

UFRRJ
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM
MATEMÁTICA E COMPUTACIONAL

DISSERTAÇÃO

Estudo da Atividade de Inibidores da Anidrase
Carbônica β de *Mycobacterium Tuberculosis* por
Docagem Molecular, Análise de Componentes
Principais e de Clusters

Winnie Mariana Adão Lopes Teixeira

2025



UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS PROGRAMA DE PÓS-
GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

ESTUDO DA ATIVIDADE DE INIBIDORES DA ANIDRASE CARBÔNICA
 β DE *MYCOBACTERIUM TUBERCULOSIS* POR DOCAGEM
MOLECULAR, ANÁLISE DE COMPONENTES PRINCIPAIS E DE
CLUSTERS

WINNIE MARIANA ADÃO LOPES TEIXEIRA

Sob orientação de
Carlos Mauricio Rabello de Sant'Anna

e co-orientação de
Felipe Leite Coelho da Silva

Dissertação submetida como
requisito parcial para obtenção do
grau de Mestre no Programa de
Pós-Graduação em Modelagem
Matemática e Computacional,
Área de Concentração em
Modelagem Matemática e
Computacional.

Seropédica, RJ, Brasil
Dezembro de 2025

Universidade Federal Rural do Rio de Janeiro
Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada
com os dados fornecidos pelo(a) autor(a)

T266e Teixeira, Winnie Mariana Adão Lopes , --
 Estudo da Atividade de Inibidores da Anidrase
 Carbônica Beta de Mycobacterium Tuberculosis por
 Docagem Molecular, Análise de Componentes Principais
 e de Clusters / Winnie Mariana Adão Lopes Teixeira.
 Seropédica, 2025.
 91 f.

 Orientador: Carlos Mauricio Rabello Sant'Anna.
 Coorientador: Felipe Leite Coelho Da Silva.
 Dissertação (Mestrado). -- Universidade Federal
 Rural do Rio de Janeiro, PPGMMC, 2025.

 1. Anidrase Carbônica beta. 2. Análise de
 Componentes Principais. 3. Clusters. 4. Planejamento
 de Fármacos. 5. Tuberculose e sulfonamidas. I.
 Sant'Anna, Carlos Mauricio Rabello, --, orient. II.
 Da Silva, Felipe Leite Coelho , -, coorient. III
 Universidade Federal Rural do Rio de Janeiro. PPGMMC.
 IV. Título.



MINISTÉRIO DA EDUCAÇÃO

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

INSTITUTO DE CIÊNCIAS EXATAS

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

Seropédica-RJ, 22 de dezembro de 2025.

WINNIE MARIANA ADÃO LOPES TEIXEIRA

Dissertação submetida como requisito parcial para a obtenção de grau de Mestra, no Programa de Pós-Graduação em Modelagem Matemática e Computacional PPGMMC, área de Concentração em Modelagem Matemática e Computacional.

DISSERTAÇÃO APROVADA EM 22/12/2025

Carlos Mauricio Rabello de Sant Anna Dr^o (Orientador, Presidente da Banca-UFRRJ)

Ronaldo Malheiros Gregório Dr^o (membro interno-UFRRJ)

Nelilma Correia Romeiro Dr^a (membro externo à Instituição-UFRJ)



ATA Nº ata/2025 - ICE (12.28.01.23)
(Nº do Documento: 7441)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 29/12/2025 13:22)
CARLOS MAURICIO RABELLO DE SANT ANNA
PROFESSOR DO MAGISTERIO SUPERIOR
DQF (11.39.00.25)
Matricula: ###204#4

(Assinado digitalmente em 31/12/2025 08:04)
RONALDO MALHEIROS GREGORIO
PROFESSOR DO MAGISTERIO SUPERIOR
DeptTL/IM (12.28.01.00.00.90)
Matricula: ###696#7

(Assinado digitalmente em 07/01/2026 16:55)
NELILMA CORREIA ROMEIRO
ASSINANTE EXTERNO
CPF: ###.###.877-##

Visualize o documento original em <https://sipac.ufrrj.br/documentos/> informando seu número: 7441, ano: 2025,
tipo: ATA, data de emissão: 29/12/2025 e o código de verificação: 50e34fc81f

AGRADECIMENTOS

Primeiramente gostaria de agradecer a Deus por ter me sustentado e proporcionado a oportunidade de cursar o mestrado em modelagem matemática. Também gostaria de agradecer a paciência, dedicação, ensinamentos e motivação do professor Carlos Mauricio que me apresentou a esse mundo de química medicinal e o pioneirismo do professor Eliezer Barreiro que me causou grande admiração e felicidade em saber que o Brasil tem pesquisas tão relevantes para a humanidade. Também agradecer meu co-orientador Felipe Leite pela paciência e dedicação me mostrando o quão aplicações estatísticas podem ser interessantes. Agradeço ao programa PPGMMC e aos professores.

Sou grata aos meus pais, Wilza e Francisco e minha irmã Ariane por terem me ajudado, incentivado e dado condições nessa caminhada além de terem batalhado arduamente para que eu chegasse até aqui. Ao Mateus dos Santos Rodrigues que me apoiou, me motivou e acalmou nos momentos mais tempestuosos nesse processo. Agradeço também aos professores Marcela Santos e Aleksandro de Melo por terem estimulado e apoiado a continuação dos meus estudos.

Os autores agradecem à CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, ao CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico e à FAPERJ – Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro pelo apoio fornecido ao longo deste trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

RESUMO

TEIXEIRA, Winnie Mariana Adão Lopes. Estudo da Atividade de Inibidores da Anidrase Carbônica β de *Mycobacterium tuberculosis* por Docagem Molecular, Análise de Componentes Principais e de Clusters. 91f. Dissertação (Mestrado em Modelagem Matemática e Computacional). Instituto de Ciências Exatas, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2025.

A tuberculose é uma doença infecciosa grave causada pela micobactéria *Mycobacterium tuberculosis*. Ela é tratável e curável, porém existem casos em que surgem cepas multirresistentes aos medicamentos de primeira e segunda linha, o que leva à necessidade de se identificar novos candidatos a antibióticos contra essa micobactéria. Esse trabalho tem como objetivo usar um conjunto de compostos da literatura com atividade inibitória sobre a enzima anidrase carbônica, um alvo promissor para ação sobre *M. tuberculosis*, para se desenvolver modelos de classificação da atividade inibitória por métodos de modelagem molecular e análise estatística. Os resultados preliminares mostram que a técnica de docagem molecular levou a resultados promissores em termos da classificação dos compostos como sendo de baixa ou de alta atividade inibitória. Foram explorados os dados calculados relacionadas a esses compostos e seus complexos com a enzima para encontrar um conjunto de variáveis de menor dimensionalidade que tenha um potencial classificador de forma que possam ser usados para o estudo de compostos candidatos a antibióticos contra essa micobactéria. Para isso, foi usada a técnica de análise de componentes principais para reduzir a dimensionalidade preservando características do conjunto e a análise de cluster para separar o conjunto de variáveis de acordo com a similaridade dos elementos a partir dos dados disponibilizados pelos componentes principais. A clusterização confirmou os padrões dos resultados da análise de componentes principais.

Palavras-chave: Modelagem Matemática, tuberculose, docagem molecular, inibidores da anidrase carbônica, análise de componentes principais, *Clusters*.

ABSTRACT

TEIXEIRA, Winnie Mariana Adão Lopes. Development of classification models inhibitory activity of Mycobacterium Tuberculosis β -carbonic anhydrase. 2025. 91p. Dissertation (Master in Mathematical and Computational Modeling). Instituto de Ciências Exatas, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2025.

Tuberculosis is a serious infectious disease caused by the mycobacterium *Mycobacterium tuberculosis*. It is treatable and curable disease, however, there are cases in which multidrug-resistant strains emerge, leading to the need to identify new antibiotic candidates against this mycobacterium. This work aims to use a set of compounds from the literature with inhibitory activity on the carbonic anhydrase enzyme, a promising target for action against *M. tuberculosis*, to develop models for classifying inhibitory activity using molecular modeling and statistical analysis methods. Preliminary results show that the molecular docking technique led to promising results in terms of classifying compounds as having low or high inhibitory activity. Calculated data related to these compounds and their complexes with the enzyme were explored to find a set of lower-dimensional variables with classification potential that can be used to study antibiotic candidate compounds against this mycobacterium. To achieve this, principal component analysis was used to reduce dimensionality while preserving set characteristics, and cluster analysis was employed to separate the set of variables according to the similarity of elements based on the data provided by the principal components. The clustering confirmed the patterns of the principal component analysis results.

Keywords: Mathematical modeling, tuberculosis, molecular docking, carbonic anhydrase inhibitors, principal component analysis, clusters.

SUMÁRIO

1	INTRODUÇÃO	1
1.1	Planejamento de fármacos	4
1.1.1	Anidrase carbonica β	7
1.2	Modelagem molecular	8
1.2.1	Spartan'24	8
1.2.2	Gold 2023.2.0	9
1.3	Análise multivariada de dados	11
1.3.1	Matriz correlação.....	11
1.3.2	Análise com componentes principais e clusters.....	13
2	OBJETIVOS.....	15
3	REVISÃO BIBLIOGRÁFICA.....	16
4	METODOLOGIA.....	19
4.1	Seleção e construção da anidrase β -carbonica da Mycobacterium Tuberculosis	19
4.2	Seleção e construção dos compostos.....	21
4.2.1	Docagem molecular	26
4.3	Análise de dados multivariados	27
4.3.1	Conjunto de dados	27
4.4	Análise de componentes principais.....	31
4.5	Análise de <i>cluster</i>	33
5	DISCUSSÃO E RESULTADOS.....	35
5.1	Análise qualitativa da interação ligante-enzima	35
5.2	Análise exploratória de dados	52

5.2.1	Matriz correlação.....	53
5.3	Análise de componentes principais e análise de cluster	57
5.3.1	Análise de componentes principais.....	58
5.3.2	Análise de Clusters (Agrupamento).....	62
6	CONCLUSÃO.....	67
7	Referências bibliográficas.....	69
Anexo		74

Lista de Figuras

Figura 1.1 – Coeficiente de incidência (casos por 100 mil hab.) e número de casos novos de tuberculose. Brasil, 2013 a 2024 Brasil (BRASIL, 2025).	2
Figura 1.2 – Proporção de casos novos de tuberculose pulmonar por etnia/cor no Brasil de 2013 a 2024. Fonte: (BRASIL, 2025).	2
Figura 1.3 – Casos novos de tuberculose diagnosticados em populações em situação de vulnerabilidade no Brasil de 2015 a 2025.....	3
Figura 1.4 – Etapas envolvidas no processo de descoberta e desenvolvimento de fármacos.	5
Figura 1.5 – Representação do modelo "chave-fechadura".....	6
Figura 4.1 – Fluxograma das etapas do estudo	19
Figura 4.2 – estrutura experimental da CA de <i>M. tuberculosis</i> 2A5V.....	20
Figura 4.3 – Comparação da AC de <i>M. tuberculosis</i> 2A5V e a AC humana 1CA2 (Fonte: RCSB PDB).....	20
Figura 4.4. Séries de moléculas 23 e 24, conforme Abdoli et al., 2023.....	25
Figura 4.4 – Modelo comparativo da CA de <i>M. tuberculosis</i> . Figura gerada com o programa PyMol	21
Figura 5.1 – Sobreposição do modelo comparativo da CA de <i>M. tuberculosis</i> ...	35
Figura 5.2 – Sobreposição das poses de docagem no programa GOLD.....	37
Figura 5.3 – Melhor solução de docagem no programa GOLD do ligante 13, com a função GoldScore. Átomos de H omitidos para maior clareza.....	37
Figura 5.4 – Matriz correlação do conjunto de dados com a função GoldScore..	53
Figura 5.5 – Matriz correlação do conjunto de dados com a função ChemScore...	53
Figura 5.6 – Matriz correlação do conjunto de dados com a função ChemPLP...	54
Figura 5.7 – Matriz correlação do conjunto de dados com a função ASP.....	54

Figura 5.8 – Matriz correlação do conjunto de dados com a função GoldScore para os compostos ionizados.....	55
Figura 4.9 – Matriz correlação do conjunto de dados com a função ChemScore para os compostos ionizados.....	55
Figura 4.10 – Matriz correlação do conjunto de dados com a função ChemPLP para os compostos ionizados.....	56
Figura 4.11 –Matriz correlação do conjunto de dados com a função ASP para os compostos ionizados.....	56
Figura 4.12 – Gráfico cotovelo função ChemPLP.....	58
Figura 4.13 – Contribuição das variáveis no primeiro componente principal da função GoldScore.....	62

Lista de Tabelas

Tabela 4.1 – Estruturas químicas dos compostos inibidores da CA β e constante de inibição em MtCA2, extraídas de (CARTA, 2009).....	22
Tabela 4.2 – Estruturas químicas dos compostos inibidores da CA β e constante de inibição em MtCA2, extraídas de (CARTA, 2009).....	25
Tabela 4.3 – Tabela 4.3. Substituintes dos inibidores da MtCA2 e respectivas constantes de inibição (séries 23 e 24 de Abdoli et al., 2023; números entre parênteses são os adotados neste trabalho).....	25
Tabela 4.4 – Relação dos descritores moleculares obtidos com o método PM6 (programa Spartan'24).....	28
Tabela 4.5 – Relação dos descritores moleculares obtidos com o programa GOLD 2023.2.0.....	28
Tabela 5.1 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função GoldScore.....	38
Tabela 5.2 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ChemPLP.....	40
Tabela 5.3 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ChemScore.....	42
Tabela 5.4 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ASP.....	43
Tabela 5.5 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função GoldScore, para moléculas ionizadas.....	45
Tabela 5.6 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ChemPLP, para moléculas ionizadas.....	47
Tabela 5.7 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ChemScore, para moléculas ionizadas.....	48
Tabela 5.8 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ASP, para moléculas	

ionizadas.....	50
Tabela 5.9 – Percentual de acerto a partir dos parâmetros estipulados nas funções.....	52
Tabela 5.10 - PCA da função GoldScore.....	58
Tabela 5.11 - PCA da função ChemPLP.....	59
Tabela 5.12 - PCA da função ChemScore.....	59
Tabela 5.13 - PCA da função ASP.....	60
Tabela 5.14 - PCA da função GoldScore com moléculas ionizadas.....	60
Tabela 5.15 - PCA da função ChemPLP com moléculas ionizadas.....	60
Tabela 5.16 - PCA da função ChemScore com moléculas ionizadas.....	61
Tabela 5.17 - PCA da função ASP com moléculas ionizadas.....	61
Tabela 5.18 - Acertos no agrupamento de cada componente principal.....	63
Tabela 5.19 - Acertos no agrupamento de cada componente principal com moléculas ionizadas.....	63

Lista de Abreviações e Siglas

ACP	Análise de componentes principais
AC	Anidrase Carbônica
CP	Componente principal
TB	Tuberculose
HIV	Virus da imunodeficiência humana
MDR	Tuberculose resistente rifampicina e isoniazida
MSF	Médicos sem fronteira
mtCA	Anidrase Carbônica da <i>Mycobacterium tuberculosis</i> .
P&D	Pesquisa de descoberta e desenvolvimento de fármacos
PDB	Protein Data Bank
Pré-XDR	Tuberculose pré-extensivamente fármacorresistente
SBDD	Planejamento de fármaco baseado na estrutura do receptor (<i>Structure-based drug design</i>)
RR	Tuberculose resistente a rifampicina
SUS	Sistema Único de Saúde
XDR	Tuberculose extremamente resistente
WHO	<i>World Health Organization</i>

A tuberculose (TB) é uma doença infecciosa e transmissível, causada pelo bacilo *Mycobacterium tuberculosis* que afeta principalmente os pulmões, mas que também pode ser extrapulmonar. Segundo o "*Global Tuberculosis 2025 report*" (relatório global da tuberculose 2025) da Organização Mundial da Saúde a TB foi a principal causa de morte por agente infeccioso no mundo em 2024. O número global estimado de pessoas diagnosticadas com TB em 2024 foi de 10,7 milhões, com 1,23 milhão de mortes, estando entre as 10 principais causas de morte no mundo (OMS, 2025).

No Brasil, segundo o Boletim epidemiológico de tuberculose de março de 2025 do Ministério da Saúde, 84.308 pessoas com casos novos de TB foram diagnosticadas e tratadas e mais de 6 mil óbitos, mostrando aumento de casos e mortalidade no período pós pandemia. Houve 1.047 casos novos de tuberculose resistente a fármacos (BRASIL, 2025).

A transmissão ocorre através do ar quando pessoas com a tuberculose ativa pulmonar ou de laringe tosem, espirram, falam ou cospem, mas o risco de transmissão começa a diminuir já após os 15 primeiros dias de tratamento. A TB geralmente é curável por meio de tratamento e evitável por meio de medidas preventivas como a imunização pela BCG, pelo tratamento preventivo da tuberculose, que trata a doença em portadores da micobactéria antes de se tornar ativa, e pelo controle e prevenção da disseminação. (BRASIL, 2017b; BRASIL, 2024).

Podemos observar na figura 1.1, o coeficiente de incidência e o número de casos novos no Brasil vem crescendo em relação a 2013, com exceção dos anos 2020 e 2021, que foi o período mais ativo da pandemia de Covid-19.

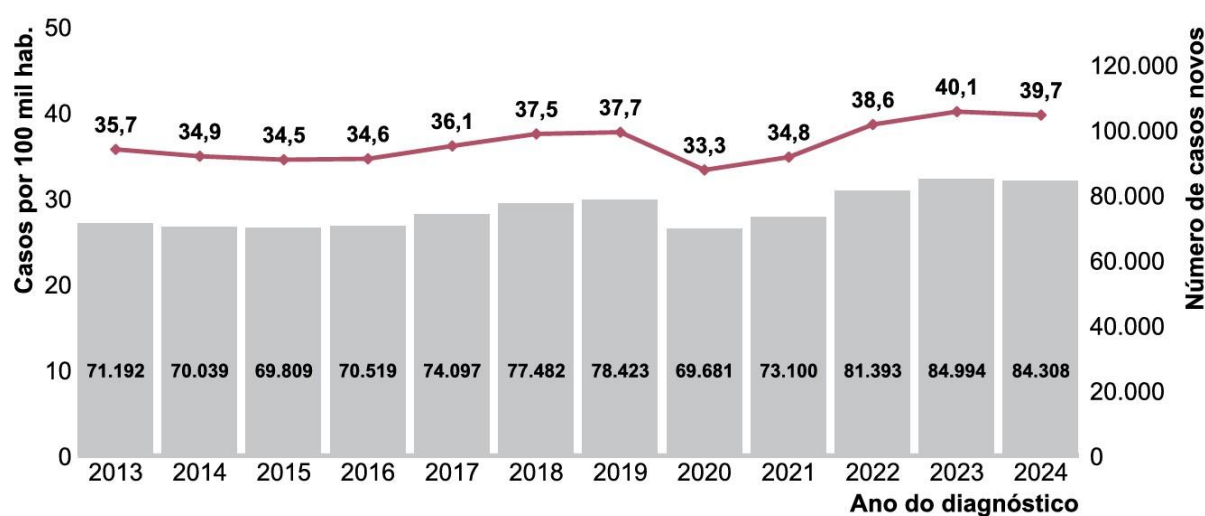


Figura 1.1 – Coeficiente de incidência (casos por 100 mil hab.) e número de casos novos de tuberculose. Brasil, 2013 a 2024 Brasil (BRASIL, 2025).

Dentro desse cenário, foi estabelecido que o grupo prioritário que é composto por pessoas que são portadoras do vírus HIV, pessoas em contato com portadores da tuberculose ativa, afetadas pela desnutrição, tabagismo, diabetes, indivíduos que são dependentes do uso do álcool e crianças menores de 2 anos (BRASIL, 2022b). A TB também é considerada parte do grupo de enfermidades causadoras e perpetuadoras da pobreza, da fome e das iniquidades sociais (BRASIL, 2024). Podemos observar que na figura 1.2 que a proporção de casos novos de TB tem maior aumento nas pessoas que se autodeclaram pretas e pardas. A figura 1.3 apresenta o gráfico mostrando o quantitativo de novos casos de pessoas em vulnerabilidade social.

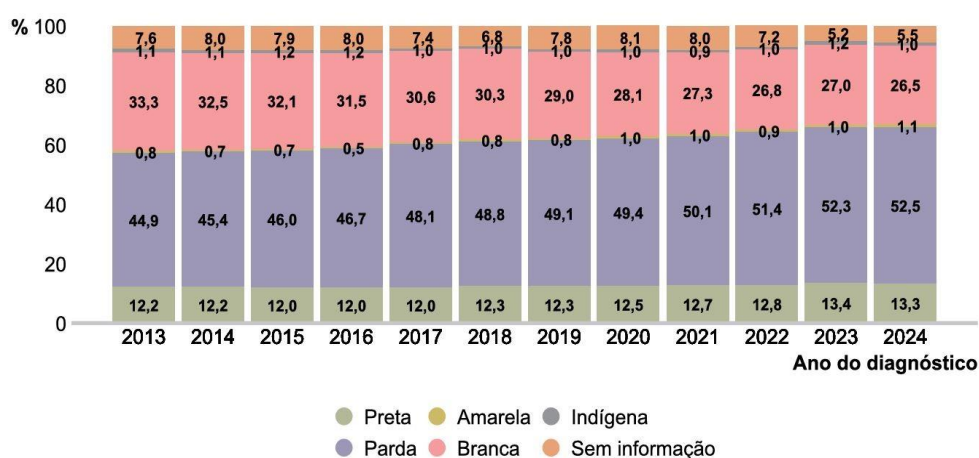


Figura 1.2 – Proporção de casos novos de tuberculose pulmonar por etnia/cor no Brasil de 2013 a 2024. Fonte: (BRASIL, 2025).

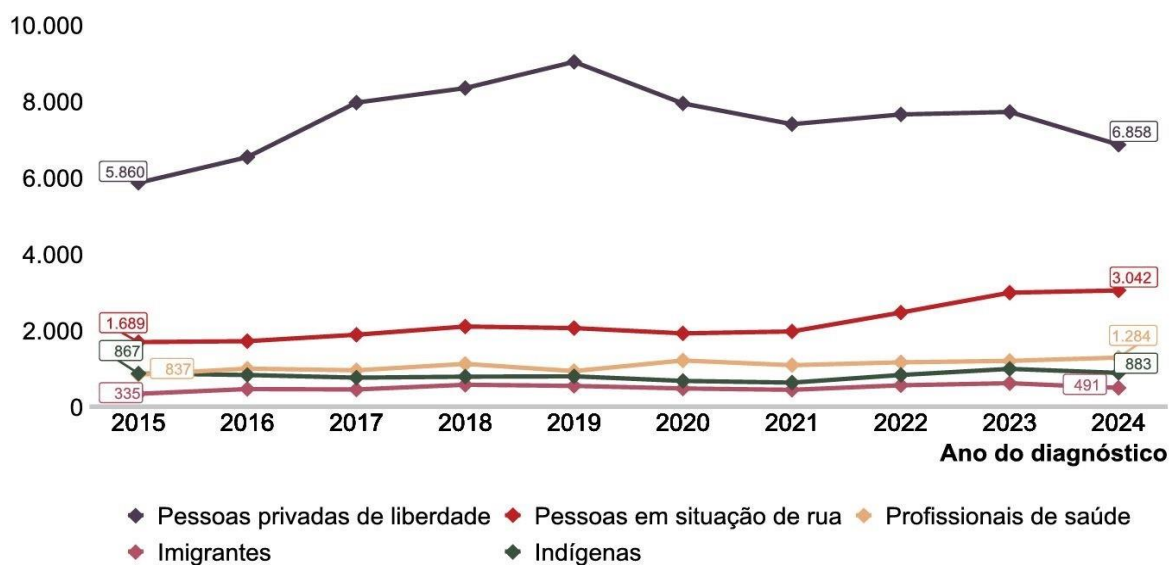


Figura 1.3 – Casos novos de tuberculose diagnosticados em populações em situação de vulnerabilidade no Brasil de 2015 a 2025. Fonte: (BRASIL, 2025).

Os fatores citados acima são dados que reafirmam a tuberculose como uma preocupação legítima e urgente. Em 2017 o Brasil apresentou o “Plano Nacional pelo Fim da Tuberculose como Problema de Saúde Pública” (BRASIL, 2017a) com o compromisso das estratégias globais estipuladas pela OMS. Nesse plano nacional a meta é que até 2035 haja menos de 10 casos por 100 mil habitantes e 1 óbito a cada 100 mil habitantes. Para atingir essas metas foram estipulados três pilares: 1. Prevenção e cuidado integrado e centrado no paciente; 2. Políticas arrojadas e sistema de apoio e 3. Intensificação da pesquisa e inovação (BRASIL, 2017a).

Como um dos objetivos do pilar 1 é prezar pelos princípios da universalidade, da integralidade e da equidade do cuidado, o diagnóstico e o tratamento dessa doença estão disponíveis no SUS (Sistema Único de Saúde) (BRASIL, 2021) sendo o tratamento realizado com antibióticos de primeira linha que são a rifampicina (R), a isoniazida (H), a pirazinamida (P) e o etambutol (E). São usados no esquema básico conhecido como 2RHZE/4RH, ou seja, com duração de seis meses sendo dois meses com a combinação dos quatro antibióticos e quatro meses com a rifampicina e a isoniazida que pode se estender em casos de complicações ou em pessoas em grupos de risco. Em caso intolerância a algum fármaco, adota-se outros esquemas combinando esses medicamentos, mas excluindo o medicamento que causa intolerância (ARBEX, 2010a).

Entretanto, em alguns casos podem surgir cepas bacterianas resistentes ao tratamento, dada pela constante multiplicação em indivíduos que não fazem o tratamento, ou fazem o tratamento inadequado ou interrompem o tratamento, o que resulta na disseminação do bacilo resistente no ambiente, tornando o tratamento mais custoso e com maior duração (BRASIL, 2024). Por isso, é importante frisar que tratamento não pode ser interrompido para evitar que a cepa se torne resistente e contamine outras pessoas. Quando a bactéria *M. tuberculosis* apresenta resistência aos medicamentos de primeira linha, ela causa a chamada tuberculose multirresistente. O tratamento será feito com os medicamentos de segunda linha, ou seja, com a estreptomicina, o etambutol, a terizidona, a pirazinamida e uma quinolona (levofloxacina ou ofloxacina), podendo ser utilizada a amicacina no lugar da estreptocina, em caso de necessidade (ARBEX, 2010b).

As cepas resistentes que causam a TB são classificadas em cinco categorias pela OMS. São elas a TB multirresistente (MDR) que é quando a micobactéria é resistente a rifampicina e isoniazida, a TB resistente a rifampicina (RR), a TB resistente a isoniazida, a TB pré-extensivamente fármacorresistente (pré-XDR), onde o patógeno é resistente à rifampicina e a qualquer fluoroquinolona, e a TB extremamente resistente (XDR), que é classificada quando a TB é resistente a rifampicina, a qualquer fluoroquinolona e pelo menos à bedaquilina ou linezolida (OMS, 2024).

Com isso, a resistência da TB leva a outros tipos de tratamento, mas com menor taxa de cura. A doença pode se manifestar de maneiras mais graves em portadores de HIV, ou seja, do vírus da imunodeficiência humana (MSF, 2024). Podemos encontrar os tratamentos feitos em cada caso de tuberculose em território nacional em (BRASIL, 2022).

1.1 Planejamento de fármacos

Em vista disso, a busca por novos medicamentos para combater esse patógeno é de grande importância. O processo de pesquisa e desenvolvimento de fármacos é custoso, demorado e de grande complexidade (GUIDO et al., 2008). Esse processo é separado em duas fases: a primeira é chamada de fase de descoberta ou pré clínica, sendo o procedimento de identificar e otimizar moléculas que possam representar

novas entidades químicas (NCE, do inglês *New Chemical Entities*); a segunda fase é o desenvolvimento, onde se fazem os testes necessários para garantir a segurança, eficiência e eficácia do medicamento para o ser humano. O esquema representado na figura (GUIDO et al., 2010) descreve as etapas dessa elaboração.

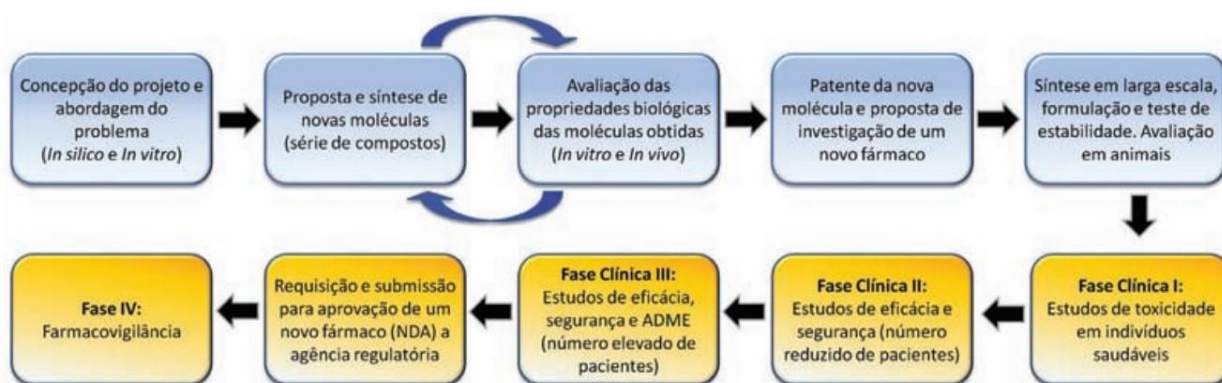


Figura 1.4 – Etapas envolvidas no processo de descoberta e desenvolvimento de fármacos (ADME- Absorção, Distribuição, Metabolismo e Excreção; NDA- solicitação para um novo fármaco, do inglês *New Drug Application*). Fonte: Guido, 2010.

Após a identificação da patologia, se forem conhecidos alvos validados para o tratamento dela, a próxima etapa é descobrir e testar compostos selecionados no alvo biológico em testes apropriados (LOMBARDINO, 2004).

O avanço tecnológico e científico tem auxiliado na descoberta de inovações terapêuticas, com contribuições das áreas da química, biologia e bioquímica. Com a identificação de alvos moleculares validados para uma patologia, pode-se analisar as interações destes com novos compostos como parte do processo de busca de candidatos a fármacos. Nesse processo, podem ser usados métodos computacionais como, por exemplo, a modelagem molecular, a docagem molecular, a triagem virtual, entre outros (GUIDO et al, 2010).

A química medicinal, na qual esse projeto se insere, tem caráter multidisciplinar, envolvendo biologia molecular e estrutural, informática, bioquímica, síntese orgânica e farmacologia, entre outros (GUIDO et al, 2010). Nela, se projetam, preparam e avaliam moléculas candidatas a fármacos. Essas moléculas, em geral, se ligam em sítios específicos em alvos proteicos, desencadeando uma resposta biológica, sendo

determinantes nesse processo forças intermoleculares e propriedades físico-químicas, entre outros (BARREIRO; RODRIGUES *et al*, 1996; BARREIRO *et al*, 2014).

O modelo "chave-fechadura" proposto por Emil Fisher em 1894 descreve simplificadaamente essa interação onde a substância bioativa, geralmente chamada ligante, é considerada a chave e a enzima a fechadura. Nesse conceito é preciso ter a "chave" que possua o encaixe perfeito com a "fechadura" como representado na figura 1.5. Na prática, essa interação depende em grande parte do volume molecular, das forças intermoleculares e também de aspectos dinâmicos, já que tanto a substância bioativa quanto a enzima são estruturas flexíveis, que se ajustam conforme interagem. (MOREIRA, 2015; SANTOS, 2021).

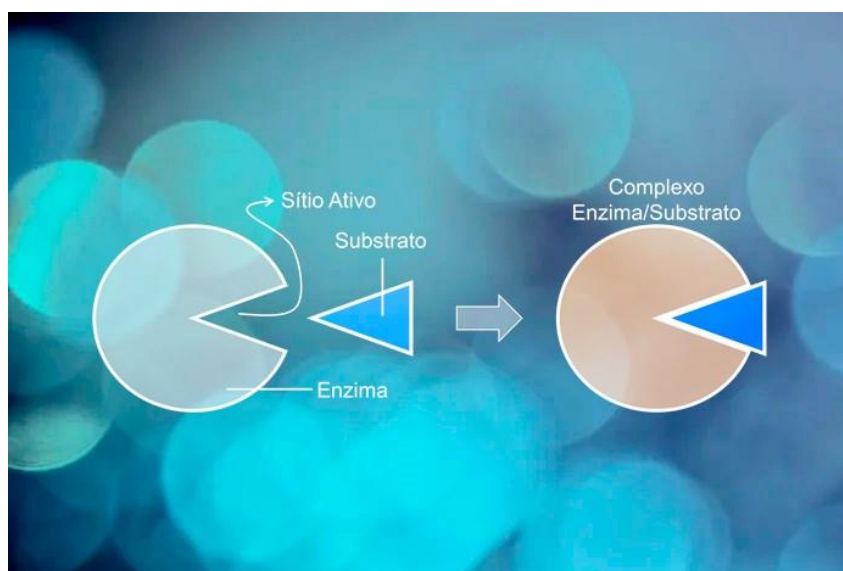


Figura 1.5 – Representação do modelo "chave-fechadura". Fonte: (SANTOS, 2021).

A ênfase desse trabalho que será a parte do processo da interpretação do mecanismo de ação molecular de compostos bioquimicamente ativos contra alvos presentes no patógeno causador da TB e das relações entre a estrutura química e a atividade.

A ênfase desse trabalho será a interpretação do mecanismo de ação molecular de compostos bioquimicamente ativos contra uma enzima presente no patógeno causador da TB, a anidrase carbonica β , e das relações entre a estrutura química dos compostos e a atividade. Serão usados dados experimentais de ensaios *in vitro* disponibilizados na literatura para desenvolver um estudo da atividade inibitória desses compostos na enzima através da simulação computacional.

1.1.1 Anidrase carbônica β

Foi comprovado que as anidrases carbônicas (CA) são alvos moleculares promissores para o desenvolvimento de novos candidatos a fármacos anti-TB, com um potencial real para o desenvolvimento de agentes antimicobacterianos com um mecanismo de ação diversificado em comparação com os dos medicamentos utilizados clinicamente para os quais muitas cepas apresentam resistência a múltiplos medicamentos e extensa resistência a múltiplos medicamentos (ASPATWAR, 2019).

Em seres humanos, a reação catalisada pela CA é essencial para o transporte de CO₂ no sangue, entre outras funções. Essa enzima também é encontrada em plantas, animais e bactérias sendo essencial para seu funcionamento fisiológico. As CAs humana são da classe α , ou seja, anidrases carbônicas α , enquanto as CAs da *M. tuberculosis* (MtCA) são da classe β , ou seja, anidrases carbônicas β , que não estão presentes em mamíferos (COVARRUBIAS, 2005).

M. tuberculosis contém três genes que codificam diferentes isoformas da MtCA, Rv1284 (que codifica a isoforma MtCA 1), Rv3588c (que codifica a isoforma MtCA 2) e Rv3273 (que codifica a isoforma MtCA 3) (CARTA, 2009).

As CAs catalisam a hidratação reversível do CO₂ para formar bicarbonato, uma reação necessária para muitas funções, incluindo a assimilação de carbono e a homeostase do pH (Covarrubias, 2006). Em experimentos *in vitro*, o gene Rv3588c de *M. tuberculosis* se mostrou essencial para o crescimento do patógeno e dependente do pH em 8,4 na sua forma ativa.

O sequenciamento dos genes possibilitou o melhor conhecimento da estrutura e potenciais inibidores, entre eles as sulfonamidas e seus bioisómeros, que em testes *in vitro* e *in vivo* se mostraram eficientes na inibição das β -CAs (ASPATWAR, 2019). Deve ser destacado que já há sulfonamidas que apresentam um amplo efeito antibacteriano, como a *p*-aminobenzeno-sulfonamida (sulfanilamida), mas estas atuam por um outro mecanismo conhecido, a inibição da enzima di-hidropteroato sintetase, que impede as bactérias de sintetizar o ácido fólico, um metabólito essencial para o seu crescimento e replicação (GOODMAN, A. 1996).

1.2 Modelagem molecular

O processo de planejamento de fármacos auxiliado por computador (CADD - *Computer Assisted Drug Design*) envolve o cálculo de propriedades moleculares relacionadas à atividade biológica, combinado com o uso de ferramentas para visualização gráfica, acesso a banco de dados, entre outras. O CADD reúne conhecimentos da Química Orgânica, Bioquímica, Biologia Molecular e Farmacologia, aplicados no estudo e descoberta de fármacos (CARVALHO, 2003).

A modelagem molecular desempenha papel essencial no CADD, pois pode fornecer informações sobre a estrutura da proteína-alvo, das moléculas bioativas (chamadas ligantes quando interagem com o alvo) e dos complexos de interação entre elas. Há programas computacionais para produzir modelos otimizados baseados em diferentes critérios, como a partir da minimização de energia do sistema (Barreiro *et al*, 2014). A modelagem molecular possui também ferramentas de desenho e visualização tridimensional que auxiliam a preparação das estruturas. Para o cálculo e minimização de energia podem ser usados diferentes métodos, destacando-se os métodos baseados em campos de força, como a mecânica molecular, e os baseados na mecânica quântica (SANT'ANNA *et al*, 2009; CARVALHO, 2003). O método de docagem molecular é usado para se propor estruturas para os complexos de interação ligante-receptor. Nos subtópicos a seguir são descritos os programas que foram usados para realização desse trabalho.

1.2.1 Spartan'24

O programa Spartan'24 (Wavefunction, Inc.) é um eficiente recurso para construir modelos de estruturas moleculares de modo fácil, tendo uma interface funcional e simples. A partir dessa interface, podemos usar modelos clássicos (mecânica molecular) e também modelos quânticos, incluindo métodos semiempíricos, funcionais de densidade, de Hartree-Fock e de pós-Hartree-Fock, como o de Møller-Plesset. É importante analisar o custo computacional a fim de escolher o método com o melhor custo-benefício. Esses cálculos foram importantes nesse trabalho pois através dele conseguimos otimizar a geometria das moléculas de modo que ela fosse a mais estável, ou seja, com a energia minimizada. E para calcular

as propriedades físico-químicas das moléculas como veremos na metodologia (SPARTAN, 2024).

Essas estruturas otimizadas foram usadas na etapa de docagem molecular, um método do planejamento de fármacos baseado na estrutura do alvo (SBDD, do inglês *Structure Based Drug Design*), que permite a definição do modo de interação das estruturas dos ligantes no sítio ativo da enzima.

1.2.2 Gold 2023.2.0

O programa GOLD (*Genetic Optimisation for Ligand Docking*) realiza a docagem molecular por meio de um algoritmo genético para acoplamento de ligantes flexíveis em sítios ativos de proteínas. Ele quantifica essa interação através de funções de pontuação, que foram otimizadas a fim de prever modos de interação (conformações e orientações) dos ligantes. O programa disponibiliza quatro funções de pontuação: ChemScore, GoldScore, ChemPLP e ASP (GOLD, 2022). A pontuação é o negativo da soma dos termos de energia de interação; quanto melhor a interação, mais negativa é a energia (ou seja, mais energia é liberada) e, portanto, maior a pontuação. As funções de pontuações são calculadas (como descrito no manual do GOLD Doking) por:

A função de pontuação GoldScore contém termos de energia para ligações de hidrogênio proteína-ligante, para energia de van der Waals (vdW) proteína-ligante, para energia de vdW interna do ligante, energia de tensão torsional do ligante (torção interna).

A aptidão da função ChemPLP é calculada como uma soma de diferentes potenciais:

$$\begin{aligned} fitness_{ChemPLP} &= fitness_{\{PLP\}} - (f_{\{chem-hb\}} + f_{\{chem-cho\}} + f_{\{chem-met\}}) \\ fitness_{PLP} &= -(w_{\{PLP\}} \cdot f_{\{PLP\}} + W_{\{lig-clash\}} \cdot f_{\{lig-clash\}} + w_{\{lig-tors\}} \cdot f_{\{lig-tors\}} \\ &\quad + f_{\{chem-cov\}} + w_{\{prot\}} \cdot f_{\{chem-prot\}} + w_{\{cons\}} \cdot f_{\{cons\}}) \end{aligned}$$

Essa função de aptidão inclui um potencial para interação covalente ($f_{\text{chem-cov}}$), considera cadeias laterais flexíveis ($f_{\text{chem-prot}}$) e moléculas de água explícitas, bem como lida com restrições (f_{cons}). $f_{\text{lig-tors}}$ é um potencial para torções do ligante, $f_{\text{lig-clash}}$ potencial de colisão de átomos pesados do ligante, $f_{\text{chem-hb}}$ e $f_{\text{chem-cho}}$, são potenciais dependentes da distância e do ângulo para ligações de hidrogênio, respectivamente, e $f_{\text{chem-met}}$ é um potencial para interação com metais. Os diferentes w representam os coeficientes calculados pelo programa para cada termo.

A função de aptidão ChemScore calcula a energia livre de interação ($\Delta G_{\text{binding}}$) como:

$$\Delta G_{\text{binding}} = \Delta G_0 + \Delta G_{\text{hbond}} + \Delta G_{\text{metal}} + \Delta G_{\text{lipo}} + \Delta G_{\text{rot}}$$

$$\begin{aligned}\Delta G_0 &= v_0 \\ \Delta G_{\text{hbond}} &= v_1 P_{\text{hbond}} \\ \Delta G_{\text{metal}} &= v_2 P_{\text{metal}} \\ \Delta G_{\text{lipo}} &= v_3 P_{\text{lipo}} \\ \Delta G_{\text{rot}} &= v_4 P_{\text{rot}}\end{aligned}$$

A função de aptidão ChemScore inclui $\Delta G_{\text{binding}}$ e mais alguns termos:

$$\text{fitness}_{\text{Chemscore}} = \Delta G_{\text{binding}} + P_{\text{clash}} + c_{\text{internal}} P_{\text{internal}} + (c_{\text{covalent}} P_{\text{covalent}} + P_{\text{constraint}})$$

onde, v_i são os coeficientes de regressão, os termos P são os tipos de contribuições físicas para a interação e c são coeficientes de ponderação. P_{hbond} é o termo para ligação de hidrogênio, P_{rot} para ligações rotativas congeladas do ligante, P_{metal} é para ligação com metais e P_{lipo} é para a contribuição lipofílica ligante-proteína. P_{clash} é o termo para colisões, P_{internal} é o termo para rotações internas do ligante, P_{covalent} é o termo para interações covalentes e $P_{\text{constraint}}$ é o termo para restrições.

Por fim, a função de aptidão ASP tem os seguintes termos:

$$\text{fitness}_{\text{ASP}} = -C_S(S_{\text{map}} + S_{\text{hbond}} + S_{\text{metal}}) - c_{\text{int}} E_{\text{int}} - c_{\text{clash}} E_{\text{clash}}$$

onde C_s é um fator de escala, c são coeficientes, e S_{map} é um potencial estatístico de interação, baseado em observações experimentais, determinado sobre todas as combinações de átomos da proteína e átomos do ligante dentro de um raio de 6,0 Å. É adicionada uma correção se a docagem for feita em alvos que contêm metal (S_{metal}); ao se adicionar essa correção, uma correção de ligação de hidrogênio é incluída por padrão (S_{hbond}). O termo de energia interna (E_{int}) é necessário para evitar a docagem de conformações de ligantes de alta energia, enquanto o termo de colisão (E_{clash}) deve evitar colisões proteína-ligante quando o potencial S_{map} for muito fraco para gerar repulsão suficiente entre os átomos da proteína e do ligante e, ao mesmo tempo, evitar a sobreposição entre átomos sem potencial (ou seja, com poucas observações experimentais, o que acaba gerando um potencial igual a zero).

1.3 Análise multivariada de dados

"A estatística é um conjunto de métodos adequados para recolher, explorar, descrever e interpretar conjuntos de dados numéricos" - essa é a definição exposta por Silvestre (2007) em seu livro, que conclui que o conjunto de dados é a base da estatística fazendo assim com que se tenha maior compreensão sobre o objeto de estudo. A análise estatística se mostra eficaz pois nos permite reduzir dados multivariados, agrupar dados, classificar, discriminar e identificar relações entre conjuntos de dados (PEARSON, 2011; SILVESTRE, 2007).

1.3.1 Matriz correlação

Para avaliar a relação entre as variáveis pode-se calcular a matriz correlação "A correlação mensura a direção e o grau da relação linear entre duas variáveis quantitativas" Moore (2007 apud FIGUEREDO, 2009).

Considere, uma amostra de tamanho n com valores de $x_1, x_2, \dots, x_n$. Temos a média amostral, que calcula o valor central das variáveis, como sendo

$$\bar{x} = \sum_{i=1}^n \frac{X_i}{n}$$

$$C = \begin{vmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1p} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2p} \\ c_{31} & c_{32} & c_{33} & \cdots & c_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{pj} & s_j^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(n-1)} & \vdots & \vdots & c_{pp} \end{vmatrix}$$

e a variância amostral, que calcula o valor da dispersão entre as variáveis em relação ao valor central, como sendo

Sejam p variáveis em uma amostra de n valores para cada variável, sendo $\bar{x}$ a média amostral e s_j^2 a variância amostral para a j -ésima variável. Assim, temos o cálculo para a covariância amostral entre as variáveis X_j e X_k é dada por

onde x_{ik} é o valor da variável para a i -ésima observação multivariada. Sendo assim,

$$C_{jk} = \sum_{i=1}^n \frac{(x_{ij} - \bar{x}_{Ij})(x_{ik} - \bar{x}_k)}{n-1}$$

temos a matriz de covariância denotada por

A matriz covariância mede quantidade da variação entre as variáveis. Para calcular os coeficientes da matriz correlação, usa-se

$$r_{jk} = \frac{c_{jk}}{(s_j s_k)}$$

sendo, $c_{kj} = c_{jk}$, $r_{kj} = r_{jk}$, $c_{jj} = s_j^2$ e $r_{jj} = 1$. Gerando assim a matriz

$$R = \begin{vmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{vmatrix}$$

Quando as observações da amostra estão padronizadas, ou seja, o desvio padrão igual a 1 e a média zero, a matriz covariância é a matriz correlação.

É considerada uma correlação perfeita se o coeficiente de correlação for 1 ou -1, e que não há correlação linear entre as variáveis se for obtido o valor zero (FIGUEREDO FILHO, 2009). É considerada correlação pequena caso o coeficiente seja a partir 0,10, média a partir 0,30, e grande 0,5 em diante (COHEN, 1988).

1.3.2 Análise com componentes principais e clusters

O método de análise de componentes principais (PCA) é uma técnica multivariada de redução de dimensionalidade. O quantitativo da matriz de dados pode interferir no resultado desejado pois dificulta a visualização de padrões de associações e pode levar a altos níveis de correlação e multicolinearidade pelo fato de ocorrer redundância (LATTIN, 2011). Dessa forma, PCA é um método de redução de dimensionalidade para retirar a redundância nas covariáveis e possibilitar que a função de predição tenha maior poder preditivo (IZBICKI, 2020).

Considere que as variáveis conjunto de dados estejam normalizadas. Nesse método as variáveis são medidas por sua variância e por sua correlação, eliminando variáveis redundantes de modo que se encontre combinações lineares desse conjunto com um quantitativo menor de variáveis e que descreva bem esse conjunto onde os coeficientes dessa combinação linear são os autovetores da matriz covariância desses dados.

A PCA principais um método de multivariado que busca encontrar uma combinação linear onde cada variável X_p , considerando uma amostra com n indivíduos e com p variáveis, está associado a um coeficiente λ_p na expectativa do coeficiente de parte das variáveis serem pequenas a ponto de serem representadas pelas variáveis de maiores λ_p . Gerando a pontuação do componente principal e λ_p os autovalores da matriz covariância com os dados padronizados, ou seja, a matriz correlação dos dados. Também são inseridas restrições de maneira que o PC_1 tem a maior variância possível, os demais componentes principais têm a maior variância possível onde a $var(CP_1) \geq var(CP_2) \cdots \geq var(CP_p)$ e a correlação entre eles é igual a zero (JAMES, 2013; IZBICKI, 2020). Garantindo que os primeiros PCs tenham maior interpretação do conjunto.

A análise de clusters (agrupamento) consiste em separar objetos em conjuntos de acordo com a dissimilaridade e unir eles a partir da similaridade (KAUFMAN; ROUSSEEUW, 1990). Para isto, foi desenvolvido métodos eficazes para atingir este objetivo este trabalho focará em dois métodos não hierárquicos o K-média e o PAM (Partitioning Around Medoids).

As técnicas combinadas oferecem benefícios dado ao tratamento de dados que o PCA oferece fazendo com que a análise de cluster apresente agrupamentos mais eficazes evidenciando melhor as características do conjunto (THIPATHI, 2025). Os resultados da PCA podem ser aplicados em regressão linear, redes neurais artificiais, Análise Discriminante linear, entre outros de acordo com a adequação do conjunto (BRO; SMILDE, 2025).

O objetivo deste trabalho foi desenvolver um estudo baseado em modelagem molecular e análise de componentes principais e de clusters para identificar as propriedades relacionadas à atividade de inibidores da anidrase carbônica de *M. tuberculosis*. O estudo foi baseado em dados experimentais de atividade disponíveis na literatura e dados gerados por modelagem molecular.

Os resultados poderão ser usados para o desenvolvimento de modelos qualitativos que possibilitam prever a atividade de outros compostos, candidatos a inibidores da anidrase carbônica de *M. tuberculosis*.

Para atingir esse objetivo foi necessário primeiro identificar e selecionar as estruturas apropriadas para o estudo da interação com a isoforma 2 da anidrase carbônica de *M. tuberculosis* (MtCA2) descritas nos artigos de referência. Foi usado o programa Spartan'24 com a finalidade de desenhar e otimizar as estruturas dos ligantes da enzima de modo que elas tenham energia mínima. Após isso, foi feita a análise das interações através da docagem molecular com o programa GOLD. Os dados gerados pela docagem molecular foram usados para classificar os ligantes em relação à medida experimental da atividade, a constante de inibição K_i , que quantifica a capacidade de inibição das moléculas sobre a enzima da micobactéria.

Além disso, os dados extraídos dos métodos *in silico* foram analisados de modo a identificar correlações para que sejam identificadas as variáveis independentes que possam determinar a variável dependente do conjunto, K_i .

Por fim, para encontrar as variáveis com maior relevância na interação ligante-receptor foram aplicados os métodos combinados de PCA, com o objetivo de reduzir a dimensionalidade do conjunto de variáveis, e o método de clusterização, a fim de agrupar as moléculas a partir das variáveis dos componentes principais mostrando quais conjuntos de variáveis estão relacionados à K_i .

Revisão Bibliográfica

O boletim epidemiológico do Ministério da Saúde proporcionou informações relevantes para a motivação do trabalho, incluindo dados atuais sobre a tuberculose no Brasil, compromisso do governo, estratégia de combate da doença, entre outros (Brasil, 2024). O boletim da OMS (2023) trouxe dados sobre a tuberculose mundial. Ambos trazem informações e estratégias sobre desafios e metas para combater a doença. Formas de tratamento foram abordadas em nota técnica do Ministério da Saúde (Brasil, 2022), e as interações medicamentosas no organismo em artigos do Jornal Brasileiro de Pneumologia (Arbex et al., 2010a, Arbex, 2010b).

O processo de planejamento e desenvolvimento de fármacos foi abordado no artigo de Guido (2010). Nele, se apresentam estratégias a partir de informações sobre a fisiopatologia das doenças, o conhecimento metabólico e a triagem de alvos moleculares para se planejar novos candidatos a fármacos, como a do SBDD (*structure-based drug design*), que visa descobrir medicamentos a partir da estrutura do alvo.

Uma ferramenta importante aplicada no SBDD é a modelagem molecular, (Sant'Anna, 2009). Entre os métodos da modelagem molecular estão os métodos clássicos, como a mecânica molecular e a dinâmica molecular; métodos quânticos, como os métodos semi-empíricos, os métodos *ab initio* e os métodos de funcional de densidade, além dos métodos comparativos de modelagem de proteínas.

A estratégia do SBDD depende do conhecimento da estrutura 3D da proteína. A estrutura em 3D da MtCA foi determinada por cristalografia de raios X (Covarrubias et al., 2005). O texto descreve particularidades da estrutura como, por exemplo, informações sobre o sítio ativo e ligantes. O artigo de Carta e colaboradores (2019) traz um conjunto de sulfonamidas/sulfamatos testados e com valores experimentais de atividade inibitória sobre a MtCA. Aspartwar e colaboradores (2019) expõe a importância da MtCA para o desenvolvimento de novos candidatos a fármacos antituberculose.

Abdoli e colaboradores (2023) estudaram um conjunto de estruturas derivadas da acetazolamida, testadas e com valores de atividade suficientemente bons de inibição da MtCA, de modo que são candidatas a serem analisadas no processo de pesquisa e desenvolvimento de fármacos. Com a mesma finalidade, Cau e colaboradores (2016) descreveram ácidos fenólicos e ésteres, Masesca e colaboradores (2013b) ditiocarbamatos, Aspartwar e colaboradores (2020) tiocarbamatos e ditiocarbamatos, além de apresentar um estudo de toxicidade, Carta e colaboradores (2009) descreveram um conjunto de sulfonamidas já conhecidas e um conjunto de estrutura derivadas dessas sulfonamidas e, por fim, Maresca e colaboradores (2013a) ácidos carboxílicos.

Os textos consultados disponibilizaram informações valiosas e cruciais para o trabalho, trazendo informações iniciais para as primeiras etapas do processo de pesquisa, descrevendo a inibição da β -CA por esses ligantes. Entretanto, não aplicaram métodos de modelagem molecular a fim de analisar as melhores interações entre eles. No entanto, há uma estrutura de um complexo MtCA-inibidor obtida por Covarrubias e colaboradores (2009).

As CA são enzimas dependentes de Zn^{2+} , que catalisam a reação reversível de conversão do dióxido de carbono e água em bicarbonato no sítio ativo da enzima. Sendo assim, por meio da docagem molecular nesse sítio podemos prever o modo de interação entre os compostos que são inibidores da atividade enzimática. Conceitualmente, um inibidor enzimático é uma substância eficiente em interferir na reação de catálise enzimática de modo que retarda ou reduz o processo ou a especificidade biológica da reação (Marques, 2008).

Após essas informações coletadas, foi feita uma revisão que contribuísse para análise estatística. Azevedo e colaboradores (2022) desenvolveram um modelo de predição de atividade inibitória sobre a subunidade B da DNA girase de micobactérias usando a técnica de regressão linear múltipla. Já Macedeo e colaboradores (2014) fizeram estudos sobre inibidores da enzima diidroorotato desidrogenase de *Plasmodium falciparum* usando análise de agrupamentos hierárquicos nos modelos de QSAR. Bandeira e colaboradores (2025) fizeram um modelo de previsão da atividade inibitória sobre enzimas de *Leishmania donovani* com Regressão Linear Múltipla.

A análise multivariada de dados possibilita a obtenção de informações a partir de um conjunto de dados por meio da coleta e análise de dados, para isto existem

algumas técnicas. Os Autores Richard A. Johnson e Dean W. Wichern (2007) explicam métodos de análise multivariada de dados incluindo componentes principais e clusters abordando suas aplicações e métodos de verificação. Ronh Wehrens (2011) mostra a teoria e aplicação desses métodos aplicados a quimiometria. Rasmus Bro e Age K. Smilde (2014) apresentam a teoria e aplicação de componentes principais detalhadamente, mostrando que é uma técnica de redução de dimensionalidade. Usando como conjunto de dados vinhos tintos os autores abordam a teoria de PCA, escolha de quantidade de PCs, interpretação gráfica, aplicação entre outros. Hervé Abdi e Lynne J. Williams (2010) apresentam a mesma proposta, porém sua abordagem teoria é explicada na forma matricial assim como Kuan Hongyu e colaboradores (2015).

Leonard Kaufman e Peter J. Roussew (1990) apresentam em sua obra a relevância da análise de agrupamento e abordam métodos de particionamento e métodos hierárquico, sendo eles o método PAM, K-Médias, CLARA, AGNES, Funny entre outros. E Shailesh Tripathi e colaboradores 2025 expõe seu estudo sobre segmentação de mercado usando a técnica combinada de PCA e Clusters.

Nesse trabalho, foram utilizadas estruturas extraídas do banco de dados *RCSB Protein Data Bank* (RCSB PDB) que fornece online estruturas cristalográficas 3D em diversos formatos, como o PDB (acesso em www.rcsb.org/). Moléculas com atividade inibitória sobre a CA de *M. tuberculosis* foram obtidas da literatura (CARTA, 2009; ABDOLI, 2023), com o objetivo de estudar a interação enzima-ligante através da docagem molecular e calcular descritores dos complexos.

Inicialmente foi verificado se havia uma correlação qualitativa entre os dados experimentais e somente os calculados na docagem molecular. Numa segunda etapa, foram também incluídos os dados das moléculas isoladas obtidos por cálculos semi-empíricos, que foram usados em uma análise estatística, como descrito no fluxograma abaixo na figura 4.1.

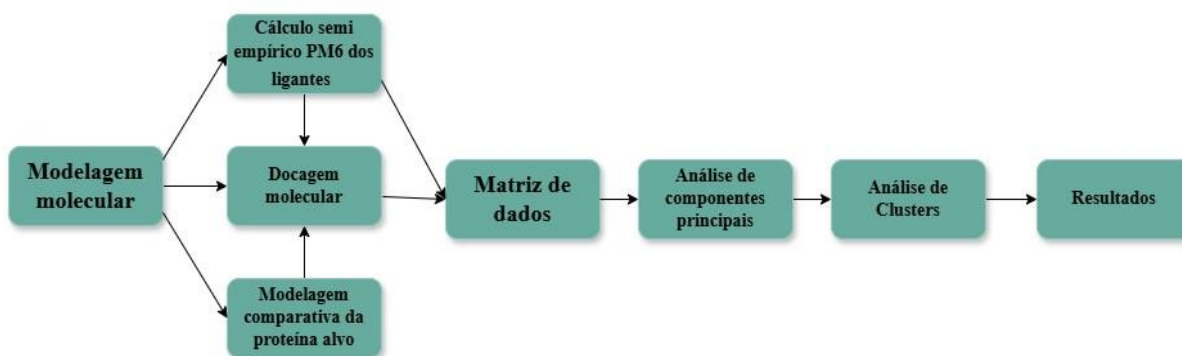


Figura 4.1 – Fluxograma da 2ª etapa do estudo.

4.1 Seleção da anidrase carbônica da *M. tuberculosis*

No estudo de docagem molecular foi usada inicialmente a estrutura da isoforma 2 da MtCA (MtCA2) disponível no banco de dados Protein Data Bank (www.rcsb.org/), com o código 2A5V (disponível em <https://www.rcsb.org/structure/2A5V>), obtida por difração de raios X com resolução de 2,20 Å.

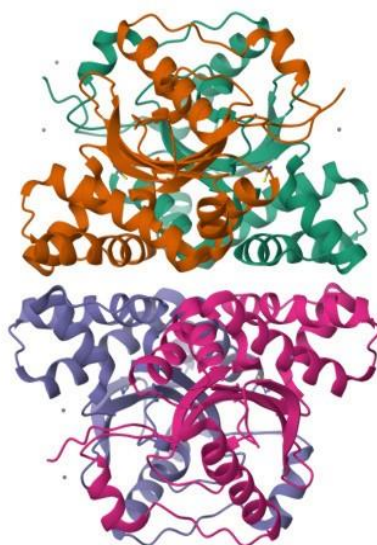


Figura 4.2 – estrutura experimental da CA de *M. tuberculosis* 2A5V (Fonte: RCSB PDB).

A figura 3.3 mostra a comparação da sequência da MtCA com a de uma CA humana (código PDB: 1CA2), mostrando uma identidade de apenas 21,79%, calculado pelo servidor UniProt (acesso em: www.uniprot.org). Essa baixa identidade sugere uma boa chance de ser possível desenvolver inibidores seletivos para a MtCA, reduzindo possíveis efeitos colaterais no uso em humanos.

1CA2_1	-----SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSL	43
2A5V_1	HHHHHHGPNTNPVAAWKALKE-GNER-----FVAGRPHPSQSVDHRAGLAAGQ	48
1CA2_1	KPLSVSYDQATSL-----RIL-----NNGHAFNVEFDDSQ-----	73
2A5V_1	KPTAVIFGCADSRVAAEIIFDQGLGDMFVVRTAGHVIDSAVLGSI EYAVTVLNVP	103
1CA2_1	-----DKAVLKGGPLDGT YRL-----IQFHFWGSLDGQGSEHTV	108
2A5V_1	LIVVLGHDSCGAVNAALAAINDGTLPGGYVRDVVERVAPSVLLGRRDGLSRVDEF	158
1CA2_1	DKKKYAAELHLVHWNTKYGDFGKAVQQPDGLAVLGIFLKVGSAKPGQLQKVVDVLD	163
2A5V_1	EQRHVHETVAIILMA--RSSAISE-RIAGGSLAIVGVITYQLDDGRA-----	200
1CA2_1	SIKTKGKSADFTNFDPRLLPESLDYWTYPGSLTTPPLLECVTWI V LKEP I SVSS	218
2A5V_1	-----VLRDHI GNIG	210
1CA2_1	EQV LKFRKLNFNNGEGEPEELMVDNWRPAQPLKNRQ I KASFK	259
2A5V_1	EEV-----	213

Figura 4.3 – Comparação da AC de *M. tuberculosis* 2A5V e a AC humana 1CA2 (Fonte: RCSB PDB).

No entanto, a estrutura 2A5V não se mostrou adequada para o estudo de docagem molecular, por conter um ligante co-cristalizado muito pequeno no seu sítio ativo, o íon tiocianato, o que deixou a cavidade do sítio muito restrita para o recebimento dos inibidores maiores no estudo de docagem. Isso aconteceu porque na docagem os modos de interação dos ligantes são avaliados somente através de mudanças na conformação (variação de ângulos de torção das ligações químicas) e da posição e orientação da molécula no sítio, enquanto a proteína é mantida rígida.

Para se obter uma estrutura da MtCA2 com uma cavidade do sítio mais volumosa, foi usada uma estratégia empregando a modelagem comparativa por meio do servidor Swiss Model (<https://swissmodel.expasy.org/>). Foi usada a sequência primária da estrutura 2A5V e, como molde, a cadeia A da estrutura de código 1G5C no PDB (resolução de 2,10 Å), que é de uma CA de *Methanothermobacter thermautotrophicus* contendo em seu sítio ativo um ligante co-cristalizado mais volumoso, o ácido 4-(2-hidroxietil)-1-piperaziniletanosulfônico (Hepes), uma substância usada para tamponar o meio de cristalização. Embora a identidade global entre as sequências primárias das duas espécies seja baixa (24,1%), a identidade da região entre os resíduos 1 a 117, que contém os resíduos do sítio ativo (considerando uma distância de até 8,0 Å do íon Zn(II)) é de 30,2%, portanto essa região está no limite para obter um modelo aceitável.

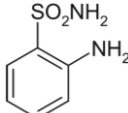
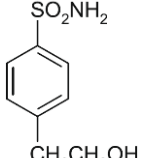
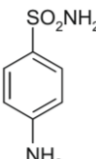
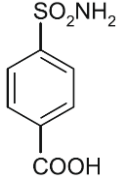
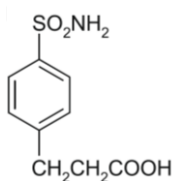
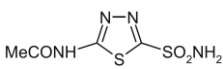
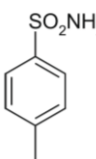
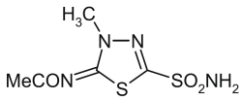
Dessa forma, obtivemos uma estrutura da MtCA2 com o sítio ativo em uma conformação mais aberta, que se mostrou adequada para o estudo de docagem (Fig. 3.3). É importante que o sítio ativo, que é onde acontece a interação enzima-ligante, seja suficientemente aberto para que seja possível uma boa interação com os ligantes.

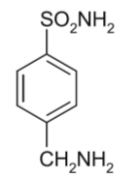
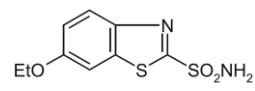
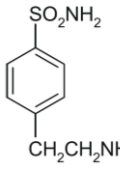
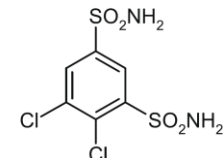
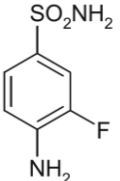
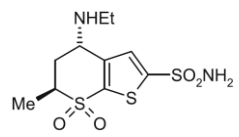
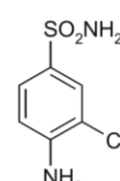
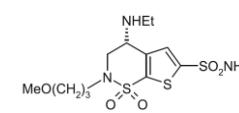
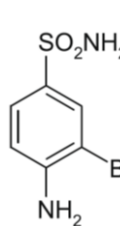
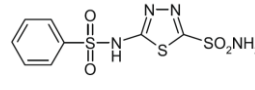
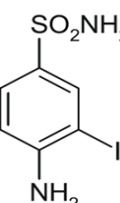
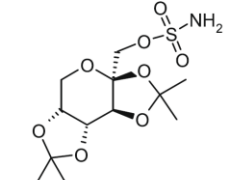
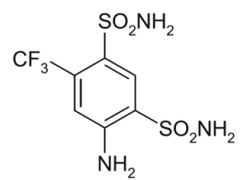
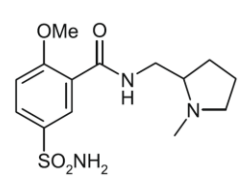
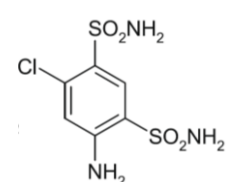
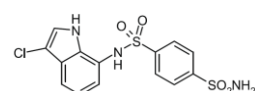
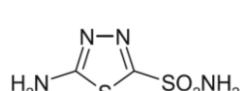
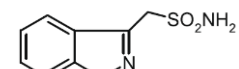
4.2 Seleção e construção dos compostos

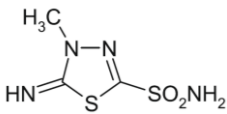
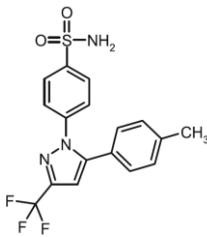
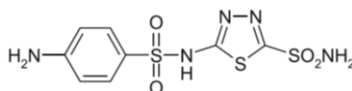
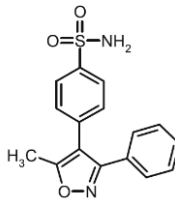
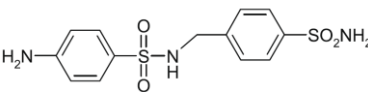
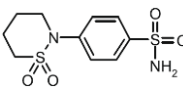
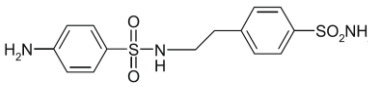
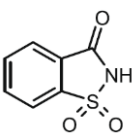
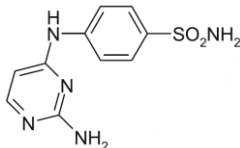
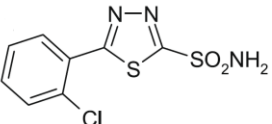
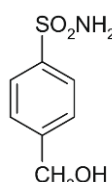
As estruturas das sulfonamidas/sulfamatos foram obtidas da literatura (CARTA et al., 2009; ABDOLI et al., 2023), que continha os dados da inibição *in vitro* mostrando que eles têm atividade inibitória sobre a enzima e, portanto, potencial para combater o patógeno com um mecanismo de ação diversificado em relação aos antibióticos clássicos (tabelas 3.1, 3.2 e 3.3). As atividades

inibitórias estavam expressas como valores de K_i (constante de inibição) relacionados à MtCA2, sendo algumas das moléculas já conhecidas e já usadas como medicamentos de uso clínico, sendo a acetazolamida (AAZ), metazolamida (MZA), etoxizolamida (EZA), diclorofenamida (DCP), dorzolamida (DZA), brinzolamida (BRZ), benzolamida (BZA), topiramato (TPM), sulpirida (SLP), indisulam (IND), zonaisamida (ZNS), celecoxibe (CLX), valdecoxibe (VLX), sultiame (SLT) e sacarina (SAC). Os compostos de **1-22**, derivados mais simples desses compostos (Tabela 3.1), e os compostos **23a-f** e **24a-f** (Tabela 3.3) foram extraídos do artigo de CARTA e colaboradores (2009). Na tabela 3.2 estão as moléculas **3a-3l**, que foram extraídas do artigo de ABDOLI e colaboradores (2023). A numeração em cada composto na tabela representa o número da molécula que será usada na análise de clusters.

Tabela 4.1 – Estruturas químicas dos compostos inibidores da CA β e constante de inibição MtCA2 (CARTA, 2009).

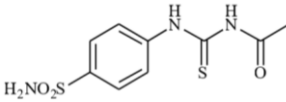
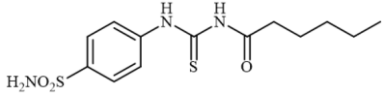
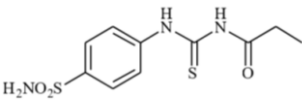
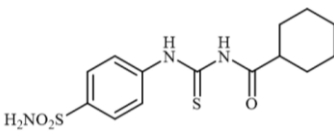
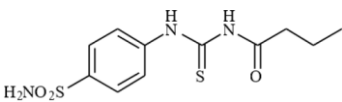
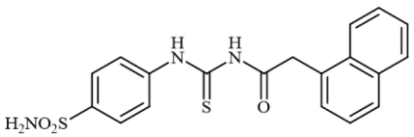
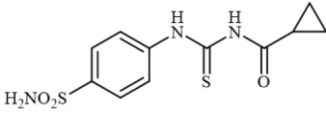
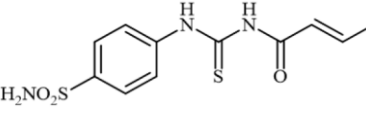
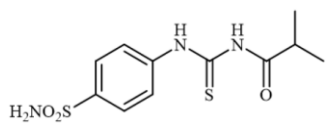
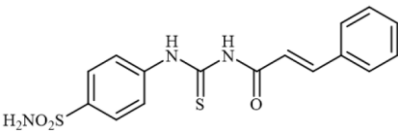
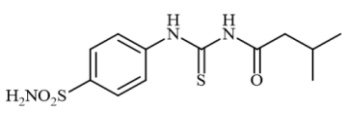
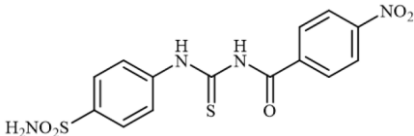
Composto	Estrutura	K_i (μ M)	Composto	Estrutura	K_i (μ M)
1		33,7	21		34,5
2		29,6	22		39,2
3		28,4	23 (AAZ)		0,009
4		38,9	24 (MZA)		0,66

5		30,7	25 (EZA)		0,02 7
6		29,1	26 (DCP)		2,01
7		28,1	27 (DZA)		0,09 9
8		27,7	28 (BRZ)		0,12
9		31,6	29 (BZA)		0,46 7
10		32,4	30 (TPM)		0,47 4
11		29,6	31 (SPL)		0,26 6
12		32,5	32 (IND)		0,71 7
13		2,09	33 (ZNS)		0,87 6

14		2,38	34 (CLX)		0,71 3
15		0,978	35 (VLX)		0,68 2
16		3,21	36 (SLT)		0,66 4
17		2,29	37 (SAC)		0,79 2
18		2,63			
19		45,2			
20		38,3			

* As siglas entre parênteses se referem à nomenclatura nos artigos de origem: acetazolamida (AAZ), metazolamida (MZA), etozizolamida (EZA), diclorofenamida (DCP), dorzolamida (DZA), brinzolamida (BRZ), benzolamida (BZA), topiramato (TPM), sulpirida (SLP), indisulam (IND), zonaisamida (ZNS), celecoxibe (CLX), valdecoxibe (VLX), sultiame (SLT) e sacarina (SAC).

Tabela 4.2 – Estruturas químicas dos compostos inibidores da CA β e constante de inibição em MtCA2, extraídas de (CARTA, 2009)

C	Molécula	K_i	C	Molécula	K_i
38 (3a)		0,792	44 (3g)		0,0211
39 (3b)		0,0044	45 (3h)		0,0034
40 (3c)		0,0054	46 (3i)		0,0084
41 (3d)		0,0076	47 (3j)		0,0046
42 (3e)		0,0042	48 (3k)		0,0092
43 (3f)		0,0571	49 (3l)		0,0081

Entre os compostos há duas séries, **23** e **24** na numeração do artigo original de Abdoli et al., 2023, cuja estrutura geral é apresentada na figura 3.4, com os respectivos grupos R descritos na Tabela 3.3.

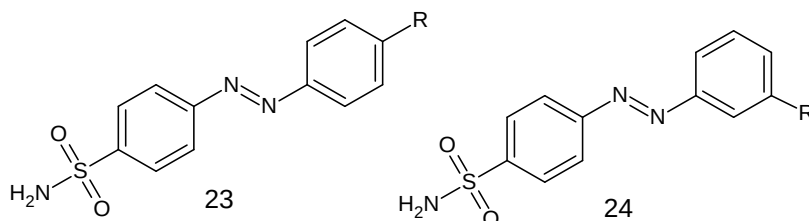


Figura 4.4. Séries de moléculas 23 e 24, conforme Abdoli et al., 2023.

Tabela 4.3. Substituintes dos inibidores da MtCA2 e respectivas constantes de inibição (séries 23 e 24 de Abdoli et al., 2023; números entre parênteses são os adotados neste trabalho).

Composto	R	$K_i(\mu\text{M})$	Composto	R	$K_i(\mu\text{M})$
(50) 23a	OH	0,678	(56) 24a	OH	6,48
(51) 23b	NH ₂	0,955	(57) 24b	NH ₂	1,98
(52) 23c	NHMe	0,346	(58) 24d	NMe ₂	2,13
(53) 23d	NMe ₂	5,48	(59) 24e	NHCH ₂ SO ₃ Na	6,55
(54) 23e	NHCH ₂ SO ₃ Na	0,059	(60) 24f	N(Me)CH ₂ SO ₃ Na	6,9
(55) 23f	N(Me)CH ₂ SO ₃ Na	0,045			

Na primeira etapa, as estruturas 3D dos compostos foram preparadas com o programa comercial (código fechado) Spartan'24 (Wavefunction, Inc.), usando o método quântico semi-empírico PM6 (*Parametric Method 6*) (STEWART, 2007). Esse cálculo faz a minimização da energia para gerar geometrias moleculares otimizadas, além de produzir diferentes descritores de cada molécula como a energia, volume, polaridade, entre outros, que estão enumerados na tabela 3.4, a seguir.

4. 2.1 Docagem molecular

Para a análise da interação entre a estrutura da MtCA2 com os compostos, foi utilizado o método da docagem molecular através do programa comercial GOLD 2023.2.0 (CCDC) disponível gratuitamente para universidades através do financiamento da CAPES, explorando-se suas funções de pontuação. Essa estratégia é uma das usadas no chamado planejamento de fármacos baseado na estrutura do alvo (SBDD, do inglês *Structure-Based drug design*) que é quando se conhece a estrutura de complexos do tipo receptor-ligante, criando a possibilidade de gerar informações para o planejamento novas entidades químicas (NCE, do inglês *New Chemical Entities*).

O programa GOLD usa um algoritmo genético, que permite a determinação dos modos de interação (ou poses) entre a molécula do ligante e o receptor, que no caso

é o sítio ativo da MtCA2. Cada pose recebe uma pontuação de ajuste, dependendo da função escolhida.

Inicialmente, o que se buscou encontrar foi um modelo que permitisse uma correlação entre os dados de modelagem (pontuação de ajuste) e os dados experimentais de K_i . Quanto menor for o valor de K_i , maior será a capacidade de inibição da atividade enzimática. Uma boa interação enzima-ligante libera muita energia, que é, portanto, um valor negativo de energia. A pontuação do GOLD é o negativo da energia de interação entre a enzima e o ligante. Assim, quanto mais positivo for o valor da pontuação, melhor é a interação.

4.3 Análise exploratória de dados multivariados

A análise exploratória pode ser definida como a arte de olhar para um ou mais conjuntos de dados em um esforço para entender a estrutura subjacente dos dados contidos ali (PEARSON, 2011). Sendo assim, através dela podemos entender os dados, identificar padrões, ter orientação de tomadas de decisões, melhorar e verificar a precisão dos dados (PEARSON, 2011). Nesta etapa, serão descritos os passos da análise dos conjuntos de dados multivariados, as técnicas de Análise de Componentes Principais (ACP) e Análise de Clusters.

4.3.1 Conjunto de dados

O conjunto de dados é importante para fazer a análise estatística e assim entender o conjunto, identificar se existem relações ou associações entre os conjuntos de dados de modo que consigamos verificar se esses dados permitirão prever características de outro conjunto (SILVESTRE, 2007). Assim, obtivemos as matrizes de dados através de dados coletados da literatura e calculados em programas de modelagem molecular, essas matrizes foram compostas por 60 objetos que são os ligantes e as variáveis de cada matriz tem um valor específico de acordo com a função usada no programa GOLD.2023.3 (CCDC).

A análise estatística foi feita com o programa R Studio 2025.09.02 (Posit) versão de código aberto; esse programa é ambiente de desenvolvimento integrado (IDE) que utiliza a linguagem R de programação que é amplamente utilizada para estatística, ciência de dados e análise multivariada e visualização gráfica.

O conjunto de variáveis contém como variável dependente a constante de inibição (K_i) e, como variáveis independentes, os dados calculados no programa Spartan'24 (Tabela 3.4), que quantificam propriedades físico-químicas dos ligantes, e no programa GOLD.2023.3 (CCDC) (Tabela 3.5), que quantificam a interação proteína-ligante, incluindo a pontuação e os termos de cada função de pontuação.

Tabela 4.4 – Relação dos descritores moleculares obtidos com o método PM6 (programa Spartan'24).

Variável	Definição	Variável	Definição
Número de tautômeros		Min. ElPlot	Potencial eletrostático mínimo
Energia		Max. ElPlot	Potencial eletrostático máximo
Energia de solvatação		Acc. P-area	Área polar acessível ao solvente
HOMO	Energia do HOMO	Polarizabilidade	
LUMO	Energia do LUMO	HBD	Número de doadores de ligação de H
Momento de dipolo		HBA	Número de aceitadores de ligação de H
Area	Área superficial	Ovalidade	Medida de desvio da forma esférica
Volume		Acc. Area	Área acessível ao solvente
PSA	Área superficial polar		

Tabela 4.5 – Relação dos descritores moleculares obtidos com o programa GOLD 2023.2.0

Variáveis	Definição	Variáveis	Definição
GoldScore		ChemPLP	
GoldScore	Maior pontuação da triplicata das docagens feitas com a função	PLP.Fitness	Maior pontuação da triplicata das docagens.
Hbond.ext	Energia de ligação de hidrogênio proteína-ligante	ChemScore.HBond	Contribuição de ligação de hidrogênio proteína-ligante (ChemScore).

VDW.Ext	Energia de Van der Waals proteína-ligante	ChemScore.Metal	Contribuição de ligação com metais (ChemScore).
torsion.int	Energia de deformação torcional do ligante.	PLP.ligand.clash	Contribuição de colisão proteína-ligante (PLP).
VDW.int	Energia de van der Waals do ligante.	PLP.ligand.torsion	Penalidade de tensão torsional interna do ligante (PLP)
DG	Energia livre.	PLP.Part.buried	Contribuição de interações oclusas (PLP).
HBond	Contribuição de ligação de hidrogênio proteína-ligante.	PLP.Part.Hbond	Contribuição de ligações de hidrogênio (PLP).
Lipo	Contribuição lipofílica proteína-ligante .	PLP.Part.metal	Contribuição de ligação com metais (PLP).
Metal	Contribuição de ligação com metais.	PLP.Part.nonpolar	Contribuição de interação apolar (PLP).
		PLP.Part.repulsive	Contribuição de interação repulsiva (PLP).
		PLP.Protein.Rotated.Torsion	Contribuição de torção de ângulos da proteína (PLP).
		ChemScore.CHOScore	Contribuição de ligações de hidrogênio fracas CH---O
		ChemScore.Internal.Correction	Contribuição de colisão interna e de torção do ligante

Chemscore		ASP	
Chemscore	Maior pontuação da triplicata das docagens feitas com a função Chemscore.	ASP Fitness	Maior pontuação da triplicata das docagens feitas com a função ASP.
DEClash	Penalidade de colisão proteína-ligante.	ASP	Potencial estatístico calculado mais o termo de colisão ChemScore e o termo de energia interna.
DEInternal	Penalidade de deformação de torção do ligante.	DEClash	Penalidade de colisão proteína-ligante para o valor ASP.

DG	Variação de energia livre.	de DEInternal	Contribuição de ligação de hidrogênio intramolecular do ligante para o valor ASP
HBond	ligação de hidrogênio proteína-ligante.	de internal.correction	Deslocamento da energia interna do ligante.
Lipo	Contribuição lipofílica proteína-ligante .	map	Potencial estatístico total calculado como uma soma de todas as combinações de átomos da proteína e do ligante.
Metal	Ligação de metais.	RotatedTorsions	Termo de congelamento de ligações rotativas para o valor ASP
Part.repulsive	Interação repulsão.	de	

Posto isso observa-se que o quantitativo da matriz de dados pode interferir no resultado desejado pois dificulta a visualização de padrões de associações e pode levar a altos níveis de correlação e multicolinearidade pelo fato de redundância substância (LATTIN, 2011). Dessa forma é necessário foi recorrer ao método de redução de dimensionalidade para retirar a redundância nas covariáveis e possibilitar que a função de predição tenha maior poder preditivo (IZBICKI, 2020).

Por essa razão, a primeira seleção de variáveis foi a baseada na matriz correlação cuja é uma matriz que mostra a correlação das variáveis duas a duas. É considerada uma correlação perfeita se o coeficiente de correlação for 1 ou -1, e que não há correlação linear entre as variáveis se for obtido o valor zero (FIGUEREDO FILHO, 2009). E o critério é considerado correlação pequena acaso o coeficiente seja a partir 0,10, seja média a partir 0,30, e grande 0,5 em diante (COHEN, 1988). A partir disso obtivemos na matriz correlação algumas variáveis que obtiveram correlação muito baixa com as outras variáveis, então foram eliminadas variáveis com nenhuma correlação acima de 0,5 ou abaixo de -0,5. Podemos ver as correlações nos capítulos 4.

A matriz correlação assim como as próximas etapas que contém métodos estatísticos foram calculados por intermédio do RStudio 4.4.1 (2024) cujo é um programa que faz cálculos estatísticos, constrói gráficos etc. cujo é um software de

código aberto (ALLARIE, 2012). Além disso, ele se mostra uma ferramenta eficaz para análise de dados.

A eliminação das variáveis pouco correlacionadas foi útil para o método de APC, pois cada variável influencia no resultado. Os componentes principais nos proporcionam combinações lineares do conjunto onde se tem a maior variabilidade entre os dados, menor perda de informação e sem a multicolinearidade entre os dados (HONGYU, et al, 2015; ABDI et al, 2010; IZBICK, 2020).

4.4 Análise de componentes principais

A Análise de Componentes Principais (*Principal Component Analysis*) foi técnica de redução de dimensionalidade escolhida. Pois, mesmo depois da matriz correlação ainda obtivemos um conjunto de dados extenso, desse modo foi possível reduzir a dimensionalidade sem incluir a variável dependente (k_i), considerando que o objetivo é determinar K_i a partir das demais variáveis (WEHRNS, 2011).

Tendo em vista que temos um conjunto de dados bem variados com escalas diferentes os dados foram padronizados, de modo que o desvio é igual a 1 e a média igual a zero, trazendo as variáveis na mesma escala para evitar que as variáveis de alta variância não afete o modelo (JAMES, 2013).

Os dados padronizados substituem cada elemento da coluna x_i por

$$\bar{x}_i = \frac{(x_i - \mu_i)}{\delta}$$

onde, x_i é o valor da variável, μ_i é a média das variáveis x_i e δ é o desvio padrão das variáveis x_i e $\bar{x}_i$ a variável padronizada (FIGUEREDO FILHO, 2009).

As variáveis $x_1, x_2, x_3, \dots, x_p$ de n objetos, que formam a matriz de dados $X_{n \times p}$ e $X_n = [X_1, X_2, X_3, \dots, X_p]$ são variáveis na forma vetorial. Após a padronização desses dados, ou seja, o desvio igual a 1 e média igual a zero, as variáveis ficam na mesma escala para evitar que as variáveis de alta variância afetem o modelo. Temos o componente principal (CP) dada pela combinação linear:

$$PC_1 = \lambda_1 \cdot X_1 + \lambda_2 \cdot X_2 + \lambda_3 \cdot X_3 + \dots + \lambda_p \cdot X_p.$$

onde λ_i são os coeficientes (também chamados pesos ou *loadings*) associados a cada variável. A primeira componente (CP₁) é a combinação linear das variáveis

originais com a variância maximizada. Os seguintes CP possuem mesma restrição, ou seja, $var(CP_1) \geq var(CP_2) \geq \dots \geq var(CP_p)$, todas as CPs possuem a correlação zero entre elas e a soma dos seus quadrados de λ_i é igual a um (JAMES, 2013; IZBICKI, 2020).

As duas primeiras restrições garantem que cada componente seja diferente e que tenha maiores combinações que expressem as características do conjunto. A variância entre os dados é importante pois assim o CP consegue representar o conjunto com menor quantidade de variáveis. Os coeficientes são obtidos através dos autovetores da matriz covariância, que é a matriz correlação com os dados padronizados, esses autovalores. E os autovalores dessa matriz são as variâncias dos Componentes principais (MANLY; ALBERTO, 2017).

A quantidade de componentes principais de PC para cada matriz de dados os métodos empregados foram variância de Kaiser e o *Scree Plot* (Gráfico cotovelo). Na variância de Kaiser são escolhidas as componentes com autovalores maiores que 1 e com a variância total maior ou igual 80% (WEHRNS, 2011; HONGYU, 2015 apud KAISER, 1958; FRAGA, 2015).

O gráfico cotovelo que relaciona os autovalores com os componentes principais correspondentes, visualmente, ao identificar um “cotovelo” no gráfico se encontra o número para decidir o número de PCs (WEHRNS, 2011). Esse cotovelo representa o ponto de inflexão no gráfico, após esse ponto a variância acumulada dos CPs variam pouco, mostrando a quantidade de CPs relevante a serem analisados. O critério de escolha de variáveis de cada componente foi feito pelo gráfico das contribuições das variáveis, que apresenta o percentual de contribuição de cada variável e a média das contribuições foram selecionadas todas as variáveis com a contribuição acima da média. As contribuições são calculadas através de

$$Contribuição_{PC} = \frac{Cos^2_{varPC}}{\sum_{var} Cos^2_{PC}} * 100.$$

Onde $\sum_{var} Cos^2_{PC}$ é o somatório do cosseno ao quadrado das variáveis no componente principal (MELO; PETRY *et al*, 2025);

Encontrados os conjuntos de PCseles foram submetidos à análise de cluster a fim de verificar qual grupo de variáveis estaria relacionada ao K_i e assim, encontrar as combinações lineares que estão relacionadas com K_i .

4.5 Análise de *cluster*

A técnica de análise de *cluster* foi empregada para agrupar os dados dois métodos foram implementados o método de k-médias e o método PAM (Partitioning Around Medoids) com o objetivo de expor quais componentes principais de cada função obtinham os dados que conseguem estar relacionada com a variável independente. A escolha de um método de agrupamento depende dos dados trabalhados e é recomendado a execução de mais de uma técnica e analisar qual se enquadra melhor a partir entendimento dos dados. O agrupamento pode ser usada como uma instrumento exploratório conseguindo inferir ou confirmar uma hipótese (KAUFMAN; ROUSSEEUW, 2009).

O **método k-médias** é um método de separação não hierárquico baseado em minimizar a distância euclidiana das variáveis dos centros C_i . Considerando $X = \{x_1, x_2, \dots, x_n\}$ o conjunto das variáveis e d a distância nesse conjunto formam-se centroides (centros) $C_1, C_2, \dots, C_k$ de modo que se associa as variáveis ao centro mais próximo formando k grupos onde a intercessão entre eles é vazia e a união deles formam o conjunto X . Os centroides são atualizados até a soma dos quadrados da distância euclidiana entre os objetos e os centros sejam minimizados (PALMA, 2018). De maneira que

$$\sum_{i,j \in C_k} d^2(C_i, x_j)$$

seja minimizada, $|C_k|$ é a cardinalidade de C_k e x_i e x_j são covariáveis. O centroide é calculado pela média das variáveis dentro de X_i (PALMA, 2018; IZBICKI *et al*, 2020).

O método PAM é um método não hierárquico, também conhecido como k-medoides. Ao invés de ser centrado na média essa técnica é centrada em objetos representativos (medoide), para k *clusters* seleciona-se k objetos que estão centralmente localizados no grupo e particionamento é se estabelece com as variáveis tendo a distância média ou a dissimilaridade é minimizada ao redor do medoide (KAUFMAN *et al*, 1990).

A quantidade de cluster foi definida pelo gráfico Silhueta que evidencia a quantidade de grupos ideal. Neste gráfico se examina a largura da silhueta alta, se ela está próxima de um elas estão muito bem agrupadas (WEHRENS,2011)

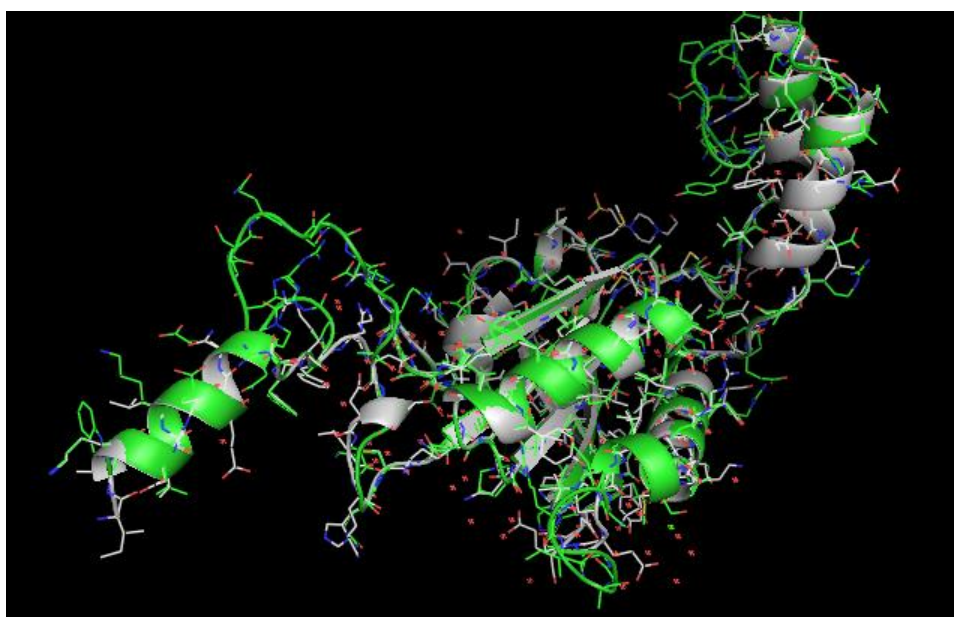
Os agrupamentos de cada CP foram examinados comparando as moléculas e seus coeficientes de inibição dentro de cada cluster e foram comparadas as duas técnicas. Foi considerado como bom resultado os CPs com acerto próximo a 80%, que mostraram uma separação melhor. O cálculo do erro foi:

$$erro = \frac{\text{quantidade de moléculas no grupo "errado"}}{\text{Quantidade total das moléculas.}}$$

Com isso, temos que os CPs geram um bom tratamento de dados, reduz ruídos e sintetiza as informações e os clusters proporcionam a separação de grupos assim a técnica combinada melhora a estimativa do agrupamento, aumentando a eficácia na análise dos dados (TRIPATHI; 2025).

5.1 Análise qualitativa da interação ligante-enzima

O modelo da MtCA obtido por modelagem comparativa apresentou boa sobreposição com o molde usado, a CA de *Methanothermobacter thermautotrophicus* contendo o ligante cocrystalizado ácido 4-(2-hidroxietil)-1-piperaziniletanossulfônico, com um RMSD baixo (2,31 Å), mostrando que, mesmo com a baixa identidade geral entre as sequências, foi possível obter um modelo com boa semelhança com o molde (figura 4.1). Esse modelo, quando sobreposto com a estrutura cristalográfica original da MtCA2 (2A5V), apresentou diferenças conformacionais consideráveis (RMSD = 7,18 Å), principalmente com aumento do volume da cavidade do sítio ativo, como desejado para o estudo de docagem molecular dos inibidores selecionados (figura 4.1).



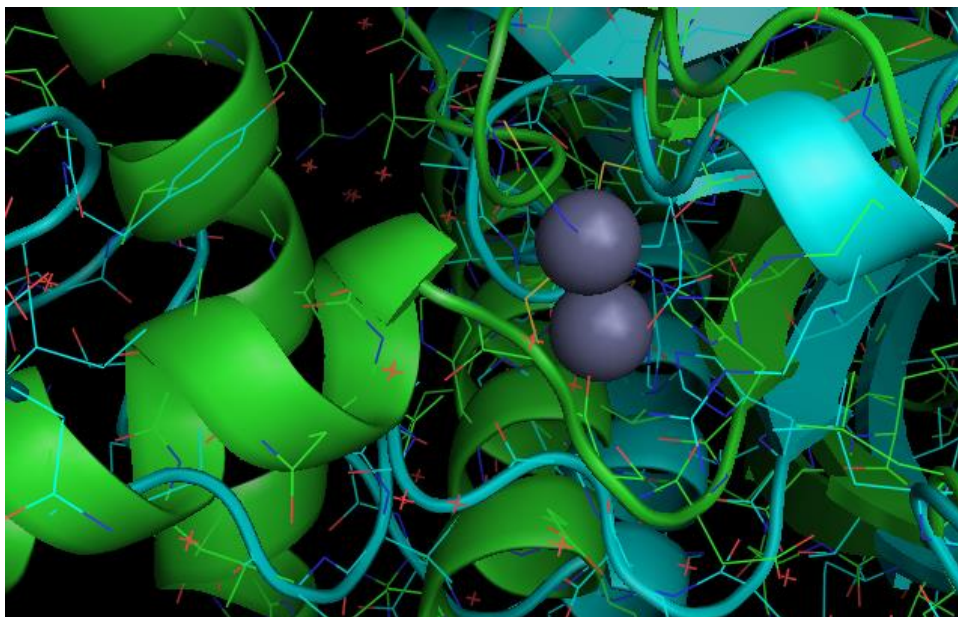


Figura 5.1. Sobreposição do modelo comparativo da CA de *M. tuberculosis* (verde) com o molde, a estrutura da CA de *Methanothermobacter thermautotrophicus* (código no PDB: 1G5C, em branco) (no alto). Sobreposição do sítio catalítico da estrutura cristalográfica de MtCA (código no PDB: 2A5V, em verde) com o modelo da MtCA (em ciano) gerado por modelagem comparativa a partir do molde 1G5C; íons Zn(II) localizados no sítio catalítico representados como esferas de cor cinza (embaixo). Figuras geradas com o programa PyMol v 0.99 (DeLano Scientific LLC).

Com esse modelo foi realizada a docagem molecular. Como exemplo, na figura 5.2 é mostrada a sobreposição de todas as poses encontradas para o composto **13** com a função Goldscore. No programa do GOLD as cores usadas para representar os átomos são o vermelho para o oxigênio (O), o amarelo para o enxofre (S), o cinza para o carbono (C) e o azul para o nitrogênio (N). A esfera mostra o cofator do sítio ativo, que no caso é o íon Zn^{2+} .

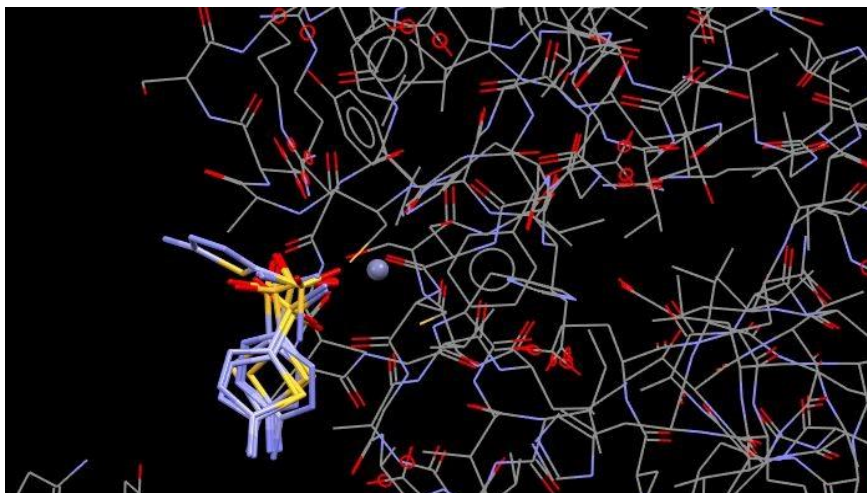


Figura 5.2 – Sobreposição das poses de docagem no programa GOLD do ligante **13** com a função GoldScore. Átomos de H omitidos para maior clareza.

Na figura 5.3 pode se observar a interação de melhor pontuação obtida para o ligante **13**. Pode-se observar nessa pose que a função GoldScore indica a proximidade entre o N do grupo sulfonamida e o íon Zn^{2+} , como esperado pela capacidade de coordenação desse grupo com o íon metálico.

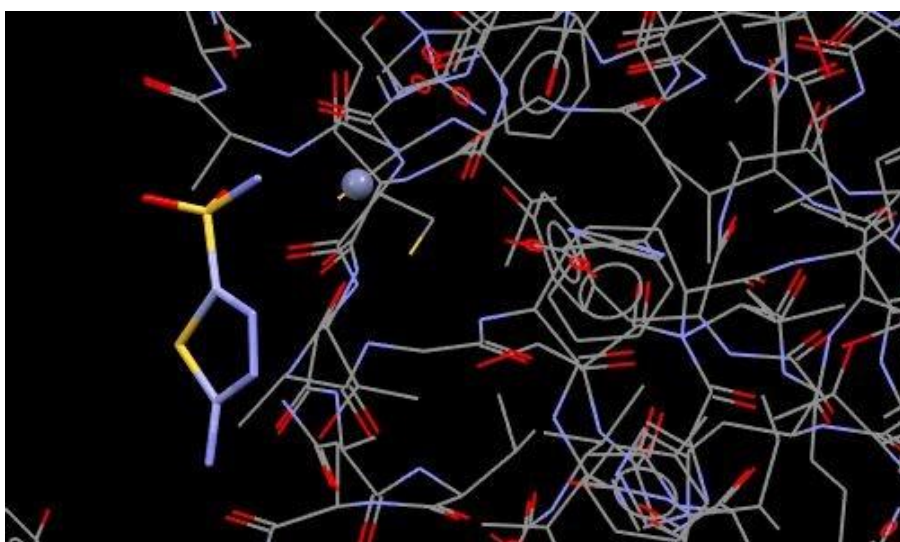


Figura 5.3 – Melhor solução de docagem no programa GOLD do ligante **13**, com a função GoldScore. Átomos de H omitidos para maior clareza.

Inicialmente, foi avaliado se a pontuação da docagem estava relacionada com os dados de K_i . Nas tabelas a seguir são apresentados não somente esses dados,

mas também as distâncias entre o N da sulfonamida e o íon Zn^{2+} em angstroms (Å), que é uma unidade de medida igual a 10^{-10} m. Considerando uma distância ≤ 3 Å uma boa interação, a maior parte das moléculas estão interagindo bem.

A constante de inibição é um dado experimental que classifica o quanto as moléculas são ativas na inibição da enzima, sendo quanto menor for, mais ativo é o ligante. Como todos os compostos são inibidores optou-se por escolher um valor arbitrário para classificar os compostos em inibidores fortes e fracos. Inicialmente o valor de 10 μM foi avaliado, mas observou-se que o valor de $K_i \leq 7$ μM para classificar um inibidor como forte permitia uma melhor separação na série de compostos. Assim, observamos na tabela 4.1 que o valor $\geq 49,7$ da maior pontuação na análise em triplicata com a função GoldScore está associado a 86,7% dos compostos que têm $K_i \leq 7$ μM .

Em seguida analisamos as pontuações resultantes da função ChemPLP. Assim, observamos na tabela 5.2 que para essa função o valor ≥ 47 da maior pontuação na análise em triplicata está associado a 76,8% dos compostos que têm $K_i \leq 7$ μM .

Tabela 5.1 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função GoldScore.

Comp.	K_i (μM)	Pont.1	Dist(Å)	Pont.2	Dist (Å)	Pont.3	Dist (Å)	Pont. Máx.
1	33,7	40,3001	2,434	40,6569	2,450	40,3444	2,428	40,657
2	29,6	40,5638	2,494	40,6043	2,433	40,2943	2,394	40,604
3	28,4	49,8697	2,171	50,4906	2,246	50,3474	2,465	50,491
4	38,9	41,1711	2,453	41,1896	2,391	40,3634	2,276	41,190
5	30,7	45,0733	2,266	44,4633	8,775	45,2313	8,383	45,231
6	29,1	46,2052	2,400	46,5218	2,297	46,4832	2,423	46,522
7	28,9	40,7432	2,378	41,2267	2,482	41,0023	2,460	41,227
8	27,7	42,7786	2,417	42,7777	2,364	43,1716	2,436	43,172
9	31,6	43,5742	2,426	43,4084	2,299	43,6683	2,419	43,668
10	32,4	45,6632	2,546	45,9065	2,558	45,9053	2,499	45,907
11	29,6	49,1879	2,364	48,4298	2,114	49,0522	2,157	49,188
12	32,5	49,0586	2,370	48,9624	2,366	47,7105	2,478	49,059
13	2,09	38,2114	2,444	39,3136	2,448	37,4006	2,387	39,314
14	2,38	39,9777	2,294	49,1538	7,165	48,3868	6,928	49,154
15	0,978	54,8403	4,146	56,0327	2,413	54,0880	4,123	56,033

16	3,21	57,5587	3,836	59,0474	2,343	59,0827	4,201	59,083
17	2,29	64,1401	4,191	65,8772	2,467	65,0827	2,538	65,877
18	2,63	54,6386	2,407	54,6247	2,287	55,0989	2,427	55,099
19	45,2	49,0545	4,085	49,0968	4,119	50,0363	3,765	50,036
20	38,3	43,1290	7,822	42,9350	2,077	41,5206	2,343	43,129
21	34,5	44,4938	2,345	45,1522	2,537	45,2573	2,148	45,257
22	39,2	41,6191	2,428	42,1410	2,394	42,6090	7,674	42,609
23	0,009	42,3937	4,036	42,0077	3,889	42,0637	4,078	42,394
24	0,66	49,8127	2,413	49,8823	2,439	49,4386	2,413	49,882
25	0,027	51,7512	2,474	51,7358	2,412	51,1890	2,508	51,751
26	2,01	49,7632	2,382	49,5901	2,197	49,8757	2,324	49,876
27	0,099	56,8320	2,505	56,7955	2,536	56,4055	2,500	56,832
28	0,127	55,3141	8,470	55,3361	9,008	54,1959	8,371	55,336
29	0,467	55,3048	2,493	54,5872	2,459	53,0458	4,194	55,305
30	0,474	48,3429	2,313	48,6722	2,281	48,8591	2,315	48,859
31	0,266	53,4221	3,489	56,0510	1,959	55,2899	2,063	56,051
32	0,717	60,5799	7,849	60,4006	7,819	60,6900	2,383	60,690
33	0,876	44,2963	2,471	43,3420	2,465	44,3066	2,476	44,307
34	0,713	55,9210	4,497	51,6039	4,482	51,9898	4,534	55,921
35	0,682	57,0778	2,525	56,6592	2,416	57,9715	2,557	57,971
36	0,664	54,3966	2,511	51,9841	4,033	52,0928	3,784	54,397
37	0,792	37,8619	3,139	36,1628	4,336	36,1781	4,285	37,862
38	0,0044	57,0742	1,940S*	54,5515	2,317N	59,3400	2,354S	59,3400
39	0,0054	60,0127	2,112S	56,8536	2,573N	56,3488	2,472S	60,0127
40	0,0076	60,6476	1,952S	60,9019	2,034S	58,5293	2,424N	60,9019
41	0,0042	61,3308	1,880S	59,2009	1,924	59,8036	2,132S	61,3308
42	0,0571	58,5716	2,397S	61,4454	1,880S	60,5756	2,030S	61,4454
43	0,0211	60,8345	1,937S	61,1588	1,882S	60,4216	1,790S	61,1588
44	0,0034	60,5086	1,906S	60,9969	2,122S	59,4917	2,464N	60,9969
45	0,0084	58,3205	1,939S	58,3916	2,062S	61,7952	1,939S	61,7952
46	0,0046	69,2301	2,445S	66,5366	2,292S	65,8922	2,225S	69,2301
47	0,0092	59,6870	2,014S	58,6387	2,433N	59,4234	2,639N	59,6870
48	0,0081	66,7683	1,942S	66,1499	2,193S	63,9368	3,747N	66,7683
49	0,0048	68,9389	2,307S	70,2748	2,250S	70,7244	2,269S	70,7244
50	0,678	49,9679	3,822	50,8903	2,549	50,1192	3,740	50,8903
51	0,955	51,1579	2,382	47,9104	4,175	51,5248	2,410	51,5248
52	0,346	48,6293	4,392	48,9783	4,49	51,9099	2,414	51,9099
53	5,48	54,0303	2,367	50,4856	4,427	42,5821	6,349	54,0303
54	0,059	52,2573	11,817	49,2915	12,406	52,8136	12,283	52,8136
55	0,045	56,3746	12,21	57,2474	12,324	56,8366	12,418	57,2474
56	6,48	49,1939	2,312	50,1526	2,133	48,4149	4,255	50,1526
57	1,98	52,631	2,337	52,7787	2,252	49,4022	4,468	52,7787

58	2,13	52,7187	4,141	52,5348	4,059	52,5628	4,142	52,7187
59	6,55	62,1926	11,331	55,4747	4,112	54,7323	4,17	62,1926
60	6,9	64,7031	7,404	52,9392	8,668	61,0899	2,516	64,7031

*A letra N e S na distância de alguns elementos sinaliza que o átomo de menor distância da molécula com o Zn^{2+} sítio ativo foi no nitrogênio (N) ou no enxofre (S).

Em seguida analisamos as pontuações resultantes da função ChemPLP, Assim, observamos na tabela 4.2 que para essa função o valor ≥ 47 da maior pontuação na análise em triplicata está associado a 76,8% dos compostos que têm $K_i \leq 7\mu M$.

Tabela 5.2 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ChemPLP.

Comp.	K_i (μM)	Pont.1	Dist(Å)	Pont.2	Dist (Å)	Pont.3	Dist (Å)	Pont. Máx.
1	33,7	38,8314	2,575	38,8559	2,528	39,2271	2,508	39,2271
2	29,6	38,2461	2,608	38,1661	2,511	37,442	7,812	38,2461
3	28,4	55,5766	9,157	55,5075	9,055	53,1686	9,466	55,5766
4	38,9	40,4483	2,576	41,0238	2,509	40,3974	2,608	41,0238
5	30,7	44,6014	2,488	42,0056	2,564	44,1272	2,436	44,6014
6	29,1	47,4193	2,475	47,5346	3,555	47,892	3,578	47,8920
7	28,9	41,1033	2,607	41,3570	2,575	42,5593	2,467	42,5593
8	27,7	40,8767	2,576	41,3778	2,606	44,3047	2,463	44,3047
9	31,6	41,8562	2,395	42,8759	2,339	40,4802	2,576	42,8759
10	32,4	40,6863	2,64	41,2345	2,606	40,9867	2,486	41,2345
11	29,6	47,8306	2,455	54,3304	2,608	49,7161	2,543	54,3304
12	32,5	50,5308	2,283	45,2816	2,570	43,7725	2,500	50,5308
13	2,09	31,1859	4,165	31,2884	4,161	31,1202	4,183	31,2884
14	2,38	40,3549	2,542	39,6333	7,292	39,1481	7,331	40,3549
15	0,978	46,6292	8,158	44,9945	11,448	46,5566	8,174	46,6292
16	3,21	67,4229	9,715	59,7084	7,505	58,0164	7,663	67,4229
17	2,29	61,3504	8,018	62,3862	7,591	59,2126	7,820	62,3862
18	2,63	57,1152	2,538	54,6732	2,330	56,6228	2,575	57,1152
19	45,2	53,8576	3,907	53,9777	3,737	54,0002	3,836	54,0002
20	38,3	44,2714	2,432	44,3315	2,503	44,5336	2,462	44,5336
21	34,5	46,9987	3,609	46,8499	2,438	46,9476	2,479	46,9987
22	39,2	42,6749	2,575	41,6654	2,541	41,1212	2,497	42,6749
23	0,009	38,3549	3,812	37,6973	3,705	38,4111	3,917	38,4111
24	0,66	45,7116	2,542	45,2705	2,468	45,5877	2,509	52,3542
25	0,027	52,6961	2,454	53,2524	2,446	52,6844	2,455	53,2524

26	2,01	49,3709	2,479	49,9856	2,511	49,8368	2,510	49,9856
27	0,099	57,2810	2,495	57,1956	2,503	57,4658	2,511	57,4658
28	0,127	62,2178	2,506	52,3846	6,830	52,3542	6,805	62,2178
29	0,467-	45,9391	8,13	46,0055	8,166	45,9764	8,024	46,0055
30	0,474	47,3001	2,576	44,6874	2,467	47,7913	2,607	47,7913
31	0,266	57,1349	2,385	58,9576	3,735	58,3432	3,673	58,9576
32	0,717	65,5512	8,35	65,5514	8,298	64,9187	8,313	65,5514
33	0,876	49,8063	4,397	46,1988	2,512	46,3854	2,552	49,8063
34	0,713	52,6460	10,438	52,4137	9,970	54,5563	9,836	54,5563
35	0,682	58,1712	2,164	58,5181	2,509	58,9115	2,541	58,9115
36	0,664	52,0586	2,449	52,0031	2,378	52,1742	2,415	52,1742
37	0,792	36,0931	7,158	36,0276	10,152	36,1132	10,180	36,1132
38	0,004	51,5850	5,505N	52,1270	2,575N	52,5774	2,659N	52,5774
39	0,0054	51,2478	2,142N	53,0249	2,231N	53,9953	2,385N	53,9953
40	0,0076	57,6691	2,408N	54,6166	2,498N	58,1163	2,580N	58,1163
41	0,0042	55,8068	2,515N	57,9720	2,576N	57,2865	2,682N	57,9720
42	0,0571	53,3905	2,564S	54,7423	2,454N	54,4318	2,690N	54,7423
43	0,0211	58,3066	1,932N	53,8352	2,542N	56,8102	2,391N	58,3066
44	0,0034	54,7700	2,473S	54,9165	2,544S	59,0532	2,596N	59,0532
45	0,0084	58,3769	2,562S	61,3041	2,542S	62,2137	2,600N	62,2137
46	0,0046	65,4568	7,829N	65,9909	6,548S	69,2649	5,626S	69,2649
47	0,0091	53,8998	4,113N	57,1043	2,561N	54,5431	2,509N	57,1043
48	0,0081	64,1458	2,491S	66,0451	2,538S	63,3136	6,096N	66,0451
49	0,0048	67,3421	5,524S	56,0090	11,757N	58,8422	6,512S	67,3421
50	0,678	51,4497	2,6	41,9143	8,078	41,5374	8,113	51,4497
51	0,955	41,6434	8,053	41,9304	8,142	41,5786	8,189	41,9304
52	0,346	42,5166	7,967	40,4111	8,512	43,0297	7,819	43,0297
53	5,48	41,687	7,292	48,6777	5,693	43,5865	7,960	48,6777
54	0,059	48,9007	12,818	45,7846	12,919	45,6865	11,504	48,9007
55	0,045	56,8476	13,415	57,2921	13,351	48,1244	9,575	57,2921
56	6048	54,0036	4,481	53,9405	4,442	53,7946	4,370	54,0036
57	1098	51,1566	4,576	51,2846	4,458	51,1692	4,612	51,2846
58	2,13	46,6498	9,019	51,4534	4,452	51,0842	2,669	51,4534
59	6,55	46,9382	9,329	57,5866	17,026	50,9157	7,652	57,5866
60	6,9	53,7901	2,314	53,3232	2,054	51,1176	2,489	53,7901

Em seguida analisamos as pontuações resultantes da função ChemScore, Assim, observamos na tabela 4,3 que para essa função o valor ≥ 27 da maior pontuação na análise em triplicata está associado a 47% dos compostos que têm $K_i \leq 7\mu M$. Com isso, vemos que a pontuação nessa função não é um bom separador,

Porém essa função não foi descartada nesse primeiro momento, passando pelas demais etapas de análise.

Tabela 5.3 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ChemScore.

Comp.	K_i (μM)	Pont.1	Dist(Å)	Pont.2	Dist (Å)	Pont.3	Dist (Å)	Pont. Máx.
1	33,7	27,8417	2,641	26,6398	2,543	27,473	2,508	27,8417
2	29,6	26,7163	9,7574	26,8214	2,507	26,7449	2,156	26,8214
3	28,4	24,9375	2,333	25,1739	2,257	24,7094	2,428	25,1739
4	38,9	27,7619	2,511	27,754	2,322	27,8754	2,475	27,8754
5	30,7	28,0882	2,434	28,1657	2,282	28,4428	2,21	28,4428
6	29,1	27,4979	2,299	28,8522	2,42	28,0938	2,396	28,8522
7	28,9	26,3244	2,306	26,4908	2,579	26,2347	2,399	26,4908
8	27,7	27,6394	2,437	27,8981	2,42	27,7358	2,502	27,8981
9	31,6	27,8449	2,445	27,7008	2,431	27,3305	2,229	27,8449
10	32,4	27,4564	2,196	27,8437	2,332	27,8755	2,437	27,8755
11	29,6	28,9854	2,576	27,998	2,543	29,0524	2,595	29,0524
12	32,5	29,8777	2,43	31,403	2,512	31,2787	2,607	31,403
13	2,09	22,1097	2,303	22,1097	2,303	22,1097	2,303	22,1097
14	2,38	23,1314	2,512	23,3776	2,408	22,9044	2,272	23,3776
15	0,978	26,5449	4,145	26,0714	4,099	26,3911	4,058	26,5449
16	3,21	30,7699	2,306	29,368	2,161	24,5851	8,384	30,7699
17	2,29	27,9131	1,947	25,3478	8,751	25,1363	8,717	27,9131
18	2,63	32,2891	2,435	30,6527	2,546	31,599	2,609	32,2891
19	45,2	27,4177	3,502	27,3415	3,461	27,0161	3,746	27,4177
20	38,3	27,1283	2,43	26,5366	2,379	26,6136	2,554	27,1283
21	34,5	27,0872	2,237	26,7142	2,496	26,3262	2,123	27,0872
22	39,2	23,7354	2,387	23,5129	2,449	24,7065	2,537	24,7065
23	0,678	21,7928	7,716	21,6984	7,313	21,585	7,917	21,7928
24	0,955	26,0493	7,703	26,3551	7,666	26,3957	7,608	26,3957
25	0,346	22,9232	10,632	24,8868	8,233	22,8712	10,579	24,8868
26	5,48	26,6325	2,21	24,3754	6,386	24,631	2,504	26,6325
27	0,059	19,6859	2,498	21,4337	3,66	19,919	11,252	21,4337
28	0,045	21,7175	2,301	19,4692	10,555	22,7798	2,634	22,7798
29	6,48	26,36	4,231	28,6441	2,2	28,7674	2,54	28,7674
30	1,98	31,5912	2,559	31,8267	2,364	27,5028	4,534	31,8267
31	2,13	31,2978	2,431	27,6962	2,271	30,8433	2,544	31,2978
32	6,55	25,4247	2,296	24,0601	2,046	23,6998	2,592	25,4247
33	6,9	27,2959	2,31	28,0247	2,318	28,5723	2,645	28,5723
34	0,0044	28,3168	2,573	29,2556	4,927	30,8344	2,593	30,8344
35	0,0054	26,1432	2,543	27,6242	2,442	27,1291	2,434	27,6242
36	0,0076	26,122	2,659	26,2678	4,961	26,5951	2,645	26,5951
37	0,0042	28,1518	2,206	26,8542	2,483	26,9341	2,575	28,1518
38	0,0571	26,4451	2,795	26,7726	2,648	26,4134	2,555	26,7726

39	0,0211	27,174	2,145	27,0573	2,556	28,2323	2,385	28,2323
40	0,0034	20,8701	2,511	24,008	2,371	24,7492	2,228	24,7492
41	0,0084	23,0321	2,450s	23,1267	2,516s	26,6218	2,578	26,6218
42	0,0046	31,1874	2,119	28,7557	9,479	27,2094	5,638s	31,1874
43	0,0092	26,0438	2,373	25,3633	2,686	25,878	2,619	26,0438
44	0,0081	27,7307	2,255	27,5973	2,058	25,7043	2,477	27,7307
45	0,0048	27,5854	6,167s	26,1256	2,56	27,8856	2,677	27,8856
46	0,009	22,5151	3,939	21,6404	3,622	21,6965	3,304	22,5151
47	0,127	26,6578	2,358	22,9645	3,765	27,576	2,596	27,576
48	0,467	23,509	4,582	22,9645	3,765	27,5206	2,548	27,5206
49	0,713	25,1852	11,66	25,4677	8,816	25,6247	9,085	25,6247
50	2,01	30,822	2,5	30,2014	2,255	30,8869	2,51	30,8869
51	0,099	30,1217	2,602	30,1791	2,299	30,2168	2,423	30,2168
52	0,027	28,2363	2,523	27,9647	2,568	28,0265	2,558	28,2363
53	0,717	32,4519	8,507	32,1269	8,489	32,0625	8,537	32,4519
54	0,66	22,588	2,439	22,3587	2,555	22,3412	2,492	22,588
55	0,792	20,6774	4,372	20,6907	4,329	20,6947	4,373	20,6947
56	0,266	29,268	2,554	28,9305	2,467	28,2161	2,572	29,268
57	0,664	29,953	2,471	30,836	2,396	29,4716	2,374	30,836
58	0,474	22,3778	2,31	21,252	2,396	20,4394	2,453	22,3778
59	0,682	38,118	2,251	37,3826	2,646	37,4526	2,458	38,118
60	0,876	28,0433	2,43	27,9045	2,323	27,8302	2,521	28,0433

Tabela 5.4 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ASP.

Comp.	$K_i (\mu M)$	Pont.1	Dist(Å)	Pont.2	Dist (Å)	Pont.3	Dist (Å)	Pont. Máx.
1	33,7	16,0726	9,427	16,313	9,351	16,4614	9,294	16,4614
2	29,6	16,4807	7,777	16,6438	7,934	16,3911	8,054	16,6438
3	28,4	18,4344	2,253	17,3171	10,581	17,0059	9,502	18,4344
4	38,9	16,5599	7,696	16,3839	7,932	15,7514	8,244	16,5599
5	30,7	16,9598	7,937	17,1127	7,335	16,9588	8,074	17,1127
6	29,1	18,2328	7,626	18,1031	7,79	17,9694	9,003	18,2328
7	28,9	17,8769	7,986	17,5602	8,376	18,1207	8,028	18,1207
8	27,7	17,9827	4,687	17,3215	4,738	17,8221	7,36	17,9827
9	31,6	15,6426	7,768	15,7124	8,248	15,6725	7,722	15,7124
10	32,4	17,5314	7,048	17,7514	8,251	17,7002	8,447	17,7514
11	29,6	23,1343	2,273	24,2583	2,198	23,4743	3,129	24,2583
12	32,5	20,2699	2,462	20,6035	2,089	20,1343	2,268	20,6035
13	2,09	11,9943	2,356	16,5174	2,211	16,4359	2,234	16,5174
14	2,38	17,0168	5,189	16,7336	5,325	16,8	5,475	17,0168
15	0,978	20,9353	12,107	21,0092	13,383	21,1708	11,55	21,1708
16	3,21	24,1754	8,133	24,0864	8,429	23,7884	11,202	24,1754
17	2,29	23,5890	13,762	23,3853	12,915	23,7663	13,926	23,7663

18	2,63	22,2344	3,785	22,2693	3,663	21,1997	3,778	22,2693
19	45,2	19,2028	3,253	18,6138	6,461	19,0874	3,557	19,2028
20	38,3	16,6629	8,017	16,0912	8,038	17,321	8,085	17,3210
21	34,5	16,9609	7,635	17,6806	7,179	17,6815	7,786	17,6815
22	39,2	14,9514	7,251	15,8927	7,058	14,2054	7,189	15,8927
23	0,678	19,1006	7,862	19,311	7,704	19,4401	7,099	19,4401
24	0,955	20,2772	7,885	20,8044	7,943	19,8286	7,888	20,8044
25	0,346	19,1143	8,024	20,3047	7,71	20,3262	7,714	20,3262
26	5,48	21,1045	7,667	20,938	2,333	21,2177	7,629	21,2177
27	0,059	26,2365	11,934	25,6779	12,315	25,5572	11,348	26,2365
28	0,045	21,6428	11,781	22,9435	11,848	23,4117	11,835	23,4117
29	6048	19,6207	4,569	19,6846	4,235	20,1633	4,687	20,1633
30	1098	21,4538	4,363	21,7606	4,209	20,3778	5,182	21,7606
31	2,13	20,3580	7,051	19,8583	7,709	20,6655	8,661	20,6655
32	6,55	25,2980	10,734	24,4748	10,111	23,3454	6,12	25,2980
33	6,9	23,1301	10,295	23,7874	10941	24,055	9,479	24,0550
34	0,0044	21,2462	6,274	20,0831	7,756	20,3113	5,02	21,2462
35	0,0054	21,4372	3,94	21,7236	6,683	21,126	8,656	21,7236
36	0,0076	22,2714	7,579	20,9269	9,332	21,5274	2,47	22,2714
37	0,0042	20,8299	4,347	21,2196	2,506	20,8383	6,161	21,2196
38	0,0571	19,4261	3,922s	20,3799	3,763	19,7162	2,937s	20,3799
39	0,0211	20,7437	3,544	20,6462	8,291	20,1613	2,145	20,7437
40	0,0034	22,3812	2,418s	22,4437	8,986	22,2232	8,299	22,4437
41	0,0084	22,7136	2,507s	22,9949	7,091	23,8559	7,627	23,8559
42	0,0046	26,1218	2,750s	25,9819	10,669	26,3022	8,522	26,3022
43	0,0092	20,6773	5,481s	20,9658	10,249	20,7827	2,503	20,9658
44	0,0081	23,7818	9,150s	23,2958	11,081	23,7749	2,267	23,7818
45	0,0048	25,4581	6,558	26,5965	5,806s	26,7513	6,236	26,7513
46	0,009	17,7399	3,417	18,0142	3,438	17,29	8,771	18,0142
47	0,127	21,1645	6,026	21,6076	6,453	22,1491	5,189	22,1491
48	0,467	20,1791	13,256	19,7189	11,988	19,8796	7,091	20,1791
49	0,713	24,7964	10,255	22,5488	11,142	24,7247	10,946	24,7964
50	2,01	20,3755	3,096	20,1464	2,443	19,127	2,285	20,3755
51	0,099	19,4678	2,446	19,4945	2,326	19,9224	2,438	19,9224
52	0,027	19,0332	2,564	18,4309	2,177	18,3902	2,334	19,0332
53	0,717	20,9524	8,18	18,9338	9,051	20,7584	9,459	20,9524
54	0,66	17,5017	2,327	17,5947	2,581	16,4664	2,361	17,5947
55	0,792	17,6851	10,245	17,594	10,063	17,452	10,271	17,6851
56	0,266	22,0673	9,197	20,2593	9,113	19,6296	2,864	22,0673
57	0,664	18,8564	2,166	18,3971	2,954	19,3838	2,325	19,3838
58	0,474	16,1170	4,708	17,734	4,469	16,7724	3,761	17,7340
59	0,682	18,5406	11,898	18,6685	11,943	19,0133	11,902	19,0133
60	0,876	17,0792	2,267	17,0446	2,235	17,5168	2,302	17,5168

Após essa análise concluímos que a pontuação de ajuste obtida pela docagem está correspondendo qualitativamente ao valor de K_i para a função GoldScore, com acerto na classificação dos inibidores (fortes x fracos) para 86,67% dos casos, e para a função ChemScore com acerto de 76,67% dos casos.

Como é possível que o grupo sulfonamida seja desprotonado após coordenação com o íon Zn^{2+} , formando um ânion no N do grupo sulfonamida, todas as moléculas foram também otimizadas nesse estado aniônico e avaliadas por docagem molecular da mesma forma que a usada para as moléculas com o grupo neutro. Os resultados são apresentados nas tabelas 5.5 a 5.8. Nas tabelas foram considerados 63 ligantes, pois as moléculas 15, 16 e DCP apresentam dois grupos sulfonamida na sua estrutura. Logo esses ligantes foram duplicados na tabela, mas com H^+ retirado de grupos SO_2NH_2 diferentes. Sendo assim, a numeração dos compostos 1 a 60 das tabelas 5.1 a 5.4, quando ionizados, passa a ser nas tabelas 5.5 a 5.8 os números apresentados entre parênteses: 1 (61), 2 (62), 3 (63), 4 (64), 5 (65), 6 (66), 7 (67), 8 (68), 9 (69), 10 (70), 11 (71), 12 (72), 13 (73), 14 (74), 15 (75), 15.2 (76), 16 (77), 16.2 (78), 17 (79), 18 (80), 19 (81), 20 (82), 21 (83), 22 (84), 23 (85), 24 (86), 25 (87), 26 (88), 27 (89), 28 (90), 29 (91), 30 (92), 31 (93), 32 (94), 33 (95), 34 (96), 35 (97), 36 (98), 37 (99), 38 (100), 39 (101), 40 (102), 41 (103), 42 (104), 43 (105), 44 (106), 45 (107), 46 (108), 47 (109), 48 (110), 49 (111), 50 (112), 50.2 (113), 51 (114), 52 (115), 53 (116), 54 (117), 55 (118), 56 (119), 57 (120), 58 (121), 59 (122), 60 (123).

Tabela 5.5– dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função GoldScore, para moléculas ionizadas.

Comp.	K_i (μM)	Pont.1	Dist(Å)	Pont.2	Dist (Å)	Pont.3	Dist (Å)	Pont. Máx.
61	33,7	37,3474	4,142	38,0109	3,961	38,0539	3,812	38,054
62	29,6	44,6725	7,249	44,6604	7,259	44,4983	6,544	44,673
63	28,4	50,1512	9,582	49,6253	9,657	45,1281	4,293	50,151
64	38,9	37,9727	2,537	38,1754	2,554	38,2353	4,293	38,235
65	30,7	44,3567	8,738	44,4449	8,366	43,802	8,89	44,445
66	29,1	43,216	4,182	47,1902	8,723	42,9998	4,2	47,190
67	28,9	44,8027	7,257	44,8585	7,36	44,6941	7,367	44,859
68	27,7	45,2729	6,832	44,7949	6,764	45,0862	7,076	45,273
69	31,6	43,8611	7,134	43,4146	7,808	42,6346	7,452	43,861
70	32,4	45,0459	6,799	44,92	6,63	44,9231	7,305	45,046
71	29,6	47,2434	2,22	46,4215	2,203	47,1882	2,198	47,243
72	32,5	46,9979	2,294	47,2271	2,364	48,3942	2,441	48,394
73	2,09	41,5922	6,503	38,3921	6,664	42,0499	6,509	42,050

74	2,38	42,8236	2,212	41,5137	2,291	41,7805	2,311	42,824
75	0,978	52,5887	9,215	52,6343	9,163	48,3072	8,786	52,634
76	0,978	61,4753	2,392	58,4066	2,303	60,9003	2,307	61,475
77	3,21	44,4476	4,483	43,9448	4,482	44,2491	4,326	44,448
78	3,21	60,0288	4,357	60,9611	4,137	61,2175	4,139	61,218
79	2,29	61,9267	4,076	63,4784	4,213	60,6149	4,542	63,478
80	2,63	52,9012	9,601	51,5244	4,053	51,8935	4,222	52,901
81	45,2	48,9076	3,933	48,1113	4,309	48,2661	3,928	48,908
82	38,3	41,4956	7,868	41,9274	3,977	41,9392	6,664	41,939
83	34,5	42,0037	4,251	42,1929	4,124	41,9636	4,105	42,193
84	39,2	42,1306	6,701	41,9418	6,598	42,0981	6,709	42,131
85	0,678	47,7003	4,407	47,6546	4,496	47,8858	4,438	47,886
86	0,955	50,6247	12,566	42,3850	4,417	42,2869	4,410	50,625
87	0,346	49,9064	8,6805	49,3709	4,457	48,6786	4,556	49,906
88	5,48	43,232	4,519	41,7051	4,74	43,8083	4,519	43,808
89	0,059	58,504	4,281	54,6446	11,809	56,3908	13,045	58,504
90	0,045	52,9037	11,566	53,4727	11,61	53,5089	11,576	53,509
91	6,48	47,8007	3,813	47,3600	4,972	46,683	3,908	47,801
92	1,98	50,3974	3,151	50,1825	3,054	49,9963	3,021	50,397
93	2,13	49,2301	4,322	52,0500	4,252	48,8303	4,341	52,050
94	6,55	63,9861	63,9861	52,5275	4,972	53,9581	4,763	63,986
95	6,9	45,23	11,445	53,5573	11,493	44,5511	11,232	53,557
96	0,0044	57,5214	2,149S	57,0650	2,272	53,9347	2,131	57,521
97	0,0054	54,5203	4,420N	59,4752	2,216	55,0190	4,357N	59,475
98	0,0076	60,3901	2,161S	59,5018	1,998	57,6029	2,051	60,390
99	0,0042	57,5035	1,766S	52,3239	2,299N	58,1501	2,233	58,150
100	0,0571	61,0492	2,052S	62,3377	2,203	54,2617	4,509N	62,338
101	0,0211	60,2905	2,163S	54,0174	2,466	55,0037	4,529N	60,291
102	0,0034	60,1237	2,059S	59,1449	2,374	60,3325	2,036	60,333
103	0,0084	58,7491	2,260S	58,1844	2,165	58,6497	2,133	58,749
104	0,0046	64,8803	2,347S	70,6556	2,508	67,8282	2,383	70,656
105	0,0092	55,3277	4,200N	54,4991	4,158N	57,6956	4,318N	57,696
106	0,0081	66,9839	2,173S	64,4144	2,094	62,4004	1,981	66,984
107	0,0048	69,7865	2,473S	67,3357	2,444	67,4746	2,374	69,787
108	0,009	41,0937	4,071	40,918	4,117	41,6909	3,999	41,691
109	1,27	54,7409	3,294	56,4885	3,31	54,5327	7,241	56,489
110	0,467	53,3577	4,467	53,3318	2,422	51,2801	4,366	53,358
111	0,713	50,8262	4,512	51,4647	4,721	50,987	4,605	51,465
112	2,01	49,2753	2,212	48,9432	2,472	48,9093	2,258	49,275
113	2,01	47,7977	2,488	47,6206	2,354	46,0307	3,807	47,798
114	0,099	51,9124	3,506	52,8609	3,598	52,9562	3,716	52,956
115	0,027	47,6931	3,989	47,5938	3,907	47,85	3,976	47,850
116	0,717	58,9488	7,927	58,5276	7,993	58,5357	7,769	58,949
117	0,66	42,9901	1,917	42,9259	1,894	42,8456	1,745	42,990
118	0,792	35,8516	3,5449	37,7875	3,002	36,0121	3,464	37,788

119	0,266	53,3107	3,665	51,1094	3,509	53,78	53,7767	53,780
120	0,664	50,5755	4,125	50,518	4,222	50,2576	4,223	50,576
121	0,474	41,334	1,7	41,7341	2,948	41,1799	1,973	41,734
122	0,682	55,6178	4,264	54,2697	4,374	55,6868	4,264	55,687
123	0,876	38,1703	3,599	37,861	3,592	38,2754	3,589	38,275

Tabela 5.6 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ChemPLP, para moléculas ionizadas.

Comp.	K_i (μM)	Pont.1	Dist(Å)	Pont.2	Dist (Å)	Pont.3	Dist (Å)	Pont. Máx.
61	33,7	35,5020	3,549	35,4939	3,46	34,1336	34,1336	35,5020
62	29,6	38,3941	8,043	38,4636	8,088	38,3825	8,055	38,4636
63	28,4	49,1116	2,41	48,5183	2,475	50,0166	2,547	50,0166
64	38,9	36,9932	3,585	37,1444	3,486	36,4659	3,595	37,1444
65	30,7	41,2577	4,176	38,6954	2,555	39,7673	2,39	41,2577
66	29,1	45,6982	3,606	45,4147	3,839	45,6564	3,637	45,6982
67	28,9	42,0531	8,081	42,4131	8,144	41,7377	7,358	42,4131
68	27,7	41,9441	8,121	41,5597	8,187	41,4947	8,122	41,9441
69	31,6	40,5879	4,082	41,1463	4,099	39,8774	7,758	41,1463
70	32,4	38,9279	8,414	40,046	8,099	41,4643	3,846	41,4643
71	29,6	47,8584	2,466	48,5521	2,597	48,1673	2,585	48,5521
72	32,5	46,0199	4,265	47,1100	4,594	44,4698	2,702	47,1100
73	2,09	32,3400	6,902	32,9546	6,760	32,34	6,869	32,9546
74	2,38	36,8369	2,640	36,8627	4,130	36,8299	2,598	36,8627
75	0,978	42,7079	9,760	43,1699	9,709	43,3479	9,755	43,3479
76	0,978	48,8434	10,650	48,4198	11,846	48,8074	10,729	48,8434
77	3,21	41,8466	4,015	40,8750	9,759	41,9155	9,726	41,9155
78	3,21	59,6595	7,418	60,0674	7,609	59,7161	7,504	60,0674
79	2,29	57,7020	8,087	58,9154	7,708	58,3798	7,155	58,9154
80	2,63	51,8475	3,531	52,5684	3,753	51,8898	2,586	52,5684
81	45,2	50,7510	2,480	50,7736	2,455	50,7505	3,447	50,7736
82	38,3	41,3359	4,223	41,3665	4,206	39,7303	2,493	41,3665
83	34,5	45,3375	3,750	44,7851	3,699	45,2744	3,737	45,3375
84	39,2	41,9667	4,208	39,3930	2,517	39,1308	2,341	41,9667
85	0,009	37,2944	2,465	37,3448	2,458	37,0625	2,573	37,3448
86	0,66	41,0163	2,233	40,4670	2,609	40,9622	2,636	41,0163
87	0,027	48,0915	2,135	47,9966	2,154	48,1209	2,082	48,1209
88	2,01	48,6807	2,476	47,8017	2,476	47,5739	2,423	48,6807
89	2,01	46,6429	2,575	48,5889	4,152	47,8778	2,565	48,5889
90	0,099	48,1571	3,845	48,7841	3,582	47,1691	3,787	48,7841
91	0,127	53,0534	3,475	52,1605	3,510	50,0106	2,735	53,0534
92	0,467	47,1998	8,025	45,4088	8,242	46,4798	8,319	47,1998
93	0,474	42,2941	2,607	36,6468	3,491	36,6056	3,535	42,2941
94	0,266	54,0258	2,204	56,2395	3,790	55,948	3,745	56,2395

95	0,717	63,0413	8,326	62,9618	8,190	62,8022	8,336	63,0413
96	0,876	41,3045	2,544	47,3041	4,361	47,2381	4,341	47,3041
97	0,713	52,8060	5,87	51,9073	7,731	52,3474	9,678	52,8060
98	0,682	51,3640	2,078	54,5759	4,852	43,7528	11,211	54,5759
99	0,664	42,5524	2,36	48,3257	2,644	44,0664	2,407	48,3257
100	0,792	32,7664	6,963	32,8136	6,971	32,8514	6,967	32,8514
101	0,0044	46,7192	2,531N	48,4074	2,501N	47,3933	2,406N	48,4074
102	0,0054	50,1135	2,426N	49,9752	2,365N	51,5523	3,709N	51,5523
103	0,0076	54,4406	2,531N	52,5025	5,026N	54,2674	2,642N	54,4406
104	0,0042	55,5070	3,644N	53,2708	2,510N	52,4435	2,527N	55,5070
105	0,0571	51,3699	4,674N	51,4394	2,127N	51,6493	2,488N	51,6493
106	0,0211	51,1122	2,572S	53,0464	3,786N	48,5771	2,552S	53,0464
107	0,0034	54,4109	2,569S	52,5148	2,509S	56,0112	2,102N	56,0112
108	0,0084	53,9970	2,506S	53,8649	2,519S	55,4999	2,604S	55,4999
109	0,0046	66,2644	6,666S	67,492	3,614S	65,3808	7,455S	67,4920
110	0,0092	52,3329	2,550N	55,4527	3,609N	52,3281	2,167N	55,4527
111	0,0081	64,3603	2,738S	61,7475	6,018S	60,5459	5,862S	64,3603
112	0,0048	57,4472	5,491S	60,9601	6,196S	59,3331	5,564S	60,9601
113	0,678	40,2057	6,587	40,4655	6,421	40,4285	6,455	40,4655
114	0,955	43,6671	6,654	43,4623	6,68	43,6391	6,756	43,6671
115	0,346	41,3971	12,672	40,3171	13,247	41,6757	12,405	41,6757
116	5,48	46,0572	11,671	44,0507	6,784	44,128	6,716	46,0572
117	0,059	42,7131	12,987	46,8326	13,887	43,4186	12,878	46,8326
118	0,045	42,6202	7,497	39,5631	8,678	45,6379	12,052	45,6379
119	6048	46,3672	4,543	46,446	4,225	46,6158	4,308	46,6158
120	1098	43,6453	6,738	43,672	6,805	43,5904	6,856	43,6720
121	2,13	48,0392	4,595	47,9423	4,611	48,0931	4,624	48,0931
122	6,55	47,8332	10,644	47,7342	10,609	47,4941	10,619	47,8332
123	6,9	45,5702	11,54	44,8974	14,793	45,5384	11,764	45,5702

Tabela 5.7 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ChemScore, para moléculas ionizadas.

Comp.	K_i (μM)	Pont.1	Dist (Å)	Pont.2	Dist (Å)	Pont.3	Dist (Å)	Pont. Máx.
61	33,7	22,3620	2,662	25,3088	2,591	22,7700	2,596	25,3088
62	29,6	26,9332	7,953	27,0034	8,025	26,9830	7,9780	27,0034
63	28,4	21,9914	2,619	21,6425	2,591	21,3127	2,7120	21,9914
64	38,9	23,7413	2,599	22,9743	2,611	23,2629	2,6790	23,7413
65	30,7	24,1920	2,638	24,2637	2,596	23,6391	2,5170	24,2637
66	29,1	24,9250	9,546	25,2895	9,067	25,2304	9,3000	25,2895
67	28,9	26,8390	7,226	26,8959	7,971	26,9170	7,8390	26,9170
68	27,7	27,7421	7,795	27,8108	7,835	27,8789	7,9710	27,8789
69	31,6	27,6240	8,119	27,6863	7,945	27,4618	7,3800	27,6863
70	32,4	27,2080	8,089	27,2328	7,944	27,0828	8,1210	27,2328

71	29,6	28,4373	2,493	28,9694	2,969	28,4844	2,4710	28,9694
72	32,5	31,0749	2,649	31,6040	2,454	31,1732	2,6780	31,6040
73	2,09	21,3256	8,16	21,2391	8,167	21,3256	8,1600	21,3256
74	2,38	23,3436	2,527	23,8149	2,492	24,3764	2,3210	24,3764
75	0,978	25,7325	9,089	25,5614	9,225	25,6777	9,1360	25,7325
76	0,978	30,384	9,891	30,5360	9,757	30,3767	9,6770	30,536
77	3,21	24,8006	9,757	24,0933	9,595	25,0009	9,6860	25,0009
78	3,21	30,3897	11,68	30,1789	12,178	30,2738	11,3630	30,3897
79	2,29	21,9215	9,218	24,3559	9,951	25,8398	2,6200	25,8398
80	2,63	28,4034	9,632	28,3233	12,610	28,3401	12,6520	28,4034
81	45,2	25,8636	2,496	25,6910	2,472	25,7581	2,4310	25,8636
82	38,3	22,3545	2,605	22,5503	2,695	22,1148	2,5650	22,5503
83	34,5	23,1365	2,446	23,5698	2,592	23,1707	2,4900	23,5698
84	39,2	22,4118	3,995	22,6129	3,726	22,6098	3,6380	22,6129
85	0,678	20,1488	5,422	19,7423	5,062	20,2779	5,4080	20,2779
86	0,955	24,0114	7,682	24,7228	13,100	24,1313	7,5710	24,7228
87	0,346	22,6262	12,135	22,4007	10,408	22,993	7,0850	22,9930
88	5,48	20,0396	9,67	20,0717	11,282	20,7867	11,4730	20,7867
89	0,059	14,8938	12,219	16,9586	10,720	14,3825	10,4600	16,9586
90	0,045	17,3550	11,362	17,3923	11,708	17,5375	11,4190	17,5375
91	6,48	25,6078	2,604	25,1197	2,474	26,3179	2,5740	26,3179
92	1,98	26,1068	12,167	25,7067	11,391	25,7772	12,1330	26,1068
93	2,13	24,6295	2,261	22,4934	11,075	26,3585	4,3470	26,3585
94	6,55	19,2486	2,559	22,5163	2,497	21,1198	2,469	22,5163
95	6,9	17,2482	11,287	17,5322	9,725	18,0471	11,0530	18,0471
96	0,0044	26,0886	3,915	26,3901	2,574	25,7129	2,4980	26,3901
97	0,0054	24,0721	2,628	22,7150	4,037	23,4748	2,4750	24,0721
98	0,0076	22,7981	4,071	22,7150	4,223	23,5420	4,1360	23,5420
99	0,0042	23,6308	9,414	23,4698	3,439	23,6358	9,4920	23,6358
100	0,0571	23,5431	9,563	23,4185	4,203	23,4494	9,6210	23,5431
101	0,0211	22,6500	9,421	23,6122	9,489	22,9807	9,4600	23,6122
102	0,0034	21,2031	9,575	19,6221	10,815	20,5655	9,7780	21,2031
103	0,0084	22,7010	9,828	22,5866	10,492	23,2364	10,7930	23,2364
104	0,0046	26,1426	11,51	24,8217	11,080	27,3836	8,9220	27,3836
105	0,0092	21,6801	2,685	24,5751	2,541	23,3347	2,5980	24,5751
106	0,0081	24,2172	10,523	25,5027	10,241	23,9330	10,0360	25,5027
107	0,0048	22,2108	11,557	25,6226	2,030	21,7047	11,6720	25,6226
108	0,009	16,9062	2,543	17,0178	2,469	16,7836	2,4470	17,0178
109	1,27	18,7330	6,784	18,5977	9,706	17,3509	10,5030	18,7330
110	0,467	23,7345	4,761	24,8287	2,440	23,3980	4,6550	24,8287
111	0,713	25,9483	10,328	25,8784	6,973	25,9130	9,8970	25,9483
112	2,01	28,7466	2,439	28,7396	2,073	29,0085	2,5090	29,0085
113	2,01	27,7675	2,682	27,1768	2,361	27,9313	2,5390	27,9313
114	0,099	21,9998	3,938	20,1555	4,972	20,5356	5,0270	21,9998
115	0,027	19,8353	2,51	19,8278	2,492	20,0381	2,4780	20,0381

116	0,717	28,4827	8,864	30,9067	8,686	31,9182	2,5730	31,9182
117	0,66	23,4587	2,524	21,6726	2,670	20,7715	2,5430	23,4587
118	0,792	21,3446	4,393	21,2224	4,419	21,1918	4,408	21,3446
119	0,266	23,9809	2,583	25,4141	2,240	23,3957	2,3140	25,4141
120	0,664	25,4055	2,497	26,6876	2,646	27,5320	2,5380	27,5320
121	0,474	16,0362	2,413	17,3457	2,400	17,6109	2,5070	17,6109
122	0,682	32,9352	2,497	32,8984	2,508	31,8183	2,3890	32,9352
123	0,876	24,1764	2,531	23,7254	2,503	23,9921	2,4810	24,1764

Tabela 5.8 – dados experimentais de K_i e as pontuações e distâncias entre o N da sulfonamida e o íon Zn^{2+} dada pela função ASP, para moléculas ionizadas.

Comp.	K_i (μM)	Pont.1	Dist (Å)	Pont.2	Dist (Å)	Pont.3	Dist (Å)	Pont. Máx.
Comp.	33,7	17,0388	10,707	17,6651	10,444	16,4695	9,977	17,6651
61	29,6	16,3288	7,717	16,2017	7,49	16,6741	7,673	16,6741
62	28,4	18,8462	9,956	17,4382	9,681	18,3772	10,394	18,8462
63	38,9	15,6266	8,229	15,9893	8,125	16,3613	7,472	16,3613
64	30,7	17,0432	7,509	16,1451	10,718	16,7553	7,657	17,0432
65	29,1	18,2625	7,71	17,6695	8,335	18,2344	8,708	18,2625
66	28,9	18,3672	5,654	18,5161	5,901	18,7073	9,323	18,7073
67	27,7	18,4447	13,171	17,6953	8,999	17,4700	7,285	18,4447
68	31,6	16,7865	6,807	16,8688	6,684	16,8144	5,613	16,8688
69	32,4	17,8757	8,035	18,1996	7,229	17,8758	6,844	18,1996
70	29,6	23,9914	2,166	24,749	2,235	24,0645	2,415	24,7490
71	32,5	18,2294	2,274	19,7681	2,218	20,1107	3,407	20,1107
72	2,09	14,4549	2,182	12,1392	2,413	12,3605	7,998	14,4549
73	2,38	16,2352	5,406	16,2958	2,485	16,7655	3,277	16,7655
74	0,978	17,8858	7,753	17,6801	7,929	17,9368	7,798	17,9368
75	0,978	21,9340	4,493	21,4771	3,746	21,3495	7,858	21,9340
76	3,21	15,3507	7,765	15,2793	7,377	16,7321	7,541	16,7321
77	3,21	21,9962	7,348	22,7607	10,896	22,5162	7,035	22,7607
78	2,29	24,3310	13,846	23,5203	10,524	23,7873	10,433	24,3310
79	2,63	24,0610	2,348	22,5883	9,793	24,3984	2,254	24,3984
80	45,2	19,2986	2,338	20,1283	2,293	19,7406	2,598	20,1283
81	38,3	17,3598	7,418	16,2728	7,789	16,9065	4,985	17,3598
82	34,5	18,1762	7,688	18,0602	8,03	18,0907	7,993	18,1762
83	39,2	16,3064	10,405	15,9034	10,563	16,3187	10,535	16,3187
84	0,678	18,9008	5,233	18,4197	5,071	17,9906	6,422	18,9008
85	0,955	19,7584	6,25	19,0383	5,481	19,5961	6,618	19,7584
86	0,346	19,2315	7,646	21,4056	7,406	18,0072	11,657	21,4056
87	5,48	21,5692	6,845	21,3379	6,845	21,4056	7,07	21,5692
88	0,059	25,1667	10,971	25,997	11,946	22,5418	10,767	25,997
89	0,045	20,3148	11,22	20,1978	9,67	21,1286	11,263	21,1286
90	6,48	18,3498	6,181	19,7755	7,728	18,7509	4,842	19,7755

91	1,98	20,5066	7,049	20,5126	6,595	19,5961	6,727	20,5126
92	2,13	20,0969	8,472	20,3914	6,998	19,7539	8,213	20,3914
93	6,55	24,4108	10,818	24,8576	10,131	25,0191	10,521	25,0191
94	6,9	20,4027	11,567	21,1083	11,715	20,0206	11,309	21,1083
95	0,0044	21,3250	9,344	19,8364	10,47	20,1841	10,644	21,3250
96	0,0054	20,7180	7,474	20,946	10,597	21,7456	8,91	21,7456
97	0,0076	21,4676	5,567	21,8365	2,311	20,8918	9,067	21,8365
98	0,0042	21,1992	9,236	21,3379	9,106	21,9631	9,493	21,9631
99	0,0571	20,0261	10,198	19,8561	9,888	20,8691	9,654	20,8691
100	0,0211	20,4004	2,903	20,1978	10,326	21,2356	8,327	21,2356
101	0,0034	23,1153	9,022	20,1978	7,8	22,0204	7,818	23,1153
102	0,0084	23,1375	7,945	23,0419	8,536	24,1260	7,555	24,1260
103	0,0046	26,6053	14,514	25,8651	14,577	26,6617	11,251	26,6617
104	0,0092	21,3721	9,345	19,597	9,047	20,4217	9,825	21,3721
105	0,0081	21,9766	15,325	24,4589	10,605	23,8900	10,362	24,4589
106	0,0048	25,3915	10,414	25,7193	11,61	24,8898	10,492	25,7193
107	0,009	16,5460	2,418	16,7136	2,234	17,3293	2,424	17,3293
108	1,27	19,7129	7,773	20,7692	5,636	20,3011	10,426	20,7692
109	0,467	20,0328	10,719	19,7672	9,898	20,7746	4,377	20,7746
110	0,713	26,1113	5,988	23,3142	6,503	23,4162	5,997	26,1113
111	2,01	19,4870	2,567	18,59	3,296	20,7199	3,987	20,7199
112	2,01	20,2189	2,288	19,7089	2,081	20,4087	2,216	20,4087
113	0,099	19,9903	5,489	17,4009	17,444	18,2259	6,626	19,9903
114	0,027	17,2098	11,039	18,2853	19,0907	18,2027	2,143	18,2853
115	0,717	19,8881	9,207	21,7626	8,594	21,6946	8,54	21,7626
116	0,66	17,3143	2,488	17,0984	2,309	17,8943	2,621	17,8943
117	0,792	15,8960	7,534	16,7747	8,055	16,7260	7,801	16,7747
118	0,266	22,2671	9,129	21,5736	10,506	22,5225	8,576	22,5225
119	0,664	20,5754	9,989	18,1666	2,278	19,0247	9,418	20,5754
120	0,474	17,2009	8,497	16,6177	6,065	16,7677	7,889	17,2009
121	0,682	20,0010	2,12	18,5095	11,883	18,2451	11,726	20,0010
122	0,876	15,3109	2,507	15,4408	2,286	15,9284	2,418	15,9284

Após essa análise concluímos que a pontuação de ajuste das funções GoldScore, ChemPLP e ASP funcionou adequadamente como critério de separação entre inibidores muito ativos (K_i igual ou inferior a $7,0 \mu\text{M}$) e os pouco ativos, apresentando um bom grau de acerto (Tabela 4.9). Foi tentado estipular um valor limite de pontuação para a função ChemScore, porém nenhum número se apresentou um bom separador.

Ao se considerar as moléculas na forma de ânions, houve uma queda no desempenho, embora com valores ainda bons de percentual de acerto na

classificação dos compostos para as funções GoldScore, ChemPLP e ASP. Possivelmente isso se deve aos parâmetros de cada função que podem não ter sido tão bem ajustados para moléculas contendo grupos sulfonamida no estado aniônico.

A função ChemScore não apresentou resultados bons em relação à separação nesse primeiro momento, mas não foi descartada a fim de ser analisado melhor nas próximas etapas. Todas as funções evidenciaram interação ligante-receptor pela pontuação e pela distância do ligante com o sítio ativo.

Tabela 5.9 – Percentual de acerto a partir dos parâmetros estipulados nas funções para classificação dos compostos como muito ou pouco ativos.

Função	GoldScore	ChemPLP	ChemScore	ASP
Moléculas neutras	86,6%	76,6%	36,6%	83,3%
Moléculas ionizadas	73%	74%	-	76,6%

Os resultados obtidos indicam que os dados gerados pelo GOLD têm boa qualidade e poderão permitir que se analise a interação proteína-ligante no nível molecular de forma que se consiga encontrar um modelo de correlação com as atividades que auxilie na busca de novos inibidores para a MtCA2, candidatos a fármacos para o tratamento da TB.

5.2 Análise exploratória de dados

Como mencionado em 3.3.1, temos 4 conjuntos de dados, um para cada função de maneira que eles tenham 60 objetos e, além das variáveis das moléculas isoladas obtidas pelos cálculos semiempíricos, para a função GoldScore são 35 variáveis, para ChemScore 35 variáveis, para ChemPLP 36 variáveis e para ASP 34 variáveis. Em relação ao conjunto formado de íons, eles possuem 63 objetos e as mesmas variáveis. São conjuntos de alta dimensionalidade, então inicialmente fizemos a análise das matrizes correlação dos dados.

5.2.1 Matriz correlação

Podemos observar na matriz correlação se temos variáveis que estão tendo uma correlação apreciável, a partir do critério do coeficiente acima de 0,5 e abaixo de -0,5. Assim, foram eliminadas as variáveis que não estava tendo nenhuma correlação boa com as demais variáveis.

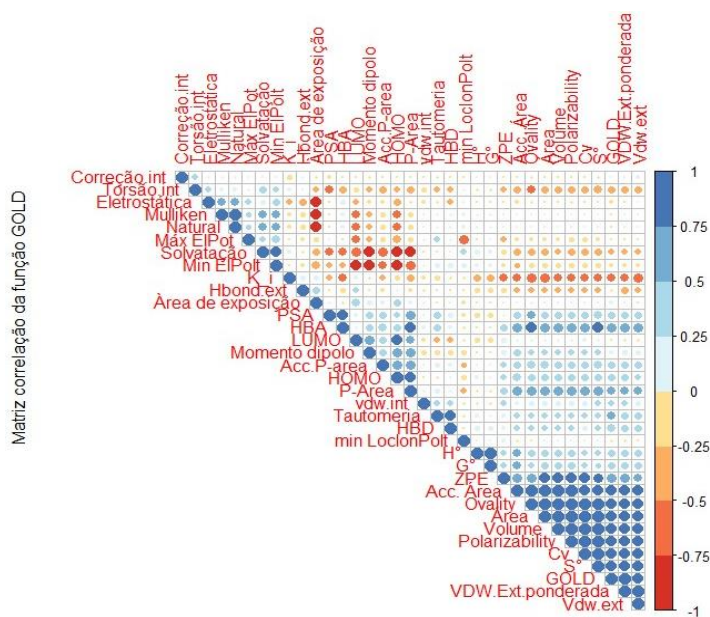


Figura 5.4 – Matriz correlação do conjunto de dados com a função GoldScore.

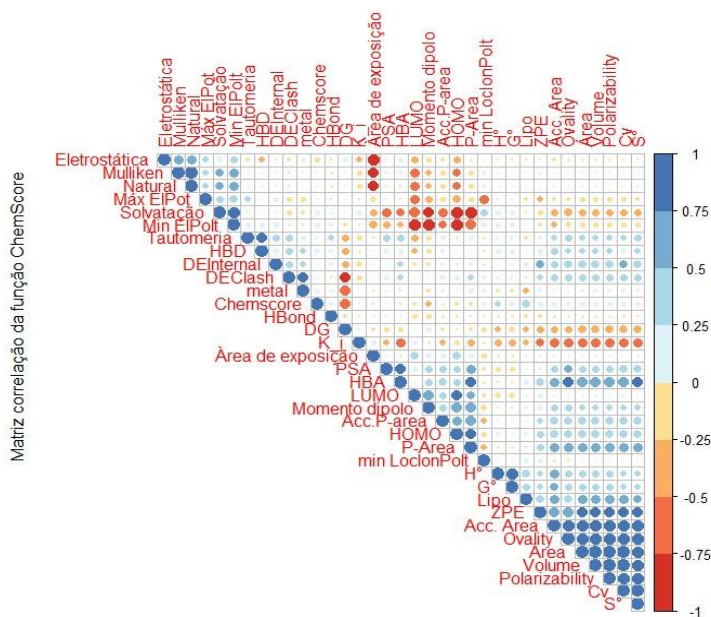


Figura 5.5 – Matriz correlação do conjunto de dados com a função ChemScore.

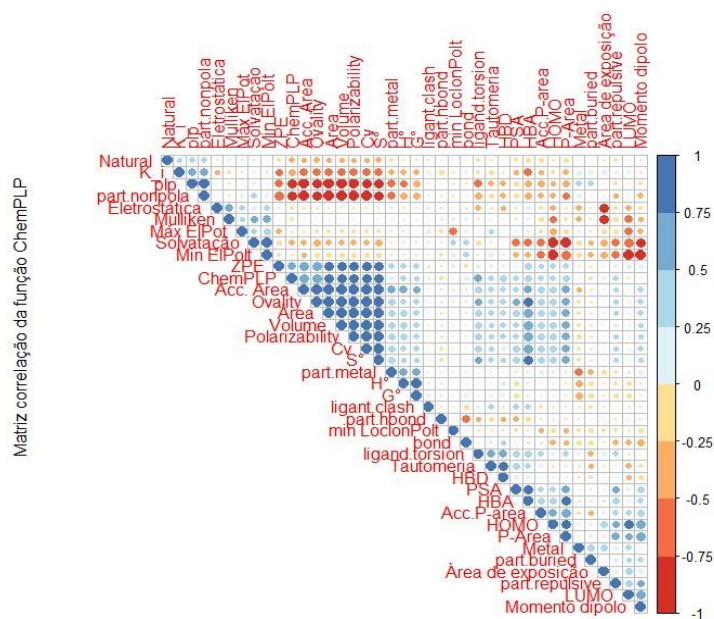


Figura 5.6 – Matriz correlação do conjunto de dados com a função ChemPLP.

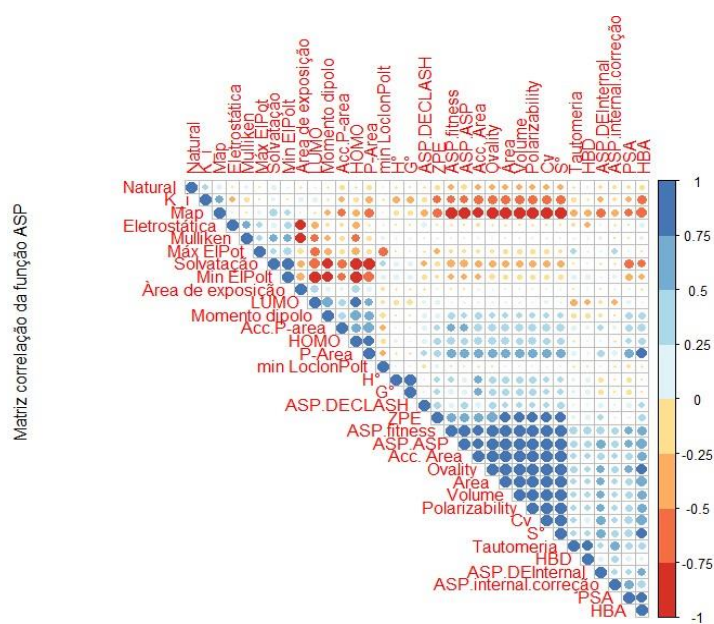


Figura 5.7 – Matriz correlação do conjunto de dados com a função ASP.

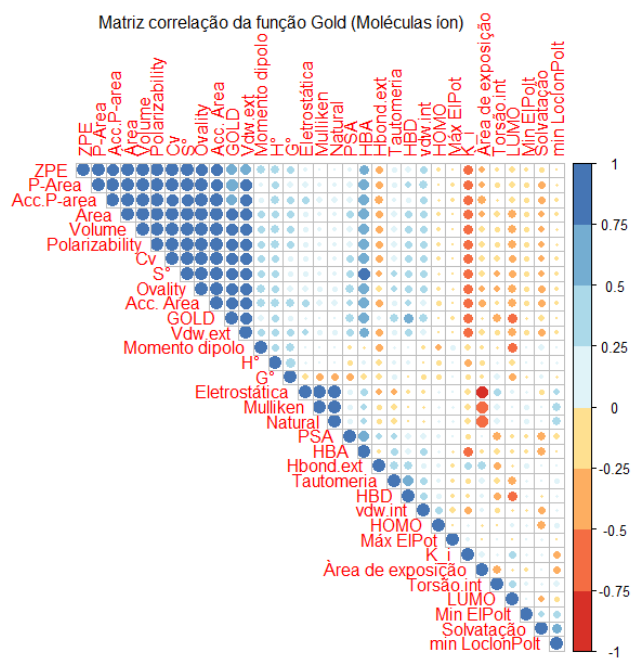


Figura 5.8 – Matriz correlação do conjunto de dados com a função GoldScore para os compostos ionizados.

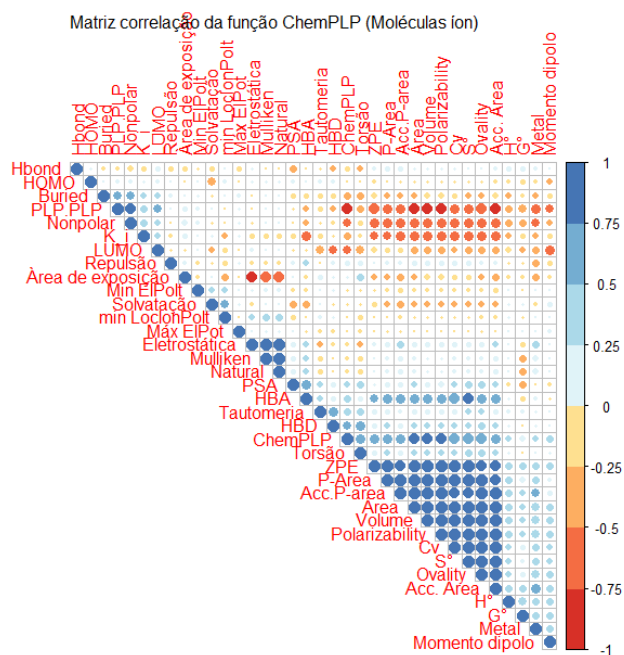


Figura 5.9 – Matriz correlação do conjunto de dados com a função ChemPLP para os compostos ionizados.

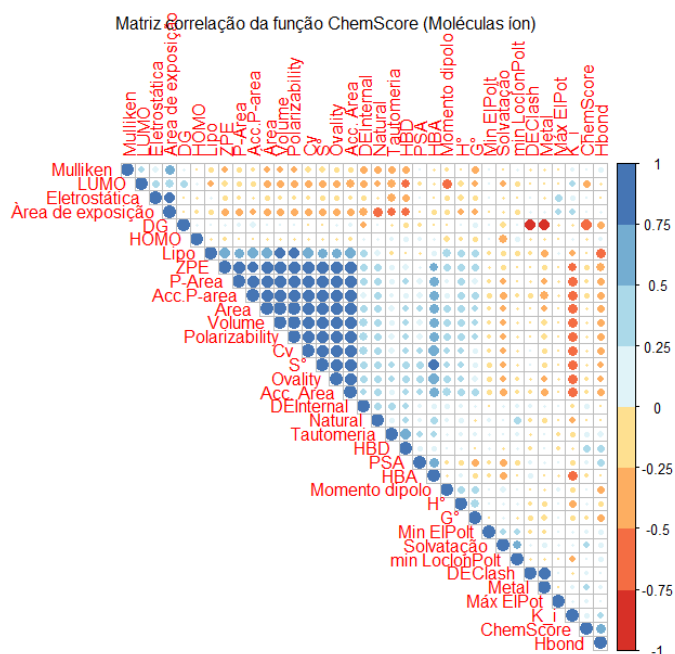


Figura 5.10 – Matriz correlação do conjunto de dados com a função ChemScore para os compostos ionizados.

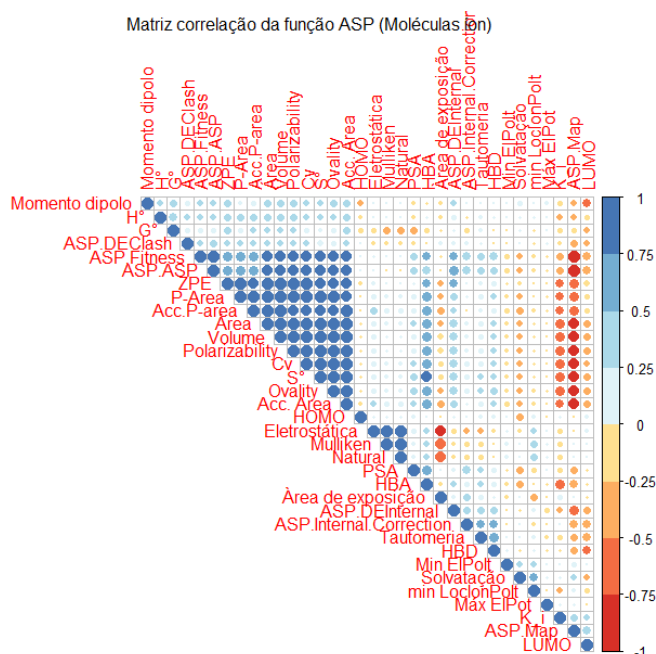


Figura 5.11 – Matriz correlação do conjunto de dados com a função ASP para os compostos ionizados.

Ao avaliar os gráficos de correlação a escala de cores mostram que algumas variáveis são pouco correlacionadas e outras que tem correlação próximo a 1. A técnica de componentes principais é adequada quando há multicolinearidade porém, as variáveis pouco correlacionadas podem reduzir a eficácia do modelo. Dada a eliminação dessas variáveis o conjunto da dados estava apropriado para a técnica de PCA.

5.3 Análise de componentes principais e análise de cluster

A técnica combinada de PCA e análise de clusters oferecem mais vantagens do que as técnicas separadas. A PCA reduz a dimensionalidade, reduz os ruídos e elimina as correlações e a análise de agrupamento separa em grupos de acordo com as características dos conjuntos e após a aplicação da primeira técnica os grupos conseguem ter uma melhor estimativa (TRIPATHI, 2025).

5.3.1 Análise de componentes principais

Nesta etapa foi calculado o desvio padrão, proporção da variância, autovalores, exposto nas tabelas 4.10 a 4.17, e gerado o gráfico cotovelo. Estas informações foram usadas para definir a quantidade de componentes principais. Pelo critério de Kaiser foi considerado componentes com autovalores maiores que 1 e com a variância total maior ou igual 80% e pelo gráfico cotovelo o ponto de inflexão mostrado no gráfico.

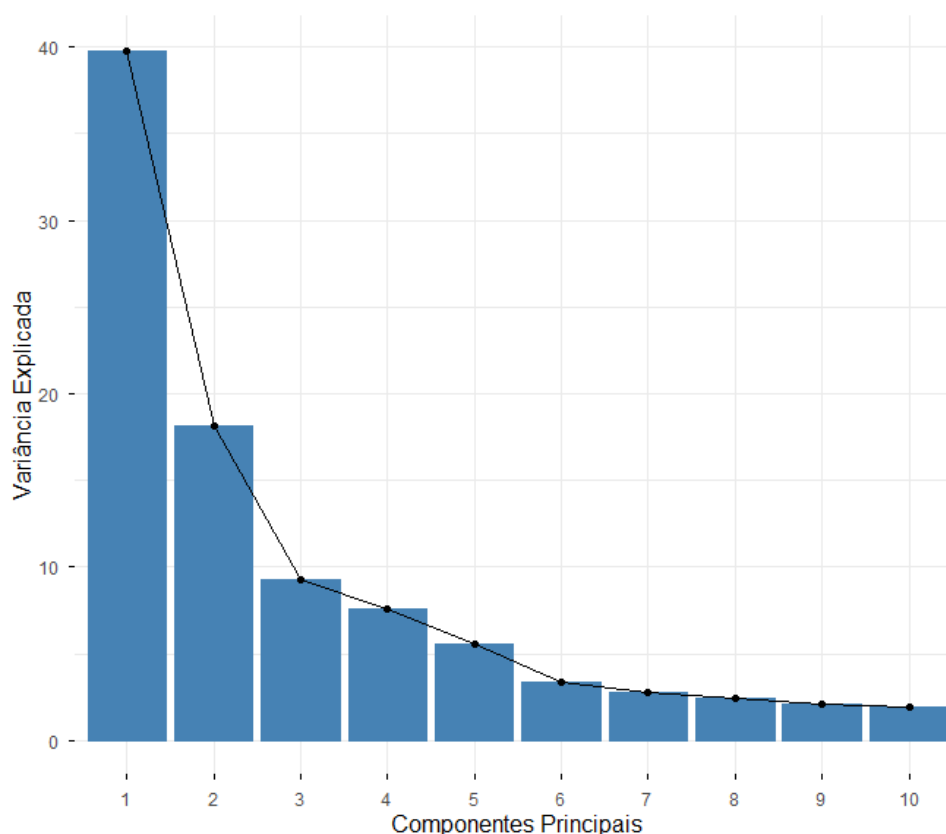


Figura 5.12 – Gráfico cotovelo função ChemPLP.

Tabela 5.10 - PCA da função GoldScore.

Componente principal	Desvio padrão	Proporção da Variância	Proporção acumulada
PC1	3,6214	0,4371	0,4371
PC2	2,3808	0,1890	0,6261
PC3	1,6151	0,0869	0,7130
PC4	1,5067	0,0757	0,7887
PC5	1,2874	0,0552	0,8440
PC6	1,0181	0,0346	0,8785
PC7	0,9056	0,0273	0,9058

Onde os autovalores são 13,11418; 5,6683; 2,6085; 2,2701; 1,6573; 1,0364; 0,8200 respectivamente.

Para a função GoldScore foram analisados os 5 primeiros componentes principais, pois a variância acumulada em PC6 foi de 84,4% e o autovalor foi 1,6573, confirmado pelo gráfico cotovelo.

Tabela 5.11 - PCA da função ChemPLP.

Componente principal	Desvio padrão	Proporção da Variância	Proporção acumulada
PC1	3,6240	0,3980	0,3980
PC2	2,4483	0,1816	0,5796
PC3	1,75051	0,09286	0,67250
PC4	1,58521	0,07615	0,74864
PC5	1,35019	0,05524	0,80389
PC6	1,05337	0,03362	0,83751

Os autovalores são 13,133; 5,994; 3,064; 2,512; 1,823; 1,109; 0,9.188, respectivamente. Para a função ChemPLP foram 5 PCs, dado a variância de acumulada em 80,3% e autovalor igual a 1,109.

Tabela 5.12 - PCA da função ChemScore.

Componente principal	Desvio padrão	Proporção da Variância	Proporção acumulada
PC1	3,3661	0,3541	0,3541
PC2	2,4137	0,1821	0,5362
PC3	1,7066	0,0910	0,6272
PC4	1,6881	0,0891	0,7162
PC5	1,4250	0,0635	0,7797
PC6	1,1765	0,0432	0,8229

Os autovalores são 11,330; 5,825; 2,912; 2,849; 2,030; 1,384; 1,255; 1,153; 0,633. Para a função ChemScore foram 6 PCs, com a variância acumulada em 82,2% e autovalor igual a 1,384.

Tabela 5.13 - PCA da função ASP.

Componente principal	Desvio padrão	Proporção da Variância	Proporção acumulada
PC1	3,6536	0,43060	0,4306
PC2	2,3692	0,18110	0,6117
PC3	1,71829	0,09524	0,70693
PC4	1,52833	0,07535	0,78227
PC5	1,32860	0,05694	0,83921
PC6	1,10520	0,03940	0,8786

Os autovalores são 13,348; 5,613; 2,952; 2,335; 1,765; 1,221; 0,828; 0,606. Para a função ASP foram escolhidos os 5 primeiros componentes principais com variância acumulada em 83,9% e autovalor 1,765 no PC5.

Tabela 5.14 - PCA da função GoldScore com moléculas ionizadas.

Componente principal	Desvio padrão	Proporção da Variância	Proporção acumulada
PC1	3,4708	0,5238	0,5238
PC2	1,7476	0,1328	0,6565
PC3	1,5025	0,0981	0,7546
PC4	1,3572	0,0800	0,8347

Os autovalores são 12,046; 3,054; 2,257; 1,842; 1,039; 0,757; 0,434; 0,334. Foram escolhidos os 4 primeiros PCs com 83,4% da variância e com autovalor de 1,842 no PC4.

Tabela 5.15 - PCA da função ChemPLP com moléculas ionizadas.

Componente principal	Desvio padrão	Proporção da Variância	Proporção acumulada
PC1	3,5749	0,4733	0,4733
PC2	1,9024	0,1340	0,6074
PC3	1,6183	0,0970	0,7044
PC4	1,4126	0,0739	0,7783
PC5	1,1336	0,0476	0,8259

Os autovalores são 12,780; 3,619; 2,618; 1,995; 1,285; 1,039; 0,854. Foram selecionados 5 PCs, com variância acumulada em 82,5% e com autovalor de 1,285.

Tabela 5.16: PCA da função ChemScore com moléculas ionizadas.

Componente principal	Desvio padrão	Proporção da Variância	Proporção acumulada
PC1	3,3888	0,4253	0,4253
PC2	1,9463	0,1403	0,5656
PC3	1,6367	0,0992	0,6648
PC4	1,5226	0,0859	0,7507
PC5	1,2791	0,0606	0,8113

Os autovalores são 11,483; 3,788; 2,678; 2,318; 1,636; 1,236; 0,9385. Escolheu-se os 5 primeiros PCs com variância acumulada em 81,1% e autovalor de 1,636 no quinto PC.

Tabela 5.17: PCA da função ASP moléculas com moléculas ionizadas.

Componente principal	Desvio padrão	Proporção da Variância	Proporção acumulada
PC1	3,6168	0,5031	0,5031
PC2	1,9474	0,1459	0,6490
PC3	1,5151	0,0883	0,7373
PC4	1,4181	0,0774	0,8146
PC5	1,0497	0,0424	0,8570

Os autovalores são 11,483; 3,788; 2,678; 2,318; 1,636; 1,236; 0,9385. Escolheu-se os 5 primeiros PCs com variância acumulada em 81,1% e autovalor de 1,636 no quinto PC. 13,081; 3,792; 2,295; 2,011; 1,101; 0,8301; 0,718; 0,580. O conjunto de dados que contém a função ASP foi considerado os 4 primeiros PCs como autovalor de 2,011 em PC4.

O próximo passo foi identificar as variáveis com maiores contribuições no PC, podemos observar nos gráficos da PC as variáveis que possuem os maiores autovetores da matriz correlação (carga ou coeficiente de ponderação), foram selecionados as que apresentavam um coeficiente de ponderação maior ou igual a

média dos coeficientes (linha pontilhada vermelha no gráfico). Como exemplo, a figura 4.9 mostra a contribuição das variáveis do primeiro componente principal da função GoldScore em porcentagem representada pela barra azul.

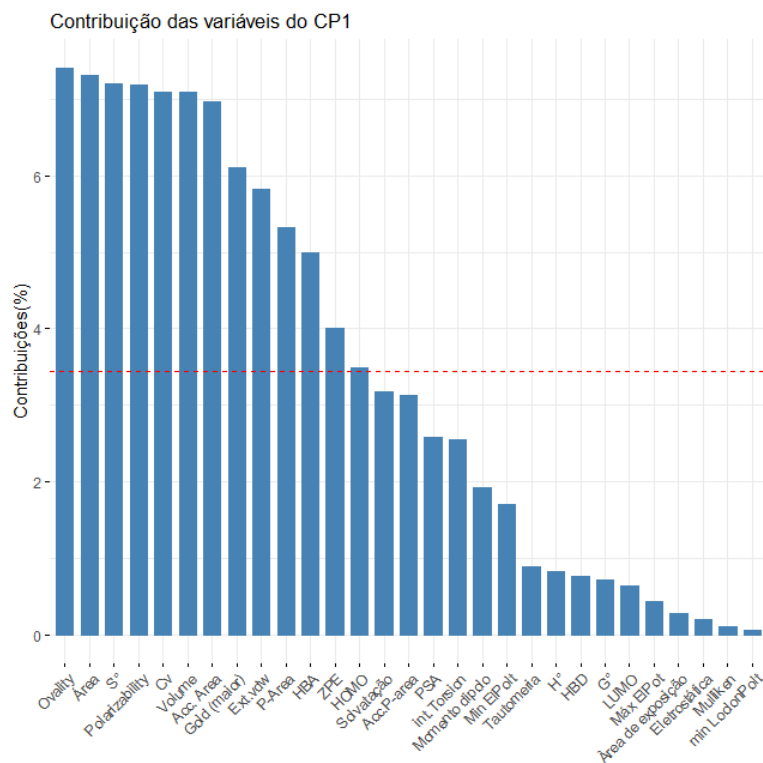


Figura 5.13 – Contribuição das variáveis no primeiro componente principal da função GoldScore.

As variáveis de contribuição acima da linha vermelha (média) foram selecionadas para fazer parte do componente principal e esses dados foram submetidos a análise de agrupamento.

5.3.2 Análise de Clusters (Agrupamento)

Nesse tópico é exposto que componentes principais que obtiveram uma boa separação na análise de cluster, ou seja, os clusters separaram as moléculas mais ativas das menos ativas por intermédio do conjunto de dados do PCA de acordo com suas características.

A análise de *cluster* foi feita com cinco componentes principais tanto no método de k-médias e PAM. Definimos por separar em 2 *clusters*, pois queríamos que os dados fossem agrupados em moléculas pouco ativas e moléculas muito ativas, o critério Silhueta confirmou que esse seria o número ótimo de grupos. Esses

dois métodos foram comparados. A tabela 4.5 mostra o percentual de acerto, nas informações em que há um traço mostra um resultado inconclusivo, pois ficaram 59 moléculas em um grupo e uma em outro. Os resultados abaixo de 60% mostraram grupos mal definidos. Para definir os acertos, bastou olhar os grupos e comparar o K_i de cada molécula dentro do grupo e perceber se eles estavam se dividindo referente o valor da constante de inibição.

Tabela 5.18 – Acertos no agrupamento de cada componente principal.

PC	GoldScore		ChemPLP		ChemScore		ASP	
	KM	PAM	KM	PAM	KM	PAM	KM	PAM
PC1	86,6%	86,6%	85%	85%	-	40%	86,6%	85%
PC2	21,6%	31,6%	33,3%	33,3%	31,6%	31,6%	31,6%	31,6%
PC3	45,0%	58,3%	60%	53,3%	-	41,6%	68,3%	68,3%
PC4	-	51,6%	51,6%	51,6%	55,0%	48,3%	25,0%	43,3%
PC5	-	48,3%	60,0%	56,6%	50%	46,6%	26,6%	31,6%

Tabela 5.19 – Acertos no agrupamento de cada componente principal das moléculas íons.

PC	GoldScore		ChemPLP		ChemScore		ASP	
	KM	PAM	KM	PAM	KM	PAM	KM	PAM
PC1	85,7%	85,7%	84,1%	84,1%	79,6%	77,7%	84,1%	84,1%
PC2	34,9%	34,9%	61,9%	58,7%	46,6%	38,0%	39,6%	50,7%
PC3	31,7%	41,2%	85,7%	79,4%	50,7%	46,0%	41,2%	36,5%
PC4	61,9%	46,0%	60,3%	52,3%	60,1%	58,7%	30,5%	61,9%
PC5	-	-	56,6%	76,1%	39,6%	39,6%	-	-

As tabelas 5.18 e 5.19 mostram que os dois métodos deram resultados parecidos na maior parte do estudo. Para a função ChemScore que na docagem não pareceu promissora no primeiro momento foi encontrado um componente principal após o tratamento de dados. As variáveis selecionadas mostraram uma frequente aparição nos componentes principais mostrando que elas são importantes na relação com a inibição, como a área, polarizabilidade, ovalidade, a própria pontuação mostra características que podem influenciar o resultado.

As equações das componentes principais são representadas por:

Goldscore:

$$PC_1 = 0.265.x_1 + 0.263.x_2 + 0.261.x_3 + 0.257.x_4 + 0.256.x_5 + 0.x_6 + 0.25.x_7 + \\ + 0.238.x_8 + 0.243.x_9 + 0.253.x_{10} + 0.218.x_{11} + 0.215.x_{12} + 0.194.x_{13}.$$

Onde, x_1 = Ovalidade, x_2 = Área, x_3 = Polarizabilidade, x_4 = S0, x_5 = volume, x_6 = CV, x_7 = Acc. Área, x_8 = ext.vdw, x_9 = Pontuação GoldScore, x_{10} = vdw.ponderada, x_{11} = p-área, x_{12} = HBA, x_{13} = ZPE.

ChemPLP:

$$PC_1 = 0.265.x_1 + 0.263.x_2 + 0.263.x_3 + 0.262.x_4 + 0.261.x_5 + 0.26.x_6 + 0.259.x_7 + \\ - 0.24.x_8 - 0.231.x_9 + 0.219.x_{10} + 0.213.x_{11} + 0.206.x_{12} + 0.197.x_{13}.$$

Onde, x_1 = Área, x_2 = Ovalidade, x_3 = Polarizabilidade, x_4 = Volume, x_5 = S0, x_6 = Cv, x_7 = Acc.Área, x_8 = apolar, x_9 = PLP, x_{10} = ChemPLP, x_{11} = P-Área, x_{12} = HBA, x_{13} = ZPE.

ASP:

$$PC_1 = 0.261.x_1 + 0.26.x_2 + 0.259.x_3 + 0.257.x_4 + 0.261.x_5 + 0.26.x_6 + 0.254.x_7 + \\ - 0.254.x_8 + 0.252.x_9 + 0.25.x_{10} + 0.221.x_{11} + 0.218.x_{12} + 0.188.x_{13}.$$

Onde, x_1 = Ovalidade, x_2 = S0, x_3 = Área, x_4 = Cv, x_5 = Polarizabilidade, x_6 = Volume, x_7 = ASP.ASP, x_8 = ASP.Map, x_9 = Pontuação ASP, x_{10} = Acc.área, x_{11} = P-Área, x_{12} = HBA, x_{13} = ZPE.

As equações das componentes principais dos conjuntos de moléculas ionizadas são representadas por:

Goldscore:

$$PC_1 = 0.285.x_1 + 0.282.x_2 + 0.282.x_3 + 0.281.x_4 + 0.280.x_5 + 0.279.x_6 + 0.275.x_7 + \\ + 0.268.x_8 + 0.263.x_9 + 0.258.x_{10} + 0.248.x_{11} + 0.244.x_{12}.$$

Onde, x_1 = Área, x_2 = Polarizabilidade, x_3 = Volume, x_4 = Ovalidade, x_5 = s°, x_6 = CV, x_7 = Acc. Área, x_8 = P-Área, x_9 = Acc.P-área, x_{10} = ZPE, x_{11} = VDW.ext, x_{12} = Pontuação GoldScore.

ChemPLP:

$$PC_1 = 0.275.x_1 + 0.274.x_2 + 0.273.x_3 + 0.271.x_4 + 0.269.x_5 + 0.268.x_6 + 0.265.x_7 + \\ -0.258.x_8 + 0.252.x_9 + 0.265.x_{10} - 0.233.x_{11} + 0.220.x_{12} - 0.195.x_{13}.$$

Onde, x_1 = Área, x_2 = Polarizabilidade, x_3 = Volume, x_4 = Ovalidade, x_5 = S° , x_6 = CV, x_7 = Acc.Área, x_8 = P-área, x_9 = ZPE, x_{10} = Acc.área, x_{11} = PLP.PLP, x_{12} = ChemPLP, x_{13} = apolaridade.

$$PC_3 = 0.447.x_1 - 0.425.x_2 - 0.377.x_3 + 0.404.x_3 - 0.287.x_5 - 0.253.x_6 - 0.210.x_7.$$

Onde, x_1 = Eletrostática, x_2 = Área de exposição, x_3 = Mulliken, x_4 = Torsão, x_5 = Tautomeria, x_6 = HBD, x_7 = ChemPLP.

ChemScore:

$$PC_1 = 0.2175.x_1 + 0.2933.x_2 + 0.2915.x_3 + 0.2855.x_4 + 0.2855.x_4 + 0.2768.x_5 + \\ + 0.2695.x_6 + 0.2919.x_7 + 0.1971.x_8 + 0.2718.x_9 + 0.2865.x_{10} + 0.2866.x_{11}.$$

Onde, x_1 = lipo, x_2 = Área, x_3 = volume, x_4 = Ovalidade, x_5 = Acc.área, x_6 = P-área, x_7 = polarizabilidade, x_8 = HB, x_9 = ZPE, x_{10} = CV, x_{11} = S° .

ASP:

$$PC_1 = 0.271.x_1 + 0.269.x_2 + 0.268.x_3 + 0.268.x_4 + 0.267.x_5 + 0.268.x_6 + 0.258.x_7 \\ - 0.253.x_8 + 0.244.x_9 + 0.244.x_{10} + 0.258.x_{11} + 0.243.x_{12} - 0.243.x_{13}.$$

Onde, x_1 = Área, x_2 = Polarizabilidade, x_3 = Volume, x_4 = S° , x_5 = Ovalidade, x_6 = CV, x_7 = Acc.área, x_8 = P-área, x_9 = Pontuação ASP, x_{10} = ZPE, x_{11} = Acc.Área, x_{12} = ASP.ASP, x_{13} = ASP.Map.

Podemos notar que a função ChemScore se mostrou pouco promissora nas análises, com exceção dos PCs das moléculas ionizadas. Ao se analisar o PC_1 vemos que não foi incluída nenhuma variável dos complexos, mas houve uma boa separação na clusterização e nos componentes principais. As demais funções resultaram em bons PCs, que atingiram o objetivo da análise. Através desse estudo vimos que algumas variáveis como área, volume, ovalidade, volume entre outros são frequentes na maior parte dos PCs, mostrando que essas variáveis parecem ser relevantes na inibição da enzima por essas moléculas.

Retomando o objetivo de encontrar o conjunto de variáveis que descrevesse bem o conjunto de modo que estivesse relacionado com a variável dependente usamos o método de clusterização a fim de verificar quais PCs classificariam as moléculas em muito ativas e pouco ativas. Usando K-médias e PAM percebemos que alguns PCs apresentaram uma boa separação.

Conclusões

Neste trabalho foram combinados métodos de modelagem molecular, análise de componentes principais e análise de clusters para encontrar um conjunto de dados de baixa dimensionalidade que esteja associado à classificação qualitativa da atividade inibitória da anidrase carbônica β da *M. tuberculosis* de 60 inibidores da classe das sulfonamidas selecionados da literatura.

Inicialmente, foi preciso usar o procedimento de modelagem comparativa para se obter um modelo 3D da MtCA2 em uma conformação de sítio ativo mais aberta em comparação com a estrutura cristalográfica, de forma a permitir o estudo de docagem molecular. O estudo de docagem molecular mostrou que as funções GoldScore, ChemPLP e ASP do programa GOLD produziram pontuações para as poses de interação dos compostos com a enzima que qualitativamente puderam separar os inibidores em muito ativos e pouco ativos em comparação com os dados experimentais de K_i sobre a enzima.

Foram preparadas matrizes de dados contendo, além dos dados experimentais de atividade das sulfonamidas, os dados de descritores de variáveis relacionadas às moléculas, obtidos por cálculos quânticos semiempíricos, e de descritores relacionados à interação receptor-ligante, obtidos pela docagem molecular.

A quantidade de descritores encontrados estava entre 34 e 40 dependendo da função de pontuação de docagem escolhida, o que resultou em uma matriz de dados com grande quantidade de variáveis. A técnica de componentes principais se mostrou eficaz na redução da dimensionalidade pois encontrou conjuntos de variáveis em menor quantidade, mas sem perda de informação. Em seguida, a análise de clusters mostrou quais componentes principais geravam subgrupos de descritores que estavam bem relacionados com a classificação da inibição em muito ativas ou pouco ativas, chegando a uma separação com 85% ou mais de acerto

quando os descritores moleculares eram combinados com os das funções de pontuação GoldScore, ChemPLP e ASP.

Os resultados deste estudo confirmaram que a modelagem molecular combinada com as técnicas de componentes principais e de análise de clusters podem ser úteis para classificação de moléculas como inibidores da enzima estudada. Através dos métodos estatísticos conseguiu-se reduzir a dimensionalidade de modo que não houvesse muita perda de informação do conjunto e também a análise de clusters nos informou bons conjuntos para classificar K_i . Isso sugere, portanto, que o procedimento pode ser explorado em estudos de triagem virtual de compostos do grupo das sulfonamidas para se encontrar candidatos a inibidores da anidrase carbônica β da *M. tuberculosis*.

Referências bibliográficas

1. ABDI, Hervé; WILLIAMS, Lynne J. **Principal component analysis**. Wiley interdisciplinary reviews: computational statistics, v. 2, n. 4, p. 433-459, 2010.
2. ABDOLI, M.; BONARDI, A.; PAOLETTI, N.; ASPATWAR, A.; PARKKILA, S.; GRATTERI, P.; SUPURAN, C. T.; ŽALUBOVSKIS, R. **Inhibition studies on human and mycobacterial carbonic anhydrases with N-((4-sulfamoylphenyl)carbamothioyl) amides**. Molecules, v. 28, n. 10, p. 4020, 2023. DOI: 10.3390/molecules28104020.
3. ALLAIRE, J. **RStudio: integrated development environment for R**. Boston, MA, v. 770, n. 394, p. 165-171, 2012.
4. ANDRICOPULO, Adriano D.; GUIDO, Rafael VC; OLIVA, Glaucius. **Virtual Screening and its Integration with Modern Drug Design Technologies**. Current Medicinal Chemistry, v. 15, n. 1, p. 37-46, 2008.
5. ARBEX, Marcos A.; VARELLA, Marília de C. L.; SIQUEIRA, Hélio R.; MELLO, Fernando Augusto Fiúza de. **Drogas Antituberculose: Interações Medicamentosas, Efeitos Adversos e Utilização em Situações Especiais - Parte 1: Fármacos de Primeira Linha**. Jornal Brasileiro de Pneumologia, v. 36, p. 626-640, 2010.
6. ARBEX, Marcos A.; VARELLA, Marília de C. L.; SIQUEIRA, Hélio R.; MELLO, Fernando Augusto Fiúza de. **Drogas Antituberculose: Interações Medicamentosas, Efeitos Adversos e Utilização em Situações Especiais - Parte 2: Fármacos de Segunda Linha**. Jornal Brasileiro de Pneumologia, v. 36, p. 641-656, 2010.
7. ASPATWAR, Ashok et al. **Mycobacterium tuberculosis β -carbonic anhydrases: Novel targets for developing antituberculosis drugs**. International Journal of Molecular Sciences, v. 20, n. 20, p. 5153, 2019.
8. ASPATWAR, Ashok; HAMMAREN, Milka; PARIKKA, Matalena; PARKKILA, Seppo; CARTA, Fabrizio; BOZDAG, Murat; VULLO, Daniela; SUPURAN, Claudiu T. **In Vitro Inhibition of Mycobacterium Tuberculosis β -Carbonic Anhydrase 3 with Mono- and Dithiocarbamates and Evaluation of Their Toxicity Using Zebrafish Developing Embryos**. Journal of Enzyme Inhibition and Medicinal Chemistry, v. 35, n. 1, p. 65-71, 2020.
9. AZEVEDO, Nathalia de et al. **Desenvolvimento de modelos de predição de atividade inibitória sobre a DNA Girase de micobactérias baseados em estudos de modelagem molecular**. Dissertação (Mestrado em modelagem matemática computacional), UFRRJ. Rio de Janeiro, 2021.

10. BANDEIRA, Soraya de O. et al. **Uso de Regressão Linear Múltipla para Previsão da Atividade Inibitória da Enzima N-miristoiltransferase de Leishmania donovani**. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, p. 1-8, 2025.
11. BARREIRO, Eliezer J.; FRAGA, Carlos Alberto Manssour *et al.* **Química Medicinal: As Bases Moleculares da Ação dos Fármacos**. Artmed Editora, 2014.
12. BARREIRO, E. J. *et al.* **Modelagem molecular: uma ferramenta para o planejamento racional de fármacos em química medicinal**. Química Nova, v. 20, n. 1, p. 1-12, 1997.
13. BAXTER, C. A.; MURRAY, C. W.; CLARK, D. E.; WESTHEAD, D. R.; ELDRIDGE, M. D. **Flexible docking using Tabu search and an empirical estimate of binding affinity**. Proteins: Structure, Function, and Genetics v. 33, n. 3, p. 367–382, 1998.
14. BHOLOWALIA, P.; KUMAR, A. **EBK-Means: A clustering technique based on Elbow Method and K-Means in WSN**. *International Journal of Computer Applications*, v. 105, n. 9, p. 17-24, 2014.
15. BRASIL. Ministério da Saúde. **Boletim Epidemiológico – Tuberculose 2024**. Disponível em: <<https://www.gov.br/aids/pt-br/central-de-conteudo/boletimsepidemiologicos/2024/boletimepidemiologico-tuberculose-2024/view>>. Acesso em: 27 mar. 2024.
16. BRASIL. Ministério da Saúde. **Secretaria de Vigilância em Saúde. Cartilha para o Agente Comunitário de Saúde: tuberculose**. Brasília, DF: MS, 2017.
17. BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde, Departamento de Doenças de Condições Crônicas e Infecções Sexualmente Transmissíveis. **Brasil Livre da Tuberculose: Plano Nacional pelo Fim da Tuberculose como Problema de Saúde Pública: Estratégias para 2021-2025**. Brasília: Ministério da Saúde, 2021.
18. BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância das Doenças Transmissíveis. **Brasil Livre da Tuberculose: Plano Nacional pelo Fim da Tuberculose como Problema de Saúde Pública**. Brasília: Ministério da Saúde, 2017.
19. BRASIL. Ministério da Saúde do Brasil. **Nota Técnica sobre as Mudanças no Tratamento da Tuberculose no Brasil para Adultos e Adolescentes – Versão 2**. 2022.
20. BRUNTON, Laurence L.; CHABNER, Bruce A.; KNOLLMANN, Björn C. (Orgs.).
21. **As bases farmacológicas da terapêutica de Goodman & Gilman**. Tradução de Augusto Langeloh... et al. Revisão técnica: Almir Lourenço da Fonseca. 12. ed. Porto Alegre: AMGH, 2017. Dados eletrônicos.
22. CARTA, Fabrizio; MARESCA, Alfonso; COVARRUBIAS, Adrian S.; MOWBRAY, Sherry L. et al. **Carbonic Anhydrase Inhibitors: Characterization and Inhibition Studies of the Most Active β -Carbonic Anhydrase from Mycobacterium Tuberculosis, rv3588c**. Bioorganic & Medicinal Chemistry Letters, v. 19, n. 23, p. 6649-6654, 2009.
23. CARVALHO, Ivone et al. **Introdução a modelagem molecular de fármacos no curso experimental de química farmacêutica**. Química Nova, v. 26, n. 3, p. 428-438, 2003.

24. COELHO, Augusto Leite. **Química Inorgânica Descritiva**. Fortaleza: EdUECE, 2015. 136 p.
25. COATES, R. F. W. **The Advanced Sonar Course: Further Topics in Underwater Acoustics**. 1ª Edition, Seiche, 2001.
26. COHEN, J. **Statistical power analysis for the behavioral sciences**. 2. ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
27. COVARRUBIAS, A. S. et al. **Structure and Function of Carbonic Anhydrases from Mycobacterium Tuberculosis**. Journal of Biological Chemistry, v. 280, n. 19, p. 18782-18789, 2005.
28. COVARRUBIAS, A. S.; BERGFORS, T.; JONES, T. A.; HÖGBOM, M. **Structural mechanics of the pH-dependent activity of beta-carbonic anhydrase from Mycobacterium tuberculosis**. Journal of Biological Chemistry, v. 281, n. 8, p. 4993-4999, 2006. DOI:10.1074/jbc.M510756200.
29. **Crystal Structure of M. Tuberculosis Beta Carbonic Anhydrase, rv3588c, Tetrameric Form**. Disponível em: <<https://www.rcsb.org/structure/2a5v>>. Acesso em: 12 jan. 2024.
30. DE AZEVEDO, Nathalia et al. **Desenvolvimento de um Modelo Preditivo para a Atividade Inibitória de Tiazolopiridinil-ureias sobre a Subunidade B da DNA Girase de Micobactérias por Meio de Descritores Moleculares**. Revista Virtual de Química, v. 14, n. 6, 2022.
31. ELDRIDGE, M. D.; MURRAY, C. W.; AUTON, T. R.; PAOLINI, G. V.; MEE, R. P. **Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes**. Journal of Computer-Aided Molecular Design, v. 11, p. 425–445, 1997.
32. FIGUEIREDO FILHO, Dalson Britto; SILVA JÚNIOR, José Alexandre. **Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r)**. Revista Política Hoje, v. 18, n. 1, p. 115-146, 2009.
33. FIGUEIREDO FILHO, D. B. **Análise multivariada para os cursos de Administração, Ciências Contábeis e Economia**. São Paulo: Atlas, 2009.
34. GUIDO, R. V. C. et al. **Planejamento de Fármacos, Biotecnologia e Química Medicinal: Aplicações em Doenças Infecciosas**. Estudos Avançados, v. 24, n. 24, p. 81-98, 2010.
35. GUIDO, Rafael VC; ANDRICOPULO, Adriano D.; OLIVA, Glaucius. **Planejamento de Fármacos, Biotecnologia e Química Medicinal: Aplicações em Doenças Infecciosas**. Estudos Avançados, v. 24, p. 81-98, 2010.
36. HONGYU, Kuang; SANDANIELO, Vera Lúcia Martins; DE OLIVEIRA JUNIOR, Gilmar Jorge. **Análise de componentes principais: resumo teórico, aplicação e interpretação**. E&S Engineering and science, v. 5, n. 1, p. 83-90, 2016.
37. INMETRO. **Quadro Geral de Unidades de Medida: Resolução CONMETRO nº 12/88**. 4. ed. Rio de Janeiro: Ed. SENAI, 2007.
38. IZIBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. São Carlos, SP: Rafael Izibicki, 2020. 1. ed.
39. JAMES, Gareth et al. **An introduction to statistical learning**. New York: springer, 2013.

40. JOHNSON, Richard A.; WICHERN, Dean W. **Applied multivariate statistical analysis**. 6. ed. Upper Saddle River: Pearson Prentice Hall, 2007.
41. KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding groups in data: an introduction to cluster analysis**. John Wiley & Sons, 2009.
42. KIRTON, S. B.; MURRAY, C. W.; VERDONK, M. L.; TAYLOR, R. D. **Prediction of binding modes for ligands in the cytochromes P450 and other heme-containing proteins**. *Proteins: Structure, Function, and Bioinformatics*, v. 58, p. 836-844, 2005.
43. LATTIN, J.; CARROLL, J. D.; GREEN, P. E. **Análise de dados multivariados**. São Paulo: Cengage Learning, 2011.
44. LOMBARDINO, Joseph G.; LOWE III, John A. **The Role of the Medicinal Chemist in Drug Discovery—Then and Now**. *Nature Reviews Drug Discovery*, v. 3, n. 10, p. 853-862, 2004.
45. LU, T. Multiwfn. **Software manual**. Version, v. 3, n. 6, 2014.
46. MARQUES, Paulo R. B.; OLIVEIRA, Yamanaka, Hideko. **Biossensores Baseados no Processo de Inibição Enzimática**. *Química Nova*, v. 31, p. 1791-1799, 2008.
47. MARESCA, Alfonso; CARTA, Fabrizio; VULLO, Daniela; SUPURAN, Claudiu T. **Dithiocarbamates Strongly Inhibit the -Class Carbonic Anhydrases from Mycobacterium Tuberculosis**. *Journal of Enzyme Inhibition and Medicinal Chemistry*, v. 28, n. 2, p. 407-411, 2013.
48. MARESCA, Alfonso; VULLO, Daniela; SCOZZAFAVA, Andrea; MANOLE, Gheorghe; SUPURAN, Claudiu T. **Inhibition of the -Class Carbonic Anhydrases from Mycobacterium Tuberculosis with Carboxylic Acids**. *Journal of Enzyme Inhibition and Medicinal Chemistry*, v. 28, n. 2, p. 392-396, 2013.
49. MÉDICOS SEM FRONTEIRAS. **Tuberculose. 2022**. Disponível em: <<https://www.msf.org.br/o-que-fazemos/atividades-medicas/tuberculose/>>. Acesso em: 03 abr. 2024.
50. MELO, Giane Lavarda; PETRY, Mirta Teresinha; MARTINS, Juliano Dalcin; VILLA, Bruna de. **Análise de componentes principais e contribuições das variáveis e estações meteorológicas na determinação da evapotranspiração de referência no Estado do Rio Grande do Sul**. *Revista Brasileira de Climatologia*, v. 37, n. 21, p. 474–492, 2025.
51. MOREIRA, C., **Enzima**. *Revista Ciência Elementar*, v. 3, n.3, p.162, 2015.
52. MOOIJ, Wijnand.T; VERDONK, Marcel L.. **General and targeted statistical potentials for protein-ligand interactions**. *Proteins*. v.61, n.2, p.272-287, 2005.
53. MUNZLINGER, Elizabete; NARCISO, Fabricio B.; DE QUEIROZ, José Eustáquio R. **Sistematização de revisões bibliográficas em pesquisas da área de IHC**. Sociedade Brasileira de Computação, 2012.
54. ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Global Tuberculosis Report 2023. Genebra: World Health Organization, 2023**. Licença: CC BY-NC-SA 3.0 IGO.
55. ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Global tuberculosis report 2025**. Geneva: World Health Organization; 2025. Licence: CC BY-NC-SA 3.0 IGO.
56. PAIVA, Fernando Fernandes; BARBATO, Daniela Maria Lemos; PAIVA, Mirella Lopez Martini Fernandes; JOÃO, Herbert Alexandre; MUNIZ, Sérgio Ricardo.

- Orientações motivacionais de alunos do ensino médio para física: considerações psicométricas.** Revista Brasileira de Ensino de Física, v. 40, n. 3, p. e3404.1 – e3204.9, 2018.
57. PALMA, Luann Farias. **Agrupamento de dados: k-médias.** Trabalho acadêmico (Bacharelado em Ciências Exatas e Tecnológicas) — Centro de Ciências Exatas e Tecnológicas, Universidade Federal do Recôncavo da Bahia, Cruz das Almas, 2018.
 58. POSIT TEAM. **RStudio: Integrated Development Environment for R.** Boston, MA: Posit Software, PBC, 2025. Disponível em: <http://www.posit.co/>
 59. SANTANNA, Carlos M. et al. **Métodos de Modelagem Molecular para Estudo e Planejamento de Compostos Bioativos: Uma Introdução.** Revista Virtual de Química, v. 1, n. 1, p. 49-57, 2009.
 60. STEWART, J.J. **Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements.** J. Mol. Model. 2007 Dec;13(12):1173-213. doi: 10.1007/s00894-007-0233-4. Epub 2007 Sep 9.
 61. SANTOS, Lucianna Helene; PASLAUSKI, Gustavo; TEIXEIRA, Gabriel de Paula (Org.). **Bioinformática na Prática: Desvendando a Biologia com Códigos.** BioInfo, 2022. DOI: 10.51780/978-6-599-275326-09.
 62. STEWART, J.J. **Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements.** J. Mol. Model. 2007 Dec; v.13, n. 12, p.1173-213, 2007. doi: 10.1007/s00894-007-0233-4.
 63. SILVESTRE, António. **Análise de Dados e Estatística Descritiva.** Escolar Editora, 2007.
 64. SUPURAN, Claudiu T.; SUPURAN, Ghiorgi. **Inhibition Studies on Human and Mycobacterial Carbonic Anhydrases with Sulfamic Acid Derivatives.** Journal of Medicinal Chemistry, v. 63, n. 7, p. 2338–2344, 2021.
 65. VULLO, Daniela; CAU, Ylenia; MORI, Mattia; SUPURAN, Claudiu T. **Inhibition of Mycobacterial Carbonic Anhydrases with Hydroxycarboxylic Acids and Their Derivatives.** Journal of Enzyme Inhibition and Medicinal Chemistry, v. 33, n.
 66. WEHRENS, Ron. **Chemometrics with R: Multivariate Data Analysis in the Natural and Life Sciences.** 1. ed. Cham: Springer, 2011.

Análise de componentes principais

Bibliotecas usadas

```
library(readxl) #Ler arquivos do excel  
library(corrplot) #visualização da matriz de correlação  
library(RColorBrewer)  
library(stargazer) #pacote para converter tabela em latex  
library(ggplot2)  
library(ggcorrplot)  
library(tinytex)  
library(factoextra)
```

Funções usadas:

Leitura de variáveis em excel:

```
dados_red <- read_xlsx("CHEMPLP.xlsx", sheet = 3)  
dados1 = dados_red[,3:35]
```

Verificando correlação

```
matcor = cor(dados1)  
print(matcor, digits = 3)
```

Padronização das variáveis

```
dados_padro = scale(dados1)  
round(apply(dados_padro,2,var),4)
```

Verificando correlação (dados padronizados)

```
matvar = var(dados_padro)
print(matvar, digits = 3)
matcor = cor(dados1)
print(matcor, digits = 3)
```

Autovalores e desvio padrão

```
eigen(matcor)$values
acpcor = prcomp(dados1, scale = TRUE)
acpcor
acpcor$sdev
acpcor$sdev^2
summary(acpcor)
```

Gráfico cotovelo

```
plot(1:ncol(dados1), acpcor$sdev^2, type = "b", xlab = "Número de componentes",
ylab = "Autovalores", pch = 20, cex.axis = 1.3, cex.lab = 1.3)
```

```
library(factoextra) # Screeplot 2
```

```
library(ggplot2)
```

```
x11();fviz_eig(acpcor)+
```

```
labs(
```

```
  title = "Gráfico de Scree",          # Título do gráfico
```

```
  x = "Componentes Principais",        # Título do eixo X
```

```
  y = "Variância Explicada"            # Título do eixo Y
```

```
) +
```

```
theme(
```

```
  plot.title = element_text(hjust = 0.5), # Centraliza o título
```

```
  axis.title.x = element_text(size = 12), # Ajusta o tamanho da fonte do eixo X
```

```
  axis.title.y = element_text(size = 12)  # Ajusta o tamanho da fonte do eixo Y
```

```
)
```

Contribuição das componentes por variável

```
x11()
fviz_contrib(acpcor,choice = "var", axes = 1)+
  labs(
    title = "Contribuição das variáveis do CP1",
    y = "Contribuições(%)"
  )
```

Análise de cluster

Variância

```
V <- sapply(dados1, var)
V
stargazer(V,
  title="Variância",
  align=T,
  type = "latex")
```

```
apply(dados1, 2, var)
```

Média

```
Med <- apply(dados1, 2, mean)
Med
stargazer(Med,
  title="Média",
  align=T,
  type = "latex")
```

```
x11();boxplot(Med)
```

Matriz correlação e padronização dos dados

Número de Clusters

```
fviz_nbclust(dados_padro, kmeans, method = "silhouette") +  
  ggtitle("Número ideal de clusters - Método da Silhueta") +  
  xlab("Número de clusters (k)") +  
  ylab("Valor médio da silhueta")
```

Cluster PAM

```
classif = cbind(dados_padro, cluster = pam_res$cluster)  
classif
```

Cluster K-médias

```
classif = cbind(dados_padro, cluster = km.res$cluster)  
classif
```