dos Santos, R. N.; Rostirolla, D. C.; Martinelli, L. K.; Ducati, R. G.; Timmers, L. F.; Basso, L. A.; Santos, D. S.; Guido, R. V.; Andricopulo, A. D.; et al.. “Discovery of new inhibitors of *mycobacterium tuberculosis* InhA enzyme using virtual screening and a 3D-pharmacophore-based approach”, *Journal of Chemical Information and Modeling*, vol. 53–9, 2013, pp. 2390–2401.

[PKB+14] Pan, P.; Knudson, S. E.; Bommineni, G. R.; Li, H.-J.; Lai, C.-T.; Liu, N.; Garcia-Diaz, M.; Simmerling, C.; Patil, S. S.; Slayden, R. A.; et al.. “Time-dependent diaryl ether inhibitors of InhA: Structure-activity relationship studies of enzyme inhibition, antibacterial activity, and in vivo efficacy”, *Journal ChemMedChem*, vol. 9–4, 2014, pp. 776–791.

[PMD+10] Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. “How to improve r&d productivity: the pharmaceutical industry’s grand challenge”, *Nature Reviews Drug Discovery*, vol. 9–3, 2010, pp. 203–214.

[PMF+09] Papaleo, E.; Mereghetti, P.; Fantucci, P.; Grandori, R.; De Gioia, L. “Free-energy landscape, principal component analysis, and structural clustering to identify representative conformations from molecular dynamics simulations: the myoglobin case”, *Journal of Molecular Graphics and Modelling*, vol. 27–8, 2009, pp. 889–899.

[QDPRNdS14] Quevedo, C. V.; De Paris, R.; Ruiz, D. D.; Norberto de Souza, O. “A strategic solution to optimize molecular docking simulations using Fully-Flexible Receptor Models”, *Expert Systems With Applications*, vol. 41–16, 2014, pp. 7608–7620.

[QDS+96] Quemard, A.; Dessen, A.; Sugantino, M.; Jacobs, W. R.; Sacchettini, J. C.; Blanchard, J. S. “Binding of catalase-peroxidase-activated isoniazid to wild-type and mutant mycobacterium tuberculosis enoyl-acp reductases”, *Journal of the American Chemical Society*, vol. 118–6, 1996, pp. 1561–1562.

[Que11] Quevedo, C. V. “Desenvolvimento de um filtro de descritores moleculares geométricos para gerar um ranqueamento em banco de dados de ligantes”, Dissertação de Mestrado, Faculdade de Informática – PUCRS, Porto Alegre, RS, Brasil, 2011, 84p.

[RFS10] Rajan, A.; Freddolino, P. L.; Schulten, K. “Going beyond clustering in md trajectory analysis: an application to villin headpiece folding”, 2010.

[RGB+98] Rozwarski, D. A.; Grant, G. A.; Barton, D. H.; Jacobs, W. R.; Sacchettini, J. C. “Modification of the NADH of the isoniazid target (InhA) from *mycobacterium tuberculosis*”, *Science*, vol. 279–5347, 1998, pp. 98–102.

[Ric77] Richards, F. M. “Areas, volumes, packing and protein structure”, *Annual Review of Biophysics and Bioengineering*, vol. 6, 1977, pp. 151–176.

[Rog11] Rognan, D. “Docking methods for virtual screening: Principles and recent advances”, *Virtual Screening: Principles, Challenges, and Practical Guidelines*, 2011, pp. 153–176.

[RTPRM05] Rodriguez, A.; Tomas, M. S.; Perez, J. J.; Rubio-Martinez, J. “Assessment of the performance of cluster analysis grouping using pharmacophores as molecular descriptors”, *Journal of Molecular Structure: THEOCHEM*, vol. 727–1, 2005, pp. 81–87.

[RVS+99] Rozwarski, D. A.; Vilchèze, C.; Sugantino, M.; Bittman, R.; Sacchettini, J. C. “Crystal structure of the *Mycobacterium tuberculosis* enoyl-ACP reductase, InhA, in complex with NAD+ and a C16 fatty acyl substrate”, *Journal of Biological Chemistry*, vol. 274–22, 1999, pp. 15582–15589.

[SBBW12] Scannell, J. W.; Blanckley, A.; Boldon, H.; Warrington, B. “Diagnosing the decline in pharmaceutical r&d efficiency”, *Nature Reviews Drug discovery*, vol. 11–3, 2012, pp. 191–200.

[SBSNdS05] Schroeder, E. K.; Basso, L. A.; Santos, D. S.; Norberto de Souza, O. “Molecular dynamics simulation studies of the wild-type, I21V, and I16T mutants of isoniazid-resistant *mycobacterium tuberculosis* enoyl reductase (InhA) in complex with NADH: toward the understanding of NADH-InhA different affinities”, *Biophysical Journal*, vol. 89–2, 2005, pp. 876–884.

[SGH+08] Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S.; Brudz, S.; Sullivan, J. P.; Muhlich, J.; Serrano, M.; et al.. "ChemBank: a small-molecule screening and cheminformatics resource database", *Nucleic Acids Research*, vol. 36–suppl 1, 2008, pp. D351–D359.

[SIBW11] Seidel, T.; Ibis, G.; Bendix, F.; Wolber, G. "Strategies for 3d pharmacophore-based virtual screening", *Drug Discovery Today: Technologies*, vol. 7–4, 2011, pp. e221–e228.

[SKK02] Schmitt, S.; Kuhn, D.; Klebe, G. "A new method to detect related function among proteins independent of sequence and fold homology", *Journal of Molecular Biology*, vol. 323–2, 2002, pp. 387–406.

[SMdG07] Sciabola, S.; Morao, I.; de Groot, M. J. "Pharmacophoric fingerprint method (topp) for 3d-qsar modeling: application to cyp2d6 metabolic stability", *Journal of Chemical Information and Modeling*, vol. 47–1, 2007, pp. 76–84.

[SMN+13] Shirude, P. S.; Madhavapeddi, P.; Naik, M.; Murugan, K.; Shinde, V.; Nandishaiah, R.; Bhat, J.; Kumar, A.; Hameed, S.; Holdgate, G.; et al.. "Methyl-Thiazoles: a novel mode of inhibition with the potential to develop novel inhibitors targeting InhA in *mycobacterium tuberculosis*", *Journal of Medicinal Chemistry*, vol. 56–21, 2013, pp. 8533–8542.

[SP04] Sierk, M. L.; Pearson, W. R. "Sensitivity and selectivity in protein structure comparison", *Protein Science*, vol. 13–3, 2004, pp. 773–785.

[SPNW04] Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. "Recognition of functional sites in protein structures", *Journal of Molecular Biology*, vol. 339–3, 2004, pp. 607–633.

[STB+06] Sullivan, T. J.; Truglio, J. J.; Boyne, M. E.; Novichenok, P.; Zhang, X.; Stratton, C. F.; Li, H.-J.; Kaur, T.; Amin, A.; Johnson, F.; et al.. "High affinity InhA inhibitors with activity against drug-resistant strains of *mycobacterium tuberculosis*", *ACS Chemical Biology*, vol. 1–1, 2006, pp. 43–53.

[STTC07] Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. "Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms", *Journal of Chemical Theory and Computation*, vol. 3–6, 2007, pp. 2312–2334.

[SW81] Solis, F. J.; Wets, R. J.-B. "Minimization by random search techniques", *Mathematics of Operations Research*, vol. 6–1, 1981, pp. 19–30.

[SWM11] Shindler, M.; Wong, A.; Meyerson, A. W. “Fast and accurate k-means for large datasets”. In: Advances in Neural Information Processing Systems, 2011, pp. 2375–2383.

[TA08] Totrov, M.; Abagyan, R. “Flexible ligand docking to multiple receptor conformations: a practical alternative”, *Current Opinion in Structural Biology*, vol. 18–2, 2008, pp. 178–184.

[TCM+08] Tintori, C.; Corradi, V.; Magnani, M.; Manetti, F.; Botta, M. “Targets looking for drugs: a multistep computational protocol for the development of structure-based pharmacophores and their applications for hit discovery”, *Journal of Chemical Information and Modeling*, vol. 48–11, 2008, pp. 2166–2179.

[TCX+12] Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. “Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets”, *Journal of Chemical Information and Modeling*, vol. 52–11, 2012, pp. 2884–2901.

[TK03] Teodoro, M. L.; Kavraki, L. E. “Conformational flexibility models for the receptor in structure based drug design”, *Current Pharmaceutical Design*, vol. 9–20, 2003, pp. 1635–1648.

[TSK06] Tan, P.-N.; Steinbach, M.; Kumar, V. “Introduction to data mining”. , Addison-Wesley, 2006.

[Tub14] World Health Organization Global Tuberculosis. “Global tuberculosis control 2013”, Relatório Técnico, 2014, 171p.

[TUST10] Tabei, Y.; Uno, T.; Sugiyama, M.; Tsuda, K. “Single versus multiple sorting in all pairs similarity search.” In: *ACML*, 2010, pp. 145–160.

[TvG94] Torda, A. E.; van Gunsteren, W. F. “Algorithms for clustering molecular dynamics configurations”, *Journal of Computational Chemistry*, vol. 15–12, 1994, pp. 1331–1340.

[TWH01] Tibshirani, R.; Walther, G.; Hastie, T. “Estimating the number of clusters in a data set via the gap statistic”, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63–2, 2001, pp. 411–423.

[UST04] Uno, T.; Sugiyama, M.; Tsuda, K. “Efficient construction of neighborhood graphs by the multiple sorting method”, *Journal of Machine Learning Research*, vol. 5–1, 2004, pp. 1–15.

[vGB90] van Gunsteren, W. F.; Berendsen, H. J. “Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry”, *Angewandte Chemie International Edition in English*, vol. 29–9, 1990, pp. 992–1023.

[VGK05] Velec, H. F.; Gohlke, H.; Klebe, G. “DrugScoreCSD knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction”, *Journal of Medicinal Chemistry*, vol. 48–20, 2005, pp. 6296–6303.

[VMF+09] Velayati, A. A.; Masjedi, M. R.; Farnia, P.; Tabarsi, P.; Ghanavi, J.; ZiaZarifi, A. H.; Hoffner, S. E. “Emergence of new forms of totally drug-resistant tuberculosis bacilli: super extensively drug-resistant tuberculosis or totally drug-resistant strains in iran”, *Chest Journal*, vol. 136–2, 2009, pp. 420–425.

[Was08] Waszkowycz, B. “Towards improving compound selection in structure-based virtual screening”, *Drug Discovery Today*, vol. 13–5, 2008, pp. 219–226.

[WCG11] Waszkowycz, B.; Clark, D. E.; Gancia, E. “Outstanding challenges in protein–ligand docking and structure-based virtual screening”, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1–2, 2011, pp. 229–259.

[WGLM98] Wermuth, C.; Ganellin, C.; Lindberg, P.; Mitscher, L. “Glossary of terms used in medicinal chemistry (iupac recommendations 1998)”, *Pure and Applied Chemistry*, vol. 70–5, 1998, pp. 1129–1143.

[WLG+07] Wang, F.; Langley, R.; Gulten, G.; Dover, L. G.; Besra, G. S.; Jacobs, W. R.; Sacchettini, J. C. “Mechanism of thioamide drug action against tuberculosis and leprosy”, *The Journal of Experimental Medicine*, vol. 204–1, 2007, pp. 73–78.

[WLW03] Wang, R.; Lu, Y.; Wang, S. “Comparative evaluation of 11 scoring functions for molecular docking”, *Journal of Medicinal Chemistry*, vol. 46–12, 2003, pp. 2287–2303.

[WMNdSR09] Winck, A. T.; Machado, K. S.; Norberto de Souza, O.; Ruiz, D. D. “FReDD: supporting mining strategies through a flexible-receptor docking database”. In: *Advances in Bioinformatics and Computational Biology*, Springer, 2009, pp. 143–146.

[WQM+12] Winck, A. T.; Quevedo, C. V.; Machado, K. S.; Norberto de Souza, O.; Ruiz, D. D. “A comparative analysis of public ligand databases based on molecular descriptors”. In: *Advances in Bioinformatics and Computational Biology*, Springer, 2012, pp. 156–167.

[WR10] Weill, N.; Rognan, D. “Alignment-free ultra-high-throughput comparison of druggable protein- ligand binding sites”, *Journal of Chemical Information and Modeling*, vol. 50–1, 2010, pp. 123–135.

[Yan10] Yang, S.-Y. “Pharmacophore modeling and applications in drug discovery: challenges and recent advances”, *Drug Discovery Today*, vol. 15–11, 2010, pp. 444–450.

[YAR11] Yuriev, E.; Agostino, M.; Ramsland, P. A. “Challenges and advances in computational docking: 2009 in review”, *Journal of Molecular Recognition*, vol. 24–2, 2011, pp. 149–164.

[Yur14] Yuriev, E. “Challenges and advances in structure-based virtual screening”, *Future Medicinal Chemistry*, vol. 6–1, 2014, pp. 5–7.

[ZS04] Zhang, Y.; Skolnick, J. “Spicker: Approach to clustering protein structures for near-native model selection”, *Journal of Computational Chemistry*, vol. 25, 2004, pp. 865–871.

APÊNDICE A – HIPÓTESES FARMACOFÓRICAS

Este capítulo apresenta as 25 hipóteses farmacofóricas utilizadas na ferramenta ZINCPharmer para a seleção do conjunto de 957 ligantes. As Figuras A.1, A.2, A.3, A.4 e A.5 ilustram as hipóteses conforme o volume da cavidade de ligação do substrato.

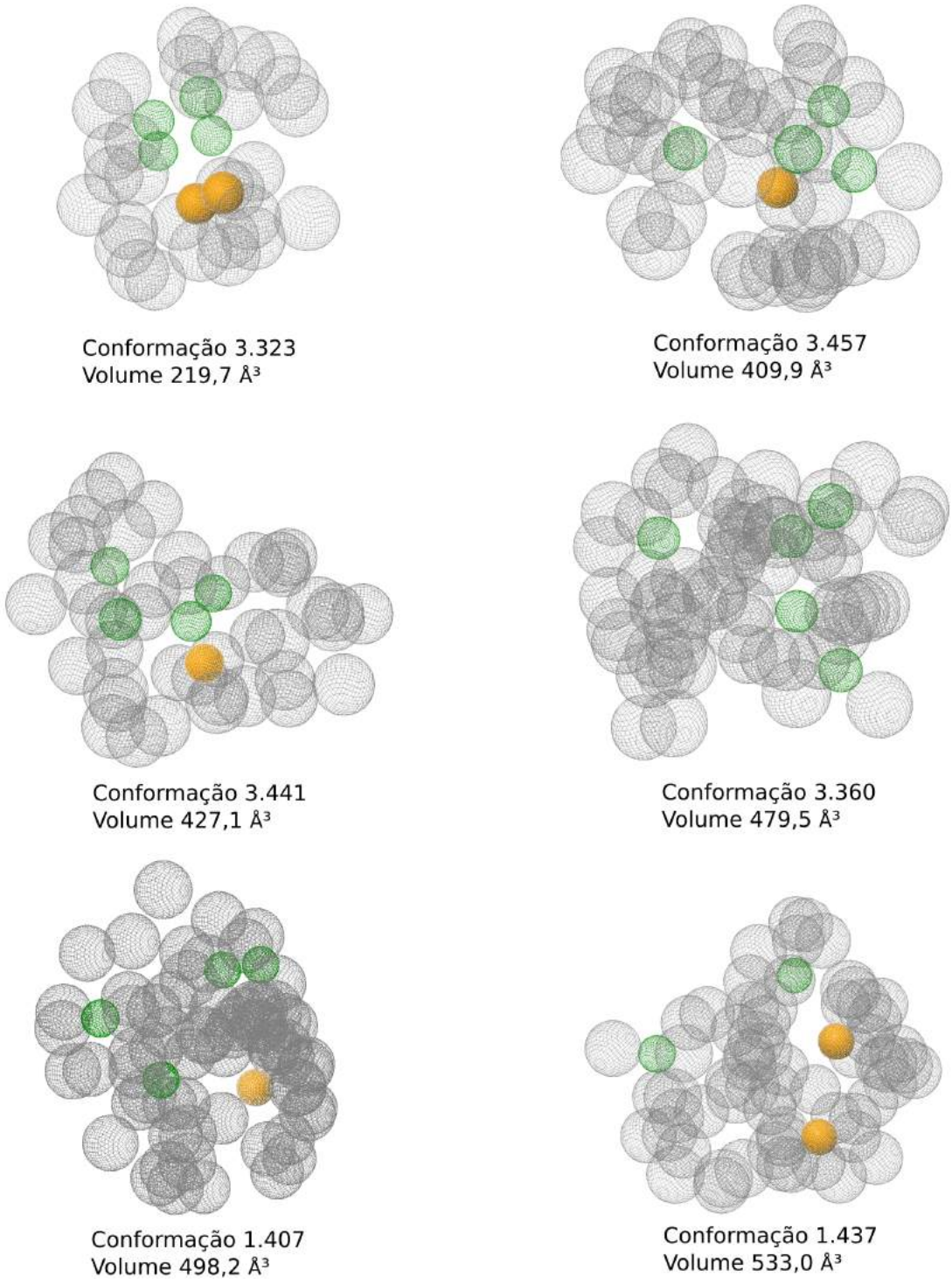
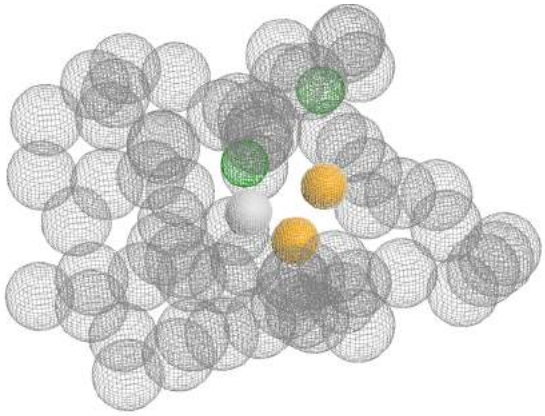
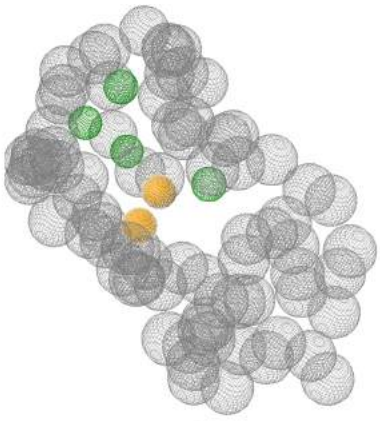


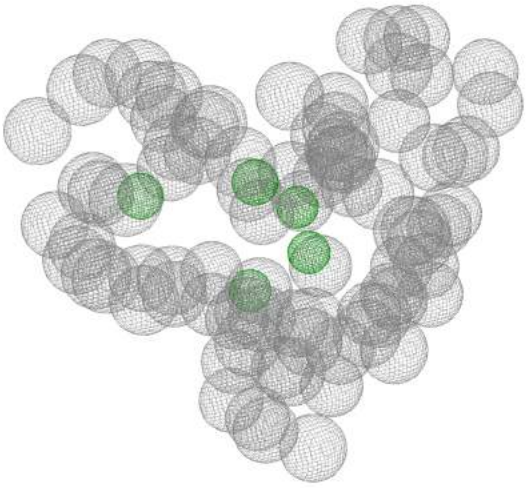
Figura A.1 – Hipóteses farmacofóricas das conformações 1.407, 1.437, 3.323, 3.360, 3.441 e 3.457 do modelo FFR.



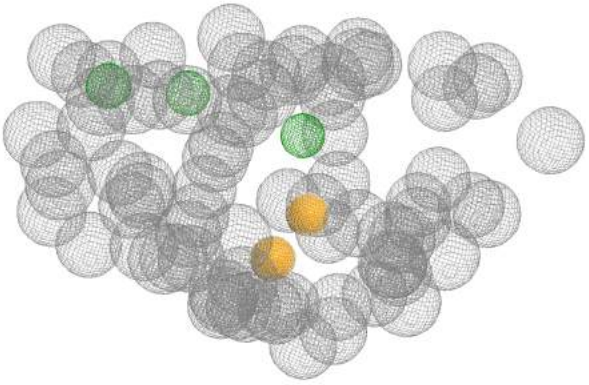
Conformação 5.561
Volume 575,5 Å³



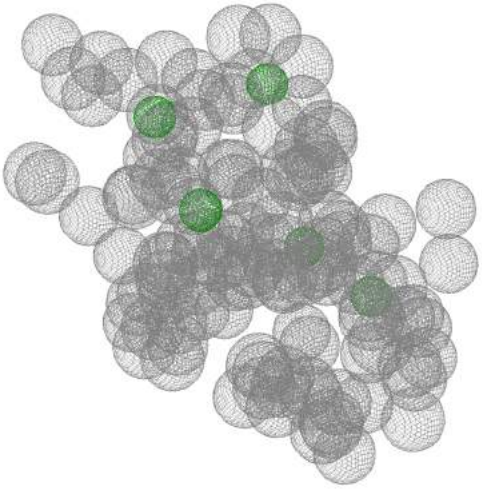
Conformação 18.540
Volume 674,8 Å³



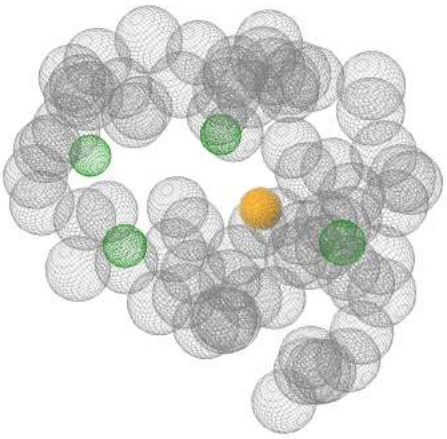
Conformação 17.503
Volume 729,0 Å³



Conformação 1.099
Volume 803,2 Å³

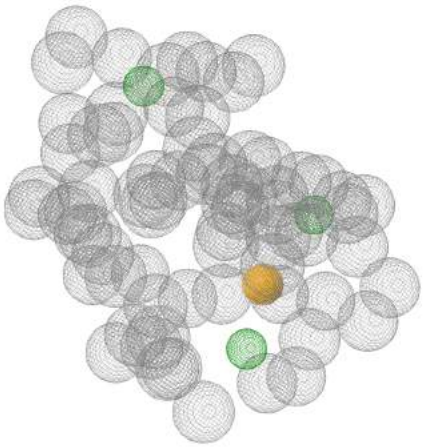


Conformação 17.618
Volume 924,3 Å³

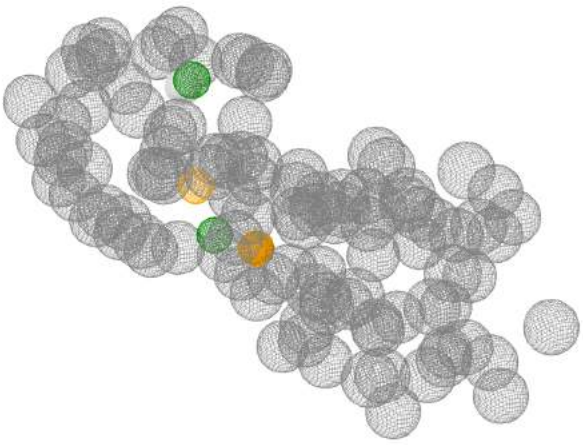


Conformação 14.968
Volume 978,3 Å³

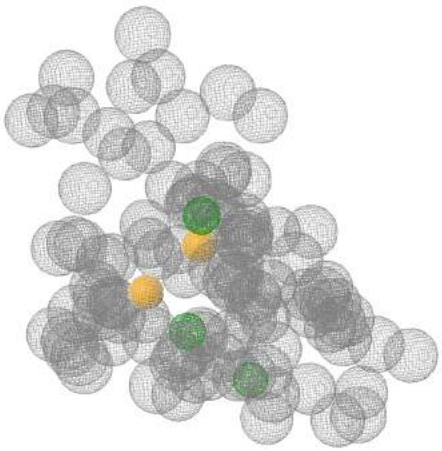
Figura A.2 – Hipóteses farmacofóricas das conformações 1.099, 5.561, 14.968, 17.503, 17.618 e 18.540 do modelo FFR.



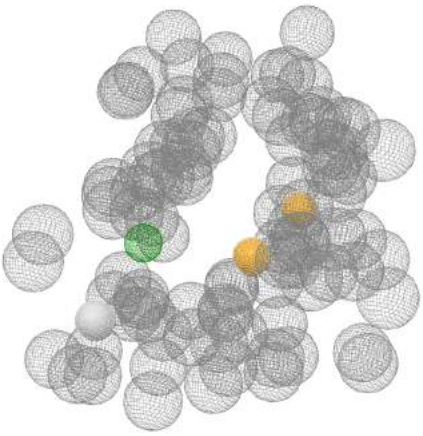
Conformação 14.933
Volume 981,4 Å³



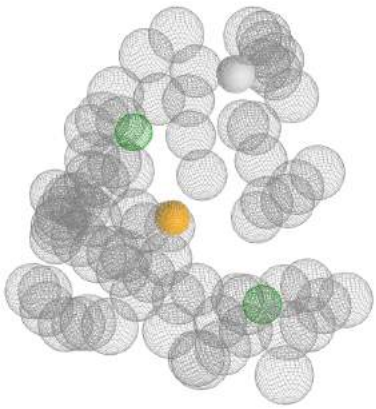
Conformação 776
Volume 999,2 Å³



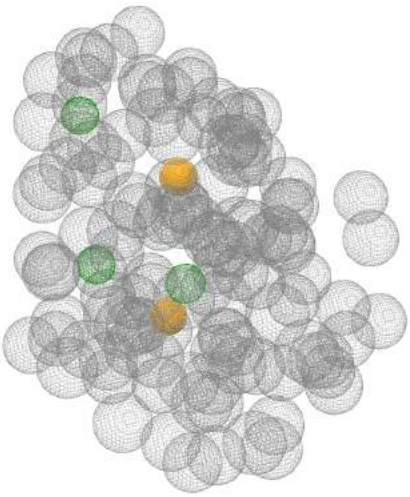
Conformação 2.065
Volume 1.052,5 Å³



Conformação 15.035
Volume 1.075,4 Å³



Conformação 3.029
Volume 1.084,0 Å³



Conformação 19.093
Volume 1.118,1 Å³

Figura A.3 – Hipóteses farmacofóricas das conformações 776, 2.065, 3.029, 14.933, 15.035 e 19.093 do modelo FFR.

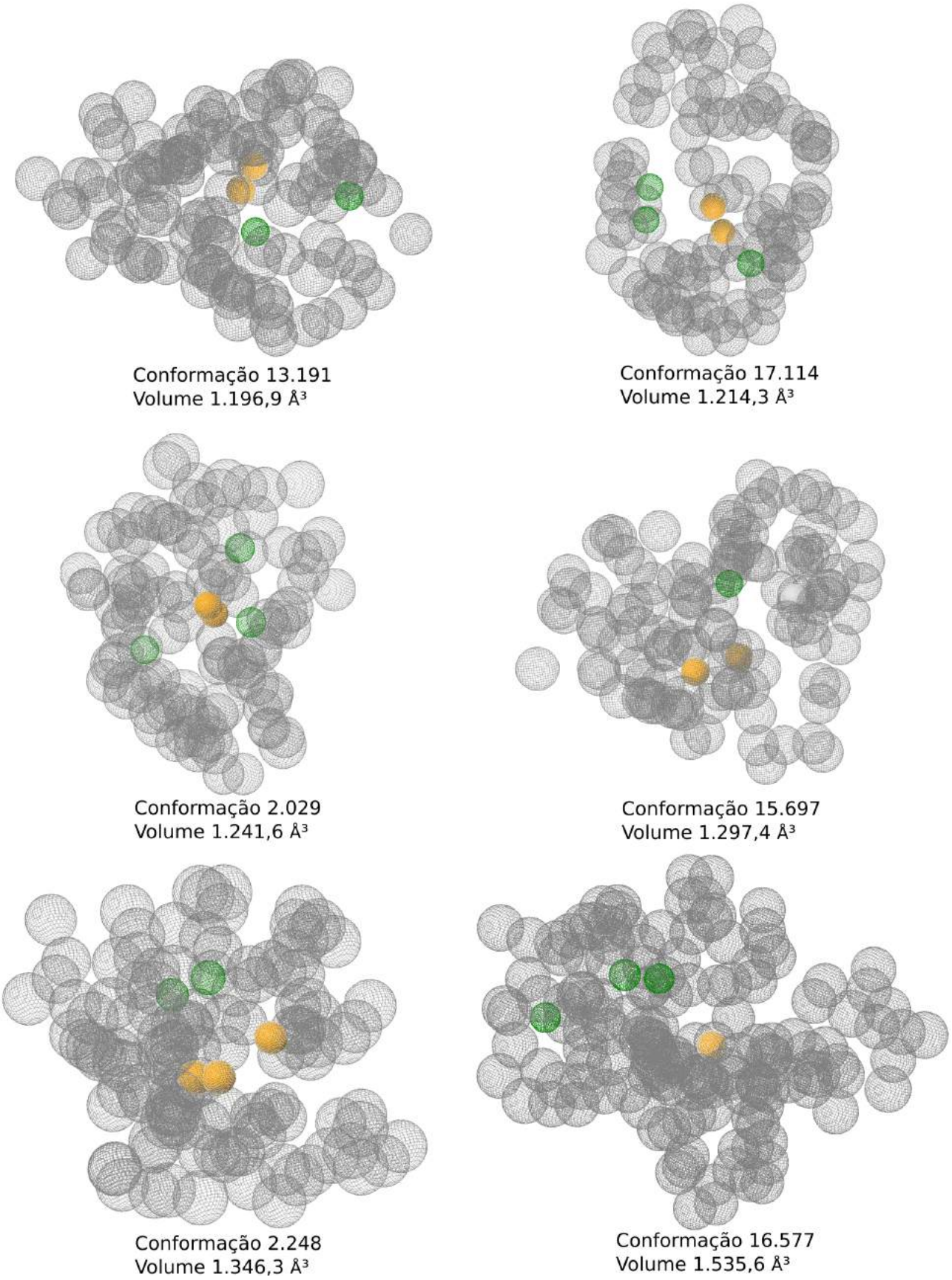


Figura A.4 – Hipóteses farmacofóricas das conformações 2.029, 2.248, 13.191, 15.697, 16.577 e 17.114 do modelo FFR.

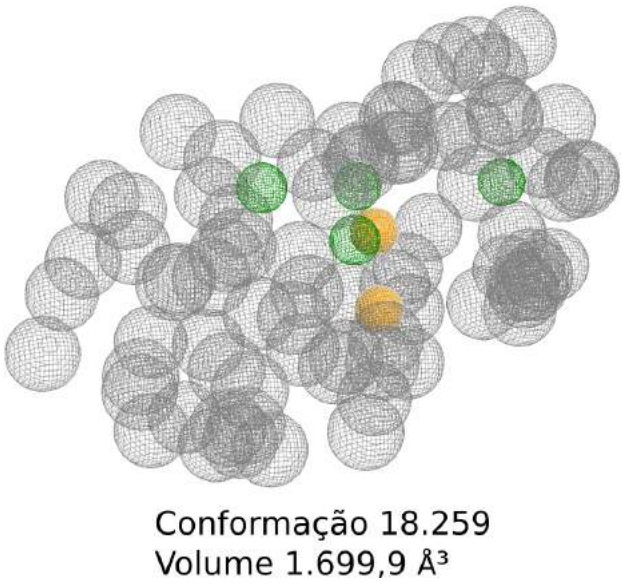


Figura A.5 – Hipótese farmacofórica da conformação 18.259 do modelo FFR.

Naturalmente, existem dezenas de combinações de hipóteses farmacofóricas possíveis para cada modelo farmacofórico extraído das estruturas representativas. A limitação temporal do doutoramento permitiu a investigação de uma hipótese para cada modelo farmacofórico. As estruturas identificadas na primeira seleção já representam as estruturas testadas. Ou seja, não houve a formação de conjuntos por tentativa e erro para encontrar as hipóteses mais adequadas.