UNIVERSIDADE DE SÃO PAULO
Faculdade de Ciências Farmacêuticas
Programa de Pós-Graduação em Fármaco e Medicamentos
Área de Insumos Farmacêuticos

Reposicionamento de fármacos para o tratamento da doença de Chagas:
desenvolvimento de modelos computacionais baseados em ligantes ativos
fenotipicamente contra o *Trypanosoma cruzi* e na inibição da diidroorotato
desidrogenase do parasita

Marcos Antonio Ferreira Junior

Dissertação para obtenção do Título de Mestre

Orientadora: Prof.^a Dr.^a Elizabeth Igne Ferreira
Co-orientador: Prof. Dr. Carlos Mauricio Rabello de Sant'Anna

São Paulo
2020

UNIVERSIDADE DE SÃO PAULO
Faculdade de Ciências Farmacêuticas
Programa de Pós-Graduação em Fármaco e Medicamentos
Área de Insumos Farmacêuticos

Reposicionamento de fármacos para o tratamento da doença de Chagas:
desenvolvimento de modelos computacionais baseados em ligantes ativos
fenotipicamente contra o *Trypanosoma cruzi* e na inibição da diidroorotato
desidrogenase do parasita

Marcos Antonio Ferreira Junior

Versão corrigida da Dissertação conforme resolução CoPGr 6018.

Dissertação para obtenção do Título de Mestre

Orientador: Prof.^a Dr.^a Elizabeth Igne Ferreira
Co-orientador: Prof. Dr. Carlos Mauricio Rabello de Sant'Anna

São Paulo
2020

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Ficha Catalográfica elaborada eletronicamente pelo autor, utilizando o programa desenvolvido pela Seção Técnica de Informática do ICMC/USP e adaptado para a Divisão de Biblioteca e Documentação do Conjunto das Químicas da USP

Bibliotecária responsável pela orientação de catalogação da publicação:
Marlene Aparecida Vieira - CRB - 8/5562

F383r Ferreira Junior, Marcos Antonio
Reposicionamento de fármacos para o tratamento da doença de Chagas: desenvolvimento de modelos computacionais baseados em ligantes ativos fenotipicamente contra o Trypanosoma cruzi e na inibição da diidroorotato desidrogenase do parasita / Marcos Antonio Ferreira Junior. - São Paulo, 2020. 199 p.

Dissertação (mestrado) - Faculdade de Ciências Farmacêuticas da Universidade de São Paulo. Departamento de Farmácia.
Orientador: Ferreira, Elizabeth Igne
Coorientador: Sant Anna, Carlos Maurício Rabello

1. Trypanosoma cruzi. 2. Doença de Chagas. 3. Reposicionamento de fármacos. 4. Aprendizado de máquinas. 5. Diidroorotato desidrogenase. I. T. II. Ferreira, Elizabeth Igne, orientador. III. Sant Anna, Carlos Maurício Rabello, coorientador.

Marcos Antonio Ferreira Junior

Reposicionamento de fármacos para o tratamento da doença de Chagas:
desenvolvimento de modelos computacionais baseados em ligantes ativos
fenotipicamente contra o *Trypanosoma cruzi* e na inibição da diidroorotato
desidrogenase do parasita

Comissão Julgadora
da
Dissertação para obtenção do Título de Mestre

Prof.^a Dr.^a Elizabeth Igne Ferreira
Orientadora/ Presidente

1o. examinador

2o. examinador

3o. examinador

4o. examinador

São Paulo, _____ de _____ de 20____.

AGRADECIMENTOS

À professora Elizabeth Igne Ferreira pela oportunidade de ser seu aluno, pela enorme paciência e dedicação durante esta jornada de aprendizado, por todas as discussões científicas e pessoais, pela amizade, por sempre instigar o meu desenvolvimento e curiosidade científica e por sempre apoiar todas as minhas iniciativas.

Ao professor Carlos Maurício Sant'Anna pela oportunidade de ser seu aluno, pelos ensinamentos, paciência, disponibilidade, pelas nossas agradáveis conversas pessoais e profissionais e por todos os esforços que tornaram este trabalho possível.

Aos meus pais, Marcos e Luzia, pelos esforços e sacrifícios que realizaram para o meu desenvolvimento pessoal e profissional, e ao meu irmão, Rafael, por ser sempre um dos meus grandes motivadores.

À Mariane, pelo companheirismo, atenção, carinho, apoio e compreensão durante todos estes anos de formação. À Macarena, por deixar os nossos dias mais leves.

Aos amigos Natanael, Hatylas, Kátia, Isabela, Gabriela, Jéssica M., Jéssica E., Larissa, Cristiano, Alessandra, Lisandra, Carlos, Bruno B., Elza, Ediliz, Romulo, Amanda, Renata, Bruno T., Fernando, Carla, Jacque, Priscila, Gabriel, Mariah, Luan, Renato, Rafael, Eloísa, Leandro, Juan, Guilherme, Thiago, Thiego, Josiel, Danilo C., Dominique, Gênesis, Aline, Bruno e Danilo M., pelo apoio, pela convivência de todos os dias e por me ajudarem a ver as dificuldades e oportunidades através de novos pontos de vista.

À Carolina Borsoi Moraes, pela colaboração científica e pelos esforços para a realização do ensaio experimental de inibição do *T. cruzi*.

Aos pesquisadores, Alessandro Nascimento, Lucio Freitas, Alan Talevi, Alan Fairlamb, Antonia Tavares, Humberto Gomes, Julio Martin e Bruno Neves, pelas discussões e contribuições científicas.

Aos amigos do LAPEN, Jeanine, Fredson, Luccas, Renan, Rodrigo, Soraya, Diego, Charles, Marina, Cynthia, João e do bloco 13, Elis, Fernando, Nuno, Maurício, Glaúcio, Ricardo, Stanley, Michael, Alfredo e Thais pelas conversas e pelo auxílio nas diferentes etapas da formação.

Ao David, ao Alexandre e à Kelma, pela paciência e pela ajuda imensurável nas diferentes etapas desta formação.

À Universidade de São Paulo e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela infraestrutura necessária para o desenvolvimento deste trabalho.

RESUMO

FERREIRA JUNIOR, M. A. **Reposicionamento de fármacos para o tratamento da doença de Chagas: desenvolvimento de modelos computacionais baseados em ligantes ativos fenotipicamente contra o *Trypanosoma cruzi* e na inibição da diidroorotato desidrogenase do parasita.** 2020. 191f. Dissertação (Mestrado) – Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, São Paulo, 2020.

A doença de Chagas é uma enfermidade crônica causada pelo parasita *Trypanosoma cruzi* (*T. cruzi*), que afeta aproximadamente oito milhões de pessoas mundialmente e carece de tratamentos seguros e eficazes. Neste trabalho, modelos computacionais baseados nas estratégias LBDD (*ligand-based drug design*) e SBDD (*structure-based drug design*) foram desenvolvidos para auxiliar no reposicionamento de fármacos e candidatos a fármaco para o tratamento da doença de Chagas. O reposicionamento, *i.e.*: a identificação de novos usos terapêuticos para moléculas conhecidas, é uma abordagem promissora para reduzir os investimentos, os riscos de falha e o tempo necessário para desenvolver um novo medicamento. Na estratégia LBDD, os algoritmos de aprendizado de máquina floresta aleatória, *Naive Bayes*, máquinas de vetores de suporte e redes neurais probabilísticas foram explorados para o desenvolvimento de modelos baseados em dados fenotípicos de inibição do *T. cruzi*. O NB – FeatMorgan, modelo de melhor performance, exibiu uma alta capacidade preditiva nas validações iniciais (validação cruzada quádrupla e validação externa), que foi confirmada com a segunda validação externa utilizando uma nova série teste (AUC ROC = 0,85, EF1 = 14,3 e TCB = 80%). Dos 269 fármacos selecionados na triagem virtual com o NB – FeatMorgan, 19 (7,1%) já possuíam atividade tripanocida reportada na literatura ($IC_{50} < 5 \mu M$ e $SI > 10$). Ainda, 17 fármacos foram obtidos e testados experimentalmente na inibição do *T. cruzi*. Notavelmente, a desloratadina e a flibanserina exibiram eficácia promissora na inibição do parasita ($A_{max} = 91\%$ e 92% , respectivamente). Na estratégia SBDD, desenvolveu-se um fluxo de triagem virtual para a seleção de inibidores da diidroorotato desidrogenase do *T. cruzi* (TcDHODH). Estudos de reancoragem e triagem virtual retrospectiva evidenciaram que a função de pontuação ASP apresentou alta performance na predição do modo de ligação de compostos na enzima e na filtragem inicial de moléculas completamente inativas. Entretanto, a sua capacidade preditiva foi baixa/intermediária na identificação de moléculas potentes na inibição da TcDHODH. Deste modo, um modelo de aprendizado de máquinas baseado no algoritmo *Naive Bayes* foi desenvolvido utilizando derivados do orotato testados na enzima com o objetivo de refinar os resultados da triagem virtual provenientes da função ASP. O modelo empregou a entalpia de interação entre o ligante e o alvo, calculada pelo método semi-empírico PM7, como descritor, assim como a lipofilicidade do ligante e a variação de energia associada às perdas entrópicas resultantes das restrições conformacionais do ligante ao se complexar com a proteína. O NB – TcDHODH – 8, modelo de melhor performance, apresentou alta capacidade preditiva na validação cruzada quádrupla, a qual foi confirmada na validação externa com a série teste (AUC ROC = 0,79 e AUC BEDROC = 0,85). Os modelos desenvolvidos nas estratégias LBDD e SBDD deste trabalho poderão auxiliar no desenvolvimento novas terapias seguras e eficazes para o tratamento da doença de Chagas. Os compostos selecionados poderão ser explorados na terapêutica da doença ou como pontos de partida para novos projetos de P&D de fármacos anti-*T. cruzi*. Os modelos computacionais poderão, ainda, ser utilizados em novas triagens virtuais a partir de bibliotecas de interesse.

Palavras-chaves: Doença de Chagas. *Trypanosoma cruzi*. Reposicionamento de fármacos. Aprendizado de máquina. Diidroorotato desidrogenase. Ancoragem molecular. Método semi-empírico. Entalpia de interação.

ABSTRACT

FERREIRA JUNIOR, M. A. **Drug repurposing for the treatment of Chagas disease: development of computational models based on compounds phenotypically active against *Trypanosoma cruzi* and on the inhibition of the parasite dihydroorotate dehydrogenase**. 2020. 191f. Dissertação (Mestrado) – Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, São Paulo, 2020.

Chagas disease is a chronic condition caused by the parasite *Trypanosoma cruzi* (*T. cruzi*), which affects ca. 8 million people around the world and lacks safe and efficacious treatments. In this work, computational models based on the LBDD (ligand-based drug discovery) and on the SBDD (structure-based drug discovery) strategies were developed with the aim of being employed in the repurposing of drugs and drug candidates for the treatment of Chagas disease. Drug repurposing, *i.e.*: the identification of novel therapeutic uses for known compounds, is a promising approach to reduce drug R&D costs, the high attrition ratios and to expedite the development timelines of novel therapies. In the LBDD approach, the machine learning algorithms random forest, Naive Bayes, support vector machine and probabilistic neural networks were used to develop computational models based on phenotypic data of *in vitro* *T. cruzi* inhibition. The NB – FeatMorgan model, which displayed the best performance, showed a high predictive power in the initial validations (5-fold cross-validation and external validation), which was confirmed at a second external validation with another chemical series (AUC ROC = 0,85 and EF1 = 14,3). Interestingly, 19 out of 269 drugs selected as potentially active during the virtual screening campaign with this model (7.1%) had already been described as active against the parasite in the literature ($IC_{50} < 5 \mu M$ and $SI > 10$). 17 drugs have been experimentally tested for the *T. cruzi* inhibition. Notably, desloratadine and flibanserine exhibited a high efficacy against the parasite ($A_{max} = 91\%$ e 92% , respectively). In the SBDD approach, a workflow for identifying *T. cruzi* dihydroorotate dehydrogenase (TcDHODH) inhibitors was developed. Redocking and retrospective virtual screening studies evidenced that the ASP scoring function exhibited a high performance for predicting the binding mode of compounds in the TcDHODH and for eliminating completely inactive molecules. Notwithstanding, the ASP function displayed a low/intermediate capacity to identify compounds with a high affinity at TcDHODH. In this sense, a machine learning model was developed, based on orotic acid derivatives tested against the enzyme, with the objective of being employed to reclassify the results from an initial virtual screening using the ASP function. The descriptors explored in this work were the ligand-protein interaction enthalpy, the ligand lipophilicity and the energy variation associated with entropic loss due to the ligand complexation with the target protein. The best reclassification model, NB – TcDHODH – 8, exhibited a high performance in the initial 5-fold cross validation, which was confirmed in the external validation using the test set (AUC ROC = 0,79 and AUC BEDROC = 0,85). The SBDD and LBDD models developed in this work can contribute to the development of novel safe and efficacious therapies for the treatment of Chagas disease. The compounds selected in the virtual screening campaigns could be used for the treatment of Chagas disease or as starting points for novel anti-*T. cruzi* drug R&D initiatives. The validated models could also be used for novel virtual screening campaigns.

Keywords: Chagas disease. *Trypanosoma cruzi*. Drug repurposing. Machine learning. Dihydroorotate dehydrogenase. Molecular docking. Semi-empirical method. Interaction enthalpy.

SUMÁRIO

CAPÍTULO 1	13
1 INTRODUÇÃO	14
1.1 A DOENÇA DE CHAGAS	14
1.2 PESQUISA E DESENVOLVIMENTO DE NOVOS FÁRMACOS PARA O TRATAMENTO DA DOENÇA DE CHAGAS	16
1.3 O REPOSICIONAMENTO DE FÁRMACOS COMO UMA ESTRATÉGIA PARA ACELERAR A P&D DE NOVOS TRATAMENTOS PARA A DOENÇA DE CHAGAS	19
1.4 USO DE MODELOS COMPUTACIONAIS PARA IDENTIFICAR COMPOSTOS A SEREM REPOSICIONADOS PARA A DOENÇA DE CHAGAS	23
CAPÍTULO 2	25
2 OBJETIVOS E JUSTIFICATIVA	26
2.1 OBJETIVOS ESPECÍFICOS	27
CAPÍTULO 3	29
3 ABORDAGEM LBDD: DESENVOLVIMENTO DE MODELOS BASEADOS NA ESTRUTURA DE COMPOSTOS ATIVOS FENOTIPICAMENTE CONTRA O <i>TRYPANOSOMA CRUZI</i>	30
3.1 INTRODUÇÃO	30
3.2 MATERIAL E MÉTODOS	36
3.3 RESULTADOS	48
3.4 DISCUSSÃO	79
3.5 CONCLUSÕES	89
CAPÍTULO 4	93
4 ABORDAGEM SBDD: DESENVOLVIMENTO DE MODELOS BASEADOS NA INIBIÇÃO DA ENZIMA DIIDROOROTATO DESIDROGENASE DO <i>TRYPANOSOMA CRUZI</i>	94

4.1	INTRODUÇÃO	94
4.2	MATERIAL E MÉTODOS.....	102
4.3	RESULTADOS.....	113
4.4	DISCUSSÃO.....	143
4.5	CONCLUSÕES.....	149
CAPÍTULO 5.....		151
5	CONCLUSÕES E PERSPECTIVAS	152
REFERÊNCIAS.....		155
APÊNDICE A - COMPOSIÇÃO DOS BANCOS DE DADOS UTILIZADOS NO DESENVOLVIMENTO E VALIDAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA DA ABORDAGEM LBDD		175
APÊNDICE B – RESULTADOS DA VALIDAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA DA ABORDAGEM LBDD		178
APÊNDICE C – LISTA DE FÁRMACOS APROVADOS SELECIONADOS PELO MODELO NB – FEATMORGAN NA ABORDAGEM LDBB		181
APÊNDICE D – LISTA DE FÁRMACOS EM DESENVOLVIMENTO CLÍNICO SELECIONADOS PELO MODELO NB – FEATMORGAN NA ABORDAGEM LDBB		187
APÊNDICE E – LISTA DE METABÓLITOS DE FÁRMACOS SELECIONADOS PELO MODELO NB – FEATMORGAN NA ABORDAGEM LDBB		194
APÊNDICE F – RESULTADOS DA VALIDAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA DA ABORDAGEM SBDD		199

CAPÍTULO 1

1 INTRODUÇÃO

1.1 A DOENÇA DE CHAGAS

A doença de Chagas ou tripanossomíase americana é uma enfermidade crônica causada pelo parasita *Trypanosoma cruzi* (*T. cruzi*) e classificada como uma doença tropical negligenciada (DTN) pela Organização Mundial da Saúde (OMS). Segundo a OMS, as DTNs são um conjunto de doenças que afetam principalmente comunidades pobres e que causam consequências devastadoras para mais de 1 bilhão de pessoas, em aspectos sociais, econômicos e de saúde (OMS, 2021). As DTNs se caracterizam pela ausência de terapias seguras e eficazes e pela carência no investimento em programas de pesquisa e desenvolvimento (P&D) de novos medicamentos para o seu tratamento (MATHERS; EZZATI; LOPEZ, 2007).

O *T. cruzi* pode ser transmitido através do contato de mucosas ou lesões na pele com as fezes de insetos hematófagos da subfamília *Triatominae* (triatomíneos), popularmente conhecidos como “barbeiro”, que estejam infectados com o parasita (DIAS et al., 2016; SANGENIS et al., 2016). Além disso, a transmissão pode ocorrer por transfusão de sangue, pela via congênita, transplante de órgãos ou ingestão de alimentos contaminados com o *T. cruzi* (PEREIRA; NAVARRO, 2013). A OMS (2020) estima que oito milhões de pessoas estejam infectadas pelo *T. cruzi* ao redor do mundo e que mais de 10.000 mortes ocorram ao ano devido a esta doença. Embora exista maior concentração de portadores do parasita em países da América Latina, dentre eles o Brasil, a doença de Chagas se tornou um problema de saúde global devido à maior incidência de casos em regiões não endêmicas da Europa, da Ásia e da América do Norte. Esta mudança no perfil epidemiológico está associada ao fenômeno da globalização, que possibilitou o fluxo migratório de pacientes contaminados, e às vias de transmissão independentes do inseto vetor (LIU; ZHOU, 2015; OMS, 2020).

O ciclo de vida do *T. cruzi* em humanos compreende a forma tripomastigota, infecciosa e majoritariamente presente na corrente sanguínea, e a forma amastigota, proliferativa e de ocorrência no interior das células. Os amastigotas intracelulares podem ser transformados em tripomastigotas após o término do ciclo proliferativo e serem liberados na corrente sanguínea para infectar novas células, reiniciando o ciclo de infecção em humanos. Na transmissão por insetos vetores, a infecção ocorre

através de tripomastigotas metacíclicos, os quais invadem as células de humanos e transformam-se em amastigotas (ONYEKWELU, 2019). Clinicamente, a doença pode ser dividida em uma fase aguda, com alta carga parasitária na corrente sanguínea, e uma fase crônica, caracterizada pela presença de parasitas em diferentes tecidos. A fase aguda, com duração de 4 a 8 semanas, tende a se desenvolver sem manifestações clínicas ou somente com a presença de alguns sintomas brandos, como mal-estar, febre, edema, hepatomegalia e esplenomegalia, o que dificulta o diagnóstico da doença. Entretanto, alguns pacientes podem desenvolver sintomas mais graves e potencialmente fatais neste estágio, como meningoencefalite ou miocardite. Após a fase aguda, as pessoas infectadas continuam sendo portadoras do *T. cruzi* durante toda a vida, período denominado de fase crônica, porém a maioria apresenta apenas uma forma assintomática da doença. Aproximadamente 30 a 40% dos infectados desenvolvem sintomas graves associados a alterações no coração e no trato gastrointestinal. As manifestações envolvendo o coração incluem parada cardíaca, arritmias ventriculares, insuficiência cardíaca, tromboembolismo e morte súbita. Por sua vez, as manifestações gastrointestinais incluem a dilatação do cólon, do esôfago, do intestino delgado ou da vesícula biliar e alterações na motilidade de diversas porções do sistema, sintomas que são estão associados a danos no sistema nervoso entérico (BERN, 2015; MATSUDA; MILLER; EVORAL, 2009; NUNES et al., 2013).

Os únicos fármacos disponíveis para o tratamento da doença de Chagas, benznidazol e nifurtimox (Figura 1), foram desenvolvidos há mais de 40 anos (lançados em 1972, pela Roche, e em 1967, pela Bayer, respectivamente). Apesar de ambas as moléculas apresentarem características que restringem seu uso na clínica, como eficácia limitada e baixa segurança, elas continuam sendo empregadas atualmente devido à ausência de terapias modernas (MORENO et al., 2010). O benznidazol e o nifurtimox são efetivos no tratamento de infecções agudas e daquelas reativadas devido à imunossupressão do indivíduo, porém apresentam eficácia limitada em pacientes na fase crônica (OMS, 2020). Os efeitos adversos observados durante o tratamento com o nifurtimox incluem perda de peso, alterações psicológicas, náusea, vômito, cólica, diarreia e excitabilidade ou sonolência. No caso do benznidazol, os efeitos incluem dermatite, hipersensibilidade cutânea, edema generalizado, febre, dor muscular e articular, depressão da medula óssea, trombocitopenia ou agranulocitose. Efeitos tumorigênicos, carcinogênicos e

mutagênicos foram observados em estudos de segurança para ambos os compostos (CASTRO; DE MECCA; BARTEL, 2006). No Brasil, somente o benznidazol possui registro como medicamento na Agência Nacional de Vigilância Sanitária - ANVISA (DIAS et al., 2016). A ausência de fármacos seguros e eficazes para o combate ao *T. cruzi* evidencia a necessidade de desenvolvimento de novas terapias para o tratamento da doença de Chagas.

Figura 1 – Estrutura química dos fármacos benznidazol e nifurtimox.

Fonte: Produção do autor.

1.2 PESQUISA E DESENVOLVIMENTO DE NOVOS FÁRMACOS PARA O TRATAMENTO DA DOENÇA DE CHAGAS

A P&D de novos fármacos representa um processo longo, complexo e que exige altos investimentos. O tempo médio para que um novo medicamento chegue ao mercado pode variar entre 12 a 15 anos e necessitar de investimentos entre USD 1 bilhão e USD 2,56 bilhões (DIMASI; GRABOWSKI; HANSEN, 2016; HUGHES et al., 2011; WOUTERS; MCKEE; LUYTEN, 2020). O processo de P&D pode ser dividido nas fases pré-clínica e clínica. A fase pré-clínica compreende a identificação e a validação de novos alvos terapêuticos, assim como a identificação de compostos ativos no alvo de interesse e a otimização das propriedades dessas moléculas através de modificações moleculares. Nesta etapa, os compostos são avaliados em modelos *in vitro* e estudos *in vivo*, de forma iterativa, de modo a identificar um candidato a fármaco com perfis toxicológico, farmacocinético e de eficácia adequados para o uso em humanos. Na fase clínica, o candidato é avaliado em estudos de segurança, farmacocinética e eficácia em humanos (estudos clínicos fase I, II e III) para comprovar o seu potencial benefício para o tratamento da doença de interesse. Além disso, estudos clínicos fase IV podem ser realizados após a aprovação para, por exemplo, a confirmação da eficácia em um número maior de pacientes ou a expansão de uso

para novas indicações terapêuticas (HUGHES et al., 2011; STROVEL et al., 2016). A alta taxa de falhas dos projetos durante o desenvolvimento representa uma das principais causas deste cenário e pode ser devido à ausência de eficácia terapêutica, à farmacocinética inadequada ou à alta toxicidade dos candidatos durante os estudos pré-clínicos ou clínicos (WARING et al., 2015). O número de projetos de pesquisa focados na terapêutica de DTNs, em particular para a doença Chagas, é bastante reduzido comparativamente a outras áreas terapêuticas. Estudos que avaliaram o lançamento mundial de novas moléculas durante 30 anos (1975 a 2004) demonstraram que somente 10 dos 1556 novos princípios ativos lançados no período, *i.e.*: menos de 1%, foram desenvolvidos especificamente para uma DTN (CHIRAC; TORREELE, 2006; TROUILLER et al., 2002). O nível de investimento em doenças das populações negligenciadas está associado à baixa expectativa de retorno financeiro para estes projetos (JOHNSTON; FORD; TAYLOR, 2014).

Diversas iniciativas de colaboração entre governos, organizações filantrópicas, academia e empresas privadas têm surgido com o objetivo de reverter este cenário e fomentar o desenvolvimento de novos fármacos para as DTNs, como por exemplo a *Drugs for Neglected Diseases Initiative* (DNDi). Particularmente, o DNDi considera a doença de Chagas como uma patologia prioritária em seu portfólio e elaborou proposta dos critérios essenciais, que deveriam ser apresentados por um novo fármaco para a doença, denominado o perfil do produto alvo ou TPP (do inglês, *target product profile*) (CHATELAIN; IOSET, 2011). A existência de um TPP definido permite direcionar os esforços de pesquisa na área, essencial frente aos limitados recursos disponíveis. As características de uma terapia ideal para Chagas incluem: i) eficácia nas fases aguda e crônica da doença, ii) eficácia e segurança superiores ao benznidazol nas doses terapêuticas, iii) ausência de genotoxicidade, teratogenicidade ou efeito pró-arrítmico, iv) ausência de potencial de interação medicamentosa com outros fármacos, v) administração oral, vi) regime posológico menor do que 7 dias para a fase aguda e do que 30 dias para a fase crônica e vii) estabilidade por 5 anos na zona climática IV, quente e úmida (DNDi, 2020). Considerando o TPP para a doença de Chagas, o DNDi elaborou também uma proposta de critérios para a identificação de novos compostos ativos (*hits*) e sua progressão a moléculas líderes durante o processo de P&D (CHATELAIN, 2015). Para que um novo hit seja considerado aceito pelo DNDi, o composto deve apresentar alta potência (concentração de composto capaz de inibir em 50% a proliferação inferior a 5 μ M) e

eficácia (máxima atividade antiparasitária superior a 90%) na inibição do *T. cruzi* e índice de seletividade frente à célula hospedeira superior a 10.

Notavelmente, diversos avanços e conquistas foram observados na área da doença de Chagas nos últimos anos, como o surgimento de novos métodos experimentais *in vitro* e *in vivo* para avaliar a eficácia de compostos e a realização de estudos clínicos com novos compostos – o posaconazol e o pró-fármaco do ravuconazol (E1224). Entretanto, ainda há diversos desafios a serem superados, como o desenvolvimento de métodos diagnósticos capazes de identificar a cura da doença, o melhor entendimento da translação de experimentos pré-clínicos para a patologia em humanos e a necessidade de identificação de novos candidatos a fármaco para o tratamento da doença (CHATELAIN, 2015). A dificuldade na P&D de novos antichagásicos pode ser evidenciada pelo reduzido número de projetos no portfólio do DNDi para o tratamento da doença de Chagas (DNDi, 2021). Somente uma nova entidade química, o fexinidazol, ativa preferencialmente em doença do sono, ou tripanossomíase africana, e novas formulações e/ou regimes posológicos do benznidazol estão em desenvolvimento clínico. Os demais projetos encontram-se ainda na fase pré-clínica, seja na etapa de identificação de novos compostos ativos ou na otimização de moléculas líderes.

A P&D de fármacos para a doença de Chagas pode ser baseado em duas abordagens diferentes para avaliar a atividade de compostos contra o *T. cruzi*: a modulação de alvos terapêuticos específicos, que sejam promissores no combate ao *T. cruzi* e a inibição do parasita completo em ensaios fenotípicos (GILBERT, 2013).

A estratégia de desenvolvimento baseada em alvos terapêuticos tem como fundamento a premissa de que a modulação de biomacromoléculas específicas do organismo pode afetar a progressão da doença. Neste caso, o entendimento dos processos fisiológicos e dos mecanismos moleculares envolvidos no desenvolvimento da patologia de interesse é essencial para substanciar a escolha racional de potenciais alvos terapêuticos. A escolha de um alvo promissor é essencial para reduzir os altos índices de falha na clínica devido à ausência de eficácia terapêutica (REAUME, 2011). A grande expansão desta estratégia de P&D é fruto dos avanços científicos em diversas áreas da química e da biologia, como a bioquímica, a biologia molecular, o sequenciamento genético e a cristalografia de proteínas (PALMER, 2016; TODD; ANDERSON; GROUNDWATER, 2009). No caso da doença de Chagas, diversos alvos terapêuticos promissores têm sido estudados, como a cruzipaina, a

tripanotona redutase, a anidrase carbônica, a esqualeno sintase, a *N*-miristoiltransferase e a diidroorotato desidrogenase. Entretanto, ainda não há compostos capazes de modular estes alvos que progrediram para a fase de pesquisa clínica e que foram eficazes no tratamento da doença de Chagas (CHELESKI et al., 2010; GILBERT, 2002; VILLALTA; RACHAKONDA, 2019).

A estratégia de desenvolvimento baseada em ensaios fenotípicos avalia os efeitos biológicos gerados por compostos mesmo na ausência de conhecimentos prévios sobre os mecanismos de ação. Deste modo, esta abordagem permite a exploração de novas patologias para as quais ainda não existam alvos terapêuticos validados, de doenças complexas e multifatoriais para as quais mais de um alvo esteja envolvido ou para aquelas cuja fisiopatologia ainda não seja completamente estudada (PALMER, 2016). Swinney e Anthony (2011) observaram que projetos baseados em métodos fenotípicos foram os responsáveis pelo maior número de lançamentos de fármacos com novos mecanismos de ação nos Estados Unidos entre os anos de 1999 e 2008. Os projetos baseados em métodos fenotípicos dependem da disponibilidade de ensaios translacionais, cujos resultados sejam capazes de prever a resposta clínica em humanos. Os resultados observados em ensaios fenotípicos *in vitro* para doenças infecciosas, como infecções bacterianas, parasitárias e virais, apresentam alta translação para os efeitos em modelos pré-clínicos *in vivo* e na eficácia em humanos durante o desenvolvimento clínico (MOFFAT et al., 2017). A maior translação dos ensaios justifica o grande número de agentes infecciosos disponíveis no mercado cujos pontos de partida foram ensaios fenotípicos (SWINNEY; ANTHONY, 2011). Ademais, ensaios fenotípicos foram amplamente utilizados no desenvolvimento do fexinidazol e do SCYX-7158 para o tratamento da doença do sono, causada pelo parasita *Trypanosoma brucei*, comprovando a efetividade da estratégia para doenças parasitárias (GILBERT, 2013).

1.3 O REPOSICIONAMENTO DE FÁRMACOS COMO UMA ESTRATÉGIA PARA ACELERAR A P&D DE NOVOS TRATAMENTOS PARA A DOENÇA DE CHAGAS

O reposicionamento de fármacos consiste na identificação de novos usos terapêuticos para compostos que foram previamente desenvolvidos para outras indicações (RUDRAPAL; KHAIRNAR; JADHAV, 2020). Esta abordagem também

pode ser denominada reproposta, resgate ou redescobrimto. Os compostos reposicionados podem ser fármacos aprovados, candidatos a fármaco em desenvolvimento ou mesmo moléculas que tenham falhado para a sua indicação inicial (JIN; WONG, 2014).

A estratégia de reposicionamento pode resultar em diversas vantagens no processo de P&D, quando comparada ao desenvolvimento de novos compostos a partir de séries químicas para as quais pouca ou nenhuma informação biológica está disponível. As vantagens incluem menores riscos de falha, investimentos de desenvolvimento e tempo necessário para chegar ao mercado (ASHBURN; THOR, 2004; KLUG; GELB; POLLASTRI, 2016; PUSHPAKOM et al., 2019). Alguns autores sugerem que os investimentos necessários podem ser até 40% menores para compostos reposicionados (CHONG; SULIVAN, 2007) e que o tempo de P&D varie entre 3 e 12 anos (ASHBURN; THOR, 2004). As moléculas reposicionadas apresentam perfis farmacocinético e toxicológico conhecidos e otimizados para o uso em humanos, visto que foram submetidas a criteriosas avaliações durante a etapa pré-clínica de P&D, antes de terem progredido para os estudos clínicos para a indicação primária. Além disso, os inúmeros dados disponíveis nas espécies pré-clínicas e em humanos possibilitam uma projeção mais assertiva do potencial risco-benefício do composto para a nova indicação desde o início do projeto.

Os compostos podem ser reposicionados nas etapas pré-clínicas ou clínicas do processo de P&D para a nova indicação, dependendo de suas propriedades, das otimizações necessárias para a nova doença e dos estudos previamente realizados para a indicação primária. Caso o mesmo produto e a posologia já avaliados em humanos, incluindo formulação, via de administração e doses, sejam adequadas para a nova indicação, o composto poderia ser testado diretamente em estudos clínicos de eficácia – reposicionamento na fase clínica II (OPREA; OVERINGTON, 2015). Por outro lado, se o medicamento original for modificado, será necessário avaliar a segurança e a farmacocinética do novo produto – reposicionamento na etapa pré-clínica ou na fase clínica I. Por exemplo, novos estudos de segurança pré-clínicos podem ser necessários se uma nova via de administração for utilizada (FETRO; SCHERMAN, 2020). De acordo com as informações disponíveis na literatura e os estudos já realizados para o candidato, é possível definir os experimentos complementares necessários para que o composto chegue ao mercado. Caso as otimizações para a nova indicação exijam modificações na estrutura do composto, ele

se tornará um ponto de partida para um novo projeto de P&D de química medicinal – reposicionamento na etapa pré-clínica como uma molécula líder. Neste cenário, o projeto de reposicionamento apresenta a vantagem de ser iniciado a partir de séries químicas promissoras, que apresentam propriedades de segurança e de farmacocinética otimizadas para o uso em humanos (KLUG; GELB; POLLASTRI, 2016).

Os relatos de fármacos ativos experimentalmente em modelos pré-clínicos de doenças infecciosas suportam o potencial desta abordagem para superar a falta de tratamentos eficazes para as DTNs. A eflornitina (Figura 2) e o nifurtimox, inicialmente em desenvolvimento para doenças oncológicas e infecções pelo *T. cruzi*, foram reposicionados para o combate ao *T. brucei gambiense*. Além destes, a miltefosina (Figura 2), em desenvolvimento para indicações oncológicas, foi reposicionada para a terapêutica de leishmaniose (FERREIRA; ANDRICOPULO, 2016). No caso da doença de Chagas, antifúngicos azólicos, como o posaconazol e o voriconazol (Figura 2), foram identificados como inibidores da citocromo P450 14-alfa-esterol desmetilase do parasita (TcCYP51) e, deste modo, como potenciais candidatos para o seu tratamento. Apesar da ausência de eficácia de compostos desta classe no combate ao *T. cruzi* durante os estudos clínicos, a sua utilização na terapêutica de Chagas continua sendo estudada, particularmente em combinação com outros fármacos (TORRICO et al., 2018). Algumas análises sugerem que diversos fatores podem ter levado à falha do posaconazol e do E1224 (pró-fármaco do voriconazol), como diferenças nas linhagens do parasita pelas quais os pacientes estavam infectados, o tipo de desfecho clínico avaliado, assim como farmacocinética e duração do tratamento inadequados para a indicação (CHATELAIN, 2015; FERREIRA; ANDRICOPULO, 2016).

Figura 2 - Estrutura química dos fármacos eflornitina, miltefosina, posaconazol e ravuconazol.

Fonte: Produção do autor.

Os diversos trabalhos da literatura voltados ao reposicionamento de fármacos para a doença de Chagas evidenciam o potencial desta estratégia em acelerar a identificação de terapias seguras e eficazes para o seu tratamento (ENGEL et al., 2010; DE RYCKER et al., 2016; PLANER et al., 2014). Por outro lado, nenhum estudo reportado avaliou de maneira sistemática todos os fármacos e candidatos a fármacos disponíveis em ensaios de inibição do parasita. Deste modo, é razoável assumir que ainda há um número significativo compostos que poderiam ser ativos contra o *T. cruzi*, porém ainda não foram testados. Soma-se a tal fato, a dificuldade de acesso a fármacos ainda em desenvolvimento clínico e a metabólitos, praticamente inviabiliza a realização de triagens experimentais extensivas com bibliotecas destes compostos (CORSELLO et al., 2017; DE RYCKER et al., 2016; ENGEL et al., 2010). Neste contexto, métodos computacionais que permitam a identificação de moléculas com

maior potencial de serem ativos contra o *T. cruzi* podem acelerar e viabilizar iniciativas de reposicionamento para a doença de Chagas (EKINS et al., 2011; JIN; WONG, 2014).

1.4 USO DE MODELOS COMPUTACIONAIS PARA IDENTIFICAR COMPOSTOS A SEREM REPOSICIONADOS PARA A DOENÇA DE CHAGAS

As primeiras iniciativas de reposicionamento originaram de observações inesperadas, que foram posteriormente investigadas e resultaram em fármacos seguros e eficazes para a nova indicação. A sildenafil, por exemplo, inicialmente em desenvolvimento para hipertensão, foi reposicionada para o tratamento de disfunção erétil a partir da análise dos dados de estudos clínicos. A talidomida, desenvolvida inicialmente como um sedativo e removida do mercado devido a efeitos teratogênicos, foi reposicionada para o tratamento de mieloma múltiplo e eritema nodoso em hanseníase (PUSHPAKOM et al., 2019). Estes casos de sucesso estimularam o interesse da comunidade científica e de indústrias farmacêuticas em explorar oportunidades de reposicionamento de maneira sistemática.

Métodos computacionais que permitem integrar informações provenientes de diferentes áreas do conhecimento, como bioinformática, quimioinformática e biologia de sistemas, podem ser empregados para gerar hipóteses de reposicionamento para serem testadas experimentalmente. Embora as análises computacionais não substituam a realização de ensaios *in vitro* e *in vivo*, elas podem auxiliar na identificação e priorização dos compostos mais promissores, otimizando, assim, os investimentos e o tempo do processo de reposicionamento (NEVES et al., 2018). Dentre os métodos utilizados na área, podem-se citar a análise de perfis de expressão gênica, transcriptômica e proteômica, a avaliação de similaridade estrutural com moléculas bioativas, a ancoragem molecular em alvos terapêuticos promissores, a análise de prontuários médicos eletrônicos, resultados de estudos clínicos e registros de farmacovigilância com ferramentas de mineração de dados e processamento de linguagem natural, assim como a análise de redes e vias de sinalização (OPREA; OVERINGTON, 2015; PUSHPAKOM et al., 2019; TALEVI; BELLERA, 2020).

Na área da doença de Chagas, observou-se aumento significativo na diversidade e volume de dados disponíveis, que podem ser utilizados como base para iniciativas de reposicionamento. Notavelmente, o surgimento de métodos de triagem

experimental de alta capacidade para o *T. cruzi* permitiu que bibliotecas com milhares ou até milhões de compostos fossem avaliadas e resultou na descoberta de novas moléculas e séries químicas ativas contra o parasita (ALONSO-PADILLA; RODRIGUEZ, 2014; CARMODY et al., 2010; KHARE et al., 2016; PEÑA et al., 2015). A maioria destes ensaios é realizada com o parasita completo e, deste modo, faz parte das estratégias de *phenotypic-based drug discovery*. Os dados gerados podem ser utilizados para o desenvolvimento de modelos baseados na estrutura de ligantes bioativos, conhecidos como LBDD (do inglês, *ligand-based drug design*). Os modelos LBDD baseiam-se na premissa de que compostos com estrutura ou propriedades similares apresentam alto potencial de exibirem um perfil de atividade biológica também similar. Brevemente, estes modelos utilizam cálculos estatísticos e métodos de aprendizado de máquina para estabelecer a relação entre os descritores das moléculas e a atividade biológica de interesse (JIN; WONG, 2014; MITCHELL, 2014). Esta abordagem já se mostrou eficaz na descoberta de compostos bioativos para o tratamento de doenças infecciosas, como a doença do sono (SCHMIDT et al., 2014), a tuberculose (EKINS et al., 2010; 2013; EKINS, FREUNDLICH, 2011; PERIWAL et al., 2011), infecções bacterianas (SPECK-PLANCHE, CORDEIRO, 2014), a malária (MAHMOUDI et al., 2006; MARRERO-PONCE et al., 2005, 2006; ZHANG et al., 2012) e mesmo para a doença de Chagas (CASTILLO-GARIT et al., 2011; EKINS et al., 2015; GARCIA-DOMENECH et al., 2015; MELO-FILHO et al., 2019).

Além da abordagem LBDD, a grande e crescente disponibilidade de estruturas cristalográficas de novos alvos terapêuticos promissores para o desenvolvimento de novas terapias para Chagas permite a exploração de modelos computacionais de reposicionamento baseados na estrutura do alvo de interesse, método conhecido como SBDD (*structure-based drug design*). Estes métodos utilizam informações sobre a estrutura da biomacromolécula de interesse para a identificação de compostos que apresentem propriedades eletrostáticas e estereoquímicas complementares ao alvo. Conforme maior a complementariedade e o número de interações de um composto com um alvo, maior o seu potencial de modulá-lo de maneira potente e seletiva (FERREIRA et al., 2015). Modelos SBDD já foram utilizados em iniciativas de reposicionamento para identificar novos tratamentos para a doença de Chagas através da modulação de diferentes alvos terapêuticos em potencial, como a putrescina permease (ALBERCA et al., 2018), a cruzipaina (BELLERA et al., 2015) e a diidrofolato redutase-timidilato sintetase (JUÁREZ-SALDIVAR et al. 2020).

CAPÍTULO 2

2 OBJETIVOS E JUSTIFICATIVA

A doença Chagas é uma enfermidade crônica que afeta oito milhões de pessoas ao redor do mundo e responsável pela morte de aproximadamente 10.000 pacientes infectados por ano. A limitada eficácia do benznidazol e do nifurtimox, únicos fármacos disponíveis no mercado para o seu tratamento, aliada à toxicidade e aos efeitos adversos causados por estas terapias, evidenciam a necessidade de novos tratamentos eficazes e seguros para a doença de Chagas.

O reposicionamento de fármacos tem se mostrado como uma estratégia promissora para reduzir os investimentos, os riscos de falha e o tempo de desenvolvimento de novos medicamentos. Isto é possível pois as moléculas reposicionadas apresentam propriedades de farmacocinética e segurança otimizadas para o uso em humanos. Além disso, os experimentos pré-clínicos e clínicos realizados para a indicação original podem ser reaproveitados e possibilitar que menos estudos sejam necessários para a nova indicação. Essas vantagens são particularmente relevantes para a doença de Chagas, que carece de investimentos em novas terapias.

Neste cenário, o presente trabalho teve como objetivo identificar fármacos aprovados e candidatos a fármacos em desenvolvimento clínico com o potencial de serem reposicionados para o tratamento da doença de Chagas. A estratégia empregada consistiu em desenvolver e validar modelos computacionais capazes de identificar moléculas ativas contra o *T. cruzi* utilizando métodos LBDD e SBDD. Os modelos foram posteriormente empregados na triagem virtual visando à identificação de moléculas ativas contra o parasita que pudessem ser reposicionadas para a terapêutica da doença de Chagas. Esta abordagem sistemática apresenta um escopo expandido, quando comparada a trabalhos anteriores, muitas vezes limitados exclusivamente a bibliotecas de fármacos aprovados ou, alguns casos, de fármacos e candidatos a fármaco. Na abordagem LBDD, buscou-se ainda complementar os trabalhos previamente realizados através da utilização de dados de moléculas bioativas previamente inexplorada para o desenvolvimento dos modelos preditivos para Chagas. Na estratégia SBDD, modelos capazes de identificar inibidores da diidroorotato desidrogenase do *T. cruzi* (TcDHODH) foram desenvolvidos, um novo alvo terapêutico promissor para o combate ao parasita, porém ainda pouco explorado.

O racional científico do potencial da inibição da TcDHODH no combate ao *T. cruzi* será apresentado na seção 4.1.1.

2.1 OBJETIVOS ESPECÍFICOS

2.1.1 Abordagem LBDD: Desenvolvimento de modelos baseados na estrutura de compostos ativos fenotipicamente contra o *Trypanosoma cruzi*

- Integrar e preparar um conjunto de compostos com atividade tripanocida *in vitro* (fenotípica) contra *T. cruzi*;
- Desenvolver modelos de QSAR (*quantitative structure activity relationship*) para predição da atividade tripanocida através da combinação de fingerprints e métodos de aprendizado de máquina;
- Validar os modelos, utilizando os métodos de validação cruzada quádrupla, teste de aleatorização e validação externa, visando a selecionar aquele de maior capacidade preditiva;
- Comparar a capacidade preditiva do modelo selecionado com um método de referência (busca por similaridade estrutural);
- Realizar a triagem virtual de fármacos aprovados, candidatos a fármaco em desenvolvimento clínico e metabólitos de fármacos utilizando o modelo selecionado;
- Analisar os compostos selecionados com base na sua via de administração em humanos, status regulatório e similaridade com compostos da série treino e
- Determinar a atividade tripanocida *in vitro* contra a forma amastigota do *T. cruzi* para compostos selecionados pelo modelo de QSAR.

2.1.2 Abordagem SBDD: Desenvolvimento de modelos baseados na inibição da enzima diidroorotato desidrogenase do *Trypanosoma cruzi* (TcDHODH)

- Identificar e compilar informações de compostos avaliados *in vitro* na inibição da TcDHODH e de estruturas cristalográficas da enzima;
- Determinar através de estudos de reancoragem a capacidade do protocolo em prever a pose e conformação de ligantes co-cristalizados com a TcDHODH;

- Avaliar a capacidade do algoritmo de ancoragem molecular em ranquear compostos ativos com uma maior pontuação do que compostos inativos (enriquecimento);
- Desenvolver modelos de aprendizado de máquinas capaz de identificar inibidores potentes da TcDHODH utilizando, dentre outros descritores, a entalpia de interação do ligante com o alvo calculada por métodos semi-empíricos;
- Validar os modelos, utilizando os métodos de validação cruzada quádrupla, o teste de aleatorização e a validação externa, visando a selecionar aquele de maior capacidade preditiva;
- Realizar a triagem virtual de fármacos aprovados e candidatos a fármaco em desenvolvimento clínico utilizando o modelo desenvolvido e
- Analisar os compostos selecionados com base na sua via de administração em humanos e o status regulatório.

CAPÍTULO 3

3 ABORDAGEM LBDD: DESENVOLVIMENTO DE MODELOS BASEADOS NA ESTRUTURA DE COMPOSTOS ATIVOS FENOTIPICAMENTE CONTRA O *TRYPANOSOMA CRUZI*

3.1 INTRODUÇÃO

3.1.1 Modelos preditivos baseados na estrutura de ligantes bioativos

Os métodos de planejamento de fármacos baseados em ligantes, também conhecidos como LBDD (*ligand-based drug design*), consistem em analisar as características estruturais e físico-químicas de compostos visando a compreender a sua relação com a propriedade ou atividade farmacológica de interesse. Estes métodos podem, ainda, ser classificados como pertencentes à área de quimioinformática, a qual compreende o armazenamento, o processamento, a representação, a visualização e a análise de dados químicos utilizando ferramentas computacionais com vistas a resolver problemas na área de química (LO et al., 2018; ALVES et al., 2018). Algumas das abordagens comumente empregadas na área incluem buscas por similaridade, modelos farmacofóricos e estudos de relações quantitativas entre a estrutura química e a atividade biológica/propriedade (do inglês, *quantitative structure-activity relationship* e *quantitative structure-property relationship*, também conhecidos como QSAR e QSPR, respectivamente) (ACHARYA et al., 2011; LEE; HUANG; JUAN, 2011; OU-YANG et al., 2012).

A abordagem de QSAR/QSPR consiste em aplicar métodos matemáticos e estatísticos de análise de dados para desenvolver modelos capazes de prever a atividade biológica ou as propriedades de novos compostos com base nas suas características estruturais. As principais etapas desta abordagem incluem: i) integrar, preparar e curar bancos de dados para os quais a atividade/propriedade de interesse é conhecida; ii) desenvolver e validar os modelos utilizando metodologias de validação interna e externa e iii) aplicar os modelos validados na triagem virtual de bibliotecas de compostos de modo a priorizar aqueles que serão testados experimentalmente (TROPSHA, 2010). Na P&D de fármacos, modelos de QSAR/QSPR têm sido amplamente utilizados como ferramentas para identificar novos compostos bioativos ou na otimização estrutural de séries químicas (ALVES et al., 2018). Além da atividade farmacológica, diversas propriedades de interesse farmacêutico podem ser

modeladas, incluindo absorção, distribuição, metabolismo e excreção (ADME), toxicidade e propriedades físico-químicas de compostos (CHERKASOV et al., 2014; LAVECCHIA, 2015; LO et al., 2018). O objetivo final do desenvolvimento de modelos preditivos é a extrapolação das correlações identificadas de modo a prever as propriedades de moléculas ainda não testadas experimentalmente, otimizando, assim, o processo de P&D (LO et al., 2019).

O desempenho de modelos de QSAR/QSPR depende das moléculas utilizadas no seu desenvolvimento, dos descritores empregados para representá-las, da natureza da propriedade modelada e dos algoritmos matemáticos utilizados para identificar as relações entre os descritores e atividade/propriedade em análise (ACHARYA et al., 2011; MITCHELL, 2014). Os trabalhos pioneiros em QSAR foram desenvolvidos na década de 1960 com um número limitado de compostos de séries congêneres e métodos de regressão multivariada simples (CHERKASOV et al., 2014; NEVES et al., 2018). Hansch et al. (1962), por exemplo, correlacionaram a atividade de auxinas na regulação do crescimento de plantas com a constante eletrostática de Hammett e a lipofilicidade (coeficiente de partição em octanol/água ou logP). A capacidade preditiva dos modelos tem sido otimizada através do desenvolvimento de novos descritores moleculares e da aplicação de novos métodos matemáticos e estatísticos de modelagem (CHERKASOV et al., 2014; SELASSIE; MEKAPATI; VERMA, 2002). Além disso, houve expansão significativa no volume de resultados químicos e biológicos disponíveis na literatura e em bancos de dados on-line, como o ChEMBL (GUALTON et al., 2017), o PubChem BioAssay (WANG et al., 2012) e o Binding DB (LIU et al., 2007). Este fenômeno somente foi possível graças aos avanços científicos observados nos últimos anos, como o surgimento da síntese combinatória e métodos de triagem experimental de alta capacidade. Estes avanços têm impacto na área de quimioinformática, visto que a capacidade preditiva de modelos LBDD melhora conforme a maior quantidade e qualidade dos dados disponíveis (VAMATHEVAN et al., 2019). Somado a isso, os novos descritores moleculares e métodos matemáticos têm fomentado o desenvolvimento de modelos de alta performance, com amplo domínio de aplicabilidade, *i.e.*: podem ser utilizados para a predição de compostos pertencentes a um espaço químico diverso (CHERKASOV et al., 2014; FERREIRA; ANDRICOPULO, 2018; MITCHELL, 2014).

As relações entre a atividade biológica/propriedade de compostos e os descritores moleculares são complexas e muitas vezes não podem ser capturadas de

maneira adequada pelos algoritmos lineares amplamente empregados nos primeiros estudos de QSAR/QSPR (ACHARYA et al., 2011; LO et al., 2018). Neste sentido, a performance de modelos preditivos LBDD poderia ser otimizada através da implementação de métodos de aprendizado de máquina, os quais consistem na utilização de algoritmos computacionais estatísticos para identificar correlações presentes nos dados (lineares e/ou não lineares) e utilizá-las na predição das propriedades de novos compostos (CASSIDY, 2020; VAMATHEVAN, 2019).

Os algoritmos de aprendizado de máquina podem ser categorizados em supervisionados e não supervisionados. Nos métodos supervisionados, a variável de interesse (ex.: atividade biológica ou propriedade) é conhecida para os compostos utilizados no desenvolvimento do modelo. Os métodos supervisionados podem, ainda, ser classificados em algoritmos de regressão, quando a variável modelada é contínua (ex.: IC₅₀ em um determinado ensaio biológico) ou de classificação, no caso de propriedades/atividades categóricas (ex.: composto ativo ou inativo).

Alguns métodos de aprendizado de máquina amplamente empregados são máquinas de vetores de suporte (do inglês, *support vector machine*), árvores de decisão (ex.: floresta aleatória ou *random forest*), redes neurais artificiais (*artificial neural networks*), kNN (do inglês, *k-nearest neighbours*) e métodos probabilísticos bayesianos (ex.: *Naive Bayes*) (SVETNIK et al., 2003; LO et al., 2018). Por outro lado, os modelos não supervisionados são desenvolvidos sem a inclusão de dados sobre a propriedade modelada. Deste modo, os algoritmos identificam os padrões existentes em compostos de um banco de dados somente a partir dos descritores moleculares. Estes métodos possuem a vantagem de poderem ser aplicados na ausência de informações prévias sobre a atividade biológica ou classificação dos compostos (SCHEEDER; HEIGWER; BOUTROS, 2018; VAMATHEVAN, 2019). Conforme descrito anteriormente, modelos LBDD já foram utilizados com sucesso na P&D de compostos bioativos para o tratamento de doenças infecciosas, como a doença do sono, a tuberculose, infecções bacterianas, a malária e mesmo para a doença de Chagas.

3.1.2 Descritores moleculares

Os descritores moleculares, segundo Todeschini e Consonni (2000, p. XI, tradução nossa) são o “resultado de um processo lógico e matemático que transforma

informações químicas codificadas na representação simbólica de uma molécula em um número útil ou o resultado de um experimento padronizado”. A representação das propriedades e da estrutura química de compostos de uma maneira numérica permite a sua interpretação e processamento por algoritmos computacionais (ALVES et al., 2018; MITCHELL, 2014).

Os descritores podem ser obtidos de maneira experimental ou gerados *in silico*. Notavelmente, o número de descritores reportados na literatura é elevado e está em expansão como resultado de diversas pesquisas na área (LEACH; GILLET, 2007). Segundo Mauri, Consonni e Todeschini (2016), descritores moleculares devem: i) ser invariáveis com relação à rotulagem e numeração dos átomos, ii) ser invariáveis com relação à roto-translação da molécula, iii) ser obtidos através de um método não ambíguo e iv) ser aplicáveis para os compostos de interesse. Critérios adicionais também foram sugeridos por estes autores com o objetivo de otimizar a aplicabilidade de descritores moleculares, os quais incluem: i) permitir a interpretação estrutural, ii) apresentar boa correlação com pelo menos uma propriedade experimental, iii) não possuir correlação trivial com outros descritores moleculares, iv) não serem baseados em propriedades experimentais, v) preferencialmente contínuos, vi) preferencialmente apresentarem um baixo nível de degeneração, vii) preferencialmente simples, viii) preferencialmente aplicáveis para ampla classe de moléculas, ix) preferencialmente capazes de discriminar isômeros e x) preferencialmente apresentar valores calculáveis em uma faixa de valores adequada para as moléculas de interesse.

Os descritores codificam informações de diferentes níveis sobre os compostos, que podem variar desde propriedades unidimensionais ou 1D (ex.: massa molecular e propriedades físico-químicas), bidimensionais ou 2D (ex.: índices topológicos e conectividade entre átomos), tridimensionais ou 3D (ex.: propriedades eletrostáticas e estéricas da estrutura tridimensional de um composto), 4D (expansão dos descritores 3D que consideram a dinâmica conformacional dos compostos) até descritores n dimensionais ou nD (ALVES et al., 2018; NEVES et al., 2018; LO et al. 2018). Descritores 1D têm a limitação de serem degenerados, de modo que compostos distintos podem apresentar o mesmo valor. Por sua vez, descritores 3D ou superiores, apesar de codificarem maior complexidade estrutural, estão sujeitos a ambiguidades durante a geração da conformação tridimensional e/ou no alinhamento dos compostos (LO et al., 2018). Neste contexto, descritores estruturais 2D são os

mais frequentemente utilizados na literatura para o desenvolvimento de modelos LBDD (CHERKASOV et al., 2014; LO et al., 2018).

Os *fingerprints* moleculares (em português, impressões digitais moleculares) são descritores que representam as características estruturais de compostos em um vetor numérico (GAO et al., 2020). Cada componente do vetor, também conhecido como *bit*, representa uma característica estrutural específica. Os métodos utilizados para gerar os *fingerprints* fragmentam as moléculas de acordo com regras pré-estabelecidas e registram os fragmentos em *bits* do vetor numérico. Os *fingerprints* podem ser classificados em (i) *fingerprints* baseados em dicionários de subestruturas, (ii) *fingerprints* topológicos baseados em caminhos lineares (do inglês, *path-based fingerprint*), (iii) *fingerprints* farmacóforos e (iv) *fingerprints* circulares (GAO et al., 2020; KRISTENSEN; NIELSEN; PEDERSEN, 2013). Os *fingerprints* topológicos são gerados a partir da conectividade linear entre dois ou mais átomos, enquanto os circulares consideram todas as conexões de cada átomo em um ambiente circular, ambos até um limite pré-determinado de ligações. Os *fingerprints* farmacofóricos, por sua vez, registram os elementos das moléculas pelas suas características farmacofóricas (RINIKER; LANDRUM, 2013). Para cada composto, o *fingerprint* é gerado através da análise da presença de cada *bit* na sua estrutura. Os vetores podem conter informações sobre a frequência com que cada característica estrutural ocorre ou indicar apenas a presença ou ausência desta característica de maneira binária (*binary fingerprint*) (O'BOYLE; SAYLE, 2016).

3.1.3 Algoritmos de aprendizado de máquina

3.1.3.1 Máquinas de vetores de suporte

No método máquinas de vetores de suporte ou SVM (do inglês, *support vector machine*), os descritores dos compostos são inicialmente projetados em um espaço de características multidimensionais. O algoritmo busca definir um hiperplano que leva à melhor separação das classes em análise. Os descritores podem ser transformados utilizando funções de kernel (ex.: linear, polinomial, sigmoidal e radial) de modo a melhorar a separação dos dados. Os compostos utilizados para definir o hiperplano são denominados vetores de suporte. A predição de propriedades para novos

compostos é determinada com base no espaço multidimensional em que eles forem projetados (LAVECCHIA, 2015; LEACH; GILLET, 2007).

3.1.3.2 Floresta aleatória

O método de floresta aleatória ou RF (do inglês, *random forest*), consiste em um conjunto de árvores de decisão, no qual cada árvore é construída com uma seleção de compostos da Série treino e cada nó utiliza uma fração dos descritores originais, ambos selecionados de maneira aleatória. A predição associada a um composto baseia-se em um sistema de votação, sendo atribuída a classe com a maioria dos votos. Árvores de decisão compreendem uma série de regras sequenciais, também conhecidas como nós, que separam os compostos de um grupo de maneira a atingir a melhor segmentação das classes de interesse. Dois ou mais subgrupos são gerados em cada nó, os quais podem ser novamente segmentados. O processo termina quando uma classe é atribuída a um subgrupo ou quando o modelo não é mais capaz de segmentar os dados de maneira efetiva. Comparativamente à utilização de uma única árvore de decisão, o método RF apresenta menor propensão ao sobreajuste, o qual ocorre quando o modelo possui alta capacidade preditiva para os compostos da série treino, porém, baixa para compostos desconhecidos (LEACH; GILLET, 2007).

3.1.3.3 Naive Bayes

O método de *Naive Bayes* (NB) se baseia no teorema de Bayes, o qual correlaciona a probabilidade de o evento ocorrer (e.g.: atividade biológica) com uma ou mais possíveis causas (e.g.: descritores estruturais ou físico-químicos dos compostos). A probabilidade de um evento A acontecer dada a existência de uma característica B pode ser calculada utilizando a equação (1), na qual $P(B|A)$ representa a probabilidade de a característica B existir dada a ocorrência do evento A e $P(A)$ e $P(B)$ representam, respectivamente, as probabilidades do evento A e da característica B ocorrerem (LAVECCHIA, 2015).

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (1)$$

3.1.3.4 Redes neurais probabilísticas

Redes neurais artificiais são métodos de aprendizado constituídos por nós, também conhecidos como neurônios, organizados em diversas camadas e conectados em formato de rede. A primeira camada é utilizada para inserir os descritores dos compostos e a classificação determinada pelo modelo é obtida na última camada. As camadas intermediárias analisam os dados e correlacionam os descritores com as propriedades em avaliação (MITCHELL, 2014). Diferentes estruturas de conexão têm sido exploradas, podendo os neurônios estar conectados apenas às camadas posteriores ou em ambas as direções (LAVECCHIA, 2015). O método de redes neurais probabilísticas ou PNN (do inglês, *probabilistic neural network*), é um tipo de rede neural artificial de base radial com quatro camadas, nas quais os neurônios são conectados apenas às camadas seguintes. Este método se baseia na comparação dos descritores de novos compostos com os padrões existentes nos dados utilizados durante o desenvolvimento do modelo (NIWA, 2003; 2004).

3.2 MATERIAL E MÉTODOS

3.2.1 Equipamentos e programas

3.2.1.1 Equipamentos

- Estação de trabalho com processador Intel i7-2670QM CPU @ 2.20 GHz, 8 GB RAM e placa gráfica Intel HD Graphics 3000.
- Estação de trabalho com 6 x Intel i7-6800K CPU @ 3.40 GHz, 16 GB RAM e placa de vídeo NVIDIA GTX 1080 8 GB.

3.2.1.2 Programas

- KNIME v3.6.0 (University of Konstanz, Suíça)
- Python 3.8.2 (Python Software Foundation, Estados Unidos)
- The Chemistry Development Kit (CDK) (STEINBECK et al., 2003)

- RDKit (RDKit, 2018)
- Enalos (NovaMechanics, Chipre)

3.2.2 Construção e padronização dos bancos de dados

Compostos testados em ensaios fenotípicos contra o parasita *T. cruzi* foram identificados a partir de uma busca nas bases de dados PubMed, Google Acadêmico, ChEMBL (GUALTON et al., 2017) e PubChem BioAssay (WANG et al., 2012) entre janeiro de 2017 e dezembro de 2018, na língua inglesa e em português. As palavras-chave *Trypanosoma cruzi*, Chagas, *Chagas disease*, *phenotypic assay*, *phenotypic screening*, HTS, *high throughput screening*, HCS, *high content screening*, *trypanosomatid*, suas combinações e correspondentes em português foram empregadas. As referências dos artigos encontrados foram avaliadas manualmente com objetivo expandir os resultados. Os dados das publicações de Peña et al. (2015), Kaiser et al. (2015a) e do Instituto Broad (CARMODY et al., 2010) foram organizados em bancos de dados, conforme descrito a seguir, e utilizados para o desenvolvimento e validação dos modelos computacionais. Estas bibliotecas foram selecionadas porque continham o maior número moléculas com dados de concentração-resposta em ensaios de inibição o *T. cruzi*.

Os bancos de dados foram processados com base nas práticas propostas por Fourches, Muratov e Tropsha (2010, 2015, 2016). Os contra-íons, as duplicatas, os compostos organometálicos, os polímeros, as misturas e os sais inorgânicos foram removidos. Adicionalmente, os grupamentos químicos foram normalizados, as cargas foram neutralizadas, os átomos de hidrogênio foram adicionados de forma explícita e as estruturas resultantes foram avaliadas visualmente.

Os compostos das publicações de Peña et al. (2015) e Kaiser et al. (2015a) foram classificados como ativos ou inativos com base nos critérios estabelecidos pelo DNDi (CHATELAIN, 2015; CHATELAIN; IOSET, 2018). Estes critérios consideram a concentração de composto capaz de inibir em 50% a proliferação do parasita (IC_{50}) e a concentração citotóxica de composto que inibe em 50% a proliferação da célula hospedeira (CC_{50}). Moléculas com alta capacidade de inibir a proliferação do parasita ($IC_{50} < 5 \mu M$) e com atividade seletiva frente a célula hospedeira (índice de seletividade >10 , calculado pela relação entre os valores de CC_{50} e IC_{50}) foram classificadas como

ativas. No caso dos dados de publicação de Peña et al. (2015), como os resultados de dois ensaios ortogonais de inibição do *T. cruzi* estavam disponíveis, somente os compostos que cumpriram os critérios para ambos os métodos foram classificados como ativos. Visando a evitar a existência de compostos classificados incorretamente como inativos devido à variabilidade dos ensaios biológicos, removeram-se as moléculas com IC₅₀ entre 5 µM e 10 µM (ACEVEDO; SCOTTI; SCOTTI, 2018). Os compostos com índice de seletividade (SI) inferior a 10 também foram removidos, mesmo que apresentassem IC₅₀ < 5 µM, para evitar a existência de falsos ativos no banco de dados. Os compostos do Instituto Broad (CARMODY et al., 2010) que apresentaram inibição do crescimento do parasita inferior a 20% na concentração de 3,75 µM foram classificados como inativos neste trabalho.

3.2.2.1 Série treino e Série teste A

Com base na publicação de Peña et al. (2015), utilizaram-se 271 compostos ativos e 297.961 inativos do Instituto Broad (CARMODY et al., 2010) para compor um banco de dados inicial. O número de compostos inativos foi reduzido para 271 através de uma amostragem aleatória com vistas a equilibrar o número de moléculas em cada classe (1:1 de ativos e inativos). Esta estratégia foi empregada para evitar que a proporção desbalanceada de ativos e inativos enviesasse o aprendizado dos algoritmos de modo a favorecer a predição da classe com maior número de moléculas (CAPUZZI et al., 2018; NEVES et al., 2016; ZAKAHROV et al., 2014). Os 542 compostos resultantes foram divididos para formar a Série treino e a Série teste A, compostas respectivamente por 80% (432 compostos) e 20% (110) do banco de dados inicial. A divisão foi feita através de um sorteio aleatório para o conjunto de compostos de cada categoria, visando a manter a mesma proporção de ativos e inativos para cada série.

3.2.2.2 Série teste B

Com base na publicação de Kaiser et al. (2015a), utilizaram-se 21 compostos ativos e 297.691 inativos do Instituto Broad (CARMODY et al., 2010) para compor o banco de dados inicial. Destes, os 271 inativos pertencentes à Série treino e à Série teste A foram removidos. O número de compostos inativos foi reduzido para 400,

através de uma seleção aleatória, para atingir uma porcentagem de moléculas ativas próxima a 5% do total do banco de dados, analogamente ao realizado em outros trabalhos (ALBERCA et al., 2016, 2018; GARTNER et al., 2017). Esta proporção baseia-se nos cálculos de Truchon e Bayly (2007), os quais demonstraram que as métricas de enriquecimento para avaliar modelos classificatórios, como a área sob a curva ROC (*Receiver Operating Characteristics*), são influenciadas pelo tamanho dos bancos de dados e pela proporção de ativos e inativos. Bancos de dados maiores e com baixa proporção de ativos tendem a reduzir os erros nos cálculos das métricas de enriquecimento.

3.2.2.3 Séries de triagem

As Séries de triagem A e B, compostas respectivamente por 2.367 fármacos aprovados e 1.618 candidatos a fármaco em desenvolvimento clínico, foram geradas a partir dos dados da plataforma *Drug Repurposing Hub*, disponibilizada pelo Instituto Broad (CORSELLO et al., 2017). A Série de triagem C, composta por 2.331 metabólitos formados a partir de fármacos aprovados, foi organizada com os dados do *DrugBank* (WISHART et al., 2018). A Série de triagem C não contém os metabólitos que haviam sido, posteriormente, desenvolvidos/testados como fármacos individualmente, pois estes já estavam presentes na Séries de triagem A e B.

3.2.3 Descritores estruturais

Os *fingerprints* moleculares (em português, impressões digitais moleculares) do tipo *hashed* foram utilizados como descritores da estrutura química das moléculas ativas e inativas. Os fingerprints foram gerados com 1024 bits utilizando os métodos *Extended Connectivity Fingerprint* com diâmetro 6 (ECFP6), *Functional Class Fingerprint* com diâmetro 6 (FCFP6) e *Extended* do CDK (STEINBECK et al., 2003), e os métodos Morgan e FeatMorgan, ambos com raio 3, do RDKit (MORGAN, 1965; ROGERS; HAHN, 2010).

3.2.4 Métodos de aprendizado de máquina e hiperparâmetros

Os algoritmos de aprendizado de máquina utilizados neste trabalho foram o floresta aleatória (RF), o *Naive Bayes* (NB), as redes neurais probabilísticas (PNN) e as máquinas de vetores de suporte (SVM) (SVETNIK et al., 2003; LO et al., 2018). Os hiperparâmetros foram otimizados pelo método de busca extensivo, no qual todas as variações pré-determinadas de cada hiperparâmetro e suas combinações são testadas, com o objetivo de atingir o maior valor de área sob a curva ROC (métrica descrita em detalhe na próxima sessão). Os hiperparâmetros otimizados, assim como os valores testados, foram C (0,1, 1 e 10), gamma (10^{-6} , 10^{-5} , 10^{-4} e 0) e as funções de kernel (linear, radial, sigmoide e radial) para o SVM. Para o algoritmo RF, otimizou-se o número de árvores (50, 100, 200 e 400). No caso do PNN, os hiperparâmetros de teta menos (0,1, 0,2 e 0,3) e teta mais (0,3, 0,4 e 0,5) foram otimizados. Os valores testados foram selecionados visando a explorar variações próximas às configurações pré-estabelecidas de cada algoritmo. Os demais hiperparâmetros foram mantidos em suas configurações originais.

3.2.5 Métricas para a avaliação dos modelos

A comparação da classificação predita para cada composto com o dado experimental resulta em quatro categorias: compostos ativos classificados corretamente (verdadeiro positivos ou VP), compostos ativos classificados como inativos (falso negativos ou FN), compostos inativos classificados corretamente (verdadeiro negativo ou VN) e compostos inativos classificados como ativos (falso positivos ou FP). A análise do número de compostos em cada classe fornece métricas para avaliar a performance dos modelos.

A sensibilidade (SE ou SENS.) representa a proporção de compostos ativos preditos corretamente (Equação 2). A especificidade (SP ou ESPEC.) representa a proporção de compostos inativos classificados corretamente (Equação 3). O valor preditivo positivo (PPV) e o valor preditivo negativo (NPV) representam a proporção de verdadeiros positivos e negativos, respectivamente, dentre os compostos classificados como ativos e inativos (Equações 4 e 5, respectivamente). A acurácia (AC) avalia a proporção de compostos, ativos ou inativos, classificados corretamente (Equação 6). A taxa de classificação balanceada (TCB), utilizada para balancear os

erros em ambas as classes (ativos e inativos) para bancos de dados desbalanceados, foi calculada como a média da sensibilidade e da especificidade (Equação 7) (PERIWAL et al., 2011).

$$\text{Sensibilidade} = \text{VP} / (\text{VP} + \text{FN}) \quad (2)$$

$$\text{Especificidade} = \text{VN} / (\text{VN} + \text{FP}) \quad (3)$$

$$\text{PPV} = \text{TP} / (\text{TP} + \text{FP}) \quad (4)$$

$$\text{NPV} = \text{TN} / (\text{TN} + \text{FN}) \quad (5)$$

$$\text{Acurácia} = (\text{VP} + \text{VN}) / (\text{VP} + \text{FN} + \text{VN} + \text{FP}) \quad (6)$$

$$\text{TCB} = (\text{SP} + \text{SE}) / 2 \quad (7)$$

A avaliação do desempenho dos modelos incluiu, ainda, a análise da curva ROC (Receiver Operating Characteristics), em particular do valor da área sob a curva (AUC ROC), do fator de enriquecimento para 1% ou 10% (EF1 e EF10, respectivamente) das moléculas com maior pontuação de predição (razão do número de moléculas ativas presentes nas fração superior da lista de compostos ordenados pela pontuação atribuída pelo modelo com relação a uma distribuição aleatória) e da área sob a curva *Boltzmann-enhanced discrimination of ROC* (AUC BEDROC), com $\alpha = 5$ (GEPPERT; VOGT; BAJORATH, 2010; TRUCHON; BAYLY, 2007). A curva ROC permite avaliar o desempenho do modelo em diferentes valores de pontuação de predição. Um modelo ideal apresentaria o AUC ROC igual a 1, enquanto a classificação aleatória resultaria em um valor de 0,5 (GARTNER et al., 2017). O fator de enriquecimento foi calculado utilizando a equação (8), na qual $\text{EF}_{x\%}$ representa o fator enriquecimento para uma porcentagem do banco de dados ($x\%$ do total); N_{total} e $N_{x\%}$ representam respectivamente o número total de compostos no banco de dados e na seleção ($x\%$ do total) e $\text{Ativos}_{x\%}$ e $\text{Ativos}_{\text{total}}$ representam respectivamente o número de compostos ativos no banco de dados completo e entre as moléculas selecionadas (EMPEREUR-MOT et al., 2015).

$$EF_{x\%} = Ativos_{x\%} * N_{total} / Ativos_{total} * N_{x\%} \quad (8)$$

3.2.6 Métodos de validação da capacidade preditiva dos modelos

Os métodos de validação cruzada quádrupla (*5-fold cross validation*), teste de aleatorização e validação externa são amplamente empregados para avaliar a robustez e a capacidade preditiva de modelos de aprendizado de máquina (TROPSHA; GRAMATICA; GOMBAR, 2003).

3.2.6.1 Validação cruzada quádrupla

A validação cruzada quádrupla consiste em dividir, de maneira aleatória para cada categoria (ativos e inativos), a série treino em cinco conjuntos (*folds*) contendo 20% dos dados cada. De maneira sequencial, um dos conjuntos é removido da série treino e um novo modelo computacional é gerado com os dados restantes, processo que se repete cinco vezes. Cada modelo é utilizado para prever a atividade dos compostos que não foram incluídos no seu desenvolvimento.

3.2.6.2 Teste de aleatorização

O teste de aleatorização consiste em misturar de maneira aleatória a variável independente dos dados (classificação dos compostos entre ativos ou inativos) e gerar novos modelos computacionais. Estes modelos devem apresentar capacidade preditiva significativamente reduzida, caso exista uma correlação verdadeira entre as variáveis.

3.2.6.3 Validação externa

A validação externa consiste em avaliar a capacidade preditiva de modelos para compostos ausentes na Série treino. Neste trabalho, as séries externas foram a Série teste A e Série teste B. A primeira validação externa (Série teste A) teve como objetivo auxiliar na seleção do modelo com a maior capacidade preditiva dentre todos os explorados. A segunda validação externa (Série teste B) foi realizada apenas para

o modelo de maior capacidade preditiva nas validações anteriores e teve como objetivo comprovar a sua performance com uma nova série externa.

3.2.7 Domínio de aplicabilidade

O domínio de aplicabilidade (DA) foi determinado com base na similaridade estrutural dos compostos para os quais a predição será feita com aqueles utilizados no desenvolvimento do modelo (Série treino). O cálculo foi realizado empregando a funcionalidade *Applicability Domain* da Enalos (AFANTITIS et al., 2011; MELGRAKI et al., 2010). Para cada modelo, o DA foi determinado com o mesmo *fingerprint* estrutural utilizado para a geração do modelo de predição.

Inicialmente, a distância euclidiana média para cada par de compostos da série treino foi determinada utilizando a equação (9), na qual A e B representam o número de fragmentos presentes em cada composto e C equivale ao número de fragmentos presentes em ambas as moléculas (CERETO-MASSAGUÉ et al, 2015).

$$\text{Distância euclidiana} = \sqrt{(A + B - 2C)} \quad (9)$$

A média de todas as distâncias euclidianas foi determinada. As distâncias que se mostraram abaixo deste valor inicial foram selecionadas para calcular uma nova média de distâncias (D) e o seu desvio padrão (DP). Estes parâmetros foram utilizados para estabelecer o valor referência do domínio de aplicabilidade (VRDA) através da equação (10) (ZHANG et al., 2006). Neste trabalho, o parâmetro Z foi mantido nas suas configurações padrão (Z = 0,5).

$$\text{VRDA} = D + Z * \text{DP} \quad (10)$$

A distância euclidiana entre cada composto avaliado pelo modelo e o seu vizinho mais próximo da série treino (composto com a menor distância) foi comparada com o VRDA. Compostos com distâncias superiores ao VRDA foram considerados fora do domínio de aplicabilidade e, portanto, com predições menos confiáveis. Estes compostos foram removidos das análises de capacidade preditiva dos modelos. A

porcentagem de compostos classificados como pertencentes ao domínio de aplicabilidade foi utilizada para estabelecer a cobertura do modelo.

3.2.8 Comparação do modelo de aprendizado de máquinas com métodos de triagem virtual por similaridade estrutural

A capacidade preditiva do melhor modelo desenvolvido neste trabalho foi comparada à triagem virtual por similaridade estrutural (BENDER; GLEN, 2004; MAGGIORA et al., 2014; RINIKER; LANDRUM, 2013), visando, assim, a avaliar se o modelo desenvolvido apresentava uma performance superior à de abordagens mais simples e computacionalmente menos custosas (MERZ et al., 2015; STUMPFE; BAJORATH, 2020). O índice de similaridade de Tanimoto foi selecionado por ser amplamente utilizado na busca de novos compostos ativos e como comparador no desenvolvimento de modelos preditivos (LAUFKÖTTER; MIYAO; BAJORATH, 2019; LIU et al., 2019; WILLETT, 2006). A comparação de capacidade preditiva foi realizada com a Série teste B, a qual já havia sido utilizada para comprovar a performance do modelo NB – FeatMorgan. O índice de Tanimoto de cada composto em análise foi calculado com relação às moléculas ativas da Série treino. No cálculo, utilizou-se o mesmo *fingerprint* estrutural do modelo de aprendizado de máquina a ser comparado. O valor de similaridade foi calculado utilizando a equação (11), na qual A e B representam os fragmentos de cada composto e C os fragmentos presentes em ambas as moléculas (BAJUSZ; RÁCZ; HÉBERGER, 2015).

$$\text{Índice de similaridade Tanimoto} = C / (A + B - C) \quad (11)$$

3.2.9 Fluxo geral do desenvolvimento, validação e aplicação dos modelos preditivos

Os modelos computacionais foram gerados considerando-se os princípios de boas práticas propostos pela OECD (*Organization for Economic Co-operation Development*) para o desenvolvimento e a validação de modelos de QSAR (relação estrutura-atividade quantitativa): (i) um desfecho claro, (ii) algoritmo não ambíguo, (iii) domínio de aplicabilidade definido, (iv) métricas adequadas de robustez e poder

preditivo e, se possível, (v) capacidade de interpretar as correlações entre os descritores e o desfecho em análise (OECD, 2007).

A Figura 3 apresenta um fluxograma com as principais etapas deste trabalho. Inicialmente, a Série treino foi utilizada para gerar diversos modelos através de combinações dos algoritmos computacionais com os *fingerprints* estruturais. Estes modelos foram validados utilizando os métodos de validação cruzada quádrupla, o teste de aleatorização e a validação externa com a Série teste A. A otimização de hiperparâmetros foi realizada durante a validação cruzada quádrupla com vistas a otimizar a métrica AUC ROC. Os 20 modelos classificatórios gerados e foram avaliados conforme as métricas anteriormente descritas (SE, SP, PPV, NPV, EX, TCB, AUC ROC, AUC BEDROC, EF1 e EF10).

O modelo NB - FeatMorgan, gerado a partir do método de aprendizado *Naive Bayes* e do *fingerprint* FeatMorgan, apresentou a maior capacidade preditiva nas validações iniciais e, deste modo, foi selecionado para as etapas seguintes. A performance do modelo NB – FeatMorgan foi confirmada em uma segunda validação externa, realizada com a Série teste B. Posteriormente, a capacidade preditiva deste modelo foi comparada à do método de triagem virtual por similaridade estrutural pelo coeficiente de Tanimoto.

Figura 3 – Fluxograma do desenvolvimento e validação de modelos computacionais

Fonte: Produção do autor.

3.2.10 Triagem virtual e análise preliminar dos fármacos selecionados

A triagem virtual de fármacos aprovados, de candidatos a fármaco em desenvolvimento clínico e de metabólitos de fármacos (Série de triagem A, B e C, respectivamente) utilizando o modelo NB – FeatMorgan possibilitou a identificação dos compostos com alto potencial de serem ativos contra o *T. cruzi*. Somente os compostos pertencentes ao domínio de aplicabilidade do modelo foram mantidos. Posteriormente, a similaridade estrutural das moléculas selecionadas foi determinada com relação aos compostos ativos da Série treino, utilizando o coeficiente de Tanimoto e o *fingerprint* FeatMorgan. O valor de similaridade e a identidade da molécula ativa de maior similaridade foram registrados para cada molécula selecionada.

O potencial de um composto ser reposicionado para o tratamento da doença de Chagas depende de diversos parâmetros além de sua eficácia na inibição do parasita, incluindo propriedades de farmacocinética e toxicidade adequadas para a indicação terapêutica. Neste contexto, as informações disponíveis para as moléculas selecionadas foram compiladas, com o objetivo de elaborar uma análise complementar do seu potencial para o reposicionamento. Os bancos de dados utilizados foram *DrugBank* (WISHART et al., 2018), *Global Data* (2020), *PubMed*, *Repurposing Hub* (CORSELLO et al., 2017) e *NCATS Inxight: Drugs* (NCATS, 2020).

No caso de fármacos e candidatos a fármaco, os critérios avaliados compreenderam a situação regulatória do composto e as vias de administração das formas farmacêuticas disponíveis no mercado ou em desenvolvimento.

3.2.11 Avaliação experimental de fármacos selecionados pelo modelo NB – FeatMorgan

3.2.11.1 Fármacos aprovados selecionados para a avaliação experimental

Dos 269 fármacos selecionados na triagem virtual, 17 foram obtidos para a avaliação experimental com base na sua disponibilidade. Amostras de flibanserina, dexclorfeniramina, mononitrato de isossorbida, ranitidina, aciclovir, bronfeniramina, sildenafil, betaistina, nimesulida, ambroxol, minoxidil, risedronato, sertralina, desloratadina, prazosina, leflunomida e doxazosina foram doadas pelo Aché Laboratórios (Guarulhos, São Paulo, Brasil). O benznidazol, fármaco de referência, foi adquirido da Sigma Aldrich (St. Louis, MO, Estados Unidos).

3.2.11.2 Ensaio de inibição *in vitro* do *T. cruzi*

Os experimentos de inibição *in vitro* do *T. cruzi* foram realizados pela pesquisadora Carolina Borsoi Moraes, do Instituto de Ciências Biomédicas da USP, seguindo um protocolo de triagem de alto conteúdo (*high content screening*) previamente validado (FRANCO, 2018; MORAES et al., 2014). O experimento foi realizado com células de osteossarcoma humano, U2OS, infectadas com o clone Y do parasita. As células hospedeiras foram adicionadas a placas de 384 poços e incubadas por 24 horas em meio DMEM completo. Em seguida, o *T. cruzi* foi adicionado ao meio do ensaio na densidade de 20 tripomastigotas por célula hospedeira semeada. Após 24 horas, os fármacos, os controles positivo e negativo (células infectadas e não infectadas com o parasita, respectivamente, tratadas com uma solução contendo 1% de dimetilsulfóxido – DMSO) e o benznidazol foram adicionados e as placas incubadas por 72 horas. Todas as incubações foram realizadas a 37 °C e em atmosfera com 5% de CO₂. Posteriormente, as placas foram fixadas com uma solução de paraformaldeído 4% em PBS (PFA) e marcadas com o corante de DNA DRAQ5 (*Deep Red AnthraQuinone 5*, Biostatus, Reino Unido). As

imagens foram obtidas utilizando o equipamento *IN Cell Analyzer 2500* (Cytiva, Estados Unidos). O programa *Investigator* (GE Healthcare, Estados Unidos) foi empregado para determinar o número de células hospedeiras e as taxas de infecção pelo parasita (relação entre o número de células infectadas sobre o número total de células). A porcentagem de células hospedeiras foi normalizada pelo número de células no controle negativo. A taxa de infecção, por sua vez, foi normalizada pelos valores observados para os controles negativo e positivo, conforme a equação (12).

$$\text{Taxa de infecção normalizada} = 1 - \frac{(\text{resultado} - \text{controle positivo})}{(\text{controle negativo} - \text{controle positivo})} \quad (12)$$

Os fármacos foram testados em 10 concentrações em um formato de diluição seriada de duas vezes a partir da concentração máxima de 50 µM. O benznidazol, por sua vez, foi avaliado em 16 concentrações a partir do máximo de 400 µM, utilizando o mesmo formato de diluição seriada. Os experimentos foram realizados em duplicata. As curvas sigmóides de concentração-resposta foram geradas com o programa GraphPad Prism v8.4.3 através de um ajuste com inclinação variável. Os valores de IC₅₀ e CC₅₀ foram determinados, respectivamente, como a concentração que inibe em 50% a infecção do parasita com relação à atividade máxima observada e como a concentração citotóxica capaz de reduzir em 50% o número de células hospedeiras com relação ao controle negativo. A atividade máxima (A_{max}) foi determinada como a máxima atividade antiparasitária normalizada observada na curva de concentração-resposta. Os valores de IC₅₀ e A_{max} representam, respectivamente, a potência e a eficácia dos compostos na inibição do parasita. O índice de seletividade foi calculado pela razão entre a toxicidade (CC₅₀) e a atividade anti-*T. cruzi* (IC₅₀).

3.3 RESULTADOS

3.3.1 Análise da literatura e identificação de compostos testados experimentalmente contra o *T. cruzi*

A revisão da literatura resultou na identificação de diversos trabalhos que reportaram a triagem experimental de bibliotecas de compostos em ensaios de inibição *in vitro* do *T. cruzi*. A Tabela 1 apresenta um resumo das 13 principais

publicações encontradas, incluindo informações sobre o protocolo do ensaio, a composição da biblioteca e os resultados experimentais disponibilizados. Observou-se que a identidade das moléculas inativas raramente foi reportada de maneira explícita e o número de compostos com dados de IC₅₀ disponíveis é bastante reduzido, o que limita o potencial de mapear todos os compostos já testados contra o parasita e de gerar modelos preditivos para o *T. cruzi*.

Tabela 1 – Resumo das principais publicações reportando a triagem experimental de compostos em ensaios fenotípicos de inibição *in vitro* do parasita *T. cruzi*

Biblioteca avaliada	Linhagem do parasita (Célula hospedeira)	Ensaio	Número de compostos com IC ₅₀ reportados (Toxicidade disponível?)	Referência
303.224 compostos do Instituto Broad	Tulahuen CL2 (3T3)	%Prolif.	4020 compostos (Sim)	CARMODY et al., 2010
1,8 milhão de compostos da GSK	Tulahuen (3T3 e H9c2)	HCS (Am/Cel., %Infec.) e %Prolif.	580 compostos (Sim)	PEÑA et al., 2015
400 compostos do Malaria Box	Tulahuen (MRC-5)	%Prolif.	400 compostos (Sim)	KAISER et al., 2015 ^a
100 fármacos	Tulahuen (L6)	%Prolif.	100 compostos (Não)	KAISER et al., 2015b
727 compostos do NIH Clinical collection + 421 fármacos aprovados	Silvio X10/7 A1 (Vero)	HCS (Am/Cel. e %Infec.)	69 compostos (Sim)	DE RYCKER et al., 2016
909 fármacos, a maioria aprovados pelo FDA	CA-I/72 (BESM)	HCS (Am/Cel.)	55 compostos (Sim)	ENGEL et al., 2010
2.000 compostos bioativos da MicroSource Discovery Systems	Tulahuen (3T3)	%Prolif.	35 compostos (Sim)	PLANER et al., 2014
376 compostos da Pathogen Box	Tulahuen (3T3)	HCS (%Infec.)	31 compostos (Sim)	DUFFY et al., 2017
~150.000 compostos da Diversity Library do Calibr Scripps Research	Tulahuen (H9c2)	HCS (Am/Cel. e %Infec.)	21 compostos (Sim)	ROQUERO et al., 2019
160.000 compostos da Novartis Academic Collaboration Library	CA-I/72 (C2C12)	HCS (Am/Cel., %Infec.)	15 compostos (Não)	NEITZ et al., 2015
741 compostos bioativos + 685 compostos da Malaria Box	Tulahuen (3T3)	HCS (%Infec.)	15 compostos (Não)	SYKES e AVERY, 2015
7.680 compostos da biblioteca ReFRAME	CA-I/72 (C2C12)	HCS (Am/Cel.)	101 compostos (Sim)	BERNATCHE Z et al., 2020
~ 3 milhões de compostos do GNF (Novartis)	CL (3T3)	%Prolif.	1 composto + 5 análogos (Sim)	KHARE et al., 2016

O desfecho avaliado nos ensaios foi o número de amastigotas por célula (Am/Cel) ou a porcentagem de células infectadas (%Infec.) no ensaio de *high content screening* (HCS) ou a proliferação do parasita avaliada pela atividade de β-galactosidase (%Prolif.). Vero – Células epiteliais de rim de macaco. 3T3 – Fibroblastos de camundongo. C2C12 – Mioblasto de camundongo. L6 – Mioblasto de ratos. MRC-5 – Fibroblastos de pulmão de humanos. H9c2 – Célula muscular cardíaca de rato. BESM – Músculo esquelético bovino.

Os trabalhos que avaliaram experimentalmente bibliotecas contendo fármacos aprovados foram comparados com relação aos compostos ativos identificados em cada triagem. Nesta análise, os critérios de atividade estabelecidos no estudo original foram mantidos. Conforme demonstrado na Figura 4 observou-se que cinco fármacos (prazosina, clomipramina, flufenazina, azelastina e fluoxetina) foram identificados como ativos nos três trabalhos comparados, evidenciando uma pequena redundância dos resultados. Por outro lado, 103 compostos foram descritos como ativos contra o *T. cruzi* em somente uma das três publicações avaliadas – 36 por de Rycker et al. (2016), 30 por Engel et al. (2010) e 37 por Planer et al. (2014). Ainda, foi possível constatar que 20 dos 97 fármacos classificados como ativos por Engel et al. (2010) e de Rycker et al. (2016) estavam presentes na biblioteca testada no trabalho de Planer et al. (2014), porém não foram selecionados para a determinação de IC₅₀. Como os compostos inativos dos trabalhos de Engel et al. (2010) e de Rycker et al. (2016) não foram descritos de maneira explícita, não foi possível realizar comparações adicionais. Ressalta-se, ainda, que, conforme o esperado, nenhuma publicação identificada no presente levantamento bibliográfico reportou a avaliação experimental de todos os compostos (fármacos, candidatos a fármaco e metabólitos) que poderiam ser reposicionados para Chagas.

Figura 4 – Diagrama de Venn dos fármacos ativos contra *T. cruzi* identificados em diferentes campanhas de triagem experimental

A – Fármacos ativos descritos por Engel et al. (2010). B – Fármacos ativos descritos por de Rycker et al. (2016). C – Fármacos ativos descritos por Planer et al. (2014). Fonte: Produção do autor.

3.3.2 Desenvolvimento e validação dos modelos

Os modelos de aprendizado de máquina foram desenvolvidos a partir de combinações dos algoritmos de aprendizado e dos *fingerprints* estruturais. Os hiperparâmetros dos algoritmos foram otimizados e, posteriormente, os modelos foram validados utilizando os métodos de validação cruzada quádrupla, teste de aleatorização e validação externa. Os hiperparâmetros que resultaram nos maiores valores de AUC ROC para os modelos baseados no algoritmo PNN foram $\text{teta menos} = 0,1$ e $\text{teta mais} = 0,4$ para o *fingerprint* FCFP6 e $\text{teta menos} = 0,1$ e $\text{teta mais} = 0,5$ para os demais *fingerprints*. No caso do RF, o número ótimo de árvores de decisão foi 50 para o FCFP6 e 400 para os demais. Na otimização do SVM, a função de kernel linear resultou na melhor performance para os *fingerprints* FCFP6 e o Extended, a função sigmoide para o ECFP6 e o FeatMorgan e a função radial para o Morgan. Os valores otimizados de gamma foram 0 para os *fingerprints* ECFP6, Morgan e FeatMorgan, 10^{-6} para o FCFP6 e 10^{-4} para o Extended. Finalmente, o hiperparâmetro $c = 10$ resultou na melhor performance para todos os modelos de SVM.

3.3.2.1 Validação cruzada quántupla

A performance dos diferentes modelos computacionais na validação cruzada quántupla está sumarizada nas Figuras 5, 6 e 7 e detalhada na Tabela 16. A análise dos resultados permitiu constatar que ambos os parâmetros explorados (método de aprendizado e *fingerprints* estruturais), incluindo suas combinações, influenciaram a capacidade preditiva dos modelos obtidos. De modo geral, vários dos modelos desenvolvidos com os algoritmos NB, RF e SVM apresentaram boas performances, o que pode ser evidenciado pelos valores de sensibilidade, especificidade e taxa de classificação balanceada próximos ou superiores a 80% e pelas áreas sob as curvas ROC e BEDROC próximas ou superiores a 0,8. Os modelos gerados a partir do PNN em combinação com os *fingerprints* Extended, FCFP6 e FeatMorgan apresentaram desempenho mediano, com TCB entre 55% e 70%. Esta baixa performance pode ser constatada pelos valores reduzidos de sensibilidade (< 48,6% para os *fingerprints* ECFP6, FCFP6, Morgan e FeatMorgan) ou especificidade (< 43,5% para o *fingerprint* Extended). Surpreendentemente, os modelos gerados a partir do algoritmo PNN apresentaram um alto valor de AUC ROC e AUC BEDROC, os quais são métricas que consideram apenas o valor de pontuação atribuído aos compostos, porém não a classificação final, *i.e.*: potencialmente ativos ou inativos. Como um todo, estes resultados evidenciam que o PNN foi capaz de ordenar os compostos de maneira eficaz, porém não conseguiu classificá-los como ativos ou inativos corretamente (Figura 6).

Figura 5 – Sensibilidade, especificidade, poder preditivo negativo e positivo dos modelos de aprendizado de máquina na validação cruzada quíntupla

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado de máquinas e de *fingerprints* estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). A capacidade preditiva na validação cruzada quíntupla foi avaliada utilizando os parâmetros: (A) especificidade e sensibilidade e (B) poder preditivo positivo e negativo. NB – *Naive Bayes*. RF - Floresta aleatória. PNN - Redes neurais probabilísticas. SVM – Máquinas de vetores de suporte. Fonte: Produção do autor.

Figura 6 – Áreas sob as curvas ROC e BEDROC e taxa de classificação balanceada dos modelos de aprendizado de máquina na validação cruzada quíntupla

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado de máquina e de *fingerprints* estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). A capacidade preditiva na validação cruzada quíntupla foi avaliada utilizando os parâmetros: (A) área sob a curva BEDROC (AUC BEDROC), (B) área sob a curva ROC (AUC ROC) e (C) a taxa de classificação balanceada (TCB). NB – *Naive Bayes*. RF - Floresta aleatória. PNN - Redes neurais probabilísticas. SVM - Máquinas de vetores de suporte. Fonte: Produção do autor.

Figura 7 – Curvas ROC dos modelos de aprendizado de máquina na validação cruzada quádrupla

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado de máquinas e de *fingerprints* estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). A capacidade preditiva na validação cruzada quádrupla foi avaliada utilizando a curva ROC. Fonte: Produção do autor.

3.3.2.2 Teste de aleatorização

A performance dos diferentes modelos computacionais no teste de aleatorização está sumarizada nas Figuras 8, 9 e 10 e detalhada na Tabela 17. Esta etapa resultou em modelos com a capacidade preditiva comprometida, conforme esperado. Os valores das áreas sob as curvas ROC e BEDROC foram próximos de 0,5 e a taxa de classificação balanceada de 50%. A perda de performance permitiu confirmar que a relação entre a atividade biológica modelada e os descritores

estruturais dos compostos não foi obtida ao acaso (TROPSHA; GRAMATICA; GOMBAR, 2003).

Figura 8 – Sensibilidade, especificidade, poder preditivo negativo e positivo dos modelos de aprendizado de máquina no teste de aleatorização

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado de máquinas e de *fingerprints* estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). Previamente, as classes dos compostos (ativos e inativos) foram aleatorizadas. A capacidade preditiva no teste de aleatorização foi avaliada utilizando os parâmetros: (A) especificidade e sensibilidade e (B) poder preditivo positivo e negativo. NB – *Naive Bayes*. RF - Floresta aleatória. PNN - Redes neurais probabilísticas. SVM - Máquinas de vetores de suporte. Fonte: Produção do autor.

Figura 9 – Áreas sob as curvas ROC e BEDROC e taxa de classificação balanceada dos modelos de aprendizado de máquina no teste de aleatorização

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado de máquinas e de *fingerprints* estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). Previamente, as classes dos compostos (ativos e inativos) foram aleatorizadas. A capacidade preditiva no teste de aleatorização foi avaliada utilizando os parâmetros: (A) área sob a curva BEDROC (AUC BEDROC), (B) área sob a curva ROC (AUC ROC) e (C) a taxa de classificação balanceada (TCB). NB – *Naive Bayes*. RF - Floresta aleatória. PNN - Redes neurais probabilísticas. SVM - Máquinas de vetores de suporte. Fonte: Produção do autor.

Figura 10 – Curvas ROC dos modelos de aprendizado de máquina no teste de aleatorização

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado de máquinas e de *fingerprints* estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). Previamente, as classes dos compostos (ativos e inativos) foram aleatorizadas. A capacidade preditiva no teste por aleatorização foi avaliada utilizando a curva ROC. Fonte: Produção do autor.

3.3.2.3 Validação externa com a Série teste A

A performance dos diferentes modelos computacionais na validação externa com a Série teste A está sumarizada nas Figuras 11, 12 e 13 e detalhada na Tabela 18. A maioria dos compostos da série externa foram classificadas como pertencentes ao domínio de aplicabilidade dos modelos pelo método de similaridade estrutural, o que possibilitou que, em média, 97% - 98% das moléculas fosse utilizada para a

análise de desempenho dos métodos. Alguns dos modelos desenvolvidos com os algoritmos NB, RF e SVM apresentaram alta capacidade preditiva, novamente evidenciada pelos valores de sensibilidade, especificidade e taxa de classificação balanceada próximos ou superiores a 80% e pelas áreas sob as curvas ROC e BEDROC próximas ou superiores a 0,8. De modo geral, os modelos que apresentaram alta performance nesta etapa foram aqueles que já haviam demonstrado uma alta performance na validação cruzada. Novamente, o algoritmo PNN resultou em modelos que conseguiram ordenar os compostos ativos e inativos de maneira eficaz (altos valores de AUC ROC e AUC BEDROC), porém com baixa performance na classificação final atribuída às moléculas (baixa sensibilidade, especificidade, taxa de classificação balanceada, entre outros).

A comparação entre os 20 modelos gerados evidenciou que nenhum deles apresentou performance superior aos demais em todas as métricas avaliadas. Por esta razão, a estratégia adotada foi selecionar inicialmente os modelos com desempenho consistente para todas as validações e parâmetros avaliados. Neste sentido, somente aqueles que exibiram valores de sensibilidade, especificidade, TCB, PPP e PPN superiores a 70% nas validações cruzada e externa foram mantidos (NB – FeatMorgan, RF – FCFP6, RF – Extended, RF – FeatMorgan, SVM – ECFP6, SVM – FCFP6, SVM – Extended, SVM – Morgan e SVM – FeatMorgan). Posteriormente, os modelos selecionados na primeira etapa foram comparados com base nos valores de AUC ROC e AUC BEDROC. Esta análise permitiu constatar que o NB – FeatMorgan apresentou os maiores valores de AUC ROC e AUC BEDROC tanto na validação cruzada quanto na validação externa. Por esta razão, este modelo foi considerado aquele de melhor performance neste trabalho e, deste modo, selecionado para a confirmação de sua capacidade preditiva com uma segunda série externa (Série teste B) e para as triagens virtuais.

Figura 11 – Sensibilidade, especificidade, poder preditivo negativo e positivo dos modelos de aprendizado de máquina na validação externa com a Série teste A

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado de máquinas e de *fingerprints* estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). A capacidade preditiva na validação externa com a série teste A foi avaliada utilizando os parâmetros: (A) especificidade e sensibilidade e (B) poder preditivo positivo e negativo. NB – *Naive Bayes*. RF - Floresta aleatória. PNN - Redes neurais probabilísticas. SVM – Máquinas de vetores de suporte. Fonte: Produção do autor.

Figura 12 – Áreas sob as curvas ROC e BEDROC e taxa de classificação balanceada dos modelos de aprendizado de máquina na validação externa com a Série teste A

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado de máquinas e de *fingerprints* estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). A capacidade preditiva na validação externa com a série teste A foi realizada utilizando os parâmetros: (A) área sob a curva BEDROC (AUC BEDROC), (B) área sob a curva ROC (AUC ROC) e (C) a taxa de classificação balanceada (TCB). NB – *Naive Bayes*. RF - Floresta aleatória. PNN - Redes neurais probabilísticas. SVM - Máquinas de vetores de suporte. Fonte: Produção do autor.

Figura 13 – Curvas ROC dos modelos de aprendizado de máquina na validação externa com a Série teste A

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado de máquinas e de *fingerprints* estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). A capacidade preditiva na validação externa com a série teste A foi avaliada utilizando a curva ROC. Fonte: Produção do autor.

3.3.2.4 Validação externa com a Série teste B

A performance do modelo NB - FeatMorgan na validação externa com a Série teste B está sumarizado na Figura 14. A maioria dos compostos desta biblioteca (98,5%) foi classificada como pertencente ao domínio de aplicabilidade e, por isso, pôde ser utilizada nesta etapa. O modelo apresentou alta capacidade preditiva, conforme esperado, a qual pôde ser comprovada pelos valores de taxa de classificação balanceada, sensibilidade, especificidade próximos ou superiores a 80%

e pela área sob a curva ROC de 0,85. Embora a área sob a curva BEDROC tenha sido consideravelmente inferior ao resultado observado nas validações anteriores, o modelo foi capaz de enriquecer as frações de maior pontuação com moléculas ativas, conforme evidenciado pelos fatores de enriquecimento para 1% (EF1 = 14,3) e 10% (EF10 = 4,8) do banco de dados.

Figura 14 – Performance do modelo NB – FeatMorgan na validação externa com a Série teste B

O modelo foi desenvolvido a partir da Série treino utilizando o algoritmo *Naive Bayes* e o *fingerprint* estrutural FeatMorgan. Os parâmetros avaliados foram: (A) sensibilidade (SENS.), especificidade (ESPEC.), taxa de classificação balanceada (TCB), poder preditivo positivo (PPV) e negativo (NPV) e a porcentagem da série teste dentro do domínio de aplicabilidade (Cov.); (B) curva ROC; (C) fator de enriquecimento e (D) áreas sob as curvas ROC e BEDROC. EF1 e EF10 – Fatores de enriquecimento para 1% e 10% dos compostos com maior pontuação. Fonte: Produção do autor.

3.3.3 Comparação da performance do modelo NB – FeatMorgan com métodos de triagem virtual por similaridade estrutural

A capacidade preditiva da triagem virtual por similaridade estrutural e sua comparação com os resultados obtidos para o modelo NB – FeatMorgan estão

sumarizadas na Figura 15. A busca por similaridade foi capaz de enriquecer as frações de maior pontuação (EF1 = 9,5 e EF10 = 2,8) e apresentou uma performance superior à seleção aleatória (AUC ROC > 0,5). Por outro lado, o método de similaridade resultou em uma performance inferior àquela do modelo NB – FeatMorgan. Como ambas as abordagens foram desenvolvidas a partir das mesmas bibliotecas de compostos (Série treino), este resultado evidencia o ganho de capacidade preditiva proveniente do desenvolvimento de modelos de aprendizado de máquina para a triagem virtual de compostos.

Figura 15 – Performance do modelo NB – FeatMorgan em comparação com o índice de similaridade Tanimoto na seleção de compostos ativos da Série teste B

O modelo NB – FeatMorgan e o índice de similaridade Tanimoto foram utilizados para realizar a triagem virtual de compostos da Série teste B. A performance dos métodos foi comparada utilizando os parâmetros: (A) Curvas ROC, (B) áreas sob as curvas ROC e BEDROC e (C) Fator de enriquecimento para 1% e 10% dos dados com maior classificação (EF1 e EF10, respectivamente). Fonte: Produção do autor.

3.3.4 Triagem virtual e análise dos compostos selecionados

O modelo NB - FeatMorgan foi utilizado na triagem virtual de bibliotecas de fármacos aprovados, candidatos a fármaco em desenvolvimento clínico e metabólitos de fármacos (Séries de triagem A, B e C, respectivamente). Os compostos de cada biblioteca foram organizados em ordem decrescente da pontuação gerada pelo modelo (Figura 16). Os compostos de menor pontuação que ainda foram classificados como ativos nas Séries de triagem A, B e C apresentaram os valores de 0,16, 0,18 e 0,16, respectivamente. Para os fármacos aprovados, 2.367 compostos foram incluídos na análise, 2.227 (94%) foram considerados como pertencentes ao domínio de aplicabilidade, dos quais 331 (14,9%) foram classificados como ativos. No caso dos candidatos a fármaco, 1.618 foram avaliados, 1.524 (94,2%) foram considerados dentro do domínio de aplicabilidade, dos quais 332 (21,8%) foram classificados como ativos. Para os metabólitos de fármacos, 2.331 compostos foram avaliados, 2.197 (94,2%) estavam dentro do domínio de aplicabilidade, dos quais 217 (9,9%) classificados como ativos.

A análise detalhada dos compostos selecionados levou à exclusão de respectivamente 63, 40 e 11 moléculas das Séries de triagem A, B e C. Estas moléculas foram excluídas porque não foi possível confirmar que as mesmas cumpriam os requisitos das suas respectivas bibliotecas (ex.: moléculas no estágio pré-clínico que haviam sido incluídas na lista candidatos a fármaco em desenvolvimento clínico). Constatou-se, ainda, que o isavuconazol, presente inicialmente na Série de triagem B, já havia sido aprovado no momento em que esta análise foi realizada. Portanto, esta molécula foi inserida na relação de fármacos aprovados. Os 269 fármacos, 291 compostos em desenvolvimento clínico e 206 metabólitos selecionados estão apresentados nas Tabelas 19, 20 e 21.

Figura 16 – Pontuação das Séries de triagem A, B e C na triagem virtual utilizando o modelo NB – FeatMorgan

A triagem virtual foi realizada com o modelo NB – FeatMorgan e as pontuações foram organizadas de maneira decrescente. A linha tracejada representa a divisão entre os compostos classificados como ativos ou inativos. A – Série de triagem A, composta por fármacos aprovados. B – Série de triagem B, compostos por candidatos a fármaco em desenvolvimento clínico. C – Série de triagem C, composta por metabólitos de fármacos. Fonte: Produção do autor.

A análise dos fármacos aprovados evidenciou ainda que 35 dos compostos selecionados já haviam sido testados em ensaios de inibição *in vitro* do *T. cruzi* e apresentado atividade biológica considerável contra o parasita, detalhada a seguir. Destes 35 compostos, 19 possuíam potência e seletividade dentro dos critérios estabelecidos pelo DNDi ($IC_{50} < 5 \mu M$ e $SI > 10$) e, por isso, teriam sido classificados como ativos neste trabalho (Tabela 2). Notavelmente, o benznidazol e o nifurtimox, aprovados para o tratamento da doença de Chagas, além de outros fármacos contendo o grupamento nitro (nitrofurantoína, furazolidona e nitrofurazona), estão entre as moléculas selecionadas. Ainda, 9 dos 19 fármacos são antifúngicos azólicos ou triazólicos, os quais atuam através da inibição da enzima TcCYP51. Além das

moléculas anteriormente descritas, a lista dos 19 fármacos continha também a ziprasidona ($IC_{50} = 1,58 \mu M$ e $SI = 15,8$), a primaquina ($IC_{50} = 0,30 \mu M$ e $SI > 26,3$), o triantereno ($IC_{50} = 1,58 \mu M$ e $SI = 12,6$), o alopurinol ($IC_{50} = 2,51 \mu M$ e $SI > 20$) e o entecavir ($IC_{50} = 2,2 \mu M$ e $SI > 45$). Estes compostos não pertencem às classes químicas anteriormente descritas e, por isso, têm maior potencial de atuarem na inibição do *T. cruzi* por novos mecanismos de ação.

No caso dos 16 fármacos restantes, 6 compostos (Tabela 3) apresentaram alta potência contra o parasita ($IC_{50} \leq 5 \mu M$), porém com um índice de seletividade inferior ao critério estabelecido ($SI < 10$). As outras 10 moléculas tiveram potência intermediária ($5 \mu M < IC_{50} < 10 \mu M$) e um índice de seletividade também inferior ao adequado. Estas dez últimas moléculas são a prazosina ($IC_{50} = 5,01 \mu M$ e $CC_{50} = 10 \mu M$), a meclizina ($IC_{50} = 5,01 \mu M$ e $CC_{50} = 20 \mu M$), a doxazosina ($IC_{50} = 5,01 \mu M$ e $CC_{50} = 15,9 \mu M$), o imatinibe ($IC_{50} = 5,01 \mu M$ e $CC_{50} = 15,9 \mu M$), o estanozolol ($IC_{50} = 5,01 \mu M$ e $CC_{50} = 15,9 \mu M$), a clomipramina ($IC_{50} = 6,31 \mu M$ e $CC_{50} = 15,9 \mu M$), a desloratadina ($IC_{50} = 6,31 \mu M$ e $CC_{50} = 15,9 \mu M$), a duloxetina ($IC_{50} = 6,31 \mu M$ e $CC_{50} = 15,9 \mu M$), o triclosan ($IC_{50} = 7,94 \mu M$ e $CC_{50} = 10 \mu M$) e o haloperidol ($IC_{50} = 7,0 \mu M$ e $CC_{50} = 33,0 \mu M$) (DE RYCKER et al., 2016; ENGEL et al., 2010; KAISER et al., 2015b). A ausência de dados experimentais em *T. cruzi* para todos os fármacos selecionados na triagem virtual limita a capacidade de expandir esta análise e obter valores mais precisos do poder preditivo do modelo.

Tabela 2 – Fármacos selecionados pelo modelo NB – FeatMorgan que são descritos como ativos contra o *T. cruzi* com $IC_{50} \leq 5 \mu M$ e $SI > 10$ (continua)

Fármaco (Pontuação no modelo) Atividade biológica da literatura	Composto ativo da Série treino de maior similaridade estrutural ao fármaco
<p>Ziprasidona (1,7) $IC_{50} = 1,58 \mu M^a$; $CC_{50} = 25,1 \mu M^a$</p>	<p>TCMDC-143362 Tanimoto = 0,24</p>
<p>Primaquina (0,4) $IC_{50} = 0,3 \mu M^b$; $CC_{50} = 7,9 \mu M^b$</p>	<p>TCMDC-143603 Tanimoto = 0,47</p>

Tabela 2 – Fármacos selecionados pelo modelo NB – FeatMorgan que são descritos como ativos contra o *T. cruzi* com $IC_{50} \leq 5 \mu M$ e $SI > 10$ (continuação)

Fármaco (Pontuação no modelo) Atividade biológica da literatura	Composto ativo da Série treino de maior similaridade estrutural ao fármaco
Triamtereno (3,31) $IC_{50} = 1,66 \mu M^b$; $CC_{50} = 18,4 \mu M^b$	TCMDC-143569 Tanimoto = 0,22
Alopurinol (1,07) $IC_{50} = 2,8 \mu M^b$; $CC_{50} > 50 \mu M^b$	TCMDC-143083 Tanimoto = 0,23
Entecavir (2,4) $IC_{50} = 2,2 \mu M^c$; $CC_{50} > 100 \mu M^c$	TCMDC-143641 Tanimoto = 0,19
Voriconazol (5,2) $IC_{50} = 0,025 \mu M^a$; $CC_{50} = 50,1 \mu M^a$	TCMDC-143082 Tanimoto = 0,39
Bifonazol (2,7) $IC_{50} = 0,003 \mu M^d$ e $0,01 \mu M^a$; $CC_{50} = 39,3 \mu M^d$ e $15,8 \mu M^a$	TCMDC-143275 Tanimoto = 0,25
Itraconazol (2,1) $IC_{50} = 0,004 \mu M^d$ e $0,01 \mu M^a$; $CC_{50} = 1,1 \mu M^d$ e $15,8 \mu M^a$	TCMDC-143082 Tanimoto = 0,32

Tabela 2 – Fármacos selecionados pelo modelo NB – FeatMorgan que são descritos como ativos contra o *T. cruzi* com $IC_{50} \leq 5 \mu M$ e $SI > 10$ (continuação)

Fármaco (Pontuação no modelo) Atividade biológica da literatura	Composto ativo da Série treino de maior similaridade estrutural ao fármaco
Miconazol (3,3) $IC_{50} = 0,04 \mu M^d$ e $0,032 \mu M^a$; $CC_{50} = 15,4 \mu M^d$ e $7,9 \mu M^a$	TCMDC-143590 Tanimoto = 0,37
Econazol (4,1) $IC_{50} = 0,04 \mu M^d$ e $0,02 \mu M^a$; $CC_{50} = 15,6 \mu M^d$ e $10 \mu M^a$	TCMDC-143423 Tanimoto = 0,38
Clotrimazol (1,2) $IC_{50} = 0,006 \mu M^a$; $CC_{50} = 12,6 \mu M^a$	TCMDC-143423 Tanimoto = 0,28
Tioconazol (3) $IC_{50} = 0,064 \mu M^d$; $CC_{50} = 19,5 \mu M^d$	TCMDC-143590 Tanimoto = 0,34
Cetoconazol (4,2) $IC_{50} = 0,005 \mu M^a$ e $0,270 \mu M^d$; $CC_{50} = 31,6 \mu M^a$ e $51 \mu M^d$	TCMDC-143590 Tanimoto = 0,35
Terconazol (2,7) $IC_{50} < 0,4 \mu M^e$; $CC_{50} = 24 \mu M^e$	TCMDC-143082 Tanimoto = 0,36

Tabela 2 – Fármacos selecionados pelo modelo NB – FeatMorgan que são descritos como ativos contra o *T. cruzi* com $IC_{50} \leq 5 \mu M$ e $SI > 10$ (conclusão)

Fármaco (Pontuação no modelo) Atividade biológica da literatura	Composto ativo da Série treino de maior similaridade estrutural ao fármaco
Nitrofurantoina (4,1) $IC_{50} = 4,3 \mu M^d$; $CC_{50} = 90,3 \mu M^d$	TCMDC-143517 Tanimoto = 0,31
Furazolidona (1,2) $IC_{50} < 0,4 \mu M^e$; $CC_{50} > 50 \mu M^e$	TCMDC-143577 Tanimoto = 0,31
Nitrofurazona (1) $IC_{50} = 1 \mu M^a$; $CC_{50} = 15,8 \mu M^a$	TCMDC-143571 Tanimoto = 0,35
Nifurtimox (2,6) $IC_{50} = 0,19 \mu M^d$; $CC_{50} = 87 \mu M^d$	TCMDC-143517 Tanimoto = 0,30
Benznidazol (1,0) $IC_{50} = 0,62 \mu M^b$; $CC_{50} > 25 \mu M^b$	TCMDC-143126 Tanimoto = 0,28

^a (DE RYCKER et al., 2016). ^b (PLANER et al., 2014). ^c (FRANCO et al., 2019). ^d (KAISER et al., 2015b). ^e (ENGEL et al., 2010). IC_{50} e CC_{50} – Concentração de fármaco capaz de inibir em 50% a proliferação do parasita ou da célula hospedeira, respectivamente. SI - índice de seletividade, calculado pela relação entre os valores de CC_{50} e IC_{50} . TCMDC – Compostos da biblioteca Tres Cantos Medicines Development Campus.

Tabela 3 – Fármacos selecionados pelo modelo NB – FeatMorgan que são descritos como ativos contra o *T. cruzi* com $IC_{50} \leq 5 \mu M$ e $SI < 10$

Fármaco (Pontuação no modelo) Atividade biológica da literatura	Composto ativo da Série treino de maior similaridade estrutural ao fármaco
Sertralina (0,3) $IC_{50} = 1,6 \mu M^a$; $CC_{50} = 7,9 \mu M^a$	TCMDC-143409 Tanimoto = 0,2
Fluoxetina (0,3) $IC_{50} = 4,0 \mu M^a$; $CC_{50} = 15,8 \mu M^a$	TCMDC-143310 Tanimoto = 0,24
Paroxetina (1,1) $IC_{50} = 4,3 \mu M^a$; $CC_{50} = 18,8 \mu M^a$	TCMDC-143334 Tanimoto = 0,23
Lomerizina (0,2) $IC_{50} = 4,0 \mu M^b$; $CC_{50} = 15,8 \mu M^b$	TCMDC-143504 Tanimoto = 0,33
Pirimetamina (1,2) $IC_{50} = 3,8 \mu M^a$; $CC_{50} = 28,5 \mu M^a$	TCMDC-143220 Tanimoto = 0,26
Niclosamida (0,8) $IC_{50} = 2,5 \mu M^c$ e $CC_{50} = 2,3 \mu M^c$	TCMDC-143609 Tanimoto = 0,31

^a (PLANER et al. 2014). ^b (DE RYCKER et al., 2016). ^c (KAISER et al., 2015b). IC_{50} e CC_{50} – Concentração de fármaco capaz de inibir em 50% a proliferação do parasita ou da célula hospedeira, respectivamente. SI - índice de seletividade, calculado pela relação entre os valores de CC_{50} e IC_{50} . TCMDC – Compostos da biblioteca Tres Cantos Medicines Development Campus.

Na análise das bibliotecas de compostos em desenvolvimento clínico e de metabólitos, constatou-se que a disponibilidade de dados experimentais de inibição do *T. cruzi* era significativamente menor para estes grupos de compostos em comparação com os fármacos aprovados. Dentre os candidatos a fármaco selecionados pelo modelo, 4 compostos (Tabela 4) já haviam sido descritos como ativos contra o parasita ($IC_{50} < 5 \mu M$), porém somente a dapivirina apresentava um índice de seletividade dentro dos critérios utilizados como referência neste trabalho ($SI > 63$). Além disso, o OVI-123 e o AZD5238 não estavam mais em desenvolvimento clínico no momento desta análise. A descontinuação do AZD5238, o qual estava em avaliação para o tratamento de tumores sólidos, foi devido a problemas de segurança durante os estudos clínicos (BOSS et al., 2010). Não foram encontrados relatos relacionados ao OVI-124. Por outro lado, o PHA-848125 e a dapivirina estão sendo ativamente estudados em humanos. O PHA-848125 está em fase II para o tratamento de tumores diversos, incluindo carcinoma hepatocelular metastático, câncer de cólon e pancreático. A dapivirina, por sua vez, está em fase III para o tratamento da síndrome da imunodeficiência adquirida (HIV).

Na análise dos metabólitos, não foi possível identificar na literatura composto selecionado pelo modelo que tenha sido reportado como ativo contra o *T. cruzi*. Por esta razão, adotou-se a estratégia de avaliar se os metabólitos selecionados eram derivados de fármacos conhecidamente ativos contra o parasita. Embora os metabólitos sejam novas entidades químicas e apresentem propriedades farmacológicas próprias, há diversos casos em que o metabólito possui um perfil de atividade biológica similar ao do fármaco original, o que pode ser explicado pela similaridade estrutural entre ambos. Esta análise preliminar foi focada somente nos 35 fármacos aprovados com $IC_{50} < 10 \mu M$ na inibição do *T. cruzi*, descritos anteriormente. Ainda, os bancos de dados descritos por Loi, Smith e Dalvie (2013) e por Douguet (2018) foram utilizados como as principais fontes para obter dados de farmacocinética dos compostos. Caso os metabólitos sejam ativos e estejam em concentrações biologicamente significativas em humanos, eles podem contribuir para a atividade anti-*T. cruzi* do fármaco de origem. A Tabela 5 apresenta um resumo dos metabólitos selecionados para os quais o fármaco original apresenta atividade anti-*T. cruzi* e, também, havia dados disponíveis sobre nível de exposição sistêmica em humanos com relação ao composto original.

Todos os compostos selecionados nas triagens virtuais foram comparados às moléculas ativas da Série treino utilizando o coeficiente de Tanimoto. Esta análise teve como objetivo identificar os compostos ativos utilizados no desenvolvimento do modelo que apresentassem a maior similaridade estrutural às moléculas selecionadas. Estes dados foram registrados nas Tabelas 2, 3, 4, 19, 20 e 21. Notavelmente, os compostos de Séries de triagem A, B e C que apresentaram os maiores coeficientes de similaridade com uma molécula da Série treino tiveram índices de Tanimoto iguais a 0,47, 0,43 e 0,48, respectivamente. Estes resultados evidenciam que as moléculas classificadas como ativas pelo modelo são estruturalmente distintas dos compostos ativos utilizados no seu desenvolvimento.

Tabela 4 – Candidatos a fármaco selecionados pelo modelo NB – FeatMorgan que são descritos como ativos contra o *T. cruzi* com $IC_{50} \leq 5 \mu M$

Candidato a fármaco (Pontuação no modelo) Atividade biológica da literatura	Status	Composto ativo da Série treino de maior similaridade ao fármaco
PHA-848125 (6,66) $IC_{50} = 1,0 \mu M^a$; $CC_{50} = 1,3 \mu M^a$	Fase II	TCMDC-143542 Tanimoto = 0,26
Dapivirina (1,41) $IC_{50} = 1,25 \mu M^b$; $CC_{50} > 64 \mu M^b$	Fase III	TCMDC-143109 Tanimoto = 0,26
OVI-123 (1,86) $IC_{50} = 1,41 \mu M^c$; $CC_{50} = 5,0 \mu M^c$	Inativo	TCMDC-143410 Tanimoto = 0,18
AZD5438 (3,66) $IC_{50} = 1,1 \mu M^a$; $CC_{50} = 2,7 \mu M^a$	Inativo	TCMDC-143382 Tanimoto = 0,26

^a (DUFFY et al., 2017). ^b (VENKATRAJ et al., 2014). ^c (PEÑA et al., 2015). IC_{50} e CC_{50} – Concentração de composto capaz de inibir em 50% a proliferação do parasita ou da célula hospedeira, respectivamente. SI - índice de seletividade, calculado pela relação entre os valores de CC_{50} e IC_{50} . TCMDC – Compostos da biblioteca Tres Cantos Medicines Development Campus.

Tabela 5 – Metabólitos selecionados pelo modelo NB – FeatMorgan que são derivados de fármacos ativos contra o *T. cruzi* com $IC_{50} \leq 10 \mu M$ (continua)

Metabólito (Pontuação no modelo) Exposição sistêmica do metabólito comparada ao fármaco de origem	Fármaco de origem Atividade biológica da literatura
Oxipurinol (1,89) Exposição metabólito/fármaco = 37,8-58,3	Alopurinol $IC_{50} = 2,8 \mu M^a$; $CC_{50} > 50 \mu M^a$
Norclomipramina (1,18) Exposição metabólito/fármaco = 0,37	Clomipramina $IC_{50} = 6,31 \mu M^b$; $CC_{50} = 15,9 \mu M^b$
3-Hidroxisdesloratadina (0,83) Exposição metabólito/fármaco = 0,5	Desloratadina $IC_{50} = 6,31 \mu M^b$; $CC_{50} = 15,9 \mu M^b$
4-Hidroxiduloxetina (0,38) Exposição metabólito/fármaco = 0,19	Duloxetina $IC_{50} = 6,31 \mu M^b$; $CC_{50} = 15,9 \mu M^b$
5,6-Diidroxiduloxetina (0,42) Exposição metabólito/fármaco = 0,22
N-desmetilimatinibe (7,96) Exposição metabólito/fármaco = 0,16	Imatinibe $IC_{50} = 5,01 \mu M^b$ e $CC_{50} = 15,9 \mu M^b$

Tabela 5 – Metabólitos selecionados pelo modelo NB – FeatMorgan que são derivados de fármacos ativos contra o *T. cruzi* com $IC_{50} \leq 10 \mu M$ (conclusão)

Metabólito (Pontuação no modelo) Exposição sistêmica do metabólito comparada ao fármaco de origem	Fármaco de origem Atividade biológica da literatura
Hidroxi-itraconazol (0,25) Exposição metabólito/fármaco = 1,4	Itraconazol $IC_{50} = 0,004 \mu M^c$ e $0,01 \mu M^b$; $CC_{50} = 1,1 \mu M^c$ e $15,8 \mu M^b$
Voriconazol- <i>N</i> -óxido (2,65) Exposição metabólito/fármaco = 1,6	Voriconazol $IC_{50} = 0,025 \mu M^b$; $CC_{50} = 50,1 \mu M^b$

^a (PLANER et al., 2014). ^b (DE RYCKER et al., 2016). ^c (KAISER et al., 2015b). A exposição do metabólito/fármaco foi calculada utilizando as respectivas áreas sob as curvas de concentração plasmática versus tempo (DOUGUET, 2018; LOI; SMITH; DALVIE, 2013; TURNHEIM; KRIVANEK; OBERBAUER, 1999). IC_{50} e CC_{50} – Concentração de composto capaz de inibir em 50% a proliferação do parasita ou da célula hospedeira, respectivamente.

A análise das vias de administração dos 269 fármacos aprovados e dos 291 candidatos a fármacos em desenvolvimento clínicos selecionados pelo modelo NB – FeatMorgan está sumarizada na Figura 17. Os compostos que poderiam ser administrados em humanos por mais de uma via de administração, incluindo a oral, foram classificados como pertencentes ao grupo de administração oral. Notavelmente, a maior parte dos fármacos da Série de triagem A e B (77,7% e 81,8%, respectivamente) pode ser administrada pela via oral em humanos. Além destes, 9,7% dos fármacos aprovados e 14,4% dos candidatos a fármaco são administrados por vias que têm potencial de resultar em exposição sistêmica, como a intravenosa e a subcutânea. No caso dos demais compostos, as principais vias de administração foram a inalatória, a ocular, a oftálmica e a tópica.

Figura 17 – Vias de administração dos fármacos e candidatos a fármacos selecionados na triagem virtual

Fonte: Produção do autor.

3.3.5 Avaliação experimental de fármacos selecionados na triagem virtual

Os 17 fármacos selecionados pelo modelo NB – FeatMorgan e obtidos para a avaliação experimental foram testados em diferentes concentrações para determinar a sua atividade antiparasitária (Figura 18). Neste experimento, o benznidazol, fármaco de referência, apresentou eficácia ($A_{\max} = 98,7\%$) e seletividade ($SI > 15$) em linha com os valores previamente reportados (MORAES et al., 2014), porém uma potência ca. 10 vezes inferior ($IC_{50} = 26,1 \mu M$ versus $2,4 \mu M$). Os fármacos que foram ativos contra *T. cruzi* estão apresentados na Figura 19. Notavelmente, todos estes compostos exibiram um baixo índice de seletividade frente à célula hospedeira. Embora nenhum composto avaliado tenha apresentado potência e seletividade dentro dos critérios sugeridos pelo DNDi ($IC_{50} < 5 \mu M$ e $SI > 10$), alguns exibiram uma eficácia adequada ($A_{\max} > 90\%$), como a flibanserina e a desloratadina ($A_{\max} = 92\%$ e 91% , respectivamente). Além disso, a prazosina e a doxazosina também apresentaram alta eficácia ($A_{\max} = 88\%$ e 84% , respectivamente), embora inferior ao critério de referência.

Figura 18 – Curvas de inibição *in vitro* do *T. cruzi* para os fármacos selecionados pelo modelo NB – FeatMorgan e avaliados experimentalmente

Os fármacos foram testados em células U2OS infectadas com o clone Y do *T. cruzi* em um ensaio de alto conteúdo (FRANCO, 2018; MORAES et al., 2014). O desfecho avaliado foi a taxa células hospedeiras infectadas pelo parasita. Fonte: Produção do autor.

Figura 19 – Fármacos selecionados pelo NB – FeatMorgan que apresentaram inibição do parasita *T. cruzi* na avaliação experimental

Os fármacos foram testados em células U2OS infectadas com o clone Y do *T. cruzi* em um ensaio de alto conteúdo (FRANCO, 2018; MORAES et al., 2014). O desfecho avaliado foi a taxa células hospedeiras infectadas pelo parasita. IC_{50} e CC_{50} – Concentração de fármaco capaz de inibir em 50% a proliferação do parasita ou da célula hospedeira, respectivamente. SI - índice de seletividade, calculado pela relação entre os valores de CC_{50} e IC_{50} . A_{max} – atividade máxima, determinada como a máxima atividade antiparasitária normalizada. Fonte: Produção do autor.

A potência (IC_{50}) dos fármacos desloratadina, prazosina, doxazosina, sertralina e risedronato já havia sido reportada na literatura, porém esta é a primeira vez que a eficácia (A_{max}) destes compostos é relatada. Os valores de IC_{50} observados neste trabalho foram similares aos de publicações de outros grupos para a desloratadina ($IC_{50} = 4,6 \mu M$ versus $6,31 \mu M$), a prazosina ($IC_{50} = 6,6 \mu M$ versus $5,01$

μM) e a doxazosina ($\text{IC}_{50} = 11,2 \mu\text{M}$ versus $5,01 \mu\text{M}$) (DE RYCKER et al., 2016). Por outro lado, dois fármacos foram 10-15 vezes mais potente neste experimento quando comparados aos valores da literatura, o risedronato ($\text{IC}_{50} = 10,2 \mu\text{M}$ versus $123 \mu\text{M}$) (MARTIN et al., 2001) e a sertralina ($\text{IC}_{50} = 0,107 \mu\text{M}$ versus $1,6 \mu\text{M}$) (PLANER et al., 2014). Vale ressaltar que esta é a primeira vez que a atividade anti-*T. cruzi* da flibanserina e da dexclorfeniramina são descritas.

3.4 DISCUSSÃO

Neste trabalho, modelos de aprendizado de máquina capazes de identificar compostos ativos contra o *T. cruzi* foram desenvolvidos e validados. O desenvolvimento compreendeu a análise de diferentes combinações dos descritores estruturais, *i.e.*: os *fingerprints* Extended, ECFP6, FCFP6, Morgan e FeatMorgan, e dos algoritmos de aprendizado de máquina *Naïve Bayes*, floresta aleatória, máquinas de vetores de suporte e redes neurais probabilísticas, assim como a otimização dos seus hiperparâmetros. O trabalho foi conduzido considerando as premissas estabelecidas pela OECD para métodos de QSAR (OECD, 2007) e as boas práticas de processamento dos dados propostas por Fourches, Muratov e Tropsha (2010, 2015, 2016) para a obtenção de modelos de QSAR preditivos. Posteriormente, o modelo de melhor performance foi utilizado para a triagem virtual de fármacos, candidatos a fármaco em desenvolvimento clínico e metabólitos de fármacos com o objetivo de identificar oportunidades de reposicionamento para o tratamento da doença de Chagas. Os principais diferenciais deste trabalho foram a utilização de bancos de dados anteriormente inexplorados para o desenvolvimento de modelos preditivos para *T. cruzi* e a aplicação dos modelos na triagem virtual sistemática de diferentes bibliotecas voltadas ao reposicionamento. Por esta razão, esta abordagem tem o potencial de identificar novos compostos a serem reposicionados.

A primeira etapa deste trabalho consistiu em mapear os estudos da literatura que reportaram a triagem experimental de compostos na inibição do *T. cruzi*. Constatou-se, inicialmente, que nenhum trabalho havia avaliado experimentalmente todos as moléculas com o potencial de serem reposicionadas para a terapêutica de Chagas, *i.e.*: todos os fármacos, candidatos a fármaco e metabólitos. Além disso, as triagens experimentais com bibliotecas contendo fármacos e candidatos a fármaco resultaram na identificação de moléculas ativas diferentes, mesmo nos casos em que

as mesmas moléculas estavam presentes na biblioteca inicial. Este resultado sugere que compostos previamente descritos como inativos contra o *T. cruzi* ainda deveriam ser considerados para novas iniciativas de reposicionamento. A diferença de resultados de triagens experimentais de inibição do *T. cruzi* pode ser devido a diversos fatores, como os protocolos experimentais do ensaio, a cepa do parasita, os critérios de seleção de compostos ativos o tipo de célula hospedeira, entre outros. Estes dados suportam a hipótese de que novos estudos visando ao reposicionamento de compostos para a terapêutica de Chagas são necessários e podem levar à identificação de novos candidatos promissores para o tratamento da doença.

Nas etapas iniciais de validação (validação cruzada quádrupla e externa com a Série teste A), os algoritmos RF, NB e SVM resultaram em modelos de alta performance. Os resultados observados foram similares aos reportados para modelos preditivos descritos na literatura para, por exemplo, *Mycobacterium tuberculosis* (EKINS et al., 2010; 2013; EKINS, FREUNDLICH, 2011; PERIWAL, 2011), *Schistosoma mansoni* (NEVES et al., 2016) e, inclusive, *T. cruzi* (ACEVEDO; SCOTTI, L.; SCOTTI, M.T., 2018; EKINS et al., 2015; MELO-FILHO et al., 2019). Conforme destacado por Liu et al. (2019), a comparação de diferentes modelos visando a selecionar aquele de maior capacidade preditiva é uma tarefa complexa, pois raramente um único modelo apresenta performance superior aos demais em todas as métricas avaliadas. Este mesmo padrão também foi observado no presente trabalho. O modelo NB – FeatMorgan foi selecionado como aquele de maior capacidade preditiva por ter apresentado valores de sensibilidade, especificidade, TCB, PPP e PPN superiores a 70% e os maiores valores de AUC ROC e AUC BEDROC nas validações cruzada e externa (Série teste A). A métrica AUC ROC representa a eficiência do modelo em separar compostos ativos de inativos e pode ser interpretada como a probabilidade de que uma molécula ativa seja classificada com uma pontuação maior do que uma molécula inativa (EMPEREUR-MOT et al., 2015; LÄTTI; NIINIVEHMAS; PENTIKÄINEN, 2016). Por sua vez, a AUC BEDROC permite avaliar a capacidade de reconhecer precocemente os compostos ativos em uma triagem virtual (TRUCHON; BAYLY, 2007). Como as diferentes implementações dos *fingerprints* circulares (MORGAN, 1965; ROGERS; HAHN, 2010) influenciam a performance dos modelos computacionais, novos projetos na área devem cogitar explorações similares às realizadas neste trabalho, em particular incluindo o *fingerprint* FeatMorgan. Notavelmente, embora *fingerprints* circulares já tenham sido

empregados no desenvolvimento de modelos preditivos para *T. cruzi*, esta é a primeira vez que o FeatMorgan foi utilizado neste contexto.

O modelo NB – FeatMorgan apresentou consistência na capacidade preditiva durante a validação com a Série teste B, confirmando o potencial da estratégia adotada para a sua seleção. Notavelmente, a Série teste B foi estruturada de maneira desbalanceada (400 inativos e 21 ativos) com objetivo de simular uma triagem virtual real. Esta estratégia foi baseada nos trabalhos de Alberca et al. (2016, 2018) e Gartner et al. (2017) e teve como objetivo determinar de maneira mais confiável as métricas de enriquecimento, como o fator de enriquecimento e a AUC ROC. Assim como demonstrado por Truchon e Bayly (2007), a AUC ROC apresenta menor variabilidade conforme maior o número de compostos ativos e menor a relação de ativos para inativos. Vale ressaltar que algumas métricas, como a acurácia, por exemplo, são inadequadas para avaliar bancos de dados desbalanceados, pois tendem a favorecer a classe com o maior número de compostos (BOUGHORBEL; JARRAY; EL-AMBARI, 2017). Nestes casos, métricas que equilibram os erros em ambas as classes, como a taxa de classificação balanceada, devem ser utilizados na avaliação dos modelos (PERIWAL et al., 2011).

Como a capacidade preditiva de modelos de aprendizado de máquina depende diretamente dos dados utilizados no seu desenvolvimento, o NB – FeatMorgan tem o potencial de identificar compostos que não teriam sido identificados em triagens virtuais de trabalhos anteriores (ACEVEDO; SCOTTI, L.; SCOTTI, M.T., 2018; EKINS et al., 2015; MELO-FILHO et al., 2019). Este diferencial está associado ao fato de a Série treino deste trabalho conter moléculas ativas que nunca haviam sido exploradas para o desenvolvimento de modelos preditivos para *T. cruzi*. Isto ocorre pois os algoritmos analisam e extraem as correlações existentes entre os descritores moleculares e a atividade biológica das moléculas da série treino e utilizam estas informações para prever a atividade de compostos não testados (JORDAN, 2018; VAMATHEVAN et al., 2019; YANG et al., 2019). Outro diferencial da série treino deste trabalho está associado ao fato de que as moléculas ativas foram testadas em dois ensaios ortogonais de inibição do parasita. A maior confiabilidade dos dados e a menor probabilidade de existirem falsos positivos influencia diretamente a qualidade dos modelos obtidos (FOURCHES; MURATOV; TROPSHA, 2015). Notavelmente, há relatos na literatura de baixa reprodutibilidade de dados experimentais para compostos identificados como ativos contra o *T. cruzi*. Andriani et al. (2011), por

exemplo, reportou que durante o teste de 23 moléculas classificadas como ativas ($IC_{50} < 1,2 \mu M$ e índice de seletividade > 100) pelo Instituto Broad (CARMODY et al., 2010), muitos dos compostos foram menos potentes ou completamente inativos, apesar do uso de condições experimentais similares. No caso de ensaios que avaliam a inibição do *T. cruzi*, a complexidade dos protocolos e o grande número de parâmetros experimentais constituem os principais causadores da grande variabilidade nos resultados entre diferentes laboratórios (CHATELAIN; IOSET, 2018). Fatores como a carga parasitária de inoculação *in vitro*, o tempo de incubação com os compostos teste e a sequência de manuseio, por exemplo, podem ter grande impacto nos resultados observados, mesmo para compostos de referência (YANG et al., 2017).

A comparação da performance de novos modelos computacionais com métodos de referência, que sejam computacionalmente menos custosos e cuja implementação seja mais simples, é uma boa prática discutida e empregada na literatura (LAUFKÖTTER; MIYAO; BAJORATH, 2019; MERZ et al., 2015; STUMPFE; BAJORATH, 2020). Neste sentido, o modelo NB – FeatMorgan foi comparado ao método de triagem virtual baseado em similaridade estrutural, o qual prioriza compostos conforme maior a proporção de fragmentos compartilhados com uma das moléculas ativas de referência (STOKES et al., 2020). A performance intermediária do método de similaridade, superior à seleção aleatória, justifica o seu amplo uso em campanhas de triagem virtual. Por outro lado, o modelo NB – FeatMorgan apresentou maior capacidade preditiva, embora tenha sido desenvolvido a partir dos mesmos compostos que foram utilizados para o método de similaridade. Este resultado confirma o ganho de performance na seleção compostos ativos através da implementação e validação de algoritmos de aprendizado de máquina. Esta melhor performance pode estar associada à capacidade do algoritmo de identificar características estruturais (fragmentos) dentro da estrutura química que contribuem positivamente ou negativamente para a atividade. Por outro lado, a análise de Tanimoto apenas atribui uma pontuação com base no número de fragmentos comuns (similaridade) entre duas moléculas.

A triagem virtual com o modelo NB – FeatMorgan resultou na seleção de 269 fármacos, 291 candidatos a fármaco e 206 metabólitos. Considerando a alta capacidade preditiva do modelo, estas listas possuem um grande potencial de estarem enriquecidas com moléculas ativas, que poderão servir como base para o desenvolvimento de novos tratamentos para a doença de Chagas. A análise das

informações disponíveis na literatura para os compostos selecionados permitiu evidenciar que alguns já haviam sido testados em ensaios de inibição do parasita e apresentado uma atividade biológica promissora ($IC_{50} < 10 \mu M$). Vale ressaltar que o número de fármacos aprovados que já foram testados contra o *T. cruzi* é substancialmente maior do que o de candidatos a fármaco ainda em desenvolvimento e de metabólitos. Este resultado pode ser consequência da menor disponibilidade e maior dificuldade de acesso a amostras destes compostos (CORSELLO et al., 2017; PUSHPAKOM et al., 2019). Para estes compostos, o uso de modelos preditivos para realizar seleção preliminar de moléculas com maior potencial de serem ativas pode otimizar ou mesmo viabilizar iniciativas de reposicionamento. Embora os fármacos aprovados tenham sido mais explorados em iniciativas de reposicionamento para Chagas, diversos fatores suportam a necessidade de novos estudos para estes compostos, como a inexistência de um trabalho no qual todas as moléculas disponíveis no mercado tenham sido testadas, a ausência da identidade de compostos inativos nas publicações (impossibilitando determinar quais moléculas já foram testadas ao menos uma vez), a grande variabilidade dos ensaios biológicos para *T. cruzi* e, conseqüentemente, a possível existência de falsos negativos. Em um contexto geral, resultados falso negativos reconhecidamente resultam em perdas econômicas significativas para os projetos, as quais são difíceis de serem calculadas (BIBETTE, 2012). No caso de projetos de reposicionamento para DTNs, essas perdas têm um impacto ainda mais significativo, visto o maior valor de um fármaco, quando comparado a moléculas pouco estudadas, e o reduzido investimentos em P&D para estas indicações.

Dentre os 269 fármacos aprovados selecionados, 35 (13%) já haviam sido descritos como ativos contra o *T. cruzi*, 25 com $IC_{50} < 5 \mu M$ e 10 com IC_{50} entre $5 \mu M$ e $10 \mu M$. Destes compostos, 19 cumpririam os critérios de atividade propostos pelo DNDi ($IC_{50} < 5 \mu M$ e SI > 10), adotados como referência neste trabalho (CHATELAIN; IOSET, 2018). Embora não seja possível estabelecer um valor definitivo da taxa de ativos (*hit rate*) da triagem virtual neste momento, pode-se concluir que a porcentagem de ativos será, no mínimo, igual ou superior a 7,1% (19 dos 269 fármacos). Esta porcentagem já se mostra significativamente maior que os níveis reportados na literatura para triagens experimentais para o *T. cruzi*, que variam entre 0,17% para bibliotecas focadas em diversidade estrutural e 0,23% para produtos naturais (CHATELAIN; IOSET, 2018). Nas triagens do Instituto BROAD (CARMODY et al.,

2010) e da GSK (PEÑA et al., 2015), por exemplo, os *hit rates* foram 0,85% e 0,13%, respectivamente (CHATELAIN; IOSET, 2018). Além disso, vale ressaltar que critérios menos restritivos para classificação de moléculas como ativas são comumente descritos na literatura, seja referente à potência ($IC_{50} < 10 \mu M$) ou à seletividade ($SI > 5$) dos compostos (KATSUNO et al., 2015; KHARE et al., 2016). Se aplicados na análise dos dados deste trabalho, estes critérios poderiam resultar em *hit rates* inclusive superiores.

A lista de fármacos selecionados pelo modelo inclui, entre outros compostos, o benznidazol e o nifurtimox, os quais são utilizados na clínica para o tratamento da doença de Chagas. Embora a seleção destes compostos e de outros conhecidamente ativos não gere novidade para a literatura, este resultado serve para confirmar a capacidade preditiva do modelo, visto que essas moléculas não estavam presentes na Série treino. Além disso, 9 dos 18 fármacos ativos que foram selecionados na triagem virtual são conhecidos inibidores da enzima citocromo P450 14-alfa-esterol desmetilase do parasita (TcCYP51), como o voriconazol, o cetoconazol e o bifonazol (DE RYCKER et al., 2016). A seleção de inibidores da TcCYP51 e moléculas contendo o grupamento nitro pelo modelo NB – FeatMorgan é razoável, visto que compostos ativos pertencentes a estas classes não foram removidos da Série treino. Deste modo, os modelos correlacionaram as características estruturais destas séries químicas com a atividade contra o *T. cruzi* (JORDAN, 2018; VAMATHEVAN et al., 2019; YANG et al., 2019).

A inibição da TcCYP51 não tem se mostrado um mecanismo de ação interessante para o desenvolvimento de novas terapias para a doença de Chagas, visto que os candidatos avaliados em estudos clínicos não foram eficazes como monoterapia para a doença (CHATELAIN, 2017). Por esta razão, a identificação e a remoção dos compostos que inibam a TcCYP51 é uma prática comum nos protocolos de triagem experimental contra o *T. cruzi*, habitualmente realizada após a confirmação das moléculas ativas (CHATELAIN; IOSET, 2018; RILEY et al., 2015). Por outro lado, a remoção de todos os compostos que inibam a TcCYP51 poderia ocasionar a perda de novas moléculas que atuem em um alvo terapêutico promissor para o combate ao *T. cruzi* e cuja ação na TcCYP51 seja uma atividade secundária do composto. Deste modo, compostos que apresentaram inibição da TcCYP51 não foram removidos da Série treino.

No caso dos candidatos a fármaco em desenvolvimento clínico, mesmo que o composto tenha sido descontinuado para a indicação primária, ainda seria possível considerar o seu reposicionamento para o tratamento de doença de Chagas. Nestes casos, seria necessário avaliar o perfil de segurança e farmacocinética da molécula, assim como os resultados que motivaram a sua descontinuação. Por exemplo, o AZD5238, candidato a fármaco selecionado como potencialmente ativo pelo NB – FeatMorgan e descrito na literatura como capaz de inibir o *T. cruzi* ($IC_{50} = 1,1 \mu M$), foi descontinuado por toxicidade excessiva nos estudos clínicos (BOSS et al., 2010). Deste modo, é muito provável que o composto não apresente um perfil de segurança adequado para o uso na terapêutica da doença de Chagas.

Na análise dos resultados da triagem virtual, observou-se, ainda, baixa similaridade estrutural entre os compostos selecionados e as moléculas ativas de série treino, inclusive no caso dos 35 fármacos com $IC_{50} < 10 \mu M$ na inibição do parasita. Este resultado confirma a capacidade do modelo de extrapolar as informações aprendidas a partir da Série treino para a identificação de compostos ativos pertencentes a novas séries químicas, uma propriedade de interesse para modelos preditivos (STOKES et al., 2020; WALTERS; MURCKO, 2020).

Os fármacos avaliados experimentalmente neste trabalho apresentaram valores de potência e seletividade insuficientes e, por isso, não seriam bons candidatos para o tratamento da doença de Chagas como monoterapias. Por outro lado, a flibanserina, a desloratadina, a prazosina e a doxazosina apresentaram alta eficácia, a qual estava acima ou próxima dos critérios de atividade propostos pelo DNDi ($A_{max} > 90\%$). Embora a importância de compostos com alta eficácia na eliminação do parasita seja amplamente discutida na literatura (CHATELAIN, 2015; CHATELAIN; IOSET, 2018; DUFFY et al., 2017), nem todas as publicações incluem este dado ao descrever novas moléculas com atividade anti-*T. cruzi*. Neste contexto, este trabalho contribui com a literatura ao reportar a atividade antiparasitária de novos fármacos, assim como apresentando os dados de eficácia para moléculas previamente testadas. Notavelmente, o posaconazol e o ravuconazol, compostos que falharam em estudos clínicos para o tratamento da doença de Chagas, apresentaram grande variação de eficácia (de 27 a 93% para o posaconazol e de 22 a 93% para o ravuconazol), quando testados em diferentes cepas e linhagens do *T. cruzi* (MORAES et al., 2014). Por outro lado, os valores de eficácia do benznidazol e do nifurtimox foram consistentemente elevados (93 - 100% e 95 - 100%, respectivamente). Neste

sentido, os fármacos identificados como promissores neste trabalho poderão ser testados em painéis com diferentes linhagens do *T. cruzi* com vistas a confirmar a sua atividade antiparasitária. Além disso, como a potência e a seletividade de compostos em ensaios anti-*T. cruzi* podem variar de acordo com a célula hospedeira, seria adequado testá-los também em modelos que utilizem outras linhagens celulares para avaliar a influência desse parâmetro no índice de seletividade. As células U2OS, utilizadas no presente trabalho, são descritas como mais sensíveis a efeitos citotóxicos quando comparadas a células Vero (epitélio renal de macacos) ou THP-1 (células monocíticas de humanos) (FRANCO et al., 2019). Embora os fármacos testados não sejam bons candidatos ao reposicionamento como monoterapias, eles podem ser testados em combinações visando a obter tratamentos seguros e eficazes, conforme discutido a seguir.

A combinação de fármacos pode ser utilizada como uma estratégia para aumentar as probabilidades de sucesso de iniciativas de reposicionamento (PLANER et al., 2014; TALEVI; BELLERA, 2020; SUN; SANDERSON; ZHENG, 2016). A interação farmacodinâmica entre dois compostos pode causar efeitos aditivos, sinérgicos ou antagônicos, respectivamente quando a atividade biológica combinada é igual, superior ou inferior à soma das atividades individuais. Interações farmacocinéticas, por sua vez, influenciam a exposição e o metabolismo dos compostos. Neste sentido, as interações entre diferentes fármacos podem ser exploradas no desenvolvimento de uma combinação com eficácia superior às monoterapias ou na redução das concentrações necessárias de cada composto para atingir a eficácia clínica (JIA et al., 2009). A estratégia de combinação poderia, por exemplo, permitir o reposicionamento de compostos ativos que não apresentassem, como monoterapia exposição sistêmica adequada para combater o parasita nas doses aprovadas ou em estudo clínico. Além disto, a redução de doses poderia permitir o uso de compostos ativos que não apresentem perfil de segurança adequado nas concentrações eficazes, quando administrados individualmente. Particularmente, a redução das doses do benznidazol e do nifurtimox, cujos usos na clínica são limitados devido a problemas de toxicidade, representaria um avanço na terapêutica da doença de Chagas. Diversas publicações já reportaram o efeito sinérgico de combinações do benznidazol com o alopurinol (MAZZETI et al., 2019), a atorvastatina (ARAUJO-LIMA et al., 2018) ou a clomipramina (GARCÍA et al., 2016; STRAUSS et al., 2018) e do posaconazol com o anlodipino ou a clemastina (PLANER et al., 2014) na inibição do

T. cruzi. Ressalta-se, ainda, que combinações de fármacos que atuem através de diferentes mecanismos de ação podem reduzir ou atrasar o desenvolvimento de resistência às terapias pelo parasita (JIA et al., 2009). Entretanto, o desenvolvimento de novas combinações deve levar em consideração os possíveis efeitos adversos indesejados oriundos da combinação, especialmente aqueles associados a maior toxicidade (MAGRO; MORETTI; LEONE, 2012).

É importante salientar que esta é a primeira vez que modelos preditivos para *T. cruzi* foram utilizados no reposicionamento de fármacos através da identificação de metabólitos de fármacos e de candidatos a fármaco que sejam potencialmente ativos contra o parasita. Estratégias de reposicionamento baseadas neste princípio já foram empregadas com sucesso na busca de novos tratamentos para a tuberculose (KIGONDU et al., 2014). Como os metabólitos são entidades químicas diferentes do fármaco original, eles possuem perfis farmacológicos e de farmacocinética próprios, os quais podem ser similares ou significativamente diferentes daqueles do composto de origem. Diversos metabólitos descritos na literatura são estáveis e podem ser encontrados na circulação sanguínea e em tecidos periféricos em concentrações biologicamente relevantes. Em alguns casos, a exposição sistêmica do metabólito é equivalente ou mesmo superior à do fármaco original. A descoberta da atividade farmacológica de metabólitos já estimulou o desenvolvimento destas moléculas como novos princípios ativos, posteriormente à aprovação do fármaco original, como é o caso da nortriptilina e da desipramina, respectivamente provenientes da amitriptilina e da imipramina (OBACH, 2013; OPREA; OVERINGTON, 2015). Neste cenário, seria possível que um metabólito fosse ativo contra o parasita, mesmo sendo derivado de um fármaco inativo, caracterizando, então, um bioprecursor. Considerando o limitado número de fármacos aprovados, quando comparado ao espaço químico total e o grande número de metabólitos reportados na literatura, esta abordagem tem o potencial de expandir substancialmente o número de compostos com o potencial de ser reposicionados. Embora a obtenção de amostras de metabólitos continue sendo desafiadora, os inúmeros benefícios inerentes ao reposicionamento justificam a exploração desta estratégia. Neste sentido, a lista de metabólitos gerada pelo NB – FeatMorgan, potencialmente enriquecida com molécula ativas contra o parasita, contribui com a priorização de futuros esforços nesta área.

Considerando a fisiopatologia da doença de Chagas e a distribuição do parasita pelo organismo, novas substâncias para o seu tratamento necessitam

apresentar exposição sistêmica adequada no plasma e nos reservatórios do parasita em tecidos periféricos para exercer sua atividade biológica. Em um cenário ideal, as concentrações plasmáticas atingidas pelos fármacos na posologia aprovada para uso em humanos seriam suficientes para inibir o *T. cruzi* de maneira eficaz. Neste cenário, seria possível iniciar estudos clínicos de eficácia fase II/ III logo após a validação do seu efeito em modelos pré-clínicos *in vitro* e *in vivo* da doença. Por esta razão, buscou-se avaliar de modo preliminar se os fármacos selecionados são utilizados na clínica através de vias comumente associadas com exposição sistêmica da substância. Notavelmente, a maioria dos fármacos selecionados das Séries de triagem A e B pode ser administrada pela via oral ou por vias parenterais, que resultam na exposição sistêmica dos compostos (ex.: intravenosa e subcutânea). A via oral é considerada preferencial para o desenvolvimento de novas moléculas para tratar a doença de Chagas, visto que a mesma resulta em maior adesão ao tratamento e pode ser realizada fora de ambientes hospitalares, proporcionando, assim, maior acesso em locais que carecem de infraestrutura (SILVA et al., 2019). A administração parenteral, por sua vez, seria aceitável no caso de tratamentos de curta duração (DNDi, 2020). Caso a posologia aprovada ou em estudo em humanos não resulte em uma exposição sistêmica do composto suficiente para a nova indicação, seria necessário realizar novos estudos de segurança (fase I), utilizando doses maiores, antes de iniciar etapa de eficácia. Se a farmacocinética da molécula for inadequada para o uso pretendido, ainda seria possível explorar novas formulações e vias de administração visando a otimizar a posologia do medicamento (OPREA; OVERINGTON, 2015; TALEVI; BELLERA, 2020). Os benefícios inerentes ao reposicionamento, especialmente as reduções significativas no tempo e nos investimentos para que um composto seja aprovado, continuariam presentes mesmo no caso em que estas etapas adicionais sejam necessárias.

O presente trabalho visa a contribuir e fomentar novas pesquisas voltadas ao reposicionamento de fármacos para o tratamento da doença de Chagas. As listas de compostos selecionados pelo modelo NB – FeatMorgan são enriquecidas com compostos potencialmente ativos e podem servir como base para novas triagens experimentais. Os compostos podem ser avaliados em diferentes ensaios para comprovar a sua atividade antiparasitária, seja como monoterapia ou em combinação com outras moléculas. Os fármacos, candidatos a fármaco e metabólitos ativos contra o *T. cruzi* deverão ser avaliados com relação aos seus perfis de segurança e

farmacocinética em humanos visando a comprovar o seu potencial para a nova indicação. Como os níveis de toxicidade aceitáveis para um fármaco variam de acordo com a doença, compostos que apresentam um risco-benefício positivo para o tratamento de indicação terminal (ex.: câncer) podem não ser considerados seguros para a doença de Chagas (OPREA; OVERINGTON, 2015). Caso as moléculas ativas não sejam adequadas para o reposicionamento, elas podem ser usadas pontos de partida promissores para novos projetos otimização estrutural, visto que pertencem a espaços químicos já explorados na terapêutica (DE RYCKER et al., 2016), em franca ligação com a Química Medicinal. O modelo NB - FeatMorgan poderá ser aplicado em novas triagens virtuais com outras bibliotecas visando a expandir o número de compostos potencialmente ativos contra o *T. cruzi*. Neste cenário, o reposicionamento de fármacos continua sendo uma estratégia promissora para o desenvolvimento de terapias para doenças que carecem que tratamentos seguros e eficazes e de investimentos no desenvolvimento de novos produtos.

3.5 CONCLUSÕES

A doença de Chagas é uma enfermidade que afeta milhões de pessoas em todo mundo e carece de tratamentos seguros e eficazes. Na estratégia LBDD deste trabalho, modelos de aprendizado de máquina foram desenvolvidos com o objetivo de auxiliar na identificação de fármacos, candidatos a fármaco e metabólitos com o potencial de serem reposicionados para a terapêutica da doença. Diversos modelos desenvolvidos com os algoritmos *Naive Bayes*, floresta aleatória e máquinas de vetores de suporte apresentaram alta performance nas validações iniciais (cruzada quádrupla e série teste A), com valores de AUC ROC próximos ou superiores a 0,8 e de sensibilidade, especificidade e taxa de classificação balanceada próximos ou superiores a 80%. A capacidade preditiva do melhor modelo, o NB – FeatMorgan, foi confirmada na segunda validação externa, utilizando a Série teste B (AUC ROC = 0,85, EF1 = 14,3 e TCB = 80%). Merece ressaltar que a performance deste modelo foi superior à do método de busca por similaridade estrutural com o coeficiente de Tanimoto (AUC ROC = 0,69 e EF = 9,5), utilizado como comparador.

A triagem virtual empregando o NB – FeatMorgan resultou na seleção de 269 fármacos aprovados, 291 candidatos a fármaco em desenvolvimento clínico e 206 metabólitos de fármacos, os quais apresentam um alto potencial de serem ativos

contra o *T. cruzi*. Ressalta-se que 35 dos 269 fármacos selecionados já haviam sido descritos como ativos contra o parasita ($IC_{50} < 10 \mu M$). Destes, 19 apresentavam, ainda, atividade biológica dentro dos critérios propostos pelo DNDi para novas séries químicas ativas contra o *T. cruzi* ($IC_{50} < 5 \mu M$ e $SI > 10$). Alguns destes fármacos, como a ziprasidona, a primaquina, o triantereno, o alopurinol e o entecavir não pertencem às classes químicas já testadas em estudos clínicos para Chagas (compostos nitro-heterocíclicos e antifúngicos azólicos/triazólicos). Por esta razão, eles têm o potencial de serem ativos através de novos mecanismos de ação. Embora a seleção de fármacos conhecidamente ativos não adicione novidade à literatura, isto permite validar a capacidade preditiva do modelo, uma vez que as moléculas selecionadas não estavam presentes na série treino nem nas séries externas de validação. No cenário em que somente os 19 dos 269 fármacos fossem ativos, o modelo teria gerado um *hit rate* de 7,1%, o qual já seria superior aos valores habitualmente reportados na literatura.

Os 17 fármacos avaliados experimentalmente contra o *T. cruzi* apresentaram potência e seletividade insuficientes para serem utilizados como monoterapias no tratamento da doença de Chagas. Entretanto, estes compostos ainda poderiam ser explorados em combinação com outras moléculas ou mesmo servir como pontos de partida promissores para novos projetos de pesquisa na área. Especialmente, a flibanserina e a desloratadina apresentaram alta eficácia ($A_{max} > 90\%$), característica considerada essencial para novos compostos a serem utilizados no combate ao parasita.

O reduzido número de candidatos a fármacos e metabólitos já avaliados contra o *T. cruzi*, resultado da maior dificuldade de acesso a amostras destes compostos, evidencia ainda mais as vantagens de empregar modelos preditivos para priorizar os compostos destas classes para serem testados experimentalmente. Ainda, a análise de metabólitos com atividade anti-*T. cruzi* e presentes em concentrações biologicamente relevantes expande o número de candidatos em potencial ao reposicionamento para Chagas.

O presente trabalho contribui com o desenvolvimento de novas terapias para a doença de Chagas através do modelo LBDD preditivo para *T. cruzi* desenvolvido e validado, das listas de compostos potencialmente ativos contra o parasita e dos novos dados experimentais para os compostos testados. Os fármacos, candidatos a fármaco e metabólitos selecionados neste trabalho podem servir como base para novos

estudos experimentais que visem ao reposicionamento de moléculas para o tratamento da doença de Chagas. Por sua vez, o modelo NB - FeatMorgan poderá ser aplicado em novas triagens virtuais para selecionar moléculas potencialmente ativas contra o parasita a partir de novas bibliotecas de interesse.

CAPÍTULO 4

4 ABORDAGEM SBDD: DESENVOLVIMENTO DE MODELOS BASEADOS NA INIBIÇÃO DA ENZIMA DIIDROOROTATO DESIDROGENASE DO *TRYPANOSOMA CRUZI*

4.1 INTRODUÇÃO

4.1.1 A enzima diidroorotato desidrogenase do *T. cruzi* (TcDHODH) como um alvo terapêutico para o tratamento da doença de Chagas

Os nucleotídeos de pirimidina são compostos essenciais para diversos processos fisiológicos, como a síntese de material nucleico (RNA e DNA), ATP, cAMP, cGMP, glicoproteínas e lipídeos da membrana celular (INAOKA et al., 2017; PINHEIRO; EMERY; NONATO, 2013). Estes nucleotídeos podem ser obtidos por vias de salvação, que utilizam produtos do metabolismo de outros nucleotídeos e bases pirimidínicas ou nucleosídeos, ou pela via de síntese *de novo*, que emprega substratos mais simples, como aminoácidos e glicose (AVELAR, 2014; PINHEIRO; EMERY; NONATO, 2013). A diidroorotato desidrogenase (DHODH), que catalisa a oxidação do diidroorotato em orotato/ácido orótico (Figura 20), é uma enzima do ciclo de síntese *de novo* de pirimidinas (REIS et al., 2017). Devido a sua relevância para processos essenciais do organismo e presença em diferentes espécies, a inibição da DHODH tem sido explorada como um mecanismo de ação promissor em diversas áreas terapêuticas, como doenças autoimunes, câncer e doenças infecciosas (REIS et al., 2017).

Figura 20 – Representação da conversão enzimática do diidroorotato em orotato através da ação da diidroorotato desidrogenase

Fonte: Adaptado de MUNIER-LEHMAN et al. (2013).

A DHODH é considerada uma flavoenzima, que possui em seu sítio catalítico um mononucleotídeo de flavina (FMN), e pode ser categorizada nas classe 1 ou 2, dependendo da sua estrutura, localização celular e especificidade de substratos (REIS et al., 2017). As DHODHs de classe 1 estão presentes no citoplasma celular como um homodímero (classe 1A) ou tetrâmero (classe 1B). Enzimas da classe 2 estão presentes como monômeros na região interna da membrana mitocondrial, no caso de eucariotos, ou da membrana celular, nos procariotos (BOSCHI et al., 2019). A oxidação do diidroorotato é uma reação acoplada à redução de diferentes substratos, dependendo da classe da enzima: fumarato para enzimas da classe 1A, nicotinamida adenina dinucleotídeo oxidada (NAD^+) para a classe 1B e ubiquinona/quinonas do ciclo respiratório para a classe 2 (BOSCHI et al., 2019; INAOKA et al., 2016).

A DHODH de humanos (hDHODH) pertence à classe 2 e tem sido explorada como um alvo terapêutico para doenças autoimunes e oncologia. A teriflunomida e seu bioprecursor, a leflunomida, são inibidores da hDHODH aprovados para o tratamento de artrite reumatoide e esclerose múltipla, respectivamente (LOLLI et al., 2018). O brequinar, por sua vez, é um inibidor da hDHODH em estudo clínico fase II para leucemia mieloide aguda. Na área de anti-infecciosos, inibidores da DHODH de parasitas do gênero *Trypanosoma* e *Leishmania*, os quais possuem a DHODH de classe 1, e de *Helicobacter pylori*, *Toxoplasma gondii*, *Plasmodium falciparum* e *Schistosoma mansoni*, que apresentam DHODH de classe 2, já foram descritos na literatura (BOSCHI et al., 2019). O composto DSM265, inibidor seletivo da DHODH de *Plasmodium falciparum*, está em desenvolvimento clínico e apresentou resultados promissores em estudos de fase 1a e 1b (BOSCHI et al., 2019; GLOBALDATA, 2020). Estes relatos evidenciam o potencial de desenvolver novos anti-infecciosos através da inibição da DHODH do patógeno. Particularmente, a DHODH do *T. cruzi* (TcDHODH) pertence à classe 1A, a qual carece de inibidores potentes e seletivos para a sua modulação (INAOKA et al., 2016).

O potencial da inibição das vias de síntese *de novo* de pirimidinas e da enzima TcDHODH como estratégias terapêuticas para o combate ao *T. cruzi* têm sido evidenciados em diversos estudos genéticos com o parasita (INAOKA et al., 2016; PINHEIRO; EMERY; NONATO, 2013). O *knock-out* da TcDHODH resultou na redução da sobrevivência e proliferação do *T. cruzi*, mesmo na presença de nucleosídeos de pirimidina, substratos para a via de salvação (ANNOURA et al., 2005). O *knock-out* da carbamoil fosfato-sintase sintase II, outra enzima da via de síntese *de novo* de

pirimidina, também resultou em uma redução significativa da proliferação do *T. cruzi*, em particular da forma amastigota (HASHIMOTO et al., 2012). Estas evidências sugerem que a via de síntese *de novo* de pirimidinas é essencial para a sobrevivência do parasita, em particular na forma intracelular (INAOKA et al., 2016; PINHEIRO; EMERY; NONATO, 2013). A reduzida participação das vias de salvação para a obtenção de pirimidina no *T. cruzi* pode estar associada à ausência da atividade da uracil fosforibosiltransferase e da uridina quinase na forma amastigota do parasita, enzimas da via de salvação (EL-SAYED et al., 2005; HAMMOND; GUTTERIDGE, 1892; INAOKA et al., 2016). Além disso, as células hospedeiras apresentam em seu citoplasma os derivados pirimidínicos principalmente na forma de nucleotídeos, enquanto o *T. cruzi* utiliza preferencialmente bases pirimidínicas e nucleosídeos nas vias de salvação (HASHIMOTO et al., 2012; PINHEIRO; EMERY; NONATO, 2013). No caso da TcDHODH, a sua função na regulação do equilíbrio redox da célula, através do metabolismo do fumarato em succinato, também pode contribuir para a atividade anti-*T. cruzi* dos seus inibidores (ANNOURA et al., 2005; TAKASHIMA et al., 2002).

Neste contexto, a inibição da TcDHODH tem sido evidenciada como uma nova estratégia promissora no combate ao *T. cruzi*. Como os inibidores da TcDHODH descritos na literatura (CHELESKI et al., 2010; INAOKA et al., 2016; 2017) ainda não foram suficientemente otimizados e não há nenhum candidato a fármaco com este mecanismo de ação disponível até o momento, métodos computacionais foram explorados no presente trabalho, visando à identificação de novos compostos capazes de inibir a TcDHODH. Considerando os benefícios da estratégia de reposicionamento, buscou-se inicialmente avaliar a existência de fármacos e candidatos a fármacos com o potencial de atuarem como inibidores da enzima.

4.1.2 Os métodos de modelagem molecular baseados na estrutura do alvo terapêutico

Os métodos computacionais de planejamento de fármacos baseados na estrutura de biomacromoléculas alvo (SBDD, do inglês *structure-based drug design*) têm sido amplamente empregados em projetos de descoberta de novas moléculas que objetivam a modulação de alvos terapêuticos específicos (*target-based drug discovery*) (FERREIRA et al., 2015; GILBERT, 2013). Nestes casos, a ação dos

compostos ocorre através da sua complexação (que pode envolver interações ou até formação de ligações) com regiões importantes de uma biomacromolécula, de forma a modular a sua atividade biológica (ex.: ligação no sítio catalítico da enzima). Ferramentas SBDD foram utilizadas, por exemplo, no desenvolvimento dos fármacos ritonavir e saquinavir, inibidores de protease do vírus da imunodeficiência humana – HIV (VAN MONTFORT; WORKMAN, 2017).

Em grande parte, o processo de complexação depende da existência de complementaridade eletrostática e estereoquímica entre o ligante e o alvo (FERREIRA et al., 2015). As ferramentas de SBDD podem ser empregadas para auxiliar no desenvolvimento de novos ligantes a partir de informações tridimensionais da estrutura do alvo terapêutico. Dentre as diversas aplicações possíveis, podem-se citar a avaliação de características estruturais da biomacromolécula; a identificação de regiões mais propensas à formação de interações; a análise de interações existentes em complexos experimentais e, principalmente, o planejamento de modificações estruturais nos ligantes de modo a otimizar a sua complementaridade eletrostática e estereoquímica com a biomacromolécula (FERREIRA et al., 2015).

Os métodos de modelagem devem considerar que a afinidade de um ligante por uma biomacromolécula é resultado de contribuições entálpicas e entrópicas favoráveis e desfavoráveis à complexação, que resultam na variação da energia livre de Gibbs padrão (ΔG^0) do processo é calculada pela equação (13), em que ΔH^0 e ΔS^0 são a variação de entalpia padrão e a variação de entropia padrão, respectivamente, e T é a temperatura absoluta do sistema (DU et al., 2016).

$$\Delta G^0 = \Delta H^0 - T * \Delta S^0 \quad (13)$$

O nível de afinidade ligante-biomacromolécula depende da estabilidade do complexo formado e está diretamente correlacionado com a magnitude da variação de energia entre os estados complexado e dissociado. A constante de equilíbrio de associação K_{as} pode ser calculada através da equação (14) na qual R representa a constante universal dos gases, e T é a temperatura absoluta do sistema (DU et al., 2016). Pela equação, fica claro que o equilíbrio da associação de um ligante à biomacromolécula alvo é favorável ($K_{as} > 1$) quando a energia livre de Gibbs padrão diminui no processo, ou seja, quando ΔG^0 é negativo.

$$\Delta G^0 = -RT \ln K_{as} \quad (14)$$

Os principais ganhos entálpicos da complexação de um ligante com uma biomacromolécula são oriundos de novas interações intermoleculares formadas no complexo. Além disso, deve-se considerar também os ganhos entrópicos devido à liberação de moléculas do solvente para o meio e os ganhos entálpicos devido à formação de novas interações entre as moléculas do solvente liberado e do meio. Por outro lado, durante a formação do complexo ocorrem perdas energéticas devido a penalidades entálpicas (mudanças conformacionais) e entrópicas (redução das entropias translacional, rotacional e torcional, principalmente) tanto para o ligante quanto para a biomacromolécula. Somado a isso, há, ainda, perdas entálpicas devido às interações intermoleculares de moléculas do solvente com o ligante/a biomacromolécula que são desfeitas (DU et al., 2016; GUIMARÃES, 2012).

Neste contexto, as modificações estruturais de compostos bioativos, que têm como objetivo otimizar a afinidade e a seletividade pelo alvo terapêutico, buscam melhorar os ganhos energéticos e reduzir as penalidades entrópicas e entálpicas do processo de complexação (DU et al., 2016). Embora o cálculo de todos os parâmetros entrópicos e entálpicos com recursos computacionais mais acessíveis seja bastante custoso e lento (por exemplo, através do uso de ferramentas da termodinâmica estatística), alguns parâmetros importantes podem ser estimados satisfatoriamente mesmo com recursos modestos, como as interações intermoleculares entre ligante e biomacromolécula, as perdas entálpicas devido a mudanças conformacionais resultantes da formação dos complexos e as perdas entrópicas de dessolvatação do ligante (COLODETTE et al., 2020; GUIMARÃES, 2012; OLIVEIRA et al., 2006).

Os sistemas podem ser estudados utilizando diferentes níveis de teoria, como os métodos clássicos de mecânica molecular ou os métodos quânticos, dependendo da precisão desejada e dos recursos computacionais disponíveis (SANT'ANNA, 2009). Os métodos de mecânica molecular implementam simplificações dos termos de energia através do uso de modelos de mecânica clássica para representar os átomos e suas ligações (VANOMMESLAEGHE; GUVENCH; MACKERELL, 2014). Nestes modelos, campos de força contendo diversas funções de energia são utilizados para se avaliar as penalidades energéticas ao se afastar da estrutura mais estável (SANT'ANNA, 2009). As funções de energia incluem termos ligados, como estiramento de ligações, deformações angulares e torção de ângulos diedros, e

termos não ligados, como interações eletrostáticas e interações de van der Waals (MASETTI; ROCCHIA, 2014). Os métodos quânticos, por sua vez, são baseados na resolução da equação de Schrödinger (15), na qual Ψ representa a função de onda, E denota a energia total do sistema e $\hat{H}$ é o operador Hamiltoniano. Como esta equação somente pode ser resolvida de maneira exata para sistemas com apenas 1 elétron, diversas aproximações são empregadas para permitir a sua resolução aproximada para sistemas mais complexos. Os métodos semi-empíricos, por exemplo, incluem aproximações na resolução que reduzem o custo computacional, como a parametrização de determinados termos utilizando dados experimentais e o cálculo explícito incluindo somente os elétrons de valência (RYDE; SÖDERHJELM, 2016).

$$\hat{H}\psi = E\psi \quad (15)$$

4.1.3 Os métodos de ancoragem molecular

Os métodos de ancoragem molecular são utilizados para simular computacionalmente a complexação de um ligante no alvo terapêutico. As suas principais aplicações na descoberta de fármacos incluem a predição do modo de ligação, a identificação de compostos ativos em uma biblioteca e, potencialmente, a predição do valor absoluto de afinidade de uma molécula pelo alvo molecular (GUEDES; PEREIRA; DARDENNE, 2018; WASZKOWYCZ; CLARK; GANCIA, 2011). Os métodos de ancoragem compreendem um algoritmo de busca, utilizado para gerar variações na conformação, orientação e posição das estruturas, e uma função de pontuação, empregada para predizer a energia de interação entre o ligante e a biomacromolécula para as soluções geradas (WASZKOWYCZ; CLARK; GANCIA, 2011).

Os primeiros métodos de ancoragem consideravam o ligante e o alvo molecular como rígidos, de modo que variações na conformação do ligante eram exploradas através de sua inclusão de maneira explícita no banco de dados. O aumento da capacidade computacional permitiu a implementação de métodos semirrígidos, nos quais o ligante é considerado flexível e os átomos do alvo são mantidos fixos. Estes métodos são amplamente aplicados em estudos de ancoragem

molecular atualmente, visto o balanço entre a capacidade preditiva e o custo computacional. Nos métodos semirrígidos, a flexibilidade dos ligantes pode ser explorada de maneira sistemática, através da geração de um conjunto extenso de conformações que abrangem de forma representativa todo o espaço conformacional, ou de maneira estocástica, pelo uso de métodos de amostragem aleatória, como o Monte Carlo ou algoritmos evolutivos (ex.: algoritmos genéticos). A exploração simultânea da flexibilidade do ligante e do alvo pode ser realizada por simulações de dinâmica molecular (CHAPUT; MOUAWAD, 2017; WASZKOWYCZ; CLARK; GANCIA, 2011; COUPEZ; LEWIS, 2006). Devido ao alto custo computacional da dinâmica molecular, diversas estratégias têm sido estudadas para considerar a flexibilidade do alvo na ancoragem molecular semirrígida, como a utilização de diversas conformações de uma mesma biomacromolécula (*ensemble docking*), a variação das conformações das cadeias laterais dos aminoácidos utilizando bibliotecas de rotâmeros ou a redução das penalidades associadas à repulsão de van der Waals, também conhecido como *soft docking* (SALMASO; MORO, 2018).

As diversas soluções geradas pelo algoritmo de busca são avaliadas pela função de pontuação, que têm como objetivo classificar as soluções, para que seja possível a seleção daquelas que melhor prevejam ou reproduzam o modo de ligação do ligante no alvo. Um outro objetivo é usar os resultados da função de pontuação para prever a afinidade dos ligantes pelo alvo, mas, devido à alta complexidade do processo, às limitações e simplificações aplicadas nos cálculos e à amostragem limitada das conformações do ligante e da biomacromolécula, os resultados apresentam uma acurácia reduzida e a área ainda está em desenvolvimento (COUPEZ; LEWIS, 2006; PANTSAR; POSO, 2018). Classicamente, as funções pontuação são frequentemente segmentadas em: funções baseadas em campos de força, funções empíricas e funções baseadas em conhecimento (do inglês, *knowledge-based*) (GUEDES; PEREIRA; DARDENNE, 2018). As funções de pontuação baseadas em campos de força calculam a soma das contribuições das energias ligadas (termos intramoleculares, como os associados a distâncias e ângulos das ligações, assim como os ângulos diedros) e não-ligadas (termos intermoleculares, como interações eletrostáticas e de van der Waals) associadas à formação dos complexos proteína:ligante (P:L). As interações de van der Waals são comumente modeladas utilizando funções como a de Lennard-Jones e as interações eletrostáticas pela função de Coulomb (FERREIRA et al., 2015; SALMASO; MORO, 2018). As

funções de pontuação empíricas habitualmente são modelos de regressão gerados a partir da correlação de dados experimentais de afinidade com descritores considerados relevantes para o processo de formação do complexo P:L, como número de ligações de hidrogênio, interações lipofílicas, energia de dessolvatação e termos entrópicos (COUPEZ; LEWIS, 2006). As funções de pontuação baseadas em conhecimento (do inglês, *knowledge-based*) consideram os pares de interações atômicas entre o ligante e a biomacromolécula. A função visa a identificar as interações favoráveis a partir da análise de sua frequência presentes em co-cristais obtidos experimentalmente (FERREIRA et al., 2015; WASZKOWYCZ; CLARK; GANCIA, 2011).

4.1.4 Funções de pontuação empíricas baseadas em métodos de aprendizado de máquinas

Diversos estudos demonstraram que os métodos de ancoragem molecular comumente resultam em alta acurácia para a predição do modo de ligação, porém têm uma performance significativamente reduzida na identificação de moléculas ativas e na predição do nível de afinidade absoluta de um composto por uma biomacromolécula. A baixa performance das funções de pontuação pode ser ocasionada por limitações na predição da energia de solvatação e das variações de energia associadas às perdas e aos ganhos entrópicos dos ligantes e das biomacromoléculas, assim como por uma amostragem ineficiente das conformações dos componentes do sistema (GUEDES; PEREIRA; DARDENNE, 2018). Neste contexto, diversas pesquisas na área de modelagem molecular têm explorado algoritmos de aprendizado de máquina no desenvolvimento de novas funções de pontuação empíricas (AIN et al., 2015; LI et al., 2020; WASZKOWYCZ; CLARK; GANCIA, 2011).

As funções de pontuação empíricas foram inicialmente desenvolvidas utilizando métodos de regressão linear múltipla (RLM) para correlacionar os descritores do processo de complexação com os dados de afinidade experimentais. O método de RLM é considerado um algoritmo de aprendizado de máquinas do tipo linear (GUEDES; PEREIRA; DARDENNE, 2018). O desenvolvimento de funções de pontuação utilizando algoritmos de aprendizado de máquina não lineares, como a floresta aleatória, as máquinas de vetores de suporte e o *Naïve Bayes* tem o potencial

de aprimorar a performance das predições ao possibilitar a identificação de correlações não lineares no banco de dados. De fato, alguns autores já reportaram maior capacidade preditiva para funções baseadas em algoritmos de aprendizado de máquinas, quando comparadas a funções clássicas, como os trabalhos de Wójcikowski, Ballester e Siedlecki (2017), Wang e Zhang (2017) e Shen et al. (2020).

4.2 MATERIAL E MÉTODOS

4.2.1 Equipamentos e programas

4.2.1.1 Equipamentos

- Estação de trabalho com processador Intel i7-2670QM CPU @ 2.20 GHz, 8 GB RAM e placa gráfica Intel HD Graphics 3000.
- Estação de trabalho com 6 x Intel i7-6800K CPU @ 3.40 GHz, 16 GB RAM e placa de vídeo NVIDIA GTX 1080 8 GB.
- Servidor Dell Poweredge M620 com 2 x Intel(R) Xeon(R) CPU E5-2630 v2 @ 2.60GHz e 64 GB RAM.

4.2.1.2 Programas

- GOLD Suite 5.2 (Cambridge Crystallographic Data Centre, Cambridge, Reino Unido)
- Maestro v2019-4 (Schrödinger, LLC, Nova Iorque, EUA)
- LigPrep v2019-4 (Schrödinger, LLC, Nova Iorque, EUA)
- Epik v2019-4 (Schrödinger, LLC, Nova Iorque, EUA)
- Glide v2019-4 (Schrödinger, LLC, Nova Iorque, EUA)
- MOPAC2016 (Stewart Computational Chemistry, Colorado Springs, EUA)
- OpenBabel v2.3.1 (O'BOYLE et al., 2011)
- KNIME v3.6.0 (University of Konstanz, Zurique, Suíça)
- PyMOL v2.4.0 (Schrödinger, LLC, Nova Iorque, EUA)
- Python v3.8.2

4.2.2 Estruturas cristalográficas e biblioteca de compostos

4.2.2.1 Estruturas cristalográficas de TcDHODH

As estruturas cristalográficas de TcDHODH foram obtidas do *Protein Data Bank* (PDB) (BERMAN et al., 2000). Os 45 cristais utilizados neste trabalho, apresentados na Tabela 6, foram selecionados pois haviam sido co-cristalizados com inibidores de atividade biológica conhecida. Os cristais apresentam alta resolução, a qual é $< 2 \text{ \AA}$ para 38 das estruturas e $< 3 \text{ \AA}$ para as 7 restantes.

4.2.2.2 Biblioteca de inibidores da TcDHODH

Obtiveram-se 73 inibidores da TcDHODH das publicações de Inaoka et al. (2016, 2017). Estes compostos incluem o orotato, produto da reação enzimática catalisada pela enzima, e mais 72 análogos que tiveram a sua capacidade de inibir a TcDHODH avaliada nas mesmas condições experimentais. Dos 73 inibidores, 45 haviam sido co-cristalizados com a TcDHODH e possuíam estrutura cristalográfica disponível no PDB, conforme detalhado na sessão anterior. Como foram explorados modelos classificatórios neste trabalho, foi necessário estabelecer um critério para separar os compostos de alta afinidade de moléculas menos potentes. Compostos com uma constante de inibição (K_i) $< 500 \text{ nM}$ foram considerados de alta afinidade e aqueles com $K_i > 500 \text{ nM}$ como de baixa afinidade. Este critério foi estabelecido com o objetivo de balancear a relação entre o número de moléculas em ambas as classes, evitando, assim, que os algoritmos de aprendizado de máquinas empregados enviesassem a predição para a classe com maior número de moléculas (CAPUZZI et al., 2018; NEVES et al., 2015; ZAKAHROV et al., 2014). Além disso, 1.400 *decoys*, compostos presumidamente inativos e com propriedades físico-químicas similares aos compostos ativos na TcDHODH, foram gerados utilizando a plataforma *Directory of Useful Decoys, Enhanced* (DUD-E) (MYSINGER et al., 2012). O número de *decoys* incluídos na biblioteca de compostos foi determinado com base nas análises de Truchon e Bayly (2007), que sugerem que as métricas de enriquecimento (utilizadas para avaliar modelos preditivos classificatórios) apresentam menor variabilidade quando o número de compostos ativos representa menos de 5% de toda a biblioteca de moléculas.

4.2.2.3 Bibliotecas de fármacos aprovados e candidatos a fármaco

2.367 fármacos aprovados e 1.618 candidatos a fármaco em desenvolvimento clínico foram obtidos da plataforma Drug Repurposing Hub, disponibilizada pelo Instituto Broad (CORSELLO et al., 2017).

4.2.3 Processamento dos ligantes e das proteínas

O processamento das estruturas para a modelagem molecular levou em consideração as recomendações apresentadas na literatura (FORLI, 2015; SASTRY et al., 2013). Os ligantes foram preparados utilizando as ferramentas LigPrep e Epik da Schrödinger (SHELLEY et al., 2007). Inicialmente, os sais foram removidos e as coordenadas tridimensionais dos compostos foram geradas a partir dos seus SMILES (*Simplified Molecular Input Line-Entry System*). Em seguida, diferentes estruturas do mesmo composto foram criadas para representar os enantiômeros de centros quirais com a estereoquímica não definida, assim como os possíveis estados de protonação em pH $7,0 \pm 2,0$ e seus tautômeros (GREENWOOD et al., 2010). A geometria dos ligantes foi otimizada com o campo de força OPLS3e (ROOS et al., 2019). As estruturas cristalográficas foram preparadas utilizando a ferramenta Protein Preparation Wizard da Schrödinger (SASTRY et al., 2013). Inicialmente, buscou-se avaliar se havia átomos, cadeias laterais ou resíduos ausentes. Em seguida, as moléculas de água distantes dos ligantes e cofatores do cristal ($> 5 \text{ \AA}$) e os artefatos de cristalização foram removidos, a ordem das ligações foi atribuída e os hidrogênios foram adicionados (FRIESNER et al., 2004; HALGREN et al., 2004). A determinação do estado de protonação dos aminoácidos e a otimização das ligações de hidrogênio do complexo P:L foram realizadas com as ferramentas Epik (SHELLEY et al., 2007) e PROPKA (LI; ROBERTSON; JENSEN, 2005). Brevemente, o processo consistiu em explorar tautômeros de resíduos histidina, estados de protonação de histidina, ácido aspártico e glutâmico e rotações das cadeias laterais das asparaginas, glutaminas e histidinas e dos hidrogênios de grupos hidroxila e tiol. Após a otimização, removeram-se as moléculas de água que realizavam menos de 3 ligações de hidrogênio com outros compostos do cristal.

4.2.4 Análise preliminar das estruturas cristalográficas

As interações entre os inibidores da TcDHODH e a enzima foram avaliadas visualmente e utilizando o servidor *Protein-Ligand Interaction Profiler* – PLIP (SALENTIN et al., 2015), capaz de identificar as interações intermoleculares não covalentes como ligações de hidrogênio, interações hidrofóbicas, interações de empilhamento π - π , interações cátion- π , interações iônicas e ligações de halogênio. Os 45 co-cristais foram sobrepostos com base nos carbonos alfa (C_α) das proteínas, utilizando o programa Maestro da Schrödinger, e analisados visualmente.

4.2.5 Estudo de reancoragem: Avaliação da capacidade dos métodos de ancoragem molecular em reproduzir as poses experimentais dos inibidores na TcDHODH

A reancoragem é uma estratégia que consiste em remover um ligante co-cristalizado de uma estrutura cristalográfica e avaliar a capacidade dos algoritmos computacionais em reproduzir a pose experimental do co-cristal. Este estudo foi realizado para as estruturas cristalográficas 3W7C, 3W7D, 3W7O, 3W7P e 3W84, selecionados por apresentarem ótimas resoluções (1,75 Å, 1,52 Å, 1,68 Å, 1,70 Å e 1,93 Å, respectivamente) e ligantes co-cristalizados que apresentam alta afinidade pela enzima ($K_i \leq 200$ nM). As funções de pontuação ASP (MOOIJ; VERONK, 2005), ChemPLP (KORB; STÜTZLE; EXNER, 2009), ChemScore (VERDONK et al., 2003) e GoldScore (JONES et al., 1997) do programa GOLD e a *standard precision* (SP) do Glide (FRIESNER et al., 2004; HALGREN et al., 2004) foram empregadas. No GOLD, a esfera de ancoragem foi gerada tendo a posição do ligante nativo da estrutura cristalográfica como centro e expandida com um raio de 10, 15 ou 20 Å. No Glide, a caixa de ancoragem foi gerada como um cubo de aresta 10, 15 ou 20 Å, também centrada no ligante nativo. O algoritmo genético de busca do GOLD foi utilizado no modo padrão de eficiência de amostragem. As ancoragens moleculares, executadas em triplicata, foram realizadas no modo semirrígido, no qual o ligante é considerado flexível e a posição da maioria dos átomos da enzima é mantida fixa. No programa GOLD, as geometrias dos grupos hidroxí dos aminoácidos Ser, Thr e Tyr e das aminas dos aminoácidos Lys foram otimizadas durante a ancoragem. No programa Glide, todos os átomos da biomacromolécula alvo foram considerados rígidos. Nos estudos

de reancoragem, a importância das moléculas de água para as interações do ligante com a proteína foi avaliada através da realização das modelagens utilizando as estruturas nas quais estas moléculas haviam sido mantidas ou removidas. Todos os resultados foram avaliados considerando o valor da raiz do desvio quadrático médio (RMSD, do inglês *Root Mean Square Deviation*) entre as coordenadas atômicas da ancoragem em comparação com os dados experimentais. O RMSD foi calculado utilizando a equação (16), na qual δ representa a distância euclidiana entre cada átomo da estrutura do ligante no co-cristal do seu correspondente na estrutura gerada durante a ancoragem molecular e N representa o número total de átomos considerados. Os valores de RMSD foram obtidos através do servidor *on-line* DockRMSD (BELL; ZHANG, 2019).

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta^2} \quad (16)$$

4.2.6 Avaliação da capacidade dos algoritmos de ancoragem molecular em identificar compostos ativos na TcDHODH

A capacidade do método de ancoragem molecular em identificar compostos ativos na TcDHODH foi avaliada utilizando a função de pontuação ASP, devido à sua melhor performance no estudo de reancoragem. As funções de pontuação ChemPLP, ChemScore e GoldScore também foram avaliadas para comparação. As estruturas cristalográficas empregadas foram a 3W84 e a 3W7J, as quais, com base nos resultados do estudo de reancoragem, tiveram todas as águas co-cristalizadas removidas antes do início das modelagens. O procedimento foi realizado utilizando a mesma metodologia do estudo de reancoragem (4.2.5). Os resultados foram avaliados com base nas métricas de enriquecimento: área sob a curva ROC (*Receiver Operating Characteristics*) (AUC ROC), área sob a curva de discriminação de ROC aprimorada por Boltzmann (*Boltzmann-enhanced discrimination of ROC*, AUC BEDROC), com $\alpha = 5$ (GEPPERT; VOGT; BAJORATH, 2010; TRUCHON; BAYLY, 2007), e o fator de enriquecimento para 1% e 10% dos compostos com maior pontuação (EF1 e EF10). A avaliação da capacidade preditiva do método foi realizada com as duas bibliotecas de compostos, descritas a seguir.

4.2.6.1 Série de ancoragem A

Biblioteca composta pelos 73 inibidores da TcDHODH descritos por Inaoka et al. (2016, 2017), classificados como compostos de alta ($K_i < 500$ nM) ou baixa ($K_i > 500$ nM) afinidade pela enzima reportada na literatura.

4.2.6.2 Série de ancoragem B

Biblioteca composta pelos 73 inibidores da TcDHODH descritos por Inaoka et al. (2016, 2017), classificados como compostos de alta ou baixa afinidade, além dos 1.400 *decoys* gerados através do servidor DUD-E (MYSINGER et al., 2012), presumidamente inativos.

4.2.7 Desenvolvimento e validação do modelo de reclassificação

Nesta etapa do trabalho, um modelo de aprendizado de máquina foi desenvolvido com o objetivo de reclassificar os resultados da ancoragem molecular obtidos com a função ASP, de modo a aprimorar a capacidade preditiva do protocolo de triagem virtual. Os descritores utilizados no desenvolvimento do modelo incluíram: a entalpia de interação intermolecular entre o ligante e o alvo, calculada utilizando o método semi-empírico PM7 (STEWART, 2013), a lipofilicidade dos ligantes (clogP) e o termo E_{tor} da função de pontuação ChemPLP (KORB; STÜTZLE; EXNER, 2009). O valor de E_{tor} estima as perdas entrópicas associadas às restrições conformacionais impostas ao ligante como consequência de sua inserção no sítio ativo do alvo terapêutico (COLODETTE et al., 2020; KORB; STÜTZLE; EXNER, 2009; MANNHOLD et al., 2009)

4.2.7.1 Séries de compostos

Os 73 inibidores de TcDHODH descritos por Inaoka et al. (2016, 2017) foram utilizados para o desenvolvimento de validação dos modelos. Os compostos foram divididos através de um sorteio aleatório para cada classe (compostos de alta ou baixa afinidade) para compor a série treino e a série teste, na proporção 80% e 20%, respectivamente.

4.2.7.2 Entalpia de interação pelo método semi-empírico PM7

Os cálculos da entalpia de interação entre os ligantes e a proteína alvo foram realizados com base nas metodologias descritas por Oliveira et al. (2006), Rocha e Sant'Anna (2019) e Colodette et al. (2020). A entalpia de interação (ΔH_{int}) foi calculada utilizando a equação (17), na qual $\Delta H_f^{ligante}$, $\Delta H_f^{proteína}$ e $\Delta H_f^{complexo}$ representam as entalpias de formação do ligante, da proteína e do complexo proteína:ligante (P:L).

$$\Delta H_{int} = \Delta H_f^{complexo} - (\Delta H_f^{ligante} + \Delta H_f^{proteína}) \quad (17)$$

As coordenadas estruturais dos 73 inibidores da TcDHODH no sítio de ativo da enzima foram obtidas de duas maneiras: (i) através do PDB para os 45 compostos que foram co-cristalizados experimentalmente e (ii) através de ancoragem molecular para os 28 compostos restantes. Os 28 compostos foram ancorados em diferentes co-cristais: 6b no 3W1Q; 4c no 3W1U; 5n no 3W3O; 5bb, 5cc, 5t no 3W6Y; 5q e 5r no 3W7D; 6f, 5z e 5w no 3W7J; 5x e 5y no 3W7M; 4z no 3W7O; 4bb, 4cc e 4w no 3W7P; 6d no 3W7Q; 4aa no 3W85; 12a, 12b, 12d, 6a e 6c no 3W86; 13a e 6e no 3W87; 13d no 3W88 e 4u no 4JD4. Esta seleção foi inicialmente baseada na comparação entre cada um dos 28 compostos e os ligantes co-cristalizados, com o objetivo de identificar pares de moléculas que fossem isômeros de posição (ex.: 13a versus 13b) ou que apresentassem modificações em apenas um dos substituintes do anel (ex.: 4b versus 4c). No caso dos compostos em que nenhuma dessas condições foi identificada (6a, 6c, 6d, 6e e 6f), a seleção do cristal foi realizada através de um sorteio aleatório. A ancoragem molecular foi realizada utilizando a função de pontuação ASP, conforme metodologia empregada no estudo de reancoragem (4.2.5). Neste estudo, utilizou-se a esfera de ancoragem com raio 15 Å. A pose de maior pontuação para cada ancoragem foi utilizada para os cálculos.

Visando a otimizar a eficiência e os custos computacionais dos cálculos, somente os aminoácidos próximos ao ligante foram considerados. Para este objetivo, a proteína foi recortada com o programa Maestro, em duas seleções diferentes, de modo a manter o ligante, o cofator da enzima (flavina mononucleotídeo) e todos os aminoácidos com, pelo menos, um átomo a um raio de 5 ou 7 Å dos ligantes. Em cada caso, a seleção dos aminoácidos foi realizada de modo a garantir que os mesmos

resíduos estivessem presentes em todos os complexos. Os diferentes tamanhos de recorte da proteína foram empregados para estudar a influência deste parâmetro no modelo em desenvolvimento. Após o recorte, átomos de hidrogênio foram adicionados às ligações peptídicas truncadas para completar o estado de valência dos átomos da proteína.

A entalpia de formação do complexo P:L ($\Delta H_f^{complexo}$), do ligante isolado ($\Delta H_f^{ligante}$) e do sítio de ligação sem o ligante ($\Delta H_f^{proteína}$) foram calculadas pelo método PM7 (STEWART, 2013), disponível no programa MOPAC2016, após otimização da geometria dos complexos utilizando o método EF (*Eigenvector Following*) (BAKER, 1986). Durante as otimizações, os átomos da cadeia principal da proteína tiveram as suas coordenadas fixadas visando a evitar a desestruturação da proteína truncada. As otimizações foram realizadas em três etapas sequenciais: primeiramente somente os átomos de hidrogênio de todo o sistema, depois todos os átomos do ligantes e os átomos de hidrogênio da proteína e, finalmente, a otimização das cadeias laterais dos aminoácidos, o ligante, o cofator e todos os átomos de hidrogênio. O método dos orbitais moleculares localizados MOZYME foi empregado para otimizar a eficiência dos cálculos para os complexos P:L e para a proteína isolada (STEWART, 1996).

O modelo de solvatação de contínuo COSMO (*Conductor-like Screening Model*) foi empregado para simular o meio através do uso de uma constante dielétrica (ϵ) (KLAMT; SCHÜÜMANN, 1993). Calculou-se a entalpia de formação do ligante utilizando $\epsilon = 78,4$, de modo a simular os efeitos do meio aquoso. Como o sítio de ligação e o complexo P:L estão imersos na proteína, há necessidade de empregar uma constante dielétrica adequada. Embora diferentes valores tenham sido propostos na literatura, não há consenso com relação a uma única constante que seja adequada para todas as situações. Li et al. (2013) sugeriram que a constante dielétrica seria entre 6-7 para sítios de ligação no interior da proteína e entre 20-30, em uma região mais próxima à sua superfície. A constante dielétrica igual a 4,0 tem sido amplamente utilizada na literatura (BLEIZIFFER; SCHALLER; RINIKER, 2018; DUAN et al., 2003; GILSON; HONIG, 1988; LI et al., 2013). Neste trabalho, duas constantes dielétricas foram exploradas nos cálculos, $\epsilon = 4,0$ e 25,0, com vistas a determinar aquela que resultaria no modelo de maior capacidade preditiva.

4.2.7.3 Parâmetros de lipofilicidade e perda de energia pela restrição conformacional do ligante

Os ganhos entálpicos provenientes das interações intermoleculares formadas entre o ligante e a proteína são estimados de maneira eficiente através do cálculo de ΔH_{int} . Entretanto, a complexação de um ligante com uma proteína é influenciada por diversas contribuições entálpicas e entrópicas favoráveis e desfavoráveis ao processo (GUIMARÃES, 2012), além da entalpia de interação, conforme discutido anteriormente. Deste modo, a inclusão de termos que representem as demais contribuições entálpicas e entrópicas para o processo de complexação tem o potencial de aprimorar a capacidade preditiva dos modelos (WANG et al. (1994). Neste trabalho, as perdas entrópicas associadas às restrições conformacionais impostas ao ligante como consequência de sua inserção no sítio ativo da TcDHODH foram estimadas utilizando o termo de E_{tor} da função de pontuação ChemPLP (KORB; STÜTZLE; EXNER, 2009), conforme proposto por Colodette et al. (2020), o qual estima as perdas entrópicas em razão de restrições conformacionais do ligante devidas à complexação ao alvo. A lipofilicidade dos ligantes foi estimada a partir do coeficiente de partição (clogP) determinado pelo método de Mannhold (MANNHOLD et al., 2009).

4.2.7.4 Algoritmo de aprendizado de máquina

O algoritmo de aprendizado de máquina *Naive Bayes* foi utilizado para o desenvolvimento do modelo. Este algoritmo foi escolhido por apresentar uma boa performance em estudos descritos na literatura (ZHU et al., 2020) e, também, no capítulo sobre LBDD do presente trabalho. Além disso, o *Naive Bayes* não possui hiperparâmetros a serem otimizados e apresenta baixo custo computacional (LEE et al., 2019). A descrição detalhada dos princípios do *Naive Bayes* pode ser encontrada na sessão 3.1.3 deste trabalho.

4.2.7.5 Métodos de validação dos modelos de aprendizado de máquina

Os métodos de validação cruzada quádrupla (*5-fold cross validation*), teste de aleatorização e validação externa (TROPSHA; GRAMATICA; GOMBAR, 2003) foram

empregados para avaliar o desempenho dos modelos desenvolvidos, conforme a metodologia descrita na sessão 3.2.6 deste trabalho.

4.2.7.6 Métricas de avaliação dos modelos

A capacidade preditiva dos modelos classificatórios foi avaliada utilizando as métricas sensibilidade, especificidade, taxa de classificação balanceada, poder preditivo positivo e negativo, AUC ROC, AUC BEDROC e os fatores de enriquecimento para 1% e 10% das moléculas de maior pontuação (EF1 e EF10) (GEPPERT; VOGT; BAJORATH, 2010; TRUCHON; BAYLY, 2007), descritas na sessão 3.2.5 deste trabalho.

4.2.8 Fluxo geral do desenvolvimento, validação e aplicação dos modelos de reclassificação da ancoragem molecular em TcDHODH

Os modelos de aprendizado de máquina do trabalho SBDD foram desenvolvidos buscando seguir as mesmas premissas do estudo LBDD, descritas na sessão 3.2.9, em particular as recomendações da OECD (2007), as quais incluem: (i) um desfecho claro, (ii) algoritmo não ambíguo, (iii) domínio de aplicabilidade definido, (iv) métricas adequadas de robustez e poder preditivo e, se possível, (v) capacidade de interpretar as correlações entre os descritores e o desfecho em análise.

O fluxograma resumido das etapas do trabalho SBDD está apresentado na Figura 21. Inicialmente, estudos de reancoragem foram realizados para avaliar a capacidade das funções de pontuação em prever o modo de ligação na TcDHODH e identificar compostos ativos na enzima. Posteriormente, um modelo de aprendizado de máquinas para a reclassificação dos resultados obtidos na ancoragem molecular com a função ASP foi desenvolvido e validado.

Figura 21 – Fluxograma do desenvolvimento e validação de modelos computacionais para a identificação de inibidores da TcDHODH

Fonte: Produção do autor.

4.2.9 Triagem virtual de fármacos e candidatos a fármaco

A triagem virtual foi realizada com as bibliotecas de fármacos aprovados e de candidatos a fármaco em desenvolvimento clínico. Os compostos foram inicialmente ancorados no cristal 3W7J de TcDHODH, no qual as moléculas de água haviam sido removidas, utilizando a função de pontuação ASP, com uma esfera de ancoragem de raio 15 Å centralizada no ligante nativo, utilizando a metodologia previamente empregada no estudo de reancoragem (4.2.5). As 100 moléculas de maior pontuação de cada biblioteca (100 fármacos aprovados e 100 candidatos a fármaco) foram selecionadas para a predição de atividade com o modelo de aprendizado de máquinas.

A entalpia de interação foi calculada conforme o procedimento descrito na seção 4.2.7.2. A proteína foi recortada de modo a manter todos os resíduos de aminoácidos com pelo menos 1 átomo a um raio de 7 Å dos ligantes. A constante dielétrica $\epsilon = 25,0$ foi empregada para o cálculo da energia de formação do sítio de ligação e do complexo P:L. As otimizações de geometria foram realizadas em três etapas sequenciais: primeiramente somente os átomos de hidrogênio, depois todos os átomos do ligantes e os átomos de hidrogênio da proteína e finalmente as cadeias laterais dos aminoácidos, o ligante, o cofator e todos os átomos de hidrogênio. As perdas entrópicas (E_{tor}) e a lipofilicidade dos ligantes (clogP) foram respectivamente estimadas através da função de pontuação ChemPLP e do coeficiente de partição

calculado pelo método de Mannhold. O modelo de aprendizado de máquinas NB – TcDHODH - 8, baseado no algoritmo *Naive Bayes* e previamente validado, foi utilizado para reclassificar as 200 moléculas inicialmente selecionadas. Para os compostos classificados como ativos, a situação regulatória e as vias de administração das formas farmacêuticas disponíveis no mercado/em desenvolvimento foram analisadas a partir dos bancos de dados *DrugBank* (WISHART et al., 2018), *GlobalData* (2020), *PubMed*, *Repurposing Hub* (CORSELLO et al., 2017) e *NCATS Inxight: Drugs* (NCATS, 2020).

4.3 RESULTADOS

4.3.1 Estruturas cristalográficas e inibidores da TcDHODH

A análise da literatura permitiu identificar 55 estruturas cristalográficas de TcDHODH depositadas no PDB. As 45 estruturas que haviam sido co-cristalizadas com inibidores da enzima com atividade biológica conhecida estão apresentadas na Tabela 6. Os inibidores da TcDHODH identificados na literatura e utilizados neste trabalho estão apresentados nas Tabelas 7 e 8. Vale ressaltar que o próprio orotato, produto da reação enzimática catalisada pela TcDHODH, inibe a enzima com $K_i = 5,51 \mu\text{M}$ (INAOKA et al., 2016).

Tabela 6 – Estruturas cristalográficas de TcDHODH co-cristalizadas com inibidores de atividade biológica conhecida (continua)

PDB	R _{fator}	R _{livre}	Fator de temperatura	Resolução (Å)	Ligante	TcDHODH K_i (nM)
3W7J	0,15	0,18	12,5	1,58	5u	24
3W6Y	0,20	0,28	23,5	2,68	5v	33
3W84	0,20	0,25	17,7	1,93	5p	44
3W7D	0,15	0,17	15,8	1,52	5k	46
3W7L	0,14	0,19	16,3	1,88	5g	106
3W7C	0,14	0,20	15,6	1,75	5l	110
3W7E	0,14	0,17	14,6	1,56	5f	128
3W7P	0,14	0,18	15,5	1,70	4y	130
3W73	0,15	0,18	16,2	1,78	5c	138

Tabela 6 – Estruturas cristalográficas de TcDHODH co-cristalizadas com inibidores de atividade biológica conhecida (conclusão)

PDB	R _{fator}	R _{livre}	Fator de temperatura	Resolução (Å)	Ligante	TcDHODH K _i (nM)
3W3O	0,15	0,19	20,5	1,96	5h	142
3W7N	0,18	0,26	37,7	2,39	5aa	154
3W85	0,15	0,20	20,1	2,00	4v	169
3W7M	0,17	0,25	30,0	2,40	5s	175
3W7Q	0,14	0,18	20,5	1,83	4dd	180
3W7O	0,14	0,18	15,9	1,68	4x	200
3W7K	0,15	0,18	17,2	1,61	5m	206
3W7H	0,14	0,18	17,0	1,67	5b	222
3W7I	0,14	0,18	15,7	1,69	5j	223
4JDB	0,17	0,21	22,2	1,82	5i	291
3W75	0,14	0,17	14,7	1,47	4t	297
3W83	0,19	0,30	21,6	2,80	5o	305
3W88	0,14	0,17	14,6	1,40	13c	377
3W1U	0,15	0,19	16,7	1,85	4b	427
3W87	0,15	0,19	13,9	1,43	13b	481
3W74	0,17	0,23	23,9	1,90	4m	497
3W76	0,14	0,17	16,0	1,58	4r	581
3W1T	0,14	0,18	13,6	1,68	4j	616
3W7G	0,14	0,17	14,8	1,55	4q	714
3W2N	0,15	0,19	14,8	1,96	4i	739
3W1X	0,14	0,16	12,3	1,45	4o	749
3W86	0,14	0,18	17,8	1,50	12c	763
3W72	0,14	0,17	16,7	1,55	5a	763
3W71	0,14	0,18	18,1	1,68	4e	788
3W2U	0,21	0,27	23,1	2,25	4l	886
3W2K	0,14	0,17	14,7	1,54	4h	1.020
3W2M	0,14	0,17	13,7	1,58	4k	1.070
3W22	0,16	0,21	23,9	1,98	4d	1.100
3W70	0,20	0,28	25,5	2,60	4p	1.110
3W23	0,14	0,18	15,0	1,48	4f	1.610
3W2J	0,14	0,17	13,0	1,42	4g	1.970
3W2L	0,13	0,17	14,7	1,64	4s	2.660
3W1R	0,16	0,18	13,0	1,58	4a	3.070
4JD4	0,14	0,16	16,2	1,51	4n	3.090
2E6A	0,16	0,19	11,3	1,64	Orotato	5.510
3W1Q	0,15	0,19	14,3	1,85	1	> 19.700

R_{fator} e R_{livre} representam o quanto os modelos de refinamento explicam os dados experimentais utilizados no seu desenvolvimento (R_{fator}) ou utilizados somente para a sua validação (R_{livre}). K_i – Constante de inibição aparente. A estrutura dos ligantes está apresentada nas Tabelas 7 e 8. Os compostos e as estruturas cristalográficas foram descritos por Inaoka et al. (2016, 2017).

Tabela 7 – Derivados do orotato que capazes de inibir a TcDHODH (continua)

 4a $K_i = 3.070$ nM	 4b $K_i = 427$ nM	 4c $K_i = 1.190$ nM	 4d $K_i = 1.100$ nM
 4e $K_i = 788$ nM	 4f $K_i = 1.610$ nM	 4g $K_i = 1.970$ nM	 4h $K_i = 1.020$ nM
 4i $K_i = 739$ nM	 4j $K_i = 616$ nM	 4k $K_i = 1.070$ nM	 4l $K_i = 886$ nM
 4m $K_i = 497$ nM	 4n $K_i = 3.090$ nM	 4o $K_i = 749$ nM	 4p $K_i = 1.110$ nM
 4q $K_i = 714$ nM	 4r $K_i = 581$ nM	 4s $K_i = 2.660$ nM	 4t $K_i = 297$ nM
 4u $K_i = 7.780$ nM	 4v $K_i = 169$ nM	 4w $K_i = 2.110$ nM	 4x $K_i = 200$ nM
 4y $K_i = 130$ nM	 4z $K_i = 818$ nM	 4aa $K_i = 2.340$ nM	 4bb $K_i = 1.610$ nM
 4cc $K_i = 896$ nM	 4dd $K_i = 180$ nM	 5a $K_i = 763$ nM	 5b $K_i = 222$ nM

Tabela 7 – Derivados do orotato que capazes de inibir a TcDHODH (conclusão)

5c $K_i = 138$ nM	5f $K_i = 128$ nM	5g $K_i = 106$ nM	5h $K_i = 142$ nM
5i $K_i = 291$ nM	5j $K_i = 223$ nM	5k $K_i = 46$ nM	5l $K_i = 110$ nM
5m $K_i = 206$ nM	5n $K_i = 882$ nM	5o $K_i = 305$ nM	5p $K_i = 44$ nM
5q $K_i = 847$ nM	5r $K_i = 3.410$ nM	5s $K_i = 175$ nM	5t $K_i = 317$ nM
5u $K_i = 24$ nM	5v $K_i = 33$ nM	5w $K_i = 272$ nM	5x $K_i = 788$ nM
5y $K_i = 259$ nM	5z $K_i = 311$ nM	5aa $K_i = 154$ nM	5bb $K_i = 561$ nM
5cc $K_i = 1.760$ nM	6a $K_i = 2.960$ nM	6b $K_i = 2.040$ nM	6c $K_i = 7.440$ nM
6d $K_i = 5.570$ nM	6e $K_i = 6.750$ nM	6f $K_i = 2.610$ nM	1 $K_i > 19.700$ nM

K_i – Constante de inibição aparente frente à TcDHODH. Os compostos foram descritos por Inaoka et al. (2016, 2017).

Tabela 8 – Derivados do orotato que capazes de inibir a TcDHODH

12a $K_i = 1.400$ nM	12b $K_i = 2.140$ nM	12c $K_i = 763$ nM	12d $K_i = 2.210$ nM
13a $K_i = 1.070$ nM	13b $K_i = 481$ nM	13c $K_i = 377$ nM	13d $K_i = 197$ nM

K_i – Constante de inibição aparente frente à TcDHODH. Os compostos foram descritos por Inaoka et al. (2016, 2017).

4.3.2 Análise do modo de ligação e das interações intermoleculares entre os inibidores derivados do orotato e a TcDHODH

A análise dos 45 co-cristais contendo os inibidores da TcDHODH permitiu identificar as principais interações formadas entre os ligantes e a enzima. O modo de ligação e as interações intermoleculares do ligante 5l no sítio ativo da TcDHODH (estrutura 3W7C no PDB) está apresentado na Figura 22 e na Tabela 9, respectivamente, de modo a ilustrar os principais achados. Neste cristal, o núcleo 2,6-dioxo-1,2,3,6-tetraidro-pirimidina realiza diversas ligações de hidrogênio com as cadeias laterais dos resíduos aminoácidos Asn67, Asn127, Asn194 e Ser195. O carboxilato, por sua vez, recebe ligações de hidrogênio dos nitrogênios peptídicos dos resíduos Met69, Gly70 e Leu71 e faz uma interação iônica com a cadeia lateral do resíduo Lys43. Além das interações com a proteína, o ligante também realiza uma interação de empilhamento paralelo π – π com o cofator da enzima, o mononucleotídeo de flavina (FMN).

Figura 22 – Estrutura cristalográfica do composto 5I co-cristalizado com a TcDHODH

A estrutura cristalográfica experimental foi obtida do PDB (3W7C). O inibidor 5I e o cofator estão representados com os átomos de carbono nas cores verde e ciano, respectivamente. A superfície da proteína (A), suas estruturas secundárias (B) e os carbonos dos seus aminoácidos (C) estão representados na cor cinza. Asn – asparagina. Ser – serina. Lys – Lisina. Leu – leucina. Gly – glicina. Met – metionina. Fonte: Produção do autor.

Tabela 9 – Interações intermoleculares entre o composto 5l e a TcDHODH

Resíduo	Tipo de interação	Região do aminoácido	Número de interações	Distância (Å)
Leu71	Interação hidrofóbica	Cadeia lateral	2	3,77 e 3,73
Asn67	Ligação de hidrogênio	Cadeia lateral	2	2,98 e 2,91
Ser68	Ligação de hidrogênio	Cadeia lateral	1	3,80
Met69	Ligação de hidrogênio	Cadeia principal	1	3,10
Gly70	Ligação de hidrogênio	Cadeia principal	1	2,94
Leu71	Ligação de hidrogênio	Cadeia principal	1	3,83
Asn127	Ligação de hidrogênio	Cadeia lateral	1	2,96
Asn194	Ligação de hidrogênio	Cadeia lateral	2	2,78 e 2,73
Ser195	Ligação de hidrogênio	Cadeia lateral	1	2,80
Lys43	Interação iônica	Cadeia lateral	1	4,09

A estrutura cristalográfica experimental foi obtida do PDB (3W7C). As interações intermoleculares entre o ligante e a proteína foram identificadas através da plataforma *on-line* PLIP. Para as ligações de hidrogênio, a distância reportada representa a distância entre o átomo doador e o aceptor. No caso das interações hidrofóbicas, a distância considerada é aquela entre os átomos de carbono. Para as interações eletrostáticas, a distância computada foi aquela entre os centros carregados. Asn – asparagina. Ser – serina. Lys – Lisina. Leu – leucina. Gly – glicina. Met – metionina.

Como os ligantes reportados por Inaoka et al. (2016, 2017) são derivados do orotato e possuem o núcleo 2,6-dioxo-1,2,3,6-tetraidro-pirimidina, todos apresentam o mesmo modo de interação na TcDHODH ao redor deste núcleo. A Figura 23, que apresenta a sobreposição das 45 estruturas cristalográficas com base nos C α da proteína, evidencia esta observação. Deste modo, as variações nas interações intermoleculares com a proteína estão associadas aos substituintes presentes na posição 5 do núcleo 2,6-dioxo-1,2,3,6-tetraidro-pirimidina. Observou-se, ainda, que há uma alça flexível próxima ao sítio ativo, formada pelos aminoácidos Ser129, Cys130, Pro131, Asn132, Val133, Pro134, Gly135, Lys136, Pro137, Gln138 e Val139, a qual apresenta variações em sua conformação nos diferentes cristais. Notavelmente, no cristal 2E6A (orotato co-cristalizado com a TcDHODH), esta alça está em uma posição mais próxima do inibidor, reduzindo, assim, o tamanho do sítio de ligação.

Figura 23 – Sobreposição das estruturas cristalográficas de TcDHODH.

Os inibidores da TcDHODH e os cofatores da enzima (FMN) estão representados com os átomos de carbono nas cores verde e ciano, respectivamente. Exceto pelo cristal 2E6A, cujas estruturas secundárias da proteína estão representadas em roxo, as demais estruturas secundárias estão apresentadas na cor cinza. (A) Comparação do cristal 2E6A, seu ligante e cofator com as 44 estruturas cristalográficas restantes (verificar Tabela 6). (B) Sobreposição das estruturas cristalográficas de TcDHODH, seus ligantes e cofatores, exceto pelo PDB 2E6A. Fonte: Produção do autor.

4.3.3 Estudo de reancoragem: Avaliação da capacidade dos métodos de ancoragem molecular em reproduzir as poses experimentais dos inibidores na TcDHODH

Os resultados do estudo de reancoragem foram avaliados considerando a média dos valores de RMSD (Tabela 10) e o número de poses ancorados que apresentaram $\text{RMSD} > 2\text{\AA}$ (Figura 24). O limite de $\text{RMSD} \leq 2\text{\AA}$ como referência para determinar o sucesso na reprodução do modo de ligação experimental é amplamente

utilizado na literatura (HEVENER et al., 2010; MAGALHAES; BARBOSA; DARDENNE, 2004; TEN BRINK; EXNER, 2009). Conforme o esperado, diversas poses obtidas durante o estudo de reancoragem apresentaram baixos valores de RMSD, sugerindo que os métodos de ancoragem foram capazes de reproduzir as poses experimentais dos inibidores baseados em orotato.

A influência das moléculas de água na reancoragem foi avaliada através da comparação dos resultados obtidos para os cristais nos quais estas haviam sido mantidas ou removidas. Observou-se menor número de poses com RMSD > 2 Å nas modelagens em que as moléculas de água foram mantidas. Entretanto, alguns algoritmos de ancoragem tiveram perda de performance mais significativa com esta remoção (Figura 24). Por exemplo, o número total de poses incorretas para as ancoragens com o Glide SP aumentou de 6 para 18 com a remoção das águas. Por outro lado, piora menos expressiva foi observada para a função GoldScore (de 9 para 10). A função ASP, por sua vez, foi aquela que apresentou o melhor desempenho nas ancoragens na ausência de moléculas de água, porém somente quando o raio da esfera de ancoragem foi inferior a 15 Å.

Tabela 10 – Valores de RMSD obtidos no estudo de reancoragem dos inibidores de TcDHODH (continua)

Função de pontuação	Raio da esfera de ancoragem (Å)	Moléculas de água	Estrutura cristalográfica				
			3W7J	3W84	3W7C	3W7P	3W85
ASP	10	Presentes	0,39 (0,14)	0,41 (0,21)	0,49 (0,06)	0,55 (0,32)	0,75 (0,24)
		Removidas	1,07 (0,33)	1,07 (0,44)	0,62 (0,26)	0,35 (0,13)	2,16 (1,11)
	15	Presentes	0,35 (0,03)	0,47 (0,14)	1,02 (0,80)	0,64 (0,23)	0,93 (0,05)
		Removidas	0,83 (0,55)	1,61 (1,71)	0,54 (0,07)	0,50 (0,07)	1,71 (1,20)
	20	Presentes	0,72 (0,63)	1,33 (1,34)	0,62 (0,18)	0,66 (0,31)	2,17 (1,23)
		Removidas	4,42 (0,38)	3,10 (2,13)	0,73 (0,16)	0,51 (0,05)	1,69 (1,34)
ChemPLP	10	Presentes	0,66 (0,66)	2,26 (3,16)	1,09 (0,79)	0,31 (0,02)	0,71 (0,08)
		Removidas	0,38 (0,19)	1,83 (1,57)	0,80 (0,06)	0,59 (0,09)	2,75 (0,02)
	15	Presentes	0,29 (0,06)	6,22 (4,94)	0,63 (0,20)	0,32 (0,02)	0,69 (0,15)
		Removidas	1,26 (0,44)	2,82 (3,39)	1,06 (0,73)	0,59 (0,09)	2,76 (0,01)
	20	Presentes	0,29 (0,04)	0,72 (0,62)	1,45 (0,86)	0,34 (0,07)	1,83 (1,68)
		Removidas	3,12 (1,48)	3,80 (3,51)	0,61 (0,06)	0,53 (0,17)	2,82 (0,10)

Tabela 10 – Valores de RMSD obtidos no estudo de reancoragem dos inibidores de TcDHODH (conclusão)

Função de pontuação	Espaço de ancoragem (Å)	Moléculas de água	Estrutura cristalográfica				
			3W7J	3W84	3W7C	3W7P	3W85
ChemScore	10	Presentes	0,20 (0,03)	0,61 (0,41)	0,96 (0,82)	0,39 (0,10)	1,32 (0,87)
		Removidas	1,90 (1,78)	1,78 (1,35)	0,62 (0,11)	0,63 (0,06)	2,65 (0,09)
	15	Presentes	0,21 (0,01)	3,39 (0,30)	1,49 (0,80)	0,41 (0,09)	1,35 (1,19)
		Removidas	0,64 (0,66)	2,59 (2,49)	0,60 (0,38)	0,66 (0,03)	2,59 (0,24)
	20	Presentes	0,19 (0,05)	3,52 (5,10)	1,92 (0,01)	0,35 (0,04)	1,75 (0,75)
		Removidas	3,89 (0,18)	2,28 (2,51)	0,88 (0,88)	0,54 (0,02)	1,35 (1,23)
GoldScore	10	Presentes	0,77 (0,01)	0,47 (0,02)	0,37 (0,09)	0,31 (0,01)	2,88 (0,02)
		Removidas	1,18 (0,65)	1,35 (0,21)	0,40 (0,02)	0,73 (0,13)	2,85 (0,02)
	15	Presentes	0,60 (0,29)	0,47 (0,05)	0,34 (0,03)	0,33 (0,01)	2,88 (0,01)
		Removidas	0,34 (0,03)	1,27 (0,06)	0,46 (0,07)	0,78 (0,07)	2,88 (0,03)
	20	Presentes	0,78 (0,04)	0,52 (0,10)	0,50 (0,04)	0,32 (0,04)	3,05 (0,14)
		Removidas	1,42 (1,91)	1,36 (0,14)	0,48 (0,06)	0,80 (0,11)	2,93 (0,13)
Glide XP	10	Presentes	0,27 (0)	3,51 (0)	0,20 (0)	0,82 (0)	1,12 (0)
		Removidas	3,41 (0)	3,5 (0)	0,81 (0)	0,82 (0)	3,01 (0)
	15	Presentes	0,68 (0)	0,49 (0)	0,29 (0)	0,98 (0)	1,10 (0)
		Removidas	3,41 (0)	3,52 (0)	0,39 (0)	0,99 (0)	3,12 (0)
	20	Presentes	0,91 (0)	0,54 (0)	0,39 (0)	3,30 (0)	1,15 (0)
		Removidas	4,26 (0)	0,44 (0)	0,59 (0)	3,16 (0)	3,13 (0)

A reancoragem foi realizada com os cristais 3W7J, 3W84, 3W7C, 3W7P e 3W85, com as moléculas de água removidas ou mantidas, utilizando diferentes tamanhos do raio da esfera de ancoragem (ASP, ChemPLP, ChemScore ou GoldScore) ou da aresta da caixa de ancoragem (Glide SP). RMSD – Raiz do desvio quadrático médio (do inglês *Root Mean Square Deviation*). As simulações foram realizadas em triplicatas e os dados estão apresentados como o valor médio (desvio padrão). Os valores de RMSD < 2 Å estão destacados.

Figura 24 – Performance dos algoritmos de ancoragem molecular na reancoragem de ligantes na TcDHODH

A reancoragem foi realizada com os cristais 3W7J, 3W84, 3W7C, 3W7P e 3W85, com as moléculas de água removidas ou mantidas, utilizando diferentes tamanhos do raio da esfera de ancoragem (ASP, ChemPLP, ChemScore ou GoldScore) ou da aresta da caixa de ancoragem (Glide SP). RMSD – Raiz do desvio quadrático médio (RMSD, do inglês *Root Mean Square Deviation*). As simulações foram realizadas em triplicata e o número de resultados com RMSD > 2 Å foram comparados. Fonte: Produção do autor.

Considerando o alto desempenho da função de pontuação ASP mesmo na ausência das moléculas de água, buscou-se avaliar mais detalhadamente os tipos de erros presentes nas poses com os piores valores de RMSD para esta função. A análise foi realizada para as ancoragens em esferas de raio 15 Å, considerando-se o seu maior potencial para acomodar maior variedades de compostos comparativamente à esfera de raio 10 Å. Conforme apresentado na Figura 25, mesmo as poses com RMSD > 2 Å (RMSD = 3,10 Å para o cristal 3W85 e 3,58 Å para o cristal 3W84) conseguiram replicar parcialmente as interações moleculares conhecidas para o complexo P:L, como as ligações de hidrogênio do núcleo 2,6-dioxo-1,2,3,6-tetraidro-pirimidina com os resíduos Asn67, Asn127, Asn194 e Ser195 e as interações do carboxilato na posição 6 com os resíduos Met69, Gly70, Leu71 e Lys43. Os anéis fenila e naftila dos inibidores, substituintes na posição 5 do anel, ocuparam uma região próxima àquela que ocupam nos ligantes co-cristalizados, porém associados a uma pose diferente. Conclui-se, portanto, que o desempenho da função ASP foi razoável para as ancoragens em cristais sem as moléculas de água e, por isso, esta função de pontuação foi empregada nas etapas seguintes deste trabalho.

Figura 25 – Reancoragem para os cristais de TcDHODH utilizando a função ASP e a esfera de ancoragem de raio 15 Å

A reancoragem foi realizada com os cristais 3W84 e 3W85, cujas moléculas de água haviam sido removidas, utilizando a função ASP e uma esfera de ancoragem com raio 15 Å. As modelagens foram realizadas em triplicata e as poses com os maiores valores de RMSD estão apresentadas (RMSD = 3,58 Å para a 3W84 e 3,10 Å para a 3W85). RMSD – Raiz do desvio quadrático médio (RMSD, do inglês *Root Mean Square Deviation*). Os átomos de carbono do ligante ancorado na proteína estão representados em verde. As estruturas cristalográficas obtidas do PDB estão com os átomos de carbono do inibidor na cor ciano, os átomos de carbono do cofator FMN na cor azul e as estruturas secundárias da proteína na cor cinza. Fonte: Produção do autor.

4.3.4 Avaliação da capacidade dos algoritmos de ancoragem molecular em identificar compostos ativos na TcDHODH

As funções de pontuação ASP, ChemPLP, ChemScore e GoldScore foram avaliadas com relação a sua capacidade de identificar inibidores da TcDHODH em uma triagem virtual com as Séries de ancoragem A e B, compostas respectivamente pelos 73 inibidores da enzima descritos por Inaoka et al. (2016, 2017) e pelos mesmos 73 inibidores em conjunto com 1.400 *decoys* (Séries de ancoragem A e B). Embora a função de pontuação ASP tenha apresentado a melhor performance na reancoragem, optou-se por testar todas as demais funções disponíveis no programa GOLD em paralelo, de modo a comparar a performance dos algoritmos.

Nas modelagens com a Série de ancoragem A (Figuras 26 e 27), observou-se capacidade preditiva intermediária (AUC ROC entre 0,60 e 0,69) ou próxima à seleção aleatória (AUC ROC entre 0,50 e 0,59), dependendo da função de pontuação e do cristal empregados. As melhores performances foram observadas para as ancoragens em esferas com raio de 10 Å ou 15 Å. Ainda, a função ASP foi a única que apresentou valores de AUC ROC entre 0,60 e 0,69 para ambos os cristais, quando se empregaram esferas com raio 10 Å e 15 Å.

Figura 26 – Curvas ROC da ancoragem molecular na TcDHODH com a Série de ancoragem A

As ancoragens foram realizadas com as funções de pontuação ASP, ChemPLP, ChemScore e GoldScore, nos cristais 3W84 e 3W7J, cujas moléculas de água haviam sido removidas, utilizando esferas de ancoragem de diferentes raios (10 Å, 15 Å e 20 Å). A Série de ancoragem A era composta pelos 73 inibidores de TcDHODH descritos por Inaoka et al. (2016, 2017). Fonte: Produção do autor.

Figura 27 - Performance das funções de pontuação ASP, ChemPLP, ChemScore e GoldScore na ancoragem molecular na TcDHODH com a Série de ancoragem A

As ancoragens foram realizadas com as funções de pontuação ASP, ChemPLP, ChemScore e GoldScore, nos cristais 3W84 e 3W7J, cujas moléculas de água haviam sido removidas, utilizando esferas de ancoragem de diferentes raios (10 Å, 15 Å e 20 Å). EF10 – Fator de enriquecimento para 10% das moléculas com maior pontuação de predição. AUC ROC e AUC BEDROC – área sob as curvas ROC e BEDROC. A Série de ancoragem A era composta pelos 73 inibidores de TcDHODH descritos por Inaoka et al. (2016, 2017). Fonte: Produção do autor.

Na análise utilizando a Série de ancoragem B, composta pelos 73 inibidores da TcDHODH e pelos 1.400 *decoys* (Figuras 28 e 29), observou-se alta capacidade preditiva para as funções de pontuação ASP, ChemPLP e ChemScore e intermediária para a função GoldScore .

Figura 28 – Curvas ROC da ancoragem molecular na TcDHODH com a Série de ancoragem B

As ancoragens foram realizadas com as funções de pontuação ASP, ChemPLP, ChemScore e GoldScore, nos cristais 3W84 e 3W7J, cujas moléculas de água haviam sido removidas, utilizando esferas de ancoragem de diferentes raios (10 Å, 15 Å e 20 Å). A Série de ancoragem B era composta pelos 73 inibidores de TcDHODH e os 1.400 decoys. Fonte: Produção do autor.

Figura 29 – Performance das funções de pontuação ASP, ChemPLP, ChemScore e GoldScore na ancoragem molecular na TcDHODH com a Série de ancoragem B

As ancoragens foram realizadas com as funções de pontuação ASP, ChemPLP, ChemScore e GoldScore, nos cristais 3W84 e 3W7J, cujas moléculas de água haviam sido removidas, utilizando esferas de ancoragem de diferentes raios (10 Å, 15 Å e 20 Å). Nesta modelagem, os 73 inibidores da TcDHODH e os 1.400 decoys (Série de ancoragem B) foram avaliados. EF1 e EF10 – Fatores de enriquecimento para 1% e 10% das moléculas com maior pontuação de predição, respectivamente. AUC ROC e AUC BEDROC – área sob as curvas ROC e BEDROC. Fonte: Produção do autor.

Os resultados obtidos nesta seção sugerem que as funções ASP, ChemPLP e ChemScore seriam adequadas para a etapa inicial de uma triagem virtual, na qual o objetivo principal seja remover as moléculas completamente inativas. Por outro lado, como estas mesmas funções apresentam uma capacidade preditiva baixa/intermediária na seleção de ativos quando somente os 73 inibidores da TcDHODH foram avaliados (Série de triagem A), observa-se a oportunidade de desenvolver um método capaz de aprimorar a seleção de moléculas com alta

afinidade pela enzima. Deste modo, desenvolveu-se, neste trabalho, uma nova função de reclassificação baseada em um modelo de energia livre de Gibbs (WANG et al., 1994) que pudesse ser utilizada no refinamento dos resultados provenientes da ancoragem com funções de pontuação padronizadas no programa GOLD. Diferentemente do trabalho original, em que foi utilizado um método de correlação linear múltipla dos parâmetros calculados para predição das atividades, aqui empregamos o algoritmo de aprendizado de máquina *Naive Bayes*, capaz de identificar correlações não lineares entre os parâmetros (equivalentes àqueles do trabalho original) e a atividade biológica. O modelo desenvolvido foi validado através dos métodos estatísticos.

4.3.5 Desenvolvimento de um modelo de reclassificação por energia livre de Gibbs para a identificação de inibidores de TcDHODH

Os descritores utilizados para o desenvolvimento do modelo de energia livre de Gibbs para a reclassificação dos resultados da ancoragem molecular foram a entalpia de interação do ligante com a proteína (ΔH_{int} , calculada pela equação 17), assim como as perdas entrópicas devido a restrições conformacionais do ligante durante a complexação com o alvo (E_{tor}) e a lipofilicidade ($clogP$) do ligante. Os descritores foram selecionados com base em trabalhos que descreveram modelos preditivos para diversos alvos terapêuticos, como proteínas kinase (PI4KIII β e PK-C), fosfodiesterase (PDE4) e tirosina fosfatases (Shp2), empregando abordagens similares (COLODETTE et al., 2020; OLIVEIRA et al., 2006; ROCHA; SANT'ANNA, 2019; WANG et al., 2004). Como o método de cálculo semi-empírico do ΔH_{int} necessita das coordenadas do complexo P:L, utilizou-se a função de pontuação ASP para ancorar na TcDHODH os 28 ligantes que não possuíam co-cristais experimentais disponíveis na literatura. A função ASP foi selecionada para esta etapa devido à sua melhor performance no estudo de reancoragem. Conforme esperado, os 28 compostos apresentaram modo de ligação similar ao dos 45 ligantes co-cristalizados. Em particular, os núcleos 2,6-dioxo-1,2,3,6-tetraidro-pirimidina realizaram as mesmas ligações de hidrogênio com as cadeias laterais dos resíduos Asn67, Asn127, Asn194 e Ser195 e com a cadeia principal dos resíduos Met69, Gly70 e Leu71, anteriormente identificadas através do servidor PLIP.

Devido ao alto custo computacional para realizar os cálculos com todos os átomos da proteína, foi necessário recortá-la e incluir apenas os aminoácidos mais próximos aos ligantes. Neste trabalho, optou-se por explorar a influência dos aminoácidos incluídos no cálculo na performance do modelo através da seleção dos resíduos com, pelo menos, um átomo a 5 Å ou a 7 Å do ligante. A Tabela 11 apresenta os resíduos mantidos em cada tamanho de recorte da proteína e do complexo P:L.

As entalpias de interação foram calculadas a partir da diferença entre a entalpia de formação do complexo P:L e as entalpias de formação do ligante e da proteína isolados, conforme descrito previamente (eq. 8). A Tabela 12 apresenta os valores de ΔH_{int} para os cálculos com as diferentes constantes dielétricas no modelo de solvatação ($\epsilon = 4,0$ e $25,0$), inclusão de diferentes resíduos da proteína (resíduos de aminoácidos com um átomo a 5 Å ou 7 Å dos ligantes) e otimização da geometria de diferentes átomos estão apresentados. Os parâmetros que representam a lipofilicidade (clogP) e as perdas entrópicas do ligante (E_{tor}) durante a complexação com a TcDHODH estão compilados na Tabela 13.

Tabela 11 – Aminoácidos da TcDHODH utilizados nos cálculos da entalpia de interação pelo método semi-empírico

Tamanho do recorte (Å)	Número de aminoácidos	Aminoácidos
5	29	Lys43, Ser44, Thr46, Arg50, Gly52, Asn53, Pro54, Asn67, Ser68, Met69, Gly70, Leu71, Pro72, Ser99, Gly100, Leu101, Asn127, Ser129, Cys130, Pro131, Gln138, Lys164, Asn194, Ser195, Asn198, Pro213, Lys214, Gly218 e Gly219
7	62	Phe169, Asp170, Ile171, Ala19, Gly20, Lys43, Ser44, Cys45, Thr46, Arg50, Gly52, Asn53, Pro54, Arg57, Tyr58, Ile66, Asn67, Ser68, Met69, Gly70, Leu71, Pro72, Asn73, Ser97, Ile98, Ser99, Gly100, Leu101, Ser102, Glu105, Asn106, Met109, Asn127, Leu128, Ser129, Cys130, Pro131, Asn132, Pro137, Gln138, Val139, Ala140, Tyr141, Tyr149, Lys164, Pro166, Pro167, Val193, Asn194, Ser195, Val196, Gly197, Asn198, Ile211, Pro213, Lys214, Gln215, Phe217, Gly218, Gly219, Leu220 e Gly221

Os 73 complexos P:L e as respectivas proteínas sem o ligante foram recortados de modo a manter apenas os resíduos próximos ao ligante. Todos os aminoácidos com pelo menos um átomo a uma distância de 5 Å ou 7 Å dos ligantes foram incluídos nos cálculos. Lys – Lisina. Ser – Serina. Thr – Treonina. Arg – Arginina. Asn – Asparagina. Pro – Prolina. Met – Metionina. Gly – Glicina. Leu – Leucina. Cys – Cisteína. Gln – Glutamina. Phe – Fenilalanina. Asp – Asparagina. Ile – Isoleucina. Ala – Alanina. Tyr – Tirosina. Gly – Ácido glutâmico. Met – Metionina. Val – Valina.

Tabela 12 – Entalpia de interação dos derivados de orotato com a TcDHODH calculada pelo método semi-empírico PM7 (continua)

Ligante	ΔH_{int} (kcal/mol)							
	$\epsilon = 4,0$				$\epsilon = 25,0$			
	Otimização somente do ligante		Otimização também dos aminoácidos		Otimização somente do ligante		Otimização também dos aminoácidos	
	5 Å	7 Å	5 Å	7 Å	5 Å	7 Å	5 Å	7 Å
12a	-78,0	28,8	-155,7	-47,8	-76,6	15,6	-145,6	-85,2
12b	-80,2	26,5	-159,3	-66,0	-72,5	7,6	-156,1	-91,5
12c	-89,0	15,3	-180,2	-78,9	-95,7	-1,8	-169,4	-104,6
12d	-194,4	-80,7	-290,3	-167,8	-193,4	-91,9	-250,1	-182,0
13a	-142,1	-69,0	-199,6	-114,5	-141,9	-82,0	-187,0	-133,9
13b	-142,1	-63,9	-198,2	-113,9	-140,8	-84,9	-193,3	-142,5
13c	-227,7	-74,3	-352,6	-223,7	-221,0	-86,0	-320,4	-249,5
13d	-228,5	-58,1	-340,7	-225,5	-221,6	-78,1	-312,0	-232,9
4a	-126,5	-6,5	-150,9	-70,7	-120,5	-16,1	-139,4	-67,0
4aa	-157,3	-75,2	-210,8	-114,5	-145,9	-94,6	-190,9	-138,1
4b	-135,2	-48,4	-179,5	-111,0	-135,4	-60,9	-176,5	-136,6
4bb	-132,9	-44,3	-200,6	-108,3	-130,8	-58,3	-182,2	-128,4
4c	-140,3	-52,4	-197,1	-108,8	-143,9	-55,2	-187,3	-133,0
4cc	-135,3	-45,9	-196,7	-111,3	-132,4	-60,7	-194,5	-135,2
4d	-163,5	-93,8	-221,7	-138,7	-157,4	-106,0	-209,3	-146,0
4dd	-115,0	-18,7	-200,1	-81,5	-104,0	-19,2	-167,3	-114,3
4e	-162,3	-63,1	-224,0	-151,7	-157,5	-79,0	-203,0	-163,3
4f	-86,4	35,5	-187,4	-78,4	-82,3	7,2	-155,2	-99,6
4g	-76,4	51,6	-165,9	-29,0	-74,1	46,7	-156,0	-44,5
4h	-140,5	4,8	-196,7	-85,4	-138,6	-9,4	-184,7	-106,9
4i	-166,4	-69,8	-221,6	-134,8	-164,5	-82,7	-209,2	-145,1
4j	-182,5	-60,0	-217,6	-103,3	-184,5	-75,5	-214,4	-139,4
4k	-146,6	12,0	-214,5	-98,1	-144,8	-1,2	-202,9	-114,5
4l	-174,5	-90,4	-222,5	-124,9	-160,3	-83,8	-197,9	-100,4
4m	-159,6	-76,1	-216,5	-131,7	-158,5	-89,2	-202,9	-138,8
4n	-166,7	-65,9	-213,7	-122,1	-165,5	-79,1	-201,5	-138,5
4o	-62,2	85,4	-186,1	-56,9	-57,0	76,5	-167,6	-69,7
4p	-184,4	-144,0	-219,5	-156,7	-176,9	-151,4	-189,6	-174,6
4q	-75,5	43,9	-161,5	-51,5	-72,1	24,2	-144,2	-75,1
4r	-113,7	-29,2	-196,3	-104,2	-112,0	-39,0	-173,6	-115,1
4s	-52,9	84,5	-155,4	-32,9	-46,6	72,8	-143,9	-61,0
4t	-73,9	69,0	-167,5	-77,1	-71,6	58,9	-158,5	-101,9
4u	-159,0	-63,8	-207,7	-126,1	-159,3	-68,7	-189,1	-129,8
4v	-155,7	-82,3	-202,0	-115,7	-155,7	-94,9	-181,0	-140,4
4w	-135,8	-46,1	-208,8	-109,0	-134,0	-48,9	-190,0	-148,1
4x	-211,9	-62,1	-332,7	-220,5	-198,6	-69,8	-313,2	-231,4
4y	-286,4	-189,2	-365,9	-264,9	-266,0	-186,1	-337,3	-283,3
4z	-88,6	56,8	-200,4	-95,7	-61,2	45,9	-179,6	-114,1
5a	-87,8	12,8	-168,4	-76,5	-82,2	3,1	-150,6	-96,0
5aa	-145,7	-83,6	-210,6	-114,9	-145,9	-74,7	-198,6	-134,1
5b	-138,2	-15,5	-184,9	-96,7	-135,0	-30,8	-172,6	-119,6
5bb	-198,4	-143,4	-220,4	-148,7	-196,9	-136,6	-195,9	-178,1

Tabela 12 – Entalpia de interação dos derivados de orotato com a TcDHODH calculada pelo método semi-empírico PM7 (conclusão)

Ligante	ΔH_{int} (kcal/mol)							
	$\epsilon = 4,0$				$\epsilon = 25,0$			
	Otimização somente do ligante		Otimização também dos aminoácidos		Otimização somente do ligante		Otimização também dos aminoácidos	
	5 Å	7 Å	5 Å	7 Å	5 Å	7 Å	5 Å	7 Å
5c	-152,9	-82,5	-209,5	-136,8	-147,6	-94,7	-200,8	-163,4
5cc	-195,9	-147,9	-232,6	-164,8	-200,6	-148,7	-212,8	-178,1
5f	-176,3	-82,9	-222,1	-138,0	-177,3	-96,6	-201,2	-146,1
5g	-131,2	-18,3	-185,2	-91,4	-127,9	-35,6	-172,4	-83,2
5h	-133,2	-44,7	-199,2	-116,5	-129,4	-69,6	-182,1	-131,8
5i	-144,7	-61,3	-199,1	-117,1	-138,3	-64,9	-187,2	-130,6
5j	-113,3	-6,7	-196,2	-89,7	-105,3	-14,2	-158,4	-113,2
5k	-122,9	17,2	-193,4	-74,3	-109,7	4,4	-175,7	-104,3
5l	-118,7	-51,6	-186,2	-89,9	-129,2	-68,8	-174,2	-136,9
5m	-122,0	-55,3	-178,1	-91,9	-129,1	-73,3	-172,7	-135,0
5n	-133,7	-62,2	-199,4	-128,6	-126,3	-77,3	-182,2	-141,6
5o	-159,7	-90,5	-234,4	-125,3	-161,2	-91,7	-208,8	-176,9
5p	-164,1	-102,0	-207,5	-133,0	-158,9	-109,9	-185,2	-147,3
5q	-145,1	-2,7	-203,3	-91,3	-135,1	-4,4	-191,3	-90,9
5r	-121,0	14,1	-182,3	-67,2	-106,3	17,7	-169,3	-74,5
5s	-156,8	-29,2	-213,8	-92,0	-133,3	-42,9	-192,8	-106,7
5t	-204,8	-146,4	-230,4	-154,8	-199,2	-137,8	-200,9	-170,2
5u	-273,8	-188,6	-354,6	-251,9	-267,1	-197,9	-335,4	-270,7
5v	-359,8	-308,2	-405,0	-322,7	-345,5	-305,4	-369,2	-330,7
5w	-112,8	-10,7	-181,3	-103,4	-102,5	-31,4	-164,3	-128,7
5x	-148,8	-28,7	-207,4	-89,6	-148,5	-23,6	-195,7	-93,8
5y	-151,2	-27,6	-206,4	-86,5	-150,9	-29,4	-193,7	-110,4
5z	-109,6	-37,1	-193,0	-104,6	-98,0	-39,3	-170,9	-114,3
6a	-70,2	28,2	-149,1	-41,0	-76,5	13,9	-158,1	-78,3
6b	59,5	116,2	-27,3	63,7	36,8	100,2	-42,5	45,5
6c	-70,2	41,0	-158,6	-37,3	-81,7	15,4	-153,3	-77,5
6d	-114,2	-24,9	-195,4	-84,3	-99,6	-11,9	-170,3	-102,2
6e	-138,0	-61,7	-198,2	-101,9	-138,0	-78,4	-189,7	-141,7
6f	-91,1	-24,7	-157,8	-102,2	-70,1	-18,6	-140,0	-97,6
1	-138,2	-62,8	-216,4	-102,0	-141,0	-78,2	-206,2	-145,3
Orotato	-263,0	-277,1	-255,7	-265,0	-253,3	-273,7	-248,5	-252,5

A entalpia de interação (ΔH_{int}) entre os ligantes e a TcDHODH foi calculada utilizando a equação (17), a qual resulta da diferença entre as entalpias de formação do complexo P:L ($\Delta H_f^{\text{complexo}}$) e a energia de formação do ligante e da proteína isolados ($\Delta H_f^{\text{ligante}}$ e $\Delta H_f^{\text{proteína}}$, respectivamente). As entalpias de formação foram calculadas utilizando o método semi-empírico PM7. Antes dos cálculos, os complexos P:L e as proteínas foram recortados de modo a manter somente os aminoácidos próximos ao ligante (resíduos com pelo menos um átomo a uma distância de 5 Å ou 7 Å dos ligantes). O modelo de solvatação contínua COSMO foi empregado para calcular as entalpias de interação utilizando constantes dielétrica $\epsilon = 4,0$ (EPS4) e 25,0 (EPS25).

Tabela 13 – Parâmetros de perdas entrópicas durante a complexação e lipofilicidade (como clogP) dos derivados do orotato

Ligante	clogP	E _{tor}	Ligante	clogP	E _{tor}
12a	2,45	1,65	4z	2,12	1,37
12b	2,45	1,78	5a	2,67	1,22
12c	2,45	1,3	5aa	2,67	2,21
12d	2,34	2,26	5b	2,67	1,35
13a	2,89	2,51	5bb	2,67	1,56
13b	2,89	2,35	5c	3,11	1,47
13c	2,78	1,92	5cc	2,78	2,12
13d	2,78	1,47	5f	2,67	1,11
4a	2,23	1,12	5g	2,67	1,77
4aa	2,23	1,46	5h	2,67	1,56
4b	2,34	1,19	5i	2,67	1,06
4bb	2,12	1,39	5j	2,67	1,54
4c	2,45	1,38	5k	2,67	1,18
4cc	2,23	1,4	5l	2,67	1,16
4d	2,67	1,12	5m	2,67	1,97
4dd	2,67	1,36	5n	2,67	1,92
4e	2,89	2,76	5o	2,67	1,3
4f	2,12	1,28	5p	2,67	2,01
4g	2,01	1,42	5q	2,67	1,93
4h	2,23	1,19	5r	3,22	2,95
4i	2,01	1,79	5s	2,67	1,13
4j	2,01	1,47	5t	2,67	1,35
4k	2,01	1,64	5u	2,56	1,11
4l	1,79	1,35	5v	2,56	2,31
4m	1,9	1,7	5w	2,56	1,47
4n	2,23	1,44	5x	2,67	1,33
4o	2,34	1,4	5y	2,78	1,2
4p	2,78	1,6	5z	2,78	1,72
4q	2,23	1,25	6a	1,9	1,69
4r	2,23	1,12	6b	1,57	1,06
4s	2,23	1,23	6c	2,23	2,05
4t	2,23	1,31	6d	2,67	2,61
4u	2,23	2,09	6e	2,67	2,14
4v	2,23	1,52	6f	2,45	2,9
4w	2,23	1,2	1	2,01	1,21
4x	2,12	1,42	Orotato	1,35	0,91
4y	2,12	1,22			

E_{tor} – Perdas entrópicas, calculadas utilizando a função ChemPLP. clogP – coeficiente de partição n-octanol/água, calculado pelo método de Mannhold (MANNHOLD et al., 2009).

Os modelos de aprendizado de máquinas foram desenvolvidos utilizando o algoritmo *Naive Bayes* e validados através dos métodos validação cruzada quádrupla, teste de aleatorização e validação externa. Ao total, oito modelos de aprendizado de máquinas foram desenvolvidos, cada um utilizando uma das oito entalpias de

interação, calculadas por métodos diferentes, conforme detalhado na Tabela 14. Além da entalpia de interação, os descritores clogP e E_{tor} foram empregados em todos os modelos. Na validação cruzada quádrupla (Figuras 30 e 31 e Tabela 22), os modelos apresentaram valores de AUC ROC entre 0,75 e 0,81, evidenciando, assim, capacidade preditiva superior ao das funções de pontuação da ancoragem molecular (AUC ROC entre 0,45 e 0,69). Esse aumento de performance também foi evidenciado pela métrica AUC BEDROC, visto que os modelos de aprendizado de máquina apresentaram valores entre 0,66 e 0,74 e as funções de pontuação entre 0,42 e 0,69 (Figura 30 e Tabela 22). O modelo de maior capacidade preditiva, o NB – TcDHODH – 8, foi desenvolvido utilizando a entalpia de interação calculada com maior número de resíduos da proteína (7 Å), a constante dielétrica mais polar ($\epsilon = 25,0$) e na qual a geometria das cadeias laterais dos resíduos de aminoácidos foi também otimizada. Este modelo apresentou os maiores valores de AUC ROC (0,81), AUC BEDROC (0,74), TCB (77,4%), sensibilidade (76%), PPV (73,1%) e NPV (81,3%) dentre todos aqueles avaliados. No teste de aleatorização (Figuras 32 e 33 e Tabela 23), todos os modelos apresentaram redução significativa de performance, próxima à seleção aleatória. Este resultado confirma que as relações identificadas entre os descritores e a atividade de interesse na validação inicial não ocorreram ao acaso.

Tabela 14 – Métodos de cálculo das entalpias de interação entre o ligante e a proteína para os diferentes modelos de aprendizado de máquina de reclassificação

Modelo	Entalpia de interação		
	Constante dielétrica (ϵ)	Átomos otimizados	Raio do recorte da proteína (Å)
NB - TcDHODH - 1	4	Ligante e átomos de hidrogênio da proteína	5
NB - TcDHODH - 2			7
NB - TcDHODH - 3		Ligante, cofator, todos os hidrogênios do sistema e as cadeias laterais dos aminoácidos	5
NB - TcDHODH - 4			7
NB - TcDHODH - 5	25	Ligante e átomos de hidrogênio da proteína	5
NB - TcDHODH - 6			7
NB - TcDHODH - 7		Ligante, cofator, todos os hidrogênios do sistema e as cadeias laterais dos aminoácidos	5
NB - TcDHODH - 8			7

Oito modelos de aprendizado de máquinas foram desenvolvidos utilizando como descritores a entalpia de interação entre o ligante e a proteína (ΔH_{int}), calculada pelo método semi-empírico PM7, a lipofilicidade (clogP) e as perdas entrópicas do ligante decorrentes da complexação com a TcDHODH (E_{tor}). A entalpia de interação foi calculada de diferentes maneiras, com variações nos resíduos da proteína incluídos nos cálculos, os átomos que tiveram a sua geometria otimizada e a constante dielétrica do modelo de solvatação COSMO. Cada modelo de aprendizado de máquinas foi desenvolvido a partir de uma única entalpia de interação específica.

Figura 30 – Performance dos modelos de aprendizado de máquina de reclassificação na validação cruzada quintupla

Os modelos de aprendizado de máquina foram desenvolvidos a partir da série teste e dos descritores da entalpia de interação do complexo P:L (ΔH_{int}), de lipofilicidade e das perdas entrópicas do ligante (clogP e E_{tor} , respectivamente). Diferentes modelos foram obtidos a partir de variações no método de cálculo da ΔH_{int} , que foram: os resíduos da proteína incluídos nos cálculos, os átomos que tiveram a sua geometria otimizada e a constante dielétrica do modelo de solvatação COSMO. A capacidade preditiva na validação cruzada quintupla foi avaliada utilizando os parâmetros: (A) curva ROC, (B e C) áreas sob as curvas ROC e BEDROC, respectivamente, e (D) taxa de classificação balanceada (TCB).
Fonte: Produção do autor.

Figura 31 – Sensibilidade, especificidade, poder preditivo positivo e negativo dos modelos de aprendizado de máquina de reclassificação na validação cruzada quádrupla

Os modelos de aprendizado de máquina foram desenvolvidos a partir da série teste e dos descritores da entalpia de interação do complexo P:L (ΔH_{int}), de lipofilicidade e das perdas entrópicas do ligante ($clogP$ e E_{tor} , respectivamente). Diferentes modelos foram obtidos a partir de variações no método de cálculo da ΔH_{int} , que foram: os resíduos da proteína incluídos nos cálculos, os átomos que tiveram a sua geometria otimizada e a constante dielétrica do modelo de solvatação COSMO. A capacidade preditiva na validação cruzada quádrupla foi avaliada utilizando os parâmetros: (A) Sensibilidade e especificidade e (B) Poder preditivo positivo e negativo. Fonte: Produção do autor.

Figura 32 – Performance dos modelos de aprendizado de máquina de reclassificação no teste de aleatorização

Os modelos de aprendizado de máquina foram desenvolvidos a partir da série teste e dos descritores da entalpia de interação do complexo P:L (ΔH_{int}), de lipofilicidade e das perdas entrópicas do ligante ($clogP$ e E_{tor} , respectivamente). Diferentes modelos foram obtidos a partir de variações no método de cálculo da ΔH_{int} , que foram: os resíduos da proteína incluídos nos cálculos, os átomos que tiveram a sua geometria otimizada e a constante dielétrica do modelo de solvatação COSMO. A capacidade preditiva no teste de aleatorização foi avaliada utilizando os parâmetros: (A) curva ROC, (B e C) áreas sob as curvas ROC e BEDROC, respectivamente, e (D) taxa de classificação balanceada (TCB). Fonte: Produção do autor.

Figura 33 – Sensibilidade, especificidade, poder preditivo positivo e negativo dos modelos de aprendizado de máquina de reclassificação no teste de aleatorização

Os modelos de aprendizado de máquina foram desenvolvidos a partir da série teste e dos descritores da entalpia de interação do complexo P:L (ΔH_{int}), de lipofilicidade e das perdas entrópicas do ligante (clogP e E_{tor} , respectivamente). Diferentes modelos foram obtidos a partir de variações no método de cálculo da ΔH_{int} , que foram: os resíduos da proteína incluídos nos cálculos, os átomos que tiveram a sua geometria otimizada e a constante dielétrica do modelo de solvatação COSMO. A capacidade preditiva no teste de aleatorização foi avaliada utilizando os parâmetros: (A) Sensibilidade e especificidade e (B) Poder preditivo positivo e negativo. Fonte: Produção do autor.

O modelo NB – TcDHODH - 8 foi posteriormente testado na validação externa com a Série teste (Figura 34). Nesta etapa, observou-se, também, alta capacidade preditiva, com valores de AUC ROC de 0,79 e AUC BEDROC de 0,85, assim como TCB, PPV e NPV próximos a 70%. Deste modo, o modelo NB – TcDHODH – 8 foi considerado validado para a reclassificação de resultados de ancoragem molecular utilizando a função de pontuação ASP.

Figura 34 – Performance do melhor de aprendizado de máquina de reclassificação NB – TcDHODH – 8 na validação externa

O modelo de aprendizado de máquina de melhor performance na validação cruzada foi testado com relação à sua capacidade preditiva de identificar inibidores da TcDHODH na série teste (validação externa). As entalpias de interação (ΔH_{int}) foram calculadas pelo método semi-empírico PM7. Os complexos P:L foram recortados de modo a incluir todos os resíduos com pelo menos 1 átomo à distância de 7 Å dos ligantes. O modelo de solvatação contínua COSMO foi empregado com a constante dielétrica de $\epsilon = 25,0$ (EPS25). As entalpias de interação foram calculadas após a otimização das cadeias laterais dos aminoácidos, do ligante, do cofator e dos hidrogênios da proteína. Os parâmetros avaliados foram: (A) sensibilidade (SENS.), especificidade (ESPEC.), taxa de classificação balanceada (TCB), poder preditivo positivo e negativo (PPV e NPV, respectivamente); (B) a curva ROC e (C) as áreas sob as curvas ROC e BEDROC. Fonte: Produção do autor.

4.3.6 Triagem virtual de fármacos e candidatos a fármaco

A triagem virtual dos 2.367 fármacos aprovados e 1.618 candidatos a fármaco em desenvolvimento clínico foi realizada utilizando a ancoragem molecular com a pontuação ASP e, em seguida, o modelo de aprendizado de máquinas de reclassificação NB – TcDHODH - 8. Os 100 fármacos e 100 candidatos a fármaco de maior pontuação na função ASP foram posteriormente submetidos ao modelo de reclassificação, o qual utiliza informações sobre a entalpia de interação entre os compostos e a proteína, calculados pelo método semi-empírico PM7, e termos para

representar a lipofilicidade do ligante e as perdas de energia conformacional das moléculas durante devido ao processo de complexação. Oito fármacos e dois candidatos a fármaco foram selecionados como ativos pelo modelo, os quais estão apresentados na Tabela 15. Na análise das informações disponíveis para estes compostos na literatura, foi possível constatar que cinco dos fármacos aprovados são adequados para a administração oral em humanos. No caso das moléculas em desenvolvimento, observou-se que ambas (uridina 5'-trifosfato e lometrexol) haviam sido descontinuadas no momento da análise. Como essas moléculas ainda poderiam ser reposicionadas para a terapêutica de Chagas, mesmo tendo sido descontinuadas para a indicação primária, optou-se por mantê-las como potenciais oportunidades de reposicionamento para a inibição da TcDHODH.

Tabela 15 – Fármacos e candidatos a fármaco selecionados como potencialmente ativos na TcDHODH através da ancoragem molecular e do modelo de reclassificação (continua)

Composto	Situação regulatória	Via de administração	ΔH_{int} (kcal/mol)	clogP	E_{tor}
Raltitrexede	Aprovado	Intravenosa	-240,00	2,56	1,20
Ixazomibe	Aprovado	Oral	-228,91	2,12	1,97
Levofolinato	Aprovado	Oral	-228,33	2,12	0,91
Pralatrexato	Aprovado	Intravenosa	-192,70	2,67	2,06
Pemetrexede	Aprovado	Intravenosa	-188,96	2,45	1,05

Tabela 15 – Fármacos e candidatos a fármaco selecionados como potencialmente ativos na TcDHODH através da ancoragem molecular e do modelo de reclassificação (conclusão)

Composto	Situação regulatória	Via de administração	ΔH_{int} (kcal/mol)	clogP	E_{tor}
Fosamprenavir	Aprovado	Oral	-184,81	2,67	4,21
Baicalina	Aprovado	Oral	-183,45	2,56	0,77
Silibinina	Aprovado	Oral	-147,02	3,11	1,31
Uridina 5'-trifosfato	Estudo clínico fase III	Inalatório	-416,90	0,25	1,56
Lometrexol	Estudo clínico fase II	Intravenosa	-272,91	2,56	1,24

Os compostos foram ancorados na TcDHODH (PDB: 3W7J) utilizando a função de pontuação ASP, a esfera de ancoragem de raio 15 Å e o modo padrão de busca do algoritmo genético. Os 100 fármacos aprovados e 100 candidatos a fármaco de maior pontuação foram selecionados para os cálculos de entalpia de interação e para a reclassificação utilizando o modelo de aprendizado de máquinas previamente validado. ΔH_{int} – Entalpia de interação do complexo P:L, calculada conforme a equação (17). As entalpias de formação da proteína e do complexo P:L foram calculadas com o método PM7, utilizando a constante dielétrica $\epsilon = 25,0$ e otimizando a geometria das cadeias laterais dos aminoácidos, todos os hidrogênios da proteína e os átomos do ligante e do cofator FMN. Os ligantes classificados como ativos pelo modelo de aprendizado de máquinas estão apresentados nesta tabela.

4.4 DISCUSSÃO

A doença de Chagas é uma enfermidade que afeta milhões de pessoas em todo o mundo e que carece de fármacos seguros e eficazes para o seu tratamento. Neste contexto, o objetivo da abordagem SBDD deste trabalho foi desenvolver e validar métodos capazes de identificar compostos ativos na TcDHODH, alvo terapêutico promissor para o combate ao *T. cruzi*. Os modelos foram posteriormente empregados na seleção de fármacos e candidatos a fármaco com o potencial de serem reposicionados para o tratamento da doença de Chagas através da inibição da TcDHODH.

Na análise da literatura, constatou-se que há um número limitado inibidores da TcDHODH descritos. Os compostos mais potentes disponíveis até o momento são derivados do orotato, produto da reação catalisada pela enzima, com substituições na posição 5 do núcleo 2,6-dioxo-1,2,3,6-tetraidro-pirimidina (INAOKA et al., 2016, 2017). Notavelmente, alguns destes análogos apresentam alta afinidade pela TcDHODH e seletividade frente à hDHODH. Entretanto estas moléculas ainda foram pouco caracterizadas com relação às propriedades de farmacocinética, segurança e eficácia em animais e humanos. Além disso, esta série química potencialmente necessitará de otimizações estruturais para resulte em um novo fármaco. Deste modo, apesar da P&D de inibidores da TcDHODH a partir dos derivados de orotato ser uma estratégia promissora no combate ao *T. cruzi*, a busca por oportunidades de reposicionamento pode permitir o desenvolvimento de novas terapias para a doença de Chagas de maneira acelerada e com menores investimentos e riscos de falha.

No trabalho de Yoshino et al. (2015), as principais interações intermoleculares existentes entre os derivados do orotato e a TcDHODH foram analisadas com vistas à determinação de um farmacóforo comum para esta série de compostos, o que poderia auxiliar no planejamento de novos análogos. Os principais achados foram ligações de hidrogênio com os resíduos Asn67, Asn194 e Lys43, assim como uma interação de empilhamento paralelo π com o mononucleotídeo de flavina. No presente trabalho, outras interações além das descritas por Yoshino e colaboradores também foram identificadas com alta frequência nos complexos P:L (> 95% dos complexos), como ligações de hidrogênio com as cadeias laterais da Ser195 e da Asn127 e com a cadeia principal dos resíduos Met69, Gly70 e Leu71. Estas diferenças podem ter sido causadas pelos critérios estabelecidos para caracterizar cada tipo de interação, como

a distância e o ângulo entre átomos. No presente trabalho, utilizou-se a plataforma PLIP para identificar as interações intermoleculares, a qual foi desenvolvida com critérios mais permissivos visando a permitir a análise mesmo de estruturas cristalográficas de baixa resolução (SALENTIN et al., 2015). As interações identificadas utilizando a plataforma PLIP estão de acordo com aquelas descritas por Inaoka et al. (2017).

A alta capacidade dos métodos de ancoragem molecular em prever o modo de interação experimental dos inibidores na TcDHODH está em linha com a acurácia reportada na literatura para estudos similares (WANG et al., 2016; WARREN et al., 2006). Neste trabalho, observou-se piora na capacidade dos métodos de ancoragem na predição do modo de interação como resultado da remoção de todas as moléculas de água presentes no co-cristal experimental. Dentre as funções de pontuação, a ASP apresentou menor perda de performance na ausência de moléculas de água. Nestes casos, os resultados de reancoragem com a função ASP foram considerados satisfatórios em esferas de ancoragem de raio igual ou inferior a 15 Å. A completa remoção de moléculas de água, apesar de reduzir as informações do modo de interação com o no alvo, é bastante empregada em estudos de ancoragem molecular, visto que reduz o custo computacional e o número de incertezas nas predições, assim como permite identificar compostos que resultariam no deslocamento destas moléculas de água devido a sua complexação com a proteína (FORLI, 2015; FRIESNER et al., 2004; SCIOR et al., 2012). Considerando as vantagens dessa estratégia e o desempenho satisfatório da função ASP no estudo de reancoragem, optou-se por utilizar, nas etapas posteriores, os cristais com a completa remoção das moléculas de água.

A predição da afinidade absoluta de um ligante por uma biomacromolécula utilizando métodos computacionais é uma tarefa desafiadora e que ainda está em desenvolvimento (GUEDES; PEREIRA; DARDENNE, 2018). Neste sentido, o uso de modelos para identificar compostos ativos sem a atribuição do valor da atividade biológica esperada, *i.e.*: métodos classificatórios, tem se mostrado como uma alternativa interessante ao desenvolvimento de modelos de regressão (AIN et al., 2015). Por este motivo, optou-se por explorar inicialmente métodos classificatórios no presente trabalho. Na primeira análise, os métodos de ancoragem molecular foram avaliados com relação às suas capacidades de selecionar compostos ativos a partir das Séries de Ancoragem A e B. As funções de pontuação ASP, ChemPLP,

ChemScore e GoldScore apresentaram alta performance na seleção de compostos ativos a partir da biblioteca contendo os derivados de orotato e os 1.400 *decoys* (Série de ancoragem B). Por outro lado, observou-se redução significativa da capacidade preditiva quando somente os 73 inibidores da TcDHODH foram avaliados (Série de ancoragem A). Estes dados sugerem que os métodos de ancoragem poderiam ser empregados para enriquecer o conjunto de moléculas selecionados com inibidores da enzima através da eliminação de compostos completamente inativos. Entretanto, as funções de pontuação exploradas não seriam capazes de diferenciar, de maneira eficaz, compostos de alta afinidade de outras moléculas que apresentem baixa inibição da TcDHODH.

Devido à complexidade e às limitações dos modelos computacionais empregados para a triagem virtual de compostos, combinações de métodos podem ser utilizadas em uma única campanha de triagem virtual com o objetivo de aprimorar a acurácia dos resultados (GUEDES; PEREIRA; DARDENNE, 2018; WANG; LIN, 2013). Considerando as limitações das funções de pontuação em identificar inibidores potentes da TcDHODH, conforme descrito anteriormente, optou-se por desenvolver um modelo de aprendizado de máquinas que seria utilizado para o refinamento dos resultados da primeira triagem virtual com métodos de ancoragem molecular. O modelo de reclassificação foi desenvolvido por meio do algoritmo de aprendizado de máquinas *Naive Bayes* utilizando alguns parâmetros associados à energia livre de Gibbs de interação, conforme sugerido por outros (WANG et al., 1994), incluindo a entalpia de interação com a proteína e o ligante, calculada neste trabalho com o método semi-empírico PM7. Os métodos quanto-mecânico (QM) semi-empíricos apresentam bom desempenho na quantificação da entalpia de interação intermolecular entre o ligante e a biomacromolécula e, deste modo, podem resultar em modelos de alta capacidade preditiva (CAVASOTTO; AUCAR, 2020; YILMAZER; KORTH, 2013). Por esta razão, optou-se por empregar o ΔH_{int} como um dos descritores para o modelo de reclassificação deste trabalho. Como os cálculos semi-empíricos necessitam de propostas iniciais para as geometrias dos complexos P:L para serem realizados, empregou-se a função ASP para gerar as poses iniciais, antes dos cálculos, visto a sua alta performance tanto no estudo de reancoragem quanto na identificação de compostos completamente inativos. Segundo Wang et al. (1994), a inclusão de parâmetros que representem a lipofilicidade e as perdas entrópicas dos ligantes devido à complexação aprimoraram a performance dos modelos

desenvolvidos. Por esta razão, outros descritores específicos foram incluídos para o desenvolvimento do modelo de reclassificação neste trabalho: a lipofilicidade do ligante, estimada pelo valor de clogP , e efeitos entrópicos associados à perda de movimentos torcionais do ligante após a complexação, estimados com o parâmetro E_{tor} da função de pontuação ChemPLP. Esta estratégia foi baseada em trabalhos que haviam reportado modelos de alta capacidade preditiva, utilizando abordagens similares (COLODETTE et al., 2020; OLIVEIRA et al., 2006; ROCHA; SANT'ANNA, 2019). O presente trabalho apresenta como diferencial o fato de ter empregado um algoritmo de aprendizado de máquinas, capaz de identificar correlações não lineares entre os dados. Deste modo, há o potencial de atingir maior capacidade preditiva quando comparado aos estudos previamente citados.

Notavelmente, diversos modelos apresentaram alta performance na validação cruzada quádrupla. Associado a isto, a completa perda de capacidade preditiva no teste de aleatorização confirma a existência de uma correlação genuína entre os descritores utilizados e a atividade biológica na enzima. Finalmente, o modelo de melhor performance, o NB – TcDHODH – 8, teve a sua capacidade preditiva confirmada na validação externa utilizando a série teste. Os modelos desenvolvidos, apesar de não empregarem o método de regressão linear múltipla, pertencem ao grupo de funções de pontuação empíricas. A performance de funções empíricas é amplamente influenciada pelos descritores utilizados para representar o processo de complexação do ligante com o alvo, pela qualidade do banco de dados experimentais da etapa de desenvolvimento e pelos algoritmos utilizados para extrair as correlações entre os descritores e a atividade biológica na biomacromolécula (GUEDES; PEREIRA; DARDENNE, 2018). De fato, os diferentes parâmetros explorados para o cálculo da entalpia de interação influenciaram a capacidade preditiva dos modelos obtidos. O NB – TcDHODH – 8 foi desenvolvido utilizando a entalpia de interação calculada com o maior número de resíduos e a otimização da geometria de um maior número de átomos. Embora este resultado motive a exploração de complexos P:L com mais átomos, os cálculos se tornariam extremamente custosos computacionalmente e praticamente inviáveis com o equipamento disponível, razão pela qual esta exploração não foi avaliada. A ΔH_{int} empregada no desenvolvimento do NB – TcDHODH – 8 foi calculada com a constante dielétrica mais polar ($\epsilon = 25,0$) para a proteína e o complexo P:L. Segundo Li et al. (2013), sítios de ligação mais

próximos da superfície e com resíduos ionizados seriam melhor modelados por uma constante dielétrica mais elevada ($\epsilon = 20 - 30$).

A alta performance do modelo NB – TcDHODH – 8 nas diferentes etapas de validação sugere que o mesmo pode ser empregado para identificar inibidores potentes de TcDHODH no fluxo de triagem virtual. A busca por inibidores da TcDHODH, realizada através da ancoragem molecular com a função ASP e a função de reclassificação NB – TcDHODH – 8, resultou na seleção de oito fármacos e dois candidatos a fármaco potencialmente ativos na enzima. Estas moléculas possuem grande potencial de serem ativos na TcDHODH e, deste modo, poderão servir como base para o desenvolvimento de novos tratamentos para a doença de Chagas.

Ressalta-se que o presente trabalho apresenta como limitação o fato de ter sido desenvolvido a partir de uma única série de inibidores, os derivados de orotato, a qual pode não ser suficientemente representativa do espaço químico de todos os compostos que têm o potencial de atuar na TcDHODH. Neste contexto, seria ideal determinar o domínio de aplicabilidade (DA) do modelo NB – TcDHODH – 8, de modo a identificar o conjunto de moléculas para as quais as predições seriam mais confiáveis (AIN et al., 2015). Entretanto, a determinação do DA para métodos de ancoragem molecular é uma área menos estabelecida na literatura quando comparada, por exemplo, aos métodos de QSAR. O DA depende das propriedades dos complexos P:L e dos ligantes da série treino (AIN et al., 2015). Zilian e Sottriffer (2013) demonstraram que a performance de funções de pontuação desenvolvidas com o algoritmo floresta aleatória podia ser explicada pela sobreposição dos descritores na série teste com a série treino, em uma análise de componentes principais. Das, Krein e Breneman (2010), por sua vez, desenvolveram um método para determinar o DA com base na comparação do sítio catalítico dos complexos P:L da série treino com a série teste, denominado *Property-Encoded Shape Distribution*. Neste contexto, a capacidade preditiva do fluxo de triagem do presente trabalho poderá ser otimizada através da implementação de um método para avaliar o DA do modelo de reclassificação. Além disso, o DA do modelo poderá ser expandido através da inclusão dados de novas séries químicas, que sejam provenientes de novas publicações na literatura ou, idealmente, de uma triagem virtual prospectiva, seguida da avaliação experimental dos melhores compostos. No momento, devido a inexistência de inibidores de TcDHODH de outras séries químicas, ainda não foi possível realizar estas otimizações. No cenário em que o modelo NB – TcDHODH - 8

fosse aplicável somente a derivados do orotato, este, ainda, representaria ganhos significativos para a terapêutica de Chagas, visto que esta é a única série de compostos descritos com atividade na TcDHODH.

Neste trabalho, a seletividade para a DHODH do *T. cruzi* versus a de humanos não foi explorada, porém seria possível desenvolver modelos preditivos para identificar compostos que preferencialmente inibam a enzima do parasita, empregando, por exemplo, metodologia similar à de Rocha e Sant'Anna (2019). Os derivados de orotato descritos por Inaoka et al. (2016, 2017) substanciam a possibilidade de obter compostos seletivos para a TcDHODH. O composto 5u, por exemplo, apresenta seletividade de 37.210 vezes para a DHODH do *T. cruzi* ($K_i = 24$ nM e $893 \mu\text{M}$ respectivamente na TcDHODH e na hDHODH). Notavelmente, a TcDHODH apresenta diferenças estruturais quando comparada à hDHODH que podem ser exploradas para a identificação de inibidores seletivos (PINHEIRO; EMERY; NONATO, 2013). Embora diversos aminoácidos do sítio catalítico sejam iguais em ambas as enzimas, os resíduos equivalentes à Met69, Leu71, Cys130 e Ser195 da TcDHODH foram substituídos pela Tyr147, Phe149, Ser215 e Thr285 na hDHODH (INAOKA et al., 2017). No caso da seletividade dos derivados de orotato pela TcDHODH, o menor volume da Leu71 nesta enzima permite a acomodação dos substituintes volumosos na posição 5 do núcleo 2,6-dioxo-1,2,3,6-tetraidro-pirimidina, enquanto a Phe149 da hDHODH apresenta maior impedimento estérico à complexação destes ligantes (INAOKA et al., 2016). Vale ressaltar que os inibidores da hDHODH aprovados como fármacos ou em desenvolvimento clínico, como a leflunomida, a teriflunomida e o brequinar, se ligam ao sítio de ligação de ubiquinona da hDHODH (TIWARI; DUBEY, 2018). Diferentemente das enzimas de classe 2, que apresentam um sítio de ligação para o diidroorotato e outro para o aceptor de elétrons, as enzimas da classe 1A não possuem o sítio de ligação da ubiquinona e realizam ambas as reações de oxidação e redução no mesmo sítio catalítico (INAOKA et al., 2016). Por esta razão, estes fármacos e candidatos a fármaco que inibem a hDHODH não seriam capazes de inibir a TcDHODH de maneira similar.

A TcDHODH apresenta uma alça flexível próxima ao sítio catalítico, que compreende os resíduos Leu128 a Ala140 (CHELESKI et al., 2010). A conformação desta alça influencia o volume disponível para a ligação de compostos na enzima, que varia entre $178 \text{ \AA}^3$, quando a proteína está em uma posição fechada, a $694 \text{ \AA}^3$, na posição aberta (INAOKA et al., 2016). Na análise dos 45 co-cristais disponíveis no

PDB, observou-se uma alta flexibilidade nas conformações destes resíduos para os diferentes compostos. A maior diferença foi constatada para o cristal contendo o orotato (PDB 2E6A) quando comparado aos demais co-cristais. Isto ocorre, pois a complexação dos derivados do orotato mantêm a enzima em uma conformação mais aberta (YOSHINO et al., 2015). Notavelmente, as diferentes conformações da alça Leu128-Ala140, mesmo nas comparações somente entre os análogos com substituintes na posição 5 do núcleo 2,6-dioxo-1,2,3,6-tetraidro-pirimidina, sugere um encaixe induzido na TcDHODH. Deste modo, a capacidade preditiva do fluxo de triagem poderia ser aprimorada através da ancoragem molecular em diferentes cristais da TcDHODH que apresentem conformações distintas. Ainda, simulações de dinâmica molecular poderiam ser empregadas para determinar a conformação de menor energia dos complexos P:L propostos, antes da realização dos cálculos de entalpia de interação semi-empíricos e da reclassificação com o modelo NB – TcDHODH – 8.

Os modelos do presente trabalho poderão ser otimizados através da incorporação de novos descritores relevantes para o processo de complexação dos ligantes com a TcDHODH. Os fármacos e candidatos a fármaco selecionados como potencialmente ativos poderão ser avaliados experimentalmente para a comprovação da sua atividade. Notavelmente, a triagem virtual do presente trabalho teve como limitação o fato de ter calculado a entalpia de interação somente para os compostos com maior pontuação na ancoragem molecular com a função ASP. Por esta razão, novos fármacos e candidatos a fármaco poderão ser submetidos ao modelo de reclassificação NB – TcDHODH – 8 visando à identificação de novos compostos com o potencial de serem reposicionados como inibidores da TcDHODH. Os resultados da avaliação experimental permitirão, ainda, validar a capacidade preditiva do fluxo de triagem virtual. Novas bibliotecas de compostos poderão ser avaliadas utilizando este fluxo de triagem virtual validado, visando à identificação de novos inibidores da TcDHODH.

4.5 CONCLUSÕES

A doença de Chagas é uma enfermidade que afeta milhões de pessoas em todo mundo e carece de tratamentos seguros e eficazes. A enzima diidroorotato desidrogenase do *T. cruzi* (TcDHODH) tem sido evidenciada como um alvo

terapêutico promissor para o desenvolvimento de novos compostos para o combate ao parasita. Na estratégia SBDD deste trabalho, um fluxo de triagem virtual para a identificação de compostos ativos nesta enzima foi desenvolvido e validado. Este fluxo compreende a ancoragem molecular no alvo utilizando uma função de pontuação e a reclassificação dos resultados com um modelo de aprendizado de máquinas.

O estudo de reancoragem demonstrou que a função de pontuação ASP do programa GOLD apresenta alta capacidade de predição do modo ligação de inibidores da TcDHODH e alta performance de filtragem de compostos inativos. Por outro lado, esta mesma função possui capacidade preditiva baixa/intermediária na separação de compostos potentes na inibição da enzima comparativamente a inibidores com baixa atividade.

Os modelos de aprendizado de máquinas, desenvolvidos com o objetivo de reclassificar os resultados provenientes da triagem virtual com a função ASP com um modelo de energia livre de interação apresentaram alta capacidade preditiva. Os modelos foram baseados em parâmetros calculados, incluindo a entalpia de interação, a lipofilicidade do ligante (estimada pelo valor de $\log P$) e os efeitos entrópicos associados à perda de movimentos torcionais do ligante na complexação. Oito modelos foram explorados através de variações no método de cálculo da entalpia de interação pelo método semi-empírico PM7. O NB - TcDHODH – 8, modelo de melhor performance, foi obtido quando maior número de aminoácidos da proteína foi incluído nas simulações, a geometria de suas cadeias laterais foi otimizada junto com os ligantes e constante dielétrica mais polar foi empregada ($\epsilon = 25,0$). Este modelo apresentou AUC ROC = 0,81 e AUC BEDROC = 0,74 na validação cruzada quádrupla e AUC ROC = 0,79 e AUC BEDROC = 0,85 na validação externa.

O presente trabalho contribui com o desenvolvimento de novas terapias para a doença de Chagas através da validação de um fluxo para a seleção de compostos ativos na TcDHODH. Estes métodos têm aplicação no reposicionamento de fármacos e candidatos e fármacos para a terapêutica de Chagas através da identificação das moléculas com maior potencial de serem ativas no alvo. Além disso, os fármacos aprovados e candidatos a fármaco selecionados como potencialmente ativos poderão ser avaliados experimentalmente para comprovar o seu potencial de inibir a TcDHODH e, deste modo, serem reposicionados para a terapêutica de Chagas. Os métodos poderão, ainda, ser utilizados na triagem virtual de novas bibliotecas de interesse, em particular, no planejamento de novos análogos do orotato.

CAPÍTULO 5

5 CONCLUSÕES E PERSPECTIVAS

A doença de Chagas é uma enfermidade que afeta milhões de pessoas em todo mundo e carece de tratamentos seguros e eficazes. Neste trabalho, modelos computacionais foram desenvolvidos com o objetivo de auxiliar na identificação de fármacos, candidatos a fármaco e metabólitos de fármacos com potencial de serem reposicionados para a terapêutica da doença. A estratégia de reposicionamento foi explorada, pois tem se mostrado promissora em reduzir os investimentos em pesquisa, os riscos de falha e o tempo de desenvolvimento de novos medicamentos. Estas vantagens são particularmente interessantes para doenças tropicais negligenciadas, para as quais os investimentos em P&D são insuficientes. Os modelos foram desenvolvidos utilizando métodos baseadas na estrutura de ligantes ativos fenotipicamente na inibição do *T. cruzi* (abordagem LBDD) e métodos baseados na estrutura da diidroorotato desidrogenase do parasita (abordagem SBDD), explorada como um alvo terapêutico promissor para o combate ao parasita.

Na estratégia LBDD, modelos de aprendizado de máquina classificatórios foram desenvolvidos explorando combinações de algoritmos de aprendizado de máquina (*Naïve Bayes*, floresta aleatória, redes neurais probabilísticas e máquinas de vetores de suporte) com descritores estruturais bidimensionais (*fingerprints* moleculares). O estudo resultou na seleção do modelo NB – FeatMorgan como aquele de maior capacidade preditiva, gerado a partir da combinação do método de *Naïve Bayes* com o *fingerprint* FeatMorgan. Este modelo apresentou alta performance durante as validações iniciais, confirmadas na segunda validação externa com uma nova série teste (AUC ROC = 0,85, TCB = 80% e EF1 = 14,3). Notavelmente, este modelo apresentou capacidade preditiva superior à seleção de composto por similaridade estrutural, método empregado como comparador. A triagem virtual utilizando o NB – FeatMorgan levou à seleção de 269 fármacos aprovados, 291 candidatos a fármaco em desenvolvimento clínico e 206 metabólitos de fármacos, os quais são potenciais candidatos a reposicionamento para o tratamento da doença de Chagas. Dentre os 269 fármacos, 19 já haviam sido descritos na literatura como ativos contra o *T. cruzi* ($IC_{50} < 5 \mu M$ e SI > 10). Embora nenhum dos 17 fármacos avaliados experimentalmente tenha apresentado potência e seletividade dentro das especificações estabelecidas, alguns deles exibiram alta eficácia na inibição do parasita ($A_{max} > 90\%$). Os compostos ativos poderão ser explorados como pontos de

partida para novos projetos de P&D de fármacos anti-*T. cruzi* ou em terapias combinadas.

Na estratégia SBDD, métodos de ancoragem molecular e modelos de reclassificação foram explorados para a identificação de compostos ativos na TcDHODH, alvo terapêutico promissor para o tratamento da doença de Chagas. A função de pontuação ASP apresentou capacidade preditiva adequada para a identificação do modo de ligação de compostos na TcDHODH e para a filtragem inicial de moléculas completamente inativas. Por outro lado, visto o seu baixo desempenho na separação de moléculas potentes de outras com baixa atividade biológica, desenvolveu-se um modelo classificatório para refinar os resultados da triagem com a função ASP. O modelo foi desenvolvido através do algoritmo de aprendizado *Naive Bayes* e dos descritores da entalpia de interação entre o ligante e o alvo (calculada pelo método semi-empírico PM7), assim como da lipofilicidade do ligante e efeitos entrópicos associados à perda de movimentos torcionais do ligante após a complexação (estimados com o parâmetro E_{tor} da função de pontuação ChemPLP). A exploração de diferentes parâmetros nos cálculos de entalpia de interação pelo método semi-empírico permitiu a obtenção de modelos de alta performance. O NB – TcDHODH – 8, modelo de maior capacidade preditiva, apresentou AUC ROC = 0,79 e AUC BEDROC = 0,85 na validação externa com a série teste. Este modelo foi aplicado em conjunto com a ancoragem molecular utilizando a função ASP para a seleção de fármacos aprovados e candidatos a fármaco com o potencial de serem reposicionados para a terapêutica de Chagas através da inibição da TcDHODH. Os compostos selecionados poderão ser testados experimentalmente visando a confirmar a sua capacidade de inibirem a TcDHODH.

Os modelos LBDD e SBDD desenvolvidos e validados neste trabalho podem contribuir para o desenvolvimento de novos tratamentos seguros e eficazes para a doença de Chagas. Estes modelos poderão ser utilizados em novos triagens virtuais para selecionar moléculas potencialmente ativas contra o parasita a partir de bibliotecas de interesse. Em particular, a seleção preliminar de moléculas com maior potencial de serem ativas utilizando modelos preditivos tem grande impacto no caso de amostras difíceis de serem obtidas e para ensaios nos quais apenas um número limitado de compostos pode ser testado. Ressalta-se, ainda, que anualmente novos fármacos são aprovados e novos candidatos a fármacos progridem para a etapa de desenvolvimento clínico, aumentando assim o número de moléculas que poderiam ser

estudadas para o reposicionamento. Além disso, os compostos identificados como potencialmente ativos contra o *T. cruzi* e na TcDHODH poderão servir como base para novas iniciativas de triagem experimental visando ao reposicionamento de compostos para a terapêutica da doença de Chagas.

REFERÊNCIAS

- ACEVEDO, C. H.; SCOTTI, L.; SCOTTI, M. T. In silico studies designed to select sesquiterpene lactones with potential antichagasic activity from an in-house asteraceae database. **ChemMedChem.**, v. 13, n. 6, p. 634-645, 2018.
- ACHARYA, C. et al. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. **Curr. Comput. Aided Drug Des.**, v. 7, n. 1, p. 10-22, 2011.
- AFANTITIS, A. et al. Ligand-based virtual screening procedure for the prediction and the identification of novel β -amyloid aggregation inhibitors using kohonen maps and counterpropagation artificial neural networks. **Eur. J. Med. Chem.**, v. 46, n. 2, p. 497-508, 2011.
- AIN, Q. U. et al. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. **Wiley Interdiscip. Rev. Comput. Mol. Sci.**, v. 5, n. 6, p. 405-424, 2015.
- ALBERCA, L. N. et al. Discovery of novel polyamine analogs with anti-protozoal activity by computer guided drug repositioning. **J. Comput. Aided Mol. Des.**, v. 30, n. 4, p. 305-321, 2016.
- ALBERCA, L. N. et al. Cascade ligand- and structure-based virtual screening to identify new trypanocidal compounds inhibiting putrescine uptake. **Front. Cell Infect. Microbiol.**, v. 8, 173, 2018.
- ALONSO-PADILLA, J.; RODRIGUEZ, A. High throughput screening for anti-Trypanosoma cruzi drug discovery. **PLoS Negl. Trop. Dis.**, v. 8, n. 12, e3259, 2014.
- ALVES, V. M.; BRAGA, R. C.; MURATOV, E. N.; ANDRADE, C. H. Quimioinformática: uma introdução. **Quím. Nova**, v. 41, n. 2, p. 202-212, 2018.
- ANDRIANI, G. et al. Activity *in vivo* of anti-trypanosoma cruzi compounds selected from a high throughput screening. **PLoS Negl. Trop. Dis.**, v. 5, n. 8, e1298, 2011.
- ANNOURA, T. et al. The origin of dihydroorotate dehydrogenase genes of kinetoplastids, with special reference to their biological significance and adaptation to anaerobic, parasitic conditions. **J. Mol. Evol.**, v. 60, n. 1, p. 113-127, 2005.
- ASHBURN, T. T.; THOR, K. B. Drug repurposing: identifying and developing new uses for existing drugs. **Nat. Rev. Drug. Discov.**, v. 3, n. 8, p. 673-683, 2004.
- ARAUJO-LIMA, C. F. et al. Repurposing strategy of atorvastatin against trypanosoma cruzi: in vitro monotherapy and combined therapy with benznidazole exhibit synergistic trypanocidal activity. **Antimicrob. Agents Chemother.**, v. 62, n. 9, e00979-18, 2018.
- AVELAR, L. A. A. **Síntese de inibidores das enzimas cruzaina e diidroorotato desidrogenase de Trypanosoma cruzi.** 2014. 211 f. Dissertação (Mestrado em

Química Orgânica e Biológica) - Instituto de Química de São Carlos, Universidade de São Paulo, São Carlos, 2014.

BAJUSZ, B.; RÁCZ, A.; HÉBERGER, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? **J. Cheminform.**, v. 7, 20, 2015.

BAKER, J. An algorithm for the location of transition states. **J. Comp. Chem.**, v. 7, p. 385-395, 1986.

BELL, E. W.; ZHANG, Y. DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. **J. Cheminform.**, v. 11, 40, 2019.

BELLERA, C. L. et al. Computer-guided drug repurposing: Identification of trypanocidal activity of clofazimine, benidipine and saquinavir. **Eur. J. Med. Chem.**, v. 93, p. 338-348, 2015.

BENDER, A., GLEN, R. C. Molecular similarity: a key technique in molecular informatics. **Org. Biomol. Chem.**, v. 2, n. 22, p. 3204-3218, 2014.

BERMAN, H. M. et al. The Protein Data Bank. **Nucleic Acids Res.**, v. 28, n. 1, p. 235-242, 2000.

BERN, C. Chagas Disease. **N. Engl. J. Med.**, v. 373, n. 5, p. 456-466, 2015.

BERNATCHEZ, J. A. et al. High-throughput screening of the ReFRAME library identifies potential drug repurposing candidates for *Trypanosoma cruzi*. **Microorganisms**, v. 8, n. 4, 472, 2020.

BIBETTE, J. Gaining confidence in high-throughput screening. **Proc. Natl. Acad. Sci. USA**, v. 109, n. 3, p. 649-650, 2012.

BLEIZIFFER, P.; SCHALLER, K.; RINIKER, S. Machine learning of partial charges derived from high-quality quantum-mechanical calculations. **J. Chem. Inf. Model.**, v. 58, n. 3, p. 579-590, 2018.

BOSCHI, D. et al. Dihydroorotate dehydrogenase inhibitors in anti-infective drug research. **Eur. J. Med. Chem.**, v. 183, 111681, 2019.

BOSS, D. S. et al. Safety, tolerability, pharmacokinetics and pharmacodynamics of the oral cyclin-dependent kinase inhibitor AZD5438 when administered at intermittent and continuous dosing schedules in patients with advanced solid tumours. **Ann. Oncol.**, v. 21, n. 4, p. 884-894, 2010.

BOUGHORBEL, S.; JARRAY, F.; EL-AMBAR, M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. **PLoS One**, v. 2, n. 6, e0177678, 2017.

CARMODY, L. C. et al. Identification of small-molecule inhibitors of *Trypanosoma cruzi* infection. In: NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **Probe**

Reports from the NIH Molecular Libraries Program. Bethesda: National Center for Biotechnology Information, 2010. Disponível em: <<https://www.ncbi.nlm.nih.gov/books/NBK55071/>>. Acessado em: 25 out. 2020.

CAPUZZI, S. J. et al. Computer-aided discovery and characterization of novel ebola virus inhibitors. **J. Med. Chem.**, v. 61, n. 8, p. 3582-3594, 2018.

CASSIDY, J. W. Applications of machine learning in drug discovery I: target discovery and small molecule drug design, artificial intelligence in oncology drug discovery and development. In: CASSIDY, J. W.; TAYLOR, B. **Artificial intelligence in oncology drug discovery and development.** Londres: IntechOpen, 2020. cap. 2, 14 p. Disponível em: <<https://www.intechopen.com/books/artificial-intelligence-in-oncology-drug-discovery-and-development/applications-of-machine-learning-in-drug-discovery-i-target-discovery-and-small-molecule-drug-design>>. Acessado em: 16 fev. 2021.

CASTILLO-GARIT, J. A. et al. Ligand-based discovery of novel trypanosomicidal drug-like compounds: In silico identification and experimental support. **Eur. J. Med. Chem.**, v. 46, n. 8, p. 3324-3330, 2011.

CASTRO, J. A.; DE MECCA, M. M.; BARTEL, L. C. Toxic side effects of drugs used to treat Chagas' disease (American trypanosomiasis). **Hum. Exp. Toxicol.**, v. 25, n. 8, p. 471-479, 2006.

CAVASOTTO, C. N.; AUCAR, M. G. High-throughput docking using quantum mechanical scoring. **Front. Chem.**, v. 8, 246, 2020.

CERETO-MASSAGUÉ, A. et al. Molecular fingerprint similarity search in virtual screening. **Methods**, v. 71, p. 58-63, 2015.

CHAPUT, L.; MOUAWAD, L. Efficient conformational sampling and weak scoring in docking programs? Strategy of the wisdom of crowds. **J. Cheminform.**, v. 9, 37, 2017.

CHATELAIN, E. Chagas disease drug discovery: toward a new era. **J. Biomol. Screen.**, v. 20, n. 1, p. 22-35, 2015.

CHATELAIN, E. Chagas disease research and development: Is there light at the end of the tunnel? **Comput. Struct. Biotechnol. J.**, v. 15, p. 98-103, 2017.

CHATELAIN, E.; IOSET, J. R. Drug discovery and development for neglected diseases: the DNDi model. **Drug. Des. Devel. Ther.**, v. 5, p. 175-181, 2011.

CHATELAIN, E.; IOSET, J. R. Phenotypic screening approaches for Chagas disease drug discovery. **Expert. Opin. Drug Discov.**, v. 13, n. 2, p. 141-153, 2018.

CHELESKI, J. et al. Novel insights for dihydroorotate dehydrogenase class 1A inhibitors discovery. **Eur. J. Med. Chem.**, v. 45, n. 12, p. 5899-5909, 2010.

CHERKASOV, A. et al. QSAR modeling: where have you been? Where are you going to? **J. Med. Chem.**, v. 57, n. 12, p. 4977-5010, 2014.

CHIRAC, P.; TORREELE, E. Global framework on essential health R&D. **Lancet**, v. 367, n. 9522, p. 1560–1561, 2006.

CHONG, C. R.; SULLIVAN, D. J. New uses for old drugs. **Nature**, v. 448, n. 7154, p. 645–646, 2007.

COLODETTE, N. M. et al. Novel phosphatidylinositol 4-kinases III beta (PI4KIII β) inhibitors discovered by virtual screening using free energy models. **J. Comput. Aided Mol. Des.**, v. 34, n. 10, p. 1091–1103, 2020.

CORSELLO, S. M. et al. The Drug Repurposing Hub: a next-generation drug library and information resource. **Nat. Med.**, v. 23, n. 4, p. 405–408, 2017.

COUPEZ, B.; LEWIS, R. A. Docking and scoring--theoretically easy, practically impossible? **Curr. Med. Chem.**, v. 13, n. 25, p. 2995–3003, 2006.

DAS, S.; KREIN, M. P.; BRENNEMAN, C. M. Binding affinity prediction with property-encoded shape distribution signatures. **J. Chem. Inf. Model.**, v. 50, n. 2, p. 298–308, 2010.

DE RYCKER, M. et al. Identification of trypanocidal activity for known clinical compounds using a new *Trypanosoma cruzi* hit-discovery screening cascade. **PLoS Negl. Trop. Dis.**, v. 10, n. 4, e0004584, 2016.

DIAS, J. C. P. et al. Il consenso brasileiro em doença de Chagas, 2015. **Epidemiol. Serv. Saúde**, v. 25, p. 7–86, 2016. Número especial.

DIAS, J. V. L. et al. Conhecimentos sobre triatomíneos e sobre a doença de Chagas em localidades com diferentes níveis de infestação vetorial. **Cien. Saude Colet.**, v. 21, n. 7, p. 2293–2303, 2016.

DIMASI, J. A.; GRABOWSKI, H. G.; HANSEN, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. **J. Health. Econ.**, v. 47, p. 20–33, 2016.

DRUGS FOR NEGLECTED DISEASES INITIATIVE (DNDi). Chagas disease: Target product profile. Disponível em: <<https://www.dndi.org/diseases-projects/chagas/chagas-target-product-profile/>>. Acessado em: 25 out. 2020.

DRUGS FOR NEGLECTED DISEASES INITIATIVE (DNDi). R&D portfolio: December 2020. Disponível em: < <https://dndi.org/wp-content/uploads/2020/12/DNDi-RD-Portfolio-December-2020.pdf>>. Acessado em: 16 fev. 2021.

DOUGUET, D. Data sets representative of the structures and experimental properties of FDA-approved drugs. **ACS Med. Chem. Lett.**, v. 9, n. 3, p. 204–209, 2018.

DU, X. et al. Insights into protein-ligand interactions: mechanisms, models, and methods. **Int. J. Mol. Sci.**, v. 17, n. 2, 144, 2016.

DUAN, Y. et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. **J. Comput. Chem.**, v. 24, n. 16, p. 1999-2012, 2003.

DUFFY, S. et al. Screening the medicines for malaria venture Pathogen Box across multiple pathogens reclassifies starting points for open-source drug discovery. **Antimicrob. Agents Chemother.**, v. 61, n. 9, e00379-17, 2017.

EKINS, S. et al. A collaborative database and computational models for tuberculosis drug discovery. **Mol. Biosyst.**, v. 6, n. 5, p. 840-851, 2010.

EKINS, S. et al. In silico repositioning of approved drugs for rare and neglected diseases. **Drug Discov. Today**, v. 16, n. 7-8, p. 298-310, 2011.

EKINS, S.; FREUNDLICH, J. S. Validating new tuberculosis computational models with public whole cell screening aerobic activity datasets. **Pharm. Res.**, v. 28, n. 8, p. 1859–1869, 2011.

EKINS, S. et al. Bayesian models leveraging bioactivity and cytotoxicity information for drug discovery. **Chem. Biol.**, v. 20, n. 3, p. 370–378, 2013.

EKINS, S. et al. Machine learning models and pathway genome data base for *Trypanosoma cruzi* drug discovery. **PLoS Negl. Trop. Dis.**, v. 9, n. 6, e0003878, 2015.

EL-SAYED, N. M. et al. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. **Science**, v. 309, n. 5733, p. 409-415, 2005.

EMPEREUR-MOT, C. et al. Predictiveness curves in virtual screening. **J. Cheminform.**, v. 7, 52, 2015.

ENGEL, J. C. et al. Image-based high-throughput drug screening targeting the intracellular stage of *Trypanosoma cruzi*, the agent of Chagas' disease. **Antimicrob. Agents Chemother.**, v. 54, n. 8, p. 3326-3334, 2010.

FERREIRA, L. G. et al. Molecular docking and structure-based drug design strategies. **Molecules**, v. 20, n. 7, p. 13384-13421, 2015.

FERREIRA, L. G.; ANDRICOPULO, A. D. Drug repositioning approaches to parasitic diseases: a medicinal chemistry perspective. **Drug Discov. Today**, v. 21, n. 10, p. 1699-1710, 2016.

FERREIRA, L. G.; ANDRICOPULO, A. D. Chemoinformatics approaches to structure- and ligand-based drug design. **Front. Pharmacol.**, v. 9, 1416, 2018. Editorial.

FETRO, C.; SCHERMAN, D. Drug repurposing in rare diseases: Myths and reality. **Therapie**, v. 75, n. 2, p. 157-160, 2020.

FORLI, S. Charting a path to success in virtual screening. **Molecules**, v. 20, n. 10, p. 18732-18758, 2015.

FOURCHES, D.; MURATOV, E.; TROPSHA, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. **J. Chem. Inf. Model.**, v. 50, n. 7, p. 1189–1204, 2010.

FOURCHES, D.; MURATOV, E.; TROPSHA, A. Curation of chemogenomics data. **Nat. Chem. Biol.**, v. 11, n. 8, 535, 2015.

FOURCHES, D.; MURATOV, E.; TROPSHA, A. Trust, but verify II: a practical guide to chemogenomics data curation. **J. Chem. Inf. Model.**, v. 56, n. 7, p. 1243–1252, 2016.

FRANCO, C. H. **Descoberta de compostos com amplo espectro contra *Trypanosoma cruzi* por meio de triagem de alto conteúdo**. 2018. 221 f. Tese (Doutorado em Ciências) - Escola Paulista de Medicina, Universidade Federal de São Paulo (UNIFESP), São Paulo, 2018.

FRANCO, C. H. et al. Drug discovery for Chagas disease: impact of different host cell lines on assay performance and hit compound selection. **Trop. Med. Infect. Dis.**, v. 4, n. 2, 82, 2019.

FRIESNER, R. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. **J. Med. Chem.**, v. 47, n.7. p. 1739-1749, 2004.

GAO, K. et al. Are 2D fingerprints still valuable for drug discovery? **Phys. Chem. Chem. Phys.**, v. 22, n. 16, p. 8373-8390, 2020.

GARCÍA, M. C. et al. Clomipramine and benznidazole act synergistically and ameliorate the outcome of experimental Chagas disease. **Antimicrob. Agents Chemother.**, v. 60, n. 6, p. 3700–3708, 2016.

GARCIA-DOMENECH, R. et al. Predicting antiprotozoal activity of benzyl phenyl ether diamine derivatives through QSAR multi-target and molecular topology. **Mol. Divers.**, v. 19, n. 2, p. 357-366, 2015.

GARTNER, M. E. et al. Development and validation of a computational model ensemble for the early detection of BCRP/ABCG2 substrates during the drug design stage. **J. Chem. Inf. Model.**, v. 57, n. 8, p. 1868-1880, 2017.

GEPPERT, H.; VOGT, M.; BAJORATH, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. **J. Chem. Inf. Model.**, v. 50, n. 2, p. 205–216, 2010.

GILBERT, I. H. Inhibitors of dihydrofolate reductase in Leishmania and trypanosomes. **Biochim. Biophys. Acta**, v. 1587, n. 2-3, p. 249-257, 2002.

GILBERT, I. H. Drug discovery for neglected diseases: molecular target-based and phenotypic approaches. **J. Med. Chem.**, v. 56, n. 20, p. 7719-7726, 2013.

GILSON, M. K.; HONIG, B. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. **Proteins**, v. 4, n. 1, p. 7-18, 1988.

GLOBALDATA. Disponível em: <<https://pharma.globaldata.com/>>. Acessado em: 25 out. 2020.

GREENWOOD, J. R. et al. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. **J. Comput. Aided Mol. Des.**, v. 24, n. 6-7, p. 591-604, 2010.

GAULTON, A. et al. The ChEMBL database in 2017. **Nucleic Acids Res.**, v. 45, n. D1, p. D945-D954, 2017.

GUEDES, I. A.; PEREIRA, F. S. S.; DARDENNE, L. E. Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. **Front. Pharmacol.**, v. 9, 1089, 2018.

GUIMARÃES, C. R. W. As múltiplas contribuições para a complexação proteína-ligante: consequências em drug design. **Rev. Virtual Quim.**, v. 4, n. 4, p. 348-364, 2012.

HALGREN, T. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. **J. Med. Chem.**, v. 47, n. 7, p. 1750-1759, 2004.

HAMMOND, D. J.; GUTTERIDGE, W. E. UMP synthesis in the kinetoplastida. **Biochim. Biophys. Acta.**, v. 718, n. 1, p. 1-10, 1982.

HANSCH, C. et al. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. **Nature**, v. 194, p. 178-180, 1962.

HASHIMOTO, M. et al. Critical importance of the de novo pyrimidine biosynthesis pathway for *Trypanosoma cruzi* growth in the mammalian host cell cytoplasm. **Biochem. Biophys. Res. Commun.**, v. 417, n. 3, p. 1002-1006, 2012.

HEVENER, K. E. et al. Validation of molecular docking programs for virtual screening against dihydroorotate synthase. **J. Chem. Inf. Model.**, v. 49, n. 2, p. 444-460, 2009.

HUGHES, J. P. et al. Principles of early drug discovery. **Br. J. Pharmacol.**, v. 162, n. 6, p. 1239-1249, 2011.

INAOKA, D. K. et al. The open form inducer approach for structure-based drug design. **PLoS One**, v. 11, n. 11, e0167078, 2016.

INAOKA, D. K. et al. Design and synthesis of potent substrate-based inhibitors of the *Trypanosoma cruzi* dihydroorotate dehydrogenase. **Bioorg. Med. Chem.**, v. 25, n. 4, p. 1465-1470, 2017.

JIA, J. et al. Mechanisms of drug combinations: interaction and network perspectives. **Nat. Rev. Drug Discov.**, v. 8, n. 2, p. 111-128, 2009.

JIN, G.; WONG, S. T. C. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. **Drug Discov. Today**, v. 19, n. 5, p. 637-644, 2014.

JOHNSTON, K. L.; FORD, L.; TAYLOR, M. J. Overcoming the challenges of drug discovery for neglected tropical diseases: The A-WOL experience. **J. Biomol. Screen.**, v. 19, n. 3, p. 335-343, 2014.

JONES, G. et al. Development and validation of a genetic algorithm for flexible docking. **J. Mol. Biol.**, v. 267, n. 3, p. 727-48, 1997.

JORDAN, A. M. Artificial intelligence in drug design - the storm before the calm? **ACS Med. Chem. Lett.**, v. 9, n. 12, p. 1150–1152, 2018.

JUÁREZ-SALDIVAR, A. et al. Computational drug repositioning for Chagas disease using protein-ligand interaction profiling. **Int. J. Mol. Sci.**, v. 21, n. 12, 4270, 2020.

KAISER, M. et al. Repurposing of the open access Malaria Box for kinetoplastid diseases identifies novel active scaffolds against trypanosomatids. **J. Biomol. Screen.**, v. 20, n. 5, p. 634-645, 2015a.

KAISER, M. et al. Antiprotozoal activity profiling of approved drugs: a starting point toward drug repositioning. **PLoS One**. v. 10, n. 8, e0135556, 2015b.

KATSUNO, K. et al. Hit and lead criteria in drug discovery for infectious diseases of the developing world. **Nat. Rev. Drug Discov.**, v. 14, n. 11, p. 751-758, 2015.

KHARE, S. et al. Proteasome inhibition for treatment of leishmaniasis, Chagas disease and sleeping sickness. **Nature**, v. 537, n. 7619, p. 229-233, 2016.

KIGONDU, E. M. et al. Pharmacologically active metabolites, combination screening and target identification-driven drug repositioning in antituberculosis drug discovery. **Bioorg. Med. Chem.**, v. 22, n. 16, p. 4453-4461, 2014.

KLAMT, A.; SCHÜÜMANN, G.; COSMO, A. New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and its Gradient. **J. Chem. Soc., Perkin Trans. 2**, n. 5, p. 799-805, 1993.

KLUG, D. M.; GELB, M. H.; POLLASTRI, M. P. Repurposing strategies for tropical disease drug discovery. **Bioorg. Med. Chem. Lett.**, v. 26, n. 11, p. 2569–2576, 2016.

KORB, O.; STÜTZLE, T.; EXNER, T. E. Empirical scoring functions for advanced protein-ligand docking with PLANTS. **J. Chem. Inf. Model.**, v. 49, n.1, p. 84-96, 2009.

KRISTENSEN, T. G.; NIELSEN, J.; PEDERSEN, C. N. Methods for similarity-based virtual screening. **Comput. Struct. Biotechnol. J.**, v. 5, e201302009, 2013.

LÄTTI, S.; NIINIVEHMAS, S.; PENTIKÄINEN, O. T. Rocker: Open source, easy-to-use tool for AUC and enrichment calculations and ROC visualization. **J. Cheminform.**, v. 8, n. 1, 45, 2016.

LAUFKÖTTER, O.; MIYAO, T.; BAJORATH, J. Large-scale comparison of alternative similarity search strategies with varying chemical information contents. **ACS Omega**, v. 4, n. 12, p. 15304-15311, 2019.

LAVECCHIA, A. Machine-learning approaches in drug discovery: methods and applications. **Drug Discov. Today**, v. 20, n. 3, p. 318-331, 2015.

LEACH, A. R.; GILLET, V. J. Analysis of high-throughput screening data. In: ____ **An introduction to cheminformatics**. Netherlands: Springer, 2007. cap. 7, p. 141-158.

LEE, C. H.; HUANG, H. C.; JUAN, H. F. Reviewing ligand-based rational drug design: the search for an ATP synthase inhibitor. **Int. J. Mol. Sci.**, v. 12, n. 8, p. 5304-5318, 2011.

LEE, J. et al. Development of predictive models for identifying potential S100A9 inhibitors based on machine learning methods. **Front. Chem.**, v. 7, 779, 2019.

LI, H.; ROBERTSON, A. D.; JENSEN, J. H. Very fast empirical prediction and rationalization of protein pKa values. **Proteins**, v. 61, n. 4, p. 704-721, 2005.

LI, L. et al. On the dielectric "constant" of proteins: smooth dielectric function for macromolecular modeling and its implementation in DelPhi. **J. Chem. Theory Comput.**, v. 9, n. 4, p. 2126-2136, 2013.

LI, H. et al. Machine-learning scoring functions for structure-based drug lead optimization. **WIREs Comput. Mol. Sci.**, v. 10, e1465, 2020.

LIU, T. et al. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. **Nucleic Acids Res.**, v. 35, n. Database issue, p. D198-201, 2007.

LIU, Q.; ZHOU, X. N. Preventing the transmission of american trypanosomiasis and its spread into non-endemic countries. **Infect. Dis. Poverty**, v. 4, 60, 2015.

LIU, S. et al. Practical model selection for prospective virtual screening. **J. Chem. Inf. Model.** v. 59, n. 1, p. 282-293, 2019.

LO, Y. C. et al. Machine learning in chemoinformatics and drug discovery. **Drug Discov. Today**, v. 23, n. 8, p. 1538-1546, 2018.

LO, Y. O. et al. Artificial Intelligence-Based Drug Design and Discovery. In: STEFANIU, A.; RASUL, A.; HUSSAIN, G. **Cheminformatics and Its Applications**. Londres: IntechOpen, 2019. cap. 10, 17 p. Disponível em: <<https://www.intechopen.com/books/cheminformatics-and-its-applications>>. Acessado em: 25 out. 2020.

LOI, C. M.; SMITH, D. A.; DALVIE, D. Which metabolites circulate? **Drug Metab. Dispos.**, v. 41, n. 5, p. 933-951, 2013.

MAGALHAES, C. S.; BARBOSA, H. J. C.; DARDENNE, L. E. A genetic algorithm for the ligand-protein docking problem. **Genet. Mol. Biol.**, v. 27, n. 4, p. 605-610, 2004.

LOLLI, M. L. et al. Use of human Dihydroorotate Dehydrogenase (hDHODH) Inhibitors in Autoimmune Diseases and New Perspectives in Cancer Therapy. **Recent Pat. Anticancer Drug Discov.**, v. 13, n. 1, p. 86-105, 2018.

MAGGIORA, G. et al. Molecular similarity in medicinal chemistry. **J. Med. Chem.**, v. 57, n. 8, p. 3186-3204, 2014.

MAGRO, L.; MORETTI, U.; LEONE, R. Epidemiology and characteristics of adverse drug reactions caused by drug-drug interactions. **Expert. Opin. Drug Saf.**, v. 11, n. 1, p. 83-94, 2012.

MAHMOUDI, N. et al. Identification of new antimalarial drugs by linear discriminant analysis and topological virtual screening. **J. Antimicrob. Chemother.**, v. 57, p. 489-497, 2006.

MANNHOLD, R.; PODA, G. I.; OSTERMANN, C.; TETKO, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. **J. Pharm. Sci.**, v. 98, n. 3, p. 861-893, 2009.

MARRERO-PONCE, Y. et al. Ligand-based virtual screening and in silico design of new antimalarial compounds using nonstochastic and stochastic total and atom-type quadratic maps. **J. Chem. Inf. Model.**, v. 45, n. 4, p. 1082-1100, 2005.

MARRERO-PONCE, A. et al. Non-stochastic quadratic fingerprints and LDA-based QSAR models in hit and lead generation through virtual screening: theoretical and experimental assessment of a promising method for the discovery of new antimalarial compounds. **Eur. J. Med. Chem.**, v. 41, n. 4, p. 483-493, 2006.

MARTIN, M. B. et al. Bisphosphonates inhibit the growth of *Trypanosoma brucei*, *Trypanosoma cruzi*, *Leishmania donovani*, *Toxoplasma gondii*, and *Plasmodium falciparum*: a potential route to chemotherapy. **J. Med. Chem.**, v. 44, n. 6, p. 909-916, 2001.

MASETTI, M.; ROCCHIA, W. Molecular mechanics and dynamics: numerical tools to sample the configuration space. **Front. Biosci. (Landmark Ed)**, v. 19, p. 578-604, 2014.

MATHERS, C. D.; EZZATI, M.; LOPEZ, A. D. Measuring the burden of neglected tropical diseases: the global burden of disease framework. **PLoS Negl. Trop. Dis.**, v. 1, n. 2, e114, 2007.

MATSUDA, N. M.; MILLER, S. M.; EVORAL, P. R. B. The chronic gastrointestinal manifestations of Chagas disease. **Clinics**, v. 64, n. 12, p. 1219-1224, 2009.

MAURI, A.; CONSONNI, V.; TODESCHINI, R. Molecular Descriptors. In: LESZCZYNSKI, J. **Handbook of Computational Chemistry**. Dordrecht: Springer, 2016.

MAZZETI, A. L. et al. Synergic effect of allopurinol in combination with nitroheterocyclic compounds against *Trypanosoma cruzi*. **Antimicrob. Agents Chemother.**, v. 63, n. 6, e02264-18, 2019.

MELGRAKI, G. et al. In silico exploration for identifying structure-activity relationship of MEK inhibition and oral bioavailability for isothiazole derivatives. **Chem. Biol. Drug Des.**, v. 76, n. 5, p. 397-406, 2010.

MELO-FILHO, C. C. et al. Discovery of new potent hits against intracellular *Trypanosoma cruzi* by QSAR-based virtual screening. **Eur. J. Med. Chem.**, v. 163, p. 649-659, 2019.

MERZ, K. M. et al. Letter from the editors. **J. Chem. Inf. Model.**, v. 55, n. 4, p. 719-720, 2015.

MITCHELL, J. B. Machine learning methods in chemoinformatics. **Wiley Interdiscip. Rev. Comput. Mol. Sci.**, v. 4, n. 5, p. 468-481, 2014.

MOFFAT, J. G. et al. Opportunities and challenges in phenotypic drug discovery: an industry perspective. **Nat. Rev. Drug Discov.**, v. 16, n. 8, p. 531-543, 2017.

MOOIJ, W. T.; VERDONK, M. L. General and targeted statistical potentials for protein-ligand interactions. **Proteins**, v. 61, n. 2, p. 272-287, 2005.

MORAES, C. B. et al. Nitroheterocyclic compounds are more efficacious than CYP51 inhibitors against *Trypanosoma cruzi*: implications for Chagas disease drug discovery and development. **Sci. Rep.**, v. 4, 4703, 2014.

MORENO, M. et al. *Trypanosoma cruzi* benznidazole susceptibility in vitro does not predict the therapeutic outcome of human Chagas disease. **Mem. Inst. Oswaldo Cruz**, v. 105, n. 7, p. 918-924, 2010.

MORGAN, H. L. The generation of a unique machine description for chemical structures - a technique developed at chemical abstracts service. **J. Chem. Doc.**, v. 5, n. 2, p. 107-113, 1965.

MUNIER-LEHMAN, H. et al. On dihydroorotate dehydrogenases and their inhibitors and uses. **J. Med. Chem.**, v. 56, n. 8, p. 3148-3167, 2013.

MYSINGER, M. M. et al., Directory of Useful Decoys, Enhanced (DUD-E): better ligands and decoys for better benchmarking. **J. Med. Chem.**, v. 55, n. 14, p. 6582-6594, 2012.

NCATS. NCATS Inxight: Drugs. Disponível em: <<https://drugs.ncats.io/>>. Acessado em: 25 out. 2020.

NEITZ, R. J. et al. Lead identification to clinical candidate selection: drugs for Chagas disease. **J. Biomol. Screen.**, v. 20, n. 1, p. 101-111, 2015.

NEVES, B. J. et al. Discovery of new anti-schistosomal hits by integration of QSAR-based virtual screening and high content screening. **J. Med. Chem.**, v. 59, n. 15, p. 7075-7088, 2016.

NEVES, B. J. et al. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. **Front. Pharmacol.**, v. 9, 1275, 2018.

NIWA, T. Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. **J. Chem. Inf. Comput. Sci.**, v. 43, n. 1, p. 113-119, 2003.

NIWA, T. Prediction of biological targets using probabilistic neural networks and atom-type descriptors. **J. Med. Chem.**, v. 47, n. 10, p. 2645-2650, 2004.

NUNES, M. C. P. et al. Chagas disease: an overview of clinical and epidemiological aspects. **J. Am. Coll. Cardiol.**, v. 62, n. 9, p. 767-776, 2013.

OBACH, R. S. Pharmacologically active drug metabolites: impact on drug discovery and pharmacotherapy. **Pharmacol. Rev.**, v. 65, n. 2, p. 578-640, 2013.

O'BOYLE, N. M. et al. Open Babel: An open chemical toolbox. **J. Cheminform.**, v. 3, 33, 2011.

O'BOYLE, N. M.; SAYLE, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. **J. Cheminform.**, v. 8, 36, 2016.

OECD. Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. In: _____ **OECD Series on Testing and Assessment**. Paris: OECD Publishing, 2007. n. 69, 154 p.

OLIVEIRA, F. G. et al. Molecular docking study and development of an empirical binding free energy model for phosphodiesterase 4 inhibitors. **Bioorg. Med. Chem.**, v. 14, n. 17, p. 6001-6011, 2006.

ONYEKWELU, K. C. Life cycle of *Trypanosoma cruzi* in the invertebrate and the vertebrate hosts. In: DE SOUZA, W. **Biology of *Trypanosoma cruzi***. Londres: IntechOpen, 2019. cap. 2, 19 p. Disponível em: <<https://www.intechopen.com/books/biology-of-em-trypanosoma-cruzi-em-/life-cycle-of-em-trypanosoma-cruzi-em-in-the-invertebrate-and-the-vertebrate-hosts>>. Acessado em: 16 fev. 2021.

OPREA, T. I.; OVERINGTON, J. P. Computational and practical aspects of drug repositioning. **Assay Drug Dev. Technol.**, v. 13 n. 6, p. 299-306, 2015.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). Chagas disease (American trypanosomiasis). Disponível em: <<http://www.who.int/chagas/epidemiology/en/>>. Acessado em: 25 out. 2020.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). Ending the neglect to attain the sustainable development goals: A road map for neglected tropical diseases 2021–2030. **Report**. Genebra, 2021. Disponível em: <<https://apps.who.int/iris/bitstream/handle/10665/332094/WHO-UCN-NTD-2020.01-eng.pdf>>. Acessado em: 16 fev. 2021.

OU-YANG, S. S. et al. Computational drug discovery. **Acta Pharmacol. Sin.**, v. 33, n. 9, p. 1131-1140, 2012.

PANTSAR, T.; POSO, A. Binding affinity via docking: fact and fiction. **Molecules**, v. 23, n. 8, 1899, 2018

PALMER, M. Phenotypic screening. In: CZECHTIZKY, W.; HAMLEY, P. **Small molecule medicinal chemistry: strategies and technologies**. New Jersey: John Wiley & Sons, 2016, cap. 10, p. 281-304.

PEÑA, I. et al. New compound sets identified from high throughput phenotypic screening against three kinetoplastid parasites: an open resource. **Sci. Rep.**, v. 5, 8771, 2015.

PEREIRA, P. C. M.; NAVARRO, E. C. Challenges and perspectives of Chagas disease: a review. **J. Venom Anim. Toxins Incl. Trop. Dis.**, v. 19, n. 1, 34, 2013.

PERIWAL, V. et al. Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. **BMC Res. Notes**, v. 4, 504, 2011.

PINHEIRO, M. P.; EMERY, F. S.; NONATO, M. C. Target sites for the design of anti-trypanosomatid drugs based on the structure of dihydroorotate dehydrogenase. **Curr. Pharm. Des.**, v. 19, n. 14, p. 2615-2627, 2013.

PLANER, J. D. et al. Synergy testing of FDA-approved drugs identifies potent drug combinations against *Trypanosoma cruzi*. **PLoS Negl. Trop. Dis.**, v. 8, n.7, e2977, 2014.

PUSHPAKOM, S. et al. Drug repurposing: progress, challenges and recommendations. **Nat. Rev. Drug Discov.**, v. 18, n. 1, p. 41-58, 2019.

RDKit. RDKit: Open-source cheminformatics. Disponível em: <<http://www.rdkit.org>>. Acessado em: 25 out. 2020.

REAUME, A. G. Drug repurposing through nonhypothesis driven phenotypic screening. **Drug Disc. Today: Therap. Strateg.**, v. 8, n. 3-4, p. 85-88, 2011.

REIS, R. A. G. et al. The dihydroorotate dehydrogenases: Past and present. **Arch. Biochem. Biophys.**, v. 632, p. 175-191, 2017.

RILEY, J. et al., Development of a fluorescence-based *Trypanosoma cruzi* CYP51 inhibition assay for effective compound triaging in drug discovery programmes for Chagas disease. **PLoS Negl. Trop. Dis.**, v. 9, n. 9, e0004014, 2015.

RINIKER, S.; LANDRUM, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. **J. Cheminform.**, v. 5, n. 1, 26, 2013.

ROCHA, S. F. L. S.; SANT'ANNA, C. M. R. A procedure combining molecular docking and semiempirical method PM7 for identification of selective Shp2 inhibitors. **Biopolymers**, v. 110, n. 11, e23320, 2019.

ROGERS, D.; HAHN, M. Extended-connectivity fingerprints. **J. Chem. Inf. Model.**, v. 50, n. 5, p. 742–754, 2010.

ROOS, K. et al. OPLS3e: extending force field coverage for drug-like small molecules. **J. Chem. Theory Comput.**, v. 15, n. 3, p. 1863-1874, 2019.

ROQUERO, I. et al. Novel chemical starting points for drug discovery in leishmaniasis and Chagas disease. **Int. J. Parasitol. Drugs Drug Resist.**, v. 10, p. 58-68, 2019.

RUDRAPAL, M.; KHAIRNAR, S. J.; JADHAV, A. G. Drug repurposing (dr): an emerging approach in drug discovery. In: BADRIA, F. A. **Drug repurposing - hypothesis, molecular aspects and therapeutic applications**. Londres: IntechOpen, 2020. cap. 1, 20 p. Disponível em: <<https://www.intechopen.com/books/drug-repurposing-hypothesis-molecular-aspects-and-therapeutic-applications/drug-repurposing-dr-an-emerging-approach-in-drug-discovery>>. Acessado em: 16 fev. 2021.

RYDE, U.; SÖDERHJELM, P. Ligand-binding affinity estimates supported by quantum-mechanical methods. **Chem. Rev.**, v. 116, n. 9, p. 5520-5266, 2016.

SALENTIN, S. et al. PLIP: fully automated protein-ligand interaction profiler. **Nucleic Acids Res.**, v. 43, n. W1, p. W443-W447, 2015.

SALMASO, V.; MORO, S. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: an overview. **Front. Pharmacol.**, v. 9, 923, 2018.

SANGENIS, L. H. et al. Chagas disease transmission by consumption of game meat: systematic review. **Rev Bras Epidemiol.**, v. 19, n. 4, p. 803-811, 2016.

SANT'ANNA, C. M. R. Métodos de modelagem molecular para estudo e planejamento de compostos bioativos: uma introdução. **Rev. Virtual Quím.**, 2009, v. 1, n. 1, p. 49-57, 2009.

SASTRY, G. M. et al. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. **J. Comput. Aided Mol. Des.**, v. 27, n. 3, p. 221-234, 2013.

SCHEEDER, C.; HEIGWER, F.; BOUTROS, M. Machine learning and image-based profiling in drug discovery. **Curr. Opin. Syst. Biol.**, v. 10, p. 43–52, 2018.

SCHMIDT, T. J. et al. In silico prediction and experimental evaluation of furanoheliangolide sesquiterpene lactones as potent agents against *Trypanosoma brucei rhodesiense*. **Antimicrob. Agents Chemother.**, v. 58, n. 1, p.325-332, 2014.

SCIOR, T. et al. Recognizing pitfalls in virtual screening: a critical review. **J. Chem. Inf. Model.**, v. 52, n. 4, p. 867-881, 2012.

SELASSIE, C. D.; MEKAPATI, S. B.; VERMA, R. P. QSAR: then and now. **Curr. Top. Med. Chem.**, v. 2, n. 12, p. 1357-1379, 2002.

SHELLEY, J. C. et al. Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. **J. Comput. Aided, Mol. Des.**, v. 21, n. 12, p. 681-691, 2007.

SHEN, C. et al. Can machine learning consistently improve the scoring power of classical scoring functions? Insights into the role of machine learning in scoring functions. **Brief. Bioinform.**, 2020. [Epub ahead of print]. doi:10.1093/bib/bbz173.

SILVA, A. C. C. et al. Chagas disease treatment and rational drug discovery: a challenge that remains. **Front. Pharmacol.**, v.10, 873, 2019.

SPECK-PLANCHE, A.; CORDEIRO, M. N. D. S. Simultaneous virtual prediction of anti-*Escherichia coli* activities and ADMET profiles: A chemoinformatics complementary approach for high-throughput screening. **ACS Comb. Sci.**, v. 16, n. 2, p. 78-84, 2014.

STEINBECK, C. C. et al. The chemistry development kit (CDK): an open-source java library for chemo- and bioinformatics. **J. Chem. Inf. Comput. Sci.**, v. 43, n. 2, p. 493–500, 2003.

STEWART, J. J. P. Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. **Int. J. Quantum Chem.**, v. 58, n. 2, p. 133–146, 1996.

STEWART, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. **J. Mol. Model.**, v. 19, n. 1, p. 1-32, 2013.

STOKES, J. M. et al. A deep learning approach to antibiotic discovery. **Cell**, v. 180, n. 4, p. 688-702, 2020.

STROVEL, J. et al. Early drug discovery and development guidelines: for academic researchers, collaborators, and start-up companies. In: MARKOSSIAN, S. et al., editors. **Assay guidance manual**. Bethesda: Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2016. cap. 1, 35 p. Disponível em: <<https://www.ncbi.nlm.nih.gov/books/NBK92015/>>. Acessado em: 16 fev. 2021.

STRAUSS, M. et al. In vitro and in vivo drug combination for the treatment of *Trypanosoma cruzi* infection: a multivariate approach. **Exp. Parasitol.**, v. 189, p. 19–27, 2018.

STUMPFE, D.; BAJORATH, J. Current trends, overlooked issues, and unmet challenges in virtual screening. **J. Chem. Inf. Model.**, v. 60, n. 9, p. 4112–4115, 2020.

SUN, W.; SANDERSON, P. E.; ZHENG, W. Drug combination therapy increases successful drug repositioning. **Drug Discov. Today**, v. 21, n. 7, p. 1189-1195, 2016.

SVETNIK, V. et al. Random forest: a classification and regression tool for Compound classification and QSAR modeling. **J. Chem. Inf. Comput. Sci.**, v. 43, n. 6, p. 1947–1958, 2003.

SWINNEY, D. C.; ANTHONY, J. How were new medicines discovered? **Nat. Rev. Drug Discov.**, v. 10, n. 7, p. 507-519, 2011.

SYKES, M. L.; AVERY, V. M. Development and application of a sensitive, phenotypic, high-throughput image-based assay to identify compound activity against *Trypanosoma cruzi* amastigotes. **Int. J. Parasitol. Drugs Drug Resist.**, v. 5, n. 3, p. 215-228, 2015.

TALEVI, A.; BELLERA, C. L. Challenges and opportunities with drug repurposing: finding strategies to find alternative uses of therapeutics. **Expert Opin. Drug Discov.**, v. 15, n. 4, p. 397-401, 2020.

TAKASHIMA, E. et al. Characterization of the dihydroorotate dehydrogenase as a soluble fumarate reductase in *Trypanosoma cruzi*. **Mol. Biochem. Parasitol.**, v. 122, n. 2, p. 189-200, 2002.

TEN BRINK, T.; EXNER, T. E. Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results. **J. Chem. Inf. Model.**, v. 49, n. 6, p. 1535-1546, 2009.

TIWARI, K.; DUBEY, V. K. Fresh insights into the pyrimidine metabolism in the trypanosomatids. **Parasit. Vectors**, v. 11, n. 1, 87, 2018.

TODD, A.; ANDERSON, R.; GROUNDWATER, P.W. Rational drug design — identifying and characterizing a target. **Pharm. J.**, v. 283, n. 7559, p.19-20 2009.

TODESCHINI, R.; CONSONNI, V. **Handbook of molecular descriptors**. 1st ed. Weinheim: WILEY-VCH Verlag Gmbh, 2000.

TORRICO, F. et al. E1224 Study Group. Treatment of adult chronic indeterminate Chagas disease with benznidazole and three E1224 dosing regimens: a proof-of-concept, randomised, placebo-controlled trial. **Lancet Infect. Dis.**, v. 18, n. 4, p. 419-430, 2018.

TROPSHA, A. Best practices for QSAR model development, validation, and exploitation. **Mol. Inform.**, v. 29, n. 6-7, p. 476-488, 2010.

TROPSHA, A.; GRAMATICA, P.; GOMBAR, V. K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. **QSAR Comb. Sci.**, v. 22, n. 1, p. 69-77, 2003.

TROUILLER, P. et al. Drug development for neglected diseases: a deficient market and a public-health policy failure. **Lancet**, v. 359, n. 9324, p. 2188-2194, 2002.

TRUCHON, J. F.; BAYLY, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. **J. Chem. Inf. Model.**, v. 47, n. 2, p. 488-508, 2007.

TURNHEIM, K.; KRIVANEK, P.; OBERBAUER, R. Pharmacokinetics and pharmacodynamics of allopurinol in elderly and young subjects. **Br. J. Clin. Pharmacol.**, v. 48, n. 4, p. 501-509, 1999.

VAMATHEVAN, J. et al. Applications of machine learning in drug discovery and development. **Nat. Rev. Drug Discov.**, v. 18, n. 6, p. 463-477, 2019.

VAN MONTFORT, R. L. M.; WORKMAN, P. Structure-based drug design: aiming for a perfect fit. **Essays Biochem.**, v. 61, n. 5, p. 431-437, 2017.

VANOMMESLAEGHE, K.; GUVENCH, O.; MACKERELL, A. D. J. Molecular mechanics. **Curr. Pharm. Des.**, v. 20, n. 20, p. 3281-3292, 2014.

VENKATRAJ, M. et al. From human immunodeficiency virus non-nucleoside reverse transcriptase inhibitors to potent and selective antitrypanosomal compounds. **Bioorg. Med. Chem.**, v. 22, n. 19, p. 5241-5248, 2014.

VERDONK, M. L. et al. Improved protein-ligand docking using GOLD. **Proteins**, v. 52, n. 4, p. 609-623, 2003.

VILLALTA, F.; RACHAKONDA, G. Advances in preclinical approaches to Chagas disease drug discovery. **Expert. Opin. Drug Discov.**, v. 14, n. 11, p. 1161-1174, 2019.

WALTERS, W. P.; MURCKO, M. Assessing the impact of generative ai on medicinal chemistry. **Nat. Biotechnol.**, v. 38, n. 2, p. 143-145, 2020.

WANG, S. et al. Protein kinase C. Modeling of the binding site and prediction of binding constants. **J. Med. Chem.**, v. 37, n. 9, p. 1326-1338, 1994.

WANG, Y. et al. PubChem's BioAssay Database. **Nucleic Acids Res.**, v. 40, n. Database issue, p. D400-D412, 2012.

WANG, Z. et al. Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. **Phys. Chem. Chem. Phys.**, v. 18, n. 18, p. 12964-12975, 2016.

WANG, J. C.; LIN, J. H. Scoring functions for prediction of protein-ligand interactions. **Curr. Pharm. Des.**, v. 19, n. 12, p. 2174-2182, 2013.

WANG, C.; ZHANG, Y. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. **J. Comput. Chem.**, v. 38, n. 3, p. 169-177, 2017.

WARING, M. J. et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. **Nat. Rev. Drug Discov.**, v. 14, n. 7, p. 475–486, 2015.

WARREN, G. L. et al. A critical assessment of docking programs and scoring functions. **J. Med. Chem.**, v. 49, n. 20, p. 5912-5931, 2006.

WASZKOWYCZ, B.; CLARK, D. E.; GANCIA, E. Outstanding challenges in protein–ligand docking and structure-based virtual screening. **Wiley Interdiscip. Rev. Comput. Mol. Sci.**, v. 1, n. 2, p. 229-259, 2011.

WILLETT, P. Similarity-based virtual screening using 2D fingerprints. **Drug Discov. Today**, v. 11, n. 23-24, p.1046-1053, 2006.

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

WÓJCIKOWSKI, M.; BALLESTER, P. J.; SIEDLECKI, P. Performance of machine-learning scoring functions in structure-based virtual screening. **Sci. Rep.**, v. 7, 46710, 2017.

WOUTERS, O. J.; MCKEE, M.; LUYTEN, J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. **JAMA**, v. 323, n. 9, p. 844–853, 2020.

YANG, G. et al. Evaluation of parameters impacting drug susceptibility in intracellular *Trypanosoma cruzi* assay protocols. **SLAS Discov.**, v. 22, n. 2, p. 125-134, 2017.

YANG, X. et al. Concepts of artificial intelligence for computer-assisted drug discovery. **Chem. Rev.**, v. 119, n. 18, p. 10520-10594, 2019.

YILMAZER, N. D.; KORTH, M. Comparison of molecular mechanics, semi-empirical quantum mechanical, and density functional theory methods for scoring protein-ligand interactions. **J. Phys. Chem. B**, v. 117, n. 27, p. 8075-8084, 2013.

YOSHINO, R. et al. Pharmacophore modeling for anti-Chagas drug design using the fragment molecular orbital method. **PLoS One**, v. 10, n. 5, e0125829, 2015.

ZAKAHROV, A. V. et al. QSAR modeling of imbalanced high-throughput screening data in PubChem. **J. Chem. Inf. Model.**, v. 54, n. 3, p. 705-712, 2014.

ZHANG, S. et al. A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. **J. Chem. Inf. Model.**, v. 46, n. 5, p. 1984-1995, 2006.

ZHANG, L. et al. Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. **J. Chem. Inf. Model.**, v. 53, n. 2, p. 475-492, 2012.

ZHU, J. et al. Integrating machine learning-based virtual screening with multiple protein structures and bio-assay evaluation for discovery of novel GSK3 β inhibitors. **Front. Pharmacol.**, v. 11, 566058, 2020.

ZILIAN, D.; SOTRIFER, C. A. SFCscore(RF): a random forest-based scoring function for improved affinity prediction of protein-ligand complexes. **J. Chem. Inf. Model.**, v. 53, n. 8, p. 1923-1933, 2013.

APÊNDICE A - COMPOSIÇÃO DOS BANCOS DE DADOS UTILIZADOS NO DESENVOLVIMENTO E VALIDAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA DA ABORDAGEM LBDD

Série treino: TCMDC-143349, TCMDC-143245, TCMDC-143201, TCMDC-143373, TCMDC-143184, TCMDC-143220, TCMDC-143425, TCMDC-143255, TCMDC-143424, TCMDC-143566, TCMDC-143345, TCMDC-143594, TCMDC-143239, TCMDC-143231, TCMDC-143404, TCMDC-143445, TCMDC-143306, TCMDC-143219, TCMDC-143340, TCMDC-143382, TCMDC-143556, TCMDC-142900, TCMDC-143233, TCMDC-143586, TCMDC-143289, TCMDC-143292, TCMDC-143158, TCMDC-143287, TCMDC-143237, TCMDC-143499, TCMDC-143290, TCMDC-143505, TCMDC-143251, TCMDC-143578, TCMDC-143472, TCMDC-143522, TCMDC-143571, TCMDC-143196, TCMDC-143257, TCMDC-143297, TCMDC-143543, TCMDC-143285, TCMDC-143641, TCMDC-143510, TCMDC-143517, TCMDC-143513, TCMDC-143136, TCMDC-143296, TCMDC-143420, TCMDC-143633, TCMDC-143603, TCMDC-143514, TCMDC-143609, TCMDC-143199, TCMDC-143180, TCMDC-125222, TCMDC-143577, TCMDC-143563, TCMDC-143500, TCMDC-143104, TCMDC-143521, TCMDC-143324, TCMDC-143187, TCMDC-143434, TCMDC-143440, TCMDC-143534, TCMDC-143163, TCMDC-143114, TCMDC-143309, TCMDC-143465, TCMDC-143336, TCMDC-143298, TCMDC-143553, TCMDC-143549, TCMDC-143599, TCMDC-143437, TCMDC-143614, TCMDC-143492, TCMDC-143405, TCMDC-143275, TCMDC-143083, TCMDC-143190, TCMDC-143156, TCMDC-143542, TCMDC-143464, TCMDC-143411, TCMDC-143610, TCMDC-143461, TCMDC-143481, TCMDC-143535, TCMDC-143332, TCMDC-143186, TCMDC-143479, TCMDC-143151, TCMDC-143161, TCMDC-143126, TCMDC-143604, TCMDC-143421, TCMDC-143616, TCMDC-143152, TCMDC-143504, TCMDC-143593, TCMDC-143191, TCMDC-143379, TCMDC-143529, TCMDC-143415, TCMDC-143315, TCMDC-143495, TCMDC-143474, TCMDC-143463, TCMDC-143159, TCMDC-143304, TCMDC-143105, TCMDC-143615, TCMDC-143224, TCMDC-143423, TCMDC-143300, TCMDC-143302, TCMDC-143207, TCMDC-143559, TCMDC-143409, TCMDC-143477, TCMDC-143262, TCMDC-143222, TCMDC-143142, TCMDC-143541, TCMDC-143286, TCMDC-143497, TCMDC-143087, TCMDC-143198, TCMDC-143149, TCMDC-143082, TCMDC-143606, TCMDC-143321, TCMDC-143502, TCMDC-143148, TCMDC-143253, TCMDC-143258, TCMDC-143200, TCMDC-143417, TCMDC-143273, TCMDC-143279, TCMDC-143197, TCMDC-143476, TCMDC-123621, TCMDC-143422, TCMDC-143209, TCMDC-143467, TCMDC-143608, TCMDC-143561, TCMDC-143381, TCMDC-143555, TCMDC-143221, TCMDC-143511, TCMDC-143153, TCMDC-143164, TCMDC-143507, TCMDC-143354, TCMDC-143120, TCMDC-143084, TCMDC-143626, TCMDC-143371, TCMDC-143546, TCMDC-143494, TCMDC-143284, TCMDC-143385, TCMDC-143548, TCMDC-143446, TCMDC-143194, TCMDC-143569, TCMDC-143484, TCMDC-143334, TCMDC-143256, TCMDC-143520, TCMDC-143293, TCMDC-143620, TCMDC-143389, TCMDC-143108, TCMDC-139489, TCMDC-143410, TCMDC-143310, TCMDC-143317, TCMDC-143433, TCMDC-143617, TCMDC-143625, TCMDC-143135, TCMDC-143179, TCMDC-143466, TCMDC-143328, TCMDC-143432, TCMDC-143590, TCMDC-143272, TCMDC-143282, TCMDC-143314, TCMDC-143362, TCMDC-143109, TCMDC-143308, TCMDC-143346, TCMDC-143526, TCMDC-143103, TCMDC-143545, TCMDC-143490, TCMDC-143612, TCMDC-143331, TCMDC-143455, TCMDC-143247, TCMDC-143601, TCMDC-143403, TCMDC-143426, TCMDC-140766, TCMDC-143552, TCMDC-143130, TCMDC-143598, TCMDC-143439, TCMDC-143322, TCMDC-143384, CID 14728590, CID 14730905, CID 14735926, CID 24836741, CID 14727998, CID 49819262, CID 14737651, CID 14735142, CID 4258175, CID 24798714, CID 14744787, CID 22406544, CID 14739070, CID 852730, CID 4255875, CID 14725345, CID 7966313, CID 24801480, CID 847106, CID 26665507, CID 24822978, CID 46500226, CID 24809265, CID 4245377, CID 24787513, CID 24799843, CID 14731130, CID 845816, CID 24793992, CID 14725228, CID 24812276, CID 24826493, CID 49648263, CID 24820652, CID 24799020, CID 24837598, CID 24782461, CID 24796141, CID 860862, CID 24785147, CID 850268, CID 851997, CID 49648451, CID 22404828, CID 7968671, CID 14736923, CID 24811280, CID 14737546, CID 852024, CID 24818538, CID 56319214, CID 4254580, CID 24784604, CID 17406742, CID 842442, CID 864446, CID 17387697, CID 4265136, CID 7976898, CID 22408677, CID 49722643, CID 17409651, CID 24813019, CID 4240825, CID 4257354, CID 17504621, CID 26659144, CID 26661939, CID 49676236, CID 47200360, CID 22413547, CID 24830132, CID 49815759, CID 49822343, CID 849692, CID 24787514, CID 4248798, CID 851575, CID 17385994, CID 22405573, CID 862504, CID 24817028, CID 22413272, CID 4245076, CID 24826066, CID 49669442, CID 24830697, CID 49645395, CID

26658911, CID 22407061, CID 49828415, CID 14729840, CID 846324, CID 858609, CID 26662016, CID 848988, CID 51085956, CID 3715445, CID 7972827, CID 49815767, CID 22411113, CID 49666892, CID 853665, CID 24788880, CID 24803483, CID 26730709, CID 24789325, CID 24825330, CID 49713794, CID 4257302, CID 49822151, CID 865921, CID 49641619, CID 51090105, CID 4250962, CID 22415301, CID 56320403, CID 22405586, CID 17407101, CID 17410394, CID 17413597, CID 17510152, CID 17512857, CID 49735994, CID 22401961, CID 26725544, CID 49646163, CID 17401547, CID 49827870, CID 22410856, CID 49722111, CID 47199424, CID 49721711, CID 24810547, CID 861183, CID 47197625, CID 17509954, CID 22405273, CID 24819308, CID 17509805, CID 24791934, CID 47200751, CID 17431437, CID 17505440, CID 49730673, CID 56321847, CID 49670391, CID 49821484, CID 858131, CID 26730467, CID 24803310, CID 17510991, CID 56323076, CID 24790812, CID 49722531, CID 49736128, CID 24800621, CID 848808, CID 24818423, CID 864165, CID 24791628, CID 17403078, CID 46500492, CID 16952834, CID 3714787, CID 51089858, CID 3713346, CID 26659712, CID 4255327, CID 47196742, CID 26671210, CID 24807156, CID 865074, CID 51090245, CID 14746151, CID 24789439, CID 49643351, CID 7967923, CID 22410673, CID 49646404, CID 866003, CID 11532946, CID 26660709, CID 26728243, CID 14722729, CID 17412339, CID 17412711, CID 47199633, CID 49644552, CID 852070, CID 17511809, CID 24783829, CID 56316993, CID 56315036, CID 7968751, CID 47203327, CID 24809086, CID 7978454, CID 49716516, CID 47197986, CID 49822024, CID 4255165, CID 17516589, CID 3716352, CID 26665919, CID 49731290, CID 56316517, CID 24803441, CID 47197744, CID 7967133, CID 24805639, CID 49678991, CID 17402251, CID 47197822, CID 56323609 e CID 26669538.

Série teste A: TCMDC-143131, TCMDC-143268, TCMDC-143536, TCMDC-143468, TCMDC-143496, TCMDC-143299, TCMDC-143531, TCMDC-143125, TCMDC-143074, TCMDC-143274, TCMDC-143551, TCMDC-143643, TCMDC-143646, TCMDC-143391, TCMDC-143524, TCMDC-143430, TCMDC-143602, TCMDC-143506, TCMDC-143637, TCMDC-143631, TCMDC-143325, TCMDC-143613, TCMDC-143319, TCMDC-143414, TCMDC-143203, TCMDC-143562, TCMDC-143329, TCMDC-143416, TCMDC-143387, TCMDC-143288, TCMDC-143530, TCMDC-143519, TCMDC-143081, TCMDC-143235, TCMDC-143071, TCMDC-143539, TCMDC-143301, TCMDC-143178, TCMDC-143182, TCMDC-143564, TCMDC-143413, TCMDC-143632, TCMDC-143550, TCMDC-143150, TCMDC-143498, TCMDC-143127, TCMDC-143137, TCMDC-143622, TCMDC-143623, TCMDC-143244, TCMDC-143143, TCMDC-143333, TCMDC-143193, TCMDC-143276, TCMDC-143537, CID 14745924, CID 22411353, CID 22406326, CID 4247826, CID 14734979, CID 14728366, CID 24823312, CID 24824031, CID 22408602, CID 24788876, CID 4242477, CID 7972279, CID 56314748, CID 22407023, CID 7973167, CID 47200652, CID 3714793, CID 24814613, CID 24826803, CID 24823607, CID 22406300, CID 26659628, CID 56314722, CID 49648864, CID 3716538, CID 7971916, CID 22403864, CID 24835129, CID 49727097, CID 26659760, CID 3712308, CID 49724362, CID 853675, CID 22411410, CID 24796847, CID 49737211, CID 24793142, CID 26670342, CID 26731067, CID 56322157, CID 49677171, CID 49644986, CID 24780349, CID 17432524, CID 17433347, CID 24819999, CID 24823180, CID 7974137, CID 22412775, CID 49671914, CID 49731703, CID 49681126, CID 3717871, CID 49718830 e CID 7975546.

Série teste B: MMV008416, MMV665820, MMV020885, MMV666080, MMV665909, MMV665994, MMV007384, MMV001230, MMV001246, MMV006861, MMV006319, MMV396663, MMV007907, MMV085203, MMV006172, MMV665914, MMV006825, MMV001241, MMV020243, MMV008160, MMV007764, CID 842457, CID 843368, CID 843427, CID 843434, CID 843528, CID 844993, CID 846485, CID 846687, CID 847631, CID 848078, CID 848302, CID 851545, CID 853491, CID 855625, CID 856106, CID 856237, CID 856309, CID 857655, CID 858019, CID 858592, CID 859110, CID 859672, CID 862166, CID 862771, CID 863774, CID 864273, CID 865151, CID 865940, CID 3712103, CID 3713958, CID 3715557, CID 3715929, CID 3716883, CID 3716999, CID 3717348, CID 4241711, CID 4244643, CID 4246746, CID 4246920, CID 4247714, CID 4247739, CID 4251637, CID 4251843, CID 4252186, CID 4256188, CID 4256229, CID 4256952, CID 4257192, CID 4258114, CID 4258194, CID 4258504, CID 4258572, CID 4259694, CID 4260675, CID 4260804, CID 4261706, CID 4261962, CID 4262303, CID 4263602, CID 4263753, CID 4264176, CID 4265052, CID 7965427, CID 7965474, CID 7969369, CID 7971961, CID 7972009, CID 7972417, CID 7972452, CID 7973704, CID 7975423, CID 7976335, CID 7976497, CID 7978052, CID 14718891, CID 14721237, CID 14723629, CID 14723872, CID 14724050, CID 14725354, CID 14725994, CID 14726473, CID 14726701, CID 14726890, CID 14727267, CID 14727853, CID 14727884, CID 14728016, CID 14729247, CID 14731053, CID 14733014, CID 14733486, CID 14733570, CID 14734820, CID 14735030, CID

14735245, CID 14735763, CID 14737058, CID 14737277, CID 14737819, CID 14738437, CID
 14741832, CID 14742216, CID 14742653, CID 14742945, CID 14745687, CID 14745770, CID
 14746120, CID 14746173, CID 14747072, CID 16952555, CID 16952751, CID 16953194, CID
 17385961, CID 17386832, CID 17401391, CID 17402099, CID 17402969, CID 17406702, CID
 17407207, CID 17408630, CID 17408871, CID 17409336, CID 17409480, CID 17409912, CID
 17410504, CID 17410723, CID 17411036, CID 17412916, CID 17415781, CID 17431481, CID
 17431615, CID 17431915, CID 17431930, CID 17433352, CID 17433371, CID 17433390, CID
 17433574, CID 17433995, CID 17506492, CID 17507297, CID 17507342, CID 17507346, CID
 17507717, CID 17507901, CID 17508089, CID 17508140, CID 17508275, CID 17508791, CID
 17509103, CID 17509326, CID 17509481, CID 17513960, CID 17517404, CID 22400948, CID
 22402443, CID 22403198, CID 22403612, CID 22405076, CID 22407588, CID 22407653, CID
 22407724, CID 22409590, CID 22410035, CID 22410831, CID 22411272, CID 22411968, CID
 22412926, CID 22413169, CID 22413295, CID 22414535, CID 22414613, CID 22415825, CID
 22416103, CID 24707945, CID 24708134, CID 24781821, CID 24782333, CID 24782704, CID
 24783617, CID 24784362, CID 24784782, CID 24785701, CID 24787303, CID 24788090, CID
 24788399, CID 24788708, CID 24788890, CID 24789703, CID 24790556, CID 24790577, CID
 24791140, CID 24792112, CID 24793621, CID 24794022, CID 24796171, CID 24796730, CID
 24796938, CID 24797799, CID 24798712, CID 24799311, CID 24799884, CID 24801204, CID
 24801758, CID 24803623, CID 24804108, CID 24804623, CID 24804652, CID 24805282, CID
 24806338, CID 24807535, CID 24807605, CID 24808569, CID 24809829, CID 24812714, CID
 24813154, CID 24815009, CID 24815413, CID 24815977, CID 24816784, CID 24816987, CID
 24817207, CID 24817682, CID 24817774, CID 24818278, CID 24818299, CID 24819418, CID
 24819768, CID 24819900, CID 24819914, CID 24820380, CID 24820454, CID 24821503, CID
 24821548, CID 24822393, CID 24822956, CID 24823485, CID 24824878, CID 24826003, CID
 24826050, CID 24826632, CID 24826846, CID 24827116, CID 24827248, CID 24827634, CID
 24827790, CID 24829543, CID 24830479, CID 24831232, CID 24833715, CID 24835284, CID
 24835543, CID 24836177, CID 24837782, CID 26657713, CID 26660386, CID 26662438, CID
 26662736, CID 26662817, CID 26665590, CID 26666681, CID 26669073, CID 26669624, CID
 26669966, CID 26670140, CID 26670460, CID 26671254, CID 26671450, CID 26724129, CID
 26724856, CID 26729526, CID 26730704, CID 26730909, CID 26730918, CID 26730958, CID
 26731065, CID 46500514, CID 47194415, CID 47195206, CID 47195355, CID 47195820, CID
 47195946, CID 47197676, CID 47197799, CID 47198000, CID 47199104, CID 47199492, CID
 47199746, CID 47201637, CID 47201873, CID 47202500, CID 47203267, CID 47203664, CID
 47204522, CID 47204673, CID 49640890, CID 49642230, CID 49644040, CID 49644077, CID
 49644084, CID 49644399, CID 49645534, CID 49646005, CID 49646121, CID 49646824, CID
 49648296, CID 49667269, CID 49668981, CID 49669053, CID 49670041, CID 49670869, CID
 49673867, CID 49673881, CID 49674150, CID 49674193, CID 49674968, CID 49676006, CID
 49676261, CID 49676946, CID 49678119, CID 49678958, CID 49679905, CID 49680314, CID
 49713889, CID 49714218, CID 49715151, CID 49715549, CID 49715595, CID 49715934, CID
 49716586, CID 49716945, CID 49717632, CID 49719500, CID 49720384, CID 49720734, CID
 49720762, CID 49723614, CID 49723698, CID 49723760, CID 49723944, CID 49723973, CID
 49724077, CID 49725397, CID 49727187, CID 49728790, CID 49729711, CID 49729974, CID
 49730675, CID 49730730, CID 49732909, CID 49733203, CID 49733650, CID 49733871, CID
 49734265, CID 49736536, CID 49736545, CID 49816255, CID 49816349, CID 49816507, CID
 49816596, CID 49816955, CID 49817722, CID 49818214, CID 49818370, CID 49818804, CID
 49819323, CID 49820156, CID 49820366, CID 49820691, CID 49821646, CID 49822083, CID
 49823805, CID 49824434, CID 49825185, CID 49825975, CID 49826090, CID 49826532, CID
 49827576, CID 49828574, CID 50085575, CID 50085810, CID 51086611, CID 51086885, CID
 51087071, CID 51087390, CID 51087399, CID 51087658, CID 51090238, CID 56315087, CID
 56315266, CID 56315437, CID 56315632, CID 56317403, CID 56317689, CID 56317886, CID
 56318624, CID 56319401, CID 56322012, CID 56322878 e CID 5632323.

APÊNDICE B – RESULTADOS DA VALIDAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA DA ABORDAGEM LBDD

Tabela 16 – Performance dos modelos de aprendizado de máquina na validação cruzada quíntupla

Método	Fingerprint	SE (%)	SP (%)	PPV (%)	NPV (%)	AC (%)	TCB (%)	AUC ROC	AUC BEDROC	EF1	EF10
NB	ECFP6	69,0	74,1	72,7	70,5	71,5	71,5	0,81	0,85	1,9	1,9
	FCFP6	78,2	74,1	75,1	77,3	76,2	76,2	0,84	0,87	1,9	1,9
	Extended	78,2	55,6	63,8	71,9	66,9	66,9	0,76	0,77	1,9	1,7
	Morgan	63,4	76,9	73,3	67,8	70,1	70,1	0,79	0,84	1,9	1,9
	FeatMorgan	72,7	83,3	81,3	75,3	78,0	78,0	0,88	0,92	1,9	1,9
PNN	ECFP6	8,8	100,0	100,0	52,3	54,4	54,4	0,82	0,86	1,9	1,9
	FCFP6	48,6	92,1	86,1	64,2	70,4	70,4	0,83	0,86	1,9	2,0
	Extended	92,6	43,5	62,1	85,5	68,1	68,1	0,79	0,82	1,9	1,9
	Morgan	3,7	100,0	100,0	50,9	51,9	51,9	0,85	0,89	1,9	1,9
	FeatMorgan	42,6	95,8	91,1	62,5	69,2	69,2	0,86	0,89	1,9	2,0
RF	ECFP6	61,1	77,3	72,9	66,5	69,2	69,2	0,77	0,82	1,9	1,9
	FCFP6	71,3	74,1	73,3	72,1	72,7	72,7	0,80	0,85	1,9	1,9
	Extended	79,2	75,9	76,7	78,5	77,5	77,5	0,86	0,88	1,9	2,0
	Morgan	64,4	79,2	75,5	69,0	71,8	71,8	0,81	0,85	1,9	2,0
	FeatMorgan	74,5	81,0	79,7	76,1	77,8	77,8	0,87	0,91	1,9	2,0
SVM	ECFP6	75,0	75,0	75,0	75,0	75,0	75,0	0,75	0,78	0,0	1,6
	FCFP6	77,3	81,9	81,1	78,3	79,6	79,6	0,80	0,82	0,0	1,6
	Extended	72,2	75,0	74,3	73,0	73,6	73,6	0,74	0,67	0,0	1,4
	Morgan	75,0	78,2	77,5	75,8	76,6	76,6	0,77	0,76	0,0	1,5
	FeatMorgan	79,2	80,6	80,3	79,5	79,9	79,9	0,80	0,78	0,0	1,6

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado e *fingerprints* estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). A capacidade preditiva na validação cruzada quíntupla foi avaliada utilizando os parâmetros: sensibilidade (SE), especificidade (SP), valor preditivo positivo e negativo (PPV e NPV), Acurácia (AC), taxa de classificação balanceada (TCB), área sob as curvas ROC e BEDROC (AUC ROC e AUC BEDROC, respectivamente), e os fatores de enriquecimento para 1% e 10% dos dados com maior classificação (EF1 e EF10, respectivamente). NB – *Naive Bayes*. RF - Floresta aleatória. PNN - Redes neurais probabilísticas. SVM - Máquinas de vetores de suporte.

Tabela 17 – Performance dos modelos de aprendizado de máquina no teste de aleatorização

Método	Fingerprint	SE (%)	SP (%)	PPV (%)	NPV (%)	AC (%)	TCB (%)	AUC ROC	AUC BEDROC	EF1	EF10
NB	ECFP6	47,7	46,8	47,2	47,2	47,2	47,2	0,46	0,45	0,5	0,7
	FCFP6	38,0	55,1	45,8	47,0	46,5	46,5	0,49	0,52	0,9	1,0
	Extended	14,8	78,7	41,0	48,0	46,8	46,8	0,49	0,55	0,9	1,2
	Morgan	38,4	59,3	48,5	49,0	48,8	48,8	0,47	0,47	1,4	0,9
	FeatMorgan	57,9	40,3	49,2	48,9	49,1	49,1	0,49	0,48	0,5	1,0
PNN	ECFP6	69,4	28,7	49,3	48,4	49,1	49,1	0,47	0,46	0,9	0,9
	FCFP6	69,0	30,6	49,8	49,6	49,8	49,8	0,49	0,50	0,9	1,0
	Extended	37,0	62,0	49,4	49,6	49,5	49,5	0,51	0,50	1,9	0,9
	Morgan	78,7	20,8	49,9	49,5	49,8	49,8	0,49	0,48	0,9	1,0
	FeatMorgan	38,0	65,3	52,2	51,3	51,6	51,6	0,53	0,53	0,9	1,0
RF	ECFP6	59,7	44,0	51,6	52,2	51,9	51,9	0,52	0,51	1,4	1,0
	FCFP6	51,9	50,0	50,9	50,9	50,9	50,9	0,53	0,56	1,9	1,2
	Extended	53,2	53,2	53,2	53,2	53,2	53,2	0,54	0,54	1,9	1,1
	Morgan	58,8	42,6	50,6	50,8	50,7	50,7	0,55	0,58	0,9	1,4
	FeatMorgan	58,8	48,6	53,4	54,1	53,7	53,7	0,53	0,53	1,4	1,1
SVM	ECFP6	43,1	60,6	52,2	51,6	51,9	51,9	0,52	0,54	0,5	1,0
	FCFP6	52,8	48,6	50,7	50,7	50,7	50,7	0,51	0,48	0,9	1,1
	Extended	53,7	48,6	51,1	51,2	51,2	51,2	0,51	0,51	0,9	1,1
	Morgan	42,6	54,2	48,2	48,5	48,4	48,4	0,48	0,43	0,5	0,9
	FeatMorgan	49,1	53,2	51,2	51,1	51,2	51,2	0,51	0,53	0,9	1,0

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado e fingerprints estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). A capacidade preditiva no teste de aleatorização foi avaliada utilizando os parâmetros: sensibilidade (SE), especificidade (SP), valor preditivo positivo e negativo (PPV e NPV), acurácia (AC), taxa de classificação balanceada (TCB), área sob as curvas ROC e BEDROC (AUC ROC e AUC BEDROC, respectivamente), e os fatores de enriquecimento para 1% e 10% dos dados com maior classificação (EF1 e EF10, respectivamente). NB – *Naive Bayes*. RF - Floresta aleatória. PNN - Redes neurais probabilísticas. SVM - Máquinas de vetores de suporte.

Tabela 18 – Performance dos modelos de aprendizado de máquina na validação externa com a Série teste A

Método	Fingerprint	SE (%)	SP (%)	PPV (%)	NPV (%)	TCB (%)	AUC ROC	AUC BEDROC	EF1	EF10	Cov (%)
NB	ECFP6	81,8	73,6	76,3	79,2	77,7	0,85	0,89	1,8	1,6	97
	FCFP6	80,0	65,4	71,0	75,6	72,7	0,80	0,82	1,8	1,6	97
	Extended	69,1	75,0	74,5	69,6	72,0	0,81	0,80	1,8	1,6	97
	Morgan	58,2	98,1	97,0	69,3	78,1	0,90	0,94	1,8	1,8	98
	FeatMorgan	78,2	86,8	86,0	79,3	82,5	0,89	0,92	1,8	1,8	98
PNN	ECFP6	7,3	100,0	100,0	50,5	53,6	0,91	0,94	1,8	1,8	97
	FCFP6	49,1	88,5	81,8	62,2	68,8	0,85	0,88	1,8	1,6	97
	Extended	96,4	53,8	68,8	93,3	75,1	0,85	0,86	1,8	1,8	97
	Morgan	3,6	100	100	50,0	51,8	0,94	0,97	1,8	1,8	98
	FeatMorgan	38,2	96,2	91,3	60,0	67,2	0,90	0,92	1,8	1,8	98
RF	ECFP6	63,6	90,6	87,5	70,1	77,1	0,88	0,90	1,8	1,8	97
	FCFP6	70,9	82,7	81,3	72,9	76,8	0,81	0,81	1,8	1,5	97
	Extended	78,2	78,8	79,6	77,4	78,5	0,89	0,88	1,8	1,8	97
	Morgan	70,9	84,9	83,0	73,8	77,9	0,89	0,94	1,8	1,8	98
	FeatMorgan	81,8	79,2	80,4	80,8	80,5	0,89	0,92	1,8	1,8	98
SVM	ECFP6	80,0	88,7	89,8	81,0	84,3	0,84	0,81	0,0	0,7	97
	FCFP6	80,0	78,8	80,0	78,8	79,4	0,79	0,80	0,0	0,0	97
	Extended	80,0	78,8	80,0	78,8	79,4	0,79	0,82	0,0	0,0	97
	Morgan	76,4	90,6	89,4	78,7	83,5	0,83	0,89	0,0	0,9	98
	FeatMorgan	80,0	81,1	81,5	79,6	80,6	0,81	0,81	0,0	0,0	98

Os modelos foram desenvolvidos a partir da Série treino utilizando diferentes combinações dos algoritmos de aprendizado e fingerprints estruturais (ECFP6, FCFP6, Extended, Morgan e FeatMorgan). A capacidade preditiva na validação externa com a Série teste A foi avaliada utilizando os parâmetros: sensibilidade (SE), especificidade (SP), valor preditivo positivo e negativo (PPV e NPV), taxa de classificação balanceada (TCB), área sob as curvas ROC e BEDROC (AUC ROC e AUC BEDROC, respectivamente), e os fatores de enriquecimento para 1% e 10% dos dados com maior classificação (EF1 e EF10, respectivamente). NB – *Naive Bayes*. RF - Floresta aleatória. PNN - Redes neurais probabilísticas. SVM - Máquinas de vetores de suporte.

**APÊNDICE C – LISTA DE FÁRMACOS APROVADOS SELECIONADOS PELO
MODELO NB – FEATMORGAN NA ABORDAGEM LDBB**

Tabela 19 – Fármacos aprovados classificados como ativos pelo modelo NB – FeatMorgan (continua)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
Isavuconazol	8,46	Oral	TCMDC-143082 (0,31)
Abemaciclibe	8,22	Oral	TCMDC-143477 (0,21)
Imatinibe	7,41	Oral	TCMDC-143433 (0,27)
Risperidona	6,78	Oral	TCMDC-143222 (0,25)
Nitroxolina	6,61	Oral	TCMDC-143633 (0,23)
Topiroxostate	5,86	Oral	TCMDC-143461 (0,25)
Vardenafila	5,47	Oral	TCMDC-143620 (0,25)
Voriconazol	5,20	Oral	TCMDC-143082 (0,39)
Crizotinibe	5,08	Oral	TCMDC-143409 (0,22)
Lamotrigina	4,77	Oral	TCMDC-143426 (0,23)
Nimorazol	4,67	Oral	TCMDC-143149 (0,21)
Fluconazol	4,49	Oral	TCMDC-143614 (0,45)
Nizatidina	4,47	Oral	TCMDC-143577 (0,20)
Ranitidina	4,47	Oral	TCMDC-143577 (0,20)
Vonoprazan	4,35	Oral	TCMDC-143190 (0,37)
Pirantel	4,35	Oral	TCMDC-143201 (0,22)
Cetoconazol	4,17	Oral	TCMDC-143590 (0,35)
Sertaconazol	4,16	Tópica	TCMDC-143590 (0,30)
Nitrofurantoina	4,13	Oral	TCMDC-143517 (0,31)
Ruxolitinibe	4,11	Oral	TCMDC-143314 (0,21)
Ganciclovir	4,09	Oral	TCMDC-143083 (0,25)
Econazol	4,07	Tópica	TCMDC-143423 (0,38)
Sulconazol	4,07	Tópica	TCMDC-143423 (0,38)
Bromexina	3,95	Oral	TCMDC-143297 (0,20)
Dantroleno	3,91	Oral	TCMDC-143233 (0,33)
Cetanserina	3,89	Oral	TCMDC-143222 (0,26)
Oxiconazol	3,86	Tópica	TCMDC-143590 (0,34)
Ribociclibe	3,84	Oral	TCMDC-143382 (0,22)
Aciclovir	3,76	Oral	TCMDC-143083 (0,22)
Canagliflozina	3,72	Oral	TCMDC-143220 (0,32)
Ácido zoledrônico	3,63	Intravenosa	TCMDC-125222 (0,34)
Palbociclibe	3,62	Oral	TCMDC-143382 (0,26)
Isoconazol	3,61	Tópica	TCMDC-143590 (0,38)
Cloroxina	3,59	Tópica	TCMDC-143180 (0,29)
Iodoquinol	3,59	Oral	TCMDC-143180 (0,29)
Dabrafenibe	3,56	Oral	TCMDC-143617 (0,24)
Luliconazol	3,52	Tópica	TCMDC-143590 (0,31)
Betrixabana	3,45	Oral	TCMDC-143156 (0,38)
Fasudila	3,41	Intravenosa	TCMDC-143302 (0,24)
Triantereno	3,31	Oral	TCMDC-143569 (0,22)
Miconazol	3,29	Tópica	TCMDC-143590 (0,37)
Ripasudila	3,24	Oftálmica	TCMDC-143382 (0,23)
Plerixafor	3,22	Subcutânea	TCMDC-143484 (0,27)
Flibanserina	3,17	Oral	TCMDC-143292 (0,24)
Mizolastina	3,09	Oral	TCMDC-143279 (0,27)
Penciclovir	3,06	Tópica	TCMDC-143304 (0,18)

Tabela 19 – Fármacos aprovados classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
Penciclovir	3,06	Tópica	TCMDC-143304 (0,18)
Tiabendazol	3,06	Oral	TCMDC-143292 (0,36)
Vismodegibe	3,03	Oral	TCMDC-143552 (0,32)
Morinidazol	3,01	Intravenosa	TCMDC-143149 (0,2)
Bromfeniramina	3,00	Oral	TCMDC-143222 (0,35)
Clorfenamina	3,00	Intravenosa	TCMDC-143222 (0,35)
Dexclorfeniramina	3,00	Oral	TCMDC-143222 (0,35)
Tinidazol	2,98	Oral	TCMDC-143578 (0,26)
Pramipexol	2,96	Oral	TCMDC-143606 (0,26)
Tioconazol	2,96	Tópica	TCMDC-143590 (0,34)
Sildenafil	2,94	Oral	TCMDC-143620 (0,28)
Enasidenibe	2,94	Oral	TCMDC-143476 (0,26)
Azatioprina	2,87	Oral	TCMDC-143209 (0,17)
Milrinona	2,87	Intravenosa	TCMDC-143620 (0,26)
Nicotinamida	2,86	Intravenosa	TCMDC-143504 (0,38)
Ornidazol	2,84	Oral	TCMDC-143578 (0,23)
Letrozol	2,84	Oral	TCMDC-143082 (0,26)
Timolol	2,78	Oral	TCMDC-143617 (0,21)
Talipexol	2,75	Oral	TCMDC-143606 (0,23)
Bifonazol	2,69	Tópica	TCMDC-143275 (0,25)
Terconazol	2,66	Tópica	TCMDC-143082 (0,36)
Adefovir	2,66	Oral	TCMDC-143336 (0,18)
Nifurtimox	2,64	Oral	TCMDC-143517 (0,3)
Leflunomida	2,61	Oral	TCMDC-143535 (0,35)
Regadenoson	2,6	Intravenosa	TCMDC-143612 (0,18)
Fenticonazol	2,51	Tópica	TCMDC-143590 (0,36)
Piribedil	2,5	Oral	TCMDC-143625 (0,32)
Olanzapina	2,49	Oral	TCMDC-143109 (0,23)
Arotinolol	2,45	Oral	TCMDC-143201 (0,22)
Domperidona	2,45	Oral	TCMDC-143135 (0,23)
Entecavir	2,42	Oral	TCMDC-143641 (0,19)
Zaleplon	2,42	Oral	TCMDC-143371 (0,24)
Azosemida	2,35	Oral	TCMDC-143219 (0,24)
Tiquizium	2,33	Oral	TCMDC-143201 (0,2)
Metronidazol	2,33	Oral	TCMDC-143578 (0,24)
Nitisinona	2,26	Oral	TCMDC-143609 (0,25)
Proxodolol	2,25	Oral	TCMDC-143424 (0,34)
Lanoconazol	2,23	Tópica	TCMDC-143423 (0,25)
Disopiramida	2,11	Oral	TCMDC-142900 (0,33)
Vareniclina	2,09	Oral	TCMDC-143641 (0,19)
Olprinona	2,06	Intravenosa	TCMDC-143087 (0,19)
Itraconazol	2,06	Oral	TCMDC-143082 (0,32)
Irsogladina	2,04	Oral	TCMDC-143297 (0,19)
Rilpivirina	1,99	Oral	TCMDC-143109 (0,23)
Vandetanibe	1,97	Oral	TCMDC-143603 (0,22)
Butoconazol	1,96	Tópica	TCMDC-143590 (0,42)
Lofepamina	1,91	Oral	TCMDC-143222 (0,25)
Cleboprida	1,9	Oral	TCMDC-143484 (0,33)
Olmutinibe	1,87	Oral	TCMDC-143545 (0,25)
Climbazol	1,86	Tópica	TCMDC-143590 (0,4)
Dalfampridina	1,85	Oral	TCMDC-143136 (0,24)

Tabela 19 – Fármacos aprovados classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
Sapropterina	1,84	Oral	TCMDC-143606 (0,18)
Meglumina	1,84	Intravenosa	TCMDC-143614 (0,12)
Bupropiona	1,81	Oral	TCMDC-143601 (0,25)
Fimasartana	1,8	Oral	TCMDC-143289 (0,21)
Acenocoumarol	1,78	Oral	TCMDC-143601 (0,25)
Desloratadina	1,77	Oral	TCMDC-143222 (0,29)
Flutrimazol	1,72	Tópica	TCMDC-143423 (0,32)
Duloxetina	1,72	Oral	TCMDC-143599 (0,26)
Fomepizol	1,71	Intravenosa	TCMDC-143440 (0,19)
Xipamida	1,7	Oral	TCMDC-143108 (0,24)
Secnidazol	1,7	Oral	TCMDC-143578 (0,24)
Abacavir	1,69	Oral	TCMDC-143606 (0,22)
Dasatinibe	1,68	Oral	TCMDC-143362 (0,25)
Ziprasidona	1,68	Oral	TCMDC-143362 (0,24)
Metirapona	1,66	Oral	TCMDC-143114 (0,29)
Fenoverina	1,64	Oral	TCMDC-143625 (0,25)
Ivacaftor	1,61	Oral	TCMDC-143490 (0,32)
Risedronato	1,61	Oral	TCMDC-143190 (0,32)
Brigatinibe	1,59	Oral	TCMDC-143109 (0,24)
Pixantrona	1,57	Intravenosa	TCMDC-143474 (0,21)
Sufentanila	1,56	Intravenosa	TCMDC-143222 (0,38)
Cloramfenicol	1,55	Tópica	TCMDC-143601 (0,3)
Clorzoxazona	1,53	Oral	TCMDC-139489 (0,19)
Oxiquinolina	1,53	Tópica	TCMDC-143633 (0,28)
Carbinoxamina	1,52	Oral	TCMDC-143542 (0,32)
Clorquinaldol	1,51	Tópica	TCMDC-143308 (0,28)
Efinaconazol	1,49	Tópica	TCMDC-143614 (0,4)
Betaistina	1,48	Oral	TCMDC-143599 (0,29)
Mirabegrona	1,44	Oral	TCMDC-143164 (0,22)
Anastrozol	1,44	Oral	TCMDC-143614 (0,23)
Amifampridina	1,43	Oral	TCMDC-143136 (0,2)
Rupatadina	1,43	Oral	TCMDC-143222 (0,32)
Nicorandil	1,4	Oral	TCMDC-143612 (0,3)
Idelalisibe	1,4	Oral	TCMDC-143279 (0,21)
Atomoxetina	1,4	Oral	TCMDC-143495 (0,25)
Clometiazol	1,39	Oral	TCMDC-143477 (0,24)
Droperidol	1,34	Intravenosa	TCMDC-143222 (0,26)
Fostamatinibe	1,31	Oral	TCMDC-143382 (0,19)
Parecoxibe	1,29	Intravenosa	TCMDC-143569 (0,27)
Piperaquina	1,27	Oral	TCMDC-143542 (0,25)
Omarigliptina	1,26	Oral	TCMDC-143114 (0,22)
Pirimetamina	1,25	Oral	TCMDC-143220 (0,26)
Piperazina	1,23	Oral	TCMDC-143601 (0,08)
Nimesulida	1,22	Oral	TCMDC-143608 (0,29)
Clotrimazol	1,21	Oral	TCMDC-143423 (0,28)
Nilutamida	1,21	Oral	TCMDC-143609 (0,25)
Clonazepam	1,21	Oral	TCMDC-143616 (0,33)
Furazolidona	1,2	Oral	TCMDC-143577 (0,31)
Azul de metileno	1,18	Intravenosa	TCMDC-143437 (0,23)
Metimazol	1,16	Oral	TCMDC-143381 (0,17)
Barnidipino	1,14	Oral	TCMDC-143484 (0,26)

Tabela 19 – Fármacos aprovados classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
Nicotina	1,14	Oral	TCMDC-143421 (0,3)
Anagliptina	1,12	Oral	TCMDC-143612 (0,24)
Empagliflozina	1,12	Oral	TCMDC-143328 (0,19)
Sulfafurazol	1,11	Oral	TCMDC-143345 (0,26)
Paroxetina	1,11	Oral	TCMDC-143334 (0,23)
Bretlium	1,09	Intravenosa	TCMDC-143609 (0,21)
Mosaprida	1,07	Oral	TCMDC-143504 (0,26)
Alopurinol	1,07	Oral	TCMDC-143083 (0,23)
Melperona	1,06	Oral	TCMDC-143222 (0,35)
Pirazinamida	1,03	Oral	TCMDC-143571 (0,21)
Didanosina	1,03	Oral	TCMDC-143083 (0,16)
Ambroxol	1,02	Oral	TCMDC-143507 (0,22)
Benznidazol	1,02	Oral	TCMDC-143126 (0,28)
Valaciclovir	1,01	Oral	TCMDC-143083 (0,17)
Amrinona	1,00	Intravenosa	TCMDC-143461 (0,23)
Cloropiramina	0,98	Oral	TCMDC-143247 (0,33)
Trapidil	0,98	Oral	TCMDC-143495 (0,2)
Histamina	0,98	Subcutânea	TCMDC-143317 (0,22)
Dasabuvir	0,98	Oral	TCMDC-143345 (0,2)
Mebidrolina	0,96	Oral	TCMDC-143389 (0,26)
Nitrofurazona	0,95	Tópica	TCMDC-143571 (0,35)
Stanozolol	0,95	Oral	TCMDC-143614 (0,15)
Sulfadiazina	0,94	Oral	TCMDC-143382 (0,22)
Sulfamerazina	0,94	Tópica	TCMDC-143345 (0,23)
Sulfametazina	0,92	Oral	TCMDC-143345 (0,23)
Mirtazapina	0,92	Oral	TCMDC-143421 (0,23)
Afloqualona	0,89	Oral	TCMDC-143258 (0,18)
Blonanserina	0,86	Oral	TCMDC-143556 (0,29)
Bosutinibe	0,85	Oral	TCMDC-143603 (0,32)
Mercaptopurina	0,83	Oral	TCMDC-143083 (0,24)
Fenazopiridina	0,81	Oral	TCMDC-143610 (0,2)
Oxatomida	0,81	Oral	TCMDC-143292 (0,24)
Isoniazida	0,81	Oral	TCMDC-143535 (0,27)
Pasiniazida	0,81	Oral	TCMDC-143535 (0,27)
Niclosamida	0,8	Oral	TCMDC-143609 (0,31)
Bromperidol	0,79	Oral	TCMDC-143222 (0,27)
Haloperidol	0,79	Oral	TCMDC-143222 (0,27)
Penbutolol	0,78	Oral	TCMDC-143604 (0,26)
Perampanel	0,77	Oral	TCMDC-143389 (0,22)
Etoricoxibe	0,77	Oral	TCMDC-143345 (0,24)
Roflumilaste	0,76	Oral	TCMDC-143535 (0,33)
Tegafur	0,76	Oral	TCMDC-143626 (0,21)
Tioguanina	0,75	Oral	TCMDC-143083 (0,18)
Clomipramina	0,75	Oral	TCMDC-143222 (0,24)
Mecamilamina	0,74	Oral	TCMDC-143481 (0,13)
Flucitosina	0,73	Oral	TCMDC-143626 (0,19)
Anagrelida	0,72	Oral	TCMDC-143479 (0,22)
Adenosina	0,71	Intravenosa	TCMDC-143083 (0,19)
Vidarabina	0,71	Tópica	TCMDC-143083 (0,19)
Doxazosina	0,71	Oral	TCMDC-143104 (0,22)
Dapiprazol	0,69	Oftálmica	TCMDC-143495 (0,27)

Tabela 19 – Fármacos aprovados classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
Tezacaftor	0,68	Oral	TCMDC-143507 (0,24)
Perexilina	0,67	Oral	TCMDC-143612 (0,14)
Dacarbazina	0,66	Intravenosa	TCMDC-143083 (0,17)
Moclobemida	0,64	Oral	TCMDC-143504 (0,3)
Prazosina	0,64	Oral	TCMDC-142900 (0,26)
Axitinibe	0,64	Oral	TCMDC-143604 (0,23)
Imidafenacina	0,63	Oral	TCMDC-143424 (0,25)
Teobromina	0,61	Oral	TCMDC-143083 (0,17)
Zonisamida	0,61	Oral	TCMDC-143599 (0,23)
Mononitrato de isossorbida	0,56	Oral	TCMDC-143149 (0,14)
Flutamida	0,56	Oral	TCMDC-143545 (0,3)
Metenamina	0,55	Oral	TCMDC-142900 (0,08)
Clofazimina	0,54	Oral	TCMDC-143279 (0,24)
Mexiletina	0,53	Oral	TCMDC-143495 (0,23)
Piridoxal	0,52	Oral	TCMDC-143126 (0,18)
Ciclopirox	0,51	Tópica	TCMDC-143191 (0,21)
Benperidol	0,48	Oral	TCMDC-143222 (0,28)
Minoxidil	0,46	Oral	TCMDC-143292 (0,24)
Meclizina	0,46	Oral	TCMDC-143504 (0,31)
Becanamicina	0,46	Oftálmica	TCMDC-143083 (0,14)
Sorafenibe	0,46	Oral	TCMDC-143199 (0,27)
Ipragliflozina	0,46	Oral	TCMDC-143354 (0,23)
Desipramina	0,45	Oral	TCMDC-143382 (0,19)
Torasemida	0,42	Oral	TCMDC-143349 (0,24)
Trientina	0,42	Oral	TCMDC-143382 (0,09)
Tranilcipromina	0,42	Oral	TCMDC-143371 (0,24)
Cinacalcete	0,42	Oral	TCMDC-143548 (0,2)
Propatilnitrato	0,42	Oral	TCMDC-143601 (0,14)
Primaquina	0,4	Oral	TCMDC-143603 (0,47)
Trimetazidina	0,39	Oral	TCMDC-142900 (0,23)
Buclicina	0,39	Oral	TCMDC-143504 (0,34)
Guanfacina	0,38	Oral	TCMDC-143279 (0,24)
Fenazona	0,35	Ótica	TCMDC-143148 (0,15)
Floxuridina	0,34	Intravenosa	TCMDC-143083 (0,21)
Idoxuridina	0,34	Tópica	TCMDC-143083 (0,21)
Ácido clodrônico	0,34	Oral	TCMDC-143224 (0,14)
Sertralina	0,34	Oral	TCMDC-143409 (0,2)
Ibrutinibe	0,34	Oral	TCMDC-143511 (0,27)
Isocarboxazida	0,33	Oral	TCMDC-123621 (0,24)
Mecloretamina	0,31	Tópica	TCMDC-143426 (0,13)
Fluoxetina	0,31	Oral	TCMDC-143310 (0,24)
Dipiridamol	0,3	Oral	TCMDC-143292 (0,2)
Voglibose	0,3	Oral	TCMDC-143614 (0,12)
Riociguat	0,3	Oral	TCMDC-143253 (0,24)
Triclocarban	0,29	Tópica	TCMDC-143405 (0,31)
Metcranio	0,29	Oral	TCMDC-143332 (0,21)
Mafenida	0,29	Tópica	TCMDC-143382 (0,2)
Tipiracil	0,29	Oral	TCMDC-143289 (0,16)
Azacitidina	0,27	Intravenosa	TCMDC-143083 (0,23)
Netupitanto	0,26	Oral	TCMDC-143247 (0,26)

Tabela 19 – Fármacos aprovados classificados como ativos pelo modelo NB – FeatMorgan (conclusão)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
Belinostate	0,26	Intravenosa	TCMDC-143615 (0,31)
Hexoprenalina	0,25	Oral	TCMDC-143409 (0,14)
Triclosan	0,25	Tópica	TCMDC-143151 (0,3)
Reboxetina	0,24	Oral	TCMDC-143104 (0,23)
Uracila	0,24	Oral	TCMDC-143410 (0,13)
Teprenona	0,24	Oral	TCMDC-143432 (0,12)
Tofacitinibe	0,23	Oral	TCMDC-143314 (0,2)
Tafamidis	0,23	Oral	TCMDC-143220 (0,2)
Mitotano	0,22	Oral	TCMDC-143609 (0,23)
Gefitinibe	0,22	Oral	TCMDC-143426 (0,21)
Flumazenil	0,21	Intravenosa	TCMDC-143616 (0,23)
Avanafila	0,21	Oral	TCMDC-143298 (0,27)
Ambenônio	0,2	Oral	TCMDC-143130 (0,28)
Dinitrato de isossorbida	0,2	Oral	TCMDC-143149 (0,16)
Indoramina	0,2	Oral	TCMDC-143256 (0,34)
Revaprazana	0,19	Oral	TCMDC-143109 (0,34)
Niraparibe	0,19	Oral	TCMDC-143382 (0,24)
Emorfazona	0,17	Oral	TCMDC-143239 (0,23)
Sotalol	0,16	Oral	TCMDC-143286 (0,21)
Lomerizina	0,16	Oral	TCMDC-143504 (0,33)
Linsidomina	0,16	Intravenosa	TCMDC-143379 (0,15)

TCMDC – Compostos da biblioteca Tres Cantos Medicines Development Campus.

**APÊNDICE D – LISTA DE FÁRMACOS EM DESENVOLVIMENTO CLÍNICO
SELECIONADOS PELO MODELO NB – FEATMORGAN NA ABORDAGEM
LDBB**

Tabela 20 – Candidatos a fármaco em desenvolvimento clínico classificados como ativos pelo modelo NB – FeatMorgan (continua)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
Ravuconazol	8,82	Oral	TCMDC-143082 (0,38)
CYT-387	7,79	Oral	TCMDC-143239 (0,28)
XL-019	7,21	Oral	TCMDC-143617 (0,29)
TG-02	6,86	Oral	TCMDC-143349 (0,27)
CYC-116	6,84	Oral	TCMDC-143617 (0,26)
PHA-848125	6,66	Oral	TCMDC-143542 (0,26)
SCH-900776	6,15	Intravenosa	TCMDC-143426 (0,22)
Sipatrigina	6,1	Intravenosa	TCMDC-143426 (0,36)
ARQ-092	6,1	Oral	TCMDC-143569 (0,24)
TRV-130	5,85	Intravenosa	TCMDC-142900 (0,26)
BMS-833923	5,71	Oral	TCMDC-143136 (0,27)
PU-H71	5,65	Oral	TCMDC-143510 (0,21)
XL-228	5,28	Intravenosa	TCMDC-143426 (0,21)
Ravoxertinibe	5,27	Oral	TCMDC-143409 (0,25)
GDC-0994	5,27	Oral	TCMDC-143409 (0,25)
Acalisibe	5,25	Oral	TCMDC-143321 (0,18)
Pritelivir	4,9	Oral	TCMDC-143219 (0,26)
Loreclezol	4,9	Oral	TCMDC-143082 (0,38)
BIIB-021	4,77	Oral	TCMDC-143298 (0,22)
MK-3697	4,76	Oral	TCMDC-143084 (0,3)
MK-2295	4,71	Oral	TCMDC-143463 (0,24)
AT-13148	4,67	Oral	TCMDC-143614 (0,22)
RAF-265	4,63	Oral	TCMDC-143382 (0,22)
Ocinaplon	4,57	Oral	TCMDC-142900 (0,23)
NVP-BGJ398	4,54	Oral	TCMDC-143545 (0,3)
AMG-319	4,53	Oral	TCMDC-143180 (0,22)
GZ-644282	4,45	Intravenosa	TCMDC-143285 (0,22)
Sanazol	4,39	Intravenosa	TCMDC-143284 (0,41)
NSI-189	4,21	Oral	TCMDC-143499 (0,37)
ABT-202	4,08	Oral	TCMDC-143422 (0,43)
JNJ-42165279	4,05	Oral	TCMDC-143161 (0,32)
AEE-788	4,04	Oral	TCMDC-143502 (0,22)
Lazabemida	4,01	Oral	TCMDC-143219 (0,38)
VX-745	4	Oral	TCMDC-143151 (0,3)
Zibotentan	3,91	Oral	TCMDC-143345 (0,22)
Tecastemizol	3,83	Oral	TCMDC-143279 (0,31)
Indiplon	3,77	Oral	TCMDC-143201 (0,23)
AZD5438	3,66	Oral	TCMDC-143382 (0,26)

Tabela 20 – Candidatos a fármaco em desenvolvimento clínico classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
Talarozol	3,66	Oral	TCMDC-143476 (0,29)
Broxiquinolina	3,59	Oral	TCMDC-143180 (0,29)
TAK-715	3,59	Oral	TCMDC-143569 (0,34)
YM-511	3,52	Oral	TCMDC-143609 (0,27)
JNJ-37822681	3,5	Oral	TCMDC-143179 (0,28)
Linifanibe	3,49	Oral	TCMDC-143135 (0,23)
BIBX-1382	3,49	Oral	TCMDC-143617 (0,21)
Y-39983	3,48	Oftálmico	TCMDC-143314 (0,25)
Orteronel	3,46	Oral	TCMDC-143336 (0,3)
CS-917	3,3	Oral	TCMDC-143569 (0,25)
Metatinibe	3,29	Oral	TCMDC-143433 (0,29)
LGK-974	3,27	Oral	TCMDC-143345 (0,28)
Tirasemtive	3,24	Oral	TCMDC-143381 (0,21)
Mocetinostate	3,18	Oral	TCMDC-143617 (0,24)
Tegobuvir	3,1	Oral	TCMDC-143190 (0,23)
LY-2228820	3,09	Oral	TCMDC-143220 (0,24)
SGX523	3,09	Oral	TCMDC-143231 (0,25)
2,4-dinitroclorobenzeno	3,05	Tópico	TCMDC-143609 (0,33)
Defactinibe	3,04	Oral	TCMDC-143382 (0,26)
NS-8	2,98	Oral	TCMDC-143422 (0,25)
Berzosertibe	2,96	Intravenosa	TCMDC-143282 (0,23)
Nicaravena	2,93	Intravenosa	TCMDC-143324 (0,31)
Gaboxadol	2,92	Oral	TCMDC-143346 (0,25)
adoprazine	2,9	Oral	TCMDC-143542 (0,3)
PRT062070	2,9	Oral	TCMDC-143382 (0,26)
Befuralina	2,89	Oral	TCMDC-143499 (0,34)
Apatinibe	2,85	Oral	TCMDC-143535 (0,29)
Ranirestate	2,83	Oral	TCMDC-143253 (0,25)
Motesanibe	2,81	Oral	TCMDC-143535 (0,3)
LY-2784544	2,77	Oral	TCMDC-143321 (0,22)
SR-27897	2,76	Oral	TCMDC-143521 (0,35)
Foropafante	2,75	Oral	TCMDC-143247 (0,24)
JNJ-38877605	2,74	Oral	TCMDC-143553 (0,22)
Roscovitina	2,73	Oral	TCMDC-143379 (0,23)
PF-02545920	2,66	Oral	TCMDC-143599 (0,26)
Bavisante	2,66	Oral	TCMDC-143499 (0,36)
AZD6765	2,66	Oral	TCMDC-143599 (0,29)
Cotinina	2,64	Oral	TCMDC-143529 (0,39)
Zardaverina	2,6	Inalatório	TCMDC-143534 (0,27)
GSK461364	2,58	Intravenosa	TCMDC-143161 (0,29)
Baricitinibe	2,5	Oral	TCMDC-143314 (0,19)
Iniparibe	2,48	Intravenosa	TCMDC-143609 (0,37)
CUDC-907	2,47	Oral	TCMDC-143239 (0,22)

Tabela 20 – Candidatos a fármaco em desenvolvimento clínico classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
L-778123	2,47	Intravenosa	TCMDC-143179 (0,29)
Voxtalisibe	2,46	Oral	TCMDC-143610 (0,24)
Danusertibe	2,43	Intravenosa	TCMDC-143434 (0,25)
AP26113	2,4	Oral	TCMDC-143109 (0,25)
Arofilina	2,38	Oral	TCMDC-143279 (0,23)
Pevonedistate	2,37	Intravenosa	TCMDC-143502 (0,21)
PRT-062607	2,34	Oral	TCMDC-143314 (0,29)
GS-6201	2,31	Oral	TCMDC-143292 (0,21)
Abametapir	2,27	Tópico	TCMDC-143440 (0,22)
YM-976	2,27	Oral	TCMDC-143306 (0,29)
serdemetan	2,26	Oral	TCMDC-143256 (0,27)
NGD-94-1	2,25	Oral	TCMDC-143426 (0,21)
E7449	2,24	Oral	TCMDC-143389 (0,22)
ODM-201	2,24	Oral	TCMDC-143306 (0,25)
BAY-1841788	2,21	Oral	TCMDC-143190 (0,27)
Quisinostate	2,2	Oral	TCMDC-143272 (0,24)
Etanidazol	2,18	Intravenosa	TCMDC-143417 (0,22)
N6022	2,17	Intravenosa	TCMDC-143321 (0,26)
NVP-HSP990	2,15	Oral	TCMDC-143220 (0,25)
Tanzisertibe	2,1	Oral	TCMDC-143379 (0,26)
Sulmazol	2,1	Oral	TCMDC-143593 (0,23)
ASP-2535	2,07	Oral	TCMDC-143445 (0,21)
GSK1070916	2,06	Intravenosa	TCMDC-143620 (0,25)
Tesevatinibe	2,04	Oral	TCMDC-143603 (0,21)
Suritozol	2,04	Oral	TCMDC-143555 (0,24)
Capadenosona	2,04	Oral	TCMDC-143220 (0,32)
RWJ-67657	1,99	Oral	TCMDC-143409 (0,27)
Difenciprona	1,98	Tópico	TCMDC-143286 (0,19)
Basimglurante	1,98	Oral	TCMDC-143190 (0,22)
LCL-161	1,97	Oral	TCMDC-143334 (0,27)
Brivanibe	1,97	Oral	TCMDC-143439 (0,22)
Prexasertibe	1,96	Intravenosa	TCMDC-143586 (0,2)
PFK-158	1,95	Intravenosa	TCMDC-143553 (0,28)
Latrepirdina	1,93	Oral	TCMDC-143222 (0,22)
Branaplam	1,93	Oral	TCMDC-143201 (0,23)
Anpirtolina	1,92	Oral	TCMDC-143382 (0,27)
CEP-32496	1,91	Oral	TCMDC-143578 (0,2)
Talazoparibe	1,89	Oral	TCMDC-143614 (0,23)
OVI-123	1,86	Intravenosa	TCMDC-143410 (0,18)
Tozasertibe	1,85	Intravenosa	TCMDC-143545 (0,23)
Dovitinibe	1,83	Oral	TCMDC-143292 (0,29)
acadesine	1,81	Intravenosa	TCMDC-143083 (0,2)
MLN2480	1,79	Oral	TCMDC-143156 (0,25)

Tabela 20 – Candidatos a fármaco em desenvolvimento clínico classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
Bupicomida	1,79	Oral	TCMDC-143190 (0,25)
Uridina	1,79	Oral	TCMDC-143083 (0,24)
Duvelisibe	1,78	Oral	TCMDC-143083 (0,19)
AZD1446	1,77	Oral	TCMDC-143262 (0,24)
TG100-115	1,76	Inalatório	TCMDC-143606 (0,23)
Taranabante	1,76	Oral	TCMDC-143548 (0,29)
Quizartinibe	1,76	Oral	TCMDC-143239 (0,21)
PNU-142633	1,74	Oral	TCMDC-143598 (0,24)
Pozaniclina	1,73	Oral	TCMDC-143201 (0,23)
Opicapona	1,73	Oral	TCMDC-143541 (0,22)
Rucaparibe	1,72	Oral	TCMDC-143190 (0,24)
T-62	1,67	Oral	TCMDC-143556 (0,33)
ENMD-2076	1,64	Oral	TCMDC-143434 (0,2)
Bitopertina	1,6	Oral	TCMDC-143159 (0,24)
CI-923	1,58	Oral	TCMDC-143389 (0,25)
Vipadenante	1,58	Oral	TCMDC-143221 (0,22)
Bropirimina	1,57	Oral	TCMDC-143569 (0,23)
Kelina	1,57	Oral	TCMDC-143262 (0,22)
AZD1080	1,56	Oral	TCMDC-143440 (0,23)
m-clorofenilpiperazina	1,56	Oral	TCMDC-143104 (0,27)
ETC-159	1,55	Oral	TCMDC-143569 (0,24)
Pibenzimol	1,55	Intravenosa	TCMDC-143292 (0,4)
LXR-623	1,54	Oral	TCMDC-143082 (0,25)
rimcazole	1,53	Oral	TCMDC-143389 (0,24)
OSI-027	1,52	Oral	TCMDC-143262 (0,21)
INC-280	1,51	Oral	TCMDC-143463 (0,25)
Muscimol	1,51	Infusão cerebral	TCMDC-143120 (0,18)
MK-8033	1,51	Oral	TCMDC-143553 (0,23)
Tenilsetam	1,5	Oral	TCMDC-143219 (0,26)
GDC-0941	1,49	Oral	TCMDC-143620 (0,21)
Etimizol	1,49	Oral	TCMDC-143304 (0,21)
R306465	1,48	Oral	TCMDC-143465 (0,27)
JNJ-40411813	1,46	Oral	TCMDC-143409 (0,22)
Enzastaurina	1,44	Oral	TCMDC-143389 (0,33)
Pafuramidina	1,43	Oral	TCMDC-143569 (0,27)
Hidroxiflutamida	1,42	Intraprostática	TCMDC-143545 (0,3)
AS-703026	1,41	Oral	TCMDC-143253 (0,24)
Dapivirina	1,41	Vaginal	TCMDC-143109 (0,26)
Balicatibe	1,41	Oral	TCMDC-143434 (0,25)
Resminostate	1,41	Oral	TCMDC-143549 (0,23)
BMS-927711	1,41	Oral	TCMDC-143381 (0,43)
ABT-491	1,4	Oral	TCMDC-143190 (0,21)

Tabela 20 – Candidatos a fármaco em desenvolvimento clínico classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
CC-223	1,4	Oral	TCMDC-143465 (0,22)
Amuvatinibe	1,38	Oral	TCMDC-143625 (0,28)
Ozanimode	1,35	Oral	TCMDC-143382 (0,23)
Selinexor	1,34	Oral	TCMDC-143561 (0,21)
RS-127445	1,32	Oral	TCMDC-143426 (0,21)
Piridoxamina	1,3	Oral	TCMDC-143308 (0,18)
AT-7519	1,26	Intravenosa	TCMDC-143253 (0,23)
ABT-639	1,25	Oral	TCMDC-143499 (0,28)
AZD7762	1,25	Intravenosa	TCMDC-143306 (0,27)
Tioguanina	1,24	Oral	TCMDC-143083 (0,19)
LY450108	1,22	Oral	TCMDC-143535 (0,4)
AZD1480	1,21	Oral	TCMDC-143306 (0,17)
PH-797804	1,2	Oral	TCMDC-143328 (0,27)
ONO-4059	1,2	Oral	TCMDC-143492 (0,24)
IOWH032	1,19	Oral	TCMDC-143103 (0,26)
Uprosertibe	1,19	Oral	TCMDC-143190 (0,28)
Enprofilina	1,18	Oral	TCMDC-143626 (0,2)
Pirazinoilguanidina	1,18	Oral	TCMDC-143304 (0,21)
GSK-1292263	1,17	Oral	TCMDC-143282 (0,28)
MK-0354	1,17	Oral	TCMDC-143289 (0,16)
BTS-72664	1,15	Oral	TCMDC-143415 (0,27)
DMP-543	1,15	Oral	TCMDC-143224 (0,29)
KW-2449	1,14	Oral	TCMDC-143494 (0,33)
Carboxiamidotriazol	1,12	Oral	TCMDC-143561 (0,26)
Telenzepina	1,11	Oral	TCMDC-143362 (0,22)
AZD4547	1,11	Oral	TCMDC-143434 (0,26)
Xaliprodena	1,06	Oral	TCMDC-143465 (0,23)
SAR-245409	1,05	Oral	TCMDC-143108 (0,31)
GTx-007	1,05	Oral	TCMDC-143545 (0,3)
Tenidape	1,05	Oral	TCMDC-143542 (0,28)
Tepotinibe	1,04	Oral	TCMDC-143620 (0,26)
Lomeguatribe	1,02	Oral	TCMDC-143297 (0,22)
Alvelestate	1,01	Oral	TCMDC-143221 (0,22)
Fostamatinibe	1	Oral	TCMDC-143298 (0,22)
APTO-253	0,98	Intravenosa	TCMDC-143159 (0,26)
Didanosina	0,96	Oral	TCMDC-143410 (0,18)
ABT-751	0,95	Oral	TCMDC-143345 (0,27)
Pilaralisibe	0,95	Oral	TCMDC-143199 (0,22)
SB-939	0,95	Oral	TCMDC-143257 (0,25)
PF-04691502	0,93	Oral	TCMDC-143610 (0,23)
Intepirdina	0,93	Oral	TCMDC-143382 (0,24)
Decernotinibe	0,92	Oral	TCMDC-143362 (0,18)
Nibentana	0,91	Intravenosa	TCMDC-143282 (0,26)

Tabela 20 – Candidatos a fármaco em desenvolvimento clínico classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
Lexibulina	0,91	Oral	TCMDC-143190 (0,25)
Broxaterol	0,9	Inalatório	TCMDC-143614 (0,22)
PRX-08066	0,89	Oral	TCMDC-143179 (0,24)
Rivaniclina	0,89	Oral	TCMDC-143309 (0,24)
Leteprininim	0,88	Oral	TCMDC-143186 (0,26)
MSX-122	0,88	Oral	TCMDC-143379 (0,2)
TBA-354	0,87	Oral	TCMDC-143522 (0,27)
Misonidazol	0,87	Intravenosa	TCMDC-143149 (0,23)
Osemozotana	0,86	Oral	TCMDC-143382 (0,2)
P276-00	0,86	Intravenosa	TCMDC-143190 (0,23)
Fexinidazol	0,83	Oral	TCMDC-143522 (0,3)
Priralfinamida	0,83	Oral	TCMDC-143328 (0,39)
CH5132799	0,82	Oral	TCMDC-143617 (0,23)
Nepicastate	0,81	Oral	TCMDC-143641 (0,23)
SB-2343	0,78	Oral	TCMDC-143617 (0,2)
CX-516	0,78	Oral	TCMDC-143425 (0,28)
Azimilida	0,77	Oral	TCMDC-143505 (0,36)
Tasisulam	0,76	Intravenosa	TCMDC-143114 (0,29)
Tirapazamina	0,76	Intravenosa	TCMDC-143410 (0,12)
Isamoltana	0,74	Intravenosa	TCMDC-143253 (0,19)
SNS-314	0,73	Intravenosa	TCMDC-143405 (0,29)
Hupertzine A	0,72	Oral	TCMDC-143087 (0,18)
CX-516	0,72	Oral	TCMDC-143625 (0,39)
PF-03084014	0,71	Oral	TCMDC-143423 (0,22)
Setanaxibe	0,7	Oral	TCMDC-143306 (0,22)
BI-847325	0,66	Oral	TCMDC-143484 (0,32)
PF-562271	0,65	Oral	TCMDC-143529 (0,22)
DMP-777	0,65	Oral	TCMDC-143625 (0,25)
Idalopirdina	0,63	Oral	TCMDC-143455 (0,24)
Imetridina	0,63	Oral	TCMDC-143087 (0,19)
Indibulina	0,61	Oral	TCMDC-143529 (0,29)
CC-401	0,6	Oral	TCMDC-143598 (0,29)
RGB-286638	0,59	Intravenosa	TCMDC-143433 (0,23)
PCI-24781	0,57	Oral	TCMDC-123621 (0,28)
Tipifarnibe	0,57	Oral	TCMDC-143409 (0,25)
Linsitinibe	0,57	Oral	TCMDC-143463 (0,22)
Aminocrisena	0,56	Intravenosa	TCMDC-143476 (0,16)
Ibipinabante	0,56	Oral	TCMDC-143130 (0,27)
Dazoxibena	0,56	Oral	TCMDC-143590 (0,37)
Ebselena	0,54	Oral	TCMDC-143490 (0,16)
Nisoxetina	0,53	Oral	TCMDC-143221 (0,22)
Filorexante	0,49	Oral	TCMDC-143507 (0,22)
AMG-208	0,48	Oral	TCMDC-143231 (0,25)

Tabela 20 – Candidatos a fármaco em desenvolvimento clínico classificados como ativos pelo modelo NB – FeatMorgan (conclusão)

Fármaco	Pontuação no modelo NB – FeatMorgan	Via de administração	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)
Celgosivir	0,46	Oral	TCMDC-143510 (0,14)
PF-3274167	0,45	Oral	TCMDC-143590 (0,29)
Remimazolam	0,45	Intravenosa	TCMDC-143253 (0,23)
Pexidartinibe	0,44	Oral	TCMDC-143179 (0,19)
Rabusertibe	0,42	Intravenosa	TCMDC-143507 (0,23)
Fagomina	0,42	Oral	TCMDC-143382 (0,1)
ARQ-621	0,42	Intravenosa	TCMDC-143409 (0,27)
Halopemida	0,4	Oral	TCMDC-143279 (0,27)
AZD8931	0,39	Oral	TCMDC-143257 (0,23)
RX-3117	0,39	Oral	TCMDC-143410 (0,2)
Emetina	0,39	Subcutânea	TCMDC-143586 (0,22)
PK-11195	0,38	Oral	TCMDC-143190 (0,27)
N- α -metilhistamina	0,36	Subcutânea	TCMDC-143641 (0,14)
Prinaberel	0,36	Oral	TCMDC-143255 (0,21)
VS-4718	0,35	Oral	TCMDC-143310 (0,29)
Talnetante	0,35	Oral	TCMDC-143569 (0,33)
Filgotinibe	0,35	Oral	TCMDC-143620 (0,27)
Broxuridina	0,34	Intravenosa	TCMDC-143083 (0,21)
Tideglusibe	0,34	Oral	TCMDC-140766 (0,18)
Pleconaril	0,33	Oral	TCMDC-143362 (0,25)
Ecabapida	0,31	Oral	TCMDC-143434 (0,28)
GW-842166	0,3	Oral	TCMDC-143114 (0,24)
URB-597	0,3	Oral	TCMDC-143598 (0,38)
Telatinibe	0,29	Oral	TCMDC-143219 (0,27)
Ácido picolínico	0,29	Oral	TCMDC-143599 (0,31)
MK-5108	0,29	Oral	TCMDC-143253 (0,24)
PX-12	0,28	Intravenosa	TCMDC-143083 (0,19)
Fananserina	0,27	Oral	TCMDC-143542 (0,34)
Silmitasertibe	0,27	Oral	TCMDC-143529 (0,21)
Delanzomibe	0,26	Intravenosa	TCMDC-143445 (0,35)
LCQ-908	0,26	Oral	TCMDC-143179 (0,26)
SQ-109	0,26	Oral	TCMDC-143285 (0,12)
AZD-3839	0,25	Oral	TCMDC-143409 (0,19)
AZD-5363	0,23	Oral	TCMDC-143279 (0,24)
Butofilolol	0,22	Oral	TCMDC-143614 (0,24)
Loxoribina	0,2	Intramuscular	TCMDC-143083 (0,25)
Alpelisibe	0,19	Oral	TCMDC-143417 (0,27)
JNJ-26990990	0,18	Oral	TCMDC-143599 (0,28)
Presatovir	0,18	Oral	TCMDC-143529 (0,25)

TCMDC – Compostos da biblioteca Tres Cantos Medicines Development Campus.

**APÊNDICE E – LISTA DE METABÓLITOS DE FÁRMACOS SELECIONADOS
PELO MODELO NB – FEATMORGAN NA ABORDAGEM LDBB**

Tabela 21 – Metabólitos classificados como ativos pelo modelo NB – FeatMorgan
(continua)

Metabólito	Pontuação no modelo NB – FeatMorgan	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)	Fármaco de origem
DBMET00025	7,96	TCMDC-143433 (0,27)	Imatinibe
DBMET02354	5,57	TCMDC-143222 (0,21)	Ranitidina
DBMET02651	5,17	TCMDC-143590 (0,34)	Cetoconazol
DBMET01498	5,01	TCMDC-143179 (0,18)	Topiroxostate
DBMET01500	5,01	TCMDC-143466 (0,18)	Topiroxostate
DBMET01185	4,24	TCMDC-143082 (0,48)	Voriconazol
DBMET00072	4,22	TCMDC-143082 (0,37)	Voriconazol
DBMET00959	4,18	TCMDC-143601 (0,2)	Nizatidina
DBMET01735	4,18	TCMDC-143601 (0,2)	Ranitidina
DBMET00520	3,85	TCMDC-143593 (0,25)	Nicotina
DBMET01899	3,83	TCMDC-143620 (0,26)	Barnidipino
DBMET02605	3,83	TCMDC-143641 (0,2)	Aciclovir
DBMET02380	3,71	TCMDC-143382 (0,25)	Palbociclibe
DBMET02659	3,59	TCMDC-143590 (0,35)	Cetoconazol
DBMET02650	3,54	TCMDC-143590 (0,32)	Cetoconazol
DBMET02624	3,54	TCMDC-143578 (0,24)	Metronidazol
DBMET00416	3,21	TCMDC-143083 (0,17)	Azatioprina
DBMET02371	3,1	TCMDC-143382 (0,25)	Palbociclibe
DBMET01186	3,1	TCMDC-143614 (0,47)	Voriconazol
DBMET00271	3,07	TCMDC-143083 (0,17)	Theofilina
DBMET01948	3,04	TCMDC-143545 (0,23)	Piperaquina
DBMET02020	2,98	TCMDC-143426 (0,21)	Ibrutinibe
DBMET00527	2,98	TCMDC-143529 (0,37)	Nicotina
DBMET02658	2,96	TCMDC-143479 (0,32)	Cetoconazol
DBMET01768	2,87	TCMDC-143135 (0,23)	Tiabendazol
DBMET00526	2,85	TCMDC-143324 (0,29)	Nicotine
DBMET00265	2,79	TCMDC-143109 (0,22)	Olanzapina
DBMET01680	2,78	TCMDC-143164 (0,24)	Indiplona
DBMET01856	2,7	TCMDC-143426 (0,23)	Vemurafenibe
DBMET02375	2,69	TCMDC-143382 (0,23)	Palbociclibe
DBMET00751	2,68	TCMDC-143282 (0,28)	Dantrolena
DBMET00071	2,65	TCMDC-143082 (0,37)	Voriconazol
DBMET02995	2,65	TCMDC-143526 (0,23)	Ziprasidona
DBMET00038	2,65	TCMDC-143282 (0,23)	Acenocumarol
DBMET02350	2,64	TCMDC-143298 (0,26)	Prazosina
DBMET02653	2,6	TCMDC-143590 (0,33)	Cetoconazol
DBMET00438	2,58	TCMDC-125222 (0,22)	Lumefantrina
DBMET00308	2,49	TCMDC-143109 (0,19)	Olanzapina

Tabela 21 – Metabólitos classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Metabólito	Pontuação no modelo NB – FeatMorgan	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)	Fármaco de origem
DBMET02353	2,49	TCMDC-143601 (0,2)	Ranitidina
DBMET03100	2,48	TCMDC-143279 (0,22)	Selumetinibe
DBMET02179	2,43	TCMDC-143608 (0,25)	Etoperidona
DBMET00182	2,4	TCMDC-143273 (0,24)	Sildenafil
DBMET00026	2,34	TCMDC-143504 (0,27)	Imatinibe
DBMET01335	2,32	TCMDC-143504 (0,2)	Indapamida
DBMET01855	2,25	TCMDC-143289 (0,28)	Vemurafenibe
DBMET02378	2,24	TCMDC-143382 (0,25)	Palbociclibe
DBMET01945	2,23	TCMDC-143603 (0,26)	Piperaquina
DBMET01892	2,23	TCMDC-143556 (0,31)	Blonanserina
DBMET01868	2,15	TCMDC-143476 (0,23)	Rilpivirina
DBMET01536	2,06	TCMDC-143156 (0,4)	Betrixabana
DBMET01338	2,04	TCMDC-143569 (0,27)	Valdecocibe
DBMET02828	2,03	TCMDC-143461 (0,14)	Anastrozol
DBMET01099	2,02	TCMDC-123621 (0,27)	Carbamazepina
DBMET00736	2,02	TCMDC-143440 (0,19)	Fomepizol
DBMET01197	2,02	TCMDC-143504 (0,3)	Flunarizina
DBMET01331	2	TCMDC-143504 (0,21)	Indapamida
DBMET00103	1,95	TCMDC-143614 (0,27)	Haloperidol
DBMET02372	1,94	TCMDC-143492 (0,21)	Palbociclibe
DBMET01894	1,93	TCMDC-143556 (0,34)	Blonanserina
DBMET01621	1,93	TCMDC-143292 (0,16)	Bromopirimina
DBMET02660	1,9	TCMDC-143082 (0,25)	Cetoconazol
DBMET01947	1,89	TCMDC-143603 (0,21)	Piperaquina
DBMET00414	1,89	TCMDC-143087 (0,15)	Alopurinol
DBMET01620	1,88	TCMDC-143556 (0,23)	Brofaromina
DBMET01533	1,87	TCMDC-143156 (0,39)	Betrixabana
DBMET00036	1,85	TCMDC-143601 (0,23)	Acenocumarol
DBMET00380	1,83	TCMDC-143608 (0,27)	Nifedipino
DBMET02657	1,81	TCMDC-143590 (0,31)	Cetoconazol
DBMET01833	1,76	TCMDC-143601 (0,25)	Nicardipino
DBMET02635	1,74	TCMDC-142900 (0,25)	Feniramina
DBMET00949	1,72	TCMDC-143382 (0,17)	Piperazina
DBMET02504	1,72	TCMDC-143599 (0,23)	Duloxetina
DBMET00977	1,72	TCMDC-143609 (0,32)	Clorpropamida
DBMET02191	1,66	TCMDC-143499 (0,32)	Revefenacina
DBMET01349	1,66	TCMDC-143345 (0,25)	Valdecocibe
DBMET00277	1,65	TCMDC-143601 (0,25)	Bupropiona
DBMET01556	1,65	TCMDC-143601 (0,25)	Bupropiona
DBMET02304	1,64	TCMDC-143495 (0,25)	Atomoxetina
DBMET02824	1,63	TCMDC-143180 (0,14)	Hidralazina
DBMET02815	1,61	TCMDC-123621 (0,17)	Hidralazina
DBMET02996	1,58	TCMDC-143499 (0,28)	Ziprasidona

Tabela 21 - Metabólitos classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Metabólito	Pontuação no modelo NB – FeatMorgan	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)	Fármaco de origem
DBMET00630	1,57	TCMDC-143083 (0,19)	Azatioprina
DBMET01697	1,53	TCMDC-143142 (0,25)	Loperamida
DBMET00920	1,48	TCMDC-143310 (0,12)	Mofetil micofenolato
DBMET02634	1,46	TCMDC-143571 (0,24)	Niacina
DBMET02666	1,45	TCMDC-143300 (0,25)	Cetoconazol
DBMET01124	1,44	TCMDC-143317 (0,3)	Bortezomibe
DBMET02630	1,43	TCMDC-143504 (0,3)	Niacina
DBMET02566	1,41	TCMDC-143603 (0,25)	Proclorpromazina
DBMET00533	1,41	TCMDC-143324 (0,25)	Nicotina
DBMET00817	1,4	TCMDC-143608 (0,25)	Clomipramina
DBMET02319	1,38	TCMDC-143490 (0,29)	Ivacaftor
DBMET00001	1,35	TCMDC-143109 (0,21)	Aminofenazona
DBMET01174	1,32	TCMDC-143425 (0,17)	Adefovir
DBMET01771	1,31	TCMDC-143255 (0,2)	Triamtereno
DBMET02768	1,3	TCMDC-143590 (0,18)	Metimazol
DBMET02668	1,29	TCMDC-143082 (0,26)	Cetoconazol
DBMET01622	1,29	TCMDC-143255 (0,16)	Bropirimina
DBMET00326	1,26	TCMDC-143641 (0,16)	Teofilina
DBMET02662	1,23	TCMDC-143082 (0,24)	Cetoconazol
DBMET01627	1,22	TCMDC-143142 (0,24)	Clorpromazina
DBMET01613	1,19	TCMDC-143422 (0,29)	Aripiprazol
DBMET00244	1,18	TCMDC-143279 (0,22)	Clomipramina
DBMET01718	1,17	TCMDC-143309 (0,28)	Nicotina
DBMET01864	1,15	TCMDC-143109 (0,2)	Rilpivirina
DBMET01635	1,14	TCMDC-143222 (0,23)	Clomipramina
DBMET00643	1,14	TCMDC-143309 (0,26)	Nicotina
DBMET01199	1,13	TCMDC-143609 (0,32)	Flunarizina
DBMET01865	1,12	TCMDC-143476 (0,22)	Rilpivirina
DBMET02654	1,08	TCMDC-143590 (0,33)	Cetoconazol
DBMET01012	1,06	TCMDC-143417 (0,45)	Nitazoxanida
DBMET02655	1,06	TCMDC-143590 (0,34)	Cetoconazol
DBMET00399	1,05	TCMDC-143120 (0,21)	Zaleplona
DBMET02579	1,04	TCMDC-143507 (0,22)	Tezacaftor
DBMET02232	1,02	TCMDC-143349 (0,26)	Torasemida
DBMET00613	1	TCMDC-143529 (0,27)	Nicotina
DBMET01342	0,99	TCMDC-143569 (0,27)	Valdecóxibe
DBMET01557	0,99	TCMDC-143601 (0,26)	Bupropiona
DBMET00350	0,98	TCMDC-143279 (0,29)	Astemizol
DBMET02150	0,96	TCMDC-143504 (0,29)	Moclobemida
DBMET00780	0,95	TCMDC-143083 (0,18)	Mercaptopurina
DBMET00364	0,95	TCMDC-143279 (0,23)	Domperidona
DBMET02244	0,92	TCMDC-143566 (0,24)	Ranolazina
DBMET02177	0,9	TCMDC-143186 (0,24)	Etoferidona

Tabela 21 – Metabólitos classificados como ativos pelo modelo NB – FeatMorgan (continuação)

Metabólito	Pontuação no modelo NB – FeatMorgan	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)	Fármaco de origem
DBMET02317	0,89	TCMDC-143490 (0,31)	Ivacaftor
DBMET01019	0,88	TCMDC-143614 (0,25)	Lorcaserina
DBMET01344	0,87	TCMDC-143569 (0,28)	Valdecoxibe
DBMET00909	0,87	TCMDC-143120 (0,26)	Metirapona
DBMET02764	0,84	TCMDC-143253 (0,11)	Metimazol
DBMET01780	0,84	TCMDC-143614 (0,18)	Bufuralol
DBMET02816	0,84	TCMDC-143136 (0,16)	Hidralazina
DBMET02787	0,83	TCMDC-143222 (0,25)	Desloratadina
DBMET01532	0,83	TCMDC-143190 (0,25)	Atalureno
DBMET02656	0,82	TCMDC-143590 (0,34)	Cetoconazol
DBMET00626	0,81	TCMDC-143601 (0,24)	Fendimetrazina
DBMET00493	0,81	TCMDC-143083 (0,15)	Didanosina
DBMET02636	0,8	TCMDC-143586 (0,26)	Feniramina
DBMET01089	0,8	TCMDC-143109 (0,24)	Lidocaina
DBMET01964	0,79	TCMDC-143161 (0,25)	Benidipino
DBMET00095	0,79	TCMDC-143477 (0,27)	Ticlopidina
DBMET01686	0,78	TCMDC-143159 (0,22)	Laquinimode
DBMET01528	0,77	TCMDC-143222 (0,29)	Rupatadina
DBMET02028	0,76	TCMDC-143410 (0,21)	Ibrutinibe
DBMET00656	0,76	TCMDC-143614 (0,18)	Capecitabina
DBMET01756	0,72	TCMDC-143476 (0,26)	Tacrina
DBMET02480	0,71	TCMDC-143297 (0,17)	Triamtereno
DBMET02058	0,71	TCMDC-143626 (0,2)	Tegafur
DBMET01712	0,7	TCMDC-143315 (0,23)	Nefazodona
DBMET01379	0,67	TCMDC-143247 (0,24)	Netupitante
DBMET01957	0,67	TCMDC-143126 (0,19)	Imipramina
DBMET02884	0,67	TCMDC-143282 (0,27)	Acalabrutinibe
DBMET00618	0,67	TCMDC-142900 (0,25)	Ticlopidina
DBMET02568	0,65	TCMDC-143484 (0,26)	Fentanil
DBMET01631	0,64	TCMDC-143309 (0,24)	Cinarizina
DBMET00513	0,62	TCMDC-143609 (0,13)	Levocarnitina
DBMET01734	0,61	TCMDC-143126 (0,19)	Mianserina
DBMET01826	0,61	TCMDC-143126 (0,19)	Mianserina
DBMET01347	0,61	TCMDC-143569 (0,24)	Valdecoxibe
DBMET01814	0,59	TCMDC-143108 (0,31)	Encainida
DBMET00037	0,57	TCMDC-143282 (0,23)	Acenocumarol
DBMET01559	0,57	TCMDC-143641 (0,26)	Cloroquina
DBMET02739	0,57	TCMDC-143641 (0,26)	Cloroquina
DBMET00415	0,55	TCMDC-143273 (0,19)	Pirazinamida
DBMET00935	0,54	TCMDC-143104 (0,24)	Vandetanibe
DBMET00691	0,53	TCMDC-143331 (0,17)	Prazosina
DBMET00091	0,53	TCMDC-143253 (0,24)	Ticlopidina
DBMET01636	0,52	TCMDC-143142 (0,23)	Clomipramina

Tabela 21 – Metabólitos classificados como ativos pelo modelo NB – FeatMorgan (conclusão)

Metabólito	Pontuação no modelo NB – FeatMorgan	Composto da série treino de maior similaridade (Coeficiente de Tanimoto)	Fármaco de origem
DBMET00099	0,52	TCMDC-143362 (0,25)	Prasugrel
DBMET01827	0,51	TCMDC-143221 (0,18)	Mirtazapina
DBMET01888	0,5	TCMDC-143556 (0,28)	Blonanserina
DBMET02826	0,5	TCMDC-143614 (0,24)	Astrozol
DBMET01063	0,48	TCMDC-143324 (0,32)	Bortezomibe
DBMET01669	0,48	TCMDC-143520 (0,24)	Flosequinana
DBMET02231	0,47	TCMDC-143349 (0,25)	Torasemida
DBMET00096	0,46	TCMDC-143253 (0,26)	Ticlopidina
DBMET01558	0,45	TCMDC-143616 (0,2)	Bupropiona
DBMET00421	0,43	TCMDC-143292 (0,2)	Udenafila
DBMET02506	0,42	TCMDC-143599 (0,22)	Duloxetina
DBMET03064	0,42	TCMDC-143136 (0,23)	Cloroquina
DBMET00327	0,42	TCMDC-143180 (0,19)	Clorzoxazona
DBMET01893	0,42	TCMDC-143556 (0,3)	Blonanserina
DBMET02172	0,4	TCMDC-143545 (0,26)	Etoferidona
DBMET02098	0,39	TCMDC-143298 (0,21)	Fostamatinibe
DBMET02503	0,38	TCMDC-143599 (0,24)	Duloxetina
DBMET00412	0,38	TCMDC-143083 (0,22)	Nelarabina
DBMET02973	0,33	TCMDC-143422 (0,16)	Prometazina
DBMET02003	0,33	TCMDC-143626 (0,19)	Tegafur
DBMET00264	0,32	TCMDC-143603 (0,19)	Olanzapina
DBMET00395	0,3	TCMDC-143253 (0,24)	Triazolam
DBMET02169	0,3	TCMDC-143083 (0,15)	Etoferidona
DBMET01207	0,29	TCMDC-142900 (0,24)	Ticlopidina
DBMET02740	0,29	TCMDC-143476 (0,28)	Cloroquina
DBMET01488	0,28	TCMDC-143114 (0,27)	Betahistina
DBMET00688	0,27	TCMDC-142900 (0,25)	Betahistina
DBMET01525	0,25	TCMDC-143253 (0,23)	Etizolam
DBMET00374	0,25	TCMDC-143082 (0,31)	Itraconazol
DBMET00363	0,24	TCMDC-142900 (0,23)	Disopiramida
DBMET00923	0,24	TCMDC-143424 (0,23)	Zonisamida
DBMET02170	0,21	TCMDC-143083 (0,15)	Etoferidona
DBMET01890	0,21	TCMDC-143556 (0,29)	Blonanserina
DBMET00341	0,19	TCMDC-143371 (0,28)	Fentanil
DBMET01082	0,18	TCMDC-143324 (0,29)	Bortezomibe
DBMET01085	0,18	TCMDC-143324 (0,29)	Bortezomibe
DBMET00536	0,18	TCMDC-143517 (0,05)	Esmolol
DBMET00249	0,17	TCMDC-143608 (0,26)	Flutamida
DBMET01757	0,16	TCMDC-143556 (0,25)	Tacrina

TCMDC – Compostos da biblioteca Tres Cantos Medicines Development Campus. DBMET – Compostos do banco de dados de metabólitos do DrugBank (WISHART et al., 2018).

APÊNDICE F – RESULTADOS DA VALIDAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA DA ABORDAGEM SBDD

Tabela 22 – Performance dos modelos de aprendizado de máquina de reclassificação na validação cruzada quádrupla

Modelo	SE (%)	SP (%)	PPV (%)	NPV (%)	TCB (%)	AUC ROC	AUC BEDROC
NB - TcDHODH - 1	64,0	75,8	66,7	73,5	69,9	0,76	0,68
NB - TcDHODH - 2	76,0	69,7	65,5	79,3	72,8	0,76	0,73
NB - TcDHODH - 3	56,0	81,8	70,0	71,1	68,9	0,75	0,67
NB - TcDHODH - 4	60,0	78,8	68,2	72,2	69,4	0,78	0,71
NB - TcDHODH - 5	76,0	72,7	67,9	80,0	74,4	0,76	0,66
NB - TcDHODH - 6	76,0	69,7	65,5	79,3	72,8	0,76	0,73
NB - TcDHODH - 7	48,0	84,8	70,6	68,3	66,4	0,76	0,72
NB - TcDHODH - 8	76,0	78,8	73,1	81,3	77,4	0,81	0,74

Os modelos de aprendizado de máquina foram desenvolvidos a partir da série teste e dos descritores da entalpia de interação do complexo P:L (ΔH_{int}), de lipofilicidade e das perdas entrópicas do ligante ($clogP$ e E_{tor} , respectivamente). Diferentes modelos foram obtidos a partir de variações no método de cálculo da ΔH_{int} , que foram: os resíduos da proteína incluídos nos cálculos, os átomos que tiveram a sua geometria otimizada e a constante dielétrica do modelo de solvatação COSMO. A capacidade preditiva na validação cruzada quádrupla foi avaliada utilizando os parâmetros: sensibilidade (SE), especificidade (SP), valor preditivo positivo e negativo (PPV e NPV), taxa de classificação balanceada (TCB), área sob as curvas ROC e BEDROC (AUC ROC e AUC BEDROC, respectivamente).

Tabela 23 – Performance dos modelos de aprendizado de máquina de reclassificação no teste de aleatorização

Modelo	SE (%)	SP (%)	PPV (%)	NPV (%)	TCB (%)	AUC ROC	AUC BEDROC
NB - TcDHODH - 1	28,6	60,0	40,0	47,4	44,3	0,42	0,45
NB - TcDHODH - 2	25,0	70,0	43,8	50,0	47,5	0,42	0,45
NB - TcDHODH - 3	53,6	43,3	46,9	50,0	48,5	0,39	0,34
NB - TcDHODH - 4	28,6	53,3	36,4	44,4	41,0	0,41	0,41
NB - TcDHODH - 5	28,6	50,0	34,8	42,9	39,3	0,42	0,45
NB - TcDHODH - 6	25,0	70,0	43,8	50,0	47,5	0,40	0,44
NB - TcDHODH - 7	53,6	36,7	44,1	45,8	45,1	0,39	0,34
NB - TcDHODH - 8	32,1	60,0	42,9	48,6	46,1	0,42	0,41

Os modelos de aprendizado de máquina foram desenvolvidos a partir da série teste e dos descritores da entalpia de interação do complexo P:L (ΔH_{int}), de lipofilicidade e das perdas entrópicas do ligante ($clogP$ e E_{tor} , respectivamente). Diferentes modelos foram obtidos a partir de variações no método de cálculo da ΔH_{int} , que foram: os resíduos da proteína incluídos nos cálculos, os átomos que tiveram a sua geometria otimizada e a constante dielétrica do modelo de solvatação COSMO. A capacidade preditiva no teste de aleatorização foi avaliada utilizando os parâmetros: sensibilidade (SE), especificidade (SP), valor preditivo positivo e negativo (PPV e NPV), taxa de classificação balanceada (TCB), área sob as curvas ROC e BEDROC (AUC ROC e AUC BEDROC, respectivamente).