UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

VICTOR HENRIQUE RABESQUINE NOGUEIRA

Descoberta de novos inibidores para doenças infecciosas: estudos integrados de modelagem molecular e aprendizado de máquina para a malária e o COVID-19

São Carlos
2024

VICTOR HENRIQUE RABESQUINE NOGUEIRA

Descoberta de novos inibidores para doenças infecciosas: estudos integrados de modelagem molecular e aprendizado de máquina para a malária e o covid-19

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

Área de concentração: Física Biomolecular
Orientador: Prof. Dr. Rafael Victório de Carvalho Guido
Coorientador: Dr. Alexandre Victor Fassio

Versão corrigida
(versão original disponível na Unidade que aloja o Programa)

São Carlos
2024

Nogueira, Victor Henrique Rabesquine
Descoberta de novos inibidores para doenças infecciosas: estudos integrados de modelagem molecular e aprendizado de máquina para a malária e o COVID-19 / Victor Henrique Rabesquine Nogueira; orientador Rafael Victório de Carvalho Guido; co-orientador Alexandre Victor Fassio - versão corrigida -- São Carlos, 2024.
176 p.

Tese (Doutorado - Programa de Pós-Graduação em Física Biomolecular) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2024.

1. Aprendizado de máquina. 2. Modelos generativos. 3. Malária. 4. Modelagem molecular. 5. COVID-19. I. Guido, Rafael Victório de Carvalho, orient. II. Fassio, Alexandre Victor, co-orient. III. Título.

À minha família, com muito amor e gratidão
pelo apoio incondicional durante toda minha jornada.

AGRADECIMENTOS

Primeiramente, gostaria de agradecer imensamente a minha família, não tenho palavras para descrever o quanto sou grato por estarem comigo em todos os momentos, dando todo suporte, amor e carinho que precisei durante todos esses anos, sempre acreditando em mim e não me deixando desistir. Amo muito vocês!

Ao Prof. Dr. Rafael Guido, por ter aceitado ser meu orientador, por ter promovido discussões enriquecedoras e por ter tido toda a paciência e compreensão ao longo desse processo. Obrigado por ser humano.

Gostaria de agradecer também ao meu coorientador, Dr. Alexandre Fassio, pela amizade, por ter expandido meus horizontes e compartilhando seus conhecimentos sobre aprendizado de máquinas e sobre a vida.

A todo pessoal do lab, em especial ao grupo da malária (Camila B., Camila R., Carol, Gabi, Giovana, Igor, Mariana, Sarah e Vinícius), obrigado pelas músicas, pelas experiências trocadas, pelos bolos, pelas risadas, pelo compartilhamento da veia jornalística e pelos momentos mais divertidos, tornando esse período tão desafiador em algo mais leve e suportável. Hopi Hoers, vocês estarão sempre no meu coração!

A todo o Keiser Lab, por terem me acolhido me feito sentir parte do grupo, em especial ao Prof. Dr. Michael Keiser, por ter me recebido em seu laboratório e me guiado no mundo do aprendizado de máquinas, e ao Rish, o melhor mentor que alguém poderia ter e um amigo para a vida toda.

A todo o grupo de brasileiros que encontrei na UCSF. Vocês não apenas me acolheram, mas também me proporcionaram o calor e o apoio de uma verdadeira família, mesmo estando tão longe de casa. Quero expressar minha gratidão especial à Jordana, cuja presença iluminou meu caminho e fez com que minha estadia fosse não apenas suportável, mas também memorável (além de ser a melhor anfitriã desse mundo todo!); ao André, por estar sempre disposto a compartilhar seus conhecimentos, sempre pronto para todo e qualquer passeio e gosto impecável por jogos e filmes; à Juliana, por todas as aventuras e viagens compartilhadas e por cozinhar tão bem; à Daniela, pela companhia e por todas as idas ao student's food market; e à Rebeca, por ser a pessoa mais alto astral e realista de todas, além de uma ótima organizadora de eventos.

Ao Murilo, o melhor e mais leal dos amigos que alguém poderia sonhar em ter. Obrigado por estar sempre ao meu lado, independentemente do momento, obrigado por todo suporte, por toda compreensão, por todos as vezes que não deixou a peteca cair, pela gastronomia impecável e pelo melhor senso de humor de todos.

Agradeço à CAPES, pelo financiamento no Brasil e no exterior. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, processos de número 88887.357974/2019-00 e 88887.695330/2022-00.

Agradeço também ao Serviço de Pós-Graduação, por todo o auxílio prestado ao longo desses anos, e às bibliotecárias do IFSC, por serem super atenciosas e sempre dispostas a ajudar, em especial à Neusa, por seu trabalho impecável e sua simpatia contagiante.

“Picture a wave in the ocean. You can see it, measure it
– its height, the way the sunlight refracts when it passes through
– and it's there, and you can see it, and you know what it is: it's a wave.
And then it crashes on the shore and it's gone. But the water is still there.
The wave was just a different way for the water to be for a little while.
The wave returns to the ocean, where it came from, where it's supposed to be.”

Chidi Anagonye – The Good Place

Temporada 4, episódio 13.

RESUMO

NOGUEIRA, V. H. R. **Descoberta de novos inibidores para doenças infecciosas:** estudos integrados de modelagem molecular e aprendizado de máquina para a malária e o COVID-19. 2023. 176 p. Tese (Doutorado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2024.

Os processos de descoberta e desenvolvimento de fármacos são extremamente complexos, podendo levar mais de 15 anos desde a identificação do alvo até a comercialização de um medicamento com mais de $1 bi investidos ao longo desses anos. Os métodos computacionais voltados para descoberta e desenvolvimento de fármacos têm auxiliado a reduzir tanto os custos quanto o tempo de tal processo, além de maximizarem a chance de sucesso. Tais métodos têm sido essenciais para a coleta, o pré-processamento, a análise e a inferência de dados e têm desempenhado um importante papel para a descoberta de novas entidades químicas (NCE, do inglês, *new chemical entities*) nas últimas décadas. A busca por NCEs por métodos computacionais está associada à utilização de representações moleculares que podem influenciar diretamente na qualidade dos resultados. Neste trabalho métodos computacionais foram utilizados para a avaliação de representações moleculares e para a descoberta de novos potenciais inibidores para doenças infecciosas. Portanto, esta tese foi dividida em três partes: i) avaliação da robustez de uma representação molecular utilizando autocodificadores variacionais (VAE, do inglês, *variational autoencoder*); ii) geração inibidores do parasito causador da malária usando VAE; e iii) busca por inibidores da M^{pro} de SARS-CoV-2, combinando abordagens computacionais e ensaios experimentais. I) À medida que esforços para melhorar a robustez das representações moleculares avançam, avança a importância de métodos rigorosos para testá-las e validá-las. Usando um VAE para gerar amostras anômalas de uma representação molecular em cadeia conhecido como SELFIES, sua suposição fundamental – serem sempre válidas quando convertidas em outra representação em cadeia (*e.g.*, SMILES) – foi verificada como não verdadeira neste trabalho. Os resultados obtidos mostraram que regiões específicas no espaço latente do VAE, explorado de forma contínua e radial, foram particularmente eficazes na geração de SELFIES que desafiaram a suposição fundamental. A organização da validade no espaço latente ajuda a entender melhor os fatores que afetam a confiabilidade da representação molecular. Portanto, neste trabalho propomos o VAE e a abordagem de geração de anomalias associada como uma ferramenta eficaz para avaliar a robustez de representações moleculares. II) A malária é uma das doenças tropical infecciosas com a maior taxa de mortalidade no mundo. A OMS estima que 619 mil mortes tenham ocorrido em decorrência da doença em 2021, sendo cerca de 76 % desse total mortes de crianças com menos de cinco anos de idade. O tratamento recomendado para a doença é feito com terapia combinada com derivados de artemisinina. Entretanto, o surgimento de cepas resistentes do parasito aos tratamentos disponíveis torna necessário a descoberta de novos antimaláricos. Neste

trabalho é apresentado um modelo generativo utilizando VAE, treinado com um conjunto de 4 milhões de moléculas, utilizando SELFIES como representação molecular, e refinado com conjuntos distintos para a geração de novos inibidores do parasita. O modelo generativo apresentou parâmetros de qualidade satisfatórios que permitiram a proposição de mais de 100 mil moléculas como candidatas a inibidores do parasita. Diante disso, aplicamos diversos filtros classificatórios para a seleção de um subconjunto de 20 moléculas para a aquisição e/ou síntese e validação experimental. III) A pandemia de COVID-19 fez com que o mundo mudasse drasticamente. O SARS-CoV-2, vírus responsável por causar a doença, já matou aproximadamente 7 milhões de pessoas ao redor do mundo, sendo mais de 705 mil no Brasil (dados de outubro/2023). Embora as vacinas e os tratamentos tenham diminuído consideravelmente o impacto da pandemia, novas variantes virais continuam a surgir e fazem com que esforços para encontrar agentes terapêuticos sejam extremamente necessários. Nesse contexto, um estudo de triagem virtual de compostos naturais brasileiros para seleção e teste experimental de candidatos a inibidores da protease principal (M^{pro}) do vírus. Os 10 melhores *hits* virtuais foram submetidos a 100 ns de simulações de dinâmica molecular para verificação da estabilidade do modo de interação e cálculo de energia livre de ligação por MM-GBSA e metadinâmica. O composto mais promissor, taxifolina (NuBBE_139), foi avaliado experimentalmente *in vitro* contra a M^{pro} e mostrou valor de IC_{50} de 820 μM e um perfil de inibição competitivo. Em suma, embora contenha três partes distintas, esta tese focou na aplicação de métodos computacionais em química medicinal. Cada abordagem contribuiu sinergicamente para avanços na área ao utilizar diferentes ferramentas computacionais para alvos diferentes. Os autocodificadores variacionais geraram moléculas inéditas, com propriedades semelhantes a antimaláricos conhecidos, mas diferentes o bastante para possibilitar a proposição de novas classes de compostos com esse fim. A utilização dessa arquitetura também revelou falhas nos SELFIES, testando sua validade na conversão para SMILES e revelando, pela primeira vez, a metodologia como uma maneira eficiente de geração de anomalias representacionais. Além disso, métodos tradicionais, como triagem virtual e dinâmica molecular, permitiram a identificação de um inibidor competitivo da Mpro de SARS-CoV-2, com IC50 na faixa de submilimolar.

Palavras-chave: Aprendizado de máquina. Modelos Generativos. Malária. Modelagem molecular. COVID-19.

ABSTRACT

NOGUEIRA, V. H. R. **Discovery of new inhibitors for infectious diseases:** integrated molecular modeling and machine learning studies for malaria and COVID-19. 2023. 176 p. Ph. D. Thesis (Doctor in Sciences) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2024.

The drug discovery and development processes are extremely complex and can take more than 15 years from target identification to the commercialization of a drug, with more than $1 billion invested over those years. In this sense, computational methods aimed at drug discovery and development have helped to reduce both the costs and time of this process, as well as maximize the chance of success. These methods have been essential for data collection, pre-processing, analysis, and inference. Additionally, they have been crucial in discovering new chemical entities (NCE) in recent decades. The search for NCEs aided by computational methods is associated with using molecular representations that directly influence the quality of results. In this work, computational methods were used for the evaluation of molecular representations robustness and the discovery of potential new inhibitors for infectious diseases. Hence, this thesis was divided into three parts: i) assessment of the robustness of a molecular representation using variational autoencoders (VAE); ii) generating inhibitors for the malaria parasite using VAE; and iii) searching for inhibitors of SARS-CoV-2 M^{pro}, combining computational approaches and experimental assays. I) As efforts to improve the robustness of molecular representations progress, the importance of rigorous methods to test and validate them increases. Using a VAE to generate anomalous samples from a popular molecular representation called SELFIES, its fundamental assumption – always being valid when converted to another string representation (e.g., SMILES) – was tested. The results showed that specific regions in the latent space of the VAE, explored continuously and radially, are particularly effective in generating SELFIES that challenge this assumption. This organization of validity in the latent space helps better understand the factors affecting the reliability of the molecular representation. This work proposes VAE and the associated anomaly generation approach as an effective tool for assessing the robustness of molecular representations. II) Malaria is one of the infectious tropical diseases with the highest mortality rate globally. WHO estimates that 619,000 deaths occurred due to the disease in 2021, with about 76 % of this total being deaths of children under the age of five. The recommended treatment for the disease relies on artemisinin-derivatives combined therapy (ACT). However, the increasing resistance of the parasite to the available treatments makes the search for new antimalarials necessary. In this work, we built a generative model using variational autoencoder architecture, trained with a set of 4 million molecules using SELFIES as a molecular representation and refined with two different sets for generating new parasite inhibitors. The model showed satisfactory quality parameters that allow the generation of more than 100,000 molecules as inhibitor candidates.In light of that, we applied several filters to sort and select the 20 most promising compounds for synthesis and experimental validation. III) The COVID-19 pandemic has caused drastic changes worldwide. The SARS-CoV-2 virus, responsible for causing the disease, has already killed, approximately, 7 million people worldwide, with more than 705,000 in Brazil (data from October 2023). Although vaccines and treatments have considerably improved the pandemic's impact, new viral variants continue to emerge, making efforts

to find new therapeutic agents extremely needed. In this context, we conducted a virtual screening study of Brazilian natural compounds for the selection and experimental testing of inhibitor candidates for SARS-CoV-2 M^{pro} protease. The top 10 virtual hit compounds were subjected to 100 ns molecular dynamics simulations to check their binding mode and stability, and to compute the binding free energy by MM-GBSA and metadynamics studies. The most promising compound, taxifolin (NuBBE_139), was experimentally assessed *in vitro* against M^{pro} protease and showed an IC_{50} value of 820 μM and a competitive inhibition profile. In summary, although it contains three distinct parts, this thesis focuses on the application of computational methods in medicinal chemistry. Each approach contributed synergistically to advances in the field by using different computational tools for different targets. Variational autoencoders generated novel molecules, with properties similar to known antimalarials but sufficiently different to make it possible to propose new classes of compounds for this purpose. The use of this architecture also revealed flaws in SELFIES, testing its validity in the conversion to SMILES and revealing, for the first time, the methodology as an efficient way of generating representational anomalies. In addition, traditional methods such as virtual screening and molecular dynamics enabled the identification of a competitive inhibitor of SARS-CoV-2 Mpro, with an IC50 in the submillimolar range.

Keywords: Machine Learning. Generative models. Malaria. Molecular Modeling. COVID-19.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CADD	desenvolvimento de fármacos assistido por computadores, do inglês, *computer-aided drig design*
QSAR	relação quantitativa estrutura-atividade, do inglês, quantitative structure–activity relationship
ML	aprendizado de máquina, do inglês, *machine learning*
DL	aprendizado profundo, do inglês, *deep leaning*
ACT	terapia combinada de derivados de artemisinina
AE	autocodificador, do inglês, *autoencoder*
VAE	autocodificador variacional, do inglês, *variational autoencoder*
ELBO	evidence lower bound objective
KL	Kullback-Leibler
GAN	rede generativa adversária, do inglês, *generative adversarial network*
AAE	autocodificador adversário
LSTM	long-short term memory
ECFP	extended-connectivity fingerprints
SMILES	simplified molecular input line entry system
SELFIES	SELF-referencIng Embedded Strings
MMV	Medicines for Malaria Venture
GRU	gated recurrent unit
t-SNE	t-distributed stochastic neighbor embedding
PCA	análise de componentes principais, do inglês, *principal component analysis*
AMC	antimaláricos ChEMBL
QED	quantitative estimation of drug-likeness
NR	naive random
SR	shuffle random
ITDR	index-token distribution random matrix
SARS-Cov-2	síndrome respiratória aguda grave do coronavírus 2

COVID-19	doença do coronavírus de 2019, do inglês, *coronavirus disease*
M^{pro}	protease principal, do inglês, *main protease*
NuBBE	Núcleo de Bioensaios, Biossíntese e Ecofisiologia de Produtos Naturais
MD	dinâmica molecular, do inglês, *molecular dynamics*
ARE	elemento responsivo a antioxidante, do inglês, *antioxidant response element*

Sumário

PREFÁCIO

Os métodos computacionais desempenham um papel central em acelerar e aprimorar o processo de descoberta de fármacos. Esses métodos têm potencial de reduzir os custos e o tempo associados ao desenvolvimento de novos medicamentos bem como aumentar a probabilidade de descoberta de compostos com potencial terapêutico. A aplicação de simulações computacionais e a análises de dados permite a identificação de maneira eficiente de candidatos a fármacos. Além disso, facilita a compreensão das interações que os ligantes estabelecem com os aminoácidos do sítio de ligação no alvo biológico e ajuda a predizer suas propriedades farmacológicas. Por exemplo, métodos de docagem molecular permitem que sejam realizadas triagens virtuais de vastas coleções de compostos contra proteínas-alvo específicas, identificando potenciais candidatos a ligantes/inibidores cuja afinidade pelo alvo seja elevada.

Modelos como os de relação quantitativa estrutura-atividade (QSAR) ajudam a prever a atividade biológica de um novo composto tendo como base a sua estrutura química bem como ajudam a guiar o processo de planejamento de novas moléculas. Já os algoritmos de inteligência artificial e, especificamente, os de aprendizado de máquina permitem analisar dados biológicos, realizar diagnósticos de doenças de maneira bastante rápida e confiável, identificar padrões até então desconhecidos e transformá-los em informação útil, gerar estruturas inéditas, identificar conformações mais prováveis, predizer propriedades químicas e biológicas, incluindo efeitos adversos, dentre outras aplicações.

Nesta tese foram utilizados e aplicados diferentes métodos computacionais no estudo de moléculas e na busca de novos candidatos a inibidores. Esses resultados estão descritos e discutidos em três capítulos distintos. O primeiro capítulo aborda o uso de autocodificadores variacionais como arquitetura de um modelo generativo para moléculas cujo objetivo é inibir a atividade do parasito da malária. O segundo capítulo apresenta a aplicação dos autocodificadores variacionais na busca em hiper esferas do espaço latente para gerar anomalias representacionais e testar representações moleculares. Esses estudos foram desenvolvidos numa parceria entre a USP e pesquisadores da Universidade da Califórnia em São Francisco (UCSF). O terceiro capítulo combina docagem e dinâmica molecular com ensaios experimentais para a

descoberta de candidatos a inibidores da protease principal (M^{pro}) de SARS-CoV-2. Esse trabalho foi conduzido em parceria entre a USP e pesquisadores da Fundação Oswaldo Cruz (FIOCRUZ, Recife).

CAPÍTULO 1: INTRODUÇÃO

1 INTRODUÇÃO

O processo de descoberta de fármacos é extremamente complexo e lento, o qual envolve diversas etapas que devem ser seguidas a fim de potencializar a chance de sucesso.[1-2] Da identificação do alvo até a comercialização do fármaco, mais de 15 podem se passar. Além disso, a depender das características de cada alvo e objetivos escolhidos, mais de $ 1 bilhão podem ser investidos durante todo o processo (Figura 1).[1,3]

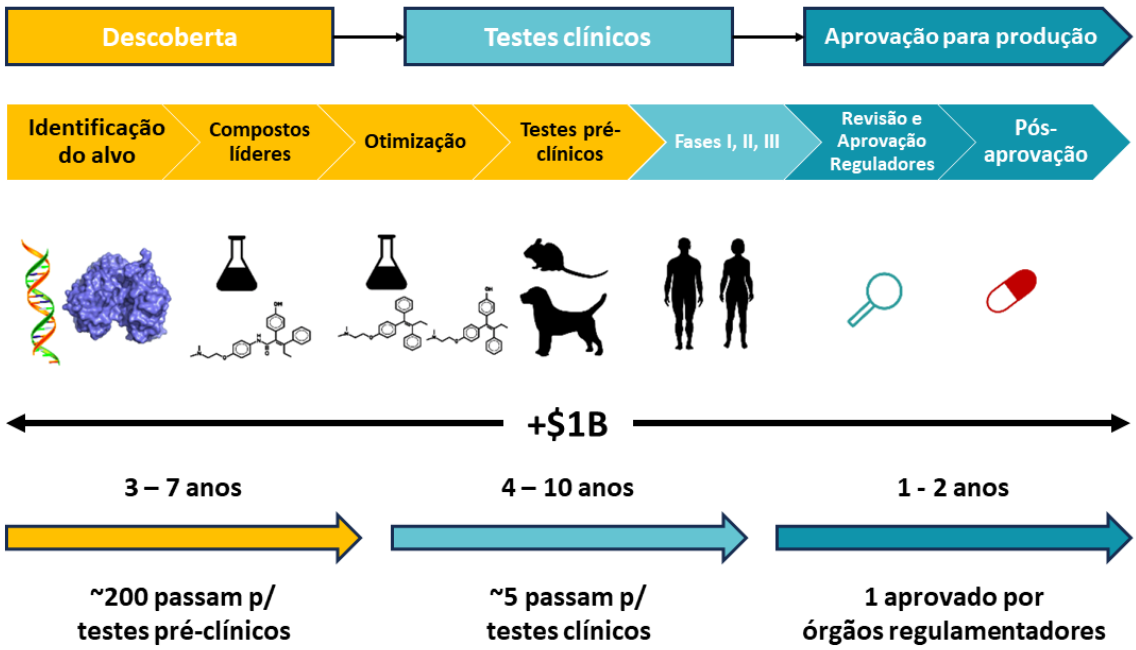


Figura 1 – Processo de descoberta e desenvolvimento de fármacos. O início do processo inclui etapas de identificação do alvo em questão, bem como a identificação de compostos líderes, suas otimizações e os ensaios pré-clínicos. Esses estágios duram, geralmente, de três a sete anos, com centenas de milhares de compostos sendo não classificados para as etapas futuras. Poucos candidatos a fármacos chegam às fases clínicas, as quais podem levar mais de 10 anos. O composto melhor sucedido segue então para os órgãos regulamentadores para serem aprovado e, posteriormente, comercializados. O monitoramento da atividade do fármaco continua mesmo após a sua comercialização. Durante todo o processo, mais de $ 1 bilhão podem ser investidos.
Fonte: Adaptada de ZHANG *et al*.[1]

Nos últimos anos, a utilização de metodologias de desenvolvimento de fármacos assistido por computadores (CADD, do inglês, *computer-aided drug design*) têm sido cada vez mais utilizadas.[3-4] Em especial, com o aumento do poder computacional e dos conjuntos de dados, a utilização de abordagens que se utilizam de aprendizado de máquina tem crescido exponencialmente.[5-6]

1.1 Aprendizado de Máquina

O aprendizado de máquina (*ML*, do inglês, *machine learning*), compreendido como uma subárea da inteligência artificial, possibilita que computadores tenham a capacidade de aprender, de forma autônoma, a reconhecer padrões a partir de um conjunto de dados inicial e que possam extrapolar este conhecimento aprendido para novos dados. Tais técnicas computacionais têm sido amplamente utilizadas na área de descoberta e desenvolvimento de fármacos.[7-12] Em especial, as redes neurais, ou o aprendizado profundo (*DL*, do inglês, *deep learning*), que é um tipo de aprendizado de máquina o qual utiliza diversas camadas internas para obter informações relevantes e aprender os padrões que estão representados pelos dados,[7] têm ganhado bastante destaque em temas como representação de moléculas,[13] processamento e análise de imagens,[14] predição de propriedades farmacológicas de moléculas,[9,15-16] interação molécula-alvo,[15] modelo de reação entre moléculas[17] e geração de novas moléculas.[18-22]

Os modelos de DL para geração de novas moléculas foram propostos como uma alternativa ao planejamento realizado por humanos, a fim de acelerar a descoberta de fármacos.[21] Em geral, esses modelos são capazes de gerar moléculas que seguem uma distribuição de características de um conjunto de treinamento, seja por meio da exploração do espaço químico disponível,[22] seja focando em alvos e propriedades específicas.[23] Além disso, os modelos generativos também têm sido utilizados como métodos alternativos para a otimização de propriedades biológicas e moleculares, focados principalmente, em mudanças nas propriedades físico-químicas das moléculas em estudo.[24-25]

Existem diversas abordagens para a geração de moléculas que se utilizam de diferentes arquiteturas de redes neurais, entre elas destaca-se os autocodificadores variacionais, uma arquitetura de modelo generativo utilizada na nesta tese.

1.2 Autocodificadores variacionais

O objetivo de um autocodificador (do inglês, *autoencoder* ou AE) é construir um espaço latente de dimensão reduzida que contém representações comprimidas, nas quais cada elemento pode ser reconstruído como sua entrada inicial. A parte responsável por mapear o dado de entrada, que tem alta dimensionalidade, para uma representação de baixa dimensionalidade é chamado de “codificador”, enquanto que a porção que é responsável por compreender o mapeamento e reconstruir a entrada original a partir dessa representação de dimensão reduzida é chamada de “decodificador” (Figura 2A).[26-27]

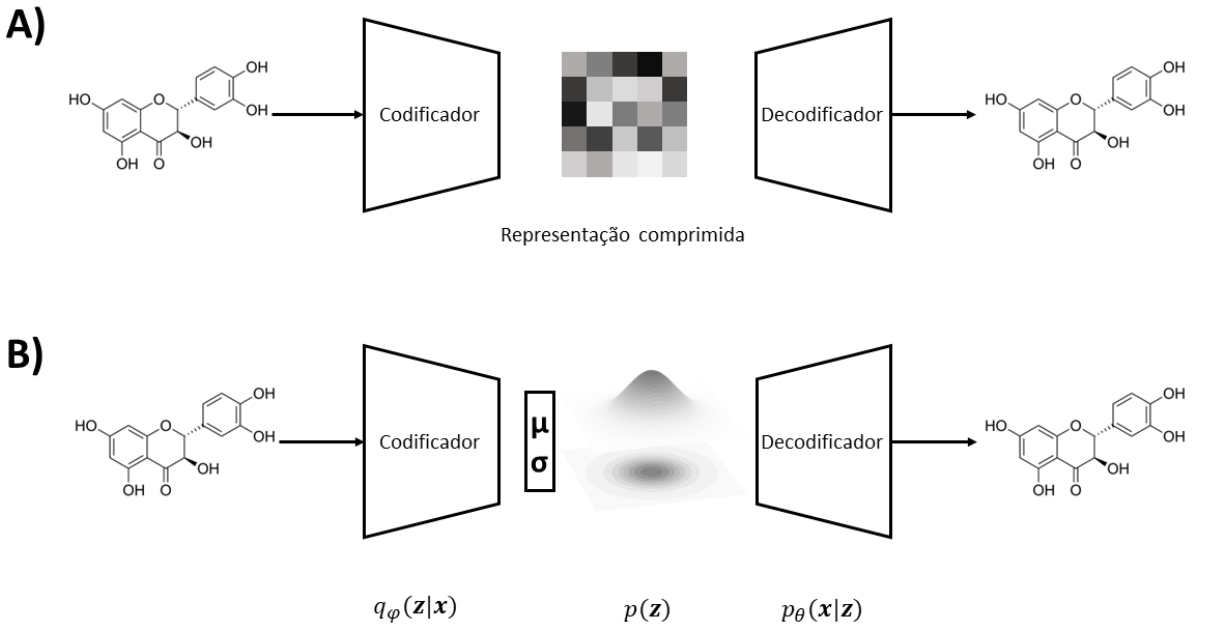


Figura 2 – Autocodificador e autocodificador variacional. A) Um autocodificador recebe uma molécula, codifica-a numa representação comprimida e a decodifica de volta. B) Um autocodificador variacional mapeia a molécula em parâmetros de uma distribuição estatística como o espaço latente, numa representação numérica contínua.
Fonte: Elaborada pelo autor.

No contexto de moléculas, os AE comprimem uma molécula x num código fixo no espaço latente z e tende a resumir as regras de mapeamento explicitas, uma vez que o número de parâmetros ajustáveis pode ser muito maior que o número de moléculas fornecidas para o treinamento. As representações criadas no espaço latente para os AE são geralmente determinísticas e discretas, de modo que essas regras explicitas fazem a decodificação de pontos aleatórios desse espaço latente

uma tarefa desafiadora e, algumas vezes, impossível. Uma alternativa para essa limitação é a utilização de métodos como os autocodificadores variacionais.

Os autocodificadores variacionais (do inglês, *variational autoencoders*, ou VAE) foram apresentados em 2013, por Kingma e Welling,[27-28] e ganharam popularidade rapidamente por serem uma ferramenta robusta para a criação de modelos generativos de textos, imagens e sons.[29-31] Em vez de mapear as entradas (que nesta tese são representadas pelas moléculas) num único ponto como fazem os AEs, os VAEs mapeiam os dados em distribuições de probabilidade condicionais para formar o espaço latente. Isso cria um gargalo de informação latente, onde a informação é expressa de forma mais difusa e incerta no espaço latente. Essa abordagem garante continuidade representacional e interpretação generativa.[32]

Por meio de propagação direta, o modelo pode ser decomposto em três partes (Figura 2B). A primeira é uma rede codificadora, a qual aproxima a distribuição *posterior* sobre a variável latente z condicionada ao dado de entrada x, $q_\varphi(z|x)$. Isso distingue, efetivamente, as regiões do espaço latente pela probabilidade de observação como uma função dos dados condicionados x. A segunda é uma distribuição *prior* não informativa, $p(z)$, que geralmente é a distribuição gaussiana multivariada padrão. A terceira parte é uma rede decodificadora, que produz uma distribuição de verossimilhança (*likelihood*) sobre os dados condicionada na *posterior* amostrada da variável latente z, $p_\theta(x|z)$, gerando, efetivamente, a reconstrução da entrada na amostragem. O treinamento do VAE otimiza em conjunto os parâmetros θ e φ do codificador e do decodificador, maximizando a *evidence lower bound objective* (*ELBO*) no logaritmo da verossimilhança da distribuição de dados, de modo a se ter o seguinte:

$$\begin{aligned} \mathcal{L}\,(x;\theta,\varphi) \;&=\; \mathbb{E}_{q_\varphi(z|x)}[log p_\theta(x|z)] \;-\; D_{KL}(q_\varphi(z|x)||p(z)) \\ &\leq\; log p_\theta(x) \end{aligned} \tag{1}$$

Cada um dos dois termos da equação (1) tem uma função específica no resultado que se obtém. O primeiro termo maximiza o logaritmo da probabilidade de se reconstruir o dado x condicionado à variável latente z, como esperado da distribuição *posterior*. Simultaneamente, o segundo termo minimiza a divergência de *Kullback-Leibler* (KL)[33] entre as distribuições *prior* e *posterior*, funcionando como um regularizador e que minimiza a variância ao penalizar o codificador quando ele se

aproxima das distribuições *posterior* que divergem da *prior*. Em outras palavras, o termo de reconstrução avalia se as amostras decodificadas correspondem à entrada, enquanto o termo de regularização investiga se o espaço latente sobreajusta aos dados de treinamento.

Quando treinado com textos como dados de entrada (ou cadeias de caracteres), os VAEs podem experimentar um problema de "desaparecimento" do KL.[34] Enquanto o VAE é feito para aprender e gerar dados como textos, por exemplo, usando ambos os contextos locais e características globais, ele tende, à medida que o treinamento avança, a depender apenas do contexto local, ignorando as características globais para a composição do resultado de saída nova. Para contornar esse problema, um parâmetro, β, que pondera o termo de regularização, pode ser adicionado à ELBO.

$$\begin{aligned} \mathcal{L}\,(x;\theta,\varphi) \;&=\; \mathbb{E}_{q_\varphi(z|x)}[log p_\theta(x|z)] \;-\; \beta D_{KL}(q_\varphi(z|x)||p(z)) \\ &\leq\; log p_\theta(x) \end{aligned} \tag{2}$$

De modo geral, os VAEs têm sido empregados para várias finalidades, abrangendo diversas áreas distintas. Por exemplo, em biologia estrutural, Sevgen *et al.* utilizaram uma abordagem de VAE, com transferência de aprendizado, para o planejamento de proteínas usando de incorporações latentes de dimensionalidade baixa e de decodificação generativa para projetarem as sequências condicionalmente.[35] Na agricultura, Wu e Xu propuseram um VAE "adversário" combinado com aprendizado residual a fim de gerar imagens de doenças de folhas de tomate, para mitigar a escassez de dados.[36] Na química, Tempke e Musho[37] demonstraram a geração de novas reações químicas por meio da amostragem do espaço latente de um VAE treinado com dados de reações em fase gasosa.

Além disso, diferentes representações moleculares também já foram exploradas para geração molecular e *design de novo*. Lee e Min[38] utilizaram matrizes gráficas como entradas para seus modelos, combinadas com um processo de otimização multiobjetivo para geração molecular. Jin *et al*.[39] empregaram o VAE de *junction-trees*, um processo que consiste de duas etapas, o qual cria uma estrutura baseada em árvore que representa as subestruturas químicas e, em seguida, as integra usando uma rede neural gráfica de passagem de mensagens. Hadipour *et al*.[40] combinaram a representação SMILES e a técnica de redução de dimensionalidade

PCA em um VAE para incorporar propriedades moleculares e agrupar grandes conjuntos de dados de moléculas pequenas.

Outras arquiteturas também têm sido utilizadas para a geração de dados em geral. As redes generativas adversárias (GANs, do inglês, *generative adversarial networks*) combinam duas redes neurais distintas: uma geradora e uma discriminadora.[41] O discriminador ajuda a diferenciar dados reais de dados falsos enquanto o gerador é responsável por criar dados falsos e tentar enganar o discriminador. Os autocodificadores adversários (AAE, do inglês, *adversarial auto encoder*) são basicamente um modelo híbrido entre os VAEs e as GANs, um codificador, um decodificador e um discriminador, o qual é responsável por distinguir as distribuições obtidas do espaço latente daquelas da distribuição *prior*.[42] Os modelos baseados no algoritmo de *long-short term memory* (LSTM) aprendem padrões em dados sequenciais e usam esse conhecimento para gerar novas sequências.[43] Suas camadas ocultas funcionam como uma memória e os dados novos são gerados a partir de sementes ou entradas sobre os quais o modelo prediz quais são os próximos elementos da sequência.

1.3 Representações moleculares

O uso de aprendizado de máquina para aplicações em química exige a conversão das estruturas químicas para um formato que seja adequado para a máquina interpretar e processar. Embora os nomes comuns das moléculas sejam fáceis de se lembrar, por exemplo, cafeína, eles carregam pouca ou nenhuma informação acerca da estrutura e das propriedades da molécula.[18] Nesse contexto, diversas notações diferentes têm sido desenvolvidas a fim de se obter representações moleculares adequadas e que possam retornar resultados razoáveis às diferentes aplicabilidades dos métodos de ML.

Em aprendizado de máquina existe uma concepção geral que o desempenho de um modelo está diretamente relacionado com a qualidade da representação dos dados utilizada, isto é, uma boa representação permite que a tarefa em questão seja aprendida mais facilmente.[13, 44] Essa mesma concepção aplica-se também às moléculas, uma vez que identificar características estruturais é importante para evidenciar atividades biológicas e relações entre as suas propriedades. Dessa forma,

uma boa representação molecular influencia diretamente na facilidade de se aprender uma tarefa.

Representar dados químicos de uma forma concisa e não ambígua, compreensível para humanos e computadores, não é uma tarefa fácil.[45] Isso se torna especialmente verdade quando se trata de representações de moléculas. Enquanto existe um número significativo de métodos que são adequados para representar moléculas orgânicas pequenas e simples, quando se considera moléculas que possuam anéis, ligações, valências, componentes inorgânicos ou simetrias, problemas relativos à sua complexidade começam a aparecer. Essas complexidades podem resultar em representações não canônicas (isto é, muitas formas de representação para a mesma molécula dentro de um mesmo método), não-únicas, conflitantes (muitas moléculas diferentes que são representadas da mesma forma), ou erradas (assumem um número errado de hidrogênios implícitos na molécula ou que falham em representar o estado tautomérico correto).

De acordo com Chuang *et al.*[13] para o aprendizado de tarefas que envolvam moléculas, as representações moleculares utilizadas devem ter quatro características essenciais:

i) **Expressiva**: as representações devem capturar a diversidade e variedade dos conjuntos de dados químicos e devem ser capazes de distinguir as diferenças sutis entre as moléculas;

ii) **Parcimônia**: as representações devem manter a expressividade sem eliminar informações críticas, o modelo deve ser capaz de aprender os padrões apesar de possíveis ruídos no conjunto de dados;

iii) **Invariante**: uma vez que o mesmo dado de entrada deve, consistentemente, gerar o mesmo dado de saída, representações moleculares devem ser invariantes a aspectos tais como numeração atômica ou conformações moleculares, bem como a rototranslação e a estrutura química. A invariância limita o espaço de funções que podem ser aprendidas, adequando-se melhor a um domínio de aplicação específico;

iv) **Interpretável**: as representações que podem ser mapeadas em uma forma estrutural permitem que especialistas possam avaliar os padrões aprendidos pelo modelo e confrontá-los com o conhecimento existente.

A seguir, serão apresentadas introduções sobre representações moleculares que foram diretamente utilizadas neste trabalho a fim de treinar os modelos de aprendizado de máquina empregados.

1.3.1 Grafos

A estrutura química de uma molécula pode ser facilmente representada numa superfície de duas dimensões, como num papel ou uma tela. Essa representação gráfica pode ser convenientemente convertida numa estrutura molecular em forma de grafo, que é a representação interpretável por máquinas mais comum.[46]

Um grafo é definido por *G=(V,E)*, onde *V* é um conjunto de nós (ou vértices) e *E* é um conjunto de arestas (do inglês, *edges*). Cada aresta pertencente ao conjunto de arestas *E* e conecta um par de nós de *V*. A representação de estruturas químicas de moléculas na forma de grafos baseia-se na ideia de mapear os átomos e as ligações que compões a molécula. Quando se pensa em estruturas de moléculas na forma de grafos, *V* é interpretado como o conjunto de todos os átomos dessa molécula e *E* é o conjunto de ligações entre esses átomos.

Fazer o mapeamento desse conceito abstrato para uma representação concreta com a qual um computador seja capaz de interpretar requer que os conjuntos de nós e arestas sejam estruturados de alguma maneira. Normalmente, esse processo é feito por meio de matrizes ou vetores que podem incluir informações sobre como os átomos estão conectados, as características desses átomos e as características das ligações químicas.[46] A Figura 3A ilustra a representação da estrutura molecular da fenilalanina.

1.3.2 Fingerprint

As assinaturas moleculares (do inglês, *molecular fingerprints*) são vetores binários que contêm elementos ordenados nos quais cada elemento mapeia propriedades estruturais e físico-químicas de uma molécula, denotando a presença ou ausência de determinada característica ou subestrutura.[13] Existem diferentes tipos de assinaturas moleculares,[47-51] mas devido principalmente a implementações em bibliotecas de Python de código aberto, os ECFPs [49] (do inglês, *extended-connectivity*

fingerprints) são as assinaturas moleculares mais amplamente utilizadas.[45] Essas assinaturas moleculares se baseiam no algoritmo de *Morgan*.[52]

Nos ECFP, os átomos que não são hidrogênios são codificados em várias camadas circulares até que um determinado diâmetro estabelecido seja alcançado. O primeiro passo para se obter um ECFP é usar informações acerca de cada átomo de uma molécula que resulta num descritor de átomo e seu entorno, que é invariável em função de como o átomo foi numerado na molécula. A Figura 3B ilustra um *fingerprint* genérico para a fenilalanina.

Inicialmente, são atribuídos valores inteiros aos átomos como identificadores. Esses identificadores iniciais são reunidos num conjunto inicial de assinaturas. Em seguida, cada átomo coleta seu próprio identificados e os identificadores dos átomos adjacentes num vetor (os vizinhos são ordenados por meio de seus próprios identificadores e a ordem das ligações). Uma função *hash* é então aplicada para reduzir o vetor em um novo e único identificador inteiro. Uma vez que todos os átomos tenham gerado seus próprios identificadores, eles substituem seus identificadores antigos por esses novos identificadores. Os novos identificadores são então adicionados ao conjunto de assinaturas. Essa iteração é repetida por um número de vezes pré-determinado. Quando esse número de iterações é atingido, identificadores duplicados são removidos do conjunto e os identificadores inteiros remanescentes no conjunto de assinaturas define uma assinatura ECFP.

Embora os ECFPs sejam amplamente utilizados para tarefas como QSAR, ao usar vetores de propriedades moleculares não há garantias de que eles serão reversíveis, ou seja, não é possível deduzir a estrutura da molécula a partir desse vetor.[45]

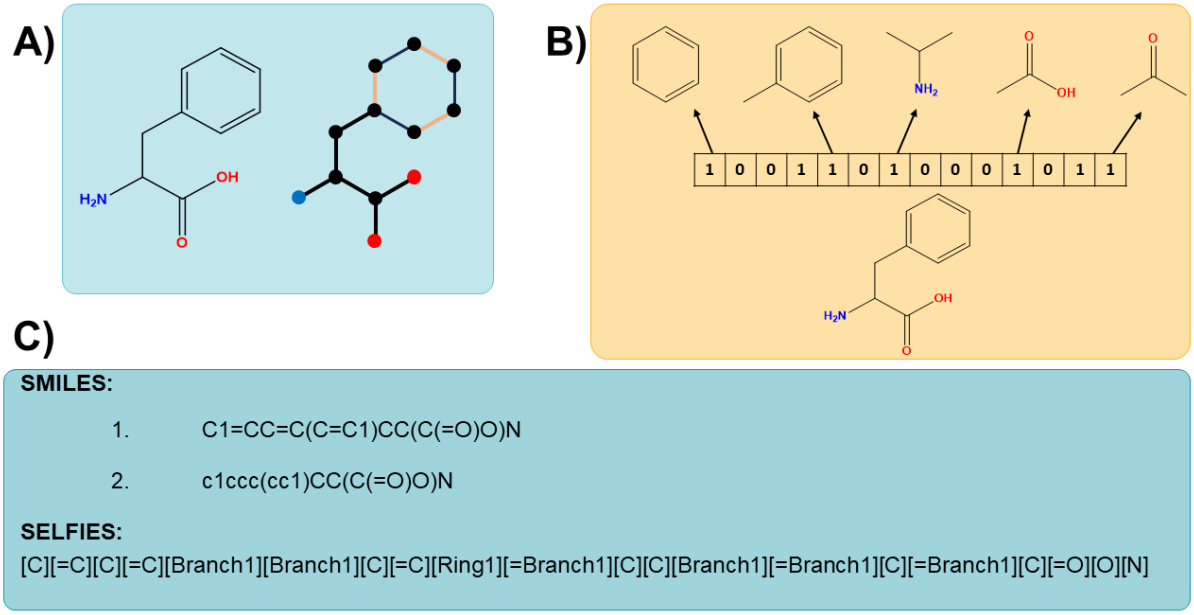


Figura 3 – Representações moleculares. A) Representação clássica, em duas dimensões, de uma fenilalanina, ao lado de sua representação na forma de grafos. B) Ilustração do funcionamento de um *fingerprint* genérico para a fenilalanina, onde cada "1" na sequência representa uma subestrutura da molécula. C) Duas representações lineares em cadeias de caracteres (*strings*) distintas: SMILES e SELFIES.
Fonte: Elaborada pelo autor.

1.3.3 SMILES

Introduzidos na década de 1980,[53-54] os SMILES (do inglês, *simplified molecular input line entry system)* se tornaram a representação linear alfanumérica com maior popularidade e aceitação na comunidade quimio-informática. Durante seu desenvolvimento, os autores focaram em implementar a teoria de grafos aplicada a moléculas para garantir uma especificação rigorosa de estruturas, com uma gramática que fosse natural e minimalista.[55] Na gramática dos SMILES, os átomos são representados pelos símbolos dos elementos químicos, enquanto as ligações simples, duplas e triplas são representadas pelos caracteres "–", "=" e "#", respectivamente. Normalmente, as ligações simples e os hidrogênios são omitidos. As ramificações são representadas por parênteses "()" que "confinam" a ramificação e os anéis são indicados por um número colocado imediatamente após o átomo o inicial desse anel e após seu último átomo. As ramificações podem ainda ser posicionadas dentro de outras ramificações, adicionando-se mais parêntesis. A aromaticidade de uma molécula (ou uma subestrutura) pode ser representada de diferentes maneiras: alternando símbolos de ligação simples e dupla, ou escrevendo os átomos aromáticos

em caixa baixa (*e.g.*, um anel benzeno pode ser escrito como C1=CC=CC=C1, C1=C–C=C–C=C1 ou c1ccccc1). A Figura 3C mostra um exemplo de SMILES que contém essas características.

1.3.4 SELFIES

Embora o uso dos SMILES tenha se tornado amplamente difundido, essa representação apresenta algumas limitações tais como uma molécula não ser unicamente descrita por um único SMILES e pequenas variações na sua sequência poderem produzir moléculas inválidas.[13,56-57] Em modelos generativos, por exemplo, a cadeia de caracteres SMILES resultante pode ser inválida devido à ausência de um par de parêntesis (indicando a ramificação) ou de um par de números (indicando o anel), na ordem correta. Por exemplo, embora “CCC)O(NC” possua um par de parêntesis, a ordem deles não está correta e, consequentemente, esse SMILES é inválido. O mesmo problema acontece para “C1CCNC2”, embora haja dois números confinando o anel, o segundo número é diferente do primeiro e resulta numa cadeia de SMILES inválido.

Para superar essas limitações, em 2020 foi introduzida uma nova representação molecular baseada em cadeia de caracteres, os SELFIES (do inglês, *SELF-referencIng Embedded Strings*), a qual foi concebida com a premissa de que toda cadeia de caracteres pode ser traduzida para uma molécula válida, uma vez que toda e qualquer combinação de símbolos que estão presentes no alfabeto dos SELFIES podem ser mapeados para um grafo químico válido, eliminando moléculas semântica e sintaticamente inválidas.[56,58]

Nos SELFIES, as subunidades da representação em cadeia são chamadas de *tokens*, nos quais os átomos também são representados pelos símbolos dos elementos químicos, e ligações duplas e triplas também são indicadas por “=” e “#”, entretanto, todos os *tokens* são confinados por colchetes. Assim, [C][C][C][C] representa uma molécula de butano, enquanto que [C][C][=C][C] se traduz para uma molécula de 2-buteno.

As ramificações e anéis são definidos por *tokens* especiais, por exemplo [Branch1] e [Ring1], respectivamente. Seu tamanho vai ser definido pelo *token* que vier imediatamente depois deles na cadeia de caracteres. Para que isso seja possível,

o *token* subsequente é sobrecarregado com um número, isto é, ele armazena a informação de um valor numérico que indica qual é o tamanho do anel ou da ramificação (Tabela 1).

Tabela 1 - Lista de símbolos (*tokens*) SELFIES que são sobrecarregados com um valor numérico que aparece depois de um *token* referente a um anel ou uma ramificação. Todos os símbolos que não estão listados são sobrecarregados com o índice "0".

Índice	Símbolo	Índice	Símbolo
0	[C]	8	[#Branch2]
1	[Ring1]	9	[O]
2	[Ring2]	10	[N]
3	[Branch1]	11	[=N]
4	[=Branch1]	12	[=C]
5	[#Branch1]	13	[#C]
6	[Branch2]	14	[S]
7	[=Branch2]	15	[P]

Fonte: Adaptada de LO *et al*. [57]

Ao analisar a cadeia de caracteres SELFIES “[C][N][Branch1][Ring2][C][C][C][C][C][C][C]” é possível verificar que há um símbolo de ramificação na terceira posição ([Branch1]). O *token* subsequente ([Ring2]) indica o tamanho da ramificação (definido como índice do *token* + 1). Assim, o SMILES correspondente é “CN(CCC)CCCC”.

Para garantir validade semântica aos SELFIES, isto é, garantir a validade da valência das ligações químicas, SELFIES utiliza uma gramática formal (ou teoria de máquinas de estados finita).[57, 59] A gramática formal deriva os símbolos das moléculas e cada passo de derivação ela pode mudar o seu estado. Como o estado define as regras para o próximo passo de derivação, isso acaba funcionando como memória mínima que codifica para restrições físicas e garante que apenas moléculas válidas sejam derivadas. Nesse contexto, SELFIES podem ser entendidas como uma linguagem de programação simples, e uma cadeia de caracteres SELFIES é um programa que cria um grafo molecular ao ser executado.[57]

CAPÍTULO 2: OBJETIVOS

2 OBJETIVOS

2.1 Objetivo geral

Gerar candidatos a compostos líderes contra o *P. falciparum* utilizando métodos de aprendizado de máquina, a partir de uma base de dados de compostos selecionados cuja atividade antiplasmodial já foi avaliada. Adicionalmente, utilizar os autocodificadores variacionais para a geração de anomalias representacionais e modos anômalos ainda desconhecidos para dados. Além disso, descobrir novos candidatos a inibidores da protease principal (M^{pro}) de SARS-CoV-2 como candidatos a compostos líderes para COVID-19.

2.2 Objetivos específicos

Malária

- Construir um banco de dados diverso com moléculas reportadas na literatura com promissora atividade antiplasmodial;
- Treinar um autocodificador variacional capaz de gerar novas moléculas;
- Refinar o modelo com características essenciais para a formação de um candidato a antimalárico;
- Treinar um modelo de regressão para predizer a potência dos compostos gerados;
- Avaliar as propriedades das moléculas obtidas.

Geração de anomalias

- Utilização de uma representação molecular cujos modos de invalidade sejam desconhecidos;
- Explorar o espaço latente do autocodificador variacional por meio de hiper esferas;
- Comparar o modelo de aprendizado de máquina com modelos que não usam desse tipo de abordagem computacional;
- Estabelecer um domínio de aplicabilidade para a geração de anomalias;

- Sugerir modificações que levem à validade da representação.

COVID-19

- Filtrar compostos de uma biblioteca de produtos naturais extraídos da biodiversidade brasileira;
- Realizar triagem virtual com os compostos dessa biblioteca que tenham características de fármacos;
- Realizar simulações de dinâmica molecular para determinar os compostos mais promissores;
- Calcular a energia livre de ligação dos compostos no alvo;
- Determinar afinidade pelo alvo, experimentalmente;
- Determinar o tipo de inibição do composto.

CAPÍTULO 3: BUSCA DE INIBIDORES PARA MALÁRIA UTILIZANDO APRENDIZADO DE MÁQUINA

3 BUSCA DE INIBIDORES PARA MALÁRIA UTILIZANDO APRENDIZADO DE MÁQUINA

3.1 Panorama geral

A malária é uma doença tropical infecciosa bastante antiga[60] causada por parasitas intracelulares do gênero *Plasmodium*. Embora existam mais de 150 espécies desse protozoário que infectam diferentes mamíferos, répteis e aves, apenas cinco são responsáveis por causar a doença em humanos: *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale* e *P. knowlesi*.[61] Além dessas, existem duas outras espécies que estão associadas à casos recentes de infecções zoonóticas, o *P. cynomolgi* e o *P. simum*, entretanto seus impactos clínicos e globais ainda precisam ser melhor compreendidos.[62-63]

De acordo com a Organização Mundial da Saúde (OMS), em seu relatório publicado em 2022, estima-se que em 2021 cerca de 247 milhões de casos da doença tenham acontecido nas 84 regiões endêmicas da malária, um aumento de 2 milhões de casos quando comparado ao ano de 2020 (245 milhões de casos reportados), com a maior parte desse aumento concentrado em países africanos.[64] A incidência da doença, isto é, a quantidade de casos a cada 1000 habitantes, reduziu de 82 em 2000, para 57 em 2019, antes de aumentar para 59 em 2020 e se manter nesse valor em 2021.[64] Tal aumento está diretamente associado à pandemia de COVID-19.[64] O relatório da OMS também indica que a região mais afetada pela doença foi a África Subsaariana, que concentrou 95% dos casos, com o *P. falciparum* sendo a espécie dominante na região.[65] A Figura 4A mostra a distribuição de casos estimados para o ano de 2020, ao redor do mundo.

Globalmente, as mortes decorrentes da malária reduziram no período de 2000 a 2019, saindo de 897 mil em 2000, para 568 mil em 2019. Para 2021, estimam-se que 619 mil mortes tenham sido causadas em decorrência da doença. Essa diferença no número de mortos entre 2019 e 2021 está também associada à pandemia de COVID-19, uma vez que muitos serviços essenciais para a manutenção da vida das pessoas em regiões afetadas pela doença foram interrompidos e/ou prejudicados nesse período.[65] Das mais de 600 mil mortes, 76% aconteceram em crianças com idade menor do que 5 anos, um percentual menor quando comparado com 2000

(87%), mas ainda sim extremamente preocupante. O relatório da OMS também aponta que 96% de todas as mortes estimadas para o ano de 2021 ocorreram na África Subsaariana. A Figura 4B mostra a distribuição de mortes causadas pelo *P. falciparum*, o principal causador das mortes pela doença,[66] em todo o mundo no ano de 2020.

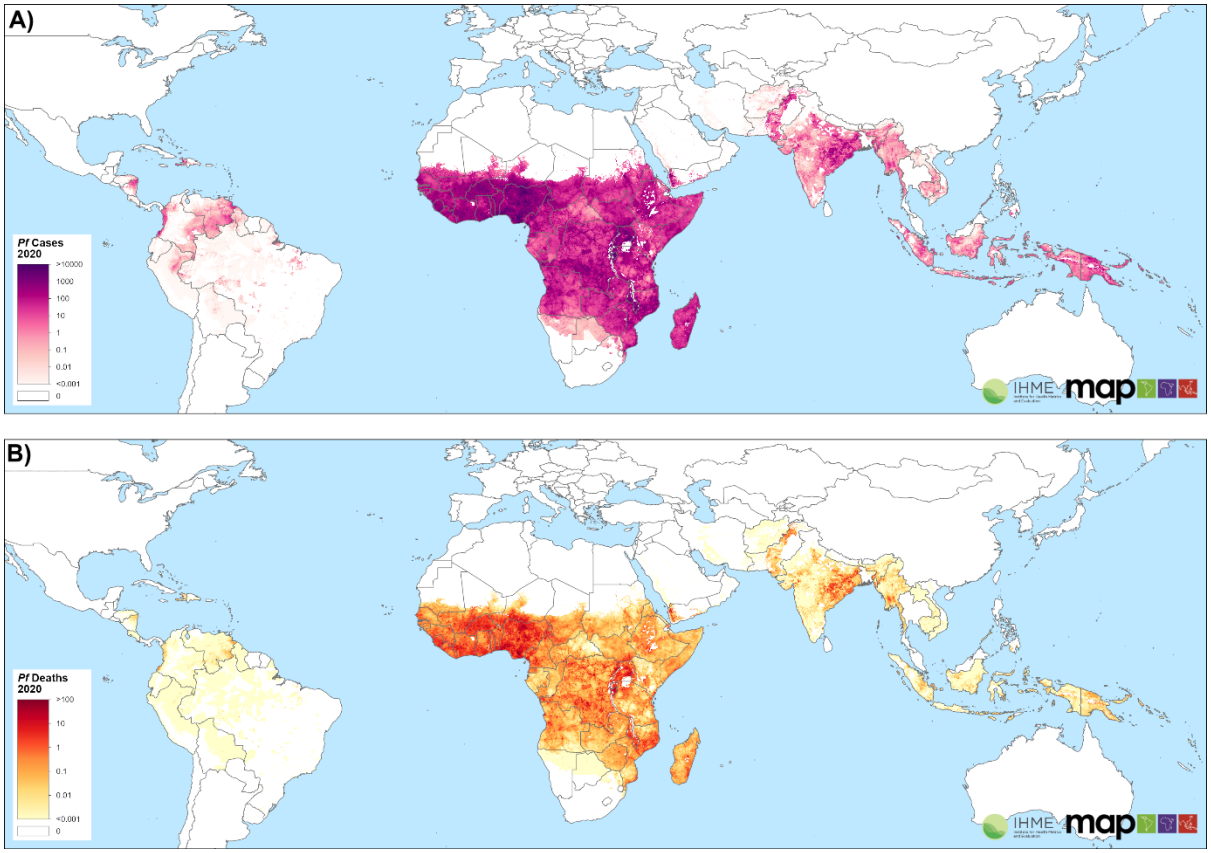


Figura 4 – Estatísticas *P. falciparum*. A) Incidência de casos do parasito em todo o planeta para o ano de 2020. B) Número de mortes causadas por malária apos a infecção por *P. falciparum* no ano de 2020.

Fonte: THE MALARIA ATLAS PROJECT.[67]

No Brasil, a região endêmica compreende a bacia Amazônica. Entretanto, a espécie com maior incidência em território nacional é o *P. vivax*, responsável por 84% dos casos estimados da doença para 2021 (Figura 5). Para esse mesmo ano, o relatório da OMS reportou 58 mortes decorrentes da doença.

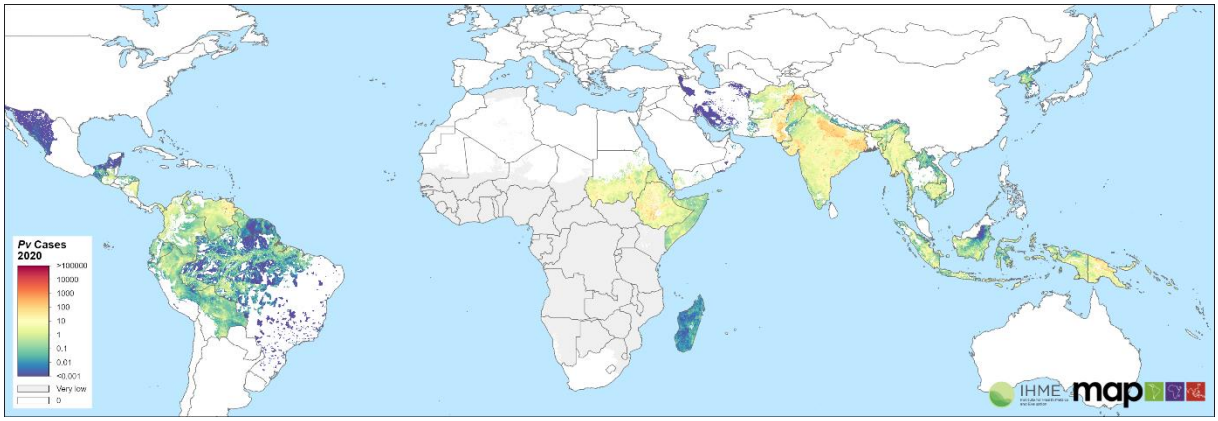


Figura 5 – Estatísticas *P. vivax*. Incidência de casos estimados de infecções do parasito ao redor do mundo para o ano de 2020.
Fonte: THE MALARIA ATLAS PROJECT.[67]

De todas as espécies, o *P. falciparum* é o principal causador das mortes pela doença.[66] Esse parasito possui um ciclo de vida bastante complexo sendo transmitido para hospedeiros humanos por mosquitos fêmea do gênero *Anopheles* (Figura 6).[68]

Ao se alimentar do sangue de um hospedeiro humano, as fêmeas desses mosquitos introduzem esporozoítos do parasito na pele de quem estão se alimentando, os quais posteriormente atingem as células do fígado. Dentro do fígado, o parasito sofre maturação, formando esquizontes (ou sua forma dormente, hipnozoítos, no caso do *P. vivax* e *P. ovale*). Esses esquizontes liberam merozoítos que invadem os glóbulos vermelhos durante o estágio assexuado de reprodução no sangue do hospedeiro. Dentro de cada glóbulo vermelho, o parasita assume a forma de um anel e evolui para um trofozoíto, caracterizado pela presença de um vacúolo digestivo ácido responsável por degradar a hemoglobina e liberar peptídeos e aminoácidos essenciais para a síntese de proteínas. Em seguida, esses parasitas amadurecem para a forma de esquizontes multinucleados que, em seguida, geram milhares de merozoítos que são liberados na corrente sanguínea após o rompimento dos eritrócitos. Esses merozoítos liberados infectam novos glóbulos vermelhos e o ciclo intraeritrocítico é reiniciado.[69] Uma pequena porcentagem de merozoítos se diferencia em gametócitos masculinos e femininos, dando início à fase de transmissão sexual. Esses gametócitos são ingeridos por um outro mosquito fêmea, continuam o ciclo do parasita no inseto vetor.[69]

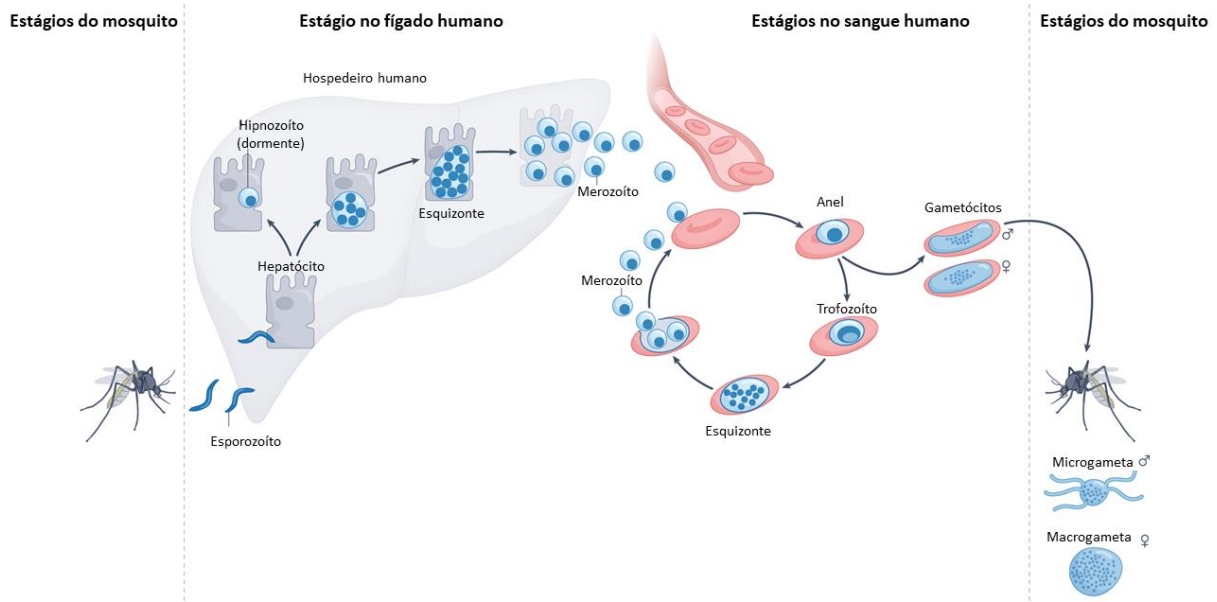


Figura 6 – Ciclo de vida do parasito da malária humana. O ciclo de vida do parasito envolve um vetor invertebrado (mosquito *anófeles*) e um hospedeiro secundário vertebrado. Dentro do hospedeiro humano, o parasito passa por estágios de desenvolvimento assexuado hepático e intraeritrocítico. Parte desses últimos inicia o desenvolvimento sexual em gametócitos. Quando maduros, os gametócitos estão prontos para infectar novos mosquitos.
Fonte: Adaptada de SIQUEIRA-NETO *et al.*[70]

Os primeiros sintomas da doença geralmente são inespecíficos e aparecem durante o estágio de reprodução assexuado do parasito, isto é, quando estão dentro dos glóbulos vermelhos, podendo incluir mal-estar, dores de cabeça, fadiga e desconforto abdominal, que são seguidos por febre irregular.[69] Também podem ocorrer náuseas, vômitos e hipotensão postural.[71] Sintomas como convulsões seguidas por coma também estão associados à infecção por *P. falciparum* (malária cerebral).[72]

A evolução da doença e as sequelas deixadas dependem de alguns fatores, como a espécie de *Plasmodium* com a qual o hospedeiro é infectado, exposições prévias ao parasito, comorbidades do paciente e a qualidade do tratamento conduzido.[73] As estatísticas mostram que somente uma parcela dos pacientes desenvolve todas as manifestações clínicas da malária, podendo chegar ao estágio severo da doença, uma vez que seu curso é contido ou pelo tratamento aplicado de maneira eficaz, ou por um sistema imunológico funcional ou mesmo pela morte precoce do infectado.[74] No entanto, como alguns processos patológicos podem continuar ocultos por um longo período de tempo, como é o caso das espécies que formam hipnozoítos, alguns pacientes podem experimentar consequências também a longo prazo.[75]

Atualmente, a OMS recomenda como “padrão ouro” para o tratamento da malária causada por *P. falciparum*, a utilização da terapia combinada de derivados de artemisinina (ACT, do inglês, *artemisinin-based combination therapy*), a qual faz uso de um derivado da artemisinina, um composto de ação rápida, junto a um outro antimalárico que possua um mecanismo de ação diferente.[64,76-78] No caso de malária causada por *P. vivax*, o tratamento preconizado também é uma terapia combinada, contudo, baseada na associação de cloroquina com primaquina ou tafenoquina. Entretanto, o aumento na resistência aos fármacos antimaláricos tem ameaçado a efetividade do tratamento, controle e eliminação da doença.[79] Os recentes casos de resistência ao tratamento da malária utilizando derivados de artemisinina, reportados em países do sudeste asiático e na África,[80-82] ameaçam a redução do número de mortes e o controle da doença conseguido até então.[68]

Dessa forma, fica evidente a necessidade de se investigar e desenvolver novos compostos com possível atividade antimalárica de modo a encontrar alternativas aos tratamentos disponíveis. Nesse contexto, o aprendizado de máquina é uma alternativa atrativa para a aceleração dos processos de descoberta e desenvolvimento de novos candidatos a fármacos antimaláricos.[83-84]

3.1.1 Aprendizado de máquina e a malária

Diversas abordagens que utilizam de aprendizado profundo têm sido utilizadas para auxiliar no controle e combate à malária. Por exemplo, Arshadi *et al.*[84] desenvolveram o *DeepMalaria*, um processo que usa aprendizado profundo capaz de predizer as propriedades inibitórias de compostos baseados em suas estruturas na forma de SMILES, auxiliando na triagem virtual baseada em ligantes independente do alvo. Os autores utilizaram um modelo de grafos treinado com uma base de dados da GlaxoSmilthKline (GSK), focando em compostos macrocíclicos devido à escassez de informação acerca de compostos desse tipo como antimaláricos.

Neves *et al.*[9] desenvolveram um protocolo de modelagem utilizando aprendizado profundo para construir modelos de relação quantitativa estrutura-atividade (QSAR) binários e contínuos baseados em grandes conjuntos de dados, a fim de predizer a atividade de compostos ainda não testados contra o parasita causador da malária bem como sua citotoxicidade.

Lima *et al.*[85] utilizaram um modelo integrativo de triagem virtual assistida por inteligência artificial para identificar novos candidatos a antimaláricos. Esse trabalho combinou modelos baseados na forma para buscar inibidores da proteína quinase 7 de *P. falciparum* com modelos de aprendizado de máquina para selecionar inibidores do estágio sanguíneo do parasito.

O desenvolvimento de métodos de aprendizado de máquina para o diagnóstico de malária tem se tornado cada vez mais comum. Siłka *et al*.[73] desenvolveram uma nova arquitetura de rede neural convolucional para detectar o parasita causador da malária a partir de amostras de sangue, com uma acurácia maior do que 99%. Nessa abordagem, o modelo foi treinado com imagens de microscopia de esfregaços de sangue infectado ou não com *P. falciparum*. O modelo retorna a informação de maneira que facilita um diagnóstico visual rápido e válido.

Algumas abordagens combinam modelos generativos com métodos mais tradicionais de aprendizado de máquina para auxiliar no diagnóstico e tratamento da malária. É o caso do trabalho de Amin *et al.*[86] que combinaram redes generativas adversariais (GAN)[41] semi-supervisionadas com transferência de aprendizado para um modelo classificatório utilizado para tornar o diagnóstico e tratamento da malária mais conveniente. O modelo utilizou imagens de microscopia de células de sangue infectadas.

Num trabalho mais recente, Tan e Liang desenvolveram um *framework* baseado em transformadores[87] e GANs para a classificação multi-classe do parasito da malária e auxiliar no diagnóstico. No *framework*, as GANs foram utilizadas para gerar mais amostras de treinamento a partir de imagens de células multi-classes, com o objetivo de melhorar a robustez do modelo resultante, enquanto os transformadores foram utilizados para realizar a classificação.

Embora os modelos generativos tenham sido utilizados para geração de moléculas, pouquíssimos trabalhos tiveram a malária como alvo. Na verdade, o único exemplo encontrado na literatura foi publicado por Godinez *et al*.[88] que utilizaram as *junction-trees* como representação molecular para construir um modelo, na arquitetura de autocodificador variacional, capaz de gerar moléculas com as propriedades bioativas desejadas para a inibição do parasito.

Nesta tese, dois modelos de aprendizado profundo foram utilizados, cada um treinado com uma representação molecular diferente, a fim de se obter novos compostos com potencial de inibição da atividade do *P. falciparum*: i) um modelo

generativo, utilizando uma arquitetura de autocodificador variacional pré-treinado com um conjunto grande e genérico de moléculas e, posteriormente, refinado com conjuntos cujas moléculas possuem propriedades específicas de interesse, utilizando SELFIES como representação molecular; e ii) um modelo de regressão treinado com uma base de dados de moléculas já testadas contra o parasito da malária, a fim de proporcionar mais uma etapa de filtragem para as moléculas geradas, utilizando grafos como representação molecular.

3.2 METODOLOGIA

3.2.1 Seleção das moléculas para a criação da base de dados

As moléculas com atividade antiplasmodial já testadas e reportadas na literatura foram obtidas a partir da base de dados ChEMBL[89] (versão 29) e armazenadas em um banco de dados *MySQL* local. Apenas as tabelas "*activities*", "*assays*", "*compound_properties*", "*compound_records*" "*compound_structures*", "*docs*" e "*molecule_dictionary*" da base de dados foram utilizadas, de modo a permitir a seleção das informações de interesse acerca das moléculas. Um filtro limitando a busca apenas pelo organismo *P. falciparum* também foi adicionado.

Apenas moléculas cujas atividades estivessem reportadas em valores de IC_{50}, EC_{50}, ou K_i foram selecionadas. As unidades permitidas durante a seleção foram nM e µM, com posterior conversão para nM. Moléculas cujo valor de atividade inibitória, convertido para pIC_{50} ($-\log IC_{50}$), fosse abaixo de 5,523 (equivalente a 3.000 nM) foram consideradas como inativas; para valores acima de 6,000 (equivalente a 1.000 nM), as moléculas foram consideradas ativas; moléculas cujo valor de atividade estivesse entre 5,523 e 6,000 foram descartadas. Moléculas duplicadas foram removidas por comparação das estruturas na forma de SMILES.[53]

3.2.2 Padronização das moléculas

Todas as estruturas das moléculas utilizadas neste capítulo foram selecionadas foram padronizadas para o formato de SMILES canônico, utilizando o protocolo disponibilizado pela equipe do *Malaria Inhibitor Prediction platform*.[90] O algoritmo

envolve 6 etapas. A primeira delas consiste em quebrar ligações covalentes entre oxigênios ou nitrogênio e átomos de metais do grupo I e II. Em seguida, o algoritmo neutraliza as cargas das moléculas, adicionando ou removendo prótons. Após essa etapa, uma série de regras de padronização são aplicadas para a transformação das moléculas, tais como a mudança de hidrogênios entre heteroátomos, especialmente em anéis, ou então tautômeros, ou então mudanças para moléculas com cargas formais positivas ou negativas, cuja estrutura pode ser neutralizada após sucessivos rearranjos de ligações simples e duplas. Depois dessa etapa, o algoritmo procede para mais uma neutralização, em caso de alguma carga ter sido exposta pela etapa anterior. A etapa seguinte descarta quaisquer componentes de sais e solvatos. E, por fim, a molécula padronizada é retornada.

3.2.3 Conjunto de dados para o autocodificador variacional

Para compor o conjunto de treinamento foram combinadas moléculas de duas bases de dados: ChEMBL v.29[91-92] e ZINC15.[93-95] Para ambas as bases de dados, apenas a informação acerca da estrutura das moléculas representadas como SMILES foi armazenada. Inicialmente, todos os 2.105.463 compostos disponíveis no *ChEMBL* v.29 foram selecionados. Da base de dados ZINC15 foram selecionados apenas compostos com características *drug-like* e disponíveis em estoque, resultando num total de 7.099.779 moléculas. Desse conjunto, 2 milhões de moléculas foram selecionadas aleatoriamente. Para garantir que não houvesse repetição de estruturas, foi verificada a presença dos compostos do *ChEMBL* no subconjunto selecionado do ZINC15 e todas as duplicatas foram removidas, resultando num total de 3.982.553 moléculas.

O conjunto passou por um processo de conversão para duas dimensões, isto é, todos os caracteres “@”, “@@”, “\” e “/” os quais indicam, na sintaxe de SMILES, são indicativos de quiralidade, foram removidos. Além disso, o conjunto passou por um processo de padronização, conforme descrito em acima.

Esse conjunto foi submetido a um processo de filtragem contendo duas etapas:

i) filtragem por tokens raros – os SMILES foram tokenizados e apenas os tokens que estavam presentes em mais de 300 moléculas foram mantidos;

ii) comprimento da molécula (em unidade de tokens) – apenas as moléculas que possuíam até 104 tokens de comprimento permaneceram no conjunto.

Esse processo resultou em 3.916.274 moléculas no conjunto total. Todos os SMILES foram convertidos para SELFIES por meio da biblioteca SELFIES 2.1.1.[57] Após essa conversão, o comprimento máximo das moléculas foi de 112 tokens. Para garantir que todas as moléculas tivessem o mesmo comprimento, tokens de *"padding"* foram adicionados àquelas cujo comprimento fosse menor do que 112. Assim, ao final do processo, o conjunto de tokens resultante continha 50 elementos (Tabela 2). A esses tokens foram atribuídos números inteiros. Para o treinamento do modelo, 85% desse conjunto foi utilizado, sendo o restante usado como conjunto de validação.

Tabela 2 – Tokens de SELFIES com os respectivos inteiros associados. O processo de tokenização das moléculas resultou num vocabulário de 50 tokens distintos, capazes de escrever todo o conjunto de moléculas.

Inteiro	Token	Inteiro	Token	Inteiro	Token	Inteiro	Token	Inteiro	Token
0	'pad'	11	[=Branch2]	22	[=S+1]	33	[I]	44	[Ring1]
1	[#Branch1]	12	[=C]	23	[=S]	34	[N+1]	45	[Ring2]
2	[#Branch2]	13	[=N+1]	24	[Br]	35	[N-1]	46	[S+1]
3	[#C-1]	14	[=N-1]	25	[Branch1]	36	[NH1]	47	[S-1]
4	[#C]	15	[=N]	26	[Branch2]	37	[N]	48	[SH0]
5	[#N+1]	16	[=O]	27	[C-1]	38	[O-1]	49	[S]
6	[#N]	17	[=P+1]	28	[CH0]	39	[OH0]		
7	[#S]	18	[=PH1]	29	[CH1]	40	[O]		
8	[18F]	19	[=P]	30	[C]	41	[P+1]		
9	[2H]	20	[=Ring1]	31	[Cl]	42	[PH1]		
10	[=Branch1]	21	[=Ring2]	32	[F]	43	[P]		

Fonte: Elaborada pelo autor.

3.2.4 Parâmetros do autocodificador variacional

Para gerar as moléculas, autocodificador variacional (VAE) foi treinado usando TensorFlow v.2.10.0.[96] Para ambos os componentes do modelo (codificador e decodificador), a função de ativação utilizada foi a tangente hiperbólica.

O codificador, responsável por mapear os SELFIES em um espaço latente, possui uma camada de incorporação (*Embedding layer*) que mapeia os tokens

SELFIES representados como vetores de inteiros em vetores densos. Essa camada é seguida por quatro camadas LSTM (*Long-Short Term Memory layer*),[43] com 256 neurônios ocultos cada, que processam as sequências advindas camada anterior. Em seguida, há uma camada de achatamento (*Flatten layer*), responsável por preparar a saída para a criação de uma distribuição latente e, por fim, uma camada densa (*Dense layer*), que produz os parâmetros da distribuição latente (utilizando valores médios e log da variância), num espaço 196-dimensional.

O decodificador, que é o responsável por gerar moléculas a partir de pontos do espaço latente, é composto por uma camada densa (*Dense layer*), que mapeia os pontos no espaço latente de volta para o espaço original dos SELFIES, seguida por uma camada de redimensionamento (*Reshape layer*) para preparar as saídas para a camada seguinte. Quatro camadas de unidades recorrentes com portas (*GRU layers*)[97] para processar e gerar sequências de tokens SELFIES, com 256 neurônios ocultos cada, seguem a camada anterior. Por fim, uma camada densa (*Dense layer*) que gera distribuições não normalizadas de tokens SELFIES para cada posição na sequência decodificada – a probabilidade do modelo para os dados.

O treinamento do modelo buscou a minimização da função objetivo, a qual é composta por dois termos: a entropia cruzada esparsa *softmax* com perda de *logits*[98-99] como termo de reconstrução, e a divergência de *Kullback-Leibler* (KL)[33] como termo de regularização, ponderada por um fator β, anelado[34] usando uma função linear após três épocas. Foi utilizado o otimizador de Adam e a taxa de aprendizado utilizada for de 0,0001. Os parâmetros foram inicializados utilizando *Glorot* uniforme,[100] e o treinamento aconteceu utilizando um *batch* de tamanho 256. O modelo final convergiu após 19 épocas, devido ao protocolo de parada antecipada (*early stopping*),[101-102] com um paciente igual a 8. O processo de treinamento desse modelo, em tempo de parede, levou 325 horas, numa única GPU NVIDIA TITAN. Para fazer as análises de quimioinformática, foi utilizada as funções da biblioteca *RDKit*.[103] Todos os hiperparâmetros foram escolhidos após uma série de testes com parâmetros diferentes.

3.2.5 Transferência de aprendizado

O modelo convergido foi refinado com 2 conjuntos de moléculas distintos: i) o conjunto contendo as 28.395 moléculas testadas contra malária, e classificadas como

ativas (15.821) e inativas (12.574), curadas no banco de dados descrito acima; ii) o conjunto MMV – Saint Jude,[104] contendo 305.643 moléculas, das quais 2.507 são ativas contra o parasito e o restante, inativo. A base de dados MMV – St. Jude possui moléculas com propriedades antiplasmodial testadas por meio de triagem de médio desempenho contra cepas 3D7 de *P. falciparum.* Esses ensaios foram conduzidos pela equipe do MMV e do hospital St. Jude.[104]

Os parâmetros treináveis do codificador foram mantidos “congelados”, enquanto o decodificador foi treinado com os novos conjuntos. Todos os hiper parâmetros, quantidade de tokens disponível no conjunto de tokens, comprimento das moléculas, e protocolos de parada antecipada foram mantidos conforme descrito nas seções 3.2.3 e 3.2.4.

3.2.6 Amostragem

As amostragens foram feitas de maneira gananciosas[105] a partir das distribuições de tokens SELFIES resultantes do decodificador em termos da posição na sequência para gerar SELFIES, isto é, em cada posição da sequência, foi escolhido o token que parecesse ser o mais provável, com base nas distribuições de probabilidade.

As buscas no espaço latente do VAE foram feitas por meio de matrizes de vetores aleatórios 196-dimensionais, uniformemente amostrados desse espaço, as quais foram fornecidas ao decodificador. Para cada um dos modelos refinados, foram amostradas 50.000 moléculas.

3.2.7 Classificação das moléculas

Para predizer valores de IC_{50} para as moléculas geradas, um modelo de regressão for feito utilizando as moléculas cuja atividade contra *Plasmodium falciparum* é conhecida e que compõem o banco descrito acima, porém, para esta abordagem, não foram descartadas nem uma das 32.176 moléculas.

Para a definição do modelo, foi utilizada a biblioteca *DeepChem*.[106] Os SMILES foram convertidos para vetores que descrevem cada átomo da molécula utilizando a

classe *ConvMolFeaturizer()*.[107] O conjunto foi então dividido, aleatoriamente, em um conjunto de treinamento (com 80% das moléculas iniciais) e conjunto de teste (com 20% das moléculas iniciais) usando a classe *RandomSplitter()*. O modelo foi criado usando a classe *GraphConvModel()* no modo de regressão, para uma única tarefa (predizer valores de IC_{50}), com três camadas convolucionais, cada uma com 128 unidades. A fração das conexões entre neurônios que são aleatoriamente desativadas durante cada passagem de treinamento foi definida como 15%. A taxa de aprendizado utilizada foi de 0,001, com um tamanho de *batch* de 100 e o modelo treinado por 200 épocas. O desempenho do modelo foi avaliado pelo coeficiente de determinação.[108]

3.2.8 Espaço químico e propriedades físico-químicas

Para a determinação do espaço químico das moléculas, foi feita uma conversão da estrutura de SMILES para *fingerprints* de Morgan.[52] Para calcular os *fingerprints* de Morgan foi utilizado o *software* livre de quimioinformática RDKit.[103] O raio utilizado para o cálculo dos descritores foi de 2, com um vetor de 2.048 bits.

A visualização do espaço químico foi realizada utilizando-se os algoritmos t-SNE (*t-distributed stochastic neighbor embedding*)[109] a PCA (*principal component analysis*)[110-111] implementados na biblioteca Scikit-learn.[112] Tais algoritmos permitem a visualização de dados hiperdimensionais em duas ou três dimensões a partir de técnicas de redução de dimensionalidade. Em outras palavras, os *fingerprints*, que inicialmente possuíam 2.048 bits, foram convertidos em vetores de apenas dois valores (duas dimensões), permitindo que fosse possível realizar a análise do espaço químico visualmente.

3.3 RESULTADOS

3.3.1 Base de dados extraída do ChEMBL29

O diagrama entidade-relacionamento (Figura 7) mostra a base de dados resultante da seleção de tabelas do ChEMBL29, com o filtro para selecionar apenas os compostos que foram testados contra *P. falciparum*. No diagrama é possível verificar o tipo de relacionamento que cada uma das tabelas apresenta.

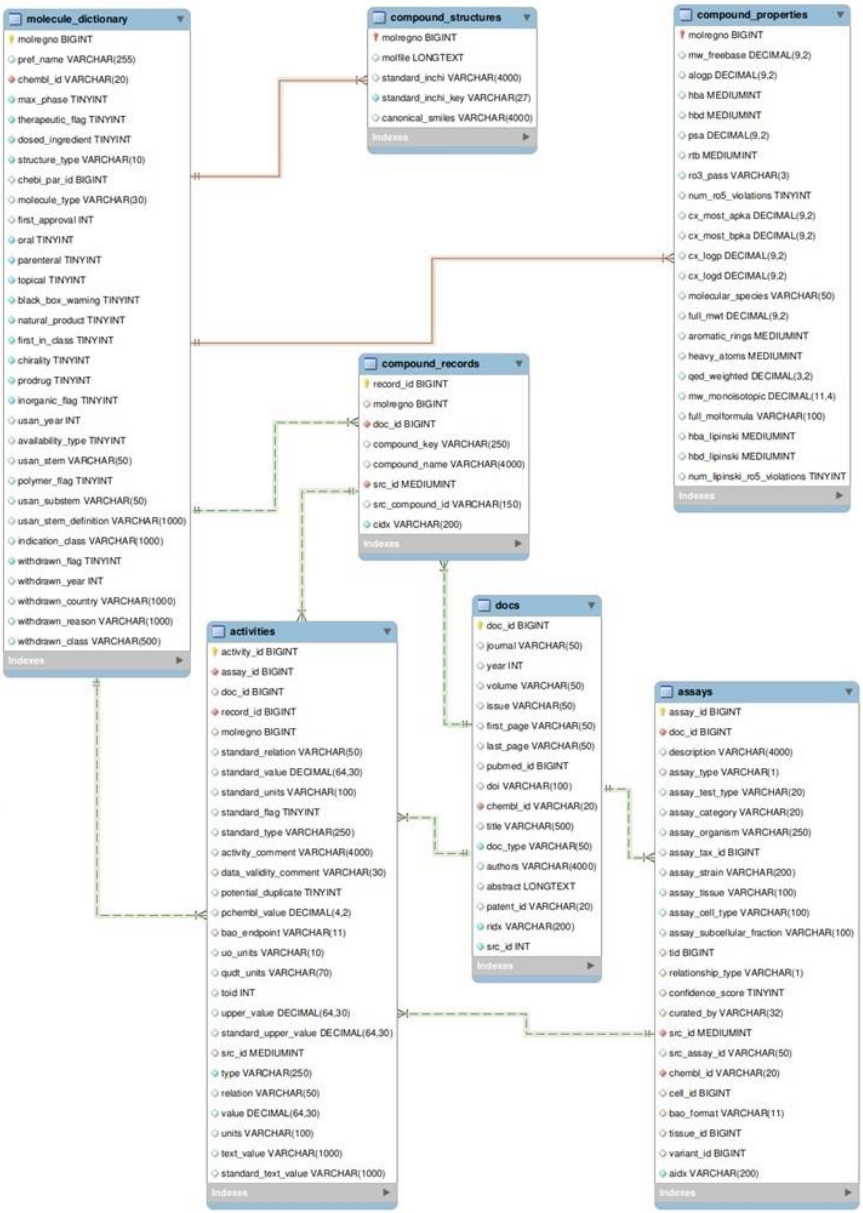


Figura 7 – Diagrama Entidade-Relacionamento com o subconjunto de tabelas extraídas do ChEMBL29 para compor o banco de moléculas testadas contra *P. falciparum*. As linhas tracejadas, em verde, representam relações do tipo não-identificadas e as linhas sólidas, em laranja, representam as relações do tipo identificadas.
Fonte: Elaborada pelo autor.

As características das diferentes tabelas que compõem o diagrama entidade-relacionamento estão descritas nos **ANEXOS**, a qual também apresenta os tipos de relacionamento entre cada uma das tabelas que compõem o banco de dados.

As moléculas resultantes desse processo de seleção do ChEMBL29 compuseram um conjunto denominado “antimaláricos ChEMBL” (ou AMC). As moléculas desse conjunto passaram por novas etapas de classificação e filtragem. Para isso, todos os valores de atividade, reportados na base como “nM”, foram

convertidos para pIC_{50}. Para esse valor menor ou igual a 5,523 (3.000 nM) a molécula foi classificada como inativa, para o valor maior ou igual a 6,000 (1.000 nM) a molécula foi classificada com ativa. Moléculas cujo pIC_{50} estivesse entre esses valores foram descartadas. Ao final dessas etapas, 28.395 moléculas únicas foram selecionadas, das quais 12.574 foram consideradas inativas e 15.821 sendo consideradas ativas, com cerca de 4 mil moléculas descartadas.

3.3.2 Análise exploratória dos conjuntos de treinamento e refinamento

As moléculas do conjunto AMC, tanto as ativas quanto as inativas, compartilham das mesmas regiões do espaço químico (Figura 8A). Embora compartilhem essa característica, quando comparadas par-a-par, os compostos se mostraram bastante diversos entre si, o que fica evidente pelo *heatmap* que indica a similaridade das moléculas (calculada pelo coeficiente de Tanimoto) numa escala de cores na qual quanto mais similar uma molécula é da outra, mais intensa a cor vermelha se mostra no mapa (Figura 8B). É evidente que existem regiões cuja similaridade entre os compostos é alta. Entretanto, o conjunto se mostra adequado tanto para a realização do treinamento do modelo de regressão, tanto para predizer valores de pIC_{50} quanto para o refinamento do modelo generativo inicial (conferindo características de antimaláricos já testados às moléculas geradas).

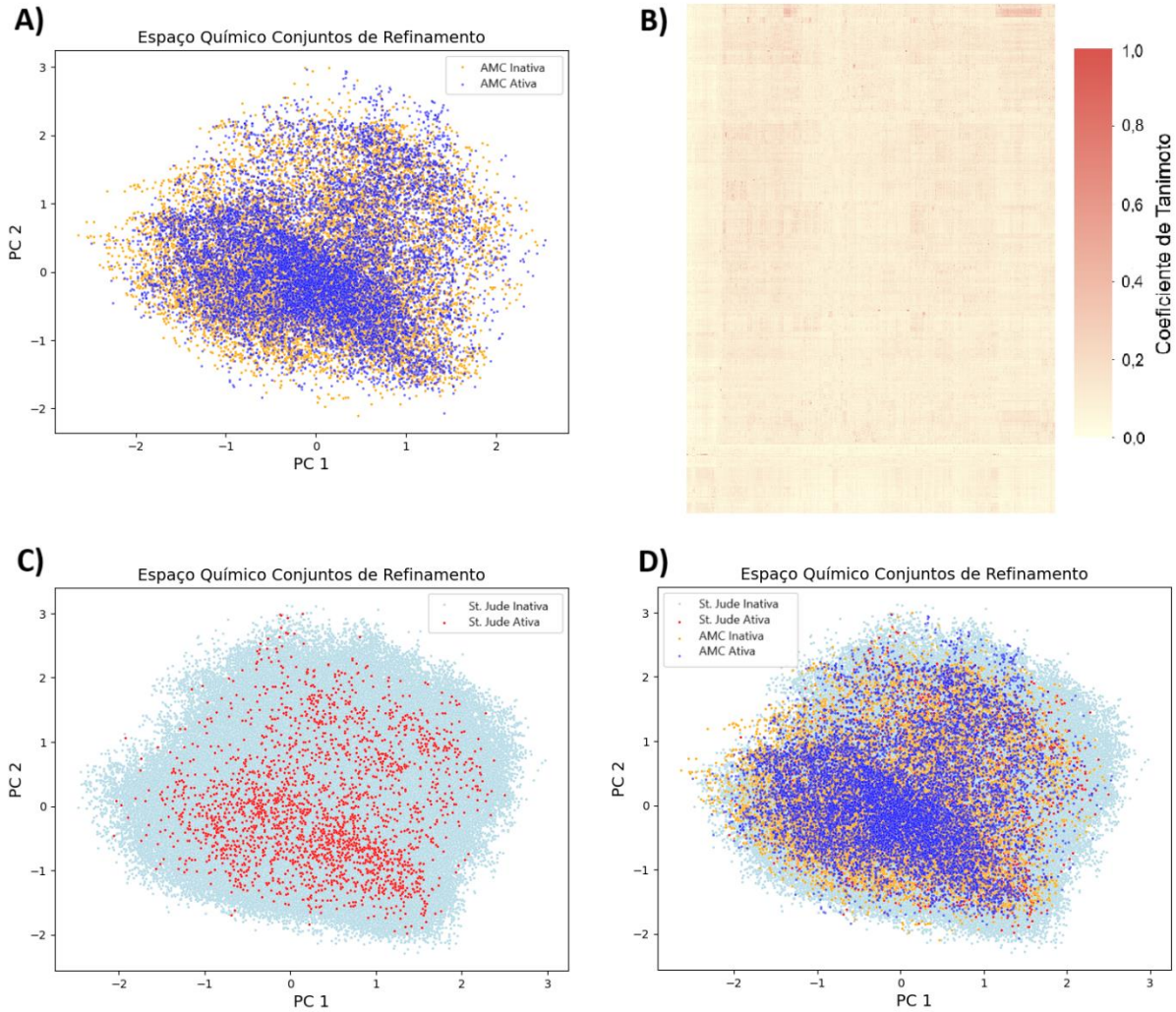


Figura 8 – Espaço químico dos conjuntos de refinamento e análise de similaridade interna com os compostos ativos e inativos dos antimaláricos ChEMBL. Análise das componentes principais (PCA) representando o espaço químico ocupado pelas moléculas dos conjuntos de refinamento. Em laranja estão representadas as moléculas inativas do conjunto antimaláricos ChEMBL, enquanto que em azul escuro estão as moléculas válidas desse mesmo conjunto. B) O *heatmap* indicam a similaridade entre as estruturas, pelo Coeficiente de Tanimoto, numa escala de cor na qual quanto mais vermelha for a região, maior será a similaridade. O eixo vertical do *heatmap* representa as moléculas ativas, enquanto que o horizontal, as inativas. C) Projeção das principais componentes para o conjunto MMV – St. Jude. Em azul claro estão respresentadas as moléculas inativas, e em vermelho, as moléculas ativas desse conjunto. D) Sobreposição dos espaços químicos de ambos os conjuntos utilizados para o refinamento, seguindo a mesma distribuição de cores descrita anteriormente.

Fonte: Elaborada pelo autor.

O as moléculas do conjunto MMV – St. Jude, tanto as classificadas como ativas quanto as inativas, também apresentaram esse comportamento de se sobreporem nas mesmas regiões do seu espaço químico (Figura 8C), indicando que elas compartilham das mesmas características, embora tenham atividades inibitórias distintas. Além disso, tanto as moléculas do MMV – St. Jude quanto as moléculas do

AMC se sobrepõem nas mesmas regiões do espaço químico (Figura 8D) (embora não existam moléculas duplicadas entre os conjuntos).

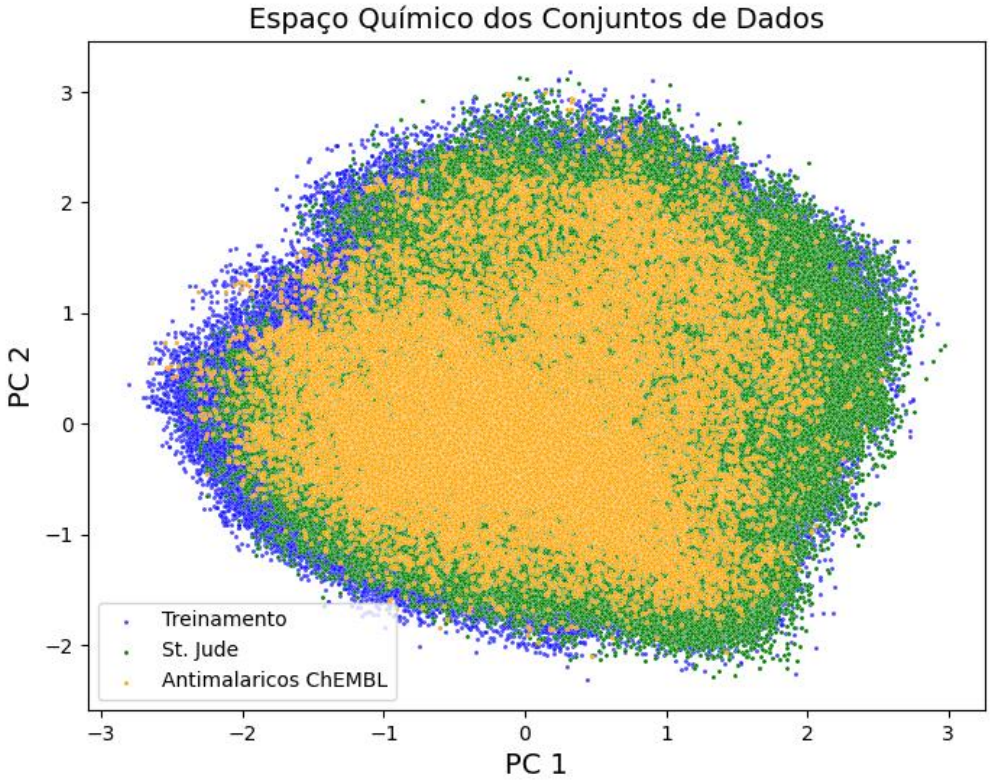


Figura 9 – Espaço químico dos conjuntos de dados. Análise das componentes principais (PCA) representando o espaço químico ocupado pelas moléculas dos conjuntos de pré-treinamento que contém moléculas do ChEMBL e do ZINC15 (em azul), de refinamento MMV – St. Jude (em verde) e de refinamento/treinamento do modelo classificatório antimaláricos ChEMBL (em amarelo).
Fonte: Elaborada pelo autor.

Ao fazer a análise das componentes principais de todos os conjuntos utilizados nas etapas de pré-treinamento e refinamento (Figura 9) é possível verificar que os espaços químicos de ambos os conjuntos se sobrepõem. Embora essa não seja uma característica necessária para os conjuntos de pré-treinamento e de refinamento, uma vez que o conjunto de pré-treinamento pode possuir apenas características gerais da tarefa desejada,[113-114] a sobreposição dos espaços dos dados de ambos os conjuntos pode permitir que o modelo tenha um desempenho melhor,[115] uma vez que o modelo inicialmente aprenderá as características gerais acerca da tarefa e, posteriormente, ele aprende características mais específicas dos dados de refinamento.

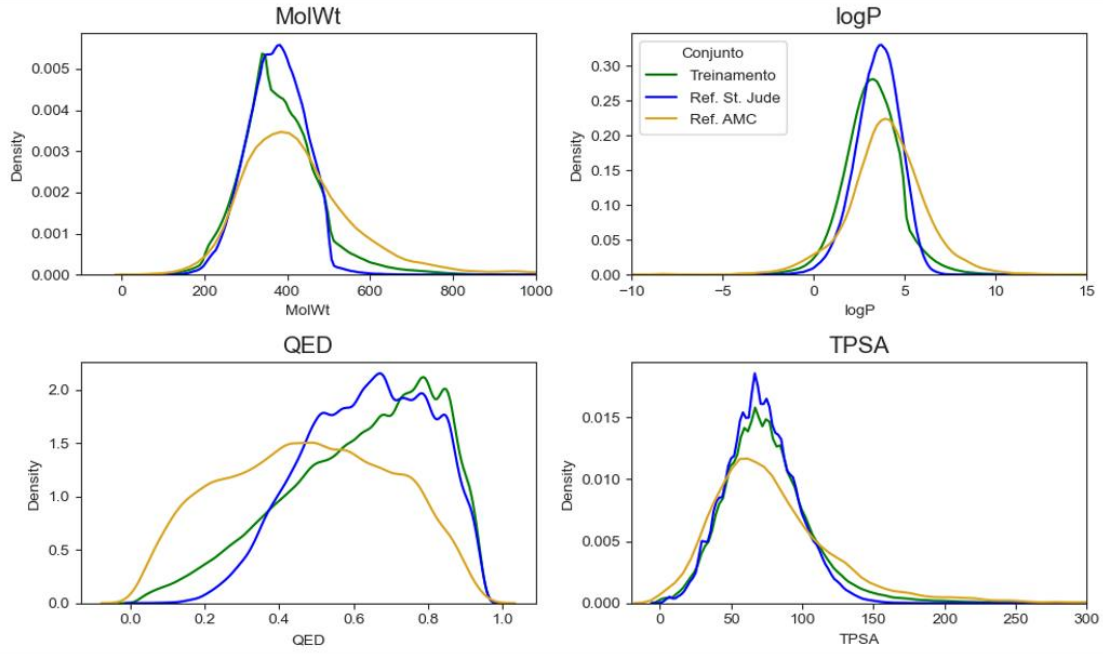


Figura 10 – Análise exploratória dos dados. Gráficos das probabilidades estimadas (Kernel Density Estimation) das propriedades físico-químicas das moléculas nos conjuntos. A curva em verde representa a probabilidades para o conjunto de pré-treinamento do VAE inicial, enquanto a curva em azul representa as probabilidades para o conjunto de refinamento MMV – St. Jude e a curva em amarelo, as probabilidades estimadas para o conjunto de refinamento antimaláricos ChEMBL (AMC). MolWt representa a massa molecular; logP o coeficiente de partição octanol-água; QED é o potencial do composto ser um fármaco (*quantitative estimation of drug-likeness*); e TPSA, a área de superfície topológica polar das moléculas.

Fonte: Elaborada pelo autor.

A Figura 10 mostra quatro propriedades físico-químicas das moléculas dos conjuntos de pré-treinamento e de refinamento utilizados. Esse tipo de representação considera a densidade de probabilidade que determinada característica seja encontrada em cada ponto da curva, cuja área é igual a um. Assim, é possível ver que a maior parte das moléculas pertencentes aos três conjuntos possuem a massa molecular menor do 500 Da, um coeficiente de partição octanol-água cujo pico é menor do que 5 e uma área de superfície topológica polar menor do 200 $Å^2$. Além disso, os conjuntos também foram avaliados com relação ao parâmetro *QED* (do inglês, *quantitative estimation of drug-likeness*),[116] um parâmetro que mede o potencial de um composto tem de ser considerado um fármaco, numa escala de 0,0 (o qual indica que todas as propriedades são desfavoráveis) a 1,0 (que indica que todas as propriedades da molécula são favoráveis). Na Figura 10 é possível observar distribuições cujos picos dos valores de QED estão mais próximos de 1,0 (acentuados entre 0,4 e 0,9) para os conjuntos de treinamento e MMV – St. Jude, enquanto para o

conjunto AMC as probabilidades estão distribuídas ao longo de todo o espectro permitido. Tal fato é um reflexo direto do tamanho do conjunto analisado e da aleatoriedade das moléculas geradas.

3.3.3 Predição da atividade inibitória

O modelo de grafos convolucionais foi treinado utilizando o conjunto de dados contendo as 32.176 moléculas filtradas do ChEMBL, cuja atividade inibitória contra o *P. falciparum* foi medida e reportada na base de dados. Tal modelo teve por objetivo principal predizer a atividade antiplasmodial das moléculas em termos de pIC_{50}. Para isso, o conjunto foi dividido em conjunto de treinamento e conjunto de teste (80% e 20% do conjunto original, respectivamente).

Após as 200 épocas de treinamento, o modelo alcançou um coeficiente de determinação de Pearson de 0,874 para o conjunto de treinamento. Esse coeficiente mede a força e a direção da relação linear entre os valores preditos e os valores reais, de modo que tal resultado indica uma forte correlação positiva, indicando que o modelo é capaz de entender os padrões e tendências do conjunto de treinamento e replicá-los com sucesso.

Já para o conjunto de teste, o modelo atingiu um coeficiente de determinação de Pearson de 0,639. Nesta aplicação, o coeficiente quantifica a habilidade que o modelo possui de generalizar e fazer predições acuradas para dados desconhecidos. Embora o coeficiente obtido para o teste tenha sido menor do que aquele obtido para a etapa de treinamento, ele ainda indica uma correlação positiva e moderada. Tal fato sugere que o modelo, de fato, possui poder de predição razoável e é capaz de proporcionar uma percepção significativa quando aplicado a dados novos.

3.3.4 Compostos Gerados

Após a convergência de cada um dos modelos refinados, 50 mil matrizes de vetores aleatórios 196-dimensionais, uniformemente amostrados do espaço latente, foram decodificadas em SELFIES. A originalidade (do inglês, *uniqueness*) dessas cadeias de caracteres SELFIES foi verificada comparando cada cadeia internamente ao conjunto, isto é, comparando-a com o conjunto gerado e excluindo quaisquer duplicatas. Tal procedimento resultou em aproximadamente 99% de originalidade

tanto para o conjunto gerado a partir do modelo refinado com o conjunto MMV – St. Jude quanto para o gerado pelo modelo refinado com os antimaláricos ChEMBL (Tabela 3).

Como era de se esperar, dada a premissa inicial da representação molecular utilizada,[58] a validade para ambos os conjuntos foi igual a 100%. Essa característica foi confirmada ao converter todos os SELFIES gerados para SMILES e verificá-los com o RDkit, recuperando moléculas sempre válidas.

Tabela 3 – Resultados dos conjuntos de 50 mil moléculas geradas: validade aferida pela conversão dos SELFIES gerados em SMILES que foram conferidos com RDKit; e originalidade (uniqueness) interna para cada um dos conjuntos gerados.

Conjunto de Refinamento	Validade (%)	Originalidade (%)
Antimaláricos ChEMBL	100	98,87
MMV – St. Jude	100	98,74

Fonte: Elaborada pelo autor.

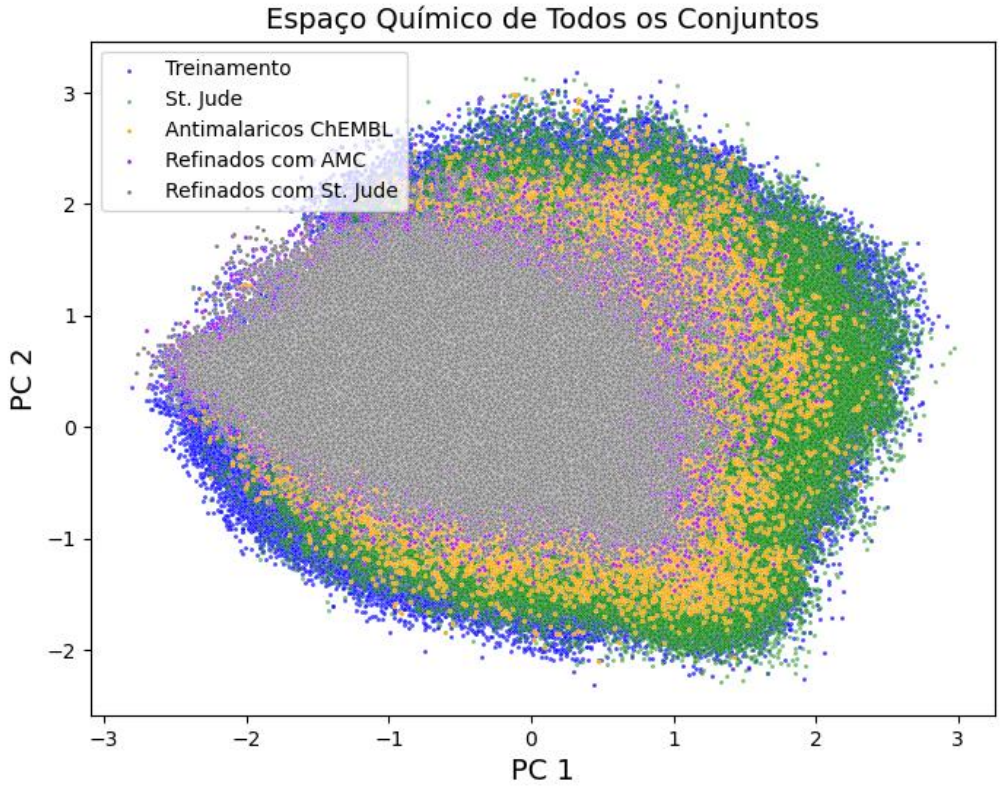


Figura 11 – Espaço químico de todos os conjuntos de dados. Análise das componentes principais (PCA) representando o espaço químico ocupado pelas moléculas dos conjuntos de treinamento que contém moléculas do ChEMBL e do ZINC15 (em azul), de refinamento MMV – St. Jude (em verde) e de refinamento/treinamento do modelo classificatório antimaláricos ChEMBL (em amarelo), e dos conjuntos gerados, com as moléculas geradas pelo modelo refinado com os antimaláricos ChEMBL em violeta e as moléculas geradas com o modelo refinado com MMV – St. Jude em cinza.

Fonte: Elaborada pelo autor.

A utilização de um conjunto de dados grande para treinar um modelo generativo se dá devido a necessidade de que esse modelo seja capaz de aprender características gerais acerca dos dados fornecidos a eles.[117-119] Existe uma sobreposição bastante grande entre os conjuntos gerados e os conjuntos de treinamento e refinamento (Figura 11), indicando que o modelo foi capaz de capturar e reproduzir as propriedades estruturais desses conjuntos.

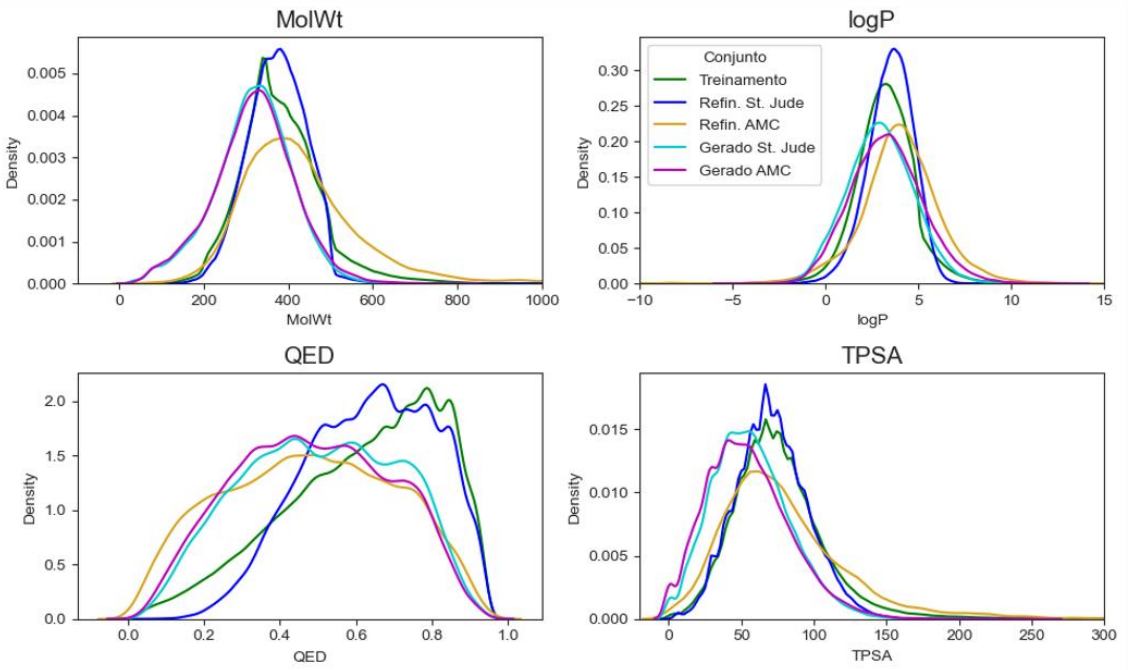


Figura 12 – Análise exploratória dos dados. Comparação dos gráficos das probabilidades estimadas (Kernel Density Estimation) das propriedades físico-químicas das moléculas nos conjuntos de treinamento e refinamento com os conjuntos gerados. A curva em verde representa a probabilidades para o conjunto de treinamento do VAE inicial, enquanto a curva em azul representa as probabilidades para o conjunto de refinamento MMV – St. Jude e a curva em amarelo, as probabilidades estimadas para o conjunto de refinamento antimaláricos ChEMBL (AMC). A curva em ciano representa as moléculas greadas pelo modelo refinado com MMV – St. Jude, enquanto que a curva em magenta representa as probabilidades estimadas para o conjunto gerado pelo modelo refinado com os antimaláricos ChEMBL. MolWt representa a massa molecular; logP o coeficiente de partição octanol-água; QED é o potencial do composto ser um fármaco (*quantitative estimation of drug-likeness*); e TPSA, a área de superfície topológica polar das moléculas.
Fonte: Elaborada pelo autor.

Assim como para os conjuntos de pré-treinamento e de refinamento, uma análise de quatro propriedades físico-químicas (e.g., massa molecular, log P. QED e TPSA) das moléculas geradas também foi conduzida (Figura 12). Ambos os conjuntos gerados apresentaram uma densidade de probabilidade maior de se encontrar moléculas com massa molecular menor do que 400 Da, com um valor de log P menor

que 5 e com uma área de superfície topológica polar menor que 100 Å^2. O parâmetro QED para ambos os conjuntos apresentou densidades de probabilidades distribuídas ao longo do espectro permitido, com o conjunto gerado pelo modelo refinado com MMV – St. Jude apresentando probabilidades levemente maiores de encontrar moléculas cujo parâmetro QED foi maior que 0,6 quando comparado ao conjunto gerado pelo modelo refinado com AMC. Esse perfil apresentado para ambos os conjuntos pode ser justificado tanto pelo tamanho deles (50 mil elementos) bem como à aleatoriedade das moléculas geradas, isto é, não foi aplicada restrição ou vieses que forçassem os vetores amostrados para decodificar moléculas a partir do espaço latente apenas com características desejadas.

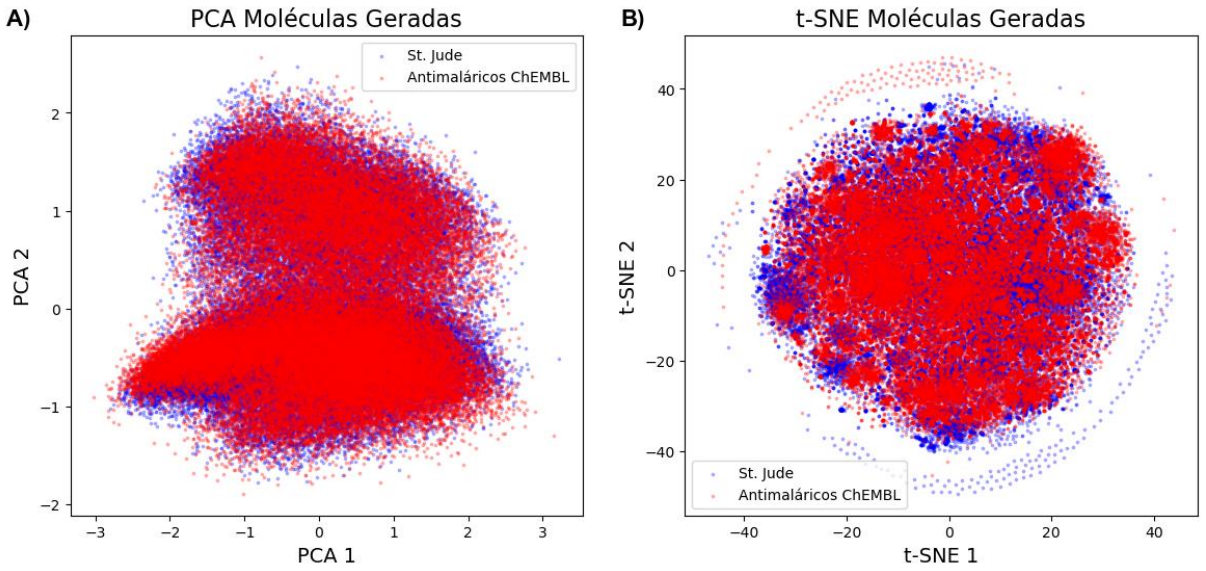


Figura 13 – Espaço químico dos conjuntos gerados (em azul, moléculas geradas a partir do modelo refinado com MMV – St. Jude, em vermelho, geradas pelo modelo refinado com antimaláricos ChEMBL), mostradas por meio da A) PCA e B) da distribuição t-SNE.
Fonte: Elaborada pelo autor.

Com o objetivo de comparar os espaços químicos dos conjuntos gerados por ambos os modelos refinados, foi feita a análise das duas primeiras componentes principais para esses conjuntos. A Figura 13A mostra que tanto as moléculas geradas ao refinar o modelo utilizando o AMC quanto as geradas usando St. Jude residem no mesmo espaço químico, indicando características semelhantes entre elas e indo ao encontro das propriedades vistas acima. Entretanto, como a fração geral das variâncias explicada pelas duas primeiras componentes principais para esses conjuntos não é muito grande (menos de 50%), distribuições de t-SNE foram feitas

para comparar ambos os espaços químicos (Figura 13B). Essas distribuições mostraram que existem regiões do espaço no qual as moléculas pertencentes a cada um dos conjuntos não se sobrepõem, indicando que existem diferenças efetivas entre as moléculas geradas pelos modelos refinados com os conjuntos diferentes. Além disso, é possível observar que as moléculas geradas a partir do conjunto MMV – St. Jude apresentaram uma cobertura maior do espaço quando comparado com aquelas geradas utilizando o AMC. Essa característica vai de encontro ao esperado, uma vez que o primeiro conjunto de refinamento mencionado é maior e apresenta diversidade estrutural interna significativa.

Embora exista diferença entre os conjuntos gerados e as porções do espaço químico que eles ocupam, uma análise qualitativa das distribuições de propriedades físico-químicas das moléculas pertencentes a esses conjuntos com relação ao espaço químico ocupado revelou que ambos se comportam de maneira muito semelhante (Figura 14). Ambos os conjuntos mostraram distribuição homogênea de massas moleculares, com um perfil geral de baixos valores para essa propriedade (Figura 14A e 14E). Ambos os conjuntos também apresentaram distribuição homogênea de coeficientes de partição octanol-água, com valores predominantemente abaixo de 5 (Figura 14B e 14F). As moléculas geradas utilizando ambos os conjuntos de refinamento apresentaram um certo gradiente de QED ao longo do espaço, com o AMC contendo duas regiões (à esquerda e à direita gráfico t-SNE) cujo potencial do composto ser um fármaco foi mais elevado (Figura 14C) enquanto que o MMV – St. Jude apresentou gradiente mais suave ao longo do espaço (Figura 14G). Por fim, ambos os conjuntos gerados apresentaram uniformidade na distribuição dos valores de pIC_{50} preditos para suas moléculas (Figura 14D e 14H), com o conjunto gerado pelo AMC mostrando-se levemente mais potente, uma vez que esse conjunto possui quantidade maior de moléculas classificadas como ativas (mais de 15 mil moléculas com $pIC_{50} > 6,0$).

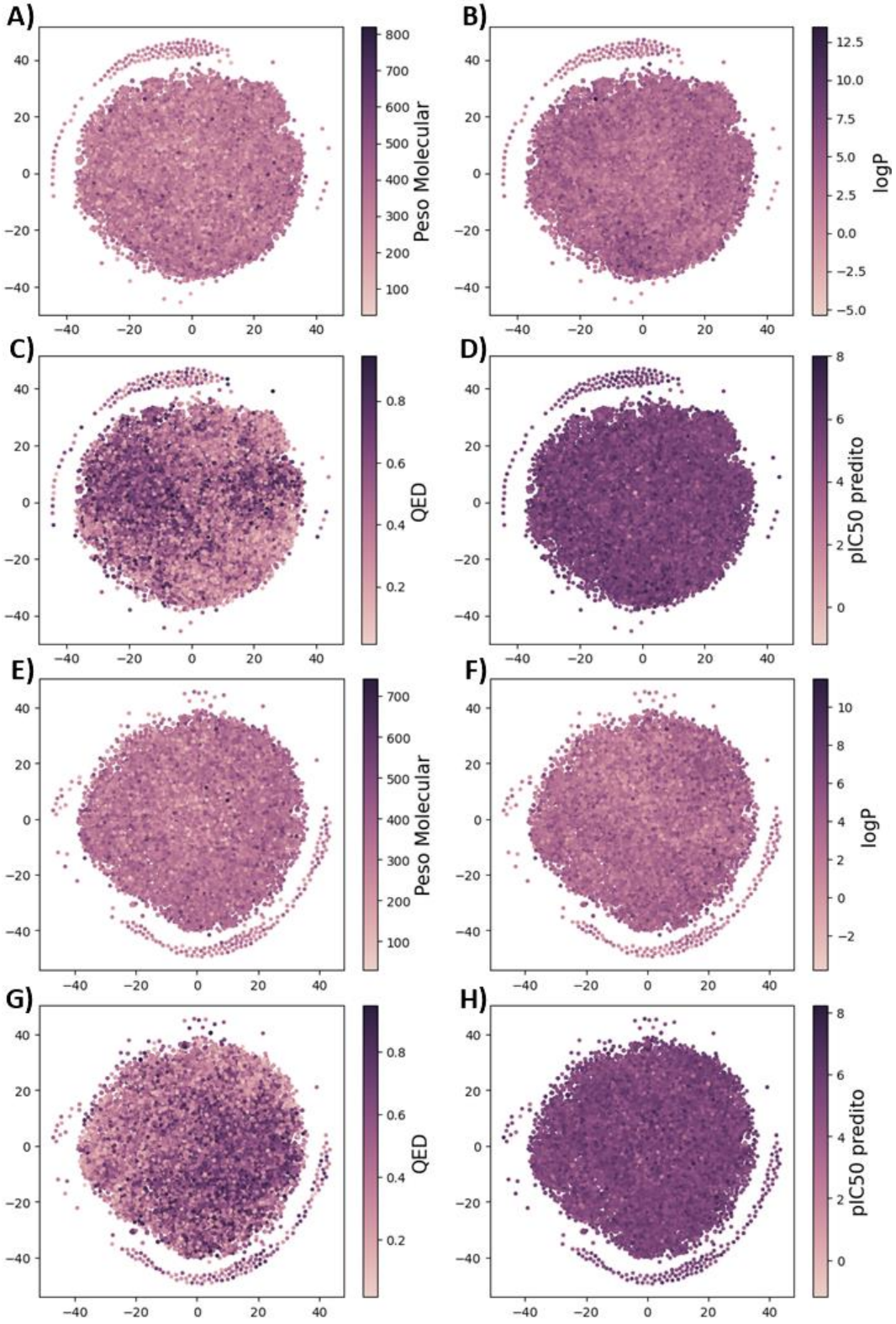


Figura 14 – Distribuição t-SNE coloridas pelas propriedades físico-químicas das moléculas geradas pelo modelo refinado com antimaláricos ChEMBL (A – D) e MMV – St. Jude (E – F). A) e E) Distribuição das massas moleculares de cada composto do conjunto gerado. B) e F) Distribuição do coeficiente de partição octanol-água. C) e G) Potencial do composto ser um fármaco (*quantitative estimation of drug-likeness*, ou, *QED*). D) e H) pIC_{50} predito utilizando o modelo de regressão treinado.

Fonte: Elaborada pelo autor.

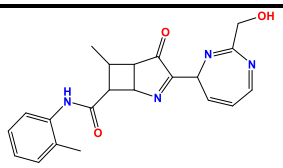
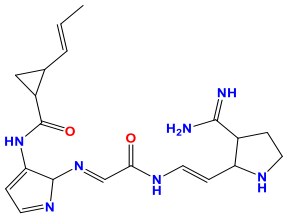
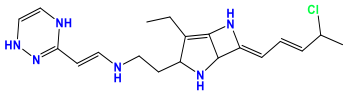
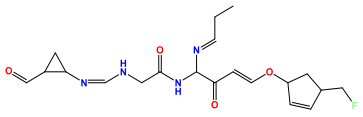
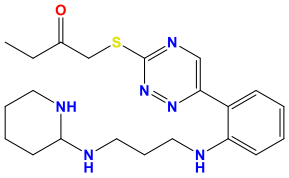
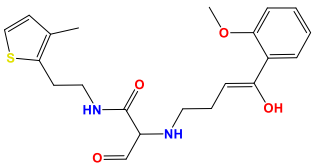
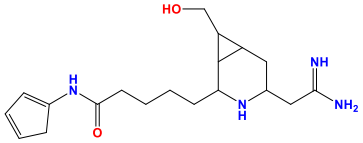
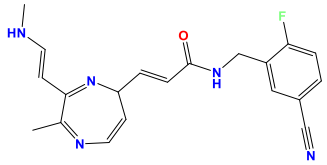
3.3.5 Análise visual dos compostos

As 50 mil moléculas geradas por cada um dos modelos refinados foram filtradas de modo que apenas moléculas que passassem pelos seguintes critérios fossem levadas a diante com as análises: apenas moléculas cujo pIC_{50} predito fosse maior que 7, cuja massa molecular estivesse entre 350 e 450 Da e cujo valor de log P fosse menor que 3 foram selecionadas.

Após a filtragem, restaram 40 moléculas obtidas por meio da amostragem no espaço gerado ao refinar o modelo com os antimaláricos ChEMBL e 70 moléculas geradas ao refinar o modelo com a base MMV – St. Jude. Além das propriedades físico-químicas desses compostos, eles foram inspecionados visualmente, a fim de remover aqueles que, à luz das propriedades calculadas, parecessem promissores ou cujas moléculas análogas a eles poderiam ser candidatos a compostos líderes contra a malária. Nessa etapa, critérios como a complexidade dos anéis presentes na estrutura, a existências de grupos reativos (como os aldeídos) e a presença de microciclos complexos foram levados em consideração para a exclusão de moléculas.

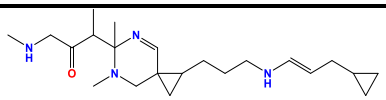
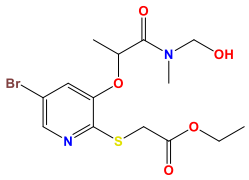
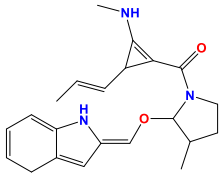
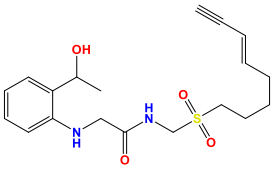
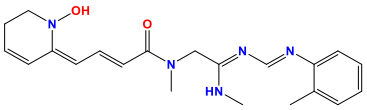
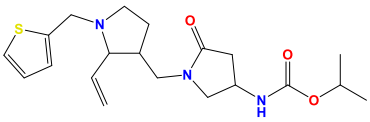
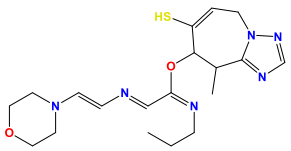
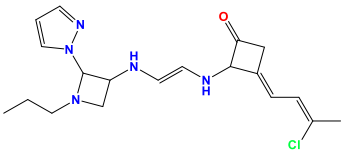
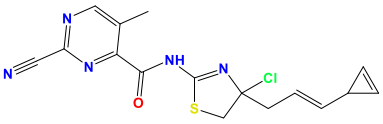
A inspeção resultou em 20 moléculas (oito do conjunto refinado com antimaláricos ChEMBL e 12 do conjunto refinado com MMV – St. Jude) com as seguintes características: tamanho adequado para modificações e otimizações futuras, massas moleculares entre 352 e 413, valor de log P entre 0,22 e 2,92, acessibilidade sintética entre 3,33 e 5,80 e valor de pIC_{50} predito entre 7,00 e 8,12 (Tabela 4 e Tabela 5).

Tabela 4 – Moléculas geradas com antimaláricos ChEMBL. Moléculas geradas ao refinar o modelo com os antimaláricos ChEMBL e filtrar por massa molecular entre 350 Da e 450 Da, logP < 3 e pIC50 > 7. Aqui, MW significa “massa molecular”; logP é o logaritmo coeficiente de partição octanol-água das moléculas; QED indica o potencial do composto ser um fármaco (quanto mais próximo de de 1,0, maior o potencial); SAScore é a acessibilidade sintética (numa escala de 1 a 10, na qual 1 significa ser mais fácil de sintetizar); TPSA é a área de superfície polar topológica; e pIC_{50} é o valor de potência contra o *Plasmodium falciparum* predito pelo modelo para o composto.

Código	Estrutura	MW	logP	QED	SAScore	TPSA	pIC_{50}
LaBEFar_VAE_01		378,4	1,61	0,83	4,68	103,5	7,02
LaBEFar_VAE_02		397,5	0,22	0,23	5,72	144,8	7,04
LaBEFar_VAE_03		374,9	2,13	0,33	5,75	72,5	7,07
LaBEFar_VAE_04		406,5	1,13	0,12	5,22	100,1	7,08
LaBEFar_VAE_05		413,6	2,92	0,38	3,86	91,8	7,17
LaBEFar_VAE_06		402,5	2,87	0,23	3,47	87,7	7,18
LaBEFar_VAE_07		346,5	1,41	0,25	4,91	111,2	7,24
LaBEFar_VAE_08		365,4	2,40	0,76	4,15	89,6	7,57

Fonte: Elaborada pelo autor.

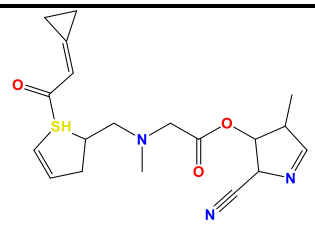
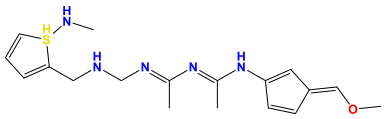
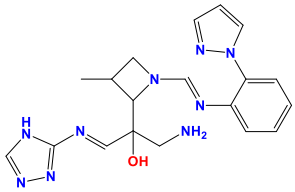
Tabela 5 – Moléculas geradas com antimaláricos St. Jude. Moléculas geradas ao refinar o modelo com os antimaláricos ChEMBL e filtrar por massa molecular entre 350 Da e 450 Da, logP < 3 e pIC50 > 7. Aqui, MW significa "massa molecular"; logP é o logaritmo coeficiente de partição octanol-água das moléculas; QED indica o potencial do composto ser um fármaco (quanto mais próximo de de 1,0, maior o potencial); SAScore é a acessibilidade sintética (numa escala de 1 a 10, na qual 1 significa ser mais fácil de sintetizar); TPSA é a área de superfície polar topológica; e pIC_{50} é o valor de potência contra o *Plasmodium falciparum* predito pelo modelo para o composto.

Código	Estrutura	MW	logP	QED	SAScore	TPSA	pIC_{50}
LaBEFar_VAE_9		374,6	2,83	0,54	5,59	56,73	7,00
LaBEFar_VAE_10		405,3	1,88	0,37	3,33	85,80	7,04
LaBEFar_VAE_11		365,5	1,54	0,78	5,31	57,36	7,04
LaBEFar_VAE_12		378,5	2,00	0,40	3,41	95,50	7,09
LaBEFar_VAE_13		381,5	2,82	0,45	3,90	80,53	7,11
LaBEFar_VAE_14		391,5	2,86	0,73	3,99	61,88	7,18
LaBEFar_VAE_15		404,5	2,28	0,45	5,11	77,13	7,19
LaBEFar_VAE_16		375,90	2,54	0,73	5,26	62,19	7,21
LaBEFar_VAE_17		359,8	2,56	0,51	4,42	91,03	7,28

(continua)

(continuação)

Tabela 5 – Moléculas geradas com antimaláricos St. Jude. Moléculas geradas ao refinar o modelo com os antimaláricos ChEMBL e filtrar por massa molecular entre 350 Da e 450 Da, logP < 3 e pIC50 > 7. Aqui, MW significa “massa molecular”; logP é o logaritmo coeficiente de partição octanol-água das moléculas; QED indica o potencial do composto ser um fármaco (quanto mais próximo de de 1,0, maior o potencial); SAScore é a acessibilidade sintética (numa escala de 1 a 10, na qual 1 significa ser mais fácil de sintetizar); TPSA é a área de superfície polar topológica; e pIC_{50} é o valor de potência contra o *Plasmodium falciparum* predito pelo modelo para o composto.

Código	Estrutura	MW	logP	QED	SAScore	TPSA	pIC_{50}
LaBEFar_VAE_18		375,5	1,98	0,44	5,80	82,76	7,33
LaBEFar_VAE_19		361,51	2,49	0,18	5,43	70,04	7,64
LaBEFar_VAE_20		393,46	1,06	0,41	4,99	133,6	8,12

Fonte: Elaborada pelo autor.

Uma busca nas plataformas *SciFinder*^{n}, *ChEMBL*, *ZINC15*, *Enamine* e *eMolecules* indicou que nenhuma dessas moléculas geradas estão disponíveis para compra imediata. Tal fato, aliado a ausência dessas moléculas nos conjuntos de pré-treinamento e de refinamento indicam que o modelo possui elevada capacidade de gerar moléculas únicas. Assim, será feita uma busca por colaboradores que sejam capazes de sintetizar essas moléculas pré-selecionadas bem como sintetizar análogos. Nesta tarefa, plataformas como *Manifold* da *Postera*[120] e *Brood* da *OpenEye*[121] serão bastante úteis, uma vez que elas são capazes de sugerir rotas sintéticas bem como fornecedores dos reagentes para cada composto submetido, além disso, essas plataformas permitem a análise de substituintes para as moléculas fornecidas.

Embora não tenha sido possível encontrar moléculas com disponibilidade imediata, a consulta ao SciFinder^n foi a única que retornou compostos similares, em

termos do coeficiente de Tanimoto, com similaridades variando entre 61% (para o composto LaBEFar_VAE_13) e 83% (para o LaBEFar_VAE_06) (Figura 15). Apesar de uma molécula com 83% de similaridade ter sido encontrada (CAS RF: 1287207-83-3), as diferenças estruturais entre eles ainda são grandes. Entretanto, ainda é possível utilizá-la como ponto de partida ou como medida de comparação e guia para ensaios que serão realizados no futuro. A tabela completa com as moléculas similares pode ser encontrada nos **ANEXOS**.

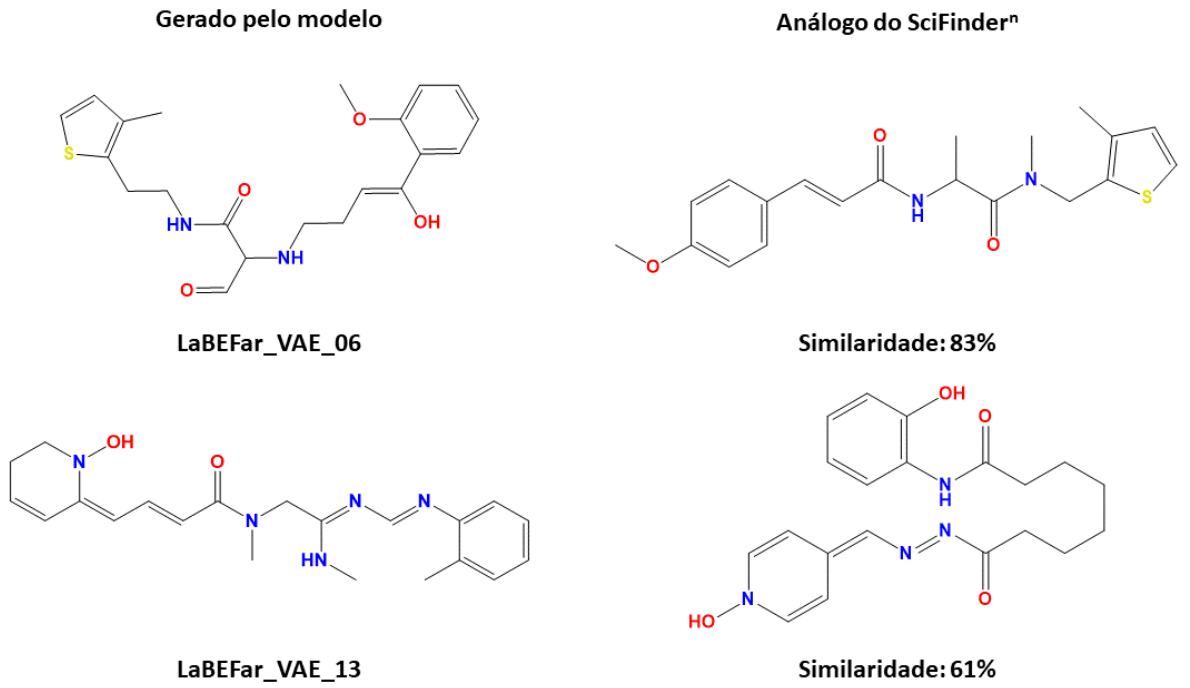


Figura 15 - Análogos encontrados pela busca na plataforma SciFinderⁿ. À esquerda, os compostos gerados pelo modelo generativo cujos respectivos análogos com a maior e menor similaridade (à direita) estão disponíveis para compra e foram listados na plataforma de busca de moléculas.

Fonte: Elaborada pelo autor.

3.4 DISCUSSÃO

Os antimaláricos são alternativas terapêuticas essenciais para a manutenção e combate à doença. O surgimento de mecanismos de resistência às terapias disponíveis tem colocado em risco a atual situação de controle da malária.[80-82] Assim, a busca constante por novos compostos que tenham o potencial de inibir o *P. falciparum* se faz urgente e necessária. Nesse contexto, a utilização de métodos computacionais e, principalmente, de modelos generativos de aprendizado profundo,

pode ajudar a alavancar e acelerar o processo de busca por novos potenciais inibidores da atividade do parasito. Esses tipos de modelos podem, inclusive, contribuir para a descoberta de novas classes de compostos que não foram ainda investigadas.

O primeiro ponto a ser levado em consideração ao utilizar esses modelos são os dados sobre os quais eles serão treinados. Neste trabalho, foram utilizadas moléculas retiradas de bases de dados distintas na forma de SELFIES, cujas propriedades físico-químicas se mostraram bastante interessantes, uma vez que apresentaram características que obedecem às regras de Lipinski.[122-123] Tais regras servem como guia para predizer se uma molécula biologicamente ativa terá uma boa biodisponibilidade oral. Essas características, mesmo não sendo fornecidas diretamente para o modelo, auxiliaram-no a compreender e reproduzir as propriedades de interesse nos conjuntos gerados (Figura 12), com as moléculas geradas seguindo, de fato, a distribuição de propriedades dos conjuntos de refinamento. É válido mencionar que os perfis do parâmetro QED para ambos os conjuntos gerados e para o AMC se assemelham. Isso é um reflexo direto do tamanho do conjunto, dado que a existência de uma quantidade menor de moléculas, se o conjunto não possuir um viés de propriedades, permite que as densidades de probabilidade fiquem mais distribuídas ao longo de todo o espectro da propriedade possível.

Embora o modelo de regressão tenha apresentado um coeficiente de determinação de Pearson de 0,87, indicando uma forte correlação positiva para o modelo, é necessário levar em conta que os dados nos quais ele foi treinado não eram “puros” no quesito “cepa na qual o composto foi testado”, isso é, os valores de atividades reportados na base de dados curada do ChEMBL v.29 foram obtidos a partir de diferentes ensaios que usaram cepas com características diferentes de resistência e sensibilidade à antimaláricos, bem como ensaios diferentes para a determinação da propriedade biológica. A escolha de combinar esses dados foi feita para que houvesse um conjunto de dado de tamanho adequado para o treinamento do modelo, mas é esperado que as próximas etapas de desenvolvimento do modelo contemplem uma diferenciação entre as cepas bem como a predição daquela que mais se adequa ao composto investigado. Verras *et al.*[104] já demonstrou que a combinação de dados de

diferentes cepas de *P. falciparum* pode gerar resultados decentes no processo de classificação e ordenamento de potenciais antimaláricos. Além disso, uma validação externa, utilizando os compostos do conjunto MMV – St. Jude de moléculas testadas apenas na cepa 3D7, resultou numa taxa de acerto de cerca de 73%. É válido dizer que, uma vez que o conjunto em questão não possuía valores de atividade reportados, o critério para definir moléculas ativas e inativas foi o mesmo utilizado anteriormente ($pIC_{50} \leq 5{,}523 \rightarrow$ inativa; $pIC_{50} \geq 6{,}000 \rightarrow$ ativa; $5{,}523 < pIC_{50} < 6{,}000 \rightarrow$ descartar).

Um aspecto crucial a ser considerado quando se trabalha com modelos generativos de moléculas que utilizam cadeias de caracteres como o tipo dos dados de entrada é a validade das moléculas que são geradas. Diversos modelos generativos da literatura que utilizam a representação molecular em cadeia de caracteres mais difundida (*e.g.*, SMILES) reportam validade das moléculas geradas abaixo de 100%.[22,58,119,124-127] Esse problema se dá principalmente à necessidade da combinação par-a-par que essa representação possui e que os modelos generativos nem sempre são capazes de compreender e reproduzir. Neste trabalho, a validade das moléculas geradas foi total devido à representação SELFIES ter sido escolhida para as entradas do modelo. Os SELFIES surgiram em 2020 e têm ganhado cada vez mais destaque em quimioinformática,[128-129] provando ser uma representação extremamente robusta, apropriada para aplicação em modelos de aprendizado de máquina e com potencial bastante grande de substituir as representações em cadeias de caracteres mais populares. Novas versões dessa representação estão em desenvolvimento e devem melhorar ainda mais o desempenho de modelos que a utilizam.[130]

Ao todo, dois conjuntos com 50 mil moléculas cada foram geradas (utilizando cada um dos conjuntos de refinamento). Ambos os conjuntos gerados apresentaram características que seguem a distribuição de propriedades dos dados sobre os quais foram treinados. Uma vez que o conjunto MMV – St. Jude possuía quantidade maior de moléculas com diversidade estrutural, era intuitivo esperar também maior diversidade para os dados gerados e refinados pelo modelo construído a partir dele. Esse fato ficou evidenciado pela representação t-SNE (Figura 13) com as moléculas geradas com MMV – St. Jude agrupadas em mais regiões do espaço do que aquelas ocupadas por moléculas geradas usando o AMC. Ainda assim, ambos apresentaram

os mesmos perfis de propriedades físico-químicas, condizentes com as características necessárias para se obter um potencial candidato a fármaco antimalárico.

Embora a distribuição de propriedades físico-químicas e o espaço químico tenham se sobreposto em praticamente todas as regiões, o modelo foi capaz de gerar algumas moléculas fora do espaço químico usado para treiná-lo e refiná-lo, margeando-os (Figura 11). Isso revela um potencial aprendido de extrapolar as restrições impostas pela informação inicialmente fornecida para o modelo. Uma vez que modelos VAE treinados utilizando SELFIES como representação molecular apresentam um espaço latente mais denso (*i.e.*, diversidade de estruturas cerca de 100 vezes maior que modelos treinados com outras representações)[56] decodificar moléculas cuja projeção no espaço químico esteja fora do conjunto de treinamento não é inesperado. Apesar disso, uma análise criteriosa dessas moléculas e da região do espaço à qual elas pertencem será feita a fim de determinar precisamente suas características e como elas se relacionam com os demais conjuntos.

Além das propriedades mencionadas anteriormente (tamanho adequado para modificações e otimizações futuras, massas moleculares entre 352 e 413 Da, valor de log P entre 0,22 e 2,92, acessibilidade sintética entre 3,33 e 5,80 e pIC_{50} predito entre 7,00 e 8,12), a maior parte das moléculas selecionadas após a inspeção visual possuía um ou mais nitrogênios básicos (*e.g.*, p*K*a > 7), 16 dos 20 compostos apresentaram pelo menos um nitrogênio básico em sua estrutura, de acordo com a estimativa feita pelo *software Datawarrior*[131] (**ANEXOS**). Essa característica estrutural tem se mostrado essencial para novos candidatos a antimaláricos de estágio avançado, cuja atividade contra o *P. falciparum* se encontram nas faixas de baixo μM e nM.[132-133]

Além disso, uma análise qualitativa externa de predição de atividade das 20 moléculas selecionadas foi feita por meio da plataforma MAIP (*Malaria Inhibitor Predictor*, do ChEMBL).[134] Essa plataforma utiliza um modelo de consenso para predizer inibidores de malária no seu estágio sanguíneo, e classifica os compostos submetidos a ela de forma não global, ou seja, as classificações feitas por ela valem apenas para o conjunto de moléculas submetidos e servem como base para o aprimoramento de tais moléculas.

A plataforma classificou esses 20 compostos de acordo com maior potencial de serem inibidores da malária. Dentre eles, a plataforma indicou o composto

LaBEFar_VAE_07 como o mais promissor para inibir a malária no estágio de desenvolvimento sanguíneo, o qual foi gerado utilizando o AMC como conjunto de refinamento. Tal composto foi o sexto melhor pontuado em termos de pIC_{50} predito pelo modelo e possui dois nitrogênios básicos (Tabela 4).

O LaBEFar_VAE_07 possui um grupo (3-azabiciclo[4.1.0]heptan-7-il)metanol, o qual é reportado na literatura como associado à inibição da recaptação de serotonina, dopamina e noradrenalina.[135-136] Até onde se tem conhecimento, não há relatos de antimaláricos que contenham esse grupo, o qual pode servir como ponto de partida para o desenvolvimento de uma nova classe de inibidores da malária.

A análise das demais moléculas selecionadas indicou que embora as propriedades físico-químicas, bem como a atividade predita sejam satisfatórias para um candidato a antimalárico, a acessibilidade sintética não é tão boa quanto possível, dada a presença de grupos que podem dificultar o processo de síntese dos compostos bem como sua disponibilidade imediata. Esse último fato foi verificado pela inexistência desses compostos nas plataformas utilizadas.

Uma estratégia imediata para contornar esse problema é a utilização de uma combinação de *softwares*, como o *BROOD*, da *OpenEye*,[121] e o *Manifold*, da *Postera*.[120] A primeira ferramenta gera análogos de compostos líderes ou *hits* trocando fragmentos da molécula por fragmentos que têm forma e propriedades eletrostáticas similares, mas que podem ter outras propriedades distintas. A segunda propõe rotas sintéticas para a obtenção dos compostos que são submetidos a ela, bem como os passos necessários e uma lista de fornecedores que possuem os reagentes que devem ser utilizados durante a síntese. Nas Figura 16 e Figura 17 são apresentados dois exemplos: o composto LaBEFar_VAE_08 e o LaBEFar_VAE_18, cuja acessibilidade sintética de ambos é relativamente desfavorável, mas com bons valores de atividade preditos.

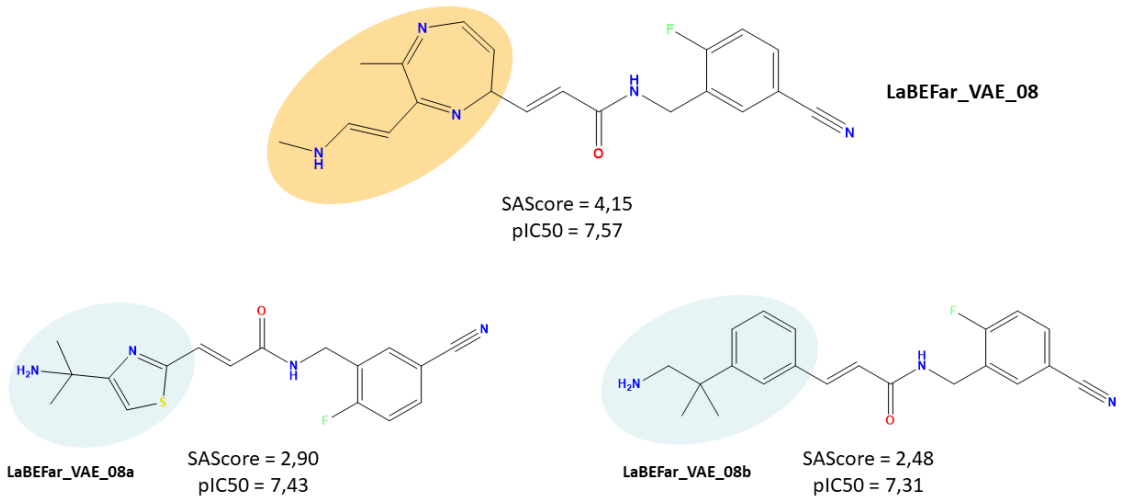


Figura 16 – Substituição do grupo (E)-N-metil-2-(2-metil-5H-1,4-diazepin-3-il)eten-1-amina (em amarelo) do composto gerado LaBEFar_VAE_08, com o resultado da substituição sugeridos pelo software BROOD, da OpenEye, destacados em azul.
Fonte: Elaborada pelo autor.

Ao submeter o LaBEFar_VAE_08 (Figura 16) ao *Manifold*, uma rota contendo nove passo foi retornada. Tal rota pode ser impraticável, a depender dos colaboradores que o sintetizarão bem como dos preços dos reagentes para obtê-lo. Entretando, os análogos resultantes do *BROOD*, LaBEFar_VAE_08a (pIC_{50} predito de 7,43) e LaBEFar_VAE_08b (pIC_{50} predito de 7,31), possuem valores do parâmetro SAScore de 2,90 e 2,48, respectivamente, que são consideravelmente menores que o LaBEFar_VAE_08 (SAScore = 5,80). Esse fato ficou evidenciado ao submetê-los ao *Manifold*, que retornou rotas com dois e um passos, respectivamente.

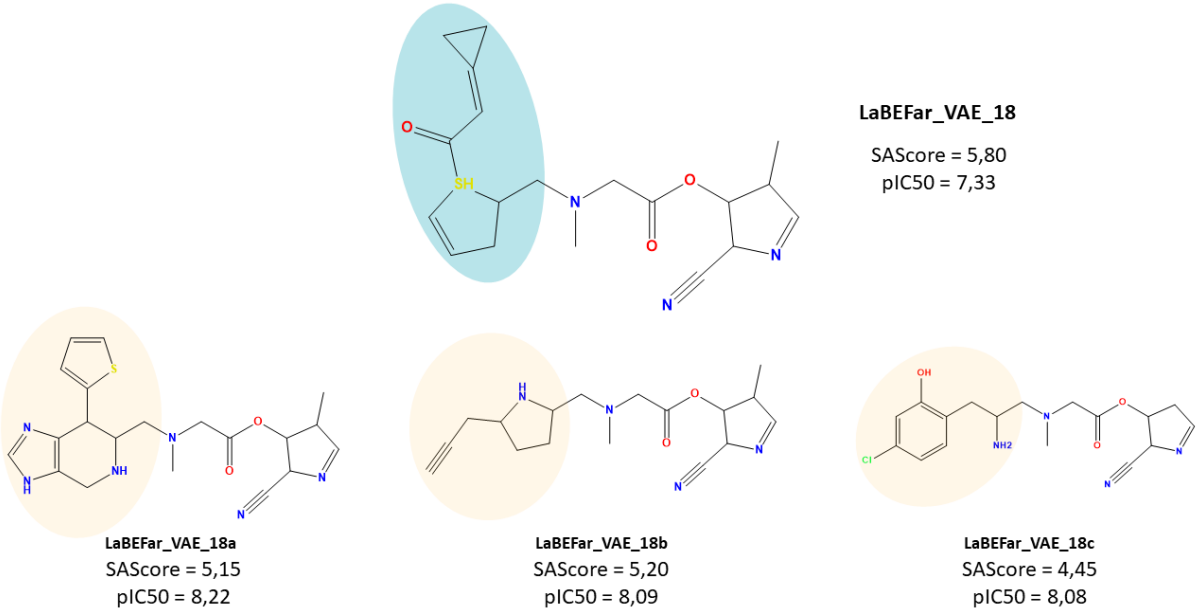


Figura 17 – Substituição de grupo 2-ciclopropilidene-1-(2,3-di-hidro-1l4-tiofen-1-il)etan-1-ona (em azul) do composto gerado LaBEFar_VAE_18, com o resultado da substituição sugeridos pelo software BROOD, da OpenEye, destacado em laranja.
Fonte: Elaborada pelo autor.

O composto LaBEFar_VAE_18 (Figura 17) possui um grupo metilenociclopropano, o qual aumenta significativamente a dificuldade de síntese da molécula. Embora não seja um grupo frequentemente visto em fármacos, principalmente em antimaláricos, esse grupo já foi reportado em moléculas que estão em fase clínica de estudo para o combate do vírus do herpes simples e do citomegalovírus em humanos.[137-138] Esse fato reforça a capacidade do modelo de extrapolar as propriedades e características das moléculas fornecidas durante o treinamento e de gerar classes novas com o propósito de inibir a malária.

Submetendo o LaBEFar_VAE_18 ao *Manifold*, a plataforma retornou uma rota sintética contendo oito passos. Entretanto, os análogos LaBEFar_VAE_18a, LaBEFar_VAE_18b e LaBEFar_VAE_18c, obtidos pelo *BROOD*, obtiveram rotas com sete, seis e cinco passos, respectivamente, o que torna suas sínteses mais factíveis. Além disso, é interessante notar que houve um aumento da atividade predita para esses três compostos análogos.

Do total de 100 mil moléculas geradas, 20 passaram por todas as etapas de filtragem e inspeção visual, a fim se se obter apenas aqueles que fossem mais promissores. Embora a inspeção visual tenha resultado em compostos com grande potencial, a seleção feita pelas propriedades físico-químicas calculadas resultou numa

série de moléculas que possuíam subestruturas incomuns e/ou instáveis quando se pensa em fármacos. Isso se deu devido a características intrínsecas ao autocodificadores variacionais. Os autocodificadores podem produzir moléculas que são válidas quimicamente, mas que possuem substituintes cuja presença é indesejável em um fármaco, seja por sua estabilidade ou devido à dificuldade de sintetizá-las. É o caso, por exemplo, de grupos como os enóis éteres, os anidridos, as aziridinas, as enaminas, os ciclo-heptatrienos, entre outros.[22] Uma alternativa para contornar esse problema é treinar o modelo para que ele seja capaz de predizer propriedades estruturais indesejáveis. O modelo construído e treinado nesta tese não possuía nem uma etapa de ordenação proposital das propriedades físico-químicas do espaço latente, isto é, a organização desse espaço se deu única e exclusivamente pelo treinamento do codificador e do decodificador simultaneamente, maximizando a função objetivo escolhida. Uma organização inerente a essa arquitetura se faz presente, de modo que decodificar vetores 196-dimensionais das vizinhanças das moléculas geradas resulta em novas moléculas cujas propriedades são semelhantes à “inicial”. Entretanto, se a intenção fosse criar um gradiente de propriedades nesse espaço, seria necessário incorporar essa tarefa à função objetivo do modelo e otimizá-la durante o treinamento. Essa abordagem já se mostrou promissora na literatura e deve ser uma evolução natural deste trabalho durante os próximos passos do aprimoramento do modelo utilizado.[22]

É importante dizer que, no momento, as moléculas já podem ser geradas sob demanda, isto é, é possível gerar uma quantidade extremamente grande de novas moléculas antes de saturar todo o espaço químico do modelo. Dessa forma, é possível guiar a busca e, consequentemente, a exploração dos potenciais compostos do modelo seguindo uma distribuição de propriedades desejadas, sem necessariamente precisar informar o modelo sobre essas propriedades no momento da geração. Tal geração sob demanda poderá facilitar a resposta positiva dos colaboradores que irão sintetizar essas moléculas.

3.5 CONCLUSÕES E PERSPECTIVAS

A utilização de modelos generativos de aprendizado profundo para o desenvolvimento de novas estruturas moleculares oferece um futuro promissor (e não distante) para o ramo da descoberta de fármacos frente aos custos para o desenvolvimento de novos medicamentos, que aumentam continuamente. O crescente avanço do poder computacional disponível, aliado ao também crescente acesso a dados de quimioinformática, tornam a utilização de abordagens de aprendizado de máquina cada vez mais factíveis e necessárias.

Neste trabalho, o foco principal foi a utilização de um modelo generativo para gerar moléculas com potencial de inibir a atividade do *P. falciparum*. O modelo utilizou uma arquitetura de autocodificadores variacionais (VAE) combinadas com camadas LSTM e GRU. Esse modelo foi capaz de gerar compostos 100% válidos, com originalidade maior que 98% e com grau de novidade bastante elevado. Isso foi possível, principalmente, devido à utilização dos SELFIES como representação molecular. Além disso, o modelo foi capaz de extrapolar o espaço químico no qual foi treinando e incorporar às moléculas geradas grupos que até o presente momento não foram explorados contra o parasito causador da malária (ou não foram reportados na literatura), indicando um potencial de desenvolvimento de novas classes inéditas.

Das 100 mil moléculas geradas, 20 foram selecionadas após filtragem e inspeção visual para destacar. A inspeção visual revelou compostos com alto potencial, porém, a seleção baseada nas propriedades físico-químicas calculadas resultou em moléculas com subestruturas indesejáveis para aplicações farmacêuticas, devido a características intrínsecas aos VAEs. Estes são capazes gerar moléculas quimicamente válidas, mas com substituintes incomuns para fármacos. Uma alternativa para etapas futuras seria treinar o modelo para prever essas propriedades indesejáveis e penalizar quando foram incorporadas a ele bem como garantir que exista um gradiente de propriedades no seu espaço latente. Apesar disso, o modelo permite a geração sob demanda de moléculas em grande quantidade, facilitando a busca e exploração de potenciais compostos seguindo uma distribuição de propriedades desejadas. Este avanço promissor pode agilizar a resposta dos colaboradores na síntese dessas moléculas.

Uma outra abordagem que será feita numa etapa futura é utilizar o modelo generativo como uma ferramenta para gerar fragmentos que contenham *scaffolds* cuja

atividade contra o *P. falciparum* seja conhecida, bem como a utilização dos *scaffolds* presentes nas moléculas melhores pontuadas para a busca de substituintes de fácil acesso sintético ou fácil aquisição. O sucesso de tal abordagem já foi demonstrado por Saramago ***et al***.[139] que combinou modelos generativos e classificatórios para encontrar inibidores da protease principal de SARS-CoV-2.

Por fim, é importante dizer que alinhar a metodologia utilizada aqui com outras abordagens disponíveis é uma forma de maximizar o potencial de sucesso de encontrar um composto que possua propriedades físico-químicas atrativas, sinteticamente acessível e potente inibidor do parasita. Essas moléculas auxiliarão no combate da doença bem como auxiliarão no controle da parasitose.

CAPÍTULO 4: GERAÇÃO DE ANOMALIAS

4 GERAÇÃO DE ANOMALIAS

4.1 Panorama geral

Os conjuntos de dados do mundo real, desde a telecomunicação até a saúde, são constantemente contaminados por dados anômalos ou dados discrepantes, que se desviam significativamente do esperado e precisam ser removidos antes de serem aplicados numa modelagem.[140-142] Isso exige o desenvolvimento de modelos robustos (e implementáveis) de detecção de dados anômalos para supervisionar uma pré-limpeza desses dados ou que sejam capazes de acionar alarmes em sistemas que processam dados dinamicamente, como navegação na rede, detecção de *spams* ou de fraude de cartão de crédito.[143] No entanto, as anomalias podem ser o objeto de interesse em alguns casos, o que faz com que a investigação passe da detecção e remoção de anomalias para sua geração e incorporação ativas. Por exemplo, em algumas áreas, desenvolver modelos capazes de detectar anomalias pode ser uma tarefa bastante complicada dada a escassez de dados de treinamento, limitando seu potencial de previsão. Nesses casos, a geração de anomalias para preencher conjuntos de dados de treinamento sintéticos pode ser uma abordagem promissora para atenuar a escassez desses dados e o desequilíbrio de classes decorrente dessa escassez.[144]

Em outros casos, a própria natureza dos possíveis dados anômalos pode ser desconhecida devido à ausência de uma fonte de dados no mundo real ou mesmo à ineficácia de uma formulação geral ou de abordagens capazes de elaborar possíveis critérios que indiquem anomalias.[145] Por exemplo, os processos de fabricação geralmente consistem em etapas de filtragem para eliminar peças defeituosas anômalas.[146] No entanto, isso depende do conhecimento dos critérios de indicação da peça defeituosa, os quais podem ser insuficientes devido à raridade da produção de peças defeituosas. Nesses casos, utilizar o aprendizado de máquina para a geração de anomalias pode revelar possíveis modos de falha nos dados. Além disso, essa capacidade de geração pode permitir a construção e a exploração de um "espaço de dados anômalos" para revelar critérios até então desconhecidos, ou até mesmo inconcebíveis para a natureza anômala, e que podem ser utilizados para a modelagem da detecção de anomalias. No caso específico da fabricação de peças, o

conhecimento prévio dos critérios anômalos de uma nova peça defeituosa pode permitir a sua detecção após a produção. Da mesma forma, explorar o espaço de dados anômalos também pode revelar modos de falha no tipo de representação utilizada para representar os dados ao detectar um comportamento inesperado, que desafia as suposições sobre a representação e, portanto, desvia-se da norma com relação à funcionalidade da representação. Esses tipos de falhas podem ser considerados "anomalias de representação", pois não indicam natureza anômala nos dados determinados semanticamente, mas expõem brechas em algum aspecto da definição da representação – por exemplo, o mecanismo de mapeamento da representação para os dados correspondentes.

As anomalias de representação são aplicáveis ao tipo de representação dos dados de moléculas, por exemplo. Como dito anteriormente (seção 1.3), moléculas podem ser representadas de diferentes maneiras. No entanto, se em algum caso uma representação não conseguir mapear corretamente os dados das moléculas correspondentes, é possível considerá-las como anomalias de representação sob os critérios dessa representação específica. Além da curadoria do conjunto de treinamento sintético por meio da geração de anomalias [semânticas], esse cenário também apresenta uma área de aplicação para a geração de anomalias representacionais que podem, eventualmente, orientar os esforços de desenvolvimento de uma representação. Este trabalho investigou a aplicação do aprendizado de máquina para a geração de anomalias de representação a fim de explorar modos de falha em uma representação de cadeia de moléculas, os SELFIES, e, consequentemente, testar a robustez da representação. Aqui também é apresentada uma comparação com um conjunto de metodologias para geração de dados anômalos que não utilizam de aprendizado de máquina.

4.1.1 Trabalhos relacionados

O aprendizado de máquinas para detecção de anomalias tem sido amplamente explorado usando diferentes métodos. Nas metodologias que não envolvem redes neurais profundas é comum encontrar trabalhos acerca de detecção de anomalias que se utilizam de abordagens como *z-score*,[147] distância de *Mahalanobis*,[148] *local outlier factor* (ou *LOF*),[149] k-vizinhos mais próximos[150] e *support vector machines.*[151] Os

métodos que utilizam redes neurais profundas, tais como os autocodificadores variacionais, podem detectar anomalias ao aprender a reconstruir representações de dados normais comprimidos, o que resulta em reconstruções de dados com perdas relativas.[152-155]

Vários métodos de aprendizado de máquinas têm sido aplicados para detectar anomalias em dados biológicos. Por exemplo, Michael-Pitschaze[156] e seus colaboradores aplicaram modelos de linguagens de proteínas para detectar proteínas anômalas por meio de aprendizado de representação. Para representar proteínas, eles extraíram a última camada de uma rede neural codificadora e aplicaram uma função de pontuação para anomalia a fim de identificar proteínas humanas semelhantes a príons e classificar proteínas virais do proteoma do hospedeiro. Na mesma linha, Czibula e seus colaboradores[157] introduziram o *AnomalP*, um método para detectar conformações anômalas de proteínas usando aprendizado profundo. O modelo construído empregou vários autocodificadores para determinar se uma certa proteína ou conformação é estruturalmente diferente com respeito a sua super-família. De maneira análoga, Tiwari e seus colaboradores[158] aplicaram autocodificadores variacionais com o objetivo de detectar anomalias na integridade de colunas de cromatografia para proteína A.

Enquanto o aprendizado de máquina para detecção de anomalias tem sido explorado extensivamente, sua aplicação para a geração de anomalias tem sido relativamente limitada. Para mitigar a escassez de dados de treinamento para classificação de séries temporais anômalas, Laptev[144] utilizou as regiões latentes discrepantes de um autocodificador variacional para gerar dados anômalos sintéticos de séries temporais. Similarmente, porém aplicado a um contexto biológico, Uzolas *et al.*,[159] empregaram redes adversárias condicionais para sintetizar imagens anormais de bandeamento de cromossomos e propuseram o método para detecção, simulação e aumento de dados citogenéticos.

4.1.2 Estudo de caso: SELFIES como representação molecular

Os SELFIES são uma representação molecular em cadeia desenvolvida para ser 100% robusta, de modo que cada cadeia SELFIES seja convertida em uma molécula SMILES válida.[56] A representação aborda mecanismos que induzem

fragilidades responsáveis por tornar um SMILES inválido, tais como as restrições de sintaxe em pares para anéis e ramificações, ao removê-las completamente. Para problemas como átomos com valência excedente em SMILES inválidos, os SELFIES não tratam cada ligação como um token separado, mas sim usam tokens confinados por colchetes que acoplam essas ligações aos próprios átomos, formando um único token que respeita as regras de ligação para o átomo em questão. O método converte SELFIES para SMILES processando tokens por meio de "vetores de regras" para correção da ordem de ligação para que os átomos "*tokenizados*" obedeçam às regras de valência e sejam estruturalmente compatíveis em geral. O módulo implementa uma tabela de "regra de derivação" para processar tokens SELFIES – um a um e em sequência – em tokens SMILES, o qual constrói a cadeia de caracteres SMILES correspondente em paralelo. Dado esse processamento sequencial, o processo de conversão corretiva de um token SELFIES depende da composição atômica de seu token, do token anterior, do sucessivo e da ordem de ligação ou do tipo de anel/ramificação – o que pode ser considerado como uma "vizinhança do token". Durante o processo de conversão, a composição atômica e a fórmula empírica não são estritamente preservadas, uma vez que alguns tokens inadequados e irreparáveis (por meio de regras de derivação) podem ser descartados para garantir a validade da molécula final. Dada essa propriedade de validade, os SELFIES são propostos como uma alternativa adequada aos SMILES para representar moléculas em modelos generativos, a fim de maximizar o tamanho do espaço químico válido gerado.

Considerando a alegação de que SELFIES são 100% válidos,[56] se, hipoteticamente, for encontrada uma cadeia de caracteres de SELFIES que se converta num SMILES invalido, violando as regras químicas, será encontrada uma anomalia de representação categórica, a qual testa efetivamente a robustez a representação ao desafiar as suposições nominais sobre validade. Baseado nesse critério de validade, é possível considerar que essas anomalias pertencem a um espaço de SELFIES inválidos fora da distribuição. Além disso, esse critério de anomalia centrado na validade baseia-se exclusivamente na capacidade do módulo SELFIES de converter uma cadeia de caracteres SELFIES em um SMILES válido por meio de seus mecanismos de mapeamento definidos. Portanto, a composição de tokens de uma cadeia SELFIES que implica significado químico, está fora de consideração, e uma classificação binária "normal versus anômala" é estabelecida somente em termos do sucesso da conversão da cadeia. Assim, esses SELFIES são

considerados anômalos no nível da representação, e qualquer significado químico implícito é desconsiderado para fins de definição de anomalias. A hipótese de invalidade dos SELFIES foi testada utilizando as propriedades generativas de um autocodificador variacional, condicionado a um conjunto de treinamento, um conjunto de tokens, a escolha da arquitetura do modelo e a abordagem de exploração do espaço latente, usando a versão mais recente disponível do módulo SELFIES 2.1.1.[57]

4.2 METODOLOGIA

4.2.1 Conjunto de dados

Inicialmente, um milhão de moléculas foram selecionadas, aleatoriamente, da base de dados *ChEMBL* v29.[91-92] Apenas a informação da estrutura das moléculas representadas como SMILES foi armazenada. As esse conjunto passou por dois filtros, resultando num conjunto de treinamento contendo 400 mil moléculas:

i) tokens raros – os SMILES foram tokenizadas e apenas os tokens que estavam presentes em mais de 300 moléculas foram mantidos;
ii) comprimento da molécula (em unidade de tokens) – apenas as moléculas que possuíam até 80 tokens de comprimento permaneceram no conjunto.

Os SMILES foram convertidos para SELFIES utilizando a biblioteca SELFIES 2.1.1.[57] Após essa conversão, o comprimento máximo das moléculas foi de 91 tokens. Para garantir que todas as moléculas tivessem o mesmo comprimento, tokens de *"padding"* foram adicionados àquelas cujo comprimento fosse menor do que 91. Assim, ao final do processo, o conjunto de tokens resultante continha 54 elementos (Tabela 6). Esses parâmetros definem a dimensionalidade de entrada para o codificador (Tabela 7).

Tabela 6 – Tokens de SELFIES com os respectivos inteiros associados. O processo de tokenização das moléculas resultou num vocabulário de 54 tokens distintos, capazes de escrever todo o conjunto de treinamento.

Inteiro	Token	Inteiro	Token	Inteiro	Token	Inteiro	Token	Inteiro	Token
0	'pad'	11	[=C]	22	[=Se]	33	[Cl]	44	[O]
1	.	12	[=N+1]	23	[B−1]	34	[F]	45	[P+1]
2	[#Branch1]	13	[=N–1]	24	[B]	35	[I-1]	46	[PH0]
3	[#Branch2]	14	[=N]	25	[Br−1]	36	[I]	47	[P]
4	[#*C*-1]	15	[=O]	26	[Br]	37	[K+1]	48	[Ring1]
5	[#*C*]	16	[=PH0]	27	[Branch1]	38	[N+1]	49	[Ring2]
6	[#N+1]	17	[=P]	28	[Branch2]	39	[N-1]	50	[S+1]
7	[#N]	18	[=Ring1]	29	[C-1]	40	[NH1]	51	[S]
8	[#S]	19	[=Ring2]	30	[C]	41	[N]	52	[Se]
9	[=Branch1]	20	[=S+1]	31	[Cl+3]	42	[Na+1]	53	[Si]
10	[=Branch2]	21	[=S]	32	[Cl-1]	43	[O-1]		

Fonte: Elaborada pelo autor.

Tabela 7 – Propriedades dos conjuntos antes e depois do processo de filtragem.

	Conjunto de Dados (x1000)	Comprimento máximo (tokens)	Tamanho do conjunto de tokens
Antes dos filtros	1.000	1.281	205
Depois dos filtros	400	91	54

Fonte: Elaborada pelo autor.

4.2.2 Parâmetros do autocodificador variacional

Um autocodificador variacional foi treinado em TensorFlow v.2.10.0,[96] com uma arquitetura inspirada pelo trabalho de Gómez-Bombarelli e seus colaboradores.[22] O codificador consiste de uma camada *Embedding* que processa os tokens SELFIES representados como vetores de inteiros; três camadas *Convolutional* unidimensional (filtros: 9,9,11; tamanho do kernel: 9,9,10); uma camada *Flatten* e uma camada *Dense* que retorna parâmetros que descrever uma distribuição latente da posterior aproximada 196-dimensional. O decodificador consiste de uma camada *Dense* seguida de três camadas *GRU* (do inglês, *Gated Recurrent Unit*) com 256 neurônios ocultos cada; e uma camada *Dense* final. Ele recebe um vetor latente amostrado e que retorna distribuições não-normalizadas de tokens SELFIES para cada posição na sequência decodificada (probabilidade do modelo para os dados). Para gerar

SELFIES, foram amostrados por ganância a partir dessas distribuições de tokens SELFIES em termos da posição na sequência. O modelo foi treinado usando entropia cruzada esparsa *softmax* com perda de *logits*,88,149 divergência de Kullback-Leibler (KL),[33] e otimizador de Adam.161 Os parâmetros foram inicializados usando um esquema *Glorot* uniforme[100] e foram treinados com um *batch* de tamanho 128. O autocodificador variacional utilizado aqui foi um β-VAE, cujo fator β (que é o peso da perda KL) foi anelado[34] usando uma função linear após três épocas, com o modelo final convergindo após 15 épocas num tempo de parede de 74 horas numa única GPU NVIDIA RTX 3090. Para determinar a convergência, foi estabelecido um protocolo de *early stopping* com um paciente igual a oito. A biblioteca RDKit forneceu as funções de quimioinformática utilizadas.[103]

4.2.3 Modelos nulos

Além dos autocodificadores variacionais, como critério de comparação, foram empregados modelos que não usam aprendizado profundo (que serão chamados de modelos nulos) para a geração de SELFIES inválidos. Como o SELFIES é uma sequência de tokens e o decodificador do VAE gera SELFIES ao gerar distribuições de probabilidade de tokens por posição da sequência, de forma análoga, foi explorada a geração aleatória de SELFIES por posição para modelos nulos usando o conjunto de dados de treinamento. Esses modelos também testam a robustez da representação uma vez que os autores afirmam que "cada SELFIES corresponde a uma molécula válida, mesmo em sequências totalmente aleatórias".[56] A seguir, serão apresentados três modelos gerativos nulos, que variam de acordo com o grau de informação que o modelo possui acerca da distribuição de tokens na sequência de cada molécula do conjunto de treinamento dos SELFIES.

4.2.3.1 *Naive andom*

O primeiro passo do modelo *Naive Random* (NR) para gerar SELFIES de forma ingênua é amostrar aleatoriamente um tamanho de sequência (contagem de tokens na sequência de uma molécula), isto é, o modelo sorteia um número aleatório entre um e 91 (contagem mínima e máxima de tokens, respectivamente, no conjunto de

treinamento). Em seguida, cada posição da sequência é preenchida com um token amostrado uniformemente do conjunto de tokens. A Figura 18A ilustra esse processo. Esse modelo é chamado de *naive random* por ser ingenuamente aleatório, ou seja, porque a sequência é gerada sem nenhum conhecimento prévio das distribuições de tokens no conjunto de treinamento SELFIES, com a seleção do tamanho da sequência e a atribuição posicional dos tokens sendo aleatórias.

4.2.3.2 *Shuffle random*

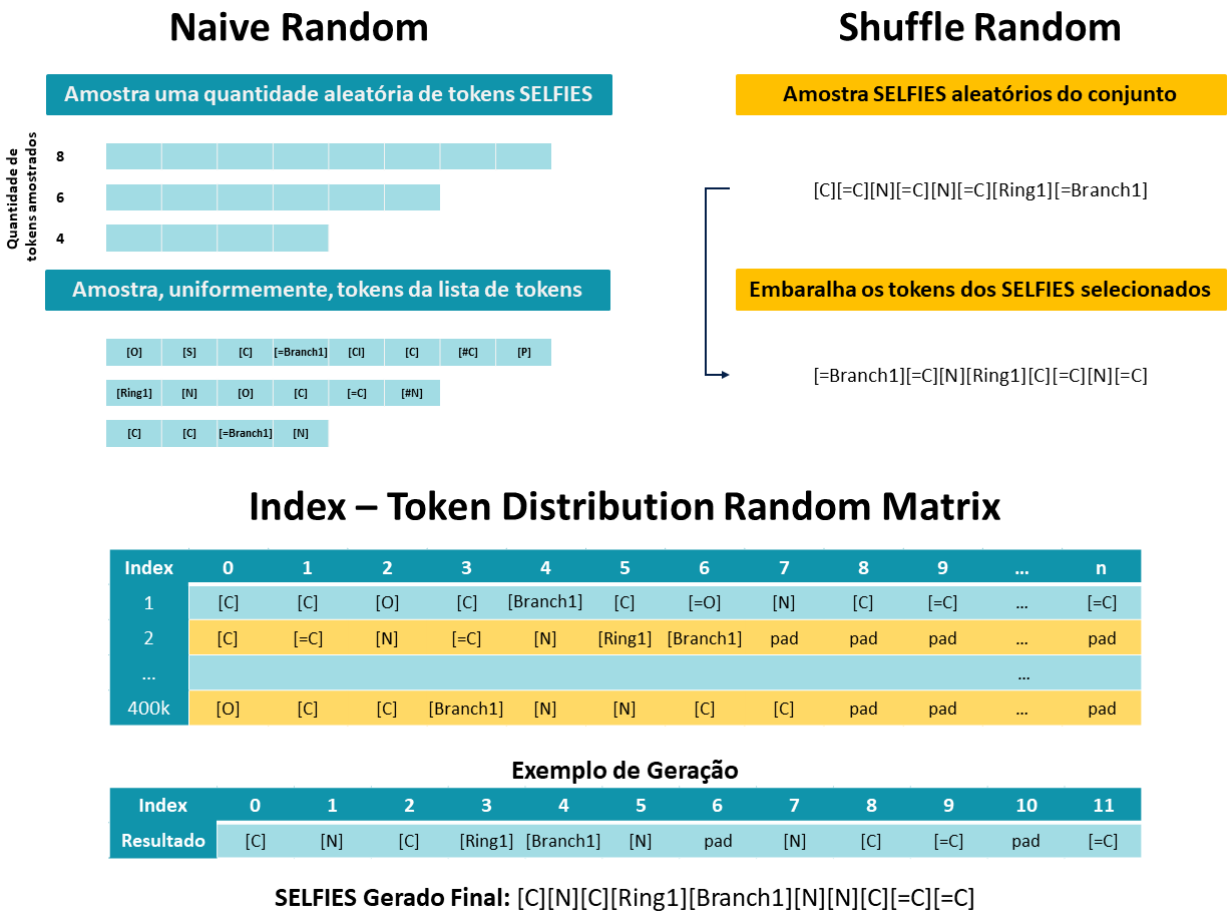
O modelo *shuffle random* (SR) gera SELFIES extraindo sequências do conjunto de treinamento e embaralhando os tokens dessa sequência internamente, o que coloca os tokens aleatoriamente em novas posições da sequência (Figura 18B). Como os SELFIES são amostrados a partir do conjunto de treinamento antes do embaralhamento, esse modelo é mais representativo da distribuição de treinamento dos SELFIES em comparação com o NR, uma vez que o processo de embaralhamento preserva a composição da sequência, ao mesmo tempo em que altera o arranjo posicional dos tokens.

4.2.3.3 *Index-token distribution random matrix*

Como as cadeias de caracteres SELFIES descrevem essencialmente uma distribuição de tokens sobre posições em uma sequência, um conjunto de dados de SELFIES pode ser interpretado como algo que estabelece uma distribuição de tokens por posição em várias sequências. Em uma sequência, a disposição horizontal dos tokens transmite características relacionais relativas à conectividade (ligações) e à estrutura (pontos de abertura e fechamento de anéis e ramificações). De forma análoga, em um conjunto de dados SELFIES, a disposição vertical dos tokens descreve coletivamente uma distribuição de tokens para cada posição em todas as sequências. Portanto, cada posição da sequência observa uma determinada frequência de tokens, o que transmite propriedades globais do conjunto de dados. Por exemplo, as sequências dos SELFIES podem ser tendenciosas no sentido de apresentarem longas cadeias de carbono em suas caudas. Nesse caso, as distribuições de tokens das últimas posições no conjunto de dados indicariam uma

alta frequência de tokens de carbono – o que expressa uma propriedade do próprio conjunto de dados.

Para adaptar essa interpretação à geração de SELFIES, o modelo *index-token distribution random* (ITDR) considera o conjunto de treinamento SELFIES como uma matriz, na qual as linhas indicam as sequências e as colunas indicam tokens por posição de sequência. Isso define uma distribuição multinomial de tokens por posição que pode ser amostrada para gerar SELFIES em uma base posicional (Figura 18C). A partir da matriz, as sequências são geradas por meio da amostragem uniforme de um token por posição (ou seja, uma coluna), e os caracteres de "*padding*" são ignorados para resultar na sequência final (Figura 18C). Esse modelo também é mais representativo com relação à distribuição de treinamento, uma vez que ele coleta amostras de tokens diretamente da distribuição de tokens por posição na sequência.



Index – Token Distribution Random Matrix

Index	0	1	2	3	4	5	6	7	8	9	...	n
1	[C]	[C]	[O]	[C]	[Branch1]	[C]	[=O]	[N]	[C]	[=C]	...	[=C]
2	[C]	[=C]	[N]	[=C]	[N]	[Ring1]	[Branch1]	pad	pad	pad	...	pad
...											...	
400k	[O]	[C]	[C]	[Branch1]	[N]	[N]	[C]	[C]	pad	pad	...	pad

Exemplo de Geração

Index	0	1	2	3	4	5	6	7	8	9	10	11
Resultado	[C]	[N]	[C]	[Ring1]	[Branch1]	[N]	pad	[N]	[C]	[=C]	pad	[=C]

SELFIES Gerado Final: [C][N][C][Ring1][Branch1][N][N][C][=C][=C]

Figura 18 – Exemplos de processos de geração do SELFIES usando os modelos nulos. A) O modelo naive random amostra um tamanho para a sequência de tokens (entre 1 e 91) e, em seguida, preenche cada posição da sequência com um token amostrado do conjunto de tokens. B) O modelo shuffle random amostra uma sequência SELFIES do conjunto de treinamento e embaralha os tokens dentro dessa sequência para gerar um novo SELFIES. C) O modelo index-token distribution random matrix interpreta a conjunto de treinamento como uma matriz para amostrar, aleatoriamente, tokens a partir das distribuições de

tokens por posição na sequência. Na matriz, n = número (ou contagem) máximo de tokens na sequência. Um exemplo de SELFIES gerado que descarta os tokens de padding está mostrado na figura.

Fonte: Elaborada pelo autor.

4.3 RESULTADOS

As buscas no espaço latente do VAE foram feitas de maneira radial por meio de amostragem aleatória de pontos em superfícies hiper esféricas 196-dimensional, centradas em zero, decodificados para conjuntos cadeias SELFIES. Esses conjuntos decodificados, de tamanho fixo (10.192 cadeias), foram avaliados quanto à porcentagem de validade em função do seu raio gerador (raio da hiper esfera da qual os pontos foram decodificados). Como o objetivo principal dessa exploração é testar a robustez representacional do SELFIES por meio do VAE, a minimização da porcentagem de validade desafia a suposição de validade de 100% dos SELFIES e testa a sua robustez. Implicitamente, a minimização da validade também atesta a capacidade do VAE como gerador de anomalias de uma representação. Além disso, um grande conjunto de cadeias de caracteres SELFIES inválidas é desejável para aprimorar a obtenção de informações acerca dos modos de falha da representação e quais são os fatores que contribuem para os critérios anômalos. Idealmente, também é esperado observar um espaço latente organizado com regiões de alta e baixa taxas de validade agrupadas.

Para avaliar a validade dos SELFIES, eles foram convertidos em SMILES e a conversibilidade dos SMILES em gráficos moleculares foi testada usando o RDKit. Se os SMILES convertidos violassem alguma regra de valência, o RDKit apontaria um erro, identificando o átomo causador desse erro e o número pelo qual a valência explícita do átomo excede a permissão máxima. A fim de garantir consistência e impor uma verificação adicional, os SELFIES foram avaliados usando diferentes plataformas (*ChemWriter* e *Smival*), juntamente com uma inspeção manual de cadeias selecionadas (Figura 19). Essa abordagem de teste obedece à reivindicação e às prescrições do método SELFIES, que define a validade em termos de conformidade com a regra dos SMILES correspondentes de uma cadeia de caracteres SELFIES, quando convertidos pelo módulo.

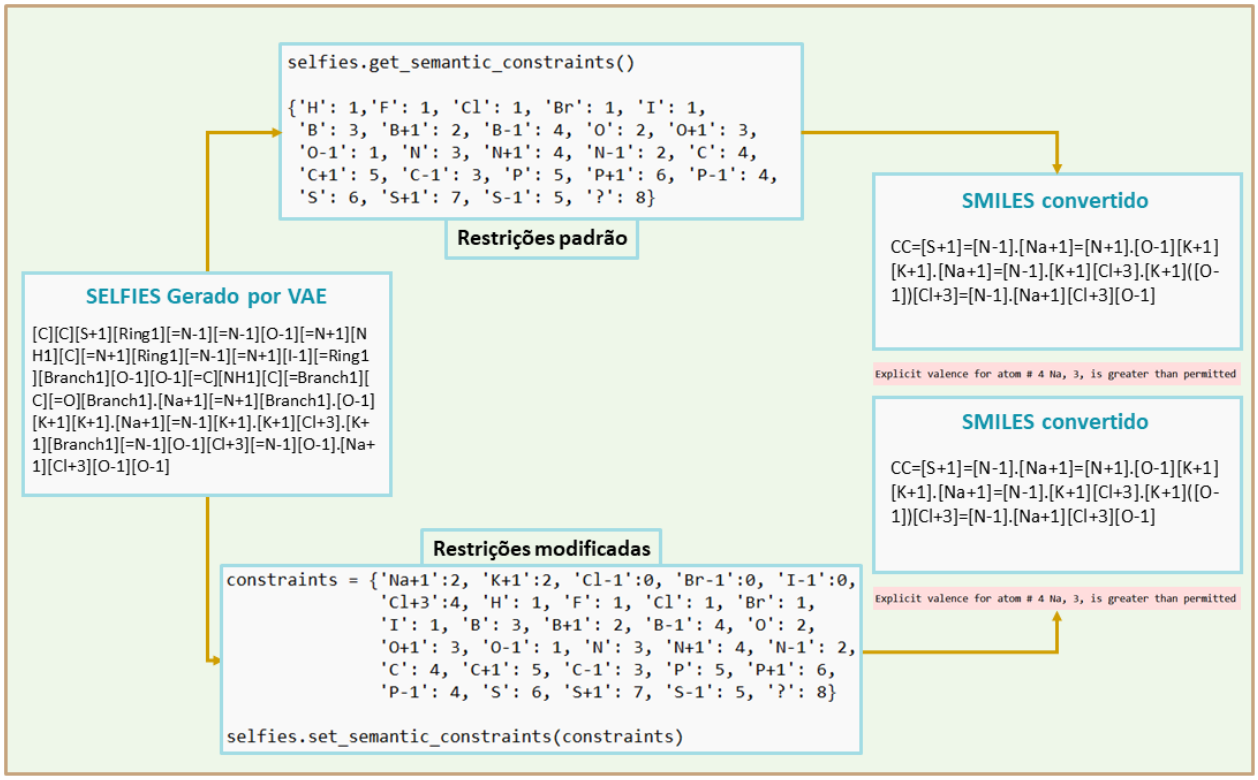


Figura 19 – Verificação da validade dos SELFIES. Exemplo de cadeia de caracteres SELFIES gerado por autocodificador variacional e convertido para SMILES usando as configurações padrão e modificada do módulo SELFIES. A validade da cadeia foi verificada usando o RDKit.
Fonte: Elaborada pelo autor.

Inicialmente, o processo de conversão de SELFIES para SMILES usou as configurações padrão do módulo para restrições de valência dos átomos, o que fez com que o RDKit retornasse erros explícitos de valência para algumas cadeias de caracteres SELFIES geradas (Figura 20A). Para destacar os erros e demonstrar sua capacidade de correção, os SELFIES inválidos foram modificados manualmente a fim de validar seus SMILES correspondentes, alterando as cargas formais associadas aos átomos tokenizados identificados pelo RDKit no SELFIES (Figura 20B). Por exemplo, a valência explícita do átomo tokenizado [Na+1] é 2, mas em um SELFIES inválido gerado (Figura 20A), as ligações associadas e a carga formal do token impõem que a valência explícita seja três, o que excede a permissão máxima em um. Para resolver

esse problema, modificamos manualmente a carga formal de [Na+1]. Da mesma forma, o [Cl-1] também foi modificado e neutralizado para tratar de erros de valência. Consequentemente, o módulo converteu os SELFIES modificados resultantes em um SMILES válido, com o sódio fazendo uma ligação dupla e ionizado, o cloro corretamente ligado, mas com alguns outros tokens que foram descartados, retornando uma estrutura geral desconectada (Figura 20B). Esse processo de modificação de valência/carga é efetivamente consistente com e equivalente ao processo de correção da ordem de ligação que é seguido pelo módulo SELFIES, o qual altera a ordem de ligação dos átomos tokenizados para obedecer às regras de valência e converter as cadeias de caracteres SELFIES em SMILES válidos.

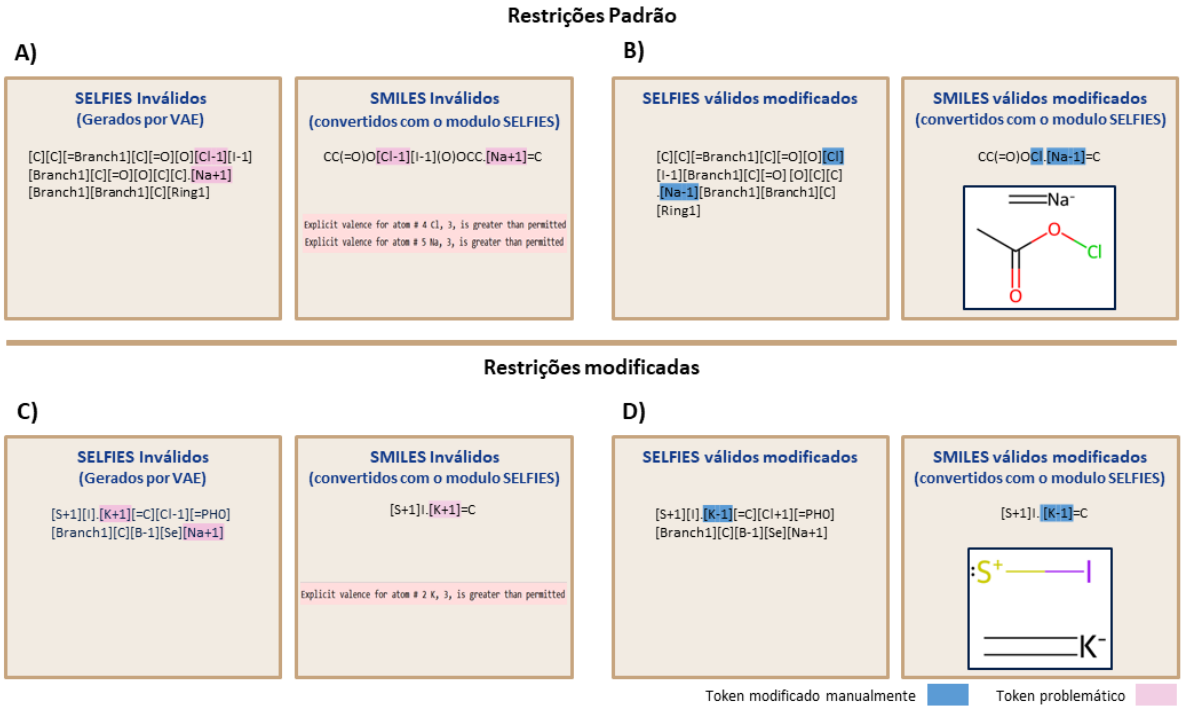


Figura 20 – Avaliação da validade de SELFIES por modificações de token e configurações das restrições. Restrições padrão: A) SELFIES convertidos em SMILES com erros de valência, B) os mesmos SELFIES com tokens problemáticos modificados manualmente para validar os SMILES convertidos resultantes. Restrições modificadas: C) SELFIES convertidos para SMILES com erros de valência e D) os mesmos SELFIES com tokens problemáticos modificados manualmente para validar os SMILES convertidos resultantes. Os tokens problemáticos e os tokens modificados manualmente estão destacados em rosa e azul, respectivamente.
Fonte: Elaborada pelo autor.

Ao filtrar várias cadeias de caracteres SELFIES inválidas geradas pelo VAE e repetir esse processo de modificação corretiva usando a configuração de parâmetros

padrão do módulo SELFIES para restrições de valência, foi possível identificar um conjunto de seis tokens de átomo "problemáticos" (três cátions e três ânions) que foram consistentemente responsáveis pela invalidade geral da cadeia de caracteres (Tabela 8). Isso ocorreu porque a valência imposta a eles pela carga formal e pelos átomos vizinhos excedeu as permissões máximas estabelecidas. Portanto, o RDKit os sinalizou, repetidamente, como fontes de erro (invalidando toda a cadeia) dado que suas valências explícitas eram sempre maiores do que o permitido.

Além das restrições padrão, o módulo SELFIES permite que os usuários definam restrições personalizadas para atender às necessidades específicas do contexto estudado e adaptadas às cadeias de caracteres SELFIES que estão sendo avaliadas, por meio da função *set_semantic_constraints()*. O módulo constrói, dinamicamente, suas regras de derivação usando esse conjunto de restrições especificadas, as quais estabelecem o número máximo permitido de ligações que cada átomo pode formar em uma molécula.[57] Essa função foi utilizada definindo, explicitamente, as restrições de valência para os seis tokens de átomos problemáticos identificados. Consequentemente, os processos de conversão SELFIES para SMILES foram totalmente resolvidos para quatro dos seis tokens problemáticos, resultando em SMILES válidos, e sem o RDKit indicar suas valências explícitas associadas como excedentes. No entanto, essa personalização não conseguiu resolver os erros de valência associados aos dois tokens de átomo restantes – designados aqui como o conjunto de tokens problemáticos filtrados (Figura 19 e Figura 20C).

Tabela 8 – Conjunto de tokens SELFIES problemáticos encontrados depois da análise dos erros no processo de conversão SELFIES para SMILES usando as configurações padrão do módulo. Em vermelho estão destacados os tokens que permaneceram problemáticos depois da customização das restrições do módulo; os tokens cujo problema foi resolvido após a customização estão coloridos em azul.

	Token SELFIES problemático
1	[Na+1]
2	[K+1]
3	[Cl+3]
4	[Cl-1]
5	[Br-1]
6	[I-1]

Fonte: Elaborada pelo autor.

Uma vez que o módulo opera num dado SELFIES e o converte para um SMILES inválido sem acusar erro ao utilizar as restrições de valência personalizadas para os tokens problemáticos, ficam estabelecidos os modos de falha para testar a robustez da representação, com as ocorrências de SELFIES inválidos consideradas como anomalias da representação sob a suposição de 100% de validade. Esses erros também puderam ser corrigidos manualmente, resultando em SELFIES convertidos em SMILES válidos (Figura 20 D).

A geração de cadeias de caracteres SELFIES radialmente indicou que os SELFIES inválidos sempre contêm pelo menos um token problemático, enquanto os SELFIES válidos podem ou não conter esses tokens problemáticos (Tabela 9). Tais tokens só invalidam um SELFIES com base na valência e nas características associadas aos seus tokens vizinhos. Dessa forma, um SELFIES inválido pode ser convertido em uma forma válida, seja editando o próprio token problemático que o constitui ou ajustando a valência ou composição atômica de seus tokens vizinhos.

Tabela 9 – Exemplos de SELFIES dos conjuntos de treinamento e gerados por VAE, indicados por validade e presença de token problemático (PTP), com os tokens problemáticos destacados em negrito. O conjunto de treinamento contém 100% de cadeias válidas, independentemente de conter ou não os tokens problemáticos. As células com SELFIES válidos e inválidos estão destacadas de azul e vermelho, respectivamente. Os SELFIES apresentados for retirados de um conjunto de 10 mil SELFIES decodificados de uma hiper esfera de de raio = 180,0.

Conjunto	Validade	PTP	Exemplo
Treinamento	Válido	Sim	[C][C][C][O][C][=Branch1][C][=O][C][=C][C][=C][C][=C][Ring1][=Branch1][O][S][=Branch1][C][=O][=Branch1][C][=O][N-1][C][=Branch1][C][=O][N][C][=N][C][Branch1][Ring1][O][C][=C][C][Branch1][Ring1][O][C][=N][Ring1] [#Branch2].**[Na+1]**
		Não	[C][C][O][C][=C][C][=C][C][=C][Ring1][=Branch1][C][N][C][C][N][Branch1][N][C][C][=C][C][=C][Branch1][C][C][O][Ring1][=Branch1][C][Branch1][Ring2][C][C][O][C] [Ring1][S]
Gerado por VAE	Válido	Sim	[C][=C][C][=N][C][=Ring1][Ring1][N][Ring2][Ring1][C][=Branch1][O-1][=Ring1][Ring1][=Branch1][C][O-1][O-1][O-1][Ring1][Ring1][Branch1][C][Cl][=C][Ring2][NH1][C][=N+1][Ring1][O-1][C][=O][O-1][NH1]**[Na+1]**[Ring1][C] [=O][Ring1][O-1][O-1]
		Não	[C][O][C][=Branch1][C][=O][C][=C][NH1][C][=N+1][C][=C][C][=C][Ring1][=Branch1][Cl][O][C][=Ring1][#Branch1]
	Inválido	Sim	[C-1][#N+1][C-1][=N-1][=N-1][=N-1][=N-1][=P][=Ring1][=N-1][=P][=N-1][=N-1][=N+1][=N-1][Cl+3][=N-1][O-1][#N][=C][Ring2][Ring2][Ring1][C][O-1][=N+1][O-1][=N+1][=N+1][=N-1][C][O-1].[O-1].[O-1]..**[K+1]**[=N-1][=N-1][Cl]**[K+1]**.[Cl][Cl+3][Branch1][=N+1][O-1][=Ring1[[=N-1][=N-1].[O-1][Ring1]..[O-1][=Ring1][Cl+3].[O-1].**[K+1][K+1]**.**[K+1]**.[O-1][O-1][O-1].[O-1][O-1]

Fonte: Elaborada pelo autor.

Ambas as cadeias de caracteres SELFIES classificados como válidas e inválidas geradas por VAE apresentaram tokens problemáticos. Entretanto, os tokens problemáticos em SELFIES inválidos foram tratados para permitir a conversão para SMILES válidos – condicionados por fatores relativos aos tokens da vizinhança, tais como a disposição de estruturas desconexas, indicados por “pontos”.

Tabela 10 – Quantidade e comparação da presença de tokens problemáticos (PTP) nos conjuntos de treinamento e gerado por VAE. No conjunto de treinamento, 100% dos SELFIES são válidos, independentemente de PTP. O conjunto gerado por VAE foi decodificado dentro do volume de uma hiper esfera de raio = 180,0.

Conjunto	**Validade**	**PTP**	**Quantidade de SELFIES**	
			Subtotal	**Total**
Treinamento	Válido	Sim	1.986 (0,50%)	400.000
		Não	398.014 (99,50%)	
Gerado por VAE	Válido	Sim	239.656 (13,06%)	1.834.560
		Não	850.458 (46,36%)	
	Inválido	Sim	744.446 (40,58%)	

Fonte: Elaborada pelo autor.

A quantidade de cadeias de caracteres SELFIES discriminada pela presença tokens problemáticos nos conjuntos treinamento e gerados pelo modelo de aprendizado de máquina foi contabilizada e está indicada na Tabela 10. Explorando superfícies hiper esféricas com raios entre 0,0 e 180,0, em intervalos de 1,0, com 10.192 SELFIES gerados por raio, o conjunto de SELFIES gerados contou com mais de 1,8 milhão de SELFIES. Embora menos de 1% do conjunto de treinamento (que continha 100% de SELFIES válidos) continha tokens problemáticos, aproximadamente, 59% do conjunto gerado continha tokens problemáticos.

Além de explorar o autocodificador variacional como gerador de anomalias de representação ao decodificar cadeias de caracteres SELFIES como uma função do raio gerador latente e avaliar a validade das cadeias geradas, também foi mapeado o perfil de validade por raio em estágios intermediários de treinamento do modelo nas épocas 0 (modelo não treinado, iniciado com distribuição uniforme *Glorot*), 1, 5 e 7

(Figura 21). Isso ilustra o progresso da capacidade de discriminação e agrupamento do VAE para SELFIES por validade no decorrer do treinamento. Os modelos nas épocas 0 e 1 não conseguiram superar o melhor modelo nulo (*naive random*) na minimização da validade, isto é, atingir um número maior de cadeias inválidas dentro do conjunto. Por outro lado, os modelos nas épocas 5, 7 e 15 (convergência) superaram o modelo nulo ao variar o raio generativo.

Para garantir uma ampla cobertura do espaço latente e avaliar o comportamento global dessa propriedade, a porcentagem de validade dos conjuntos de SELFIES gerados como uma função de um raio gerador variando entre 0,0 e 1.000,0, foi verificada (Figura 21B). A Figura 21B indica uma organização espacial latente de três pontos/regiões-chave considerando o modelo convergido (15ª época).

i) Fronteira normal-anômala (SELFIES válidos-inválidos): R = 13;
   - A região R < 13,0 é composta puramente por SELFIES válidos, porque as cadeias decodificadas nessa região demostram 100% de validade, enquanto pontos decodificados em R > 13 geraram conjuntos de SELFIES cuja validade é variável.

ii) Mínimo global para a porcentagem de validade: R = 61;
   - No modelo final convergido, o VAE minimizou a validade do conjunto de SELFIES decodificado para 11,24% nesse raio, com a porcentagem de invalidade maximizada a 88,76%. Implicitamente, esse raio indicou o limite superior de desempenho do modelo como um gerador de anomalias de representação.

iii) Domínio de aplicabilidade como um gerador de anomalia representacional, relativo ao modelo nulo com melhor desempenho: R > 28;
   - Nesse domínio radial, o VAE superou o melhor gerador nulo (*naive random*) na tarefa de tarefa de minimizar a porcentagem de validade dos conjuntos SELFIES gerados. Os conjuntos SELFIES gerados em todos os raios < 28 apresentam porcentagens de validade menores do que as geradas pelo modelo nulo (*naive random*), que gerou conjuntos de SELFIES com uma porcentagem de validade de 73,88% (Tabela 11). Devido ao desempenho superior na minimização da validade, esse domínio radial estabeleceu, efetivamente, o domínio de aplicabilidade do VAE como um gerador de anomalias para uma representação molecular.

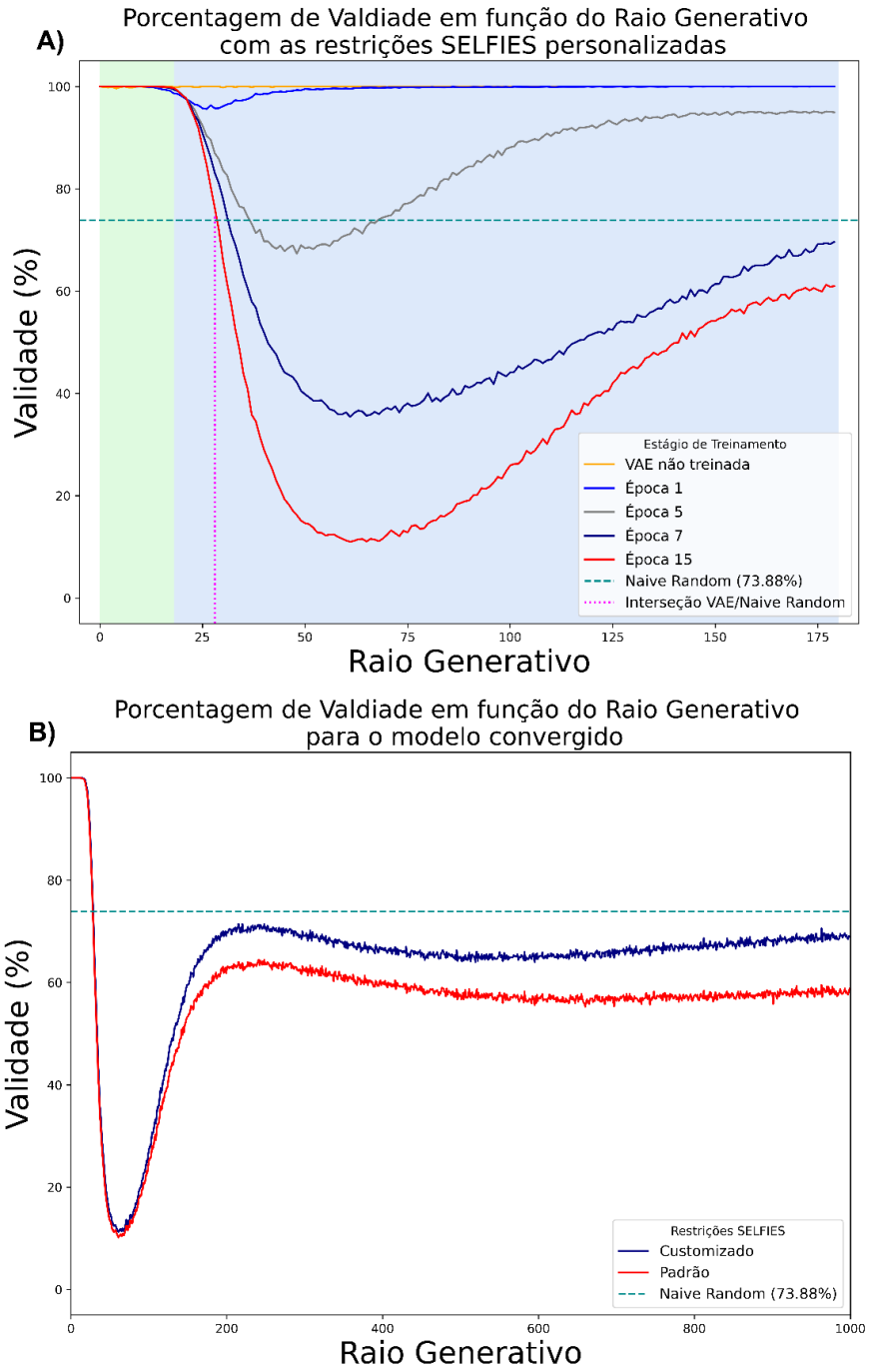


Figura 21 – Conjuntos SELFIES gerados de tamanho 10.192 para cada raio (R), exceto R < 6,0 (a área de superfície da esfera e a densidade molecular são muito baixas para gerar 10.192 cadeias SELFIES exclusivas dessa região). A) Porcentagem de validade nos conjuntos SELFIES gerados em função do raio, de R = 0,0 a R = 180,0, em vários estágios de treinamento (do modelo não treinado aos modelos treinados na 1ª, 5ª, 7ª e 15ª épocas). A região verde indica o espaço normal puramente válido de SELFIES, enquanto a região azul indica o espaço dos SELFIES anômalos (inválidos). A linha pontilhada em magenta indica o ponto no qual o VAE começa a superar o modelo *naive random* na minimização da validade. B) Porcentagem de validade por raio gerador, de R = 0,0 a R = 1000,0 no modelo final convergido, para as restrições padrão (vermelho) e personalizadas (azul). Em ambos os gráficos, a linha tracejada em ciano escuro indica o modelo nulo *naive random*.
Fonte: Elaborada pelo autor.

Como era de se esperar, dadas as características da função *prior* do VAE, as cadeias de SELFIES consideradas como válidas têm uma maior probabilidade de se agruparem em torno da origem do espaço. Em contrapartida, os SELFIES inválidos estão mais distribuídos numa região mais externa do espaço, distante da origem. Tal fato está mostrado na Figura 22, a qual apresenta uma análise das componentes principais para um conjunto de 10.192 cadeias de SELFIES gerados dentro de uma hiper esfera de raio 61,0.

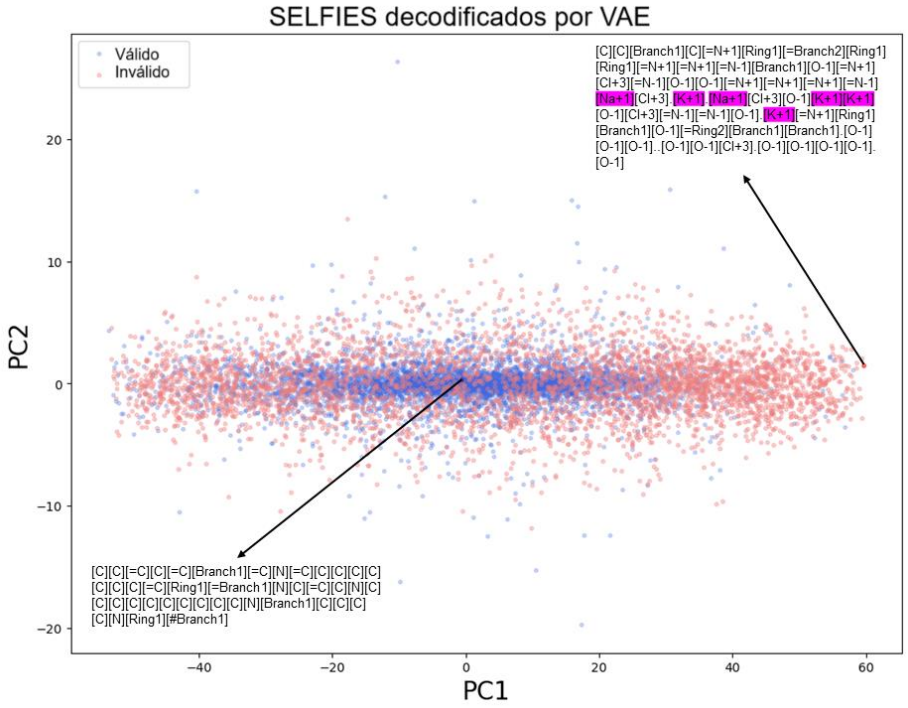


Figura 22 – Análise de componentes principais (PCA) de um vetor 196-dimensional (amostrado dentro do volume da hiper esfera de raio R = 61,0) decodificado para 10.192 cadeias SELFIES. Os pontos estão coloridos de acordo com a validade (azul para válidos, vermelho para inválidos). Os SELFIES válidos estão concentrados no centro do gráfico, enquanto os SELFIES inválidos estão espalhados pela periferia. As setas indicam exemplos de SELFIES válidos e inválidos (com o token problemático destacado) decodificados a partir de pontos correspondentes projetados do espaço de dimensão alta.

Fonte: Elaborada pelo autor.

Tabela 11 – Porcentagem de validade dos conjuntos de SELFIES gerados pelos diferentes tipos de modelos para uma amostra de 10,192 cadeias de caracteres. O melhor modelo nulo em minimizar a validade está destacado em negrito, assim como o o raio generativo para o VAE que minimiza a validade.

Tipo de Modelo	Gerador	Validade (%)
Nulo	Naive Random	**73,88**
	Shuffle Random	99,71
	Index-Token Distribution Random	99,91
Aprendizado Profundo	VAE (raio generativo = 61)	**11,24**

Fonte: Elaborada pelo autor.

Para avaliar a capacidade e aplicabilidade de geração de anomalias de representação do VAE, foram gerados conjuntos de 10.192 cadeias de caracteres SELFIES, utilizando os três modelos nulos e a validade percentual foi comparada entre os modelos. Do total de SELFIES gerados pelos modelos *shuffle random* e *index-token distribution matrix*, praticamente 100% das cadeias foram consideradas válidas (99,71% e 99,91%, respectivamente), o que os torna métodos fracos para testar a robustez da representação. Em contrapartida, o modelo *naive random* resultou em 73,88% de validade para seu conjunto de cadeias SELFIES gerado, mostrando-se um teste significativamente melhor que os modelos nulos na tarefa de desafiar a robustez da representação molecular (Tabela 11, Figura 23). Sendo o melhor modelo nulo para geração de anomalias, o *naive random* serviu como linha de base para validar a relevância e o domínio de aplicabilidade do VAE como um método superior de gerar anomalias de representação molecular. Em termos de presença de tokens problemáticos nos conjuntos gerados, tanto os modelos nulos quanto o VAE foram capazes de gerar SELFIES válidas e inválidas contendo tais tokens (Tabela 12).

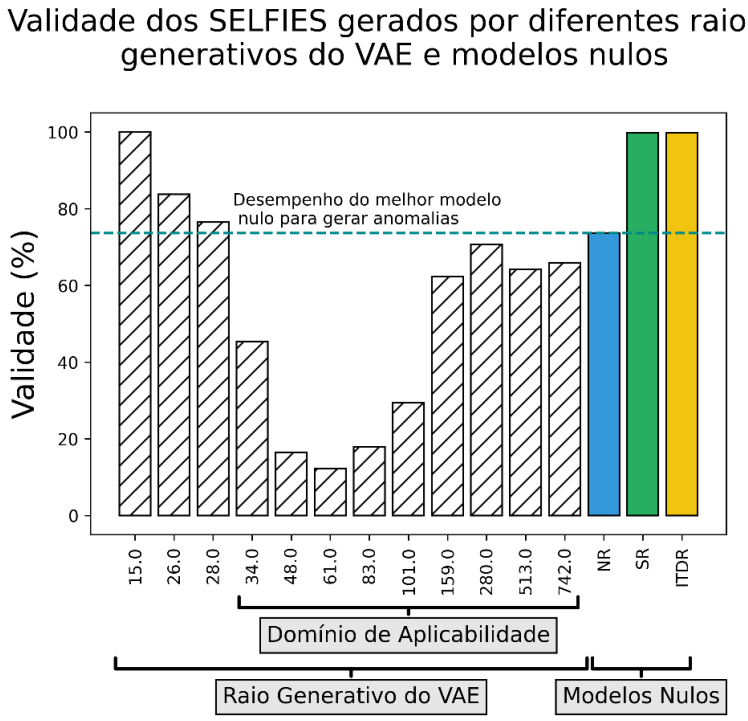
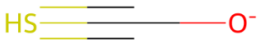
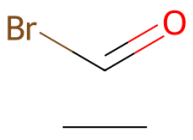


Figura 23 – Gráfico de barras resumindo a porcentagem de validade de conjuntos de SELFIES gerados para os quatro modelos aplicados. As barras hachuradas indicam diferentes raios generativos do VAE. A barra azul representa o modelo *naive random* (NR), a verde, o *shuffle random* (SR), e a amarela, *index-token matrix distribution random* (ITDR). Baixas percentuais de validade indicam melhor desempenho em gerar anomalias representacionais. As barras abaixo da linha azul tracejada indicam os raios generativos que têm um desempenho melhor do que o melhor modelo nulo (NR) e representam o domínio de aplicabilidade do VAE.

Fonte: Elaborada pelo autor.

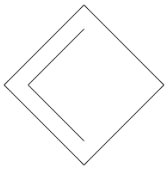
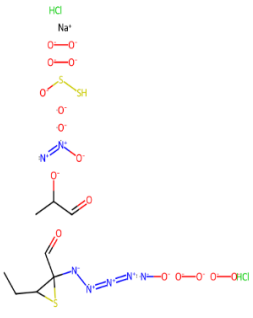
Tabela 12 – Exemplos de SELFIES válidos e inválidos gerados, com os SMILES correspondentes, contendo os tokens problemáticos, para os quatro modelos utilizados (NR = naive random, SR = shuffle random, ITDR = index-token distribution random). As células contendo SELFIES válidos e inválidos (e SMILES convertidos) estão coloridas em azul e vermelho, respectivamente. Os tokens problemáticos estão destacados em negrito em cada cadeia SELFIES e SMILES. As estruturas dos SELFIES convertidos para SMILES válidos estão representados ao lado.

Gerador		Exemplos de SELFIES, SMILES e estrutura	
NR	Válido	SELFIES: [C][#Branch2][I]**[Na+1]**[#S][O-1][Branch2][I-1][C][P+1] SMILES: C(#S)[O-1]	
	Inválido	SELFIES: **[Na+1]**[P][Si][Ring1][N+1][Br][=O][N+1][NH1][P+1][Ring2]**[Na+1]**[N][P+1][=N][=PH0] SMILES: [Na+1]P=[Si]Br	
SR	Válido	SELFIES: [C][#Branch1][=O][C][Ring1][Branch1][=C][O][C][C][=C][Branch1][N][F]**[Na+1]**[=Branch1][C][C][C][O-1][=C][=C][C][=C][#Branch1][C][C][C][Branch1][C][=Branch1][Ring1].[Br][C][=O][=C] SMILES: CC.BrC=O	

(continua)

(continuação)

Tabela 12 – Exemplos de SELFIES válidos e inválidos gerados, com os SMILES correspondentes, contendo os tokens problemáticos, para os quatro modelos utilizados (NR = naive random, SR = shuffle random, ITDR = index-token distribution random). As células contendo SELFIES válidos e inválidos (e SMILES convertidos) estão coloridas em azul e vermelho, respectivamente. Os tokens problemáticos estão destacados em negrito em cada cadeia SELFIES e SMILES. As estruturas dos SELFIES convertidos para SMILES válidos estão representados ao lado.

Gerador		Exemplos de SELFIES, SMILES e estrutura	
ITDR	Inválido	SELFIES: [=C][#Branch1][C][=O][=C]**[K+1]**[C][Ring1][Branch1][P][S][=C][=Branch2][C][=Branch1][=O][=O][=Branch1][C][=C].[C][C][C][=C][Branch1][C][Branch1][O][Ring1][#Branch2][C][Ring1][=Branch1][N][C][=C][C][S][=C][C][=Branch1][=C][Cl][O-1][=O][C][O][N][Branch1][C]	
		SMILES: C1(=O)C**[K+1]**C1PS=C2OOC=C.CC C=CO2	
	Válido	SELFIES: [C][C][=C][C][#Branch1][C][Ring1][=O][Ring1][Branch2][C][O][C][Branch1][C] [#C][Ring2][C][Branch1][C][=Branch2][Branch1][C][=O][S][N][O][S][=C] [Branch1][Branch1][F][C][Ring1][S][=N] [Ring1][Ring1][=Branch1][Ring1][=C][O][=C]**[Na+1]**[= C]	
		SMILES: C1C=C=C1	
	Inválido	SELFIES: [C][=C][O][C][C][C][C][=Branch1][=Branch1][C][C][Branch1][C][#Branch1] [Ring1][N][Branch1][C][=C][=C][Branch1][=C][Branch1][Ring1][C][C].[C][C][C][=C] [Branch1][=Branch2][C][Ring2][Ring1][C] **[Na+1]**[C][N][Ring1][Cl][=O][Ring1][O][Ring2][Ring1][=O][C]	
		SMILES: C12=COCCCC1C3C.CCC=C(C2**[Na+1]** C=N)O3	
VAE	Válido	SELFIES: [C][C][C][S][C][Ring1][Ring1][=Branch2][N+1][=Branch1][C][=O][=C][=N+1] [=N-1][=N-1][N+1][=N+1][=N+1][Ring1][O-1]. [O-1][=S][S][Cl-1][Branch1][Branch1]. [O-1][O-1][Branch1][# C][O-1][O-1][=N-1] [O-1].[O-1].[O-1][O-1][O-1][Cl+3].[O-1][O -1][Cl+3][Cl+3][O-1][O-1][O-1][O-1] [Branch1].[O-1][=N+1][Ring1][Ring2][=N+1][Branch1][Cl].[Cl].**[Na+1]**.[C]. [O-1][=C][Ring2][=Ring1][Ring1][C] [=Branch1][C][=O][=Branch1].[O-1][Branch1].[O-1][=N-1].[Cl].[O-1][O-1]**[K+1]**	
		SMILES: CCC1SC1(C=O)[N-1][N+1]= [N+1]=[N+1].[O-1]SS.[O-1][O-1].[O-1] .[O-1][O-1].[O-1][O-1].[O-1][N+1]=[N+1].Cl.[Na+1].C2. [O-1]C2C=O.[O-1].[O-1] [N-1].Cl.[O-1][O-1]	
	Inválido	SELFIES: [O][=C][N][Branch2][Ring1][Branch1][N+1][Ring1][=Ring1][Ring1][=Ring1] [Ring1][=Ring1][Ring1][=N-1][=N-1] [=N-1][=N+1][=N-1][N+1].[N+1].[O-1].[O-1][=Ring1][=N-1][O-1][O-1][O-1][Ring1].[O-1][=N+1][=N-1][N+1].**[Na+1]**[Cl+3] [=N-1][Cl][O-1][O-1][O-1][O-1][Cl+3] [Branch1].[Cl].[O-1].[O-1].[O-1][=Ring1][O][Ring2][=Ring2][Ring1][=N][Ring2][Ring1][Branch1][NH1][N][Ring2][=N-1] [#Branch2].**[K+1]**.[O-1][Branch1][C][O-1][O-1].[Ring1][=N+1][Ring1][Ring1]	
		SMILES: O=CN=[N+1].[N+1].[O-1]1.[O-1]1.[O-1] [N+1]=[N-1].**[Na+1]**[Cl+3]=[N-1].Cl. [O-1].[O-1].[O-1].**[K+1]**.[O-1]C2[O-1]. [N+1]2	

Fonte: Elaborada pelo autor.

A partir da análise das cadeias válidas geradas pelos modelos foi possível observar que o módulo SELFIES lida com as cadeias que contêm os tokens problemáticos de duas maneiras: 1) ele transformou a cadeia descartando tokens um a um até que o token problemático fosse acomodado em um ponto da estrutura de modo a formar uma desconexão (Figura 24A); 2) ele descartou o token problemático para tornar a cadeia de caracteres válida (Figura 24B).

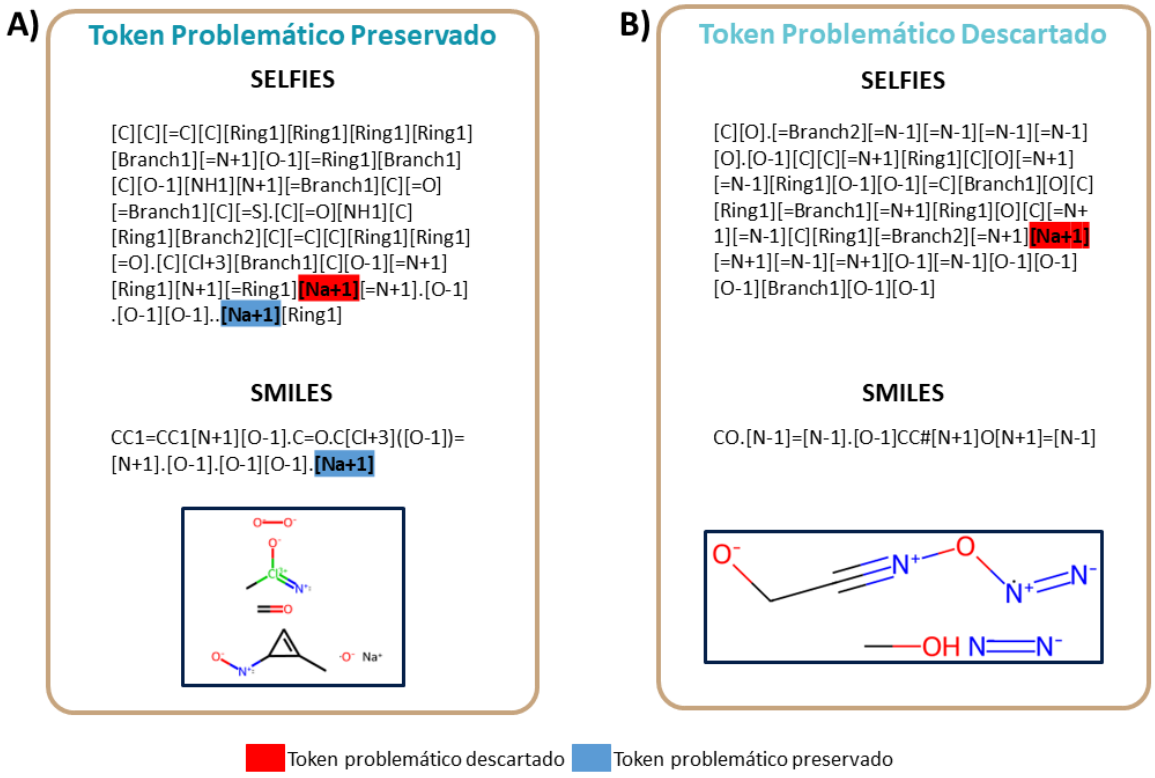


Figura 24 – Dois casos de SELFIES válidas contendo tokens problemáticos. A) SELFIES gerados por VAE (raio gerador = 98,0) com tokens problemáticos preservados (em azul) ao converter os SELFIES em SMILES. B) SELFIES gerados por VAE (raio gerador = 98,0) com token problemático descartado (em vermelho) na conversão de SELFIES para SMILES, exibidos junto com a estrutura molecular.
Fonte: Elaborada pelo autor.

4.4 DISCUSSÃO

As representações moleculares em cadeias de caracteres descrevem moléculas usando uma sequência linear de tokens. Essa abordagem limita a expressão química, dada a natureza inerentemente gráfica das moléculas. Contudo, o uso de tokens como blocos de construção é vantajoso para a redução da

dimensionalidade eficiência de memória e compatibilidade de máquina.[161, 162, 163] O potencial de modelos generativos é frequentemente limitado por resultados inválidos na forma de SMILES e grafos.[46, 55, 57, 164-165] No entanto, uma vez garantida a validade máxima da representação, o modelo generativo de variável latente pode focar em aprender espaços químicos acessíveis maiores. Nesse sentido, os SELFIES têm um papel importante, pois abordam as causas comuns que resultam na invalidade dos SMILES por meio da criação de tokens especializados e mecanismos corretivos.[56] Dado o sucesso dos SELFIES e o progresso no desenvolvimento da representação, a necessidade de investigar métodos avançados para testar a robustez de uma representação molecular deve ser enfatizado.

Para testar a robustez da representação dos SELFIES, os quatro modelos generativos estudados tentaram maximizar as permutações das posições dos tokens nas sequências para descobrir possíveis modos de falha ao converter uma cadeia de SELFIES para SMILES. Como a distribuição de treinamento dos SELFIES é puramente válida, a descoberta de um SELFIES inválido (ou anômalo) implica explorar espaços de distribuições de tokens que divergem do "espaço válido" de distribuições do conjunto de treinamento. Ao divergir desse espaço válido de distribuição de tokens, o que aconteceu foi a imposição de fatores de estresse ao mecanismo de conversão corretiva do SELFIES.

As moléculas do mundo real variam de produtos naturais com estruturas químicas complexas a moléculas de medicamentos com baixa massa molecular e sinteticamente acessíveis, cujas representações de cadeia exploram permutações para sequenciamento de tokens.[22, 166] Para o tamanho do conjunto de tokens utilizado e o comprimento máximo dos SELFIES, esse espaço foi equivalente a 91^{54} cadeias de caracteres possíveis. Aplicar os autocodificadores variacionais a essas sequências converteu sua natureza inerentemente discreta em representações contínuas, o que também capturou proeminências, minimizou a complexidade e permitiu explorar continuamente um espaço latente organizado para uma ampla gama de novas sequências usando métricas de distância.

Realizar a análise de amostragem-decodificação em hiper esferas no espaço latente revelou um domínio radial (R > 28) que superou, consistentemente, o melhor modelo nulo (*naive random*) na tarefa de minimizar a porcentagem de validade em

conjuntos de SELFIES gerados de tamanho fixo (10.192). O perfil de validade percentual versus raio generativo atingiu um mínimo global de 11,24% no raio igual a 61, com o domínio parabólico indicando regiões latentes de variação máxima, além do qual a tendência de validade se estabilizou. Foi possível inferir que o VAE distribuiu a validade dos SELFIES de forma unimodal sobre o raio generativo, concentrando o treinamento (de cadeias válidas normais) em regiões latentes sugeridas pela *prior* gaussiana unimodal. Isso implicou numa organização de dois domínios radiais latentes principais: (i) SELFIES puramente válidos em R < 13; (ii) porcentagem de validade diminuindo monotonicamente em 13 < R < 61. Como o componente de perda *KL* da função objetivo do VAE traz as representações latentes do conjunto de treinamento em direção à origem e o conjunto de treinamento era composto por dados puramente válidos, era de se esperar que fosse observada 100% de validade em regiões em torno da origem. Além disso, apenas 0,5% do conjunto de treinamento SELFIES continha tokens problemáticos e o VAE só começou a introduzir tokens problemáticos nos conjuntos SELFIES gerados em raios maiores que 13 (Figura 25), sendo a maior porcentagem de cadeias contendo tokens problemáticos (88,76%) vista no mínimo global. Portanto, os SELFIES gerados dentro do limite normal-anômalo (R < 13) são mais representativas do conjunto de treinamento em termos de validade e presença de tokens problemáticos.

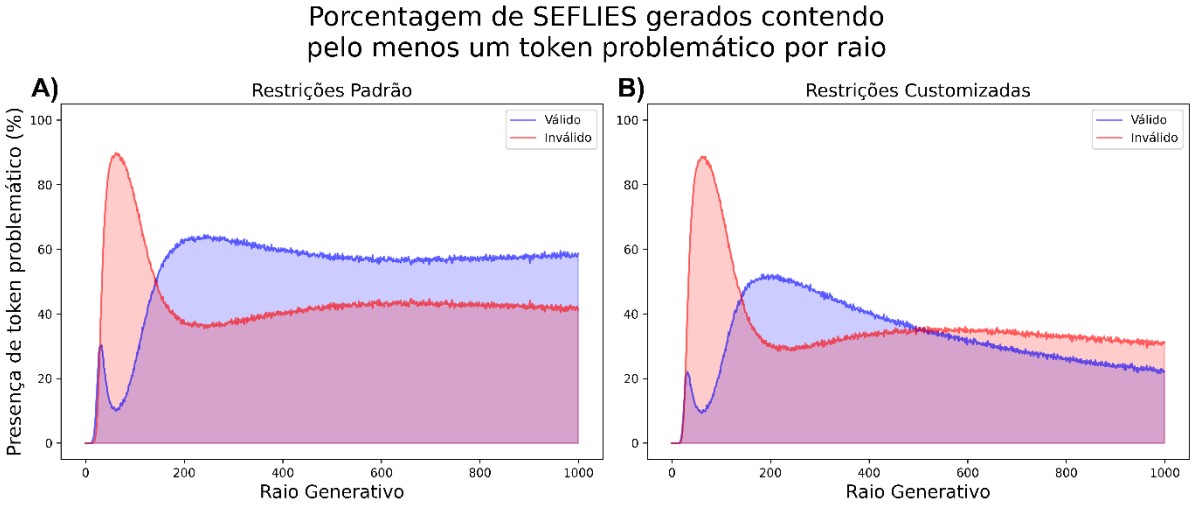


Figura 25 – Porcentagem de cadeias SELFIES geradas contendo pelo menos um token problemático em função do raio generativo e estado de validade usando as restrições A) padrão e B) customizada. Um total de 10.192 cadeias únicas de SELFIES foram gerados para cada raio, exceto para R < 6,0.

Fonte: Elaborada pelo autor.

Entre os três modelos nulos, o *naive random* superou significativamente o *shuffle random* e o *index-token distribution random* na tarefa de minimizar a porcentagem de validade nos conjuntos SELFIES gerados. Isso ocorreu principalmente porque os SELFIES gerados por esse modelo têm menos informação acerca da distribuição de SELFIES válidos do conjunto de treinamento, isto é, o procedimento de amostragem ingênua não tem conhecimento prévio sobre as distribuições de tokens do conjunto de treinamento por posição e sobre a composição de tokens dentro das sequências, por isso, recebe esse nome. Ao contrário dos outros dois modelos nulos, ele fez uma amostragem aleatória de tokens presente no conjunto de tokens, com o número de posições (contagem de tokens numa sequência) também selecionado aleatoriamente dentro do intervalo de contagem de tokens das sequências do conjunto de treinamento. Dado seu grau de aleatoriedade, o modelo *naive random* foi o melhor modelo nulo possível para explorar ao máximo o espaço de sequências admissíveis.

Ao rastrear fontes de erro em SELFIES inválidos, dois tokens problemáticos ([Na+1] e [K+1]) foram identificados. Esses tokens, consistentemente, estavam ligados em excesso em termos de permissões de valência e, consequentemente, causavam invalidade geral em sua representação SMILES convertida. No entanto, os erros puderam ser corrigidos após um processo de correção de valência manual consistente com a abordagem de correção de ligação feita pelo módulo. Assim, ficou evidente que a presença de tokens problemáticos foi um dos dois pré-requisitos para a invalidade da cadeia. Isso ocorreu porque cada SELFIES inválido continha pelo menos um token problemático cujas ligações associadas foram consistentemente sinalizadas por excederem as permissões de valência nos SMILES convertidos. Portanto, os tokens problemáticos não invalidaram um SELFIES apenas por existirem na sequência, uma vez que esse processo também estava condicionado à vizinhança do token problemático. De um conjunto de mais de 1,8 milhão de SELFIES geradas pelo VAE, 13,06% dessas cadeias continham pelo menos um token problemático. Ao rastrear as fontes de validade nessas cadeias de caracteres, ficou evidente que ou o módulo descartou completamente os tokens problemáticos, ou descartou seus tokens vizinhos para introduzir um ponto de desconexão no local do token no SMILES convertido. É presumível então que o módulo, tanto nas suas configurações padrão ou usando configurações personalizadas, não possui recurso capaz de lidar com a

correção de valências/cargas associadas a esses dois tokens problemáticos. Isso pôde ser atribuído a uma deficiência no código fonte do módulo ou a alguma lacuna conceitual na sua formulação, considerando que os SELFIES demonstraram 100% de robustez no subconjunto de dados gerados sem esses tokens problemáticos. É importante mencionar que os mesmos erros foram observados durante a utilização da versão anterior do módulo (versão 2.1.0).

Tanto os conjuntos SELFIES gerados por VAE quanto os gerados por *naive random* conseguiram revelar o conjunto de tokens problemático, consequentemente, testando a robustez da representação e descobrindo a causa da invalidade da cadeia de caracteres. No entanto, o VAE gerou uma porcentagem significativamente maior (88,76%) de SELFIES inválidos em comparação com o modelo *naive random* (26,12%). Isso indicou uma margem de desempenho superior a 63% na tarefa de minimizar a porcentagem de validade nos conjuntos de SELFIES gerados. Como a invalidade foi duplamente condicionada à presença de tokens problemáticos e à vizinhança dos tokens, o VAE pôde explorar um número maior de possibilidades de selecionar e colocar incorretamente tokens na vizinhança daqueles que resultaram na invalidação dos SELFIES. Além disso, a organização radial da validade permitiu uma geração direcionada de conjuntos do SELFIES pela porcentagem de validade desejada, possibilitando a investigação dos padrões de validação e invalidação do SELFIES ao longo de um *continuum*. De fato, como diferentes raios (superfícies de hiper esferas) decodificam conjuntos do SELFIES com porcentagens de validade variadas, o VAE pode ser interpretado como uma coleção de geradores – cada um com sua própria propriedade geradora definida em um parâmetro de validade.

Embora o VAE supere os modelos nulos na minimização da validade em seu domínio de aplicabilidade e ofereça vantagens como a organização radial contínua, a abordagem apresenta algumas limitações. Por ser um modelo de aprendizado profundo, com mais de um milhão de parâmetros no decodificador, o processo de geração é computacionalmente custoso. Holisticamente, a amostragem e a geração de vetores aleatórios por meio da decodificação são aproximadamente 100 vezes mais lentas que o modelo nulo mais rápido, o *shuffle random*. Em termos de consumo de memória, o decodificador consumiu 1600 vezes a quantidade de espaço consumida pelo gerador nulo menos custoso, o *naive random* (Tabela 13).

Tabela 13 – Consumo de tempo e memória para gerar um conjunto de 10.192 cadeias de caracteres SELFIES. Os modelos *shuffle random* e *naive random* foram usados como base para calcular a razão de consumo de tempo e memória, respectivamente. (NR = naive random, SR = shuffle random, ITDR = index-token distribution random, VAE = Autocodificador Variacional).

Modelo	Razão de Tempo	Razão de Memória	
SR	1,00	1482,78	
NR	1,50	1,00	
ITDR	6,40	1596,67	
VAE	102,55	Parâmetros do modelo inteiro	2370,69
		Parâmetros do decodificador	1659,48

Fonte: Elaborada pelo autor.

Foi possível gerar conjuntos SELFIES de tamanho 10.192 em 994 raios, de 6 a 1.000, uniformemente espaçados, com uma originalidade de 1,0 (fração de cadeias geradas que são exclusivas), com uma originalidade menor para raios menores (R < 6). No entanto, é esperado que o modelo se aproxime de um estado de saturação em raios maiores, o que tende a diminuir a originalidade. Isso deve ocorrer porque o número de cadeias únicas que podem ser geradas pela amostragem dentro ou na superfície das hiper esferas é limitado tanto pela geometria (volume e área de superfície) quanto pelos gradientes (sensibilidade de saída do decodificador às perturbações no vetor decodificado)

É importante enfatizar a condicionalidade dos resultados obtidos devido ao conjunto de treinamento (subconjunto ChEMBL v.29), aos hiper parâmetros do modelo e à abordagem de busca de variáveis latentes utilizados. No artigo original do SELFIES,[56] os autores treinaram um modelo VAE com SELFIES para demonstrar o aprendizado de um espaço latente químico 100% válido. Eles usaram como conjunto de dados de referência para treinamento do modelo o QM9[167-168] e decodificaram moléculas a partir da perspectiva de hiperplanos, em um espaço latente de 241 dimensões. Além disso, eles também testaram e afirmaram ter obtido 100% de robustez ao introduzirem mutações aleatórias (localmente, em posições selecionadas da sequência) em SELFIES e avaliaram sua validade. Neste trabalho, foi oferecido um caso alternativo de um autocodificador variacional que é capaz de aprender um espaço latente revelando modos de falha representacional ao gerar cadeias anômalas de SELFIES na decodificação hiper esférica.

Em aprendizado de máquina, os resultados do modelo geralmente são condicionados por hiper parâmetros tais como escolhas de arquitetura. No caso do

VAE, a busca exaustiva do espaço de hiper parâmetros pode levar a uma explosão combinatória. Considerando o modelo estudado, embora os valores exatos para a validade mínima e o raio correspondente devam depender dos hiper parâmetros, é esperado que a relação validade versus raio generativo deva sempre apresentar um único mínimo (que é global) independentemente dos hiper parâmetros, a menos que seja intencionalmente forçada a ser de outra forma. Além disso, se os critérios para gerar anomalias forem conhecidos de antemão, o VAE também poderá ser modificado para incorporar informações rotuladas, tais como rótulos de classes binárias normais e anômalas. Essas variantes de modelo permitiriam o condicionamento direcionado do espaço latente baseado nos critérios desejados.

4.5 CONCLUSÕES

Este estudo aplicou quatro modelos (três modelos nulos: *naive random*, *shuffle random* e index-token *distribution random*; e um modelo de aprendizagem profunda não supervisionada: autocodificador variacional) a um conjunto de dados de 400 mil cadeia de caracteres SELFIES, uma popular representação molecular, a fim de comparar quantitativamente a capacidade máxima de geração SELFIES anômalos, com critérios de anomalia definidos pela invalidade dos SELFIES – um resultado que desafia as premissas básicas da representação ser 100% válida. O estudo descobriu que o VAE supera o desempenho dos modelos nulos na tarefa de minimizar a porcentagem de validade em conjuntos SELFIES gerados pela decodificação de amostras numa exploração radial do espaço latente, estabelecendo um domínio de aplicabilidade do VAE como um gerador de anomalias representacionais superior aos demais modelos. Portanto, o trabalho serviu a dois propósitos principais: (i) gerou anomalia variacional profunda, e sua metodologia associada, como um método efetivo de testar a robustez de representações de moléculas – embora possa ser utilizado para outros tipos de dados – e revelar critérios de anomalias previamente desconhecidos; e ii) demonstrou uma prova de conceito com um estudo de caso usando SELFIES, um marco na direção do desenvolvimento da maximização da robustez de representações moleculares.

CAPÍTULO 3: BUSCA DE INIBIDORES PARA A M^{PRO} DE SARS-CoV-2

5 BUSCA DE INIBIDORES PARA A M^{pro} DE SARS-CoV-2

5.1 Panorama geral

No final de 2019, células do epitélio respiratório de pacientes de Wuhan, na China, que sofriam de pneumonia devido a causas desconhecidas, permitiram a identificação de um novo vírus, o coronavírus da síndrome respiratória aguda grave 2 (do inglês, *severe acute respiratory syndrome coronavirus 2*, ou, SARS-CoV-2), o qual foi responsável por causar a doença do coronavírus 2019 (ou COVID-19).[169] Essa doença se espalhou de maneira extremamente rápida ao redor do mundo e fez com que a Organização Mundial da Saúde declarasse pandemia desse novo coronavírus, em março de 2020. Além disso, a OMS recomendou a adoção de medidas de contenção da disseminação da doença para todos os países.[170-171] Em maio de 2023, foi oficialmente decretado fim do o estado de Emergência de Saúde Pública de Importância Internacional (ESPII).

De acordo com a OMS, em todo o mundo, do início da pandemia até a primeira semana epidemiológica de novembro, mais de 771 milhões de casos foram confirmados, os quais resultaram em mais de 6,977 milhões de mortes acumuladas para o mesmo período. No Brasil, no início de novembro de 2023, mais de 37 milhões de casos também já haviam sido confirmados, dos quais mais de 705 mil resultaram em óbitos, colocando o país em segundo lugar em número de mortes e em sexto em número de casos.[171] A Figura 26 mostra a distribuição de mortes ao redor do mundo desde o começo da pandemia até novembro de 2023.

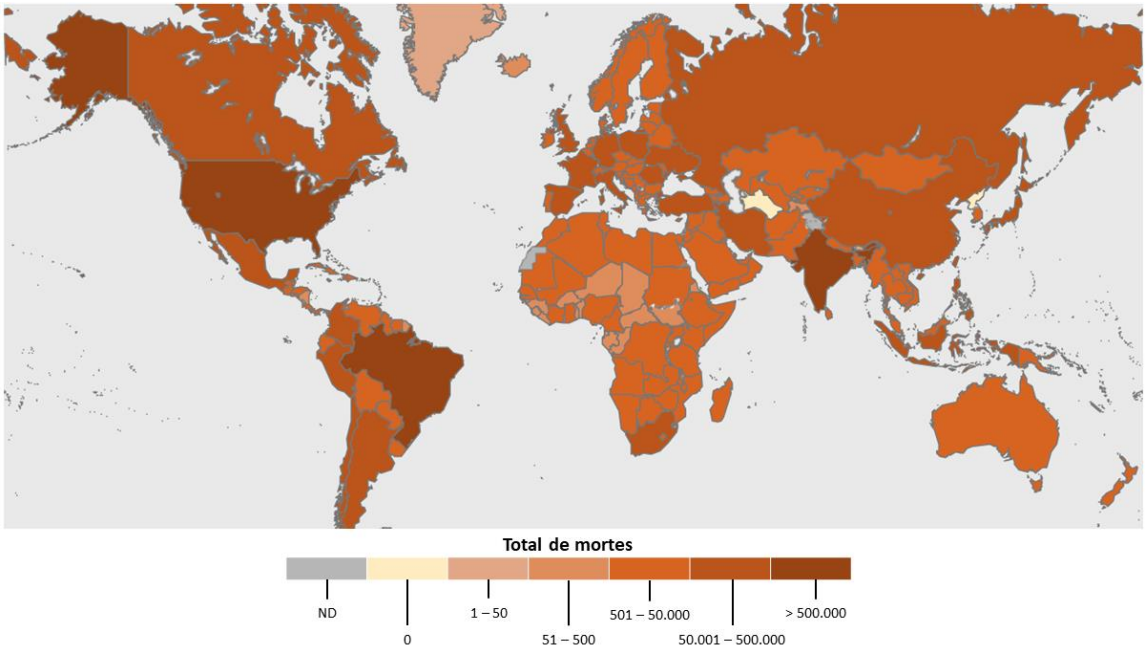


Figura 26 – Total de mortes causadas por COVID-19 desde o começo da pandemia, em 2023, até a primeira semana de novembro de 2023, numa escala de cor na qual quanto mais intenso o vermelho, maior o número de mortos.
Fonte: Adaptada de WORLD HEALTH ORGANIZATION.[171]

O SARS-CoV-2 é um betacoronavírus pertencente à família *Coronaviridae*, assim como o SARS-CoV e o MERS-CoV (*Middle East* respiratory syndrome coronavirus), os quais foram responsáveis pelos surtos de 2002 e 2012, respectivamente, e por tornar os coronavírus uma preocupação significativa para a saúde pública no século XXI.[172] Embora seja provável que nunca se estabeleça como e quando exatamente o SARS-CoV-2 foi transmitido à população humana pela primeira vez, os dados disponíveis até o momento sustentam a hipótese de que esse vírus passou por circulação enzoótica antes de se propagar para os humanos.[173]

Assim como a maior parte dos vírus de RNA, os coronavírus evoluem rapidamente (em questão de meses ou anos) a uma frequência que é observável e mensurável, com a sua evolução sendo impulsionada pela taxa na qual as mutações são geradas e se disseminam nas populações.[174] As mutações adaptativas que ocorrem no genoma viral podem alterar significativamente o comportamento do vírus, como é o caso da mutação D614G na proteína *spike* do SARS-CoV-2, em que uma única troca de aminoácidos resultou numa cepa cuja transmissibilidade é muito mais elevada.[175]

A COVID-19, inicialmente, compromete o sistema respiratório e sua disseminação entre as pessoas ocorre principalmente através de gotículas expelidas

durante tosse e espirro.[176] Muitas transmissões ocorrem em situações de contato próximo com indivíduos pré-sintomáticos, assintomáticos ou sintomáticos. Além disso, procedimentos que geram aerossol e a contaminação de superfícies pelo vírus também representam formas significativas de propagação da doença.[176] Os sintomas costumam aparecer entre cinco e seis dias depois da exposição ao vírus e geralmente duram entre 11 e 14 dias.[177] Dentre os sintomas mais comuns para a doença estão a tosse, febre, falta de ar e dor de garganta, entretanto, também podem ser observados sintomas gastrointestinais, fadiga, aumento da produção de escarro e dor de cabeça. Os sintomas mais graves podem incluir choque séptico, acidose metabólica e disfunção da coagulação, os quais podem levar à morte.[178]

Diante da gravidade dos sintomas e consequências associados à COVID-19, a busca por estratégias eficazes de prevenção e controle tornou-se imprescindível e, no contexto do combate à pandemia, um enorme esforço feito pela comunidade científica para a descoberta e desenvolvimento de medidas que incluem vacinas e medicamentos para a doença.[179] O desenvolvimento e implementação das vacinas específicas para COVID-19 representam um marco significativo para a ciência global, com vários imunizantes recebendo autorização para uso emergencial em menos de um ano do começo da pandemia.[179] Essas vacinas têm demonstrado eficácia em reduzir a gravidade da doença, prevenir hospitalizações e, principalmente, salvar vidas.[180] No Brasil, no começo de novembro de 2023, seis vacinas com tecnologias diferentes tinham sido autorizadas pela Agência Nacional de Vigilância Sanitária (ANVISA): a Comirnaty (Pfizer/Wyeth), a Comirnaty bivalente (Pfizer) e a Spikevax bivalente, que utilizam RNA mensageiro, a Janssen Vaccine (Janssen-Cilag) e a Oxford/Covishield (Fiocruz e Astrazeneca), que utilizam adenovírus, e a Coronavac (Butantan), que utiliza antígeno do vírus inativado.[181]

Para além da vacinação, a necessidade por tratamentos complementares por meio de fármacos, especialmente aqueles que tenham amplo espectro e possam ser administrados oralmente, também tem sido crucial[179] e movimentado os esforços de diversos pesquisadores ao redor do mundo, permitindo que grandes progressos também fossem alcançados nessa direção.[176] Assim, ao final de outubro de 2023, a ANVISA listava seis medicamentos aprovados para uso emergencial e registro sanitário com esse propósito: Remdesivir, Molnupiravir, Paxlovid, Baricitinibe, Sotrovimabe e Tocilizumabe.[182] O Remdesivir é um antiviral injetável de amplo espectro, disponível apenas para os pacientes hospitalizados, que reduz a replicação

do vírus ao inibir a polimerase de RNA (*RdRp*, do inglês, *RNA-dependent RNA polymerase*).[182-184] O Molnupiravir é um pró-fármaco antiviral de amplo espectro de uso oral que adiciona mutações deletérias ao vírus por meio da RNA polimerase.[182,185-186] O Paxlovid é um antiviral de uso oral que combina dois compostos diferentes, o nimartrelvir, que inibe a protease principal do vírus impedindo o processo de clivagem das suas poliproteínas, e o ritonavir, que inibe o citocromo P450 3A4.[182,187] O Baricitinibe é, originalmente, um medicamento para o tratamento de artrite reumatoide[188] que foi reposicionado para o tratamento de pacientes de COVID-19 hospitalizados, agindo na inibição das janus quinases envolvidadas em processos inflamatórios, hematopoiese e em função imunológica.[182,189-190] O Sotrovimabe é um anticorpo monoclonal, de uso restrito a hospitais, que neutraliza o vírus ao se ligar à proteína *spike*.[182,191] O Tocilizumabe é um anticorpo monoclonal bloqueador de receptores da interleucina-6, inicialmente utilizado para o tratamento de artrite, que tem sido empregado em pacientes hospitalizados que estão recebendo corticoides e aumentando a taxa de sobrevivência deles.[182,192-193] A Figura 27 mostra as estruturas das moléculas que compõem os quatro medicamentos não-anticorpos aprovados pela ANVISA.

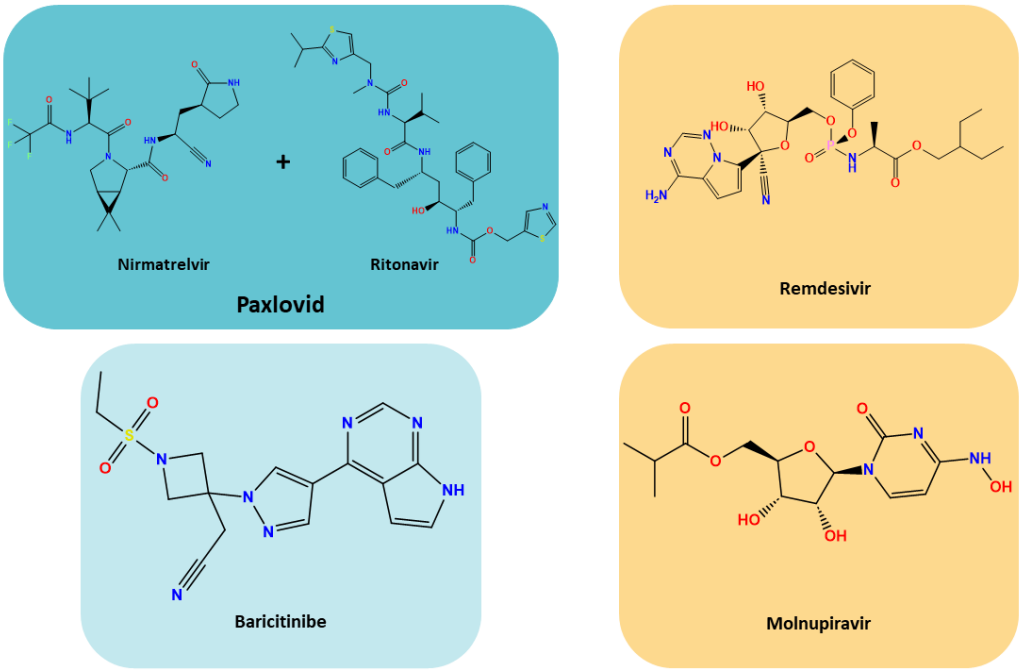


Figura 27 – Medicamentos aprovados pela ANVISA para o tratamento da COVID-19. Em azul escuro está o Paxlovid, comercializado pela Pfizer, cujo alvo é a protease principal do SARS-CoV-2. Em amarelo, os compostos que têm como alvo a de RNA polimerase, com o Remdesivir comercializado pela Gilead, e o Molnupiravir. Em azul claro, o composto inibidor da janus quinases.

Fonte: Elaborada pelo autor.

Embora diferentes vacinas tenham sido desenvolvidas, aprovadas e distribuídas pelo mundo em tempo recorde[194] e algumas alternativas para o tratamento de pacientes infectados pelo vírus estejam disponíveis,[185] a susceptibilidade que o SARS-CoV-2 tem às mutações exige que a eficácia e cobertura dos protocolos de vacinação e tratamentos sejam constantemente verificados. Assim, a busca por tratamentos alternativos e efetivos ainda deve ser considerada uma prioridade no controle e manutenção da doença.[174,195]

As estratégias para a descoberta de fármacos para COVID-19 podem ser divididas em duas categorias: i) aquelas que visam fatores do hospedeiro ou ii) aquelas que visam as proteínas do vírus, que são importantes para seu ciclo ou para a infecção.[179] A estratégia utilizada neste trabalho teve como alvo uma proteína essencial para o ciclo do vírus, a protease principal (M^{pro}) de SARS-CoV-2.

5.1.1 Protease principal (M^{pro}) de SARS-CoV-2

O SARS-CoV-2 é um Vírus (+)ssRNA (do inglês, *positive-sense single-stranded RNA viroses*), isto é, possui uma única fita de RNA simples de sentido positivo, cujo genoma contém, aproximadamente, 30.000 pares de base.[196] Esse genoma codifica para 29 proteínas diferentes, sendo quatro delas estruturais (S, *spike*; E, *envelope*, M, *membrane*; e N, *nucleocapsid*), as quais, além de conferir estrutura, desempenham funções importantes na interação do vírus com os receptores celulares, incluindo a entrada viral e a interação com anticorpos.[197-199] Além disso, o genoma do SARS-CoV-2 codifica para 16 proteínas não-estruturais (nsp1 – nsp16) que compõem o complexo replicação e transcrição,[199-201] e nove proteínas acessórias importantes para a interação vírus-hospedeiro que afetam sua imunidade bem como a proliferação viral.[199]

Uma proteína extremamente conservada entre todos os coronavírus conhecidos é a proteína não-estrutural 5 (nsp5), também chamada protease principal (M^{pro}) ou proteína similar a quimotripsina (3CLpro). Essa é uma proteína não estrutural pertencente à família das cisteíno proteases responsáveis pela clivagem proteolítica das poliproteínas virais e é essencial para o processo de replicação do vírus,[200] além de compartilhar de 96% de identidade sequencial com a M^{pro} de SARS-CoV (responsável pelo surto do vírus em 2002).[202] Essa proteína possui três domínios estruturais característicos.[203, 204] Os domínios I (resíduos 8 ao 101) e II (resíduos 102

ao 184) contêm barris β que formam uma estrutura similar à quimotripsina. Entre esses dois domínios há uma fenda na qual se localiza o sítio ativo e a díade catalítica Cys145-His41.[205] Além disso, uma alça (resíduos 185 ao 200) conecta os domínios II e III (resíduos 201 ao 306) sendo este último formado por cinco hélices α que adotam uma estrutura globular.[205] A Figura 28 ilustra a estrutura da protease principal de SARS-CoV-2.

A M^{pro} tem função central na replicação e transcrição viral, o que a torna um alvo bastante interessante para a descoberta e desenvolvimento de candidatos a fármacos antivirais. Além disso, o sítio de clivagem da M^{pro} é bem conhecido e, até o momento, não há nenhuma protease humana com um sítio ativo similar que tenha sido descrita. Essa característica única favorece a descoberta de inibidores seletivos.[206] Diante disso, a M^{pro} de SARS-CoV-2 tem sido investigada como alvo molecular de diversos estudos de triagens experimental e virtual.[207-212] Além disso, o desenvolvimento surpreendentemente rápido de um medicamento aprovado cujo alvo é essa proteína188 fortaleceu ainda mais a perspectiva de utilizá-la como base na busca por tratamentos eficazes para a COVID-19.

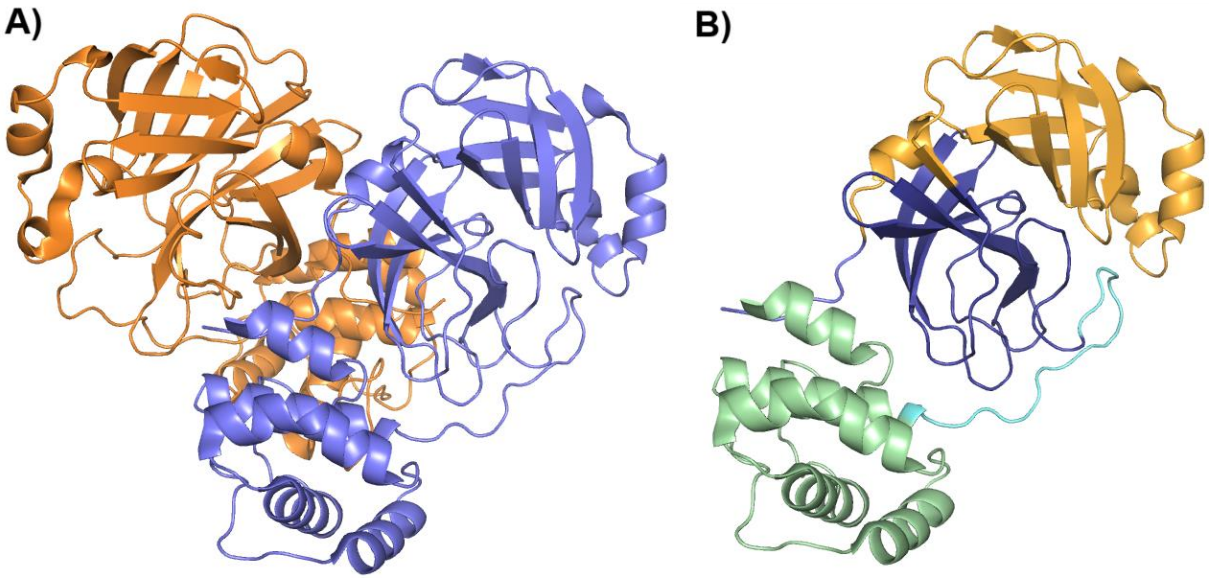


Figura 28 – Estrutura tridimensional da M^{pro} de SARS-Cov-2 (PDB ID: 7SFH), resolvida por difração de raios-x. A) Representação do dímero da proteína, com cada monômero de uma cor. B) Monômero colorido com base nos domínios estruturais. Em laranja está representado o domínio I, em azul, o domínio II, e em verde, o domínio III. A alça que liga os domínios II e III está colorida em ciano.

Fonte: Elaborada pelo autor.

5.1.2 Produtos naturais

Os métodos computacionais proporcionam um ganho substancial de tempo, além da redução de custos nos estágios iniciais do desenvolvimento de fármacos, tornando possível a triagem de bases de dados com um número bastante grande de compostos.[213] Neste contexto, um número expressivo de artigos (> 300 na base de dados PubMed – termos da busca: covid-19; M^{pro}; *virtual screening*) foram publicados entre 2020 e 2023 em periódicos com seletiva política editorial que utilizaram a M^{pro} como alvo molecular em campanhas de triagem virtual. As triagens virtuais permitem a avaliação de extensas bibliotecas de compostos em busca de potenciais inibidores para um alvo específico.[213] Essas bibliotecas abrangem uma variedade de moléculas, incluindo substâncias já conhecidas, compostos sintéticos e produtos naturais provenientes de diversas fontes.[214]

Os produtos naturais têm sido utilizados através dos séculos como agentes terapêuticos,[215] e desempenharam um papel essencial na descoberta de fármacos, principalmente em pesquisas para o tratamento de câncer e doenças infecciosas, embora diversas outras áreas também tenham se beneficiado bastante.[215-217] Os produtos naturais são altamente otimizados durante o processo evolutivo, estruturalmente falando, para desempenhar funções biológicas específicas,[215-216] como a regulação de mecanismos de defesa endógenos e competição com outros organismos, explicando sua alta relevância para doenças infecciosas e câncer.[216]

Comparados aos compostos sintéticos, os produtos naturais apresentam algumas vantagens no processo de descoberta de fármacos, por exemplo, a diversidade e complexidade das estruturas que eles possuem e que não são comumente em bibliotecas de compostos sintéticos.[216,218] No geral, os compostos extraídos de produtos naturais possuem um menor número de nitrogênios e halogênios em sua estrutura, entretanto, o número de oxigênios costuma ser maior, bem como um menor coeficiente de partição octanol-água e um maior número de doadores e aceptores de ligações de hidrogênio.[215-216]

O Brasil possui uma biodiversidade extremamente rica, abrigando entre 10% e 20% de todas as espécies vivas do mundo.[219] Dadas as características promissoras de produtos naturais e lavando-se em conta que eles são uma das principais fontes de compostos para medicamentos e cosméticos,[219] a disponibilidade de bibliotecas com compostos dessa natureza é bastante benéfica para realização de triagens *in*

vitro e *in silico*.[220] Nesse sentido, o Núcleo de Bioensaios, Ecofisiologia e Biossíntese de Produtos Naturais (NuBBE) desenvolve e fornece, gratuitamente, uma base de dados denominada $NuBBE_{DB}$ com informações de mais de 2.100 produtos naturais e derivados identificados em espécies encontradas na biodiversidade brasileira.[219-220]

Assim, este trabalho visou a descoberta de potenciais inibidores da atividade da protease principal do SARS-CoV-2, por meio de triagem virtual utilizando compostos extraídos da biodiversidade brasileira.

5.2 METODOLOGIA

5.2.1 Estrutura do alvo

Para a realização dos estudos de triagem virtual, a estrutura tridimensional da M^{pro} de SARS-CoV-2 foi obtida do repositório RCSB PDB[221] (http://www.rcsb.org/). Foi utilizada a estrutura cujas coordenadas foram depositadas sob o código PDB ID 6LU7 (resolução 2,16 Å),[205] disponível logo no início da pandemia. A identidade sequencial entre a M^{pro} (PDB ID 6LU7) e a cepa circulante no Brasil nos primeiros meses de pandemia (B.1.1) é de 100%. Essa estrutura contém uma molécula ligada ao sítio catalítico (N-[(5-metilisoxazol-3-il)carbonil]alanil-L-valil-N~1~-((1R,2Z)-4-(benziloxi)-4-oxo-1-{[(3R)-2-oxopirrolidin-3-il]metil}but-2-enil)-L-leucinamida) entre os domínios I e II da M^{pro}, o qual foi usado neste trabalho.

5.2.2 Biblioteca de compostos

Para a triagem de compostos contra o alvo, foi utilizado um subconjunto de moléculas do banco de dados $NuBBE_{DB}$.[219-220] O subconjunto selecionado apresentou moléculas com valor de log P ≤ 3; massa molecular ≤ 600 Da; número de átomos doadores e aceptores de hidrogênio ≤ 5 e ≤ 10, respectivamente; área de superfície polar total (TPSA) ≤ 300 Å²; e número de ligações rotacionáveis ≤ 25. No total, o subconjunto selecionado apresentou 772 moléculas distintas.

5.2.3 Triagem virtual

A triagem virtual foi realizada com as 772 moléculas do banco de dados $NuBBE_{DB}$ utilizando o servidor online *MTiScreenOpen*.[222] Foi utilizado o ligante cristalizado no sítio catalítico entre os domínios I e II da M^{pro} como referência para a definição do sítio de ligação da estrutura. O software AutoDockTools-1.5.6[223] foi usado para centrar a caixa de busca nas coordenadas X = -22, Y = 3 e Z = -29 e dimensões de arestas de 22 Å x 20 Å x 18 Å. O servidor retornou as melhores 4468 poses, isto é, para cada uma das 772 moléculas o *software* encontrou até 9 modos de ligação diferentes. O programa considerou que poses do mesmo composto que apresentaram valores de desvio quadrático médio (RMSD) ≤ 2,5 Å haviam convergido. Nestes casos, o processo de modelagem foi terminado antes de explorar os 9 modos de ligação pré-ajustados. De acordo com a afinidade de ligação à protease do vírus, os 10 compostos melhores posicionados foram escolhidos para dar seguimento às análises.

5.2.4 Dinâmica molecular

Para as 10 moléculas mais bem pontuadas na etapa de triagem virtual, foram realizados 100 ns de simulações de dinâmica molecular usando o pacote de programas *AMBER18*[224] para a avaliação do modo de ligação predito bem como estimar a energia de ligação. Inicialmente, todos os hidrogênios foram adicionados com o *software reduce*,[225] pertencente ao pacote *AmberTools*.[224] Todos os ligantes foram parametrizados com o campo de força *GAFF* (*General Amber Force Field*)[226] com cargas atômicas do tipo *AM1-BCC*,[227] enquanto que o receptor foi parametrizado utilizando o campo de força *AMBER FF14SB*.[228] Todos os complexos foram envolvidos por uma caixa cúbica de água (e.g., *TIP3P*[229]) com 12 Å de distância mínima entre a borda da caixa de solvatação e o átomo mais próximo da estrutura. Além disso, quatro átomos de sódio na forma iônica foram adicionados aos sistemas para mantê-los eletronicamente estáveis. Energias livres foram calculadas com MM-GBSA, utilizando a ferramenta *pymdpbsa* do pacote *AMBER18*.[224]

5.2.5 Metadinâmica

As energias livres de ligação relativas entre a M^{pro} e os ligantes também foram calculadas por metadinâmica.[230] Nesse trabalho, foi estabelecida colaboração com os pesquisadores Matheus Victor Ferreira Ferras e o Prof. Dr. Roberto Dias Lins Neto da Fundação Oswaldo Cruz, Recife. A escolha das variáveis coletivas (CVs) foi feita satisfazendo o estabelecido por Laio e Gervasio (2008)[231] para melhor caracterizar todos os pequenos eventos relevantes para o processo. A metadinâmica foi realizada com um viés para duas CVs: CV1 e CV2. A primeira representa a distância entre o centro de massa dos átomos do ligantes e os carbonos-α dos resíduos num raio de 0,35 nm no *frame* inicial. A segunda representa o número de contatos entre os átomos dos ligantes e os átomos dos resíduos num raio de 0,35 nm. A exploração do espaço de fase das CVs foi feita adicionando potenciais gaussianos de 0,25 kJ/mol de altura e 0,2 nm de profundidade a cada 2 ps. O ponto inicial correspondeu ao último *frame* da simulação de dinâmica molecular. A amostragem e cálculo das CVs foram feitas por 10 ns com o plugin PLUMED 2.3.5[232-233] do software GROMACS v. 4.6.7.[234] Os ligantes foram separados da proteína ao longo dos caminhos das CVs e a superfície de energia livre foi recursivamente reconstruída usando a ferramenta *sum_hills*.

5.2.6 Ensaio de atividade inibitória e competitividade

A clonagem do gene referente à M^{pro}, a expressão da proteína e sua purificação foram feitas de acordo com o protocolo descrito por Noske *et al*.[235]

A atividade da proteína foi determinada em ensaio experimental baseado em FRET, usando o dabcyl-KTSAVLQ/SGFRKME-Edans-NH_2 como substrato fluorogênico.[235] Os experimentos foram conduzidos em tampão de 20 mM Tris·HCl (pH 7,3), 1 mM EDTA, 1 mM DDT, 0,01% Triton. A intensidade de fluorescência foi monitorada com um leitor de microplacas (SpectraMax Gemini EM), a cada 30 segundos, durante 30 minutos, a 37ºC, usando os comprimentos de onda para excitação e emissão de 360 nm e 460 nm, respectivamente.

Para a determinação do valor de IC_{50}, o composto selecionado foi dissolvido em DMSO e distribuído em placa de 96 poços (10 concentrações que variaram entre 12.000 a 23 μM); 0,015 μM da proteína; e 20 μM de substrato. As amostras foram pré-incubadas com as diferentes concentrações de composto a 37ºC por 30 minutos. As

reações de controle foram feitas na presença e na ausência da M^{pro}, contendo o mesmo volume de DMSO sem a presença do inibidor. Os experimentos foram realizados em triplicatas e as velocidades de reação foram calculadas por meio da inclinação da fase linear das reações.

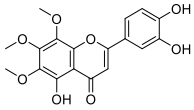
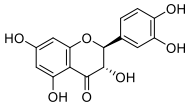
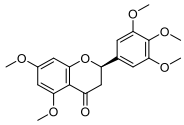
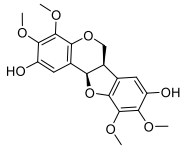
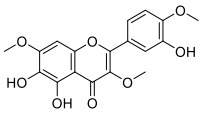
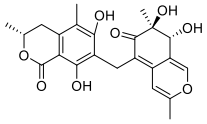
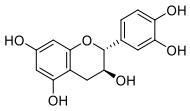
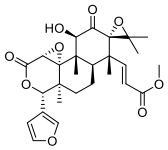
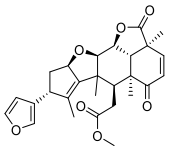
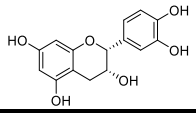
Para determinar o tipo de competitividade do inibidor, a atividade da proteína foi testada frente a quatro diferentes combinações de concentrações de composto e substrato: i) substrato na concentração de K_M, sem inibidor; ii) composto na concentração do IC_{50} determinado, sem substrato; iii) substrato na concentração de K_M, composto em 10 vezes a concentração do IC_{50}; e iv) substrato em 10 vezes a concentração de K_M, composto na concentração do IC_{50}. As amostras foram pré-incubadas com as diferentes concentrações de composto e substrato a 37ºC por 30 minutos. Os experimentos foram realizados em triplicatas.

5.3 RESULTADOS

5.3.1 Triagem virtual

No total, a base $NuBBE_{DB}$ possui 2.147 compostos, dos quais 78% são isolados de plantas, 15% são produtos semissintéticos, 5% são isolados de microrganismos, cerca de 1,6% são produtos de biotransformação e, aproximadamente, 0,2% são compostos isolados do ambiente marinho.[219] Como o foco deste trabalho foi identificar produtos naturais como candidatos a inibidores da M^{pro} para o desenvolvimento de fármacos antivirais, um conjunto de filtros moleculares para selecionar compostos com características mais próximas de fármaco-similar (do inglês, *drug-like*) foi aplicado. Assim, a aplicação dos critérios descritos na seção 5.2.2 selecionou 772 compostos. Os modos de ligação desses compostos foram modelados e a energia referente à interação com os resíduos do sítio catalítico da M^{pro} de SARS-CoV-2 foi estimada usando o servidor online *MTiScreenOpen*[222], que utiliza o software de modelagem *AutoDock*. Os resultados obtidos indicaram um espectro de energia de interação que variaram entre -9,5 $kcal.mol^{-1}$ e -1,4 $kcal.mol^{-1}$. O valor limite empregado para a seleção dos compostos como candidatos a *hits* foi -5,8 $kcal.mol^{-1}$. Esse valor refere-se à interação para o modo de ligação do inibidor cristalizado determinada pelo *Autodock Vina*.[236] Portanto, entre os *hits*, as 10 moléculas mais bem posicionadas na triagem virtual foram selecionadas (Tabela 14).

Tabela 14 – Estrutura e afinidades de interação calculadas pelos métodos de docagem molecular, MM-GBSA e metadinâmica dos compostos naturais extraídos da biodiversidade brasileira. Os compostos estão classificados de acordo com a energia predita pela docagem molecular durante a campanha de triagem virtual.

Posição	Nome	Estrutura	Classe química	Afinidade docking (kcal/mol)	MM-GBSA (kcal/mol)	Metadinâmica (kcal/mol)	Ref.
1	NuBBE_2523		Flavonoide	-9,5	-19,08	-11,49	237
2	NuBBE_139		Flavonoide	-8,9	-25,80	-23,54	238
3	NuBBE_1472		Flavonoide	-8,6	-	-19,79	239
4	NuBBE_2049		Flavonoide	-8,3	-21,68	-15,81	240
5	NuBBE_1874		Flavonoide	-8,2	-	-15,47	241
6	NuBBE_507		Policetídeo	-8,2	-9,55	-14,67	242
7	NuBBE_287		Flavonoide	-8,1	-16,88	-12,63	243, 244, 245, 246, 247, 248
8	NuBBE_1635		Terpeno	-8,1	-28,01	-9,55	249
9	NuBBE_1292		Terpeno	-8,1	-27,83	-4,21	250
10	NuBBE_866		Flavonoide	-8,0	-	-	243, 251

Fonte: Elaborada pelo autor.

5.3.2 Dinâmica molecular

Os 10 compostos mais bem posicionados na triagem virtual foram submetidos a 100 ns de simulações de dinâmica molecular (MD) com o pacote de ferramentas AMBER18, utilizando como referência a conformação obtida durante a triagem dos compostos, visando a confirmação das poses preditas pelo processo de docagem molecular e para avaliar a estabilidade do modo de interação predito bem como a energia livre referente à interação. A Figura 29 mostra como os valores de RMSD dos resíduos da M^{pro} e dos candidatos a ligantes evoluíram ao longo da trajetória da dinâmica, partindo-se das estruturas de referência (o modelo de interação predito pela docagem molecular).

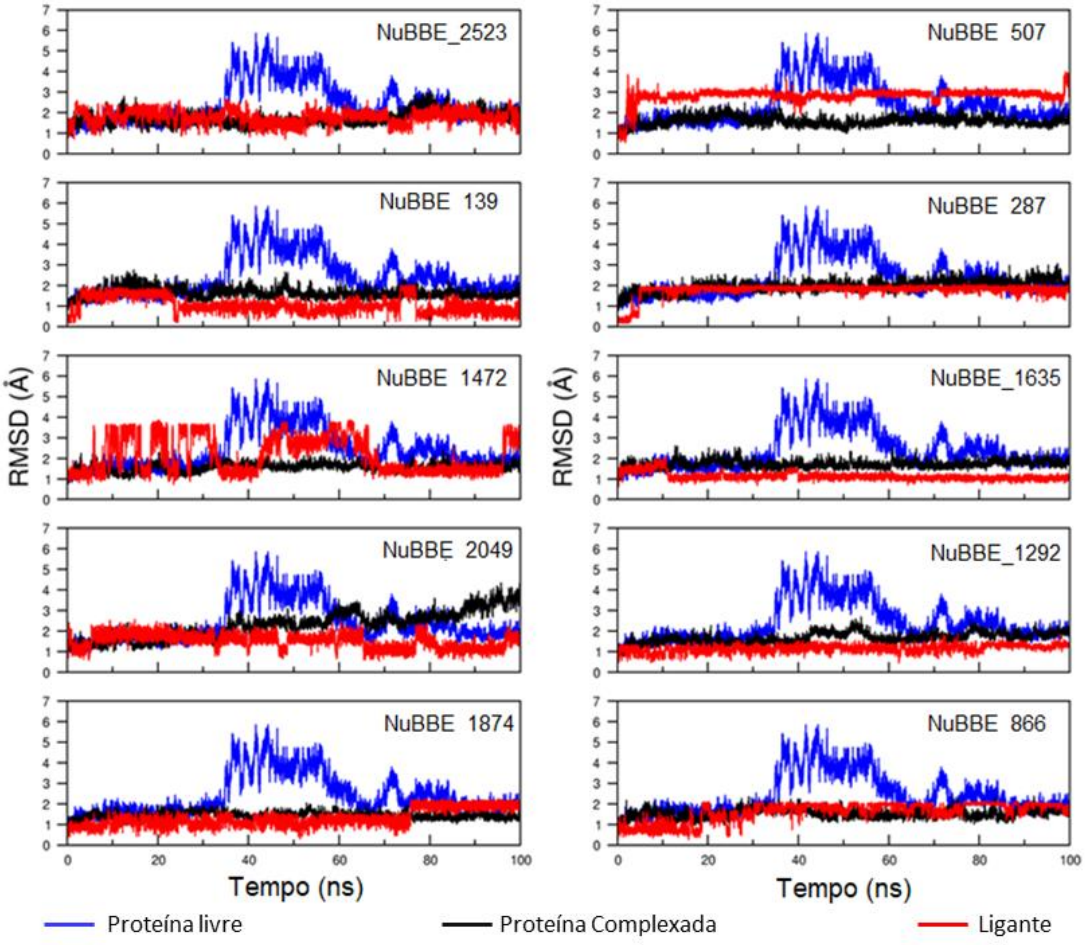


Figura 29 – Desvio quadrático médio (RMSD) durante os 100 ns de simulação de dinâmica molecular para a cadeia principal dos resíduos da estrutura da M^{pro} livre (azul), da cadeia principal dos resíduos em complexo com os ligantes (preto) e para os ligantes (vermelho).
Fonte: Elaborada pelo autor.

Os complexos formados pelos compostos NuBBE_2523, NuBBE_139, NuBBE_2049, NuBBE_287, NuBBE_1635 e NuBBE_1292 permaneceram estáveis

durante toda a simulação de dinâmica molecular, apresentando valores de RMSD em relação à pose inicial entre 2 e 3 Å. Entretanto, os compostos NuBBE_1472, NuBBE_1874, NuBBE_507 e NuBBE_866 saíram do sítio de ligação alguns nanosegundos após o início da simulação, sugerindo que os modos de ligação preditos para esses compostos pela docagem molecular não foram estáveis o suficiente para permitir que o complexo proteína-ligante se mantivesse.

A principal diferença entre os gráficos de RMSD da estrutura livre da M^{pro} com as estruturas em complexo com os ligantes é observada entre os tempos 35 e 75 ns de simulação (Figura 29). Essas diferenças nos valores de RMSD consistem principalmente nas mudanças conformacionais das alças do domínio III e da alça que liga o domínio II ao domínio III. Tais mudanças conformacionais estão em acordo com o trabalho de Suaréz e Díaz,[252] que avaliaram o comportamento da proteína na presença e na ausência de ligantes e descobriram que o domínio III é menos estável, isto é, tem maior flexibilidade conformacional, quando nenhum composto encontra-se interagindo com os resíduos do sítio de ligação entre os domínios II e III. Portanto, os experimentos de dinâmica molecular indicaram que o modo de ligação predito pelo *Autodock* para seis dos 10 hits virtuais foi estável ao longo de um tempo significativo de simulação (100 ns).

A flutuação quadrática média (RMSF) obtida para os últimos 50 ns de simulação (Figura 30) foi calculada para obter mais detalhes sobre os modos de interação preditos. Os resultados desta análise revelaram como as moléculas ligadas ao sítio ativo da proteína influenciaram na flexibilidade da cadeia principal dos resíduos da M^{pro}, com a estrutura da proteína apresentando um pequeno aumento na flexibilidade conformacional devido à presença dos ligantes. A única exceção observada refere-se ao complexo entre a proteína e a molécula NuBBE_287 (7º colocado na classificação), para o qual é possível notar um padrão de flutuação similar ao da proteína livre, indicando que as interações com o NuBBE_287 não alteraram o comportamento e flexibilidade dos resíduos da M^{pro}.

Para auxiliar na seleção dos compostos mais promissores para aquisição e validação experimental, estabelecemos uma colaboração científica com o Prof. Roberto Dias Lins Neto (Fundação Oswaldo Cruz, Recife). Nessa colaboração, simulações de metadinâmica com os 10 candidatos a *hits* foram conduzidas para avaliar as superfícies de energia livre dos ligantes ao se desassociarem do sítio ativo

da M^{pro}. A Figura 31 mostra os perfis de energia obtidos na metadinâmica para os seis compostos que não saíram do sítio nas simulações MD. A análise comparativa dos dados indica que os valores de energia calculados pelos estudos de metadinâmica estão correlacionados com os valores de energia livre de ligação calculados pela metodologia de MM-GBSA (Tabela 14) (r = 0,61, desconsiderando o NuBBE_1292 e NuBBE_1635). Especificamente, o NuBBE_139 destacou-se entre os 10 *hits* avaliados, pois apresentou perfil de dissociação mais promissor com a M^{pro} (indicado pelo poço de energia mais profundo e largo na Figura 31) e energias livres de -25,80 e -23,54 Kcal/mol, calculadas pelos métodos de metadinâmica e MM-GBSA, respectivamente (Tabela 14). Portanto, a análise integrada entre os métodos de docagem molecular e simulação indicaram que o NuBBE_139 (2º colocado na classificação na triagem virtual) seria o candidato mais promissor para avançar aos ensaios de validação experimental.

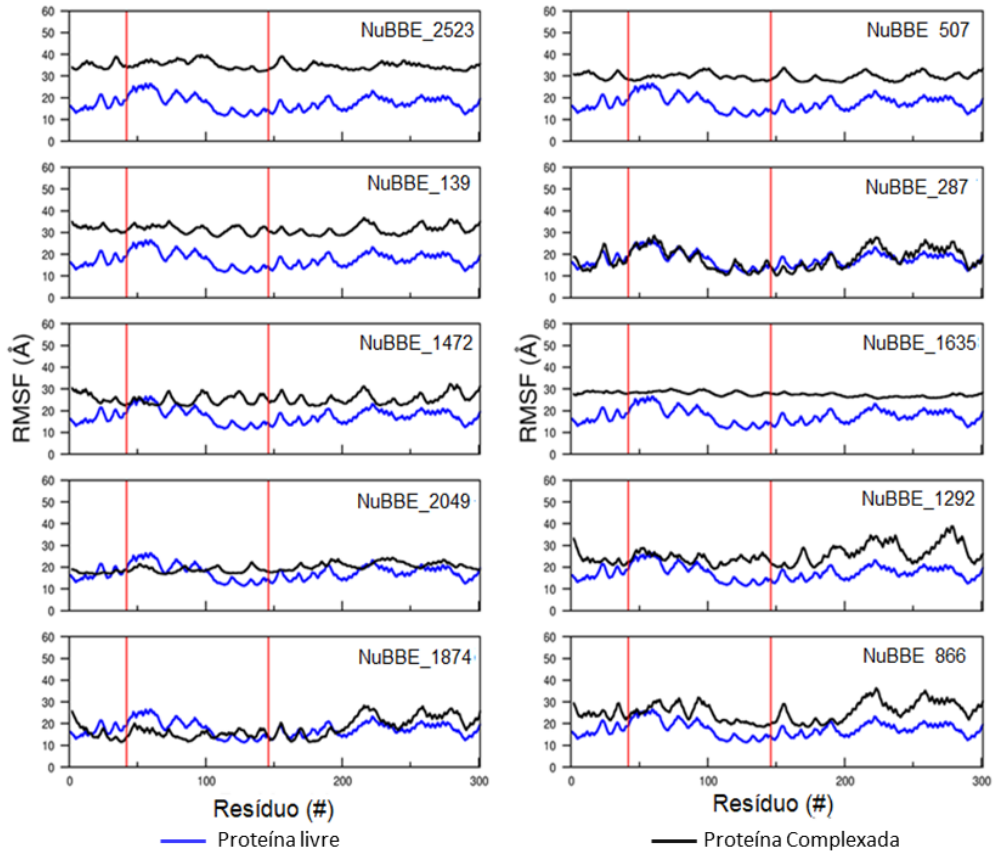


Figura 30 – Flutuação quadrática média (RMSF) para a estrutura da M^{pro} livre (em azul) e para os complexos (em preto). As linhas verticais, em vermelho, indicam a díade catalítica (H41 e C145).

Fonte: Elaborada pelo autor.

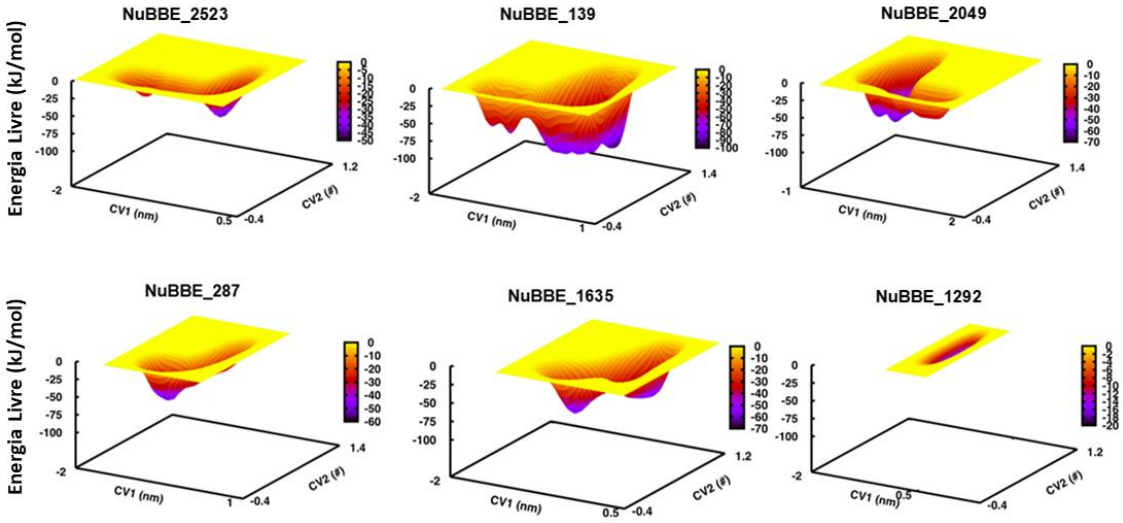


Figura 31 – Perfil da superfície de energia livre dos compostos da $NuBBE_{DB}$, ao se desassociarem do sítio ativo da M^{pro}, em função de CVs.
Fonte: Elaborada pelo autor.

A Figura 32 indica as principais interações polares e apolares entre o NuBBE_139 e os resíduos do sítio ativo da M^{pro} durante as simulações de dinâmica molecular. Esse modo de ligação é representativo de um cluster de conformações observado durante a trajetória da simulação. É possível observar a formação de interações com os resíduos chaves da díade catalítica, His41 e Cys145, além de interações polares (Gly143 e Glu166) e apolares (Met49, Asn142 e Leu27) com outros aminoácidos dos subsítios da M^{pro}.

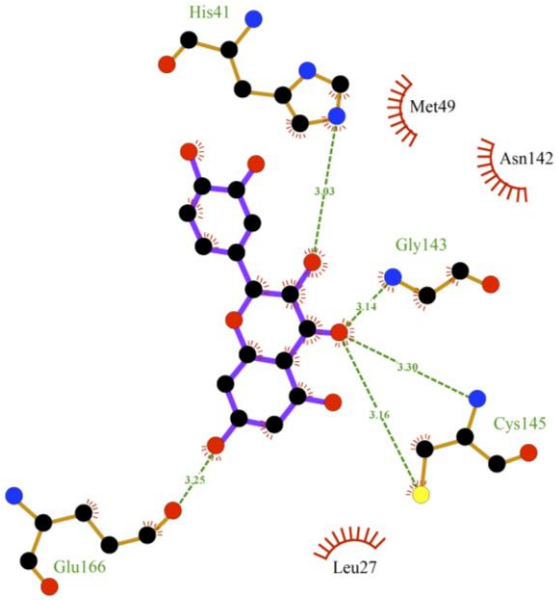


Figura 32 – Diagrama de interação entre os principais resíduos do sítio ativo da M^{pro} e o ligante NuBBE_139.
Fonte: Elaborada pelo autor.

5.3.3 Ensaios experimentais

Diante dos resultados das simulações de MD e metadinâmica apresentados acima, o composto NuBBE_139 (também conhecido como taxifolina) foi selecionado para aquisição e avaliação experimental da atividade inibitória contra a M^{pro} de SARS-CoV-2. Os estudos experimentais foram conduzidos pela doutoranda Mariana Ortiz de Godoy que verificou inibição da M^{pro} dependente da concentração de taxifolina. Como mostrado na Figura 33, a taxifolina se mostrou um inibidor submilimolar da atividade catalítica da M^{pro} (IC_{50} = 820 ± 3 μM). Para efeitos de comparação, a atividade inibitória da M^{pro} foi avaliada com um produto natural análogo da taxifolina, a quercetina, reportado na literatura como um inibidor com afinidade no baixo micromolar (K_i = 7,4 μM).[253] Como Abian *et al*. determinaram a atividade da quercetina em termos da constante de afinidade K_i, a inibição, em termos de IC_{50}, foi obtida experimentalmente utilizando o mesmo protocolo descrito na seção 5.2.3 e pode ser vista na Figura 33A. Diferentemente da taxifolina, a quercetina apresentou um IC_{50} = 19 ± 3 μM.

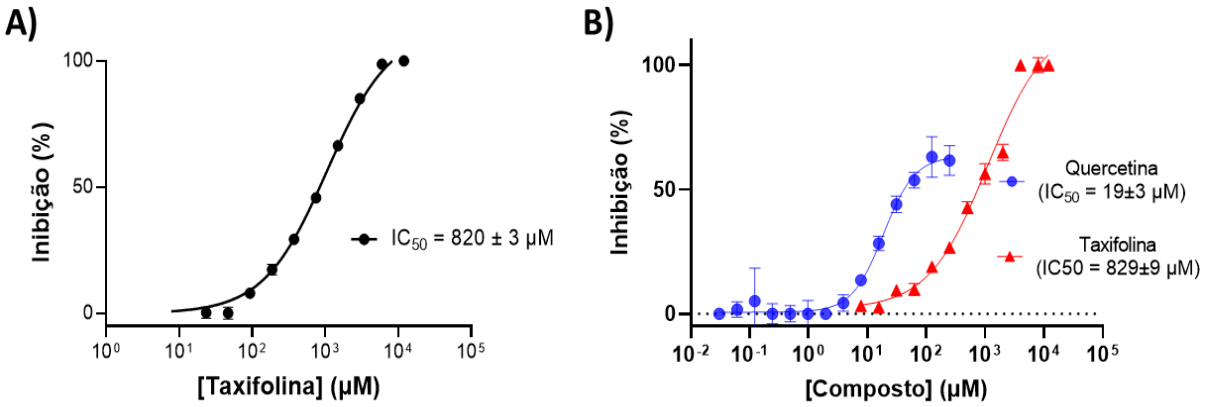


Figura 33 – A) Curva de inibição da taxifolina contra a M^{pro} de SARS-CoV-2, com IC_{50} = 820 ± 3 μM. B) Curva de inibição da taxifolina (vermelho) e do análogo quercetina (azul) contra a M^{pro} de SARS-CoV-2, com IC_{50} = 829 ± 9 μM e IC_{50} = 19 ± 3 μM, respectivamente.
Fonte: Elaborada pelo autor.

Por fim, a hipótese inicial assumida durante a realização deste estudo foi de que a taxifolina (e os demais compostos avaliados durante a triagem virtual), seriam inibidores competitivos com o substrato, ou seja, eles competiriam com o substrato da M^{pro} pelo seu sítio ativo. Para comprovar essa hipótese, um teste de inibição da proteína foi feito em quatro diferentes condições de concentração de substrato e inibidor (descritas na seção 3.3.6).

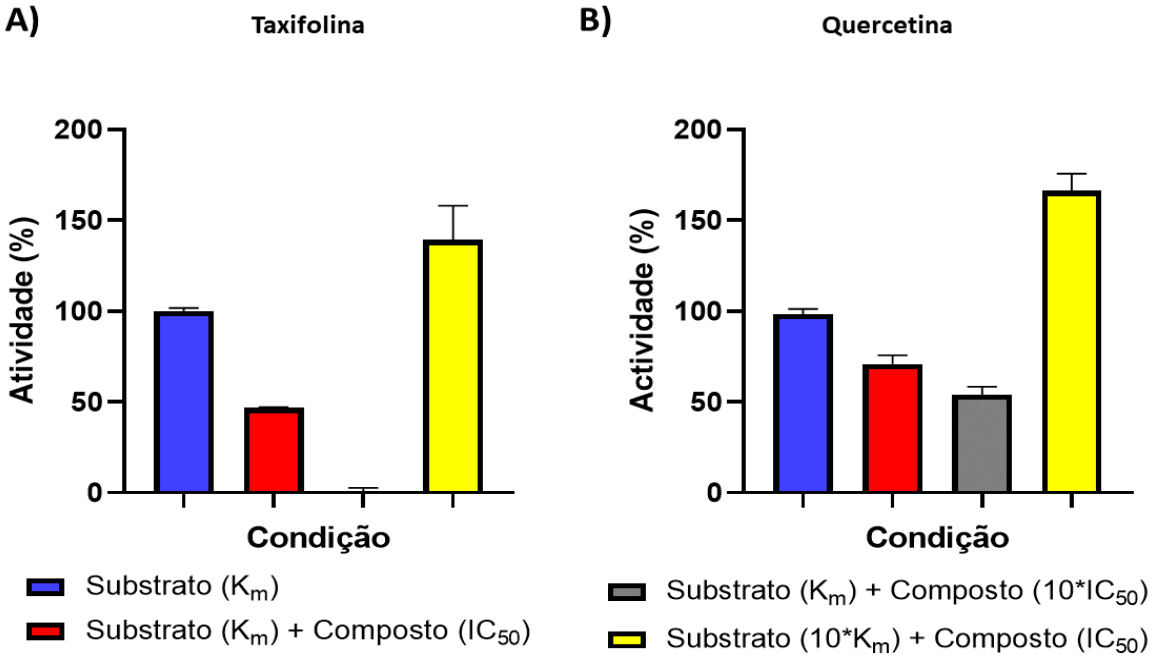


Figura 34 – Avaliação do tipo de competitividade da taxifolina (A) e da quercetina (B). Quatro condições foram testadas i) substrato na concentração de k_m, sem inibidor (azul); ii) composto na concentração do IC_{50} determinado, sem substrato (vermelho); iii) substrato na concentração de k_m, composto em 10 vezes a concentração do IC_{50}(cinza); e iv)

substrato em 10 vezes a concentração de k_m, composto na concentração do IC_{50} (amarelo).

Fonte: Elaborada pelo autor.

A Figura 34A apresenta esses resultados. Na presença apenas de substrato na concentração do valor de K_M, a proteína apresentou 100% de atividade. Quando a condição foi mudada para substrato na concentração de K_M e taxifolina na concentração do valor de IC_{50} a atividade da proteína reduziu para cerca de 50%, estando em acordo com o esperado. Quando a condição na qual a proteína foi submetida era de substrato na concentração do valor do K_M e taxifolina em 10 vezes a concentração do valor de IC_{50}, a leitura obtida foi de 100% de inibição da atividade da proteína. Por fim, quando a concentração de substrato foi 10 vezes o valor de K_M e da taxifolina foi no valor de IC_{50}, a presença de alta concentração do substrato reestabeleceu a atividade catalítica da proteína. Essa combinação de condições corroboraram a hipótese assumida inicialmente, confirmando que a taxifolina compete pelo mesmo sítio do substrato, sendo, portanto, um inibidor competitivo da M^{pro} de SARS-CoV-2.

A mesma análise foi feita com a quercetina (Figura 34B). Assim como para a taxifolina, a primeira condição (substrato na concentração do valor de K_M) indicou 100% de atividade da proteína. Entretanto, quando a concentração de substrato foi mantida no valor do K_M e a quercetina na concentração do valor de IC_{50}, a atividade da proteína não reduziu a 50%. Curiosamente, quando a concentração de quercetina foi aumentada para 10 vezes o valor de IC_{50}, a atividade da proteína caiu para 50%. Por fim, quando a concentração utilizada de substrato foi 10 vezes o valor de K_M e quercetina na concentração do valor de IC_{50}, a atividade catalítica foi reestabelecida. Esses dados sugerem que a quercetina também é um inibidor competitivo da M^{pro} de SARS-CoV-2, contudo, diferentemente da taxifolina, a quercetina não foi capaz de causar inibição maior que 50%, mesmo em concentrações tão elevadas como 10 vezes os valor do IC_{50}.

5.4 DISCUSSÃO

O NuBBE_139 é também conhecido como taxifolina, ou dihidroquercetina, e pertence à subclasse dos flavonóis, classe dos flavonoides. A taxifolina é um derivado polifenólico que foi isolado e identificado em flores de *Prerogyne nitens*

(*Caesalpinioidae*), uma espécie comum na América Latina e amplamente encontrada no Brasil.[238] Esse composto também foi isolado de extratos de galhos de *Rapanea lancifolia* (*Myrsinaceae*),[254] comumente encontrada nas regiões Sul e Sudeste do país.[255] A taxifolina possui dois centros estereogênicos no anel C, diferente de seu análogo natural quercetina, que não apresenta nenhum carbono assimétrico. O estereoisômero identificado neste trabalho consiste na (+)-taxifolina referente ao estereoisômero de configuração 2R, 3R. A Figura 35 ilustra o núcleo flavona, comum aos flavonoides, juntamente com os dois estereoisômeros da taxifolina.

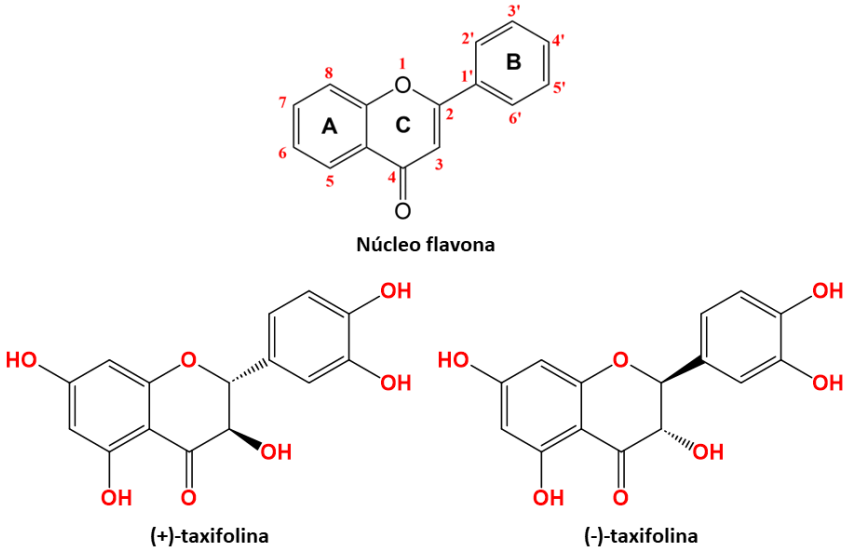


Figura 35 – Núcleo comum aos flavonoides (centro superior), que consiste de dois anéis fenil (A e B) e um heterocíclico (C) contendo o oxigênio. No canto inferior esquerdo está a estrutura do flavonoide (+)-taxifolina, utilizada neste trabalho. No canto inferior direito, o flavonoide (-)-taxifolina.

Fonte: Elaborada pelo autor.

A taxifolina (NuBBE_139) é descrita na literatura como possuindo atividade anti-inflamatória[256-257] e propriedades antioxidantes. Essa última está relacionada ao substituinte fenólico característico da classe dos flavonoides.[258-259] O potencial mutagênico da taxifolina também já foi avaliado,[260] sendo ela considerada não mutagênica e menos tóxica que seu análogo quercetina em estudo comparativo. Além de não ser mutagênico, a taxifolina possui propriedades antitumorais e pode prevenir danos genotóxicos causados por compostos carcinogênicos.[261-262] Neste sentido, a taxifolina é considerada um potencial agente quimiopreventivo, regulando a transcrição gênica por meio de um mecanismo dependente de elemento responsivo a antioxidantes (ARE, do inglês, *antioxidant response element*).[263] Nesse estudo, experimentos de transfecção transientes utilizando constructos modificados de ARE

demonstraram que a taxifolina ativou significativamente a transcrição de genes específicos que continham ARE.

Flavonoides são produtos naturais que apresentam atividade antiviral.[264] Relatos da literatura indicam que a taxifolina tem propriedades inibitórias frente aos vírus da hepatite A, *coxsackievirus* B4 e vírus da leucemia murina. Além disso, Bernatova & Liskova (2021)[265] demonstraram que a taxifolina apresentou propriedades relevantes em estudos pré-clínicos para o tratamento de pacientes hipertensos acometidos com infecção viral.

As simulações indicaram que a taxifolina interage com os resíduos da díade catalítica (His41 e Cys145) bem como com outros resíduos específicos do sítio catalítico da M^{pro} que já foram descritos em outros complexos com inibidores potentes. Por exemplo, o inibidor PF-07321332 (K_i = 3,11 nM), atual nirmatrelvir, um dos compostos presentes no Paxlovid comercializado pela *Pfizer* como medicamento contra COVID-19, estabelece contatos polares e apolares com os resíduos Gln189 e Met49 que se mostraram relevantes para o processo de reconhecimento molecular.[187] Além disso, Owen *et al.*[187] demonstraram que a interação com o resíduo Glu166 foi essencial para o aumento da potência e seletividade do nirmatrelvir. Essas interações auxiliaram na estabilização do modo de ligação do NuBBE_139 e foram fundamentais para a seleção deste composto para a avaliação experimental da atividade inibitória da M^{pro}.

Os ensaios experimentais mostraram que a taxifolina inibe a atividade catalítica da M^{pro} na faixa de submilimolar (IC_{50} = 820 ± 3 μM), enquanto seu análogo apresentou uma potência maior. Algumas características dessas moléculas podem justificar tal diferença, como as que seguem. Abian *et al.*[253] modelaram o modo de ligação para a quercetina no sítio ativo da M^{pro} por docagem molecular. Assim, um protocolo similar ao deles foi reproduzido aqui, utilizando com a mesma estrutura cristalográfica (PDB ID: 6Y2E)[206] para gerar o modo de ligação predito para a quercetina (Figura 36). A sobreposição do modo de ligação da quercetina com o modo de ligação representativo da simulação de dinâmica para a taxifolina está apresentado na Figura 36A. Quando comparados, os modos de ligação da taxifolina e da quercetina são invertidos. Por exemplo, o substituinte hidroxila ligado na posição 7 do anel A da taxifolina está em contato polar com a cadeia lateral do resíduo Glu166, enquanto o modo de ligação predito para a quercetina posiciona o anel B em contato com o resíduo Glu166 (Figura 36B). A análise da superfície da estrutura 6Y2E, na qual foi modelado o modo de

ligação da quercetina, indica que o anel A da taxifolina sofreria um choque estérico com a cadeia principal da Cys44, impedindo assim que a taxifolina assumisse o modo de ligação observado nos estudos de simulação por dinâmica molecular (Figura 36C).

Uma segunda característica entre a estrutura da quercetina e da taxifolina, que poderia explicar a diferença nas atividades inibitórias observadas, seria associada ao pH dos experimentos de inibição *in vitro*. No ensaio para verificar a atividade enzimática, Abian *et al*.[253] utilizaram pH 8, enquanto que o ensaio conduzido neste trabalho foi realizado em pH 7,4. De acordo com a base de dados *DrugBank*[266] a quercetina e a taxifolina possuem, respectivamente, valore de pK_a de 6,44 e 7,8. Portanto, no ensaio enzimático conduzido neste trabalho a atividade inibitória da taxifolina foi avaliada em um valor de pH < pK_a, enquanto que no ensaio reportado por Abian et al.[253] a atividade inibitória da quercetina foi determinada em um valor de pH > pK_a. Essas diferenças nas propriedades físico-químicas entre a quercetina e a taxifolina poderiam explicar a diferença na atividade inibitória. Isso porque a quercetina foi avaliada em um pH que favorece mais de 90% da estrutura na forma desprotonada (O^-). Por outro lado, o pH do ensaio com a taxifolina favorece mais de 50% da molécula no estado protonado (OH). A desprotonação da quercetina aconteceria primeiro no substituinte OH da posição 7 do anel A do núcleo flavonoide.[267] Essa desprotonação teria impacto significativo no modo de ligação da quercetina, que sofreria repulsão eletrostática da cadeia lateral do Glu166. Esses dados estão em acordo com o modelo predito para a quercetina o qual apresenta o anel A afastado do Glu166.

Para avaliar essa hipótese, ensaios experimentais com a taxifolina foram conduzidos em pH = 8,0 (valor de pH > pK_a), como mostra a Figura 37B. Os resultados indicaram que a mudança de pH não teve influência significativa na atividade inibitória da protease principal de SARS-CoV-2 pela taxifolina ao variar o pH para uma condição na qual a estrutura na forma desprotonada (O^-) fosse favorecida. Entretanto, a proximidade do pH = 8,0 utilizado ao pK_a da taxifolina (7,8) pode explicar esse fenômeno.

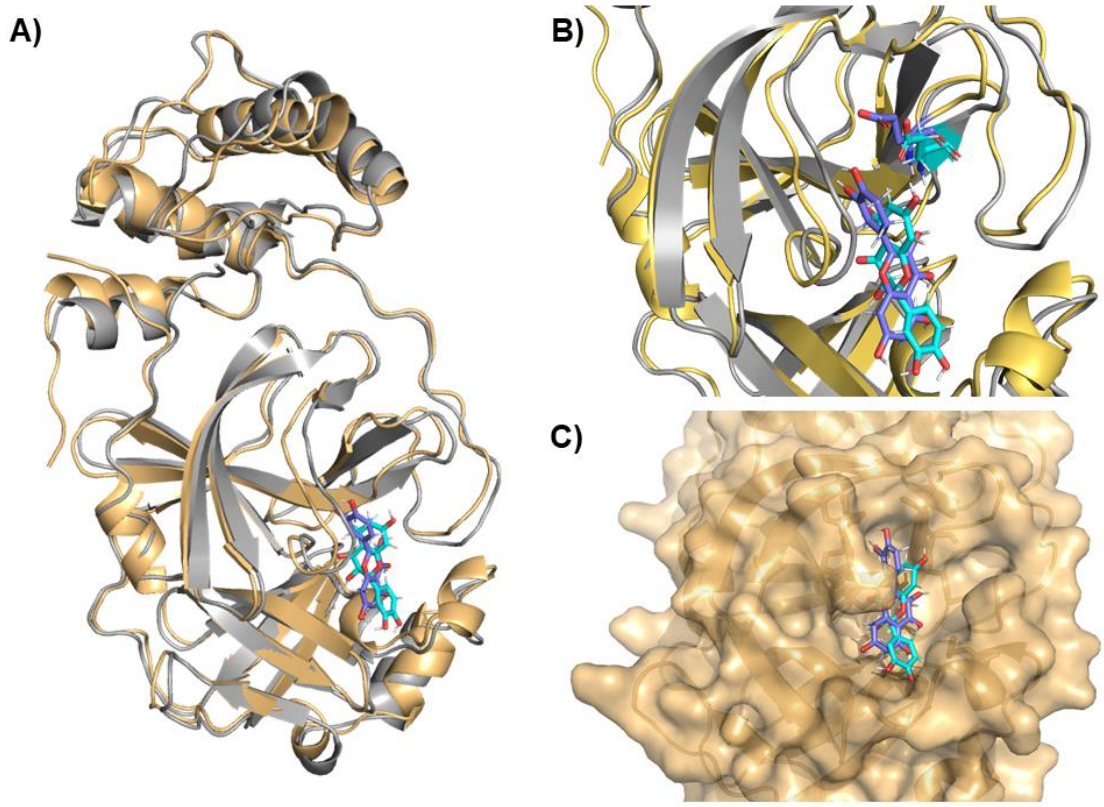


Figura 36 – Modos de ligação da quercetina e taxifolina sobrepostos. A) Em amarelo a estrutura 6Y2E, docada com a quercetina (roxo), e em cinza, a estrutura 6LU7, retirada de um *frame* representativo de *cluster* da dinâmica molecular com a taxifolina (ciano). B) Sobreposição das estruturas das proteínas e dos modos de ligação dos ligantes, com o resíduo Glu166 destacado em ciano na estrutura da 6LU7 (MD taxifolina) e em roxo na estrutura da 6Y2E (docking quercetina). C) Superfície da 6Y2E com os ligantes sobrepostos, taxifolina em ciano e quercetina em roxo.

Fonte: Elaborada pelo autor.

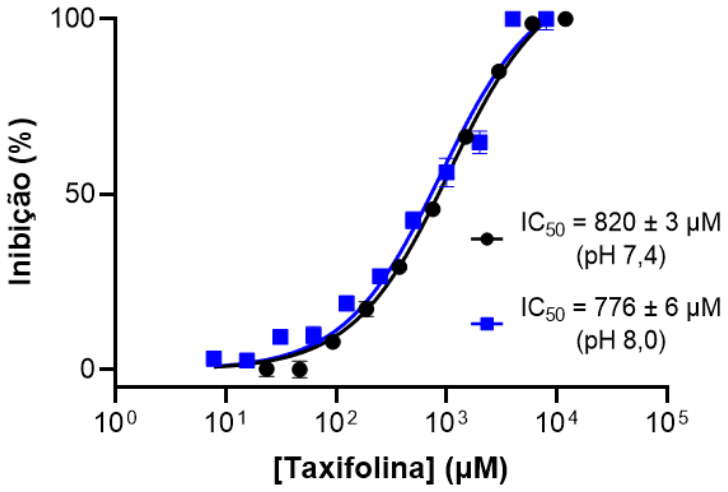


Figura 37 – Curva de inibição da taxifolina contra a protease principal de SARS-CoV-2, em diferentes condições de pH, com IC_{50} = 820 ± 3 μM em pH = 7,4 (azul), e IC_{50} = 776 ± 6 μM em pH = 8,0 (preto).

Fonte: Elaborada pelo autor.

Uma outra hipótese para a diferença observada entre os valores de atividade dos dois análogos pode estar na isomeria da taxifolina (Figura 35). Bernatova & Liskova,[265] descreveram propriedades relevantes da taxifolina nos estudos pré-clínicos para o tratamento de pacientes hipertensos acometidos com infecção viral. Neste estudo, os autores também reportaram que essa atividade era maior ao utilizar o isômero (-)-taxifolina, embora a (+)-taxifolina também tivesse atividade. Dessa forma, como o isômero utilizado aqui foi a (+)-taxifolina, a aquisição do isômero (-)-taxifolina será necessária para confirmar essa hipótese e dar prosseguimento ao estudo.

Por fim, ao realizar o estudo de triagem virtual seguida das simulações de dinâmica molecular e metadinâmica, uma suposição bastante importante foi feita: as moléculas triadas devem se comportar como um inibidor competitivo proteína. Assumir isso como verdade permitiu que o sítio catalítico da proteína fosse determinado como o alvo durante a docagem bem como o comportamento dos ligantes nesse sítio durante as simulações fosse observado. O ensaio experimental confirmou que tanto o ligante mais promissor resultante da avaliação feita aqui, a taxifolina, quanto o seu análogo reportado na literatura, a quercetina, são inibidores competitivos da M^{pro} de SARS-CoV-2.

Os resultados obtidos são um forte indício que, embora ambos os inibidores compartilhem o mesmo mecanismo de inibição competitivo com o substrato, o modo de ligação da quercetina na M^{pro} é diferente da taxifolina. Esses dados experimentais corroboram os resultados obtidos nos estudos de modelagem molecular que indicaram modos de ligações diferentes da quercetina e taxifolina no sítio catalítico da M^{pro} de SARS-CoV-2. A confirmação efetiva será feita nas próximas etapas deste trabalho, assim que for possível se obter estruturar cristalográficas que contenham os ligantes nos referidos sítios.

5.5 CONCLUSÕES

Três anos e meio se passaram desde que os primeiros casos de infecção por SARS-CoV-2, em Wuhan, foram detectados. Neste período, os números de pessoas infectadas e de vítimas fatais cresceram exponencialmente e numa velocidade alarmante. Embora a chegada das vacinas tenha melhorado substancialmente a situação da pandemia de COVID-19, ainda existe uma demanda bastante grande por

fármacos eficazes que sejam capazes de tratar a doença, principalmente considerando o potencial de mutação que o vírus possui resultando nas novas variantes que colocam em risco a atual situação de controle da doença.

A biodiversidade brasileira é uma fonte imensurável de compostos que podem ter distintas aplicações tecnológicas. Deste modo, a triagem de moléculas provenientes de produtos naturais da biodiversidade brasileira contra a M^{pro} de SARS-CoV-2 se mostrou uma estratégia atrativa para a descoberta de novos candidatos a fármacos antivirais. Os estudos integrados de triagem virtual, simulação de dinâmica molecular, metadinâmica e inibição da atividade enzimática auxiliaram a descoberta da taxifolina como um inibidor submilimolar da M^{pro}. A utilização dos ensaios experimentais corroborou a hipótese inicial de busca de inibidores competitivo, como revelou o ensaio de competitividade.

A próxima etapa deste estudo incluirá: i. determinação do modo de ligação experimental por cristalografia e difração de raios-X; ii. a investigação da atividade antiviral *in vitro* em ensaios padronizados com células infectadas com SAR-CoV-2. Além disso, serão avaliadas as atividades inibitórias dos demais compostos que se mostraram promissores nos experimentos *in sílico*.

CAPÍTULO 6: CONSIDERAÇÕES FINAIS

6 CONSIDERAÇÕES FINAIS

Embora esta tese esteja estruturada em três partes distintas, todas elas abordam um tema central: a aplicação de métodos computacionais no âmbito da química medicinal. Ao longo das seções, cada abordagem específica empregada contribuiu de modo sinérgico para uma compreensão abrangente do potencial e dos avanços que foram (e podem ser) proporcionados para a área ao integrar as ferramentas computacionais para a descoberta de novos compostos bioativos.

A utilização de métodos de vanguarda, como os modelos generativos de aprendizado profundo, usando como arquitetura os autocodificadores variacionais, permitiu a geração de conjuntos de moléculas inéditas e 100% válidas (compartilhando características comuns aos antimaláricos reportados na literatura), mas com propriedades e substituintes novos, permitindo a descoberta e proposição de novas classes de compostos como potenciais inibidores do *P. falciparum*. A alta porcentagem de validade bem como a grande diversidade das moléculas geradas deveu-se, em grande parte, à utilização de uma representação (SELFIES) e ao processamento adequados dos dados que alimentaram o modelo.

A mesma arquitetura de VAEs demonstrou ser uma ferramenta robusta para testar os modos de falhas desconhecidos para representação de dados, apresentando resultados melhores do que os obtidos pelos modelos que não são baseados em aprendizado de máquina. A utilização de dessa arquitetura permitiu encontrar os modos de falha dos SELFIES, desafiando sua premissa inicial: ser 100% válido no processo de conversão de uma cadeia SELFIES para uma cadeia SMILES válida. Os resultados mostraram a existência de um domínio de aplicabilidade organizado de forma radial no espaço latente do modelo.

Adicionalmente, pela aplicação de métodos computacionais mais tradicionais, como a triagem virtual e dinâmica molecular, foi possível encontrar uma molécula inibidora da M^{pro} de SARS-CoV-2. Essa molécula apresentou valor de IC_{50} na faixa submilimolar, demonstrando um perfil de inibidor competitivo. É válido dizer que essa molécula foi extraída da biodiversidade brasileira, destacando a contribuição que a utilização dessa fonte de recursos ainda pode proporcionar.

Em resumo, a aplicação de diversas abordagens computacionais, incluindo métodos clássicos e do estado-da-arte, permitiu a proposição de moléculas inéditas

como candidatos a inibidores do parasita causador da malária e a descoberta de um composto de origem natural como inibidor de um alvo molecular relevante no SARS-CoV-2. Compreender e aplicar tais metodologias não apenas amplia as possibilidades na química medicinal, mas também proporciona uma abordagem mais holística para enfrentar os desafios associados à busca de compostos bioativos como candidatos a fármacos para doenças infecciosas.

REFERÊNCIAS

1 ZHANG, Y. *et al*. Application of computational biology and artificial intelligence in drug design. **International Journal of Molecular Sciences**, v. 23, n. 21, p. 13568, 2022.

2 MOHS, R. C.; GREIG, N. H. Drug discovery and development: role of basic biological research. **Alzheimer's and Dementia:** translational research and clinical interventions, v. 3, n. 4, p. 651–657, 2017.

3 BERDIGALIYEV, N.; ALJOFAN, M. An overview of drug discovery and development. **Future Medicinal Chemistry**, v. 12, n. 10, p. 939–947, 2020.

4 SABE, V. T. et al. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: a review. **European Journal of Medicinal Chemistry**, v. 224, p. 113705, 2021.

5 KARTHIKEYAN, A.; PRIYAKUMAR, U. D. Artificial intelligence: machine learning for chemical sciences. **Journal of Chemical Sciences**, v. 134, n. 1, p. 2, 2022.

6 BROWN, N. *et al*. Artificial intelligence in chemistry and drug design. **Journal of Computer-Aided Molecular Design**, v. 34, n. 7, p. 709–715 2020.

7 MATER, A. C.; COOTE, M. L. Deep learning in chemistry. **Journal of Chemical Information and Modeling**, v. 59, n. 6, p. 2545–2559, 2019.

8 VAMATHEVAN, J. *et al*. Applications of machine learning in drug discovery and development. **Nature Reviews Drug Discovery**, v. 18, n. 6, p. 463–477, 2019.

9 NEVES, B. J. *et al*. Deep Learning-driven research for drug discovery: tackling malaria. **PLoS Computational Biology**, v. 16, n. 2, p. 1–21, 2020.

10 CUMMING, J. G. *et al*. Chemical predictive modelling to improve compound quality. **Nature Reviews Drug Discovery**, v. 12, n. 12, p. 948–962, 2013.

11 COVA, T. F. G. G.; PAIS, A. A. C. C. Deep learning for deep chemistry: optimizing the prediction of chemical patterns. **Frontiers in Chemistry**, v. 7, p. 1–22, 2019. DOI: 10.3389/fchem.2019.00809.

12 STÅHL, N. *et al*. Deep reinforcement learning for multiparameter optimization in de novo drug design. **Journal of Chemical Information and Modeling**, v. 59, n. 7, p. 3166–3176, 2019.

13 CHUANG, K. V.; GUNSALUS, L. M.; KEISER, M. J. Learning molecular representations for medicinal chemistry. **Journal of Medicinal Chemistry**, v. 63, n. 16, p. 8705–8722, 2020.

14 ASHDOWN, G. W. *et al*. A machine learning approach to define antimalarial drug action from heterogeneous cell-based screens. **Science Advances**, v. 6, n. 39, p. 2–10, 2020.

15 ALIPER, A. *et al*. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. **Molecular Pharmaceutics**, v. 13, n. 7, p. 2524–2530, 2016.

16 WANG, C. *et al*. Pairwise input neural network for target-ligand interaction prediction. *In:* INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE, 2014, Belfast. **Proceedings** […] Belfast: IEEE, 2014.

17 HUGHES, T. B.; MILLER, G. P.; SWAMIDASS, S. J. Modeling epoxidation of drug-like molecules with a deep machine learning network. **ACS Central Science**, v. 1, n. 4, p. 168–180, 2015.

18 SOUSA, T. *et al*. Generative deep learning for targeted compound design. **Journal of Chemical Information and Modeling**, v. 61, n. 11, p. 5343–5361, 2021.

19 KADURIN, A. *et al*. DruGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. **Molecular Pharmaceutics**, v. 14, n. 9, p. 3098–3104, 2017.

20 OLIVECRONA, M. *et al*. Molecular de-novo design through deep reinforcement learning. **Journal of Cheminformatics**, v. 9, n. 1, p. 1–14, 2017.

21 IMRIE, F. *et al*. Deep generative models for 3D linker design. **Journal of Chemical Information and Modeling**, v. 60, n. 4, p. 1983–1995, 2020.

22 GÓMEZ-BOMBARELLI, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. **ACS Central Science**, v. 4, n. 2, p. 268–276, 2018.

23 SEGLER, M. H. S. *et al*. Generating focused molecule libraries for drug discovery with recurrent neural networks. **ACS Central Science**, v. 4, n. 1, p. 120–131, 2018.

24 JIN, W. *et al*. Learning multimodal graph-to-graph translation for molecular optimization. *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 7., 2019, New Orleans. **Proceedings** [...] New Orleans: ICLR, 2019.

25 ZHOU, Z. *et al*. Optimization of molecules via deep reinforcement learning. **Scientific Reports**, v. 9, n. 1, p. 1–10, 2019.

26 KRAMER, M. A. Nonlinear principal component analysis using autoassociative neural networks. **AIChE Journal**, v. 37, n. 2, p. 233–243, 1991.

27 KINGMA, D. P.; WELLING, M. An introduction to variational autoencoders. **Foundations and Trends® in Machine Learning**, v. 12, n. 4, p. 307–392, 2019.

28 KINGMA, D. P.; WELLING, M. **Auto-encoding variational bayes**. 2013. DOI: 10.48550/arXiv.1312.6114.

29 KINGMA, D. P. *et al*. Semi-supervised learning with deep generative models. **Advances in Neural Information Processing Systems**, v. 4, p. 3581–3589, 2014.

30 KHEMAKHEM, I. *et al*. **Variational autoencoders and nonlinear ICA**: a unifying framework. 2019. DOI: 10.48550/arXiv.1907.04809.

31 PU, Y. *et al*. **Variational autoencoder for deep learning of images, labels and captions**. 2016. DOI: 10.48550/arXiv.1609.08976.

32 DOERSCH, C. **Tutorial on variational autoencoders**. 2016. DOI: 10.48550/arXiv.1606.05908.

33 KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. **Annals of Mathematical Statistics**, v. 22, n. 1, p. 79–86, 1951.

34 FU, H. *et al*. **Cyclical annealing schedule**: a simple approach to mitigating kl vanishing. 2019. DOI: 10.48550/arXiv.1903.10145.

35 SEVGEN, E. *et al*. **ProT-VAE**: protein transformer variational autoencoder for functional protein design. 2023. DOI: 10.1101/2023.01.23.525232.

36 WU, Y.; XU, L. Image generation of tomato leaf disease identification based on adversarial-VAE. **Agriculture**, v. 11, n. 10, 2021. DOI: 10.3390/agriculture11100981.

37 TEMPKE, R.; MUSHO, T. Autonomous design of new chemical reactions using a variational autoencoder. **Communications Chemistry**, v. 5, n. 1, 2022. DOI: 10.1038/s42004-022-00647-x.

38 LEE, M.; MIN, K. MGCVAE: multi-objective inverse design via molecular graph conditional variational autoencoder. **Journal of Chemical Information and Modeling**, v. 62, n. 12, p. 2943–2950, 2022.

39 JIN, W.; BARZILAY, R.; JAAKKOLA, T. **Junction tree variational autoencoder for molecular graph generation**. 2018. DOI: 10.48550/arXiv.1802.04364.

40 HADIPOUR, H. *et al*. Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means. **BMC Bioinformatics**, v. 23, n. S4, p. 132, Apr. 2022.

41 GOODFELLOW, I. J. *et al*. Generative adversarial networks. 2014. DOI: 10.48550/arXiv.1406.2661.

42 MAKHZANI, A. *et al*. **Adversarial autoencoders**. 2015. DOI: 10.48550/arXiv.1511.05644.

43 HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997.

44 BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: a review and new perspectives. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 35, n. 8, p. 1798–1828, 2013.

45 WIGH, D. S.; GOODMAN, J. M.; LAPKIN, A. A. A review of molecular representation in the age of machine learning. **WIREs Computational Molecular Science**, v. 12, n. 5, p. e1603, 2022.

46 DAVID, L. *et al*. Molecular representations in AI-driven drug discovery: a review and practical guide. **Journal of Cheminformatics**, v. 12, n. 1, p. 56, 2020.

47 DURANT, J. L. *et al*. Reoptimization of MDL keys for use in drug discovery. **Journal of Chemical Information and Computer Sciences**, v. 42, n. 6, p. 1273–1280, 2002.

48 CARHART, R. E.; SMITH, D. H.; VENKATARAGHAVAN, R. Atom pairs as molecular features in structure-activity studies: definition and applications. **Journal of Chemical Information and Computer Sciences**, v. 25, n. 2, p. 64–73, 1985.

49 ROGERS, D.; HAHN, M. Extended-connectivity fingerprints. **Journal of Chemical Information and Modeling**, v. 50, n. 5, p. 742–754, 2010.

50 SHERIDAN, R. P. *et al*. Chemical similarity using geometric atom pair descriptors. **Journal of Chemical Information and Computer Sciences**, v. 36, n. 1, p. 128–136, 1996.

51 MCGAUGHEY, G. B. *et al*. Comparison of topological, shape, and docking methods in virtual screening. **Journal of Chemical Information and Modeling**, v. 47, n. 4, p. 1504–1519, 2007.

52 MORGAN, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. **Journal of Chemical Documentation**, v. 5, n. 2, p. 107–113, 1965.

53 WEININGER, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. **Journal of Chemical Information and Modeling**, v. 28, n. 1, p. 31–36, 1988.

54 WEININGER, D.; WEININGER, A.; WEININGER, J. L. SMILES. 2. algorithm for generation of unique SMILES notation. **Journal of Chemical Information and Computer Sciences**, v. 29, n. 2, p. 97–101, 1989.

55 KRENN, M. *et al*. SELFIES and the future of molecular string representations. **Patterns**, v. 3, n. 10, p. 100588, 14 out. 2022.

56 KRENN, M. *et al*. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. **Machine Learning:** science and technology, v. 1, n. 4, p. 045024, 2020.

57 LO, A. *et al*. Recent advances in the self-referencing embedding strings (SELFIES) library. **Digital Discovery**, v. 2, n. 4, p. 897–908, 2023.

58 KRENN, M. *et al*. **SELFIES**: a robust representation of semantically constrained graphs with an example application in chemistry. 2019. DOI: 10.48550/arXiv.1905.13741.

59 HOPCROFT, J. E.; MOTWANI, RAJEEV.; ULLMAN, J. D. **Introduction to automata theory, languages, and computation**. 3rd ed. New York: Pearson, 2007.

60 SALLARES, R.; BOUWMAN, A.; ANDERUNG, C. The spread of malaria to southern Europe in antiquity: new approaches to old problems. **Medical History**, v. 48, n. 3, p. 311–328, 2004.

61 ASHLEY, E. A.; PYAE PHYO, A.; WOODROW, C. J. Malaria. **Lancet**, v. 391, n. 10130, p. 1608–1621, 2018.

62 MOURIER, T. *et al*. The genome of the zoonotic malaria parasite *Plasmodium simium* reveals adaptations to host switching. **BMC Biology**, v. 19, n. 1, p. 219, 2021.

63 BYKERSMA, A. The new zoonotic malaria: *Plasmodium cynomolgi*. **Tropical Medicine and Infectious Disease**, v. 6, n. 2, p. 46, 2021.

64 WORLD HEALTH ORGANIZATION. **World malaria report 2022**. Disponível em: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2022. Acesso em: 5 mar. 2023.

65 WORLD HEALTH ORGANIZATION. **World malaria report 2022**. Disponível em: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2022. Acesso em: 5 mar. 2023.

66 AGUIAR, A. C. C. *et al*. Discovery of marinoquinolines as potent and fast-acting *Plasmodium falciparum* inhibitors with *in vivo* activity. **Journal of Medicinal Chemistry**, v. 61, n. 13, p. 5547–5568, 2018.

67 THE MALARIA ATLAS PROJECT. **Trends in global malaria burden**. Disponível em: https://malariaatlas.org/. Acesso em: 5 set. 2023.

68 PHILLIPS, M. A. *et al*. Malaria. **Nature Reviews Disease Primers**, v. 3, n. 1, p. 17050, 2017.

69 SIQUEIRA-NETO, J. L. *et al*. Antimalarial drug discovery: progress and approaches. **Nature Reviews Drug Discovery**, v. 22, n.10, p. 807–826, 2023.

70 SIQUEIRA-NETO, J. L. *et al*. Antimalarial drug discovery: progress and approaches. **Nature Reviews Drug Discovery**, v. 22, n.10, p. 807–826, 2023.

71 WHITE, N. J. *et al*. Malaria. **Lancet**, v. 383, n. 9918, p. 723–735, 2014.

72 NEWTON, C. R. J. C.; HIEN, T. T.; WHITE, N. Neurological aspects of tropical disease: cerebral malaria. **Journal of Neurology, Neurosurgery & Psychiatry**, v. 69, n. 4, p. 433–441, 2000.

73 SIŁKA, W. *et al*. Malaria detection using advanced deep learning architecture. **Sensors**, v. 23, n. 3, p. 1501, 2023.

74 MOXON, C. A. *et al*. New insights into malaria pathogenesis. **Annual Review of Pathology:** mechanisms of disease, v. 15, p. 315–343, 2020. DOI: 10.1146/annurevpathmechdis-012419-032640.

75 ZANGHI, G.; VAUGHAN, A. M. *Plasmodium vivax* pre-erythrocytic stages and the latent hypnozoite. **Parasitology International**, v. 85, 2021. DOI: 10.1016/j.parint.2021.102447.

76 PHILLIPS, M. A. *et al*. Malaria. **Nature Reviews Disease Primers**, v. 3, n. 1, p. 17050, 2017.

77 EASTMAN, R. T.; FIDOCK, D. A. Artemisinin-based combination therapies: a vital tool in efforts to eliminate malaria. **Nature Reviews Microbiology**, v. 7, n. 12, p. 864–874, 2009.

78 WORLD HEALTH ORGANIZATION. **Guidelines for the treatment of malaria**. Disponível em: https://iris.who.int/bitstream/handle/10665/162441/9789241549127_eng.pdf. Acesso em: 5 mar. 2023.

79 MENARD, D.; DONDORP, A. Antimalarial drug resistance: a threat to malaria elimination. **Cold Spring Harbor Perspectives in Medicine**, v. 7, n. 7, p. 1–24, 2017.

80 ASHLEY, E. A. *et al*. Spread of artemisinin resistance in *Plasmodium falciparum* malaria. **New England Journal of Medicine**, v. 371, n. 5, p. 411–423, 2014.

81 HAMILTON, W. L. *et al*. Evolution and expansion of multidrug-resistant malaria in southeast Asia: a genomic epidemiology study. **Lancet Infectious Diseases**, v. 19, n. 9, p. 943–951, 2019.

82 LU, F. *et al*. emergence of indigenous artemisinin-resistant *Plasmodium falciparum* in Africa. **New England Journal of Medicine**, v. 376, n. 10, p. 991–993, 2017.

83 WINKLER, D. A. use of artificial intelligence and machine learning for discovery of drugs for neglected tropical diseases. **Frontiers in Chemistry**, v. 9, 2021. DOI: 10.3389/fchem.2021.614073.

84 ARSHADI, A. K. *et al*. Deepmalaria: artificial intelligence driven discovery of potent antiplasmodials. **Frontiers in Pharmacology**, v. 10, 2020. DOI: 10.3389/fphar.2019.01526.

85 LIMA, M. N. N. *et al*. Artificial intelligence applied to the rapid identification of new antimalarial candidates with dual-stage activity. **ChemMedChem**, v. 16, n. 7, p. 1093–1103, 2021.

86 AMIN, I. *et al*. Transfer learning-based semi-supervised generative adversarial network for malaria classification. **Computers, Materials and Continua**, v. 74, n. 3, p. 6335–6349, 2023.

87 TAN, D.; LIANG, X. Multiclass malaria parasite recognition based on transformer models and a generative adversarial network. **Scientific Reports**, v. 13, n. 1, p. 1–16, 2023.

88 GODINEZ, W. J. *et al*. Design of potent antimalarials with generative chemistry. **Nature Machine Intelligence**, v. 4, n. 2, p. 180–186, 2022.

89 MENDEZ, D. *et al*. ChEMBL: towards direct deposition of bioassay data. **Nucleic Acids Research**, v. 47, n. D1, p. D930–D940, 2019.

90 BOSC, N. *et al*. MAIP: a web service for predicting blood-stage malaria inhibitors. **Journal of Cheminformatics**, v. 13, n. 1, p. 1–14, 2021.

91 DAVIES, M. *et al*. ChEMBL web services: streamlining access to drug discovery data and utilities. **Nucleic Acids Research**, v. 43, n. W1, p. W612–W620, 2015.

92 MENDEZ, D. *et al*. ChEMBL: towards direct deposition of bioassay data. **Nucleic Acids Research**, v. 47, n. D1, p. D930–D940, 2019.

93 IRWIN, J. J.; SHOICHET, B. K. ZINC - A free database of commercially available compounds for virtual screening. **Journal of Chemical Information and Modeling**, v. 45, n. 1, p. 177–182, 2005.

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

95 IRWIN, J. J. *et al*. ZINC: a free tool to discover chemistry for biology. **Journal of Chemical Information and Modeling**, v. 52, n. 7, p. 1757–1768, 2012.

96 ABADI, M. *et al*. **TensorFlow**: large-scale machine learning on heterogeneous distributed systems. DOI: 10.48550/arXiv.1603.04467.

97 CHO, K. *et al*. **Learning phrase representations using RNN encoder-decoder for statistical machine translation**. 2014. DOI: 10.48550/arXiv.1406.1078.

98 GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge: MIT Press, 2016.

99 PANG, T. *et al*. Rethinking softmax cross-entropy loss for adversarial robustness. *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 8., 2020, Virtual. **Proceedings** [...] Virtual: ICLR, 2020.

100 GLOROT, X.; BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. *In*: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 13., 2010, Sardinia. **Proceedings** [...] Sardinia: PMLR, 2010.

101 PLAUT, D.; NOWLAN, S.; HINTON, G. **Experiments on learning by back propagation.** Pittsburgh: ERIC, 1986. Technical Report CMU-CS-86-126.

102 ORR, G. B.; MÜLLER, K.-R. **Neural networks:** tricks of the trade. 2nd ed. Berlin: Springer, 2012. (Lecture notes in computer science, v. 7700)

103 LANDRUM, G. **RDKit:** open-source cheminformatics software. Disponível em: https://www.rdkit.org/. Acesso em: 03 set. 2021.

104 VERRAS, A. *et al*. Shared consensus machine learning models for predicting blood stage malaria inhibition. **Journal of Chemical Information and Modeling**, v. 57, n. 3, p. 445–453, 2017.

105 ROCKAFELLAR, R. T.; WETS, R. J. B. **Variational analysis**. 3rd ed. Berlin: Springer, 2009.

106 RAMSUNDAR, B. *et al*. **Deep learning for the life sciences:** applying deep learning to genomics, microscopy, drug discovery, and more. Sebastopol: O'Reilly Media, 2019.

107 DUVENAUD, D. *et al*. Convolutional networks on graphs for learning molecular fingerprints. 2015. DOI: 10.48550/arXiv.1509.09292.

108 GLANTZ, S.; SLINKER, B.; NEILANDS, T. **Primer of applied regression & analysis of variance**. 3rd ed. New York: McGraw-Hill Education, 2017.

109 VAN DER MAATEN, L.; HINTON, G. Visualizing data using t-SNE. **Journal of Machine Learning Research**, v. 9, n. 86, p. 2579–2605, 2008.

110 HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 6, p. 417–441, 1933.

111 JACKSON, J. E. **A user's guide to principal components**. New York: Wiley, 1991.

112 PEDREGOSA, F. *et al*. Scikit-learn: machine learning in python. **Journal of Machine Learning Research**, v. 12, n. 85, p. 2825–2830, 2011.

113 OSIPENKO, S. *et al*. Transfer learning for small molecule retention predictions. **Journal of Chromatography A**, v. 1644, 2021. DOI: 10.1016/j.chroma.2021.462119.

114 AMABILINO, S. *et al*. Guidelines for recurrent neural network transfer learning-based molecular generation of focused libraries. **Journal of Chemical Information and Modeling**, v. 60, n. 12, p. 5699–5713, 2020.

115 LIU, Z. *et al*. Improved fine-tuning by better leveraging pre-training data. 2021. DOI: 10.48550/arXiv.2111.12292.

116 BICKERTON, G. R. *et al*. Quantifying the chemical beauty of drugs. **Nature Chemistry**, v. 4, n. 2, p. 90–98, 2012.

117 SONI, S.; ROBERTS, K. Evaluation of dataset selection for pre-training and finetuning transformer language models for clinical question answering. *In*: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 12., 2020, Marseille. **Proceedings** [...] Marseille: European Language Resources Association, 2020.

118 TEVOSYAN, A. *et al*. Improving VAE based molecular representations for compound property prediction. **Journal of Cheminformatics**, v. 14, n. 1, p. 69, 2022.

119 GALUSHKA, M. *et al*. Prediction of chemical compounds properties using a deep learning model. **Neural Computing and Applications**, v. 33, n. 20, p. 13345–13366, 2021.

120 POSTERA. **Manifold**. Disponível em: https://app.postera.ai/manifold/. Acesso em: 14 set. 2023

121 OPENEYE. Cadence molecular sciences. **BROOD 3.2.1.1**. Disponível em: http://www.eyesopen.com. Acesso em: 14 set. 2023.

122 BENET, L. Z. *et al*. BDDCS, the rule of 5 and drugability. **Advanced Drug Delivery Reviews**, v. 101, p. 89–98, 2016. DOI: 10.1016/j.addr.2016.05.007.

123 LIPINSKI, C. A. *et al*. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. **Advanced Drug Delivery Reviews**, v. 46, n. 1–3, p. 3–26, 2001.

124 DAI, H. *et al*. **Syntax-directed variational autoencoder for structured data**. 2018. DOI: 10.48550/arXiv.1802.08786.

125 LI, C. *et al*. Geometry-based molecular generation with deep constrained variational autoencoder. **IEEE Transactions on Neural Networks and Learning Systems**, p. 1–10, 2022. DOI: 10.1109/TNNLS.2022.3147790.

126 LIM, J. *et al*. Molecular generative model based on conditional variational autoencoder for de novo molecular design. **Journal of Cheminformatics**, v. 10, n. 1, p. 31–40, 2018.

127 GUIMARAES, G. L. *et al*. **Objective-reinforced generative adversarial networks (organ) for sequence generation models**. 2017. DOI: 10.48550/arXiv.1705.10843.

128 FREY, N. C. *et al*. Neural scaling of deep chemical models. **Nature Machine Intelligence**, v. 5, n. 11, p. 1297–1305, 2023.

129 NIGAM, A. *et al*. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. **Chemical Science**, v. 12, n. 20, p. 7079–7090, 2021.

130 CHENG, A. H. *et al*. Group SELFIES: a robust fragment-based molecular string representation. **Digital Discovery**, v. 2, n. 3, p. 748–758, 2023.

131 SANDER, T. *et al*. DataWarrior: an open-source program for chemistry aware data visualization and analysis. **Journal of Chemical Information and Modeling**, v. 55, n. 2, p. 460–473, 2015.

132 VALLURI, H. *et al*. Basic nitrogen (BaN) Is a key property of antimalarial chemical space. **Journal of Medicinal Chemistry**, v. 66, n. 13, p. 8382–8406, 2023.

133 BHANOT, A.; SUNDRIYAL, S. Physicochemical profiling and comparison of research antiplasmodials and advanced stage antimalarials with oral drugs. **ACS Omega**, v. 6, n. 9, p. 6424–6437, 2021.

134 BOSC, N. *et al*. MAIP: a web service for predicting blood-stage malaria inhibitors. **Journal of Cheminformatics**, v. 13, n. 1, p. 1–14, 2021.

135 WORLD INTELLECTUAL PROPERTY ORGANIZATION (Great Britain. Giorgio Bonanomi; Fabrizio Micheli. **Azabicyclo[4.1.0]heptane derivatives**. WO2010133569A1, 18 May 2010, 25 Nov. 2010. Disponível em: https://patents.google.com/patent/WO2010133569A1/en. Acesso em: 23 jan. 2022.

136 MENG, P. *et al*. Epigenetic mechanism of 5-HT/NE/DA triple reuptake inhibitor on adult depression susceptibility in early stress mice. **Frontiers in Pharmacology**, v. 13, 2022. DOI: 10.3389/fphar.2022.848251.

137 VOLLMER, K. J. *et al*. Activation of 6-alkoxy-substituted methylenecyclopropane nucleoside analogs requires enzymatic modification by adenosine deaminase-like protein 1. **Antimicrobial Agents and Chemotherapy**, v. 63, n. 10, 2019. DOI: 10.1128/AAC.01301-19.

138 ROUPHAEL, N. G. *et al*. Phase Ib trial to evaluate the safety and pharmacokinetics of multiple ascending doses of filociclovir (MBX-400, cyclopropavir) in healthy

volunteers. **Antimicrobial Agents and Chemotherapy**, v. 63, n. 9, 2019. DOI: 10.1128/AAC.00717-19

139 SARAMAGO, L. C. *et al*. AI-driven discovery of SARS-CoV-2 main protease fragment-like inhibitors with antiviral activity *in vitro*. **Journal of Chemical Information and Modeling**, v. 63, n. 9, p. 2866–2880, 2023.

140 DERAKHSHAN, A.; HARRIS, I. G.; BEHZADI, M. Detecting telephone-based social engineering attacks using scam signatures. *In*: ACM WORKSHOP ON SECURITY AND PRIVACY ANALYTICS, 2021, New York. **Proceeding** [...] New York: ACM, 2021.

141 SHAN, L. *et al*. Abnormal ECG detection based on an adversarial autoencoder. **Frontiers in Physiology**, v. 13, p. 1–14, 2022. DOI: 10.3389/fphys.2022.961724.

142 SAMARIYA, D. *et al*. Detection and explanation of anomalies in healthcare data. **Health Information Science and Systems**, v. 11, n. 1, p. 20, 2023.

143 KETEPALLI, G. *et al*. Anomaly detection in credit card transaction using deep learning techniques. *In*: INTERNATIONAL CONFERENCE ON COMMUNICATION AND ELECTRONICS SYSTEMS, 7., 2022, Coimbatore. **Proceedings** [...] Coimbatore: IEEE Press, 2022.

144 LAPTEV, N. AnoGen: deep anomaly generator. *In*: OUTLIER DETECTION DECONSTRUCTED (ODD) WORKSHOP, 5., 2018, London. **Proceedings** [...] London: ODD Press, 2018.

145 DU, H.; WANG, S.; HUO, H. XFinder: detecting unknown anomalies in distributed machine learning scenario. **Frontiers in Computer Science**, v. 3, p. 1–13, 2021. DOI: 10.3389/fcomp.2021.710384

146 SCHLEGL, T. *et al*. Scalable anomaly detection in manufacturing systems using an interpretable deep learning approach. **Procedia CIRP**, v. 104, n. 4, p. 1547–1552, 2021.

147 BAWANEH, M.; SIMON, V. Anomaly detection in smart city traffic based on time series analysis. *In*: INTERNATIONAL CONFERENCE ON SOFTWARE, TELECOMMUNICATIONS AND COMPUTER NETWORKS (SoftCOM), 2019, Split. **Proceedings** [...] Split: IEEE Press, 2019.

148 KAMOI, R.; KOBAYASHI, K. Why is the mahalanobis distance effective for anomaly detection? 2020. DOI: 10.48550/arXiv.2003.00402.

149 ALGHUSHAIRY, O. *et al*. A review of local outlier factor algorithms for outlier detection in big data streams. **Big Data and Cognitive Computing**, v. 5, n. 1, p. 1, 2020.

150 RANI, S. *et al*. Analysis of anomaly detection of Malware using KNN. *In*: INTERNATIONAL CONFERENCE ON INNOVATIVE PRACTICES IN TECHNOLOGY AND MANAGEMENT, 2., 2022, Gautam Buddha Nagar. **Proceedings** [...] Gautam Buddha Nagar: IEEE Press, 2022.

151 HOSSEINZADEH, M. *et al*. Improving security using SVM-based anomaly detection: issues and challenges. **Soft Computing**, v. 25, n. 4, p. 3195–3223, 2021.

152 COZZATTI, M.; SIMONETTA, F.; NTALAMPIRAS, S. **Variational autoencoders for anomaly detection in respiratory sounds**. 2022. DOI: 10.48550/arXiv.2208.03326.

153 PRIFTI, E. *et al*. Variational convolutional autoencoders for anomaly detection in scanning transmission electron microscopy. **Small**, v. 19, n. 16, p. 2205977, 2023.

154 IQBAL, T.; QURESHI, S. Reconstruction probability-based anomaly detection using variational auto-encoders. **International Journal of Computers and Applications**, v. 45, n. 3, p. 231–237, 2023.

155 RAMAKRISHNA, S. *et al*. Efficient out-of-distribution detection using latent space of β-VAE for cyber-physical systems. **ACM Transactions on Cyber-Physical Systems**, v. 6, n. 2, p. 1–34, 2022.

156 MICHAEL-PITSCHAZE, T. *et al*. **Detecting anomalous proteins using deep representations**. 2023. DOI: 10.1101/2023.04.03.535457.

157 CZIBULA, G.; CODRE, C.; TELETIN, M. AnomalP: an approach for detecting anomalous protein conformations using deep autoencoders. **Expert Systems with Applications**, v. 166, p. 114070, 2021. DOI: 10.1016/j.eswa.2020.114070.

158 TIWARI, A.; BANSODE, V.; RATHORE, A. S. Application of advanced machine learning algorithms for anomaly detection and quantitative prediction in protein A chromatography. **Journal of Chromatography A**, v. 1682, p. 463486, 2022. DOI: 10.1016/j.chroma.2022.463486.

159 UZOLAS, L. *et al*. Deep anomaly generation: an image translation approach of synthesizing abnormal banded chromosome images. **IEEE Access**, v. 10, p. 59090–59098, 2022. DOI: 10.1109/ACCESS.2022.3178786.

160 KINGMA, D. P.; BA, J. **Adam**: a method for stochastic optimization. 2014. DOI: 10.48550/arXiv.1412.6980.

161 HÖDL, S. *et al*. **Explainability techniques for chemical language models**. 2023. DOI: 10.48550/arXiv.2305.16192.

162 ZENG, Z. *et al*. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. **Nature Communications**, v. 13, n. 1, p. 862, 2022.

163 UCAK, U. V.; ASHYRMAMATOV, I.; LEE, J. Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization. **Journal of Cheminformatics**, v. 15, n. 1, p. 55, 2023.

164 BILODEAU, C. *et al*. Generative models for molecular discovery: recent advances and challenges. **WIREs Computational Molecular Science**, v. 12, n. 5, 2022.

165 RAGHUNATHAN, S.; PRIYAKUMAR, U. D. Molecular representations for machine learning applications in chemistry. **International Journal of Quantum Chemistry**, v. 122, n. 7, 2022. DOI: 10.1002/qua.26870.

166 KUSNER, M. J.; PAIGE, B.; HERNÁNDEZ-LOBATO, J. M. **Grammar variational autoencoder**. 2017. DOI: 10.48550/arXiv.1703.01925.

167 RUDDIGKEIT, L. *et al*. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. **Journal of Chemical Information and Modeling**, v. 52, n. 11, p. 2864–2875, 2012.

168 RICHARDS, R. J.; GROENER, A. M. **Conditional β-VAE for de novo molecular generation**. 2022. DOI: 10.48550/arXiv.2205.01592.

169 ZHU, N. *et al*. A Novel Coronavirus from Patients with Pneumonia in China, 2019. **New England Journal of Medicine**, v. 382, n. 8, p. 727–733, 2020.

170 LAI, C. *et al*. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): the epidemic and the challenges. **International Journal of Antimicrobial Agents**, v. 55, n. 3, p. 105924, 2020.

171 WORLD HEALTH ORGANIZATION. **WHO COVID-19 dashboard.** Disponível em: https://covid19.who.int/. Acesso em: 9 nov. 2023.

172 CUI, J.; LI, F.; SHI, Z.-L. Origin and evolution of pathogenic coronaviruses. **Nature Reviews Microbiology**, v. 17, n. 3, p. 181–192, 2019.

173 THE LANCET MICROBE. Searching for SARS-CoV-2 origins: confidence versus evidence. **Lancet Microbe**, v. 4, n. 4, p. e200, 2023.

174 MARKOV, P. V. *et al*. The evolution of SARS-CoV-2. **Nature Reviews Microbiology**, v. 21, n. 6, p. 361–379, 5 jun. 2023.

175 VOLZ, E. *et al*. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. **Cell**, v. 184, n. 1, p. 64- 75.e11, 2021.

176 CASCELLA, M. *et al*. **Features, evaluation, and treatment of coronavirus (COVID-19)** [Updated 2023 Aug 18]. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK554776/. Acesso em: 12 out. 2023.

177 WORLD HEALTH ORGANIZATION. **Coronavirus disease (COVID-19)**. Disponível em: https://www.who.int/news-room/fact-sheets/detail/coronavirus-disease-(covid-19). Acesso em: 11 out. 2023.

178 SOHEILI, M. *et al*. The efficacy and effectiveness of COVID-19 vaccines around the world: a mini-review and meta-analysis. **Annals of Clinical Microbiology and Antimicrobials**, v. 22, n. 1, p. 42, 2023.

179 NG, T. I. *et al*. Antiviral drug discovery for the treatment of COVID-19 infections. **Viruses**, v. 14, n. 5, p. 961, 2022.

180 TREGONING, J. S. *et al*. Progress of the COVID-19 vaccine effort: viruses, vaccines and variants versus efficacy, effectiveness and escape. **Nature Reviews Immunology**, v. 21, n. 10, p. 626–636, 2021.

181 BRASIL. Ministério da Saúde. Agência Nacional de Vigilância Sanitária. **Vacinas**: Covid-19. Disponível em: https://www.gov.br/anvisa/pt-br/assuntos/paf/coronavirus/vacinas. Acesso em: 22 out. 2023.

182 BRASIL. Ministério da Saúde. Agência Nacional de Vigilância Sanitária. **Medicamentos aprovados para tratamento da Covid-19**. Disponível em:

https://www.gov.br/anvisa/pt-br/assuntos/paf/coronavirus/medicamentos. Acesso em: 22 out. 2023.

183 GOTTLIEB, R. L. *et al*. Early remdesivir to prevent progression to severe covid-19 in outpatients. **New England Journal of Medicine**, v. 386, n. 4, p. 305–315, 2022.

184 SHEAHAN, T. P. *et al*. An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 in human airway epithelial cell cultures and multiple coronaviruses in mice. **Science Translational Medicine**, v. 12, n. 541, p. 5883, 2020.

185 MURAKAMI, N. *et al*. Therapeutic advances in COVID-19. **Nature Reviews Nephrology**, v. 19, n. 1, p. 38–52, 2023.

186 JAYK BERNAL, A. *et al*. Molnupiravir for oral treatment of covid-19 in nonhospitalized patients. **New England Journal of Medicine**, v. 386, n. 6, p. 509–520, 2022.

187 OWEN, D. R. *et al*. An oral SARS-CoV-2 M pro inhibitor clinical candidate for the treatment of COVID-19. **Science**, v. 374, n. 6575, p. 1586–1593, 2021.

188 MARKHAM, A. Baricitinib: first global approval. **Drugs**, v. 77, n. 6, p. 697–704, 2017.

189 SELVARAJ, V. *et al*. Baricitinib in hospitalised patients with COVID-19: a meta-analysis of randomised controlled trials. **eClinicalMedicine**, v. 49, p. 101489, 2022.

190 MANOHARAN, S.; YING, L. Y. Baricitinib for the management of SARS-CoV-2-infected patients: a systematic review and meta-analysis of randomised controlled trials. **Canadian Journal of Infectious Diseases and Medical Microbiology**, v. 2022, p. 1–6, 2022.

191 HEO, Y.-A. Sotrovimab: first approval. **Drugs**, v. 82, n. 4, p. 477–484, 2022.

192 ABANI, O. *et al*. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. **Lancet**, v. 397, n. 10285, p. 1637–1645, 2021.

193 ROSAS, I. O. *et al*. Tocilizumab in patients hospitalised with COVID-19 pneumonia: efficacy, safety, viral clearance, and antibody response from a randomised controlled trial (COVACTA). **eClinicalMedicine**, v. 47, p. 101409, 2022. DOI: 10.1016/j.eclinm.2022.101409.

194 BLOOM, D. E. *et al*. How new models of vaccine development for covid-19 have helped address an epic public health crisis. **Health Affairs**, v. 40, n. 3, p. 410–418, 2021.

195 KESSELHEIM, A. S. *et al*. An overview of vaccine development, approval, and regulation, with implications for COVID-19. **Health Affairs**, v. 40, n. 1, p. 25–32, 2021.

196 BAI, C.; ZHONG, Q.; GAO, G. F. Overview of SARS-CoV-2 genome-encoded proteins. **Science China Life Sciences**, v. 65, n. 2, p. 280–294, 2022.

197 WALLS, A. C. *et al*. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. **Cell**, v. 181, n. 2, p. 281- 292.e6, 2020.

198 ASTUTI, I.; YSRAFIL. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): an overview of viral structure and host response. **Diabetes and Metabolic Syndrome: Clinical Research and Reviews**, v. 14, n. 4, p. 407–412, 2020.

199 YAN, W. *et al*. Structural biology of SARS-CoV-2: open the door for novel therapies. **Signal Transduction and Targeted Therapy**, v. 7, n. 26, 2022.

200 DÖMLING, A.; GAO, L. Chemistry and biology of SARS-CoV-2. **Chem**, v. 6, n. 6, p. 1283–1295, 2020.

201 DA SILVA, S. J. R. *et al*. Role of nonstructural proteins in the pathogenesis of SARS-CoV-2. **Journal of Medical Virology**, v. 92, n. 9, p. 1427–1429, 2020.

202 RABAAN, A. A. *et al*. SARS-CoV-2, SARS-CoV, and MERS-CoV: a comparative overview. **Infezioni in Medicina**, v. 28, n. 2, p. 174–184, 2020.

203 PILLAIYAR, T. *et al*. An overview of severe acute respiratory syndrome-coronavirus (SARS-CoV) 3CL protease inhibitors: peptidomimetics and small molecule chemotherapy. **Journal of Medicinal Chemistry**, v. 59, n. 14, p. 6595–6628, 2016.

204 XUE, X. *et al*. Structures of two coronavirus main proteases: implications for substrate binding and antiviral drug design. **Journal of Virology**, v. 82, n. 5, p. 2515–2527, 2008.

205 JIN, Z. *et al*. Structure of Mpro from COVID-19 virus and discovery of its inhibitors. **Nature**, v. 582, n. 7811, p. 289–293, 2020.

206 JIN, Z. *et al*. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. **Nature**, v. 582, n. 7811, p. 289–293, 2020.

207 ZHANG, L. *et al*. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved a-ketoamide inhibitors. **Science**, v. 368, n. 6489, p. 409–412, 2020.

208 ZHANG, L. *et al*. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved a-ketoamide inhibitors. **Science**, v. 368, n. 6489, p. 409–412, abr. 2020.

209 DAI, W. *et al*. Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. **Science**, v. 368, n. 6497, p. 1331–1335, 2020.

210 HALL, D. C.; JI, H.-F. A search for medications to treat COVID-19 via in silico molecular docking models of the SARS-CoV-2 spike glycoprotein and 3CL protease. **Travel Medicine and Infectious Disease**, v. 35, p. 101646, 2020.

211 MIRZA, M. U.; FROEYEN, M. Structural elucidation of SARS-CoV-2 vital proteins: computational methods reveal potential drug candidates against main protease, Nsp12 polymerase and Nsp13 helicase. **Journal of Pharmaceutical Analysis**, v. 10, n. 4, p. 320–328, 2020.

212 YU, R. *et al*. Computational screening of antagonists against the SARS-CoV-2 (COVID-19) coronavirus by molecular docking. **International Journal of Antimicrobial Agents**, v. 56, n. 2, p. 106012, 2020.

213 PEELE, K. A. *et al*. Molecular docking and dynamic simulations for antiviral compounds against SARS-CoV-2: a computational study. **Informatics in Medicine Unlocked**, v. 19, p. 100345, 2020.

214 ANDRICOPULO, A.; GUIDO, R.; OLIVA, G. Virtual screening and its integration with modern drug design technologies. **Current Medicinal Chemistry**, v. 15, n. 1, p. 37–46, 2008.

215 BANEGAS-LUNA, A. J.; CERÓN-CARRASCO, J. P.; PÉREZ-SÁNCHEZ, H. A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data. **Future Medicinal Chemistry**, v. 10, n. 22, p. 2641–2658, 2018.

216 ATANASOV, A. G. *et al*. Discovery and resupply of pharmacologically active plant-derived natural products: a review. **Biotechnology Advances**, v. 33, n. 8, p. 1582–1614, 2015.

217 ATANASOV, A. G. *et al*. Natural products in drug discovery: advances and opportunities. **Nature Reviews Drug Discovery**, v. 20, n. 3, p. 200–216, 2021.

218 HARVEY, A. L.; EDRADA-EBEL, R.; QUINN, R. J. The re-emergence of natural products for drug discovery in the genomics era. **Nature Reviews Drug Discovery**, v. 14, n. 2, p. 111–129, 2015.

219 FEHER, M.; SCHMIDT, J. M. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. **Journal of Chemical Information and Computer Sciences**, v. 43, n. 1, p. 218–227, 2003.

220 PILON, A. C. *et al*. NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. **Scientific Reports**, v. 7, n. 1, p. 7215, 2017.

221 VALLI, M. *et al*. Development of a natural products database from the biodiversity of Brazil. **Journal of Natural Products**, v. 76, n. 3, p. 439–444, 2013.

222 BERMAN, H. M. The protein data bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235–242, 2000.

223 LABBÉ, C. M. *et al*. MTiOpenScreen: a web server for structure-based virtual screening. **Nucleic Acids Research**, v. 43, n. W1, p. W448–W454, 2015.

224 MORRIS, G. M. *et al*. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. **Journal of Computational Chemistry**, v. 30, n. 16, p. 2785–2791, 2009.

225 CASE, D.A. *et al*. **AMBER 2019**. University of California, San Francisco, 2019. Disponível em: https://ambermd.org/doc12/Amber19.pdf. Acesso em: 18 mar. 2020.

226 WORD, J. M. *et al*. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation 1 1Edited by J. Thornton. **Journal of Molecular Biology**, v. 285, n. 4, p. 1735–1747, 1999.

227 WANG, J. *et al*. Development and testing of a general amber force field. **Journal of Computational Chemistry**, v. 25, n. 9, p. 1157–1174, 2004.

228 FIORENTINI, R.; TARENZI, T.; POTESTIO, R. Fast, accurate, and system-specific variable-resolution modeling of proteins. **Journal of Chemical Information and Modeling**, v. 63, n. 4, p. 1260–1275, 2023.

229 MAIER, J. A. *et al*. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. **Journal of Chemical Theory and Computation**, v. 11, n. 8, p. 3696–3713, 2015.

230 PRICE, D. J.; BROOKS, C. L. A modified TIP3P water potential for simulation with Ewald summation. **Journal of Chemical Physics**, v. 121, n. 20, p. 10096–10103, 2004.

231 LAIO, A.; PARRINELLO, M. Escaping free-energy minima. **Proceedings of the National Academy of Sciences of the United States of America**, v. 99, n. 20, p. 12562–12566, 2002.

232 LAIO, A.; GERVASIO, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. **Reports on Progress in Physics**, v. 71, n. 12, p. 126601, 2008.

233 TRIBELLO, G. A. *et al*. PLUMED 2: new feathers for an old bird. **Computer Physics Communications**, v. 185, n. 2, p. 604–613, 2014.

234 BONOMI, M. *et al*. PLUMED: a portable plugin for free-energy calculations with molecular dynamics. **Computer Physics Communications**, v. 180, n. 10, p. 1961–1972, 2009.

235 HESS, B. *et al*. GRGMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. **Journal of Chemical Theory and Computation**, v. 4, n. 3, p. 435–447, 2008.

236 NOSKE, G. D. *et al*. A crystallographic snapshot of SARS-CoV-2 main protease maturation process: SARS-CoV-2 Mpro maturation. **Journal of Molecular Biology**, v. 433, n. 18, p. 167118, 2021.

237 TROTT, O.; OLSON, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. **Journal of Computational Chemistry**, v. 32, p. 455–461, 2010.

238 SILVA, G. A. B. *et al*. Methoxyflavones from baccharis patens. **Journal of Natural Products**, v. 48, n. 5, p. 861, 1985.

239 REGASINI, L. O. *et al*. Constituintes químicos das flores de *Pterogyne nitens* (*Caesalpinioideae*). **Quimica Nova**, v. 31, n. 4, p. 802–806, 2008.

240 ARRUDA, A. C. *et al*. Further pyrano flavones from *Neoraputia alba*. **Journal of the Brazilian Chemical Society**, v. 4, n. 2, p. 80–83, 1993.

241 FILHO, R. B.; DE MORAES, M. P. L.; GOTTLIEB, O. R. Pterocarpans from Swartzia laevicarpa. **Phytochemistry**, v. 19, n. 9, p. 2003–2006, 1980.

242 BRAZ FILHO, R.; GOTTLIEB, O. R. The flavones of *Apuleia leiocarpa*. **Phytochemistry**, v. 10, n. 10, p. 2433–2450, 1971.

243 ANDRIOLI, W. J. *et al*. Mycoleptones A-C and polyketides from the endophyte mycoleptodiscus indicus. **Journal of Natural Products**, v. 77, n. 1, p. 70–78, 2014.

244 DE SOUSA, L. R. F. *et al*. Isolation of arginase inhibitors from the bioactivity-guided fractionation of *Byrsonima coccolobifolia* leaves and stems. **Journal of Natural Products**, v. 77, n. 2, p. 392–396, 2014.

245 LEITE, A. C. *et al*. Trypanocidal activity of limonoids and triterpenes from *Cedrela fissilis*. **Planta Medica**, v. 74, n. 15, p. 1795–1799, 2008.

246 CARDOSO, C. L. *et al*. New biflavonoid and other flavonoids from the leaves of *Chimarrhis turbinata* and their antioxidant activities. **Journal of the Brazilian Chemical Society**, v. 16, n. 6 B, p. 1353–1359, 2005.

247 CASTRO-GAMBOA, I. *et al*. HPLC-EICD: an useful tool for the pursuit of novel analytical strategies for the detection of antioxidant secondary metabolites. **Journal of the Brazilian Chemical Society**, v. 14, n. 5, p. 771–776, 2003.

248 CINTRA, P. *et al*. Toxicity of *Dimorphandra mollis* to workers of Apis mellifera. **Journal of the Brazilian Chemical Society**, v. 13, n. 1, p. 115–118, 2002.

249 DE PAULA, J. R. *et al*. Sesquiterpenes, triterpenoids, limonoids and flavonoids of *Cedrela odorata* graft and speculations on the induced resistance against *Hypsipyla grandella*. **Phytochemistry**, v. 44, n. 8, p. 1449–1454, 1997.

250 VIEIRA, P. C. *et al*. The chemosystematics of Dictyoloma. **Biochemical Systematics and Ecology**, v. 16, n. 6, p. 541–544, 1988.

251 FORIM, M. R. *et al*. Chemical characterization of *Azadirachta indica* grafted on Melia azedarach and analyses of azadirachtin by HPLC-MS-MS (SRM) and meliatoxins by MALDI-MS. **Phytochemical Analysis**, v. 21, n. 4, p. 363–373, 2010.

252 DA PAZ LIMA, M. *et al*. Alkaloids from *Spathelia excelsa*: their chemosystematic significance. **Phytochemistry**, v. 66, n. 13, p. 1560–1566, 2005.

253 SUÁREZ, D.; DÍAZ, N. SARS-CoV-2 Main protease: a molecular dynamics study. **Journal of Chemical Information and Modeling**, v. 60, n. 12, p. 5815–5831, 2020.

254 ABIAN, O. *et al*. Structural stability of SARS-CoV-2 3CLpro and identification of quercetin as an inhibitor by experimental screening. **International Journal of Biological Macromolecules**, v. 164, p. 1693–1703, 2020.

255 LEITE, A. C. *et al*. Trypanocidal activity of flavonoids and limonoids isolated from *myrsinaceae* and *meliaceae* active plant extracts. **Brazilian Journal of Pharmacognosy**, v. 20, n. 1, p. 1–6, 2010.

256 FREITAS, M. F.; KINOSHITA, L. S. *Myrsine (Myrsinoideae- Primulaceae*) no sudeste e sul do Brasil. **Rodriguesia**, v. 66, n. 1, p. 167–189, 2015.

257 SUNIL, C.; XU, B. An insight into the health-promoting effects of taxifolin (dihydroquercetin). **Phytochemistry,** v. 166, p. 112066, 2019. DOI: 10.1016/j.phytochem.2019.112066.

258 CAI, C. *et al*. Effects of taxifolin on osteoclastogenesis *in vitro* and *in vivo*. **Frontiers in Pharmacology**, v. 9, p. 1286, 2018. DOI: 10.3389/fphar.2018.01286.

259 WEI, Y. *et al*. Determination of taxifolin in *Polygonum orientale* and study on its antioxidant activity. **Journal of Food Composition and Analysis**, v. 22, n. 2, p. 154–157, 2009.

260 XIE, X. *et al*. Taxifolin protects RPE cells against oxidative stress-induced apoptosis. **Molecular Vision**, v. 23, p. 520–528, 2017.

261 MAKENA, P. S. *et al*. Comparative mutagenic effects of structurally similar flavonoids quercetin and taxifolin on tester strains *Salmonella typhimurium* TA102 and *Escherichia coli* WP-2 *uvrA*. **Environmental and Molecular Mutagenesis**, v. 50, n. 6, p. 451–459, 2009.

262 RAZAK, S. *et al*. Taxifolin, a natural flavonoid interacts with cell cycle regulators causes cell cycle arrest and causes tumor regression by activating Wnt/ β-catenin signaling pathway. **BMC Cancer**, v. 18, n. 1, p. 1043, 2018.

263 CHEN, X. *et al*. Plant flavonoid taxifolin inhibits the growth, migration and invasion of human osteosarcoma cells. **Molecular Medicine Reports**, v. 17, n. 2, p. 3239–3245, 2018.

264 LEE, S. B. *et al*. The chemopreventive effect of taxifolin is exerted through ARE-dependent gene regulation. **Biological and Pharmaceutical Bulletin**, v. 30, n. 6, p. 1074–1079, 2007.

265 BADSHAH, S. L. *et al*. Antiviral activities of flavonoids. **Biomedicine and Pharmacotherapy**, v. 140, n. March, p. 111596, 2021.

266 BERNATOVA, I.; LISKOVA, S. Mechanisms modified by (−)-epicatechin and taxifolin relevant for the treatment of hypertension and viral infection: knowledge from preclinical studies. **Antioxidants**, v. 10, n. 3, p. 1–26, 2021.

267 WISHART, D. S. *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018. **Nucleic Acids Research**, v. 46, n. D1, p. D1074–D1082, 2018.

268 MUSIALIK, M. *et al*. Acidity of hydroxyl groups: an overlooked influence on antiradical properties of flavonoids. **Journal of Organic Chemistry**, v. 74, n. 7, p. 2699–2709, 2009.

ANEXOS

A1 Banco de antimaláricos

As características das diferentes tabelas que compõem o diagrama entidade-relacionamento estão descritas abaixo, juntamente com os tipos de relacionamento entre as tabelas do banco de dados.

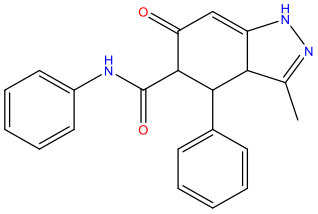
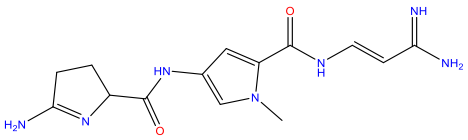
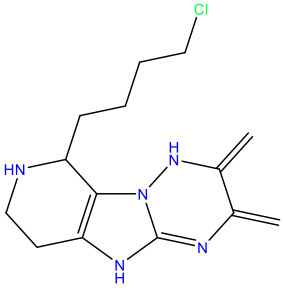
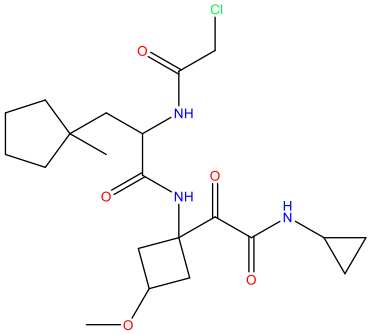
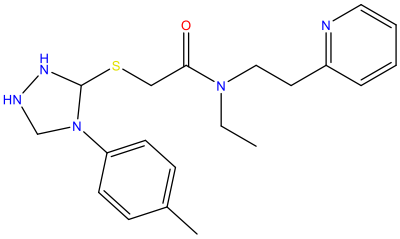
Tabela A 1 – Descrição das tabelas que compõem o banco de dados gerados a partir dos registros do ChEMBL.

Tabela	Conteúdo	# de tabelas relacionadas	Tipo de relacionamento
activities	Valores de atividade resultantes de ensaios reportados em documentos científicos.	4	1-para-muitos, não-identificado.
assays	Ensaios biológicos reportados em documentos científicos.	2	1-para-muitos, não-identificado.
compound_properties	Propriedades físico-químicas de cada composto, calculadas com a biblioteca RDKit.[103]	1	1-para-muitos, identificado.
compound_structures	Representações estruturais dos compostos, em formatos distintos.	1	1-para-muitos, identificado.
compound_records	Lista os compostos obtidos a partir de um documento científico.	3	1-para-muitos, não-identificado.
docs	Informações acerca dos documentos científicos, sejam eles artigos de periódicos científicos ou provenientes de patentes, dos quais os ensaios foram extraídos	3	1-para-muitos, não-identificado.
molecule_dictionary	Lista de compostos com identificadores associados (responsável por conectar as estruturas dos compostos e suas propriedades às outras tabelas).	4	1-para-muitos, não-identificado e identificado.

Fonte: Elaborada pelo autor.

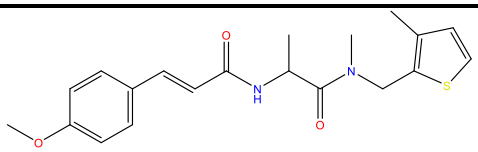
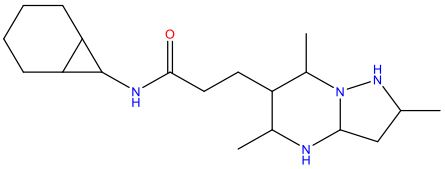
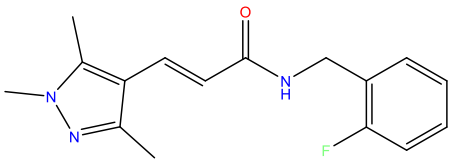
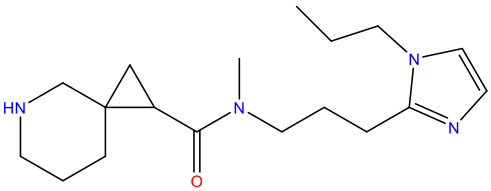
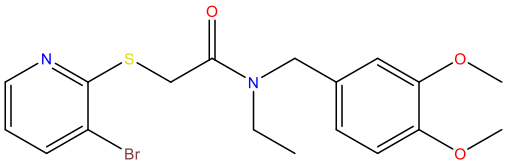
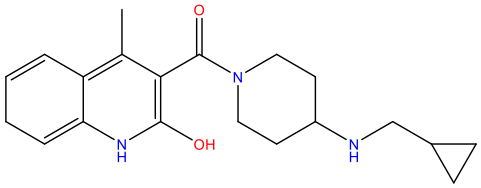
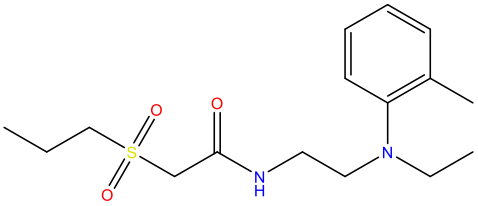
A2 Compostos similares

Tabela A 2 – Código das moléculas selecionadas pela inspeção visual, com a quantidade de nitrogênios básicos (pKa acima de 7), maior porcentagem de similaridade em termos do coeficiente de Tanimoto na busca no SciFinder^n, o número de referência das moléculas na base CAS e a molécula mais similar ao composto gerado de acordo com o SciFinder^n.

Código	Basic N	Similaridade	CAS RF	Molécula
LaBEFar_VAE_01	2	71	1303558-18-0	
LaBEFar_VAE_02	3	66	37913-78-3	
LaBEFar_VAE_03	3	62	2574124-46-0	
LaBEFar_VAE_04	2	64	1312022-33-5	
LaBEFar_VAE_05	2	70	2726884-65-5	

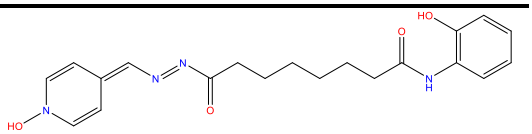
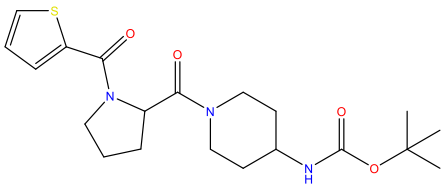
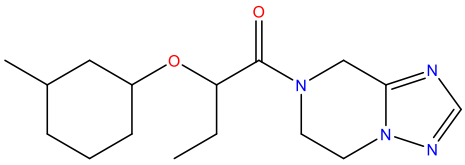
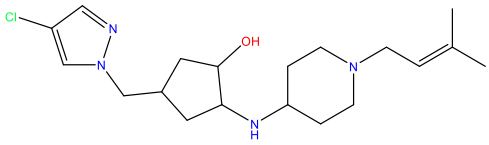
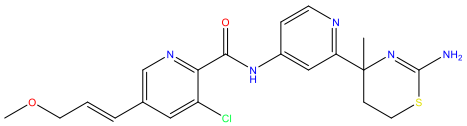
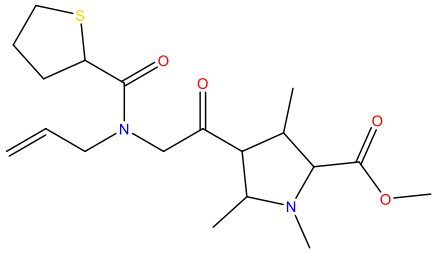
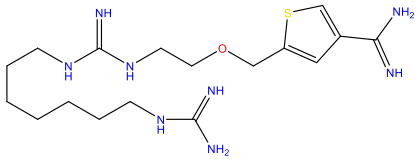
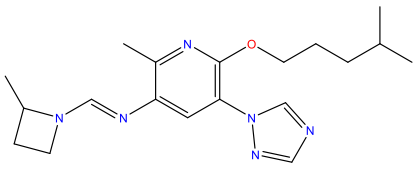
Fonte: Elaborada pelo autor.

Tabela A 3 – Código das moléculas selecionadas pela inspeção visual, com a quantidade de nitrogênios básicos (pKa acima de 7), maior porcentagem de similaridade em termos do coeficiente de Tanimoto na busca no SciFinder^{n}, o número de referência das moléculas na base CAS e a molécula mais similar ao composto gerado de acordo com o SciFinder^{n}.

Código	Basic N	Similaridade	CAS RF	Molécula
LaBEFar_VAE_06	1	83	1287207-83-3	
LaBEFar_VAE_07	2	68	2705396-29-6	
LaBEFar_VAE_08	0	74	1173352-00-5	
LaBEFar_VAE_9	4	64	2428418-24-8	
LaBEFar_VAE_10	0	81	1389709-19-6	
LaBEFar_VAE_11	1	72	2702246-86-2	
LaBEFar_VAE_12	0	80	1390436-40-4	

Fonte: Elaborada pelo autor.

Tabela A 4 – Código das moléculas selecionadas pela inspeção visual, com a quantidade de nitrogênios básicos (pKa acima de 7), maior porcentagem de similaridade em termos do coeficiente de Tanimoto na busca no SciFinder^{n}, o número de referência das moléculas na base CAS e a molécula mais similar ao composto gerado de acordo com o SciFinder^{n}.

Código	Basic N	Similaridade	CAS RF	Molécula
LaBEFar_VAE_13	2	61	1222807-13-7	
LaBEFar_VAE_14	1	80	2870445-79-5	
LaBEFar_VAE_15	1	64	1954171-12-0	
LaBEFar_VAE_16	3	66	2417910-43-9	
LaBEFar_VAE_17	0	66	1404487-09-7	
LaBEFar_VAE_18	2	64	2716639-44-8	
LaBEFar_VAE_19	3	62	2054940-16-6	
LaBEFar_VAE_20	2	66	1057623-29-6	

Fonte: Elaborada pelo autor.

A3 Trabalhos publicados e em andamento

Nosso grupo é relativamente grande e bastante diverso, de modo que tive a oportunidade de contribuir em alguns trabalhos colaborativos do grupo ao longo desses primeiros semestres de doutoramento. Essas colaborações geraram dois trabalhos que foram recentemente publicados em periódicos com seletiva política editorial:

1. Garcia, M. L. *et al*. QSAR studies on benzothiophene derivatives as Plasmodium falciparum N-myristoyltransferase inhibitors: Molecular insights into affinity and selectivity. **Drug Development Research**, v. 1, p. 1-21, 2020.

Received: 20 September 2019 | Revised: 16 December 2019 | Accepted: 20 January 2020
DOI: 10.1002/ddr.21646

RESEARCH ARTICLE

DDR WILEY

QSAR studies on benzothiophene derivatives as *Plasmodium falciparum* N-myristoyltransferase inhibitors: Molecular insights into affinity and selectivity

Mariana L. Garcia | Andrew A. de Oliveira | Renata V. Bueno | Victor H. R. Nogueira | Guilherme E. de Souza | Rafael V. C. Guido

2. Freire, M.C.L.C. *et al*. Non-Toxic Dimeric Peptides Derived from the Bothropstoxin-I Are Potent SARS-CoV-2 and Papain-like Protease Inhibitors. **Molecules,** v. 26, n. 16, p. 1-17, 2021.

molecules

MDPI

Article

Non-Toxic Dimeric Peptides Derived from the Bothropstoxin-I Are Potent SARS-CoV-2 and Papain-Like Protease Inhibitors

Marjorie C. L. C. Freire [1], Gabriela D. Noske [1], Natália V. Bitencourt [2], Paulo R. S. Sanches [2], Norival A. Santos-Filho [2], Victor O. Gawriljuk [1], Eduardo P. de Souza [3], Victor H. R. Nogueira [1], Mariana O. de Godoy [1], Aline M. Nakamura [1], Rafaela S. Fernandes [1], Andre S. Godoy [1], Maria A. Juliano [4], Bianca M. Peres [5], Cecília G. Barbosa [5], Carolina B. Moraes [6], Lucio H. G. Freitas-Junior [5], Eduardo M. Cilli [2], Rafael V. C. Guido [1,*] and Glaucius Oliva [1,*]

Manuscritos em preparação

Os capítulos dessa tese estão sendo preparados em três diferentes manuscritos, cujos títulos provisórios estão listados a seguir:

- Nogueira, V.H.R.; De Godoy, M.O.; Freire, M.C.L.C.; Souza, G. E.; Fassio, A. V.; Ferraz, M. V. F.; Oliva, G.; Lins, R. D.; Guido, R. V. C. An integrated computational approach for the discovery of SARS-CoV-2 Main Protease inhibitors from the Brazilian Biodiversity.

- Nogueira, V.H.R.; Sharma, R.; Guido, R.V.C.; Keiser, M.J. Deep Variational Anomaly Generation: An Approach to Testing Molecular Representation Robustness.

- Nogueira, V.H.R.; Rossi, G. M.; Moura, I. M. R; Maluf, S. C.; Fassio, A. V.; Sharma, R.; Keiser, M.J.; Guido, R.V.C. Beyond Model: Experimental Validation of AI-Designed Antimalarial Molecules.