UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA BIOMÉDICA

LUCAS VINÍCIUS SILVA DE ALBUQUERQUE

Desenvolvimento de uma abordagem híbrida inteligente para estimação de docking molecular entre proteínas utilizando redes de pseudo-convolução e Random Forests

Recife

2025

LUCAS VINÍCIUS SILVA DE ALBUQUERQUE

Desenvolvimento de uma abordagem híbrida inteligente para estimação de docking molecular entre proteínas utilizando redes de pseudo-convolução e Random Forests

Trabalho apresentado ao Programa de Pós-graduação em Engenharia Biomédica do Centro de Tecnologia e Geociências da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre em Engenharia Biomédica.

Área de Concentração: Computação Biomédica

Orientador: Wellington Pinheiro dos Santos

Coorientador: Luiz Alberto Lira Soares

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Albuquerque, Lucas Vinicius Silva de.
Desenvolvimento de uma abordagem híbrida inteligente para estimação de docking molecular entre proteínas utilizando redes de pseudo-convolução e Random Forests / Lucas Vinicius Silva de Albuquerque. - Recife, 2025.
70f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, Programa de Pós-Graduação em Engenharia Biomédica, 2025.
Orientação: Wellington Pinheiro dos Santos.
Coorientação: Luiz Alberto Lira Soares.

1. Docking molecular; 2. Interações proteína-proteína; 3. Redes de pseudo-convolução; 4. Random Forests; 5. Descoberta de medicamentos; 6. Afinidade de ligação. I. Santos, Wellington Pinheiro dos. II. Soares, Luiz Alberto Lira. III. Título.

UFPE-Biblioteca Central

UNIVERSIDADE FEDERAL DE PERNAMBUCO

Ata da defesa/apresentação do Trabalho de Conclusão de Curso de Mestrado do Programa de Pós-graduação em Engenharia Biomédica - CTG da Universidade Federal de Pernambuco, no dia 24 de março de 2025.

ATA Nº 98

Aos 24 dias do mês de março de 2025, às 14:00, em sessão pública realizada de forma remota, teve início a defesa/apresentação do Trabalho de Conclusão de Curso intitulada Desenvolvimento de uma abordagem híbrida inteligente para estimação de docking molecular entre proteínas utilizando redes de pseudo-convolução e Random Forests do(a) mestrando(a) LUCAS VINICIUS SILVA DE ALBUQUERQUE, na área de concentração Computação Biomédica, sob a orientação do(a) Prof.(a) WELLINGTON PINHEIRO DOS SANTOS e co-orientação do(a) Prof.(a) LUIZ ALBERTO LIRA SOARES. A Comissão Examinadora foi aprovada pelo colegiado do programa de pós-graduação em 18/03/2025, sendo composta pelos examinadores: WELLINGTON PINHEIRO DOS SANTOS, do(a) UFPE; MAGDA RHAYANNY ASSUNCAO FERREIRA, do(a) UFPE; CLARISSE LINS DE LIMA, do(a) UFPE. Após cumpridas as formalidades conduzidas pelo(a) presidente(a) da comissão, professor(a) WELLINGTON PINHEIRO DOS SANTOS, o(a) candidato(a) ao grau de Mestre(a) foi convidado(a) a discorrer sobre o conteúdo do Trabalho de Conclusão de Curso. Concluída a explanação, o(a) candidato(a) foi arguido(a) pela Comissão Examinadora que, em seguida, reuniu-se para deliberar e conceder, ao mesmo, a menção APROVADO. Para a obtenção do grau de Mestre(a) em Engenharia Biomédica, o(a) concluinte deverá ter atendido todas às demais exigências estabelecidas no Regimento Interno e Normativas Internas do Programa, nas Resoluções e Portarias dos Órgãos Deliberativos Superiores, assim como no Estatuto e no Regimento Geral da Universidade, observando os prazos e procedimentos vigentes nas normas.

Dra. CLARISSE LINS DE LIMA, UFPE

Examinadora Externa à Instituição

Dra. MAGDA RHAYANNY ASSUNCAO FERREIRA, UFPE

Examinadora Externa ao Programa

Dr. WELLINGTON PINHEIRO DOS SANTOS, UFPE

Examinador Interno

LUCAS VINICIUS SILVA DE ALBUQUERQUE

Mestrando(a)

RESUMO

Docking molecular é uma técnica computacional utilizada para prever como duas moléculas, geralmente uma proteína e um ligante, interagem entre si. Essa técnica simula o encaixe da molécula menor (ligante) no sítio ativo da molécula maior (proteína), permitindo a análise de afinidade e especificidade das interações. Essa abordagem é fundamental para a descoberta de novos medicamentos, pois auxilia na identificação de possíveis candidatos a fármacos e na compreensão dos mecanismos moleculares subjacentes a diversas doenças. Além disso, técnicas avançadas de Inteligência Artificial (IA) e aprendizado de máquina têm aprimorado a precisão e a eficiência dessas previsões, reduzindo custos e tempo no desenvolvimento de novos tratamentos. Neste trabalho, foi desenvolvida uma abordagem híbrida inteligente para a estimativa do encaixe molecular entre proteínas, integrando redes de pseudo-convolução e algoritmos de *Random Forests*. O objetivo foi melhorar a precisão na previsão da afinidade de ligação entre proteínas por meio de uma estratégia baseada em aprendizado de máquina. As redes de pseudo-convolução foram empregadas para processar sequências de aminoácidos das proteínas candidatas, fragmentando-as em segmentos menores e capturando informações estruturais relevantes. Posteriormente, os vetores resultantes foram classificados por modelos de Random Forests. A metodologia foi avaliada por meio de experimentos comparativos com abordagens tradicionais de *docking* molecular, explorando também a capacidade de generalização do modelo a diferentes tipos de proteínas e interações moleculares. Os resultados demonstraram avanços significativos, com destaque para a redução de 8113 para 11 atributos, o que aumentou a eficiência computacional sem prejuízo da acurácia. O modelo de Random Forest com 200 árvores obteve acurácia de 99,8%, índice Kappa de 0,997, sensibilidade de 0,997, especificidade de 1,000 e AUC de 1,000, evidenciando alto desempenho e contribuições relevantes para a descoberta computacional de medicamentos. A conclusão deste trabalho evidencia que a abordagem proposta, baseada em redes de pseudo-convolução e algoritmos Random Forest, obteve desempenho elevado na tarefa de predição de docking molecular, com acurácia de 99,8% e índice Kappa de 0,997. Também é destacado que a redução de atributos, de 8113 para apenas 11, possibilitou uma significativa diminuição no tempo de treinamento dos modelos, mantendo a alta performance. Por fim, são sugeridas aplicações futuras da metodologia em diferentes bases de dados e cenários de interação molecular.

Palavras-chaves: Docking molecular. Interações proteína-proteína. Redes de pseudo-convolução.

Random Forests. Descoberta de medicamentos. Afinidade de ligação.

ABSTRACT

Molecular docking is a computational technique used to predict how two molecules, usually a protein and a ligand, interact with each other. This technique simulates the fitting of the smaller molecule (ligand) into the active site of the larger molecule (protein), enabling the analysis of binding affinity and interaction specificity. This approach is essential for drug discovery, as it assists in identifying potential drug candidates and understanding the molecular mechanisms underlying various diseases. Furthermore, advanced Artificial Intelligence (AI) and machine learning techniques have enhanced the accuracy and efficiency of such predictions, reducing the costs and time involved in drug development. In this work, an intelligent hybrid approach was developed to estimate molecular docking between proteins, integrating pseudo-convolutional networks and Random Forests. The aim was to improve the accuracy of predicting protein binding affinity through a machine learning strategy. Pseudo-convolutional networks were used to process amino acid sequences of candidate proteins, segmenting them into smaller fragments and extracting structural features. Then, the resulting vectors were classified using Random Forest models. The methodology was evaluated through experiments comparing its performance with traditional molecular docking techniques, also exploring the model's ability to generalize across different types of proteins and molecular interactions. The results demonstrated significant advances, including a reduction from 8113 to 11 attributes, which increased computational efficiency without compromising accuracy. The Random Forest model with 200 trees achieved 99.8% accuracy, a Kappa index of 0.997, sensitivity of 0.997, specificity of 1.000, and AUC of 1.000, indicating high performance and promising contributions to computational drug discovery. The conclusion highlights that the proposed approach achieved outstanding performance in docking prediction, with significant reduction in training time due to dimensionality reduction. The study also suggests future applications of the methodology in different datasets and molecular interaction contexts.

Keywords: Molecular docking. Protein-protein interactions. Pseudo-convolution networks. Random Forests. Drug discovery. Binding affinity.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nota: ao longo deste texto, manter-se-ão em itálico alguns termos estrangeiros consagrados na literatura científica por não possuírem equivalentes precisos em português. Entre eles, destacam-se: docking, que se refere ao processo computacional de simulação do encaixe entre moléculas; e Random Forests, um algoritmo de aprendizado de máquina baseado em múltiplas árvores de decisão.

1.1 MOTIVAÇÃO E JUSTIFICATIVA

A Quarta Revolução Industrial, caracterizada pela integração de tecnologias avançadas, tem o potencial de impactar significativamente o desenvolvimento de nações no Sul Global, especialmente nas áreas de Engenharia Biomédica, saúde digital e indústria farmacêutica (JAYANTHI et al., 2020; BALI; BALI, 2020; KIM; HAN, 2020; MELO; ARAúJO, 2020; AJMERA; JAIN, 2019; XIONG, 2021; PANG et al., 2018; MAHOMED, 2018; CELESTI; AMFT; VILLARI, 2019). Essas tecnologias, como Inteligência Artificial (IA), Internet das Coisas (IoT) e análise de grandes dados, oferecem oportunidades únicas para melhorar o acesso, entrega e resultados dos cuidados de saúde em cenários de recursos limitados (KAUR; GARG; GUPTA, 2021; THAKARE; KHIRE; KUMBHAR, 2022; SHAFQAT et al., 2020; TURCU; TURCU, 2013; POYNER; SHERRATT, 2019; BANERJEE et al., 2020). Na Engenharia Biomédica, esses avanços podem aprimorar o design e desenvolvimento de dispositivos médicos, diagnósticos e próteses, oferecendo soluções acessíveis e personalizadas para atender necessidades específicas de saúde (AGUADO et al., 2018; PRIMICERI et al., 2018; GHOSH et al., 2018). Plataformas de saúde digital podem revolucionar a prestação de serviços de saúde, permitindo monitoramento remoto, telemedicina e gerenciamento personalizado de saúde, reduzindo a distância entre pacientes e profissionais de saúde, especialmente em áreas remotas ou carentes (AWAD et al., 2021; LOWERY, 2020; SENBEKOV et al., 2020). Além disso, a indústria farmacêutica pode se beneficiar do design e descoberta inteligente de medicamentos por meio de métodos computacionais que utilizam IA e aprendizado de máquina. Essas técnicas possibilitam a identificação mais rápida e eficiente de potenciais candidatos a medicamentos, reduzindo custos e tempo associados a métodos tradicionais (FLEMING, 2018; PAUL et al., 2021; AGRAWAL, 2018; PATEL; SHAH, 2022; ZHAVORONKOV, 2018; JIMÉNEZ-LUNA et al., 2021; ZHAVORONKOV; VANHAELEN; OPREA, 2020; DENG et

al., 2022). A integração dessas tecnologias possui um potencial imenso para o Sul Global, oferecendo oportunidades de avançar e superar os desafios existentes na área de saúde, resultando em melhor acesso a cuidados de saúde, melhores resultados para pacientes e desenvolvimento farmacêutico aprimorado.

A estimativa da afinidade entre duas proteínas possui grande importância para as indústrias de saúde e farmacêutica. As interações proteína-proteína desempenham um papel central em diversos processos biológicos e vias de doenças, tornando-as alvos atraentes para intervenções terapêuticas. A estimativa precisa da afinidade de proteínas oferece insights sobre a força e especificidade dessas interações, possibilitando a identificação de possíveis alvos de medicamentos e o design de estratégias terapêuticas eficazes (FERREIRA et al., 2015; AGNIHOTRY et al., 2020; SAIKIA; BORDOLOI, 2019; RAVAL; GANATRA, 2022; VASANT et al., 2021; PINZI; RASTELLI, 2019). O conhecimento preciso da afinidade de proteínas facilita o design racional de medicamentos, a otimização de candidatos e o desenvolvimento de abordagens de medicina personalizada (FERREIRA et al., 2015; AGNIHOTRY et al., 2020; SAIKIA; BORDOLOI, 2019; RAVAL; GANATRA, 2022; VASANT et al., 2021; PINZI; RASTELLI, 2019). Além disso, auxilia na compreensão dos mecanismos subjacentes a diversas doenças, abrindo caminho para a descoberta de novas opções de tratamento (PINZI; RASTELLI, 2019).

A redução do tempo e dos custos associados ao encaixe molecular apresenta desafios significativos no campo da descoberta de medicamentos (PARIS; RUIZ; SOUZA, 2015; QUEVEDO et al., 2014; SAIKIA; BORDOLOI, 2019; XU; LI; CAI, 2017; ALTUNTAŞ; BOZKUS; FRAGUELA, 2016; YURIEV; AGOSTINO; RAMSLAND, 2011; ALONSO; BLIZNYUK; GREADY, 2006; BELLO; MARTÍNEZ-ARCHUNDIA; CORREA-BASURTO, 2013; SALMASO; MORO, 2018; KITCHEN et al., 2004; OKIMOTO et al., 2009; TAYLOR; JEWSBURY; ESSEX, 2002; BIASI et al., 2016; DONG et al., 2018). O encaixe molecular envolve a simulação da ligação entre uma molécula pequena e uma proteína-alvo, um processo que exige extensos recursos computacionais e validação experimental. A exploração do vasto espaço conformacional e a previsão precisa das interações proteína-ligante contribuem para a natureza demorada desse método. Essas limitações prejudicam a eficiência do processo de descoberta de medicamentos e aumentam os custos devido à necessidade de iterações experimentais extensas (PARIS; RUIZ; SOUZA, 2015; QUEVEDO et al., 2014; SAIKIA; BORDOLOI, 2019; XU; LI; CAI, 2017; ALTUNTAŞ; BOZKUS; FRAGUELA, 2016; SALMASO; MORO, 2018; BIASI et al., 2016; DONG et al., 2018). Superar essas dificuldades requer abordagens inovadoras que possam agilizar o encaixe molecular, mantendo a precisão e reduzindo a dependência de procedimentos experimentais intensivos em recursos.

A inteligência computacional, mais especificamente a aprendizagem de máquina, oferece vantagens valiosas na estimativa do grau de encaixe molecular, abordando os desafios mencionados acima. Ao aproveitar conjuntos de dados em grande escala, essas técnicas permitem o desenvolvimento de modelos preditivos capazes de aprender padrões complexos e relações entre estruturas proteicas e afinidades de ligação (HECHT; FOGEL, 2009; BALLESTER; MITCHELL, 2010; YANG; CHEN; ZHANG, 2022; ASHTAWY; MAHAPATRA, 2014; ALGHAMEDY et al., 2018; HSIN; GHOSH; KITANO, 2013; TERAYAMA et al., 2018). Modelos de aprendizagem de máquina podem capturar interações não lineares e características sutis que são desafiadoras de modelar usando métodos tradicionais. Através da integração de várias características de entrada, como sequências de proteínas, informações estruturais e propriedades fisicoquímicas, esses modelos podem prever com precisão a afinidade de proteínas (MA et al., 2014; VEIT-ACOSTA; JUNIOR, 2021; LI et al., 2012; KANAKALA et al., 2023; BITENCOURT-FERREIRA; AZEVEDO, 2019; DRUCHOK et al., 2021; HECK et al., 2017). A aplicação da aprendizagem de máquina e da inteligência computacional agiliza a triagem de grandes bibliotecas químicas, acelerando a identificação de possíveis candidatos a medicamentos (MA et al., 2014; VEIT-ACOSTA; JUNIOR, 2021; LI et al., 2012; KANAKALA et al., 2023; BITENCOURT-FERREIRA; AZEVEDO, 2019; DRUCHOK et al., 2021; HECK et al., 2017). Além disso, essas técnicas facilitam a otimização do design de medicamentos e levam a processos de descoberta de medicamentos mais eficientes e econômicos. A utilização da aprendizagem de máquina e da inteligência computacional representa um caminho promissor para aprimorar a precisão e eficiência da estimativa de encaixe molecular.

Nos últimos anos, houve avanços significativos em modelos computacionais baseados em aprendizagem de máquina para estimar o encaixe molecular (JIMÉNEZ-LUNA et al., 2020; GENTILE et al., 2020; MORRONE et al., 2020; YANG et al., 2021). O campo de pesquisa testemunhou o desenvolvimento e aprimoramento de várias abordagens que utilizam arquiteturas de aprendizado profundo, como redes neurais convolucionais (CNNs), redes neurais recorrentes (RNNs) e redes neurais de grafos (GNNs). Esses modelos demonstraram um desempenho impressionante em aplicações de diagnóstico de apoio baseadas em imagens biomédicas, sinais e parâmetros clínicos, como câncer de mama, Covid-19, distúrbios mentais e em doenças neurodegenerativas como Alzheimer e Parkinson (LIMA; SILVA-FILHO; SANTOS, 2016; SANTANA et al., 2018; GOMES et al., 2020; BARBOSA et al., 2021; SANTANA et al., 2018; ESPINOLA et al., 2021a; ESPINOLA et al., 2021b; CORDEIRO; SANTOS; SILVA-FILHO, 2016b; AZEVEDO et al., 2015; OLIVEIRA et al., 2020; GOMES et al., 2023; SANTANA et al., 2022; SHIRAHIGE et al., 2022; FONSECA et al., 2022; SANTANA; SANTOS, 2022; BARBOSA et al., 2022; ESPINOLA et al., 2022; GOMES; RODRIGUES;

SANTOS, 2022; BARBOSA et al., 2022; SOUZA et al., 2021). Tais modelos têm sido utilizados com sucesso na previsão precisa das afinidades de ligação proteína-proteína. Além disso, a disponibilidade de extensos bancos de dados de estrutura e interação de proteínas facilitou a criação de modelos robustos e abrangentes de aprendizado de máquina. A integração de múltiplas modalidades de dados, incluindo sequências de proteínas, estruturas, dinâmicas e informações de ligantes, tem aprimorado ainda mais o poder preditivo desses modelos (GENTILE et al., 2020; MORRONE et al., 2020; YANG et al., 2021; FAN; SHI, 2022; AHMED; MAM; SOWDHAMINI, 2021; JIMÉNEZ-LUNA et al., 2020; STEPNIEWSKA-DZIUBINSKA; ZIELENKIEWICZ; SIEDLECKI, 2018). Além disso, abordagens de transferência de aprendizado surgiram, em que modelos pré-treinados em tarefas relacionadas são ajustados para a previsão de encaixe molecular, resultando em transferência eficiente de conhecimento e convergência mais rápida do modelo. Os modelos computacionais de ponta baseados em aprendizagem de máquina possuem um potencial significativo para estimar com precisão o encaixe molecular, contribuindo assim para o avanço da descoberta de medicamentos e o desenvolvimento de terapêuticas eficazes.

Desenvolver algoritmos inteligentes para o encaixe molecular que evitem a modelagem tridimensional de proteínas por ângulos e energias tem grande importância no campo da descoberta computacional de medicamentos. Ao depender exclusivamente de descrições de proteínas baseadas em sequências de caracteres que representam sequências de aminoácidos, esses algoritmos oferecem vantagens e possibilidades significativas. Um dos principais benefícios é a redução na complexidade computacional e no tempo necessário para a modelagem tridimensional, que pode ser intensiva em recursos e demorada. Ao utilizar diretamente sequências de proteínas, algoritmos inteligentes podem aproveitar a riqueza de informações nelas codificadas, incluindo motivos estruturais, domínios funcionais e relações evolutivas. Essa abordagem não apenas acelera o processo de encaixe molecular, mas também abre novas oportunidades para explorar uma ampla gama de interações proteína-proteína. Além disso, focar em descrições baseadas em sequências de proteínas permite uma aplicabilidade mais ampla, pois possibilita a análise de proteínas com estruturas desconhecidas ou não caracterizadas. Além disso, esses algoritmos podem aproveitar técnicas de aprendizado de máquina para aprender e prever as interações proteína-proteína exclusivamente a partir de informações de sequência, aumentando ainda mais sua precisão e potencial. Portanto, o desenvolvimento de algoritmos inteligentes que utilizam descrições de proteínas baseadas em sequências de aminoácidos representa uma direção promissora na pesquisa de encaixe molecular, possibilitando a estimativa eficiente e precisa da

afinidade proteica ao contornar os desafios associados à modelagem tridimensional(LIU; HU, 2022).

Neste trabalho, propusemos uma arquitetura híbrida inteligente para estimar o encaixe molecular entre duas proteínas usando redes profundas baseadas em pseudo-convoluções e Random Forests. Redes de pseudo-convolução consistem em um processo iterativo onde, dado um número específico de camadas, uma sequência de entrada representativa de um par de proteínas candidatas-alvo, representada como uma sequência de caracteres, é dividida em partes menores com tamanho em potência de dois. Na camada final, o conjunto de sequências obtidas é representado por matrizes de coocorrência que modelam relações de vizinhança e populações de caracteres. As matrizes correspondentes a cada segmento de caractere de saída são transformadas em vetores e então concatenadas, formando um vetor de características. Os vetores assim representados são classificados por uma Floresta Aleatória. O problema de estimativa de encaixe molecular foi modelado como um problema de classificação. Essa abordagem foi inspirada em Gomes et al. (2021a): eles melhoraram os resultados de identificação do vírus obtidos a partir de dispositivos de RT-PCR ao treinar algoritmos de aprendizado de máquina sobre um grande conjunto de dados: 347.363 sequências de DNA de vírus de 24 famílias de vírus e SARS-CoV-2. Os autores obtiveram resultados de sensibilidade e especificidade de 97% a 99%, demonstrando que o diagnóstico molecular da Covid-19 pode ser otimizado combinando RT-PCR e o método pseudo-convolucional para identificar sequências de DNA para o SARS-CoV-2 com valores de especificidade e sensibilidade maiores.

1.2 OBJETIVOS

Este trabalho teve como objetivo geral construir uma arquitetura híbrida inteligente para estimar o encaixe molecular entre duas proteínas usando redes profundas baseadas em pseudo-convoluções e Random Forests.

Como objetivos específicos, têm-se:

1. Investigar e compreender as bases teóricas das redes de pseudo-convolução e das Random Forests no contexto do encaixe molecular de proteínas.

2. Selecionar e preparar um conjunto de dados representativos contendo informações relevantes sobre proteínas candidatas-alvo, considerando sequências de aminoácidos e informações estruturais.

3. Projetar e implementar a arquitetura híbrida inteligente, integrando redes de pseudo-convolução e algoritmos de Random Forests para realizar a estimativa de encaixe molecular.

4. Realizar experimentos e análises para avaliar o desempenho da arquitetura proposta em termos de precisão, sensibilidade e especificidade na previsão de afinidades proteicas.

5. Comparar os resultados obtidos pela arquitetura proposta com abordagens tradicionais de encaixe molecular, destacando as vantagens e desvantagens de cada método.

6. Investigar a capacidade da arquitetura híbrida em lidar com diferentes tipos de proteínas, considerando diferentes classes funcionais e estruturais.

7. Explorar a capacidade de generalização da arquitetura proposta ao lidar com proteínas cujas estruturas são desconhecidas ou não caracterizadas.

8. Analisar os fatores que afetam o desempenho da arquitetura, como o tamanho do conjunto de treinamento, a complexidade das proteínas e a qualidade dos dados de entrada.

9. Propor ajustes e otimizações na arquitetura híbrida com base nos resultados dos experimentos, visando aprimorar seu desempenho e eficácia.

10. Discutir as implicações dos resultados obtidos e o potencial impacto da arquitetura híbrida no campo da descoberta de medicamentos, fornecendo direções para pesquisas futuras nessa área.

2 TRABALHOS RELACIONADOS

O processo de avaliação do potencial de ligação entre um ligante e uma proteína alvo é uma tarefa multifacetada que envolve a seleção de candidatos de bases de dados extensas para subsequente validação *in vitro* e *in vivo* (CRAMPON et al., 2022). Esse processo encontra aplicações em diversas áreas, como descoberta de medicamentos, geração de peptídeos, descoberta de ligantes de DNA e outras.

A triagem virtual é uma técnica empregada para selecionar candidatos a ligantes para proteínas alvo, com o objetivo de reduzir o número de compostos que precisam passar por fases de testes *in vitro/vivo* (CRAMPON et al., 2022). Essa abordagem ajuda a minimizar os custos associados a esses processos. Ao longo dos anos, algoritmos baseados em aprendizado de máquina (ML) foram propostos para triagem virtual. A literatura científica apresenta uma variedade de métodos, incluindo técnicas clássicas de ML como vizinhos mais próximos (K-nearest neighbors) e aumento de gradiente (gradient boosting) (CHANDAK et al., 2020), bem como métodos avançados de aprendizado em grafos (KASHYAP; DATTA, 2022), que são treinados utilizando grandes conjuntos de dados ligante-proteína adaptados para aprendizado supervisionado.

Li et al. (2023) exemplificam o processo de descoberta de novos peptídeos umami para a indústria alimentícia utilizando aprendizado de máquina e encaixe molecular. Eles identificaram com sucesso seis novos peptídeos derivados de ossos de cordeiro e determinaram que ligações de hidrogênio e interações eletrostáticas foram as principais forças envolvidas. Seu trabalho empregou uma abordagem de aprendizado profundo baseada em um modelo de rede neural que combinou redes neurais totalmente conectadas (MLP) e redes neurais recorrentes (RNN) para prever o sabor umami e avaliar o limiar de sabor umami de peptídeos desconhecidos.

Abordando a exploração limitada de metodologias de encaixe para interações entre ácidos nucleicos e agentes intercalantes de DNA, Oliveira et al. (2022) propuseram um método de aprendizado de máquina para prever alterações na temperatura de fusão do DNA após a ligação do fármaco. Eles compararam sua abordagem com os métodos Autodock, Dock6 e Consensus.

Em um estudo de reposicionamento de fármacos envolvendo medicamentos aprovados pela FDA, Yang, Li e Chang (2022) identificaram o Cobimetinib como um inibidor de A-FABP. Eles investigaram um conjunto de dados com aproximadamente 2.600 compostos e empregaram uma abordagem de aprendizado de máquina baseada em ligantes e um método

de encaixe molecular baseado em estrutura, ambos baseados em Naive Bayesian. Além disso, eles utilizaram o t-SNE para visualização de dados.

Choi e Lee (2021) desenvolveram uma abordagem para propor moléculas semelhantes a fármacos utilizando uma estratégia global de otimização de propriedades moleculares combinada com um algoritmo de aprendizado de máquina para previsão de escores de encaixe. Seu método gerou diversas moléculas inéditas com altos escores de encaixe para uma proteína específica. O modelo de aprendizado de máquina serviu como parte da função objetiva para avaliar o escore de encaixe das soluções propostas.

De maneira geral, técnicas de aprendizado de máquina têm sido aplicadas a tarefas específicas de pontuação de alvos. Diversos estudos, como os de Ricci-Lopez et al. (2021), Chandak et al. (2020), Nogueira e Koch (2019), demonstraram a aplicação de diferentes abordagens clássicas de aprendizado de máquina, incluindo Regressão Logística, Aumento de Gradiente, Vizinhos Mais Próximos (K-Nearest Neighbor), Máquina de Vetores de Suporte, Floresta Aleatória e Perceptron de Múltiplas Camadas.

A maioria das abordagens computacionais para o encaixe de proteínas encontradas na literatura científica requer a simulação do complexo em 3D antes de prever escores ou determinar compostos ativos/inativos (LIN; LI; LIN, 2020; WU et al., 2020b; ZHAO; CAO; ZHANG, 2020; AL-KHAFAJI; AL-DUHAIDAHAWI; TOK, 2021; AFTAB et al., 2020; CUI et al., 2020; NORMAN et al., 2020; WANG et al., 2020; PARVEZ et al., 2020; WU et al., 2020a). Neste estudo, propomos uma solução baseada em pseudo-convolução para representação de proteínas e previsão de encaixe de proteínas. Este método elimina a necessidade de simulação complexa em 3D.

2.1 Metodologia de Seleção dos Trabalhos

Para compor esta revisão, foi conduzida uma busca sistemática nas bases de dados Google Scholar, PubMed e Scielo, utilizando as palavras-chave: “protein-protein interaction prediction”, “sequence-based docking”, “machine learning in drug discovery” e “deep learning PPI”. O período de busca abrangeu os últimos cinco anos (2019–2024), priorizando artigos revisados por pares e com foco em abordagens computacionais para predição de interações proteína-proteína.

Os critérios de inclusão consideraram trabalhos que: ▪ Utilizam aprendizado de máquina ou técnicas computacionais para predição de interações proteína-proteína; ▪ Apresentam resultados quantitativos claros (ex.: acurácia, sensibilidade, especificidade); ▪ Descrevem meto-

dologias inovadoras ou comparações com métodos tradicionais.

Foram excluídos estudos que: ▪ Não disponibilizam detalhes metodológicos suficientes para reprodução; ▪ Focam exclusivamente em interações proteína-ligante sem considerar interações proteína-proteína; ▪ São revisões sem apresentação de novos dados experimentais ou computacionais.

2.2 Tabela Comparativa dos Trabalhos Analisados

2.3 Análise Crítica dos Trabalhos

Os trabalhos analisados demonstram a diversidade de abordagens na predição de interações proteína-proteína, variando desde métodos clássicos de aprendizado de máquina até técnicas mais recentes de aprendizado profundo. No entanto, algumas limitações são recorrentes:

- **Especialização Excessiva**: Muitos estudos focam em aplicações específicas, como sabor umami ou reposicionamento de fármacos, limitando a generalização dos métodos.
- **Dependência de Estrutura 3D**: Abordagens que requerem informações estruturais tridimensionais podem ser restritivas, especialmente quando tais dados não estão disponíveis.
- **Validação Experimental Limitada**: A ausência de validação experimental em alguns estudos dificulta a comprovação prática dos resultados obtidos.

O método proposto nesta dissertação busca superar essas limitações ao:

- Utilizar representações baseadas em sequências de aminoácidos, eliminando a necessidade de estruturas 3D;
- Aplicar uma arquitetura híbrida combinando redes de pseudo-convolução e Random Forests, visando maior acurácia e eficiência;
- Focar em uma abordagem generalista, aplicável a diversas interações proteína-proteína, com potencial para validação experimental futura.

Tabela 1 – Comparativo de abordagens para predição de interações proteína-proteína

Autores	**Abordagem/Método**	**Resultados**	**Limitações**	**Comparação com o Método Proposto**
(LI et al., 2023)	Redes neurais combinando MLP e RNN para previsão de sabor umami	Identificação de seis novos peptídeos com interações significativas	Foco limitado ao sabor umami; aplicabilidade restrita	O método proposto visa uma abordagem mais ampla, aplicável a diversas interações proteína-proteína
(OLIVEIRA et al., 2022)	ML para prever alterações na temperatura de fusão do DNA pós-ligação	Comparação com Autodock, Dock6 e Consensus; resultados promissores	Aplicação específica a interações DNA-ligante	Nosso modelo foca em interações proteína-proteína, ampliando o escopo de aplicação
(YANG; LI; CHANG, 2022)	ML baseado em Naive Bayes e docking estrutural para reposicionamento de fármacos	Identificação do Cobimetinib como inibidor de A-FABP	Base de dados limitada a compostos aprovados pela FDA	O método proposto utiliza uma abordagem mais generalista, não restrita a compostos previamente aprovados
(CHOI; LEE, 2021)	Otimização global de propriedades moleculares com ML para previsão de escores de docking	Geração de moléculas inéditas com altos escores de docking	Necessidade de validação experimental das moléculas geradas	Nosso modelo prioriza a predição de interações existentes, facilitando a validação experimental
(RICCI-LOPEZ et al., 2021)	Aplicação de métodos clássicos de ML (Regressão Logística, KNN, SVM)	Resultados variados dependendo do método e conjunto de dados	Desempenho inferior em comparação com técnicas de aprendizado profundo	O método proposto utiliza técnicas avançadas de aprendizado profundo, visando maior acurácia

3 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos e técnicas fundamentais que embasam esta dissertação. Serão abordados os princípios de **aprendizado de máquina**, com ênfase na técnica de *Random Forests*, os fundamentos de **docking molecular**, o conceito de **redes de pseudo-convolução** e, por fim, a ideia de **transfer learning** (aprendizado por transferência), conforme pertinente na literatura recente. Cada seção traz definições claras, exemplos de aplicações biomédicas e a relação de cada tópico com o tema desta pesquisa, com referências clássicas e atuais para embasar a discussão.

3.1 APRENDIZADO DE MÁQUINA

Aprendizado de máquina (AM) é um ramo da inteligência artificial que lida com algoritmos capazes de *aprender* padrões a partir de dados e experiências, melhorando seu desempenho em determinada tarefa sem serem explicitamente programados para tal. Uma definição clássica estabelece que um sistema de AM "aprende" de experiência E em relação a um conjunto de tarefas T e uma medida de desempenho P se seu desempenho em T, medido por P, melhora com E (MITCHELL, 1997). Em termos práticos, esses algoritmos constroem modelos estatísticos ou computacionais a partir de dados de treinamento, permitindo prever ou tomar decisões sobre novos dados de forma autônoma.

Os algoritmos de AM podem ser divididos em algumas categorias principais, de acordo com o tipo de aprendizado realizado:

- Aprendizado supervisionado: quando o algoritmo aprende a partir de exemplos rotulados de entrada-saída, ajustando um modelo para prever as saídas corretas para futuras entradas (e.g., classificação de imagens médicas com diagnóstico conhecido ou regressão do nível de um biomarcador) (JAMES et al., 2013).

- Aprendizado não supervisionado: quando o modelo busca extrair estrutura ou padrões dos dados sem qualquer rótulo prévio, como agrupamento (clusterização) de pacientes com perfis genéticos semelhantes, detecção de anomalias, entre outros.

- Aprendizado por reforço: quando um agente aprende por meio de interação com um ambiente, recebendo recompensas ou penalidades e melhorando sua estratégia com o

objetivo de maximizar a recompensa cumulativa (por exemplo, otimização de protocolos de tratamento adaptativos, embora esse paradigma seja menos comum em aplicações clínicas diretas).

Nos últimos anos, o aprendizado de máquina tornou-se onipresente em aplicações biomédicas, impulsionado pela disponibilidade de grandes volumes de dados e avanços em hardware de computação. Técnicas de AM têm sido empregadas em diagnóstico por imagem (por exemplo, na detecção de tumores em radiografias, tomografias e ressonâncias) com níveis de desempenho equiparáveis aos de especialistas humanos (TOPOL, 2019). Também têm contribuído para a análise de dados genômicos e de expressão gênica, auxiliando na identificação de genes associados a doenças e na medicina de precisão (TOPOL, 2019). Em problemas de descoberta de fármacos, métodos de aprendizado de máquina vêm sendo utilizados para prever propriedades bioquímicas de moléculas e suas probabilidades de interação com alvos biológicos, agilizando etapas de triagem de candidatos a medicamento (KHAMIS; GOMAA; AHMED, 2015). Dessa forma, o AM reduz custos e tempo de pesquisa ao priorizar experimentos mais promissores com base em modelos preditivos.

No contexto desta dissertação, o aprendizado de máquina é a base para desenvolver uma abordagem inteligente de predição de *docking* molecular entre proteínas. Em vez de depender exclusivamente de cálculos determinísticos ou físicos de encaixe molecular, utilizamos modelos de AM treinados em dados conhecidos de interações para inferir a afinidade ou a probabilidade de duas proteínas formarem um complexo. Assim, os próximos tópicos detalharão técnicas específicas de AM empregadas (como *Random Forests*) e como se incorporam ao problema de docking proposto.

3.2 RANDOM FORESTS

Definição e funcionamento

Random Forests (ou *Florestas Aleatórias*) são um método de aprendizado de máquina do tipo *ensemble*, introduzido por Leo Breiman em 2001 (BREIMAN, 2001), que consiste em um conjunto de múltiplas árvores de decisão construídas de forma aleatorizada e combinadas para produzir uma predição única e mais robusta. A ideia central é reduzir a variabilidade e melhorar a generalização do modelo aproveitando a sabedoria do conjunto de árvores, em contraste com

o uso de uma única árvore de decisão que tende a sobreajustar aos dados de treinamento.

Em uma *Random Forest*, cada árvore é treinada com uma amostra aleatória (com reposição) dos dados de treinamento – técnica conhecida como *bagging* (Bootstrap AGGregatING) – e, adicionalmente, em cada nó da árvore, um subconjunto aleatório dos atributos é considerado para definir a melhor divisão (*split*). Esse duplo mecanismo de aleatoriedade (nos dados e nos atributos) promove a diversidade entre as árvores do conjunto, de modo que os erros de predição de árvores individuais tendem a se cancelar quando é feita a agregação por voto majoritário (para classificação) ou média (para regressão) (BREIMAN, 2001). O resultado é um modelo geralmente mais acurado e estável do que uma única árvore, mantendo ainda a interpretabilidade parcial via medidas de importância de variáveis.

Algumas características e vantagens-chave dos *Random Forests* incluem:

- Robustez a overfitting: devido à média dos resultados de muitas árvores, o modelo final tende a ter menor sobreajuste aos dados de treino do que árvores individuais, mantendo boa capacidade de generalização.

- Trabalho com alta dimensionalidade: *Random Forests* lidam bem com grandes quantidades de variáveis de entrada, podendo operar em cenários de milhares de atributos (como dados genômicos) e identificar automaticamente quais são mais relevantes para a predição (JAMES et al., 2013).

- Importância de atributos: o algoritmo fornece estimativas de importância de cada característica no processo decisório (por exemplo, através da redução média do critério de impureza ou da acurácia fora-da-bolsa quando aquele atributo é permutado), o que pode ser valioso em contextos biomédicos para destacar biomarcadores significativos.

- Paralelismo e escalabilidade: como as árvores são construídas independentemente, o treinamento pode ser paralelizado, e o método é escalável para conjuntos de dados grandes sem necessidade de muitos ajustes de hiperparâmetros.

Aplicações biomédicas e relação com o tema

Random Forests têm encontrado ampla aplicação em problemas biomédicos e bioinformáticos, graças à sua capacidade de lidar com dados complexos, possivelmente ruidosos, e produzir modelos acurados. Por exemplo, em estudos de expressão gênica e dados de microarranjos,

onde o número de genes (atributos) é muito maior que a quantidade de amostras, métodos ensemble como *Random Forest* mostraram desempenho superior a técnicas lineares tradicionais na classificação de subtipos de câncer (JAMES et al., 2013). De maneira geral, na detecção de doenças, predição de prognósticos e identificação de fatores de risco, *Random Forests* têm sido utilizados com sucesso, frequentemente superando modelos mais simples pela habilidade de capturar interações não lineares entre variáveis clínicas (TOPOL, 2019).

No âmbito de descoberta de fármacos e biologia estrutural, *Random Forests* também vêm sendo empregados para melhorar etapas do processo de *drug design*. Em docking molecular, por exemplo, já foram propostos *scoring functions* baseados em *Random Forest* para avaliar poses de ligação proteína-ligante, competindo favoravelmente com funções de pontuação tradicionais derivadas da física (WÓJCIAKOWSKI; BALLESTER; SIEDLECKI, 2017). Essa utilização se dá porque o modelo pode aprender, a partir de exemplos conhecidos, combinações complexas de interações moleculares que levam a alta afinidade de ligação, muitas vezes capturando efeitos que funções matemáticas pré-definidas não conseguem modelar completamente.

Cabe destacar um caso aplicado relacionado: Gomes et al. (2021), ao propor uma metodologia de identificação de genomas virais baseada em pseudo-convoluções, testaram diversos algoritmos de classificação (como k-NN, redes neurais, Naive Bayes, SVM e *Random Forest*) e verificaram que o *Random Forest* obteve o melhor desempenho global (maior acurácia e índice $kappa$) na distinção de sequências do SARS-CoV-2 de outros vírus (GOMES et al., 2021b). Esse resultado reforça a adequação do *Random Forest* para tarefas biomédicas complexas, motivando sua escolha nesta dissertação.

No presente trabalho, o *Random Forest* é utilizado como componente classificatório da abordagem híbrida proposta. Após extrair atributos relevantes das sequências proteicas por meio das redes de pseudo-convolução (seção 3.4), utiliza-se um modelo *Random Forest* para receber esses atributos e aprender a decidir se um par de proteínas possui boa compatibilidade de encaixe (alta afinidade) ou não. A opção por *Random Forest* se justifica por sua robustez a dados heterogêneos, capacidade de manejar múltiplos atributos de entrada (inclusive possivelmente redundantes) e excelente desempenho observado em trabalhos correlatos (GOMES et al., 2021b). Espera-se assim aproveitar o poder de generalização do ensemble para produzir predições confiáveis de docking, mesmo em cenários nos quais os dados de treinamento possam ser limitados em quantidade ou apresentar variabilidade significativa.

3.3 DOCKING MOLECULAR

Docking molecular (ou *encaixe molecular*) é uma técnica computacional utilizada para prever a orientação preferencial e a afinidade de ligação quando duas moléculas interagem, formando um complexo estável. Em termos mais específicos, o docking busca determinar como um ligante (por exemplo, um fármaco em potencial) se acopla ao sítio de ligação de um receptor (tipicamente uma proteína) de maneira energeticamente favorável (KHAMIS; GOMAA; AHMED, 2015). A premissa fundamental é que a conformação de menor energia livre corresponde ao complexo mais provável, de modo que, ao explorar diversas orientações e conformações relativas entre as moléculas, os algoritmos de docking procuram aquela que maximize a complementaridade intermolecular (geométrica e química) e minimize a energia de interação.

O processo de docking molecular geralmente envolve as seguintes etapas principais:

- Preparação das estruturas: as coordenadas 3D da proteína alvo e do ligante devem estar disponíveis (por exemplo, obtidas experimentalmente via cristalografia de raios X ou modelagem computacional). São definidos o sítio de ligação na proteína e, em algumas abordagens, rotinas para tratar a flexibilidade molecular (rotacionar ligações, etc.).

- Amostragem de poses de encaixe: o algoritmo gera numerosas orientações e conformações possíveis do ligante em relação ao receptor. Podem ser usadas estratégias de busca sistemática, estocástica (como Algoritmos Genéticos, Monte Carlo) ou métodos determinísticos. Para docking proteína-proteína, essa amostragem envolve diversas orientações relativas das duas superfícies protéicas.

- Avaliação por função de pontuação: cada pose gerada é avaliada por uma *scoring function*, que estima a qualidade da interação molecular naquela configuração, normalmente aproximando a energia livre de ligação. Essas funções podem incorporar termos de energia molecular clássicos (van der Waals, eletrostático, hidratação) e parâmetros empíricos ajustados para correlacionar com afinidades conhecidas (TROTT; OLSON, 2010).

- Classificação e seleção das melhores poses: as poses são ordenadas conforme suas pontuações e geralmente um conjunto das melhores (mais estáveis) é retornado. Idealmente, a pose de mais alta pontuação (ou menor energia) deve corresponder ao modo de ligação

verdadeiro. Avaliações adicionais, como refinamentos ou simulações, podem ser feitas nas melhores poses para maior confiabilidade.

Essa técnica é amplamente utilizada no planejamento de fármacos, pois permite triar rapidamente, *in silico*, milhares de compostos candidatos contra um alvo biológico, priorizando aqueles com maior probabilidade de se ligarem firmemente no sítio ativo do alvo (KHAMIS; GOMAA; AHMED, 2015). Desde os primórdios do docking nas décadas de 1980 e 1990, diversas ferramentas e programas tornaram-se disponíveis, aprimorando a eficiência e acurácia das predições. Por exemplo, o software AutoDock Vina, lançado em 2010, popularizou-se por empregar um algoritmo otimizado de busca e uma função de pontuação semiempírica calibrada, alcançando melhores velocidades e acurácia de pose em benchmarks de docking quando comparado a métodos anteriores (TROTT; OLSON, 2010). Esses programas têm ajudado a identificar novos inibidores enzimáticos, orientar modificações em moléculas líderes e entender interações moleculares em nível atômico, contribuindo significativamente para a química medicinal e a biologia estrutural. Embora mais comumente associado ao encaixe de pequenos ligantes em proteínas, o conceito de docking estende-se também a interações macromoleculares, como *docking* proteína-proteína ou proteína-ácido nucleico. Nesses casos, o desafio computacional aumenta, pois envolve superfícies de interação maiores, flexibilidade conformacional significativa e funções de pontuação mais complexas para discriminar interfaces verdadeiras de falsos positivos. O docking proteína-proteína é crucial para prever complexos biomoleculares quando não se dispõe de estruturas experimentais de complexos, auxiliando na compreensão de vias de sinalização celular e interações patógeno-hospedeiro, por exemplo. No entanto, as taxas de sucesso de predição para docking proteína-proteína ainda são limitadas devido à complexidade do problema, exigindo frequentemente etapas de filtragem e refinamento adicionais, bem como conhecimento experimental para restringir as buscas.

Dada a importância do docking molecular, muito esforço tem sido dedicado a aprimorar suas duas componentes centrais: a busca conformacional e as funções de pontuação. As funções de pontuação clássicas nem sempre conseguem distinguir corretamente poses nativas de alternativas, ou predizer com alta correlação os valores de afinidade experimental, levando a falsos positivos/negativos (WÓJCIAKOWSKI; BALLESTER; SIEDLECKI, 2017). Para contornar essas limitações, metodologias de aprendizado de máquina vêm sendo incorporadas ao pipeline de docking nos últimos anos. Uma direção é desenvolver funções de pontuação baseadas em ML: algoritmos treinados em grandes bases de dados de complexos proteína-ligante com afinidades

conhecidas, aprendendo a relacionar atributos do complexo (distâncias, contatos, características físico-químicas) com a afinidade ou probabilidade de interação (KHAMIS; GOMAA; AHMED, 2015). Khamis et al. (2015) revisam diversas abordagens nesse sentido, mostrando que modelos como redes neurais e métodos de ensemble podem melhorar a acurácia de predição de afinidades em comparação com funções empíricas ou de mecânica molecular puras (KHAMIS; GOMAA; AHMED, 2015).

Resultados promissores emergiram dessa integração entre docking e ML. Wójcikowski et al. (2017) demonstraram que funções de pontuação aprendidas, incluindo modelos de *Random Forest* treinados em dados de docking, superaram funções de pontuação tradicionais em tarefas de *virtual screening* (identificação de ligantes ativos em bibliotecas de compostos) e predição de afinidades em benchmarks padronizados (WÓJCIAKOWSKI; BALLESTER; SIEDLECKI, 2017). De forma semelhante, Ragoza et al. (2017) treinaram uma rede neural convolucional 3D para avaliar diretamente complexos proteína-ligante: o modelo, denominado AtomNet, analisava representações volumétricas dos átomos do ligante e do receptor em uma possível pose e aprendia a prever a probabilidade de ligação (RAGOZA et al., 2017). Esse foi um marco por aplicar *deep learning* ao *docking*, obtendo desempenho competitivo com métodos consagrados e inaugurando uma linha de pesquisas em que redes neurais profundas são usadas para *scoring* e geração de poses. Além de substituir a função de pontuação, outra estratégia é usar ML para filtrar ou priorizar poses geradas por um algoritmo de docking tradicional, ou para refinar predições. Zhang et al. (2019) propuseram uma estratégia combinada de descoberta de fármacos em que modelos de máquina de vetor de suporte (SVM) foram treinados para distinguir ligantes verdadeiros de falsos positivos após a etapa de docking, integrando-se assim predições de docking e aprendizado para melhorar a taxa de sucesso na seleção de compostos candidatos (ZHANG et al., 2019). Guedes et al. (2021) também desenvolveram um conjunto de novas funções de pontuação híbridas (DockTScore) que combinam termos de campos de força clássicos com modelos de aprendizado de máquina (como regressão linear múltipla, SVM e *Random Forest*) treinados em grandes bases de dados de afinidades experimentais; as funções resultantes apresentaram desempenho comparável ou superior às melhores funções disponíveis em predições de energia de ligação, evidenciando a contribuição do ML na modelagem de interações moleculares (GUEDES et al., 2021).

Em suma, a tendência recente na área de docking molecular aponta para abordagens que aliam conhecimento biofísico e inteligência artificial para superar limites dos métodos convencionais. É exatamente nessa interseção que se insere o presente trabalho. Em vez de realizar

o docking de forma explícita via simulação física detalhada (o que pode ser computacionalmente custoso e sujeito a erros de pontuação), nossa abordagem busca estimar a afinidade de interação entre duas proteínas a partir de informações de sequência e modelos de aprendizado treinados. Trabalhos prévios sugerem a viabilidade de se predizer afinidade de ligação proteína-proteína utilizando apenas dados de sequência quando técnicas avançadas de representação e aprendizado são empregadas (CAMPOS; FERREIRA; SANTOS, 2022). Assim, esta dissertação explora uma metodologia na qual as sequências aminoacídicas das proteínas de interesse são convertidas em representações numéricas ricas (por redes de pseudo-convolução, descritas adiante), e então um modelo de aprendizado de máquina (*Random Forest*) infere, com base em exemplos conhecidos, quão fortemente tais proteínas provavelmente interagem. Essa "atalho" inteligente evita a necessidade de simular o encaixe tridimensional exato, mas ainda assim busca fornecer resultados acurados sobre a afinidade, conforme demonstrado em pesquisas recentes (CAMPOS; FERREIRA; SANTOS, 2022).

3.4 REDES DE PSEUDO-CONVOLUÇÃO

As redes de pseudo-convolução constituem uma abordagem inovadora de extração de características de sequências biológicas, introduzida nos últimos anos com aplicações em genômica e bioinformática (GOMES et al., 2021b). A ideia central é inspirada nas redes neurais convolucionais tradicionais (CNNs), conhecidas por capturar padrões locais em dados estruturados (como imagens ou sequências) através de filtros convolucionais aprendidos. No entanto, em vez de utilizar filtros treináveis e camadas profundas como em uma CNN típica, a técnica de pseudo-convolução realiza uma transformação determinística nas sequências, fragmentando-as e calculando padrões de coocorrência de símbolos, resultando em uma representação que pode ser interpretada como uma "imagem" característica da sequência (GOMES et al., 2021b). Por essa razão, denomina-se "pseudo-convolução" – há uma operação análoga à convolução no sentido de considerar subsequências locais, mas não envolve convoluções matemáticas padrão com kernels aprendidos.

Em linhas gerais, uma rede pseudo-convolucional realiza os seguintes passos para extrair atributos de uma sequência biológica (por exemplo, uma cadeia de nucleotídeos ou de aminoácidos) (GOMES et al., 2021b; CAMPOS; FERREIRA; SANTOS, 2022):

- Fragmentação da sequência: a sequência original é dividida em n subsequências menores,

possivelmente com sobreposição entre elas. Essa sobreposição simula a ideia de uma janela deslizante, semelhante ao campo receptivo de uma convolução. Por exemplo, em *genomas* de vírus, Gomes et al. (2021) dividiram cada sequência nucleotídica em segmentos de tamanho fixo com certo percentual de overlap (GOMES et al., 2021b). O número de segmentos (n) e o grau de sobreposição são hiperparâmetros definidos de acordo com o problema e o tamanho das sequências.

- Cálculo de matrizes de coocorrência: para cada subsequência gerada, calcula-se a frequência de ocorrência de pares de símbolos adjacentes (ou dentro de uma janela determinada) na sequência. Esses símbolos podem ser nucleotídeos (A, C, G, T, etc. para DNA/RNA) ou aminoácidos (20 tipos para proteínas, considerando o código de uma letra). As frequências são organizadas em uma matriz quadrada cuja dimensão é o tamanho do alfabeto considerado. Por exemplo, no caso de sequências de proteínas usando o alfabeto de 20 aminoácidos, obtém-se uma matriz 20×20 onde a entrada (i, j) representa a contagem (ou proporção) do par de aminoácidos (i, j) aparecer em determinada relação (e.g., consecutivos na subsequência). Essa matriz é essencialmente um mapa de coocorrência de aminoácidos, capturando relações locais entre resíduos na sequência.

- Conversão em imagem: a matriz de coocorrência calculada para cada subsequência pode ser interpretada como uma imagem, onde cada célula da matriz é um *pixel* com intensidade correspondente à frequência normalizada do par de símbolos. Frequentemente atribui-se uma escala de cores (um colormap) para transformar a matriz em uma imagem colorida (por exemplo, mapeando valores baixos para tons escuros e valores altos para tons claros, ou usando um gradiente de cores). No caso de trabalhar com dados de proteínas, pode-se gerar inicialmente duas matrizes de coocorrência separadas (uma para cada proteína em estudo) e, posteriormente, concatená-las lado a lado, formando uma única imagem composta que representa o par de proteínas (CAMPOS; FERREIRA; SANTOS, 2022). Essa imagem combinada incorpora informações de ambas as sequências e pode ressaltar padrões complementares entre as duas (p. ex., tendências de certas combinações de aminoácidos nas superfícies de interação).

- Extração de atributos quantitativos: as imagens obtidas (seja uma por sequência ou uma imagem conjunta do par) servem como base para extrair atributos numéricos. Uma abordagem é "achatar" (flatten) a matriz ou imagem em um vetor de atributos, possi-

velmente após alguma redução de dimensionalidade. Alternativamente, pode-se aplicar técnicas de visão computacional ou aprendizado profundo sobre essas imagens para obter características de alto nível. Por exemplo, Campos et al. (2022) empregaram uma rede neural convolucional pré-treinada (VGG19) para processar a imagem concatenada das matrizes de coocorrência das proteínas, obtendo assim um vetor de atributos avançados a partir do penúltimo nível da VGG19 (CAMPOS; FERREIRA; SANTOS, 2022). Essa etapa já adentra o campo de *transfer learning*, descrito na próxima seção, mas é opcional dependendo do enfoque: no método original de Gomes et al. (2021), optou-se por utilizar diretamente as frequências das matrizes (flattened) como atributos de entrada para classificadores tradicionais, após normalização (GOMES et al., 2021b).

A principal vantagem das redes de pseudo-convolução é produzir uma representação fixa e comparável de sequências de comprimentos variados, capturando informações de padrões locais de forma alinhamento-independente. Em vez de realizar um alinhamento múltiplo de sequências ou extrair manualmente descritores bioquímicos, a pseudo-convolução fornece um retrato estatístico da sequência (ou par de sequências) que pode alimentar algoritmos de aprendizado. Gomes et al. (2021) demonstraram que essa representação é poderosa: no caso de detecção de SARS-CoV-2, sua técnica de pseudo-convolução aliada a classificadores de ML alcançou sensibilidade e especificidade superiores a 99%, distinguindo com sucesso o novo coronavírus de centenas de milhares de outras sequências virais, sem necessidade de alinhamento genômico (GOMES et al., 2021b). Isso evidencia que as matrizes de coocorrência preservam características discriminativas importantes das sequências.

No contexto de docking molecular entre proteínas, as redes de pseudo-convolução oferecem uma forma de traduzir sequências de aminoácidos em informações potencialmente relacionadas à interação estrutural. A premissa é que a sequência de uma proteína carrega indícios (ainda que indiretos) sobre sua estrutura tridimensional e possíveis sítios de interação. Ao aplicar pseudo-convolução, cada proteína é convertida em uma imagem que resume padrões de aminoácidos (por exemplo, frequência de pares hidrofóbicos, cargas opostas adjacentes, etc.). Concatenando as imagens de duas proteínas em análise, obtém-se um único mosaico que, de certa maneira, coloca "lado a lado" as assinaturas sequenciais de ambas. Campos et al. (2022) empregaram exatamente essa ideia em um método de predição de afinidade proteína-proteína: geraram matrizes de coocorrência para as sequências de duas proteínas, uniram as matrizes e então utilizaram técnicas de aprendizado de máquina para regressão a fim de estimar a afini-

dade de ligação do par (CAMPOS; FERREIRA; SANTOS, 2022). Esse método obteve coeficientes de correlação (Pearson $\approx 0,66$) comparáveis a abordagens state-of-the-art de docking que utilizam estruturas, o que é notável considerando que apenas informações de sequência foram utilizadas (CAMPOS; FERREIRA; SANTOS, 2022).

Nesta dissertação, adotamos as redes de pseudo-convolução como componente de extração de características, dada sua capacidade de condensar propriedades relevantes das sequências proteicas de forma automatizada. As sequências das proteínas alvo são processadas por essa técnica, produzindo vetores de atributos que servem de entrada para o modelo *Random Forest* (seção anterior). Assim, as pseudo-convoluções atuam como uma espécie de *feature engineering* baseada em conhecimento biológico (frequências de padrões de aminoácidos), porém realizada de modo sistemático e reproduzível. A combinação dessas representações com um classificador robusto permite explorar se há correspondências entre padrões sequenciais e a propensão de duas proteínas interagirem – um tópico de grande interesse, pois, se bem-sucedido, pode viabilizar triagens rápidas de interações proteína-proteína apenas a partir de dados genômicos ou proteômicos, sem experimentação estrutural intensiva.

3.5 TRANSFER LEARNING (APRENDIZADO POR TRANSFERÊNCIA)

Transfer learning, ou aprendizado por transferência, é um paradigma de aprendizado de máquina em que o conhecimento adquirido em um problema (fonte) é reutilizado para ajudar a resolver outro problema (alvo) que tenha alguma relação ou semelhança. Diferentemente do aprendizado tradicional, em que se treina um modelo do zero para cada nova tarefa, no *transfer learning* busca-se transferir parte dos parâmetros ou estruturas de um modelo pré-treinado em um conjunto de dados grande e genérico para uma nova tarefa que tipicamente dispõe de menos dados (PAN; YANG, 2010). A hipótese subjacente é que representações internas aprendidas em tarefas amplas (por exemplo, reconhecimento de imagens genéricas ou linguagem natural) capturam características de baixo e médio nível que podem ser úteis em outras tarefas, reduzindo a necessidade de dados e tempo de treinamento no novo domínio.

Um cenário clássico de *transfer learning* ocorre em visão computacional: modelos de redes neurais convolucionais profundos treinados no banco de imagens ImageNet (que contém milhões de imagens de objetos do cotidiano, repartidas em milhares de classes) mostraram-se altamente eficazes como ponto de partida para tarefas de imagem médicas, que possuem conjuntos de treinamento muito menores (YAMASHITA et al., 2018). Ao reutilizar a arquitetura

e os pesos de uma rede já treinada (por exemplo, VGG, ResNet, Inception), seja fixando-os como extratores de características ou ajustando-os levemente (*fine-tuning*) com alguns dados da nova tarefa, os pesquisadores conseguiram resultados superiores do que treinar uma rede semelhante do zero com poucas imagens (YAMASHITA et al., 2018). Isso se explica porque as primeiras camadas de uma CNN típica aprendem filtros genéricos (bordas, texturas, formas) que são aplicáveis em praticamente qualquer conjunto de imagens, inclusive radiografias ou lâminas histológicas; assim, a rede pré-treinada já fornece uma base de detecção de padrões visuais, restando ao treinamento na base pequena apenas ajustar camadas finais para a especificidade do problema (por exemplo, detecção de pneumonia em raio-X) (YAMASHITA et al., 2018; TOPOL, 2019). Estratégias semelhantes de transferência têm sido adotadas em processamento de linguagem natural, onde modelos extensivamente treinados em enormes corpora de texto (como BERT, GPT) são refinados para tarefas biomédicas específicas (extração de informações de prontuários, classificação de artigos, etc.), obtendo excelentes resultados sem necessitar de treinar modelos gigantes do início.

Do ponto de vista formal, o aprendizado por transferência pode envolver a transferência de diferentes elementos: *features* (atributos ou representações latentes), parâmetros de modelos (pesos de redes neurais), estruturas (arquiteturas ou partes de redes) ou mesmo conhecimento relacional/linguístico. Zhuang et al. (2021) fornecem uma revisão abrangente de técnicas de *transfer learning*, destacando que elas têm se tornado onipresentes em diversas áreas da ciência de dados, especialmente quando há desequilíbrio de disponibilidade de dados entre a tarefa fonte e alvo (ZHUANG et al., 2021). Em problemas biomédicos, nos quais frequentemente há escassez de dados rotulados (pela dificuldade ou custo de obtenção, necessidade de especialistas, experimentos caros), o *transfer learning* mostrou ser uma ferramenta valiosa para aumentar a acurácia de modelos preditivos (TOPOL, 2019). Por exemplo, na análise de imagens de patologia digital, redes treinadas em ImageNet e depois ajustadas com um número limitado de imagens patológicas conseguiram identificar células cancerígenas com desempenho notável, algo que seria impraticável sem a transferência devido à limitada coleção de imagens disponíveis para treino direto.

No escopo desta dissertação, o conceito de aprendizado por transferência é relevante ao combinar-se com as redes de pseudo-convolução discutidas anteriormente. Conforme mencionado, uma das abordagens possíveis após gerar as matrizes de coocorrência das sequências proteicas é empregar uma rede neural profunda pré-treinada para extrair características dessas matrizes (tratadas como imagens). Essa estratégia foi adotada por Campos et al. (2022): os

autores utilizaram a VGG19, uma CNN profunda treinada no ImageNet, aplicando-a sobre as imagens de coocorrência concatenadas das duas proteínas em cada par, e aproveitaram os neurônios das camadas intermediárias da VGG19 como descritores numéricos das características das sequências (CAMPOS; FERREIRA; SANTOS, 2022). Em seguida, realizaram um pós-processamento para reduzir a dimensionalidade e usaram esses descritores em um modelo de regressão (no caso deles, uma *Random Forest* ou outro regressor não linear) para prever o valor de afinidade de ligação. Essa é uma forma explícita de *transfer learning*: transfere-se o aprendizado de uma rede treinada para reconhecer objetos genéricos (a VGG19 foi originalmente treinada para classificar fotos em mil categorias) para o domínio de análise de sequências proteicas, partindo do pressuposto que os filtros da rede (que detectam bordas, gradientes, texturas etc.) podem capturar padrões nas matrizes de coocorrência que estejam correlacionados com propriedades biológicas importantes.

É importante ressaltar que, embora a VGG19 (ou qualquer rede de imagem genérica) não tenha sido treinada com dados biológicos, as características de baixo nível que ela extrai (como distribuições espaciais de intensidades nas imagens de coocorrência) podem servir como um conjunto inicial de atributos. A etapa final de aprendizado (seja uma *Random Forest*, uma rede neural ou outro modelo) então ajusta a combinação dessas características para a tarefa específica de predição de docking. Em síntese, utiliza-se o *transfer learning* para evitar ter que treinar uma rede profunda de ponta a ponta com dados de docking (o que demandaria muitas amostras de interações proteína-proteína, geralmente escassas), aproveitando o "conhecimento visual" da rede pré-treinada e focando o aprendizado apenas na etapa final.

Nesta pesquisa, consideramos e exploramos o uso de aprendizado por transferência de forma similar, avaliando seu impacto na qualidade das predições de interação. A literatura indica que tal abordagem pode melhorar a expressividade dos atributos extraídos, capturando nuances que uma simples contagem de frequências talvez não evidencie (CAMPOS; FERREIRA; SANTOS, 2022). Por outro lado, também implica acrescentar complexidade ao modelo e depende da premissa de que a representação pseudo-convolucional em forma de imagem é compatível com os filtros de uma rede treinada em fotografias naturais. Assim, um dos aspectos analisados teoricamente neste capítulo é justamente quando e por que o *transfer learning* faz sentido no contexto de docking molecular: em resumo, ele é especialmente útil quando se dispõe de poucas amostras de interação para treinar (o que é comum, dadas as dificuldades de obter dados confiáveis de afinidade proteína-proteína em larga escala) e quando se consegue estabelecer uma representação dos dados de entrada que "dialogue" com modelos pré-treinados

disponíveis. No caso, as pseudo-convoluções fornecem essa ponte, convertendo sequências em imagens, e permitindo assim que modelos de visão computacional sejam aplicados – um belo exemplo de sinergia entre diferentes áreas de pesquisa em prol de resolver um problema biomédico complexo.

4 METODOLOGIA

4.1 BASES DE DADOS

Para implementar e validar nossa proposta, utilizamos dois conjuntos de dados: *Affinity Benchmark Version 3* (257) (HWANG et al., 2008) e *Negatome 2* (BLOHM et al., 2013).

O *Affinity Benchmark Version 3* é uma versão atualizada do benchmark de encaixe de proteína-proteína, conhecido como *Benchmark 2*. De acordo com os autores (HWANG et al., 2008), esta atualização incorpora 40 novos casos de teste, resultando em um aumento de 48% em comparação com o *Benchmark 2.0*. Os 124 casos de teste sem ligação na *Benchmark 3.0* são categorizados em três grupos: 88 casos de corpo rígido, 19 casos de dificuldade média e 17 casos difíceis. A classificação é baseada na extensão das mudanças conformacionais que ocorrem na interface durante a formação do complexo. A expansão do *Benchmark 3.0* não apenas fornece à comunidade científica um conjunto maior de casos de teste para avaliar métodos de encaixe, mas também facilita o desenvolvimento de novos algoritmos que requerem exemplos de treinamento abundantes. Para uma demonstração visual dos complexos da *Affinity Benchmark Version 3*, consulte a Tabela 3.

Tabela 3 – Exemplo da organização da base de dados Affinity Benchmark 3

Complex	Cat.	PDB ID 1	Protein 1	PDB ID 2	Protein 2	I-RMSD (Å)	δASA(Å2)	BM version introduced
Rigid-body (162)								
1AHW_AB:C	AA	1FGN_LH	Fab 5g9	1TFH_A	Tissue factor	0,69	1899	2
1DQJ_AB:C	AA	1DQQ_CD	Fab Hyhel63	3LZT_	HEW lysozyme	0,75	1765	2
1E6J_HL:P	AA	1E6O_HL	Fab	1A43_	HIV-1 capsid protein p24	1,05	1245	2
1JPS_HL:T	AA	1JPT_HL	Fab D3H44	1TFH_B	Tissue factor	0,51	1852	2
1MLC_AB:E	AA	1MLB_AB	Fab44.1	3LZT_	HEW lysozyme	0,6	1392	2

O *Negatome 2.0* introduziu um avanço metodológico significativo por meio da implementação de um procedimento intrincado de mineração de texto para anotação manual. Uma versão revisada do Excerbt, uma ferramenta sofisticada de mineração de texto que emprega análise semântica de sentenças, foi utilizada para identificar possíveis não interações. Inspeções manuais subsequentes revelaram que cerca de 50% dos resultados de mineração de texto com os valores de confiança mais altos correspondiam a pares de proteínas não interagentes (NIP). A expansão do banco de dados no *Negatome 2.0* é notável, com um aumento de mais de 300% em comparação com sua iteração anterior (BLOHM et al., 2013). Para uma representação visual dos complexos do Negatome 2, consulte a Tabela 4.

Uma vez que o objetivo deste estudo é propor um método capaz de prever a presença

Tabela 4 – Exemplo da organização da base de dados Negatome 2

No.	Protein A	Protein B	PMID	Evidence
1	Q6ZNK6	Q9Y4K3	15047173	MI:0019 - coimmunoprecipitation
2	Q9NR31	Q15797	17356069	MI:0019 - coimmunoprecipitation
3	P11627	P53986	20155396	MI:0411 - enzyme linked immunosorbent assay
4	P33176	Q96EK5	16225668	MI:0059 - gst pull down
5	Q9NPY3	P02745	11994479	MI:0411 - enzyme linked immunosorbent assay

ou ausência de encaixe (*docking*) entre duas proteínas, a construção de um banco de dados apropriado exige a inclusão de complexos com encaixe (rotulados como 1) bem como complexos sem encaixe (rotulados como 0). Infelizmente, os bancos de dados públicos existentes na literatura não abrangem ambas as classes simultaneamente. Consequentemente, foi empregada uma combinação de bancos de dados selecionados, cada um fornecendo instâncias das classes necessárias. Os conjuntos de dados utilizados incluem o *Affinity Benchmark 3*, composto por 257 complexos com encaixe, e o *Negatome 2*, que abrange 6184 complexos sem encaixe. Um subconjunto do banco de dados final é ilustrado na Tabela 5.

Tabela 5 – Exemplo da organização do conjunto de dados completo utilizado neste trabalho

Protein A	Protein B	Label
1FGN_LH	1TFH_A	1
1DQQ_CD	3LZT	1
1E6O_HL	1A43	1
1JPT_HL	1TFH_B	1
1MLB_AB	3LZT	1
Q6ZNK6	Q9Y4K3	0
Q9NR31	Q15797	0
P11627	P53986	0
P33176	Q96EK5	0
Q9NPY3	P02745	0

4.2 MÉTODO PROPOSTO

Neste estudo, apresentamos um método para a extração de características de sequências de ácido ribonucleico (RNA) de proteínas. Nossa abordagem, inspirada na máquina de pseudo-convolução utilizada por Gomes et al. (2021a), visa representar os RNAs de proteínas como vetores numéricos derivados de matrizes de coocorrência que capturam as características de

vizinhança dentro de uma molécula. Essa representação permite que um modelo de aprendizado de máquina classifique a presença ou ausência de encaixe entre um par de proteínas. Notavelmente, nosso método elimina a necessidade de simular a estrutura 3D do complexo proteico para a classificação de encaixe. Para validar nossa proposta, foi utilizado o conjunto de dados proposto e empregamos um modelo *Random Forest*.

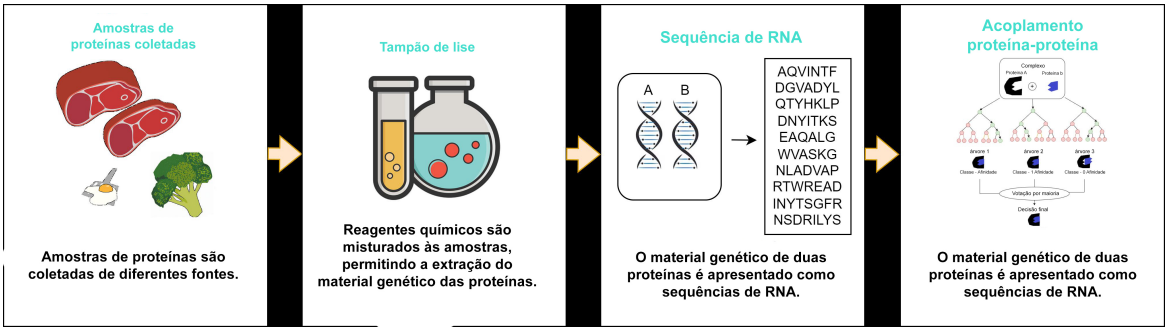


Figura 1 – Esquema geral da proposta: Inicialmente, o material genético é adquirido por meio da coleta de amostras e, posteriormente, as sequências de RNA são obtidas e armazenadas como arquivos de texto. Essas sequências de RNA passam por um processo de representação pseudo-convolucional, que envolve sua conversão em vetores numéricos. Em seguida, um modelo Random Forest é empregado para classificar essas representações em duas categorias: 0 ou 1. Aqui, 1 indica a presença de encaixe entre as amostras, enquanto 0 indica a ausência de encaixe. As amostras utilizadas neste estudo foram obtidas a partir de dois conjuntos de dados distintos: o *Affinity Benchmark 3* e o conjunto de dados *Negatome 2*.

Autoria da figura 1: O Autor

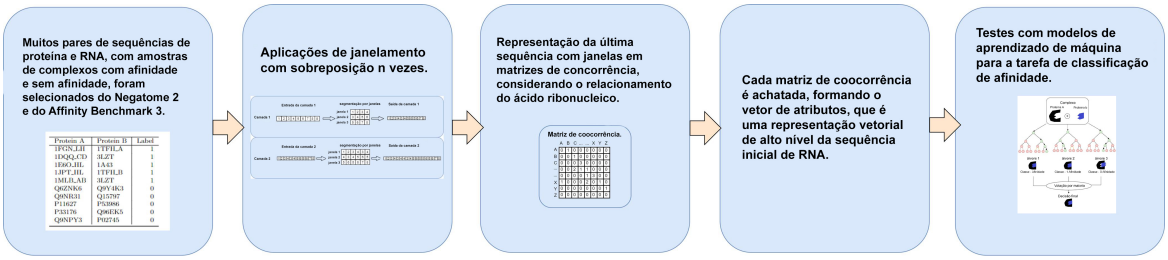


Figura 2 – Contribuição detalhada: Um total de 6.441 complexos, englobando casos de encaixe e não encaixe, foram utilizados para treinar os modelos Random Forest. Para pré-processar as sequências de RNA, cada sequência foi concatenada e submetida a subsequente segmentação em sub-sequências, considerando parâmetros de janela e sobreposição. Essas sub-sequências foram posteriormente concatenadas e tinham o potencial de passar por segmentações adicionais com base nos mesmos parâmetros de janela e sobreposição. Em seguida, a sequência concatenada resultante foi representada como matrizes de co-ocorrência de 26$\times$ 26, capturando efetivamente a distribuição dos vizinhos de RNA. Para criar o vetor de características para classificação, a forma achatada da matriz 26$\times$ 26 foi empregada como entrada.

Autoria da figura 2: O Autor

O processo de extração de características usando uma máquina pseudo-convolucional pode ser delineado da seguinte forma. Inicialmente, as sequências de RNA são concatenadas para formar uma sequência completa. Esta sequência completa é posteriormente particionada em

n sub-sequências. Para permitir sobreposição entre essas sub-sequências, o tamanho dos segmentos de sobreposição é determinado por um parâmetro passado ao método. Notavelmente, todas as n sub-sequências são então concatenadas, resultando em uma sequência estendida. Todo esse processo pode ser repetido n vezes, gerando sequências cada vez mais extensas com novas características em cada iteração.

A sequência final é representada por uma matriz de co-ocorrência de 26 por 26. Essa matriz captura a frequência de ocorrência de pares específicos de ácido ribonucleico (RNA) e fornece insights sobre a proximidade do RNA em relação a seus vizinhos correspondentes. Por exemplo, se considerarmos a sequência de RNA lida da esquerda para a direita, o elemento de RNA atual sendo analisado é denotado como E, enquanto seu vizinho adjacente é rotulado como P. Consequentemente, o elemento da matriz correspondente localizado na interseção da linha E e da coluna P é incrementado de acordo para indicar sua associação. Para uma representação visual detalhada desse processo, consulte a Figura 3.

Após a extração de características de todos os complexos, o banco de dados resultante foi particionado em conjuntos de treinamento e teste balanceados com base nos atributos de média e desvio padrão. Esse procedimento rigoroso de divisão é crucial para garantir uma avaliação robusta do modelo. A distribuição de instâncias por classe e conjunto de dados pode ser observada na Tabela 6.

Tabela 6 – Número de instâncias dos conjuntos de treinamento e teste

	Sem Docking	Docking	Total
Conjunto de Treinamento	4948	205	5153
Conjunto de Teste	1237	51	1288

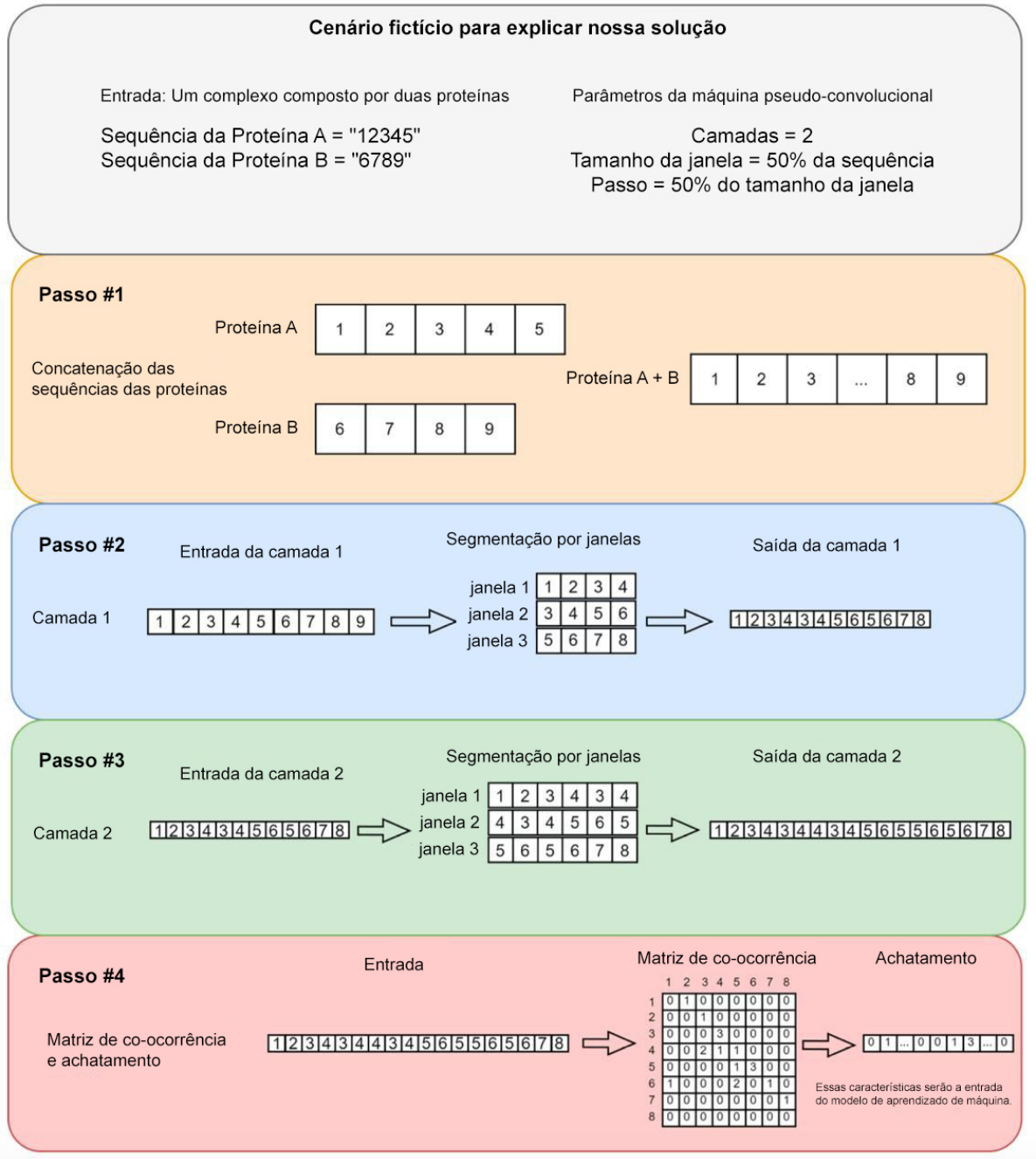


Figura 3 – Passos do método proposto: Um novo método de representação de sequências genômicas foi desenvolvido, envolvendo a análise da relação entre os ácidos ribonucleicos. Primeiro, as sequências de RNA são concatenadas, e uma janela de sobreposição é aplicada, gerando subsequências menores. Esse processo pode ser repetido n vezes, com sobreposições sucessivas. As subsequências finais são concatenadas e convertidas em uma matriz de coocorrência 26×26, onde cada célula (i, j) representa a frequência relativa com que o símbolo i é seguido do símbolo j em uma dada subsequência. A matriz inclui os 20 aminoácidos padrão mais caracteres especiais ou modificados presentes nas sequências analisadas. A matriz obtida modela a distribuição de vizinhança entre pares de símbolos na sequência e é posteriormente achatada (flattened) em um vetor de atributos numéricos, o qual serve de entrada para um classificador baseado em *Random Forests*.

Autoria da figura 3: O Autor

4.3 MÉTRICAS DE AVALIAÇÃO

Para avaliar de maneira objetiva os resultados da classificação, utilizamos os seguintes métodos: o índice κ, a *acurácia geral*, a *matriz de confusão*, a sensibilidade, a especificidade e a área sob a curva (AUC). A *matriz de confusão* para um universo de classes de interesse $\Omega = C_1, C_2, \ldots, C_m$ é uma matriz $\mathbf{T} = [t_{i,j}]m \times m$ de tamanho $m \times m$, onde cada elemento ti, j representa o número de objetos pertencentes à classe C_j, mas classificados como C_i (DUDA; HART; STORK, 2001; CORDEIRO; SANTOS; SILVA-FILHO, 2017; LIMA et al., 2015; CORDEIRO; BEZERRA; SANTOS, 2017; CORDEIRO; SANTOS; SILVA-FILHOA, 2013; CORDEIRO; SANTOS; SILVA-FILHO, 2016a; CORDEIRO; SANTOS; SILVA-FILHO, 2016b).

A *acurácia geral* é a probabilidade de que o experimento forneça resultados corretos, ou seja, classificar corretamente os pares de sequências de proteínas quanto à afinidade. Em outras palavras, é a probabilidade dos verdadeiros positivos (VP) e verdadeiros negativos (VN) entre todos os resultados. A métrica de sensibilidade indica a taxa de verdadeiros positivos, enquanto a especificidade é a taxa de verdadeiros negativos. A sigla AUC representa a Área sob a Curva ROC (Característica de Operação do Receptor), e pode ser aproximada pela média. A curva ROC, por sua vez, é um gráfico que mostra a Taxa de Verdadeiros Positivos em relação à Taxa de Falsos Positivos. Por fim, o índice Kappa é uma taxa de correlação estatística (DUDA; HART; STORK, 2001). Assim, as métricas de Acurácia, Sensibilidade, Especificidade, AUC e Índice Kappa podem ser calculadas de acordo com as equações 4.1, 4.2, 4.3, 4.4 e 4.5, respectivamente.

$$\text{Acurácia} = \rho_v = \frac{n_{TP} + n_{TN}}{n_{TP} + n_{TN} + n_{FP} + n_{FN}}, \tag{4.1}$$

$$\text{Sensibilidade} = \rho_{TP} = \frac{n_{TP}}{n_{TP} + n_{FN}}, \tag{4.2}$$

$$\text{Especificidade} = \rho_{TN} = \frac{n_{TN}}{n_{TN} + n_{FP}}, \tag{4.3}$$

$$\text{AUC} = \int_0^1 \rho_{TP} d\rho_{TN}, \tag{4.4}$$

$$\kappa = \frac{\rho_v - \rho_z}{1 - \rho_z}, \tag{4.5}$$

onde

$$\rho_z = \frac{\sum_{i=1}^{m} \left(\sum_{j=1}^{m} t_{i,j} \right) \left(\sum_{j=1}^{m} t_{j,i} \right)}{\left(\sum_{i=1}^{m} \sum_{j=1}^{m} t_{i,j} \right)^2}, \tag{4.6}$$

ρ_v é a acurácia, e $t_{i,j}$ é o elemento da matriz de confusão na posição (i, j), ou seja, o número de instâncias no conjunto de treinamento pertencentes à classe i mas classificadas como pertencentes à classe j pelo modelo de aprendizado de máquina em avaliação, para $1 \leq i, j \leq m$.

4.4 SELEÇÃO DE ATRIBUTOS E BALANCEAMENTO

A base de dados passou por duas etapas de pré-processamento: redução de atributos e balanceamento. Na primeira etapa, a seleção de atributos visou identificar as características (features) mais relevantes estatisticamente para a classificação e diminuir o tamanho do conjunto de dados necessário para avaliar a afinidade entre as proteínas. Utilizando o filtro EvolutionarySearch do software Weka, com 100 gerações e tamanho da população de 50, a computação evolucionária foi aplicada. Com isso, o número de atributos foi reduzido de 8113 para 11, focando nos mais relevantes estatisticamente.

Na segunda etapa, a base foi balanceada usando o filtro *DistributionBasedBalance*, que assume uma distribuição gaussiana dos dados e gera instâncias sintéticas baseadas nessas distribuições. Depois do balanceamento, o conjunto de dados foi dividido: 5153 instâncias (80%) para treino e 1288 instâncias (20%) para teste.

4.5 CLASSIFICADORES, MÉTRICAS E AVALIAÇÃO

Os classificadores avaliados foram: rede bayesiana e NaiveBayes; J48, random tree, random forests com 50, 100, 150 e 200 árvores; multilayer perceptron com 20, 50 e 100 neurônios na camada escondida; e support vector machine variando o parâmetro C em 0.01, 0.1 e 1 para os kernels RBF com γ de 0.01, 0.25 e 0.5, além de kernels polinomiais de graus 1, 2 e 3.

Os experimentos foram conduzidos no software Weka, na aba experimenter, utilizando validação cruzada com 10 folds e 30 repetições para cada classificador. O conjunto de treino com 5153 instâncias e 11 atributos foi empregado nos experimentos.

As métricas utilizadas para avaliar os modelos incluíram acurácia, índice kappa, sensibilidade, especificidade e área da curva ROC.

Por fim, a configuração do classificador com o melhor índice kappa foi utilizada para um teste final na aba Explorer do Weka, onde foi treinado com o conjunto de treino e avaliado com o conjunto de teste previamente separado após o balanceamento.

4.6 METODOLOGIA DE PESQUISA

O problema de predição de *docking* entre proteínas foi identificado, destacando a necessidade de uma abordagem eficiente, com o objetivo de criar um modelo de predição baseado em pseudo-convolução. A revisão bibliográfica foi realizada de forma sistemática, abrangendo a predição de *docking* de proteínas, algoritmos de aprendizado de máquina e técnicas de representação de sequências de RNA.

Para a seleção e preparação dos dados, foram escolhidas bases de dados apropriadas, como o *Affinity Benchmark 3* e o *Negatome 2*, seguidas da limpeza, pré-processamento e balanceamento dos dados para construir um conjunto de treinamento e teste representativo. A extração de características foi desenvolvida e implementada por meio de pseudo-convolução, representando as sequências de RNA das proteínas. As sequências foram divididas em sub-sequências, gerando co-matrizes de co-ocorrência para capturar as características de vizinhança.

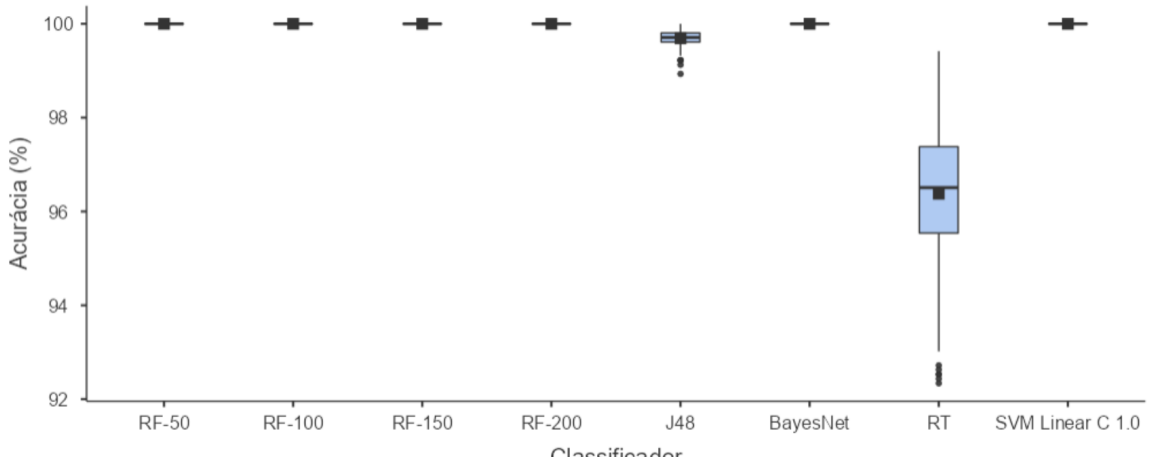
O modelo *Random Forest* foi implementado para classificar as representações de proteínas como interagindo ou não interagindo, utilizando métricas de avaliação como acurácia, sensibilidade, especificidade, AUC e índice Kappa para medir o desempenho do modelo. Os experimentos foram realizados utilizando validação cruzada para avaliar o desempenho do modelo em diferentes configurações e analisar sua robustez em relação a diferentes parâmetros e estratégias de validação.

Os resultados obtidos pelo modelo proposto foram comparados com outros métodos de predição de *docking* de proteínas da literatura, discutindo as vantagens e limitações do modelo em relação a abordagens existentes. A análise dos resultados destacou as capacidades do modelo de predição baseado em pseudo-convolução, relevando sua importância para a predição eficiente de docking de proteínas.

As principais conclusões e contribuições do estudo foram resumidas, identificando possíveis direções futuras para aprimorar e expandir a abordagem proposta. Por fim, o relatório final da pesquisa foi escrito, seguindo as normas acadêmicas e requisitos do programa de mestrado, e o artigo científico foi preparado para submissão em conferências ou periódicos relevantes da área.

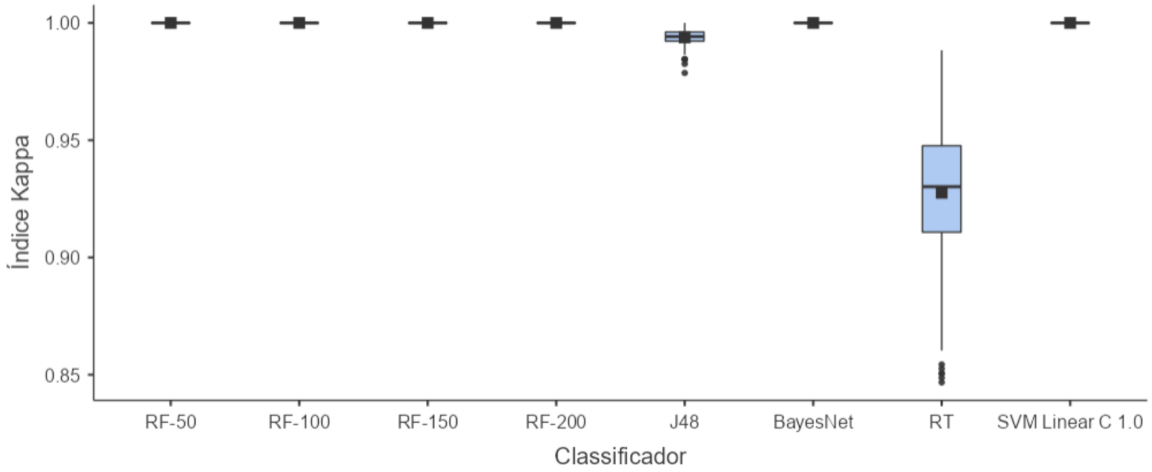
5 RESULTADOS

Para fins de comparação, diversos classificadores foram treinados utilizando a base de dados sem a aplicação de seleção de atributos, visando analisar as diferenças nos resultados obtidos com e sem a seleção de atributos. As figuras subsequentes apresentam os gráficos boxplots das métricas resultantes do treinamento dos classificadores com a base sem a seleção de atributos. Observa-se que o único classificador cujo desempenho foi inferior em comparação aos demais, quando a base de dados continha todos os atributos, foi o *Random Tree*.



Autoria da figura 4: O Autor

Figura 4 – Boxplots das acurácias obtidas com o treinamento dos classificadores utilizando a base sem seleção de atributos. RF representa o classificador Random Forest e RT o classificador Random Tree.

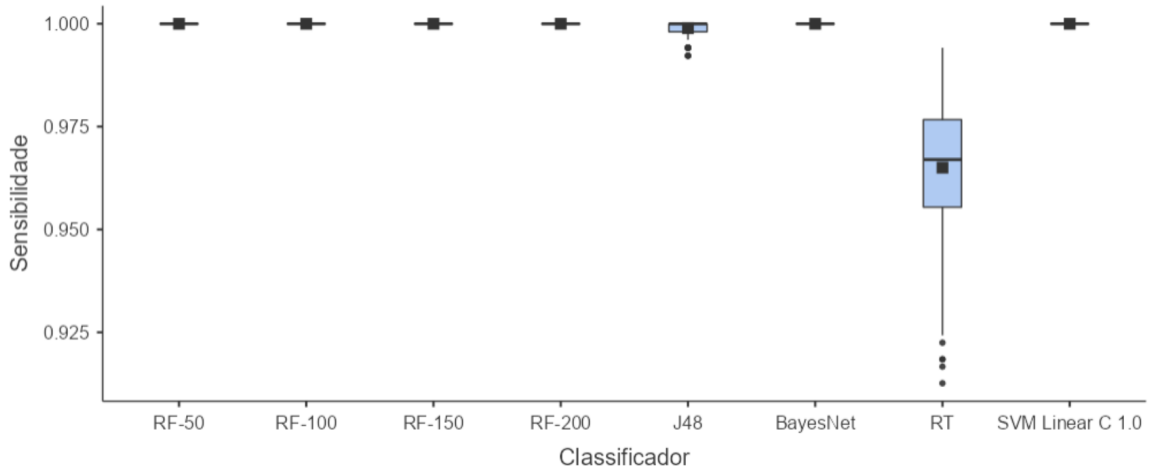


Autoria da figura 5: O Autor

Figura 5 – Boxplots dos índices kappa obtidos com o treinamento dos classificadores utilizando a base sem seleção de atributos. RF representa o classificador Random Forest e RT o classificador Random Tree.

A Tabela 7 detalha os valores obtidos de cada métrica assim como seus respectivos desvios padrões para cada configuração de classificador avaliado. As Figuras 13 a 23 mostram os gráficos boxplots das métricas para os diversos classificadores avaliados no conjunto de treino com seleção de atributos.

A partir das figuras, é possível observar que houve um leve declínio de performance nas métricas comparando sem e com seleção de atributos. Porém, a queda de performance foi pequena, principalmente quando considerando que houve uma redução de 8113 atributos para apenas 11, ou seja, uma redução de mais de 99%, reduzindo drasticamente o tempo necessário de treinamento dos modelos. Principalmente nesses casos de bases grandes e complexas, uma redução na quantidade de atributos é importantíssima para garantir a aplicabilidade dos modelos em problemas reais, onde o tempo de processamento e o custo computacional são de suma importância.

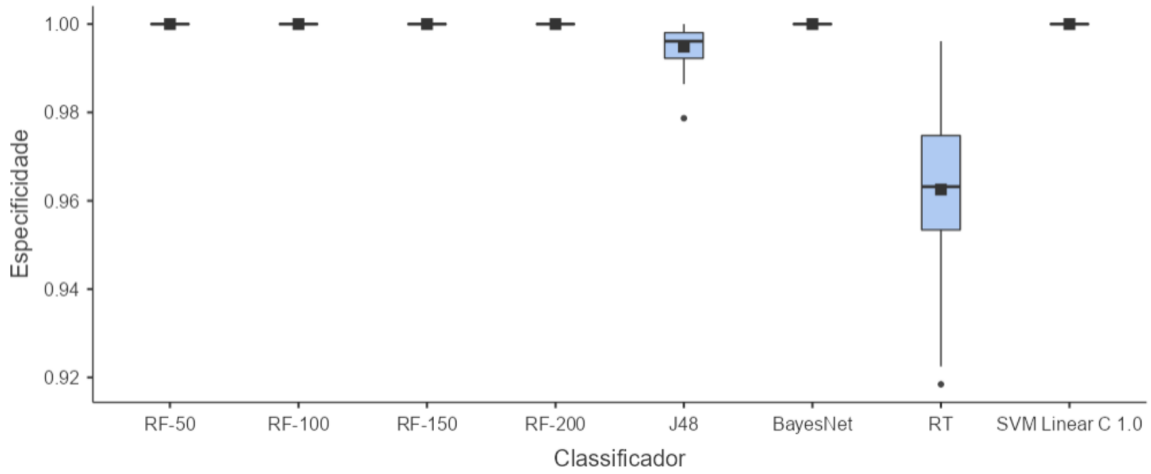


Autoria da figura 6: O Autor

Figura 6 – Boxplots das sensibilidades obtidas com o treinamento dos classificadores utilizando a base sem seleção de atributos. RF representa o classificador Random Forest e RT o classificador Random Tree.

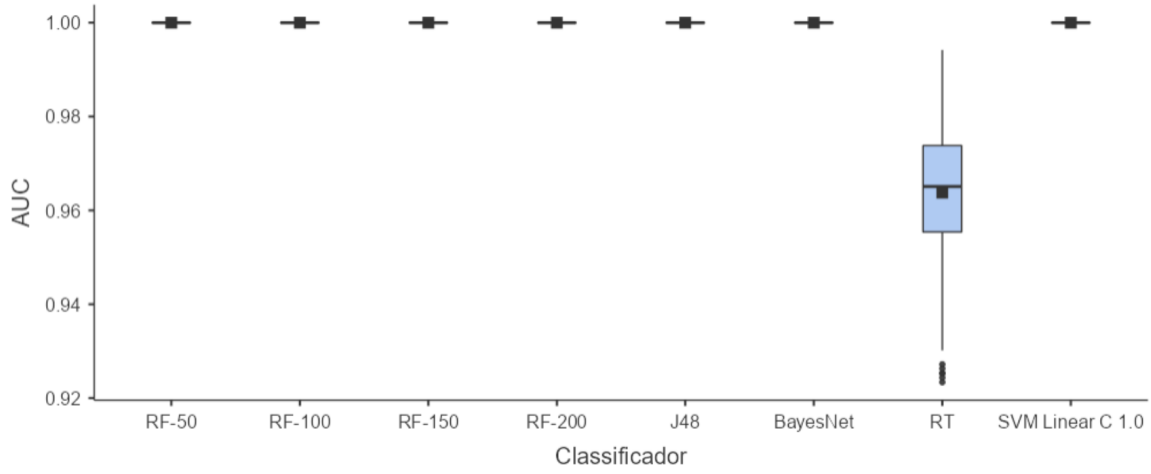
A complexidade do docking de ligantes flexíveis, abordada por Magalhães et al. (2004) (MAGALHãES; BARBOSA; DARDENNE, 2004), ressalta a necessidade de algoritmos capazes de gerenciar múltiplos graus de liberdade conformacional. Esse desafio sublinha a importância de desenvolver técnicas que possam lidar com a flexibilidade molecular, um aspecto ainda desafiador nas pesquisas atuais, como indicado por Li et al. (2015) (LI et al., 2015) e Fan et al. (2019) (FAN; FU; ZHANG, 2019). A eficácia do *docking* em sistemas complexos continua sendo uma área de pesquisa intensa.

A partir dos gráficos também observa-se que os melhores modelos estão presentes nas



Autoria da figura 7: O Autor

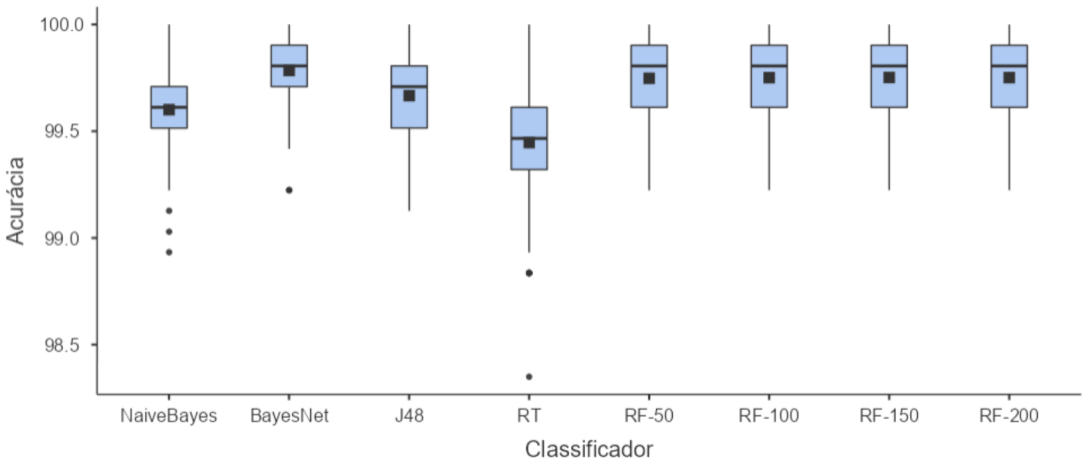
Figura 7 – Boxplots das especificidades obtidas com o treinamento dos classificadores utilizando a base sem seleção de atributos. RF representa o classificador Random Forest e RT o classificador Random Tree.



Autoria da figura 8: O Autor

Figura 8 – Boxplots das áreas embaixo da curva ROC obtidas com o treinamento dos classificadores utilizando a base sem seleção de atributos. RF representa o classificador Random Forest e RT o classificador Random Tree.

Figuras 9, 10, 11, 12 e 13.Com destaque para a rede bayesiana e *Random Forest*. As melhores configurações do classificador *SVM* foram observadas para o kernel polinomial de grau 1 (equivalente ao kernel linear), identificado como "SVM1". Esse kernel, por sua menor complexidade e custo computacional, apresentou desempenho superior em comparação com configurações mais complexas, como o kernel radial ("SVM2") e o polinomial de grau maior ("SVM3"). Esse resultado reforça que, em problemas de classificação com dados bem separáveis no espaço original, modelos mais simples como o *SVM* linear podem ser mais eficazes do que alternativas mais sofisticadas, reduzindo riscos de sobreajuste sem sacrificar a acurácia. Não houve dife-



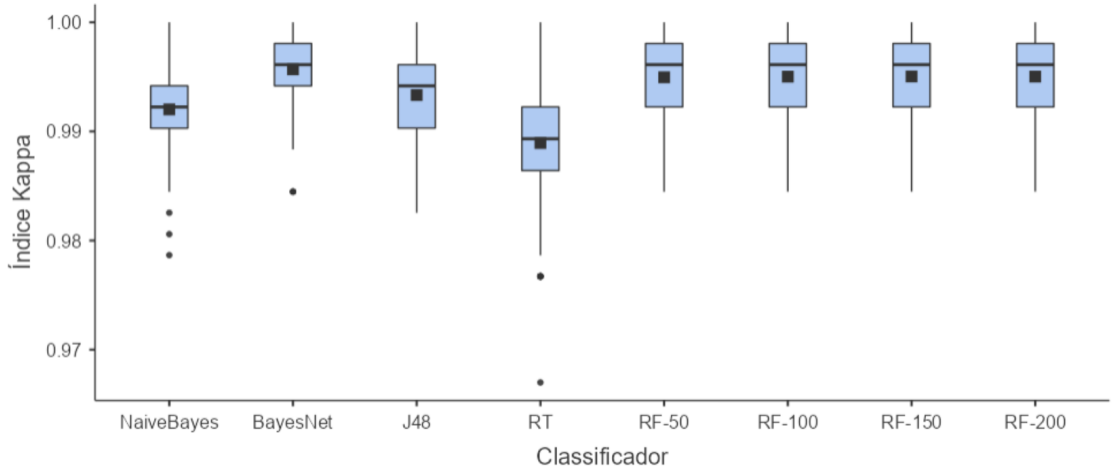
Autoria da figura 8: O Autor

Figura 9 – Boxplots das acurácias obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. RF representa o classificador Random Forest e RT o classificador Random Tree.

Autoria da figura 9: O Autor

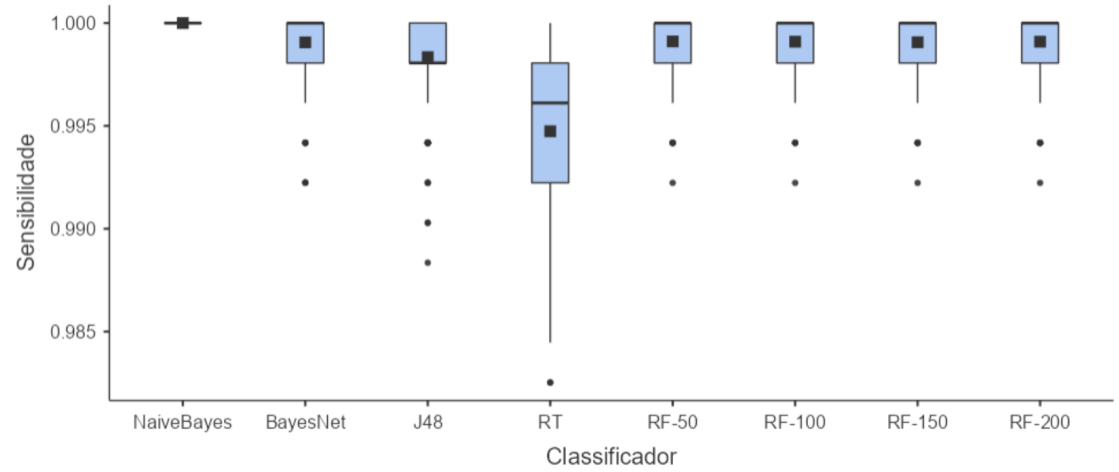
Tabela 7 – Métricas obtidas a partir da base de treinamento com seleção de atributos e balanceamento. As métricas avaliadas foram a acurácia, índice kappa, sensibilidade, especificidade e AUC.

Classificadores	**Acurácia (%)**	**Índice Kappa**	**Sensibilidade**	**Especificidade**	**AUC**
NaiveBayes	99.6 (0.181)	0.992 (0.003)	1.000 (0.000)	0.992 (0.003)	1.000 (0.000)
BayesNet	99.8 (0.151)	0.996 (0.003)	0.999 (0.001)	0.997 (0.002)	1.000 (0.000)
J48	99.7 (0.169)	0.993 (0.003)	0.998 (0.002)	0.995 (0.003)	0.997 (0.002)
RT	99.4 (0.259)	0.989 (0.005)	0.995 (0.003)	0.994 (0.003)	0.994 (0.002)
RF-50	99.7 (0.152)	0.995 (0.003)	0.999 (0.001)	0.996 (0.003)	0.999 (0.000)
RF-100	99.8 (0.148)	0.995 (0.003)	0.999 (0.001)	0.996 (0.003)	0.999 (0.000)
RF-150	99.8 (0.150)	0.995 (0.003)	0.999 (0.001)	0.996 (0.003)	0.999 (0.000)
RF-200	99.8 (0.149)	0.995 (0.003)	0.999 (0.001)	0.996 (0.003)	0.999 (0.000)
MLP 20	94.2 (1.61)	0.884 (0.032)	0.962 (0.029)	0.922 (0.030)	0.981 (0.008)
MLP 50	94.5 (1.73)	0.890 (0.035)	0.963 (0.031)	0.927 (0.029)	0.982 (0.009)
MLP 100	94.7 (1.60)	0.894 (0.032)	0.966 (0.026)	0.928 (0.028)	0.983 (0.008)
SVM 1	96.0 (0.568)	0.921 (0.012)	1.000 (0.000)	0.921 (0.012)	0.960 (0.006)
SVM 2	98.1 (0.418)	0.961 (0.009)	1.000 (0.000)	0.961 (0.009)	0.980 (0.004)
SVM 3	99.0 (0.307)	0.979 (0.006)	1.000 (0.000)	0.979 (0.006)	0.990 (0.003)
SVM 4	70.9 (2.81)	0.418 (0.056)	1.000 (0.000)	0.418 (0.056)	0.709 (0.028)
SVM 5	84.1 (2.56)	0.682 (0.050)	0.997 (0.005)	0.685 (0.050)	0.841 (0.025)
SVM 6	90.2 (1.99)	0.805 (0.039)	0.997 (0.005)	0.807 (0.040)	0.902 (0.020)
SVM 7	60.8 (2.54)	0.216 (0.050)	1.000 (0.000)	0.216 (0.051)	0.608 (0.025)
SVM 8	76.5 (2.45)	0.530 (0.049)	0.998 (0.004)	0.532 (0.049)	0.765 (0.024)
SVM 9	85.3 (2.57)	0.705 (0.049)	0.998 (0.004)	0.708 (0.049)	0.853 (0.025)
SVM 10	74.5 (3.27)	0.490 (0.065)	1.000 (0.000)	0.490 (0.065)	0.745 (0.032)
SVM 11	67.8 (11.2)	0.356 (0.225)	0.900 (0.302)	0.456 (0.264)	0.678 (0.112)
SVM 12	86.0 (2.26)	0.719 (0.045)	0.994 (0.007)	0.725 (0.045)	0.860 (0.023)
SVM 13	79.1 (10.5)	0.581 (0.210)	0.938 (0.079)	0.644 (0.263)	0.791 (0.105)
SVM 14	87.6 (2.09)	0.753 (0.042)	0.996 (0.006)	0.757 (0.042)	0.876 (0.020)
SVM 15	92.4 (1.80)	0.848 (0.036)	0.998 (0.004)	0.850 (0.036)	0.924 (0.018)
SVM 16	82.5 (2.44)	0.650 (0.049)	0.996 (0.005)	0.653 (0.049)	0.825 (0.024)
SVM 17	89.5 (1.93)	0.791 (0.038)	0.997 (0.005)	0.794 (0.039)	0.895 (0.019)
SVM 18	93.2 (1.69)	0.864 (0.034)	0.999 (0.004)	0.865 (0.034)	0.932 (0.017)



Autoria da figura 10: O Autor

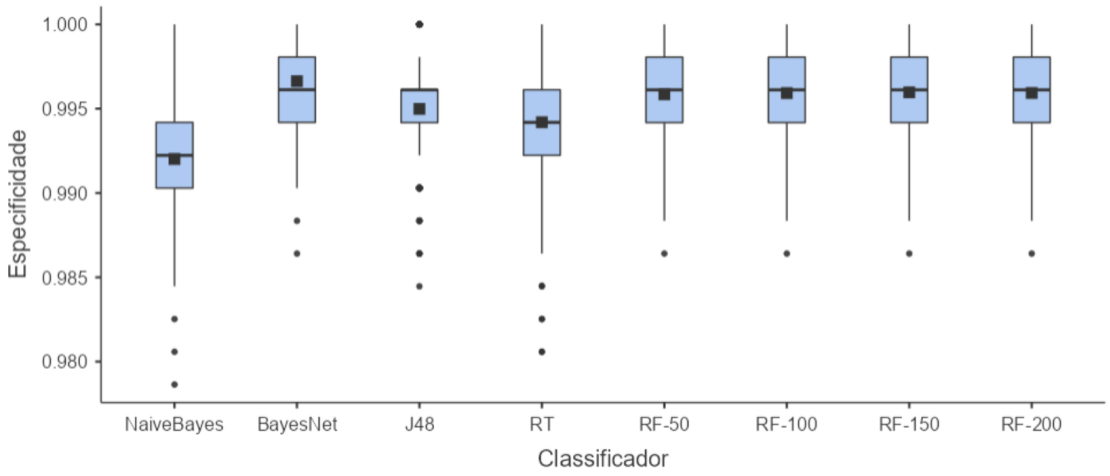
Figura 10 – Boxplots dos índices kappa obtidos com o treinamento dos classificadores utilizando a base com seleção de atributos. RF representa o classificador *Random Forest* e RT o classificador *Random Tree*.



Autoria da figura 11: O Autor

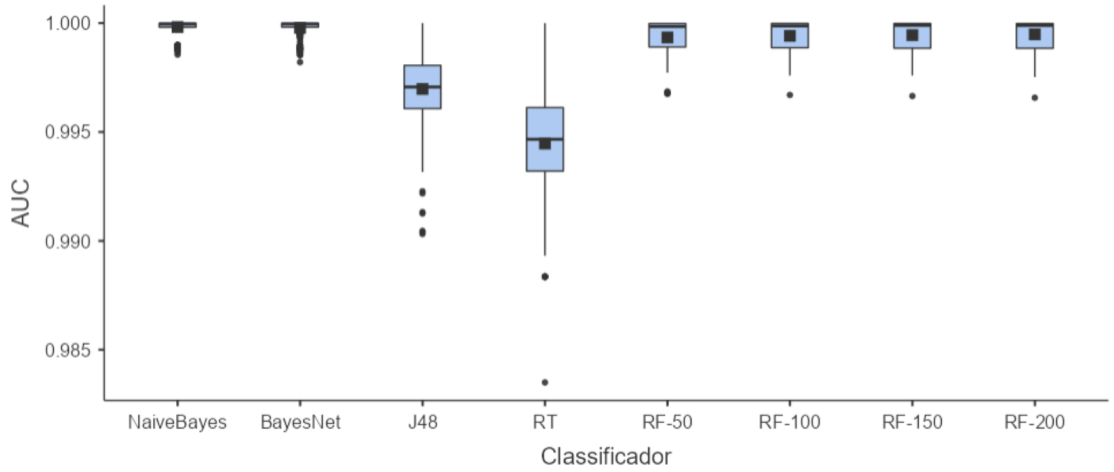
Figura 11 – Boxplots das sensibilidades obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. RF representa o classificador Random Forest e RT o classificador Random Tree.

rença significativa entre as diferentes configurações de MLP, não afetando as métricas com o aumento de neurônios na camada escondida. Alguns classificadores, apesar de terem obtido acurácia e índice kappa não tão bons, obtiveram uma sensibilidade máxima. Em contrapartida, a especificidade desses classificadores foi baixa, motivo pelo qual a acurácia e kappa também tiveram seus valores puxados para baixo. A sensibilidade foi a métrica em que a maior parte dos classificadores obtiveram ótimos resultados, chegando até a valores máximos. Isso indica que esses modelos tendem a serem muito bons em identificar casos positivos (alta sensibilidade), porém muito ruins em identificar casos negativos (especificidade) para esta base de



Autoria da figura 12: O Autor

Figura 12 – Boxplots das especificidades obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. RF representa o classificador Random Forest e RT o classificador Random Tree.

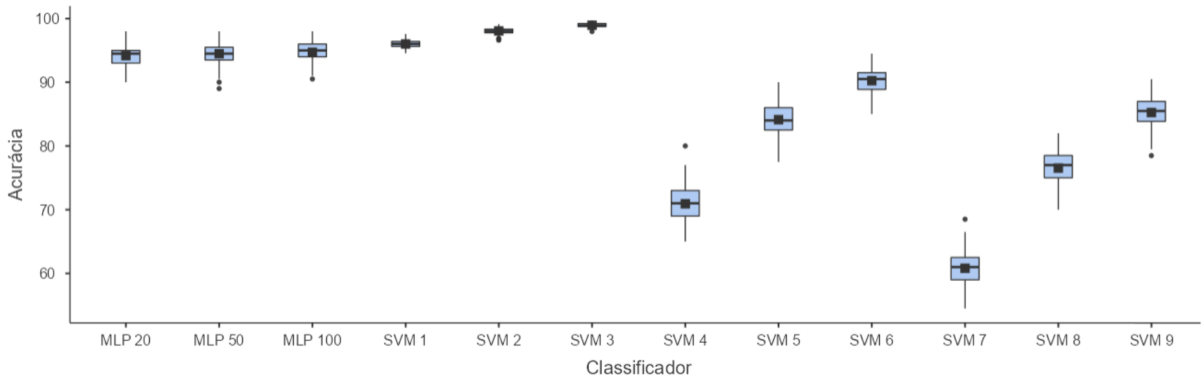


Autoria da figura 13: O Autor

Figura 13 – Boxplots das áreas embaixo da curva ROC obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. RF representa o classificador Random Forest e RT o classificador Random Tree.

docking específica. É imprescindível na escolha do melhor modelo levar em consideração esses aspectos, em que caso seja de suma importância a identificação de casos negativos é preciso que a especificidade seja priorizada na hora de escolha do melhor modelo, caso contrário a sensibilidade seria priorizada.

No presente estudo, a priorização foi do índice *kappa*. Essa escolha se justifica porque o índice *kappa* avalia o grau de concordância entre as predições do modelo e os valores reais, ajustando para o acaso, o que o torna uma métrica mais robusta em cenários com classes desbalanceadas. Enquanto métricas como acurácia podem ser infladas artificialmente



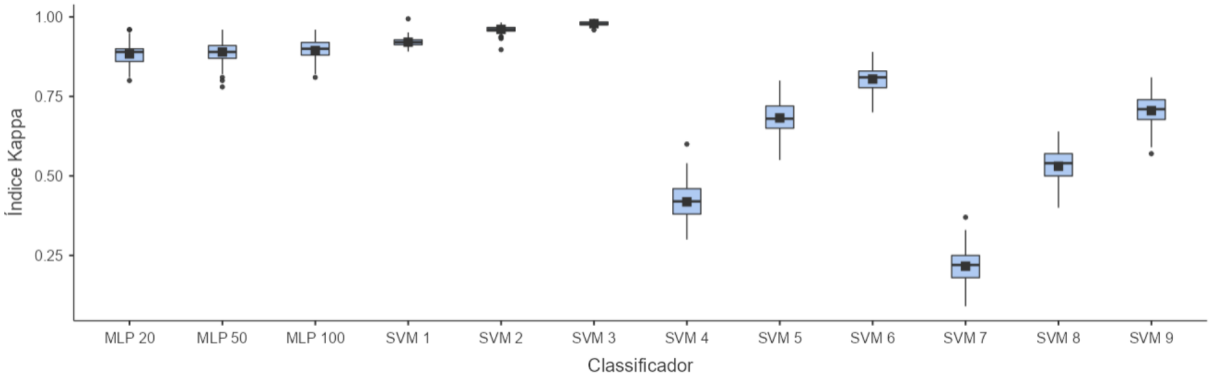
Autoria da figura 14: O Autor

Figura 14 – Boxplots das acurácias obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. SVM 1, 2 e 3 são SVM com kernels polinomiais de grau 1, com C igual a 0.01, 0.1 e 1, respectivamente. SVM 4, 5 e 6 são SVM com kernels polinomiais de grau 2, com C igual a 0.01, 0.1 e 1, respectivamente. SVM 7, 8 e 9 são SVM com kernels polinomiais de grau 3, com C igual a 0.01, 0.1 e 1, respectivamente.

quando há predominância de uma das classes, o *kappa* penaliza acordos esperados por chance, refletindo de forma mais fidedigna a capacidade discriminativa do modelo. Logo, a partir do índice *kappa*, o melhor modelo escolhido foi o *Random Forest* com 200 árvores. O teste final com o conjunto de teste separado foi então realizado com o modelo RF-200. As métricas obtidas estão na Tabela 8. É possível observar que todas as métricas obtiveram resultados altíssimos, chegando a 1 para a especificidade e AUC. Contrário da maioria dos classificadores que tenderam a obter melhores resultados na sensibilidade em vez da especificidade.

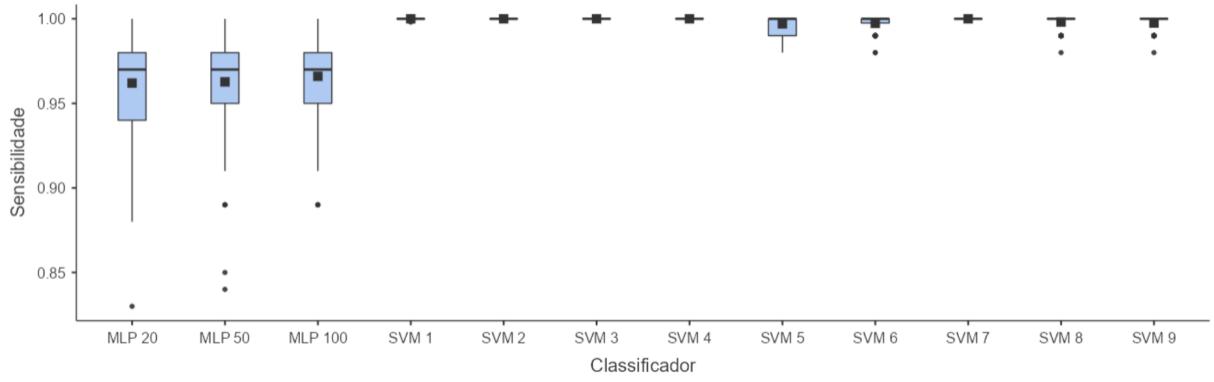
Tabela 8 – Métricas obtidas no teste final com o conjunto de teste para o classificador Random Forest com 200 árvores. As métricas avaliadas foram a acurácia, índice kappa, sensibilidade, especificidade e AUC.

Classificadores	**Acurácia (%)**	**Índice Kappa**	**Sensibilidade**	**Especificidade**	**AUC**
RF-200	99.8%	0.997	0.997	1.000	1.000



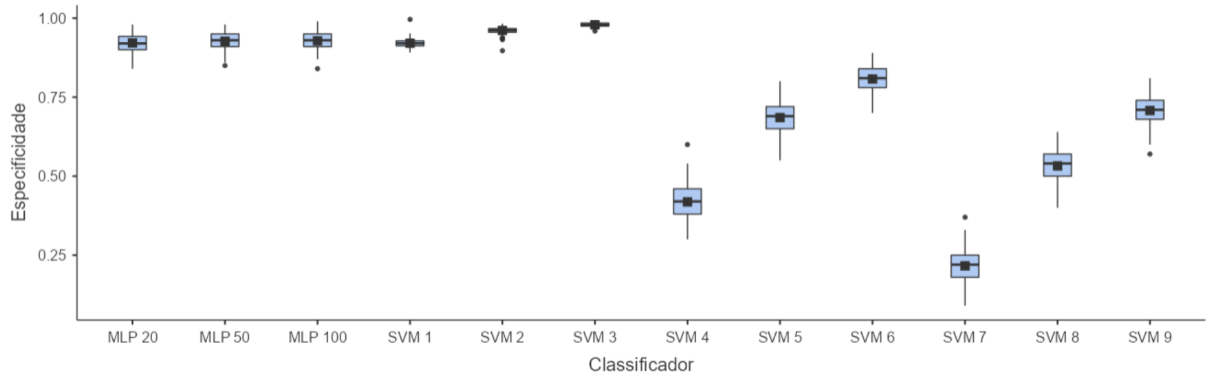
Autoria da figura 15: O Autor

Figura 15 – Boxplots dos índices kappa obtidos com o treinamento dos classificadores utilizando a base com seleção de atributos. SVM 1, 2 e 3 são SVM com kernels polinomiais de grau 1, com C igual a 0.01, 0.1 e 1, respectivamente. SVM 4, 5 e 6 são SVM com kernels polinomiais de grau 2, com C igual a 0.01, 0.1 e 1, respectivamente. SVM 7, 8 e 9 são SVM com kernels polinomiais de grau 3, com C igual a 0.01, 0.1 e 1, respectivamente.



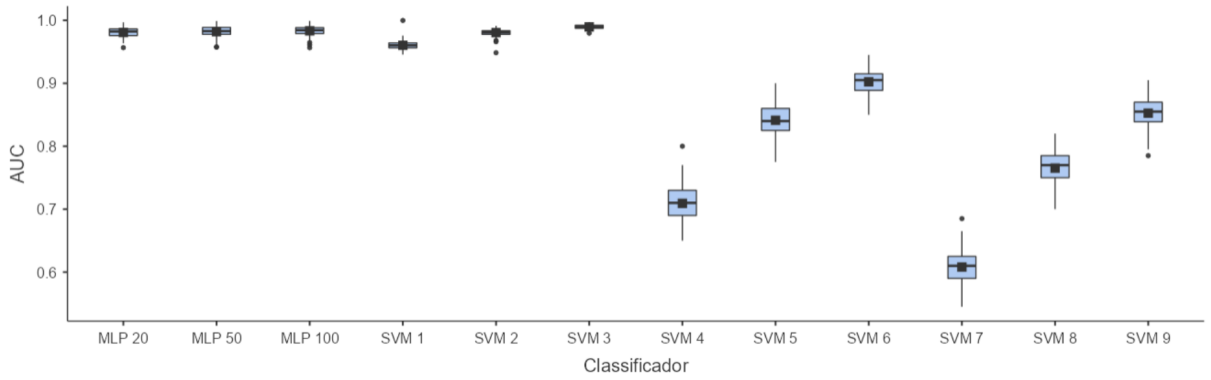
Autoria da figura 16: O Autor

Figura 16 – Boxplots das sensibilidades obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. SVM 1, 2 e 3 são SVM com kernels polinomiais de grau 1, com C igual a 0.01, 0.1 e 1, respectivamente. SVM 4, 5 e 6 são SVM com kernels polinomiais de grau 2, com C igual a 0.01, 0.1 e 1, respectivamente. SVM 7, 8 e 9 são SVM com kernels polinomiais de grau 3, com C igual a 0.01, 0.1 e 1, respectivamente.



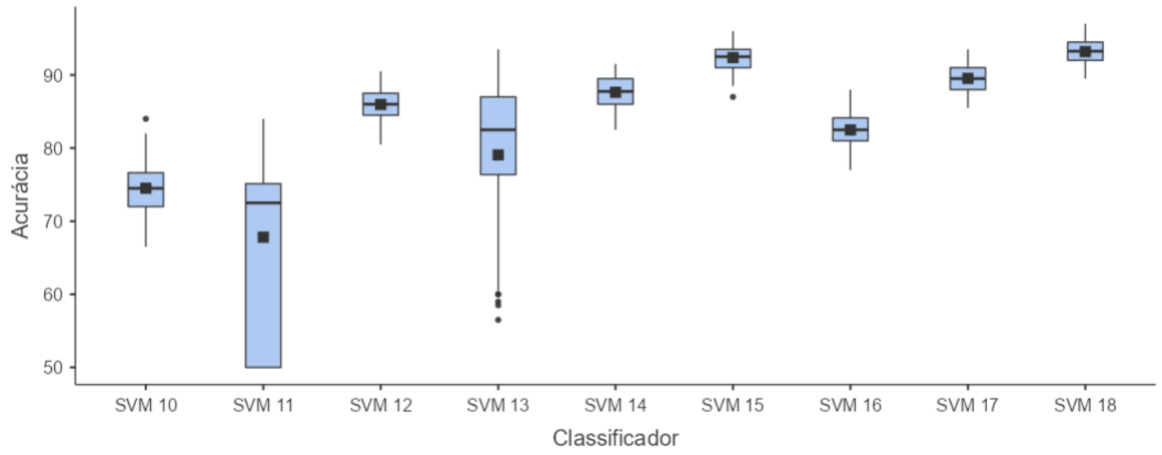
Autoria da figura 17: O Autor

Figura 17 – Boxplots das especificidades obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. SVM 1, 2 e 3 são SVM com kernels polinomiais de grau 1, com C igual a 0.01, 0.1 e 1, respectivamente. SVM 4, 5 e 6 são SVM com kernels polinomiais de grau 2, com C igual a 0.01, 0.1 e 1, respectivamente. SVM 7, 8 e 9 são SVM com kernels polinomiais de grau 3, com C igual a 0.01, 0.1 e 1, respectivamente.



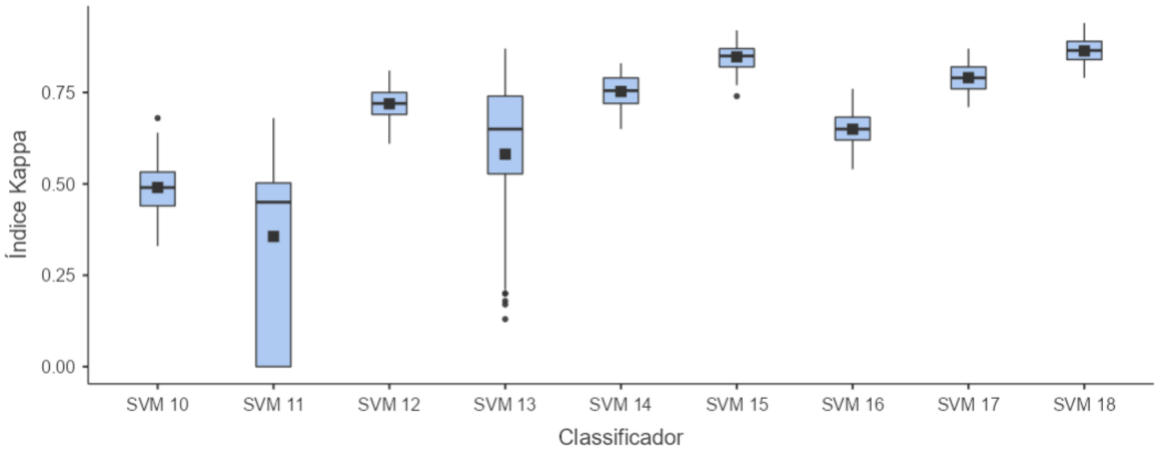
Autoria da figura 18: O Autor

Figura 18 – Boxplots das áreas embaixo da curva ROC obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. SVM 1, 2 e 3 são SVM com kernels polinomiais de grau 1, com C igual a 0.01, 0.1 e 1, respectivamente. SVM 4, 5 e 6 são SVM com kernels polinomiais de grau 2, com C igual a 0.01, 0.1 e 1, respectivamente. SVM 7, 8 e 9 são SVM com kernels polinomiais de grau 3, com C igual a 0.01, 0.1 e 1, respectivamente.



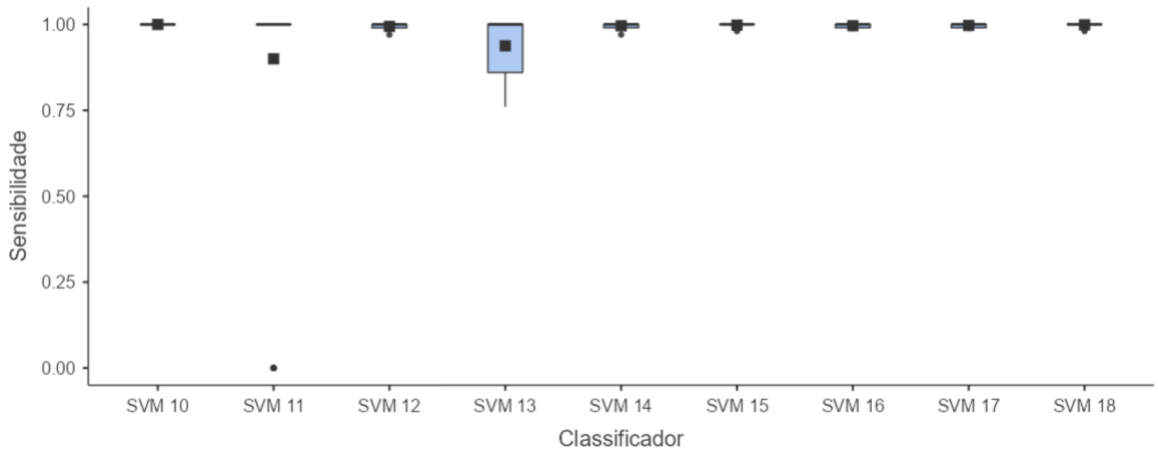
Autoria da figura 19: O Autor

Figura 19 – Boxplots das acurácias obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. SVM 10, 11 e 12 são SVM com kernels RBF com $\gamma = 0.01$, e com C igual a 0.01, 0.1 e 1, respectivamente. SVM 13, 14 e 15 são SVM com kernels RBF com $\gamma = 0.25$, e com C igual a 0.01, 0.1 e 1, respectivamente. SVM 16, 17 e 18 são SVM com kernels RBF com $\gamma = 0.5$, e com C igual a 0.01, 0.1 e 1, respectivamente.



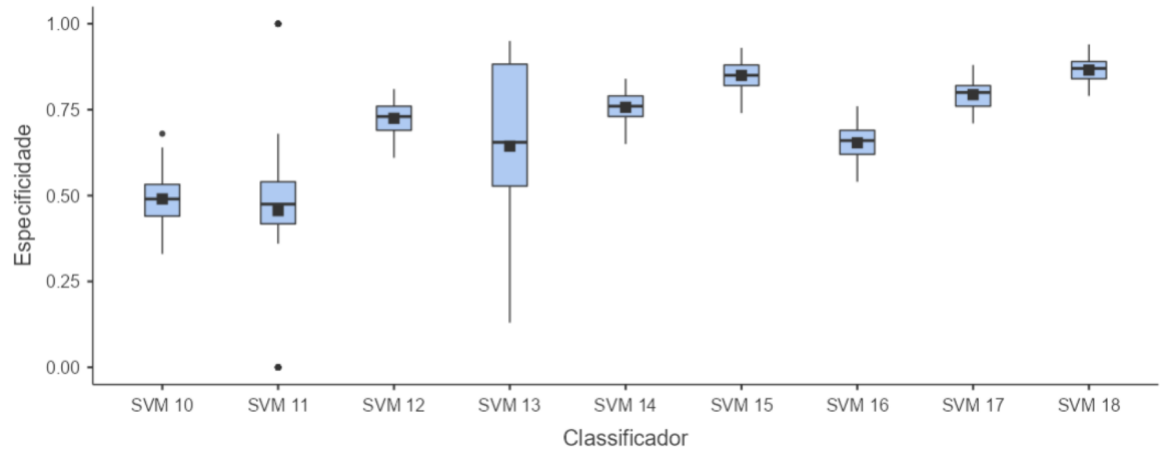
Autoria da figura 20: O Autor

Figura 20 – Boxplots dos índices kappa obtidos com o treinamento dos classificadores utilizando a base com seleção de atributos. SVM 10, 11 e 12 são SVM com kernels RBF com $\gamma = 0.01$, e com C igual a 0.01, 0.1 e 1, respectivamente. SVM 13, 14 e 15 são SVM com kernels RBF com $\gamma = 0.25$, e com C igual a 0.01, 0.1 e 1, respectivamente. SVM 16, 17 e 18 são SVM com kernels RBF com $\gamma = 0.5$, e com C igual a 0.01, 0.1 e 1, respectivamente.



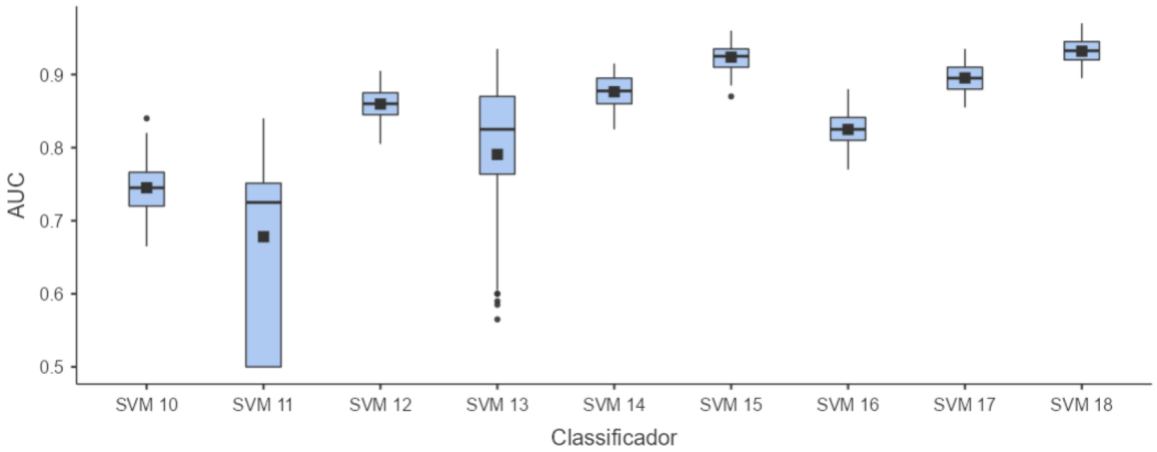
Autoria da figura 21: O Autor

Figura 21 – Boxplots das sensibilidades obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. SVM 10, 11 e 12 são SVM com kernels RBF com $\gamma = 0.01$, e com C igual a 0.01, 0.1 e 1, respectivamente. SVM 13, 14 e 15 são SVM com kernels RBF com $\gamma = 0.25$, e com C igual a 0.01, 0.1 e 1, respectivamente. SVM 16, 17 e 18 são SVM com kernels RBF com $\gamma = 0.5$, e com C igual a 0.01, 0.1 e 1, respectivamente.



Autoria da figura 22: O Autor

Figura 22 – Boxplots das especificidades obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. SVM 10, 11 e 12 são SVM com kernels RBF com $\gamma = 0.01$, e com C igual a 0.01, 0.1 e 1, respectivamente. SVM 13, 14 e 15 são SVM com kernels RBF com $\gamma = 0.25$, e com C igual a 0.01, 0.1 e 1, respectivamente. SVM 16, 17 e 18 são SVM com kernels RBF com $\gamma = 0.5$, e com C igual a 0.01, 0.1 e 1, respectivamente.



Autoria da figura 23: O Autor

Figura 23 – Boxplots das áreas embaixo da curva ROC obtidas com o treinamento dos classificadores utilizando a base com seleção de atributos. SVM 10, 11 e 12 são SVM com kernels RBF com $\gamma = 0.01$, e com C igual a 0.01, 0.1 e 1, respectivamente. SVM 13, 14 e 15 são SVM com kernels RBF com $\gamma = 0.25$, e com C igual a 0.01, 0.1 e 1, respectivamente. SVM 16, 17 e 18 são SVM com kernels RBF com $\gamma = 0.5$, e com C igual a 0.01, 0.1 e 1, respectivamente.

6 CONCLUSÃO

6.1 CONCLUSÕES GERAIS

Com base nos resultados apresentados, é possível concluir que a abordagem híbrida inteligente desenvolvida para a estimativa de *docking* molecular entre proteínas demonstrou avanços significativos em comparação com métodos tradicionais. O método propôs a integração entre redes de pseudo-convolução, responsáveis pela extração de características a partir de sequências aminoacídicas, e o classificador *Random Forest*, com o objetivo de predizer a afinidade de interação entre pares de proteínas.

A metodologia obteve métricas de desempenho excepcionais, destacando-se uma acurácia de 99,8%, sensibilidade de 0,997, especificidade de 1,000, área sob a curva ROC (AUC) de 1,000 e índice *kappa* de 0,997 — este último adotado como principal critério de decisão por ser uma métrica mais robusta em cenários com possíveis desbalanceamentos de classe, ao penalizar acertos obtidos ao acaso.

Outro ponto de destaque foi a expressiva redução do número de atributos utilizados: de 8.113 para apenas 11. Essa redução impactou diretamente no tempo de treinamento dos modelos, tornando o processo mais eficiente computacionalmente, com ganhos de desempenho e menor consumo de recursos, sem comprometer a precisão dos resultados. A seleção de atributos, portanto, contribuiu para a criação de um modelo mais leve, interpretável e escalável.

A abordagem proposta demonstrou, ainda, excelente capacidade de generalização, sendo capaz de lidar com diferentes tipos de proteínas e cenários de interação molecular, o que reforça seu potencial para ser aplicada em pipelines de triagem virtual e descoberta de fármacos. Dado o seu desempenho elevado e sua eficiência estrutural, a metodologia contribui significativamente para o avanço do estado da arte em bioinformática e engenharia biomédica, abrindo caminhos promissores para futuras investigações, incluindo o uso em terapias personalizadas, reposicionamento de fármacos e design racional de moléculas bioativas.

6.2 DIFICULDADES APRESENTADAS

A pesquisa enfrentou diversas dificuldades ao longo de sua execução, sendo a principal delas a complexidade intrínseca ao processamento e análise de grandes volumes de dados proteicos. A combinação de múltiplas bases de dados, como o Affinity Benchmark 3 e o Negatome

2, demandou um esforço significativo de limpeza, pré-processamento e balanceamento para garantir a representatividade e a qualidade dos dados utilizados. Para superar esse desafio, foi realizada uma rigorosa seleção de atributos, reduzindo o número de características de 8.113 para apenas 11, utilizando métodos de computação evolucionária. Esse processo foi essencial para minimizar a carga computacional e aumentar a eficiência das análises.

Outra dificuldade notável foi o desenvolvimento de uma metodologia que eliminasse a necessidade de simulações tridimensionais, tradicionalmente utilizadas em métodos de docking molecular. Esse desafio foi superado com o uso de redes de pseudo-convolução, que permitiram extrair informações estruturais diretamente de sequências de aminoácidos. Para implementar e validar essa abordagem, foi necessário projetar e ajustar parâmetros do modelo, o que envolveu extensos testes com diferentes configurações de algoritmos, como Random Forest e SVM. A validação cruzada com múltiplas repetições e o uso de métricas robustas de avaliação asseguraram a confiabilidade dos resultados, superando as limitações iniciais de precisão e sensibilidade.

Além disso, dificuldades técnicas relacionadas ao desempenho computacional também foram enfrentadas. A execução de experimentos com validação cruzada para um conjunto de dados de alta dimensionalidade exigiu infraestrutura computacional avançada. Para contornar essa limitação, foi adotado o uso de ferramentas de software como Weka e técnicas de balanceamento de carga para otimizar o processamento das simulações, permitindo a conclusão eficiente da pesquisa dentro dos prazos estabelecidos.

6.3 CONTRIBUIÇÕES E TRABALHOS FUTUROS

A pesquisa realizada contribuiu significativamente para a área de bioinformática e inteligência artificial aplicadas à saúde, destacando o impacto de metodologias baseadas em aprendizado de máquina na resolução de problemas complexos, como a identificação de interações moleculares e o diagnóstico precoce de doenças. Esses avanços são especialmente relevantes para áreas como descoberta de medicamentos. O trabalho resultou na publicação de resumos expandidos apresentados no VII Simpósio de Inovação em Engenharia Biomédica.

Entre os resumos publicados, destacam-se: "Estratégia de docking utilizando sistemas inteligentes baseados em aprendizado de máquina", que abordou avanços na bioinformática para estimar interações proteína-proteína, "Diagnóstico da doença de Alzheimer através de eletroencefalograma e recursos de aprendizado de máquina: uma revisão narrativa", que analisou o

uso de técnicas de aprendizado de máquina para o apoio diagnóstico da doença de Alzheimer; "Personalized Rehabilitation: 3D Printing and Molding for Cerebral Palsy Hand Orthoses", que apresentou uma abordagem para a produção de órteses personalizadas utilizando impressora 3D para pacientes com paralisia cerebral; "Predição e identificação da diabetes a partir da análise de parâmetros clínicos e laboratoriais utilizando aprendizado de máquina", que explorou diferentes algoritmos para a classificação da diabetes; e "Development of an artificial intelligence to assist the ambulation of individuals in exoskeletons", que propôs o uso de inteligência artificial para otimizar a marcha em exoesqueletos.

Para dar continuidade a essa linha de pesquisa, há pontos de melhoria e possibilidades de avanço. Recomenda-se ampliar os conjuntos de dados utilizados, integrando bases mais robustas e diversas, a fim de aumentar a generalização e a precisão dos modelos. Além disso, é essencial validar os resultados por meio de experimentos in vitro ou in vivo, aproximando a pesquisa do contexto clínico. Aplicar os modelos em cenários reais de saúde, como diagnósticos e terapias personalizadas, também se mostra como uma importante etapa futura.

Outro ponto relevante é a exploração de novos algoritmos, como redes neurais profundas, aprendizado por reforço e modelos generativos, que podem ampliar ainda mais a capacidade de predição e personalização. Adicionalmente, fomentar colaborações interdisciplinares, envolvendo áreas como farmacologia, engenharia mecânica e outras ciências da saúde, permitirá maior aplicação prática dos resultados obtidos. Essas direções fornecem uma base sólida para que futuros pesquisadores possam expandir e aplicar as metodologias desenvolvidas, promovendo avanços contínuos e relevantes na biomedicina e inteligência artificial.

REFERÊNCIAS

AFTAB, S. O.; GHOURI, M. Z.; MASOOD, M. U.; HAIDER, Z.; KHAN, Z.; AHMAD, A.; MUNAWAR, N. Analysis of SARS-CoV-2 RNA-dependent RNA polymerase as a potential therapeutic drug target using a computational approach. *Journal of Translational Medicine*, BioMed Central, v. 18, n. 1, p. 1–15, 2020.

AGNIHOTRY, S.; PATHAK, R. K.; SRIVASTAV, A.; SHUKLA, P. K.; GAUTAM, B. Molecular docking and structure-based drug design. *Computer-aided drug design*, Springer, p. 115–131, 2020.

AGRAWAL, P. Artificial intelligence in drug discovery and development. *Journal of Pharmacovigilance*, v. 6, n. 2, p. 1000e173, 2018.

AGUADO, B. A.; GRIM, J. C.; ROSALES, A. M.; WATSON-CAPPS, J. J.; ANSETH, K. S. Engineering precision biomaterials for personalized medicine. *Science translational medicine*, American Association for the Advancement of Science, v. 10, n. 424, p. eaam8645, 2018.

AHMED, A.; MAM, B.; SOWDHAMINI, R. Deelig: A deep learning approach to predict protein-ligand binding affinity. *Bioinformatics and Biology Insights*, SAGE Publications Sage UK: London, England, v. 15, p. 11779322211030364, 2021.

AJMERA, P.; JAIN, V. Modelling the barriers of health 4.0 – the fourth healthcare industrial revolution in india by tism. *Operations Management Research*, Springer, v. 12, n. 3-4, p. 129–145, 2019.

AL-KHAFAJI, K.; AL-DUHAIDAHAWI, D.; TOK, T. T. Using integrated computational approaches to identify safe and rapid treatment for sars-cov-2. *Journal of Biomolecular Structure and Dynamics*, Taylor & Francis, v. 39, n. 9, p. 3387–3395, 2021.

ALGHAMEDY, F.; BOPAIAH, J.; JONES, D.; ZHANG, X.; WEISS, H. L.; ELLINGSON, S. R. Incorporating protein dynamics through ensemble docking in machine learning models to predict drug binding. *AMIA Summits on Translational Science Proceedings*, American Medical Informatics Association, v. 2018, p. 26, 2018.

ALONSO, H.; BLIZNYUK, A. A.; GREADY, J. E. Combining docking and molecular dynamic simulations in drug design. *Medicinal research reviews*, Wiley Online Library, v. 26, n. 5, p. 531–568, 2006.

ALTUNTAŞ, S.; BOZKUS, Z.; FRAGUELA, B. B. Gpu accelerated molecular docking simulation with genetic algorithms. In: SPRINGER. *Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30–April 1, 2016, Proceedings, Part II 19*. [S.l.], 2016. p. 134–146.

ASHTAWY, H. M.; MAHAPATRA, N. R. Molecular docking for drug discovery: Machine-learning approaches for native pose prediction of protein-ligand complexes. In: SPRINGER. *Computational Intelligence Methods for Bioinformatics and Biostatistics: 10th International Meeting, CIBB 2013, Nice, France, June 20-22, 2013, Revised Selected Papers 10*. [S.l.], 2014. p. 15–32.

AWAD, A.; TRENFIELD, S. J.; POLLARD, T. D.; ONG, J. J.; ELBADAWI, M.; MCCOUBREY, L. E.; GOYANES, A.; GAISFORD, S.; BASIT, A. W. Connected healthcare: Improving patient care using digital health technologies. *Advanced Drug Delivery Reviews*, Elsevier, v. 178, p. 113958, 2021.

AZEVEDO, W. W.; LIMA, S. M.; FERNANDES, I. M.; ROCHA, A. D.; CORDEIRO, F. R.; SILVA-FILHO, A. G. da; SANTOS, W. P. dos. Fuzzy morphological extreme learning machines to detect and classify masses in mammograms. In: IEEE. *2015 IEEE International Conference on Fuzzy Systems (fuzz-IEEE)*. [S.l.], 2015. p. 1–8.

BALI, J.; BALI, R. T. India and the fourth industrial revolution: How we should approach artificial intelligence in healthcare and biomedical research? *The Journal of the Association of Physicians of India*, v. 68, n. 3, p. 72–74, 2020.

BALLESTER, P. J.; MITCHELL, J. B. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, Oxford University Press, v. 26, n. 9, p. 1169–1175, 2010.

BANERJEE, A.; CHAKRABORTY, C.; KUMAR, A.; BISWAS, D. Emerging trends in iot and big data analytics for biomedical and health care technologies. *Handbook of data science approaches for biomedical engineering*, Elsevier, p. 121–152, 2020.

BARBOSA, V. A. d. F.; GOMES, J. C.; SANTANA, M. A. de; LIMA, C. L. de; CALADO, R. B.; JUNIOR, C. R. B.; ALBUQUERQUE, J. E. d. A.; SOUZA, R. G. de; ARAÚJO, R. J. E. de; JUNIOR, L. A. R. M.; SOUZA, R. E. de; SANTOS, W. P. dos. Covid-19 rapid test by combining a random forest-based web system and blood tests. *Journal of Biomolecular Structure and Dynamics*, Taylor & Francis, v. 40, n. 22, p. 11948–11967, 2022.

BARBOSA, V. A. de F.; GOMES, J. C.; SANTANA, M. A. de; ALMEIDA, J. E. de; SOUZA, R. G. de; SOUZA, R. E. de; SANTOS, W. P. dos. Heg.IA: An intelligent system to support diagnosis of Covid-19 based on blood tests. *Research on Biomedical Engineering*, Springer, p. 1–18, 2021.

BARBOSA, V. A. de F.; SILVA, A. Félix da; SANTANA, M. A. de; AZEVEDO, R. Rabelo de; LIMA, R. d. C. Fernandes de; SANTOS, W. P. dos. Deep-wavelets and convolutional neural networks to support breast cancer diagnosis on thermography images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, Taylor & Francis, p. 1–19, 2022.

BELLO, M.; MARTÍNEZ-ARCHUNDIA, M.; CORREA-BASURTO, J. Automated docking for novel drug discovery. *Expert opinion on drug discovery*, Taylor & Francis, v. 8, n. 7, p. 821–834, 2013.

BIASI, L. D.; FINO, R.; PARISI, R.; SESSA, L.; CATTANEO, G.; SANTIS, A. D.; IANNELLI, P.; PIOTTO, S. Novel algorithm for efficient distribution of molecular docking calculations. In: SPRINGER. *Advances in Artificial Life, Evolutionary Computation and Systems Chemistry: 10th Italian Workshop, WIVACE 2015, Bari, Italy, September 22-25, 2015, Revised Selected Papers 10*. [S.l.], 2016. p. 65–74.

BITENCOURT-FERREIRA, G.; AZEVEDO, W. F. de. Machine learning to predict binding affinity. *Docking Screens for Drug Discovery*, Springer, p. 251–273, 2019.

BLOHM, P.; FRISHMAN, G.; SMIALOWSKI, P.; GOEBELS, F.; WACHINGER, B.; RUEPP, A.; FRISHMAN, D. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res*, England, v. 42, n. Database issue, p. D396–400, nov. 2013.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

CAMPOS, L. B.; FERREIRA, J. R. B. d. C.; SANTOS, W. P. d. Modelos de estimação de afinidade de proteínas para o design inteligente de fármacos baseados em pseudoconvoluções e regressores não lineares. *Research, Society and Development*, v. 11, n. 8, p. e40311831222, 2022.

CELESTI, A.; AMFT, O.; VILLARI, M. *Guest editorial special section on cloud computing, edge computing, internet of things, and big data analytics applications for healthcare industry 4.0*. [S.l.]: IEEE, 2019. 454–456 p.

CHANDAK, T.; MAYGINNES, J. P.; MAYES, H.; WONG, C. F. Using machine learning to improve ensemble docking for drug discovery. *Proteins: Structure, Function, and Bioinformatics*, Wiley, v. 88, n. 10, p. 1263–1270, maio 2020.

CHOI, J.; LEE, J. V-dock: Fast generation of novel drug-like molecules using machine-learning-based docking score and molecular optimization. *International Journal of Molecular Sciences*, MDPI AG, v. 22, n. 21, p. 11635, out. 2021.

CORDEIRO, F.; SANTOS, W.; SILVA-FILHOA, A. Segmentation of mammography by applying growcut for mass detection. *Studies in health technology and informatics*, v. 192, p. 87, 2013.

CORDEIRO, F. R.; BEZERRA, K. F.; SANTOS, W. P. dos. Random walker with fuzzy initialization applied to segment masses in mammography images. In: *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. Thessaloniki: [s.n.], 2017. p. 156–161.

CORDEIRO, F. R.; SANTOS, W. P.; SILVA-FILHO, A. G. An adaptive semi-supervised fuzzy growcut algorithm to segment masses of regions of interest of mammographic images. *Applied Soft Computing*, Elsevier, v. 46, p. 613–628, 2016.

CORDEIRO, F. R.; SANTOS, W. P.; SILVA-FILHO, A. G. A semi-supervised fuzzy growcut algorithm to segment and classify regions of interest of mammographic images. *Expert Systems with Applications*, Elsevier, v. 65, p. 116–126, 2016.

CORDEIRO, F. R.; SANTOS, W. P. d.; SILVA-FILHO, A. G. Analysis of supervised and semi-supervised growcut applied to segmentation of masses in mammography images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, Taylor & Francis, v. 5, n. 4, p. 297–315, 2017.

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

CUI, W.; AOUIDATE, A.; WANG, S.; YU, Q.; LI, Y.; YUAN, S. Discovering anti-cancer drugs via computational methods. *Frontiers in Pharmacology*, Frontiers Media SA, v. 11, p. 733, 2020.

DENG, J.; YANG, Z.; OJIMA, I.; SAMARAS, D.; WANG, F. Artificial intelligence in drug discovery: applications and techniques. *Briefings in Bioinformatics*, Oxford Academic, v. 23, n. 1, 2022.

DONG, D.; XU, Z.; ZHONG, W.; PENG, S. Parallelization of molecular docking: a review. *Current Topics in Medicinal Chemistry*, Bentham Science Publishers, v. 18, n. 12, p. 1015–1028, 2018.

DRUCHOK, M.; YARISH, D.; GARKOT, S.; NIKOLAIENKO, T.; GURBYCH, O. Ensembling machine learning models to boost molecular affinity prediction. *Computational Biology and Chemistry*, Elsevier, v. 93, p. 107529, 2021.

DUDA, R.; HART, P.; STORK, D. G. *Pattern Classification*. [S.l.]: John Wiley and Sons, 2001.

ESPINOLA, C. W.; GOMES, J. C.; PEREIRA, J. M. S.; SANTOS, W. P. dos. Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study. *Research on Biomedical Engineering*, Springer, v. 37, n. 1, p. 53–64, 2021.

ESPINOLA, C. W.; GOMES, J. C.; PEREIRA, J. M. S.; SANTOS, W. P. dos. Vocal acoustic analysis and machine learning for the identification of schizophrenia. *Research on Biomedical Engineering*, Springer, v. 37, n. 1, p. 33–46, 2021.

ESPINOLA, C. W.; GOMES, J. C.; PEREIRA, J. M. S.; SANTOS, W. P. dos. Detection of major depressive disorder, bipolar disorder, schizophrenia and generalized anxiety disorder using vocal acoustic analysis and machine learning: an exploratory study. *Research on Biomedical Engineering*, Springer, v. 38, n. 3, p. 813–829, 2022.

FAN, F. J.; SHI, Y. Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction. *Bioorganic & Medicinal Chemistry*, Elsevier, v. 72, p. 117003, 2022.

FAN, J.; FU, A.; ZHANG, L. Progress in molecular docking. In: *Quantitative Biology*. [S.l.: s.n.], 2019. v. 7.

FERREIRA, L. G.; SANTOS, R. N. dos; OLIVA, G.; ANDRICOPULO, A. D. Molecular docking and structure-based drug design strategies. *Molecules*, MDPI, v. 20, n. 7, p. 13384–13421, 2015.

FLEMING, N. How artificial intelligence is changing drug discovery. *Nature*, Nature Publishing Group, v. 557, n. 7706, p. S55–S55, 2018.

FONSECA, F. S.; TORCATE, A. S.; SILVA, A. C. G. D.; FREIRE, V. H. W.; FARIAS, G. P. D. M. D.; OLIVEIRA, J. F. L. D.; JÚNIOR, F. M. D. O.; JÚNIOR, J. C. D. S.; GOMES, D. A.; SANTOS, W. P. dos. Early prediction of generalized infection in intensive care units from clinical data: a committee-based machine learning approach. In: IEEE. *2022 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*. [S.l.], 2022. p. 1–6.

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*, ACS Publications, v. 6, n. 6, p. 939–949, 2020.

GHOSH, U.; NING, S.; WANG, Y.; KONG, Y. L. Addressing unmet clinical needs with 3d printing technologies. *Advanced healthcare materials*, Wiley Online Library, v. 7, n. 17, p. 1800417, 2018.

GOMES, J. C.; BARBOSA, V. A. de F.; SANTANA, M. A. de; BANDEIRA, J.; VALENçA, M. J. S.; SOUZA, R. E. de; ISMAEL, A. M.; SANTOS, W. P. dos. IKONOS: An intelligent tool to support diagnosis of Covid-19 by texture analysis of x-ray images. *Research on Biomedical Engineering*, Springer, p. 1–14, 2020.

GOMES, J. C.; MASOOD, A. I.; SILVA, L. H. d. S.; FERREIRA, J. R. B. da C.; JUNIOR, A. A. F.; ROCHA, A. L. d. S.; OLIVEIRA, L. C. P. de; SILVA, N. R. C. da; FERNANDES, B. J. T.; SANTOS, W. P. dos. Covid-19 diagnosis by combining rt-pcr and pseudo-convolutional machines to characterize virus sequences. *Scientific Reports*, Springer, v. 11, n. 1, p. 1–28, 2021.

GOMES, J. C.; MASOOD, A. I.; SILVA, L. H. d. S.; FERREIRA, J. R. B. d. C.; ALVAREZ, A. A. F. J.; ROCHA, A. L. d. S.; OLIVEIRA, L. C. P. d.; SILVA, N. R. C. d.; FERNANDES, B. J. T.; SANTOS, W. P. d. Covid-19 diagnosis by combining rt-pcr and pseudo-convolutional machines to characterize virus sequences. *Scientific Reports*, v. 11, n. 1, p. 1–28, 2021.

GOMES, J. C.; RODRIGUES, M. C. A.; SANTOS, W. P. dos. Asteri: Image-based representation of eeg signals for motor imagery classification. *Research on Biomedical Engineering*, Springer, v. 38, n. 2, p. 661–681, 2022.

GOMES, J. C.; SANTANA, M. A. de; MASOOD, A. I.; LIMA, C. L. de; SANTOS, W. P. dos. Covid-19's influence on cardiac function: a machine learning perspective on ecg analysis. *Medical & Biological Engineering & Computing*, Springer, p. 1–25, 2023.

GUEDES, I. A.; BARRETO, A. M. S.; MARINHO, D.; KREMMSKI, L.; KUENEMANN, M. A.; SPERANDIO, O.; DARDENNE, L. E.; MITEVA, M. A. New machine learning and physics-based scoring functions for drug discovery. *Scientific Reports*, v. 11, n. 1, p. 3198, 2021.

HECHT, D.; FOGEL, G. B. Computational intelligence methods for docking scores. *Current Computer-Aided Drug Design*, Bentham Science Publishers, v. 5, n. 1, p. 56–68, 2009.

HECK, G. S.; PINTRO, V. O.; PEREIRA, R. R.; LEVIN, N. M.; AZEVEDO, W. F de. Supervised machine learning methods applied to predict ligand-binding affinity. *Current medicinal chemistry*, Bentham Science Publishers, v. 24, n. 23, p. 2459–2470, 2017.

HSIN, K.-Y.; GHOSH, S.; KITANO, H. Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology. *PloS one*, Public Library of Science San Francisco, USA, v. 8, n. 12, p. e83922, 2013.

HWANG, H.; PIERCE, B.; MINTSERIS, J.; JANIN, J.; WENG, Z. Protein-protein docking benchmark version 3.0. *Proteins*, United States, v. 73, n. 3, p. 705–709, nov. 2008.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An Introduction to Statistical Learning*. New York: Springer, 2013. (Springer Texts in Statistics).

JAYANTHI, P.; IYYANKI, M.; MOTHKURI, A.; VADAKATTU, P. Fourth industrial revolution: an impact on health care industry. In: SPRINGER. *Advances in Artificial Intelligence, Software and Systems Engineering: Proceedings of the AHFE 2019 International Conference on Human Factors in Artificial Intelligence and Social Computing, the AHFE International Conference on Human Factors, Software, Service and Systems Engineering, and the AHFE International Conference of Human Factors in Energy, July 24-28, 2019, Washington DC, USA 10*. [S.l.], 2020. p. 58–69.

JIMÉNEZ-LUNA, J.; CUZZOLIN, A.; BOLCATO, G.; STURLESE, M.; MORO, S. A deep-learning approach toward rational molecular docking protocol selection. *Molecules*, MDPI, v. 25, n. 11, p. 2487, 2020.

JIMÉNEZ-LUNA, J.; GRISONI, F.; WESKAMP, N.; SCHNEIDER, G. Artificial intelligence in drug discovery: Recent advances and future perspectives. *Expert opinion on drug discovery*, Taylor & Francis, v. 16, n. 9, p. 949–959, 2021.

KANAKALA, G. C.; AGGARWAL, R.; NAYAR, D.; PRIYAKUMAR, U. D. Latent biases in machine learning models for predicting binding affinities using popular data sets. *ACS Omega*, ACS Publications, 2023.

KASHYAP, J.; DATTA, D. Drug repurposing for SARS-CoV-2: a high-throughput molecular docking, molecular dynamics, machine learning, and DFT study. *Journal of Materials Science*, Springer Science and Business Media LLC, v. 57, n. 23, p. 10780–10802, abr. 2022.

KAUR, A.; GARG, R.; GUPTA, P. Challenges facing ai and big data for resource-poor healthcare system. In: IEEE. *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*. [S.l.], 2021. p. 1426–1433.

KHAMIS, M. A.; GOMAA, W.; AHMED, W. F. Machine learning in computational docking. *Artificial Intelligence in Medicine*, v. 63, n. 3, p. 135–152, 2015.

KIM, K.-B.; HAN, K.-H. A study of the digital healthcare industry in the fourth industrial revolution. *Journal of Convergence for Information Technology*, Convergence Society for SMB, v. 10, n. 3, p. 7–15, 2020.

KITCHEN, D. B.; DECORNEZ, H.; FURR, J. R.; BAJORATH, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, Nature Publishing Group, v. 3, n. 11, p. 935–949, 2004.

LI, C.; HUA, Y.; PAN, D.; QI, L.; XIAO, C.; XIONG, Y.; LU, W.; DANG, Y.; GAO, X.; ZHAO, Y. A rapid selection strategy for umami peptide screening based on machine learning and molecular docking. *Food Chemistry*, Elsevier BV, v. 404, p. 134562, mar. 2023.

LI, X.-L.; ZHU, M.; LI, X.-L.; WANG, H.-Q.; WANG, S. Protein-protein interaction affinity prediction based on interface descriptors and machine learning. In: SPRINGER. *Intelligent Computing Theories and Applications: 8th International Conference, ICIC 2012, Huangshan, China, July 25-29, 2012. Proceedings 8*. [S.l.], 2012. p. 205–212.

LI, Z.; GU, J.; ZHUANG, H.; KANG, L.; ZHAO, X.; GUO, Q. Adaptive molecular docking method based on information entropy genetic algorithm. In: *Applied Soft Computing*. [S.l.: s.n.], 2015. v. 26.

LIMA, S.; AZEVEDO, W.; CORDEIRO, F.; SILVA-FILHO, A.; SANTOS, W. Feature extraction employing fuzzy-morphological decomposition for detection and classification of mass on mammograms. In: *Conference proceedings:... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*. [S.l.: s.n.], 2015. v. 2015, p. 801–804.

LIMA, S. M. de; SILVA-FILHO, A. G. da; SANTOS, W. P. dos. Detection and classification of masses in mammographic images in a multi-kernel approach. *Computer methods and programs in biomedicine*, Elsevier, v. 134, p. 11–29, 2016.

LIN, X.; LI, X.; LIN, X. A review on applications of computational methods in drug screening and design. *Molecules*, MDPI, v. 25, n. 6, p. 1375, 2020.

LIU, Y.; HU, B. Recent developments of sequence-based prediction of protein–protein interactions. *Current Opinion in Structural Biology*, Elsevier, v. 73, p. 102334, 2022.

LOWERY, C. What is digital health and what do i need to know about it? *Obstetrics and Gynecology Clinics*, Elsevier, v. 47, n. 2, p. 215–225, 2020.

MA, D.; GUO, Y.; LUO, J.; PU, X.; LI, M. Prediction of protein–protein binding affinity using diverse protein–protein interface features. *Chemometrics and Intelligent Laboratory Systems*, Elsevier, v. 138, p. 7–13, 2014.

MAGALHãES, C.; BARBOSA, H.; DARDENNE, L. A genetic algorithm for the ligand-protein docking problem. In: *Genetics and Molecular Biology - GENET MOL BIOL*. [S.l.: s.n.], 2004. v. 27.

MAHOMED, S. Healthcare, artificial intelligence and the fourth industrial revolution: Ethical, social and legal considerations. *South African Journal of Bioethics and Law*, Health and Medical Publishing Group (HMPG), v. 11, n. 2, p. 93–95, 2018.

MELO, J. A. G. de Melo e Castro e; ARAúJO, N. M. F. Impact of the fourth industrial revolution on the health sector: A qualitative study. *Healthcare informatics research*, Korean Society of Medical Informatics, v. 26, n. 4, p. 328–334, 2020.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997.

MORRONE, J. A.; WEBER, J. K.; HUYNH, T.; LUO, H.; CORNELL, W. D. Combining docking pose rank and structure with deep learning improves protein–ligand binding mode prediction over a baseline docking approach. *Journal of chemical information and modeling*, ACS Publications, v. 60, n. 9, p. 4170–4179, 2020.

NOGUEIRA, M. S.; KOCH, O. The development of target-specific machine learning models as scoring functions for docking-based target prediction. *Journal of Chemical Information and Modeling*, American Chemical Society (ACS), v. 59, n. 3, p. 1238–1252, fev. 2019.

NORMAN, R. A.; AMBROSETTI, F.; BONVIN, A. M.; COLWELL, L. J.; KELM, S.; KUMAR, S.; KRAWCZYK, K. Computational approaches to therapeutic antibody design: established methods and emerging trends. *Briefings in Bioinformatics*, Oxford University Press, v. 21, n. 5, p. 1549–1567, 2020.

OKIMOTO, N.; FUTATSUGI, N.; FUJI, H.; SUENAGA, A.; MORIMOTO, G.; YANAI, R.; OHNO, Y.; NARUMI, T.; TAIJI, M. High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations. *PLoS computational biology*, Public Library of Science San Francisco, USA, v. 5, n. 10, p. e1000528, 2009.

OLIVEIRA, A. P. S. D.; SANTANA, M. A. D.; ANDRADE, M. K. S.; GOMES, J. C.; RODRIGUES, M. C.; SANTOS, W. P. dos. Early diagnosis of parkinson's disease using eeg, machine learning and partial directed coherence. *Research on Biomedical Engineering*, Springer, v. 36, p. 311–331, 2020.

OLIVEIRA, T. A. de; MEDAGLIA, L. R.; MAIA, E. H. B.; ASSIS, L. C.; CARVALHO, P. B. de; SILVA, A. M. da; TARANTO, A. G. Evaluation of docking machine learning and molecular dynamics methodologies for DNA-ligand systems. *Pharmaceuticals*, MDPI AG, v. 15, n. 2, p. 132, jan. 2022.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, v. 22, n. 10, p. 1345–1359, 2010.

PANG, Z.; YUAN, H.; ZHANG, Y.-T.; PACKIRISAMY, M. Guest editorial health engineering driven by the industry 4.0 for aging society. *IEEE Journal of Biomedical and Health Informatics*, Institute of Electrical and Electronics Engineers, v. 22, n. 6, p. 1709–1710, 2018.

PARIS, R. D.; RUIZ, D. A.; SOUZA, O. N. D. A cloud-based workflow approach for optimizing molecular docking simulations of fully-flexible receptor models and multiple ligands. In: IEEE. *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*. [S.l.], 2015. p. 495–498.

PARVEZ, M. S. A.; KARIM, M. A.; HASAN, M.; JAMAN, J.; KARIM, Z.; TAHSIN, T.; HASAN, M. N.; HOSEN, M. J. Prediction of potential inhibitors for RNA-dependent RNA polymerase of SARS-CoV-2 using comprehensive drug repurposing and molecular docking approach. *International Journal of Biological Macromolecules*, Elsevier, v. 163, p. 1787–1797, 2020.

PATEL, V.; SHAH, M. Artificial intelligence and machine learning in drug discovery and development. *Intelligent Medicine*, Elsevier, v. 2, n. 3, p. 134–140, 2022.

PAUL, D.; SANAP, G.; SHENOY, S.; KALYANE, D.; KALIA, K.; TEKADE, R. K. Artificial intelligence in drug discovery and development. *Drug discovery today*, Elsevier, v. 26, n. 1, p. 80, 2021.

PINZI, L.; RASTELLI, G. Molecular docking: shifting paradigms in drug discovery. *International journal of molecular sciences*, MDPI, v. 20, n. 18, p. 4331, 2019.

POYNER, I. K.; SHERRATT, R. S. Improving access to healthcare in rural communities—iot as part of the solution. In: IET. *3rd IET International Conference on Technologies for Active and Assisted Living (TechAAL 2019)*. [S.l.], 2019. p. 1–6.

PRIMICERI, E.; CHIRIACÒ, M. S.; NOTARANGELO, F. M.; CROCAMO, A.; ARDISSINO, D.; CEREDA, M.; BRAMANTI, A. P.; BIANCHESSI, M. A.; GIANNELLI, G.; MARUCCIO, G. Key enabling technologies for point-of-care diagnostics. *Sensors*, MDPI, v. 18, n. 11, p. 3607, 2018.

QUEVEDO, C. V.; PARIS, R. D.; RUIZ, D. D.; SOUZA, O. N. de. A strategic solution to optimize molecular docking simulations using fully-flexible receptor models. *Expert Systems with Applications*, Elsevier, v. 41, n. 16, p. 7608–7620, 2014.

RAGOZA, M.; HOCHULI, J.; IDROBO, E.; SUNSERI, J.; KOES, D. R. Protein–ligand scoring with convolutional neural networks. *Journal of Chemical Information and Modeling*, v. 57, n. 4, p. 942–957, 2017.

RAVAL, K.; GANATRA, T. Basics, types and applications of molecular docking: a review. *IP International Journal of Comprehensive and Advanced Pharmacology*, v. 7, n. 1, p. 12–16, 2022.

RICCI-LOPEZ, J.; AGUILA, S. A.; GILSON, M. K.; BRIZUELA, C. A. Improving structure-based virtual screening with ensemble docking and machine learning. *Journal of Chemical Information and Modeling*, American Chemical Society (ACS), v. 61, n. 11, p. 5362–5376, out. 2021.

SAIKIA, S.; BORDOLOI, M. Molecular docking: challenges, advances and its use in drug discovery perspective. *Current drug targets*, Bentham Science Publishers, v. 20, n. 5, p. 501–521, 2019.

SALMASO, V.; MORO, S. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Frontiers in pharmacology*, Frontiers Media SA, v. 9, p. 923, 2018.

SANTANA, M. A. d.; PEREIRA, J. M. S.; SILVA, F. L. d.; LIMA, N. M. d.; SOUSA, F. N. d.; ARRUDA, G. M. S. d.; LIMA, R. d. C. F. d.; SILVA, W. W. A. d.; SANTOS, W. P. d. Breast cancer diagnosis based on mammary thermography and extreme learning machines. *Research on Biomedical Engineering*, SciELO Brasil, v. 34, p. 45–53, 2018.

SANTANA, M. A. de; BARBOSA, V. A. de F.; LIMA, R. de Cássia Fernandes de; SANTOS, W. P. dos. Combining deep-wavelet neural networks and support-vector machines to classify breast lesions in thermography images. *Health and Technology*, Springer, p. 1–13, 2022.

SANTANA, M. A. de; PEREIRA, J. M. S.; SILVA, F. L. da; LIMA, N. M. de; SOUSA, F. N. de; ARRUDA, G. M. S. de; LIMA, R. d. C. F. de; SILVA, W. W. A. da; SANTOS, W. P. dos. Breast cancer diagnosis based on mammary thermography and extreme learning machines. *Research on Biomedical Engineering*, Brazilian Society of Biomedical Engineering, v. 34, n. 1, p. 45–53, 2018.

SANTANA, M. A. de; SANTOS, W. P. dos. A deep-wavelet neural network to detect and classify lesions in mammographic images. *Research on Biomedical Engineering*, Springer, v. 38, n. 4, p. 1051–1066, 2022.

SENBEKOV, M.; SALIEV, T.; BUKEYEVA, Z.; ALMABAYEVA, A.; ZHANALIYEVA, M.; AITENOVA, N.; TOISHIBEKOV, Y.; FAKHRADIYEV, I. et al. The recent progress and applications of digital technologies in healthcare: a review. *International journal of telemedicine and applications*, Hindawi, v. 2020, 2020.

SHAFQAT, S.; KISHWER, S.; RASOOL, R. U.; QADIR, J.; AMJAD, T.; AHMAD, H. F. Big data analytics enhanced healthcare systems: a review. *The Journal of Supercomputing*, Springer, v. 76, p. 1754–1799, 2020.

SHIRAHIGE, L.; LEIMIG, B.; BALTAR, A.; BEZERRA, A.; BRITO, C. V. F. de; NASCIMENTO, Y. S. O. do; GOMES, J. C.; TEO, W.-P.; SANTOS, W. P. dos; CAIRRÃO, M.; FONSECA, A.; MONTE-SILVA, K. Classification of parkinson's disease motor phenotype: a machine learning approach. *Journal of Neural Transmission*, Springer, p. 1–15, 2022.

SOUZA, R. G. de; SILVA, G. dos Santos Lucas e; SANTOS, W. P. dos; LIMA, M. E. de; INITIATIVE, A. D. N. Computer-aided diagnosis of alzheimer's disease by mri analysis and evolutionary computing. *Research on Biomedical Engineering*, Springer, v. 37, p. 455–483, 2021.

STEPNIEWSKA-DZIUBINSKA, M. M.; ZIELENKIEWICZ, P.; SIEDLECKI, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, Oxford University Press, v. 34, n. 21, p. 3666–3674, 2018.

TAYLOR, R. D.; JEWSBURY, P. J.; ESSEX, J. W. A review of protein-small molecule docking methods. *Journal of computer-aided molecular design*, Springer, v. 16, p. 151–166, 2002.

TERAYAMA, K.; IWATA, H.; ARAKI, M.; OKUNO, Y.; TSUDA, K. Machine learning accelerates md-based binding pose prediction between ligands and proteins. *Bioinformatics*, Oxford University Press, v. 34, n. 5, p. 770–778, 2018.

THAKARE, V.; KHIRE, G.; KUMBHAR, M. Artificial intelligence (ai) and internet of things (iot) in healthcare: Opportunities and challenges. *ECS Transactions*, IOP Publishing, v. 107, n. 1, p. 7941, 2022.

TOPOL, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, v. 25, n. 1, p. 44–56, 2019.

TROTT, O.; OLSON, A. J. Autodock vina: improving the speed and accuracy of docking with a new scoring function and efficient optimization. *Journal of Computational Chemistry*, v. 31, n. 2, p. 455–461, 2010.

TURCU, C. E.; TURCU, C. O. Internet of things as key enabler for sustainable healthcare delivery. *Procedia-Social and Behavioral Sciences*, Elsevier, v. 73, p. 251–256, 2013.

VASANT, O. K.; CHANDRAKANT, M. A.; CHANDRASHEKHAR, K. V.; BABASAHEB, G. V.; DNYANDEV, K. M. A review on molecular docking. *International Research Journal of Pure and Applied Chemistry*, v. 22, n. 3, p. 60–68, 2021.

VEIT-ACOSTA, M.; JUNIOR, W. F. de A. The impact of crystallographic data for the development of machine learning models to predict protein-ligand binding affinity. *Current Medicinal Chemistry*, Bentham Science Publishers, v. 28, n. 34, p. 7006–7022, 2021.

WANG, X.; TERASHI, G.; CHRISTOFFER, C. W.; ZHU, M.; KIHARA, D. Protein docking model evaluation by 3d deep convolutional neural networks. *Bioinformatics*, Oxford Academic, v. 36, n. 7, p. 2113–2118, 2020.

WU, C.; LIU, Y.; YANG, Y.; ZHANG, P.; ZHONG, W.; WANG, Y.; WANG, Q.; XU, Y.; LI, M.; LI, X. et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharmaceutica Sinica B*, Elsevier, v. 10, n. 5, p. 766–788, 2020.

WU, F.; ZHOU, Y.; LI, L.; SHEN, X.; CHEN, G.; WANG, X.; LIANG, X.; TAN, M.; HUANG, Z. Computational approaches in preclinical studies on drug discovery and development. *Frontiers in Chemistry*, Frontiers Media SA, v. 8, p. 726, 2020.

WóJCIAKOWSKI, M.; BALLESTER, P. J.; SIEDLECKI, P. Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific Reports*, v. 7, 2017.

XIONG, X. Bring technology home and stay healthy: The role of fourth industrial revolution and technology in improving the efficacy of health care spending. *Technological Forecasting and Social Change*, Elsevier, v. 165, p. 120556, 2021.

XU, J.; LI, J.; CAI, Y. Molecular docking simulation based on cpu-gpu heterogeneous computing. In: SPRINGER. *Advanced Parallel Processing Technologies: 12th International Symposium, APPT 2017, Santiago de Compostela, Spain, August 29, 2017, Proceedings 12*. [S.l.], 2017. p. 27–37.

YAMASHITA, R.; NISHIO, M.; DO, R. K. G.; TOGASHI, K. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, v. 9, n. 4, p. 611–629, 2018.

YANG, C.; CHEN, E. A.; ZHANG, Y. Protein–ligand docking in the machine-learning era. *Molecules*, MDPI, v. 27, n. 14, p. 4568, 2022.

YANG, L.; YANG, G.; CHEN, X.; YANG, Q.; YAO, X.; BING, Z.; NIU, Y.; HUANG, L.; YANG, L. Deep scoring neural network replacing the scoring function components to improve the performance of structure-based molecular docking. *ACS Chemical Neuroscience*, ACS Publications, v. 12, n. 12, p. 2133–2142, 2021.

YANG, S.; LI, S.; CHANG, J. Discovery of cobimetinib as a novel a-FABP inhibitor using machine learning and molecular docking-based virtual screening. *RSC Advances*, Royal Society of Chemistry (RSC), v. 12, n. 21, p. 13500–13510, 2022.

YURIEV, E.; AGOSTINO, M.; RAMSLAND, P. A. Challenges and advances in computational docking: 2009 in review. *Journal of Molecular Recognition*, Wiley Online Library, v. 24, n. 2, p. 149–164, 2011.

ZHANG, Y.; WANG, Y.; ZHOU, W.; FAN, Y.; ZHAO, J.; ZHU, L.; LU, S.; LU, T.; CHEN, Y.; LIU, H. A combined drug discovery strategy based on machine learning and molecular docking. *Chemical Biology & Drug Design*, v. 93, n. 5, p. 685–699, 2019.

ZHAO, J.; CAO, Y.; ZHANG, L. Exploring the computational methods for protein-ligand binding site prediction. *Computational and Structural Biotechnology Journal*, Elsevier, v. 18, p. 417–426, 2020.

ZHAVORONKOV, A. Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry. *Molecular Pharmaceutics*, ACS Publications, v. 15, n. 10, p. 4311–4313, 2018.

ZHAVORONKOV, A.; VANHAELEN, Q.; OPREA, T. I. Will artificial intelligence for drug discovery impact clinical pharmacology? *Clinical Pharmacology & Therapeutics*, Wiley Online Library, v. 107, n. 4, p. 780–785, 2020.

ZHUANG, F.; QI, Z.; DU, K.; XI, D.; ZHU, Y.; ZHANG, H.; XIONG, H.; HE, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, v. 109, n. 1, p. 43–76, 2021.