

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Faculdade de Farmácia
Programa de Pós-Graduação em Ciências Farmacêuticas

Gabriel Corrêa Veríssimo

**INTEGRAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA AO
DOCKING MOLECULAR PARA PLANEJAMENTO E REALIZAÇÃO DE ENSAIOS
IN VITRO DE INIBIDORES DA ENZIMA ENOIL-ACP-REDUTASE NAD(P)H-
DEPENDENTE (FabI) DE *Staphylococcus aureus* E DE *Escherichia coli***

Belo Horizonte

2023

Gabriel Corrêa Verfssimo

**INTEGRAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA AO
DOCKING MOLECULAR PARA PLANEJAMENTO E REALIZAÇÃO DE ENSAIOS
IN VITRO DE INIBIDORES DA ENZIMA ENOIL-ACP-REDUTASE NAD(P)H-
DEPENDENTE (FabI) DE *Staphylococcus aureus* E DE *Escherichia coli***

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Farmacêuticas da Faculdade de Farmácia da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do grau de Mestre em Ciências Farmacêuticas.

Orientador: Prof. Dr. Vinícius Gonçalves Maltarollo

Coorientadora: Profa. Dra. Débora Maria Abrantes Costa

Belo Horizonte

2023

V517i Veríssimo, Gabriel Corrêa.
Integração de algoritmos de aprendizado de máquina ao *docking* molecular para planejamento e realização de ensaios *in vitro* de inibidores da enzima enoil-ACP-redutase NAD(P)H-dependente (FabI) de *Staphylococcus aureus* e de *Escherichia coli* [recurso eletrônico] / Gabriel Corrêa Veríssimo. – 2023.

1 recurso eletrônico (203 f. : il.) : pdf

Orientador: Vinícius Gonçalves Maltarollo.
Coorientadora: Débora Maria Abrantes Costa.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Faculdade de Farmácia, Programa de Pós-Graduação em Ciências Farmacêuticas.

Exigências do sistema: Adobe Acrobat Reader.

1. Aprendizado de máquina – Teses. 2. Simulação de acoplamento molecular – Teses. 3. Antibacterianos – Teses. 4. Desenho de fármacos – Teses. 5. Proteínas recombinantes – Teses. I. Maltarollo, Vinícius Gonçalves. II. Costa, Débora Maria Abrantes. III. Universidade Federal de Minas Gerais. Faculdade de Farmácia. IV. Título.

CDD: 615.4



UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE FARMÁCIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS FARMACÊUTICAS

ATA DA DEFESA DE DISSERTAÇÃO DO ALUNO GABRIEL CORRÊA VERÍSSIMO

Realizou-se, no dia 31 de outubro de 2023, às 14:00 horas, em formato remoto, a 400ª defesa de Dissertação, intitulada INTEGRAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA AO DOCKING MOLECULAR PARA PLANEJAMENTO E REALIZAÇÃO DE ENSAIOS IN VITRO DE INIBIDORES DA ENZIMA ENOIL-ACP-REDUTASE NAD(P)H-DEPENDENTE (FAB) DE STAPHYLOCOCCUS AUREUS E DE ESCHERICHIA COLI, apresentada por GABRIEL CORRÊA VERÍSSIMO, número de Registro 2021667060, graduado no curso de FARMÁCIA, como requisito parcial para a obtenção do grau de Mestre em CIÊNCIAS FARMACÊUTICAS, à seguinte Comissão Examinadora: Prof(a). Vinicius Gonçalves Maltarollo - Orientador (UFMG), Prof(a). Débora Maria Abrantes Costa (UFMG), Prof(a). Jadson Castro Gertrudes (UFOP), Prof(a). Rafaela Salgado Ferreira (UFMG).

A Comissão considerou a Dissertação:

Aprovada

Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.
Belo Horizonte, 31 de outubro de 2023.



Documento assinado eletronicamente por **Rafaela Salgado Ferreira, Professora do Magistério Superior**, em 31/10/2023, às 16:55, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Jadson Castro Gertrudes, Usuário Externo**, em 31/10/2023, às 16:55, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Débora Maria Abrantes Costa, Professora do Magistério Superior**, em 31/10/2023, às 16:56, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Vinicius Goncalves Maltarollo, Professor do Magistério Superior**, em 31/10/2023, às 16:56, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2757437** e o código CRC **FB64EF01**.

AGRADECIMENTOS

Gostaria de agradecer aos meus orientadores, Prof. Dr. Vinícius Gonçalves Maltarollo e Prof^a. Dr^a. Débora Maria Abrantes Costa, por aceitarem a orientação, pela paciência e pelos ensinamentos que com certeza contribuíram muito para minha formação acadêmica e profissional.

Gostaria também de agradecer individualmente a cada um dos meus orientadores. Primeiro, ao professor Vinícius, por ter aceitado me orientar ainda na iniciação científica, me ensinando e guiando na área computacional e sua interface com as ciências biológicas, farmacêuticas e com a química. Também agradeço a paciência e por sempre considerar e acreditar nas minhas escolhas individuais, sendo capaz de construir consensos e buscar parcerias que resultaram nesse projeto de Mestrado. Minha vontade de trabalhar com proteínas recombinantes me levou de encontro à professora Débora, que me aceitou com o presente projeto e sou muito grato por isso. Agradeço à Prof.^a Débora por me orientar em uma área que, apesar do meu grande interesse, eu tinha pouca experiência prévia. Esse trabalho só foi possível graças a sua didática e dedicação em transmitir os conhecimentos e habilidades necessárias para as metodologias desenvolvidas.

Agradeço ao Prof. Dr. Ronaldo Alves Pinto Nagem por me aceitar e receber em seu laboratório, dispor recursos imprescindíveis para a realização da minha pesquisa, discutir resultados obtidos e fornecer contribuições que iluminaram caminhos e enriqueceram a pesquisa.

Agradeço ao Jamil Silvano de Oliveira, técnico de nível superior do departamento de Bioquímica e Imunologia do ICB-UFMG, por me auxiliar na realização de diversos experimentos e permitir o uso de equipamentos que foram essenciais para o andamento da pesquisa.

Agradeço ao Dr. Thales Kronenberger pelas importantes contribuições que enriqueceram a pesquisa e a *Eberhard Karls Universität Tübingen* pelo uso das licenças dos softwares LigPrep e Glide (Schrödinger, Inc.).

Agradeço à UFMG, CAPES, CNPq e FAPEMIG pela estrutura e apoio financeiro. Agradeço, em especial, à UFMG, pela defesa da educação pública, gratuita e de qualidade, mesmo frente aos ataques e cortes orçamentários sofridos durante o desgoverno de extrema-direita que atacou a educação, a ciência e a saúde.

AGRADECIMENTOS (Continuação)

Agradeço à *OpenEye Scientific Software* pelas licenças acadêmicas dos programas OMEGA, QUACPAC e da suíte OEDocking.

Agradeço à Fernanda da Cruz e ao Leonardo Passagli, que além de serem alunos de iniciação científica excelentes que contribuíram enormemente para a realização desse trabalho, são amigos que me apoiaram muito tanto academicamente quanto psicologicamente.

Agradeço ao Mateus Serafim e ao Philipe Fernandes, pela contribuição, explicação, demonstração e realização dos ensaios de concentração inibitória mínima e de diferença da transferência de saturação por ressonância magnética nuclear. Também agradeço a amizade e os artigos que já publicamos juntos.

Agradeço ao Rafael Almeida, um grande amigo que conheci durante o Mestrado e que me ajudou muito na parte de programação. Meus códigos nunca serão os mesmos com o que eu aprendi com você, obrigado pelos ensinamentos e pelo apoio. Falando em apoio, obrigado por sempre me incentivar e acreditar no meu potencial mais do que eu mesmo. E obrigado também por dividir comigo uma das melhores experiências que eu tive no Mestrado, a de apresentar trabalho na reunião anual da SBBq.

Agradeço aos meus amigos Valtair dos Santos Júnior, Patryck Moraes, Pedro Augusto, Diana Oliveira, Roger Ryuler, Eduardo Melos, Rodrigo Martins, Vinícius Peret, Gabriel Lima e demais colegas pela amizade, pelo apoio e pela convivência. Gostaria de agradecer em especial ao Valtair, por me aguentar a tempos, dividirmos artigos juntos e estarmos sempre apoiando um ao outro. Toda vez que leio os agradecimentos da sua dissertação eu fico feliz que você se sinta bem em ser você mesmo e por poder contribuir em algo para isso. Também um agradecimento especial à Diana, Roger e Patryck, por me apoiarem durante toda a parte de expressão e purificação da FabI e por compartilharmos momentos que rimos muito (de alegria ou de nervoso). Por último, mas igualmente importante, um agradecimento especial ao Pedro, um amigo de longa data, que sempre acredita no potencial dos amigos e que também tem um potencial enorme.

AGRADECIMENTOS (Conclusão)

Agradeço à minha mãe e à minha irmã pelo apoio e incentivo de sempre.

Por último, o meu agradecimento pessoal mais importante: ao Lucas, meu companheiro que está do meu lado desde a graduação e quem mais me compreende e me apoia, mesmo nos momentos mais difíceis. Não poderia deixar de agradecer por estar do meu lado durante todo o Mestrado, especialmente porque foi um período inicialmente marcado pelo caos da pandemia de COVID-19 e, posteriormente, pelo processo de reconstrução e retomada da vida que parecia estar em suspenso. Jamais terei como te agradecer por me apoiar e suavizar o sofrimento que foi estar em isolamento social por tanto tempo, especialmente no período que eu adoeci física e mentalmente. Nunca imaginaria dizer que dois álbuns marcaram tanto um período da minha vida e significaram tanto para mim quanto o *“High as Hope”* e, especialmente, o *“Dance Fever”*. Muito bom estar vacinado e poder ouvir esses álbuns com você e sentir um alívio incomensurável.

RESUMO

O desenvolvimento de bactérias multirresistentes é um problema de saúde no mundo todo. A Organização Mundial da Saúde (OMS) classifica as cepas de *Escherichia coli* resistentes a carbapenêmicos e as cepas de *Staphylococcus aureus* resistentes a meticilina ou vancomicina, respectivamente, como nível crítico e alto de prioridade para o desenvolvimento de antibacterianos. Métodos computacionais, como o *docking* molecular, estão acelerando o desenvolvimento de novos antibacterianos. Entretanto, a capacidade preditiva do *docking* pode ser aprimorada com o uso de algoritmos de aprendizado de máquina (ML). Nesse contexto, o objetivo com este trabalho foi de integrar modelos de ML com estudos de *docking* molecular visando melhorar a capacidade preditiva da inibição das enzimas FabI de *S. aureus* e *E. coli*, além de obter as enzimas FabI recombinantes das duas espécies para ensaios *in vitro* utilizando substâncias selecionadas por triagem virtual. Para isso, 2.352 protocolos de *docking* foram validados por *redocking*, *crossdocking* e curva ROC. Onze algoritmos de ML foram utilizados para gerar 220.856.328 modelos de classificação relacionando a atividade inibitória com os *fingerprints* de interação, calculados com base nas poses do *docking*. Os melhores modelos de cada enzima foram validados e selecionados com base em diversas métricas de classificação. Os três melhores modelos de cada proteína foram utilizados em uma triagem virtual para selecionar substâncias que foram testadas contra células de *E. coli* e *S. aureus*. Para obtenção das proteínas recombinantes, células *Escherichia coli* BL21(DE3) foram transformadas por eletroporação com plasmídeos pET28a-saFabI e pET29a-ecFabI, contendo a sequência codificante das proteínas de interesse. As células foram cultivadas a 37 °C em meio Luria-Bertani suplementado com MgSO₄ e canamicina. O IPTG foi adicionado para induzir a expressão por 18 horas a 18 °C. A lise das células foi promovida por sonicação e a purificação das proteínas foi conduzida pelas cromatografias de afinidade, dessalinização e exclusão molecular. Ambas as proteínas foram obtidas com alto grau de pureza em sua forma tetramérica e estáveis. Em relação aos resultados dos modelos de aprendizado de máquina, foi possível obter, entre os três melhores modelos de cada enzima, valores de MCC_{int} variando entre 0,567 e 0,846 e MCC_{ext} variando entre 0,638 e 1,000, com os algoritmos *Support Vector Machine* (SVM) e *Multilayer Perceptron* (MLP) resultando nos melhores valores e superando em diferentes métricas os resultados de *docking*. Esses modelos foram

utilizados em uma triagem virtual, permitindo a obtenção de nove substâncias ativas nos ensaios *in vitro* contra as células de *E. coli* e *S. aureus* e os estudos de STD-NMR serão conduzidos para confirmar a interação dessas substâncias com as proteínas obtidas.

Palavras-chave: aprendizado de máquina; *docking* molecular; planejamento de antibacterianos; expressão e purificação de enzimas; enoil-ACP-redutase (FabI).

ABSTRACT

The emergence of multidrug-resistant bacteria is a severe health problem worldwide. The World Health Organization (WHO) categorizes carbapenem-resistant strains of *Escherichia coli* and methicillin/vancomycin-resistant strains of *Staphylococcus aureus* as critical and high-level priorities for antibacterial drug development. Computational methods, as molecular docking, are accelerating the development of new antibacterials; however, predictability of molecular docking can be improved with machine learning (ML). In this context, the aim with this work is to integrate ML models with molecular docking studies to enhance predictive ability of ligand inhibitory activity against both *S. aureus* and *E. coli* FabI enzymes. Additionally, the aim is also to obtain recombinant FabI enzymes from both species for *in vitro* assays using compounds selected through virtual screening., and to obtain the *S. aureus* e *E. coli* FabI recombinant enzymes for *in vitro* assays with selected compounds based on ML-based virtual screening. Therefore, 2,352 docking protocols were validated using redocking, crossdocking, and ROC curve analyses. Eleven ML algorithms were deployed, generating 220,856,328 classification models correlating inhibitory activity with interaction fingerprints, calculated based on docking poses. We selected the best models for each enzyme using several classification metrics. The top three models for each protein were then utilized in a virtual screening to select substances that were subsequently tested against both *E. coli* and *S. aureus* cells. To obtain the recombinant proteins, we transformed *Escherichia coli* BL21(DE3) cells via electroporation with pET28a-saFabI and pET29a-ecFabI plasmids, each containing the coding sequences for the proteins of interest. These cells were cultured at 37 °C in Luria-Bertani medium supplemented with MgSO₄ and kanamycin. IPTG was used to induce expression for a duration of 18 hours at 18 °C. Cell lysis was carried out by sonication, and protein purification was performed using affinity, desalting, and size exclusion chromatography. Both proteins were obtained in the form of highly pure and stable tetramers. With regard to the machine learning results, the top three models of each enzyme yielded MCC_{int} values ranging from 0.567 to 0.846, and MCC_{ext} values ranging from 0.638 to 1.000. The Support Vector Machine (SVM) and Multilayer Perceptron (MLP) algorithms produced the best results, outperforming docking in all metrics These models were used in a virtual screening, allowing nine active

compounds to be obtained in *in vitro* assays against *E. coli* and *S. aureus* cells. Subsequent STD-NMR studies will be conducted to verify the interaction of these substances with the obtained proteins.

Keywords: machine learning; molecular docking; antibacterial design; expression and purification of enzymes; enoyl-ACP-reductase (FabI).

LISTA DE FIGURAS

Figura 1 – Linha do tempo do desenvolvimento e introdução na clínica de fármacos com atividade antibacteriana (verde) e da resistência antibacteriana (vermelho).	33
Figura 2 – Via metabólica da biossíntese de ácidos graxos de tipo II. Em azul estão as enzimas de <i>S. aureus</i> e <i>E. coli</i> , em verde as isoformas presentes apenas em <i>E. coli</i> e as demais isoformas de outras espécies estão em preto.....	35
Figura 3 – Mecanismo da reação catalisada pela FabI.	36
Figura 4 – Inibidores da FabI, suas fases mais avançadas dos estudos clínicos atualmente e as empresas pesquisadoras responsáveis.....	39
Figura 5 – Representação da transferência de saturação de uma proteína diretamente excitada para um ligante.	48
Figura 6 – Exemplos dos tipos de funções de pontuação.	52
Figura 7 – Representação das duas abordagens de agrupamento hierárquico.	58
Figura 8 – Representação de uma classificação com algoritmo de kNN com $k = 5$	59
Figura 9 – Representação de uma árvore de decisão de classificação, onde a classificação é dada pela maioria dos votos, e de uma árvore de decisão de regressão, onde o valor da variável dependente é dado pela média dos valores em cada folha.	61
Figura 10 – Representação da importância de se considerar outliers na construção hiperplano (linha sólida) e suas margens (linhas pontilhadas) para a classificação de uma amostra desconhecida. Em (A), a margem que minimiza as classificações incorretas no conjunto de treinamento e, em (B), margens mais permissivas que são menos susceptíveis aos outliers e, portanto, melhores em generalização.....	62
Figura 11 – Representação de uma rede neural artificial com seus principais elementos.....	64
Figura 12 – Representação geral da metodologia utilizada no presente trabalho.	66
Figura 13 – Definição das métricas de validação utilizadas.	82
Figura 14 – Mapa do vetor pET28a-saFabI utilizado para a expressão da proteína FabI de <i>S. aureus</i>	88
Figura 15 – Mapa do vetor pET29a-ecFabI utilizado para a expressão da proteína FabI de <i>E. coli</i>	90

Figura 16 – Densidade e distribuição dos valores de RMSD de redocking e crossdocking entre todos os protocolos de acoplamento molecular que retornaram o ligante no sítio ativo. À esquerda os valores para a proteína de <i>S. aureus</i> (PDB ID: 4FS3) e à direita os valores da proteína de <i>E. coli</i> (PDB ID: 1QG6). Em azul, os valores obtidos para as proteínas de encaixe induzido e, em rosa escuro, os valores obtidos para as proteínas do PDB tratadas apenas para remoção de dupla ocupância. A linha vermelha representa $\text{RMSD}_{\text{redocking}} < 2,00 \text{ \AA}$ e $\text{RMSD}_{\text{crossdocking}} < 2,55 \text{ \AA}$	103
Figura 17 – Representação da AFN-1252 (cinza) nos sítios de ligação das enzimas FabI obtidas diretamente do PDB (rosa) e das obtidas pelo docking por encaixe induzido (azul). Em laranja, as colisões que são observadas nas estruturas de proteínas que estavam previamente complexadas com o triclosan (4FS3_IFD e 1QG6).	104
Figura 18 – Curvas ROC para os melhores protocolos de acoplamento molecular de cada proteína.	106
Figura 19 – Representação tridimensional dos resultados de redocking (A e C) e crossdocking (B e D) para as proteínas FabI de <i>S. aureus</i> (A e B) e de <i>E. coli</i> (C e D) com os melhores protocolos. Em cinza, as estruturas cristalográficas e, em rosa envelhecido, as estruturas obtidas pelos protocolos de docking.....	109
Figura 20 – Distribuição do conjunto de dados de inibidores da saFabI (N = 273) entre os subconjuntos e diferentes propriedades.	111
Figura 21 – Distribuição do conjunto de dados de inibidores da ecFabI (N = 140) entre os subconjuntos e diferentes propriedades.	111
Figura 22 – Dispersão dos modelos de aprendizado de máquina para a saFabI em função de seu valor de MCC interno e externo. Em roxo, o modelo com balanço entre maior robustez e preditividade selecionado para cada técnica e, em rosa, os demais modelos.....	115
Figura 23 – Dispersão dos modelos de aprendizado de máquina para a ecFabI em função de seu valor de MCC interno e externo. Em roxo, o modelo com balanço entre maior robustez e preditividade selecionado para cada técnica e, em rosa, os demais modelos.....	117
Figura 24 – Gráficos de radar das métricas de validação interna (rosa) e externa (roxo) calculadas para os cinco melhores modelos de aprendizado de máquina.....	119
Figura 25 – Curvas ROC para os quatro melhores modelos de aprendizado de máquina para a saFabI.	120

Figura 26 – Validação X-scrambling para os modelos de <i>S. aureus</i> (à esquerda) e de <i>E. coli</i> (à direita).	122
Figura 27 – Curvas ROC dos três melhores modelos de aprendizado de máquina para os conjuntos de treinamento e de teste das duas proteínas, comparadas as curvas ROC dos seus respectivos protocolos de docking com e sem decoys.	123
Figura 28 – Gráficos de radar dos três melhores modelos de aprendizado de máquina de cada proteína (saFabI e ecFabI) comparados com os seus respectivos protocolos de docking.	126
Figura 29 – Interpretação dos fingerprints de interação do modelo SA_MLL.	129
Figura 30 – Interpretação dos fingerprints de interação do modelo SA_MLS.	130
Figura 31 – Interpretação dos fingerprints de interação do modelo SA_SVS.	131
Figura 32 – Interpretação dos fingerprints de interação do modelo EC_MLL.	133
Figura 33 – Interpretação dos fingerprints de interação do modelo EC_SVP.	134
Figura 34 – Interpretação dos fingerprints de interação do modelo EC_SVR.	135
Figura 35 – Representação das moléculas em componentes principais obtidas pelos diferentes fingerprints de interação utilizados nos melhores modelos.	139
Figura 36 – Análise do gel SDS-PAGE 12% para o teste de expressão da saFabI após coloração com prata.	147
Figura 37 – Análise do gel SDS-PAGE 12% para o teste de expressão da ecFabI após coloração com Coomassie Blue R250.	147
Figura 38 – Cromatograma de afinidade do segundo teste de purificação com a saFabI.	149
Figura 39 – Análise do gel SDS-PAGE 12% para a cromatografia de afinidade da saFabI após coloração com Coomassie Blue R250.	150
Figura 40 – Cromatograma de exclusão molecular do segundo teste de purificação com a saFabI.	151
Figura 41 – Distribuição de tamanho por massa obtido por DLS para o segundo teste de purificação com a saFabI.	151
Figura 42 – Análise do gel SDS-PAGE 12% para as etapas de lise bacteriana e cromatografias de afinidade, dessalinização e exclusão molecular da saFabI após coloração com Coomassie Blue R250.	152
Figura 43 – Análise do gel SDS-PAGE 12% para as etapas de lise bacteriana e cromatografias de afinidade, dessalinização e exclusão molecular da ecFabI após coloração com Coomassie Blue R250.	153

Figura 44 – Cromatograma de afinidade da purificação final com a saFabI.....	154
Figura 45 – Cromatograma de dessalinização da purificação final com a saFabI..	154
Figura 46 – Cromatograma de exclusão molecular da purificação final com a saFabI.	155
Figura 47 – Distribuição de tamanho por volume obtida por DLS da amostra de saFabI após exclusão molecular.	156
Figura 48 – Cromatograma de afinidade da purificação final com a ecFabI.....	157
Figura 49 – Cromatograma de dessalinização da purificação final com a ecFabI..	157
Figura 50 – Cromatograma de exclusão molecular da purificação final com a ecFabI.	158
Figura 51 – Distribuição de tamanho por volume obtida por DLS da amostra de ecFabI após exclusão molecular.	159
Figura 52 – Modos de ligação (3D à esquerda e 2D à direita) do crizotinibe para ecFabI e saFabI.	163
Figura 53 – Modos de ligação (3D à esquerda e 2D à direita) da doxiciclina para ecFabI e saFabI.	163
Figura 54 – Modos de ligação (3D à esquerda e 2D à direita) do azul de metileno para ecFabI e saFabI.	164
Figura 55 – Modos de ligação (3D à esquerda e 2D à direita) da lincomicina para ecFabI e saFabI.	164
Figura 56 – Modos de ligação (3D à esquerda e 2D à direita) da zidovudina para ecFabI e saFabI.	165
Figura 57 – Modos de ligação (3D à esquerda e 2D à direita) da mefloquina para ecFabI e saFabI.	165
Figura 58 – Modos de ligação (3D à esquerda e 2D à direita) do triclosan para ecFabI e saFabI.	166
Figura 59 – Modos de ligação (3D à esquerda e 2D à direita) de SNL2016_38 para ecFabI e saFabI.	166
Figura 60 – Modos de ligação (3D à esquerda e 2D à direita) de EDNCl para ecFabI e saFabI.	167

LISTA DE FIGURAS DO APÊNDICE A

Figura A.1 – Sequência codificante 5' -> 3' de nucleotídeos da saFabI e a sequência de aminoácidos resultante.....	201
Figura A.2 – Sequência codificante 5' -> 3' de nucleotídeos da ecFabI e a sequência de aminoácidos resultante.....	202

LISTA DE TABELAS

Tabela 1 – Principais sistemas de expressão utilizados na produção de proteínas recombinantes e suas vantagens e desvantagens básicas.	41
Tabela 2 – Exemplos de cepas de E. coli utilizadas na expressão de proteínas recombinantes e suas principais características.	43
Tabela 3 – Número de substâncias ativas, inativas e decoys para cada conjunto de dados.	72
Tabela 4 – Identificação dos receptores gerados em termos da estrutura de proteína utilizada, do ligante originalmente associado à proteína, do volume da caixa e das dimensões da caixa.....	73
Tabela 5 – Parâmetros dos algoritmos de docking variados para geração dos protocolos.....	74
Tabela 6 – Proporção de moléculas em cada subconjunto de cada dataset.	75
Tabela 7 – Parâmetros e valores variados para a geração de fingerprints.	77
Tabela 8 – Hiperparâmetros utilizados nos modelos de aprendizado de máquina. ..	79
Tabela 9 – Descrição das camadas ocultas utilizadas nos modelos de MLP.....	80
Tabela 10 – Propriedades das proteínas recombinantes calculadas no ProtParam.	100
Tabela 11 – Cinco melhores protocolos de docking da proteína FabI de cada espécie e seus parâmetros e valores de RMSD.....	105
Tabela 12 – Valores de AUC _{ROC} , AUC _{BEDROC} (utilizando $\alpha = 160,9, 32,2$ e $16,1$) e fator de enriquecimento (EF) nas frações de 1,0, 5,0 e 10,0%. Em negrito, o protocolo com melhor desempenho de cada proteína.....	107
Tabela 13 – Número de modelos para cada técnica de aprendizado de máquina utilizado.	112
Tabela 14 – Resultados de MCC dos melhores modelos de cada técnica para a saFabI.	114
Tabela 15 – Resultados de MCC dos melhores modelos de cada técnica para a ecFabI.	116
Tabela 16 – Métricas de validação interna e externa calculadas para os cinco melhores modelos de cada proteína.....	118

Tabela 17 – Valores de AUC _{ROC} e AUC _{BEDROC} para os modelos de ML (nos conjuntos de teste e de treinamento) e protocolos de docking (no conjunto de dados total com ou sem decoys).....	123
Tabela 18 – Métricas de classificação para os valores classificados da pontuação do docking da saFabI.....	125
Tabela 19 – Métricas de classificação para os valores classificados da pontuação do docking da ecFabI.....	125
Tabela 20 – Relação entre os melhores modelos e seus respectivos fingerprints..	138
Tabela 21 – Número de moléculas fora do domínio de aplicabilidade (N _{DA}) para cada fingerprint de interação e a porcentagem do conjunto de dados que esse número representa (% _{DA}).....	140
Tabela 22 – Moléculas selecionadas na triagem virtual para a saFabI e sua classificação em ativa (1) e inativa (0) para a inibição enzimática de acordo com as predições dos modelos.....	141
Tabela 23 – Moléculas selecionadas na triagem virtual para a ecFabI e sua classificação em ativa (1) e inativa (0) para a inibição enzimática de acordo com as predições dos modelos.....	143
Tabela 24 – Resumo dos testes de purificação realizados com a saFabI.....	148
Tabela 25 – Resultados do DLS para o pico principal de cada triplicata da saFabI.....	156
Tabela 26 – Resultados do DLS para o pico principal de cada triplicata da ecFabI.....	159
Tabela 27 – MIC (µM) das moléculas contra S. aureus (ATCC 29213) e E. coli (ATCC 35218).....	160

LISTA DE ABREVIATURAS, SIGLAS E UNIDADES DE MEDIDA

(p)ppGpp	guanosina penta ou tetrafosfato
Å	Ångström
A ₂₈₀	Absorbância a 280 nm
ACP	Proteína carreadora de grupos acila
ADMET	Administração, Distribuição, Metabolismo, Excreção e Toxicidade
AI	<i>Artificial</i> intelligence (inteligência artificial)
ANN	<i>Artificial Neural Networks</i> (Redes neurais artificiais)
APS	Persulfato de amônio
AUC	<i>Area under the curve</i> (área sob a curva)
bACC	Acurácia balanceada
BEDROC	<i>Boltzmann-Enhanced Discrimination of ROC curve</i> (discriminação aprimorada de Boltzmann da curva ROC)
BraCoLi	<i>Brazilian Compound Library</i>
CAPES	Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CK	Coeficiente Kappa de Cohen
cm	Centímetros
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
Co	Cobalto
CoA	Coenzima A
crio-EM	Criomicroscopia eletrônica
Cu	Cobre
Da	Dalton
DL	<i>Deep learning</i> (aprendizado profundo)
DLS	<i>Dynamic light scattering</i> (espalhamento dinâmico da luz)
DMSO	Dimetilsulfóxido
DNA	Deoxyribonucleic acid (ácido desoxirribonucleico)
DO ₆₀₀	Densidade óptica a 600 nm
DP	Desvio padrão
DT	<i>Decision Tree</i> (Árvore de Decisão)
DTT	Ditiotreitol

ecFabI	FabI de <i>Escherichia coli</i>
EF	<i>Enrichment Factor</i> (Fator de enriquecimento)
EIFP	<i>Extended Interaction Fingerprint</i>
ESBL	<i>Extended Spectrum Beta-Lactamase</i> (β -lactamase de espectro estendido)
ext	Métrica calculada por validação externa
ϵ	Coeficiente de extinção molar ($M^{-1}cm^{-1}$)
FAFAR	Faculdade de Farmácia
Fak	Complexo de ácido graxo quinase
FAPEMIG	Fundação de Amparo à Pesquisa do Estado de Minas Gerais
FAS	<i>Fatty acid biosynthesis</i> (via de biossíntese de ácidos graxos)
FAS-I	<i>Fatty acid biosynthesis – type I</i> (via de biossíntese de ácidos graxos de tipo II)
FAS-II	<i>Fatty acid biosynthesis – type I</i> (via de biossíntese de ácidos graxos de tipo I)
FDA	<i>United States Food and Drug Administration</i>
FIFP	<i>Functional Interaction Fingerprint</i>
FPR	<i>False positive rate</i> (taxa de falsos positivos)
g	Gramma
<i>g</i>	Força- <i>g</i>
gNB	<i>Gaussian Naive Bayes</i>
GST	Glutathione-S-transferase
GUI	<i>Graphical user interface</i> (interface gráfica do utilizador)
HCA	<i>Hierarchical Clustering Analysis</i> (Análise de Agrupamento Hierárquico)
HIFP	<i>Hybrid Interaction Fingerprint</i>
His ₆	Hexahistidina
HQSAR	<i>Hologram Quantitative Structure-Activity Relationship</i>
IC ₅₀	Concentração inibitória média
ICB	Instituto de Ciências Biológicas
ID	Identificador
IFD	Encaixe induzido
InChi	Identificador químico internacional

int	Métrica calculada por validação interna <i>5-fold</i>
IPTG	Isopropil β -D-1-tiogalactopiranosídeo
KanR	Gene de resistência à canamicina
kDa	Kilodalton
kNN	<i>k-Nearest Neighbors</i>
L	Litro
LAREMAR	Laboratório de Ressonância Magnética de Alta Resolução
LB	Luria-Bertani
LBDD	<i>Ligand-based drug design</i> (planejamento de fármacos baseado na estrutura dos ligantes)
LBFSGS	<i>Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm</i>
log P	Logaritmo negativo do coeficiente de partição
LTS	<i>Long-term support</i>
M	Concentração molar (mol por litro)
MacroMol	Laboratório de macromoléculas
MALDI-TOF	<i>Matrix-assisted laser desorption ionization–time-of-flight mass spectrometry</i> (espectrometria de massas por dessorção ionizante assistida por uma matriz com a medida do tempo de voo)
MBP	<i>Maltose-binding protein</i> (proteína de ligação à maltose)
MCC	Coeficiente de correlação de Matthews
mg	Miligrama
MIC	<i>Minimal inhibitory concentration</i> (concentração inibitória mínima)
ML	<i>Machine learning</i> (aprendizado de máquina)
mL	Mililitro
MLA	<i>Multilayer Perceptron – Adam solver</i>
MLL	<i>Multilayer Perceptron – Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (LBFSGS) solver</i>
MLP	<i>Multilayer Perceptron</i>
MLS	<i>Multilayer Perceptron – Stochastic Gradient Descent (SGD) solver</i>
mM	Concentração milimolar (milimol por litro)
MRSA	<i>Methicillin-resistant Staphylococcus aureus</i> (<i>Staphylococcus aureus</i> resistente à meticilina)
MTA	<i>Material transfer agreement</i> (acordo de transferência de material)

NAD	Dinucleotídeo de nicotinamida e adenina
NADP	Fosfato de dinucleotídeo de nicotinamida e adenina
NB	<i>Naive Bayes</i>
N _{DA}	Número de moléculas fora do domínio de aplicabilidade
ng	Nanograma
Ni	Níquel
nM	Concentração nanomolar (nanomol por litro)
NMR	<i>Nuclear magnetic resonance</i> (ressonância magnética nuclear)
OMS	Organização Mundial da Saúde
P&D	Pesquisa e desenvolvimento
PAGE	<i>Polyacrylamide gel electrophoresis</i> (eletroforese em gel de poliacrilamida).
PCA	<i>Principal Component Analysis</i> (Análise de Componente Principal)
PDB	<i>Protein Data Bank</i>
PDB ID	Código identificador do <i>Protein Data Bank</i>
PGFS	<i>Policy Gradient for Forward Synthesis</i>
pH	Potencial hidrogeniônico
pKa	Logaritmo negativo da constante de acidez
QM	<i>Quantum mechanics</i> (mecânica quântica)
QSAR	<i>Quantitative structure-activity relationship</i> (relação quantitativa da estrutura e atividade)
RBF	<i>Radial Basis Function</i>
RF	<i>Random Forest</i>
RMN	Ressonância Magnética Nuclear
RMSD	<i>Root mean squared deviation</i> (raiz quadrada do desvio médio quadrático)
RNA	<i>Ribonucleic acid</i> (ácido ribonucleico)
ROC	<i>Receiver operator characteristic</i> (característica de operação do receptor)
rpm	Rotações por minuto
saFabI	FabI de <i>Staphylococcus aureus</i>
SBDD	<i>Structure-based drug design</i> (planejamento de fármacos baseado na estrutura do alvo molecular)

SBL	<i>Substrate binding loop</i> (loop de ligação do substrato)
sdf	<i>Structure-data file</i>
SDS	<i>Sodium dodecyl sulfate</i> (dodecil sulfato de sódio)
SGD	<i>Stochastic Gradient Descent</i>
SPXY	<i>Sample set partitioning based on joint X-y distance</i> (partição do conjunto de amostras baseado na distância conjunta X-y)
STD	<i>Saturation transfer difference</i> (diferença da transferência de saturação)
STD-NMR	<i>Saturation transfer difference nuclear magnetic resonance</i> (diferença da transferência de saturação por ressonância magnética nuclear)
SVL	Máquina de Vetores de Suporte com <i>kernel</i> linear
SVM	<i>Support Vector Machine</i> (Máquina de Vetores de Suporte)
SVP	Máquina de Vetores de Suporte com <i>kernel</i> polinomial
SVR	Máquina de Vetores de Suporte com <i>kernel Radial Basis Function (RBF)</i>
SVS	Máquina de Vetores de Suporte com <i>kernel</i> sigmoide
TCEP	Tris(2-carboxietil)fosfina
TCL	Triclosan
TEMED	Tetrametiletilenodiamina
TEV	Tobacco Etch Virus (vírus do mosaico do tabaco)
TNR	<i>True positive rate</i> (taxa de verdadeiros negativos ou especificidade)
TPR	<i>True positive rate</i> (taxa de verdadeiros positivos ou sensibilidade)
UFC	Unidades formadoras de colônia
UFMG	Universidade Federal de Minas Gerais
V	Volts
VISA	<i>Vancomycin-intermediate Staphylococcus aureus</i> (<i>Staphylococcus aureus</i> de sensibilidade intermediária à vancomicina)
VRSA	<i>Vancomycin-resistant Staphylococcus aureus</i> (<i>Staphylococcus aureus</i> resistente à vancomicina)
WSL	<i>Windows Subsystem for Linux</i> (Subsistema do Windows para Linux)
Zn	Zinco
ΔG	Energia livre de Gibbs

μ	Micro
μg	Micrograma
μL	Microlitro
μM	Concentração micromolar (micromol por litro)

Abreviações (em uma letra e três letras) de resíduos de aminoácidos

A	Ala	Alanina	M	Met	Metionina
C	Cys	Cisteína	N	Asn	Asparagina
D	Asp	Aspartato	P	Pro	Prolina
E	Glu	Glutamato	Q	Gln	Glutamina
F	Phe	Fenilalanina	R	Arg	Arginina
G	Gly	Glicina	S	Ser	Serina
H	His	Histidina	T	Thr	Treonina
I	Ile	Isoleucina	V	Val	Valina
K	Lys	Lisina	W	Trp	Triptofano
L	Leu	Leucina	Y	Tyr	Tirosina

SUMÁRIO

1 INTRODUÇÃO	27
2 OBJETIVOS.....	29
2.1 OBJETIVO GERAL	29
2.2 OBJETIVOS ESPECÍFICOS	29
3 REVISÃO BIBLIOGRÁFICA.....	31
3.1 O PAPEL DA FABI NA BIOSÍNTESE DE ÁCIDOS GRAXOS DE TIPO II (FAS-II) E SUA RELEVÂNCIA NO CENÁRIO DE RESISTÊNCIA ANTIBACTERIANA	31
3.2 EXPRESSÃO, PURIFICAÇÃO E CARACTERIZAÇÃO DE PROTEÍNAS-ALVO NO CONTEXTO DO DESENVOLVIMENTO DE INIBIDORES DA FABI.....	40
3.3 ESTUDOS DE ACOPLAMENTO MOLECULAR.....	49
3.3.1 Estratégias de validação do acoplamento molecular	52
3.4 ALGORITMOS DE APRENDIZADO DE MÁQUINA NO DESENVOLVIMENTO DE FÁRMACOS	55
4 MATERIAIS E MÉTODOS	65
4.1 ESQUEMA GERAL DO TRABALHO	65
4.2 COMPUTADORES E PROGRAMAS	67
4.3 SELEÇÃO E PRÉ-TRATAMENTO DOS ALVOS MOLECULARES PARA OS ESTUDOS DE ACOPLAMENTO MOLECULAR.....	68
4.4 SELEÇÃO E PRÉ-TRATAMENTO DOS CONJUNTOS DE LIGANTES PARA OS ESTUDOS DE ACOPLAMENTO MOLECULAR.....	70
4.5 SELEÇÃO E VALIDAÇÃO DOS PROTOCOLOS DE ACOPLAMENTO MOLECULAR.....	72
4.6 SEPARAÇÃO DO CONJUNTO DE DADOS EM SUBCONJUNTOS DE TREINAMENTO E DE TESTE PARA OS ALGORITMOS DE APRENDIZADO DE MÁQUINA.....	75
4.7 CÁLCULO DOS <i>FINGERPRINTS</i> DE INTERAÇÃO	76

4.8 CONSTRUÇÃO E VALIDAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA COM OS <i>FINGERPRINTS</i> DE INTERAÇÃO DAS POSES DO ACOPLAMENTO MOLECULAR.....	77
4.9 INTERPRETAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA E DAS PRINCIPAIS CARACTERÍSTICAS DE INTERAÇÃO LIGANTE-PROTEÍNA	84
4.10 TRIAGEM VIRTUAL DA BIBLIOTECA DE FÁRMACOS APROVADOS PELO FDA E DA BRACOLI	85
4.11 AVALIAÇÃO DO DOMÍNIO DE APLICABILIDADE PARA O APRENDIZADO DE MÁQUINA.....	86
4.12 OBTENÇÃO DOS VETORES DE EXPRESSÃO DAS PROTEÍNAS RECOMBINANTES	87
4.12.1 Obtenção do vetor da FabI de <i>Staphylococcus aureus</i>	87
4.12.2 Obtenção do vetor da FabI de <i>Escherichia coli</i>	90
4.13 DOSAGEM DE DNA DOS VETORES	91
4.14 PRODUÇÃO DE CÉLULAS ELETROCOMPETENTES	91
4.15 TRANSFORMAÇÃO DAS CÉLULAS DE <i>E. COLI</i> E PREPARO DO BANCO DE CÉLULAS	92
4.16 EXPRESSÃO DAS PROTEÍNAS RECOMBINANTES	93
4.17 LISE BACTERIANA.....	94
4.18 IDENTIFICAÇÃO DE PROTEÍNAS RECOMBINANTES POR SDS-PAGE	95
4.19 PURIFICAÇÃO DAS PROTEÍNAS RECOMBINANTES	96
4.20 DETERMINAÇÃO DO TEOR DE PROTEÍNAS E CÁLCULO DO RENDIMENTO FINAL	99
4.21 CARACTERIZAÇÃO DO ESTADO OLIGOMÉRICO DE PROTEÍNAS POR ESPALHAMENTO DINÂMICO DA LUZ.....	100
4.22 ENSAIOS DE CONCENTRAÇÃO INIBITÓRIA MÍNIMA.....	100
5 RESULTADOS E DISCUSSÃO.....	102
5.1 VALIDAÇÃO DOS PROTOCOLOS DE ACOPLAMENTO MOLECULAR.....	102

5.2 SEPARAÇÃO DO CONJUNTO DE DADOS EM CONJUNTOS DE TREINAMENTO E DE TESTE	109
5.3 CONSTRUÇÃO E VALIDAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA COM OS <i>FINGERPRINTS</i> DE INTERAÇÃO	112
5.4 INTERPRETAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA	127
5.5 AVALIAÇÃO DO DOMÍNIO DE APLICABILIDADE E TRIAGEM VIRTUAL DE POTENCIAIS INIBIDORES DA FABI	138
5.6 EXPRESSÃO, PURIFICAÇÃO E CARACTERIZAÇÃO DO ESTADO OLIGOMÉRICO DAS PROTEÍNAS FABI RECOMBINANTES.....	146
5.7 ENSAIOS DE CONCENTRAÇÃO INIBITÓRIA MÍNIMA COM OS LIGANTES PROMISSORES DA TRIAGEM VIRTUAL.....	160
6 CONCLUSÃO	169
REFERÊNCIAS.....	171
APÊNDICES	201
APÊNDICE A – SEQUÊNCIAS DE NUCLEOTÍDEOS DAS PROTEÍNAS RECOMBINANTES E AS SEQUÊNCIAS DE AMINOÁCIDOS RESULTANTES	201
APÊNDICE B – PROCEDIMENTO OPERACIONAL PADRÃO: METODOLOGIA PARA COLORAÇÃO DE GEL SDS-PAGE POR NITRATO DE PRATA.....	203

1 INTRODUÇÃO

O aumento da resistência à antibacterianos entre patógenos de importância clínica é um problema e uma preocupação global. Em 2019, foram estimados 1,27 milhões de óbitos atribuíveis a infecção por cepas resistentes a antimicrobianos em todo o mundo. Dentre os microrganismos causadores, *E. coli* e *S. aureus* foram, respectivamente, o primeiro e o segundo lugar entre os óbitos associados com resistência bacteriana neste ano. O impacto das cepas resistentes na saúde pública fez com que, em 2017, a Organização Mundial de Saúde (OMS) estabelecesse níveis de prioridade para o desenvolvimento de novos antibacterianos e os dois patógenos foram classificados respectivamente nos níveis crítico e alto (Organização Mundial de Saúde, 2017; Murray *et al.*, 2022). Além da mortalidade, essas duas espécies de microrganismos ainda desempenham um papel significativo em outras condições patológicas que, quando não levam ao óbito, impactam na qualidade de vida dos pacientes, por exemplo: intoxicação alimentar e infecções transmitidas por alimentos, infecções oculares, de pele e de trato urinário, pneumonia, meningite, endocardite, osteomielite e a septicemia (Szweda *et al.*, 2012; Sarowska *et al.*, 2019).

Alvos moleculares adequados contra microrganismos patogênicos envolvem enzimas ou vias bioquímicas essenciais ao ciclo de vida e/ou infecção do patógeno e que não estão presentes no hospedeiro, visando evitar efeitos adversos. Nesse sentido, a enoil-ACP redutase NAD(P)H-dependente (FabI) que catalisa a etapa limitante da via de biossíntese de ácidos graxos de tipo II (FAS-II, do inglês, *type II fatty acid biosynthesis*) constitui uma interessante fonte de alvos moleculares, uma vez que a via de tipo I é presente apenas em mamíferos e é totalmente mediada por um único polipeptídeo multifuncional que catalisa todas as etapas, enquanto que, na via de tipo II, cada enzima é responsável por uma etapa (Payne *et al.*, 2001; Nguyen *et al.*, 2013; Schiebel *et al.*, 2015; Kronenberger *et al.*, 2017; Dodge *et al.*, 2019; Chen *et al.*, 2022).

Nesse contexto, o conhecimento de alvos moleculares pode ser considerado um ponto crucial para o planejamento e descoberta de fármacos. A triagem e a identificação *in vitro* de antibacterianos promissores é fundamentada em três estratégias principais: (I) ensaios baseados no alvo molecular, (II) ensaios baseados na célula (também conhecidos como ensaios no organismo completo) e (III) ensaios baseados no alvo molecular e na célula (também conhecidos como ensaios fenotípicos) (Landeta; Mejia-Santana, 2022). Nesse panorama, métodos

computacionais têm sido cada vez mais empregados no planejamento de fármacos para reduzir custos, acelerar o processo, aumentar as taxas de sucesso e, principalmente, permitir uma redução no número de substâncias testadas, por fornecer informações que possibilitam uma filtragem prévia (triagem virtual) das substâncias mais promissoras. (Wilson; Lill, 2011; Shen *et al.*, 2019).

O *docking* molecular, ou acoplamento molecular, é uma técnica computacional que permite encontrar a combinação mais provável entre a conformação e a orientação de um possível ligante no sítio de ligação e, ainda, calcular uma pontuação que estima a afinidade de ligação entre o ligante e seu alvo, permitindo hierarquizar ligantes em função de suas pontuações (Wilson; Lill, 2011; Shen *et al.*, 2019). Ainda assim, a pontuação do *docking* é falha em estimar as reais energias de ligação e, frequentemente, protocolos de *docking* podem se tornar impróprios para uso em triagens virtuais devido à alta taxa de falsos positivos e falsos negativos (Rastelli *et al.*, 2009; Fan; Fu; Zhang, 2019). Por isso, diferentes estratégias empregando algoritmos de *machine learning* (ML, do inglês, aprendizado de máquina) estão sendo cada vez mais utilizadas para melhorar a acurácia em estudos de triagem virtual (Crampon *et al.*, 2022).

Neste contexto, foram realizados e validados estudos de *docking* molecular com a FabI de *E. coli* e de *S. aureus* integrados a diferentes algoritmos de ML para predição da inibição das duas enzimas (ecFabI e saFabI). Os modelos de ML-*docking* foram então empregados para a realização de uma triagem virtual de substâncias nas bibliotecas BraCoLi e na biblioteca de fármacos aprovados pelo FDA (*United States Food and Drug Administration*). As substâncias mais promissoras, de acordo com a triagem virtual, foram testadas nas células em ensaios *in vitro* de concentração inibitória mínima (MIC, do inglês, *minimal inhibitory concentration*). Ademais, as duas enzimas recombinantes foram obtidas por expressão em *Escherichia coli*, purificadas e caracterizadas e, juntamente com as substâncias mais promissoras nos ensaios de MIC, foram encaminhadas para ensaios de diferença da transferência de saturação por ressonância magnética nuclear (STD-NMR) para confirmação da interação entre os ligantes e as proteínas.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Aplicar algoritmos de aprendizado de máquina ao *docking* molecular, visando a realização de uma triagem virtual de substâncias e identificação das mais promissoras para avaliação da atividade antibacteriana e para comprovar a interação da enzima com os ligantes selecionados.

2.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos do trabalho compreendem:

- validar protocolos de *docking* molecular utilizando *redocking* e *crossdocking*, visando a escolha dos melhores protocolos com base nos valores de RMSD;
- construir uma biblioteca de ligantes com atividade inibitória experimentalmente definida para as enzimas FabI de *S. aureus* e de *E. coli*;
- definir o melhor protocolo de *docking* para cada enzima pela análise dos valores de RMSD e do comportamento da curva-ROC construída com a biblioteca de ligantes já ancorados;
- utilizar os complexos ligante-proteína obtidos pelo *docking* molecular para gerar *fingerprints* de interação para cada um dos ligantes da biblioteca construída;
- implementar e treinar algoritmos de aprendizado de máquina com base nos *fingerprints* de interação, gerando modelos para a predição da atividade inibitória da FabI de *S. aureus* e de *E. coli*;
- aplicar os modelos desenvolvidos em bibliotecas virtuais de substâncias químicas (quimiotecas), visando a identificação de substâncias promissoras para a inibição da FabI de *S. aureus* e de *E. coli*;
- expressar e purificar a proteína FabI recombinante de *S. aureus* e de *E. coli*;
- determinar o estado oligomérico da enzima em solução;
- determinar a concentração inibitória mínima (MIC) das substâncias selecionadas pela triagem virtual frente cepas de *S. aureus* e de *E. coli* para avaliar a atividade antibacteriana;

- confirmar o mecanismo de inibição das proteínas FabI recombinantes por ensaios de diferença da transferência de saturação por ressonância magnética nuclear (STD-NMR) com as substâncias com atividade antibacteriana determinada por MIC.

3 REVISÃO BIBLIOGRÁFICA

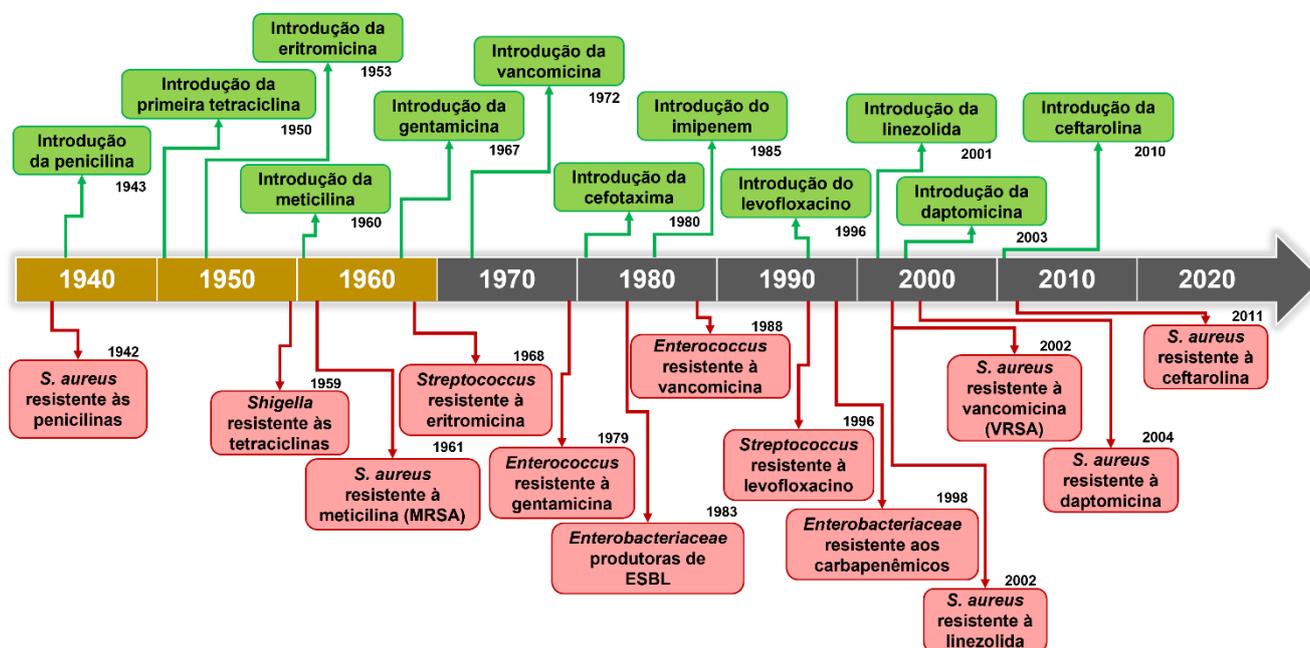
3.1 O PAPEL DA FabI NA BIOSÍNTESE DE ÁCIDOS GRAXOS DE TIPO II (FAS-II) E SUA RELEVÂNCIA NO CENÁRIO DE RESISTÊNCIA ANTIBACTERIANA

O crescimento da resistência à antibacterianos entre patógenos de importância clínica é um problema e uma preocupação global. Em 2017, a Organização Mundial de Saúde (OMS) publicou uma lista de bactérias para as quais o desenvolvimento de novos antibacterianos é urgentemente necessário. Essa lista foi dividida em três níveis de prioridade: crítico, alto e médio, incluindo, por exemplo, bactérias da família *Enterobacteriaceae* (como a *Escherichia coli*) resistentes aos antimicrobianos carbapenêmicos ou produtoras de β -lactamase de espectro estendido (ESBL) no nível crítico e, no nível de alta prioridade, as cepas de *Staphylococcus aureus* resistentes à meticilina (MRSA), à vancomicina (VRSA) ou de sensibilidade intermediária à vancomicina (VISA) (Organização Mundial de Saúde, 2017). Todas essas cepas são consideradas multirresistentes, uma vez que elas apresentam resistência aos antimicrobianos mencionados, que são o fator principal que as caracterizam, e a uma ampla gama de outros antimicrobianos pelo acúmulo de genes de resistência ao longo do tempo, especialmente em meios com alta pressão seletiva (Dunn; Connor; McNally, 2019; Chen; Huang; Shie, 2020).

Historicamente, a introdução de um novo antimicrobiano na clínica é rapidamente seguida pelo surgimento de um mecanismo de resistência específico (**Figura 1**), isso mesmo durante a era de ouro dos antibacterianos (1940-1962), onde uma série de novos fármacos de classes químicas diferentes chegaram rapidamente na clínica (Morehead; Scarbrough, 2018; Lai *et al.*, 2022). Descoberta em 1928 e produzida em escala industrial a partir de 1943, a penicilina era o antimicrobiano de escolha com alta eficácia para o tratamento de infecções por *S. aureus*. Entretanto, em 1942, antes mesmo de seu amplo uso clínico, as cepas produtoras de β -lactamases emergiram e promoveram uma pandemia durante década de 1950 (DeLeo; Chambers, 2009; Morehead; Scarbrough, 2018; Lai *et al.*, 2022). A primeira tetraciclina foi aprovada pelo FDA em 1950, mas em 1959 as primeiras cepas resistentes foram identificadas, bactérias do gênero *Shigella* pertencentes à família *Enterobacteriaceae* (Morehead; Scarbrough, 2018; Zainab *et al.*, 2020). Em 1953, a eritromicina (um macrolídeo) foi aprovada para uso e, em 1968, foram encontradas

bactérias do gênero *Streptococcus* resistentes a este fármaco. Em 1960, foram inseridas as penicilinas resistentes à β -lactamase, como a meticilina e a oxacilina, mas casos de infecção por *Staphylococcus aureus* resistentes foram reportados um ano após (Jensen; Lyon, 2009; Kronenberger *et al.*, 2017). Em 1967, a gentamicina (um aminoglicosídeo) foi introduzida, seguida pelo aparecimento de cepas resistentes de bactérias do gênero *Enterococcus* em 1979. Em 1972, a vancomicina (um glicopeptídeo) foi introduzida, sendo que, em 1988, as bactérias do gênero *Enterococcus* foram novamente as primeiras a apresentarem resistência. Em 2002, foram identificadas as primeiras cepas de *Staphylococcus aureus* resistentes (VRSA) ou de sensibilidade intermediária à vancomicina (VISA) (Morehead; Scarbrough, 2018). Em 1980, a cefotaxima, uma cefalosporina (assim como a ceftazidima e a ceftriaxona), foi introduzida e três anos depois apareceram as primeiras bactérias da família *Enterobacteriaceae* resistentes, por apresentarem o mecanismo da β -lactamase de espectro estendido (ESBL) (Lai *et al.*, 2022). Em 1985, o imipenem foi introduzido e, novamente, bactérias da família *Enterobacteriaceae* resistentes aos antimicrobianos carbapenêmicos foram identificadas pela primeira vez no ano 1998 (Morehead; Scarbrough, 2018). Outros exemplos incluem o levofloxacino que foi introduzido em 1996 e cepas resistentes apareceram no mesmo ano, e três antibacterianos que foram introduzidos nos anos de 2001 (linezolida, uma oxazolidinona), 2003 (daptomicina, um lipopeptídeo) e 2010 (ceftarolina, uma cefalosporina), todos esses levaram a identificação de cepas resistentes de *Staphylococcus aureus* no ano seguinte a sua introdução na clínica (Morehead; Scarbrough, 2018; Lai *et al.*, 2022).

Figura 1 – Linha do tempo do desenvolvimento e introdução na clínica de fármacos com atividade antibacteriana (verde) e da resistência antibacteriana (vermelho).



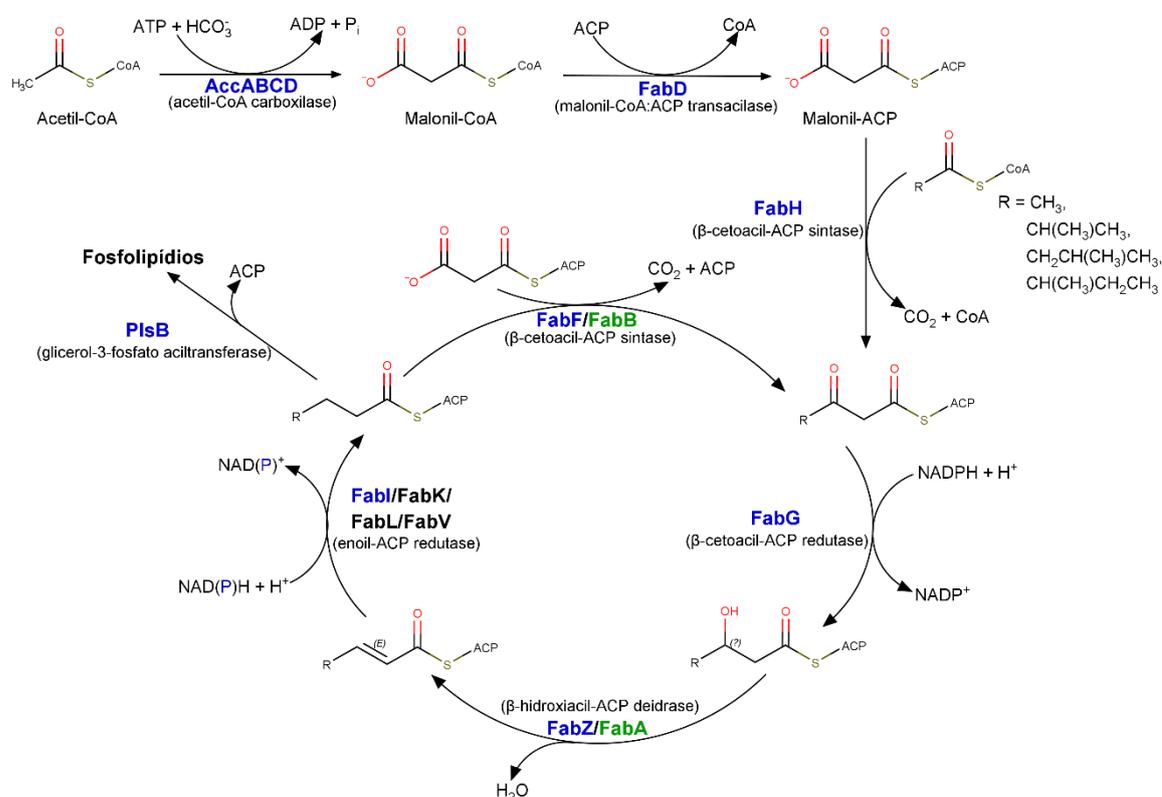
Fonte: Autoria própria.

O uso de antimicrobianos na criação de animais para consumo de carne representa 73% do uso de total desses fármacos, sendo que as classes com maiores taxas de resistência neste cenário são as tetraciclina, sulfonamidas e penicilinas (Van Boeckel *et al.*, 2019). O uso indiscriminado desses fármacos na pecuária, juntamente com o uso irracional de antimicrobianos na terapêutica, contribuíram significativamente para o aumento da resistência bacteriana (Franco *et al.*, 2009; Van Boeckel *et al.*, 2019). Em 2019, foram estimados que, em todo o mundo, 4,95 milhões de óbitos estavam de alguma forma associados com a infecção por cepas resistentes a antimicrobianos, incluindo 1,27 milhões de óbitos atribuíveis a esta causa. Dentre os microrganismos causadores, *E. coli* e *S. aureus* foram, respectivamente, o primeiro e o segundo lugar entre os óbitos associados com resistência bacteriana neste ano, destacando a importância do desenvolvimento de antibacterianos que sejam efetivos contra esses microrganismos (Murray *et al.*, 2022). Além da mortalidade, esses microrganismos ainda desempenham um papel significativo em outras condições patológicas que, quando não levam ao óbito, impactam na qualidade de vida dos pacientes, por exemplo: intoxicação alimentar e infecções transmitidas por alimentos, infecções oculares, de pele e de trato urinário, pneumonia, meningite, endocardite, osteomielite e a septicemia (Szweda *et al.*, 2012; Sarowska *et al.*, 2019).

O rápido aparecimento de cepas multirresistentes e sua relevância clínica em diversas patologias, demonstra a necessidade de busca por novos alvos moleculares para o desenvolvimento de novos antimicrobianos (Jensen; Lyon, 2009; Kronenberger *et al.*, 2017). No passado, esse processo de desenvolvimento de novos antimicrobianos tinha apenas dois alicerces principais: a identificação de produtos naturais ou semissintéticos com atividade antibacteriana ou a utilização de modificações moleculares que visassem aprimorar as classes de fármacos já existentes. Entretanto, o avanço das técnicas de sequenciamento e o desenvolvimento da bioinformática revolucionaram o processo de descoberta de novos antimicrobianos e, na atualidade, é possível varrer, por completo ou em partes, o genoma bacteriano em busca por prováveis novos alvos moleculares (Payne *et al.*, 2001).

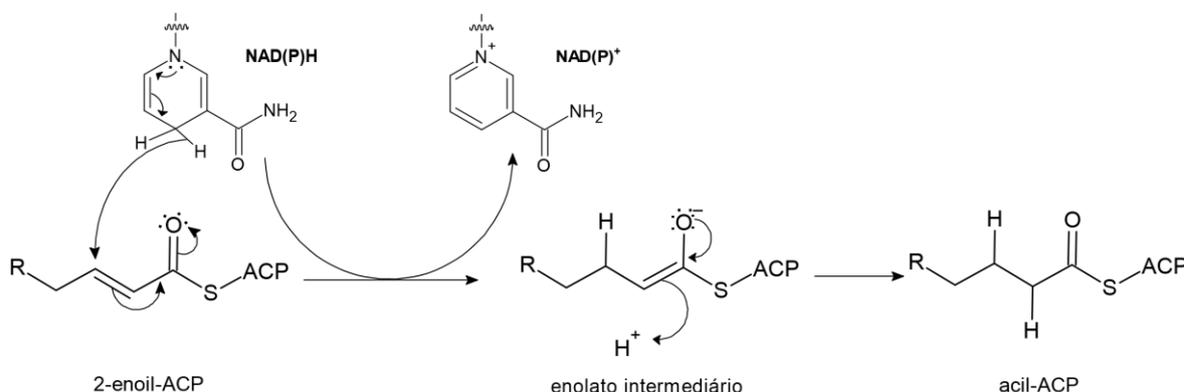
Bons alvos moleculares contra microrganismos patogênicos envolvem enzimas ou vias bioquímicas essenciais ao ciclo de vida e/ou infecção do patógeno e que não estão presentes no hospedeiro, visando evitar efeitos adversos. Nesse sentido, a via de biossíntese de ácidos graxos de tipo II (FAS-II, do inglês, *type II fatty acid biosynthesis*) constitui uma interessante fonte de alvos moleculares (**Figura 2**) (Nguyen *et al.*, 2013; Schiebel *et al.*, 2015; Kronenberger *et al.*, 2017; Dodge *et al.*, 2019; Chen *et al.*, 2022). Enquanto a biossíntese de ácidos graxos em bactérias, plantas e protozoários ocorre via FAS-II, na qual cada reação tem uma enzima individualmente responsável, em mamíferos essa biossíntese ocorre via FAS-I, mediada por um único polipeptídeo que constitui um complexo enzimático multifuncional de vários domínios que catalisa todas as reações envolvidas nesta via. Além disso, os dois sistemas FAS possuem baixa similaridade de sequências entre si (Payne *et al.*, 2001).

Figura 2 – Via metabólica da biossíntese de ácidos graxos de tipo II. Em azul estão as enzimas de *S. aureus* e *E. coli*, em verde as isoformas presentes apenas em *E. coli* e as demais isoformas de outras espécies estão em preto.



Fonte: Adaptado de Schiebel et al., 2012.

Os inibidores da enoil-ACP redutase NAD(P)H-dependente (FabI) vêm sendo estudados como antimicrobianos de alternativa para o tratamento de infecções por cepas multirresistentes. Essa utilização tem sido sustentada por ensaios *in vivo* com roedores e ensaios clínicos (Payne *et al.*, 2002; Park *et al.*, 2007b; Balemans *et al.*, 2010; Escaich *et al.*, 2011; Hafkin; Kaplan; Murphy, 2016; Bassetti *et al.*, 2020; Parker *et al.*, 2022). A FabI é uma enzima essencial para a via FAS-II, por catalisar a etapa final de redução da ligação dupla carbono-carbono em uma porção enoila covalentemente ligada a uma ACP (**Figura 3**). Nessa reação, algumas espécies apresentam preferências diferentes para a coenzima envolvida na redução, por exemplo, a FabI de *S. aureus* é NADPH-dependente, enquanto que a FabI de *E. coli* é NADH-dependente (Schiebel *et al.*, 2015). No ambiente de infecção, bactérias que não consigam incorporar ácidos graxos do hospedeiro se tornam, sob inibição da FabI, incapazes de sintetizar seus ácidos graxos necessários, promovendo uma redução significativa do crescimento bacteriano (Parsons; Rock, 2011; Frank *et al.*, 2020).

Figura 3 – Mecanismo da reação catalisada pela FabI.

Fonte: Autoria própria.

Apesar dos ácidos graxos serem essenciais para as bactérias sintetizarem seus fosfolípidios, lipoproteínas, lipopolissacarídeos e ácidos micólicos, não são todas as espécies de bactérias que conseguem utilizar os ácidos graxos produzidos pelo sistema FAS de seu hospedeiro, necessitando que a própria bactéria realize a síntese “*de novo*” (Lu; Tonge, 2008; Kronenberger *et al.*, 2017). Recentemente, foram identificados mecanismos de mutação e superexpressão da FabI e de *bypass* da via FAS-II em isolados de humanos e animais (via metabólica alternativa). Esses mecanismos têm sido objeto de estudo devido à preocupação com o potencial surgimento de cepas resistentes aos inibidores da via FAS-II (Wang *et al.*, 2023).

Entre as bactérias Gram-negativas, a síntese de FAS-II continua sendo essencial, tendo em vista que dependem de ácidos graxos β -hidroxilados para formar a estrutura do lipídio A, um lipopolissacarídeo da parede celular. Isso ocorre porque, em geral, não há mecanismos capazes de transferir cadeias de acila da coenzima A (CoA) para a proteína carreadora de grupos acila (ACP) de FAS-II, portanto, o grupo hidroxila não pode ser introduzido. Além disso, a suplementação do meio com ácidos graxos β -hidroxilados também é ineficaz, uma vez que as aciltransferases envolvidas na biossíntese do lipídio A só utilizam substratos tioésteres de ACP (Parsons; Rock, 2011).

Em contraste com as demais Gram-negativas, as bactérias da espécie *Neisseria gonorrhoeae* não demandam o lipídio A para sobrevivência e apresentam uma acil-ACP sintase (AasN) que viabiliza a ativação de ácidos graxos extracelulares e sua conversão em derivados acil-ACP e posterior alongamento da cadeia.

Entretanto, um ensaio demonstrou que a inibição da FabI levou a inibição da incorporação dos ácidos graxos exógenos e do crescimento bacteriano, validando a FabI de *Neisseria* como um alvo molecular. O mecanismo dessa inibição é desconhecido em *Neisseria*, mas acredita-se que possa ocorrer por um desequilíbrio do metabolismo de ácidos graxos devido a depleção dos níveis celulares de ACP, assim como ocorre em algumas Gram-positivas como *S. aureus* (Yao *et al.*, 2016). Em *E. coli*, de fato não há nenhum mecanismo conhecido capaz de converter ácidos graxos exógenos em ácidos graxos β -hidroxilados para incorporação em seus lipídios (Yao; Rock, 2017). Ainda assim, foram relatados na literatura casos de redução da sensibilidade aos inibidores por mutação da FabI ou por sua superexpressão (Wang *et al.*, 2023). Por fim, a eficácia do tratamento com inibidores da FabI nas demais Gram-negativas é sustentada por estudos *in vivo* em modelo de coxa neutropênica, com cepas resistentes de *Acinetobacter baumannii* (Gram-negativa), e em modelo de infecção urinária, com *E. coli* resistente ao carbapenem. Nesses modelos, o tratamento com um novo inibidor da FabI reduziu significativamente unidades formadoras de colônia (UFC) em diferentes tecidos (Parker *et al.*, 2022).

Apesar dos resultados observados em bactérias Gram-negativas, a dependência de FAS-II em bactérias Gram-positivas é bastante controversa e mecanismos de mutação e, principalmente, de *bypass* são relatados (Morvan *et al.*, 2017). Em algumas bactérias da ordem Lactobacillales, como a *Streptococcus agalactiae*, a suplementação do meio com ácidos graxos levou a uma regulação negativa de FAS-II e, mesmo na presença de inibidores dessa via, permitiu o crescimento celular em níveis comparáveis ao controle positivo (Balemans *et al.*, 2010).

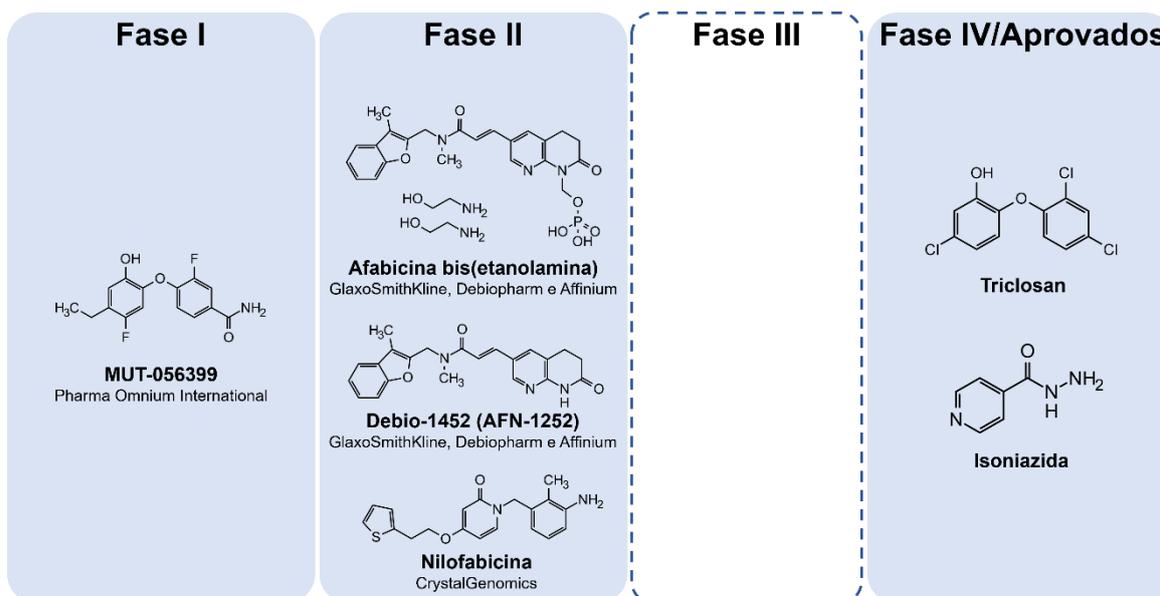
S. aureus também apresenta um mecanismo de *bypass* que envolve a fosforilação desses ácidos graxos exógenos pelo complexo de ácido graxo quinase (Fak) gerando acil-PO₄ e incorporação aos lipídios de membrana via PlsY/PlsC, preferencialmente, ou via PlsX/PlsC, mas comumente quando a via FAS-II está inibida. Contudo, a eficácia desse *bypass* é alvo de controvérsias e parece depender do local da infecção (Morvan *et al.*, 2017; Yao; Rock, 2017; Kénanian *et al.*, 2019). Essa discussão também é fomentada pelo fato de que, em *S. aureus* tratadas com inibidores da FabI, há um acúmulo de acil-ACPs de cadeia curta e uma severa depleção dos níveis de ACP livre, necessário na via de *bypass*. Entretanto, já foram reportados a presença genes mutantes que codificam FabD menos eficientes em *S.*

aureus e, conseqüentemente, reduzem o consumo de ACP e aumentam a disponibilidade de ACP livre ao custo de apresentarem menor viabilidade e menores taxas de crescimento (Parsons *et al.*, 2014; Yao; Rock, 2015; Morvan *et al.*, 2017).

Um estudo *in vivo* realizado em camundongos Balb/C mostra que, em modelos de septicemia, inibidores da FabI conseguem reduzir UFC de cepas MRSA, mas não as eliminar completamente como nos grupos sob administração de vancomicina. Além disso, o UFC tende a subir com o tempo devido a adaptação das bactérias utilizando o mecanismo de *bypass* induzido pela pressão do antibacteriano (Kénanian *et al.*, 2019). Contudo, outros estudos *in vivo* corroboram para o tratamento de infecções intraperitoneais de diferentes cepas com inibidores da FabI pelo sucesso no tratamento da infecção (Park *et al.*, 2007a; Escaich *et al.*, 2011; Kaplan *et al.*, 2012). Um estudo em modelo de coxa neutropênica, utilizado para tecidos moles, demonstrou a eficácia da inibição da FabI no tratamento desse tipo de infecção também em cepas MRSA (Banevicius *et al.*, 2013) e um segundo estudo com o mesmo modelo de infecção elucidou o mecanismo de que a bactéria ainda dependia da via de FAS-II para síntese de alguns ácidos graxos necessários, como o ácido pentadecanóico, devido à ausência de ácidos graxos ramificados no ambiente desse tipo de infecção (Frank *et al.*, 2020). Estudos clínicos também evidenciam a eficácia do tratamento de infecções de pele por *S. aureus* com inibidores da FabI, mesmo em cepas MRSA (Hafkin; Kaplan; Murphy, 2016; Bassetti *et al.*, 2020). Esses estudos reforçam a validação da FabI de *S. aureus* como um alvo molecular adequado, entretanto, com preocupações em relação a eficácia da sua inibição em tratar casos de bacteremia. Ainda assim, *S. aureus* causa predominantemente infecções de pele e de tecidos moles, que até o momento tem-se observado sucesso no tratamento em ensaios pré-clínicos (em roedores) e clínicos utilizando inibidores da FabI. Por fim, um apontamento importante levantado em relação a *S. aureus* e demais Gram-positivas, é que o uso de estratégias de tratamento que combinem inibidores da FabI com outras substâncias capazes de modular a biossíntese de ácidos graxos, como os indutores de (p)ppGpp que regulam a síntese e a distribuição de malonil-CoA, são capazes de inibir o crescimento bacteriano e geram perspectivas de tratamento em biterapia ou politerapia (Pathania *et al.*, 2021). Portanto, a inibição da FabI se mantém como uma estratégia eficaz e robusta para o tratamento de infecções por bactérias Gram-negativas e Gram-positivas.

Sendo uma alternativa promissora ao cenário de cepas multirresistentes de *S. aureus* e de *E. coli*, o planejamento de antimicrobianos visando a inibição da FabI ganha destaque na atualidade. De acordo com dados obtidos em busca na base de dados Cortellis – *Drug Discovery Intelligence* (Clarivate Analytics, 2023), existem três substâncias em estudos clínicos de Fase II, uma substância em Fase I e dois fármacos já comercialmente disponíveis: a isoniazida, utilizada no tratamento da tuberculose, e o triclosan, utilizado como ativo em sanitizantes e sabonetes e, posteriormente, banido desses produtos pelo FDA por promover aumento da resistência bacteriana (**Figura 4**) (Timmins; Deretic, 2006; U.S. Food & Drug Administration, 2016).

Figura 4 – Inibidores da FabI, suas fases mais avançadas dos estudos clínicos atualmente e as empresas pesquisadoras responsáveis.



Fonte: Autoria própria.

Além das substâncias em estudos clínicos, estão descritos na literatura uma série de inibidores com atividade *in vitro*, entre eles se encontram: diazaborinas, imidazóis, naftiridinonas, benzimidazóis, tiopiridinas, derivados do difenil éter e outras classes químicas (Kronenberger *et al.*, 2017). Essas substâncias destacam a validade do alvo molecular e a necessidade de se desenvolver novos inibidores da FabI como uma fonte de novas alternativas terapêuticas.

3.2 EXPRESSÃO, PURIFICAÇÃO E CARACTERIZAÇÃO DE PROTEÍNAS-ALVO NO CONTEXTO DO DESENVOLVIMENTO DE INIBIDORES DA FabI

Nas últimas três décadas, a triagem e identificação *in vitro* de antibacterianos promissores é fundamentada em três estratégias principais que possuem suas vantagens e desvantagens: (I) ensaios baseados no alvo molecular, (II) ensaios baseados na célula (também conhecidos como ensaios no organismo completo) e (III) ensaios baseados no alvo molecular e na célula (também conhecidos como ensaios fenotípicos) (Landeta; Mejia-Santana, 2022).

Os ensaios no alvo molecular, utilizados para identificar moléculas que possam se ligar diretamente ao alvo molecular (geralmente uma proteína), só podem ser realizados se o alvo for conhecido e possível de ser expresso e obtido em sua forma ativa. Além disso, as moléculas encontradas como ativas nesses ensaios frequentemente não são ativas na célula devido ao efluxo ou a baixa permeabilidade celular. Esse fator é ainda mais significativo em bactérias Gram-negativas, uma vez que a parede celular dessas bactérias é uma barreira ainda mais eficiente e difícil de ser ultrapassada por substâncias do que a das Gram-positivas. Por outro lado, os ensaios nas células, como a concentração inibitória mínima (MIC), apesar de serem mais utilizados e com maiores taxas de sucesso, dependem de ensaios secundários para identificar o alvo molecular e prosseguir com as modificações moleculares visando a otimização de um candidato à fármaco. Portanto, o uso de estratégias que combinem os dois tipos de triagem (ensaios fenotípicos) permite não só a identificação da atividade antibacteriana de uma substância, como também a identificação do seu alvo molecular e da sua permeabilidade frente à parede bacteriana (Landeta; Mejia-Santana, 2022).

O primeiro passo na obtenção de uma proteína-alvo é a seleção de um sistema heterólogo que irá expressar a proteína a partir de um gene. Os genes podem ser expressos em muitos sistemas diferentes, portanto, é essencial avaliar qual deles é mais vantajoso para a proteína recombinante de interesse. Em geral, o sistema de expressão ideal é aquele que produz a proteína com segurança e menor custo, garantindo ainda o enovelamento adequado, a atividade biológica e as características físico-químicas da proteína (Mir, 2004; Gomes *et al.*, 2016). Por exemplo, ainda que a *E. coli* seja o organismo mais utilizado e estabelecido na produção de proteínas recombinantes, não é um sistema de expressão recomendado para proteínas com

modificações pós-traducionais. Isso ocorre porque bactérias são limitadas em maquinário para promover adequadamente essas modificações (Terpe, 2006; Macek *et al.*, 2019). A **Tabela 1** traz resumidamente, os principais sistemas de expressão utilizados, suas vantagens e desvantagens (Gomes *et al.*, 2016; Macek *et al.*, 2019).

Tabela 1 – Principais sistemas de expressão utilizados na produção de proteínas recombinantes e suas vantagens e desvantagens básicas.

Sistema de expressão	Vantagens	Desvantagens
Sistemas de expressão procariotos		
<i>Escherichia coli</i>	<ol style="list-style-type: none"> 1. Fácil. 2. Rápido. 3. Econômico. 4. Alta taxa de crescimento. 5. Capacidade de fermentação contínua. 	<ol style="list-style-type: none"> 1. Não é capaz de remover íntrons dos transcritos. 2. Genes exógenos podem conter sequências que podem levar à terminação prematura e perda da expressão gênica. 3. Possui viés de códon. 4. Modificações pós-traducionais e glicosilações são extremamente incomuns em bactérias. 5. Produção de proteínas na forma insolúvel (corpos de inclusão). 6. Degradação de proteínas. 7. Acúmulo de endotoxinas.
<i>Bacillus subtilis</i>	<ol style="list-style-type: none"> 1. Não produz endotoxinas. 2. Facilmente transformado por bacteriófagos e plasmídeos. 3. Capaz de secretar proteínas extracelulares funcionais diretamente no meio de cultura. 	<ol style="list-style-type: none"> 1. Produção de proteases extracelulares que podem degradar proteínas heterólogas. 2. Instabilidade dos plasmídeos. 3. Expressão reduzida ou ausente da proteína de interesse.
Sistemas de expressão eucariotos		
Leveduras	<ol style="list-style-type: none"> 1. Crescimento rápido em meio de baixo custo. 2. Sistema apropriado para modificações pós-traducionais. 3. Sistema seguro de expressão. 4. Sem produção de endotoxinas. 	<ol style="list-style-type: none"> 1. Hiperglicosilação de proteínas. 2. Viés de códon. 3. Ineficiente em secretar proteínas para o meio de cultura, levando à retenção intracelular.
Fungos filamentosos	<ol style="list-style-type: none"> 1. Alto nível de expressão. 	<ol style="list-style-type: none"> 1. Sistema complexo e com pouco conhecimento sobre sua fisiologia.
Células de Inseto (baculovírus)	<ol style="list-style-type: none"> 1. Alto nível de expressão. 2. Sistema apropriado para modificações pós-traducionais. 3. Excelente ferramenta para produção de glicoproteínas. 	<ol style="list-style-type: none"> 1. Impossibilidade da expressão contínua. 2. Condições de cultivo mais exigentes.
Células de mamíferos	<ol style="list-style-type: none"> 1. Enovelamento adequado de proteínas. 2. Modificações pós-traducionais e montagem do produto adequadas. 3. Perfil adequado de glicosilação para proteínas humanas. 	<ol style="list-style-type: none"> 1. Alto custo. 2. Tecnologia complexa. 3. Potencial de contaminação por vírus de animais.
Plantas	<ol style="list-style-type: none"> 1. Fácil expansão com baixo custo. 2. As proteínas podem ser localizadas em diferentes regiões da planta e em diferentes estágios de crescimento. 	<ol style="list-style-type: none"> 1. Níveis de expressão dependem da proteína-alvo. 2. Ensaio funcionais precisam ser desenvolvidos.

Fonte: Adaptado de Gomes *et al.* (2016).

O uso de *E. coli* como sistema de expressão bem consolidado, eficiente e de baixo custo é justificável e ideal para a produção das proteínas recombinantes FabI de *E. coli* e *S. aureus*, uma vez que se trata de uma proteína bacteriana, que não demanda maquinário especial para processamento de seu transcrito ou de modificações pós-traducionais, e que é produzida naturalmente (sem modificações) pelo próprio organismo de expressão.

Em relação ao uso de *E. coli*, uma série de cepas são utilizadas para produção de proteínas recombinantes, sendo as cepas mais comuns as derivadas da BL21 e da K-12. A escolha da cepa depende principalmente de: (i) características da proteína, por exemplo, toxicidade à bactéria, presença e estabilidade de ligações dissulfeto, rendimento, localização (proteína de membrana ou citoplasmática) e necessidade de ser secretada; e (ii) características do plasmídeo construído, por exemplo, a necessidade de uso de códons raros em *E. coli*, necessidade de estabilizar genes, o gene de resistência antibacteriana, o mecanismo de indução da expressão e a presença de certos promotores (como por exemplo, o promotor para a RNA polimerase do bacteriófago T7). Na **Tabela 2** cita-se algumas das cepas mais utilizadas para expressão de proteínas recombinantes (Terpe, 2006; Tungekar; Castillo-Corujo; Ruddock, 2021).

Tabela 2 – Exemplos de cepas de *E. coli* utilizadas na expressão de proteínas recombinantes e suas principais características.

Cepas	Derivação	Características principais
BL21	B834	Deficiente nas proteases <i>lon</i> e <i>OmpT</i> (reduz a degradação da proteína recombinante).
BL21(DE3)	BL21	Deficiente nas proteases <i>lon</i> e <i>OmpT</i> . Rotineiramente utilizada para produção de proteínas recombinantes sobre o controle do promotor T7 regulado pela RNA polimerase T7 transportada pelo profago DE3 (integração cromossômica sob o controle de um promotor lacUV5). A produção da RNA polimerase T7 é inibida até que a adição de isopropil β-D-1-tiogalactopiranosídeo (IPTG) ocorra.
BL21(DE3) [pLysS]/[pLysE]	BL21(DE3)	Deficientes nas proteases <i>lon</i> e <i>OmpT</i> . Codifica o profago DE3 que carrega a RNA polimerase T7. A produção da RNA polimerase T7 é inibida até que a adição de IPTG ocorra. Além disso, essas cepas possuem sequência da lisozima do fago T7, reduzindo os níveis basais de expressão da RNA polimerase de T7 e da proteína recombinante na ausência de IPTG. Apresentam genes de resistência ao cloranfenicol.
C41/C43(DE3)	BL21(DE3)	Cepas com mutações de resistência destinadas à produção de proteínas tóxicas e/ou de membrana.
HMS174(DE3)	K-12	Cepa que estabiliza melhor certos genes-alvo cujos produtos poderiam causar a perda do profago DE3 e permite produção de proteínas heterólogas sobre o controle do promotor T7.
JM 83	K-12	Cepa para secreção de proteínas recombinantes para o periplasma.
Origami	K-12	Cepa que facilita a formação de ligações dissulfeto.
Origami B	BL21	Cepa que facilita a formação de ligações dissulfeto e é deficiente nas proteases <i>lon</i> e <i>OmpT</i> .
Rosetta	BL21	Cepa que melhora a expressão de proteínas de organismos eucariotos que contém códons raramente utilizados em <i>E. coli</i> , também é deficiente nas proteases <i>lon</i> e <i>OmpT</i> .
Rosetta-gami	BL21	Cepa que melhora a expressão de proteínas de organismos eucariotos que contém códons raramente utilizados em <i>E. coli</i> , é deficiente nas proteases <i>lon</i> e <i>OmpT</i> e facilita a formação de ligações dissulfeto.

Fonte: Adaptado de Terpe (2006).

No contexto da FabI de *E. coli* e de *S. aureus*, a utilização da cepa BL21(DE3) é ideal devido a utilização dos vetores pET29a e pET28a, respectivamente, que apresentam resistência à canamicina e controle de expressão da proteína-alvo mediado pelo promotor T7. Além disso, não se espera que a proteína seja tóxica para a bactéria, portanto, não seria demandada a utilização de cepas que apresentem

mutações de resistência destinadas a proteínas tóxicas ou que produzam a lisozima do fago T7, o que reduziria a expressão basal da proteína na ausência de IPTG.

Outra consideração importante a ser feita antes da clonagem e transformação do gene da proteína-alvo é avaliar como será realizada a purificação da proteína após sua expressão. A principal alternativa é realizar uma cromatografia de afinidade, que pode ser realizada de diferentes maneiras. Uma delas é utilizando um ligante imobilizado ou uma molécula ligada à coluna que mimetize o substrato, nesse caso, as características da proteína devem permitir que ela interaja com afinidade suficiente para que seja removida apenas na etapa de eluição. Um exemplo desse tipo de cromatografia de afinidade é o uso de coluna empacotada com resina de proteína G para purificação de anticorpos. Entretanto, no contexto de proteínas recombinantes, a estratégia mais utilizada é adição de uma cauda (*tag*) na proteína para permitir a sua purificação por afinidade. Alguns exemplos incluem caudas de proteína de ligação à maltose (MBP-*tag*), de glutathione-S-transferase (GST-*tag*) e, a mais utilizada, uma cauda de hexahistidina (His₆-*tag*) (Garcia Denegri *et al.*, 2014; Tungekar; Castillo-Corujo; Ruddock, 2021).

O uso da His₆-*tag* permite a purificação por cromatografia de afinidade com metal imobilizado na matriz da coluna cromatográfica, geralmente Ni²⁺, Co²⁺, Cu²⁺ ou Zn²⁺. A cromatografia em coluna de Ni²⁺ é o método mais comum e se baseia na alta afinidade deste metal por resíduos adjacentes de histidina. Nessa cromatografia, a amostra percorre a coluna frente a um gradiente de imidazol. Em baixas concentrações de imidazol, idealmente, apenas a proteína com a His₆-*tag* se liga à coluna, sendo posteriormente eluída devido a um aumento da concentração dessa substância que, pela sua capacidade quelante, rompe as interações da proteína com o Ni²⁺. Recomenda-se utilizar uma concentração baixa de 2 a 50 mM de imidazol no tampão de equilíbrio para evitar a ligação de proteínas de baixa afinidade ao Ni²⁺ e, após a eluição, é necessário a troca de tampão, pois o imidazol pode levar a agregação de proteínas e interfere em diversos experimentos como a cristalografia de proteínas, testes de inibição enzimática e estudos de ressonância magnética nuclear (RMN ou, no inglês, NMR) (Young; Britton; Robinson, 2012).

A localização da cauda de histidina é outro fator importante, deve-se considerar se sua exposição e clivagem são mais favoráveis na extremidade C-terminal ou N-terminal pela análise da estrutura da proteína (previamente depositada em um banco de dados, como o *Protein Data Bank* (PDB), ou obtida computacionalmente por

modelagem baseada em homologia). Além disso, características específicas da proteína devem ser avaliadas, por exemplo, o estado oligomérico ativo das proteínas FabI de *S. aureus* e de *E. coli* é o tetrâmero, entretanto, a presença de His₆-tag na extremidade C-terminal impede a transição dímero-tetrâmero na FabI de *S. aureus*, devido à orientação diferencial dos resíduos nessa extremidade. Dessa forma, a FabI de *S. aureus* com a cauda de histidina na extremidade C-terminal existe como dímero, enquanto a FabI de *E. coli* com His₆-tag C-terminal existe como tetrâmero inalteradamente (Kim *et al.*, 2017; Oliveira; Domingues, 2018). Ainda assim, a transição e manutenção do tetrâmero em ambas as proteínas é altamente dependente da ligação com a coenzima (NAD(P)H) e com o substrato ou inibidor, principalmente porque ajudam na estabilização do loop de ligação do substrato (SBL, L:195-200) juntamente com alterações na interface QR (H4/5) (Schiebel *et al.*, 2012; Maltarollo *et al.*, 2022).

Além da purificação por afinidade, existem outras estratégias de purificação que podem ser feitas na ausência de caudas, são exemplos: a cromatografia de troca iônica, a cromatografia de interação hidrofóbica e a cromatografia de exclusão molecular. Entretanto, a viabilidade desses métodos depende de fatores como estabilidade, hidrofobicidade, tamanho e carga da proteína e dos demais componentes presentes na matriz da amostra (Oliveira; Domingues, 2018). Entre esses métodos, a cromatografia de exclusão molecular é bastante usada no contexto da FabI como última etapa num fluxograma de purificação com cromatografia de afinidade (Mehboob *et al.*, 2012; Chang *et al.*, 2013; Schiebel *et al.*, 2015; Fage *et al.*, 2020; Eltschkner *et al.*, 2021; Parker *et al.*, 2022). Por permitir a separação pelo tamanho e massa molecular, a exclusão molecular é capaz de separar agregados e estados oligoméricos diferentes, permitindo a purificação da proteína ao nível de homogeneidade de tamanho e, conseqüentemente, a caracterização de seu tamanho e estado oligomérico (Oliveira; Domingues, 2018).

Uma vez purificada, a proteína precisa ser caracterizada, o nível de detalhamento da caracterização depende da aplicação final da proteína e, frequentemente não é necessário que a proteína seja totalmente caracterizada. A pureza de uma proteína pode ser confirmada por eletroforese em gel de poliacrilamida com dodecil sulfato de sódio (SDS-PAGE), uma vez que o gel desnaturante permitirá estimar aproximadamente a concentração da proteína, sua massa molecular (após desnaturação) e visualizar outras proteínas contaminantes pela presença de bandas

em faixas de massa molecular não esperadas. A confirmação da sequência da proteína ainda pode ser realizada por espectrometria de massas por dessorção ionizante assistida por uma matriz com a medida do tempo de voo (MALDI-TOF). Nessa técnica, as bandas de proteínas obtidas em um gel (como o de SDS-PAGE) são digeridas por enzimas proteolíticas e o espectro de massas dos peptídeos resultantes pode levar à identificação da proteína pela análise e comparação dos resultados obtidos com bancos de dados de proteínas (Mir, 2004; Oliveira; Domingues, 2018).

A análise por eletroforese em gel de poliacrilamida desnaturante não permite avaliar o estado oligomérico e a homogeneidade de tamanho da proteína, sendo necessário o uso de outras técnicas, como a cromatografia por exclusão molecular, para caracterizar a proteína em termos da sua homogeneidade de tamanho e estado oligomérico. Outra técnica que permite caracterizar a homogeneidade de uma amostra é o espalhamento dinâmico da luz (DLS, do inglês, *Dynamic Light Scattering*), sendo capaz de medir a flutuação da intensidade da luz espalhada por partículas em movimento browniano em uma solução. Além de fornecer uma estimativa do tamanho de uma partícula, uma vez que partículas maiores se movem lentamente e partículas menores movem-se mais rapidamente, essa técnica permite distinguir até mesmo a presença de diferentes proteínas de tamanho próximo, mas apenas quando essas proteínas apresentam formatos e raios diferentes, levando a uma diferença nos coeficientes de difusão. É importante notar que se trata de uma técnica de baixa resolução e, conseqüentemente, pode ter como limitação a dificuldade de diferenciar monômero de dímero (Costa, 2014; Minton, 2016; Oliveira; Domingues, 2018).

Por fim, a estrutura quaternária da proteína pode ser tridimensionalmente elucidada com ou sem ligantes utilizando diferentes técnicas, as duas mais empregadas são a difração de raios-X em cristais da proteína ou a ressonância magnética nuclear (RMN) da proteína em solução (Burley *et al.*, 2022). Uma terceira técnica é a criomicroscopia eletrônica (crio-EM) que, apesar de no passado ter limitações de baixas resoluções e necessidade de grandes complexos proteicos, atualmente obteve grandes avanços sendo possível a obtenção de proteínas como a apoferritina (440 – 465 kDa) com resolução de 1,54 Å e de proteínas menores como a lactato desidrogenase (145 kDa) e a isocitrato desidrogenase (93 kDa) com resoluções de 2,8 Å e 3,8 Å respectivamente (Merk *et al.*, 2016; Lyumkis, 2019; Bai, 2021).

A elucidação da estrutura tridimensional de uma proteína ou do complexo proteína-ligante por difração de raios-X demanda a obtenção da macromolécula com alto grau de pureza e de cristais de alta qualidade para medir as direções e intensidades dos raios-X difratados pelo cristal, obtendo o mapa de densidade eletrônica dos átomos desse cristal que permite a construção do modelo tridimensional da proteína por técnicas computacionais. Por ser um método que depende da densidade eletrônica dos átomos presentes no cristal, átomos de hidrogênio que possuem apenas um elétron são raramente visíveis nesta técnica, dependendo de conhecimento prévio para atribuição do estado de ionização e localização dos hidrogênios nas moléculas (Costa, 2014; Maveyraud; Mourey, 2020).

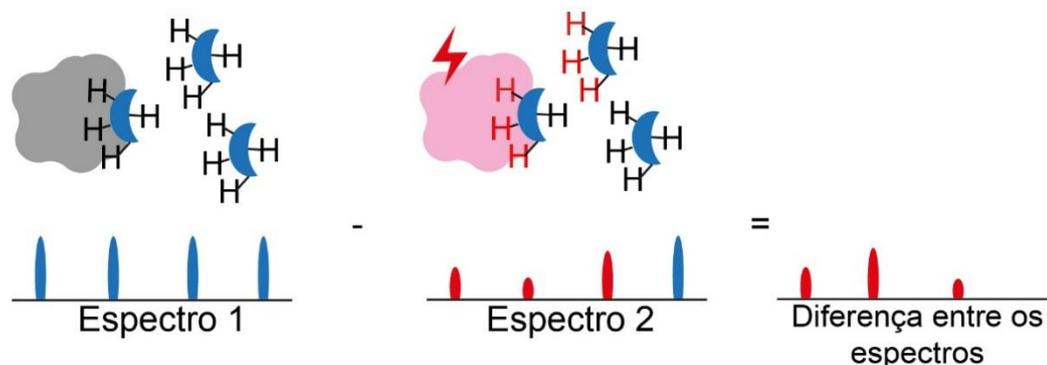
Por outro lado, nos estudos de RMN a coleta de dados geralmente começa pelos espectros de hidrogênio, devido a maior sensibilidade da espectroscopia de ^1H e sua posição relativa em relação aos átomos mais pesados pode ser obtida com alta precisão e menor esforço, mas dependendo do conhecimento prévio para a determinação da localização dos átomos mais pesados (Schirò *et al.*, 2020). Os estudos estruturais de RMN são limitados pelo tamanho, a maioria das estruturas depositadas no PDB utilizando essa técnica são pequenas proteínas, peptídeos ou domínios proteicos isolados (polímeros < 8,5 kDa) (Burley *et al.*, 2022). Em contraste com a difração de raios-X, essa técnica espectroscópica se baseia em propriedades dos núcleos atômicos, mais especificamente no seu momento angular intrínseco (ou “*spin*”) e no momento de dipolo magnético. O momento angular intrínseco tem uma energia específica para cada estado de *spin* e, quando ele interage com um campo magnético externo, resulta num diagrama de níveis de energia relacionado com os valores de dipolo magnético do núcleo. Quando os núcleos são sondados por pulsos de radiofrequência, eles promovem transições de energia e ressonam, de forma com que são registradas frequências de ressonância nuclear atômica (deslocamentos químicos) que sua intensidade é dependente dos átomos ao redor (do ambiente químico). Uma vez que as frequências de ressonância são altamente susceptíveis a alterações no ambiente químico, podem ser utilizadas para extrair informações estruturais e caracterizar eventos de interação ligante-proteína (Blaum *et al.*, 2018; Fernandes, 2022).

A limitação de tamanho é proibitiva para o uso da técnica na elucidação da estrutura tridimensional da FabI, uma vez que seu tetrâmero tem massa molecular entre 110-125 kDa. Portanto, para se contornar este fato e permitir a determinação da

capacidade de um ligante interagir ou não com a proteína, utiliza-se uma técnica conhecida como diferença da transferência de saturação (ou *Saturation Transfer Difference* – STD ou STD-NMR), que é capaz de lidar com proteínas de alto peso molecular, não depende de marcação com isótopos, a proteína pode estar em baixas concentrações e é aplicável para ligantes com afinidades baixas e médias à proteína (baixos valores em mM a altos valores em nM). Por outro lado, o STD é incapaz de fornecer qualquer informação direta sobre a estrutura da proteína, uma vez que as interações proteína-ligante são observadas inteiramente pelo ponto de vista do ligante, e é incapaz de informar diretamente a posição e orientação do ligante no sítio proteico (Gimeno *et al.*, 2017; Blaum *et al.*, 2018; Fernandes, 2022).

O STD se baseia na utilização de um pulso de RMN que excita apenas a proteína, se o ligante é capaz de interagir com a proteína, ocorre a transferência de saturação das ressonâncias de prótons da proteína para os prótons de um ligante que está trocando entre um estado livre para um estado ligado à proteína. Se essa transferência ocorre, o espectro resultante (espectro 2) se assemelha ao espectro do ligante livre (espectro 1) em termos de deslocamentos químicos, mas a intensidade dos picos individuais é atenuada de forma diferencial, dependendo do posicionamento dos prótons do ligante no sítio de ligação. A diferença entre esses dois espectros de hidrogênio indica aqueles hidrogênios do ligante que estabelecem contatos próximos à proteína em distância menor ou igual a 5 Å (**Figura 5**). Além disso, o ligante é capaz de manter a excitação específica dos prótons, enquanto a proteína retorna ao seu estado basal (Blaum *et al.*, 2018; Fernandes, 2022).

Figura 5 – Representação da transferência de saturação de uma proteína diretamente excitada para um ligante.



Fonte: Autoria própria.

3.3 ESTUDOS DE ACOPLAMENTO MOLECULAR

O processo de pesquisa e desenvolvimento (P&D) de fármacos encontra muitos desafios, tais como a limitação no número de substâncias passíveis de serem testadas experimentalmente, o alto custo, as baixas taxas de sucesso em testes pré-clínicos e clínicos, longos períodos até a chegada ao mercado e a dificuldade de se encontrar novas estruturas químicas ativas (Yang *et al.*, 2019). Para contornar esses desafios, métodos computacionais têm sido cada vez mais utilizados no planejamento de novos fármacos, tornando o processo mais rápido, reduzindo custos e aumentando as taxas de sucesso. Esses métodos podem ser empregados como parte de estratégias baseadas na estrutura do alvo molecular (conhecidos no inglês como *Structure-based Drug Design*, ou SBDD), como o *docking* molecular, ou na estrutura de ligantes (conhecidos no inglês como *Ligand-based Drug Design*, ou LBDD), como os estudos quantitativos da relação estrutura-atividade (ou QSAR, do inglês, *quantitative structure-activity relationship*). Enquanto as estratégias LBDD dependem de uma série de ligantes conhecidos, o que pode ser um fator limitante para o desenvolvimento de ligantes estruturalmente inovadores, as estratégias SBDD dependem de dados consistentes sobre as estruturas dos alvos macromoleculares ou de complexos ligante-receptor (Wilson; Lill, 2011; Shen *et al.*, 2019).

O acoplamento molecular, ou *docking* molecular, é um método computacional utilizado no planejamento de fármacos que emprega a estrutura do alvo molecular para encontrar a combinação da conformação e da orientação mais prováveis entre um alvo macromolecular (uma proteína ou ácido nucléico) e um ligante (uma molécula ou outra proteína) (Dias; De Azevedo Jr., 2008; Shen *et al.*, 2019). O acoplamento molecular entre uma molécula e seu alvo pode ser realizado de forma rígida, flexível ou semi-flexível. No *docking* rígido, a conformação dos ligantes e da proteína não muda, apenas a orientação e posição relativa entre eles são ajustadas para otimizar a interação. É um método rápido, mas com baixa capacidade preditiva e, portanto, pouco utilizado no contexto de pequenas moléculas, possuindo aplicações no contexto de interações proteína-proteína ou proteína-ácido nucléico. No *docking* flexível, a conformação de um ligante e um receptor é variada com um determinado nível de liberdade. É um método que reproduz com maior fidelidade o acoplamento real. Entretanto, tem alto custo computacional devido ao número de ligações rotacionáveis e os diversos ângulos possíveis dessas rotações. Por último, o *docking*

semi-flexível é o mais utilizado e, por isso, muitas vezes sendo conhecido apenas como *docking*. Nesse acoplamento, a conformação do receptor é rígida e inalterada, entretanto, a conformação do ligante apresenta níveis de liberdade dentro de uma faixa e critérios delimitados pelo algoritmo (McNutt *et al.*, 2021).

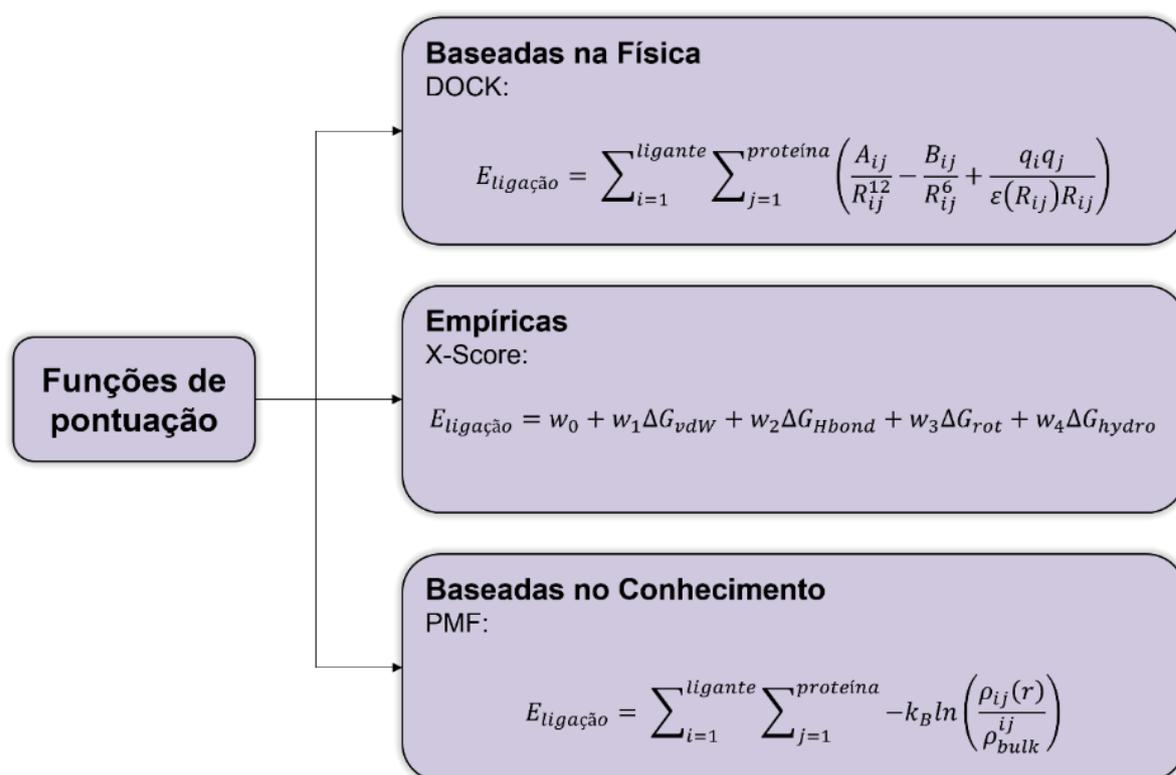
No contexto do *docking* molecular, a localização tridimensional fornecida pela combinação da conformação e da orientação do ligante é também chamada de pose. Cada pose gerada para um ligante é avaliada por uma função de pontuação (também chamada de *score*). Essa função visa estimar a afinidade de ligação entre uma molécula e seu alvo, permitindo a hierarquização e seleção dos ligantes e suas poses conforme suas afinidades (McNutt *et al.*, 2021). Portanto, o *docking* molecular se baseia no algoritmo de busca, que busca pelas poses dos ligantes, e nas funções de pontuação, que hierarquizam essas poses por afinidade (McNutt *et al.*, 2021).

Os algoritmos de busca são classificados em sistemáticos, estocásticos ou determinísticos. Os **métodos sistemáticos** de busca exploram os graus de liberdade de cada ligante de forma incremental, promovendo pequenas variações em parâmetros estruturais e mudando progressivamente a conformação dos ligantes no espaço. O aumento do número de ligações rotacionáveis, aumenta o grau de liberdade e, conseqüentemente, pode levar a um problema combinatório com o número de avaliações necessárias aumentando rapidamente (Torres *et al.*, 2019). Os programas de *docking* FRED e HYBRID utilizam algoritmos de busca que se baseiam nessa estratégia, explorando o espaço conformacional dos ligantes exaustivamente para rapidamente fazer o acoplamento ligante-proteína seguindo um modelo baseado na teoria de chave-fechadura (McGann, 2011, 2012). Os **métodos estocásticos** também realizam mudanças nos graus de liberdade dos ligantes, entretanto, essas modificações são executadas aleatoriamente, frequentemente empregando heurística e métodos iterativos como algoritmos evolucionários (ex.: algoritmo genético) e o algoritmo de Monte Carlo (Torres *et al.*, 2019). Os programas AutoDock e GOLD são exemplos de *softwares* que utilizam algoritmos estocásticos (Jones *et al.*, 1997; Morris *et al.*, 1998). Por fim, os **métodos determinísticos**, que consideram o estado anterior para promover alterações na conformação e orientação do ligante, de forma com que o novo estado tenha um valor de energia igual ou inferior ao anterior. Esses algoritmos têm o custo computacional mais elevado e os exemplos incluem as simulações de dinâmica molecular e os métodos de minimização de energia (Torres *et al.*, 2019).

Explorado o vasto espaço conformacional, uma função de pontuação irá ranquear as poses geradas. Existem várias funções de pontuação e elas podem ser divididas em três classes: as funções baseadas na física, as funções empíricas e as funções baseadas em conhecimento prévio. As **funções baseadas na física**, se referem àquelas baseadas em campos de força moleculares, modelos de solvatação e/ou em métodos de mecânica quântica. As baseadas em campos de força são as mais clássicas e estimam a energia de ligação utilizando termos de mecânica molecular, fundamentada nas leis da mecânica clássica, para as forças covalentes (tamanho, ângulo e torções das ligações covalentes) e para as forças não-covalentes (interações de van der Waals e interações eletrostáticas). Esse tipo de função geralmente não calcula o efeito de solvatação, para isso, são utilizadas as funções de modelos de solvatação. As funções baseadas em mecânica quântica têm acurácia maior que as demais, pelo uso da mecânica quântica para resolver os desafios relacionados às ligações covalentes, à polarização e à transferência de carga. Entretanto, seu custo computacional é altamente proibitivo, sendo observados algoritmos que utilizam a mecânica quântica (QM) juntamente com a mecânica molecular (MM), sendo conhecidos como estratégias QM/MM como tentativa de obter alta performance com menor processamento. As **funções empíricas**, por sua vez, estimam a energia de ligação de um complexo proteína-ligante somando fatores energéticos importantes como a presença de ligações de hidrogênio, efeito hidrofóbico, impedimentos estéricos, ângulos de torção, interações eletrostáticas e outros fatores. Entretanto, essas funções não utilizam campos de força, pois a predição da afinidade de ligação se dá pela relação desses fatores observados com os dados experimentais de outros complexos proteína-ligante. Por último, as **funções baseadas em conhecimento**, são aquelas que se baseiam na ideia da física estatística clássica, onde são observadas as distribuições de distâncias, ângulos e geometrias formados pelos átomos envolvidos nas interações. Esses dados são utilizados para estimar a afinidade de ligação de um complexo por uma função obtida pela relação de dados experimentais de complexos tridimensionais e seus valores de afinidade de ligação (Li; Fu; Zhang, 2019; Torres *et al.*, 2019). Exemplos dos tipos de funções de pontuação estão representados na **Figura 6**, onde $E_{\text{ligação}}$ representa a energia de ligação estimada pelo *docking* (pontuação), os parâmetros A e B são relacionados com os contatos entre diferentes pares de átomos dentre as combinações possíveis, R é a distância entre os centros atômicos, q é a carga parcial

em cada átomo, ϵ é a constante dielétrica, w é o peso atribuído a um elemento da equação, ΔG é a energia livre de Gibbs estimada (para vdW = interações de Van der Waals, $HBond$ = ligações de hidrogênio, rot = de ligações rotacionáveis, $hydro$ = das interações hidrofóbicas), r é o raio observado entre átomos i e j , k_B é a constante de correção e ρ é a densidade (Wang *et al.*, 1998; Jain, 2006).

Figura 6 – Exemplos dos tipos de funções de pontuação.



Fonte: Adaptado de Li; Fu; Zhang, 2019.

3.3.1 Estratégias de validação do acoplamento molecular

Os programas e algoritmos disponíveis para a realização de estudos de *docking* molecular apresentam diferentes parâmetros que podem ser variados, portanto, a escolha da melhor combinação de parâmetros deve ser feita com base em métodos de validação do acoplamento realizado. Para isso, podem ser feitas duas estratégias principais com base nos dados disponíveis: uma estratégia baseada em RMSD (*root mean squared deviation*, ou raiz quadrada do desvio médio quadrático) para avaliar capacidade de prever modo de ligação e uma estratégia baseada na curva característica de operação do receptor (curva ROC) para avaliar a capacidade de distinguir ativos de inativos.

O RMSD é uma métrica que compara, entre duas estruturas em um mesmo sistema de coordenadas, a qualidade da reprodução da pose de um ligante obtida por *docking* com o modo de ligação obtido experimentalmente. Ele é calculado pela raiz quadrada da média do somatório das diferenças entre as posições atômicas dessas duas estruturas (*a* e *b*) em um eixo tridimensional de coordenadas *x*, *y* e *z* (1). A validação de um protocolo de *docking* utilizando o RMSD pode ser feita utilizando estratégias de *redocking* e *crossdocking*. No *redocking*, o ligante cuja estrutura está experimentalmente determinada em complexo com a proteína utilizada para os estudos de *docking*, é novamente acoplado utilizando o protocolo de *docking* que se deseja avaliar. O RMSD é então calculado entre a pose do ligante obtida experimentalmente e a resultante do *docking*, sendo que, de acordo com a literatura, valores de RMSD de *redocking* inferiores à 2,00 Å são aceitáveis. No *crossdocking*, além da estrutura do complexo ligante-proteína que é utilizada para o *docking*, é necessária uma outra estrutura da proteína em conformação diferente e que esteja em complexo com outro ligante. O ligante do segundo complexo é então acoplado à estrutura de proteína do primeiro complexo e o RMSD entre a pose do ligante obtida pelo *docking* e a pose obtida experimentalmente é calculado. É importante notar que no *crossdocking*, os dois complexos proteicos devem estar sobrepostos no espaço tridimensional e que se recomenda uma análise com maior parcimônia do valor de referência para o RMSD de *crossdocking*. Ainda que se mantenha a relação de que quanto menor o valor, melhor a qualidade do protocolo de *docking*, o *crossdocking* pode ser menos sensível que o *redocking*. Isso ocorre tanto pelas alterações conformacionais que ocorrem na proteína induzidas pelo ligante presente (*induced fit*), quanto pelas características de tamanho (alta massa molecular) e flexibilidade (grande número de ligações rotacionáveis) do ligante, que podem afetar a capacidade do *docking* de prever corretamente o modo de ligação. Portanto, a análise deve se basear não somente nos valores de RMSD, mas também na análise da pose e interações do ligante e na consideração dos parâmetros mencionados (Giacoppo *et al.*, 2015; Velázquez-Libera *et al.*, 2020; Wierbowski *et al.*, 2020).

$$RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ia} - x_{ib})^2 + (y_{ia} - y_{ib})^2 + (z_{ia} - z_{ib})^2} \quad (1)$$

A validação de um protocolo de *docking* também pode ser feita por curva ROC. Essa estratégia é voltada principalmente para a validação de protocolos de *docking* para triagem virtual de substâncias com potencial atividade biológica, pois visa avaliar a capacidade do protocolo de estimar maior afinidade em ligantes que interagem com a proteína e menor afinidade em moléculas que não interagem com a proteína. Em outras palavras, permite avaliar a capacidade preditiva do modelo em estratégias de triagem virtual de substâncias. A curva ROC é uma curva dada por dois eixos: no eixo *y*, a taxa de verdadeiros positivos (TPR, *true positive rate*) e, no eixo *x*, a taxa de falsos positivos (FPR, *false positive rate*). Para construção da curva também se utiliza das moléculas experimentalmente capazes de interagir com o alvo molecular para a geração computacional dos *decoys*, moléculas consideradas inativas por serem estruturalmente diferentes. Dessa forma, as moléculas com atividade determinada experimentalmente e os *decoys* (considerados inativos), são acoplados utilizando o protocolo de *docking* e a pontuação é então calculada. A pontuação e atividade dessas moléculas é então utilizada para construção da curva ROC, sua área sob a curva (AUC, *area under the curve*) permite avaliar a performance do modelo em termos de sensibilidade e especificidade, sendo que a melhor performance é alcançada quanto mais próxima for de 1 (Adrià; Garcia-Vallvé; Pujadas, 2012; Torres *et al.*, 2019).

Como dito anteriormente, a pontuação do *docking* é uma estimativa da afinidade de um ligante com a proteína e, portanto, se trata de uma aproximação utilizada para hierarquizar ligantes de uma proteína. Isso implica dizer que a pontuação do *docking* é falha em estimar as reais energias de ligação. Além disso, mesmo como uma função de pontuação utilizada apenas para hierarquizar ligantes, frequentemente protocolos de *docking* podem se tornar impróprios para uso em triagens virtuais devido à alta taxa de falsos positivos (Rastelli *et al.*, 2009; Fan; Fu; Zhang, 2019).

Nesse contexto, algoritmos de aprendizado de máquina estão sendo cada vez mais utilizados para melhorar a acurácia ou reduzir o tempo computacional do *docking*. Esses métodos permitem até mesmo a combinação dos resultados do *docking* com dados de ligantes conhecidos. Essas estratégias híbridas entre LBDD e SBDD usam mais informação para a construção dos modelos, permitindo maior acurácia de predição (Crampon *et al.*, 2022). Um exemplo de estratégia híbrida para capturar informações tanto dos ligantes quanto do alvo molecular é a codificação dos resultados do *docking* em *fingerprints* de interação. Esses *fingerprints* são vetores que

codificam em *bits* as informações dos complexos proteína-ligante e, assim, permitem a aplicação de modelos de aprendizado de máquina para estabelecer correlações entre os *bits* e uma propriedade ou atividade biológica de interesse (Istyastono *et al.*, 2020; Fassio *et al.*, 2022).

3.4 ALGORITMOS DE APRENDIZADO DE MÁQUINA NO DESENVOLVIMENTO DE FÁRMACOS

A inteligência artificial (ou, no inglês, *artificial intelligence* – AI) envolve qualquer sistema capaz de mimetizar aspectos da inteligência humana. Dentro dessa área, o aprendizado de máquina se configura como um ramo da AI que envolve algoritmos capazes de aprender e fazer previsões com base na sua experiência, ao invés de serem explicitamente programados para tais funções (Veríssimo; Gertrudes; Maltarollo, 2023). A aplicação de algoritmos de aprendizado de máquina no desenvolvimento de fármacos vem crescendo e se tornado crucial nas últimas décadas. Isso se justifica por dois fatores: o aumento da disponibilidade de dados experimentais bem consolidados em bases de dados de livre acesso e o importante crescimento de implementações de aprendizado de máquina bem documentadas, de qualidade e de fácil utilização (Ballester, 2019).

Esses algoritmos podem ser aplicados em diversos problemas na química medicinal e no desenvolvimento de fármacos. São exemplos de tarefas e estudos passíveis de aplicação de métodos de aprendizado de máquina: estudos de QSAR e modelos de classificação que realizam a previsão da atividade biológica de substâncias químicas a partir de descritores e propriedades moleculares, determinação do domínio de aplicabilidade de modelos de regressão e classificação, obtenção e seleção de variáveis e descritores para construção de modelos preditivos, redução de dimensionalidade, identificação de possíveis atividades e alvos moleculares de substâncias, previsão de estrutura secundária e tridimensional (terciária/quaternária) de proteínas, previsão de sítios de interação em alvos moleculares, previsão de interações proteína-proteína e ligante-proteína, *docking* molecular e cálculo de funções de pontuação, triagem virtual de substâncias e outros exemplos (Lima *et al.*, 2016).

Os algoritmos de aprendizado de máquina podem ser divididos em três principais categorias de aprendizado: não-supervisionado, supervisionado ou por reforço.

O aprendizado não-supervisionado envolve dados que não possuem nenhum tipo de rótulo ou classificação prévia, sendo que seu objetivo é observar padrões nos dados, permitindo análise descritiva, agrupamento de dados e inferência de propriedades sobre os dados observados. Esse tipo de aprendizado é observado na Análise de Componente Principal (*Principal Component Analysis* – PCA) e Análise de Agrupamento Hierárquico (*Hierarchical Clustering Analysis* – HCA).

O aprendizado supervisionado, por sua vez, engloba algoritmos que trabalham com dados rotulados (por exemplo, moléculas com dados de atividade) que são utilizados para treinar modelos de classificação ou de regressão para prever o rótulo de dados não-rotulados (por exemplo, moléculas sem dados de atividade). No contexto do aprendizado supervisionado, podemos citar os seguintes exemplos de algoritmos de aprendizado de máquina: *k-Nearest Neighbors* (kNN), *Naive Bayes* (NB), *Árvore de Decisão* (*Decision Tree* – DT), *Random Forest* (RF), *Máquina de Vetores de Suporte* (*Support Vector Machine* – SVM) e *redes neurais artificiais* (*Artificial Neural Networks* – ANN), como as *Multilayer Perceptron* (MLP).

Por último, o aprendizado por reforço, utilizado quando não existem dados rotulados e o algoritmo faz decisões a cada etapa que devem ser pontuadas positiva ou negativamente. Seu objetivo é descobrir as regras a serem seguidas que permitem maximizar as recompensas (maximizar uma pontuação ou minimizar um erro). Se trata de uma estratégia de aprendizado menos utilizada quando comparado com as demais, mas é utilizada no algoritmo *Policy Gradient for Forward Synthesis* (PGFS) que permite verificar a ordem de reagentes e reações a serem utilizados em uma síntese química (Veríssimo; Gertrudes; Maltarollo, 2023).

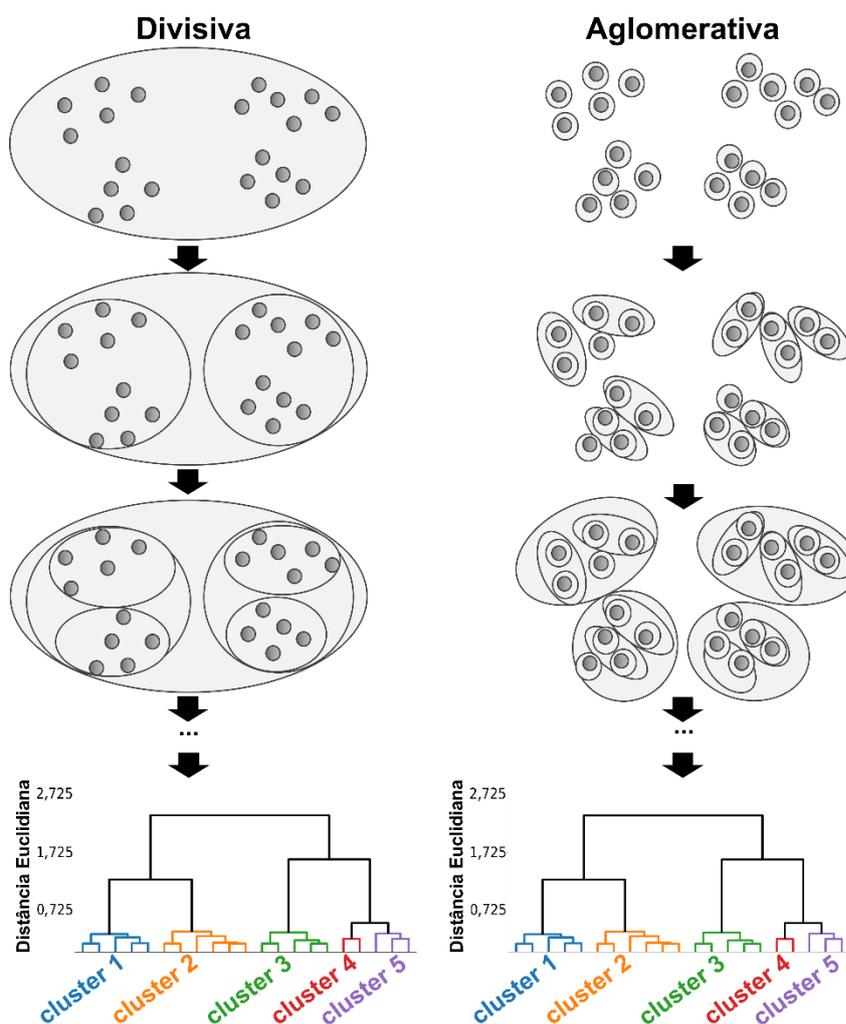
A seguir, uma explicação resumida sobre os exemplos de algoritmos mencionados, com exceção do PGFS, devido sua especificidade e baixa aplicabilidade no contexto do presente trabalho.

A análise de componentes principais (PCA) se trata de um algoritmo utilizado em aprendizado não-supervisionado para a redução de dimensionalidade. Esse algoritmo parte de uma matriz de dados para calcular a tendência de distribuição das variáveis, os graus de relação linear entre elas e, finalmente, a matriz de covariância, que permite hierarquizar as componentes principais em ordem decrescente da

porcentagem de variância explicada do conjunto de dados. Dessa forma, permite ao usuário selecionar apenas as componentes principais que expliquem em maior porcentagem a variância dos dados, realizando assim uma redução de dimensionalidade para manter apenas as componentes principais mais importantes (Greenacre *et al.*, 2022). A redução de dimensionalidade e seleção de variáveis é uma importante etapa prévia para geração de modelos de regressão e classificação com algoritmos de aprendizado supervisionado (Yoo; Shahlaei, 2018; Gu; Li; Li, 2020; Li *et al.*, 2021).

A **análise de agrupamento hierárquico (HCA)** é um método de aprendizado não-supervisionado utilizado para separar amostras em grupos (*clusters*), fornecendo visualização gráfica em dendrogramas da proximidade entre as amostras. Essa técnica se baseia na formação iterativa de grupos entre amostras, com base na distância entre elas (geralmente a distância Euclidiana), de forma com que, a amostra seja agrupada de grupos maiores à menores (abordagem divisiva) ou de grupos menores à maiores (abordagem aglomerativa), sendo a última abordagem a mais comum (**Figura 7**) (Veríssimo; Gertrudes; Maltarollo, 2023). Essa técnica é frequentemente utilizada em estudos de QSAR para separar moléculas em conjuntos de treinamento (para construir o modelo de aprendizado de máquina) e de teste (para validar estatisticamente o modelo) antes de realizar as previsões de atividades biológicas (Pirhadi; Ghasemi, 2010; Salahinejad; Ghasemi, 2014; Primi *et al.*, 2016; Veríssimo *et al.*, 2023).

Figura 7 – Representação das duas abordagens de agrupamento hierárquico.



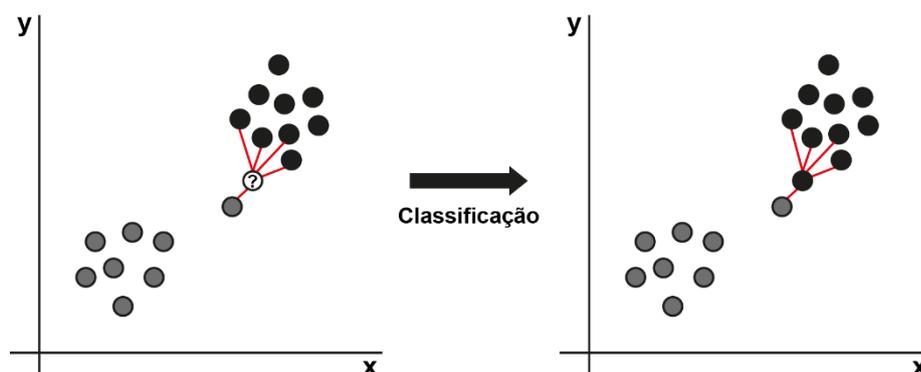
Fonte: Adaptado de Veríssimo; Gertrudes; Maltarollo (2023).

Para entrar em detalhes sobre algoritmos utilizados no aprendizado supervisionado é importante compreender como os modelos de predição são construídos. **Modelos de classificação** são aqueles utilizados para classificar uma amostra. No contexto do desenvolvimento de fármacos, pode-se classificar se uma molécula é ativa ou inativa, por exemplo. Por outro lado, **modelos de regressão** envolvem a predição do valor de uma propriedade sendo, por exemplo, a predição do valor de atividade biológica de uma substância. Para construir esses modelos é necessário que o conjunto de dados rotulados seja dividido em conjunto de treinamento e conjunto de teste. O **conjunto de treinamento** é utilizado para efetivamente construir o modelo de predição. Esse conjunto será utilizado pelo algoritmo de aprendizado de máquina para realizar as predições com base nas relações de suas variáveis independentes com a variável dependente que se deseja

avaliar. E, por fim, o **conjunto de teste** (ou conjunto de validação) é utilizado para validar o modelo estatístico pela comparação dos valores preditos com os valores experimentais. Essa validação é chamada de validação externa uma vez que os elementos do conjunto de teste não foram utilizados para a construção do modelo (Gramatica, 2020).

O **algoritmo de k-Nearest Neighbors (kNN)** é bastante utilizado em aprendizado supervisionado para prever a classe ou o valor de uma propriedade. Trata-se de um algoritmo bastante simples: o algoritmo irá calcular a distância de uma amostra que se deseja realizar a predição de sua variável dependente para todos os elementos presentes no conjunto de treinamento. Com base no número k de vizinhos mais próximos (hiperparâmetro determinado pelo usuário ou por um algoritmo de otimização de parâmetros), determinará o rótulo dessa amostra. Se for uma tarefa de classificação, a classe da variável dependente será aquela que for mais presente entre os k vizinhos mais próximos (**Figura 8**). Se for uma tarefa de regressão, o valor da variável dependente será dado pela média entre os k vizinhos mais próximos (Arian *et al.*, 2020). Esse algoritmo é utilizado em estudos de QSAR e em outras estratégias de triagem virtual de substâncias baseadas em estratégias LBDD (Gunturi; Narayanan, 2007; Tropsha; Golbraikh; Cho, 2011; Mostafa *et al.*, 2022).

Figura 8 – Representação de uma classificação com algoritmo de kNN com $k = 5$.



Fonte: Autoria própria.

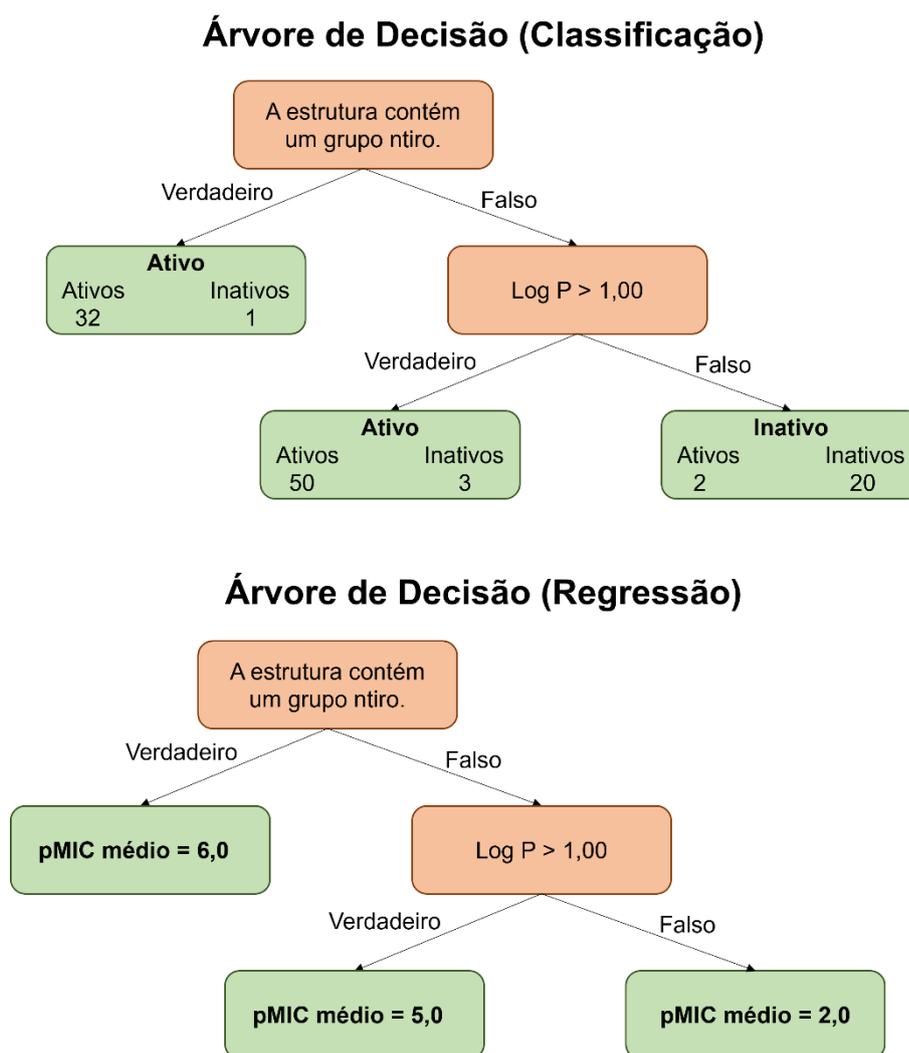
O algoritmo **Naive Bayes (NB)** é baseado no teorema de Bayes e é utilizado em aprendizado supervisionado para construção de modelos de classificação, uma vez que, diretamente e sem adaptações, não pode ser aplicado para tarefas de regressão (Frank *et al.*, 2000). O modelo de *Naive Bayes* se baseia na conversão dos dados do conjunto de treinamento em uma tabela de frequências. Nessa tabela estão

com as probabilidades de cada classe e as probabilidades de cada variável independente dentro de cada classe. Para uma amostra desconhecida, as probabilidades das suas variáveis independentes são utilizadas para calcular pontuações proporcionais a probabilidade de a amostra pertencer a cada classe (ex.: ativo ou inativo), sendo que a classe com maior probabilidade é atribuída à amostra (Mandal; Jana, 2019). Um exemplo de aplicação desses modelos é na predição da toxicidade de substâncias quando administradas em humanos (Marzo; Benfenati, 2018; Zhang *et al.*, 2019, 2020).

Os algoritmos supervisionados de **Árvore de Decisão (DT, do inglês, *Decision Tree*) e *Random Forest (RF)*** utilizados para modelos de classificação e regressão são baseados no conceito das árvores de decisão. Árvores de decisão são árvores onde cada folha representa uma decisão e cada nível (raiz ou galho) representam as diversas opções até se chegar na decisão. Em geral, o tipo mais comum desse algoritmo são as árvores binárias, onde existem apenas duas opções de caminho a cada nível da árvore (**Figura 9**). Cada escolha é baseada em uma das variáveis independentes, de acordo com o conjunto de treinamento, e podem ser utilizadas diversas medidas para a escolha de qual variável será utilizada à cada nível (métricas como a entropia, erro médio quadrático, erro médio absoluto ou outras). Por fim, ao se chegar a uma folha, a decisão final será dada com base na maioria dos votos (classificação) ou na média dos valores (regressão) encontrados entre todos os elementos do conjunto de treinamento que atendiam a todas as regras do caminho até a chegada na folha. Os modelos de *Random Forest* seguem a mesma lógica e são um conjunto de árvores de decisão, entretanto, apresentam a introdução de elementos de aleatoriedade no método. Por exemplo, utilizam um subconjunto aleatório de amostras do treinamento na construção de cada árvore e um subconjunto de variáveis independentes a cada nível. Isso é realizado com objetivo de evitar sobreajuste (*overfitting*), ou seja, para evitar que o modelo seja muito bem ajustado aos dados observados no conjunto de treinamento, mas ineficaz na predição de novos dados. Por fim, a decisão final para cada amostra é dada pela maioria dos votos ou pela média dos valores finais de todas as árvores (Veríssimo; Gertrudes; Maltarollo, 2023). Essas técnicas podem ser utilizadas para modelos de QSAR e de classificação na predição de atividade biológica de ligantes. Os modelos de RF também são frequentemente utilizados para prever a função de pontuação no *docking* molecular

(Han; Wang; Bryant, 2008; Kuz'min *et al.*, 2011; Zilian; Sottriffer, 2013; Li *et al.*, 2015; Wang; Zhang, 2017).

Figura 9 – Representação de uma árvore de decisão de classificação, onde a classificação é dada pela maioria dos votos, e de uma árvore de decisão de regressão, onde o valor da variável dependente é dado pela média dos valores em cada folha.

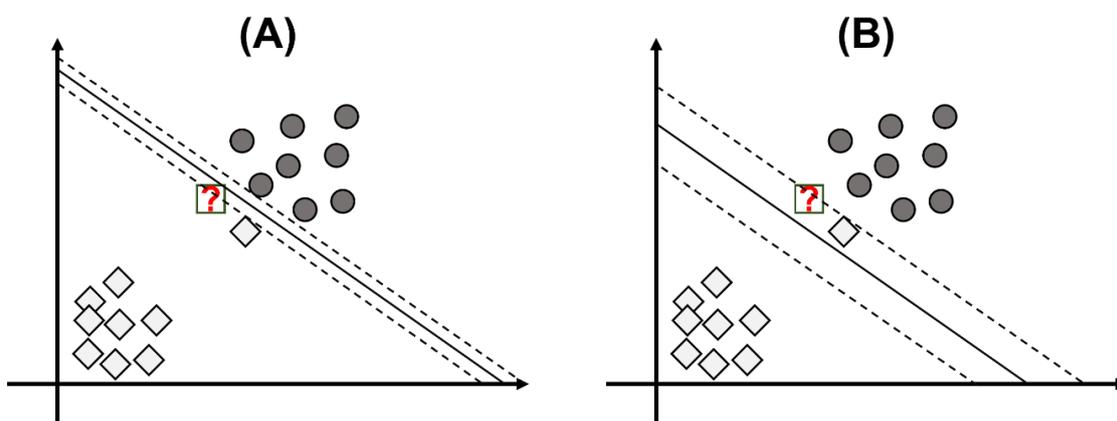


Fonte: Autoria própria.

Os algoritmos supervisionados de **Máquina de Vetores de Suporte (SVM)** podem ser utilizados para tarefas de classificação ou de regressão, mas os modelos para essas tarefas funcionam de formas diferentes. Para ambas as tarefas, o algoritmo envolve a construção de um hiperplano. Nos modelos de classificação, o objetivo é que o hiperplano separe efetivamente os elementos do conjunto de

treinamento em suas classes. Essa divisão deve ser feita para obter o menor número possível de classificações incorretas e, ao mesmo tempo, fazer com que a margem de erro tolerado pelo hiperplano seja a maior possível para reduzir a sensibilidade à *outliers* (**Figura 10**). Nos modelos de regressão, o objetivo é o oposto, ao invés de tentar obter um hiperplano que seja a maior margem possível entre duas classes permitindo violações de margem para reduzir o efeito dos *outliers*, o objetivo na regressão é construir um hiperplano que englobe o maior número possível de elementos enquanto se limita violações de margem. O hiperplano e suas margens são utilizados para construir uma equação que correlaciona as variáveis independentes com a(s) variável(is) dependente(s) analisada(s) (Veríssimo; Gertrudes; Maltarollo, 2023). Os modelos de SVM são bastante versáteis e podem ser utilizados para prever a afinidade proteína-ligante e a função de pontuação de um *docking*, podem ser aplicados em estudos de QSAR, na construção de modelos de classificação para predição de atividade biológica, podem ser utilizados para predição de propriedades ADMET, podem ser utilizados em triagens virtuais, predição de sítios ativos em proteínas e outras funções (Cai *et al.*, 2002; Shahlaei *et al.*, 2010; Ashtawy; Mahapatra, 2015; Jayaraj; Jain, 2019; Holderbach *et al.*, 2020; Yousaf *et al.*, 2021).

Figura 10 – Representação da importância de se considerar outliers na construção hiperplano (linha sólida) e suas margens (linhas pontilhadas) para a classificação de uma amostra desconhecida. Em (A), a margem que minimiza as classificações incorretas no conjunto de treinamento e, em (B), margens mais permissivas que são menos susceptíveis aos outliers e, portanto, melhores em generalização.



Fonte: Autoria própria.

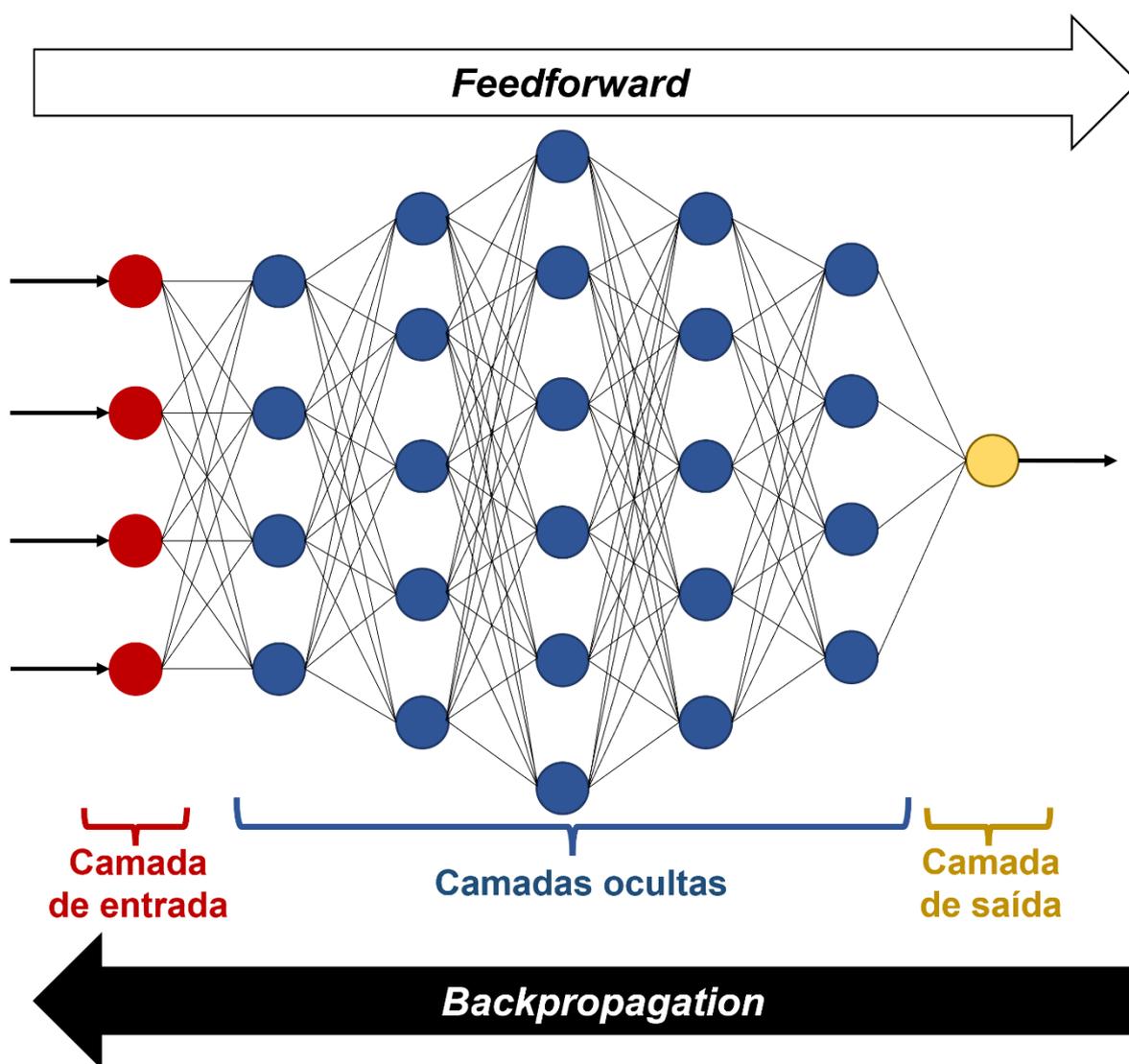
Por fim, os modelos de aprendizado de máquina baseados em **redes neurais artificiais (ANN)** envolvem um conjunto de diferentes algoritmos que utilizam diversos tipos de redes neurais. Esses algoritmos frequentemente são classificados dentro de uma subárea do aprendizado de máquina chamado de aprendizado profundo ou *deep learning* (DL), que corresponde a uma família de algoritmos que se baseia no uso de redes neurais artificiais de múltiplas camadas. Por terem performance superior em comparação com as técnicas mais tradicionais de aprendizado de máquina, esses algoritmos vêm ganhando interesse no planejamento de fármacos (Serrano *et al.*, 2018; Gentile *et al.*, 2020).

O algoritmo clássico para redes neurais artificiais é o algoritmo de ***Multilayer Perceptron (MLP)***, que consiste em uma rede de neurônios artificiais, unidades de processamento simples, organizados em camadas consecutivas que se comunicam por meio de conexões sinápticas. No MLP, a primeira camada (ou camada de entrada) recebe a informação bruta, ou seja, as variáveis de entrada, a última camada (ou camada de saída) consiste nas possíveis variáveis de resposta e, as camadas intermediárias entre a primeira e a última são chamadas de camadas ocultas, pois não são diretamente observáveis e representam os cálculos feitos ao decorrer do processamento. Cada neurônio calcula a soma ponderada de suas entradas, podendo ou não somar um viés (*bias*) a este valor. Sobre o resultado dessa conta é aplicada uma função de ativação para produzir a sua saída, esta comunicação ocorre entre cada uma das sucessivas camadas ocultas até a camada final sendo que, no MLP, essas camadas são totalmente conectadas, ou seja, todos os neurônios de uma camada se comunicam com todos os neurônios da camada anterior e da seguinte.

O treinamento de um modelo de MLP ocorre em duas etapas, denominadas *feedforward* e *backpropagation*. O *feedforward* é o processo de propagar os dados de entrada através das camadas, aplicando pesos e funções de ativação, para gerar uma previsão na camada de saída. O *backpropagation* é o processo de retroceder, calculando o erro entre as previsões e os valores reais, para ajustar gradualmente os pesos e vieses em cada camada e, assim, minimizar o erro. Isso ocorre iterativamente durante o treinamento da rede, permitindo que ela aprenda a realizar tarefas de aprendizado supervisionado, como classificação e regressão, ajustando seus parâmetros para obter melhores resultados (Günther; Fritsch, 2010; Serrano *et al.*, 2018; Desai; Shah, 2021; Veríssimo; Gertrudes; Maltarollo, 2023). A **Figura 11** representa uma rede neural artificial utilizada para MLP e seus principais elementos.

Modelos de MLP também são bastante versáteis, podendo ser utilizado para predição de afinidade proteína-ligante, classificar substâncias conforme o potencial órgão ou sistema em que irá desempenhar sua atividade biológica, identificar subtipos de câncer e para modelos de QSAR e de classificação para determinação de atividade biológica (Patra; Chua, 2010; Limbu; Dakshanamurthy, 2022; Tang; Chen, 2022; Yang *et al.*, 2022).

Figura 11 – Representação de uma rede neural artificial com seus principais elementos.



Fonte: Adaptado de Veríssimo; Gertrudes; Maltarollo (2023).

4 MATERIAIS E MÉTODOS

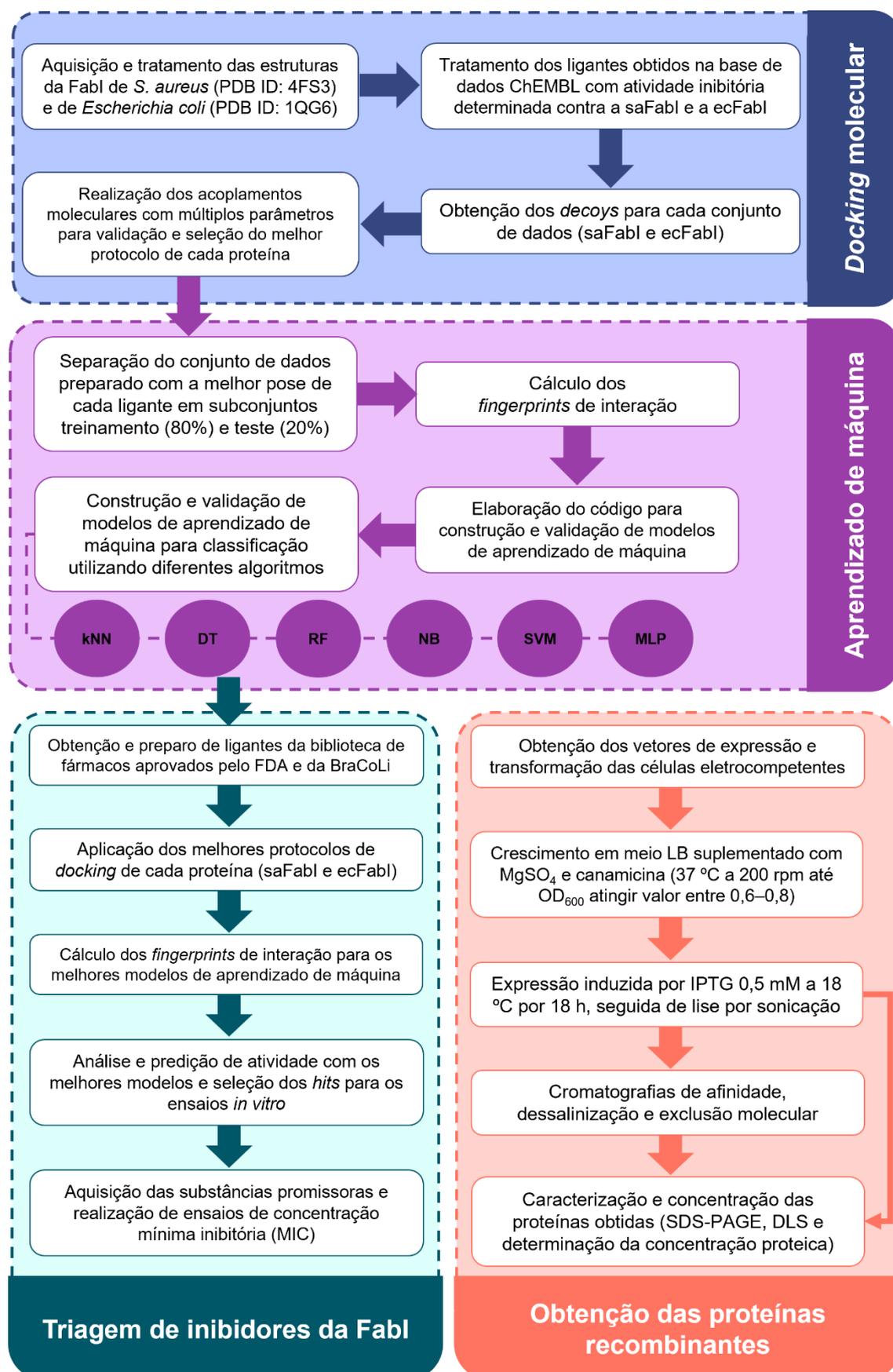
4.1 ESQUEMA GERAL DO TRABALHO

A **Figura 12** representa um esquema geral da metodologia empregada no presente trabalho. Resumidamente, foram conduzidos acoplamentos moleculares para conjuntos de ligantes com conhecida atividade inibitória da FabI de *S. aureus* e de *E. coli*. A melhor pose de cada ligante, em complexo com a sua respectiva proteína, foi utilizada para gerar *fingerprints* de interação. Os *fingerprints* e atividade biológica foram os dados de entrada para o treinamento dos modelos de aprendizado de máquina para classificação. Assim, foram utilizados os seguintes algoritmos para construção dos modelos: *k-Nearest Neighbors* (kNN), *Decision Tree* (DT), *Random Forest* (RF), *Naive Bayes* (NB), *Support Vector Machine* (SVM) e *Multilayer Perceptron* (MLP). Os melhores modelos de aprendizado de máquina foram utilizados para uma triagem virtual na biblioteca de fármacos aprovados pelo FDA disponível no ZINC (Sterling; Irwin, 2015) e na biblioteca BraCoLi (Veríssimo *et al.*, 2022a).

Ensaio de concentração mínima inibitória (MIC) avaliaram a atividade das substâncias promissoras obtidas pela triagem virtual contra as duas espécies de bactérias. Paralelamente, foi conduzido o processo de expressão, purificação e caracterização das proteínas recombinantes saFabI (FabI de *S. aureus*) e ecFabI (FabI de *E. coli*).

Em estudos futuros, as proteínas recombinantes obtidas serão utilizadas em ensaios de STD-NMR visando avaliar a capacidade de interação das substâncias obtidas pela triagem com as proteínas de interesse.

Figura 12 – Representação geral da metodologia utilizada no presente trabalho.



Fonte: Autoria própria.

4.2 COMPUTADORES E PROGRAMAS

As análises computacionais foram realizadas utilizando os recursos computacionais do Laboratório de Modelagem Molecular da Faculdade de Farmácia da Universidade Federal de Minas Gerais e de máquinas virtuais da plataforma NMRbox, disponíveis no site <https://nmrbox.nmrhub.org/> (Maciejewski *et al.*, 2017).

Os programas utilizados foram executados nos sistemas operacionais Windows 11 e Ubuntu 20.04 LTS, sendo que todos os programas que foram utilizados em Linux também foram testados no ambiente do “Subsistema do Windows para Linux” (também conhecido como “*Windows Subsystem for Linux*” ou WSL), viabilizando a execução de tarefas específicas do Linux no contexto do Windows e permitindo a integração de recursos entre os diferentes sistemas operacionais.

Todos os códigos desenvolvidos, inclusive os desenvolvidos para o aprendizado de máquina, foram escritos em Python e testados tanto em Windows quanto em Linux. Entretanto, os códigos que envolviam a utilização do pacote LUNA 0.11.6, responsável pelo cálculo dos *fingerprints* de interação, funcionaram apenas em ambiente Linux (nativo ou via WSL).

Os principais programas, ferramentas e pacotes utilizados foram: Visual Studio Code 1.82.2 (Microsoft Corporation, 2023), utilizado para a escrita dos códigos em Python para múltiplas funções, tais como, tratamento de dados em geral, algoritmos de aprendizado de máquina, geração de gráficos e automação de tarefas; KNIME 4.5.2 2 (Berthold *et al.*, 2009), utilizado para tratamento geral dos dados; PyMOL Open Source 2.6.0a0 (Schrödinger LLC, 2015) e Discovery Studio Visualizer 2021 (DASSAULT SYSTÈMES, 2021), para visualizar e gerar imagens da estrutura tridimensional das proteínas e ligantes; MAESTRO 2022-3 (Sastry *et al.*, 2013; Schrödinger Release 2022-3, 2022), utilizado para o preparo da estrutura das proteínas; LigPrep (Shelley *et al.*, 2007; Schrodinger Release 2022-3, 2022), utilizado para o preparo dos ligantes; Glide (Friesner *et al.*, 2004, 2006; Halgren *et al.*, 2004; Schrödinger Release 2022-3, 2022), utilizado para a obtenção dos complexos proteína-ligante obtidos por acoplamento molecular por encaixe induzido; Decoy Finder 2.0 (Cereto-Massagué *et al.*, 2012), utilizado para obter a estrutura dos *decoys* para a validação do acoplamento molecular; OEDocking 4.2.0.2 (“OEDOCKING 4.2.0.2”, 2023), utilizado para realização dos experimentos de acoplamento molecular; MakeReceptor GUI 4.2.0.2 (“OEDOCKING 4.2.0.2”, 2023), utilizado para definir a

caixa onde seria realizado o acoplamento molecular (preparo do receptor do OEDocking); OMEGA 4.2.1.1 (Hawkins *et al.*, 2010; OpenEye Scientific Software, 2023), utilizado para gerar as múltiplas conformações de ligante utilizadas pela suíte do OEDocking; LigRMSD 1.0 (Velázquez-Libera *et al.*, 2020), para calcular o RMSD; MASSA Algorithm 0.9.2 (Veríssimo, 2021; Veríssimo *et al.*, 2023), para realizar a separação do conjunto de dados em conjunto de treinamento e de teste; ProtParam (Swiss Institute of Bioinformatics, 2017), para o cálculo das propriedades físico-químicas das proteínas; LUNA 0.11.6 (Fassio *et al.*, 2022), para o cálculo dos *fingerprints* de interação; Beautiful Soup 4 (bs4 0.0.1) (Richardson, 2007), para automatizar a coleta de dados do PDB; RDKit 2023.3.2 (Landrum, 2013), para leitura de arquivos de moléculas e para os cálculos relativos à curva ROC; Matplotlib 3.6.2 (Hunter, 2007), para construção de gráficos; Scikit-learn 1.2.0 (Pedregosa *et al.*, 2011) e Scipy 1.11.1 (Virtanen *et al.*, 2020), para aplicação de algoritmos de aprendizado de máquina; tqdm 4.64.1 (Virtanen *et al.*, 2020; Costa-Luis *et al.*, 2023), para inclusão de barra de progresso nos algoritmos de ML; SnapGene Viewer 7.0.2, para geração da imagem dos vetores.

Os códigos escritos para construção e validação dos modelos de aprendizado de máquina variando os diferentes hiperparâmetros e os conjuntos de ligantes utilizados para treinar estes modelos estão disponíveis no GitHub (<https://github.com/gcverissimo/cheML>).

4.3 SELEÇÃO E PRÉ-TRATAMENTO DOS ALVOS MOLECULARES PARA OS ESTUDOS DE ACOPLAMENTO MOLECULAR

A primeira etapa prévia a qualquer estudo de *docking* é a escolha das estruturas adequadas dos alvos moleculares. Para isto, foi realizada uma busca no RCSB *Protein Data Bank* (PDB) de todos os identificadores PDB (PDB ID) de estruturas cristalográficas das enzimas FabI de *S. aureus* e de *E. coli* obtidas por cocrystalização com ligantes. Posteriormente, foi utilizado um código em Python para automação da coleta dos dados de resolução, da presença de mutações na sequência de aminoácidos e dos valores de R (“*R-Value Free*”, “*R-Value Work*”, “*R-Value Observed*”) dessas estruturas. Finalmente, com as estruturas de menores valores de resolução e sem presença de mutação, foi realizada uma avaliação dos valores de R e das métricas “*Clashscore*”, “*Ramachandran outliers*”, “*Sidechain outliers*” e “*RSRZ*”

outliers” para a escolha final da melhor estrutura de proteína de cada espécie. As seguintes estruturas de proteínas identificadas pelo código identificador do PDB (PDB ID) foram as escolhidas: 4FS3 (Kaplan *et al.*, 2012), FabI de *S. aureus* cristalizada com o ligante AFN-1252 e com resolução de 1,80 Å, e 1QG6 (Ward *et al.*, 1999), FabI de *E. coli* cristalizada com o triclosan e com resolução de 1,90 Å.

As duas estruturas foram tratadas para remoção de moléculas de água, correções de cargas e da dupla ocupância de átomos utilizando a ferramenta “*Protein Preparation Wizard*” do software MAESTRO 2022-3 (Sastry *et al.*, 2013; Schrödinger Release 2022-3, 2022).

As estruturas da AFN-1252 e do triclosan foram obtidas do PubChem (Kim *et al.*, 2023) e tratadas no LigPrep (Shelley *et al.*, 2007; Schrodinger Release 2022-3, 2022) utilizando o campo de força OPLS4 (Lu *et al.*, 2021) para a obtenção das estruturas tridimensionais de menor energia nos estados de ionização significativamente favoráveis em $\text{pH} = 7,4 \pm 2,0$. A obtenção dos prováveis estados de ionização nessa faixa de pH é importante, uma vez que o triclosan e outras substâncias apresentam pKa dentro dessa faixa e, conseqüentemente, possuem uma importante contribuição de diferentes protômeros que podem apresentar grandes diferenças entre seus perfis de interação com a proteína (Kronenberger *et al.*, 2019). Por fim, as estruturas geradas para os ligantes foram salvas em um arquivo “.sdf”.

Visando avaliar o efeito dos diferentes ligantes na conformação das proteínas e, principalmente, nos resultados de *redocking* e *crossdocking*, também foram realizados estudos de *induced-fit docking* (*docking* por encaixe-induzido) utilizando o programa Glide (Friesner *et al.*, 2004, 2006; Halgren *et al.*, 2004; Schrödinger Release 2022-3, 2022). Para isto, foram removidos os ligantes das duas proteínas e o estudo de encaixe induzido foi conduzido com os ligantes preparados no LigPrep. Nesse estudo, a AFN-1252 foi ancorada na FabI de *E. coli* (PDB ID: 1QG6), promovendo mudanças conformacionais e resultando na estrutura nomeada como “1QG6_IFD”. Da mesma forma, o triclosan foi ancorado na FabI de *S. aureus* (PDB ID: 4FS3), promovendo alterações conformacionais e resultando na estrutura nomeada como “4FS3_IFD”. Assim, as estruturas de FabI de *S. aureus* e *E. coli* foram consideradas em dois estados conformacionais diferentes para, posteriormente, avaliar a performance de cada um desses estados nos estudos de *docking*.

4.4 SELEÇÃO E PRÉ-TRATAMENTO DOS CONJUNTOS DE LIGANTES PARA OS ESTUDOS DE ACOPLAMENTO MOLECULAR

Foram coletados do banco de dados ChEMBL (Gaulton *et al.*, 2017) e da literatura (Zheng *et al.*, 2013; Ghattas *et al.*, 2019; Veríssimo *et al.*, 2022b) 140 ligantes únicos com dados de inibição enzimática da FabI de *E. coli* e 273 ligantes únicos com dados de inibição enzimática da FabI de *S. aureus*. Todos os 140 ligantes relativos à *E. coli* tinham atividade inibitória da FabI expressa em concentração inibitória média (IC₅₀) e, dos 273 ligantes únicos referentes à enzima de *S. aureus*, a atividade de 270 ligantes foi expressa em IC₅₀ e 3 ligantes tiveram a atividade expressa em porcentagem de inibição a 1 µM.

A obtenção dos ligantes únicos (ausência de molécula repetidas) envolveu um processo de exclusão de dados repetidos utilizando o KNIME 4.5.2 (Berthold *et al.*, 2009) e consistiu na obtenção do identificador químico internacional (InChi) e, por ele, encontrar estruturas repetidas e tratá-las. Esse processo priorizou os dados de IC₅₀ frente aos dados de porcentagem de inibição, ou seja, caso uma substância tenha atividade determinada e expressa nas duas formas pesquisadas, os dados de porcentagem de inibição foram excluídos e os dados de IC₅₀ foram mantidos. Na presença de substâncias repetidas com atividade enzimática determinada sob a mesma forma, por exemplo, a mesma molécula com diferentes valores de IC₅₀, o valor de IC₅₀ considerado para esta molécula foi a média entre as determinações encontradas.

A classificação binária (0 para inativo e 1 para ativo) das substâncias, conforme seus valores de atividade, foi necessária para construção dos modelos de aprendizado de máquina e para validação dos protocolos de *docking* por curva característica de operação do receptor (ou curva ROC, do inglês, “*receiver operator characteristic curve*”). Em relação as substâncias com atividade inibitória determinada em IC₅₀, aquelas com IC₅₀ > 1 µM foram consideradas como inativas (0), enquanto as com IC₅₀ ≤ 1 µM foram classificadas como ativas (1). As três substâncias com atividade expressa em porcentagem de inibição a 1 µM foram classificadas como inativas tendo em vista que, no artigo de referência (Takhi *et al.*, 2014), essas substâncias foram excluídas da etapa de determinação de IC₅₀ devido aos baixos valores de porcentagem de inibição na fase de triagem inicial, que variaram entre 16 e 47%. Além

disso, esse valor de porcentagem indicaria uma IC₅₀ maior do que 1 µM, sendo, portanto, classificadas como inativas (0).

Com os dados dessas moléculas, os ligantes foram preparados no LigPrep (Shelley *et al.*, 2007; Schrodinger Release 2022-3, 2022), utilizando o campo de força OPLS4 (Lu *et al.*, 2021), para a obtenção das estruturas tridimensionais de menor energia e os estados de ionização mais prováveis em pH = 7,4 ± 2,0. Por fim, as estruturas geradas para os ligantes foram salvas em um arquivo “.sdf”.

Para a validação dos protocolos de *docking* pela curva ROC, também foi necessário a obtenção e preparo de *decoys* usando o Decoy Finder 2.0 (Cereto-Massagué *et al.*, 2012). *Decoys* são moléculas que são presumidamente inativas contra um alvo molecular e são comumente utilizadas para validar protocolos de *docking*. Para isso, os ligantes classificados como ativos foram utilizados na busca de *decoys* para a enzima de cada bactéria (124 ativos para *S. aureus* e 42 ativos para *E. coli*). Essa busca foi feita entre as substâncias disponíveis na biblioteca virtual de substâncias ZINC (Irwin *et al.*, 2012).

O processo de busca pelo Decoy Finder 2.0 foi feito da seguinte forma: O *fingerprint* MACCS foi calculado para cada ativo e para cada candidato à *decoy*. Em seguida, o coeficiente de Tanimoto (Tanimoto, 1958) entre o *fingerprint* de cada ativo e de cada potencial *decoy* foi calculado. Considerando os *fingerprints* de duas moléculas (*a* e *b*), o coeficiente de Tanimoto (2) corresponde ao número de bits de *fingerprint* presentes nas duas moléculas (*c*) dividido pelo número de bits de *fingerprint* presentes em apenas uma das moléculas (Willett, 2006).

$$Tanimoto = \frac{c}{a+b-c} \quad (2)$$

Para cada molécula ativa, foram selecionados 50 *decoys* de acordo com as seguintes características: o coeficiente de Tanimoto entre um *decoy* potencial e uma molécula ativa não pode ser maior do que 0,40 e entre um *decoy* potencial e um *decoy* já escolhido não pode ser maior do que 0,80, a massa molecular não pode passar de 40 Da de diferença do ligante ativo, um *decoy* deve ter o mesmo número de aceptores e doadores de ligação de hidrogênio que o ligante ativo e, por último, o log P não pode variar mais do que 1 unidade de log entre um ativo e o *decoy* potencial. Dessa forma, os *decoys* selecionados são estruturalmente diferentes dos ligantes ativos e, portanto,

presumidamente inativos. Além disso, os *decoys* têm propriedades físico-químicas próximas, evitando o viés de que a ausência de atividade seja exclusivamente devido às propriedades moleculares (Cereto-Massagué *et al.*, 2012). Como resultado, foram gerados 6200 *decoys* para a FabI de *S. aureus* e 2100 *decoys* para a FabI de *E. coli*, que foram então preparados no LigPrep (Shelley *et al.*, 2007; Schrodinger Release 2022-3, 2022), utilizando o campo de força OPLS4 (Lu *et al.*, 2021), para a obtenção das estruturas tridimensionais de menor energia nos estados de ionização significativamente favoráveis em $\text{pH} = 7,4 \pm 2,0$. Essas estruturas geradas para os *decoys* foram salvas em um arquivo “.sdf”. A **Tabela 3** mostra a proporção de ativos, inativos e decoys para cada *dataset*.

Tabela 3 – Número de substâncias ativas, inativas e *decoys* para cada conjunto de dados.

<i>Dataset</i>	Ativos	Inativos	<i>Decoys</i>
FabI <i>E. coli</i>	42	98	2100
FabI <i>S. aureus</i>	124	149	6200

Fonte: Autoria própria.

4.5 SELEÇÃO E VALIDAÇÃO DOS PROTOCOLOS DE ACOPLAMENTO MOLECULAR

Os estudos de *docking* molecular foram realizados utilizando a suíte OEDocking 4.2.0.2 (“OEDOCKING 4.2.0.2”, 2023) para as estruturas das proteínas de *S. aureus* (4FS3 e 4FS3_IFD) e para as estruturas das proteínas de *E. coli* (1QG6 e 1QG6_IFD). Para cada proteína, foram testados diferentes protocolos de *docking*.

Utilizando a suíte OEDocking 4.2.0.2, os estudos de *docking* foram realizados da seguinte forma: inicialmente, foram construídos os receptores utilizando o programa MakeReceptor GUI 4.2.0.2 dessa mesma suíte. Com a estrutura do complexo proteína-ligante, o MakeReceptor GUI é capaz de detectar o sítio ativo e gerar, automaticamente, a caixa onde o *docking* será realizado (o receptor). Para cada proteína, foi automaticamente gerado o primeiro receptor de tamanho definido pelo programa e, com base nele, foram gerados mais seis receptores de diferentes tamanhos variando de 1 Å em 1 Å as três dimensões que definem o volume da caixa. A **Tabela 4** contém a identificação de todos os receptores gerados.

Tabela 4 – Identificação dos receptores gerados em termos da estrutura de proteína utilizada, do ligante originalmente associado à proteína, do volume da caixa e das dimensões da caixa.

Receptor	Proteína	Ligante	Volume da caixa (Å ³)	Dimensão x (Å)	Dimensão y (Å)	Dimensão z (Å)
v0	4FS3	AFN-1252	5419,00	14,00	17,33	22,33
v1	4FS3	AFN-1252	6416,00	15,00	18,33	23,33
v2	4FS3	AFN-1252	7527,00	16,00	19,33	24,33
v3	4FS3	AFN-1252	8756,00	17,00	20,33	25,33
v4	4FS3	AFN-1252	4529,00	13,00	16,33	21,33
v5	4FS3	AFN-1252	3741,00	12,00	15,33	20,33
v6	4FS3	AFN-1252	3048,00	11,00	14,33	19,33
v7	1QG6	Triclosan	3397,00	13,00	14,00	18,67
v8	1QG6	Triclosan	4130,00	14,00	15,00	19,67
v9	1QG6	Triclosan	4960,00	15,00	16,00	20,67
v10	1QG6	Triclosan	5893,00	16,00	17,00	21,67
v11	1QG6	Triclosan	2756,00	12,00	13,00	17,67
v12	1QG6	Triclosan	2200,00	11,00	12,00	16,67
v13	1QG6	Triclosan	1723,00	10,00	11,00	15,67
v14	1QG6_IFD*	AFN-1252	6010,00	16,00	16,33	23,00
v15	1QG6_IFD*	AFN-1252	7072,00	17,00	17,33	24,00
v16	1QG6_IFD*	AFN-1252	8249,00	18,00	18,33	25,00
v17	1QG6_IFD*	AFN-1252	9550,00	19,00	19,33	26,00
v18	1QG6_IFD*	AFN-1252	5059,00	15,00	15,33	22,00
v19	1QG6_IFD*	AFN-1252	4214,00	14,00	14,33	21,00
v20	1QG6_IFD*	AFN-1252	3466,00	13,00	13,33	20,00
v21	4FS3_IFD*	Triclosan	3567,00	13,33	14,33	18,67
v22	4FS3_IFD*	Triclosan	4322,00	14,33	15,33	19,67
v23	4FS3_IFD*	Triclosan	5175,00	15,33	16,33	20,67
v24	4FS3_IFD*	Triclosan	6134,00	16,33	17,33	21,67
v25	4FS3_IFD*	Triclosan	2905,00	12,33	13,33	17,67
v26	4FS3_IFD*	Triclosan	2329,00	11,33	12,33	16,67
v27	4FS3_IFD*	Triclosan	1834,00	10,33	11,33	15,67

* O IFD se refere a estrutura da proteína em estado conformacional obtido pelo encaixe induzido, como descrito na seção **4.3 Seleção e pré-tratamento dos alvos moleculares para os estudos de acoplamento molecular (p. 68)**.

Fonte: Autoria própria.

Além dos seis receptores para cada proteína, um código em Python utilizando o módulo *subprocess* foi escrito para efetuar as chamadas de linha de comando e, assim, realizar o *docking* variando também: o número de conformações máximo obtido para cada ligante utilizando o OMEGA 4.2.1.1 (Hawkins *et al.*, 2010; OpenEye Scientific Software, 2023), os algoritmos de *docking* (FRED, HYBRID e POSIT), o parâmetro de resolução do *docking* para os algoritmos FRED e HYBRID (que define de quantos em quantos angstroms a molécula poderá realizar uma translação ou uma rotação) e os parâmetros de permissão de colisões e de relaxamento das poses para o algoritmo POSIT. A definição de cada um desses parâmetros está descrita na

documentação do programa (“OEDOCKING 4.2.0.2”, 2023). Na **Tabela 5** estão relatados os demais parâmetros variados, além do tamanho do receptor já descrito na **Tabela 4**. Em conjunto, todos esses parâmetros representaram 588 protocolos por estrutura de proteína. Para cada espécie, isso representou 1176 protocolos e, considerando as 4 estruturas de proteínas utilizadas, o total de protocolos de *docking* foi de 2352. Em todos os protocolos foi considerada apenas a pose melhor ranqueada, definida como tal pelo algoritmo de *docking*.

Tabela 5 – Parâmetros dos algoritmos de *docking* variados para geração dos protocolos.

Algoritmo	Número de conformações máximo por ligante	Resolução do <i>docking</i>	Relaxamento de poses	Permissão de colisões
FRED	30	“High” “Standard” “Low”	-	-
	60			
	90			
	120			
HYBRID	30	“High” “Standard” “Low”	-	-
	60			
	90			
	120			
POSIT	1	-	“None” “Clashed” “All”	“Noclashes” “Mildclashes” “Hclashes” “Allclashes”
	30			
	60			
	90			
	120			

Fonte: Autoria própria.

Inicialmente, os acoplamentos com os 2352 protocolos foram realizados com o triclosan e a AFN-1252 obtidos do PubChem, como descrito na seção **4.3 Seleção e pré-tratamento dos alvos moleculares para os estudos de acoplamento molecular (p. 68)**, para realizar os estudos de *redocking* e *crossdocking*. A seleção dos 5 melhores protocolos de cada espécie foi feita pela escolha dos menores valores de RMSD de *redocking* e *crossdocking* que foram calculados utilizando o LigRMSD 1.0 (Velázquez-Libera *et al.*, 2020).

Para a escolha do melhor protocolo de cada espécie a ser utilizado nos algoritmos de aprendizado de máquina, foi necessária a construção da curva ROC e cálculo da área sob a curva (AUC, do inglês, *area under the curve*), dos fatores de enriquecimento e da área sob a curva utilizando a discriminação aprimorada de Boltzmann da curva ROC (BEDROC, do inglês, *Boltzmann-Enhanced Discrimination*

of ROC curve). Para a realização desses cálculos e construção do gráfico da curva ROC foram utilizados os pacotes do Python: RDKit 2023.3.2 (Landrum, 2013) e Matplotlib 3.6.2 (Hunter, 2007). Também foi necessário realizar a ancoragem molecular dos ligantes do banco de dados e dos *decoys* utilizando os 5 melhores protocolos de cada espécie. Finalmente, a curva ROC foi construída, sendo definida pela relação entre os dados experimentais de classificação de atividade (ativo ou inativo) e o valor da pontuação do acoplamento molecular dos ligantes e *decoys*.

Com base nos valores de RMSD, AUC-ROC, dos fatores de enriquecimento e de AUC-BEDROC, foi possível selecionar o protocolo de *docking* para a FabI de cada espécie com melhor predição de pose e afinidade e, assim, realizar os acoplamentos moleculares das moléculas das bibliotecas de triagem virtual para a construção dos modelos de aprendizado de máquina.

4.6 SEPARAÇÃO DO CONJUNTO DE DADOS EM SUBCONJUNTOS DE TREINAMENTO E DE TESTE PARA OS ALGORITMOS DE APRENDIZADO DE MÁQUINA

Os conjuntos de moléculas de cada espécie, 140 ligantes para a FabI de *E. coli* e 273 ligantes para a FabI de *S. aureus*, foram separados em conjunto de treinamento (80%) e conjunto de teste (conjunto de validação externa, 20%) utilizando o MASSA Algorithm 0.9.2 (Veríssimo, 2021; Veríssimo *et al.*, 2023). Essa ferramenta em Python foi desenvolvida pelo autor deste trabalho e realiza a separação racional de conjuntos de moléculas em subconjuntos de treinamento e de teste para estudos de QSAR e aprendizado de máquina. A separação é feita com base nas características estruturais, físico-químicas e biológicas das moléculas que são exploradas por análise de agrupamento hierárquico seguida de clusterização por K-modes. A **Tabela 6** mostra o número de moléculas em cada subconjunto (treinamento e teste) para cada *dataset*.

Tabela 6 – Proporção de moléculas em cada subconjunto de cada *dataset*.

<i>Dataset</i>	Moléculas no conjunto de treinamento	Moléculas no conjunto de teste
FabI <i>E. coli</i> (ecFabI)	112	28
FabI <i>S. aureus</i> (saFabI)	218	55

Fonte: Autoria própria.

4.7 CÁLCULO DOS *FINGERPRINTS* DE INTERAÇÃO

Fingerprints de interação são vetores com notações da presença/ausência ou contagem de características estruturais dos ligantes, como um *fingerprint* molecular tradicional, mas, além disso, também incluem essas notações para características da proteína e das interações presentes no complexo proteína-ligante. Os ligantes do conjunto de dados, em suas melhores poses e no seu respectivo complexo proteína-ligante, foram utilizados para calcular os *fingerprints* de interação utilizando o LUNA 0.11.6 (Fassio *et al.*, 2022).

Com o objetivo de se encontrar os melhores parâmetros para geração dos *fingerprints*, foram variados os parâmetros do algoritmo do LUNA sendo, posteriormente, avaliados com base nas métricas de validação dos modelos de aprendizado de máquina. Os parâmetros variados foram o tipo de *fingerprint* de interação (“IFP_TYPE”), o número de níveis de iteração (“IFP_NUM_LEVELS”), o incremento do raio da esfera que busca as características estruturais a cada iteração (“IFP_RADIUS_STEP”), o comprimento do *fingerprint* (“IFP_LENGTH”) e se o *fingerprint* era de contagem ou apenas da presença/ausência dos bits (“IFP_COUNT”). Os valores utilizados para cada parâmetro estão representados na **Tabela 7** e todas as combinações de valores desses quatro parâmetros foram testadas. A escolha desses parâmetros foi orientada com base nos estudos conduzidos no artigo de referência do LUNA (Fassio *et al.*, 2022). Em relação aos tipos de *fingerprints* (“IFP_TYPE”), foram testados todos os algoritmos disponíveis: *Extended Interaction Fingerprint* (EIFP), *Functional Interaction Fingerprint* (FIFP) e *Hybrid Interaction Fingerprint* (HIFP). O EIFP considera as subestruturas atômicas explicitamente definidas; o FIFP leva em conta as características farmacofóricas; e o HIFP é uma combinação dos dois algoritmos anteriores, considerando as propriedades farmacofóricas para grupos de átomos e as subestruturas explícitas sobre cada átomo. Entretanto, todos os três tipos de *fingerprint* possibilitam a interpretação das características estruturais específicas que compõem cada *bit* para cada molécula.

Tabela 7 – Parâmetros e valores variados para a geração de *fingerprints*.

Parâmetro	Valores
IFP_TYPE	EIFP, FIFP ou HIFP
IFP_NUM_LEVELS	1, 2, 3, 4, 5, 6 ou 7
IFP_RADIUS_STEP	1,4329275, 2,865855 ou 5,73171
IFP_LENGTH	1024, 2048 ou 4096
IFP_COUNT	<i>count</i> ou <i>bit</i>

Fonte: Autoria própria.

Dessa forma, foi possível obter diferentes representações dos resultados do *docking* para uso no aprendizado de máquina, de forma a envolver as informações da proteína, dos ligantes e das prováveis interações e tipos de interações presentes nestes complexos.

4.8 CONSTRUÇÃO E VALIDAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA COM OS *FINGERPRINTS* DE INTERAÇÃO DAS POSES DO ACOPLAMENTO MOLECULAR

Os *fingerprints* de interação das moléculas dos conjuntos de treinamento e de teste foram utilizados para construção e validação dos modelos de classificação. Os seguintes algoritmos de aprendizado de máquina foram utilizados para a construção dos modelos de classificação: *k-Nearest Neighbors* (kNN), Árvore de Decisão (DT), *Random Forest* (RF), *Gaussian Naïve Bayes* (gNB), Máquina de Vetores de Suporte (SVM) e *Multilayer Perceptron* (MLP). No caso do SVM também foram consideradas suas 4 opções de *kernel* disponíveis no Scikit-learn 1.2.0 (Pedregosa *et al.*, 2011), sendo elas: SVM com *kernel* linear (SVL), *Radial Basis Function* (SVR), sigmoide (SVS) ou polinomial (SVP). Da mesma forma, também foram consideradas as 3 opções de *solver* para o MLP, sendo elas: MLP com o *solver Adam* (MLA), com o *Stochastic Gradient Descent* (MLS) ou com o *Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm* (MLL). Além disso, uma Análise de Componente Principal (PCA) foi aplicada previamente aos algoritmos de classificação para manter apenas as componentes principais que explicassem 85% da variância dos *bits* de *fingerprint* e, dessa forma, efetuar uma redução de dimensionalidade prévia aos modelos de classificação.

As transformações prévias realizadas envolveram: primeiramente, uma padronização dos dados utilizando o “*sklearn.preprocessing.StandardScaler()*” e, posteriormente, uma PCA com “*sklearn.decomposition.PCA(n_components=0.85, svd_solver='full')*”. Para isso, os cálculos e os ajustes (Método “.*fit()*”) foram sempre realizados com o conjunto de treinamento e a transformação (Método “.*transform()*”) aplicada a ambos os conjuntos, de treinamento e de teste. O modelo e cálculos gerados pelo método “.*fit()*” foram salvos para garantir consistência em aplicações futuras, por exemplo, para a realização da triagem virtual.

Assim como para os parâmetros do *fingerprint* de interação do LUNA, a otimização dos hiperparâmetros dos algoritmos de aprendizado de máquina para os modelos de classificação também foi feita com base nas métricas de validação. Os hiperparâmetros testados estão descritos na **Tabela 8** e seus valores estão representados conforme escritos em Python. A **Tabela 9** traz a descrição das camadas ocultas testadas nas redes neurais (parâmetro '*hidden_layer_sizes*'). A escolha de hiperparâmetros foi realizada com base em estudos prévios, nas descrições disponíveis na documentação do Scikit-learn e nos guias do Google para aprendizado profundo (Pedregosa *et al.*, 2011; Kensert *et al.*, 2018; Bian *et al.*, 2019; Maltarollo, 2019; Godbole *et al.*, 2023). Uma observação importante é que o parâmetro *random seed* foi fixado em 2023 para garantir reprodutibilidade.

Tabela 8 – Hiperparâmetros utilizados nos modelos de aprendizado de máquina.

Algoritmo	Hiperparâmetros
kNN	'n_neighbors': [1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39]
	'weights': ['uniform', 'distance']
	'algorithm': ['ball_tree']
	'p': [2]
	'leaf_size': [15, 30, 45, 60]
	'metric': ['minkowski']
DT	'criterion': ['gini']
	'splitter': ['best']
	'min_samples_split': [2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40]
	'max_features': ['sqrt', 'log2', None]
RF	'criterion': ['gini']
	'n_estimators': [10000, 20000, 30000, 40000, 50000]
	'max_depth': [5, 10, 20, 30, 40, 50]
	'max_features': ['sqrt', 'log2', None]
gNB*	'var_smoothing': np.logspace(0, -10, num=50)
	'prob_a': np.arange(0.01, 1.00, 0.01)
	'prob_b': np.arange(0.99, 0.00, -0.01)
	'prob': [(x, y) for (x, y) in zip(prob_a, prob_b)]
SVL	'C': [0.0001, 0.001, 0.01, 0.1, 0.5, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 150, 200, 250, 500, 750, 1000, 1250, 1500]
SVR	'C': Vide SVL
	'gamma': [0.0001, 0.0005, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 5, 10]
SVS	'C': Vide SVL
	'gamma': Vide SVR
	'coef0': [-1, -0.9, -0.8, -0.7, -0.6, -0.5, -0.4, -0.3, -0.2, -0.1, -0.09, -0.08, -0.07, -0.06, -0.05, -0.04, -0.03, -0.02, -0.01, -0.009, -0.008, -0.007, -0.006, -0.005, -0.004, -0.003, -0.002, -0.001, -0.0005, -0.0001, 0, 0.0001, 0.0005, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]
SVP**	'C': Vide SVL
	'gamma': [0.0001, 0.0005, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09]
	'coef0': Vide SVS
	'degree': [2, 3, 4, 5]
MLL	'hidden_layer_sizes': A Tabela 9 traz a descrição das camadas ocultas testadas nas redes neurais.
	'activation': ['identity', 'logistic', 'tanh', 'relu']
	'alpha': [0.00001, 0.0001, 0.001, 0.01]
	'max_iter': [500]
MLA e MLS	'hidden_layer_sizes': A Tabela 9 traz a descrição das camadas ocultas testadas nas redes neurais.
	'activation': ['identity', 'logistic', 'tanh', 'relu']
	'alpha': [0.00001, 0.0001, 0.001, 0.01]
	'max_iter': [500]
	'batch_size': ['auto']
	'learning_rate': ['constant']
	'learning_rate_init': [0.001, 0.01, 0.1, 0.3]
	'early_stopping': [True]
(MLS)'momentum' / (MLA)'beta_1': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]	

Legenda: *: "np." representa a importação do módulo *numpy*. **: os valores de *gamma* testados para o SVM polinomial foram diferentes dos demais. Isso ocorreu pela remoção dos valores superiores que levavam a execuções intermináveis e demandavam cancelamento.

Fonte: Autoria própria.

Tabela 9 – Descrição das camadas ocultas utilizadas nos modelos de MLP.

Hiperparâmetro	Possíveis valores
<i>'hidden_layer_sizes':</i>	[10], [20], [40], [60], [80], [100], [10, 10], [10, 10, 10], [10, 10, 10, 10], [10, 10, 10, 10, 10], [20, 20], [20, 20, 20], [20, 20, 20, 20], [20, 20, 20, 20, 20], [40, 40], [40, 40, 40], [40, 40, 40, 40], [40, 40, 40, 40, 40], [60, 60], [60, 60, 60], [60, 60, 60, 60], [60, 60, 60, 60, 60], [80, 80], [80, 80, 80], [80, 80, 80, 80], [80, 80, 80, 80, 80], [100, 100], [100, 100, 100], [100, 100, 100, 100], [100, 100, 100, 100, 100], [100, 80], [100, 80, 60], [100, 80, 60, 40], [100, 80, 60, 40, 20], [80, 60], [80, 60, 40], [80, 60, 40, 20], [80, 60, 40, 20, 10], [60, 40], [60, 40, 20], [60, 40, 20, 10], [40, 20], [40, 20, 10], [80, 40], [80, 40, 20], [20, 10], [128], [64], [32], [16], [8], [8, 8], [8, 8, 8], [8, 8, 8, 8], [8, 8, 8, 8, 8], [16, 16], [16, 16, 16], [16, 16, 16, 16], [16, 16, 16, 16, 16], [32, 32], [32, 32, 32], [32, 32, 32, 32], [32, 32, 32, 32, 32], [64, 64], [64, 64, 64], [64, 64, 64, 64], [64, 64, 64, 64, 64], [128, 128], [128, 128, 128], [128, 128, 128, 128], [128, 128, 128, 128], [128, 64], [128, 64, 32], [128, 64, 32, 16], [128, 64, 32, 16, 8], [64, 32], [64, 32, 16], [64, 32, 16, 8], [32, 16], [32, 16, 8], [16, 8]

Fonte: Autoria própria.

Inicialmente, o objetivo era de realizar a busca de parâmetros utilizando o “*sklearn.model_selection.GridSearchCV*” mas, apesar dele exibir no terminal os avisos (*warnings*) que ocorriam durante a construção dos modelos, essa classe era incapaz de capturar os avisos e retorná-los na tabela de resultados finais. Portanto, era impossível identificar qual conjunto de parâmetros retornou o aviso durante a construção do modelo. Isso era particularmente importante para os modelos de SVM e MLP, uma vez que o aviso de falha de convergência foi um critério de exclusão do modelo (suas métricas de validação foram zeradas). Mesmo programando explicitamente a conversão desse aviso em erro e modificando o parâmetro “*error_score*”, o *GridSearchCV* não conseguia zerar as métricas ou retornar o problema de convergência na tabela. Isso erro ocorre quando o *GridSearchCV* é utilizado para trabalhar em paralelo entre os núcleos do processador e é devido ao fato de que o ambiente de *warnings* do Python é reiniciado toda vez que inicia o trabalho em um novo núcleo. Esse problema é conhecido no GitHub do Scikit-learn desde 2019, sob a numeração 12939 (<https://github.com/scikit-learn/scikit-learn/issues/12939>), e diferentes estratégias têm sido utilizadas para contorná-lo.

A estratégia desenvolvida para contornar o problema mencionado foi a de escrever um conjunto de módulos em Python, utilizando principalmente as bibliotecas Scikit-learn 1.2.0 (Pedregosa *et al.*, 2011) e RDKit 2023.3.2 (Landrum, 2013), de forma a viabilizar a construção dos modelos de aprendizado de máquina e o cálculo das métricas de validação com todas as diferentes combinações de hiperparâmetros, utilizando paralelização e permitindo o rastreamento dos avisos de convergência e

demais avisos relacionados. Ou seja, realizar não só a programação de um módulo de *Grid Search*, mas escrever um conjunto de módulos capaz de realizar as tarefas de aprendizado de máquina por completo. Além disso, utilizando o tqdm 4.64.1 (Costa-Luis *et al.*, 2023), também foi implementada uma barra de progresso com porcentagem de andamento, tempo decorrido e de término estimado para acompanhar o andamento da construção e validação dos modelos de aprendizado de máquina.

Todos os modelos foram validados internamente por validação cruzada *5-fold* e por validação externa com o conjunto de teste. Para isso, foram calculadas as seguintes métricas nas duas validações: coeficiente de correlação de Matthews (MCC), F1-Score, taxa de verdadeiros positivos ou sensibilidade (TPR), taxa de verdadeiros negativos ou especificidade (TNR), acurácia balanceada (bACC, ideal para conjuntos de dados não balanceados entre as classes), área sob a curva ROC (AUC-ROC) e o coeficiente Kappa de Cohen (CK) (Mosley, 2013; Roy; Kar; Das, 2015; Lipiński; Szurmak, 2017; Chicco; Warrens; Jurman, 2021).

A validação *5-fold* consiste na separação arbitrária do conjunto de treinamento em cinco grupos, cada um deles representando 20% desse conjunto. Em cada separação, um grupo é utilizado para validação do modelo e os demais são utilizados conjuntamente para treinar o modelo. Dessa forma, a média aritmética das cinco replicatas para as métricas calculadas com o conjunto de validação deixado de fora é utilizada como valor final da métrica de validação interna. Por outro lado, na validação externa, o modelo é construído utilizando integralmente o conjunto de treinamento. A validação externa é então realizada utilizando o conjunto de teste, que se trata de moléculas que não foram utilizadas para treinar o modelo.

A escolha dos melhores modelos de cada técnica foi feita pela análise da média entre os valores de MCC interno e externo. Posteriormente, foi realizada a escolha dos três melhores modelos da proteína de cada espécie de bactéria (saFabI e ecFabI). Nesse caso, foram consideradas todas as métricas mencionadas. A **Figura 13** traz as equações e representações das métricas mencionadas, com a exceção da AUC-ROC, pois essa métrica foi previamente descrita na seção **3.3.1 Estratégias de validação do acoplamento molecular (p. 52)**.

Figura 13 – Definição das métricas de validação utilizadas.

	Preditos ativos	Preditos inativos
Experimentalmente ativos	Verdadeiros positivos (TP)	Falsos negativos (FN)
Experimentalmente inativos	Falsos positivos (FP)	Verdadeiros negativos (TN)

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$F1-Score = \frac{2TP}{2TP + FP + FN}$$

$$TPR = \frac{TP}{TP + FN}$$

$$TNR = \frac{TN}{TN + FP}$$

$$bACC = \frac{(TPR + TNR)}{2}$$

$$CK = \frac{2 \times (TP \times TN - FP \times FN)}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

Fonte: Autoria própria.

O MCC é uma métrica que considera por completo a matriz de confusão, fornecendo uma análise completa e geral da capacidade preditiva dos modelos. Além disso, é pouco susceptível ao efeito do desbalanceamento do conjunto de dados entre as duas classes (no caso, ativo e inativo). Os valores de MCC são fixados entre -1 e +1, sendo que o valor de +1 indica predição perfeitamente correta (classificar todos os ativos como ativos e todos os inativos como inativos), -1 indica a predição perfeitamente incorreta (classificar todos os ativos como inativos e todos os inativos como ativos) e 0 significa que a performance do modelo é aleatória. Em geral, valores de MCC superiores a 0,5 são considerados como aceitáveis, indicando uma

performance robusta de predição (Marzo; Benfenati, 2018; Emmert-Streib; Moutari; Dehmer, 2019; Chicco; Jurman, 2020, 2023; Chicco; Warrens; Jurman, 2021).

O F1-Score é uma métrica de validação que avalia a performance de classificação dos modelos privilegiando a classe positiva (das moléculas ativas), porque se trata de uma média harmônica entre a taxa de verdadeiros positivos (TPR) e o valor preditivo positivo (PPV). Ela avalia três dos quatro erros fundamentais (TP, FP e FN) e é uma métrica que sofre com viés pelo desbalanceamento do conjunto de dados. Os valores de F1-Score variam de 0, quando todas as amostras da classe positiva são classificadas incorretamente, a 1, quando a classificação é perfeita $FN = FP = 0$. Em geral, assim com as demais métricas que variam de 0 a 1, como por exemplo, a acurácia balanceada (bACC), a taxa de verdadeiros positivos (TPR) e a taxa de verdadeiros negativos (TPR), é comum considerar como aceitáveis valores acima de 0,6. Uma exceção é a AUC_{ROC} onde valores acima de 0,5 são comumente aceitos. Entretanto, é importante ressaltar que, apesar de serem os valores mais comumente aceitos, esses limites de aceitabilidade podem variar dependendo das características do problema modelado (Emmert-Streib; Moutari; Dehmer, 2019; Chicco; Jurman, 2020, 2023).

Por fim, o coeficiente de Cohen Kappa (CK) é uma métrica que varia de -1 a +1 e avalia o quão bem um modelo é capaz de realizar classificações quando comparado com uma classificação aleatória. Assim, como o MCC é uma métrica balanceada e os valores de +1 indicam predição perfeitamente correta, 0 predição aleatória e -1 predição perfeitamente incorreta. Valores de CK superiores a 0,6 são resultados considerados como de concordância aceitável entre o predito e o experimental. Entretanto, os valores de CK tendem a ficar mais próximos do 0 do que o MCC, sendo sempre menores que os do MCC. Além disso, o MCC possui melhor capacidade de distinguir uma predição perfeitamente incorreta ($MCC = -1$) de uma obtida por acaso ($MCC = 0$). Isso porque o coeficiente de Cohen Kappa pode resultar em valores maiores do que -1 mesmo para predições perfeitamente incorretas, o que indica que o MCC é uma métrica mais informativa e permite uma interpretação mais realista (McHugh, 2012; Chicco; Warrens; Jurman, 2021).

Além das métricas calculadas para todos os modelos construídos, os três melhores modelos da proteína de cada espécie de bactéria (saFabI e ecFabI) foram também validados pelo método do *X-scrambling*, que é baseado no algoritmo do SCRAMBLE'N'GAMBLE (Lipiński; Szurmak, 2017) e consiste na aleatorização dos

valores de todas as variáveis independentes (no caso, dos *bits* de *fingerprint*) seguida pela construção do modelo. Espera-se que os modelos obtidos por essa técnica, por serem aleatórios, tenham valores inadequados na métrica de validação escolhida para análise. No caso, tanto o MCC interno quanto o externo foram escolhidos para análise. Espera-se encontrar baixos valores de MCC nos modelos gerados pelo *X-scrambling* e alto valor no modelo obtido pelo conjunto de dados preparado (modelo de referência), indicando que as predições do modelo de referência não foram obtidas pelo acaso. Para cada modelo foram feitas 100 randomizações conjuntas de *X-scrambling* e os valores dos dois MCC foram representados em um gráfico de dispersão.

Por fim, para comparar os resultados obtidos pelo *docking* com os resultados obtidos pelos modelos de aprendizado de máquina, foi construída também a curva ROC e realizados os cálculos de AUC-ROC, dos fatores de enriquecimento e de AUC-BEDROC utilizando a probabilidade das moléculas de serem da classe ativa. Além disso, para permitir comparação entre as duas estratégias, os valores de pontuação do *docking* foram classificados utilizando três diferentes estratégias: (i) utilizando como limite de classificação a média aritmética entre o menor valor da classe inativa e o maior valor da classe ativa, (ii) utilizando como limite de classificação a média aritmética entre a média de todos os valores da classe ativa e a média de todos os valores da classe inativa e, por último, (iii) utilizando o limiar fornecido pela curva ROC que maximiza a diferença entre a taxa de verdadeiros positivos (TPR) e a taxa de falsos positivos (FPR). A estratégia que resultou em melhores valores de MCC interno e externo para o *docking* foi utilizada para comparação com os modelos de aprendizado de máquina.

4.9 INTERPRETAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA E DAS PRINCIPAIS CARACTERÍSTICAS DE INTERAÇÃO LIGANTE-PROTEÍNA

Para interpretação dos modelos de aprendizado de máquina e identificação das variáveis que mais contribuíram para esses modelos, foi utilizado um algoritmo de permutação para definir a importância das variáveis independentes para a predição (Mi *et al.*, 2021). O conjunto de teste teve cada uma de suas variáveis independentes aleatorizadas separadamente 100 vezes antes de aplicar a transformação (PCA) e, em cada uma dessas repetições, foi calculado o MCC_{externo} utilizando os três melhores

modelos de cada proteína. A média do valor absoluto da diferença entre o MCC_{externo} do modelo de referência (melhor modelo) e cada MCC_{externo} das 100 repetições com variável aleatorizada foi utilizada para hierarquizar as variáveis independentes em termos de contribuição para o modelo de aprendizado de máquina. Dessa forma, as 10 variáveis que mais contribuíram para a alteração do valor de MCC foram selecionadas e, utilizando o LUNA, foram rastreadas as informações estruturais que cada *bit* de *fingerprint* representava. Também foram analisadas se essas variáveis estavam majoritariamente presentes entre moléculas ativas ou inativas e, com isso, uma representação em gráfico de barras foi construída com esses dados.

4.10 TRIAGEM VIRTUAL DA BIBLIOTECA DE FÁRMACOS APROVADOS PELO FDA E DA BRACOLI

Para a triagem virtual, 1913 moléculas da biblioteca BraCoLi (Veríssimo *et al.*, 2022a) e 1430 moléculas da biblioteca de fármacos aprovados pelo FDA disponível no ZINC (Sterling; Irwin, 2015) foram preparadas no LigPrep (Shelley *et al.*, 2007; Schrodinger Release 2022-3, 2022), utilizando o campo de força OPLS4 (Lu *et al.*, 2021), para a obtenção de um arquivo “.sdf” com as estruturas tridimensionais de menor energia nos estados de ionização significativamente favoráveis em $\text{pH} = 7,4 \pm 2,0$. Os ligantes dessas bibliotecas foram então acoplados as duas proteínas de interesse utilizando os melhores protocolos de *docking* para cada proteína, os *fingerprints* de interação foram calculados utilizando o LUNA 0.11.6 (Fassio *et al.*, 2022) e as predições de atividade foram realizadas com os modelos de aprendizado de máquina.

Para a análise dos resultados da triagem virtual foi realizado um consenso, sendo selecionadas as moléculas que foram ativas em pelo menos dois dos três melhores modelos de cada proteína. Para a proteína FabI de *S. aureus* um segundo filtro foi aplicado devido ao alto número de moléculas, sendo mantidas apenas as moléculas que foram ativas nos três modelos da saFabI. Todas as moléculas selecionadas tiveram suas poses no *docking* avaliadas por inspeção visual e as moléculas mais promissoras tiveram sua aquisição consultada. Foram obtidas para os ensaios *in vitro* as moléculas com disponibilidade na BraCoLi e, para a biblioteca das moléculas aprovadas pelo FDA, as disponíveis no Laboratório de Controle de Qualidade da Faculdade de Farmácia da Universidade Federal de Minas Gerais.

4.11 AVALIAÇÃO DO DOMÍNIO DE APLICABILIDADE PARA O APRENDIZADO DE MÁQUINA

O domínio de aplicabilidade para os três melhores modelos de aprendizado de máquina de cada proteína foi avaliado pela abordagem de caixa delimitadora usando análise de componentes principais (PCA) (Sahigara *et al.*, 2012; Fernandes *et al.*, 2021). Os *fingerprints* utilizados na construção dos melhores modelos foram submetidos a duas metodologias diferentes:

- (a) Redução de dimensionalidade por PCA para 3 dimensões, visando a visualização das moléculas dentro de um espaço tridimensional e fornecer uma visão do domínio de aplicabilidade.
- (b) Redução de dimensionalidade por PCA para número de dimensões que expliquem 85% da variância presente no conjunto de dados, com o objetivo de realizar o cálculo de cinco métricas de distância/similaridade (Euclidiana, Manhattan, Cosine, Wasserstein e Jaccard) entre cada molécula e o centro (coordenadas dadas pelas médias das variáveis de todas as amostras do conjunto de treinamento). Caso as moléculas tivessem mais de 95% dos seus descritores com distâncias superiores as observadas no conjunto de treinamento, ela seria considerada fora do domínio de aplicabilidade para essa métrica de distância. Posteriormente, se a molécula tivesse sido indicada como fora do domínio de aplicabilidade para a maioria das métricas (pelo menos 3 métricas), ela seria finalmente classificada como fora do domínio de aplicabilidade.

O domínio de aplicabilidade foi avaliado para os conjuntos de teste e para as bibliotecas da triagem virtual. Para isso, os cálculos e os ajustes da PCA pelo método “.*fit()*” foram sempre realizados com o conjunto de treinamento e as transformações “.*transform()*” aplicadas para todos os conjuntos trabalhados (treinamento, teste e as bibliotecas de triagem virtual).

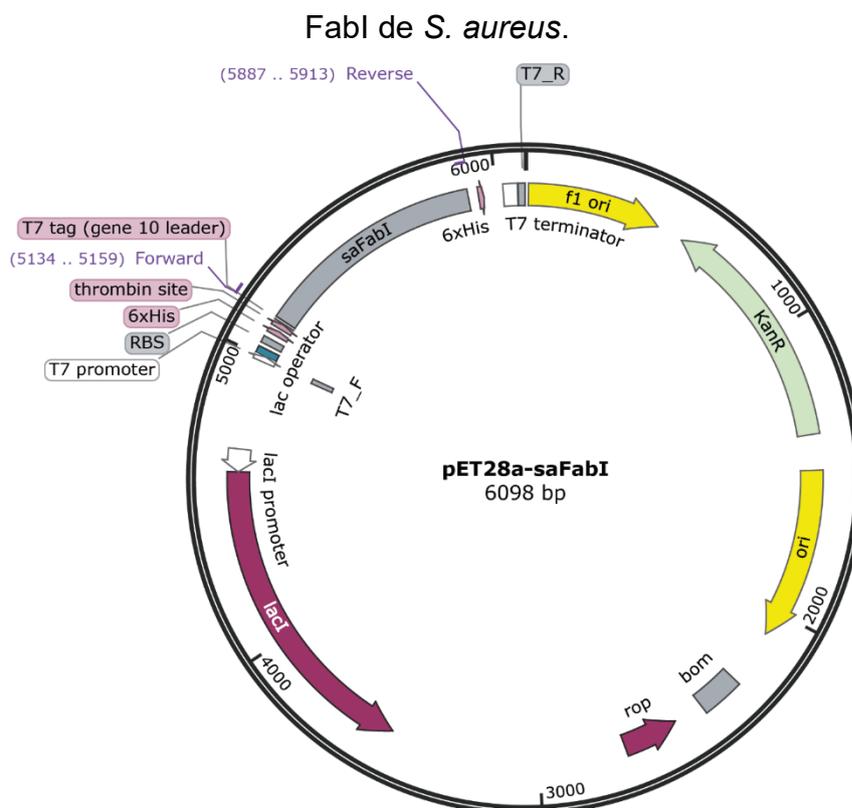
4.12 OBTENÇÃO DOS VETORES DE EXPRESSÃO DAS PROTEÍNAS RECOMBINANTES

O presente trabalho foi desenvolvido com uso do patrimônio genético e, portanto, foi cadastrado no Sistema Nacional de Gestão de Patrimônio Genético e do Conhecimento Tradicional Associado (SisGen) sob o número A59F846. Toda a metodologia de expressão, purificação e caracterização das proteínas recombinantes foi desenvolvida no Laboratório de Biologia Estrutural e Biotecnologia do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais.

4.12.1 Obtenção do vetor da FabI de *Staphylococcus aureus*

O vetor de expressão pET28a-saFabI (**Figura 14**) contendo a sequência gênica codificante da enzima saFabI (**Apêndice: Figura A.1, p. 201**) e o gene de resistência para canamicina (KanR) foi fornecido pelo Prof. Rob Lavigne e pela Profa. Joleen Masschelein, por meio de um acordo de transferência de material (MTA/*Material Transfer Agreement*) entre a *Katholieke Universiteit Leuven* e a Universidade Federal de Minas Gerais, para o Prof. Vinícius Gonçalves Maltarollo (FAFAR-UFMG) (Mattheus *et al.*, 2010; Fage *et al.*, 2020). Células de *E. coli* DH5 α , cepa utilizada para replicação de plasmídeo, foram previamente transformadas com o vetor recebido (Al-Janabi *et al.*, 2022). Dessa forma, essas células foram recebidas sendo necessária a extração do vetor para a posterior transformação da cepa de expressão BL21(DE3), que possui a RNA polimerase T7 e permite a expressão da proteína recombinante após a adição de IPTG.

Figura 14 – Mapa do vetor pET28a-saFabI utilizado para a expressão da proteína



Fonte: Representação realizada no SnapGene Viewer 7.0.2 com os arquivos cedidos pelo Prof. Rob Lavigne e pela Profa. Joleen Masschelein.

4.12.1.1 Preparo de inóculo para extração de DNA plasmidial

Primeiramente, um microtubo contendo células transformadas com o vetor pET28a-saFabI foi retirado do freezer -80°C , descongelado e seu conteúdo ($100\ \mu\text{L}$) foi inteiramente transferido para um tubo Falcon de 10 mL de meio LB (Luria-Bertani) líquido (NaCl 1% m/v, extrato de levedura 0,5% m/v, peptona 1% m/v). O tubo com o inóculo foi incubado juntamente com um tubo de controle negativo (meio LB líquido) em incubadora *shaker* digital modelo MA832 (Marconi) a 37°C e 200 rpm por 16 horas.

Após o período de incubação, procedeu-se a extração plasmidial utilizando o kit comercial *PureYield™ Plasmid Miniprep System* (Promega) conforme manual do próprio kit (Promega Corporation, 2009) e no ambiente controlado pela cabine de segurança biológica Esco Class II Airstream (Esco).

4.12.1.2 Extração do DNA plasmidial utilizando o PureYield™ Plasmid Miniprep System

Primeiramente, procedeu-se a etapa de lise celular para obtenção do DNA plasmidial, onde foram adicionados 1,5 mL do inóculo preparado no item anterior em um microtubo estéril de 2 mL sob banho de gelo. Esse microtubo foi submetido ao processo de centrifugação por 1 minuto à 4 °C e 8000 g, em minicentrífuga refrigerada Fresco 17 (Haereus), para decantação das células. Ao final dessa centrifugação, o sobrenadante foi descartado, foram adicionados mais 1,5 mL do inóculo no mesmo microtubo e uma nova centrifugação foi realizada para obter um volume final de inóculo centrifugado de 3 mL. O volume restante do inóculo foi utilizado para preparo do banco de células DH5α conforme metodologia da seção **4.15 Transformação das células de *E. coli* e preparo do banco de células (p. 92)**.

O sobrenadante final foi descartado para a secagem do *pellet*, que foi ressuscitado em 600 µL de água Mili-Q estéril livre de DNase e RNase, ainda em banho de gelo. Em seguida, 100 µL do tampão de lise (disponível no kit comercial) foram adicionados ao microtubo que foi homogeneizado cuidadosamente por inversão seis vezes. Posteriormente, 350 µL de solução neutralizante gelada (disponível no kit comercial) foram adicionados e novamente procedeu-se a homogeneização por seis inversões. Finalmente, o microtubo foi então centrifugado por 3 minutos a 4 °C e 8000 g.

O sobrenadante foi então transferido para uma minicoluna *PureYield™*, evitando sua contaminação com debris celulares. Essa minicoluna foi transferida para um tubo coletor e centrifugada a 4 °C e 8000 g por 30 segundos. O filtrado no tubo foi descartado e a coluna foi novamente posicionada no tubo coletor para a etapa de lavagem.

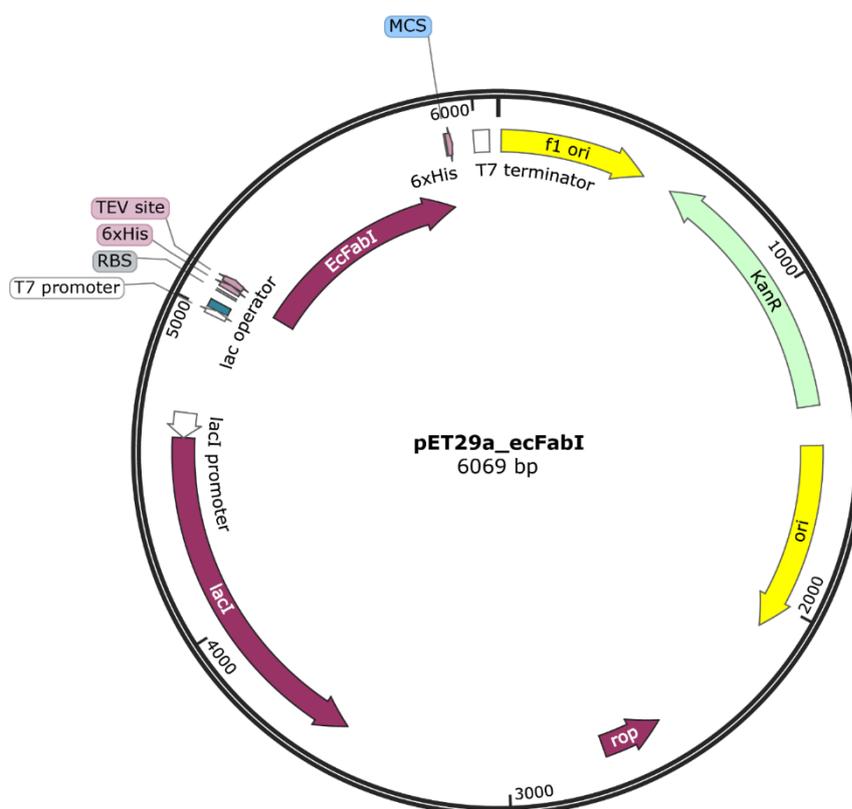
A etapa de lavagem foi realizada pela adição de 200 µL da solução removedora de endotoxinas (disponível no kit) à minicoluna, seguida de centrifugação a 4° C e 8000 g por 30 segundos e adição de 400 µL da solução de lavagem (disponível no kit), sendo repetido o processo de centrifugação após a adição da solução de lavagem.

A minicoluna foi então colocada em um microtubo estéril de 1,5 mL estéril para a etapa de eluição, onde foram adicionados 20 µL de água Mili-Q estéril livre de RNase e DNase, ficando sob repouso por 1 minuto à temperatura ambiente. O microtubo foi então centrifugado a 4 °C e 8000 g por 60 segundos, a minicoluna foi descartada e o DNA eluído foi dosado e armazenado a -20 °C.

4.12.2 Obtenção do vetor da FabI de *Escherichia coli*

O vetor de expressão pET29a-ecFabI (Figura 15) contendo a sequência gênica codificante da enzima ecFabI (Apêndice: Figura A.2, p. 202) e o gene de resistência para canamicina (KanR) foi adquirido da empresa GenOne e recebido em um criotubo.

Figura 15 – Mapa do vetor pET29a-ecFabI utilizado para a expressão da proteína FabI de *E. coli*.



Fonte: Representação realizada no SnapGene Viewer 7.0.2.

O criotubo contendo 5 µg do vetor de expressão pET29a-ecFabI foi retirado do freezer a -20°C e suavemente descongelado em banho de gelo. Em seguida, foi submetido a uma centrifugação a 6000 g por 1 minuto sob temperatura de 4 °C, utilizando a minicentrífuga refrigerada Fresco 17 (Haereus). No ambiente controlado da cabine de segurança biológica Esco Class II Airstream (Esco), o plasmídeo foi ressuspensionado em 25 µL de água Mili-Q estéril livre de DNase e RNase. Este processo foi finalizado pela agitação em agitador IKA MS3 (IKA) a baixa velocidade por 30 segundos e aquecimento em banho seco Thermo-Shaker TS-100 (Biosan) a 50 °C por 15 minutos, visando assegurar a dissolução do plasmídeo e finalizando o

preparo da solução estoque. Posteriormente, aguardou-se o tubo atingir temperatura ambiente para transferir para banho de gelo.

Com a solução estoque do plasmídeo pET29a_ecFabI pronta, procedeu-se a preparação da solução de trabalho. Para isso, 5 μ L da solução estoque foram adicionados a um microtubo com 5 μ L de água Mili-Q estéril livre de DNase e RNase. O restante da solução estoque foi armazenado à temperatura de -20 °C e o conteúdo do microtubo da solução de trabalho foi homogeneizado e mantido em banho de gelo para a realização da transformação das células.

4.13 DOSAGEM DE DNA DOS VETORES

A massa do vetor para a ecFabI sintetizada foi de 5 μ g. A solução de trabalho foi preparada de forma a obter uma concentração de 100,0 ng/ μ L.

No caso da saFabI, foi preciso determinar a concentração de DNA eluída na MiniPrep. Para isso, utilizou-se o espectrofotômetro NanoDrop® ND-1000 (NanoDrop Technologies, Inc.) que determina a absorvância a 260 nm e quantifica a concentração de DNA na amostra. Além disso, fornece o valor da razão das leituras de absorvância a 260 nm (detecção de DNA) e 280 nm (detecção de proteína), onde valores entre 1,80 e 2,00 indicam a pureza da amostra (Costa, 2014). Dessa forma, a concentração da saFabI foi de 166,1 ng/ μ L com razão 260/280 de 1,96, indicando amostra pura dentro dos limites estabelecidos, podendo ser utilizada para a transformação das células de *E. coli*.

4.14 PRODUÇÃO DE CÉLULAS ELETROCOMPETENTES

A preparação de células de *E. coli* competentes das cepas DH5 α (Life Technologies) e BL21(DE3) (*Coli Genetic Stock Center at Yale* - CGSC) foi realizada para permitir a transformação dessas células visando o estabelecimento do banco de células para expressão das proteínas recombinantes.

Inicialmente, um pré-inóculo da cepa de interesse foi feito em 5 mL de meio LB líquido, sem antibacteriano e incubado por 16 horas a 37 °C e 200 rpm. Após o crescimento, o pré-inóculo foi utilizado para preparar um inóculo de 500 mL, utilizando uma proporção de 1:100 de cultura para meio LB líquido. O inóculo com a cepa escolhida cresceu a 37 °C e 200 rpm até atingir leituras de absorvância a 600 nm entre

0,60 e 0,80, sendo realizadas em espectrofotômetro WPA CO8000 (Biochrom). Após alcançar essa leitura, a cultura foi mantida a uma temperatura de 0 °C por 20 minutos. Em todas as etapas subsequentes, sempre que possível, as células foram mantidas a uma temperatura de 0°C e todos os recipientes foram previamente resfriados em gelo antes do uso.

Assim, as células foram então submetidas à centrifugação a 4000 *g* por 15 minutos a 4 °C, e o sobrenadante foi descartado cuidadosamente. O *pellet* resultante foi ressuscitado em 500 mL de glicerol 10% v/v estéril e gelado. Outros três ciclos de centrifugação e ressuspensão foram realizados, respectivamente, em 250 mL, 20 mL e 2 mL de glicerol 10%, conforme descrito acima. Por fim, as células ressuscitadas em 2 mL de glicerol 10% gelado foram imediatamente aliqüotadas em aliqüotas de 50 µL por microtubo e, em seguida, foram armazenadas em freezer -80 °C.

4.15 TRANSFORMAÇÃO DAS CÉLULAS DE *E. coli* E PREPARO DO BANCO DE CÉLULAS

O processo de transformação das células eletrocompetentes foi feito para estabelecimento do banco de células de *E. coli* DH5α e BL21(DE3) transformadas com os vetores pET28a-saFabI e pET29a-ecFabI.

Para isso, foi descongelado, em banho de gelo, um microtubo com 50 µL de células de *E. coli* eletrocompetentes previamente preparadas para eletroporação (conforme discutido na seção **4.14 Produção de células eletrocompetentes (p. 91)**). Simultaneamente, uma cubeta de 0,2 cm foi mantida refrigerada para o procedimento. Após isso, adicionou-se 1 µL da suspensão com o vetor ao microtubo, que foi incubado em banho de gelo por 1 minuto. Em seguida, todo o conteúdo foi transferido com cuidado para a cubeta, garantindo que o líquido alcançasse a parte inferior. A cubeta foi então introduzida no eletroporador MicroPulser (Bio-Rad), onde o pulso elétrico foi aplicado para permitir a transformação.

Em cabine de segurança biológica, foram adicionados 800 µL de meio LB líquido estéril na cubeta, para permitir a recuperação da viabilidade celular. O conteúdo foi homogeneizado utilizando uma micropipeta e transferido para um novo microtubo estéril, sendo então incubado a 37 °C por 1 hora, sob agitação de 200 rpm em *shaker* digital modelo MA832 (Marconi). Após o período de incubação, o material

foi centrifugado a 1500 g por 5 min em temperatura ambiente e o sobrenadante foi descartado.

Ainda em cabine de segurança biológica, o *pellet* foi ressuspenso em 100 µL de meio LB líquido estéril e plaqueado com alça de Drigalski em uma placa de Petri contendo 20 mL de ágar LB (peptona 1% m/v, extrato de levedura 0,5% m/v, NaCl 0,5% m/v, ágar 1,5% m/v) e canamicina a 50 µg/mL (antimicrobiano de seleção). Simultaneamente, uma placa de Petri contendo 20 mL de ágar LB foi mantida aberta durante todo o processo de plaqueamento para servir como controle negativo. As placas então foram incubadas invertidas em estufa bacteriológica 520 (Fanem) a 37 °C durante 16 horas.

Após a transformação, no ambiente da cabine de segurança biológica, foram selecionadas da placa três colônias isoladas e um *pool* de colônias (três colônias muito próximas na placa). Cada uma das colônias isoladas e o *pool* foram introduzidos em tubos Falcon de 50 mL separados, contendo cada um 5 mL de meio líquido LB estéril e canamicina na concentração final de 50 µg/mL, visando a seleção das células transformadas. O controle negativo com 5 mL de meio líquido LB estéril foi mantido aberto na cabine de segurança biológica durante todo o preparo dos inóculos. Os tubos foram então fechados e incubados em *shaker* digital modelo MA832 (Marconi) sob temperatura de 37 °C e agitação de 200 rpm por 16 horas.

Após o período de incubação, sob condições estéreis, foram preparados os microtubos identificados para cada cultivo. Em cada microtubo, foram adicionados 500 µL de cada cultivo e 500 µL de solução estéril de glicerol 50% em LB líquido. Durante todo o processo, as células foram mantidas em banho de gelo e, ao final, os microtubos foram armazenados em um *ultrafreezer* a -80 °C, estabelecendo os bancos de células transformadas para expressão das proteínas recombinantes.

4.16 EXPRESSÃO DAS PROTEÍNAS RECOMBINANTES

Para a expressão das proteínas recombinantes saFabI e ecFabI foram realizados pré-inóculos em meio LB líquido contendo canamicina na concentração final de 50 µg/mL. Para isso, foram adicionadas as células BL21(DE3) transformadas com o respectivo plasmídeo de interesse ao meio LB líquido na proporção de 1:100 em um tubo Falcon, que foi incubado em *shaker* MA 832 (Marconi) por 16 horas a 37 °C, sob agitação de 200 rpm.

Em seguida, um inóculo foi feito na proporção de 1:100 em meio LB líquido com canamicina (concentração final de 50 µg/mL) e sulfato de magnésio (concentração final de 10 mM). O inóculo foi feito em Erlenmeyer chanfrado de 2 L com volume de cultura correspondendo a 50% da capacidade do Erlenmeyer, visando fornecer agitação e suprimento de oxigênio suficiente para o crescimento a 37 °C e 200 rpm. O crescimento foi monitorado em espectrofotômetro WPA CO8000 (Biochrom) e foi mantido até atingir leituras de absorbância a 600 nm entre 0,60 e 0,80. Após atingir a densidade óptica a 600 nm (DO₆₀₀) desejada, a temperatura do *shaker* foi reduzida para 18 °C e foi adicionado o indutor IPTG (isopropil-β-D-tiogalactopiranosídeo) para concentração final de 0,5 mM. A indução da expressão foi feita a 18 °C, por 18 horas sob rotação de 200 rpm. Após a expressão das proteínas, as culturas foram centrifugadas a 8000 g por 30 minutos a 4 °C em centrífuga Multifuge X3R (Thermo Scientific).

A temperatura de expressão (18 °C) foi definida com base no conhecimento de que o processo de agregação de proteínas e formação de corpos de inclusão pode ser minimizado pela redução da temperatura pós-indução (Papaneophytou; Kontopidis, 2014). Além disso, várias referências na literatura sobre expressão da FabI utilizam temperaturas mais baixas, entre 16 e 18 °C (Priyadarshi; Kim; Hwang, 2010; Schiebel *et al.*, 2012; Fage *et al.*, 2020). O tempo de expressão também foi estabelecido conforme a mesma bibliografia e ensaios previamente realizados e estabelecidos no laboratório.

4.17 LISE BACTERIANA

Após a centrifugação da cultura expressa, as células foram congeladas a -20 °C para lise posterior ou diretamente ressuspensas em tampão de lise (tris-HCl 50 mM pH 7,4, sacarose 1% p/v, Tween 20 1% v/v, glicerol 1% v/v). O volume de tampão de lise utilizado foi na proporção de 15 mL por grama de *pellet* centrifugado. Após a ressuspensão do *pellet*, foram adicionados lisozima e PMSF (fluoreto de fenilmetanosulfonil) na concentração final de 100 µg/mL e 1 mM, respectivamente. A suspensão foi mantida no gelo por 30 minutos e, em seguida, o material foi sonificado com amplitude de 30%, com 6 pulsos de 15 segundos ligado e intervalos de 1 min desligado, em sonicador Fisher Scientific Sonic Dismembrator modelo 500. Após a

lise, o extrato celular foi centrifugado a 10.000 g por 30 minutos duas vezes, transferindo o sobrenadante para um novo frasco para a segunda centrifugação.

As amostras do sobrenadante (fração solúvel) em tampão de lise foram aplicadas em gel de eletroforese para análise, enquanto as amostras do *pellet* da lise (fração insolúvel) precisaram ser solubilizadas em tampão de solubilização (tris-HCl 50 mM pH 8,0, ureia 8 M e NaCl 500 mM) para análise por SDS-PAGE.

4.18 IDENTIFICAÇÃO DE PROTEÍNAS RECOMBINANTES POR SDS-PAGE

A eletroforese em gel de poliacrilamida em condições desnaturantes com dodecil sulfato de sódio (SDS-PAGE) é um método frequentemente utilizado para determinar a massa molecular de proteínas, uma vez que a desnaturação promovida pelo detergente aniônico forte (SDS) em combinação com agente redutor (β -mercaptoetanol) e calor (100 °C) permite a desagregação das subunidades da proteína que ficam carregadas negativamente. Como a carga intrínseca das proteínas é normalizada e todas as proteínas ficam com a mesma relação massa/carga, a migração da proteína no gel de poliacrilamida se dá unicamente pelo seu tamanho.

A identificação das proteínas recombinantes e avaliação dos resultados obtidos na expressão, lise e purificação foi realizada por essa técnica e, para isso, foi primeiramente necessário a preparação do gel desnaturante de poliacrilamida.

O gel de poliacrilamida foi preparado em placas de vidro Mini-PROTEAN® (BioRad) sendo composto por dois géis: o de separação (com bis-acrilamida a 15%, tris-HCl pH 8,8 0,375 M, SDS 0,1% v/v, persulfato de amônio 0,1% v/v e tetrametiletilenodiamina 0,1% v/v) e o de empilhamento (com bis-acrilamida a 4%, tris-HCl pH 6,8 0,125 M, SDS 0,1% v/v, persulfato de amônio 0,1% v/v e tetrametiletilenodiamina 0,1% v/v). A solução de cada gel foi preparada no momento da confecção do gel de poliacrilamida. Em especial, o persulfato de amônio (APS) e a tetrametiletilenodiamina (TEMED) foram adicionados em cada solução no instante antes de verter cada gel para a placa, evitando a gelificação fora da placa. Aguardou-se a gelificação do gel de separação para adicionar o gel de empilhamento e, após a completa gelificação do gel de empilhamento, obteve-se o gel para eletroforese em condição desnaturante.

As amostras foram preparadas com tampão de amostra 4x (tris-HCl 200 mM pH 6,8, SDS 8,0% m/v, azul de bromofenol 0,4% m/v, glicerol 40% v/v e β -

mercaptoetanol 400 mM), na proporção de 1 parte de tampão de amostra 4x para 3 partes de amostra, e aquecidas sob agitação a 100 °C por 20 minutos. Posteriormente, as amostras preparadas e o padrão de peso molecular (Bio-Rad Precision Plus Protein™ Dual Color Standard) foram aplicados no gel de eletroforese preparado e submerso na cuba de eletroforese Mini-PROTEAN® Tetra Vertical Electrophoresis Cell (BioRad) em tampão de corrida 1x (glicina 192 mM, tris-HCl 25 mM pH 8,3, SDS 0,03% m/v). A corrida foi iniciada a 120 V utilizando a fonte de eletroforese PowerPac™ Basic Power Supply (BioRad).

Ao final da corrida, o gel foi retirado da placa e corado *overnight* em solução corante (Coomasie Blue R250 0,1% m/v, etanol 41,6% v/v, ácido acético 16,6% v/v) sob agitação, sendo descorado sob agitação em solução descorante (etanol 30% v/v, ácido acético 10% v/v) até que fosse possível a visualização do perfil de expressão das proteínas (o que levou até 4 horas). Alternativamente, o gel foi corado em prata para maior sensibilidade seguindo metodologia descrita no **Apêndice B (p. 203)**.

4.19 PURIFICAÇÃO DAS PROTEÍNAS RECOMBINANTES

Após a lise bacteriana, procederam-se as purificações das proteínas recombinantes saFabI e ecFabI em sistema ÄKTA pure™ (Cytiva). Para isso, foram utilizadas as colunas: HisTrap™ HP 5 mL (Cytiva) para a cromatografia de afinidade, HiPrep™ 26/10 (Cytiva) para a dessalinização e HiLoad™ 16/600 Superdex™ 200 pg (Cytiva) para a exclusão molecular. Nas cromatografias foram testados diferentes tampões utilizando a saFabI até a obtenção da proteína pura, solúvel e sem formação de agregado. A condição obtida para a proteína saFabI foi replicada para a ecFabI e a mesma análise de estado oligomérico foi realizada para ambas as proteínas visando verificar a formação de agregados.

No primeiro teste, foi utilizado o protocolo inicial de teste de purificação do laboratório. Para isso, foi realizada cromatografia de afinidade com coluna equilibrada em tampão de equilíbrio A1 (tris-HCl 50 mM, NaCl 500 mM e imidazol 30 mM, pH = 7,4) e eluição das proteínas com gradiente linear de imidazol de 30 a 500 mM em tampão de eluição B1 (tris-HCl 50 mM, NaCl 500 mM e imidazol 500 mM, pH = 7,4). As frações da cromatografia de afinidade foram reunidas e uma alíquota de 5 mL foi coletada e armazenada entre 2 e 8 °C. Outros 10 mL foram aplicados na coluna de dessalinização para troca de tampão com o tampão de dessalinização D1 (tris-HCl 50

mM e NaCl 50 mM, pH = 7,4). As frações após a dessalinização também foram recolhidas e reunidas. Entretanto, tanto a alíquota após afinidade quanto a alíquota após a dessalinização precipitaram. Sendo que a alíquota da afinidade precipitou após 5 dias e a da dessalinização precipitou *overnight*.

Visando o direcionamento da estratégia de purificação, realizou-se uma análise da literatura de protocolos de purificação da FabI a partir dos PDB IDs de proteínas FabI de *E. coli* e *S. aureus* disponíveis no PDB. Foram consultados os artigos referentes aos seguintes códigos de identificação do PDB: 4FS3 (Kaplan *et al.*, 2012); 7UMW e 7UM8 (Parker *et al.*, 2022); 4D41, 4D42, 4D43, 4D44, 4D45 e 4D46 (Schiebel *et al.*, 2015); 3GNS, 3GNT e 3GR6 (Priyadarshi; Kim; Hwang, 2010); 6YUR e 6YUU (Eltchkner *et al.*, 2021); 4NZ9 (Mehboob *et al.*, 2012); 6TBB e 6TBC (Fage *et al.*, 2020); 5CFZ, 5CG1 e 5CG2 (Jordan *et al.*, 2015). Assim, observou-se que diferentes estratégias obtiveram sucesso com diferentes tampões. Em geral, embora nas outras cromatografias tenha se observado tampões em outras faixas de pH, na afinidade foram utilizados tampões em faixas de pH entre 7,5 e 8,0, mas o pH 8,0 foi o predominante, até mesmo nas outras etapas. A grande maioria dos artigos utilizaram glicerol em concentrações de 5 a 10% nos tampões da cromatografia de afinidade, de dessalinização e de armazenamento. Alguns artigos utilizaram agentes redutores nas purificações, como o TCEP (tris(2-carboxietil)fosfina) e o DTT (ditiotreitól), principalmente na dessalinização e exclusão molecular. Com base nessas observações, um segundo teste foi realizado com o objetivo de obter a proteína pura e solúvel modificando os tampões para pH = 8,0 e adicionando glicerol nos tampões de equilíbrio, eluição e dessalinização.

No segundo teste, a cromatografia de afinidade foi realizada com coluna equilibrada em tampão de equilíbrio A2 (tris-HCl 50 mM, NaCl 500 mM, imidazol 30 mM e glicerol 10%, pH = 8,0) e eluição das proteínas com gradiente linear de imidazol de 30 a 500 mM em tampão de eluição B2 (tris-HCl 50 mM, NaCl 500 mM, imidazol 500 mM e glicerol 10%, pH = 8,0). As frações da cromatografia de afinidade foram reunidas, uma alíquota de 5 mL foi coletada e armazenada entre 2 e 8 °C e outros 10 mL foram aplicados na coluna de dessalinização para troca de tampão com o tampão de dessalinização D2 (tris-HCl 50 mM, NaCl 200 mM e glicerol 10%, pH = 8,0). As frações da coluna de dessalinização foram reunidas e concentradas em centrífuga com rotor fixo a 8000 g para aproximadamente 7 mL utilizando VivaSpin™ (Sartorius) com *cut-off* de 10 kDa e capacidade máxima de 20 mL. A amostra foi armazenada

entre 2 e 8 °C para realização da cromatografia de exclusão no dia seguinte. Dos 7 mL, somente 5 mL foram aplicados na coluna de exclusão molecular devido à capacidade máxima da coluna disponível, o restante foi armazenado entre 2 e 8 °C. A cromatografia de exclusão molecular em tampão de exclusão E1 (tris-HCl 50 mM e NaCl 200 mM, pH = 8,0) foi então executada e as frações reunidas foram concentradas com o VivaSpin™ para aproximadamente 1 mg/mL (a concentração de proteínas foi dada pela metodologia descrita na seção **4.20 Determinação do teor de proteínas e cálculo do rendimento final (p. 99)**) e armazenadas entre 2 e 8 °C até a realização dos ensaios de espalhamento dinâmico da luz (descritos na seção **4.21 Caracterização do estado oligomérico de proteínas por espalhamento dinâmico da luz (p. 100)**). Apesar da solubilidade da proteína, grande parte dela foi observada como agregado nos resultados de DLS e uma terceira purificação foi realizada com o objetivo de se obter a proteína pura, solúvel e sem agregados, pronta para a realização de ensaios de STD-NMR.

Para a realização dos ensaios de STD-NMR o ideal é utilizar como tampão substâncias que não interferem nos espectros de hidrogênio das amostras. Portanto, o Tris-HCl dos tampões de dessalinização e exclusão molecular foi substituído por tampão fosfato de potássio. Para evitar a formação de agregados, foi adicionado o DTT a 2 mM nesses mesmos tampões e a terceira metodologia de purificação envolveu:

- (a) cromatografia de afinidade realizada com coluna equilibrada em tampão de equilíbrio A2 (tris-HCl 50 mM, NaCl 500 mM, imidazol 30 mM e 10% glicerol, pH = 8,0) e eluição das proteínas com gradiente linear de imidazol de 30 a 500 mM em tampão de eluição B2 (tris-HCl 50 mM, NaCl 500 mM, imidazol 500 mM e 10% glicerol, pH = 8,0);
- (b) imediata dessalinização em coluna para troca de tampão com o tampão de dessalinização D3 (tampão fosfato de potássio 20 mM, NaCl 200 mM, glicerol 10% e DTT 2 mM, pH = 8,0);
- (c) concentração da amostra para aproximadamente 7 mL utilizando VivaSpin™ (Sartorius) com *cut-off* de 10 kDa e capacidade máxima de 20 mL;
- (d) aplicação de 5 mL de amostra em coluna de exclusão molecular e corrida em tampão de exclusão molecular E2 (tampão fosfato de potássio 20 mM, NaCl 200 mM e DTT 2 mM, pH = 8,0);

- (e) os 2 mL restantes da amostra de dessalinização, que não foram aplicados, foram armazenados de 2 a 8 °C como amostra de dessalinização de reserva para análises, como determinação do teor de proteínas e SDS-PAGE;
- (f) concentração da amostra de exclusão molecular para 5 mL (ecFabI) e 3 mL (saFabI) utilizando VivaSpin™ (Sartorius) com *cut-off* de 10 kDa e capacidade máxima de 20 mL e realização das análises do teor de proteínas, SDS-PAGE e de identificação do estado de oligomerização por DLS.

4.20 DETERMINAÇÃO DO TEOR DE PROTEÍNAS E CÁLCULO DO RENDIMENTO FINAL

A quantificação das proteínas em solução foi realizada através da medição da absorbância a 280 nm, seguindo a lei de Lambert-Beer (3).

$$A = \varepsilon * l * C \quad (3)$$

Onde A representa a absorbância a 280 nm, ε é o coeficiente de extinção molar ($M^{-1}cm^{-1}$), l é o comprimento da trajetória óptica em centímetros, e C é a concentração molar da proteína. As leituras de absorbância foram registradas utilizando o espectrofotômetro NanoDrop® ND-1000 (NanoDrop Technologies, Inc.), que gera os resultados de absorbância ajustados para o caminho óptico de 1,0 cm.

Os coeficientes de extinção molar (ε) das proteínas recombinantes foram obtidos pelo programa ProtParam (Swiss Institute of Bioinformatics, 2017) presente no servidor ExPASy Bioinformatics Resource Portal (disponível em: <https://web.expasy.org/protparam/>). O programa também fornece outros dados como a absorbância a 280 nm para solução aquosa da proteína em concentração de 1 mg/mL ($A_{280} 0,1\%$), a massa molecular da proteína (Da) e o número de aminoácidos. Os dados obtidos para as duas proteínas estão dispostos na **Tabela 10** e foram calculados com base na sequência completa da proteína, incluindo as caudas de histidina N-terminais e os sítios de clivagem para trombina (saFabI) e para protease de TEV (ecFabI), uma vez que não houve remoção dessas sequências. Com esses valores, foi possível obter a concentração final de cada proteína em $\mu mol/L$ e em mg/mL.

Tabela 10 – Propriedades das proteínas recombinantes calculadas no ProtParam.

Proteína	Número de aminoácidos	Massa molecular (Da)	ϵ ($M^{-1}cm^{-1}$)	A_{280} 0,1%
saFabI	278	30313,35	13410	0,442
ecFabI*	277	29807,03	17420	0,584

* Ainda que a FabI de E. coli apresente resíduos de cisteína, ela não é capaz de formar cistinas, uma vez que seus resíduos de cisteína estão muito distantes entre si no plano tridimensional. Portanto, utilizou-se o coeficiente de extinção molar em que todas as cisteínas estão reduzidas.

Fonte: Autoria própria.

O rendimento total das proteínas em mg/L de cultura foi calculado individualmente utilizando a concentração obtida em mg/mL e o volume de cultura utilizado no seu respectivo cultivo.

4.21 CARACTERIZAÇÃO DO ESTADO OLIGOMÉRICO DE PROTEÍNAS POR ESPALHAMENTO DINÂMICO DA LUZ

A determinação do estado oligomérico das proteínas saFabI e ecFabI em tampão de exclusão molecular E2 (tampão fosfato de potássio 20 mM, NaCl 200 mM e DTT 2 mM, pH = 8,0) foi feita por ensaios de Espalhamento Dinâmico da Luz (DLS) utilizando o equipamento Zetasizer Nano ZS90 (Malvern Panalytical). Antes das análises, as amostras de proteína foram centrifugadas a 8000 g por 10 minutos a 4 °C, para remoção de qualquer material suspenso. Os ensaios foram realizados sob temperatura constante de 25 °C em cubeta de vidro com caminho óptico de 1,0 cm e 1,5 mL de amostra em concentração de no mínimo 1,0 mg/mL. Cada amostra foi analisada em triplicata e o número de medidas por replicata foi calculado automaticamente pelo *software* Malvern Zetasizer 7.13. Os dados coletados foram exportados e foi feita a análise do estado de oligomerização das proteínas recombinantes.

4.22 ENSAIOS DE CONCENTRAÇÃO INIBITÓRIA MÍNIMA

Os ensaios de concentração inibitória mínima (MIC) foram conduzidos no Laboratório de Vírus do Departamento de Microbiologia do Instituto de Ciências Biológicas da UFMG. As moléculas consideradas promissoras na triagem virtual foram

avaliadas frente a cepas de bactérias *Staphylococcus aureus* (ATCC 29123) e *Escherichia coli* (ATCC 35218) pelo ensaio de microdiluição em caldo foi feito em microplacas de 96 poços, de acordo com o protocolo do *Clinical and Laboratory Standards Institute* (CLSI) (Clinical and Laboratory Standards Institute, 2017). Inicialmente, as moléculas foram dissolvidas em dimetilsulfóxido (DMSO) (Merck, Germany) e diluídas em meio Mueller Hinton (Oxoid, Thermo Fischer Scientific, UK) a uma concentração de 200 μM . Em seguida, 100 μL das diluições foram adicionadas em cada poço, juntamente do mesmo volume com uma suspensão bacteriana contendo $1,0 \times 10^5$ UFC/mL, ou seja, 100.000 unidades formadoras de colônia por mililitro, resultando em concentrações de 100 μM .

As microplacas foram incubadas a 37 °C por 24 horas e inspecionadas visualmente quanto à inibição do crescimento bacteriano (ausência de turbidez). Poços com ausência de crescimento foram considerados moléculas *hits* com atividade antibacteriana, sendo estas posteriormente avaliadas quanto a sua concentração inibitória mínima, pelo mesmo método e sob as mesmas condições, em concentrações entre 100 e 0,78 μM . Ambos os ensaios foram adicionados de controles de viabilidade celular (meio e bactéria), esterilidade (somente meio), veículo (DMSO), além de controles de inibição do crescimento bacteriano com os antimicrobianos (5x o valor de MIC reportado na literatura) penicilina para *S. aureus* e estreptomicina para *E. coli*. Os testes foram realizados em triplicata e em dois ensaios independentes.

As moléculas com MIC inferior a 100 μM foram selecionadas para os ensaios de diferença da transferência de saturação por ressonância magnética nuclear (STD-NMR) juntamente com as proteínas saFabI e ecFabI purificadas e caracterizadas. Os ensaios serão realizados nos laboratórios de Macromoléculas (MacroMol) e de Ressonância Magnética de Alta Resolução (LAREMAR), ambos do Departamento de Química do Instituto de Ciências Exatas da UFMG.

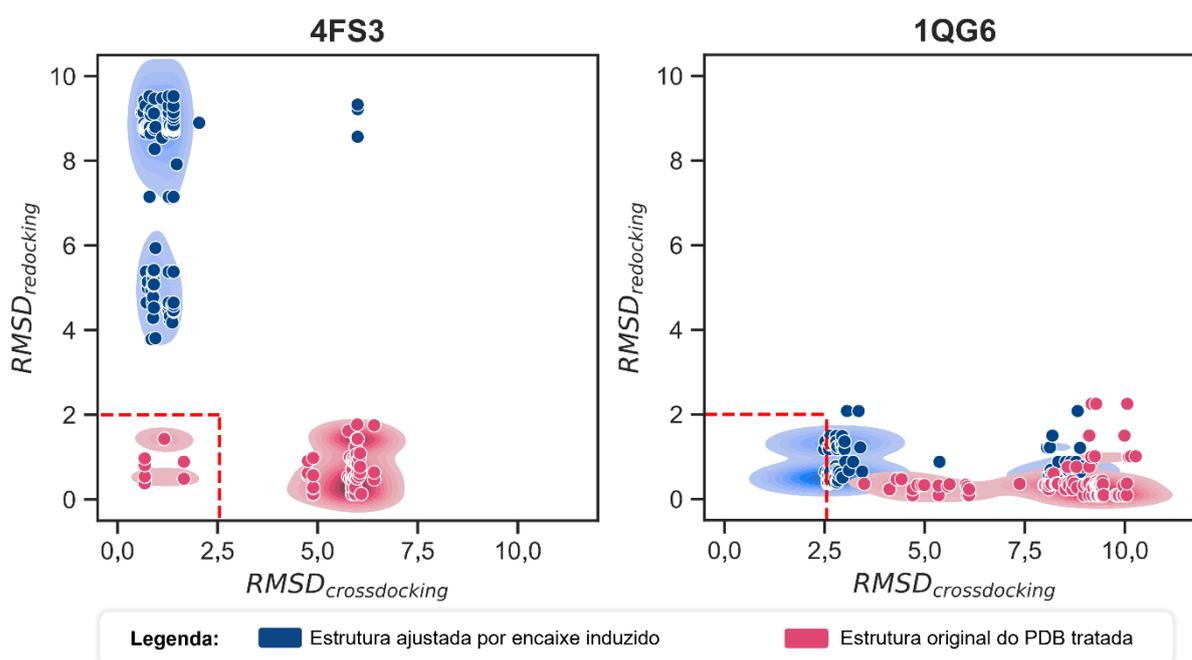
5 RESULTADOS E DISCUSSÃO

5.1 VALIDAÇÃO DOS PROTOCOLOS DE ACOPLAMENTO MOLECULAR

A validação inicial dos 2352 protocolos de acoplamento molecular foi feita pelo cálculo do RMSD de *redocking* e de *crossdocking*. O RMSD de *redocking* foi obtido pelo cálculo com o ligante cristalográfico de cada proteína e o RMSD de *crossdocking* foi obtido pelo alinhamento das estruturas cristalográficas de uma proteína na outra, seguido pela extração do ligante para o *crossdocking*. Dessa forma, a 1QG6 foi estruturalmente alinhada à 4FS3 (que foi utilizada como centro para a sobreposição estrutural) para obtenção das coordenadas tridimensionais do triclosan (TCL) para o *crossdocking* da 4FS3 e da 4FS3_IFD. E, da mesma maneira, a 4FS3 foi estruturalmente sobreposta à 1QG6 para obtenção das coordenadas da AFN-1252 para o *crossdocking* da 1QG6 e da 1QG6_IFD.

Primeiramente, para a escolha dos melhores protocolos de acoplamento molecular, foi realizada uma análise visando avaliar o efeito do acoplamento molecular por encaixe induzido (realizado conforme detalhado na seção **4.3 Seleção e pré-tratamento dos alvos moleculares para os estudos de acoplamento molecular (p. 68)**). Para isso, foram comparados os resultados de RMSD de *redocking* e *crossdocking* obtidos entre as estruturas das proteínas diretamente extraídas do PDB e as estruturas das proteínas tratadas pelo *docking* por encaixe-induzido realizado no programa Glide. Na **Figura 16** é possível perceber uma clara vantagem entre os valores de RMSD obtidos pelas estruturas 4FS3 (saFabI) e 1QG6_IFD (ecFabI), mostrando que as estruturas com a AFN-1252 resultam em poses mais precisas nos resultados de *docking*. Além disso, os protocolos utilizando o algoritmo de *docking* POSIT, que propõe a pose dos ligantes utilizando informações de similaridade bidimensional e tridimensional com o ligante originalmente presente no sítio ativo, resultaram em maiores valores de RMSD de *crossdocking*. Esse resultado foi interessante porque o maior tempo de processamento e gasto computacional desse algoritmo seria proibitivo para a realização de triagens virtuais.

Figura 16 – Densidade e distribuição dos valores de RMSD de *redocking* e *crossdocking* entre todos os protocolos de acoplamento molecular que retornaram o ligante no sítio ativo. À esquerda os valores para a proteína de *S. aureus* (PDB ID: 4FS3) e à direita os valores da proteína de *E. coli* (PDB ID: 1QG6). Em azul, os valores obtidos para as proteínas de encaixe induzido e, em rosa escuro, os valores obtidos para as proteínas do PDB tratadas apenas para remoção de dupla ocupância. A linha vermelha representa $\text{RMSD}_{\text{redocking}} < 2,00 \text{ \AA}$ e $\text{RMSD}_{\text{crossdocking}} < 2,55 \text{ \AA}$.

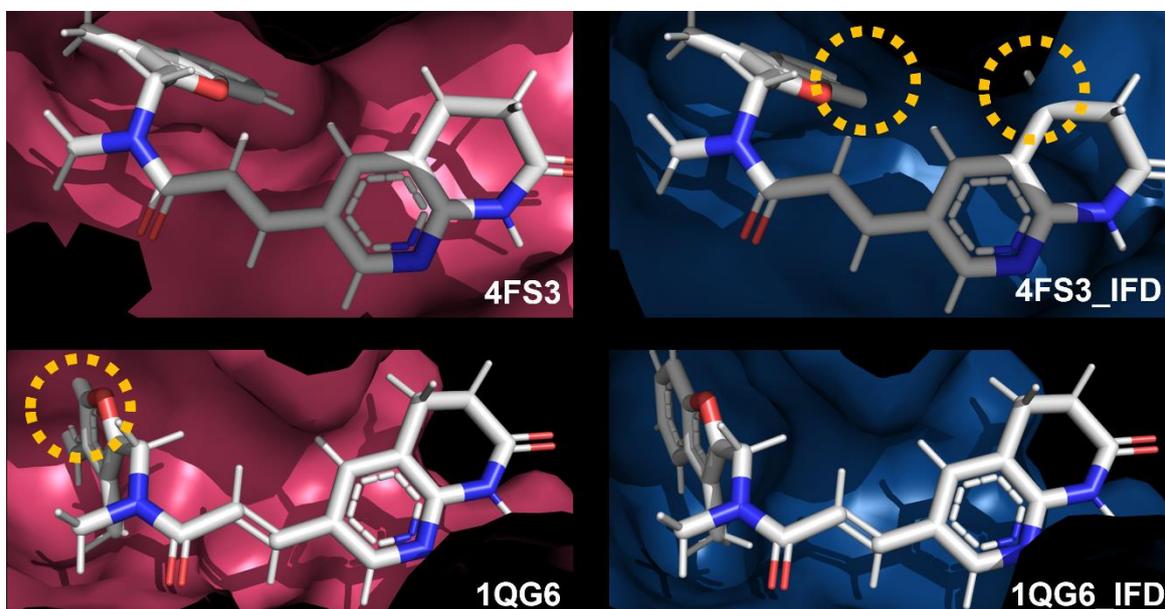


Fonte: Autoria própria.

O melhor desempenho de RMSD alcançado pelas proteínas que possuem a AFN-1252 na sua estrutura, seja pela presença na estrutura cristalográfica (4FS3) ou pela presença devido ao *docking* por encaixe-induzido (1QG6_IFD), se deve ao fato de que a AFN-1252 é uma molécula grande capaz de interagir com a FabI e promover alterações conformacionais que dão acesso a regiões que estão impedidas na estrutura das proteínas complexadas com o triclosan. A **Figura 17** ilustra precisamente essa característica, evidenciando a colisão que ocorre com a AFN-1252 quando esse ligante é acoplado na estrutura de proteínas previamente em complexo com o triclosan. Dessa forma, em estudos de *docking* com a FabI, torna-se inviável a utilização de estruturas proteicas que possuem o triclosan como ligante, devido à presença do impedimento estérico. Durante todo o desenvolvimento deste trabalho,

apenas uma estrutura de ecFabI complexada com a AFN-1252 estava disponível no PDB (PDB ID: 4JQC), entretanto, ela apresenta resolução de 2.8 Å, sendo considerada de baixa confiabilidade para uso em estudos de planejamento de fármacos (Cooper *et al.*, 2011).

Figura 17 – Representação da AFN-1252 (cinza) nos sítios de ligação das enzimas FabI obtidas diretamente do PDB (rosa) e das obtidas pelo *docking* por encaixe induzido (azul). Em laranja, as colisões que são observadas nas estruturas de proteínas que estavam previamente complexadas com o triclosan (4FS3_IFD e 1QG6).



Fonte: Autoria própria.

Os valores de RMSD de *redocking* e *crossdocking* foram utilizados para seleção dos cinco melhores protocolos de cada proteína, uma vez definida as melhores estruturas de proteínas para cada espécie (4FS3 para a saFabI e 1QG6_IFD para a ecFabI). Os parâmetros utilizados em cada um dos cinco melhores protocolos de cada espécie e os seus respectivos valores de $RMSD_{redocking}$ e $RMSD_{crossdocking}$ estão representados na **Tabela 11**. Todos esses protocolos apresentaram valores de $RMSD_{redocking}$ inferiores a 2,00 Å, indicando que as metodologias utilizadas são válidas para os estudos de acoplamento molecular, uma vez que estão dentro dos limites estabelecidos pela literatura (Kirchmair *et al.*, 2008; Yusuf *et al.*, 2008). Os valores de RMSD de *crossdocking* também foram aceitáveis mesmo que, para a proteína de *E.*

coli, estejam superiores a 2,00 Å. Como discutido no tópico **3.3.1 Estratégias de validação do acoplamento molecular (p. 52)**, é importante se considerar alguns fatores como o tamanho e presença de ligações rotacionáveis na AFN-1252, a obtenção da estrutura da proteína por encaixe induzido e o fato de que o *crossdocking* foi realizado entre proteínas diferentes que, apesar de suas similaridades, possuem particularidades e cofatores diferentes. Dessa forma, ressalta-se a importância de realizar a análise visual caso-a-caso dos resultados da triagem.

Tabela 11 – Cinco melhores protocolos de *docking* da proteína FabI de cada espécie e seus parâmetros e valores de RMSD.

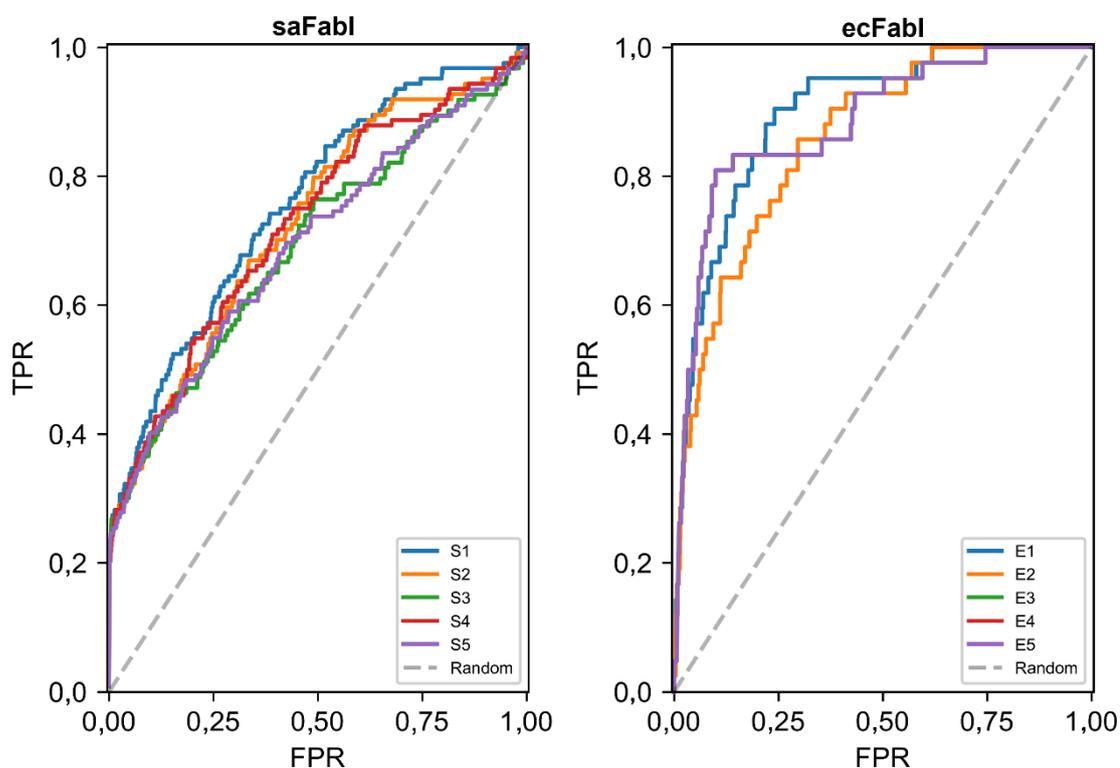
ID	Volume da caixa (Å)	Algoritmo	Número de conformações	Resolução do <i>docking</i>	RMSD _{redocking} (Å)	RMSD _{crossdocking} (Å)
4FS3 (saFabI)						
S1	8756,00	FRED	120	High	0,38	0,68
S2	4529,00	FRED	120	High	0,38	0,68
S3	3048,00	FRED	120	High	0,38	0,68
S4	4529,00	FRED	60	High	0,54	0,68
S5	3048,00	FRED	60	High	0,54	0,68
1QG6_IFD (ecFabI)						
E1	7072,00	FRED	120	Standard	0,64	2,52
E2	7072,00	FRED	90	Standard	0,64	2,52
E3	8249,00	HYBRID	90	High	1,18	2,49
E4	8249,00	HYBRID	90	Standard	1,18	2,49
E5	8249,00	HYBRID	90	Low	1,18	2,49

Fonte: Autoria própria.

Para validar a usabilidade dos protocolos de *docking* em triagens virtuais de substâncias, também foram construídas as curvas ROC (**Figura 18**) e calculados os valores de área sobre a curva ROC (AUC_{ROC}), enriquecimento e área sobre a curva BEDROC (AUC_{BEDROC}) utilizando os ligantes e *decoys*. Enquanto a área sob a curva ROC (AUC_{ROC}) faz uma avaliação geral da capacidade de distinguir entre as classes (ativos e inativos), as métricas de enriquecimento e AUC_{BEDROC} fornecem uma informação sobre essa capacidade de distinção em frações iniciais da curva, ou seja, onde se encontram as moléculas classificadas pelo *docking* com a menor pontuação (por exemplo, quando espera-se o menor ΔG para as moléculas mais ativas) ou com

a maior pontuação (por exemplo, quando o esperado é a maior probabilidade de uma molécula ser classificada como ativa). Em outras palavras, essas duas métricas avaliam a capacidade de distinção do protocolo entre as moléculas com maior probabilidade de serem biologicamente ativas de acordo com o *docking*. Essas métricas são ideais para validação de protocolos de *docking* utilizados em triagem virtual, pois simulam a escolha das mais moléculas ativas pelos resultados da pontuação do *docking*. Para os cálculos de AUC_{BEDROC} , os valores de α utilizados foram de 160,9, 32,2 e 16,1, que representam um peso de 80% no cálculo da área sob a curva para a fração inicial de 1,0%, 5,0% e 10,0%, respectivamente (Venkatraman; Chakravarthy; Kihara, 2009; Avram *et al.*, 2013; Castillo-González *et al.*, 2015). Essa mesma fração da curva, foi calculada em termos de fator de enriquecimento e os valores estão dispostos na **Tabela 12**.

Figura 18 – Curvas ROC para os melhores protocolos de acoplamento molecular de cada proteína.



Fonte: Autoria própria.

Tabela 12 – Valores de AUC_{ROC} , AUC_{BEDROC} (utilizando $\alpha = 160,9$, $32,2$ e $16,1$) e fator de enriquecimento (EF) nas frações de 1,0, 5,0 e 10,0%. Em negrito, o protocolo com melhor desempenho de cada proteína.

ID	AUC_{ROC}	AUC_{BEDROC} ($\alpha = 160,9$)	AUC_{BEDROC} ($\alpha = 32,2$)	AUC_{BEDROC} ($\alpha = 16,1$)	EF (1,0%)	EF (5,0%)	EF (10,0%)
saFabl							
S1	0,757	0,531	0,363	0,386	23,225	6,447	4,190
S2	0,726	0,521	0,345	0,360	22,424	6,124	3,707
S3	0,691	0,539	0,350	0,360	25,021	6,010	3,655
S4	0,724	0,518	0,349	0,367	23,225	6,124	3,949
S5	0,693	0,524	0,342	0,357	24,409	6,058	3,848
ecFabl							
E1	0,897	0,275	0,398	0,502	16,232	9,524	6,667
E2	0,858	0,259	0,360	0,444	16,232	8,571	5,476
E3	0,884	0,242	0,407	0,529	11,594	10,000	7,381
E4	0,884	0,242	0,407	0,529	11,594	10,000	7,381
E5	0,884	0,242	0,407	0,529	11,594	10,000	7,381

Fonte: Autoria própria.

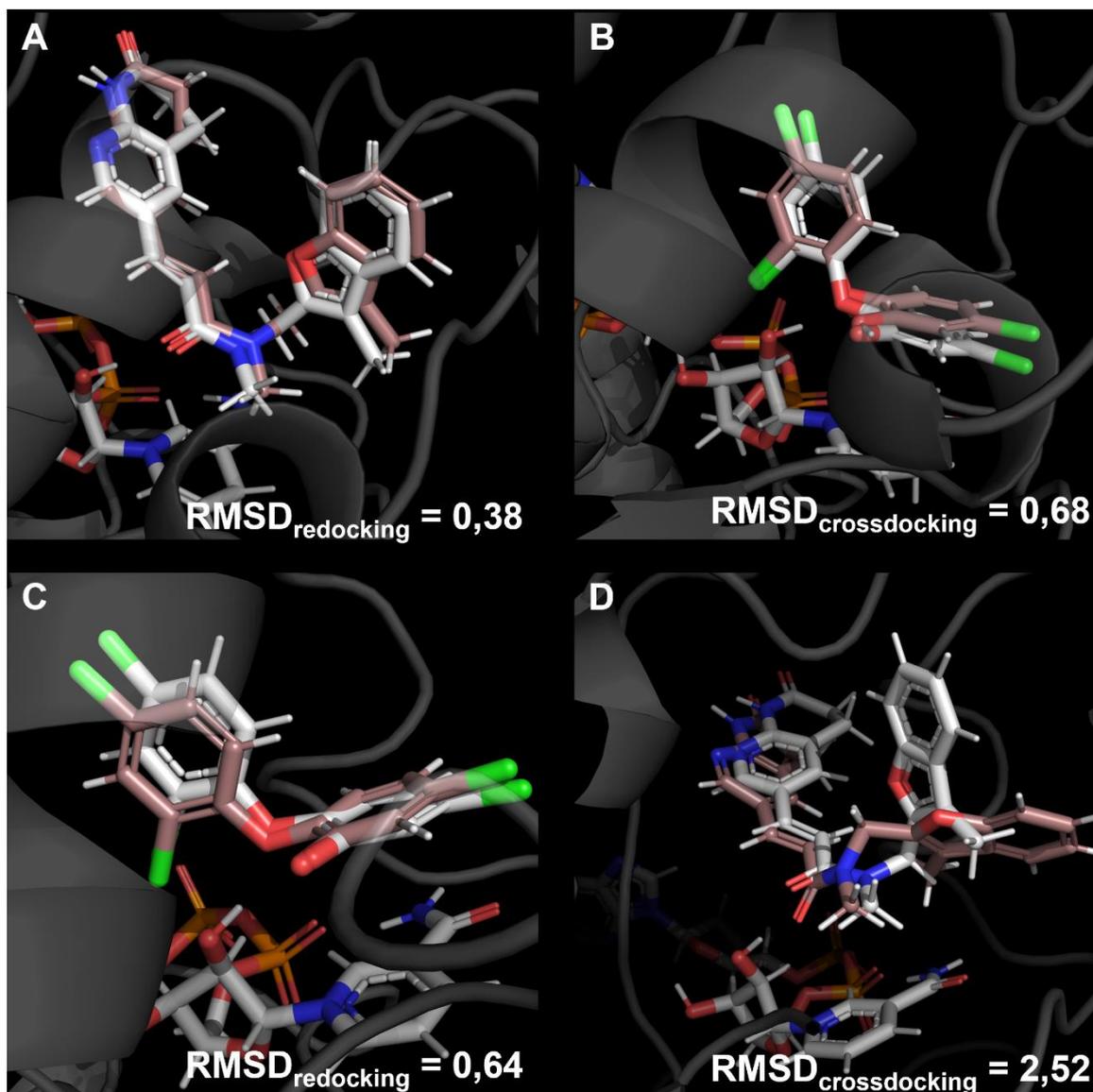
Os valores encontrados para *S. aureus* indicam que o primeiro protocolo (S1) teve performance geral superior aos demais para discriminação geral de compostos entre ativos e inativos, sugerido pelos valores de AUC_{ROC} . O fator de enriquecimento (EF) e a AUC_{BEDROC} são duas métricas que são utilizadas na análise do reconhecimento inicial de uma curva ROC. Entretanto, EF é mais sensível ao efeito do número de moléculas e ao desbalanceamento do conjunto de dados do que $BEDROC$, isso ocorre porque o fator de enriquecimento é sensível ao número absoluto de ativos e inativos entre os x% com maior probabilidade de serem ativos, enquanto, a $BEDROC$ é uma métrica ponderada fixada entre 0 e 1 que fornece um peso de 80% para as moléculas entre os x% com maior probabilidade de serem ativos, enquanto ainda leva em consideração as demais moléculas (Truchon; Bayly, 2007; Venkatraman; Chakravarthy; Kihara, 2009). Dessa forma, priorizou-se a análise dos valores de $BEDROC$ em detrimento dos valores de EF. Essa característica se torna ainda mais relevante nos protocolos S3 e S5, pois o volume da caixa onde foi realizado o acoplamento molecular resultou na incapacidade do algoritmo de *docking* de realizar o acoplamento com alguns ligantes e, portanto, esses ligantes não foram geraram pontuação no *docking* e não foram contabilizados para análise. Essa incapacidade de realizar o acoplamento foi um critério de exclusão para esses protocolos. Ainda assim, o protocolo S1 foi superior aos demais nas métricas de AUC_{ROC} e AUC_{BEDROC} e igual

ou superior no $\text{RMSD}_{\text{redocking}}$ e no $\text{RMSD}_{\text{crossdocking}}$, sendo a única exceção na $\text{AUC}_{\text{BEDROC}}$ com α de 160,9 (1%), onde o protocolo excluído S3 teve uma performance melhor (+0,008).

Em relação aos resultados de *S. aureus* o protocolo (E1) teve performance geral superior ao dos demais em termos de AUC_{ROC} . Na fração inicial de 1% da $\text{AUC}_{\text{BEDROC}}$ ($\alpha = 160,9$), esse protocolo foi o que teve melhor desempenho e, ainda que nas demais frações seu desempenho tenha sido inferior a outros protocolos, a fração inicial de 1% é a mais relevante na seleção dos dados permitindo que, em uma triagem, a escolha dos 1% mais ativos (ou com maior probabilidade de serem ativos) seja feita com maior acurácia nesse protocolo do que nos demais. Além disso, esse protocolo esteve entre os melhores valores de $\text{RMSD}_{\text{redocking}}$, obtendo valores menores (-0,54) do que os protocolos que performaram melhor em outros α de BEDROC. A obtenção de poses com menores valores de RMSD é um objetivo importante nesse estudo, uma vez que a obtenção dos *fingerprints* de interação é baseada na pose obtida pelo acoplamento molecular.

Portanto, com base na análise das métricas calculadas, os protocolos S1 e E1 foram escolhidos para realizar a obtenção dos *fingerprints* de interação neste trabalho. Na **Figura 19** estão representados os complexos proteína-ligante utilizados no *redocking* e *crossdocking* para esses dois protocolos.

Figura 19 – Representação tridimensional dos resultados de *redocking* (A e C) e *crossdocking* (B e D) para as proteínas FabI de *S. aureus* (A e B) e de *E. coli* (C e D) com os melhores protocolos. Em cinza, as estruturas cristalográficas e, em rosa envelhecido, as estruturas obtidas pelos protocolos de *docking*.



Fonte: Autoria própria.

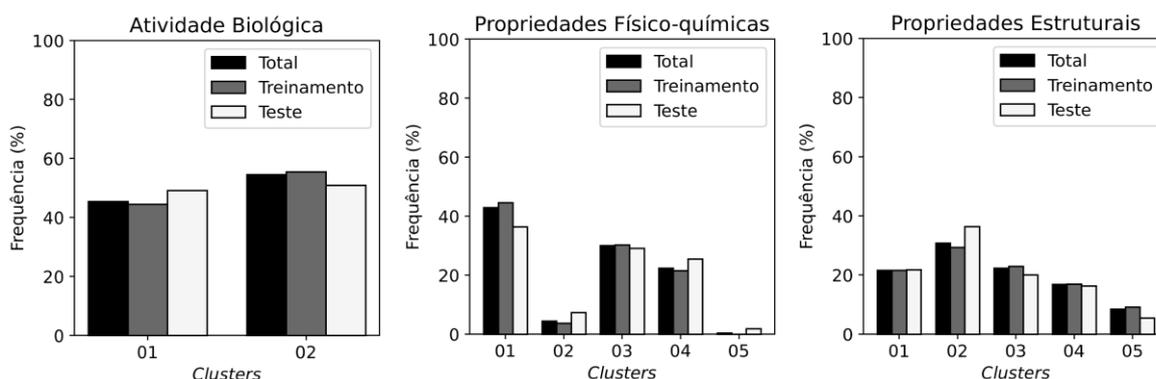
5.2 SEPARAÇÃO DO CONJUNTO DE DADOS EM CONJUNTOS DE TREINAMENTO E DE TESTE

A separação entre os conjuntos de treinamento e de teste, fundamental para a construção e validação dos modelos de aprendizado de máquina, foi realizada utilizando o programa MASSA Algorithm 0.9.2. Esse programa efetua a divisão racional do conjunto de dados e possui performance superior a separação aleatória

(não-razional) e, da mesma forma, superior ou equiparável a racional por outros algoritmos, como Kennard-Stone, SPXY (do inglês, “*sample set partitioning based on joint X-y distance*” ou simplesmente “partição do conjunto de amostras baseado na distância conjunta X-y”) e *Sphere Exclusion*, sendo capaz de realizar a separação independentemente dos descritores utilizados na construção dos modelos (Veríssimo, 2021; Veríssimo *et al.*, 2023). Essa característica do MASSA é especialmente relevante para o contexto deste estudo, uma vez que diferentes *fingerprints* de interação foram gerados devido a variação dos parâmetros de geração. A manutenção da mesma amostragem entre os diferentes descritores nos conjuntos de treinamento e de teste foi essencial para assegurar uma comparação precisa entre os diferentes modelos e parâmetros de *fingerprint*.

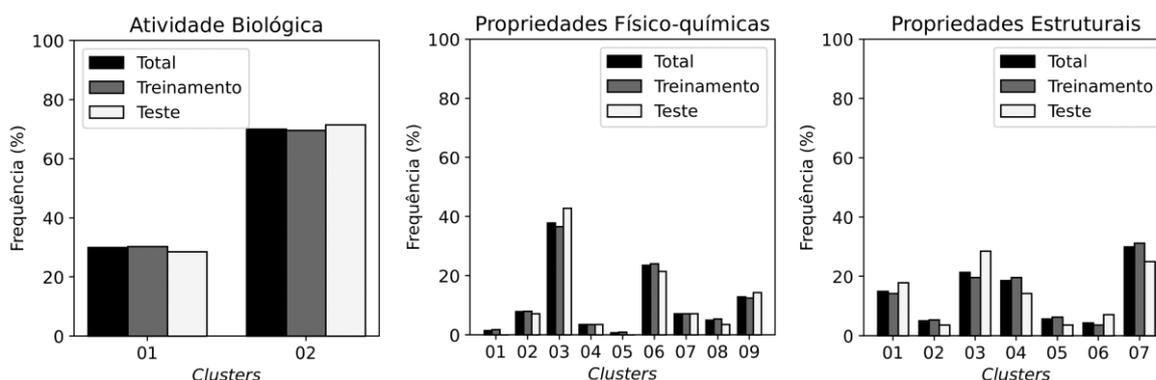
O MASSA Algorithm realiza uma separação em *clusters* baseada em HCA buscando representar a distância euclidiana entre as moléculas com base em suas propriedades estruturais (*Atom Pair fingerprint*), físico-químicas (número de aceptores de ligação de hidrogênio, número de doadores de ligação de hidrogênio, massa molecular, número de ligações rotacionáveis, fração de carbono sp^3 , área de superfície polar topológica, log P de Wildman-Crippen) e biológicas (no caso, a classificação em ativo e inativo), gerando gráficos de barras que representam a proporção de moléculas em cada um dos conjuntos entre os diferentes domínios e *clusters*. Esses gráficos de barra estão representados na **Figura 20** para o conjunto de ligantes de *S. aureus* e, na **Figura 21**, para o conjunto de ligantes de *E. coli*. As moléculas utilizadas nessas separações foram apenas aquelas com atividade biológica determinada experimentalmente. Dessa forma, os *decoys*, utilizados para a validação do *docking*, não foram utilizados nessas separações e nem mesmo para os modelos de aprendizado de máquina.

Figura 20 – Distribuição do conjunto de dados de inibidores da saFabI (N = 273) entre os subconjuntos e diferentes propriedades.



Fonte: Autoria própria.

Figura 21 – Distribuição do conjunto de dados de inibidores da ecFabI (N = 140) entre os subconjuntos e diferentes propriedades.



Fonte: Autoria própria.

A separação dos conjuntos de treinamento e de teste para os dois conjuntos de dados demonstrou a eficácia em manter a proporção dos ligantes entre os diferentes *clusters* dos domínios biológico, físico-químico e estrutural. Além disso, nenhum *cluster* com mais de 5% de representação no conjunto de dados total ficou sem representação em ambos os subconjuntos. Isso garante que o treinamento dos modelos de aprendizado de máquina seja feito com a maior diversidade químico-biológica possível e evita com que moléculas do conjunto de teste estejam fora do domínio de aplicabilidade.

5.3 CONSTRUÇÃO E VALIDAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA COM OS *FINGERPRINTS* DE INTERAÇÃO

Os cinco parâmetros de geração dos *fingerprints* de interação do LUNA variados resultaram em 378 conjuntos de descritores para cada proteína, ou seja, 756 *fingerprints* de interação foram utilizados para geração dos modelos de aprendizado de máquina. Partindo desses conjuntos de dados, foram construídos e validados interna (validação *5-fold*) e externamente mais de 220 milhões de modelos de aprendizado de máquina (220.856.328 modelos) utilizando os algoritmos kNN, DT, RF, gNB, MLP (MLL, MLS, MLA) e SVM (SVL, SVR, SVS, SVP). A **Tabela 13** traz a descrição detalhada do número de modelos em cada técnica.

Tabela 13 – Número de modelos para cada técnica de aprendizado de máquina utilizado.

Algoritmos	Número de modelos por <i>fingerprint</i>	Número de modelos para cada proteína	Número de modelos total por algoritmo
kNN	160	60.480	120.960
gNB	5.000	1.890.000	3.780.000
DT	60	22.680	45.360
RF	90	34.020	68.040
SVL	28	10.584	21.168
SVR	896	338.688	677.376
SVS	54.656	20.659.968	41.319.936
SVP	136.640	51.649.920	103.299.840
MLL	1.296	489.888	979.776
MLS	46.656	17.635.968	35.271.936
MLA	46.656	17.635.968	35.271.936
Total	292.138	110.428.164	220.856.328

Fonte: Autoria própria.

Todos os modelos gerados foram ranqueados de acordo com a média entre o MCC interno e o MCC externo para excluir potenciais modelos com baixa capacidade preditiva ou que são mais susceptíveis ao sobreajuste (*overfitting*), fenômeno que ocorre quando um modelo se ajusta bem aos dados de treinamento, mas tem baixa acurácia de predição para outros dados. Os modelos com os 20 maiores valores de MCC médio de cada algoritmo de aprendizado de máquina foram então selecionados

para as etapas posteriores. O MCC foi escolhido para a hierarquização dos modelos por ser uma métrica de qualidade mais robusta que AUC, F1-Score, acurácia e o coeficiente de Cohen Kappa, pois leva em consideração todos os parâmetros da matriz de confusão e a distribuição de predições em todas as classes, sendo menos susceptível ao viés de conjuntos de dados desbalanceados que as demais métricas (Chicco; Jurman, 2020, 2023; Zhu, 2020; Chicco; Warrens; Jurman, 2021).

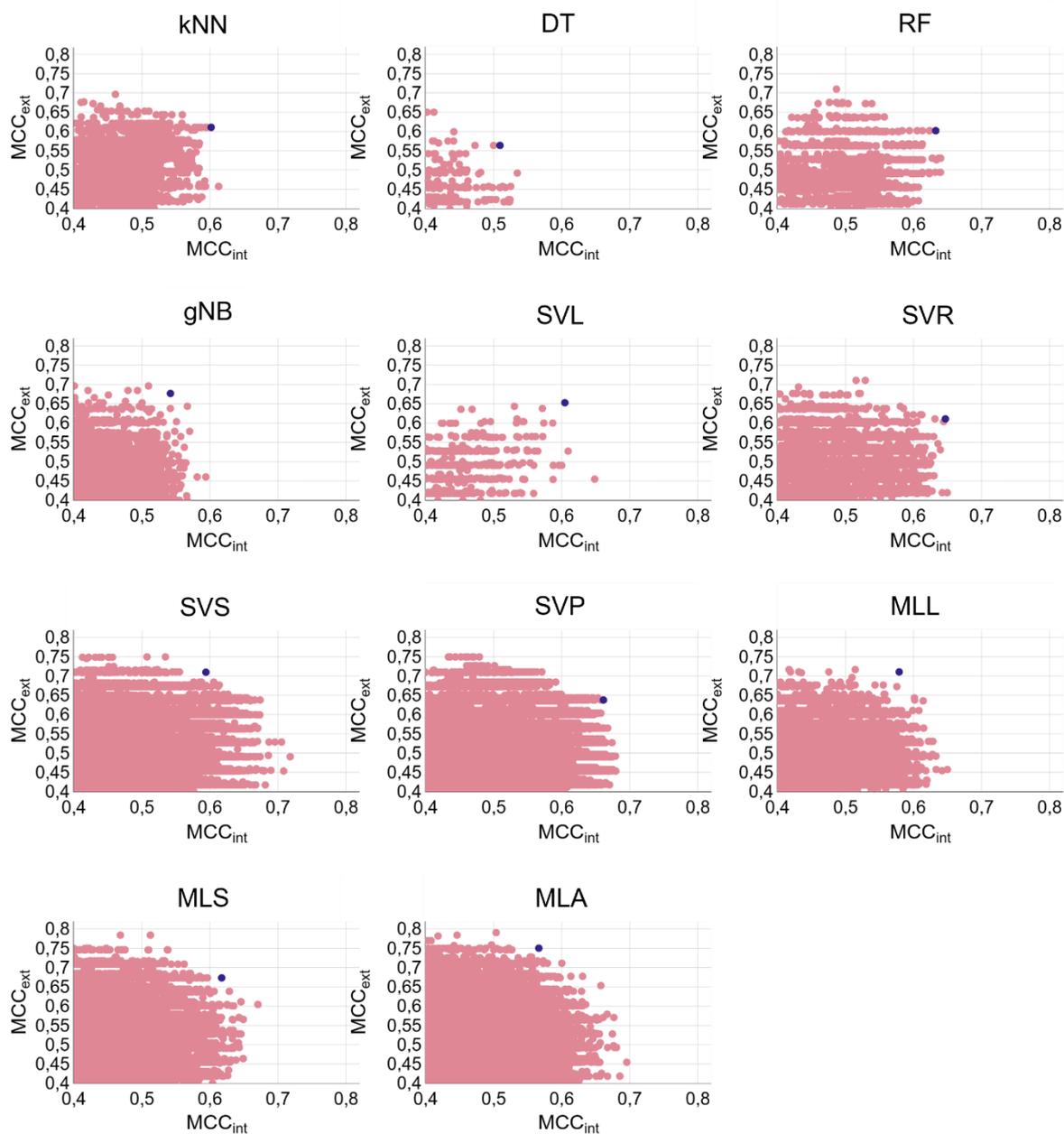
Em seguida, os valores de MCC interno e externo desses melhores modelos foram manualmente analisados para garantir que o modelo escolhido apresente não apenas a melhor média entre as essas métricas, mas também que haja um equilíbrio entre os valores obtidos na validação interna e externa. A **Tabela 14** representa os valores dos melhores modelos de cada algoritmo para a proteína FabI de *S. aureus* e a **Tabela 15** representa os valores dos melhores modelos de cada algoritmo para a proteína FabI de *E. coli*. Para garantir a visualização dos melhores modelos de cada técnica em comparação com os demais foram construídos gráficos de dispersão (**Figuras 22 e 23**) que mostram o melhor modelo de cada técnica (em roxo) e os demais modelos (em rosa).

Tabela 14 – Resultados de MCC dos melhores modelos de cada técnica para a saFabl.

Modelo	Técnica	Parâmetros do <i>Fingerprint</i>	Hiperparâmetros	MCC _{5-fold}	MCC _{ext}	$\overline{\text{MCC}}$
SA_kNN	kNN	IFP_type: EIFP, nLevels: 1, len: 2048, radius: 5,73171, notation: bits	n_neighbors: 21, weights: distance, algorithm: ball_tree, p: 2, leaf_size: 15, metric: minkowski	0,602	0,611	0,607
SA_DT	DT	IFP_type: HIFP, nLevels: 2, len: 2048, radius: 1,4329275, notation: bits	criterion: gini, splitter: best, min_samples_split: 34, max_features: None	0,510	0,564	0,537
SA_RF	RF	IFP_type: HIFP, nLevels: 2, len: 4096, radius: 2,865855, notation: bits	criterion: gini, n_estimators: 30000, max_depth: 10, max_features: sqrt	0,633	0,603	0,618
SA_gNB	gNB	IFP_type: EIFP, nLevels: 3, len: 2048, radius: 1,4329275, notation: bits	var_smoothing: 1,0, priors: array([0,76, 0,24])	0,542	0,677	0,610
SA_SVL	SVC (Linear)	IFP_type: FIFP, nLevels: 2, len: 4096, radius: 2,865855, notation: bits	C: 0,01, kernel: linear	0,605	0,653	0,629
SA_SVR	SVC (RBF)	IFP_type: FIFP, nLevels: 2, len: 4096, radius: 2,865855, notation: bits	C: 20, gamma: 0,0001, kernel: rbf	0,647	0,611	0,629
SA_SVS	SVC (Sigmoide)	IFP_type: EIFP, nLevels: 2, len: 1024, radius: 1,4329275, notation: bits	C: 3, gamma: 0,004, coef0: - 0,008, kernel: sigmoid	0,594	0,710	0,652
SA_SVP	SVC (Polinomial)	IFP_type: FIFP, nLevels: 2, len: 4096, radius: 2,865855, notation: bits	C: 50, gamma: 0,0001, coef0: -0,6, degree: 3, kernel: poly	0,662	0,638	0,650
SA_MLL	MLP (LBFGS)	IFP_type: EIFP, nLevels: 3, len: 2048, radius: 1,4329275, notation: counts	hidden_layer_sizes: [8, 8, 8, 8, 8], solver: lbfgs, activation: relu, alpha: 0,01, max_iter: 500	0,579	0,711	0,645
SA_MLS	MLP (SGD)	IFP_type: EIFP, nLevels: 2, len: 1024, radius: 1,4329275, notation: bits	hidden_layer_sizes: [128, 64, 32, 16], solver: sgd, activation: relu, alpha: 1e-05, max_iter: 500, batch_size: auto, learning_rate: constant, learning_rate_init: 0,3, momentum: 0,9, early_stopping: True	0,617	0,673	0,645
SA_MLA	MLP (Adam)	IFP_type: FIFP, nLevels: 2, len: 4096, radius: 2,865855, notation: bits	hidden_layer_sizes: [20, 10], solver: adam, activation: logistic, alpha: 0,01, max_iter: 500, batch_size: auto, learning_rate: constant, learning_rate_init: 0,3, beta_1: 0,8, early_stopping: True	0,567	0,750	0,659

Fonte: Autoria própria.

Figura 22 – Dispersão dos modelos de aprendizado de máquina para a saFabi em função de seu valor de MCC interno e externo. Em roxo, o modelo com balanço entre maior robustez e preditividade selecionado para cada técnica e, em rosa, os demais modelos.



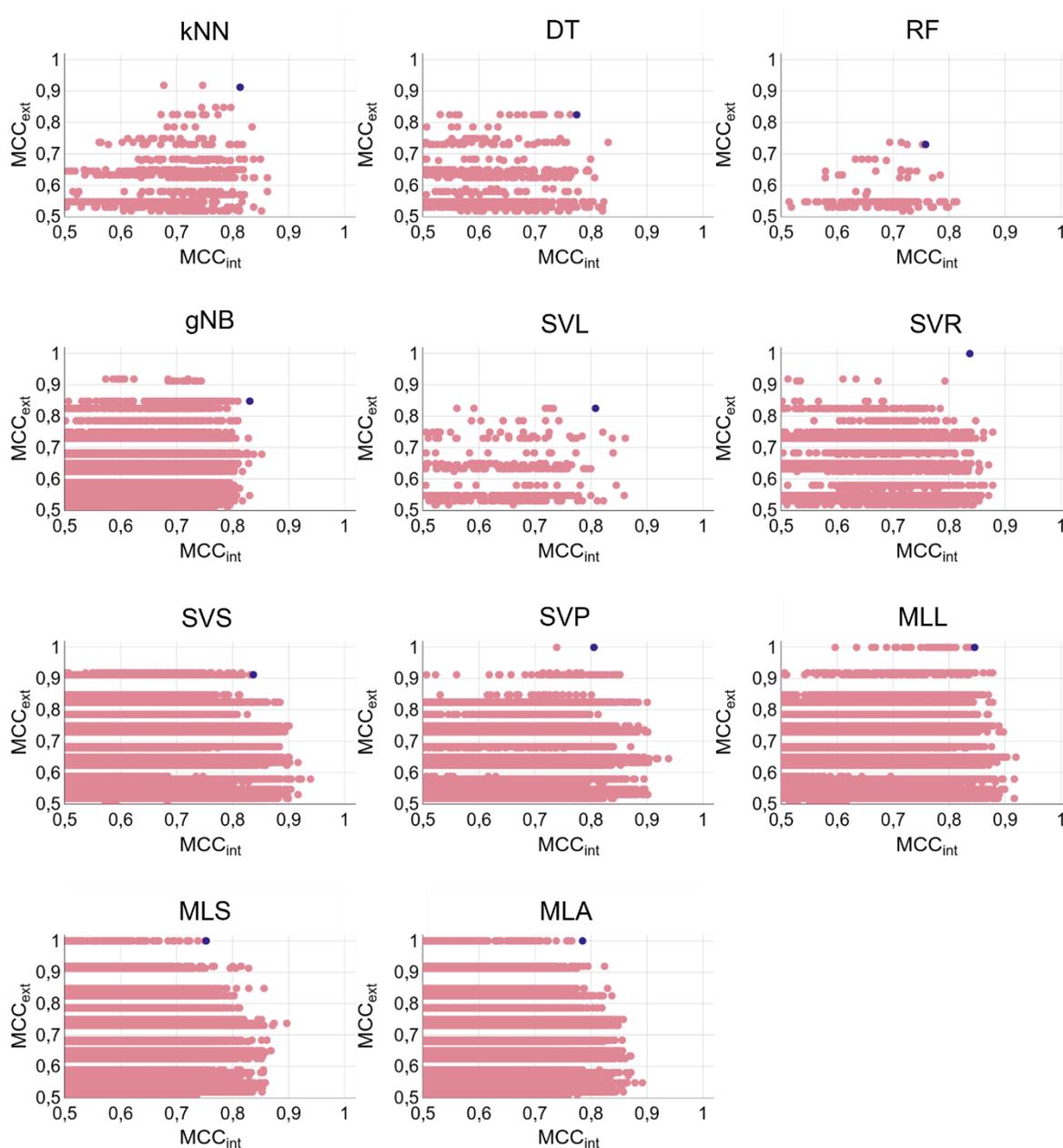
Fonte: Autoria própria.

Tabela 15 – Resultados de MCC dos melhores modelos de cada técnica para a ecFabl.

Modelo	Técnica	Parâmetros do <i>Fingerprint</i>	Hiperparâmetros	MCC _{5-fold}	MCC _{ext}	$\overline{\text{MCC}}$
EC_kNN	kNN	IFP_type: EIFP, nLevels: 3, len: 4096, radius: 1,4329275, notation: counts	n_neighbors: 11, weights: distance, algorithm: ball_tree, p: 2, leaf_size: 15, metric: minkowski	0,813	0,913	0,863
EC_DT	DT	IFP_type: HIFP, nLevels: 2, len: 1024, radius: 2,865855, notation: bits	criterion: gini, splitter: best, min_samples_split: 2, max_features: None	0,775	0,825	0,800
EC_RF	RF	IFP_type: EIFP, nLevels: 1, len: 1024, radius: 5,73171, notation: counts	criterion: gini, n_estimators: 10000, max_depth: 5, max_features: None	0,758	0,730	0,744
EC_gNB	gNB	IFP_type: FIFP, nLevels: 3, len: 1024, radius: 1,4329275, notation: bits	var_smoothing: 0,09540954763499938, priors: array([0,83, 0,17])	0,831	0,849	0,840
EC_SVL	SVC (Linear)	IFP_type: FIFP, nLevels: 3, len: 4096, radius: 1,4329275, notation: bits	C: 0,01, kernel: linear	0,808	0,826	0,817
EC_SVR	SVC (RBF)	IFP_type: FIFP, nLevels: 3, len: 4096, radius: 1,4329275, notation: bits	C: 15, gamma: 0,0001, kernel: rbf	0,837	1,000	0,919
EC_SVS	SVC (Sigmoide)	IFP_type: FIFP, nLevels: 3, len: 4096, radius: 1,4329275, notation: bits	C: 3, gamma: 0,002, coef0: 0,04, kernel: sigmoid	0,837	0,913	0,875
EC_SVP	SVC (Polinomial)	IFP_type: FIFP, nLevels: 3, len: 4096, radius: 1,4329275, notation: bits	C: 0,5, gamma: 0,002, coef0: -1, degree: 3, kernel: poly	0,806	1,000	0,903
EC_MLL	MLP (LBFGS)	IFP_type: FIFP, nLevels: 3, len: 4096, radius: 1,4329275, notation: bits	hidden_layer_sizes: [40, 40], solver: lbfgs, activation: logistic, alpha: 1e-05, max_iter: 500	0,846	1,000	0,923
EC_MLS	MLP (SGD)	IFP_type: FIFP, nLevels: 3, len: 4096, radius: 1,4329275, notation: bits	hidden_layer_sizes: [100], solver: sgd, activation: relu, alpha: 0,01, max_iter: 500, batch_size: auto, learning_rate: constant, learning_rate_init: 0,3, momentum: 0,4, early_stopping: True	0,753	1,000	0,876
EC_MLA	MLP (Adam)	IFP_type: FIFP, nLevels: 3, len: 4096, radius: 1,4329275, notation: bits	hidden_layer_sizes: [32, 32, 32], solver: adam, activation: logistic, alpha: 0,01, max_iter: 500, batch_size: auto, learning_rate: constant, learning_rate_init: 0,3, beta_1: 0,7, early_stopping: True	0,785	1,000	0,893

Fonte: Autoria própria.

Figura 23 – Dispersão dos modelos de aprendizado de máquina para a ecFabl em função de seu valor de MCC interno e externo. Em roxo, o modelo com balanço entre maior robustez e preditividade selecionado para cada técnica e, em rosa, os demais modelos.



Fonte: Autoria própria.

Os resultados representados nos gráficos de dispersão (**Figuras 22 e 23**) demonstram o sucesso na seleção dos melhores modelos de cada técnica garantindo o compromisso entre os valores de MCC_{int} e MCC_{ext} . Por ser uma métrica que varia de -1 a +1 e considera como aceitáveis valores superiores a 0,5 (Chicco; Warrens; Jurman, 2021; Chicco; Jurman, 2023), foi possível observar que todas as técnicas

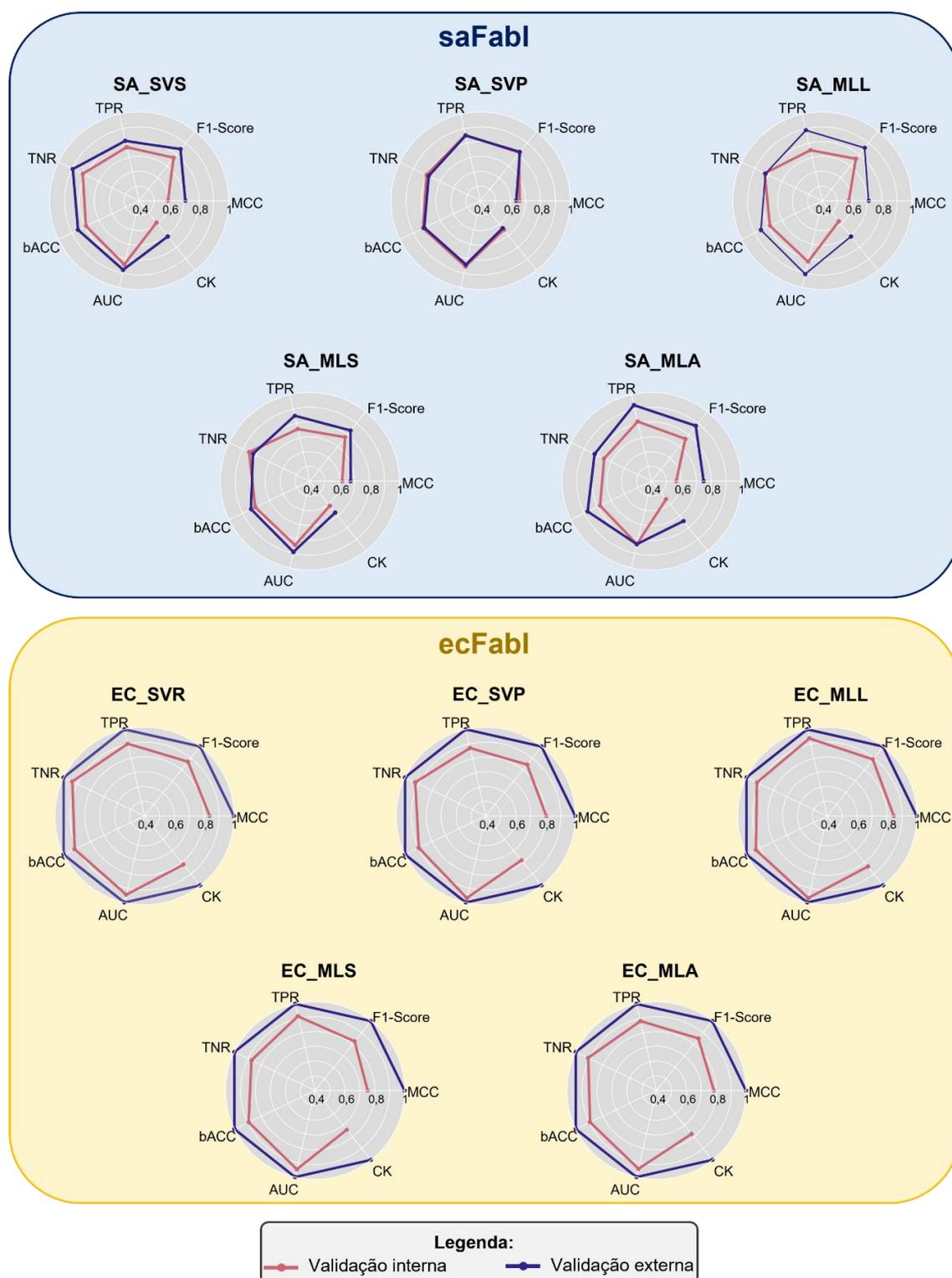
foram capazes de gerar bons modelos preditivos. Entretanto, os modelos que utilizaram os algoritmos de *Support Vector Machine* e *Multilayer Perceptron* foram os que resultaram nos melhores valores de MCC e, portanto, foram os que apresentaram melhor capacidade preditiva geral. Com base na análise dos valores de MCC (principalmente, a média entre os valores de MCC interno e externo, para garantir um equilíbrio entre essas duas validações), foram então selecionados os cinco melhores modelos de cada proteína. Com esses modelos foram analisadas as demais métricas de validação e os resultados obtidos por eles estão representados na **Tabela 16** e nos gráficos de radar da **Figura 24**.

Tabela 16 – Métricas de validação interna e externa calculadas para os cinco melhores modelos de cada proteína.

Modelo	Validação	MCC	F1-Score	TPR	TNR	bACC	AUC	CK
saFabi								
SA_SVS	Interna	0,594	0,771	0,771	0,818	0,794	0,843	0,589
	Externa	0,710	0,846	0,815	0,893	0,854	0,878	0,709
SA_SVP	Interna	0,662	0,815	0,854	0,801	0,828	0,856	0,650
	Externa	0,638	0,821	0,852	0,786	0,819	0,843	0,637
SA_MLL	Interna	0,579	0,762	0,750	0,826	0,788	0,822	0,578
	Externa	0,711	0,857	0,889	0,821	0,855	0,909	0,709
SA_MLS	Interna	0,617	0,780	0,761	0,851	0,806	0,844	0,615
	Externa	0,673	0,836	0,852	0,821	0,837	0,892	0,673
SA_MLA	Interna	0,567	0,765	0,813	0,752	0,782	0,837	0,558
	Externa	0,750	0,877	0,926	0,821	0,874	0,837	0,746
ecFabi								
EC_SVR	Interna	0,837	0,868	0,900	0,936	0,918	0,946	0,820
	Externa	1,000	1,000	1,000	1,000	1,000	1,000	1,000
EC_SVP	Interna	0,806	0,844	0,871	0,924	0,898	0,970	0,784
	Externa	1,000	1,000	1,000	1,000	1,000	1,000	1,000
EC_MLL	Interna	0,846	0,890	0,938	0,923	0,930	0,970	0,838
	Externa	1,000	1,000	1,000	1,000	1,000	1,000	1,000
EC_MLS	Interna	0,753	0,825	0,914	0,872	0,893	0,946	0,740
	Externa	1,000	1,000	1,000	1,000	1,000	1,000	1,000
EC_MLA	Interna	0,785	0,850	0,881	0,908	0,895	0,943	0,777
	Externa	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Fonte: Autoria própria.

Figura 24 – Gráficos de radar das métricas de validação interna (rosa) e externa (roxo) calculadas para os cinco melhores modelos de aprendizado de máquina.



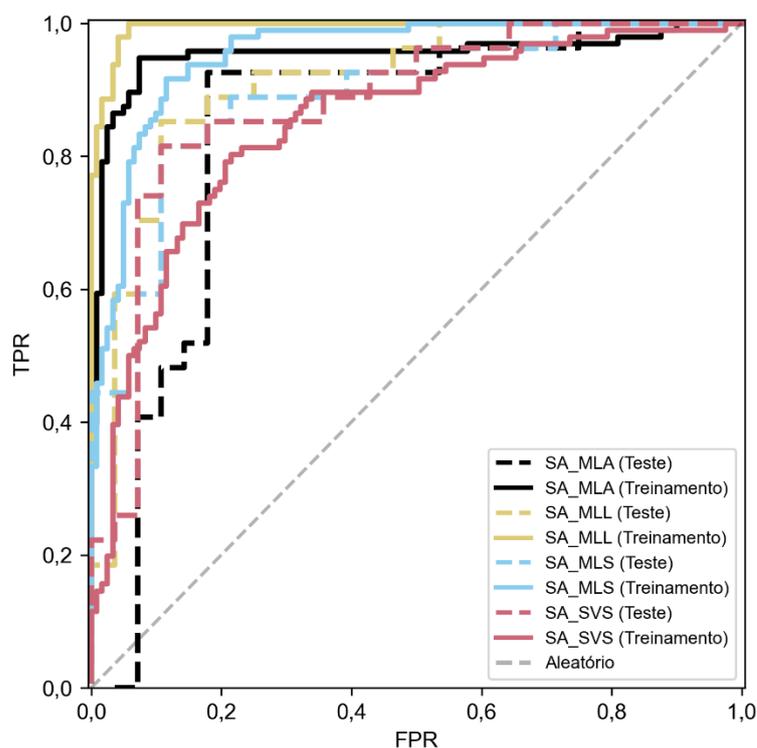
Fonte: Autoria própria.

Para realizar a triagem virtual com os modelos de aprendizado de máquina (etapa descrita na seção **4.10 Triagem virtual da biblioteca de fármacos aprovados**

pele FDA e da BraCoLi (p. 85)), optou-se pela estratégia de realizar um consenso entre os três melhores modelos de cada proteína. Para isto, procedeu-se a análise e escolha dos três melhores modelos com base nas métricas de validação calculadas.

Em relação aos modelos para a proteína de *S. aureus*, priorizou-se os dados de validação externa, porque o conjunto de treinamento é inteiramente utilizado para construção do modelo, assim como utilizado na triagem virtual. Dessa forma, o modelo SA_SVP foi o que teve a pior performance para saFabI, com TNR_{ext} de 0,786, indicando maior tendência à classificar erroneamente as moléculas negativas como positivas, MCC_{ext} de 0,638, indicando a menor capacidade preditiva geral que os demais modelos, $F1-Score_{ext}$ de 0,821 e $bACC_{ext}$ de 0,819, indicando que o maior valor de TPR_{ext} se deve ao fato do modelo classificar mais amostras como positivas. Em conclusão, o modelo tem menor acurácia e acaba por ser mais otimista que os demais, classificando erroneamente mais moléculas como ativas. Com isso, o modelo SA_SVP foi removido da seleção. Devido à dificuldade de se escolher um modelo para retirar entre os quatro restantes, foi construída a curva ROC (Figura 25) para esses modelos de aprendizado de máquina.

Figura 25 – Curvas ROC para os quatro melhores modelos de aprendizado de máquina para a saFabI.



Fonte: Autoria própria.

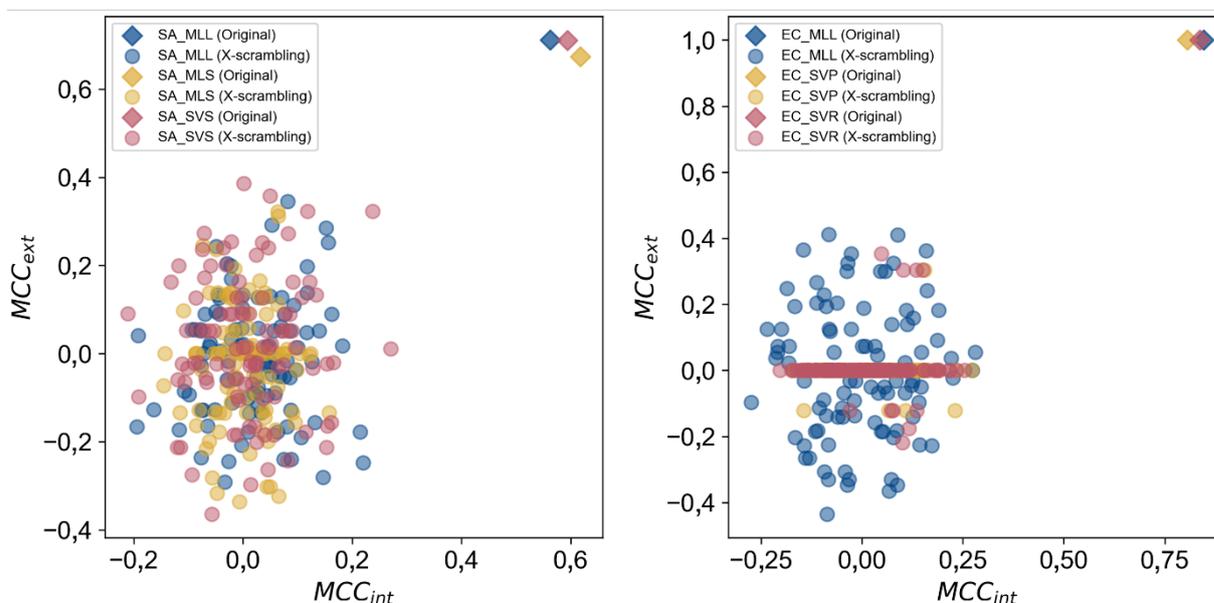
A **Figura 25** mostra que, apesar dos bons resultados das métricas de validação interna e externa do modelo SA_MLA, quando analisada as probabilidades de acerto da classe positiva (moléculas ativas) fornecidas pelo modelo e usadas para calcular a curva ROC, o modelo tem grande tendência ao erro de classificação no início da curva. Ou seja, a probabilidade fornecida para o modelo de que uma molécula seja de fato ativa não é confiável no início da curva e, portanto, esse modelo não seria adequado para uma triagem virtual. Dessa forma, esse modelo foi removido da seleção. É importante destacar que, por serem modelos de classificação, uma estratégia para contornar essa situação em uma triagem virtual seria empregar a predição de ativo/inativo (classificação) em vez de priorizar o ranqueamento pela probabilidade de ser classificado como ativo. Isso se deve ao fato de que as métricas de classificação demonstraram um desempenho notável na tarefa de classificação, garantindo uma boa acurácia e capacidade preditiva para essa função. Ainda que essa técnica tenha sido utilizada no presente estudo, uma vez que foi feito consenso entre os votos dos três modelos e não ranqueamento pela probabilidade, esse parâmetro do crescimento inicial da curva foi utilizado como critério de desempate na seleção dos três melhores modelos. Assim, os três melhores modelos para saFabl foram: SA_SVS, SA_MLL e SA_MLS.

Em relação aos modelos para a proteína de *E. coli*, todas as métricas de validação externa tiveram valores igual a 1,000 indicando classificação perfeita dos modelos. Portanto, não havia como diferenciá-los pela validação externa. Dessa forma, observou-se que o modelo EC_MLS teve os piores resultados nas métricas de MCC_{int} , $F1-Score_{int}$, TNR_{int} , $bACC_{int}$ e CK_{int} . Sendo o primeiro modelo de *E. coli* removido da seleção. O segundo modelo removido foi o EC_MLA, porque teve o pior resultado em AUC_{int} e o segundo pior resultado nas métricas MCC_{int} , TPR_{int} , TNR_{int} , $bACC_{int}$ e CK_{int} . Assim, os três melhores modelos para ecFabl foram: EC_SVR, EC_SVP e EC_MLL. Ainda que as curvas ROC para os modelos da ecFabl não precisaram ser usadas como critério de desempate, diferentemente da saFabl, as curvas desses modelos também foram construídas e serão discutidas ainda nessa seção juntamente com uma comparação dos resultados do *docking*.

Os três melhores modelos de cada proteína também foram validados por *X-scrambling*, realizado conforme mencionado na seção **4.8 Construção e validação dos modelos de aprendizado de máquina com os *fingerprints* de interação das**

poses do acoplamento molecular (p. 77). Os resultados da validação por *X-scrambling* estão dispostos na **Figura 26**.

Figura 26 – Validação *X-scrambling* para os modelos de *S. aureus* (à esquerda) e de *E. coli* (à direita).

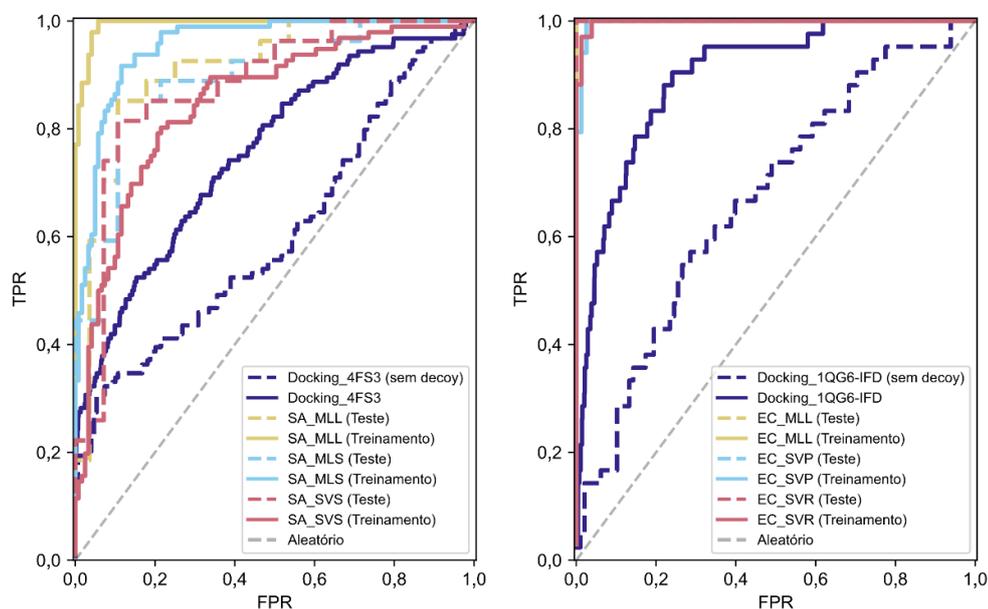


Fonte: Autoria própria.

A validação por *X-scrambling* (representada na **Figura 26**) busca comparar a qualidade dos modelos obtidos originalmente pelo conjunto de dados preparado com modelos gerados pelo acaso. Dessa forma, observa-se que os modelos obtidos pelo acaso tendem a resultar em baixos valores nas métricas de validação, (no caso, MCC_{int} e MCC_{ext}), enquanto os modelos originais têm altos valores nessas métricas. Isso indica que os modelos originais não foram obtidos ao acaso, uma vez que se distanciam dos 100 modelos obtidos pela estratégia de aleatorização.

Por fim, como última estratégia de validação, foram construídas as curvas ROC dos três melhores modelos de aprendizado de máquina de cada proteína e realizada a comparação com as curvas obtidas pelos seus respectivos protocolos de *docking* (**Figura 27**). Adicionalmente, também foram calculados os valores de BEDROC para os conjuntos de treinamento e de teste (**Tabela 17**).

Figura 27 – Curvas ROC dos três melhores modelos de aprendizado de máquina para os conjuntos de treinamento e de teste das duas proteínas, comparadas as curvas ROC dos seus respectivos protocolos de *docking* com e sem decoys.



Fonte: Autoria própria.

Tabela 17 – Valores de AUC_{ROC} e AUC_{BEDROC} para os modelos de ML (nos conjuntos de teste e de treinamento) e protocolos de *docking* (no conjunto de dados total com ou sem *decoys*).

Conjunto de dados	AUC_{ROC}	AUC_{BEDROC} ($\alpha = 160,9$)	AUC_{BEDROC} ($\alpha = 32,2$)	AUC_{BEDROC} ($\alpha = 16,1$)
saFabI				
<i>Docking_4FS3</i> (sem decoy)	0,606	1,000	0,954	0,861
<i>Docking_4FS3</i>	0,757	0,531	0,363	0,386
SA_MLL (Teste)	0,909	1,000	0,976	0,939
SA_MLL (Treinamento)	0,994	1,000	1,000	0,999
SA_MLS (Teste)	0,892	1,000	0,999	0,985
SA_MLS (Treinamento)	0,954	1,000	0,998	0,986
SA_SVS (Teste)	0,878	1,000	0,983	0,931
SA_SVS (Treinamento)	0,849	1,000	0,941	0,897
ecFabI				
<i>Docking_1QG6-IFD</i> (sem decoy)	0,671	0,762	0,603	0,548
<i>Docking_1QG6-IFD</i>	0,897	0,275	0,398	0,502
EC_MLL (Teste)	1,000	1,000	1,000	1,000
EC_MLL (Treinamento)	1,000	1,000	1,000	1,000
EC_SVP (Teste)	1,000	1,000	1,000	1,000
EC_SVP (Treinamento)	0,997	1,000	1,000	0,998
EC_SVR (Teste)	1,000	1,000	1,000	1,000
EC_SVR (Treinamento)	0,998	1,000	1,000	0,999

Fonte: Autoria própria.

Os resultados de AUC_{ROC} e de AUC_{BEDROC} evidenciam uma vantagem clara dos modelos de aprendizado de máquina em associação com o *docking*, quando comparados com o uso exclusivo do *docking*. Esta superioridade é observada tanto nos valores calculados para o conjunto de treinamento quanto para o conjunto de teste, em relação aos modelos de ambas as proteínas. Essa vantagem é também notável nas curvas ROC da ecFabI nos conjuntos de treinamento e de teste, bem como na curva ROC da saFabI no conjunto de treinamento, mesmo em termos do reconhecimento inicial.

No conjunto de teste da saFabI, embora os modelos SA_SVS e SA_MLL tenham uma vantagem geral de área sob a curva, em termos de reconhecimento inicial, a curva do *docking* com os *decoys* parece apresentar vantagem sob os demais modelos de aprendizado de máquina para o conjunto de teste. Entretanto, é importante notar que a curva ROC é altamente susceptível ao número de moléculas do conjunto de dados. Portanto, os dois conjuntos de dados não podem ser diretamente comparados com base no comportamento da curva ROC. Esta discrepância é observada ao compararmos os resultados do *docking* com e sem *decoys*, onde a remoção dos *decoys* resulta em um desempenho aparentemente reduzido na curva e, em alguns casos, uma redução no valor de AUC_{ROC} . No entanto, leva a um aumento na AUC_{BEDROC} , uma métrica mais confiável e menos susceptível ao tamanho do conjunto de dados. Dessa forma, a comparação em termos de AUC_{BEDROC} permite afirmar que os modelos de aprendizado de máquina promoveram uma grande melhoria na capacidade preditiva.

Ainda visando comparar os resultados de ML-*docking* com *docking* sozinho foi realizado os cálculos das métricas de classificação para o *docking* após uma classificação dos valores de pontuação utilizando três estratégias diferentes: (i) utilizando como limite de classificação a média aritmética entre o menor valor da classe inativa e o maior valor da classe ativa (**estratégia MinMax**, limiar de classificação: -13,154 para saFabI e -14,521 para ecFabI), (ii) utilizando como limite de classificação a média aritmética entre a média de todos os valores da classe ativa e a média de todos os valores da classe inativa (**estratégia Média**, limiar de classificação: -13,123 para saFabI e -14,543 para ecFabI) e, por último, (iii) utilizando o limiar fornecido pela curva ROC que maximiza a diferença entre a taxa de verdadeiros positivos e a taxa de falsos positivos (**estratégia ROC**, limiar de classificação: -14,298 para saFabI e -15,007 para ecFabI). Os resultados obtidos

estão representados na **Tabela 18** para a proteína de *S. aureus* e na **Tabela 19** para a proteína de *E. coli*.

Tabela 18 – Métricas de classificação para os valores classificados da pontuação do *docking* da saFabl.

Protocolo	Conjunto	MCC	F1-Score	TPR	TNR	bACC	AUC	CK	Limiar
Docking	Total*	0,121	0,496	0,468	0,651	0,559	0,606	0,120	
4FS3	Treinamento	0,129	0,497	0,474	0,653	0,564	0,617	0,128	-13,154
(MinMax)	Teste	0,089	0,490	0,444	0,643	0,544	0,566	0,088	
Docking	Total*	0,114	0,494	0,468	0,644	0,556	0,606	0,113	
4FS3	Treinamento	0,120	0,495	0,474	0,645	0,559	0,617	0,120	-13,123
(Média)	Teste	0,089	0,490	0,444	0,643	0,544	0,566	0,088	
Docking	Total*	0,336	0,447	0,306	0,946	0,626	0,606	0,267	
4FS3	Treinamento	0,348	0,451	0,309	0,950	0,63	0,617	0,277	-14,298
(ROC)	Teste	0,291	0,432	0,296	0,929	0,612	0,566	0,227	

* Total = Total sem *decoy*.

Fonte: Autoria própria.

Tabela 19 – Métricas de classificação para os valores classificados da pontuação do *docking* da ecFabl.

Protocolo	Conjunto	MCC	F1-Score	TPR	TNR	bACC	AUC	CK	Limiar
Docking	Total*	0,247	0,514	0,667	0,602	0,634	0,671	0,230	
1QG6-IFD	Treinamento	0,254	0,518	0,647	0,628	0,638	0,668	0,241	-14,521
(MinMax)	Teste	0,228	0,500	0,750	0,500	0,625	0,694	0,192	
Docking	Total*	0,225	0,500	0,643	0,602	0,622	0,671	0,211	
1QG6-IFD	Treinamento	0,227	0,500	0,618	0,628	0,623	0,668	0,217	-14,543
(Média)	Teste	0,228	0,500	0,750	0,500	0,625	0,694	0,192	
Docking	Total*	0,271	0,511	0,571	0,714	0,643	0,671	0,268	
1QG6-IFD	Treinamento	0,265	0,500	0,529	0,744	0,637	0,668	0,264	-15,007
(ROC)	Teste	0,316	0,545	0,750	0,600	0,675	0,694	0,286	

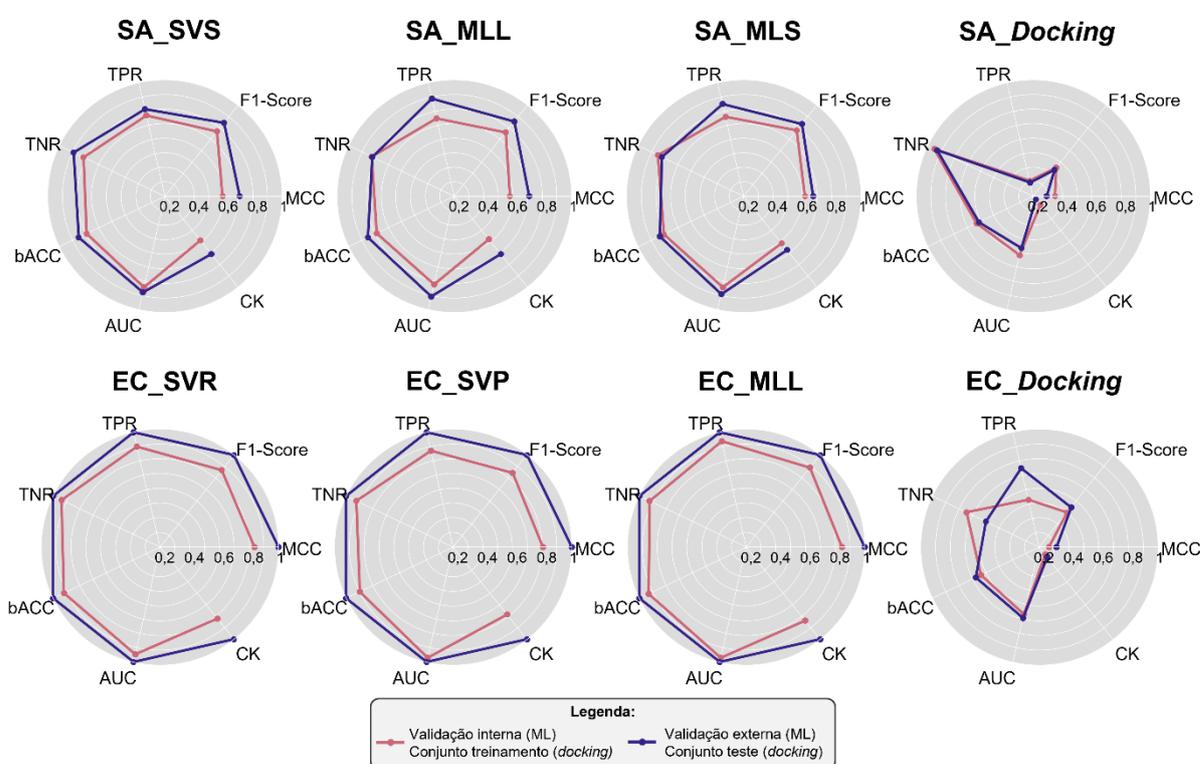
* Total = Total sem *decoy*.

Fonte: Autoria própria.

Os resultados obtidos para as métricas de classificação no *docking* mostram que, entre as estratégias testadas, a melhor estratégia para classificar os valores de pontuação do *docking* para ambas as proteínas é uso do limiar fornecido pela curva ROC que maximiza a diferença entre a taxa de verdadeiros positivos e a taxa de falsos

positivos (**estratégia ROC**), porque essa estratégia resultou em melhores valores de MCC, TNR, bACC e CK, o que indica a melhor capacidade de distinção entre falsos negativos e verdadeiros negativos, acurácia geral e capacidade preditiva geral. Quando esses resultados são comparados com os resultados do aprendizado de máquina (**Tabela 16**) é possível ver a clara superioridade dos modelos de ML. A **Figura 28** demonstra a comparação dos três melhores modelos de aprendizado de máquina com os protocolos de *docking*.

Figura 28 – Gráficos de radar dos três melhores modelos de aprendizado de máquina de cada proteína (saFabI e ecFabI) comparados com os seus respectivos protocolos de *docking*.



Fonte: Autoria própria.

Os modelos de aprendizado de máquina performaram melhor que o *docking* sozinho em todas as métricas, exceto o TNR que, para o *docking* da saFabI, teve bons resultados. Entretanto, os valores encontrados para os resultados de *docking* de ambas as proteínas indicam baixa capacidade preditiva, não sendo aplicáveis para classificação de moléculas em ativos e inativos ($MCC < 0,5$), enquanto o aprendizado de máquina resultou em modelos preditivos e, até mesmo, perfeitos com MCC_{teste} (MCC_{ext}) de 1,000. Em termos quantitativos, quando comparado com os resultados de

docking, os modelos de aprendizado de máquina representaram ganhos de $MCC_{int}/MCC_{treinamento}$ entre 0,219 e 0,314, para a saFabI, e entre 0,520 e 0,581, para a ecFabI. Em termos de MCC_{ext} (MCC_{teste}) os ganhos foram maiores, variando de 0,347 a 0,459 para a saFabI e de exatamente 0,684 para a ecFabI.

Com isso, foi possível validar os modelos de aprendizado de máquina e comprovar que o uso do ML-*docking* é uma estratégia eficaz e melhor, em termos de capacidade preditiva, do que o uso do *docking* sozinho no contexto de inibidores da saFabI e da ecFabI. Dessa forma, demonstra a capacidade de utilização desses modelos para estratégias de triagem virtual de potenciais inibidores da FabI de *S. aureus* e de *E. coli*.

5.4 INTERPRETAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Para realizar a interpretação dos modelos de aprendizado de máquina foi calculada a contribuição de cada variável (*bit de fingerprint*) em termos de redução ou aumento nos valores do $MCC_{externo}$ conforme descrito na seção **4.9 Interpretação dos modelos de aprendizado de máquina e das principais características de interação ligante-proteína (p. 84)**. Além disso, com base na média das variáveis, foi analisado se a frequência de cada variável do conjunto de dados era maior entre as moléculas ativas ou entre as inativas.

A interpretação dos *fingerprints* de interação do LUNA apresenta algumas limitações. Embora seja possível extrair dados lineares sobre os átomos, ligações, conexões e interações, assim como representações tridimensionais de cada *bit de fingerprint* em cada molécula do conjunto de dados, é difícil estabelecer correlações espaciais para *bits* que codificam informações exclusivas dos ligantes. Isso ocorre porque o algoritmo não fornece uma definição geral do que cada *bit* representa em termos de suas características tridimensionais para esse tipo específico de informação. Em vez disso, ele retorna o que foi observado em cada molécula para cada *bit* e, com alguma frequência, ocorrem colisões de *bits* (estruturas diferentes codificadas no mesmo *bit*).

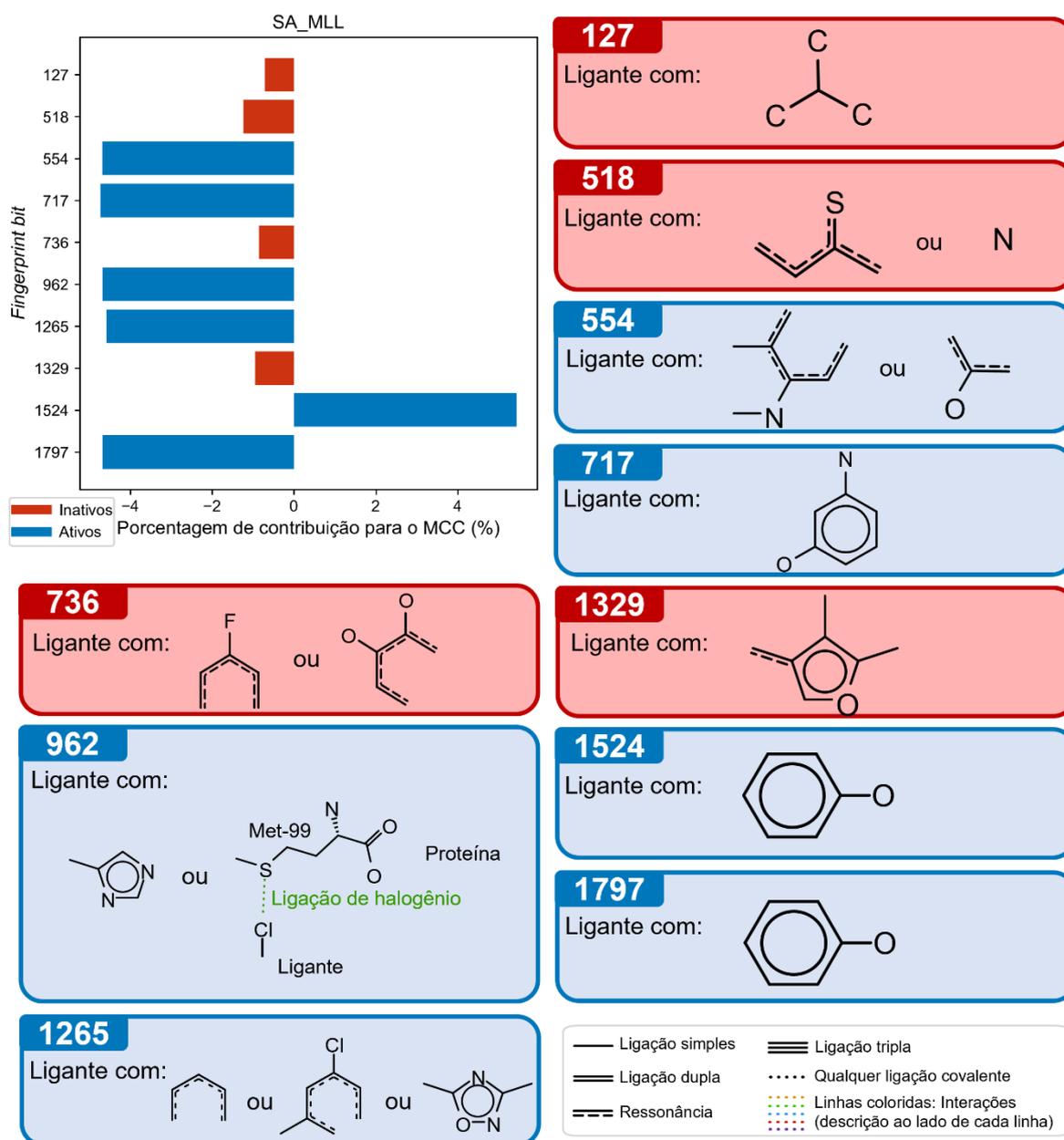
A presença de colisão de *bits* e a ausência de uma definição geral com dados tridimensionais sobre o que está sendo computado nos *bits* que contêm informações exclusivas dos ligantes dificulta a interpretação dos modelos. Portanto, com a interpretação disponível atualmente no algoritmo, é possível encontrar fragmentos

estruturalmente semelhantes em *bits* diferentes, contribuindo de maneiras diferentes para a atividade. Isso demonstra que existe um componente do espaço tridimensional que não é facilmente generalizado para *bits* que contêm exclusivamente informações estruturais do ligante.

Sobre a colisão de *bits*, esse fenômeno não apresentou grande impacto sobre a capacidade de predição dos modelos. Os modelos foram extensamente validados e apresentaram altos valores nas métricas de validação *5-fold* e externa, além de comportamento adequado para curva ROC e, por último, a validação por *X-scrambling* mostra que os modelos não foram obtidos aleatoriamente e as características estruturais definidas pelo *fingerprint* do LUNA apresentam correlação com as classes de atividade. Isso sugere que o ruído criado pela colisão em um *bit* possa ser compensado pela correlação com outros *bits* ou que os fragmentos da colisão têm alguma correlação entre si. A não interferência da colisão de *bits* na qualidade dos modelos era esperada, uma vez que no trabalho original do algoritmo do LUNA (Fassio *et al.*, 2022), foram testados *fingerprints* com comprimento de 16384. Esses *fingerprints* maiores tendem a apresentar menor número de colisões de *bits*, mas apresentaram capacidade preditiva semelhante aos de 4096 e com maior gasto computacional, principalmente, para treinar os modelos de aprendizado de máquina. Além disso, de acordo com o princípio da navalha de Occam (princípio da parcimônia), deve ser priorizado o modelo mais simples (com menor número de variáveis) entre os mais competentes para explicar um fenômeno, no caso o de classificar corretamente as moléculas entre ativos e inativos (Dresp-Langley *et al.*, 2019).

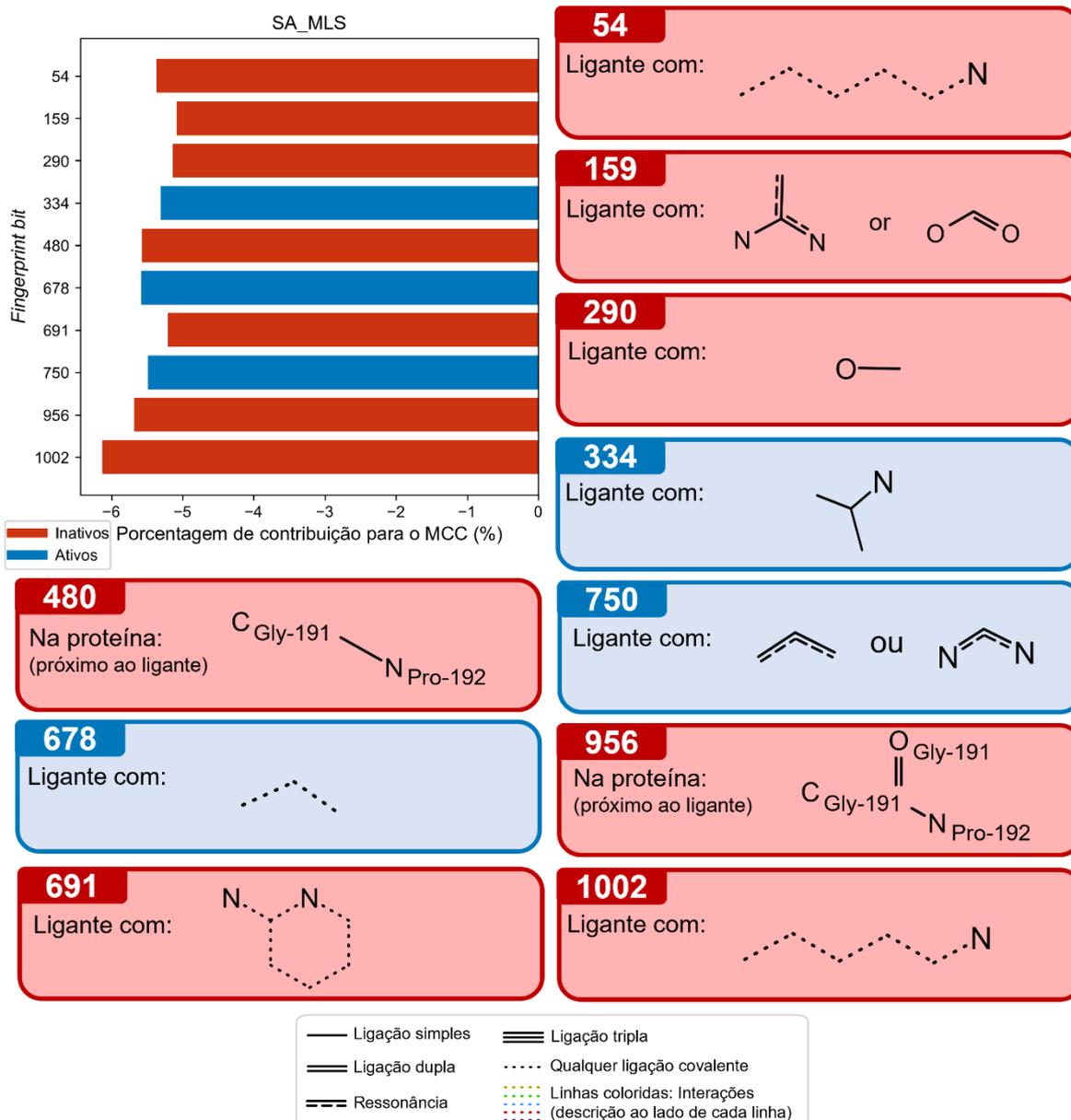
Com essas considerações, foi realizada a interpretação dos modelos buscando as características presentes entre os *bits* com maior contribuição para o valor de MCC_{ext} . Nas **Figuras 29, 30 e 31**, estão representadas as interpretações realizadas para os três melhores modelos da saFabi (SA_MLL, SA_MLS e SA_SVS, respectivamente).

Figura 29 – Interpretação dos *fingerprints* de interação do modelo SA_MLL.



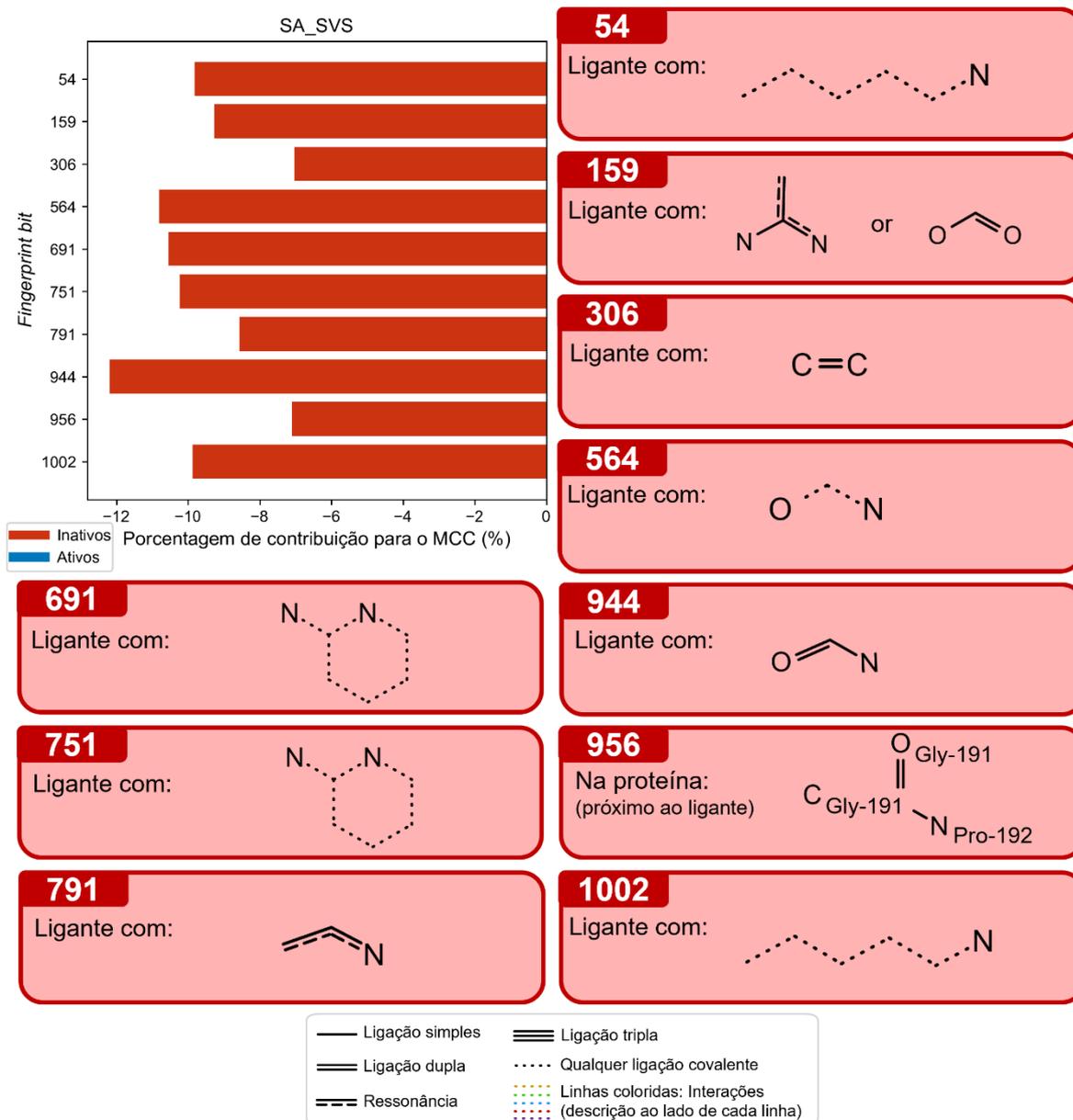
Fonte: Autoria própria.

Figura 30 – Interpretação dos *fingerprints* de interação do modelo SA_MLS.



Fonte: Autoria própria.

Figura 31 – Interpretação dos *fingerprints* de interação do modelo SA_SVS.



Fonte: Autoria própria.

Em relação a FabI de *S. aureus*, foi possível observar que a presença da Gly-191 e da Pro-192 com proximidade de até 2,86 Å (definida pelo tamanho máximo da esfera) contribui negativamente para a atividade e tem grande impacto na predição dos modelos SA_MLS e SA_SVS, como visto nos *bits* 956 (ambos os modelos) e 480 (SA_MLS). Por outro lado, a presença de uma interação do tipo ligação de halogênio com o enxofre da Met-99 contribui positivamente para a atividade e tem grande impacto na predição do modelo SA_MLL.

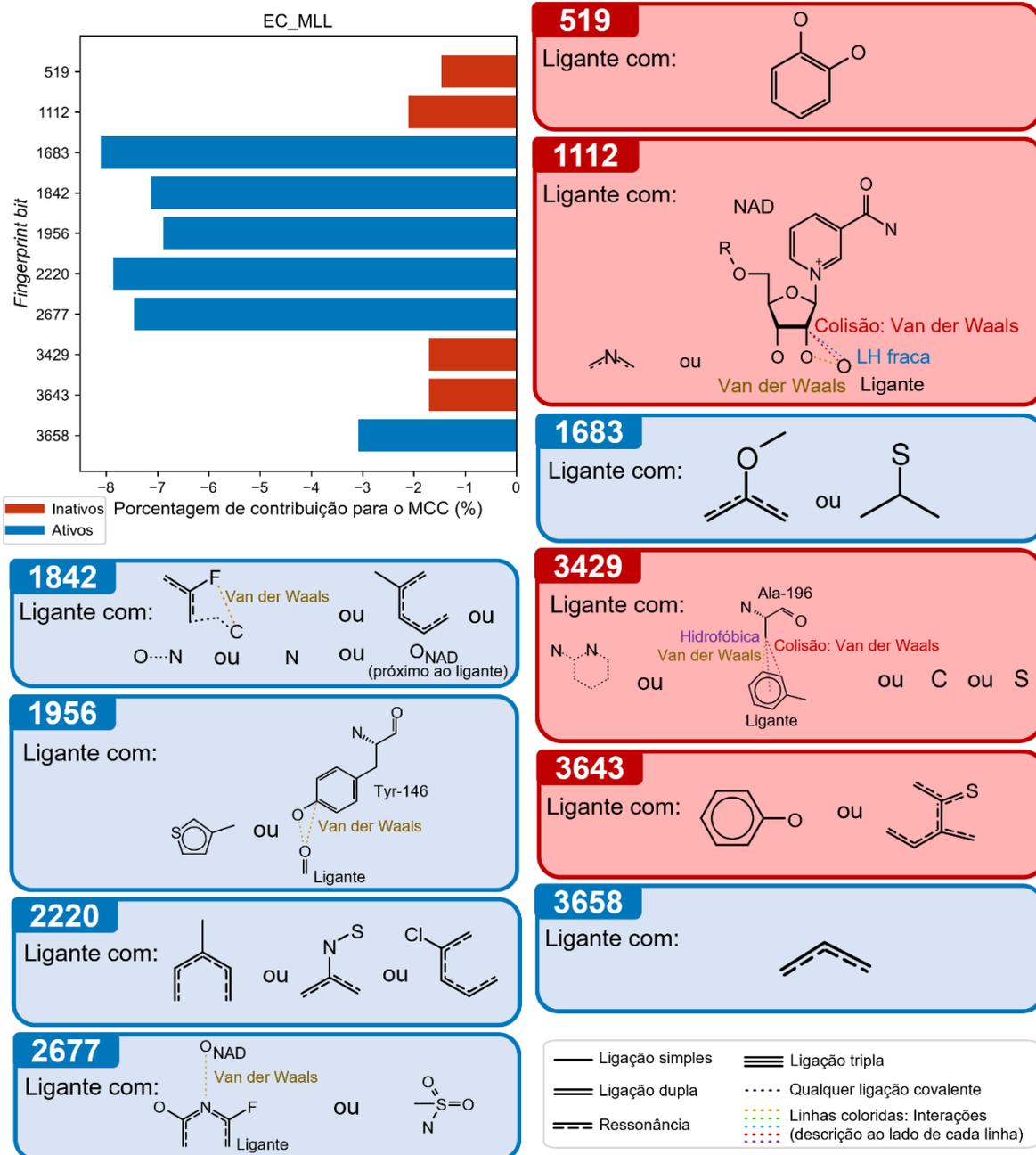
É importante observar que, em geral, grupos fenólicos (SA_MLL *bits* 717, 1524 e 1797) contribuem positivamente para a atividade com grande impacto na qualidade dos modelos, mas grupos fenólicos com hidroxila vicinal (SA_MLL *bit* 736) estão relacionados com uma menor contribuição para a atividade, ainda que com menor impacto sobre a qualidade do modelo. Curiosamente, a aleatorização dos valores do *bit* 1524 em SA_MLL levou a uma melhora da qualidade do modelo. A presença de grupos estruturalmente semelhantes em *bits* diferentes, como discutido anteriormente, está relacionada principalmente com uma limitação do algoritmo do LUNA em fornecer interpretações gerais das características tridimensionais que diferem esses *bits*. Ainda sobre grupos fenólicos, o *bit* 717 (SA_MLL) mostra grupos 3-aminofenólicos também contribuem positivamente para a atividade e tem impacto sobre os modelos de aprendizado de máquina.

Sobre grupos com nitrogênio, observa-se que grupos com cinco carbonos e um ou dois nitrogênios contribuem negativamente para modelos SA_MLS e SA_SVS (*bits* 54, 691, 751, 1002). Por outro lado, os grupos imidazol e oxadiazol têm contribuição positiva para a interação com a enzima (*bit* 956 e 1265 do modelo SA_MLL, respectivamente).

Por último, considerando o modelo SA_MLL, há uma preferência por grupos aromáticos com cloro (clorobenzeno, por exemplo) (*bit* 1265), em comparação com grupos com flúor (fluorobenzeno, por exemplo) (*bit* 736). E, ainda, anéis furano contribuem negativamente para a interação com a enzima saFabI (*bit* 1329).

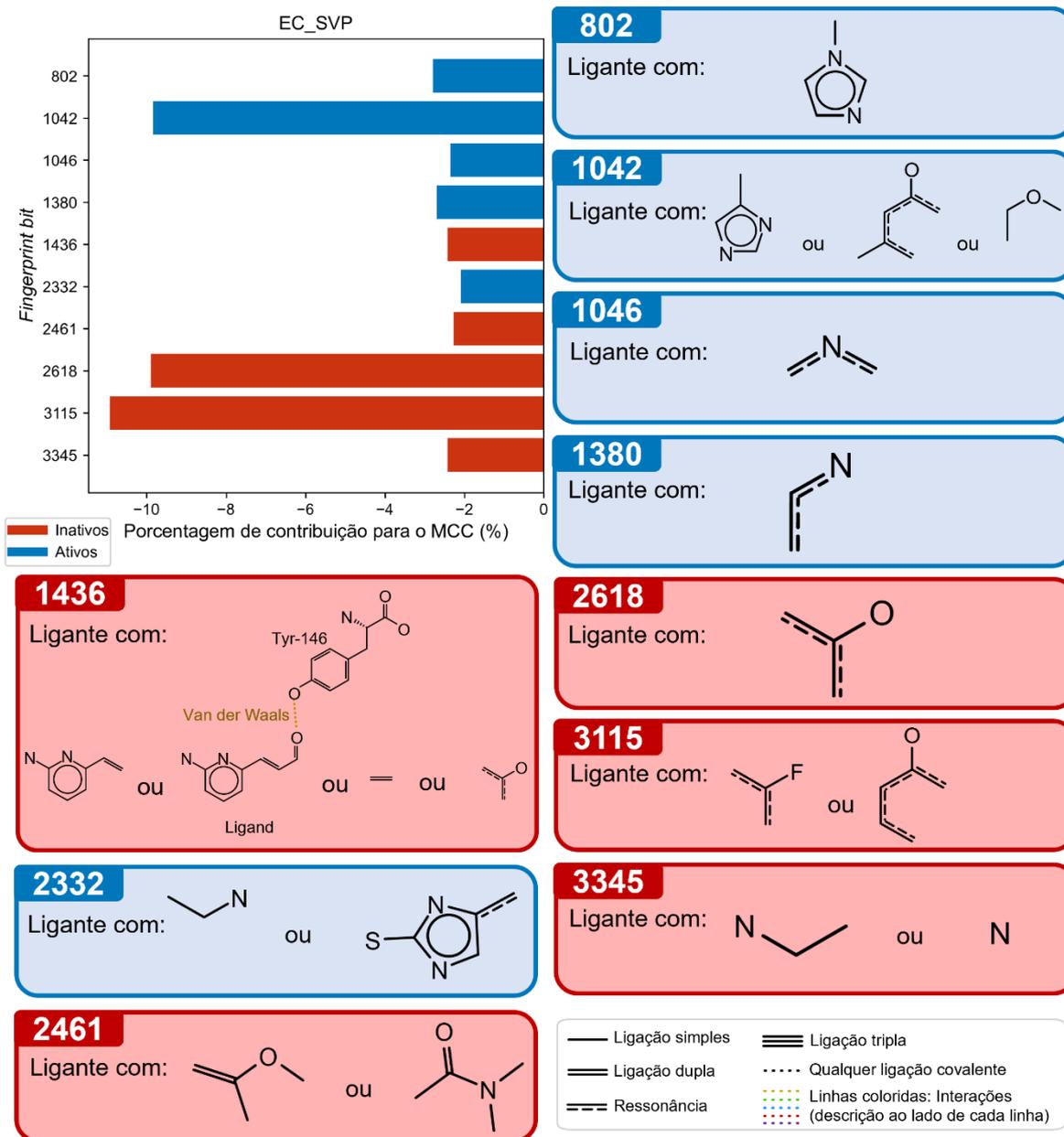
Em relação aos três melhores modelos da ecFabI, as interpretações estão representadas nas **Figuras 32, 33 e 34** (EC_MLL, EC_SVP e EC_SVR, respectivamente).

Figura 32 – Interpretação dos *fingerprints* de interação do modelo EC_MLL.



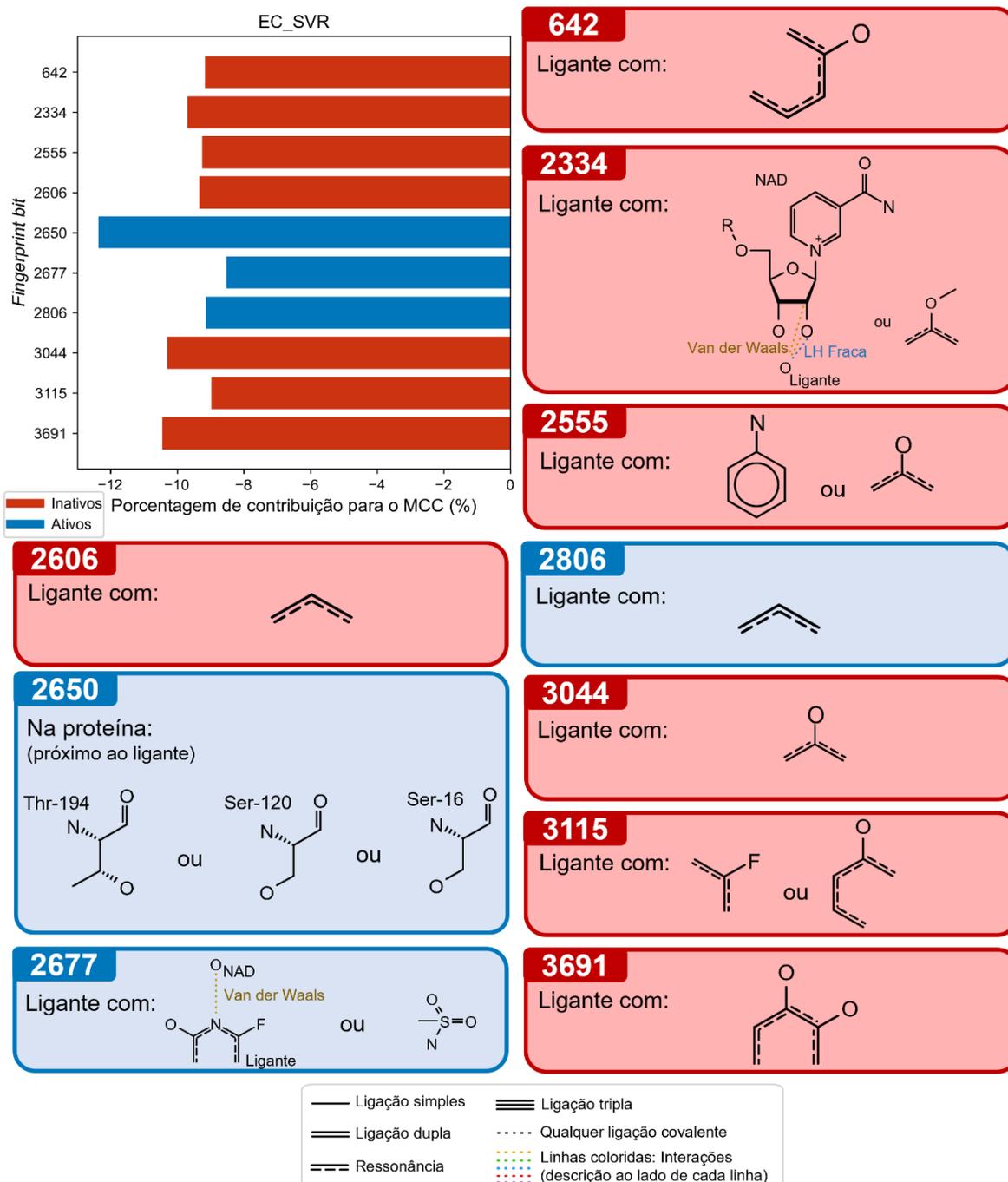
Fonte: Autoria própria.

Figura 33 – Interpretação dos *fingerprints* de interação do modelo EC_SVP.



Fonte: Autoria própria.

Figura 34 – Interpretação dos *fingerprints* de interação do modelo EC_SVR.



Fonte: Autoria própria.

Em relação, as principais observações que possuem significado estrutural para os modelos da ecFabI é possível verificar que a interação com o NAD é relevante para a atividade biológica e para os modelos de aprendizado de máquina (*bit* 2677 para a EC_MLL e EC_SVR e sua presença do NAD próximo ao ligante no *bit* 1846 do modelo EC_MLL). Colisões com o NAD também levaram a um aumento da probabilidade de que as moléculas sejam classificadas como inativas (*bit* 1112 do

modelo EC_MLL). Essa interação com a 2'-hidroxila da ribose de nicotinamida do NAD é considerada, na literatura, como importante para a atividade, corroborando para a validade dos achados para este modelo (Yang *et al.*, 2017; Maltarollo *et al.*, 2022). Contudo, com base no modelo EC_SVR e no conjunto de dados, parece haver uma preferência com que essa interação ocorra por átomos de nitrogênio do ligante (*bit* 2677), mais do que por átomos de oxigênio (*bit* 2334).

Colisões com a Ala-196 também são desfavoráveis para a atividade e desempenham papel importante na classificação do modelo EC_MLL (*bit* 3429). A interação com esse resíduo também é apontada na literatura como importante para a afinidade de ligação (Yang *et al.*, 2017).

Outro fato também conhecido pela literatura é a importância da interação com a Tyr-146 para a atividade biológica (Maltarollo *et al.*, 2022). Esse fato foi também observado na interpretação do modelo EC_MLL pela contribuição do *bit* 1956 para a atividade biológica com grande impacto sobre o MCC do modelo gerado. Entretanto, quando essa interação é realizada pelo grupo 3-(6-aminopiridin-2-il)prop-2-enal, a contribuição para a atividade é negativa. Isso ocorre, porque a presença desse grupo já indica maior probabilidade de que a molécula seja inativa (*bit* 1436 do modelo EC_SVP).

A presença dos resíduos de Thr-194, Ser-120 ou Ser-16 (*bit* 2650) em uma distância máxima de até 4,30 Å dos ligantes levou a uma maior probabilidade do ligante ser classificado como ativo pelo modelo EC_SVR. Esse modelo também aponta que 3 carbonos em ressonância (*bits* 2606 e 2806) apresentam impacto significativo no desempenho do modelo, entretanto, o *bit* 2606 está mais presente entre inativos e o *bit* 2806 está mais presente entre ativos. Essa observação permite demonstrar a limitação de que existe de fato um componente do espaço tridimensional que não é facilmente observável para *bits* que trazem informações estruturais apenas do ligante, portanto, alguma informação se perde impedindo a compreensão de como o algoritmo identifica diferencialmente esses dois *bits*.

Assim como visto na saFabI, na ecFabI a presença de um fenol com hidroxilas vicinais também tem impacto na qualidade dos modelos EC_MLL (*bit* 519) e EC_SVR (*bit* 3691) e contribui negativamente para a atividade. Outro fato observado para saFabI que se repete para a ecFabI é a presença de anéis imidazólicos que contribuem positivamente para a atividade e desempenham grande importância no desempenho do modelo EC_SVP (*bits* 802, 1042 e 2332). A presença do átomo de

flúor, ligado a um carbono em ressonância com outros dois carbonos, também contribuiu negativamente para a atividade biológica, com impacto significativo no desempenho dos modelos EC_SVP e EC_SVR (*bit* 3115). Contudo, quando o flúor é capaz de estabelecer interações intramoleculares do tipo van der Waals (*bit* 1842 do modelo EC_MLL), de acordo com a classificação de interações do LUNA, há uma maior probabilidade de que a molécula seja classificada como ativa.

Como última observação sobre os modelos da enzima ecFabI, nota-se que sulfonamidas contribuíram positivamente para a atividade biológica (EC_MLL *bit* 2677) e anilinas negativamente (EC_SVR *bit* 2555).

Por fim, foi possível observar que, em relação aos modelos construídos para as duas enzimas, o algoritmo de MLP utilizando como *solver* o algoritmo LBFGS (MLL) foi capaz de fornecer informações estruturais mais relevantes e condizentes com a bibliografia (Yang *et al.*, 2017; Maltarollo *et al.*, 2022). A aplicação deste algoritmo em conjunto com diferentes configurações de parâmetros do LUNA, voltadas para a minimização de colisões de *bits* e para a ampliação da capacidade de alocação de informações referentes às interações intermoleculares nos *fingerprints*, pode facilitar a superação das dificuldades encontradas na interpretação dos modelos. Algumas observações nesse sentido incluem: (i) o aumento do comprimento do *fingerprint* permite minimizar o efeito das colisões e alocar mais informações nos *bits* (a escolha do comprimento deve considerar um balanço entre eficiência e gasto computacional, mas neste estudo foi possível obter conclusões significativas utilizando o comprimento de 4096); (ii) o uso de esferas menores (*radius step* de 1,433 Å ou menor) em conjunto com maior número de níveis (acima de 2 níveis) pode permitir com que mais iterações do algoritmo sejam feitas, refinando melhor as informações alocadas nos *bits* de *fingerprint* e aumentando o número de *bits* que consideram mais as interações intermoleculares; (iii) a implementação de um algoritmo que seja capaz de, em uma única representação de um complexo proteína-ligante, considerar todos os *bits* e atribuir pesos para cada um deles conforme sua importância para a predição e contribuição para a atividade, por exemplo, como são os mapas de contribuição do HQSAR (*Hologram Quantitative Structure-Activity Relationship*) (Chhatbar *et al.*, 2019; Tong *et al.*, 2021) e os construídos com o RDKit (Neves *et al.*, 2020; Djokovic *et al.*, 2023).

5.5 AVALIAÇÃO DO DOMÍNIO DE APLICABILIDADE E TRIAGEM VIRTUAL DE POTENCIAIS INIBIDORES DA FabI

Os *fingerprints*, específicos para cada um dos três melhores modelos de aprendizado de máquina de cada enzima, foram utilizados para avaliar o domínio de aplicabilidade. A **Tabela 20** mostra a correlação entre os modelos e os parâmetros específicos de cada *fingerprint* que os descrevem.

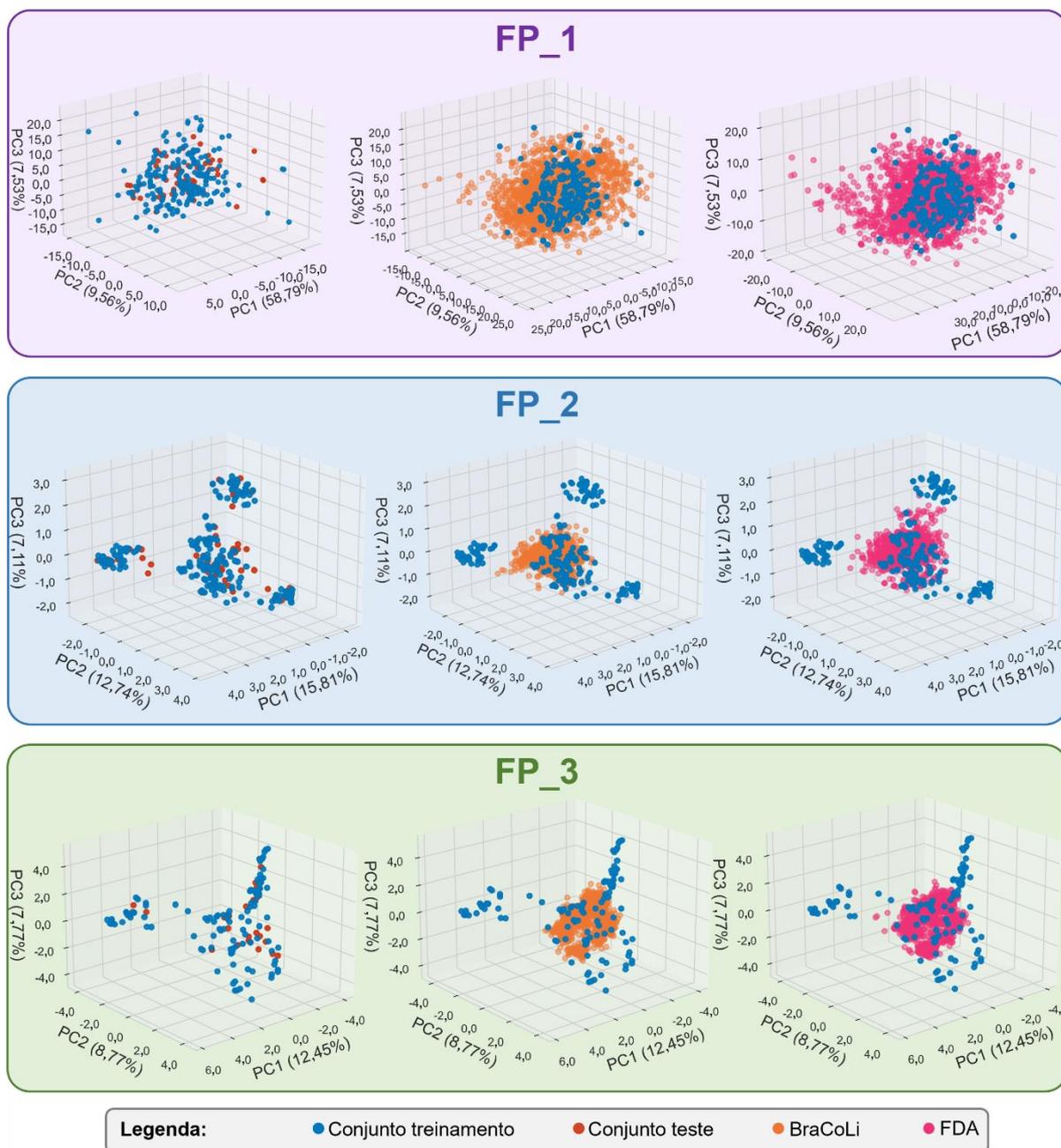
Tabela 20 – Relação entre os melhores modelos e seus respectivos *fingerprints*.

Modelo	ID do <i>fingerprint</i>	Tipo de <i>fingerprint</i>	Comprimento do <i>fingerprint</i>	Tamanho da esfera	Número de níveis	Característica computada em cada <i>bit</i>
saFabI						
SA_MLL	FP_1	EIFP	2048	1,43293	3	<i>counts</i>
SA_MLS	FP_2	EIFP	1024	1,43293	2	<i>bits</i>
SA_SVS	FP_2	EIFP	1024	1,43293	2	<i>bits</i>
ecFabI						
EC_MLL	FP_3	FIFP	4096	1,43293	3	<i>bits</i>
EC_SVP	FP_3	FIFP	4096	1,43293	3	<i>bits</i>
EC_SVR	FP_3	FIFP	4096	1,43293	3	<i>bits</i>

Fonte: Autoria própria.

Na **Tabela 20** é possível perceber que os seis melhores modelos de aprendizado de máquina resultaram em apenas três *fingerprints* diferentes. Esses *fingerprints* foram utilizados para a avaliação do domínio de aplicabilidade de acordo com metodologia de Fernandes et al. (2021) previamente descrita na seção **4.11 Avaliação do domínio de aplicabilidade para o aprendizado de máquina (p. 86)**. A **Figura 35** mostra as componentes principais, compostas pelas características dos *fingerprints* calculadas, das moléculas em um espaço tridimensional. Essa representação permite a visualização do domínio de aplicabilidade e, por ela, aparentemente todas as moléculas do conjunto de teste estão dentro do domínio de aplicabilidade dos modelos, uma vez que não há grandes distâncias entre moléculas do teste para o conjunto de treinamento. Para as bibliotecas empregadas na triagem virtual, o gráfico sugere que para o FP_2 e FP_3 não há nenhuma molécula fora do domínio de aplicabilidade. Entretanto, para o FP_1, utilizado no modelo SA_MLL, parecem existir moléculas fora do domínio de aplicabilidade.

Figura 35 – Representação das moléculas em componentes principais obtidas pelos diferentes *fingerprints* de interação utilizados nos melhores modelos.



Fonte: Autoria própria.

Para confirmar os achados das interpretações visuais da PCA com três componentes principais, procedeu-se a uma análise mais robusta envolvendo a redução de dimensionalidade com PCA para um número de componentes principais que explicassem pelo menos 85% da variância do conjunto de dados. Com as métricas de distância entre cada ponto foi possível determinar quais moléculas e quantas delas estavam fora do domínio de aplicabilidade usando a metodologia

definida na seção **4.11 Avaliação do domínio de aplicabilidade para o aprendizado de máquina (p. 86)**. A **Tabela 21** mostra quantas moléculas ficaram fora do domínio de aplicabilidade de cada *fingerprint* para cada conjunto de dados.

Tabela 21 – Número de moléculas fora do domínio de aplicabilidade (N_{DA}) para cada *fingerprint* de interação e a porcentagem do conjunto de dados que esse número representa ($\%_{DA}$).

Conjunto	N_{DA} no	$\%_{DA}$ no	N_{DA} no	$\%_{DA}$ no	N_{DA} no	$\%_{DA}$ no
	FP_1	FP_1	FP_2	FP_2	FP_3	FP_3
Teste	2	3,63 %	0	0,00 %	0	0,00 %
BraCoLi	322	16,98 %	0	0,00 %	0	0,00 %
FDA	294	18,44 %	0	0,00 %	0	0,00 %

Fonte: Autoria própria.

A **Tabela 21** confirma em parte os resultados obtidos pela interpretação dos gráficos gerados pela PCA com três componentes principais. De fato, apenas o *fingerprint* FP_1 apresenta moléculas fora do domínio de aplicabilidade para os conjuntos de dados da triagem virtual, mas os resultados obtidos destacam a presença de duas moléculas fora do domínio de aplicabilidade no conjunto de teste do FP_1. Esses resultados, mostram que, mesmo considerando todas as componentes principais utilizadas para treinar os modelos de aprendizado de máquina, poucas moléculas do conjunto de teste estão de fora do domínio de aplicabilidade. Isso demonstra o sucesso do MASSA Algorithm em separar conjuntos de dados, obtendo menos de 5 % das moléculas do conjunto de teste fora do domínio de aplicabilidade. Curiosamente, poucas moléculas também estavam fora do domínio de aplicabilidade nos conjuntos de triagem virtual, com 2 *fingerprints* (que representam 5 dos 6 melhores modelos) apresentando nenhuma molécula fora do domínio. Esse resultado é muito interessante, porque demonstra uma capacidade ímpar dos *fingerprints* do LUNA de gerar *bits* que possuem boa representatividade das características moleculares mais comuns observadas no sítio ativo dos ligantes.

Como o FP_1 é utilizado apenas pelo modelo SA_MLL, e o modelo foi capaz de prever corretamente a classe das moléculas do conjunto de teste fora do domínio de aplicabilidade (vide $MCC_{ext} = 1,000$), optou-se por seguir a triagem virtual mantendo esse modelo e estabelecendo um consenso entre ele e os demais modelos da saFabI. Um consenso entre os modelos da ecFabI também foi realizado, sendo selecionadas

as moléculas que foram ativas em pelo menos dois dos três melhores modelos de cada proteína. Para a proteína FabI de *S. aureus* um segundo filtro foi aplicado devido ao alto número de moléculas, sendo mantidas apenas as moléculas que foram ativas nos três modelos da saFabI. Todas as moléculas selecionadas tiveram suas poses no *docking* avaliadas por inspeção visual e as moléculas mais promissoras tiveram sua aquisição consultada. A estrutura dos ligantes selecionados na triagem virtual que estavam disponíveis para aquisição está representada na **Tabelas 22 e 23**, para saFabI e ecFabI, respectivamente. Esses ligantes foram então encaminhados para ensaios *in vitro* para avaliação da atividade biológica frente as bactérias *S. aureus* e *E. coli*.

Tabela 22 – Moléculas selecionadas na triagem virtual para a saFabI e sua classificação em ativa (1) e inativa (0) para a inibição enzimática de acordo com as predições dos modelos.

(continua)

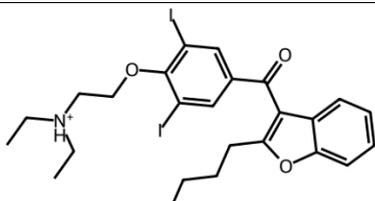
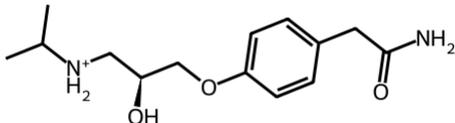
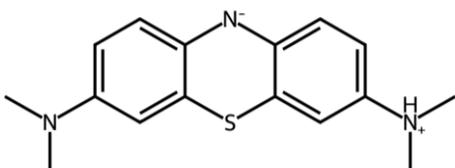
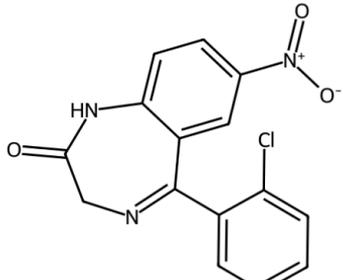
ID	Estrutura	SA_MLL	SA_MLS	SA_SVS	Total de votos
ZINC000003830212 (Amiodarona)		1	1	1	3
ZINC000000113415 (Atenolol)		1	1	1	3
ZINC000012414057 (Azul de metileno)		1	1	1	3
ZINC000003813003 (Clonazepam)		1	1	1	3

Tabela 22 – Moléculas selecionadas na triagem virtual para a saFabI e sua classificação em ativa (1) e inativa (0) para a inibição enzimática de acordo com as predições dos modelos.

(continuação)

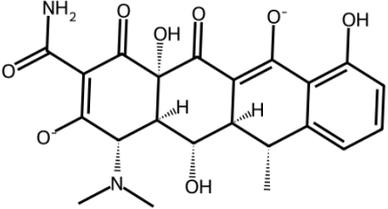
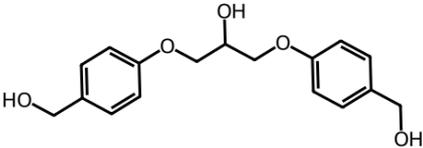
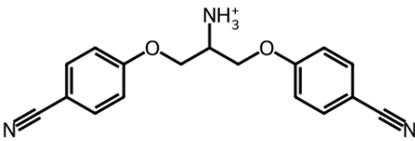
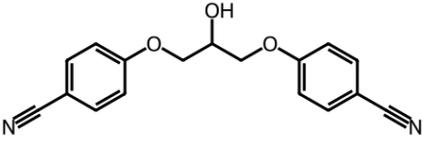
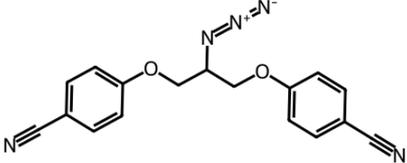
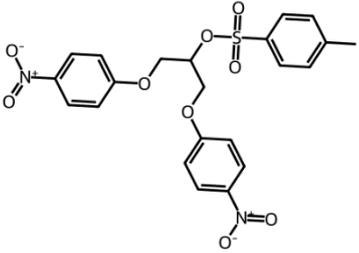
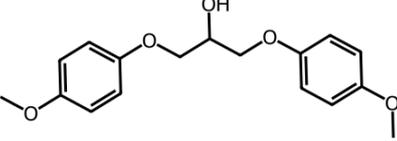
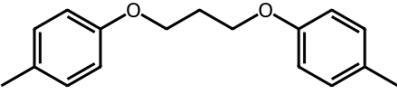
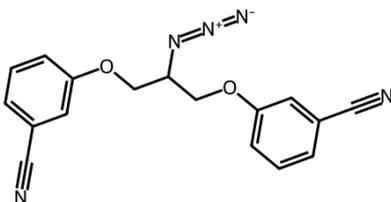
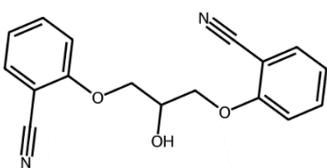
ID	Estrutura	SA_MLL	SA_MLS	SA_SVS	Total de votos
ZINC000016052277 (Doxiciclina)		1	1	1	3
RML2018_9		1	1	1	3
SNL2016_5		1	1	1	3
SNL2016_26		1	1	1	3
SNL2016_68		1	1	1	3
SNL2016_111		1	1	1	3
SNL2016_34		1	1	1	3
SNL2016_119		1	1	1	3

Tabela 22 – Moléculas selecionadas na triagem virtual para a saFabI e sua classificação em ativa (1) e inativa (0) para a inibição enzimática de acordo com as predições dos modelos.

ID	Estrutura	SA_MLL	SA_MLS	SA_SVS	(conclusão)
					Total de votos
SNL2016_67		1	1	1	3
SNL2016_24		1	1	1	3

Fonte: Autoria própria.

Tabela 23 – Moléculas selecionadas na triagem virtual para a ecFabI e sua classificação em ativa (1) e inativa (0) para a inibição enzimática de acordo com as predições dos modelos.

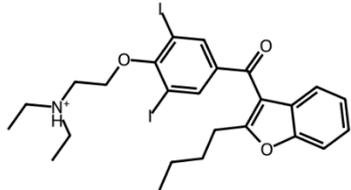
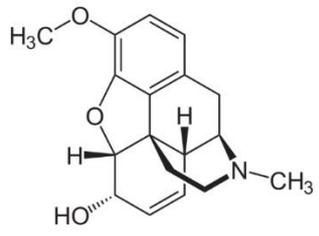
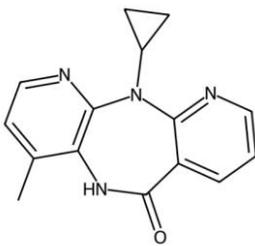
ID	Estrutura	EC_MLL	SA_SVP	SA_SVR	(continua)
					Total de votos
ZINC000003830212 (Amiodarona)		1	1	1	3
ZINC000003806721 (Codeína)		1	1	1	3
ZINC000000004778 (Nevirapina)		1	1	1	3

Tabela 23 – Moléculas selecionadas na triagem virtual para a ecFabI e sua classificação em ativa (1) e inativa (0) para a inibição enzimática de acordo com as predições dos modelos.

(continuação)

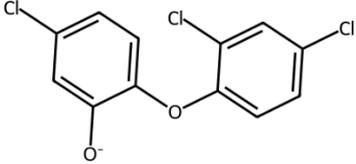
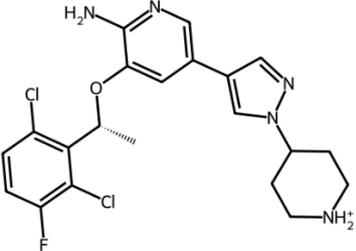
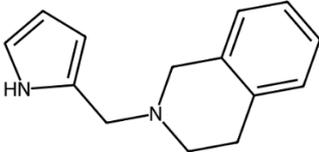
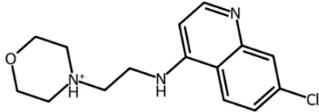
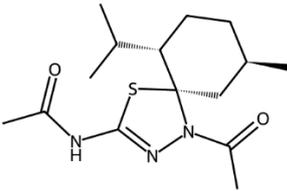
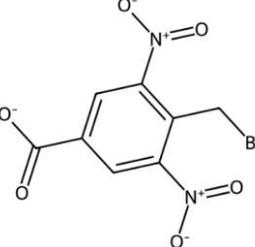
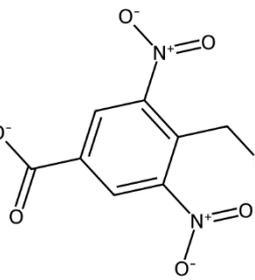
ID	Estrutura	EC_MLL	SA_SVP	SA_SVR	Total de votos
ZINC00000002216 (Triclosan)		1	1	1	3
ZINC000035902489 (Crizotinibe)		1	1	1	3
SFPB2016_25c		1	1	1	3
LCM2016_20a		1	1	1	3
PLBR2019_tiossemicarbazonas_E(S)		1	1	1	3
MSL2016_13*		1*	1*	1*	3*
EDNCI*		0*	0*	0*	0*

Tabela 23 – Moléculas selecionadas na triagem virtual para a ecFabI e sua classificação em ativa (1) e inativa (0) para a inibição enzimática de acordo com as predições dos modelos.

(continuação)

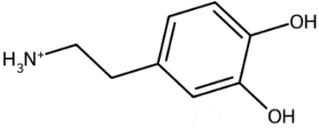
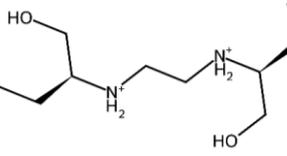
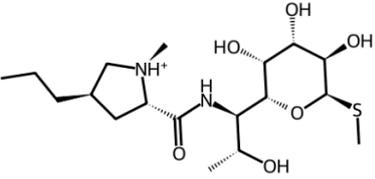
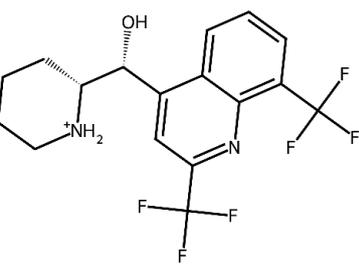
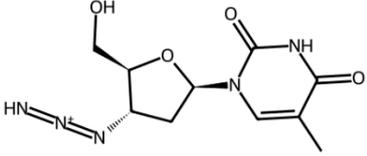
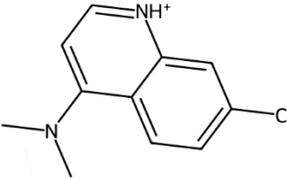
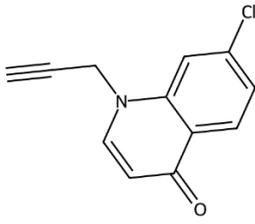
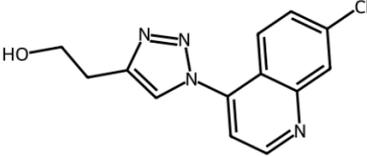
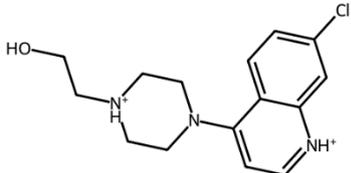
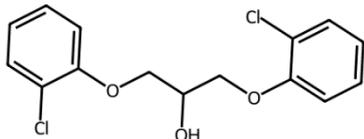
ID	Estrutura	EC_MLL	SA_SVP	SA_SVR	Total de votos
ZINC00000033882 (Dopamina)		1	1	0	2
ZINC000019364219 (Etambutol)		1	1	0	2
ZINC000003982483 (Lincomicina)		0	1	1	2
ZINC000000537964 (Mefloquina)		1	0	1	2
ZINC000003779042 (Zidovudina)		0	1	1	2
MMSA2018_47		1	0	1	2
MMSA2018_49		1	0	1	2
MMSA2018_27		1	0	1	2

Tabela 23 – Moléculas selecionadas na triagem virtual para a ecFabI e sua classificação em ativa (1) e inativa (0) para a inibição enzimática de acordo com as predições dos modelos.

ID	Estrutura	EC_MLL	SA_SVP	SA_SVR	(conclusão)
					Total de votos
LCM2016_24		1	0	1	2
SNL2016_38		1	0	1	2

* A substância MSL2016_13 não estava disponível, entretanto, um derivado com o cloro na posição do bromo (EDNCI) foi encaminhado para os testes *in vitro*. Esse derivado não constava na BraCoLi e sua predição não indicou atividade. Entretanto, devido a sua similaridade estrutural com a substância MSL2016_13 e a presença de um átomo de cloro, relatado pela interpretação dos modelos como importante para a atividade biológica, optou-se por utilizar essa substância para os ensaios *in vitro*.

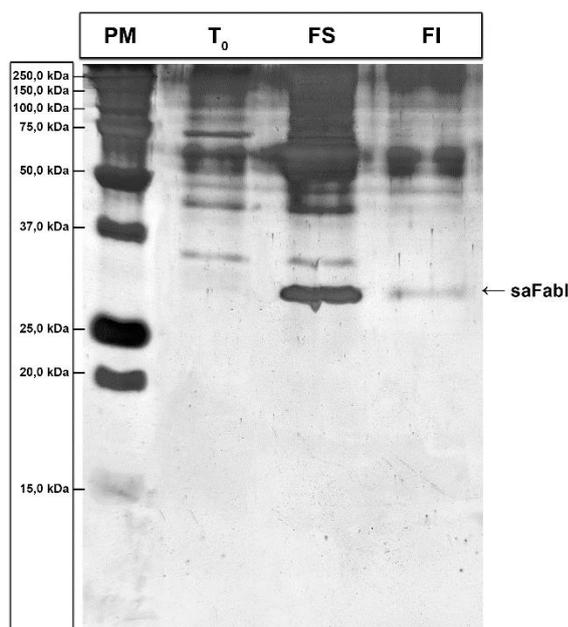
Fonte: Autoria própria.

5.6 EXPRESSÃO, PURIFICAÇÃO E CARACTERIZAÇÃO DO ESTADO OLIGOMÉRICO DAS PROTEÍNAS FabI RECOMBINANTES

As células transformadas e selecionadas por crescimento em meio ágar LB com canamicina a 50 µg/mL (antimicrobiano de seleção) foram utilizadas para a expressão das proteínas recombinantes saFabI e ecFabI. A expressão de ambas as proteínas foi realizada com base no protocolo descrito por Fage et al. (2020), com crescimento a 37 °C e 200 rpm até atingir densidade óptica a 600 nm (DO₆₀₀) entre 0,60 e 0,80, quando a temperatura foi reduzida a 18 °C e adicionado IPTG na concentração final de 0,5 mM para a indução por 18 horas sob rotação de 200 rpm.

Inicialmente, um teste de indução foi feito com um volume menor de meio de cultura, para verificar se a proteína era adequadamente expressa, e a lise foi conduzida em microtubo com volume de 1,0 mL de meio de cultura. Amostras do rompimento celular foram analisadas por SDS-PAGE e o gel foi corado em Coomassie Blue R250, entretanto, o gel da FabI de *S. aureus* precisou ser descorado e, posteriormente, corado novamente com prata. As fotografias dos géis estão dispostas na **Figura 36**, para a saFabI, e na **Figura 37**, para a ecFabI.

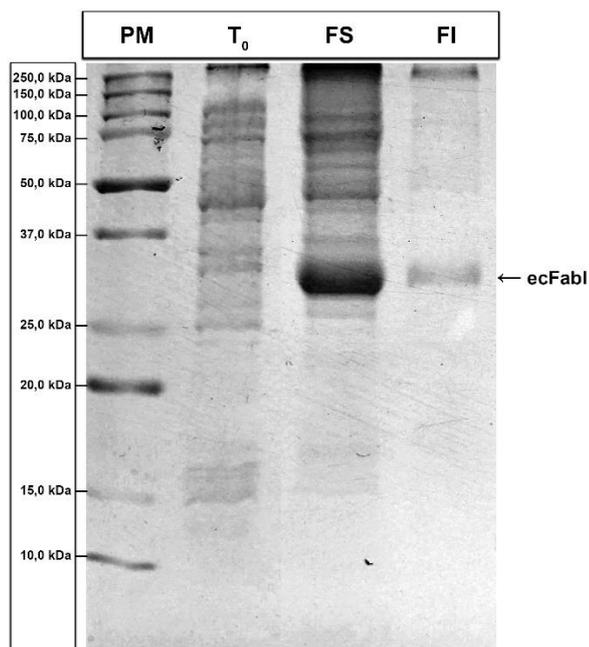
Figura 36 – Análise do gel SDS-PAGE 12% para o teste de expressão da saFabI após coloração com prata.



Legenda: PM: padrão de peso molecular, T₀: fração não induzida, FS: fração solúvel após indução, FI: fração insolúvel após indução.

Fonte: Autoria própria.

Figura 37 – Análise do gel SDS-PAGE 12% para o teste de expressão da ecFabI após coloração com Coomassie Blue R250.



Legenda: PM: padrão de peso molecular, T₀: fração não induzida, FS: fração solúvel após indução, FI: fração insolúvel após indução.

Fonte: Autoria própria.

A análise de ambos os géis demonstrou a superexpressão de proteínas com massa molecular de aproximadamente 30 kDa, possivelmente, correspondendo às proteínas recombinantes saFabI e ecFabI, que apresentam massas moleculares de 30,3 e 29,8 kDa. Tanto a banda da saFabI quanto da ecFabI, apresentaram boa resolução se distinguindo das demais bandas. Também foi possível verificar que a lise em sonicador permitiu a obtenção da proteína em fração solúvel. Ainda que se tenha observado a presença das proteínas em suas frações insolúveis, a maior quantidade delas estava na fração solúvel, pois o *pellet* da lise foi ressuscitado em tampão de solubilização com volume correspondente ao volume de amostra da fração solúvel e foram aplicados no gel volumes iguais de cada amostra.

Com o sucesso no teste de expressão da FabI procederam-se aos testes de purificação com um volume maior de cultura expressa (1 L) para tentar obter as proteínas recombinantes puras. Os testes de purificação foram conduzidos com a saFabI e a melhor condição foi reproduzida para a ecFabI, com o objetivo de se padronizar um método de purificação único para as duas enzimas. A **Tabela 24**, mostra os tampões utilizados em cada uma das tentativas e os resultados observados.

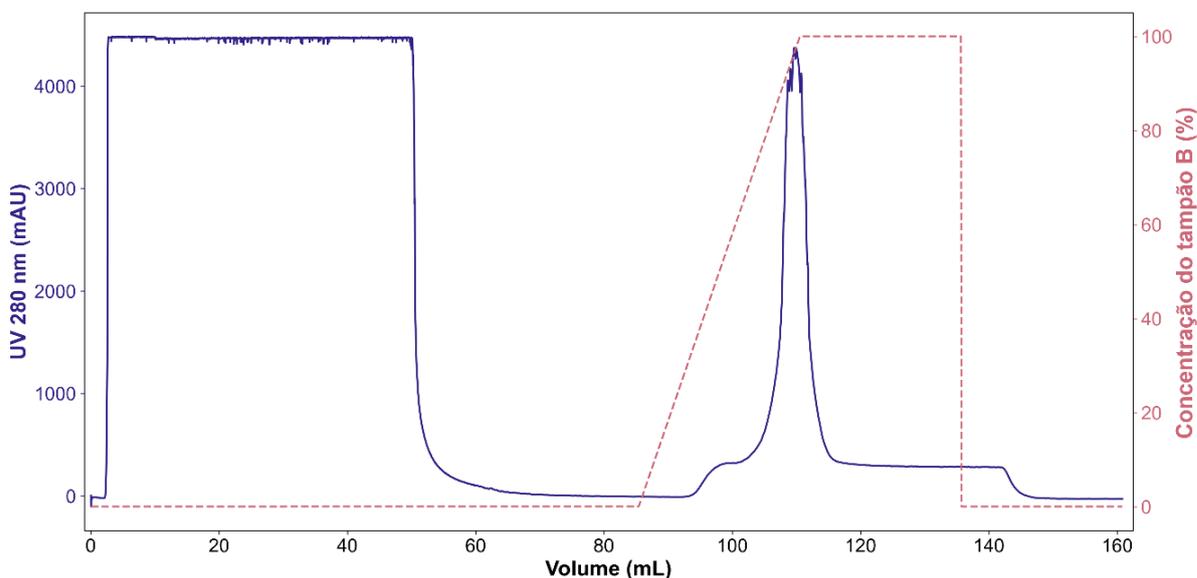
Tabela 24 – Resumo dos testes de purificação realizados com a saFabI.

Teste	Tampões de afinidade		Tampão de dessalinização	Tampão de exclusão molecular	Resultado
	Equilíbrio	Eluição			
1º	Tris-HCl 50 mM, NaCl 500 mM e imidazol 30 mM, pH = 7,4	Tris-HCl 50 mM, NaCl 500 mM e imidazol 500 mM, pH = 7,4	Tris-HCl 50 mM e NaCl 50 mM, pH = 7,4	-	Precipitação completa das amostras.
2º	Tris-HCl 50 mM, NaCl 500 mM, imidazol 30 mM e glicerol 10%, pH = 8,0	Tris-HCl 50 mM, NaCl 500 mM e glicerol 10%, pH = 8,0	Tris-HCl 50 mM, NaCl 200 mM e glicerol 10%, pH = 8,0	Tris-HCl 50 mM e NaCl 200 mM, pH = 8,0	Formação de agregados.
3º	Tris-HCl 50 mM, NaCl 500 mM, imidazol 30 mM e glicerol 10%, pH = 8,0	Tris-HCl 50 mM, NaCl 500 mM e glicerol 10%, pH = 8,0	Tampão fosfato de potássio 20 mM, NaCl 200 mM, glicerol 10% e DTT 2 mM, pH = 8,0	Tampão fosfato de potássio 20 mM, NaCl 200 mM e DTT 2 mM, pH = 8,0	Obtenção da proteína pura e sem agregados.

Fonte: Autoria própria.

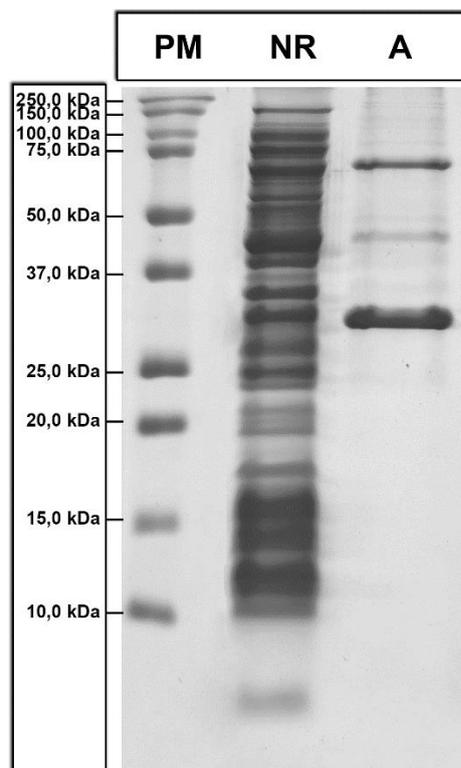
Em relação aos primeiros testes realizados para a saFabI, pode-se concluir que a proteína precipita facilmente em tampões com baixas concentrações de sal, como visto no primeiro teste. Além disso, o glicerol a 10% auxilia na estabilização da proteína, também evitando a sua precipitação. A **Figura 38** mostra o cromatograma de afinidade do segundo teste, neste é possível observar que houve a ligação da proteína na resina da coluna durante o equilíbrio em baixa concentração de imidazol e sua eluição em gradiente com concentração de imidazol elevada. Além disso, observou-se também que o imidazol utilizado apresentava alguma impureza que gerou leitura de absorbância a 280 nm. O gel de SDS-PAGE da fração não-retida e do pico da afinidade para o segundo teste estão representados na **Figura 39**, na qual é possível observar a banda da saFabI entre 25 e 37 kDa. No gel, é possível observar que houve a eluição e separação da proteína saFabI, entretanto, uma fração da proteína ficou na fração não-retida, eluída durante a fase de equilíbrio em baixa concentração de imidazol.

Figura 38 – Cromatograma de afinidade do segundo teste de purificação com a saFabI.



Fonte: Autoria própria.

Figura 39 – Análise do gel SDS-PAGE 12% para a cromatografia de afinidade da saFabI após coloração com Coomassie Blue R250.

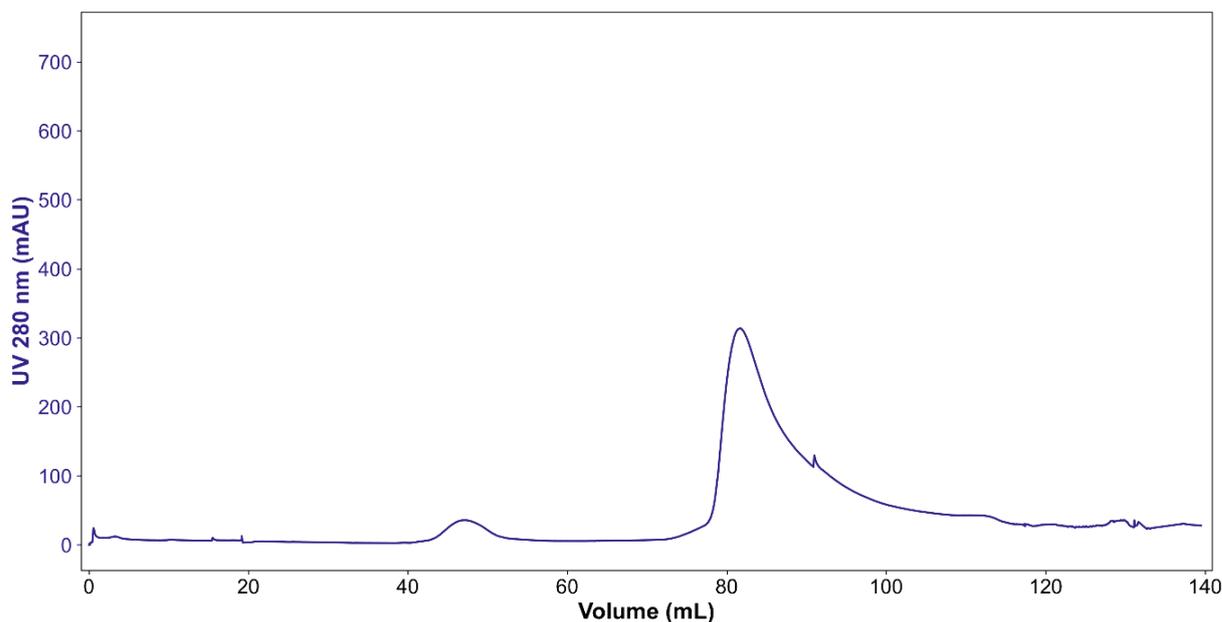


Legenda: PM: padrão de peso molecular, NR: fração não-retida e eluída durante a fase de injeção da amostra e de lavagem da coluna da afinidade, A: fração eluída durante a fase de eluição da afinidade.

Fonte: Autoria própria.

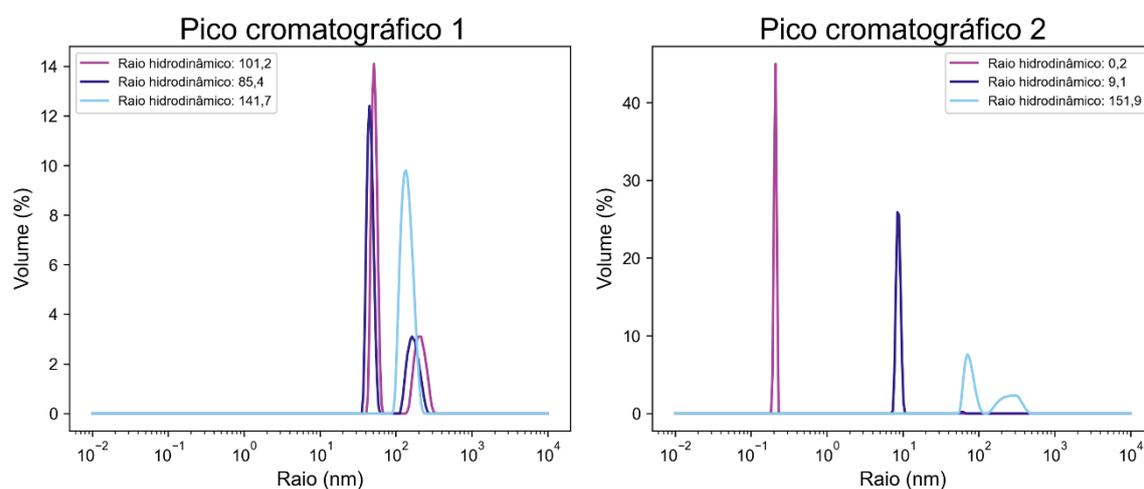
Nesse segundo teste, mesmo com a adição de glicerol, ajuste de pH para 8,0 e concentração salina mais elevada, a proteína formou agregados rapidamente. A presença desses agregados foi observada na exclusão molecular e confirmada por DLS. A **Figura 40** mostra o cromatograma de exclusão molecular resultante do segundo teste, no qual é possível observar a presença de dois picos. Nesses picos, era esperado que, no primeiro, próximo ao volume morto da coluna (47,96 mL), fosse observado um agregado de proteínas e, no segundo, próximo a 80 mL, a proteína saFabI na forma de tetrâmero. Nos resultados de DLS desses dois picos (**Figura 41**) é possível observar que, para o primeiro pico, de fato temos agregados proteicos com raios hidrodinâmicos variando na triplicata entre 85,4 e 141,7 nm. Entretanto, no segundo pico, foram observados raios hidrodinâmicos de 151,9 nm na amostra, indicando que existem agregados proteicos mesmo nesse pico.

Figura 40 – Cromatograma de exclusão molecular do segundo teste de purificação com a saFabl.



Fonte: Autoria própria.

Figura 41 – Distribuição de tamanho por massa obtido por DLS para o segundo teste de purificação com a saFabl.

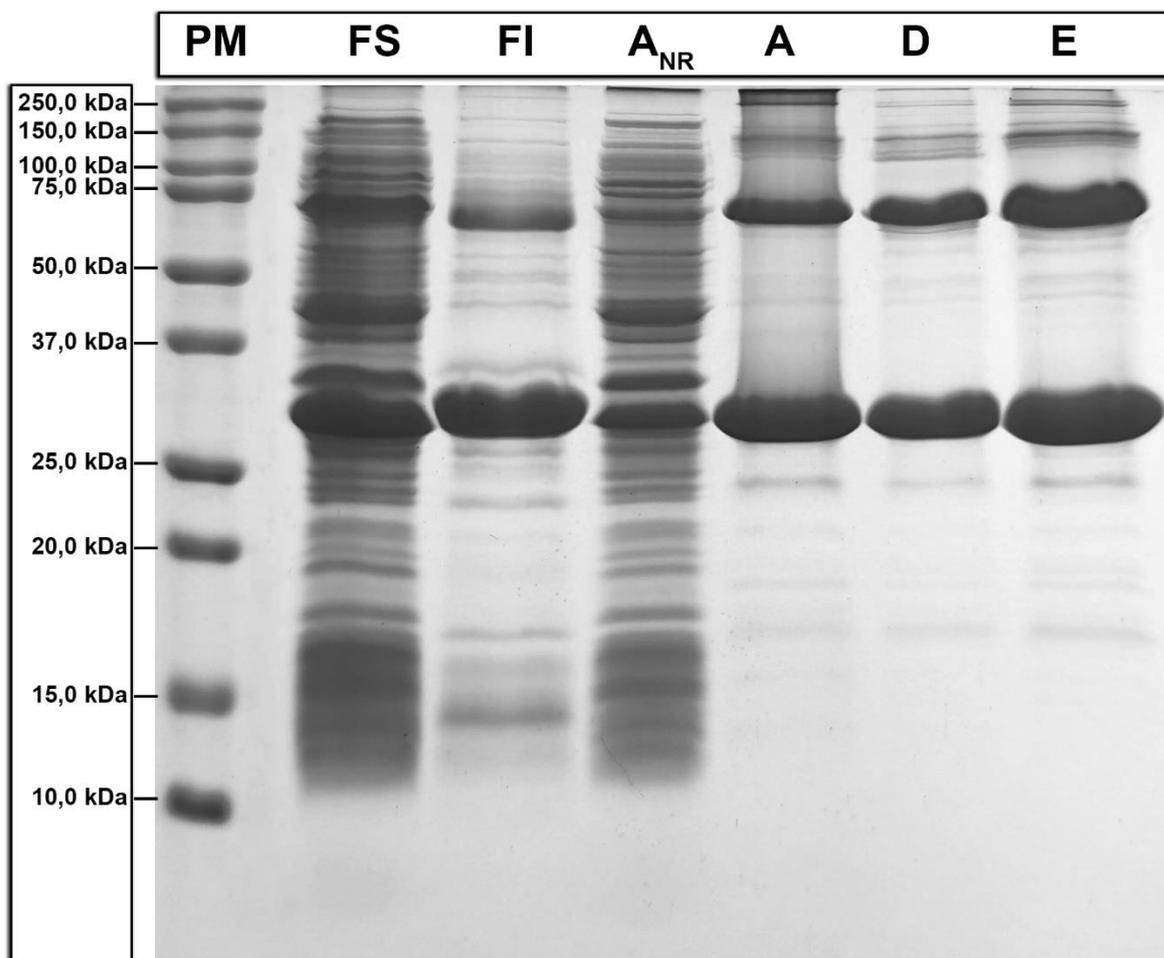


Fonte: Autoria própria.

Os resultados do segundo teste indicam que, no tampão escolhido, a proteína ainda era capaz de formar agregados com o tempo, por isso, foi adicionado o DTT em concentração final de 2 mM nos tampões de dessalinização e exclusão molecular. Além disso, os dois primeiros testes foram realizados em tampão tris-HCl porque o objetivo inicial era de realizar a cristalografia das proteínas com os ligantes da triagem

virtual. Entretanto, optou-se por realizar o STD-NMR das amostras de proteína com o ligante, uma vez que a própria UFMG dispõe de espectrômetro de RMN-600 MHz. Para isso, foi necessário que o tampão tris-HCl fosse substituído por tampão fosfato, uma vez que os hidrogênios do tris interferem na leitura no RMN. Com essas modificações foi possível obter as proteínas saFabI e ecFabI estáveis. As **Figuras 42 e 43** representam, respectivamente, os géis de SDS-PAGE da saFabI e da ecFabI, com as amostras das etapas de lise bacteriana, cromatografia de afinidade, dessalinização e exclusão molecular.

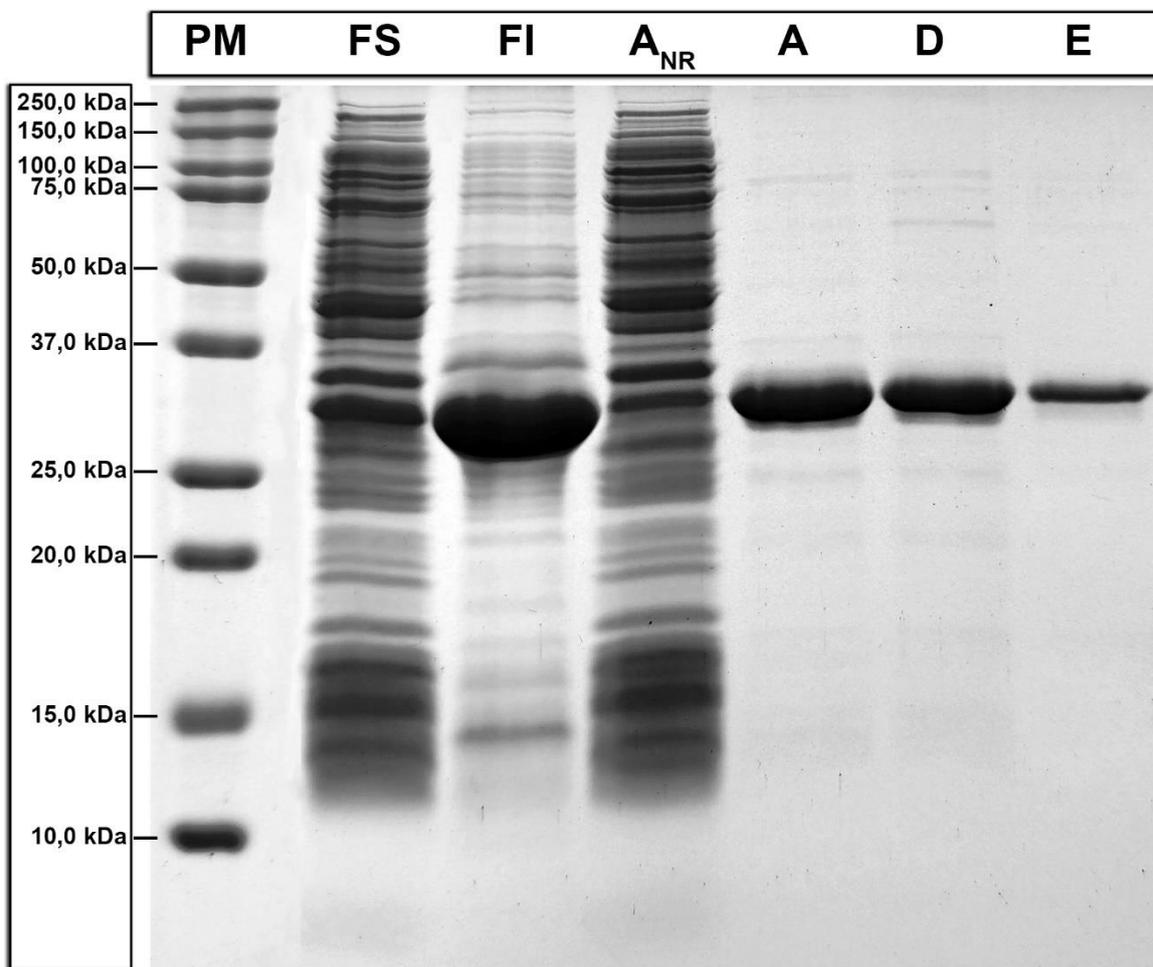
Figura 42 – Análise do gel SDS-PAGE 12% para as etapas de lise bacteriana e cromatografias de afinidade, dessalinização e exclusão molecular da saFabI após coloração com Coomassie Blue R250.



Legenda: PM: padrão de peso molecular, FS: fração solúvel após indução, FI: fração insolúvel após indução, A_{NR}: fração não-retida e eluída durante a fase de injeção da amostra e de lavagem da coluna da afinidade, A: fração eluída durante a fase de eluição da afinidade, D: amostra após dessalinização, E: amostra do pico mais proeminente na cromatografia de exclusão molecular.

Fonte: Autoria própria.

Figura 43 – Análise do gel SDS-PAGE 12% para as etapas de lise bacteriana e cromatografias de afinidade, dessalinização e exclusão molecular da ecFabI após coloração com Coomassie Blue R250.



Legenda: PM: padrão de peso molecular, FS: fração solúvel após indução, FI: fração insolúvel após indução, ANR: fração não-retida e eluída durante a fase de injeção da amostra e de lavagem da coluna da afinidade, A: fração eluída durante a fase de eluição da afinidade, D: amostra após dessalinização, E: amostra do pico mais proeminente na cromatografia de exclusão molecular.

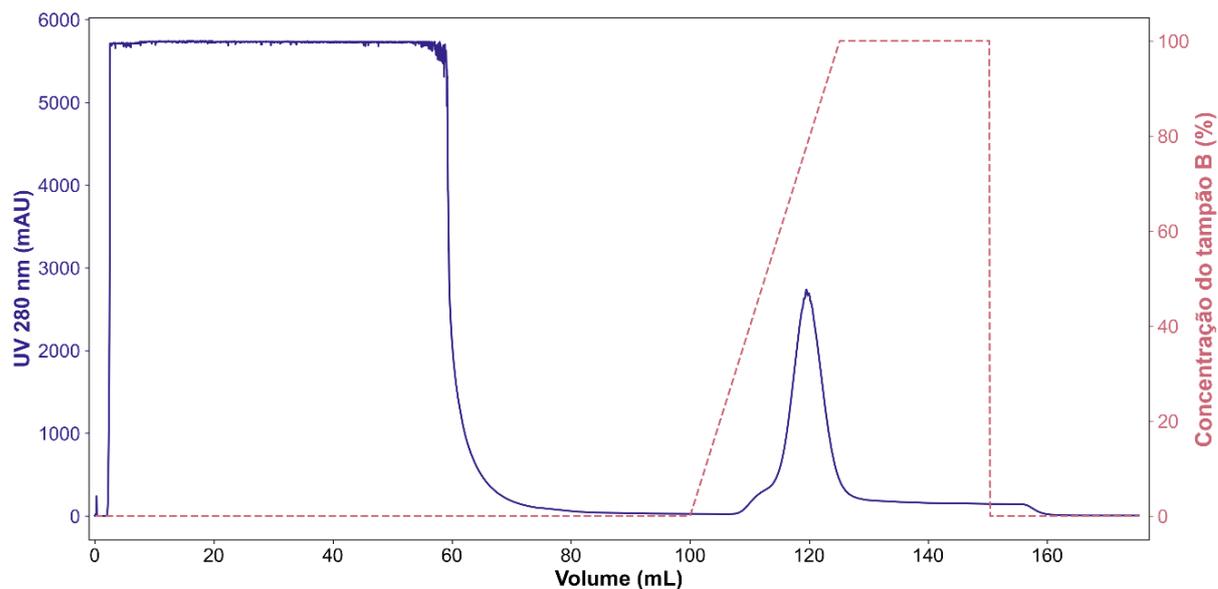
Fonte: Autoria própria.

Nos géis de SDS-PAGE foi possível observar novamente bandas de superexpressão próximas a 30 kDa, correspondendo, possivelmente, às proteínas FabI recombinantes. Para ambos os casos, um volume maior de células em tampão de lise representou uma queda na eficiência de lise pelo sonicador, com uma diminuição significativa da proteína na fração solúvel, indicando que o volume para lise em sonicador precisa ser menor e as amostras precisam ser particionadas para lise. Também foi possível observar que a cromatografia de afinidade foi responsável

pela remoção praticamente completa dos contaminantes da amostra. Os resultados das cromatografias de cada proteína serão discutidos individualmente a seguir.

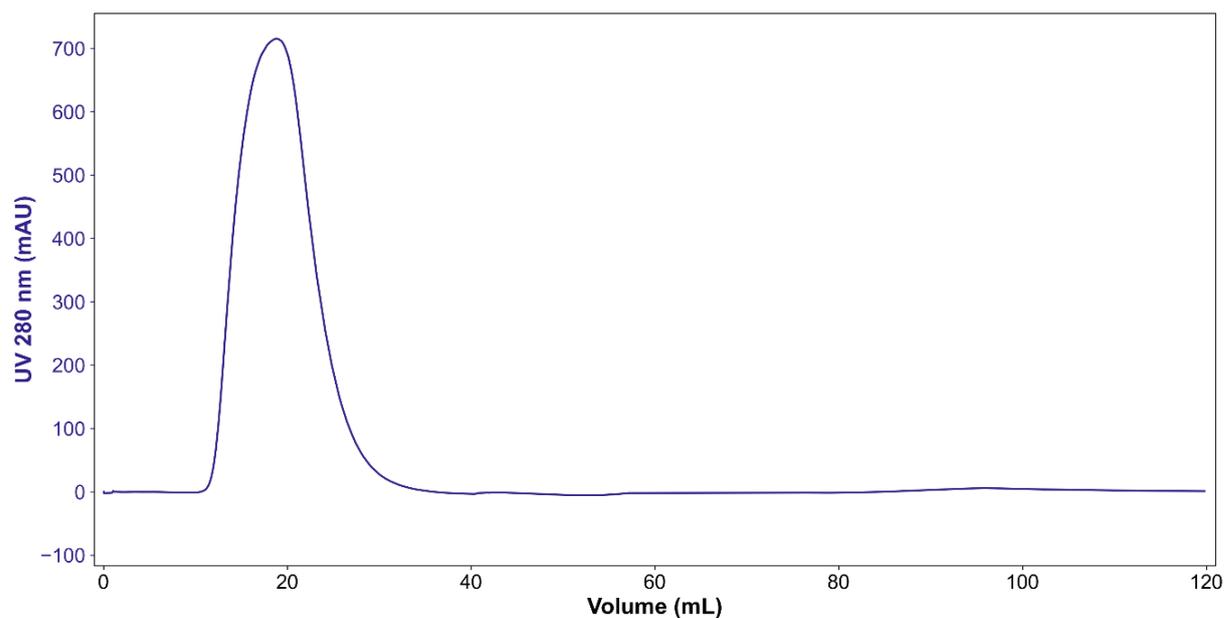
As **Figuras 44, 45 e 46**, representam, respectivamente, os cromatogramas de afinidade, dessalinização e exclusão molecular obtidos para a saFabI.

Figura 44 – Cromatograma de afinidade da purificação final com a saFabI.

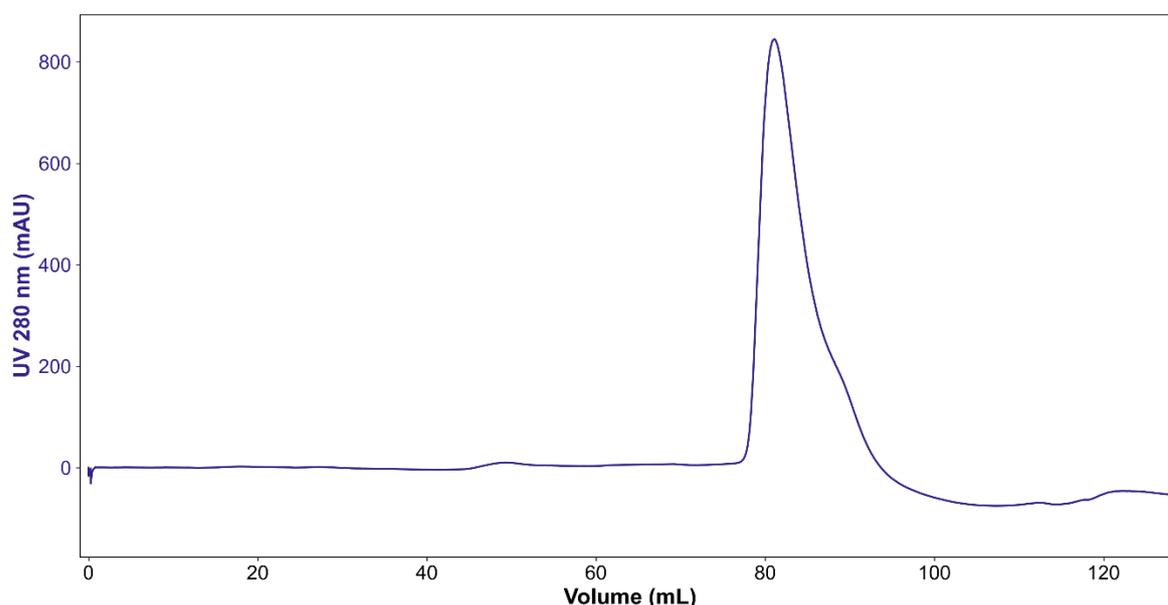


Fonte: Autoria própria.

Figura 45 – Cromatograma de dessalinização da purificação final com a saFabI.



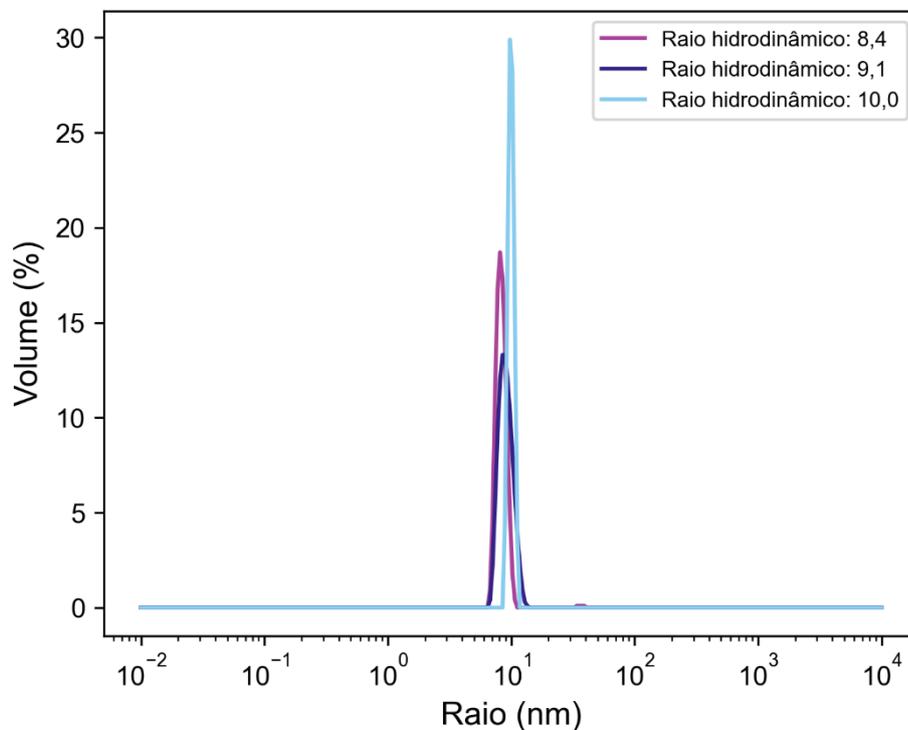
Fonte: Autoria própria.

Figura 46 – Cromatograma de exclusão molecular da purificação final com a saFabI.

Fonte: Autoria própria.

Em relação aos resultados das cromatografias para a saFabI, é possível observar novamente que o imidazol utilizado na cromatografia de afinidade apresenta absorvância a 280 nm e que parte da proteína pode ter sido perdida na fração não-retida. Em relação a algumas bandas proteicas visualizadas no SDS-PAGE (**Figura 42**) que não estão entre 25 e 37 kDa, o artigo de Fage *et al.* (2020), que utiliza o mesmo plasmídeo da saFabI que o presente trabalho, indica que são bandas da própria saFabI confirmadas por espectrometria de massas. De acordo com os autores deste artigo, essas bandas correspondem a forma de tetrâmero (entre 100 e 150 kDa) e dímero (entre 50 e 75 kDa) da proteína que são observáveis mesmo em condições desnaturantes, além do monômero já esperado entre 25 e 37 kDa. A exclusão molecular da FabI apresentou apenas um pico desprezível na região do volume morto de coluna, indicando um provável sucesso na eliminação de agregados de proteína. Entretanto, é importante notar que a ausência de agregados deve ser verificada no pico próximo a 80 mL, pois o pico próximo ao volume morto de coluna pode ocorrer mesmo na presença de DTT. Para confirmar a ausência de agregados no pico observado próximo a 80 mL e confirmar a contribuição do DTT para evitar a formação de agregados, foram realizados os ensaios de DLS. Os resultados do DLS estão representados na **Figura 47** e na **Tabela 25**.

Figura 47 – Distribuição de tamanho por volume obtida por DLS da amostra de saFabI após exclusão molecular.



Fonte: Autoria própria.

Tabela 25 – Resultados do DLS para o pico principal de cada triplicata da saFabI.

Experimento	Massa molecular		Índice de	
	estimada (kDA \pm DP)	Massa (%)	polidispersividade (%)	Polidispersão
1	99,6 \pm 8,7	99,5	8,8	Monodisperso
2	123,7 \pm 16,8	99,7	12,9	Monodisperso
3	153,5 \pm 6,5	99,9	4,4	Monodisperso
Média	125,6 \pm 27,0	99,7 \pm 0,2	8,7 \pm 4,2	Monodisperso

Fonte: Autoria própria.

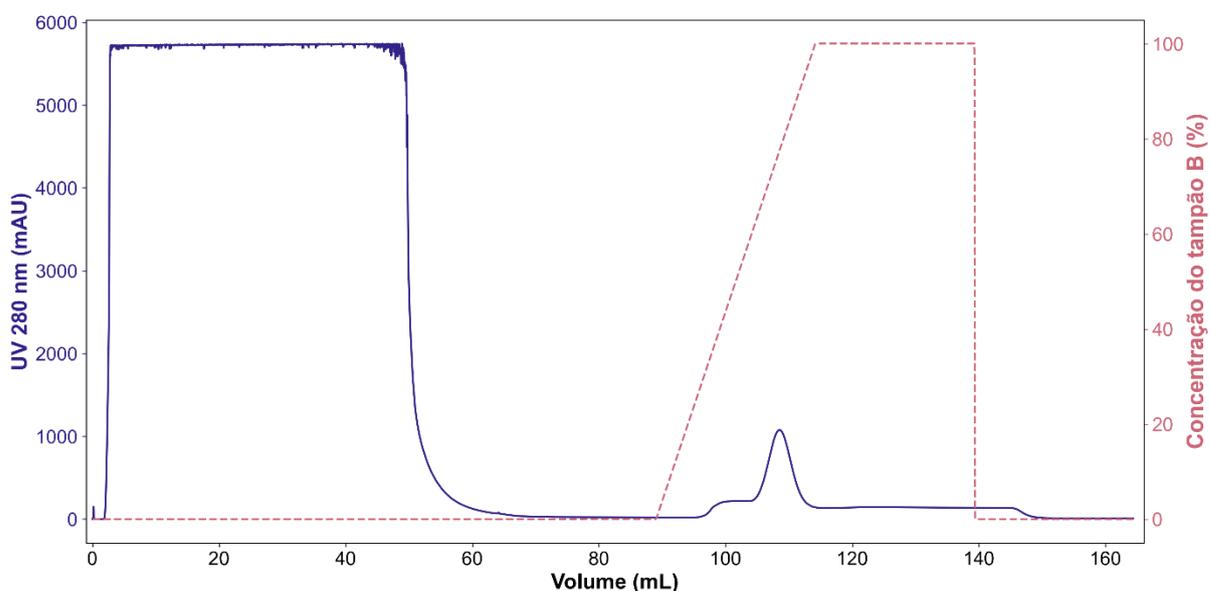
Os resultados de DLS indicaram $99,7 \pm 0,2\%$ da massa de proteína correspondendo a uma proteína com raio hidrodinâmico médio de $9,2 \pm 0,8$ e massa molecular de $125,6 \pm 27,0$. Indicando que o estado oligomérico da saFabI em solução é o tetrâmero e que essa forma é a predominante na amostra, em termos de massa. Além disso, a proteína apresentou índice de polidispersividade de $8,7 \pm 4,2\%$,

indicando que se trata de uma amostra monodispersa adequada para ensaios estruturais.

Por fim, realizando os cálculos de rendimento, foi possível obter 17,29 mg de saFabI por litro de cultura na forma de tetrâmero com $99,7 \pm 0,2\%$ de massa.

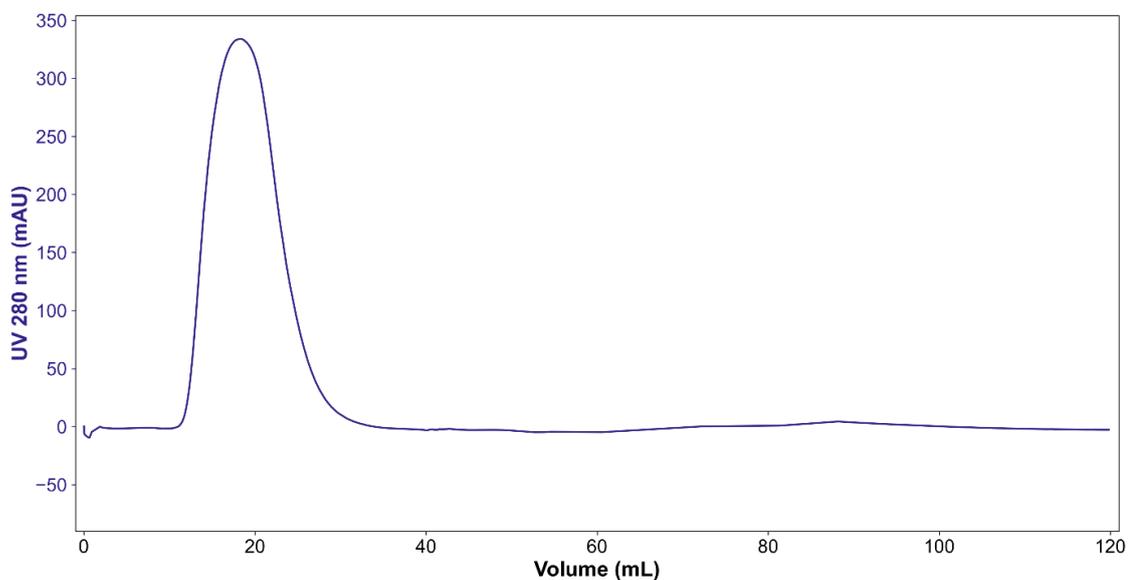
As **Figuras 48, 49 e 50** representam, respectivamente, os cromatogramas de afinidade, dessalinização e exclusão molecular obtidos para a ecFabI.

Figura 48 – Cromatograma de afinidade da purificação final com a ecFabI.



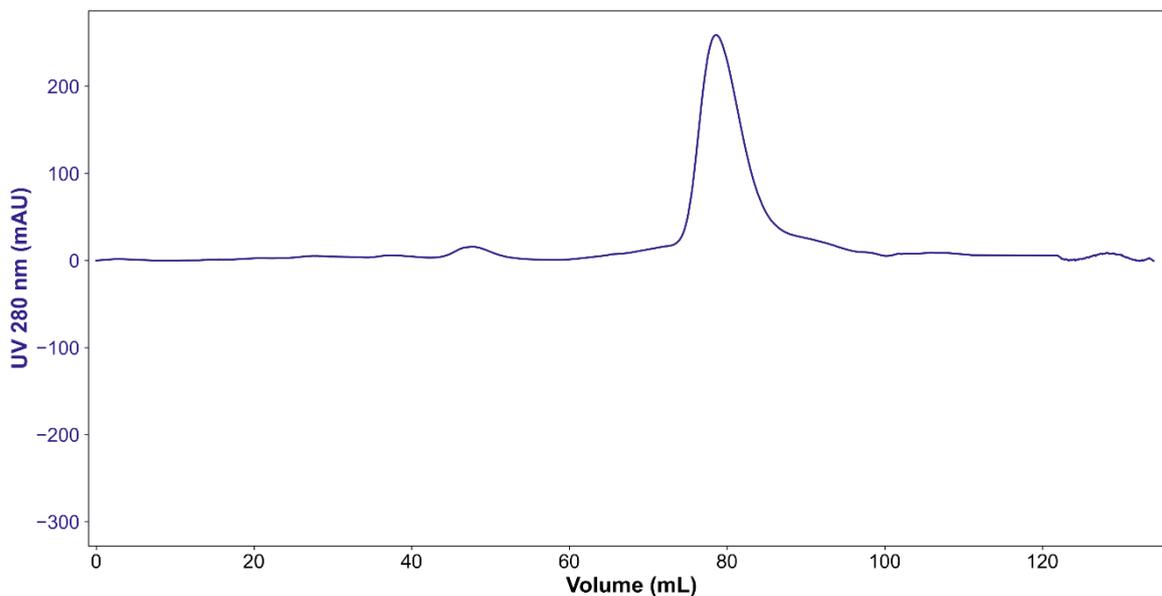
Fonte: Autoria própria.

Figura 49 – Cromatograma de dessalinização da purificação final com a ecFabI.



Fonte: Autoria própria.

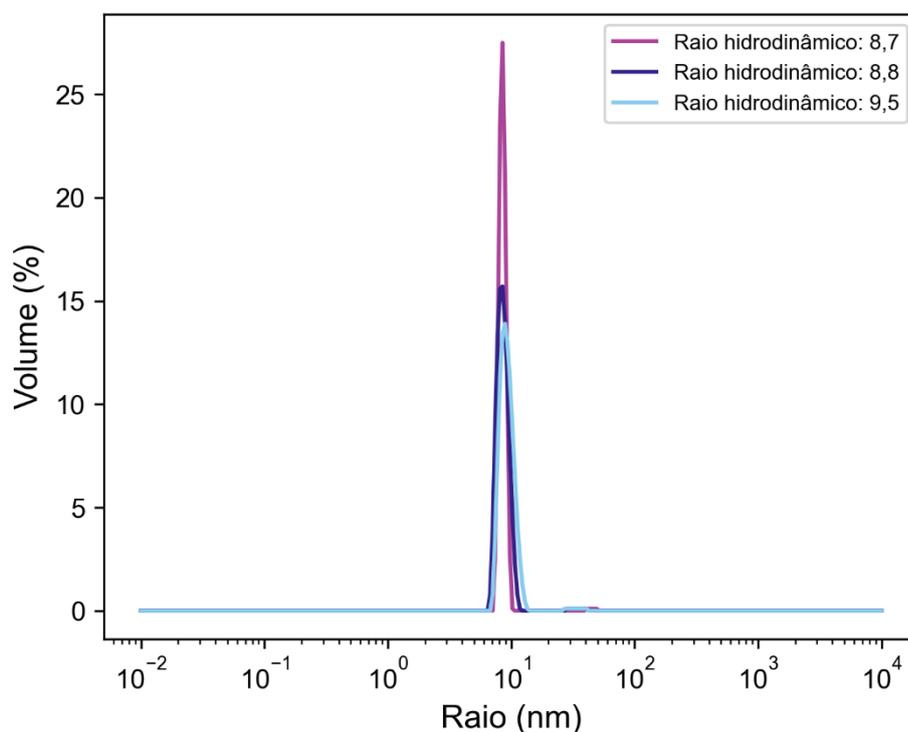
Figura 50 – Cromatograma de exclusão molecular da purificação final com a ecFabI.



Fonte: Autoria própria.

Em relação aos resultados das cromatografias para a ecFabI, é possível observar comportamento similar ao observado na saFabI. Em relação as bandas proteicas visualizadas no SDS-PAGE (**Figura 43**) as diferenças são mais evidentes, uma vez que, neste caso, a condição desnaturante parece ter sido suficiente para separar as cadeias da ecFabI. Na exclusão molecular dessa proteína foi possível observar um pico desprezível na região do volume morto de coluna, indicando um provável sucesso na eliminação de agregados de proteína. Para confirmar a ausência de agregados no pico observado próximo a 80 mL (pico proeminente) também foram realizados os ensaios de DLS. Os resultados do DLS estão representados na **Figura 51** e na **Tabela 26**.

Figura 51 – Distribuição de tamanho por volume obtida por DLS da amostra de ecFabI após exclusão molecular.



Fonte: Autoria própria.

Tabela 26 – Resultados do DLS para o pico principal de cada triplicata da ecFabI.

Experimento	Massa molecular		Índice de	
	estimada (kDA \pm DP)	Massa (%)	polidispersividade (%)	Polidispersão
1	99,6 \pm 4,0	99,4	5,4	Monodisperso
2	111,0 \pm 27,4	98,7	10,5	Monodisperso
3	123,7 \pm 56,7	98,7	12,5	Monodisperso
Média	111,4 \pm 12,0	98,9 \pm 0,4	9,5 \pm 3,7	Monodisperso

Fonte: Autoria própria.

Os resultados de DLS indicaram $98,9 \pm 0,4\%$ da massa de proteína correspondendo a uma proteína com raio hidrodinâmico médio de $9,0 \pm 0,4$ e massa molecular de $111,4 \pm 12,0$. Indicando que o estado oligomérico da ecFabI em solução é o tetrâmero e que essa forma é a predominante na amostra, em termos de massa, por $98,9\%$. Além disso, a proteína apresentou índice de polidispersividade de $9,5 \pm 3,7\%$, indicando que se trata de uma amostra monodispersa adequada para ensaios estruturais.

Por fim, realizando os cálculos de rendimento, foi possível obter **3,97 mg de ecFabI por litro de cultura** na forma de tetrâmero com $98,9 \pm 0,4\%$ de massa.

As proteínas recombinantes obtidas (saFabI e ecFabI) foram separadas para serem utilizadas nos ensaios de STD-NMR juntamente com os ligantes que tiveram atividade biológica definida nos ensaios de concentração inibitória mínima.

5.7 ENSAIOS DE CONCENTRAÇÃO INIBITÓRIA MÍNIMA COM OS LIGANTES PROMISSORES DA TRIAGEM VIRTUAL

As moléculas consideradas promissoras na triagem virtual foram avaliadas frente a cepas de bactérias *Staphylococcus aureus* (ATCC 29123) e *Escherichia coli* (ATCC 35218) pelo ensaio de microdiluição em caldo realizado em microplacas de 96 poços. Uma triagem *in vitro* inicial a 100 μM foi realizada e, com as substâncias com atividade antibacteriana, a MIC foi determinada em concentrações entre 100 e 0,78 μM . Todas as substâncias da triagem virtual (**Tabelas 22 e 23**), com exceção da substância indisponível MSL2016_13, como comentado anteriormente (vide seção **5.5 Avaliação do domínio de aplicabilidade e triagem virtual de potenciais inibidores da FabI, p. 138**), foram testadas frente as duas espécies de bactéria. As substâncias ativas na triagem prévia têm seus resultados de MIC representados na **Tabela 27**.

Tabela 27 – MIC (μM) das moléculas contra *S. aureus* (ATCC 29213) e *E. coli* (ATCC 35218).

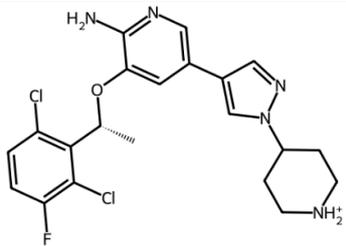
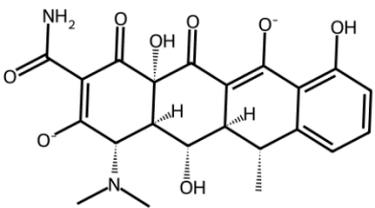
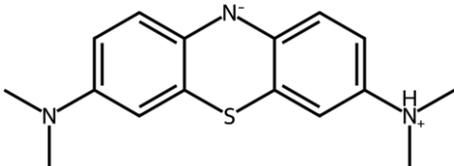
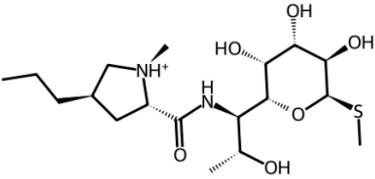
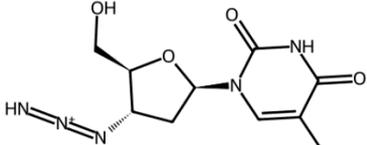
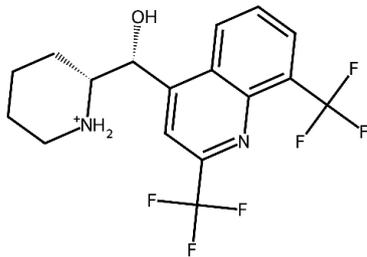
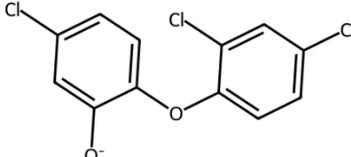
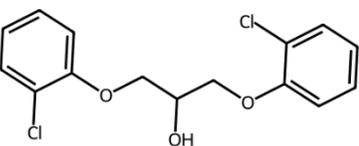
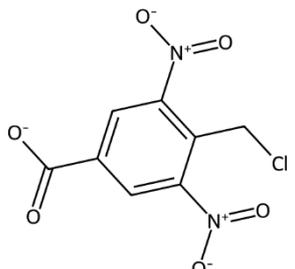
Moléculas	Estruturas	(continua)	
		MIC <i>S. aureus</i> (μM)	MIC <i>E. coli</i> (μM)
ZINC000035902489 (Crizotinibe)		50	NA
ZINC000016052277 (Doxiciclina)		< 0,78	3,125

Tabela 27 – MIC (μM) das moléculas contra *S. aureus* (ATCC 29213) e *E. coli* (ATCC 35218).

Moléculas	Estruturas	(conclusão)	
		MIC <i>S. aureus</i> (μM)	MIC <i>E. coli</i> (μM)
ZINC000012414057 (Azul de metileno)		50	NA
ZINC000003982483 (Lincomicina)		1,56	NA
ZINC000003779042 (Zidovudina)		NA	25
ZINC000000537964 (Mefloquina)		50	50
ZINC000000002216 (Triclosan)		< 0,78	< 0,78
SNL2016_38		100	NA
EDNCI		50	NA

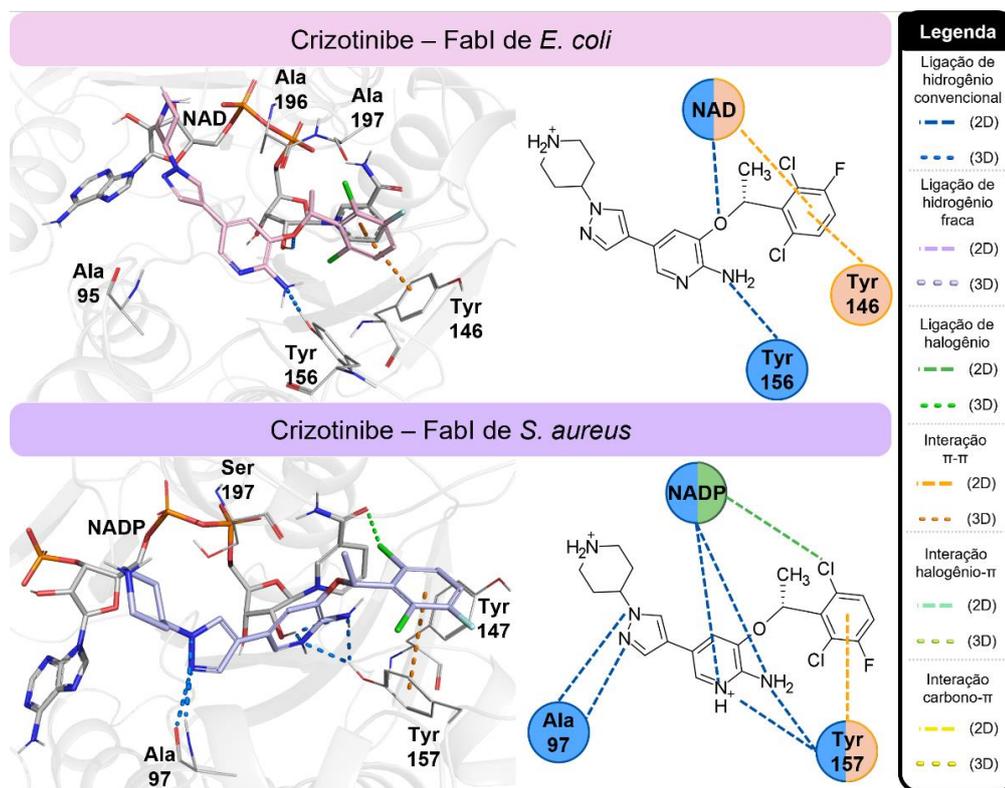
Legenda: NA: Não ativa.
Fonte: Autoria própria.

Em relação as substâncias com MIC < 100 µM, é possível observar que a maioria das substâncias foram mais ativas ou exclusivamente ativas em *S. aureus*, uma Gram-positiva. Esse fato pode se justificar pela diferença da composição da parede bacteriana, uma vez que a *E. coli*, uma Gram-negativa, com sua membrana externa pode dificultar a penetração de substâncias e justificar a retenção de moléculas no meio extracelular (Asse Junior *et al.*, 2019).

Ainda sobre os resultados de MIC, com exceção do EDNCI, as substâncias encontradas já foram desenvolvidas como antibacterianos ou possuem atividade antibacteriana conhecida. O triclosan já é um conhecido inibidor da FabI com atividade antibacteriana e foi testado para manter comparação com as demais substâncias (Timmins; Deretic, 2006). A doxiciclina e a lincomicina são, respectivamente, uma tetraciclina e uma lincosamida utilizadas para o tratamento de infecções bacterianas. O mecanismo de ação das tetraciclinas envolve a sua ligação à subunidade 30S do ribossomo bacteriano, enquanto, as lincosamidas se ligam à subunidade 50S, ambas as classes atuam inibindo a síntese proteica (Krishna; Staines, 2012). O azul de metileno é considerado seguro para uso *in vivo* e é utilizado para tratamento da metemoglobinemia, encefalopatia induzida por ifosfamida, vasoplegia acompanhada de choque séptico e de intoxicação por cianeto. Além disso, tem atividade antimalárica e antibacteriana conhecidas, mas seu mecanismo de ação está em debate, em geral, acredita-se que seu mecanismo esteja envolvido com suas propriedades redox, interferindo nas vias de transporte de elétrons em bactérias (Thesnaar *et al.*, 2021). A zidovudina (AZT), aprovada pelo FDA em 1986 para o tratamento de HIV, também tem atividade antibacteriana conhecida, até mesmo contra cepas ESBL (Antonello *et al.*, 2021). A mefloquina é utilizada como antimalárico e possui atividade antibacteriana conhecida (Capan *et al.*, 2010). O crizotinibe e a SNL2016_38 também já possuem atividade antibacteriana *in vitro* determinada após resultado de uma triagem virtual para a inibição da FabI de *S. aureus* (Asse Junior *et al.*, 2019).

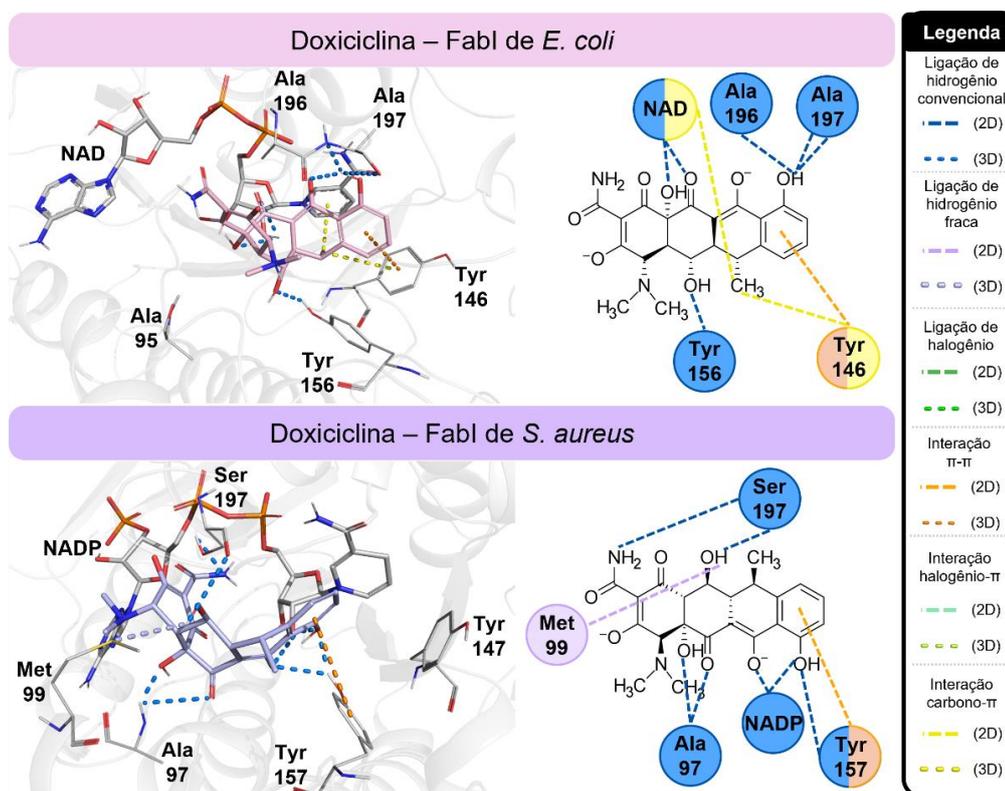
Para essas substâncias com MIC definida e potencial inibição da FabI, foram analisados os modos de ligação obtidos pelo *docking* molecular. As representações tridimensionais e bidimensionais estão apresentadas nas **Figuras 52–60**.

Figura 52 – Modos de ligação (3D à esquerda e 2D à direita) do crizotinibe para ecFabl e saFabl.



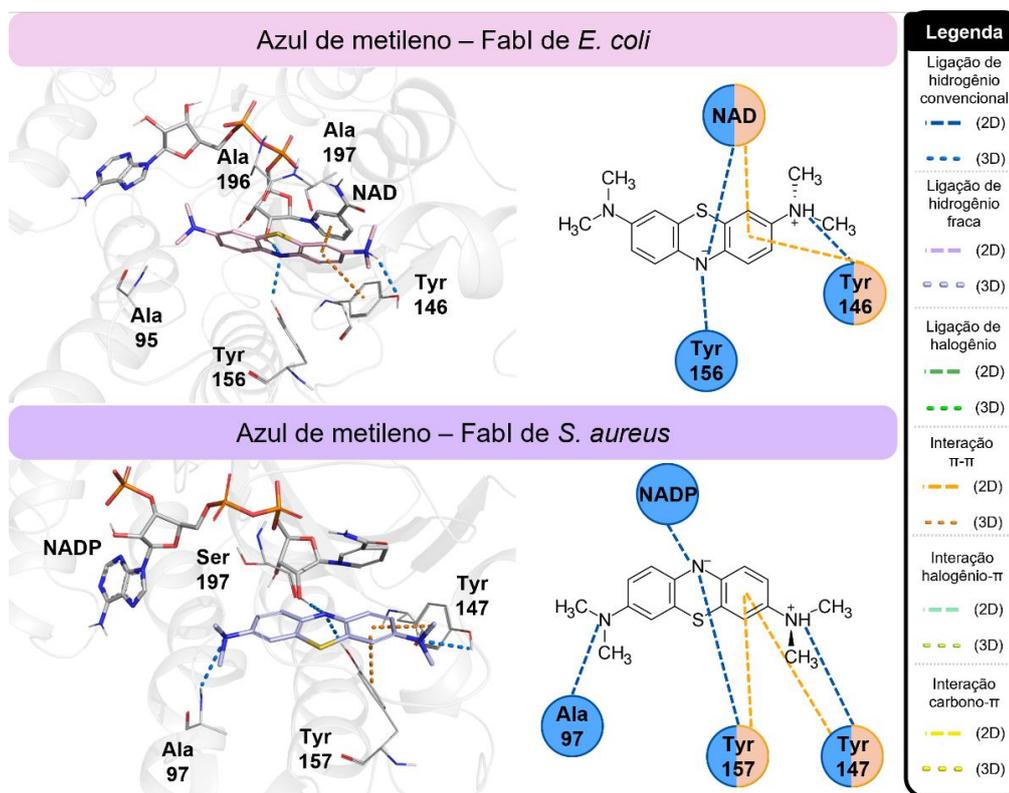
Fonte: Autoria própria.

Figura 53 – Modos de ligação (3D à esquerda e 2D à direita) da doxiciclina para ecFabl e saFabl.



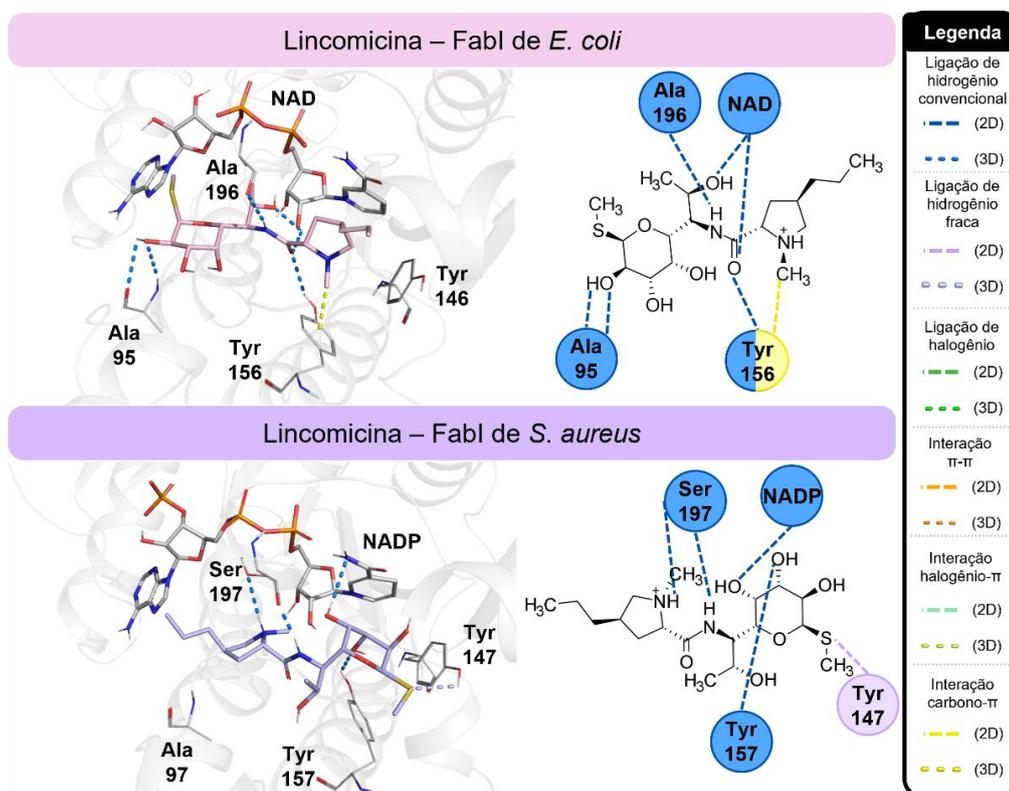
Fonte: Autoria própria.

Figura 54 – Modos de ligação (3D à esquerda e 2D à direita) do azul de metileno para ecFabI e saFabI.



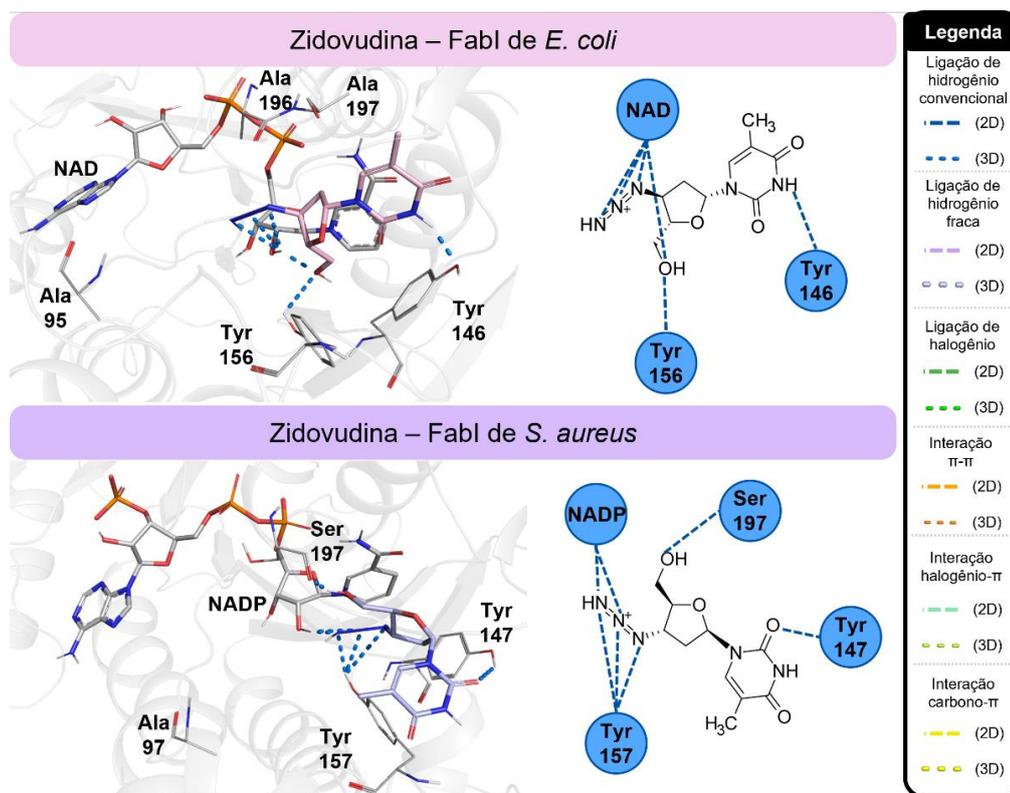
Fonte: Autoria própria.

Figura 55 – Modos de ligação (3D à esquerda e 2D à direita) da lincomicina para ecFabI e saFabI.



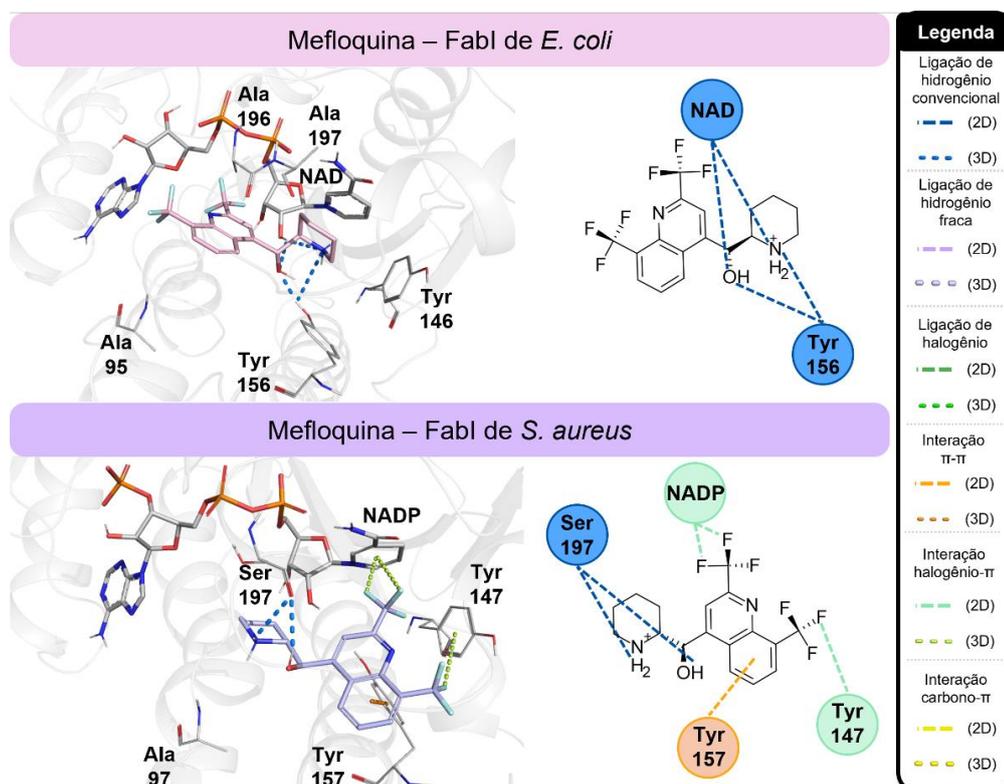
Fonte: Autoria própria.

Figura 56 – Modos de ligação (3D à esquerda e 2D à direita) da zidovudina para ecFabl e saFabl.



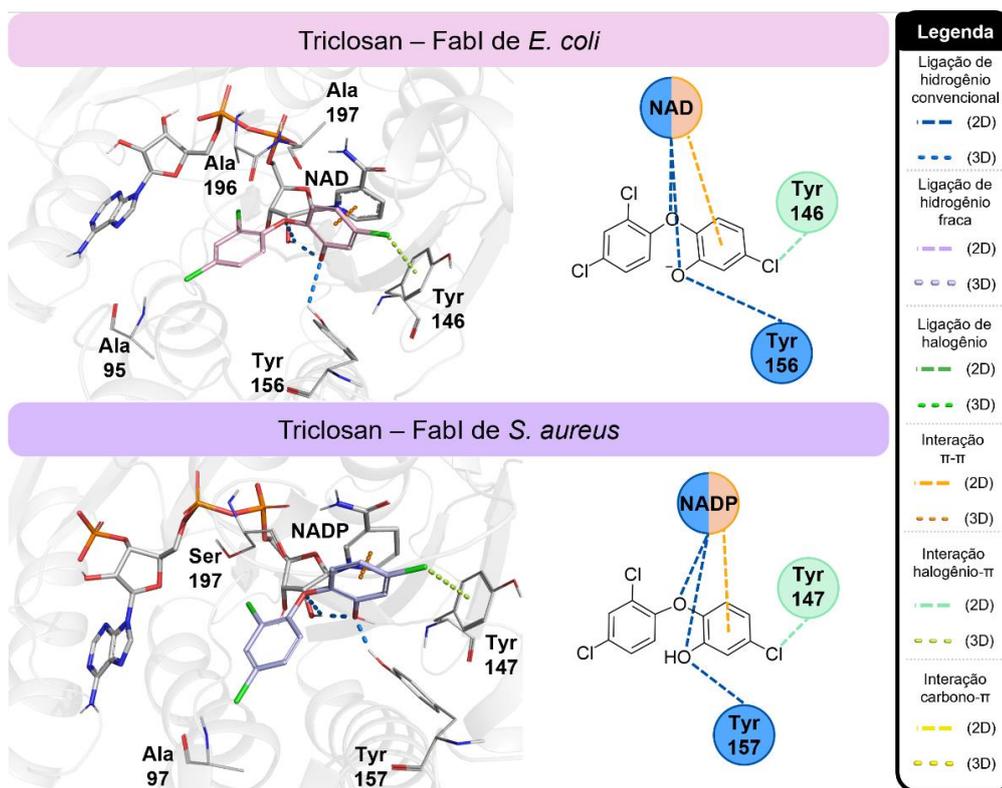
Fonte: Autoria própria.

Figura 57 – Modos de ligação (3D à esquerda e 2D à direita) da mefloquina para ecFabl e saFabl.



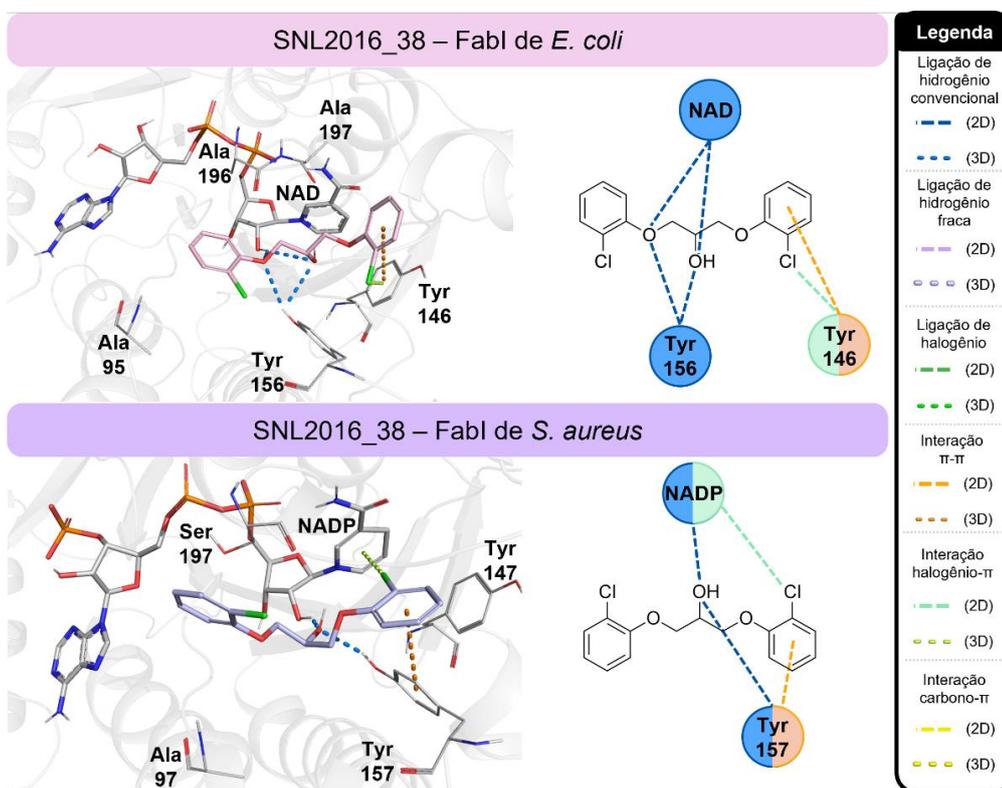
Fonte: Autoria própria.

Figura 58 – Modos de ligação (3D à esquerda e 2D à direita) do triclosan para ecFabl e saFabl.



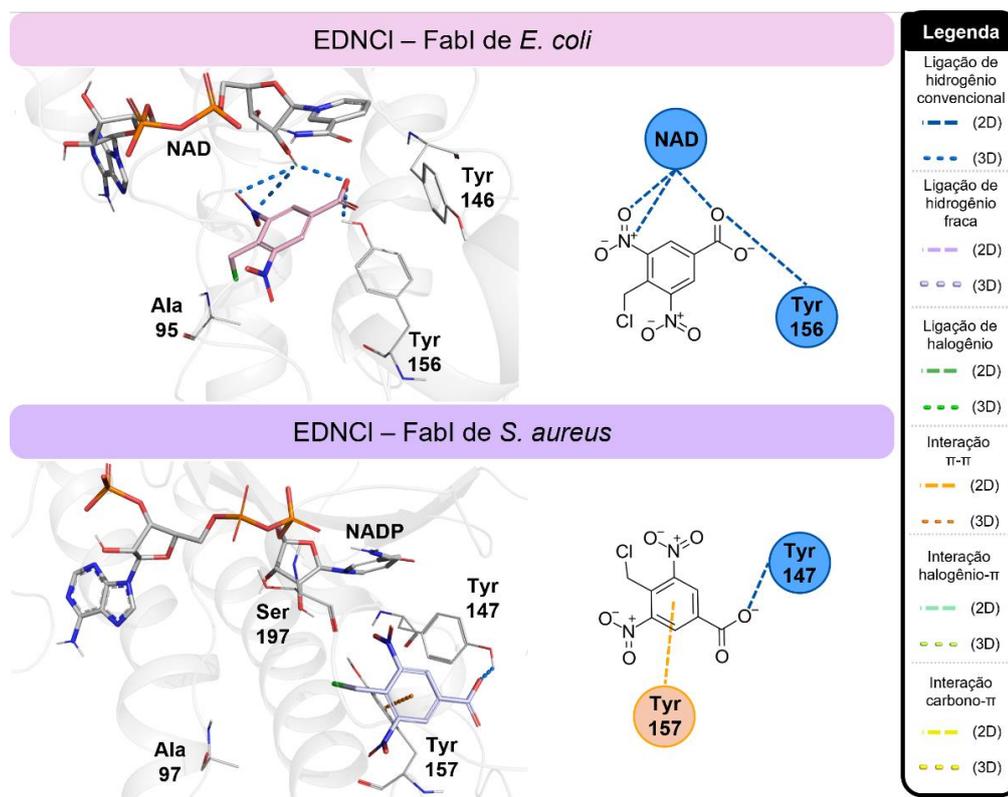
Fonte: Autoria própria.

Figura 59 – Modos de ligação (3D à esquerda e 2D à direita) de SNL2016_38 para ecFabl e saFabl.



Fonte: Autoria própria.

Figura 60 – Modos de ligação (3D à esquerda e 2D à direita) de EDNCI para ecFabl e saFabl.



Fonte: Autoria própria.

Em relação aos modos de ligação das moléculas com atividade antibacteriana determinada experimentalmente, que foram apresentados nas **Figuras 52–60**, é possível observar a importância e a conservação da interação dessas moléculas com os resíduos de Tyr-156 (ecFabl) ou Tyr-157 (saFabl), Tyr-146 (ecFabl) ou Tyr-147 (saFabl) e com a hidroxila do NADH (ecFabl) ou NADPH (saFabl). Esses resíduos são comumente considerados na literatura como os mais importantes para a atividade inibitória (Yang *et al.*, 2017; Maltarollo *et al.*, 2022). Além desses resíduos, também estiveram presentes os resíduos de Ala-95 (ecFabl), Ala-97 (saFabl), Ala-196 (ecFabl), Ala-197 (ecFabl), Ser-197 (saFabl) e Met-99 (saFabl). Os resultados encontrados corroboram também com a interpretação dos modelos de aprendizado de máquina, uma vez que entre os *bits* mais importantes para a classificação em ativo/inativo estavam interações ou colisões com alguns desses resíduos. É importante notar que, em todos os modos de ligação das diferentes substâncias, foi observada uma boa conservação na interação com os resíduos mais importantes para a atividade inibitória entre as duas enzimas (saFabl e ecFabl). Contudo, nas poses da mefloquina e da lincomicina há uma inversão no posicionamento da molécula entre

saFabI e ecFabI, dificultando a análise de qual pose é mais favorável e ressaltando a importância da realização de ensaios de cocristalização para elucidação das estruturas e conformações mais prováveis desses potenciais ligantes no sítio ativo de cada enzima.

Os resultados dos ensaios *in vitro* e as análises dos modos de ligação do *docking* molecular validam a capacidade dos modelos de aprendizado de máquina de encontrar antibacterianos. Além disso, até o presente momento, nenhum estudo na literatura relatou experimentos com as substâncias aqui apresentadas para determinação da inibição enzimática ou capacidade de interação com a FabI. Portanto, ainda não existem estudos que confirmem a sua inibição como potencial mecanismo de ação secundário ou principal na atividade antibacteriana dessas substâncias. Isso torna os resultados ainda mais promissores, uma vez que os dados *in vitro* de MIC indicam a atividade bacteriana e os resultados *in silico* de *docking* e de aprendizado de máquina indicam que há probabilidade da participação da inibição da FabI no mecanismo de ação antibacteriana dessas substâncias. Por isso, os ligantes com MIC definida e as proteínas recombinantes obtidas (saFabI e ecFabI) foram então separados para os ensaios de STD-NMR a serem realizados nos laboratórios de Macromoléculas (MacroMol) e de Ressonância Magnética de Alta Resolução (LAREMAR), ambos do Departamento de Química do Instituto de Ciências Exatas da UFMG, e para os ensaios de cocristalização e de resolução das estruturas por difração de raios-X que serão realizados no Laboratório Nacional de Luz Síncrotron do Centro Nacional de Pesquisa em Energia e Materiais (CNPEM).

6 CONCLUSÃO

O uso dos algoritmos de *docking* e variação dos seus diferentes parâmetros permitiu a validação pelos valores de RMSD de *redocking* e *crossdocking* de 2.352 protocolos de acoplamento molecular. Com a seleção do melhor protocolo, obtiveram-se valores de $\text{RMSD}_{\text{redocking}}$ de 0,38, $\text{RMSD}_{\text{crossdocking}}$ de 0,68 e AUC_{ROC} de 0,757 para a saFabI e, para a ecFabI, $\text{RMSD}_{\text{redocking}}$ de 0,64, $\text{RMSD}_{\text{crossdocking}}$ de 2,52 e AUC_{ROC} de 0,897. Embora os valores dessas métricas sejam aceitáveis, ao calcular a $\text{AUC}_{\text{BEDROC}}$ foi possível observar que esses protocolos resultaram em valores de 0,531 para saFabI e de 0,275 para a ecFabI que, juntamente com a análise do comportamento da curva ROC, permitem concluir que o uso de uma estratégia de triagem virtual exclusivamente baseada no *docking* seria pouco eficiente e induziria a seleção errônea de ligantes para ensaios *in vitro*.

A aplicação dos algoritmos de aprendizado de máquina permitiu a construção de 220.856.328 modelos de classificação utilizando diferentes algoritmos e combinações de hiperparâmetros. Os três melhores modelos de cada enzima tiveram valores de MCC_{int} variando entre 0,567 e 0,846 e MCC_{ext} variando entre 0,638 e 1,000, com os algoritmos *Support Vector Machine* (SVM) e *Multilayer Perceptron* (MLP) resultando nos melhores valores. Isso demonstra que os modelos construídos têm boa capacidade preditiva geral e robustez. Além disso, os modelos também apresentaram valores de bACC, TPR e TNR maiores do que 0,750, o que indica uma excelente acurácia entre os modelos, tanto em termos de verdadeiros positivos, quanto em termos de verdadeiros negativos.

Na comparação com os resultados de *docking*, os modelos de aprendizado de máquina superaram o uso do *docking* sozinho em todos os cenários e métricas avaliados, inclusive em termos de $\text{AUC}_{\text{BEDROC}}$, o que destaca a aplicabilidade da técnica de ML-*docking* em triagens virtuais. E, além disso, demonstra o sucesso na integração de algoritmos de *docking* com algoritmos de aprendizado de máquina.

Sobre o uso do LUNA para gerar os *fingerprints* de interação, o algoritmo permitiu a construção de uma representação das características que mais contribuíram para a classificação de ligantes em ativos e inativos. Nessa representação, além da observação de características já descritas na literatura como importantes para a atividade biológica, validando a interpretabilidade dos modelos, também foram observadas novas características que foram importantes para a

classificação, permitindo novas conclusões e contribuições para o desenvolvimento de inibidores da FabI. Também foi possível observar que o uso do MASSA foi capaz de garantir, na maioria dos casos, que as substâncias estejam dentro do domínio de aplicabilidade.

Em relação à obtenção das proteínas recombinantes, ambas as proteínas foram obtidas puras e estáveis em sua forma tetramérica. A saFabI foi obtida com $99,7 \pm 0,2\%$ de massa e rendimento de 17,29 mg/L de cultura e a ecFabI foi obtida com $98,9 \pm 0,4\%$ de massa e rendimento de 3,97 mg/L de cultura. Além disso, os resultados de exclusão molecular e do DLS indicam que as proteínas foram obtidas com índice de polidispersividade adequado para estudos de biologia estrutural, como os estudos de STD-NMR.

Por fim, a triagem virtual seguida de ensaios de MIC permitiu chegar em 9 substâncias com atividade antibacteriana contra *S. aureus* e/ou *E. coli* com provável interação com a FabI. Assim, espera-se realizar os ensaios de STD-NMR com esses ligantes e as proteínas obtidas para analisar se essas substâncias que apresentaram atividade antibacteriana de fato interagem com a FabI.

REFERÊNCIAS

ADRIÀ, C. M.; GARCIA-VALLVÉ, S.; PUJADAS, G. DecoyFinder, a tool for finding decoy molecules. **Journal of Cheminformatics**, v. 4, n. 1, p. 1–1, maio 2012.

AL-JANABI, S. S.; SHAWKY, H.; EL-WASEIF, A. A.; FARRAG, A. A.; ABDELGHANY, T. M.; EL-GHWAS, D. E. Stable, efficient, and cost-effective system for the biosynthesis of recombinant bacterial cellulose in Escherichia coli DH5 α platform. **Journal of Genetic Engineering and Biotechnology**, v. 20, n. 1, p. 1–10, dez. 2022.

ANTONELLO, R. M.; DI BELLA, S.; BETTS, J.; LA RAGIONE, R.; BRESSAN, R.; PRINCIPE, L.; MORABITO, S.; GIGLIUCCI, F.; TOZZOLI, R.; BUSETTI, M.; KNEZEVICH, A.; FURLANIS, L.; FONTANA, F.; LUZZARO, F.; LUZZATI, R.; LAGATOLLA, C. Zidovudine in synergistic combination with fosfomicin: an in vitro and in vivo evaluation against multidrug-resistant Enterobacterales. **International Journal of Antimicrobial Agents**, v. 58, n. 1, p. 1–8, jul. 2021.

ARIAN, R.; HARIRI, A.; MEHRIDEHNAVI, A.; FASSIHI, A.; GHASEMI, F. Protein kinase inhibitors' classification using K-Nearest neighbor algorithm. **Computational Biology and Chemistry**, v. 86, p. 1–7, jun. 2020.

ASSE JUNIOR, L. R.; KRONENBERGER, T.; SERAFIM, M. S. M.; SOUSA, Y. V.; FRANCO, I. D.; VALLI, M.; BOLZANI, V. da S.; MONTEIRO, G. C.; PRATES, J. L. B.; KROON, E. G.; MOTA, B. E. F.; FERREIRA, D. S.; OLIVEIRA, R. B. de; MALTAROLLO, V. G. Virtual screening of antibacterial compounds by similarity search of Enoyl-ACP reductase (FabI) inhibitors. **Future Medicinal Chemistry**, v. 12, n. 1, p. 51–68, nov. 2019.

ASHTAWY, H. M.; MAHAPATRA, N. R. Machine-learning scoring functions for identifying native poses of ligands docked to known and novel proteins. **BMC Bioinformatics**, v. 16, n. 6, p. 1–17, 17 abr. 2015.

AVRAM, S. I.; CRISAN, L.; BORA, A.; PACUREANU, L. M.; AVRAM, S.; KURUNCZI, L. Retrospective group fusion similarity search based on eROCE evaluation metric. **Bioorganic & Medicinal Chemistry**, v. 21, n. 5, p. 1268–1278, mar. 2013.

BAI, X. Seeing Atoms by Single-Particle Cryo-EM. **Trends in Biochemical Sciences**, v. 46, n. 4, p. 253–254, abr. 2021.

BALEMANS, W.; LOUNIS, N.; GILISSEN, R.; GUILLEMONT, J.; SIMMEN, K.; ANDRIES, K.; KOUL, A. Essentiality of FASII pathway for *Staphylococcus aureus*. **Nature**, v. 463, n. 7279, p. E3–E5, jan. 2010.

BALLESTER, P. J. Machine Learning for Molecular Modelling in Drug Design. **Biomolecules**, v. 9, n. 6, p. 216–218, jun. 2019.

BANEVICIUS, M. A.; KAPLAN, N.; HAFKIN, B.; NICOLAU, D. P. Pharmacokinetics, pharmacodynamics and efficacy of novel FabI inhibitor AFN-1252 against MSSA and MRSA in the murine thigh infection model. **Journal of Chemotherapy**, v. 25, n. 1, p. 26–31, fev. 2013.

BASSETTI, M.; DEL PUENTE, F.; MAGNASCO, L.; GIACOBBE, D. R. Innovative therapies for acute bacterial skin and skin-structure infections (ABSSSI) caused by methicillin-resistant *Staphylococcus aureus*: advances in phase I and II trials. **Expert Opinion on Investigational Drugs**, v. 29, n. 5, p. 495–506, maio 2020.

BERTHOLD, M. R.; CEBRON, N.; DILL, F.; GABRIEL, T. R.; KÖTTER, T.; MEINL, T.; OHL, P.; THIEL, K.; WISWEDEL, B. KNIME - the Konstanz information miner: version 2.0 and beyond. **ACM SIGKDD Explorations Newsletter**, v. 11, n. 1, p. 26–31, jun. 2009.

BIAN, Y.; JING, Y.; WANG, L.; MA, S.; JUN, J. J.; XIE, X. Q. Prediction of Orthosteric and Allosteric Regulations on Cannabinoid Receptors Using Supervised Machine Learning Classifiers. **Molecular Pharmaceutics**, v. 16, n. 6, p. 2605–2615, jun. 2019.

BLAUM, B. S.; NEU, U.; PETERS, T.; STEHLE, T. Spin ballet for sweet encounters: saturation-transfer difference NMR and X-ray crystallography complement each other in the elucidation of protein–glycan interactions. **Acta Crystallographica Section F: Structural Biology Communications**, v. 74, n. 8, p. 451–462, jul. 2018.

BURLEY, S. K.; BERMAN, H. M.; DUARTE, J. M.; FENG, Z.; FLATT, J. W.; HUDSON, B. P.; LOWE, R.; PEISACH, E.; PIEHL, D. W.; ROSE, Y.; SALI, A.; SEKHARAN, M.; SHAO, C.; VALLAT, B.; VOIGT, M.; WESTBROOK, J. D.; YOUNG, J. Y.; ZARDECKI, C. Protein Data Bank: A Comprehensive Review of 3D Structure Holdings and Worldwide Utilization by Researchers, Educators, and Students. **Biomolecules**, v. 12, n. 10, p. 1–27, out. 2022.

CAI, Y. D.; LIU, X. J.; XU, E. B.; CHOU, K. C. Support vector machines for predicting HIV protease cleavage sites in protein. **Journal of Computational Chemistry**, v. 23, n. 2, p. 267–274, jan. 2002.

CAPAN, M.; MOMBO-NGOMA, G.; MAKRISTATHIS, A.; RAMHARTER, M. Anti-bacterial activity of intermittent preventive treatment of malaria in pregnancy: Comparative in vitro study of sulphadoxine-pyrimethamine, mefloquine, and azithromycin. **Malaria Journal**, v. 9, n. 1, p. 1–5, out. 2010.

CASTILLO-GONZÁLEZ, D.; MERGNY, J. L.; DE RACHE, A.; PÉREZ-MACHADO, G.; CABRERA-PÉREZ, M. A.; NICOLOTTI, O.; INTROCASO, A.; MANGIATORDI, G. F.; GUÉDIN, A.; BOURDONCLE, A.; GARRIGUES, T.; PALLARDÓ, F.; CORDEIRO, M. N. D. S.; PAZ-Y-MIÑO, C.; TEJERA, E.; BORGES, F.; CRUZ-MONTEAGUDO, M. Harmonization of QSAR Best Practices and Molecular Docking Provides an Efficient Virtual Screening Tool for Discovering New G-Quadruplex Ligands. **Journal of Chemical Information and Modeling**, v. 55, n. 10, p. 2094–2110, out. 2015.

CERETO-MASSAGUÉ, A.; GUASCH, L.; VALLS, C.; MULERO, M.; PUJADAS, G.; GARCIA-VALLVÉ, S. DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets. **Bioinformatics**, v. 28, n. 12, p. 1661–1662, jun. 2012.

CHANG, A.; SCHIEBEL, J.; YU, W.; BOMMINENI, G. R.; PAN, P.; BAXTER, M. V.; KHANNA, A.; SOTRIFFER, C. A.; KISKER, C.; TONGE, P. J. Rational optimization of drug-target residence time: Insights from inhibitor binding to the staphylococcus aureus FabI enzyme-product complex. **Biochemistry**, v. 52, n. 24, p. 4217–4228, jun. 2013.

CHEN, A.; MINDREBO, J. T.; DAVIS, T. D.; KIM, W. E.; KATSUYAMA, Y.; JIANG, Z.; OHNISHI, Y.; NOEL, J. P.; BURKART, M. D.; KOBE, B. Mechanism-based cross-linking probes capture the Escherichia coli ketosynthase FabB in conformationally distinct catalytic states. **Acta Crystallographica Section D: Structural Biology**, v. 78, n. 9, p. 1171–1179, set. 2022.

CHEN, C. J.; HUANG, Y. C.; SHIE, S. Sen. Evolution of Multi-Resistance to Vancomycin, Daptomycin, and Linezolid in Methicillin-Resistant Staphylococcus aureus Causing Persistent Bacteremia. **Frontiers in Microbiology**, v. 11, n. 1414, p. 1–18, jul. 2020.

CHHATBAR, D. M.; CHAUBE, U. J.; VYAS, V. K.; BHATT, H. G. CoMFA, CoMSIA, Topomer CoMFA, HQSAR, molecular docking and molecular dynamics simulations study of triazine morpholino derivatives as mTOR inhibitors for the treatment of breast cancer. **Computational Biology and Chemistry**, v. 80, n. 1, p. 351–363, jun. 2019.

CHICCO, D.; JURMAN, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. **BMC Genomics**, v. 21, n. 1, p. 1–13, jan. 2020.

CHICCO, D.; JURMAN, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. **BioData Mining**, v. 16, n. 1, p. 1–23, dez. 2023.

CHICCO, D.; WARRENS, M. J.; JURMAN, G. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. **IEEE Access**, v. 9, p. 78368–78381, maio 2021.

CLARIVATE ANALYTICS. **Cortellis | Drug Discovery Intelligence**. Disponível em: <<https://www.cortellis.com/drugdiscovery/home>>. Acesso em: 10 ago. 2023.

CLINICAL AND LABORATORY STANDARDS INSTITUTE. **Performance standards for antimicrobial susceptibility testing: CLSI supplement M100**. 33 ed. Wayne: Clinical and Laboratory Standards Institute, 2017. 402 p.

COOPER, D. R.; POREBSKI, P. J.; CHRUSZCZ, M.; MINOR, W. X-ray crystallography: assessment and validation of protein–small molecule complexes for drug discovery. **Expert Opinion on Drug Discovery**, v. 6, n. 8, p. 771–782, ago. 2011.

COSTA, D. M. A. **Caracterização estrutural das enzimas cis-naftaleno dihidrodiol desidrogenase e salicilato hidroxilase recombinantes de Pseudomonas putida G7 envolvidas na degradação do naftaleno**. 2014. 156 p. Tese (Doutorado em Bioquímica e Imunologia) – Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, 2014.

COSTA-LUIS, C. da; LARROQUE, S. K.; ALTENDORF, K.; MARY, H.; RICHARDSHERIDAN; KOROBOV, M.; YORAV-RAPHAEL, N.; IVANOV, I.; BARGULL, M.; RODRIGUES, N.; CHEN, G.; LEE, A.; NEWAY, C.; CRAZYPYTHON; JC; ZUGNONI, M.; PAGEL, M. D.; MJSTEVENS777; DEKTYAREV, M.; ROTHBERG, A.; PLAVIN, A.; DILL, F.; FICHTEFOLL; STURM, G.; HEOHEO; KEMENADE, H. van; MCCRACKEN, J.; MAPLECCC; NORDLUND, M.; BOYLE, M. **tqdm: A fast, Extensible Progress Bar for Python and CLI**. Versão 4.64.1. [S.l.]: 2023. Repositório. Disponível em: <<https://zenodo.org/record/8233425>>. Acesso em: 15 set. 2023.

CRAMPON, K.; GIORKALLOS, A.; DELDOSSI, M.; BAUD, S.; STEFFENEL, L. A. Machine-learning methods for ligand–protein molecular docking. **Drug Discovery Today**, v. 27, n. 1, p. 151–164, jan. 2022.

DASSAULT SYSTÈMES. **BIOVIA Discovery Studio Visualizer**. v. 2021. [S.l.]: Dassault Systèmes, 2021. Disponível em: <<https://www.3ds.com/products-services/biovia/>>. Acesso em: 13 jul. 2023.

DELEO, F. R.; CHAMBERS, H. F. Reemergence of antibiotic-resistant *Staphylococcus aureus* in the genomics era. **The Journal of Clinical Investigation**, v. 119, n. 9, p. 2464–2474, set. 2009.

DESAI, M.; SHAH, M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). **Clinical eHealth**, v. 4, p. 1–11, jan. 2021.

DIAS, R.; AZEVEDO JUNIOR, W. Molecular Docking Algorithms. **Current Drug Targets**, v. 9, n. 12, p. 1040–1047, dez. 2008.

DJOKOVIC, N.; RAHNASTO-RILLA, M.; LOUGIAKIS, N.; LAHTELA-KAKKONEN, M.; NIKOLIC, K. SIRT2i_Predictor: A Machine Learning-Based Tool to Facilitate the Discovery of Novel SIRT2 Inhibitors. **Pharmaceuticals**, v. 16, n. 1, p. 1–24, jan. 2023.

DODGE, G. J.; PATEL, A.; JAREMKO, K. L.; MCCAMMON, J. A.; SMITH, J. L.; BURKART, M. D. Structural and dynamical rationale for fatty acid unsaturation in *Escherichia coli*. **Proceedings of the National Academy of Sciences of the United States of America**, v. 116, n. 14, p. 6775–6783, abr. 2019.

DRESP-LANGLEY, B.; EKSETH, O. K.; FESL, J.; GOHSHI, S.; KURZ, M.; SEHRING, H. W. Occam's Razor for Big Data? On Detecting Quality in Large Unstructured Datasets. **Applied Sciences**, v. 9, n. 15, p. 1–28, jul. 2019.

DUNN, S. J.; CONNOR, C.; MCNALLY, A. The evolution and transmission of multi-drug resistant *Escherichia coli* and *Klebsiella pneumoniae*: the complexity of clones and plasmids. **Current Opinion in Microbiology**, v. 51, p. 51–56, out. 2019.

ELTSCHKNER, S.; KEHREIN, J.; LE, T. A.; DAVOODI, S.; MERGET, B.; BASAK, S.; WEINRICH, J. D.; SCHIEBEL, J.; TONGE, P. J.; ENGELS, B.; SOTRIFFER, C.; KISKER, C. A Long Residence Time Enoyl-Reductase Inhibitor Explores an Extended Binding Region with Isoenzyme-Dependent Tautomer Adaptation and Differential

Substrate-Binding Loop Closure. **ACS Infectious Diseases**, v. 7, n. 4, p. 746–758, abr. 2021.

EMMERT-STREIB, F.; MOUTARI, S.; DEHMER, M. A comprehensive survey of error measures for evaluating binary decision making in data science. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 9, n. 5, p. 1–15, set. 2019.

ESCAICH, S.; PROUVENSIER, L.; SACCOMANI, M.; DURANT, L.; OXOBY, M.; GERUSZ, V.; MOREAU, F.; VONGSOUTH, V.; MAHER, K.; MORRISSEY, I.; SOULAMA-MOUZE, C. The MUT056399 inhibitor of FabI is a new antistaphylococcal compound. **Antimicrobial Agents and Chemotherapy**, v. 55, n. 10, p. 4692–4697, out. 2011.

FAGE, C. D.; LATHOUWERS, T.; VANMEERT, M.; GAO, L. J.; VRANCKEN, K.; LAMMENS, E. M.; WEIR, A. N. M.; DEGROOTE, R.; CUPPENS, H.; KOSOL, S.; SIMPSON, T. J.; CRUMP, M. P.; WILLIS, C. L.; HERDEWIJN, P.; LESCRIER, E.; LAVIGNE, R.; ANNÉ, J.; MASSCHELEIN, J. The Kalimantacin Polyketide Antibiotics Inhibit Fatty Acid Biosynthesis in *Staphylococcus aureus* by Targeting the Enoyl-Acyl Carrier Protein Binding Site of FabI. **Angewandte Chemie International Edition**, v. 59, n. 26, p. 10549–10556, jun. 2020.

FAN, J.; FU, A.; ZHANG, L. Progress in molecular docking. **Quantitative Biology**, v. 7, n. 2, p. 83–89, jun. 2019.

FASSIO, A. V.; SHUB, L.; PONZONI, L.; MCKINLEY, J.; O'MEARA, M. J.; FERREIRA, R. S.; KEISER, M. J.; DE MELO MINARDI, R. C. Prioritizing Virtual Screening with Interpretable Interaction Fingerprints. **Journal of Chemical Information and Modeling**, v. 62, n. 18, p. 4300–4318, set. 2022.

FERNANDES, P. O. **Study of the interaction between ABL kinase and active compounds against chronic myeloid leukemia through computational techniques and nuclear magnetic resonance**. 2022. Dissertação (Mestrado em

Ciências Farmacêuticas) – Faculdade de Farmácia, Universidade Federal de Minas Gerais, Belo Horizonte, 2022.

FERNANDES, P. O.; MARTINS, D. M.; DE SOUZA BOZZI, A.; MARTINS, J. P. A.; DE MORAES, A. H.; MALTAROLLO, V. G. Molecular insights on ABL kinase activation using tree-based machine learning models and molecular docking. **Molecular Diversity**, v. 25, n. 3, p. 1301–1314, ago. 2021.

FRANCO, B. E.; MARTÍNEZ, M. A.; SÁNCHEZ RODRÍGUEZ, M. A.; WERTHEIMER, A. I. The determinants of the antibiotic resistance process. **Infection and Drug Resistance**, v. 2, n. 1, p. 1–11, 2009.

FRANK, E.; TRIGG, L.; HOLMES, G.; WITTEN, I. H. Technical note: Naive Bayes for regression. **Machine Learning**, v. 41, n. 1, p. 5–25, out. 2000.

FRANK, M. W.; YAO, J.; BATTE, J. L.; GULLETT, J. M.; SUBRAMANIAN, C.; ROSCH, J. W.; ROCK, C. O. Host fatty acid utilization by staphylococcus aureus at the infection site. **mBio**, v. 11, n. 3, p. 1–14, maio 2020.

FRIESNER, R. A.; BANKS, J. L.; MURPHY, R. B.; HALGREN, T. A.; KLICIC, J. J.; MAINZ, D. T.; REPASKY, M. P.; KNOLL, E. H.; SHELLEY, M.; PERRY, J. K.; SHAW, D. E.; FRANCIS, P.; SHENKIN, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. **Journal of Medicinal Chemistry**, v. 47, n. 7, p. 1739–1749, mar. 2004.

FRIESNER, R. A.; MURPHY, R. B.; REPASKY, M. P.; FRYE, L. L.; GREENWOOD, J. R.; HALGREN, T. A.; SANSCHAGRIN, P. C.; MAINZ, D. T. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. **Journal of Medicinal Chemistry**, v. 49, n. 21, p. 6177–6196, out. 2006.

GARCIA DENEGRI, M. E.; MARUÑAK, S.; TODARO, J. S.; PONCE-SOTO, L. A.; ACOSTA, O.; LEIVA, L. Neutralisation of the pharmacological activities of Bothrops alternatus venom by anti-PLA2 IgGs. **Toxicon**, v. 86, p. 89–95, 1 ago. 2014.

GAULTON, A.; HERSEY, A.; NOWOTKA, M. L.; PATRICIA BENTO, A.; CHAMBERS, J.; MENDEZ, D.; MUTOWO, P.; ATKINSON, F.; BELLIS, L. J.; CIBRIAN-UHALTE, E.; DAVIES, M.; DEDMAN, N.; KARLSSON, A.; MAGARINOS, M. P.; OVERINGTON, J. P.; PAPADATOS, G.; SMIT, I.; LEACH, A. R. The ChEMBL database in 2017. **Nucleic Acids Research**, v. 45, n. D1, p. D945–D954, jan. 2017.

GENTILE, F.; AGRAWAL, V.; HSING, M.; TON, A.-T.; BAN, F.; NORINDER, U.; GLEAVE, M. E.; CHERKASOV, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. **ACS Central Science**, v. 6, n. 6, p. 939–949, jun. 2020.

GHATTAS, M. A.; EISSA, N. A.; TESSARO, F.; PEROZZO, R.; SCAPOZZA, L.; OBAID, D.; ATATREH, N. Structure-based drug design and in vitro testing reveal new inhibitors of enoyl-acyl carrier protein reductases. **Chemical Biology & Drug Design**, v. 94, n. 2, p. 1545–1555, ago. 2019.

GIACOPPO, J. O. S.; MANCINI, D. T.; GUIMARÃES, A. P.; GONÇALVES, A. S.; DA CUNHA, E. F. F.; FRANÇA, T. C. C.; RAMALHO, T. C. Molecular modeling toward selective inhibitors of dihydrofolate reductase from the biological warfare agent *Bacillus anthracis*. **European Journal of Medicinal Chemistry**, v. 91, p. 63–71, fev. 2015.

GIMENO, A.; SANTOS, L. M.; ALEMI, M.; RIVAS, J.; BLASI, D.; COTRINA, E. Y.; LLOP, J.; VALENCIA, G.; CARDOSO, I.; QUINTANA, J.; ARSEQUELL, G.; JIMÉNEZ-BARBERO, J. Insights on the Interaction between Transthyretin and A β in Solution. A Saturation Transfer Difference (STD) NMR Analysis of the Role of Iododiflunisal. **Journal of Medicinal Chemistry**, v. 60, n. 13, p. 5749–5758, jul. 2017.

GODBOLE, V.; DAHL, G. E.; GILMER, J.; SHALLUE, C. J.; NADO, Z. **Deep Learning Tuning Playbook**. 2023. Disponível em: <https://github.com/google-research/tuning_playbook>. Acesso em: 19 ago. 2023.

GOMES, A. R.; BYREGOWDA, S. M.; VEEREGOWDA, B. M.; BALAMURUGAN, V. An Overview of Heterologous Expression Host Systems for the Production of

Recombinant Proteins. **Advances in Animal and Veterinary Sciences**, v. 4, n. 7, p. 345–356, 2016.

GRAMATICA, P. Principles of QSAR Modeling: Comments and Suggestions From Personal Experience. **International Journal of Quantitative Structure-Property Relationships**, v. 5, n. 3, p. 61–97, jan. 2020.

GREENACRE, M.; GROENEN, P. J. F.; HASTIE, T.; D'ENZA, A. I.; MARKOS, A.; TUZHILINA, E. Principal component analysis. **Nature Reviews Methods Primers**, v. 2, n. 1, p. 1–21, dez. 2022.

GU, W.; LI, Q.; LI, Y. Law and mechanism analysis of biodegradability of polychlorinated naphthalenes based on principal component analysis, QSAR models, molecular docking and molecular dynamics simulation. **Chemosphere**, v. 243, n. 1, p. 1–8, mar. 2020.

GÜNTHER, F.; FRITSCH, S. neuralnet: Training of Neural Networks. **The R Journal**, v. 2, n. 1, p. 30–38, jun. 2010.

GUNTURI, S. B.; NARAYANAN, R. In Silico ADME Modeling 3: Computational Models to Predict Human Intestinal Absorption Using Sphere Exclusion and kNN QSAR Methods. **QSAR & Combinatorial Science**, v. 26, n. 5, p. 653–668, 1 maio 2007.

HAFKIN, B.; KAPLAN, N.; MURPHY, B. Efficacy and Safety of AFN-1252, the First Staphylococcus-Specific Antibacterial Agent, in the Treatment of Acute Bacterial Skin and Skin Structure Infections, Including Those in Patients with Significant Comorbidities. **Antimicrobial Agents and Chemotherapy**, v. 60, n. 3, p. 1695–1701, mar. 2016.

HALGREN, T. A.; MURPHY, R. B.; FRIESNER, R. A.; BEARD, H. S.; FRYE, L. L.; POLLARD, W. T.; BANKS, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. **Journal of Medicinal Chemistry**, v. 47, n. 7, p. 1750–1759, mar. 2004.

HAN, L.; WANG, Y.; BRYANT, S. H. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem. **BMC Bioinformatics**, v. 9, n. 1, p. 1–8, set. 2008.

HAWKINS, P. C. D.; SKILLMAN, A. G.; WARREN, G. L.; ELLINGSON, B. A.; STAHL, M. T. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and cambridge structural database. **Journal of Chemical Information and Modeling**, v. 50, n. 4, p. 572–584, abr. 2010.

HOLDERBACH, S.; ADAM, L.; JAYARAM, B.; WADE, R. C.; MUKHERJEE, G. RASPD+: Fast Protein-Ligand Binding Free Energy Prediction Using Simplified Physicochemical Features. **Frontiers in Molecular Biosciences**, v. 7, p. 1–14, dez. 2020.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science & Engineering**, v. 9, n. 3, p. 90–95, maio 2007.

IRWIN, J. J.; STERLING, T.; MYSINGER, M. M.; BOLSTAD, E. S.; COLEMAN, R. G. ZINC: A free tool to discover chemistry for biology. **Journal of Chemical Information and Modeling**, v. 52, n. 7, p. 1757–1768, jul. 2012.

ISTYASTONO, E. P.; RADIFAR, M.; YUNIARTI, N.; PRASASTY, V. D.; MUNGKASI, S. PyPLIF HIPPOS: A Molecular Interaction Fingerprinting Tool for Docking Results of AutoDock Vina and PLANTS. **Journal of Chemical Information and Modeling**, v. 60, n. 8, p. 3697–3702, ago. 2020.

JAIN, A. N. Scoring Functions for Protein-Ligand Docking. **Current Protein & Peptide Science**, v. 7, n. 5, p. 407–420, out. 2006.

JAYARAJ, P. B.; JAIN, S. Ligand based virtual screening using SVM on GPU. **Computational Biology and Chemistry**, v. 83, p. 1–9, nov. 2019.

JENSEN, S. O.; LYON, B. R. Genetics of antimicrobial resistance in *Staphylococcus aureus*. **Future microbiology**, v. 4, n. 5, p. 565–582, jun. 2009.

JONES, G.; WILLETT, P.; GLEN, R. C.; LEACH, A. R.; TAYLOR, R. Development and validation of a genetic algorithm for flexible docking. **Journal of Molecular Biology**, v. 267, n. 3, p. 727–748, abr. 1997.

JORDAN, C. A.; SANDOVAL, B. A.; SEROBYAN, M. V.; GILLING, D. H.; GROZIAK, M. P.; XU, H. H.; VEY, J. L. Crystallographic insights into the structure-activity relationships of diazaborine enoyl-ACP reductase inhibitors. **Acta Crystallographica Section: F Structural Biology Communications**, v. 71, n. 12, p. 1521–1530, nov. 2015.

KAPLAN, N.; ALBERT, M.; AWREY, D.; BARDOUNIOTIS, E.; BERMAN, J.; CLARKE, T.; DORSEY, M.; HAFKIN, B.; RAMNAUTH, J.; ROMANOV, V.; SCHMID, M. B.; THALAKADA, R.; YETHON, J.; PAULS, H. W. Mode of action, in vitro activity, and in vivo efficacy of AFN-1252, a selective antistaphylococcal FabI inhibitor. **Antimicrobial Agents and Chemotherapy**, v. 56, n. 11, p. 5865–5874, nov. 2012.

KÉNANIAN, G.; MORVAN, C.; WECKEL, A.; PATHANIA, A.; ANBA-MONDOLONI, J.; HALPERN, D.; GAILLARD, M.; SOLGADI, A.; DUPONT, L.; HENRY, C.; POYART, C.; FOUET, A.; LAMBERET, G.; GLOUX, K.; GRUSS, A. Permissive Fatty Acid Incorporation Promotes Staphylococcal Adaptation to FASII Antibiotics in Host Environments. **Cell Reports**, v. 29, n. 12, p. 3974- 3982.e4, dez. 2019.

KENSERT, A.; ALVARSSON, J.; NORINDER, U.; SPJUTH, O. Evaluating parameters for ligand-based modeling with random forest on sparse data sets. **Journal of Cheminformatics**, v. 10, n. 1, p. 1–10, out. 2018.

KIM, H. T.; KIM, S.; NA, B. K.; CHUNG, J.; HWANG, E.; HWANG, K. Y. Structural insights into the dimer-tetramer transition of FabI from *Bacillus anthracis*. **Biochemical and Biophysical Research Communications**, v. 493, n. 1, p. 28–33, nov. 2017.

KIM, S.; CHEN, J.; CHENG, T.; GINDULYTE, A.; HE, J.; HE, S.; LI, Q.; SHOEMAKER, B. A.; THIESSEN, P. A.; YU, B.; ZASLAVSKY, L.; ZHANG, J.; BOLTON, E. E. PubChem 2023 update. **Nucleic Acids Research**, v. 51, n. D1, p. D1373–D1380, jan. 2023.

KIRCHMAIR, J.; MARKT, P.; DISTINTO, S.; WOLBER, G.; LANGER, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection - What can we learn from earlier mistakes? **Journal of Computer-Aided Molecular Design**, v. 22, n. 3–4, p. 213–228, jan. 2008.

KRISHNA, S.; STAINES, H. M. Non-antifolate antibiotics: Clindamycin, doxycycline, azithromycin and fosmidomycin. In: STAINES, H. M.; KRISHNA, S. **Treatment and Prevention of Malaria: Antimalarial Drug Chemistry, Action and Use**. 1. ed. Basel: Springer Basel, 2012. p. 141–156.

KRONENBERGER, T.; DE OLIVEIRA FERNADES, P.; DRUMOND FRANCO, I.; POSO, A.; GONÇALVES MALTAROLLO, V. Ligand- and Structure-Based Approaches of Escherichia coli FabI Inhibition by Triclosan Derivatives: From Chemical Similarity to Protein Dynamics Influence. **ChemMedChem**, v. 14, n. 23, p. 1995–2004, dez. 2019.

KRONENBERGER, T.; ASSE JUNIOR, L. R.; WRENGER, C.; TROSSINI, G. H. G.; HONORIO, K. M.; MALTAROLLO, V. G. Studies of Staphylococcus aureus FabI inhibitors: fragment-based approach based on holographic structure–activity relationship analyses. **Future Medicinal Chemistry**, v. 9, n. 2, p. 135–151, jan. 2017.

KUZ'MIN, V. E.; POLISHCHUK, P. G.; ARTEMENKO, A. G.; ANDRONATI, S. A. Interpretation of QSAR Models Based on Random Forest Methods. **Molecular Informatics**, v. 30, n. 6–7, p. 593–603, jun. 2011.

LAI, C. K. C.; NG, R. W. Y.; LEUNG, S. S. Y.; HUI, M.; IP, M. Overcoming the rising incidence and evolving mechanisms of antibiotic resistance by novel drug delivery approaches – An overview. **Advanced Drug Delivery Reviews**, v. 181, n. 1, p. 1–19, fev. 2022.

LANDETA, C.; MEJIA-SANTANA, A. Union Is Strength: Target-Based and Whole-Cell High-Throughput Screens in Antibacterial Discovery. **Journal of Bacteriology**, v. 204, n. 4, abr. 2022.

LANDRUM, G. **RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling**. 2013.

LI, H.; LEUNG, K. S.; WONG, M. H.; BALLESTER, P. J. Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. **Molecular Informatics**, v. 34, n. 2–3, p. 115–126, fev. 2015.

LI, J.; FU, A.; ZHANG, L. An Overview of Scoring Functions Used for Protein-Ligand Interactions in Molecular Docking. **Interdisciplinary Sciences: Computational Life Sciences**, v. 11, n. 2, p. 320–328, mar. 2019.

LI, J.; LUO, D.; WEN, T.; LIU, Q.; MO, Z. Representative feature selection of molecular descriptors in QSAR modeling. **Journal of Molecular Structure**, v. 1244, p. 1–8, nov. 2021.

LIMA, A. N.; PHILOT, E. A.; TROSSINI, G. H. G.; SCOTT, L. P. B.; MALTAROLLO, V. G.; HONORIO, K. M. Use of machine learning approaches for novel drug discovery. **Expert Opinion on Drug Discovery**, v. 11, n. 3, p. 225–239, fev. 2016.

LIMBU, S.; DAKSHANAMURTHY, S. A New Hybrid Neural Network Deep Learning Method for Protein–Ligand Binding Affinity Prediction and De Novo Drug Design. **International Journal of Molecular Sciences**, v. 23, n. 22, p. 1–16, 1 nov. 2022.

LIPÍŃSKI, P. F. J.; SZURMAK, P. SCRAMBLE'N'GAMBLE: A tool for fast and facile generation of random data for statistical evaluation of QSAR models. **Chemical Papers**, v. 71, n. 11, p. 2217–2232, nov. 2017.

LU, C.; WU, C.; GHOREISHI, D.; CHEN, W.; WANG, L.; DAMM, W.; ROSS, G. A.; DAHLGREN, M. K.; RUSSELL, E.; VON BARGEN, C. D.; ABEL, R.; FRIESNER, R. A.; HARDER, E. D. OPLS4: Improving force field accuracy on challenging regimes of chemical space. **Journal of Chemical Theory and Computation**, v. 17, n. 7, p. 4291–4300, jul. 2021.

LU, H.; TONGE, P. J. Inhibitors of FabI, an Enzyme Drug Target in the Bacterial Fatty Acid Biosynthesis Pathway. **Accounts of Chemical Research**, v. 41, n. 1, p. 11–20, jan. 2008.

LYUMKIS, D. Challenges and opportunities in cryo-EM single-particle analysis. **Journal of Biological Chemistry**, v. 294, n. 13, p. 5181–5197, mar. 2019.

MACEK, B.; FORCHHAMMER, K.; HARDOUIN, J.; WEBER-BAN, E.; GRANGEASSE, C.; MIJAKOVIC, I. Protein post-translational modifications in bacteria. **Nature Reviews Microbiology**, v. 17, n. 11, p. 651–664, set. 2019.

MACIEJEWSKI, M. W.; SCHUYLER, A. D.; GRYK, M. R.; MORARU, I. I.; ROMERO, P. R.; ULRICH, E. L.; EGHBALNIA, H. R.; LIVNY, M.; DELAGLIO, F.; HOCH, J. C. NMRbox: A Resource for Biomolecular NMR Computation. **Biophysical journal**, v. 112, n. 8, p. 1529–1534, abr. 2017.

MALTAROLLO, V. G. Classification of Staphylococcus Aureus FabI Inhibitors by Machine Learning Techniques. **International Journal of Quantitative Structure-Property Relationships (IJQSPR)**, v. 4, n. 4, p. 1–14, jan. 2019.

MALTAROLLO, V. G.; SHEVCHENKO, E.; LIMA, I. D. D. M.; CINO, E. A.; FERREIRA, G. M.; POSO, A.; KRONENBERGER, T. Do Go Chasing Waterfalls: Enoyl Reductase (FabI) in Complex with Inhibitors Stabilizes the Tetrameric Structure and Opens Water Channels. **Journal of Chemical Information and Modeling**, v. 62, n. 22, p. 5746–5761, nov. 2022.

MANDAL, L.; JANA, N. D. A comparative study of naive bayes and k-NN algorithm for multi-class drug molecule classification. In: IEEE INDIA COUNCIL INTERNATIONAL CONFERENCE (INDICON), 16, 2019. **Symposium Proceedings**. Institute of Electrical and Electronics Engineers, 2019. p. 1–4.

MARZO, M.; BENFENATI, E. Classification of a Naïve Bayesian Fingerprint model to predict reproductive toxicity. **SAR and QSAR in Environmental Research**, v. 29, n. 8, p. 631–645, jul. 2018.

MATTHEUS, W.; MASSCHELEIN, J.; GAO, L. J.; HERDEWIJN, P.; LANDUYT, B.; VOLCKAERT, G.; LAVIGNE, R. The Kalimantacin/Batumin Biosynthesis Operon Encodes a Self-Resistance Isoform of the FabI Bacterial Target. **Chemistry & Biology**, v. 17, n. 10, p. 1067–1071, out. 2010.

MAVEYRAUD, L.; MOUREY, L. Protein X-ray Crystallography and Drug Discovery. **Molecules**, v. 25, n. 5, p. 1–18, fev. 2020.

MCHUGH, M. L. Interrater reliability: The kappa statistic. **Biochemia Medica**, v. 22, n. 3, p. 276–282, out. 2012.

MCGANN, M. FRED pose prediction and virtual screening accuracy. **Journal of Chemical Information and Modeling**, v. 51, n. 3, p. 578–596, mar. 2011.

MCGANN, M. FRED and HYBRID docking performance on standardized datasets. **Journal of Computer-Aided Molecular Design**, v. 26, n. 8, p. 897–906, jun. 2012.

MCNUTT, A. T.; FRANCOEUR, P.; AGGARWAL, R.; MASUDA, T.; MELI, R.; RAGOZA, M.; SUNSERI, J.; KOES, D. R. GNINA 1.0: molecular docking with deep learning. **Journal of Cheminformatics**, v. 13, n. 1, p. 1–20, dez. 2021.

MEHBOOB, S.; HEVENER, K. E.; TRUONG, K.; BOCI, T.; SANTARSIERO, B. D.; JOHNSON, M. E. Structural and enzymatic analyses reveal the binding mode of a novel series of francisella tularensis enoyl reductase (FabI) inhibitors. **Journal of Medicinal Chemistry**, v. 55, n. 12, p. 5933–5941, jun. 2012.

MERK, A.; BARTESAGHI, A.; BANERJEE, S.; FALCONIERI, V.; RAO, P.; DAVIS, M. I.; PRAGANI, R.; BOXER, M. B.; EARL, L. A.; MILNE, J. L. S.; SUBRAMANIAM, S. Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. **Cell**, v. 165, n. 7, p. 1698–1707, jun. 2016.

MI, X.; ZOU, B.; ZOU, F.; HU, J. Permutation-based identification of important biomarkers for complex diseases via machine learning models. **Nature Communications**, v. 12, n. 1, p. 1–12, maio 2021.

MICROSOFT CORPORATION. **Visual Studio Code**. Disponível em: <<https://github.com/microsoft/vscode>>. Acesso em: 22 set. 2023.

MINTON, A. P. Recent applications of light scattering measurement in the biological and biopharmaceutical sciences. **Analytical Biochemistry**, v. 501, n. 1, p. 4–22, maio 2016.

MIR, L. **Genômica**. 1. ed. São Paulo: Editora Atheneu, 2004. 1114 p.

MOREHEAD, M. S.; SCARBROUGH, C. Emergence of Global Antibiotic Resistance. **Primary Care: Clinics in Office Practice**, v. 45, n. 3, p. 467–484, set. 2018.

MORRIS, G. M.; GOODSELL, D. S.; HALLIDAY, R. S.; HUEY, R.; HART, W. E.; BELEW, R. K.; OLSON, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. **Journal of Computational Chemistry**, v. 19, n. 14, p. 1639–1662, jun. 1998.

MORVAN, C.; HALPERN, D.; KÉNIANIAN, G.; PATHANIA, A.; ANBA-MONDOLONI, J.; LAMBERET, G.; GRUSS, A.; GLOUX, K. The *Staphylococcus aureus* FASII bypass escape route from FASII inhibitors. **Biochimie**, v. 141, n. 1, p. 40–46, out. 2017.

MOSLEY, L. S. D. **A balanced approach to the multi-class imbalance problem**. 2013. Tese (Doctor of Philosophy: Industrial Engineering) – Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, 2013.

MOSTAFA, A. A.; ALHOSSARY, A. A.; SALEM, S. A.; MOHAMED, A. E. GBO-kNN a new framework for enhancing the performance of ligand-based virtual screening for drug discovery. **Expert Systems with Applications**, v. 197, n. 1, p. 1–10, jul. 2022.

MURRAY, C. J.; IKUTA, K. S.; SHARARA, F.; SWETSCHINSKI, L.; ROBLES AGUILAR, G.; GRAY, A.; HAN, C.; BISIGNANO, C.; RAO, P.; WOOL, E.; JOHNSON, S. C.; BROWNE, A. J.; CHIPETA, M. G.; FELL, F.; HACKETT, S.; HAINESWOODHOUSE, G.; KASHEF HAMADANI, B. H.; KUMARAN, E. A. P.; MCMANIGAL, B.; AGARWAL, R.; AKECH, S.; ALBERTSON, S.; AMUASI, J.; ANDREWS, J.; ARAVKIN, A.; ASHLEY, E.; BAILEY, F.; BAKER, S.; BASNYAT, B.; BEKKER, A.; BENDER, R.; BETHOU, A.; BIELICKI, J.; BOONKASIDECHA, S.; BUKOSIA, J.; CARVALHEIRO, C.; CASTAÑEDA-ORJUELA, C.; CHANSAMOUTH, V.; CHAURASIA, S.; CHIURCHIÙ, S.; CHOWDHURY, F.; COOK, A. J.; COOPER, B.; CRESSEY, T. R.; CRIOLLO-MORA, E.; CUNNINGHAM, M.; DARBOE, S.; DAY, N. P. J.; DE LUCA, M.; DOKOVA, K.; DRAMOWSKI, A.; DUNACHIE, S. J.; ECKMANNS, T.; EIBACH, D.; EMAMI, A.; FEASEY, N.; FISHER-PEARSON, N.; FORREST, K.; GARRETT, D.; GASTMEIER, P.; GIREF, A. Z.; GREER, R. C.; GUPTA, V.; HALLER, S.; HASELBECK, A.; HAY, S. I.; HOLM, M.; HOPKINS, S.; IREGBU, K. C.; JACOBS, J.; JAROVSKY, D.; JAVANMARDI, F.; KHORANA, M.; KISSOON, N.; KOBEISSI, E.; KOSTYANEV, T.; KRAPP, F.; KRUMKAMP, R.; KUMAR, A.; KYU, H. H.; LIM, C.; LIMMATHUROTSAKUL, D.; LOFTUS, M. J.; LUNN, M.; MA, J.; MTURI, N.; MUNERAHUERTAS, T.; MUSICHA, P.; MUSSI-PINHATA, M. M.; NAKAMURA, T.; NANAVATI, R.; NANGIA, S.; NEWTON, P.; NGOUN, C.; NOVOTNEY, A.; NWAKANMA, D.; OBIERO, C. W.; OLIVAS-MARTINEZ, A.; OLLIARO, P.; OOKO, E.; ORTIZ-BRIZUELA, E.; PELEG, A. Y.; PERRONE, C.; PLAKKAL, N.; PONCE-DE-LEON, A.; RAAD, M.; RAMDIN, T.; RIDDELL, A.; ROBERTS, T.; ROBOTHAM, J. V.; ROCA, A.; RUDD, K. E.; RUSSELL, N.; SCHNALL, J.; SCOTT, J. A. G.; SHIVAMALLAPPA, M.; SIFUENTES-OSORNIO, J.; STEENKESTE, N.; STEWARDSON, A. J.; STOEVA, T.; TASAK, N.; THAIPRAKONG, A.; THWAITES, G.; TURNER, C.; TURNER, P.; VAN DOORN, H. R.; VELAPHI, S.; VONGPRADITH, A.; VU, H.; WALSH, T.; WANER, S.; WANGRANGSIMAKUL, T.; WOZNIAK, T.; ZHENG, P.; SARTORIUS, B.; LOPEZ, A. D.; STERGACHIS, A.; MOORE, C.; DOLECEK, C.; NAGHAVI, M. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. **The Lancet**, v. 399, n. 10325, p. 629–655, fev. 2022.

NEVES, B. J.; BRAGA, R. C.; ALVES, V. M.; LIMA, M. N. N.; CASSIANO, G. C.; MURATOV, E. N.; COSTA, F. T. M.; ANDRADE, C. H. Deep Learning-driven research

for drug discovery: Tackling Malaria. **PLOS Computational Biology**, v. 16, n. 2, p. 1–21, fev. 2020.

NGUYEN, C.; HAUSHALTER, R. W.; LEE, D. J.; MARKWICK, P. R. L.; BRUEGGER, J.; CALDARA-FESTIN, G.; FINZEL, K.; JACKSON, D. R.; ISHIKAWA, F.; O'DOWD, B.; MCCAMMON, J. A.; OPELLA, S. J.; TSAI, S. C.; BURKART, M. D. Trapping the dynamic acyl carrier protein in fatty acid biosynthesis. **Nature**, v. 505, n. 7483, p. 427–431, dez. 2013.

OEDOCKING 4.2.0.2. Santa Fe, NM. OpenEye Scientific Software, Inc., 2023. Disponível em: <<https://docs.eyesopen.com/applications/oedocking/>>. Acesso em: 10 jun. 2023.

OLIVEIRA, C.; DOMINGUES, L. Guidelines to reach high-quality purified recombinant proteins. **Applied Microbiology and Biotechnology**, v. 102, n. 1, p. 81–92, jan. 2018.

OPENEYE SCIENTIFIC SOFTWARE. **OMEGA 4.2.1.1** Santa Fe, NM, USA. 2023. Disponível em: <<http://www.eyesopen.com>>. Acesso em: 14 jul. 2020.

ORGANIZAÇÃO MUNDIAL DE SAÚDE. **WHO publishes list of bacteria for which new antibiotics are urgently needed**. 2017. Disponível em: <<https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed>>. Acesso em: 13 jul. 2021.

PAPANEOPHYTOU, C. P.; KONTOPIDIS, G. Statistical approaches to maximize recombinant protein expression in *Escherichia coli*: A general review. **Protein Expression and Purification**, v. 94, n. 1, p. 22–32, fev. 2014.

PARK, H. S.; YOON, Y. M.; JUNG, S. J.; KIM, C. M.; KIM, J. M.; KWAK, J. H. Antistaphylococcal activities of CG400549, a new bacterial enoyl-acyl carrier protein reductase (FabI) inhibitor. **Journal of Antimicrobial Chemotherapy**, v. 60, n. 3, p. 568–574, set. 2007a.

PARK, H. S.; YOON, Y. M.; JUNG, S. J.; YUN, I. N. R.; KIM, C. M.; KIM, J. M.; KWAK, J. H. CG400462, a new bacterial enoyl–acyl carrier protein reductase (FabI) inhibitor. **International Journal of Antimicrobial Agents**, v. 30, n. 5, p. 446–451, nov. 2007b.

PARKER, E. N.; CAIN, B. N.; HAJIAN, B.; ULRICH, R. J.; GEDDES, E. J.; BARKHO, S.; LEE, H. Y.; WILLIAMS, J. D.; RAYNOR, M.; CARIDHA, D.; ZAINO, A.; SHEKHAR, M.; MUÑOZ, K. A.; RZASA, K. M.; TEMPLE, E. R.; HUNT, D.; JIN, X.; VUONG, C.; PANNONE, K.; KELLY, A. M.; MULLIGAN, M. P.; LEE, K. K.; LAU, G. W.; HUNG, D. T.; HERGENROTHER, P. J. An Iterative Approach Guides Discovery of the FabI Inhibitor Fabimycin, a Late-Stage Antibiotic Candidate with In Vivo Efficacy against Drug-Resistant Gram-Negative Infections. **ACS Central Science**, v. 8, n. 8, p. 1145–1158, ago. 2022.

PARSONS, J. B.; FRANK, M. W.; JACKSON, P.; SUBRAMANIAN, C.; ROCK, C. O. Incorporation of extracellular fatty acids by a fatty acid kinase-dependent pathway in *Staphylococcus aureus*. **Molecular Microbiology**, v. 92, n. 2, p. 234–245, abr. 2014.

PARSONS, J. B.; ROCK, C. O. Is bacterial fatty acid synthesis a valid target for antibacterial drug discovery? **Current Opinion in Microbiology**, v. 14, n. 5, p. 544–549, out. 2011.

PATHANIA, A.; ANBA-MONDOLONI, J.; GOMINET, M.; HALPERN, D.; DAIROU, J.; DUPONT, L.; LAMBERET, G.; TRIEU-CUOT, P.; GLOUX, K.; GRUSS, A. (P)ppgpp/gtp and malonyl-coa modulate *staphylococcus aureus* adaptation to fasii antibiotics and provide a basis for synergistic bi-therapy. **mBio**, v. 12, n. 1, p. 1–15, jan. 2021.

PATRA, J. C.; CHUA, K. H. K. Neural network based drug design for diabetes mellitus using QSAR with 2D and 3D descriptors. In: IEEE International Joint Conference on Neural Networks (IJCNN), 2010. **Proceedings**, 2010. p. 1–8.

PAYNE, D. J.; MILLER, W. H.; BERRY, V.; BROSKEY, J.; BURGESS, W. J.; CHEN, E.; DEWOLF, W. E.; FOSBERRY, A. P.; GREENWOOD, R.; HEAD, M. S.; HEERDING, D. A.; JANSON, C. A.; JAWORSKI, D. D.; KELLER, P. M.; MANLEY, P. J.; MOORE, T. D.; NEWLANDER, K. A.; PEARSON, S.; POLIZZI, B. J.; QIU, X.; RITTENHOUSE, S. F.;

SLATER-RADOSTI, C.; SALVERS, K. L.; SEEFELD, M. A.; SMYTH, M. G.; TAKATA, D. T.; UZINSKAS, I. N.; VAIDYA, K.; WALLIS, N. G.; WINRAM, S. B.; YUAN, C. C. K.; HUFFMAN, W. F. Discovery of a novel and potent class of fabI-directed antibacterial agents. **Antimicrobial Agents and Chemotherapy**, v. 46, n. 10, p. 3118–3124, out. 2002.

PAYNE, D. J.; WARREN, P. V.; HOLMES, D. J.; JI, Y.; LONSDALE, J. T. Bacterial fatty-acid biosynthesis: a genomics-driven target for antibacterial drug discovery. **Drug Discovery Today**, v. 6, n. 10, p. 537–544, maio 2001.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL V. AND THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER P. AND WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, n. 1, p. 2825–2830, out. 2011.

PIRHADI, S.; GHASEMI, J. B. 3D-QSAR analysis of human immunodeficiency virus entry-1 inhibitors by CoMFA and CoMSIA. **European Journal of Medicinal Chemistry**, v. 45, n. 11, p. 4897–4903, nov. 2010.

PRIMI, M. C.; MALTAROLLO, V. G.; MAGALHÃES, J. G.; MALTA DE SÁ, M.; RANGEL-YAGUI, C. O.; TROSSINI, G. H. G. Convergent QSAR studies on a series of NK 3 receptor antagonists for schizophrenia treatment. **Journal of Enzyme Inhibition and Medicinal Chemistry**, v. 31, n. 2, p. 283–294, mar. 2016.

PRIYADARSHI, A.; KIM, E. E.; HWANG, K. Y. Structural insights into *Staphylococcus aureus* enoyl-ACP reductase (FabI), in complex with NADP and triclosan. **Proteins: Structure, Function, and Bioinformatics**, v. 78, n. 2, p. 480–486, fev. 2010.

PROMEGA CORPORATION. **PureYield(TM) Plasmid Miniprep System Technical Bulletin TB374**. Madison: Promega, 2009. 11 p. Disponível em: <<https://www.promega.com.br/-/media/files/resources/protocols/technical-bulletins/101/pureyield-plasmid-miniprep-system->

protocol.pdf?rev=dcfef17bf1ff4c5d8d6e59d88e1bcc31&sc_lang=en>. Acesso em: 19 set. 2023.

RASTELLI, G.; DEGLIESPOSTI, G.; DEL RIO, A.; SGOBBA, M. Binding Estimation after Refinement, a New Automated Procedure for the Refinement and Rescoring of Docked Ligands in Virtual Screening. **Chemical Biology & Drug Design**, v. 73, n. 3, p. 283–286, mar. 2009.

RICHARDSON, L. **Beautiful soup documentation**. 2007. Disponível em: <<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>>. Acesso em: 19 set. 2023.

ROY, K.; KAR, S.; DAS, R. N. **A Primer on QSAR/QSPR Modeling: Fundamental Concepts**. 1. ed. Cham: Springer International Publishing, 2015. 121 p.

SAHIGARA, F.; MANSOURI, K.; BALLABIO, D.; MAURI, A.; CONSONNI, V.; TODESCHINI, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. **Molecules**, v. 17, n. 5, p. 4791–4810, abr. 2012.

SALAHINEJAD, M.; GHASEMI, J. B. 3D-QSAR studies on the toxicity of substituted benzenes to *Tetrahymena pyriformis*: CoMFA, CoMSIA and VolSurf approaches. **Ecotoxicology and Environmental Safety**, v. 105, n. 1, p. 128–134, jul. 2014.

SAROWSKA, J.; FUTOMA-KOLOCH, B.; JAMA-KMIECIK, A.; FREJ-MADRZAK, M.; KSIAZCZYK, M.; BUGLA-PLOSKONSKA, G.; CHOROSZY-KROL, I. Virulence factors, prevalence and potential transmission of extraintestinal pathogenic *Escherichia coli* isolated from different sources: Recent reports. **Gut Pathogens**, v. 11, n. 1, p. 1–16, fev. 2019.

SASTRY, G. M.; ADZHIGIREY, M.; DAY, T.; ANNABHIMOJU, R.; SHERMAN, W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. **Journal of Computer-Aided Molecular Design**, v. 27, n. 3, p. 221–234, abr. 2013.

SCHIEBEL, J.; CHANG, A.; LU, H.; BAXTER, M. V.; TONGE, P. J.; KISKER, C. Staphylococcus aureus FabI: Inhibition, Substrate Recognition, and Potential Implications for In Vivo Essentiality. **Structure**, v. 20, n. 5, p. 802–813, maio 2012.

SCHIEBEL, J.; CHANG, A.; MERGET, B.; BOMMINENI, G. R.; YU, W.; SPAGNUOLO, L. A.; BAXTER, M. V.; TAREILUS, M.; TONGE, P. J.; KISKER, C.; SOTRIFFER, C. A. An Ordered Water Channel in Staphylococcus aureus FabI: Unraveling the Mechanism of Substrate Recognition and Reduction. **Biochemistry**, v. 54, n. 10, p. 1943–1955, mar. 2015.

SCHIRÒ, A.; CARLON, A.; PARIGI, G.; MURSHUDOV, G.; CALDERONE, V.; RAVERA, E.; LUCHINAT, C. On the complementarity of X-ray and NMR data. **Journal of Structural Biology: X**, v. 4, n. 1, p. 1–9, jan. 2020.

SCHRÖDINGER LLC. **The PyMOL Molecular Graphics System, Version 2.0**. 2015. Disponível em: <<https://pymol.org/>>. Acesso em: 23 set. 2023.

SCHRODINGER RELEASE 2022-3. **LigPrep**. 2022. Disponível em: <<https://www.schrodinger.com/products/ligprep>>. Acesso em: 1 mar. 2023.

SCHRÖDINGER RELEASE 2022-3. **Glide**. 2022. Disponível em: <<https://www.schrodinger.com/products/glide>>. Acesso em: 1 mar. 2023.

SCHRÖDINGER RELEASE 2022-3. **Protein Preparation Wizard**. New York: Schrödinger, 2022. Disponível em: <<https://www.schrodinger.com/science-articles/protein-preparation-wizard>>. Acesso em: 1 mar. 2023.

SERRANO, A.; IMBERNÓN, B.; PÉREZ-SÁNCHEZ, H.; CECILIA, J. M.; BUENOCRESPO, A.; ABELLÁN, J. L. Accelerating Drugs Discovery with Deep Reinforcement Learning: An Early Approach. In: International Conference on Parallel Processing (ICPP), 47, 2018. **Proceedings**, n. 8, 2018. p. 1–8.

SHAHLAEI, M.; SABET, R.; ZIARI, M. B.; MOEINIFARD, B.; FASSIHI, A.; KARBAKHSR, R. QSAR study of anthranilic acid sulfonamides as inhibitors of

methionine aminopeptidase-2 using LS-SVM and GRNN based on principal components. **European Journal of Medicinal Chemistry**, v. 45, n. 10, p. 4499–4508, out. 2010.

SHELLEY, J. C.; CHOLLETI, A.; FRYE, L. L.; GREENWOOD, J. R.; TIMLIN, M. R.; UCHIMAYA, M. Epik: A software program for pKa prediction and protonation state generation for drug-like molecules. **Journal of Computer-Aided Molecular Design**, v. 21, n. 12, p. 681–691, set. 2007.

SHEN, C.; DING, J.; WANG, Z.; CAO, D.; DING, X.; HOU, T. From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. **Wiley Interdisciplinary Reviews: Computational Molecular Science**, v. 10, n. 1, p. 1–23, jan. 2019.

STERLING, T.; IRWIN, J. J. ZINC 15 – Ligand Discovery for Everyone. **Journal of Chemical Information and Modeling**, v. 55, n. 11, p. 2324–2337, nov. 2015.

SWISS INSTITUTE OF BIOINFORMATICS. **ExpASy-ProtParam tool**. Lausanne: Swiss Institute of Bioinformatics, 2017.

SZWEDA, P.; SCHIELMANN, M.; KOTLOWSKI, R.; GORCZYCA, G.; ZALEWSKA, M.; MILEWSKI, S. Peptidoglycan hydrolases-potential weapons against *Staphylococcus aureus*. **Applied Microbiology and Biotechnology**, v. 96, n. 5, p. 1157–1174, out. 2012.

TAKHI, M.; SREENIVAS, K.; REDDY, C. K.; MUNIKUMAR, M.; PRAVEENA, K.; SUDHEER, P.; RAO, B. N. V. M.; RAMAKANTH, G.; SIVARANJANI, J.; MULIK, S.; REDDY, Y. R.; NARASIMHA RAO, K.; PALLAVI, R.; LAKSHMINARASIMHAN, A.; PANIGRAHI, S. K.; ANTONY, T.; ABDULLAH, I.; LEE, Y. K.; RAMACHANDRA, M.; YUSOF, R.; RAHMAN, N. A.; SUBRAMANYA, H. Discovery of azetidine based enamide amides as potent bacterial enoyl ACP reductase (FabI) inhibitors. **European Journal of Medicinal Chemistry**, v. 84, n. 1, p. 382–394, set. 2014.

TANG, S.; CHEN, L. iATC-NFMLP: Identifying Classes of Anatomical Therapeutic Chemicals Based on Drug Networks, Fingerprints, and Multilayer Perceptron. **Current Bioinformatics**, v. 17, n. 9, p. 814–824, mar. 2022.

TANIMOTO, T. T. Elementary mathematical theory of classification and prediction. **International Business Machines Corporation**, p. 1–11, nov. 1958.

TERPE, K. Overview of bacterial expression systems for heterologous protein production: From molecular and biochemical fundamentals to commercial systems. **Applied Microbiology and Biotechnology**, v. 72, n. 2, p. 211–222, set. 2006.

THESNAAR, L.; BEZUIDENHOUT, J. J.; PETZER, A.; PETZER, J. P.; CLOETE, T. T. Methylene blue analogues: In vitro antimicrobial minimum inhibitory concentrations and in silico pharmacophore modelling. **European Journal of Pharmaceutical Sciences**, v. 157, n. 1, p. 1–10, fev. 2021.

TIMMINS, G. S.; DERETIC, V. Mechanisms of action of isoniazid. **Molecular Microbiology**, v. 62, n. 5, p. 1220–1227, dez. 2006.

TONG, J. B.; LUO, D.; FENG, Y.; BIAN, S.; ZHANG, X.; WANG, T. H. Structural modification of 4, 5-dihydro-[1, 2, 4] triazolo [4, 3-f] pteridine derivatives as BRD4 inhibitors using 2D/3D-QSAR and molecular docking analysis. **Molecular Diversity**, v. 25, n. 3, p. 1855–1872, ago. 2021.

TORRES, P. H. M.; SODERO, A. C. R.; JOFILY, P.; SILVA-JR, F. P. Key Topics in Molecular Docking for Drug Design. **International Journal of Molecular Sciences**, v. 20, n. 18, p. 1–29, set. 2019.

TROPSHA, A.; GOLBRAIKH, A.; CHO, W. J. Development of kNN QSAR Models for 3-Arylisoquinoline Antitumor Agents. **Bulletin of the Korean Chemical Society**, v. 32, n. 7, p. 2397–2404, jul. 2011.

TRUCHON, J. F.; BAYLY, C. I. Evaluating virtual screening methods: Good and bad metrics for the “early recognition” problem. **Journal of Chemical Information and Modeling**, v. 47, n. 2, p. 488–508, fev. 2007.

TUNGEKAR, A. A.; CASTILLO-CORUJO, A.; RUDDOCK, L. W. So you want to express your protein in *Escherichia coli*? **Essays in Biochemistry**, v. 65, n. 2, p. 247–260, jul. 2021.

U.S. FOOD & DRUG ADMINISTRATION. **FDA issues final rule on safety and effectiveness of antibacterial soaps**. 2016. Disponível em: <<https://www.fda.gov/news-events/press-announcements/fda-issues-final-rule-safety-and-effectiveness-antibacterial-soaps>>. Acesso em: 11 set. 2023.

VAN BOECKEL, T. P.; PIRES, J.; SILVESTER, R.; ZHAO, C.; SONG, J.; CRISCUOLO, N. G.; GILBERT, M.; BONHOEFFER, S.; LAXMINARAYAN, R. Global trends in antimicrobial resistance in animals in low- And middle-income countries. **Science**, v. 365, n. 6459, p. 1–6, set. 2019.

VELÁZQUEZ-LIBERA, J. L.; DURÁN-VERDUGO, F.; VALDÉS-JIMÉNEZ, A.; VALDÉS-JIMÉNEZ, A.; NÚÑEZ-VIVANCO, G.; CABALLERO, J. LigRMSD: a web server for automatic structure matching and RMSD calculations among identical and similar compounds in protein-ligand docking. **Bioinformatics**, v. 36, n. 9, p. 2912–2914, maio 2020.

VENKATRAMAN, V.; CHAKRAVARTHY, P. R.; KIHARA, D. Application of 3D zernike descriptors to shape-based ligand similarity searching. **Journal of Cheminformatics**, v. 1, n. 1, p. 1–19, 17 dez. 2009.

VERÍSSIMO, G. C. **MASSA Algorithm: Molecular data set sampling for training-test separation**. 2021. Disponível em: <https://github.com/gcverissimo/MASSA_Algorithm>. Acesso em: 12 set. 2023.

VERÍSSIMO, G. C.; DOS SANTOS JÚNIOR, V. S.; DE ALMEIDA, I. A. do R.; RUAS, M. S. M.; COUTINHO, L. G.; DE OLIVEIRA, R. B.; ALVES, R. J.; MALTAROLLO, V. G.

The Brazilian compound library (BraCoLi) database: a repository of chemical and biological information for drug design. **Molecular Diversity**, v. 26, n. 6, p. 3387–3397, dez. 2022a.

VERÍSSIMO, G. C.; DOS SANTOS JUNIOR, V. S.; FERNANDES, P. O.; ISHIDA, S.; KOJIMA, R.; OKUNO, Y.; GERTRUDES, J. C.; MALTAROLLO, V. G. GCN-Based Structure-Activity Relationship and DFT Studies of Staphylococcus aureus FabI Inhibitors. **International Journal of Quantitative Structure-Property Relationships**, v. 7, n. 1, p. 1–16, 4 nov. 2022b.

VERÍSSIMO, G. C.; GERTRUDES, J. de C.; MALTAROLLO, V. G. Machine learning methods in drug design. In: ROY, K. **Cheminformatics, QSAR and Machine Learning Applications for Novel Drug Development**. 1. ed. Cambridge: Academic Press, 2023. p. 329–360.

VERÍSSIMO, G. C.; PANTELEÃO, S. Q.; GERTRUDES, J. C.; FERNANDES, P. O.; KRONENBERGER, T.; HONÓRIO, K. M.; MALTAROLLO, V. G. MASSA Algorithm: automated rational sampling of training and test subsets for QSAR modelling. **Journal of Computer-Aided Molecular Design**, p. 1–20, out. 2023.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; OTHERS. SciPy 1.0: fundamental algorithms for scientific computing in Python. **Nature methods**, v. 17, n. 3, p. 261–272, fev. 2020.

WANG, C.; LIU, S.; FENG, H.; BARRETT, H.; PENG, H.; KARUNARATNE, S. H. P. P.; ZHANG, Y.; YANG, M. Effects of Triclosan on the Development of Antimicrobial Resistance in the Environment: A Review. **Current Pollution Reports 2023**, v. 9, n. 1, p. 454–467, jun. 2023.

WANG, C.; ZHANG, Y. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. **Journal of Computational Chemistry**, v. 38, n. 3, p. 169–177, jan. 2017.

WANG, R.; LIU, L.; LAI, L.; TANG, Y. SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. **Journal of Molecular Modeling**, v. 4, n. 12, p. 379–394, dez. 1998.

WARD, W. H. J.; HOLDGATE, G. A.; ROWSELL, S.; MCLEAN, E. G.; PAUPTIT, R. A.; CLAYTON, E.; NICHOLS, W. W.; COLLS, J. G.; MINSHULL, C. A.; JUDE, D. A.; MISTRY, A.; TIMMS, D.; CAMBLE, R.; HALES, N. J.; BRITTON, C. J.; TAYLOR, I. W. F. Kinetic and Structural Characteristics of the Inhibition of Enoyl (Acyl Carrier Protein) Reductase by Triclosan. **Biochemistry**, v. 38, n. 38, p. 12514–12525, set. 1999.

WIERBOWSKI, S. D.; WINGERT, B. M.; ZHENG, J.; CAMACHO, C. J. Cross-docking benchmark for automated pose and ranking prediction of ligand binding. **Protein Science**, v. 29, n. 1, p. 298–305, jan. 2020.

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

WILSON, G. L.; LILL, M. A. Integrating structure-based and ligand-based approaches for computational drug design. **Future Medicinal Chemistry**, v. 3, n. 6, p. 735–750, maio 2011.

YANG, X.; LU, J.; YING, M.; MU, J.; LI, P.; LIU, Y. Docking and molecular dynamics studies on triclosan derivatives binding to FabI. **Journal of Molecular Modeling**, v. 23, n. 1, p. 1–13, jan. 2017.

YANG, X.; WANG, Y.; BYRNE, R.; SCHNEIDER, G.; YANG, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. **Chemical Reviews**, v. 119, n. 18, p. 10520–10594, set. 2019.

YANG, X.; ZHENG, Y.; XING, X.; SUI, X.; JIA, W.; PAN, H. Immune subtype identification and multi-layer perceptron classifier construction for breast cancer. **Frontiers in Oncology**, v. 12, n. 1, p. 1–12, dez. 2022.

YAO, J.; BRUHN, D. F.; FRANK, M. W.; LEE, R. E.; ROCK, C. O. Activation of exogenous fatty acids to acyl-acyl carrier protein cannot bypass FabI inhibition in *Neisseria*. **Journal of Biological Chemistry**, v. 291, n. 1, p. 171–181, jan. 2016.

YAO, J.; ROCK, C. O. How Bacterial Pathogens Eat Host Lipids: Implications for the Development of Fatty Acid Synthesis Therapeutics. **The Journal of Biological Chemistry**, v. 290, n. 10, p. 5940–5946, 3 mar. 2015.

YAO, J.; ROCK, C. O. Exogenous fatty acid metabolism in bacteria. **Biochimie**, v. 141, n. 1, p. 30–39, out. 2017.

YOO, C. K.; SHAHLAEI, M. The applications of PCA in QSAR studies: A case study on CCR5 antagonists. **Chemical Biology & Drug Design**, v. 91, n. 1, p. 137–152, jan. 2018.

YOUNG, C. L.; BRITTON, Z. T.; ROBINSON, A. S. Recombinant protein expression and purification: A comprehensive review of affinity tags and microbial applications. **Biotechnology Journal**, v. 7, n. 5, p. 620–634, maio 2012.

YOUSAF, A.; SHEHZADI, T.; FAROOQ, A.; ILYAS, K. Protein active site prediction for early drug discovery and designing. **International Review of Applied Sciences and Engineering**, v. 13, n. 1, p. 98–105, set. 2021.

YUSUF, D.; DAVIS, A. M.; KLEYWEGT, G. J.; SCHMITT, S. An alternative method for the evaluation of docking performance: RSR vs RMSD. **Journal of Chemical Information and Modeling**, v. 48, n. 7, p. 1411–1422, jul. 2008.

ZAINAB, S. M.; JUNAID, M.; XU, N.; MALIK, R. N. Antibiotics and antibiotic resistant genes (ARGs) in groundwater: A global review on dissemination, sources, interactions, environmental and human health risks. **Water Research**, v. 187, n. 1, p. 1–17, dez. 2020.

ZHANG, H.; LIU, C. T.; MAO, J.; SHEN, C.; XIE, R. L.; MU, B. Development of novel in silico prediction model for drug-induced ototoxicity by using naïve Bayes classifier approach. **Toxicology in Vitro**, v. 65, n. 1, p. 1–11, jun. 2020.

ZHANG, H.; REN, J. X.; MA, J. X.; DING, L. Development of an in silico prediction model for chemical-induced urinary tract toxicity by using naïve Bayes classifier. **Molecular Diversity**, v. 23, n. 2, p. 381–392, maio 2019.

ZHENG, C. J.; SOHN, M. J.; LEE, S.; KIM, W. G. Meleagrins, a New FabI Inhibitor from *Penicillium chrysogenum* with at Least One Additional Mode of Action. **PLOS ONE**, v. 8, n. 11, p. 1–9, nov. 2013.

ZHU, Q. On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. **Pattern Recognition Letters**, v. 136, n. 1, p. 71–80, ago. 2020.

ZILIAN, D.; SOTRIFER, C. A. SFCscoreRF: A random forest-based scoring function for improved affinity prediction of protein-ligand complexes. **Journal of Chemical Information and Modeling**, v. 53, n. 8, p. 1923–1933, ago. 2013.

APÊNDICES

APÊNDICE A – SEQUÊNCIAS DE NUCLEOTÍDEOS DAS PROTEÍNAS RECOMBINANTES E AS SEQUÊNCIAS DE AMINOÁCIDOS RESULTANTES

Figura A.1 – Sequência codificante 5' -> 3' de nucleotídeos da saFabI e a sequência de aminoácidos resultante.

5'	Ⓟ ATG GGC AGC AGC CAT CAT CAT CAT CAC AGC AGC GGC CTG GTG CCG CGC GGC AGC CAT	60
	M G S S H H H H H S S G L V P R G S H →	
	ATG GCT AGC TTA AAT CTT GAA AAC AAA ACA TAT GTC ATC ATG GGA ATC GCT AAT AAG CGT	120
	M A S L N L E N K T Y V I M G I A N K R →	
	AGT ATT GCT TTT GGT GTC GCT AAA GTT TTA GAT CAA TTA GGT GCT AAA TTA GTA TTT ACT	180
	S I A F G V A K V L D Q L G A K L V F T →	
	TAC CGT AAA GAA CGT AGC CGT AAA GAG CTT GAA AAA TTA TTA GAA CAA TTA AAT CAA CCA	240
	Y R K E R S R K E L E K L L E Q L N Q P →	
	GAA GCG CAC TTA TAT CAA ATT GAT GTT CAA AGC GAT GAA GAG GTT ATT AAT GGT TTT GAG	300
	E A H L Y Q I D V Q S D E E V I N G F E →	
	CAA ATT GGT AAA GAT GTT GGC AAT ATT GAT GGT GTA TAT CAT TCA ATC GCA TTT GCT AAT	360
	Q I G K D V G N I D G V Y H S I A F A N →	
	ATG GAA GAC TTA CCG GGA CGC TTT TCT GAA ACT TCA CGT GAA GGC TTC TTG TTA GCT CAA	420
	M E D L R G R F S E T S R E G F L L A Q →	
	GAC ATT AGT TCT TAC TCA TTA ACA ATT GTG GCT CAT GAA GCT AAA AAA TTA ATG CCA GAA	480
	D I S S Y S L T I V A H E A K K L M P E →	
	GGT GGT AGC ATT GTT GCA ACA ACA TAT TTA GGT GGC GAA TTC GCA GTT CAA AAT TAT AAT	540
	G G S I V A T T Y L G G E F A V Q N Y N →	
	GTG ATG GGT GTT GCT AAA GCG AGC TTA GAA GCA AAT GTT AAA TAT TTA GCA TTA GAC TTA	600
	V M G V A K A S L E A N V K Y L A L D L →	
	GGT CCT GAT AAT ATT CCG GTT AAT GCA ATT TCA GCT GGT CCA ATC CGT ACA TTA AGT GCA	660
	G P D N I R V N A I S A G P I R T L S A →	
	AAA GGT GTG GGT GGT TTC AAT ACA ATT CTT AAA GAA ATC GAA GAG CGT GCA CCT TTA AAA	720
	K G V G G F N T I L K E I E E R A P L K →	
	CGT AAC GTT GAT CAA GTA GAA GTA GGT AAA ACA GCG GCT TAC TTA TTA AGT GAC TTA TCA	780
	R N V D Q V E V G K T A A Y L L S D L S →	
	AGT GGC GTT ACA GGT GAA AAT ATT CAT GTA GAT AGC GGA TTC CAC GCA ATT AAA TAA	837
	S G V T G E N I H V D S G F H A I K Ⓟ →	

Figura A.2 – Sequência codificante 5' -> 3' de nucleotídeos da ecFabI e a sequência de aminoácidos resultante.

5'	ATG CAC CAT CAC CAC CAC CAT GAA AAT CTA TAT TTC CAG GGT ATG GGC TTC CTG AGC GGT	60
	M H H H H H H E N L Y F Q G M G F L S G →	
	AAG CGC ATC CTT GTT ACC GGT GTT GCG TCC AAG CTC TCT ATC GCC TAC GGC ATT GCG CAG	120
	K R I L V T G V A S K L S I A Y G I A Q →	
	GCA ATG CAT CGT GAG GGC GCG GAG CTG GCG TTC ACC TAT CAG AAT GAT AAA TTG AAG GGT	180
	A M H R E G A E L A F T Y Q N D K L K G →	
	CGT GTA GAG GAA TTT GCG GCT CAA TTG GGC TCT GAC ATC GTG CTG CAA TGC GAC GTG GCC	240
	R V E E F A A Q L G S D I V L Q C D V A →	
	GAG GAC GCA TCG ATT GAC ACC ATG TTT GCG GAG CTG GGT AAA GTT TGG CCA AAG TTT GAT	300
	E D A S I D T M F A E L G K V W P K F D →	
	GGC TTC GTG CAC AGC ATC GGC TTC GCT CCG GGT GAC CAA CTG GAC GGC GAC TAC GTG AAC	360
	G F V H S I G F A P G D Q L D G D Y V N →	
	GCT GTG ACG CGT GAG GGC TTC AAA ATC GCG CAC GAT ATC AGC TCC TAC AGC TTT GTT GCT	420
	A V T R E G F K I A H D I S S Y S F V A →	
	ATG GCC AAG GCC TGC CGT AGC ATG CTG AAT CCG GGT TCT GCT CTG CTG ACC CTG AGC TAC	480
	M A K A C R S M L N P G S A L L T L S Y →	
	TTG GGT GCG GAG CGC GCA ATT CCG AAT TAT AAC GTG ATG GGT CTG GCG AAG GCT AGT CTG	540
	L G A E R A I P N Y N V M G L A K A S L →	
	GAA GCA AAC GTC CGT TAT ATG GCG AAC GCG ATG GGT CCG GAA GGC GTT CGT GTC AAC GCC	600
	E A N V R Y M A N A M G P E G V R V N A →	
	ATC TCA GCG GGT CCG ATC CGC ACC TTG GCG GCT TCG GGC ATT AAA GAT TTT CGT AAA ATG	660
	I S A G P I R T L A A S G I K D F R K M →	
	TTA GCG CAC TGT GAA GCG GTG ACT CCG ATC CGC AGA ACC GTG ACG ATT GAA GAT GTC GGA	720
	L A H C E A V T P I R R T V T I E D V G →	
	AAC AGC GCG GCA TTT CTG TGC AGC GAC CTG AGC GCG GGT ATT TCC GGT GAA GTT GTT CAT	780
	N S A A F L C S D L S A G I S G E V V H →	
	GTT GAT GGC GGT TTC AGC ATT GCA GCG ATG AAT GAA TTG GAG CTC AAA TGA	3' 831
	V D G G F S I A A M N E L E L K * →	

APÊNDICE B – PROCEDIMENTO OPERACIONAL PADRÃO: METODOLOGIA PARA COLORAÇÃO DE GEL SDS-PAGE POR NITRATO DE PRATA

FIXAÇÃO DO GEL

- » Prepare 100 mL (quantidade suficiente para 2 géis) de solução fixadora com:
 - 50% (v/v) de Metanol
 - 10% (v/v) de Ácido Acético
 - 200 uL de Formaldeído
- » Deixe o gel na solução fixadora overnight ou, no mínimo, por 1 hora.

COLORAÇÃO DO GEL

- » Prepare a solução de Etanol 50% (v/v).
- » Prepare a solução de **Tiosulfato de Sódio (0,02% m/v). 20 mg em 100 mL.**
- » **Guarde 2 mL dessa solução.**
- » Prepare a solução de **Nitrato de Prata (0,2% m/v). 200 mg em 100 mL.**

OBS.: Utilize sempre água MiliQ ou da melhor qualidade possível no procedimento.

OBS.: Utilize de cada solução aproximadamente 50mL para cada gel (para cobrir).

1. Com solução de etanol 50%, lave o gel 3 vezes por vinte minutos a cada lavagem utilizando 50 mL em cada.
2. Deixe o gel 1 minuto agitando com solução de Tiosulfato de Sódio.
3. Lave 3 vezes com água por 20 segundos cada lavagem.
4. Deixe na solução de Nitrato de Prata por 30 minutos.

- » Enquanto isso, prepare a solução de Revelação em 100 mL (suficiente para 2 géis):
 - 6g de Carbonato de Sódio
 - 2 mL da solução de Tiosulfato de Sódio
 - 200uL de Formaldeído

5. Após o banho em nitrato de prata, lave 3 vezes com água por 20 segundos cada lavagem.
6. Deixe em solução de Revelação até o aparecimento das bandas.
7. STOP da revelação com a adição de 5mL de ácido acético para cada 50mL de Solução de Revelação. Armazene o gel em água.