CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
CAMPUS DIVINÓPOLIS

Alex Raimundo de Oliveira

UTILIZAÇÃO DE MODELOS BASEADOS EM GRAFOS PARA A DESCOBERTA E REPOSICIONAMENTO DE FÁRMACOS A PARTIR DA ANÁLISE DE SIMILARIDADE DE MÉTRICAS EM SÍTIOS DE LIGAÇÃO

Divinópolis
2023

ALEX RAIMUNDO DE OLIVEIRA

UTILIZAÇÃO DE MODELOS BASEADOS EM GRAFOS PARA A DESCOBERTA E REPOSICIONAMENTO DE FÁRMACOS A PARTIR DA ANÁLISE DE SIMILARIDADE DE MÉTRICAS EM SÍTIOS DE LIGAÇÃO

Trabalho de Conclusão de Curso apresentado no curso de Graduação em Engenharia de Computação do Centro Federal de Educação Tecnológica de Minas Gerais como requisito parcial para obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Mestre Michel Pires da Silva
Coorientador: Doutor Eduardo Habib Bechelane Maia

DIVINÓPOLIS

2023

ALEX RAIMUNDO DE OLIVEIRA

UTILIZAÇÃO DE MODELOS BASEADOS EM GRAFOS PARA A DESCOBERTA E REPOSICIONAMENTO DE FÁRMACOS A PARTIR DA ANÁLISE DE SIMILARIDADE DE MÉTRICAS EM SÍTIOS DE LIGAÇÃO

Trabalho de Conclusão de Curso apresentado no curso de Graduação em Engenharia de Computação do Centro Federal de Educação Tecnológica de Minas Gerais como requisito parcial para obtenção do título de Bacharel em Engenharia de Computação.

Aprovado em 12 de dezembro de 2023.

gov.br Documento assinado digitalmente
MICHEL PIRES DA SILVA
Data: 15/02/2024 08:48:43-0300
Verifique em https://validar.iti.gov.br

Mestre Michel Pires da Silva

CEFET-MG Campus Divinópolis

gov.br Documento assinado digitalmente
EDUARDO HABIB BECHELANE MAIA
Data: 15/02/2024 13:28:26-0300
Verifique em https://validar.iti.gov.br

Doutor Eduardo Habib Bechelane Maia

CEFET-MG Campus Divinópolis

gov.br Documento assinado digitalmente
TIAGO ALVES DE OLIVEIRA
Data: 16/02/2024 10:08:27-0300
Verifique em https://validar.iti.gov.br

Mestre Tiago Alves de Oliveira

CEFET-MG Campus Divinópolis

Dedico aos meus pais e amigos que me auxiliaram durante o processo de construção deste trabalho.

“O ontem é história, o amanhã é um mistério, mas o hoje é uma dádiva. É por isso que se chama presente.”

Mestre Oogway

RESUMO

Um dos constantes desafios da indústria farmacêutica é a descoberta e desenvolvimento de novos fármacos, o qual está associado a altos custos de dinheiro e tempo durante seu processo. Através do uso do *Computer Assisted Drug Design* (CADD) é possível aplicar métodos para a diminuição desses custos. Dentre as estratégias presentes no CADD, como Triagem Virtual baseada em ligantes, alvos e fragmentos, estratégias para seleção e triagem de moléculas em extensas bases de dados, destaca-se a da criação e visualização de Redes de Espaços Químicos (REQ). As REQ surgem como uma recente estratégia alternativa às representações de espaços químicos baseadas em coordenadas, que são muitas vezes de difícil interpretação e visualização. Uma REQ pode ser gerada através da interpretação das relações entre moléculas, modeladas por meio de um grafo, onde os nós representam as moléculas e as arestas representam algum tipo de função que avalia a similaridade entre duas moléculas, dando destaque para as funções de Tanimoto, Par Molecular Correspondente e Variante de similaridade de Tanimoto baseada na Máxima Comum Subestrutura. As REQs, possuem a capacidade de, através das relações e propriedades das redes de grafos, prover certas percepções relacionadas às bases de estudos para os pesquisadores, tendo em vista que em uma REQ é possível atribuir uma carga considerável de informações aos nós dá rede. Este trabalho teve então como objetivo, construir uma ferramenta para a criação de uma REQ, como também descrever os passos e fundamentação teórica para tal.

Palavras-chave: REQ; funções de similaridade; CADD; reposicionamento de fármacos.

ABSTRACT

One of the constant challenges in the pharmaceutical industry is the discovery and development of new drugs. In an effort to reduce costs and time associated with this challenge, the industry utilizes Computer-Aided Drug Design (CADD). Among the strategies present in CADD, such as Virtual Screening Based on ligands, targets, and fragments, strategies for selection and screening of molecules in extensive databases, the creation and visualization of Chemical Space Networks (CSN) stands out. CSNs emerge as a recent alternative strategy to coordinate-based representations of chemical spaces, which are often difficult to interpret and visualize. A CSN can be generated by interpreting the relationships between molecules, modeled through a graph where the nodes represent molecules and the edges represent some type of function that evaluates the similarity between two molecules, with emphasis on Tanimoto functions, Molecular Matched Pair, and Tanimoto Similarity Variant based on Maximum Common Substructure. CSNs have the ability to provide insights related to study bases for researchers through the relationships and properties of graph networks, considering that a CSN can assign a considerable amount of information to the nodes of the network. This work aims to create a tool for the generation of a CSN, as well as to describe the steps and theoretical foundation for it.

Keywords: CSN; similarity functions; CADD; drug discovery.

SUMÁRIO

1 INTRODUÇÃO

Durante a maior parte da história da humanidade, os seres humanos dependiam principalmente de plantas, ervas e outros recursos naturais para o tratamento de doenças e enfermidades, o que resultava em altas taxas de mortalidade, relacionadas principalmente ao baixo conhecimento entre doenças e seus mecanismos de funcionamento. Com o surgimento dos primeiros medicamentos sintéticos no início do século XIX (SNEADER, 2010), muitas dessas doenças, que anteriormente não tinham um tratamento eficaz, como tuberculose e pneumonia, com a aplicação dos medicamentos sintéticos e naturais, começaram a ser tratadas de forma efetiva e, na maioria dos casos, completamente curadas.

Os novos medicamentos proporcionaram avanços significativos no tratamento eficaz de doenças e enfermidades, porém, uma parcela considerável da população ainda enfrenta desafios para acessar medicamentos essenciais, resultando em mais de 2 bilhões de pessoas incapazes de adquirir medicamentos básicos (CHAN, 2018). Em regiões como África e Ásia, essa problemática é ainda mais acentuada, com mais de 50% da população enfrentando dificuldades na obtenção de medicamentos (LEISINGER; GARABEDIAN; WAGNER, 2012). O custo elevado de muitos medicamentos torna-os inacessíveis para populações de baixa renda e países de renda média, sendo esse fator, provavelmente, o principal obstáculo ao acesso aos medicamentos (STEVENS; HUYS, 2017).

Ao mesmo tempo em que a demanda da população por medicamentos aumenta, observa-se um crescente ônus em termos de tempo e custo envolvidos no processo de descoberta e reposicionamento de fármacos, o que resulta em uma diminuição da quantidade desses compostos que chegam ao mercado. Essa situação é atribuída às exigências regulatórias, à repetição das mesmas metodologias e aos altos custos de produção aplicados (ZHAO; GUO, 2019). Por exemplo, o desenvolvimento de um novo medicamento para o tratamento da doença de Alzheimer pode custar até 5,6 bilhões de dólares, desde o início da pesquisa até o lançamento do novo fármaco (CUMMINGS; REIBER; KUMAR, 2018). Em geral, o desenvolvimento de novos fármacos tem um custo estimado entre 1 e 2 bilhões de dólares, podendo levar de 10 a 17 anos para estarem disponíveis para a população, levando em consideração todo o tempo de

desenvolvimento, desde a descoberta do alvo até o registro do fármaco (LEELANANDA; LINDERT, 2016). Ainda assim, a probabilidade de um candidato a fármaco chegar ao mercado após entrar na fase de ensaios clínicos diminuiu de 10% no período de 2002-2004 para aproximadamente 5% entre os anos de 2006 e 2008, representando uma queda de 50% em apenas quatro anos (ARROWSMITH, 2012).

Um fármaco é uma molécula de pequeno porte que tem como objetivo se ligar a uma proteína alvo por meio de interações energéticas entre os átomos da molécula e uma região específica da proteína, conhecida como sítio de ligação. Essa ligação desencadeia uma série de efeitos no alvo, os quais podem variar de acordo com o mecanismo de interação explorado. Esse mecanismo pode ser conduzido, por exemplo, por interações químicas, tais como ligações de hidrogênio, interações eletrostáticas, ligações covalentes ou interações hidrofóbicas.

Para obter eficácia nesse contexto, são realizadas análises em extensas bases de dados. Essas análises têm como propósito avaliar as propriedades de compostos químicos e selecionar aqueles que apresentam maior probabilidade de se ligarem de forma adequada a um alvo proteico específico. É frequente utilizar o termo ligantes ou moléculas de pequeno porte para se referir a tais compostos. São empregadas então técnicas computacionais, conhecidas como *in silico*, que possibilitam o uso do computador como uma ferramenta para agilizar e reduzir os custos envolvidos na produção de novos fármacos. A aplicação de métodos de simulação computacional permite a avaliação de diversos fatores essenciais no estudo de candidatos a fármacos, tais como toxicidade, atividade biológica e disponibilidade, mesmo antes da realização dos testes em laboratório (*in vitro*) e em organismos vivos (*in vivo*) (FERREIRA; SANTOS et al., 2015).

Durante o início da década de 60, foram realizadas análises computacionais para estabelecer relações quantificáveis entre a estrutura química de moléculas e seus efeitos farmacodinâmicos (PD, do inglês *Pharmacodynamics*) e farmacocinéticos (PK, do inglês *Pharmacokinetics*). O objetivo era fornecer estimativas da bioatividade das moléculas, marcando o primeiro uso do computador como uma ferramenta auxiliar na área da bioquímica. Essa abordagem ficou conhecida como Relações Estrutura-Atividade Quantitativas (QSARs, do inglês *Quantitative Structure–Activity Relationships*), originalmente desenvolvidas para estudar conjuntos de compostos quimicamente semelhantes.

Após o desenvolvimento das QSARs, surgiu o processo de Triagem Virtual (TV), que envolve a análise de grandes quantidades de moléculas, pontuando-as e classificando-as de acordo com sua probabilidade de ter afinidade com um alvo específico. Essa abordagem permite estender o conceito das QSARs em uma dimensão química mais ampla (BREDA et al., 2008). Existem essencialmente três tipos de abordagens utilizadas no processo de Triagem Virtual: TV baseada em estrutura (SBVS - *Structure-based Virtual Screening*), TV baseada em ligante (LBVS - *Ligand-based Virtual Screening*) e a TV baseada em fragmento (FBVS - *Fragment-Based Virtual Screening*). Existem ainda técnicas que combinam elementos da SBVS e da LBVS com o objetivo de minimizar erros e atender aos requisitos necessários na utilização das técnicas de forma separada (MAIA et al., 2020).

Além das três abordagens mencionadas anteriormente, em relação à classe de algoritmos utilizados, podemos dividir os processos de Triagem Virtual em outros tipos: aqueles baseados em Similaridade, aqueles baseados em aprendizado de máquina, aqueles que utilizam métodos qualitativos e aqueles realizados com algoritmos evolucionários (DIAS; FILGUEIRA DE AZEVEDO, 2008).

Na área da computação, é possível utilizar métodos baseados em aprendizado de máquina devido à disponibilidade de grandes volumes de dados, necessários para o treinamento dos algoritmos. Com a crescente popularização das vastas bases de dados de moléculas, o processo baseado em similaridade passou então a contar com a possibilidade da aplicação de aprendizado de máquina.

Considerando a complexidade em termos de tempo e custo do processo de descoberta e reposicionamento de fármacos, é fundamental implementar medidas e abordagens que reduzam esses impactos. Nesse contexto, diversas estratégias *in silico* têm sido desenvolvidas para classificar e quantificar ligantes. No entanto, essas abordagens muitas vezes falham em identificar relações entre múltiplos ligantes, especialmente em estudos com alvos múltiplos. Portanto, é necessário explorar soluções que permitam aos pesquisadores identificar sobreposições estruturais em grandes bases de ligantes, visando agilizar o processo de descoberta e reposicionamento de fármacos. Nesse sentido, as redes de relação química, também conhecidas como redes de espaço químico (*Chemical Space Networks* - CSN), surgem como uma alternativa promissora (MAGGIORA; BAJORATH, 2014).

Um CSN é uma representação visual das relações mensuráveis entre moléculas. Essa abordagem pode ser aplicada para mapear as interações entre alvos e/ou ligantes, sendo modelada como uma rede de grafos interconectados. Nesse modelo, as arestas são estabelecidas com base em relações moleculares, como medidas de similaridade ou semelhanças em suas propriedades físico-químicas. Os nós da rede correspondem às próprias moléculas e contêm informações atribuídas a elas, como atividade biológica, por exemplo (VOGT et al., 2016). Essa representação do espaço químico permite uma melhor compreensão das relações entre as moléculas em uma pré-análise, o que pode direcionar de forma mais eficiente os estudos quando o conjunto de exploração é extenso (RECANATINI; CABRELLE, 2020).

Conforme mencionado anteriormente, as conexões em um CSN podem ser estabelecidas de várias maneiras, utilizando diferentes métodos de pontuação. Dependendo da abordagem desejada para construir a rede, podem ser aplicados métodos amplamente utilizados na literatura, como a similaridade de Tanimoto (MAGGIORA; SHANMUGASUNDARAM, 2004), os pares moleculares correspondentes (*Matched Molecular Pairs*) (KENNY; SADOWSKI, 2005) e a variante de similaridade de Tanimoto baseada na Subestrutura Comum Máxima (ZHANG, B. et al., 2015). Cada função de pontuação pode resultar em diferentes topologias para as redes, aumentando ou reduzindo a densidade e o peso das conexões entre as moléculas. As CSNs geralmente são baseadas em funções de similaridade numérica ou subestrutural, mas também podem ser construídas de outras formas, como a combinação dessas duas abordagens (VOGT et al., 2016).

Essa representação em rede permite uma visualização mais clara das relações de similaridade entre os compostos, formando grupos ou agrupamentos de moléculas semelhantes. Essa abordagem é útil para identificar famílias de compostos com propriedades químicas e atividades biológicas semelhantes.

Ao analisar a CSN, é possível observar a formação de sub-redes ou subgrupos dentro da rede global, onde os compostos estão mais densamente conectados entre si. Esses subgrupos representam conjuntos de compostos com características estruturais e propriedades similares. Essa informação é valiosa no processo de descoberta e desenvolvimento de fármacos, pois sugere que compostos dentro de um mesmo subgrupo podem compartilhar mecanismos de ação semelhantes e, portanto, apresentar

atividades biológicas relacionadas (MAGGIORA; BAJORATH, 2014).

Além disso, as CSNs também podem revelar novas interpretações sobre a diversidade química dos compostos (MALDONADO et al., 2006). Ao analisar a distribuição dos nós e das arestas na rede, é possível identificar regiões mais densas, onde possui muitas conexões entre os nós, e regiões mais dispersas, onde possui poucas conexões entre os nós.

A partir então dos motivos previamente apresentados, o objetivo do presente estudo consiste em propor uma ferramenta para a composição e criação de Redes de Espaços Químicos. Essa ferramenta tem como intuito reduzir os custos relacionados à identificação de promissores candidatos a fármacos, viabilizando a triagem de moléculas por meio da detecção de sobreposições não óbvias entre múltiplos ligantes e alvos. Além disso, busca-se uma solução que proporcione aos usuários uma ampliação da visão e compreensão das relações, permitindo a flexibilidade na geração de diversas topologias de CSN. Para tanto, é esperado que essas topologias possam ser definidas com critérios de ponderação específicos para cada problema e/ou investigação.

2 FUNDAMENTAÇÃO TEÓRICA

Neste trabalho, tem-se por objetivo, a construção de uma ferramenta para geração e visualização de CSNs, por tanto, nos próximos capítulos desta seção, será apresentada a fundamentação teórica necessária baseada nos passos e conhecimentos teóricos necessários para sua construção.

2.1 Processo de Desenvolvimento *in silico*

Existem três grandes categorias de experimentos na área de desenvolvimento de fármacos e que são utilizados em conjunto, são eles: *in vitro*, *in vivo* e *in silico*.

a) *In vitro* (do latim "dentro do vidro) refere-se à técnica de realizar um determinado procedimento em um ambiente controlado fora de um organismo vivo. Muitos experimentos em biologia celular são conduzidos fora de organismos ou células.
b) *In vivo* (do latim "dentro do vivo") refere-se à experimentação utilizando um organismo completo e vivo. Estudos em animais e ensaios clínicos são duas formas de pesquisa *in vivo*. São geralmente mais adequados para se observar os efeitos gerais de um experimento diretamente no organismo vivo.
c) *In silico* é uma expressão usada para significar "realizado em um computador ou por meio de simulação computacional". Estes tipos de estudos são amplamente utilizados em estudos que preveem como medicamentos podem interagem como o corpo e com patógenos.

Dado o objetivo do trabalho, a seguir serão aprofundados os conceitos e abordagens do universo de desenvolvimento *in silico*.

2.1.1 *Computer-Assisted Drug Design* (CADD)

Dada a elevada quantidade de tempo e recursos financeiros investidos no desenvolvimento de novos medicamentos, os pesquisadores estão constantemente criando novos métodos que auxiliem nesse processo (HILLISCH; PINEDA; HILGENFELD, 2004). Uma das metodologias estabelecidas com o objetivo de reduzir o peso do processo de reposicionamento e descoberta de fármacos é o Desenho de Medicamentos Assistido

por Computador (CADD, do inglês *Computer-Aided Drug Design*). O CADD é um procedimento cíclico utilizado na concepção de novos medicamentos, em que todas as etapas, tanto de concepção quanto de avaliação, são realizadas por meio de programas de computador e conduzidas por especialistas em química medicinal (OGLIC et al., 2018).

O principal objetivo do CADD é avaliar a interação entre o ligante e um alvo específico. Observa-se principalmente a intensidade dessa afinidade e a configuração geométrica adotada pelo complexo após a ligação. Geralmente, essas atividades envolvem o uso de técnicas de mecânica e dinâmica molecular, bem como métodos de química quântica *ab initio*, que são aqueles independentes de conhecimentos prévios derivados de experimentos. Além disso, funções de pontuação são utilizadas para estimar quantitativamente a afinidade de ligação entre um alvo e um ligante. Esses métodos podem empregar uma variedade de técnicas de ciência da computação, como regressão linear (RUIZ-CARMONA et al., 2014), aprendizado de máquina (LIU; WANG, 2015), redes neurais (TAYARANI et al., 2013) ou outras técnicas estatísticas (HARRISON, 2010).

Nos estágios iniciais do desenvolvimento de um fármaco, é comum ter poucas informações, ou até mesmo nenhuma, sobre o alvo, os ligantes e a estrutura. Através das técnicas de CADD, é possível obter informações relacionadas às moléculas, como por exemplo quais proteínas são suscetíveis de serem os alvos em uma patogênese e quais são os possíveis ligantes ativos que podem inibir a atividade dessas proteínas. Kapetanovic (KAPETANOVIC, 2008) afirma que as técnicas de CADD têm principalmente os seguintes objetivos:

a) Agilizar o processo de descoberta de fármacos por meio do uso de simulações computacionais.
b) Otimizar e identificar novos fármacos utilizando abordagens computacionais para obter informações químicas e biológicas sobre possíveis ligantes e/ou alvos.
c) Eliminar compostos com propriedades indesejáveis e selecionar os candidatos com maior probabilidade de sucesso.

Mesmo antes dos testes em laboratório, é possível obter por meio do CADD simulações e previsões de diversos fatores, como toxicidade, atividade, biodisponibilidade e eficácia (FERREIRA; GLAUCIUS; ANDRICOPULO, 2011). Essa capacidade tem levado ao aumento do destaque do CADD. A obtenção desses fatores possibilita um melhor

planejamento e direcionamento da pesquisa, o que, neste caso, implica em menos testes *in vitro* e *in vivo*, resultando em redução do tempo e dos custos de pesquisa. Além disso, o CADD permite a análise de grandes quantidades de pequenas moléculas em um curto período de tempo, revelando como elas se ligam a alvos de interesse farmacológico mesmo antes de serem sintetizadas. Geralmente, o processo de desenvolvimento de um fármaco tem início com a identificação de alvos moleculares para um determinado composto, seguida de uma análise molecular minuciosa das principais forças de reconhecimento entre o ligante e o alvo. Em seguida, são desenhadas as moléculas mais promissoras e, por fim, realizam-se os ensaios experimentais tanto em laboratório (*in vitro*) quanto em organismos vivos (*in vivo*).

Embora as CADD auxiliem no desenvolvimento de fármacos, reduzindo custos e tempo de pesquisa, as simulações geralmente também requerem um alto custo computacional, fazendo-se necessária a utilização de máquinas de alto desempenho para a minimização do tempo associado à obtenção desses modelos computacionais. Diante disso, torna-se um desafio constante encontrar diferentes soluções que possam diminuir o custo, tempo de pesquisa, o tempo necessário para realizar as simulações e, ao mesmo tempo, aumentar sua precisão (DUFFY et al., 2012). Nesse contexto, a Triagem Virtual emerge como uma abordagem promissora.

2.1.2 Triagem Virtual

A Triagem Virtual (VS, do inglês *Virtual Screening*) é uma técnica *in silico* aplicada no processo de descoberta de fármacos. Consiste em um conjunto de métodos computacionais que classificam moléculas em um banco de dados de acordo com a previsão de suas propriedades biológicas ou químicas, sejam elas contínuas ou categóricas (MARTIN et al., 2016). Funcionando como um filtro, a VS avalia automaticamente grandes bases de estruturas moleculares tridimensionais utilizando métodos computacionais, selecionando as moléculas mais promissoras para os testes *in vitro*. Por meio do uso da VS, espera-se identificar as pequenas moléculas que têm maior probabilidade de se ligarem ao alvo molecular, geralmente uma proteína ou receptor enzimático. Portanto, a TV auxilia na identificação dos melhores candidatos capazes de se ligarem ao alvo, e somente as moléculas mais promissoras são sintetizadas. Além

disso, a TV identifica compostos que podem ser tóxicos ou possuir propriedades farmacodinâmicas e farmacocinéticas desfavoráveis. Assim, as técnicas de TV têm desempenhado um papel destacado entre as estratégias para a identificação de novas substâncias bioativas (BERMAN et al., 2013).

A TV no contexto do desenvolvimento de fármacos tem se tornado uma ferramenta indispensável para auxiliar na redução do tempo e dos custos associados à descoberta e otimização de fármacos (ZHANG, G. et al., 2018). Essa técnica desempenha um papel crucial na identificação de moléculas bioativas, uma vez que permite a seleção de compostos em um banco de dados estrutural que apresentam maior probabilidade de manifestar atividade biológica frente a um alvo de interesse. Após a identificação da bioatividade, as moléculas são submetidas a ensaios biológicos. Além disso, existem técnicas de TV que empregam métodos de aprendizado de máquina capazes de prever compostos com propriedades farmacodinâmicas, farmacocinéticas ou toxicológicas específicas, com base exclusivamente em suas características estruturais e físico-químicas derivadas da estrutura do ligante (MA et al., 2009). Dessa forma, as ferramentas de TV desempenham um papel significativo entre as estratégias para a identificação de novas substâncias bioativas, proporcionando melhorias na velocidade do processo de descoberta de fármacos, uma vez que avaliam automaticamente grandes bibliotecas de compostos por meio de simulações computacionais.

Existem essencialmente três técnicas de TV que são utilizadas atualmente, essas técnicas compreendem as TVs baseadas em estrutura, ligante e fragmento. Existe ainda uma abordagem híbrida entre as TVs baseadas em estrutura e em ligante.

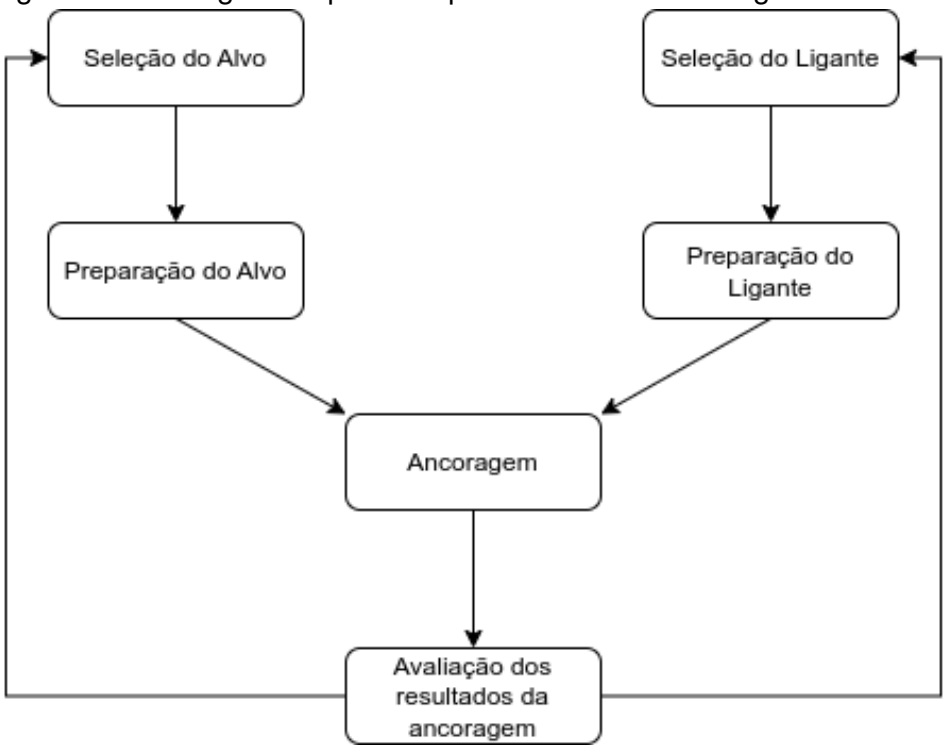
2.1.3 TV Baseada em Estrutura

A TV Baseada em Estrutura (SBVS, do inglês *Structure-based Virtual Screening*), também conhecida como TV Baseada em Alvo (TBVS, do inglês *Target-Based Virtual Screening*), tem como objetivo prever o melhor modo de interação entre duas moléculas para formarem um complexo estável. Para isso, considera-se a probabilidade de ancoragem dos ligantes candidatos com a proteína alvo, levando em consideração o grau de afinidade do complexo. Os métodos de SBVS requerem que a estrutura tridimensional do alvo seja conhecida, a fim de que as interações entre ele e cada composto possam ser

previstas. Nessa estratégia, os compostos são selecionados em uma base de dados e ranqueados de acordo com sua afinidade com o sítio do receptor (LIU; WANG, 2015).

Dentre as técnicas de SBVS, o ancoragem molecular, também conhecido como *docking*, tem sido amplamente empregado devido ao seu baixo custo computacional e aos resultados satisfatórios obtidos. Essa técnica surgiu na década de 1980, quando pesquisadores desenvolveram e testaram uma série de algoritmos capazes de explorar os alinhamentos geometricamente viáveis entre um ligante e um alvo. Existem diversos métodos para realizar o ancoragem molecular, sendo que alguns apresentam melhor desempenho em diferentes classes de alvos, sendo necessário selecionar aquele mais adequado ao alvo de interesse. O ancoragem molecular possui uma ampla variedade de usos e aplicações na área de descoberta de fármacos, incluindo estudos de estrutura-atividade, otimização de *leads*, que são compostos em estágios iniciais de desenvolvimento, busca de possíveis *leads* por TV, fornecimento de hipóteses de ligação para facilitar previsões em estudos de mutagênese, entre outras aplicações (LEELANANDA; LINDERT, 2016) .

Figura 1 – Fluxograma típico dos passos de uma ancoragem molecular.



Fonte: (MORRIS; LIM-WILBY, 2008)

A Figura 1 ilustra os principais procedimentos comuns a todos os métodos de

docking. Por tratar-se de uma técnica de TV Baseada em Estrutura (SBVS), o primeiro passo consiste em selecionar a estrutura tridimensional da macromolécula alvo e da pequena molécula a ser validada. O segundo passo envolve a preparação de cada estrutura, sendo que esse processo pode variar de acordo com o tipo de método de ancoragem molecular utilizado. Após a realização do *docking*, é necessário realizar uma análise dos resultados e selecionar as configurações de ancoragem que obtiveram as pontuações mais elevadas (MORRIS; LIM-WILBY, 2008).

O modo de ligação de um ligante em relação ao alvo pode ser definido de forma única por meio de suas variáveis de estado. Estas consistem em sua posição (traduções nos eixos x, y e z), orientação (ângulos de Euler, ângulo de eixo ou um *quartenion*) e, no caso do ligante ser flexível, sua conformação (os ângulos de torção de cada ligação rotativa). Cada uma dessas variáveis de estado representa um grau de liberdade em um espaço de busca multidimensional, e seus limites determinam a extensão da busca. O *docking* de estruturas rígidas é mais rápido devido ao menor espaço de busca, no entanto, se a conformação do ligante não estiver correta, a probabilidade de encontrar um encaixe complementar será reduzida (MORRIS; LIM-WILBY, 2008).

Todos os métodos de *docking* requerem uma função de pontuação para classificar as diversas configurações de ligação candidatas e um método de busca para explorar as variáveis de estado. As funções de pontuação podem ser empíricas, baseadas em campos de força ou baseadas em conhecimento, enquanto os métodos de busca se dividem em duas categorias principais: sistemáticos e estocásticos. Os métodos de busca estocásticos realizam alterações aleatórias nas variáveis de estado de forma iterativa até que um critério de término definido pelo usuário seja atingido, resultando em variações nos resultados da busca. Sousa et al. (SOUSA; FERNANDES; RAMOS, 2006) discutem essas classes de algoritmos com mais detalhes. Os métodos de busca também podem ser classificados de acordo com a abrangência com que exploram o espaço de busca, sendo categorizados como locais ou globais. Os métodos de busca local tendem a encontrar o mínimo local mais próximo da conformação atual, enquanto os métodos de busca global buscam o mínimo global ou o de menor energia dentro do espaço de busca definido. Os métodos de busca híbridos globais-locais têm demonstrado um desempenho ainda melhor do que os métodos globais isolados, sendo mais eficientes e capazes de encontrar energias mais baixas. Embora os métodos de TV baseados em estrutura ofereçam várias vantagens

para o desenvolvimento e reposicionamento de fármacos, é importante mencionar uma desvantagem significativa: a geração de falsos positivos, que ocorre quando a predição de ligantes para alvos moleculares seleciona estruturas que não se conectam como esperado nos ensaios *in vitro* e *in vivo*.

Conforme mencionado previamente, as técnicas de TV Baseada em Estrutura (SBVS) requerem o conhecimento prévio das estruturas tridimensionais tanto do alvo quanto do ligante. No entanto, existe outra classe de TV, conhecida como TV Baseada em Ligante, na qual não é necessário o conhecimento de ambas as estruturas. Nessa abordagem, apenas a estrutura das pequenas moléculas a serem estudadas é necessária.

2.1.4 TV Baseada em Ligante

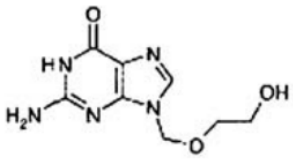
A TV Baseada em Ligante (LBVS, do inglês *Ligand-based Virtual Screening*), também conhecida como TV Baseada em Moléculas, é uma das categorias de técnicas computacionais utilizadas para a TV. Nesse tipo de abordagem, não é necessário ter conhecimento prévio da estrutura tridimensional do alvo molecular, tornando-a adequada para situações em que pouco ou nenhum conhecimento está disponível sobre a estrutura do receptor ou quando o sítio de ligação do alvo não está bem definido. É especialmente útil quando se deseja explorar uma biblioteca de moléculas conhecidas em busca de candidatos promissores para interações com o alvo, com base em similaridade estrutural, propriedades químicas ou atividade biológica relatada.

A LBVS parte do pressuposto de que ligantes que exibem alguma forma de similaridade molecular têm maior probabilidade de apresentar propriedades biológicas semelhantes. Portanto, ligantes que são similares a um ligante ativo para uma determinada proteína-alvo têm maior chance de também serem ativos, em comparação com ligantes que não possuem nenhuma forma de similaridade. Essa abordagem permite explorar relações estruturais e químicas entre moléculas para identificar candidatos promissores com base em características compartilhadas com um ligante ativo conhecido. A TV Baseada em Ligante utiliza as informações presentes nos ligantes ativos em vez da estrutura do alvo, tanto para identificação quanto para seleção de *leads*. Os métodos baseados no ligante são a única opção quando as informações da estrutura

tridimensional do alvo não estão disponíveis, pois, mesmo sem conhecer a estrutura do alvo de interesse, muitas vezes é conhecido um conjunto de ligantes que é ativo contra ele (MALDONADO et al., 2006).

Os métodos computacionais baseados em ligantes requerem dois elementos fundamentais para o seu efetivo funcionamento: uma medida de similaridade eficiente e uma função de pontuação confiável. Além disso, o método selecionado deve ser capaz de realizar a triagem de um grande número de potenciais ligantes com precisão e velocidade adequadas. As medidas de similaridade molecular frequentemente envolvem três componentes principais: os descritores, os coeficientes e o esquema de ponderação.

Figura 2 – Descritores usuais associados à dimensionalidade da representação molecular.

	Typical Representation	Typical Descriptors
1D	C8H10N5O3	Molecular weight Atom counts
2D		Fragment counts Topological indices Connectivity
3D		Molecular surface Molecular volume Interaction energies

Fonte: (MALDONADO et al., 2006)

Os descritores são utilizados para caracterizar as moléculas a serem comparadas. Eles podem ser determinados a partir da estrutura (constituição, configuração e conformação) das moléculas. Os descritores constitucionais incluem informações sobre a ordem de ligação dos átomos, a presença ou ausência de fragmentos e outras características em duas dimensões. O descritor de configuração caracteriza o arranjo tridimensional dos átomos e, por fim, os descritores conformacionais representam os arranjos espaciais termodinamicamente estáveis dos átomos. A Figura 2 apresenta alguns exemplos de descritores moleculares e suas classificações, que podem ser

calculados a partir de estruturas unidimensionais (1D): peso molecular e contagem de átomos; bidimensionais (2D): contagem de fragmentos, índices topológicos e conectividade; e tridimensionais (3D): superfície molecular, volume molecular e interações energéticas (MALDONADO et al., 2006).

Os **coeficientes de similaridade** são funções que transformam pares de representações moleculares compatíveis em números reais, geralmente normalizados para valores entre 0 e 1. Esses coeficientes proporcionam uma medida quantitativa do grau de semelhança química. Ao aplicar os conceitos de similaridade na química, é necessário estabelecer distinções entre similaridades globais e locais. A similaridade global refere-se à comparação de dois objetos completos, levando em consideração todas as características e propriedades desses objetos. Por exemplo, ao comparar duas moléculas, a similaridade global levaria em consideração todas as ligações químicas, grupos funcionais e conformações presentes nas moléculas em sua totalidade. Por outro lado, a similaridade local concentra-se em partes específicas ou subestruturas particulares das moléculas em comparação. Em vez de avaliar a molécula como um todo, apenas determinadas regiões ou elementos individuais dessas estruturas são analisados. É esperado que haja resultados contrastantes ao tratar a mesma amostra por meio de análises de similaridade global ou local (MALDONADO et al., 2006).

O terceiro componente fundamental na avaliação de similaridade entre moléculas é o **esquema de ponderação**, o qual é utilizado para atribuir diferentes graus de importância aos diversos componentes dessas representações. Ao comparar moléculas, diversas propriedades ou características podem ser consideradas, tais como estrutura, polaridade, carga, entre outras. No entanto, nem todas essas características possuem o mesmo impacto na medida de similaridade entre as moléculas. Dependendo da análise e do contexto, certas características podem ser mais relevantes do que outras. Assim sendo, o esquema de ponderação possibilita a atribuição de pesos distintos a cada característica, com base em sua importância relativa (MALDONADO et al., 2006).

Diversas técnicas para a avaliação de similaridade entre duas estruturas moleculares são descritas na literatura. No entanto, no âmbito deste trabalho, há algumas que são especialmente relevantes e merecem uma análise mais aprofundada. Essas técnicas serão exploradas na próxima seção, que aborda as Redes de Espaços Químicos e destaca a importância das métricas de similaridade para a definição e compreensão de

sua construção dentro de nossa abordagem.

Como mencionado previamente, um desafio encontrado nas abordagens estruturais é a ocorrência de resultados falso-positivos, quando compostos são selecionados por meio de simulações computacionais, mas não apresentam o comportamento esperado nos ensaios *in vitro* e *in vivo*. No entanto, as abordagens baseadas em ligantes também enfrentam essa mesma questão. Uma solução para minimizar esse tipo de problema, presente em ambas as abordagens, é a adoção de modelos computacionais híbridos, que consistem na combinação das abordagens LBVS e SBVS. Como o foco da proposta não está na definição de novos métodos para o ranqueamento de moléculas, o trabalho desenvolvido em 2018, por Muniz (MUNIZ, 2018) propõe um novo método de docagem molecular híbrido e que explora os desafios da seleção e reposicionamento de candidatos a fármacos, incluindo a obtenção de falso-positivos, durante a triagem de pequenas moléculas.

2.1.5 TV Baseada em Fragmento

A Triagem Virtual Baseada em Fragmentos (FBVS, do inglês *Fragment-Based Virtual Screening*) é uma estratégia inovadora no campo de descoberta e reposicionamento de fármacos que visa desenvolver compostos potentes a partir de fragmentos moleculares simples (OLIVEIRA et al., 2023).

A FBVS geralmente inicia com a seleção de fragmentos moleculares com peso molecular baixo, geralmente inferior a 300 Da, baixa complexidade estrutural e afinidade inicial modesta pelo alvo de interesse (DOAK; NORTON; SCANLON, 2016). Esses fragmentos são então submetidos a triagens utilizando métodos biofísicos sensíveis para identificar moléculas que apresentam alguma afinidade com o alvo. Ao contrário das abordagens convencionais que visam compostos altamente potentes desde o início, a FBVS busca identificar fragmentos que possam ser desenvolvidos em moléculas mais complexas e potentes no decorrer do processo de otimização medicinal.

Na *design* de fármacos, a FBVS oferece diversas vantagens. Essas incluem a economia de custos experimentais, a diversidade de fragmentos que podem ser explorados e a flexibilidade para desenvolver novos compostos de maneiras variadas (ERLANSON et al., 2016). Esta abordagem tem sido adotada em diversos estudos para desenvolver

inibidores para diferentes tipos de alvos.

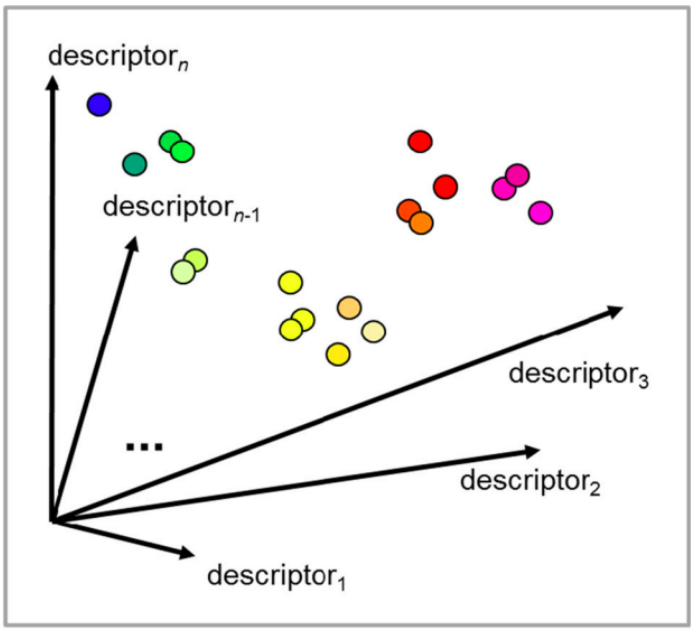
Para conduzir um experimento de triagem de fragmentos, vários procedimentos são geralmente necessários. Isso inclui a seleção de uma biblioteca de compostos, o estabelecimento de um método para a identificação de hits, a determinação das estruturas dos complexos fragmento-alvo, o desenvolvimento de um ensaio para analisar a relação estrutura-atividade e a elaboração de uma estratégia para transformar o fragmento em um inibidor potente.

2.2 Redes de Espaços Químicos

As Redes de Espaços Químicos (CSN, do inglês *Chemical Space Networks*), introduzidas durante a última década, têm sido utilizadas como uma forma de visualizar e interpretar as relações entre conjuntos de pequenas moléculas (MAGGIORA; BAJORATH, 2014). Existem outras abordagens para representar a distribuição espacial dos compostos em espaços químicos. A abordagem computacional mais popular tem sido o projeto de representações de espaço químico baseadas em coordenadas, em que as posições das moléculas são determinadas por vetores de características baseados em seus descritores. No entanto, esses espaços químicos são tipicamente de alta dimensionalidade e apresentam desafios de visualização. A Figura 3 ilustra uma representação esquemática de um espaço químico baseado em coordenadas, onde cada dimensão é representada por um descritor definido no momento de sua criação e a distância entre as moléculas, representadas pelos pontos nesses espaços, se correlaciona diretamente com a similaridade entre esses compostos. A Figura 3 demonstra ainda a dificuldade em compreender suas múltiplas dimensões. Além disso, a redução desses espaços para visualizações tridimensionais resulta em perda indesejável de informações químicas. Diante disso, as CSNs surgem como uma alternativa livre de coordenadas para representar espaços químicos (VOGT et al., 2016).

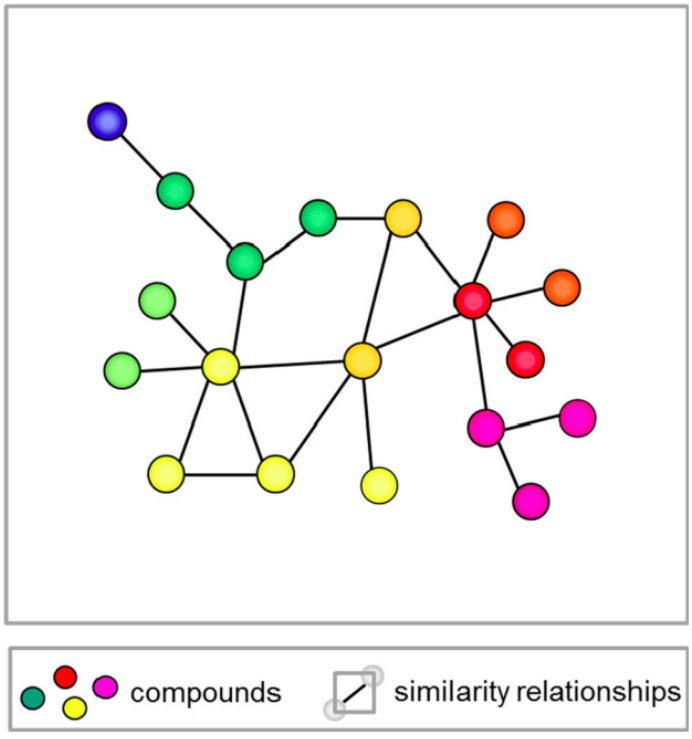
Espaços não dependentes de coordenadas podem ser construídos levando em consideração de forma explícita todos os relacionamentos em pares entre as moléculas. As Redes de Espaços Químicos (CSNs) podem ser criadas com base em uma rede molecular fundamentada na similaridade. Na Figura 4, apresentada abaixo, é possível visualizar o modelo de uma Rede de Espaços Químicos livre de coordenadas e construída com base

Figura 3 – Representação esquemática do espaço químicos baseada em coordenadas multi-dimensionais.



Fonte: (VOGT et al., 2016)

Figura 4 – Representação de uma rede de espaços químicos livre de coordenadas montada através de métricas de similaridades entre compostos.



Fonte: (VOGT et al., 2016)

em funções de similaridade. Por se tratar de uma representação livre de coordenadas a Figura 4 não possui os problemas de dimensionalidade encontrados na representação da Figura 3, ainda, as informações de semelhanças entre compostos podem ser extraídas através das conexões entre os nos, sendo assim, mais facilmente percebidas do que nos espaços baseados em descritores. Nesse tipo de CSN, a rede é estruturada utilizando conceitos de grafo, em que os nós representam os compostos e as arestas conectam os pares de nós com base em medidas de similaridade, as mesmas medidas que podem ser utilizadas para TV baseada em ligantes. Dessa forma, as CSNs se destacam por serem visualmente mais acessíveis e de mais fácil interpretação (VOGT et al., 2016).

Outra vantagem do uso desse tipo de espaço químico é a possibilidade de incluir anotações nos nós da estrutura com informações adicionais sobre os compostos. Dependendo do tipo de informação atribuída aos nós da CSN, é possível realizar diversas análises ou reduzir a abstração. Por exemplo, ao anotar os nos dos compostos com informações de potência, as CSNs podem ser utilizadas para investigar as relações de estrutura-atividade (SAR, do inglês *Structure-activity Relationship*). Para uma análise interativa, os nós podem ser diretamente vinculados à exibição da estrutura química, o que diminui o grau de abstração geralmente associado às redes. Comparadas a outras representações do espaço químico, as redes possuem a vantagem adicional de poderem ser caracterizadas e comparadas em detalhes, utilizando uma variedade de abordagens estatísticas da ciência de redes de forma geral (MAGGIORA; BAJORATH, 2014).

Recanatini (RECANATINI; CABRELLE, 2020) afirma que a primeira e mais significativa questão a ser considerada ao lidar com a construção de redes para investigar a semelhança entre moléculas está relacionada ao material utilizado para construir tais modelos, ou seja, o que é comumente denominado de "dados". Por meio dessa terminologia, ele se refere a uma ampla gama de informações, expressas em formato numérico, alfabético ou digital, coletadas em bancos de dados que frequentemente estão disponíveis publicamente. Além disso, de acordo com Recanatini, embora a comunidade de descoberta de fármacos utilize conjuntos de dados desde os primeiros métodos computacionais para o cálculo de propriedades moleculares, as informações fornecidas pelas tecnologias experimentais de alto rendimento estão crescendo a um ritmo sem precedentes. Atualmente, é possível acessar fontes de dados sobre compostos, alvos e doenças que abrangem milhões de moléculas, assim como milhares de proteínas e genes

em praticamente todas as áreas terapêuticas.

Todos os tipos de dados contidos nos bancos de dados químicos podem ser úteis para fins de projeto de fármacos em geral, mas, no que diz respeito a aplicações em redes, os dados de bioatividade, informações químicas, fármacos, produtos naturais, disponibilidade comercial e fragmentos são mais interessantes (RECANATINI; CABRELLE, 2020).

Como mencionado na Subseção 2.1.4, para explorar um pouco mais sobre o mundo das CSNs será necessário definir algumas métricas que podem ser utilizadas para a geração das arestas, métricas essas que são mais condizentes com o universo a ser explorado de acordo com esta proposta, são elas: similaridade de Tanimoto (BAJORATH, 2008), *Matched Molecular Pairs* (KENNY; SADOWSKI, 2005) e a variante de similaridade de Tanimoto baseada na Máxima Comum Subestrutura (ZHANG, B. et al., 2015).

a) Similaridade de Tanimoto

   A similaridade de Tanimoto faz parte dos conjuntos de medidas numéricas de similaridade e é comumente aplicada em Redes de Espaços Químicos utilizando *thresholds*, que são critérios estabelecidos para determinar um valor de aceitação. No caso das CSNs baseadas na similaridade de Tanimoto, o *threshold* é utilizado para filtrar as arestas que possuem valores abaixo do limite estabelecido.

   O cálculo da similaridade é realizado levando em consideração a estrutura molecular dos compostos que estão sendo analisados. Para isso, é utilizado o método de *fingerprint*, que consiste em uma representação vetorial binária de tamanho n das características estruturais da molécula. Nesse método, o valor 1 indica a presença da característica na estrutura, enquanto o valor 0 indica a ausência da característica. O valor numérico da similaridade pode ser obtido utilizando a seguinte equação:

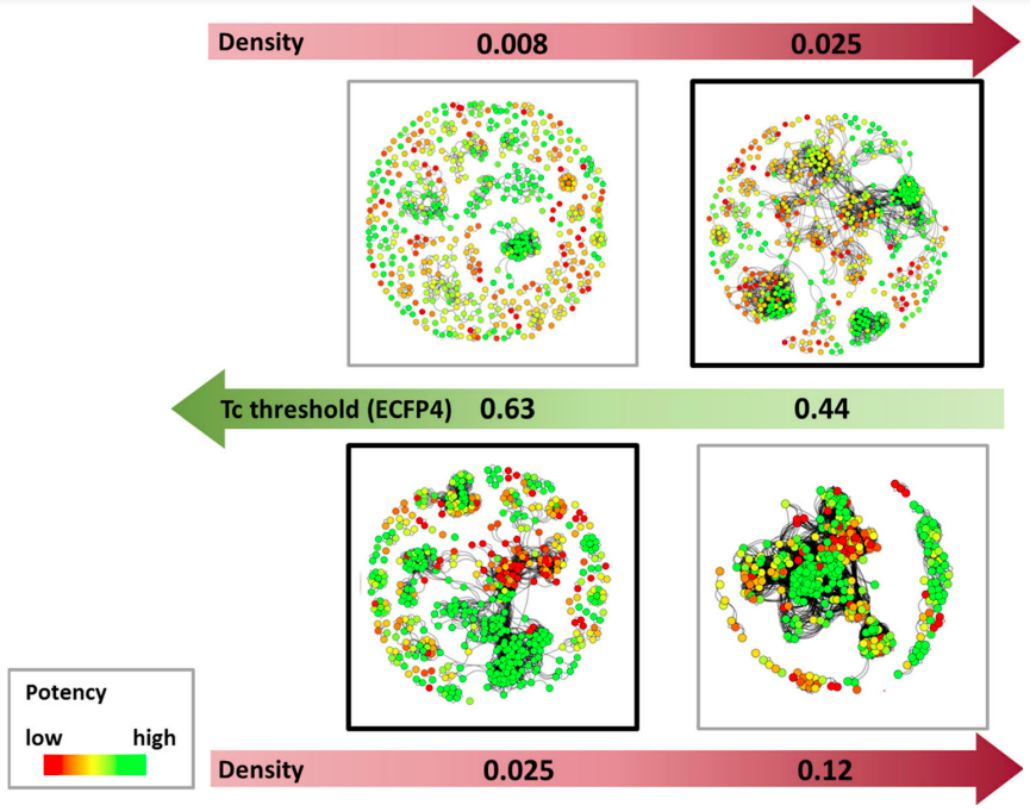
$$S_{Tan} = \frac{c}{a + b - c} \tag{2.1}$$

   em que *a* representa as propriedades da primeira molécula, *b* as propriedades da segunda molécula e *c* a quantidade de propriedades comuns entre as duas moléculas. Através do valor obtido pelo cálculo da função 2.1, é então possível determinar a similaridade entre duas moléculas (MAGGIORA;

SHANMUGASUNDARAM, 2011).

Uma vez que o *threshold* é um valor arbitrário selecionado com base no objetivo da pesquisa em questão, é viável obter diferentes configurações topológicas das CSNs, sendo que a topologia refere-se à distribuição da CSN no espaço. Além disso, a densidade das conexões na CSN pode variar de acordo com a escolha do *threshold* assim como mostra a Figura 5. É possível ainda, notar, através da Figura 5, que a densidade da CSN é inversamente proporcional ao aumento do valor de corte. Na imagem o termo ECFP4 se refere a um tipo de método utilizado para a criação de *fingerprints* moleculares.

Figura 5 – Diferentes topologias geradas a partir da alteração do *threshold* para o cáculo da similaridade de Tanimoto.



Fonte: (VOGT et al., 2016)

b) Matched Molecular Pairs

Valores de similaridade numérica são amplamente empregados na química computacional para uma variedade de propósitos. Como alternativa, medidas de similaridade baseadas em estrutura também podem ser adotadas. Nesse

contexto, a presença de uma subestrutura específica em um composto é considerada um critério de similaridade: se duas moléculas compartilham uma subestrutura definida, elas são consideradas similares; caso contrário, não são. Um *Matched Molecular Pair* (MMP) pode ser definido como um par de compostos que podem ser distinguidos pela modificação química de apenas um sítio. Em outras palavras, esses dois compostos são estruturalmente semelhantes, exceto por uma alteração em uma região específica da molécula, como a substituição de um átomo ou grupo funcional.

No contexto de uma Rede de Espaços Químicos, ao aplicar o critério de similaridade por meio de MMPs, é gerada uma CSN com densidade constante. Portanto, para uma mesma base de compostos e a utilização da mesma função de MMP, a densidade da rede será a mesma em qualquer momento em que sua construção seja realizada (KENNY; SADOWSKI, 2005).

c) Variante de similaridade de Tanimoto baseada na Máxima Comum Subestrutura

Com o objetivo de unir os benefícios das medidas contínuas de similaridade, isto é, limiares de similaridade ajustáveis, e a avaliação de similaridade baseada em estruturas, devido à sua natureza robusta e intuitiva, é interessante abordar também uma abordagem híbrida para medir a similaridade na criação de CSNs. Essa abordagem híbrida envolve a modificação do cálculo do coeficiente de Tanimoto para levar em consideração a subestrutura máxima comum (MCS) de um par de moléculas. A equação 2.1 pode ser adaptada para a seguinte forma:

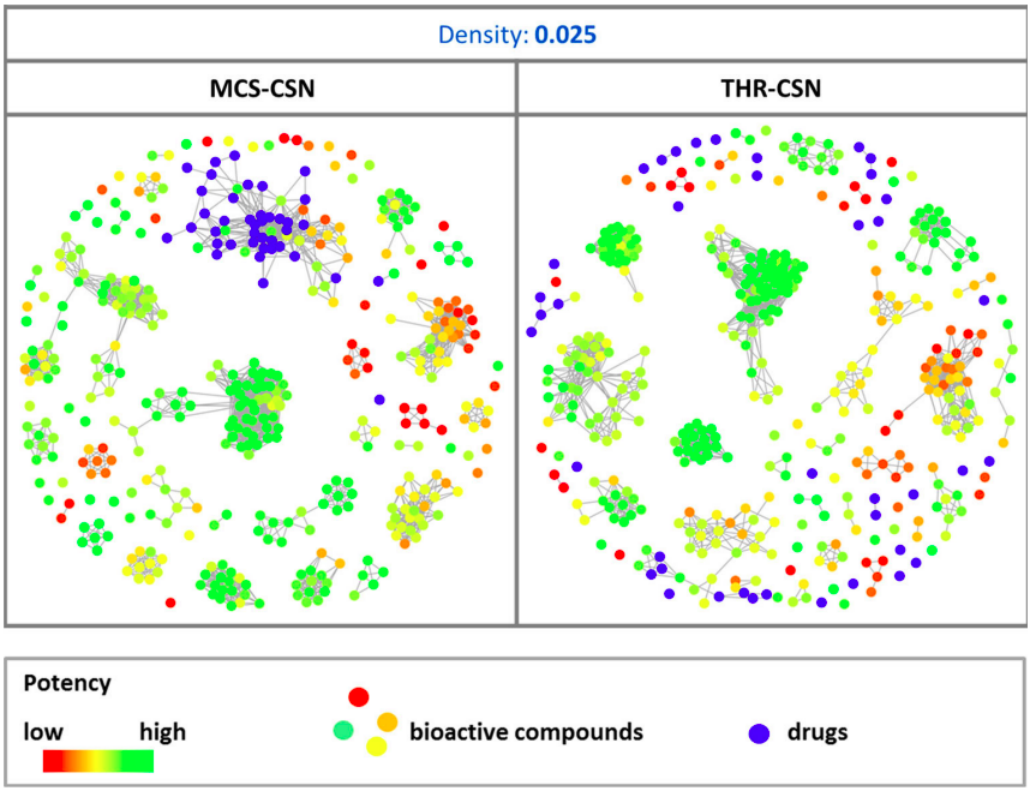
$$S_{Tan} = \frac{|MCS(A,B)|}{a + b - |MCS(A,B)|} \tag{2.2}$$

em que *a* representa as propriedades da primeira molécula, *b* as propriedades da segunda molécula, $|MCS(A,B)|$ é o cálculo da maior subestrutura em comum entre *A* e *B*, sendo *A* e *B* as respectivas moléculas (ZHANG, B. et al., 2015).

Na Figura 6 é feita uma comparação entre uma duas CSNs, uma gerada através da similaridade de tanimoto e outra gerada através da variante da MCS. É perceptível que a escolha da função de similaridade tem impacto direto na topologia e formação da rede, podendo, com o mesmo *threshold* ou não, conectar moléculas que na outra CSN poderiam não estar conectadas e vice

versa.

Figura 6 – Comparação entre as topologias geradas a partir da variante de similaridade de Tanimoto (MCS-CSN) e da similaridade de Tanimoto clássica (THR-CSN)



Fonte: (VOGT et al., 2016)

2.3 Considerações Finais

Conforme abordado no início desta seção, os métodos de Descoberta e Reposicionamento de Fármacos Assistidos por Computador (CADD, na sigla em inglês) surgem como uma estratégia para reduzir a complexidade inerente ao processo. Esses métodos permitem a avaliação das relações entre ligantes e um alvo específico. Dentre essas técnicas, destacam-se as abordagens de Seleção Virtual Baseada em Estrutura (SBVS), que realizam simulações de ancoragem molecular visando identificar a melhor conformação para a ligação de pequenas moléculas a uma determinada proteína.

Outra abordagem é a de Seleção Virtual Baseada em Ligantes (LBVS), que, em contrapartida às técnicas de SBVS, dispensa o conhecimento prévio da estrutura do alvo. Essa abordagem contribui ainda mais para a redução de custos em termos de tempo e

recursos financeiros, pois não exige um conhecimento aprofundado da estrutura do alvo para selecionar moléculas com potencial bioativo.

Na última parte da seção, é explorado o conceito de Redes de Similaridade de Compostos (CSNs), que são estruturas baseadas em grafos capazes de representar espaços químicos de pequenas moléculas livres de coordenadas. Usando técnicas de LBVS essas estruturas nos permitem visualizar relações de similaridade, as quais seriam de difícil entendimento em espaços baseados em coordenadas. Além disso, as CSNs possibilitam a inclusão de informações adicionais nos nos, além de suas estruturas ou fórmulas químicas. Portanto, as CSNs surgem como uma solução promissora para o estudo de moléculas e a seleção de possíveis fármacos, oferecendo uma representação menos abstrata e uma visão simplificada das relações entre as moléculas.

3 METODOLOGIA

Dando sequência à este trabalho, com base na fundamentação teórica, neste capítulo serão apresentadas as ferramentas e métricas necessárias para a construção de nossa solução que tem como objetivo propor uma nova ferramenta para a geração de Redes de Espaços Químicos.

3.1 Extração e Composição da Base de Dados

Existem uma série de bancos de dados disponíveis para a obtenção dos mais variados compostos. Alguns dos mais utilizados são:

a) PubChem

   Mantido pelo *National Center for Biotechnology Information* (NCBI) é uma base de dados aberta que fornece informações sobre substâncias químicas, suas propriedades e atividades biológicas. Ele abrange uma ampla gama de fontes, incluindo literatura científica, patentes e bancos de dados contribuídos (https://pubchem.ncbi.nlm.nih.gov/).

b) ZINC Database

   É uma base de dados que se concentra em fornecer compostos prontos para triagem virtual. Ele contém informações sobre compostos disponíveis comercialmente, com ênfase em estruturas prontas para docking e simulações de ligação molecular (https://zinc.docking.org/).

c) Drugbank Online

   É um banco de dados online abrangente e de acesso gratuito que contém informações sobre drogas e alvos de drogas. Como recurso de bioinformática e quimioinformática, nele são combinados dados detalhados de medicamentos aprovados pela *Food and Drug Administration* (FDA) com informações abrangentes sobre o alvo do medicamento (https://go.drugbank.com/about).

d) ChEMBL

   Outro dentre os bancos de dados mais populares que fornecem conhecimento sobre compostos bioativos, especialmente dados de ensaios de atividade e informações sobre os alvos (MENDEZ et al., 2019). Além de possuir uma

interface agradável e intuitiva, a plataforma web do ChEMBL conta com diversos recursos que facilitam o seu uso e a seleção de compostos, como filtros interativos baseados na pesquisa de um determinado termo e "Mapas de Calor de Bioatividade", que mostram a relação de bioatividade entre compostos e alvos (https://www.ebi.ac.uk/chembl/). O ChEMBL ainda conta com um *data web service* para facilitar na extração e obtenção de compostos a partir de parâmetros específicos, sua documentação está disponível em: (https://chembl.gitbook.io/chembl-interface-documentation/web-services/chembl-data-web-services).

Por esses motivos, o ChEMBL será o banco de dados escolhido para seleção das pequenas moléculas, que serão filtradas a partir de uma proteína alvo em específico, ainda a definir, e elas serão utilizadas na montagem de nossa CSN. Nessa rede, essas moléculas servirão como nos, e suas similaridades serão calculadas para a criação das arestas.

A composição da base de dados, foi feita então a partir do *web service* disponibilizado pela ChEMBL, para tal foram selecionados 500 bioativos associados às seguintes enzimas: Acetylcholinesterase, Butyrylcholinesterase e a Beta-secretase 1.

3.2 Seleção e Avaliação de Métricas de Correlação

Na Seção Seção 2.2, foram apresentados três métodos para o cálculo de similaridade entre moléculas: um baseado em similaridade numérica, outro em similaridade estrutural e um terceiro que combina os dois tipos de abordagem. A criação de CSNs usando métricas de similaridade numérica proporciona maior flexibilidade, oferecendo vantagens como a capacidade de controlar a densidade da rede gerada, aumentando ou diminuindo as conexões com base no valor do *threshold* escolhido. Isso nos permite explorar relações mais sutis entre as moléculas. No entanto, é importante salientar que uma definição inadequada do *threshold* pode levar à perda de informações importantes.

Por outro lado, a construção de uma rede baseada na estrutura química permite uma abordagem mais direta, uma vez que não é necessário definir um *threshold*. Nessa abordagem, a rede pode preservar informações detalhadas sobre as moléculas, como ligações químicas específicas, que podem ser relevantes para a compreensão das relações

estruturais e dos mecanismos de ação das moléculas. No entanto, esse método apresenta limitações na identificação de similaridades sutis entre moléculas, especialmente quando suas estruturas são bastante diferentes.

Métodos híbridos nos permitem mitigar algumas das desvantagens apresentadas por cada uma das abordagens quando utilizadas separadamente. Ao combinar essas técnicas, é possível analisar tanto a similaridade dos compostos com base em suas propriedades quanto em suas características estruturais. No entanto, a possibilidade de variação do *threshold* ainda pode ser problemática, uma vez que é necessário encontrar um ponto de equilíbrio que permita detectar similaridades sutis e, ao mesmo tempo, excluir relações irrelevantes.

Como uma estratégia mais abrangente, utilizaremos a função híbrida "Variante de similaridade de Tanimoto baseada na Máxima Comum Subestrutura"como métrica de correlação entre os compostos selecionados para construir as arestas da rede.

3.3 Modelos de Visualização e Informações Sobre Moléculas

A biblioteca RDKit (https://www.rdkit.org/) (LANDRUM; RDK, 2022) é uma ferramenta de código aberto desenvolvida com o propósito de manipular e analisar estruturas químicas. Suas funcionalidades abrangem a geração e manipulação de moléculas, cálculos de propriedades moleculares, simulações de dinâmica molecular, análise de estrutura-atividade, métricas de similaridade e diversas outras. Por meio dessa biblioteca, é possível gerar visualizações bidimensionais das estruturas moleculares dos compostos, proporcionando uma representação gráfica das características químicas. Dadas as vantagens apresentadas na Seção 2.2, para a utilização da Variante de similaridade de Tanimoto baseada na Máxima Comum Subestrutura, ela será a função de similaridade escolhida para geração das arestas da CSN, função essa que será calculada através das *features* disponíveis na RDKit.

Diversas ferramentas para a geração e visualização de redes de grafos estão disponíveis para uso, duas das mais utilizadas são a NetworkX e o Gephi. A NetworkX (https://networkx.org/) (HAGBERG; SWART; S CHULT, 2008), é uma biblioteca de código aberto em Python utilizada para a criação, manipulação e análise de redes complexas. Suas funcionalidades abrangem a construção e análise de estruturas de redes, incluindo

grafos direcionados e não direcionados. Com o auxílio da NetworkX, é possível criar redes com nos e arestas que representam uma diversidade de sistemas complexos, incluindo as Redes de Espaços Químicos. O Gephi é um *software* de código aberto para análise de gráficos e redes. Ele utiliza um motor de renderização 3D para exibir grandes redes em tempo real e acelerar a exploração. Uma arquitetura flexível e multi tarefa traz novas possibilidades para trabalhar com conjuntos de dados complexos e produzir resultados visuais valiosos. Há diversas características-chave do Gephi no contexto de exploração interativa e interpretação de redes. Ele proporciona acesso fácil e abrangente a dados de rede e permite espacialização, filtragem, navegação, manipulação e agrupamento. Finalmente, ao apresentar recursos dinâmicos do Gephi, é dado destaque a aspectos-chave da visualização dinâmica de redes (BASTIAN; HEYMANN; JACOMY, 2009). Ele disponibiliza um *toolkit* de desenvolvimento externo que permite utilizar uma série de recursos disponíveis no *software*. Esse *toolkit*, assim como o Gephi, são desenvolvidos em Java, que é uma linguagem bastante utilizada para o desenvolvimento de *apps desktop* graças ao *widget toolkit GUI* Java Swing, que já vem sendo utilizando em aplicações específicas para o desenvolvimento e análise de fármacos (MAIA et al., 2020).

Por já possuir uma clara descrição dos passos necessários para a aplicação do RDKit no contexto das CSNs, com base no artigo de Scalfani (SCALFANI; PATEL; FERNANDEZ, 2022), a RDKit será a ferrameta escolhida para manipulação dos compostos, como cálculo da métrica de similaridade e geração de estruturas 2D. No caso da ferramenta para a criação e visualização dos espaços químicos o Gephi, em conjunto do Java Swing, será a utilizada, principalmente pelo conhecimento prévio e experiência em outras ocasiões com a linguagem Java na construção de *apps desktop*.

Uma característica distintiva das Redes de Espaços Químicos (CSNs) é a possibilidade de anotar informações relacionadas aos compostos que compõem as redes nos próprios nós. Essa abordagem permite aos pesquisadores obter uma quantidade maior de dados e informações para a validação de resultados e a busca por respostas. Para a escolha das informações a serem vinculadas aos nós da rede, foi tomado como critério de seleção incluir, dentre algumas outras, as mesmas apresentadas no *Compound Report Card* de uma molécula do ChEMBL na aba de *Calculated Properties*, foram elas: *ChEMBL ID*, *Name*, *Synonyms*, *Type*, *Max Phase*, *Molecular Weight*, *AlogP*, *Polar Surface Area*, *HBA*, *HBD*, *#RO5 Violations*, *#Rotatable Bonds*, *Passes Ro3*, *QED Weighted*, *CX*

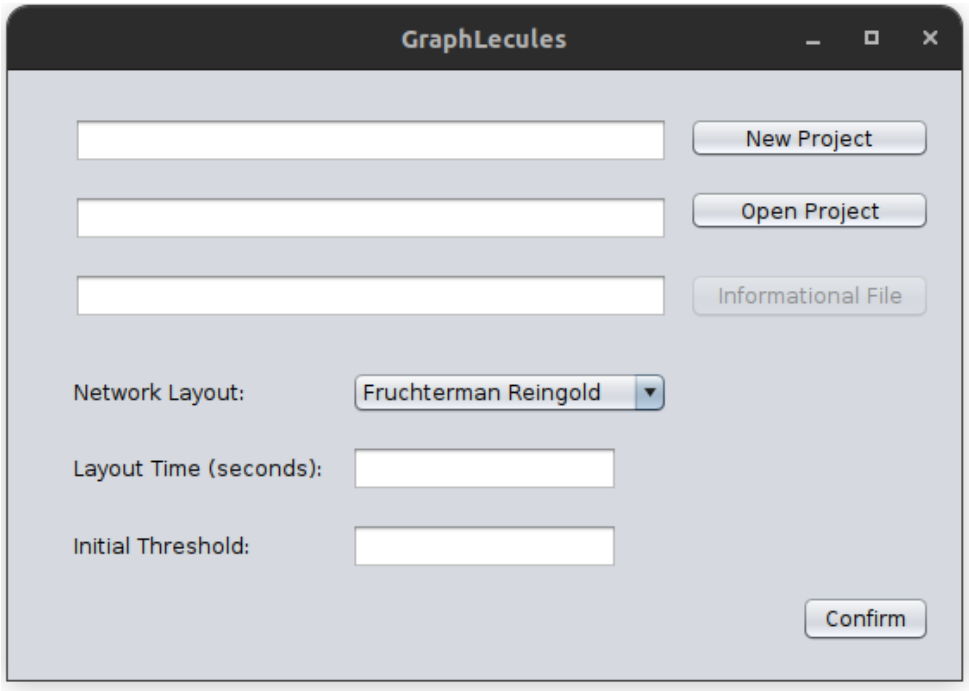
Acidic pKa, *CX Basic pKa*, *CX LogP*, *CX LogD*, *Aromatic Rings*, *Structure Type*, *Inorganic Flag*, *Heavy Atoms*, *HBA (Lipinski)*, *HBD (Lipinski)*, *#RO5 Violations (Lipinski)*, *Molecular Weight (Monoisotopic)*, *Np Likeness Score*, *Molecular Species*, *Molecular Formula*, *Smiles* e *Inchi Key*.

4 RESULTADOS

Serão discutidos e apresentados agora as *features* e resultados obtidos na versão final da ferramenta desenvolvida. A fim de facilitar, o nome "GraphLecules"(GL) será atribuído à ferramenta construída. Todos os códigos e *scripts*, assim como a base de dados utilizada, estarão disponíveis em (https://github.com/AlexR02/GraphLecules)

4.1 Tela Inicial

Figura 7 – Representação da tela inicial do GraphLecules



Fonte: Próprio Autor.

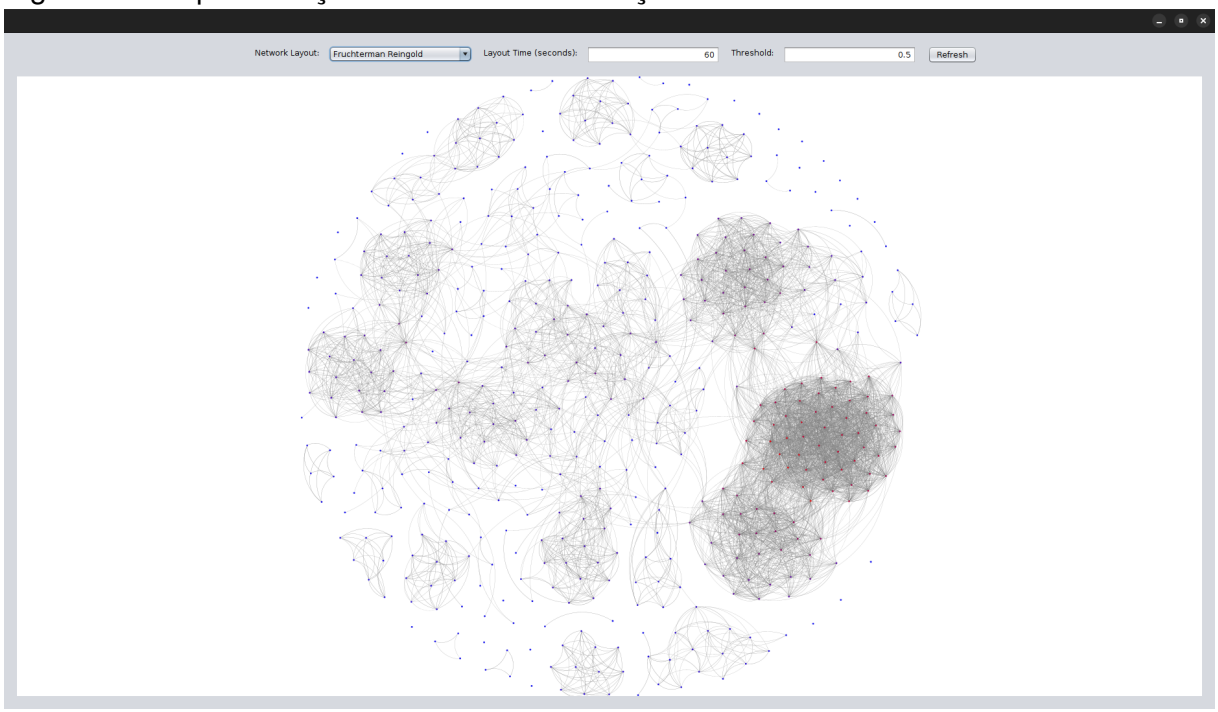
A Figura 7 é a primeira tela a ser exibida na execução do projeto. Nela é possível observar nas três primeira linhas, três botões e três caixas de texto. A primeira linha contém a caixa de texto que representa o caminho para o arquivo de um novo projeto, e o botão "*New Project*" é o botão para selecionar esse arquivo. O arquivo selecionado deve ser um arquivo do tipo csv, o padrão desse arquivo é gerado a partir do *script* "scriptAjusteFile.csv", disponível também no repositório do projeto. Para geração do arquivo, os dados dos arquivos extraídos do ChEMBL são lidos e escritos no padrão necessário para leitura pelo GL. Quando é iniciado um novo projeto o GL irá salvar criar

uma pasta na *home* do usuário e salvará nela os arquivos gerados em formato gexf, que é um tipo de arquivo específico do Gephi. Na segunda linha, a caixa de texto representa o caminho para o arquivo de um projeto já gerado pelo GL. A terceira linha é o caminho para um arquivo contendo as informações sobre as moléculas, sendo o mesmo padrão de arquivo necessário ao iniciar um novo projeto, esse arquivo só é requisitado quando se está abrindo um projeto já criado, porém o mesmo não é obrigatório, servindo apenas para o carregamento das informações atribuídas aos nós da rede, mas ainda sim, se faz necessário já que a capacidade de atribuição de informação aos nós é um dos diferenciais de uma CSN. Ainda na Figura 7 tem-se mais três linhas na metade de baixo da tela. Na primeira linha é possível escolher um *layout* específico para a geração da rede, sendo eles: Fruchterman Reingold, Force Atlas e YifanHu. Todos eles *layouts* provenientes do kit de ferramentas para desenvolvedores do Gephi. Na segunda linha representada pelo "*Layout Time (seconds)* é um parâmetro inerente ao kit do Gephi, nele é passado o tempo, em segundos, que será aplicado o layout à rede. Na terceira linha representada pelo "*Initial Threshold*"é onde será definido o limiar para podar as arestas do grafo, nele podem ser definidos valor variando de 0 a 1, onde 0 representa que uma molécula possui 0 de similaridade com outra e 1 representa que duas moléculas são totalmente similares.

4.2 Tela de Visualização da Rede

Ao clicar no botão "*Confirm*" encontrado na tela representada na Figura 7, ao final do tempo necessário para o processamento dos dados, cálculo da função de similaridade e escrita dos arquivos, e do tempo de aplicação do *layout*, a tela representada pela Figura 8 irá surgir. Nessa tela, ao topo, existem três elementos, já presentes na tela inicial, com o adicional de um botão chamado de "*Refresh*". Ao clicar neste botão, a rede será reprocessada, agora levando praticamente apenas o tempo de aplicação do layout, dado que o calculo da função de similaridade não será necessário. Abaixo, ocupando a maior parte da tela, será renderizada a CSN, nesse espaço é possível arrastar para mudar o posicionamento da rede na tela, dar zoom e clicar nós nós representados pelas moléculas. É perceptível na rede que cada nó possui uma cor específica, a cor representa o grau do nó, sendo esse grau o número de conexões que ele faz com outros nós. As cores variam do azul ao vermelho. A cor azul representa os nós que possuem a menor

Figura 8 – Representação da tela de Visualização da CSN



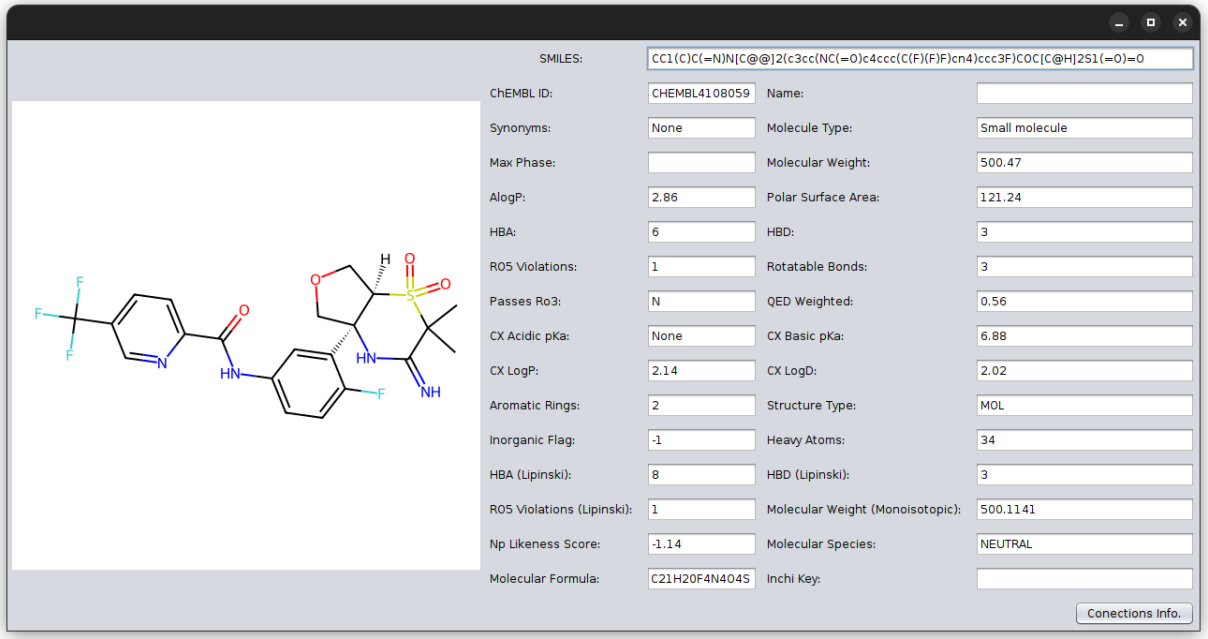
Fonte: Próprio Autor.

quantidade de conexões, sendo os que possuem uma cor azul vibrante aqueles que não possuem conexão nenhuma e a cor vermelha representa os nós que possuem a maior quantidade de conexões, sendo os que possuem uma cor vermelha mais vibrante aqueles que possuem o maior número de conexões.

4.3 Tela do *Card* de Informações de uma Molécula

Considerada uma característica inerente às CNS, a capacidade de carregar informações detalhadas sobre os nós, contribui para que elas sejam de grande contribuição na pesquisa pelo desenvolvimento de fármacos (MAGGIORA; BAJORATH, 2014). Por esse motivo também é possível clicar nos nós da rede gerada. Ao clicar em qualquer nó, será aberto uma tela contendo todas as informações que estavam disponíveis na base retirada do ChEMBL, essa tela é representada pela Figura 9. No lado esquerdo da tela é possível visualizar uma representação bi-dimensional da molécula representada pelo nó clicado, tal imagem foi gerada através de funções presentes na RDKit, o código para a geração também está disponível no repositório já mencionado. Do lado direito da tela é possível observar as informações que foram carregadas através do arquivo informacional, inserido na tela representada pela Figura 7. Ainda, no canto inferior

Figura 9 – Representação da tela do *Card* de Informações de uma Molécula



Fonte: Próprio Autor.

direito da tela, encontra-se o botão "*Conections Info.*", detalharemos sua funcionalidade na seção abaixo.

4.4 Tela das Informações de Conexão de uma Molécula

A tela das informações de conexão de uma molécula é representada pela Figura 10. Nela serão listadas todas as moléculas com as quais o nó da molécula clicada na Figura 8 está conectado. Essa lista contém as informações de "*SMILES*", "*CheMBL ID*" e da pontuação dada pela função de similaridade representada pelo campo "*Similarity*". A lista é ordenada de acordo com o valor da função de pontuação em ordem decrescente. Ao final da listagem de cada molécula tem-se o botão "*Molecule Info.*", ao clicar nesse botão ira surgir a tela representada pela Figura 9, permitindo assim comparar as moléculas conectadas ao nó.

Figura 10 – Representação da tela das Informações de Conexão de uma Molécula

Fonte: Próprio Autor.

5 CONCLUSÃO

Considerando a o objetivo principal deste trabalho, a criação de uma ferramente para a criação e visualização de redes de espaços químicos. Através da ferramenta aqui detalhada, é possível observar que:

a) É possível criar uma rede a partir da relação de similaridade entre moléculas;
b) É possível manipular essa rede e extrair informações inerentes aos nos do grafo;
c) É possível obter informações inerentes às moléculas existentes na rede;
d) É possível gerar diferentes topologias de redes através da seleção de diferentes *layouts* e diferentes *thresholds*.

É possível afirmar então que o objetivo principal foi atingido. Como continuação do trabalho, é possível ainda incluir novas funcionalidades ao projeto base, como por exemplo: opções para seleção de novas funções de similaridade, novos layouts para geração das redes, filtros para visualização mais precisa dos nos e sua relações, capacidade da escolha de cores, tanto do *background* quanto dos nos, capacidade de separação entre vizinhanças na rede, etc. Todas essas novas funcionalidades contribuiriam para uma melhor análise das relações entre as moléculas da rede como um todo. É possível também realizar trabalhos para a otimização dos cálculos da função de similaridade, ou também trabalhos para a comparação com outras alternativas existentes para geração de CSNs.

Por fim, é possível concluir que a ferramenta aqui desenvolvida é bastante promissora, embora ainda possa ser melhorada, ela já é capaz de prover a pesquisares a capacidade de gerar redes de moléculas com certa liberdade de ajustes e carga de informação, contribuindo assim para o desenvolvimento e reposicionamento de novos fármacos a partir da análise de similaridade entre moléculas.

REFERÊNCIAS

ARROWSMITH, John. A decade of change. **Nature Reviews Drug Discovery**, v. 11, n. 1, p. 17, 2012.

BAJORATH, Jürgen. **Chemoinformatics: concepts, methods, and tools for drug discovery**. [S.l.]: Springer Science & Business Media, 2008. v. 275.

BASTIAN, Mathieu; HEYMANN, Sebastien; JACOMY, Mathieu. **Gephi: An Open Source Software for Exploring and Manipulating Networks**. [S.l.: s.n.], 2009. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.

BERMAN, Helen M et al. The future of the protein data bank. **Biopolymers**, Wiley Online Library, v. 99, n. 3, p. 218–222, 2013.

BREDA, Aldo et al. Virtual Screening of Drugs: Score Functions, Docking, and Drug Design. **Current Computer-Aided Drug Design**, Bentham Science Publishers, v. 4, April, p. 265–272, 2008. DOI: 10.2174/157340908786786047.

CHAN, Margaret. **Ten Years in Public Health 2007-2017: Report by Dr Margaret Chan Director-General World Health Organization**. [S.l.]: World Health Organization, 2018.

CUMMINGS, Jeffrey; REIBER, Carl; KUMAR, Parvesh. The price of progress: Funding and financing Alzheimer's disease drug development. **Alzheimer's & Dementia: Translational Research & Clinical Interventions**, Elsevier, v. 4, p. 330–343, 2018.

DIAS, Raquel; FILGUEIRA DE AZEVEDO, Walter. **Molecular Docking Algorithms**. [S.l.]: Bentham Science Publishers, 2008. v. 9.

DOAK, Bradley C; NORTON, Raymond S; SCANLON, Martin J. The ways and means of fragment-based drug design. **Pharmacology & therapeutics**, Elsevier, v. 167, p. 28–37, 2016.

DUFFY, Bryan C et al. Early phase drug discovery: cheminformatics and computational techniques in identifying lead series. **Bioorganic & medicinal chemistry**, Elsevier, v. 20, n. 18, p. 5324–5342, 2012.

ERLANSON, Daniel A et al. Twenty years on: the impact of fragments on drug discovery. **Nature reviews Drug discovery**, Nature Publishing Group UK London, v. 15, n. 9, p. 605–619, 2016.

FERREIRA, Leonardo G; SANTOS, Raquel N et al. Molecular Docking and Structure-Based Drug Design Strategies. **Molecules**, Multidisciplinary Digital Publishing Institute, v. 20, n. 7, p. 13384–13421, 2015.

FERREIRA, Rafaela S; GLAUCIUS, Oliva; ANDRICOPULO, Adriano D. Integração das técnicas de triagem virtual e triagem biológica automatizada em alta escala: oportunidades e desafios em P&D de fármacos. **Química Nova**, SciELO Brasil, v. 34, p. 1770–1778, 2011.

HAGBERG, Aric; SWART, Pieter; S CHULT, Daniel. **Exploring network structure, dynamics, and function using NetworkX**. [S.l.], 2008.

HARRISON, Robert L. Introduction to monte carlo simulation. In: AMERICAN INSTITUTE OF PHYSICS, 1. AIP conference proceedings. [S.l.: s.n.], 2010. v. 1204, p. 17–21.

HILLISCH, Alexander; PINEDA, Luis Felipe; HILGENFELD, Rolf. Utility of homology models in the drug discovery process. **Drug discovery today**, Elsevier, v. 9, n. 15, p. 659–669, 2004.

KAPETANOVIC, IM2443682. Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. **Chemico-biological interactions**, Elsevier, v. 171, n. 2, p. 165–176, 2008.

KENNY, Peter W; SADOWSKI, Jens. Structure modification in chemical databases. **Chemoinformatics in drug discovery**, Wiley Online Library, p. 271–285, 2005.

LANDRUM, G.; RDK, contributors. **rdkit/rdkit: 2022_03_5 (Q1 2022) Release**. [S.l.: s.n.], 2022. https://github.com/rdkit/rdkit.

LEELANANDA, Sumudu P; LINDERT, Steffen. Computational methods in drug discovery. **Beilstein journal of organic chemistry**, Beilstein-Institut, v. 12, n. 1, p. 2694–2718, 2016.

LEISINGER, Klaus Michael; GARABEDIAN, Laura Faden; WAGNER, Anita Katharina. Improving access to medicines in low and middle income countries: corporate responsibilities in context. **Southern med review**, BioMed Central, v. 5, n. 2, p. 3, 2012.

LIU, Jie; WANG, Renxiao. Classification of current scoring functions. **Journal of chemical information and modeling**, ACS Publications, v. 55, n. 3, p. 475–482, 2015.

MA, Xiao H et al. Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. **Combinatorial chemistry & high throughput screening**, Bentham Science Publishers, v. 12, n. 4, p. 344–357, 2009.

MAGGIORA, Gerald M; BAJORATH, Jürgen. Chemical space networks: a powerful new paradigm for the description of chemical space. **Journal of computer-aided molecular design**, Springer, v. 28, p. 795–802, 2014.

MAGGIORA, Gerald M; SHANMUGASUNDARAM, Veerabahu. Molecular similarity measures. **Chemoinformatics: concepts, methods, and tools for drug discovery**, Springer, p. 1–50, 2004.

_____._____. **Chemoinformatics and computational chemical biology**, Springer, p. 39–100, 2011.

MAIA, Eduardo Henrique B et al. Molecular Architect: A User-Friendly Workflow for Virtual Screening. **ACS Omega**, American Chemical Society, v. 5, p. 6628–6640, 2020. DOI: 10.1021/acsomega.9b04403.

MALDONADO, Ana G. et al. **Molecular similarity and diversity in chemoinformatics: From theory to applications**. en. v. 10. [S.l.]: Springer Science e Business Media LLC, fev. 2006. P. 39–79.

MARTIN, Yvonne C et al. Glossary of terms used in computational drug design, part II (IUPAC Recommendations 2015). **Pure and Applied Chemistry**, De Gruyter, v. 88, n. 3, p. 239–264, 2016.

MENDEZ, David et al. ChEMBL: towards direct deposition of bioassay data. **Nucleic acids research**, Oxford University Press, v. 47, n. D1, p. d930–d940, 2019.

MORRIS, Garrett M.; LIM-WILBY, Marguerita. **Molecular Docking**. [S.l.]: Humana Press, 2008. P. 365–382.

MUNIZ, Heloisa dos Santos. **Métodos híbridos em docagem molecular: implementação, validação e aplicação**. 2018. Tese (Doutorado) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos.

OGLIC, Dino et al. Active search for computer-aided drug design. **Molecular informatics**, Wiley Online Library, v. 37, n. 1-2, p. 1700130, 2018.

OLIVEIRA, Tiago Alves de et al. Virtual Screening Algorithms in Drug Discovery: A Review Focused on Machine and Deep Learning Methods. **Drugs and Drug Candidates**, v. 2, n. 2, p. 311–334, 2023. ISSN 2813-2998. DOI: 10.3390/ddc2020017. Disponível em: <https://www.mdpi.com/2813-2998/2/2/17>.

RECANATINI, Maurizio; CABRELLE, Chiara. Drug research meets network science: where are we? **Journal of Medicinal Chemistry**, ACS Publications, v. 63, n. 16, p. 8653–8666, 2020.

RUIZ-CARMONA, Sergio et al. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. **PLoS computational biology**, Public Library of Science San Francisco, USA, v. 10, n. 4, e1003571, 2014.

SCALFANI, Vincent F; PATEL, Vishank D; FERNANDEZ, Avery M. Visualizing chemical space networks with RDKit and NetworkX. **Journal of Cheminformatics**, Springer, v. 14, n. 1, p. 87, 2022.

SNEADER, Walter. **A History of Pharmaceutical Sciences: A Collection of Historical Articles**. Chichester, UK: John Wiley & Sons, Ltd, 2010.

SOUSA, Sergio Filipe; FERNANDES, Pedro Alexandrino; RAMOS, Maria Joao. Protein–ligand docking: current status and future challenges. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 65, n. 1, p. 15–26, 2006.

STEVENS, Hilde; HUYS, Isabelle. Innovative approaches to increase access to medicines in developing countries. **Frontiers in medicine**, Frontiers Media SA, v. 4, p. 218, 2017.

TAYARANI, Ali et al. Artificial neural networks analysis used to evaluate the molecular interactions between selected drugs and human cyclooxygenase2 receptor. **Iranian journal of basic medical sciences**, Mashhad University of Medical Sciences, v. 16, n. 11, p. 1196, 2013.

VOGT, Martin et al. Lessons learned from the design of chemical space networks and opportunities for new applications. **Journal of computer-aided molecular design**, Springer, v. 30, p. 191–208, 2016.

ZHANG, Bijun et al. Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures. **Journal of computer-aided molecular design**, Springer, v. 29, p. 937–950, 2015.

ZHANG, Gang et al. Virtual screening of small molecular inhibitors against DprE1. **Molecules**, MDPI, v. 23, n. 3, p. 524, 2018.

ZHAO, Hui; GUO, Zhongwu. Medicinal chemistry strategies in follow-on drug discovery. **Chemical Biology & Drug Design**, v. 94, n. 5, p. 1922–1931, 2019. ISSN 1747-0277. DOI: 10.1111/cbdd.13564.