UNIVERSIDADE FEDERAL DE MINAS GERAIS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

GIOVANA DE CASTRO FIORINI MAIA

AVALIAÇÃO DO POTENCIAL PREDITIVO DE ASSINATURAS ESTRUTURAIS NA IDENTIFICAÇÃO DE INTERAÇÕES PROTEÍNA-PEPTÍDEO

Belo Horizonte
2024

GIOVANA DE CASTRO FIORINI MAIA

AVALIAÇÃO DO POTENCIAL PREDITIVO DE ASSINATURAS ESTRUTURAIS NA IDENTIFICAÇÃO DE INTERAÇÕES PROTEÍNA-PEPTÍDEO

Dissertação apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito à obtenção do título de Mestre em Bioinformática.

Orientador: Raquel Cardoso de Melo Minardi

Belo Horizonte

2024

043 Maia, Giovana de Castro Fiorini.
Avaliação do potencial preditivo de assinaturas estruturais na identificação de interações proteína-peptídeo [manuscrito] / Giovana de Castro Fiorini Maia. – 2024.
88 f. : il. ; 29,5 cm.

Orientador: Raquel Cardoso de Melo Minardi.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Aprendizado de Máquina. 3. Peptídeos. 4. Proteínas. I. Minardi, Raquel Cardoso de Melo. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Giovana de Castro Fiorini Maia

"AVALIAÇÃO DO POTENCIAL PREDITIVO DE ASSINATURAS ESTRUTURAIS NA IDENTIFICAÇÃO DE INTERAÇÕES PROTEÍNA-PEPTÍDEO"

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Profª Raquel Cardoso de Melo Minardi - Orientadora

Universidade Federal de Minas Gerais

Profª Maria Goreti de Almeida Oliveira

Universidade Federal de Viçosa

Profª Sabrina de Azevedo Silveira

Universidade Federal de Viçosa

Belo Horizonte, 10 de setembro de 2024.

Documento assinado eletronicamente por **Raquel Cardoso de Melo Minardi**, **Coordenador(a) de curso de pós-graduação**, em 11/09/2024, às 10:03, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Maria Goreti de Almeida Oliveira**, **Usuário Externo**, em 12/09/2024, às 07:15, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Sabrina de Azevedo Silveira**, **Usuário Externo**, em 21/09/2024, às 23:19, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3542514** e o código CRC **15AAA593**.

Referência: Processo nº 23072.251665/2024-17 SEI nº 3542514

Agradecimentos

É impossível atravessar a jornada da pós-graduação sozinho. Agradeço aos que estiveram ao meu lado durante esses dois anos. Inicialmente a Deus, que me abençoa todos os dias e me coloca exatamente onde eu deveria estar.

Agradeço aos meus pais, Cynthia e André, pelo amor, apoio e pelos genes que me foram passados. Às minhas irmãs, Letícia e Victória, por me fazerem querer deixar um mundo melhor para vocês. Aos meus avós, e em especial ao meu avô Mauro, que sempre me mostrou a importância do estudo e lutou para que eu chegasse onde estou. Agradeço também ao meu marido, Frederico, que convive comigo todos os dias, que me ouve, me apoia na carreira acadêmica, e muitas vezes, acreditou mais em mim do que eu mesma.

Agradeço também aos meus amigos, que fizeram desta jornada muito mais especial. Ao Guilherme, meu amigo e colega que me acompanha desde o primeiro dia da graduação em Ciências Biológicas. Aos *Rangers*, que conviveram comigo no mestrado e me proporcionaram muitos momentos de alegria e aprendizado. À todos os colegas do Laboratório de Bioinformática e Sistemas, que me ouviram quando tudo parecia dar errado, me ensinaram sobre bioinformática de proteínas, sobre a pós-graduação e sobre a vida. Também agradeço a todos os professores que me deram aula durante o mestrado e, em especial à Profa. Raquel Minardi, minha orientadora. Agradeço pela oportunidade de desenvolver essa pesquisa, pela paciência, pelo respeito de sempre e pelos ensinamentos.

Por fim, agradeço às agências de fomento à pesquisa: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Além disso, agradeço ao Programa de Pós-graduação em Bioinformática da UFMG e à Sheila e ao Tiago.

O presente trabalho contou com o apoio da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Resumo

As interações entre proteínas e peptídeos desempenham papeis vitais no funcionamento do organismo humano. Estudos *in silico* baseados em inteligência artificial que buscam compreender melhor essas interações são fundamentais para desenvolvimento de terapias baseadas em peptídeos. Entretanto, predizer interações proteína-peptídeo ainda é um desafio para a bioinformática estrutural devido principalmente à flexibilidade do peptídeo e a dificuldade em desenvolver métricas mais precisas para ranquear as posições esperadas dos ligantes no *docking* molecular. Neste trabalho buscamos avaliar o potencial preditivo de assinaturas estruturais na identificação de interações proteína-peptídeo, utilizando modelos de aprendizado de máquina. Para isso, o trabalho foi dividido em quatro partes: I. Avaliação de três ferramentas de *docking* e modelagem molecular para complexos proteína-peptídeo: HPEPDOCK, HDOCK e *AlphaFold Multimer*, II.Comparação da modelagem de peptídeos nas formas *holo* e *apo* utilizando o *AlphaFold Multimer*, III. Desenvolvimento de modelos de aprendizado de máquina para predição de complexos proteína-peptídeo reais e falsos, utilizando assinaturas estruturais da interface proteína-peptídeo, e IV. Desenvolvimento de modelos de aprendizado de máquina para predição de interação proteína-peptídeo, usando assinaturas estruturais da interface da proteína, e do peptídeo concatenadas. Os principais resultados revelam que as três ferramentas comparadas (HPEPDOCK, HDOCK e *AlphaFold Multimer*) obtiveram resultados próximos para a modelagem de complexos, conseguindo distinguir bem complexos reais de complexos falsos, sendo o HPEPDOCK a ferramenta que obteve maior mediana de valor de DockQ quando comparados os complexos preditos com os complexos experimentais (mediana de DockQ = 0.935). Em relação à comparação dos peptídeos na forma *holo* e *apo*, as modelagens ficaram semelhantes, com média de RMSD igual a 0.214. Desenvolvemos também modelos de classificação de complexos reais e falsos e obtivemos, para o melhor modelo F1-score de 0.741, usando Redes Neurais. Por fim, para os modelos de predição de interação proteína-peptídeo, obtivemos F1-score de 0.930 para o melhor modelo gerado com o algoritmo *Gradient Boosting*. Em resumo, as assinaturas estruturais se mostraram ferramentas úteis para treinar modelos de predição de interação proteína-peptídeo, e os modelos criados podem ser utilizados para facilitar a triagem virtual, auxiliando no desenvolvimento de terapias e fármacos baseados em peptídeos.

Palavras-chave: Peptídeo; Assinatura Estrutural; Aprendizado de Máquina; Interação Proteína-peptídeo; *Docking*; *Docking* proteína-peptídeo; *AlphaFold*

Abstract

Protein-peptide interactions are essential for the proper functioning of the human body, playing crucial roles in various biological processes. In silico studies based on artificial intelligence have become fundamental for the development of peptide-based therapies. However, predicting these interactions remains a challenge in structural bioinformatics, primarily due to the intrinsic flexibility of peptides and the difficulty in developing more accurate metrics for ranking the expected positions of ligands in molecular docking. In this study, we evaluated the predictive potential of structural signatures in identifying protein-peptide interactions using machine learning models. The dissertation is structured into four main parts: I. Evaluation of three docking and molecular modeling tools for protein-peptide complexes: HPEPDOCK, HDOCK, and AlphaFold Multimer; II. Comparison of peptide modeling in holo and apo forms using AlphaFold Multimer; III. Development of machine learning models for predicting real and fake protein-peptide complexes using structural signatures of the protein-peptide interface; and IV. Development of machine learning models for predicting protein-peptide interactions using concatenated structural signatures from protein interface and peptide. The main results indicate that the three docking tools evaluated (HPEPDOCK, HDOCK, and AlphaFold Multimer) showed similar performances in complex modeling, effectively distinguishing real complexes from false ones. HPEPDOCK stood out with the highest median DockQ score (0.935) when comparing predicted complexes with experimental ones. In the comparison between peptides modeled in holo and apo forms, the results were similar, with an average RMSD of 0.214. Machine learning models for classifying real and fake complexes were developed, achieving an F1-score of 0.741 for the best model, using Neural Networks. Additionally, for protein-peptide interaction prediction, the best model developed with the Gradient Boosting algorithm achieved an F1-score of 0.930. In summary, structural signatures proved to be valuable tools for training protein-peptide interaction prediction models. The developed models can be used to facilitate virtual peptide screening, contributing to the development of new peptide-based therapies and drugs.

Keywords: *Peptide; Structural signature; Machine learning; Docking; Protein-peptide interaction; Protein-peptide docking; AlphaFold.*

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

aCSM: *atomic Cutoff Scanning Matrix*

AUC: *Área sob a curva*

CAPRI: *Critical Assessment of PRediction of Interactions*

CASP: *Critical Assessment of protein Structure Prediction*

CSM: *Cutoff Scanning Matrix*

FFT: *Fast Fourier Transform*

FNAT: *Fração de contatos nativos*

GNC: *Graph Convolutional Networks*

iRMS: *Interface Root Mean Square Deviation*

KNN: *K-Nearest Neighbors*

L-RMS: *Ligand Root Mean Square Deviation*

MCC: *Matthews Correlation Coefficient*

PAE: *Predicted Align Error*

pLDDT: *Predicted Local Distance Difference Test*

PDB: *Protein Data Bank*

ROC: *Receiver Operating Characteristic*

RMSD: *Root Mean Square Deviation*

SVM: *Support Vector Machine*

Sumário

1. Introdução

1.1 Proteínas e peptídeos

As proteínas realizam inúmeras funções fisiológicas para os seres vivos, como a de catalisar reações bioquímicas (enzimas), funcionar como neutralizadores de antígenos (anticorpos), receptores celulares, transportadores, dentre outros. Essa versatilidade é atribuída à capacidade das proteínas de interagir com uma variedade de ligantes, incluindo outras proteínas, peptídeos, carboidratos, ácidos nucleicos e pequenas moléculas. Essas interações são possíveis devido à diversidade de tipos de aminoácidos que compõem as proteínas, podendo ser aromáticos ou alifáticos, hidrofílicos ou hidrofóbicos, carregados positiva, negativamente ou não carregados. (BRÄNDÉN; TOOZE, 1999)

Para a formação das proteínas, ocorrem ligações covalentes entre os aminoácidos conhecidas como ligações peptídicas. Durante esse processo, o carbono da carboxila de um aminoácido se une ao nitrogênio do grupo amina de outro aminoácido, ocasionando a liberação de uma molécula de água. Quando cadeias polipeptídicas curtas de aminoácidos (2 a 50 resíduos de aminoácido) são ligadas dessa forma, originam-se os peptídeos. Cadeias maiores são denominadas proteínas. Uma característica importante dos peptídeos é a maior flexibilidade de sua estrutura em comparação com as proteínas. Devido ao seu tamanho reduzido, as cadeias peptídicas estabelecem menos interações proteína-peptídeo, não se enovelando de forma estável e permitindo uma ampla gama de conformações. Além disso, a falta de domínios funcionais definidos resulta em uma menor estabilidade estrutural, facilitando interações com uma variedade de ligantes e uma capacidade de assumir diversas conformações. (LEHNINGER; NELSON; COX, 2017)

1.1.1 Interações proteína-peptídeo

As interações fracas entre resíduos de aminoácidos são responsáveis por estabilizar a estrutura da proteína, e consequentemente, responsáveis pela sua função. Essas interações podem ser interações de van der Waals, interações hidrofóbicas, ligações de hidrogênio, interações eletrostáticas e pontes dissulfeto, uma ligação covalente e forte, diferente das demais (MANCINI et al., 2004; MELO et al., 2007; NESHICH et al., 2003; SOBOLEV et al., 1999)

As ligações de hidrogênio ocorrem entre um hidrogênio ligado a um átomo doador (oxigênio ou nitrogênio, no caso de aminoácidos) que é atraído por um átomo aceptor de ligação de hidrogênio (outro átomo mais eletronegativo). Elas são fundamentais para a formação e estabilidade das estruturas secundárias das proteínas. Resíduos de aminoácidos com cadeia

lateral hidrofóbica, quando próximos, interagem estabelecendo interações de Van der Waals. Essas interações são importantes para estabilizar estruturas terciárias de cadeias polipeptídicas, uma vez que esses resíduos tendem a se agrupar no interior das proteínas. As interações iônicas atrativas e repulsivas são entre resíduos de aminoácidos com cargas formais. Quando os átomos possuem cargas diferentes (positiva-negativa), e estão em distâncias adequadas, são formadas interações iônicas atrativas. Quando possuem ambos cargas iguais (positiva-positiva ou negativa-negativa), interações iônicas repulsivas. Os empilhamentos aromáticos são interações entre dois resíduos que possuem anel aromático na cadeia lateral, como o triptofano (Trp) e fenilalanina (Phe). Esse tipo de interação é favorecido pelo efeito hidrofóbico gerado pelas porções apolares das moléculas. Já as pontes dissulfeto são ligações covalentes entre grupos tiol (-SH) de dois resíduos de cisteína em uma proteína. Como são ligações mais fortes, ou seja, que demandam mais energia para sua quebra, são importantes para a estabilização da estrutura da proteína principalmente em condições de pH e temperatura atípicos.

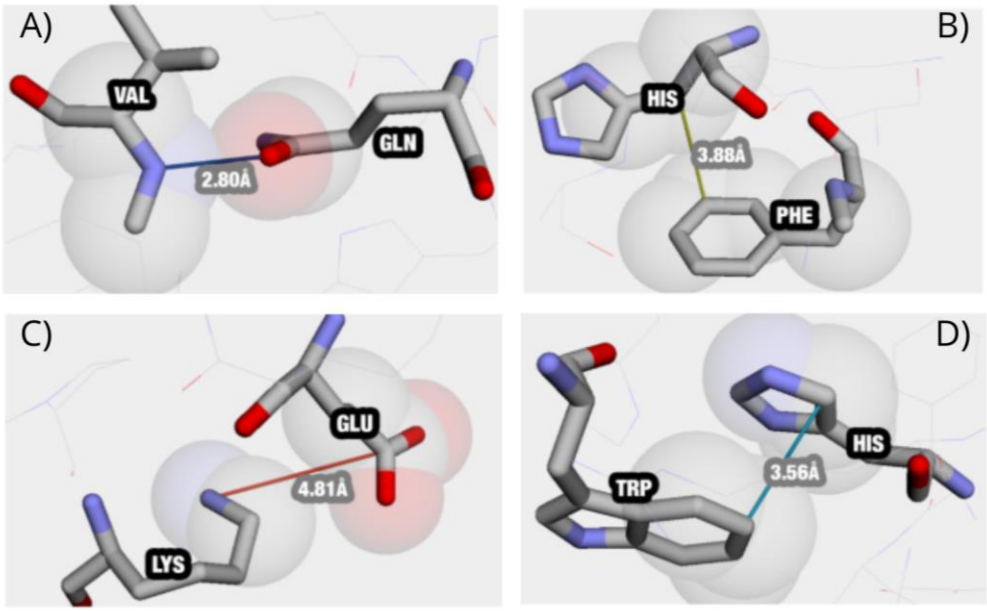


Figura 1. Exemplos de interações não covalentes entre resíduos de aminoácidos. A) Ligação de hidrogênio entre o nitrogênio da valina (roxo) e o oxigênio da glutamina (vermelho). **B)** Interação hidrofóbica entre carbonos da histidina e fenilalanina. **C)** Interação iônica entre o nitrogênio da lisina (roxo) e o ácido glutâmico. **D)** Empilhamento aromático entre os aneis aromáticos do triptofano e da histidina.

Fonte: Adaptado de SILVA et al., 2019

As distâncias em proteínas são medidas na escala de Angstroms (Å) que equivale a 10^{-10}m. Para definir cada tipo de interação entre resíduos de aminoácidos, é possível usar critérios de distâncias de corte (*cutoff*) juntamente com as propriedades atômicas. A Tabela 1 mostra o

cutoff de distância mínima e máxima que consideramos para definir cada tipo de contato, e os tipos de átomos que realizam esse contato.

Tabela 1. Tipos de contatos entre os átomos. (Traduzido de PIMENTEL et al., 2021)

Tipo de contato	Classes (resíduo 1-2)	Cutoff (min-máx)
Hidrofóbico	Hidrofóbico - Hidrofóbico	2 - 4.5 Å
Atrativo/Repulsivo	Carregado positivamente - Carregado negativamente (atrativo) Carregado negativamente - Carregado negativamente (repulsivo) Carregado positivamente - Carregado positivamente (repulsivo)	2 - 6 Å
Ligação de hidrogênio	Aceptor - Doador	≤ 3.9 Å
Empilhamento aromático	Centróide do anel - Centróide do anel	2 - 6 Å
Pontes dissulfeto	Cisteína - Cisteína	1.5 - 2.8 Å

As interações entre proteínas e peptídeos desempenham papéis vitais no funcionamento do organismo humano. Um exemplo é encontrado nos hormônios e neurotransmissores, cuja natureza peptídica lhes confere um papel fundamental na transmissão de sinais. Ao se ligarem aos seus receptores específicos nas células, essas moléculas desencadeiam uma variedade de respostas biológicas essenciais. Por exemplo, a insulina e o glucagon regulam os níveis de açúcar no sangue, e a ocitocina é crucial para a liberação do leite e o estímulo do parto. A vasopressina, por sua vez, exerce uma função antidiurética, enquanto a angiotensina está relacionada ao controle da pressão arterial.

Pelo fato de desempenharem funções moduladoras no organismo, o estudo dos peptídeos e de suas interações com ligantes ganhou visibilidade em diferentes áreas. Na medicina, por exemplo, utilizam-se terapias com peptídeos naturais ou sintéticos (produzidos por meio de desenho racional de peptídeos) para inibição de interação entre ligante-receptor, visando controle de doenças como diabetes mellitus (GEORGE et al., 2018), doenças cardiovasculares (GAO et al., 2019) e gastrointestinais (KUNKEL et al., 2011), podendo ser utilizados até mesmo como antivirais (SU et al., 2020). Já na agricultura, os peptídeos podem ser utilizados, por exemplo, como inibidores de proteases na defesa bioquímica de plantas a pragas agrícolas. (BARASHKOVA et al., 2022) (HELLINGER; GRUBER, 2019)

A bioinformática estrutural é o campo dedicado ao estudo e análise de dados relacionados às estruturas de proteínas, peptídeos e outras biomoléculas como DNA, RNA e pequenas moléculas e possui papel importante no estudo de interações proteína-peptídeo. Devido a desafios experimentais na produção de proteínas recombinantes e síntese de

peptídeos, torna-se necessário o estudo *in silico* dessas interações para economizar tempo e recursos, além de maior sucesso nos processos biológicos. (VERLI et al, 2014).

As bases de dados de complexos proteína-peptídeo reúnem e organizam informações necessárias às pesquisas de bioinformática estrutural. O *Protein Data Bank* (PDB) é a base de dados mais abrangente de estruturas e complexos de proteínas, incluindo os complexos proteína-peptídeo. Atualmente, contempla mais de 222 mil estruturas resolvidas experimentalmente, além de complexos preditos compreendidos no *AlphaFold DataBase*.

Dentre as bases de dados específicas para dados de estruturas de complexos proteína-peptídeos, a PepBDB (WEN et al., 2019), atualizada pela última vez em 2020, possui informações de estrutura e sequências de cerca de 13 mil complexos de proteína-peptídeo. Ela permite buscar por complexos de acordo com a resolução em angstrons da estrutura, método pelo qual o complexo foi resolvido experimentalmente e sequência. Em 2020, Xu e colaboradores desenvolveram um *benchmark* não redundante de complexos proteína-peptídeo que conta com estrutura e sequência de 89 complexos, a base de dados PepPro (XU et al., 2020). A Propedia (MARTINS et al., 2021) atualmente é a maior base de dados atualizada de complexos proteína-peptídeo, contando com mais de 20 mil complexos. Além de armazenar as estruturas, a ferramenta separa os complexos em *clusters* de acordo com a sua interface, sítio de ligação ou sequência, facilitando a busca do usuário por complexos e peptídeos de seu interesse. Alguns anos depois, foi lançada ainda a Propedia 2 (MARTINS et al., 2023), atualizada atualmente com mais de 49 mil complexos e com novo conjunto de dados desenvolvido utilizando assinaturas estruturais, para a classificação de peptídeos.

1.2 *Docking* molecular

Para prever, de modo teórico, o modo e a afinidade de ligação de uma macromolécula receptora com outra molécula ligante (seja ela também outra macromolécula, ou um ligante pequeno), usamos técnicas de *docking* molecular. Elas utilizam algoritmos de busca para encontrar os modos de ligação do ligante no sítio do receptor e funções de pontuação, que buscam quantificar a qualidade dos modelos obtidos. A técnica de *docking* pode ser empregada para a triagem virtual de moléculas no desenvolvimento de fármacos, por exemplo.

Existem várias ferramentas de *docking* disponíveis que usam diferentes algoritmos de busca e funções de pontuação. Em relação à amostragem, alguns algoritmos fazem uma busca sistemática por toda a proteína e exploram de forma combinatória inúmeros graus de liberdade da molécula. Outros, utilizam métodos de busca estocásticos como o algoritmo genético e

método Monte Carlo para o atracamento de ligantes. Já as funções de pontuação podem ser empíricas, baseadas em conhecimento ou em campo de força, dentre outras. (VERLI et al, 2014)

Apesar de existirem várias ferramentas de *docking* molecular, ainda existem desafios no estudo de interações proteína-ligante. Devido a flexibilidade da proteína receptora e do ligante, existe um grande número de conformações possíveis em que o ligante pode adquirir ao interagir com a proteína, o que torna o *docking* caro computacionalmente. Além disso, as métricas para ordenar os compostos, ainda não são capazes de se correlacionar significativamente com métricas experimentais de afinidade (LONDON et al., 2010). Dessa forma, avaliar o desempenho de diferentes métodos de *docking* é essencial para compreender seus prós e contras, além de auxiliar no desenvolvimento de métodos mais precisos que possam superar as limitações dos métodos existentes.

1.2.1 Redocking

Para avaliar ferramentas de *docking* molecular, são empregadas técnicas como o *redocking*. Essa técnica envolve submeter uma proteína e um ligante, cuja estrutura foi resolvida experimentalmente em complexo, a um programa de *docking*. O objetivo é verificar a capacidade do programa em posicionar o ligante no sítio correto da proteína, e adotando a pose correta, em relação àquela resolvida experimentalmente.

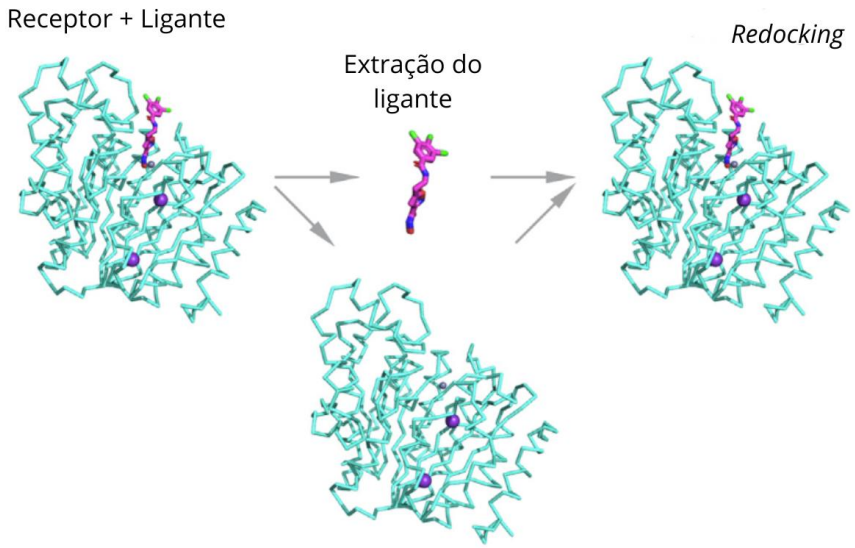


Figura 2. Esquema ilustrativo da técnica de *redocking.*

Fonte: Adaptado de TEMML et al., 2021.

Para avaliar as poses dos ligantes, é comum utilizar uma métrica chamada RMSD (*Root Mean Square Deviation*). O RMSD é a raiz quadrada do desvio quadrático médio entre átomos

considerados equivalentes em estruturas sobrepostas, ou seja, alinhadas estruturalmente. Ele consiste na raíz quadrada da soma do quadrado das diferenças entre as coordenadas x, y e z dos átomos equivalentes, dividida pelo número de átomos total (Figura 3). Quanto menor o valor de RMSD, mais similares e bem sobrepostas estão as duas estruturas comparadas. Considera-se um *redocking* bem sucedido para pequenos ligantes quando a melhor pose ranqueada pela ferramenta possui RMSD ≤ 2 Å, quando comparada com o ligante do complexo experimental (ALLEN et al., 2015).

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{i1} - x_{i2})^2 + (y_{i1} - y_{i2})^2 + (z_{i1} - z_{i2})^2}$$

Figura 3. Equação do RMSD. X, y e z correspondem a coordenadas dos átomos, e n corresponde ao número de átomos.

A métrica considerada estado da arte para avaliar similaridade de complexos proteína-proteína é o DockQ (BASU et al., 2016). Ela é uma medida de qualidade de modelos de *docking* que varia de 0 a 1, e quanto maior seu valor, mais semelhantes os complexos (modelo e nativo). O cálculo do DockQ *score* leva em consideração três métricas: o iRMS (RMSD da interface), L-RMS (RMSD do ligante) e o Fnat (fração de contatos nativos). Essas medidas são utilizadas pelo *Critical Assessment of PRediction of Interactions* (CAPRI), uma competição de modelagem de estruturas de complexos proteína-proteína (JANIN et al., 2003). Na Figura 4, é mostrada a equação utilizada para calcular o DockQ. Ele é uma média entre o iRMS, L-RMS e Fnat, sendo que os RMSD são normalizados de forma a não se tornarem muito grandes devido ao número de átomos. As distâncias d1 e d2 são constantes e foram otimizadas para os valores de 8.5Å e 1.5Å, respectivamente. (BASU et al., 2016).

$$DockQ(F_{nat}, LRMS, iRMS, d_1, d_2) = (F_{nat} + RMS_{scaled}(LRMS, d_1) + RMS_{scaled}(iRMS, d_2))/3$$

Figura 4. Equação do DockQ.

1.2.2 Docking proteína-peptídeo

Ferramentas de *docking* molecular se mostram eficientes para a descoberta e desenvolvimento de novos fármacos baseados em pequenas moléculas. Entretanto, esses métodos utilizados para pequenos ligantes, muitas vezes não são eficientes para modelar e ancorar peptídeos: moléculas maiores e mais flexíveis. Dessa forma, ferramentas cujo objetivo é realizar o *docking* de peptídeos vêm sendo desenvolvidas e refinadas a cada dia.

De acordo com CIEMNY et al., 2018, existem alguns principais desafios em relação ao *docking* de peptídeos como, por exemplo, lidar com a alta flexibilidade do peptídeo e desenvolver função de pontuação acurada para ranquear as melhores poses. Diferentes ferramentas foram desenvolvidas buscando melhorar a qualidade dos métodos de *docking* proteína-peptídeo e resolver esses problemas utilizando abordagens variadas.

A flexibilidade do peptídeo e do receptor pode ser tratada de diversas formas em ferramentas de *docking* proteína-peptídeo. Em casos de *docking* global, trata-se o receptor como rígido e o peptídeo como flexível, possibilitando busca amostral por um bom sítio de ligação e reduzindo o custo computacional (ALAM et al., 2017; DE VRIES et al., 2017). Outras ferramentas consideram a flexibilidade do receptor limitada apenas às cadeias laterais, ou somente aos resíduos do sítio de ligação, ou ainda utilizam o *ensemble docking*, que realiza a ancoragem do ligante em diferentes conformações rígidas do receptor (LONDON et al., 2011; VERDONK et al., 2003). Além da flexibilidade, muitas vezes o sítio de ligação do peptídeo ainda não é conhecido para a proteína, levando à realização do *docking global* ou *cego*, mais demorado e caro computacionalmente.

Assim como há diversas abordagens para lidar com a flexibilidade, também existem várias funções de pontuação utilizadas para classificar as poses em *docking* proteína-peptídeo, conforme apresentado na seção 1.2. Apesar disso, em alguns casos, os modelos mais bem classificados ainda não correspondem aos modelos mais próximos das poses nativas (AGRAWAL et al., 2019). Segundo os resultados da 6° edição do CAPRI, as melhores funções de pontuações foram as híbridas, ou seja, que unem outras informações ao cálculo da função de pontuação, como informações coevolutivas, por exemplo. (LENSINK et al., 2017)

1.2.3 Ferramentas de docking molecular

Existem diversas ferramentas para o *docking* molecular. Algumas delas, que serão utilizadas neste trabalho, serão apresentadas nesta seção.

O HPEPDOCK (ZHOU et al., 2018) é uma ferramenta de *docking* proteína-peptídeo que leva em consideração a flexibilidade do peptídeo e pode ser utilizada para *docking* cego global ou local, com passagem de sítio como parâmetro. É considerada uma ferramenta híbrida que utiliza o algoritmo MODPEP (YAN et al., 2017) para modelagem *de novo* da estrutura do peptídeo, o MODELLER (WEBB; SALI, 2016) para a modelagem da estrutura da proteína, e uma versão modificada do MDOCK (MA et al., 2021) para a ancoragem do peptídeo na proteína. A função de pontuação utilizada para ranquear os modelos gerados, ITScore-PP (HUANG et al., 2008), foi desenvolvida para ranquear complexos proteína-proteína e é baseada

em conhecimento (*knowledge-based*), ou seja, utilizam informações provenientes de estruturas experimentais conhecidas de complexos proteína-proteína para avaliar e prever a afinidade de ligação. Em WENG et al., 2020, o programa obteve melhor desempenho para *docking* global proteína-peptídeo quando comparado com outras 14 ferramentas. A ferramenta está disponível somente na versão servidor *web*, e pode ser acessada em: (http://huanglab.phys.hust.edu.cn/hpepdock/).

O HDOCK (YAN et al., 2017b) é uma ferramenta de *docking* proteína-proteína e proteína-DNA/RNA que utiliza uma estratégia híbrida para modelar as melhores poses de ligação em sítio de proteína. Ela modela as estruturas do receptor e ligante através do MODELLER (WEBB; SALI, 2016), e realiza o *docking* molecular utilizando um algoritmo baseado em *Fast Fourier Transform* (FFT), no qual a proteína é representada por um modelo reduzido, ou seja, cada cadeia lateral na superfície da proteína é representada somente por seu centro de massa. A função de pontuação usada pelo programa é iterativa e baseada em conhecimento, e leva em consideração a distância do ligante a átomos próximos do receptor. A ferramenta realiza *docking* global ou local, com passagem de sítio. A ferramenta está disponível em versão web e *standalone*, e pode ser acessada em: (http://hdock.phys.hust.edu.cn/).

Apesar de não ser necessariamente uma ferramenta de *docking* molecular, o *AlphaFold Multimer* se mostrou promissor para modelagem de complexos de proteínas (EVANS et al., 2021; VARGA; SCHUELER-FURMAN, 2023). Ele é um método baseado em aprendizado profundo que permite a modelagem teórica de estruturas proteicas a partir de suas sequências de aminoácidos, produzindo previsões precisas, sendo desenvolvido para prever estruturas de complexos proteicos envolvendo múltiplas cadeias. A ferramenta não possui função de pontuação específica, mas ranqueia os modelos com base no *predicted Local Alignment Distance* (plDDT), além de retornar também o *Predicted Aligned Error* (PAE), para cada modelo. O pLDDT é uma pontuação que independe da sobreposição de estruturas. Ela indica a confiabilidade do modelo produzido de acordo com as distâncias dos pares de átomos do modelo, quando comparadas com as distâncias entre aqueles das estruturas de referência. Já o PAE indica a confiabilidade das posições relativas entre pares de aminoácidos do modelo, sendo utilizada para avaliar a posição entre cadeias de proteínas.

1.3 Aprendizado de máquina em bioinformática estrutural

Com o grande aumento na quantidade de dados biológicos sendo gerados nas últimas décadas, passou-se a utilizar o aprendizado de máquina para a resolução de problemas característicos das áreas biológicas. Essas técnicas têm sido empregadas na análise de dados ômicos, na

interpretação de imagens biomédicas, no planejamento de fármacos, entre muitas outras aplicações. (MIN; LEE; YOON, 2016)

Em bioinformática estrutural, o aprendizado de máquina ganhou espaço em problemas desde a predição de modificações pós-traducionais em proteínas, até predição de sítios de ligação. Li e colaboradores (LI et al., 2020) utilizaram perceptron multicamadas na predição de amiloides, agregado de proteínas fibrinas cuja deposição anormal está relacionada com doenças como o Alzheimer. Outros trabalhos utilizaram aprendizado profundo para a predição de estrutura secundária de RNA (FEI et al., 2022), predição de localização subcelular de proteínas (KUMAR et al., 2014) e de interação entre peptídeos e receptores relacionados ao sistema imune (CHENG et al., 2021).

Um problema que desafia os estudos de bioinformática estrutural por anos é o enovelamento de proteínas, ou seja, entender como uma estrutura primária de uma proteína se conforma em uma estrutura tridimensional. Entender como uma proteína se enovela e predizer sua estrutura nativa ainda é um problema em aberto na área. Apesar disso, avanços significativos têm sido alcançados com a inteligência artificial e, principalmente o aprendizado profundo (KUMAR; SRIVASTAVA, 2024; TORRISI; POLLASTRI; LE, 2020).

A DeepMind, subsidiária da Google, introduziu o *AlphaFold* (ALQURAISHI, 2019), uma ferramenta que utiliza aprendizado profundo para modelar estruturas tridimensionais de proteínas a partir das suas sequências de aminoácidos. Dois anos mais tarde, o lançamento do *AlphaFold* 2 atingiu um marco importante ao alcançar os melhores resultados registrados até então no CASP (*Critical Assessment of protein Structure Prediction*), uma competição que avalia métodos de predição de estruturas proteicas (JUMPER et al., 2021). Além disso, foi desenvolvido o *AlphaFold Multimer* (EVANS et al., 2021; JOHANSSON-ÅKHE; WALLNER, 2022), uma versão específica da ferramenta treinada para a modelagem de complexos proteicos. Por fim, neste ano de 2024, foi lançada a mais nova versão da ferramenta, o *AlphaFold 3*, que promete ir além da predição de estruturas de proteínas e predizer também estruturas de RNA, DNA e pequenas moléculas.

Outra utilização importante do aprendizado de máquina aplicado à bioinformática é em problemas relacionados a interações proteína-ligante. Predizer sítios de ligação de proteínas é uma tarefa demorada e cara computacionalmente. A ferramentas GRaSP (SANTANA et al., 2020) utiliza uma estratégia baseada em grafos e aprendizado de máquina supervisionado para predizer resíduos pertencentes a um sítio de ligação de uma proteína, utilizando informações estruturais como acessibilidade ao solvente, características físico-químicas e contatos. Além da predição do sítio de ligação e *hotspots*, outra utilização do aprendizado de máquina é no

desenvolvimento de funções de pontuações de ferramentas de *docking*. Li e colaboradores (LI et al., 2015) utilizaram o Random Forest para melhorar a performance de classificação dos modelos de complexos proteína-ligante gerados pelo AutoDock Vina. Além disso, Yasuo e colaboradores (YASUO; SEKIJIMA, 2019) desenvolveram o SIEVE-Score: uma função de pontuação para triagem virtual de pequenas moléculas onde as energias de ligação são utilizadas em modelo de aprendizado de máquina para predizer os melhores *hits*. Todos esses avanços ilustram como o aprendizado de máquina está transformando a bioinformática, tornando os processos de predição e análise de interações proteína-ligante mais eficientes e precisos.

1.4 Assinaturas estruturais

Com o advento da inteligência artificial (IA) na bioinformática estrutural, um novo desafio emergiu na comunidade científica: a extração de características de estruturas de proteínas e que possam ser usadas como entradas apropriadas para algoritmos de aprendizado de máquina. A capacidade de converter informações estruturais complexas em formatos que os algoritmos de IA possam processar e interpretar é fundamental para avançar nas previsões de estruturas proteicas e nas interações biomoleculares. Dessa necessidade, surgiram as assinaturas estruturais: representações (vetores) que descrevem a estrutura tridimensional de uma proteína. Elas são importantes métodos para identificar características estruturais de proteínas que podem ser importantes para a compreensão de sua função e já foram utilizadas em trabalhos anteriores no grupo de pesquisa visando o melhoramento da catálise de enzimas (MARIANO et al., 2019), identificação de peptídeos terapêuticos (RODRIGUES et al., 2022) e predição de ligantes (PIRES et al., 2013).

Pires e colaboradores (PIRES et al., 2011) desenvolveram uma assinatura baseada em grafos: a assinatura *Cutoff Scanning Matrix* (CSM). O CSM calcula a distância euclidiana entre todos os pares de átomos de carbono alfa dos aminoácidos de uma proteína e gera um vetor que representa o número de pares de átomos encontrados dentro de uma determinada distância de corte. Nesse modelo, cada átomo de carbono alfa é representado como um nó, enquanto as distâncias entre eles são representadas como arestas. Proteínas com estruturas distintas possuem vetores de assinatura diferentes, o que pode ser observado na Figura 6, onde há uma comparação entre o vetor de assinatura de uma proteína globular e outra fibrosa. O CSM demonstrou grande sucesso em tarefas de classificação funcional de proteínas, alcançando uma precisão de até 99,1% em classificação estrutural, atingindo a precisão de 95,4% na classificação de famílias de proteínas (PIRES et al., 2011).

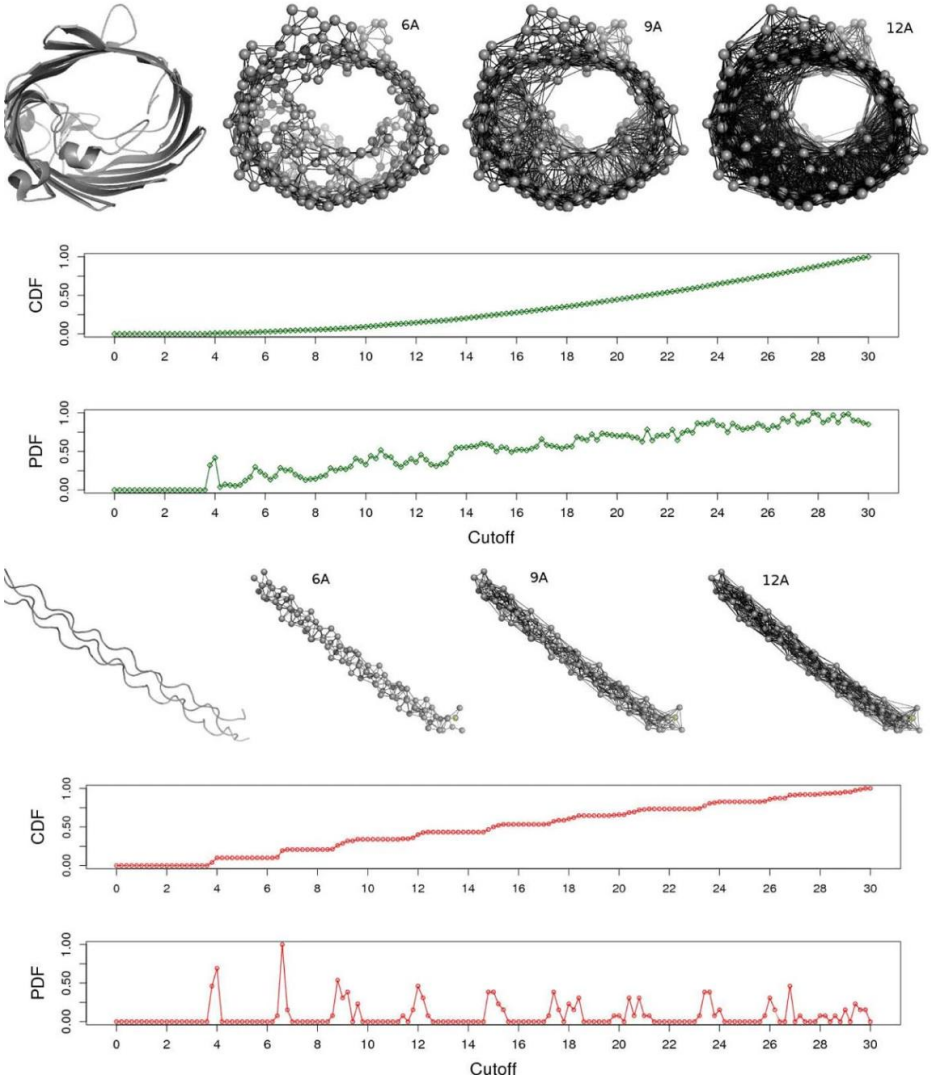


Figura 5. Assinaturas CSM de duas proteínas com estruturas distintas: globular (acima) e fibrosa (abaixo). As proteínas são representadas como grafos em três valores de cutoff distintos: 6Å, 9Å e 12Å.

Fonte: Adaptado de Pires et al, 2011

Mais tarde, o grupo de pesquisa desenvolveu também a assinatura *atomic Cutoff Scanning Matrix* (aCSM) (PIRES et al., 2013). Apesar de muito parecida com o CSM, o aCSM calcula a distância entre todos os átomos da proteína, ao invés de somente carbonos alfa. Além disso, ainda foram criadas outras duas versões do método: aCSM_HP e aCSM_all. Essas novas versões consideram os contatos entre os átomos considerando tipos físico-químicos. No aCSM_HP, considera-se a frequência dos contatos hidrofóbicos, polares e hidrofóbicos-polares, gerando 3 valores por corte, enquanto no aCSM_all considera-se as classes de átomos: hidrofóbico, positivo, negativo, aceptor, doador, aromático, sulfeto e neutro, gerando 36 valores

por corte. O aCSM foi utilizado com sucesso para prever ligantes para proteínas do *Trypanosoma cruzi* (PIRES et al., 2013) partindo da premissa de que sítios com assinaturas semelhantes fazem contatos similares e, consequentemente, permitem interação com ligantes semelhantes.

1.5 Justificativa

A aplicação de métodos de inteligência artificial na bioinformática tem demonstrado uma eficácia crescente, permitindo análises complexas e predições mais acuradas, que antes eram inviáveis, como é o caso da predição de estruturas tridimensionais de proteínas com o *AlphaFold.* Aliado a isso, o uso de assinaturas estruturais das proteínas nos permite fazer predições mais robustas do que utilizando somente informações sobre as sequências, uma vez que a estrutura tridimensional está intimamente relacionada à função que a proteína desempenha.

A pesquisa em fármacos baseados em peptídeos tem ganhado destaque, uma vez que há vantagens significativas quando comparado com as pequenas moléculas, como uma maior eficácia em inibir interações proteína-proteína. Desde o ano 2000 houve aprovação de aproximadamente 40 fármacos peptídicos no mundo (WANG et al., 2022). Embora de grande importância, o estudo dos peptídeos apresenta ainda desafios significativos. O processo de síntese química de peptídeos é dispendioso, e essas moléculas são altamente flexíveis, não possuindo domínios e estruturas definidas, como as proteínas. Isso dificulta a predição precisa de suas estruturas tridimensionais.

Diversos estudos têm sido publicados com o objetivo de predizer estruturas de peptídeos *in silico*, utilizando ferramentas que utilizam aprendizado de máquina como o *AlphaFold* (MAUPETIT; DERREUMAUX; TUFFERY, 2009; MCDONALD et al., 2023; REY et al., 2023; YAN; ZHANG; HUANG, 2017). No entanto, a predição de regiões mais flexíveis e sem estrutura secundária definida ainda representa um desafio. Uma consequência disso é uma maior dificuldade em triagem de peptídeos, uma vez que o *docking* proteína-peptídeo demanda mais tempo e poder computacional do que o *docking* de pequenos ligantes (CIEMNY et al., 2018).

Assim, a pergunta central que orientou este trabalho foi: em uma prospecção de peptídeos ligantes para um alvo específico, seria viável realizar uma triagem inicial utilizando assinaturas estruturais tanto da proteína quanto do peptídeo, antes de proceder ao *docking* molecular? Dessa forma, seria possível identificar os peptídeos mais promissores, permitindo

o uso de técnicas mais específicas e computacionalmente intensivas, como o docking molecular e a dinâmica molecular, apenas para um conjunto reduzido de candidatos potenciais.

Com base nesses aspectos e com esse questionamento, este trabalho visa avaliar o potencial do uso de assinaturas estruturais de proteínas e peptídeos para treinar modelos de aprendizado de máquina que buscam predizer peptídeos ligantes e não ligantes em alvo proteico.

2. Objetivos

2.1 Objetivo geral

Avaliar o potencial preditivo de assinaturas estruturais na identificação da interação ou ligação proteína-peptídeo, utilizando modelos de aprendizado de máquina.

2.2 Objetivos específicos

- Avaliar comparativamente ferramentas de *docking* e modelagem molecular (HPEPDOCK, HDOCK e *AlphaFold Multimer*) para predição de interação proteína-peptídeo.
- Comparar estruturas de peptídeos preditas nas formas *holo* e *apo*.
- Avaliar o uso de assinaturas estruturais para treinar e testar modelos de aprendizado de máquina para predição / identificação de complexos proteína-peptídeos reais e falsos.
- Avaliar o uso de assinaturas estruturais para treinar e testar modelos de aprendizado de máquina para predição de interação proteína-peptídeo.

3. Materiais e métodos

A seção de Materiais e Métodos será organizada em quatro partes distintas (Parte I, Parte II, Parte III e Parte IV), além da descrição da coleta de dados. Cada uma dessas partes corresponderá à metodologia empregada para alcançar os quatro objetivos específicos do projeto, sendo acompanhada por um fluxograma para ilustrar cada etapa do processo.

3.1 Coleta de dados

Para todos os experimentos realizados neste projeto, utilizamos os dados relativos às estruturas dos complexos proteína-peptídeo, da base de dados PepPro (XU et al., 2020). Essa base compreende 89 complexos proteína-peptídeo cujas estruturas foram determinadas experimentalmente através de difração de raios-x, apresentando uma resolução inferior a 2.5 Å, e disponíveis no PDB. Estes complexos englobam peptídeos com comprimentos variados, de 5 a 30 resíduos de aminoácidos, exibindo diversas estruturas secundárias.

Foi realizada uma curadoria na base de dados PepPro, na qual foram excluídos os complexos cujos receptores possuíam mais de uma cadeia com o objetivo de reduzir vieses (PDB IDs: 1UHB, 1FV1, 2OTW, 2PV2, 3BEF, 3H8A, 3MHP, 3WG5, 4ETO, 4QJA, 5CQX), assim como aqueles que não estavam presentes na base de dados Propedia (PDB IDs: 1OAI, 2BZ8, 1UTI). Esta exclusão ocorreu pois os receptores continham menos de 60 resíduos de aminoácidos, um critério de exclusão estabelecido pela Propedia. Ao final, restaram 75 complexos da PepPro, com peptídeos variando de tamanho entre 5 e 30 resíduos de aminoácidos e com funções e estruturas secundárias variadas, como mostra a Tabela S1.

3.2 Parte I - Avaliação de ferramentas de *docking* e modelagem molecular para complexos proteína-peptídeo.

Nesta seção do trabalho, descreveremos a metodologia empregada para alcançarmos o primeiro objetivo proposto: avaliar ferramentas de *docking* e modelagem molecular para complexos proteína-peptídeo. A Figura 7 ilustra o fluxograma desta seção.

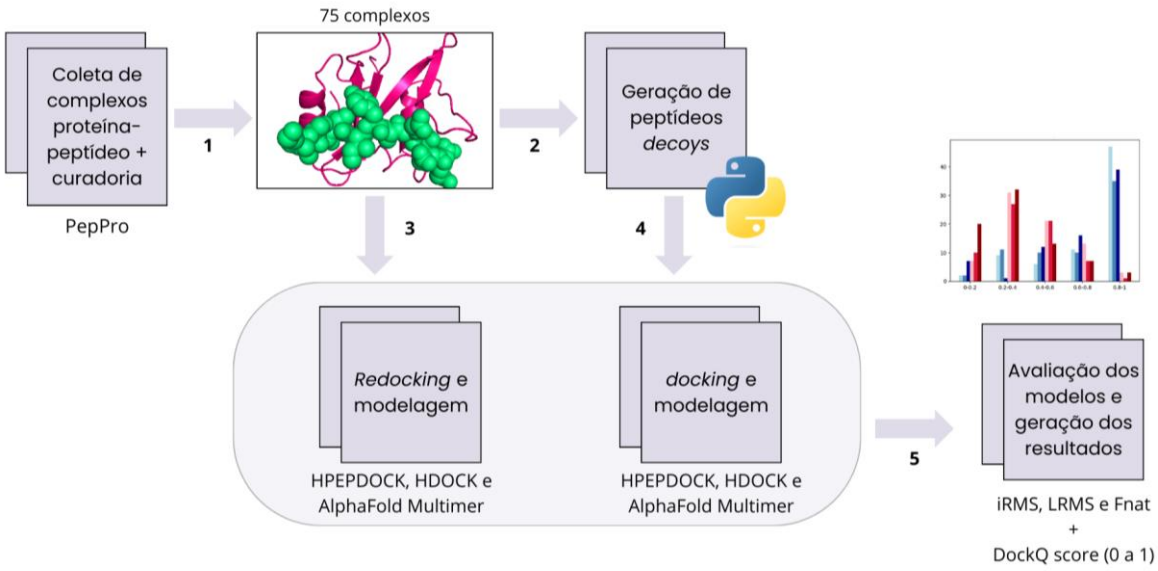


Figura 6. Fluxograma da etapa I do projeto: avaliação de ferramentas de docking e modelagem molecular para complexos proteína-peptídeo. 1. As estruturas dos 75 complexos foram obtidas da base de dados PepPro. **2.** Foi gerado um peptídeo *decoy* para cada peptídeo pertencente a cada um dos 75 complexos. **3.** e **4.** As modelagens e *dockings* foram feitas utilizando as ferramentas HPEPDOCK, HDOCK e *AlphaFold Multimer*. **5.**Ao final, foi realizada avaliação dos modelos e comparação das ferramentas utilizando o DockQ score, iRMS, L-RMS e Fnat.

Fonte: próprio autor.

3.2.1 Docking e modelagem molecular

Com o objetivo de avaliar ferramentas de *docking* e modelagem na predição de interação proteína-peptídeo, realizamos experimentos de *redocking*/modelagem com três ferramentas distintas: HPEPDOCK, HDOCK e *AlphaFold Multimer*.

3.2.2 Redockings

Realizou-se o *redocking* dos 75 complexos selecionados da base de dados PepPro em cada uma das três ferramentas descritas anteriormente, com o intuito de testá-las para *docking* de peptídeos. Nas ferramentas específicas de *docking* (HPEPDOCK e HDOCK), as entradas foram o arquivo *.pdb* dos receptores e a sequência dos peptídeos em formato *.fasta*. Os dockings foram feitos sem passagem de sítio de ligação como parâmetros de entrada (*docking* cego), com os parâmetros *default* e para ambas as ferramentas, utilizou-se versão *web server*.

Para a ferramenta de modelagem *AlphaFold Multimer*, a entrada utilizada foi um arquivo *.fasta* contendo duas cadeias: o receptor (cadeia A) e o peptídeo (cadeia B). Esses arquivos foram gerados a partir dos arquivos *.pdb* presentes na base de dados PepPro. Todos os parâmetros usados foram os *default*. Utilizou-se a versão 2.2.0 do *AlphaFold Multimer* presente

nos servidores do NMRbox (MACIEJEWSKI et al., 2017) (https://nmrbox.nmrhub.org/). As especificações da máquina escolhida foram: Molybdenum (molybdenum.nmrbox.org), contendo 128 núcleos de processamento e 128 threads AMD EPYC 7H12, 256 GB RAM, e GPU NVIDIA A100.

3.2.3 Dockings com peptídeos decoys

Para testar as ferramentas, geramos peptídeos *decoys.* As moléculas consideradas *decoys* são aquelas que não possuem a atividade biológica desejada, e por isso são usadas como controles negativos em estudos de triagem virtual e experimentos de *docking*. Essas moléculas são projetadas para se parecerem estruturalmente com os compostos ativos (ligantes) em algum grau, mas não interagem efetivamente com o receptor de interesse. (HUANG; SHOICHET; IRWIN, 2006). Os *decoys* são importantes para teste de ferramentas de *docking* molecular e treinamento e teste de modelos de aprendizado de máquina. Para a geração dos peptídeos *decoys*, realizamos a permutação dos resíduos dos peptídeos originais presentes na base de dados, por meio de script em *python*. Assim, foi possível manter a proporção dos tipos de resíduos e o mesmo tamanho dos peptídeos, mas mudando sua estrutura. A Tabela S2 mostra a sequência de cada peptídeo *decoy* gerado de acordo com o respectivo peptídeo real.

Realizou-se o *docking*/modelagem dos 75 receptores selecionados da PepPro com peptídeos *decoys*, nas mesmas ferramentas descritas anteriormente, nas mesmas condições e com os mesmos parâmetros.

3.2.4 Padronização dos arquivos

Foi necessária uma padronização dos arquivos gerados pelas três ferramentas, utilizando *scripts* em *python*. As modificações foram: mudança no nome das cadeias referentes ao peptídeo de cada complexo da PepPro (antes: cadeia P, agora: cadeia B); concatenação dos arquivos dos modelos (peptídeos) com o do receptor presente na PepPro (somente para os modelos gerados pelo HPEPDOCK) e renumeração dos resíduos (modelos do HPEPDOCK e HDOCK). No caso dos modelos gerados pelo *AlphaFold Multimer*, não foi preciso realizar modificação.

3.2.5 Avaliação dos modelos proteína-peptídeo

Para avaliar a qualidade dos *dockings*, calculou-se o DockQ (BASU et al., 2016) dos modelos dos complexos gerados, comparando com os originais presente na base de dados. A ferramenta utilizada pode ser encontrada em https://github.com/bjornwallner/DockQ/. Somente os modelos com o maior valor de DockQ de cada um dos 75 complexos (melhor modelo representando o complexo com peptídeo real e o melhor modelo representando o complexo

com peptídeo *decoy*) foi utilizado para as estatísticas e análises das ferramentas, bem como para os passos posteriores do trabalho.

Para a comparação estatística entre os resultados obtidos, foi realizado um teste-t de student comparando as médias dos valores de DockQ obtidos pelos complexos reais e falsos.

3.3 Parte II - Comparação da modelagem de peptídeos nas formas *holo* e *apo*

Com o objetivo de comparar os 75 peptídeos da PepPro modelados nas formas *holo* e *apo*, utilizamos a ferramenta *AlphaFold 2*, na versão *monomer*s. Os peptídeos na forma *holo* (complexos proteína-peptídeo) já haviam sido modelados na parte anterior do projeto (Parte I). Já para a obtenção dos peptídeos na forma *apo*, utilizamos como entrada para o *AlphaFold 2* um arquivo *.fasta* contendo a sequência do peptídeo. As especificações e parâmetros adicionais da ferramenta foram os mesmos dos descritos no item 3.2.2. Ao final, foram geradas 5 estruturas de cada peptídeo, cuja aquela que possuísse o maior pLDDT foi selecionada para a próxima etapa.

Para comparar as estruturas *holo* e *apo* de cada peptídeo, foi calculado o RMSD de todos os átomos, e dividido pelo número de resíduos de cada peptídeo (RMSD normalizado). A Figura 8 ilustra o fluxo de trabalho desta etapa.

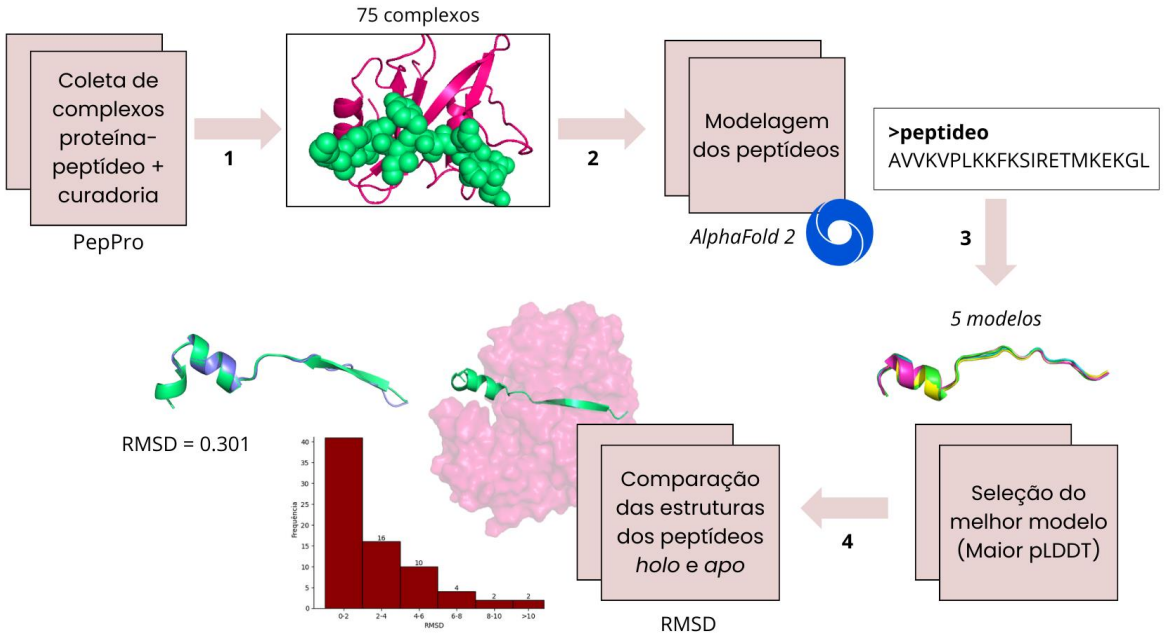


Figura 7. Fluxograma da etapa II do projeto: comparação entre peptídeos modelados de forma *holo* e *apo*. **1.**As sequências dos 75 peptídeos foram obtidas da base de dados PepPro. **2.**O AlphaFold 2 foi utilizado para a modelagem dos peptídeos na forma *apo*. **3.**Dos cinco modelos gerados, somente o de maior pLDDT foi selecionado. **4.**Foi feita a comparação das estruturas dos peptídeos *holo* e *apo* por meio do cálculo do RMSD dos carbonos alpha dos peptídeos.

Fonte: próprio autor.

3.4 Parte III - Modelos de predição de complexos reais e falsos.

A Parte III deste projeto tem como objetivo criar modelos de aprendizado de máquina para avaliar a capacidade de distinguir complexos reais (proteína receptora ligada ao peptídeo real, que já foram resolvidos experimentalmente) de complexos falsos (proteínas ligadas a peptídeos *decoys*), utilizando assinaturas estruturais. A Figura 9 apresenta um fluxograma desta etapa do trabalho.

Todos os experimentos aqui foram realizados com os 75 complexos experimentais extraídos da base de dados PepPro (reais) e, para cada complexo real, um complexo falso, resultado de docking com peptídeo *decoy*.

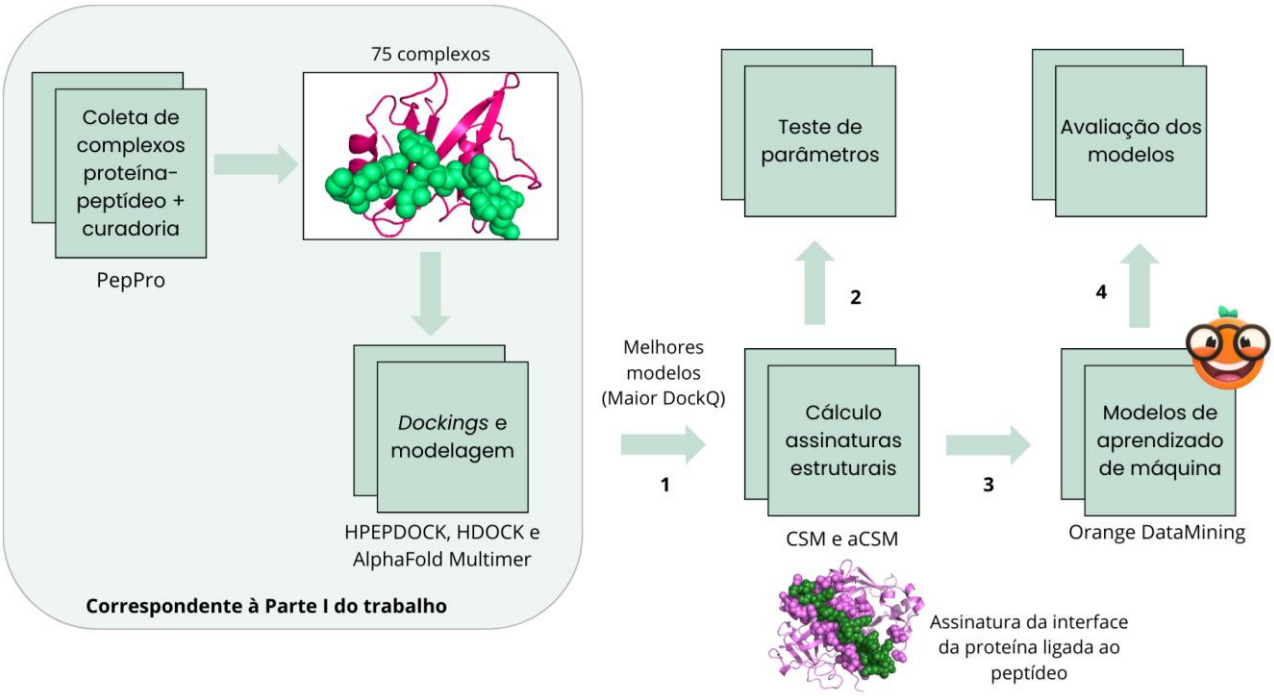


Figura 8. Fluxograma da etapa III do projeto: Modelos de predição de complexos reais e falsos. 1. Somente os melhores modelos (com maior DockQ) foram selecionados para cada um dos 75 complexos: reais e com *decoys*. **2**.Foram testados alguns parâmetros como valores de corte para definir a interface (4,5,6,6.5,7 e 8), tipos de assinaturas (CSM e aCSM_all) e parâmetros das assinaturas (*limite de corte* e *passo de corte*). **3.** As assinaturas foram calculadas para a interface da proteína ligada ao peptídeo e utilizadas para treinamento de modelos de aprendizado de máquina usando a ferramenta Orange DataMining. **4.**Foi feita a avaliação dos modelos usando as métricas AUC, F1-score, MCC, precisão, revocação e especificidade.

Fonte: próprio autor.

3.4.1 Cálculo das assinaturas estruturais

As assinaturas empregadas neste trabalho são a *Cutoff Scanning Matrix* (CSM) e a atomic *Cutoff Scanning Matrix* all (*aCSM_all*). A escolha dessas assinaturas baseia-se em estudos anteriores. Em PIRES et al., 2013, o *aCSM_all* demonstrou um desempenho superior na tarefa de predição de ligantes para sítios de proteínas devido a ser uma assinatura mais complexa e conter informações sobre propriedades físico-químicas em nível atômico. Além disso, buscamos compará-lo com a versão clássica do CSM. Todos os cálculos das assinaturas foram realizados utilizando a biblioteca SIGNA (presente em: https://github.com/lbs-ufmg/signa), com o *script* implementado de acordo com Pires. at al, 2013.

É importante salientar que nesta etapa as assinaturas estruturais foram calculadas utilizando os átomos da interface da proteína complexada com o peptídeo, como ilustra a Figura 10. Em esferas roxas, são representados os átomos da proteína que se encontram na região de interface com o peptídeo. Em esferas verdes, os átomos do peptídeo. O retângulo vermelho mostra a região utilizada para o cálculo da assinatura: os átomos da interface da proteína complexada com os átomos do peptídeo.

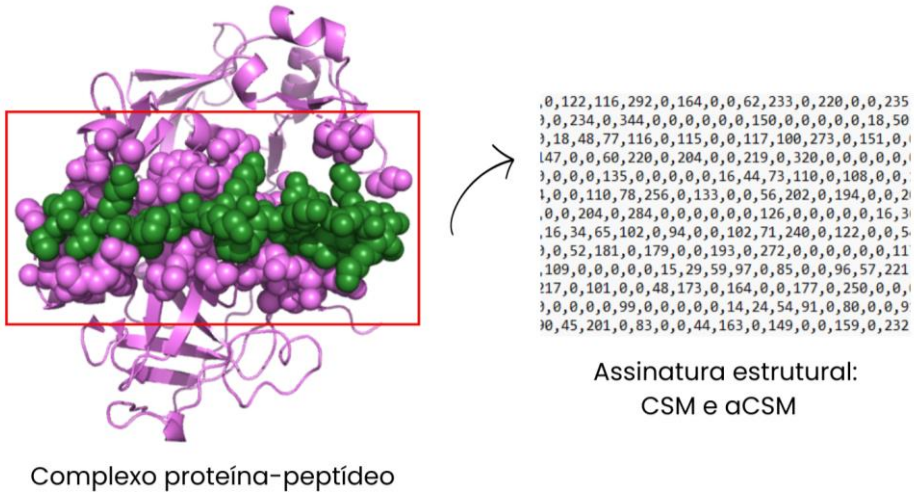


Figura 9. Representação do cálculo da assinatura estrutural dos átomos da interface da proteína complexada com os átomos do peptídeo. PDB ID: 1AVF. Cadeia A (proteína) em roxo, e cadeia B (peptídeo), em verde. O retângulo vermelho mostra que a assinatura estrutural é gerada a partir dos átomos da interface da proteína (esferas roxas) ligado ao peptídeo (esferas verdes).

Fonte: próprio autor.

3.4.2 Teste de parâmetro: interface

Nesta seção e na seção seguinte (3.4.3), realizamos testes com alguns parâmetros importantes para o cálculo da assinatura. O primeiro deles é a definição da interface, ou seja, qual seria o melhor valor de corte, em angstroms, para definir quais átomos da proteína possivelmente estão em contato com os átomos do peptídeo. Para isso, um *script* em *Python* foi criado para extrair

os átomos da proteína que estivessem a uma distância de corte do peptídeo, em angstroms, juntamente com os átomos do peptídeo, em um arquivo *.pdb*. Assim, foram gerados 6 arquivos de interfaces para cada um dos 450 complexos proteína-peptídeo. Os valores testados (4Å, 5Å, 6Å, 6.5Å, 7Å e 8Å) foram decididos com base em estudos anteriores, onde foram testados os mesmos pontos de corte para definição de interface (MARIANO et al., 2019; PIRES et al., 2011).

A assinatura *aCSM_all* com os parâmetros *default* (limite de corte = 30 e passo de corte = 0.2) foi calculada para cada um dos arquivos *.pdb* gerados e utilizada para treinamento de um modelo de classificação (real ou falso) usando redes neurais, na ferramenta Orange DataMining, com os seguintes parâmetros *default*: *Neurons in hidden layers:* 500, *activation*: "ReLu", *solver*: "*Adam", regularization* α = 0, *maximal number of iterations:* 200 e *replicable training.* O modelo foi testado com validação cruzada (10-fold).

3.4.3 Teste de parâmetros: assinaturas estruturais

Outros parâmetros a serem testados para a construção do modelo de classificação usando as assinaturas estruturais são os tipos de assinatura (CSM e *aCSM_all*), e os parâmetros limite de corte e passo de corte de cada assinatura. Os valores padrões de cada um dos parâmetros são de 30 Å e 0.2Å, respectivamente. Para uma assinatura CSM gerada com esses parâmetros, serão gerados 150 valores, um para cada intervalo de distância: 0Å-0.2Å, 0.2Å-0.4Å, 0.4Å-0.6Å, e assim por diante até 29.8Å-30Å. Já no caso da assinatura aCSM_all, serão gerados 36 valores para cada intervalo, considerando a natureza do átomo, portanto usando os mesmos parâmetros *default*, a assinatura aCSM_all gera um vetor composto por 5.400 valores.

Calculamos as assinaturas CSM e *aCSM_all* para todas as interfaces e peptídeos dos complexos proteína-peptídeo (450) usando os parâmetros limite de corte e passo de corte listados na Tabela 2. Esses parâmetros foram definidos empiricamente e consideramos a interface sendo os átomos da proteína a 6Å de distância do peptídeo.

Tabela 2. Parâmetros dos experimentos com os cálculos de assinaturas CSM e aCSM_all.

Teste	Assinatura	Limite de corte	Passo de corte
1	CSM	30	0.2
2	CSM	20	0.2
3	CSM	10	0.1
4	aCSM	30	0.2
5	aCSM	20	0.2

6	aCSM	10	0.1

Desenvolvemos um modelo de classificação no Orange DataMining para cada um dos testes apresentados na Tabela 2. Todos os modelos foram configurados com os seguintes parâmetros *default*: *Neurons in hidden layers:* 500, *activation*: "ReLu", *solver*: "*Adam", regularization* α = 0, *maximal number of iterations:* 200 e *replicable training.* Os modelos foram testados com validação cruzada (10-fold).

3.4.4 Modelos de aprendizado de máquina

Após a definição dos melhores parâmetros, passamos para a parte de desenvolvimento do modelo preditivo para realizar a tarefa de classificar complexos como reais ou falsos.

A construção dos modelos supervisionados foi feita por meio da ferramenta Orange Data Mining (DEMŠAR et al., 2004). Os algoritmos testados foram: *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), Redes Neurais, *Gradient Boosting*, Regressão Logística, *Random Forest*, *Naive Bayes* e *Decision Tree*.

Realizamos testes de modificações de hiperparâmetros de cada modelo e os melhores parâmetros estão citados abaixo.

- **KNN:** K = 7, metric = *"Euclidian"*, weight: *"Uniform"*
- **v-SVM:** *linear kernel*, *numerical tolerance*: 0.0010, *interaction limit*: 500.
- **Redes Neurais:** *Neurons in hidden layers:* 500, *activation*: "ReLu", *solver*: "L-Bfgs-B", *regularization* α = 0, *maximal number of iterations:* 100 e *replicable training.*
- ***Gradient Boosting***: *Gradient Boosting (scikit-learn)*, number of trees: 400, learning rate "0.100", *replicable training*, *growth control for limit depth of individual trees*: 2, "*do not split subsets smaller than 2*" e *fraction of training instances*: 1.00
- **Regressão logística:** *regularization type of "lasso (L1)"* e *strength c* = 0.05.
- ***Random forest*:** 10 *trees* e *minimum length of subsets*: 5
- ***Naive Bayes*:** Não possui parâmetros ajustáveis.
- ***Decision Tree*:** *"induce binary tree", Min. number of instances in leaves:* 5, "*do not split subsets smaller than 5*", *limit the maximal tree depth to 100", "Stop when majority reaches 95%".*

Os modelos foram criados somente usando os melhores parâmetros obtidos dos itens anteriores: *aCSM_all*, limite de corte = 10, passo de corte = 0.1, interface ≤ 6Å.

3.4.5 Avaliação dos modelos de aprendizado de máquina

Para avaliar todos os modelos criados foi utilizada a validação cruzada (10-fold) e para a comparação dos modelos, foram usadas as métricas: área sob a curva ROC, revocação, especificidade, acurácia, precisão, F-score e Coeficiente de Correlação de Matthews (MCC). Todas elas relacionam, de formas diferentes, as taxas de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos. A fórmula utilizada para o cálculo de cada métrica se encontra na Tabela 3.

A Curva Característica de Operação do Receptor (do inglês, ROC) é um gráfico utilizado para avaliar classificadores, cujo eixo X corresponde à taxa de falsos positivos (ou 1-especificidade) e o eixo Y corresponde à taxa de verdadeiros positivos (ou revocação). Como métrica neste trabalho, também utilizamos a área sob a curva ROC (do inglês, AUC). Essa área varia de 0 a 1, sendo 0.5 um classificador aleatório.

Tabela 3. Métricas para avaliação de modelos de aprendizado de máquina.

Método	Fórmula
Sensibilidade (revocação)	TP / (TP+FN)
Especificidade	TN / (FP+TN)
Acurácia	(TP+TN) / N
Precisão	TP / (TP+FP)
F-score	2 x (PxS) / (P+S)
MCC	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$

TP = Verdadeiros positivos, FN = Falsos negativos, TN = Verdadeiros negativos, FP = Falso positivos, P = Precisão, S = Sensibilidade, N = total de elementos

Fonte: Adaptado de MARIANO, 2021.

3.5 Parte IV - Modelos de predição de interação proteína-peptídeo

A Parte IV deste projeto tem como objetivo criar modelos de classificação para predição de interação entre proteína e peptídeo. A Figura 11 apresenta um fluxograma desta etapa do trabalho.

Assim como na Parte III, os experimentos serão realizados com os complexos modelados na primeira parte do trabalho, referentes aos complexos da base de dados PepPro.

Somente o melhor modelo de cada um dos 450 complexos (75 modelos reais + 75 modelos falsos, para cada uma das três ferramentas) será utilizado nas próximas etapas.

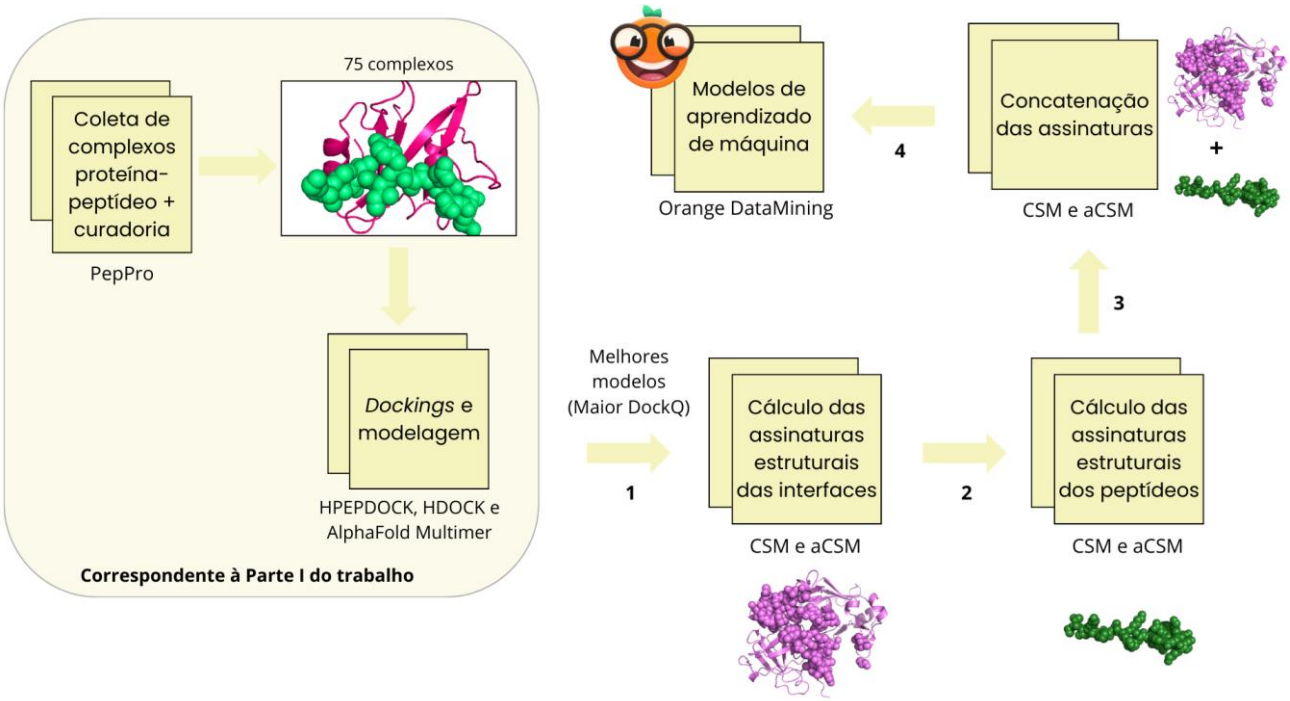


Figura 10. Fluxograma da etapa IV do projeto: Modelos de predição de interação proteína-peptídeo **1.**Somente os melhores modelos (com maior DockQ) foram selecionados para cada um dos 75 complexos: reais e com *decoys.* **2.**Foram calculadas as assinaturas somente dos átomos da interface da proteína, e separadamente, dos átomos do peptídeo modelados na forma *apo* pelo *AlphaFold Multimer*. **3.** Ambas as assinaturas foram concatenadas em somente um arquivo CSV. **4.** As assinaturas concatenadas foram utilizadas para treinamento de modelos de aprendizado de máquina usando a ferramenta Orange DataMining. Foi feita a avaliação dos modelos usando as métricas AUC, F1, MCC, precisão, revocação, especificidade.

Fonte: próprio autor.

3.5.1 Cálculo das assinaturas estruturais

As assinaturas utilizadas e a biblioteca usada para gerá-las já foram descritas na Parte III deste trabalho.

Nesta etapa, o modo com que as assinaturas foram utilizadas foi ligeiramente diferente. Enquanto na parte III, as assinaturas foram retiradas do complexo proteína-peptídeo ligado, aqui, as assinaturas serão retiradas separadamente e concatenadas ao final, como mostra a Figura 12. Em esferas roxas, são representados os átomos da proteína que se encontram na região de interface com o peptídeo. Em esferas verdes, os átomos do peptídeo. Esse peptídeo se refere ao peptídeo modelado na sua forma *apo* pelo *AlphaFold Multimer*. Ambas as assinaturas foram geradas separadamente e concatenadas em um único arquivo, gerando uma assinatura com o dobro do tamanho. Por exemplo, utilizando a assinatura a*CSM_all* com os

parâmetros de limite de corte = 30 e passo de corte = 0.2, obtém-se uma assinatura com o tamanho de 5.400 para a proteína, outra assinatura de tamanho 5.400 para o peptídeo, e concatenando ambas, tem-se uma assinatura final de tamanho 10.800.

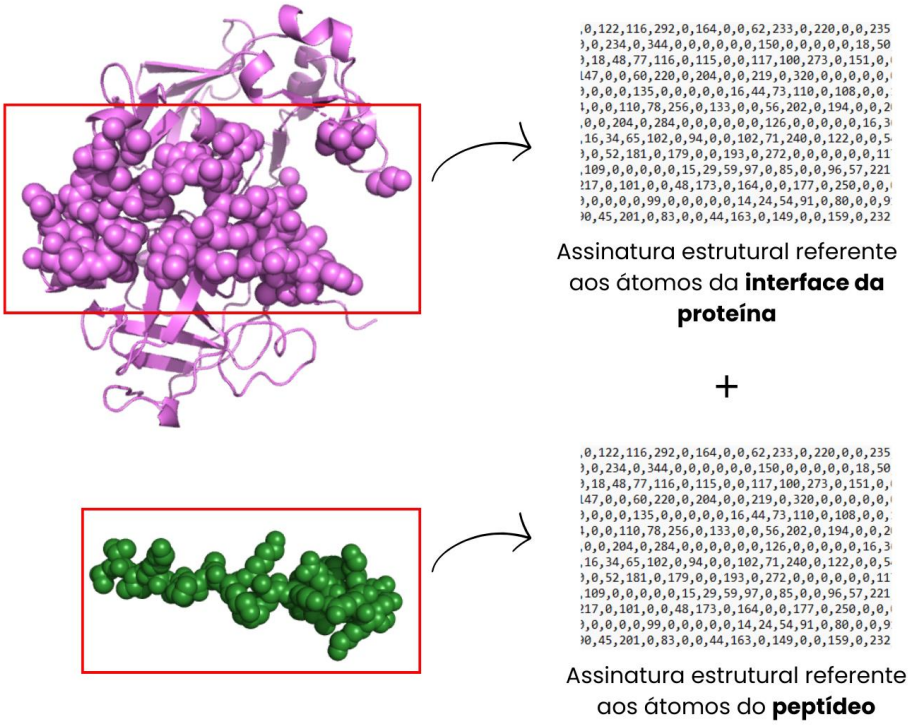


Figura 11. Representação do cálculo da assinatura estrutural da interface da proteína e do peptídeo separadamente. PDB ID: 1AVF. Cadeia A (proteína) em roxo, e cadeia B (peptídeo), em verde. É retirada a assinatura dos átomos da interface da proteína (esferas roxas) e, separadamente, é retirada a assinatura de todos os átomos do peptídeo. Essas assinaturas são concatenadas, gerando somente um arquivo.

Fonte: próprio autor.

3.5.2 Teste de parâmetros: assinaturas estruturais

Testamos a assinatura *aCSM_all* e seus parâmetros limite de corte e passo de corte segundo a Tabela 4.

Tabela 4. Parâmetros dos experimentos com os cálculos de assinaturas separadas para as proteínas (interface) e peptídeos.

	Proteína			Peptídeo		
Teste	Assinatura	Limite de corte	Passo de corte	Assinatura	Limite de corte	Passo de corte
1	aCSM	30	0.2	CSM	30	0.2

2	aCSM	20	0.2	CSM	20	0.2
3	aCSM	10	0.1	CSM	10	0.1

Desenvolvemos um modelo de classificação no Orange DataMining para cada um dos testes. Todos os modelos foram configurados com os seguintes parâmetros: *Neurons in hidden layers:* 500, *activation*: "ReLu", *solver*: "*Adam", regularization* α = 0, *maximal number of iterations:* 200 e *replicable training.* Os modelos foram testados com validação cruzada (10-fold).

3.5.3 Modelos de aprendizado de máquina

A construção dos modelos supervisionados foi feita por meio da ferramenta Orange Data Mining. Os algoritmos usados foram os mesmos utilizados na Parte III do projeto.

Os parâmetros de cada modelo estão citados abaixo:

- **KNN:** *K = 3, metric = "Euclidian", weight: "Uniform"*
- **SVM:** *C = 5, regression loss epsilon = 0.1, linear kernel, numerical tolerance*: 0.0010, *interaction limit*: 500.
- **Redes Neurais:** *Neurons in hidden layers:* 500, *activation*: "ReLu", *solver*: "L-Bfgs-B", *regularization* α = 0, *maximal number of iterations:* 100 e *replicable training.*
- ***Gradient Boosting***: *Gradient Boosting (scikit-learn),* number of trees: 200, learning rate "0.100", *replicable training, growth control for limit depth of individual trees*: 3, "*do not split subsets smaller than 2*" e *fraction of training instances*: 1.00
- **Regressão logística:** *regularization type of "lasso (L1)"* e *strength c* = 1.
- ***Random forest*:** 5 *trees* e *minimum length of subsets*: 5
- ***Naive Bayes:*** Não possui parâmetros ajustáveis.
- ***Decision Tree:*** *"induce binary tree", Min. number of instances in leaves:* 5, "*do not split subsets smaller than 10*", *limit the maximal tree depth to 100", "Stop when majority reaches 95%".*

Os modelos foram criados somente usando os melhores parâmetros obtidos dos itens anteriores: *aCSM_all*, limite de corte = 20, passo de corte = 0.2, interface de até 6Å. Além disso, foram testados e avaliados utilizando as mesmas métricas presentes na Tabela 3.

4. Resultados e discussões

A seção de resultados e discussões será dividida em quatro partes, assim como a metodologia. Cada uma das partes (I, II, III e IV) apresentará os resultados das respectivas metodologias.

Para conduzir as etapas subsequentes, utilizamos os dados da base PepPro. Optamos por esta base devido ao seu tamanho pequeno e à diversidade dos complexos proteína-peptídeo contidos. Comparativamente, a PepPro é considerada pequena em relação à Propedia 2, que atualmente abriga mais de 49 mil complexos. No entanto, escolhemos a PepPro para investigar minuciosamente o comportamento de complexos diversos em relação aos objetivos deste estudo. Apesar de seu tamanho reduzido, a PepPro contém uma variedade significativa de complexos, incluindo peptídeos de tamanhos diversos e estruturas secundárias variadas. Após concluir este estudo e com base na metodologia empregada, planejamos expandir nossas análises e previsões utilizando a base de dados Propedia 2. Acreditamos que poderemos desenvolver modelos de predição ainda mais acurados com uma base de dados mais extensa.

4.1 Parte I - Comparação de ferramentas de *docking* e modelagem molecular

Nesta primeira etapa do trabalho, avaliamos três ferramentas para *docking*/modelagem proteína-peptídeo: HPEPDOCK, HDOCK e *AlphaFold Multimer*. A escolha de cada uma dessas ferramentas para as análises se deve a cada um destes métodos representar um paradigma diferente. O HPEPDOCK é uma ferramenta para *docking* flexível proteína-peptídeo (considera peptídeos com até 30 resíduos), e que foi considerada a melhor ferramenta de docking proteína-peptídeo em um estudo considerando 13 outras ferramentas de *docking* global (WENG et al., 2020). Enquanto isso, o HDOCK foi desenvolvido para *docking* proteína-proteína, e realiza *docking* rígido. Entretanto, obteve sucesso predizendo modos de ligação proteína-peptídeo em desafios propostos no CAPRI (YAN et al., 2017a), além de ser utilizado por Rajpoot et al., 2021 para avaliar modo de ligação de peptídeo no sítio de ligação da proteína Spike do SARS-CoV-2. O *AlphaFold Multimer,* apesar de não ser uma ferramenta de *docking* molecular, tem demonstrado a capacidade de modelar complexos proteína-peptídeo, incluindo trabalhos sugerindo que a ferramenta conseguiria decidir qual o peptídeo com maior afinidade, dentre dois, para um sítio de uma proteína. (JOHANSSON-ÅKHE; WALLNER, 2022; VARGA; SCHUELER-FURMAN, 2023)

Cada *docking* realizado pelas ferramentas HDOCK e HPEPDOCK gerou 100 modelos para cada um dos 75 complexos proteína-peptídeo, enquanto cada modelagem realizada pelo *AlphaFold Multimer* gerou 25 modelos. Esses números são os padrões para cada ferramenta.

Esses modelos foram classificados em relação ao DockQ *score*, e somente o melhor modelo para cada complexo (menor valor de DockQ *score*) foi selecionado para as análises posteriores.

A ferramenta HDOCK apresentou a seguinte limitação: 7 peptídeos originais e 9 peptídeos *decoys* não conseguiram ser modelados e ancorados. Nossa hipótese é que o HDOCK realiza alinhamento de sequências e modelagem comparativa para modelar a proteína e o ligante quando as entradas são somente as sequências de aminoácido (usando HH-suite52, ClustalW53 e MODELLER). A ferramenta provavelmente não encontrou estrutura similar à do peptídeo para modelá-lo e realizar o *docking.* Além disso, alguns peptídeos, mesmo quando modelados e ancorados, tiveram sua sequência de aminoácidos cortada nas extremidades C e N terminal, provavelmente também devido a modelagem comparativa. O mesmo não ocorreu com o HPEPDOCK, que utiliza modelagem *de novo* para o peptídeo. Dessa forma, utilizamos 68 dos 75 peptídeos originais e 66 dos 75 peptídeos *decoys* ancorados pela ferramenta HDOCK para os experimentos subsequentes.

4.1.1 iRMS, L-RMS e Fnat

Para avaliar as estruturas modeladas e ancoradas, usou-se as métricas utilizadas pelo CAPRI: iRMS (RMSD da interface), L-RMS (RMSD do ligante) e Fnat (fração de contatos nativos).

O cálculo do RMSD é amplamente utilizado para avaliação de ferramentas de modelagem molecular e *docking* molecular, seja de proteína-proteína ou proteína-pequena molécula. Além de seu cálculo ser simples, ele permite uma comparação direta entre duas estruturas, independente da quantidade de resíduos ou átomos. Por outro lado, o valor do RMSD tende a ser mais alto quanto maior o número de átomos da proteína, sendo sensível a pequenas diferenças estruturais. Dessa forma, o RMSD deve ser utilizado juntamente a outras métricas para melhor avaliar a qualidade de um modelo.

Nos gráficos da Figura 13, apresentamos as frequências de complexos por intervalos de iRMSD e L-RMSD. É possível perceber que os complexos reais (barras em tons de amarelo, no gráfico à esquerda e azul, no gráfico à direita) obtiveram valores mais baixos de RMSD (entre 0 e 2), ou seja, são estruturas que ficaram mais parecidas com os complexos experimentais. Já os complexos falsos (usando peptídeos *decoys*) obtiveram maiores valores de RMSD e mais distribuídos pelos gráficos (barras em tons de laranja, no gráfico à direita e azul no gráfico à esquerda), mostrando mais baixa qualidade dos modelos, e algumas vezes, ligantes localizados longe do sítio de ligação real da proteína, como esperado. Isto indica que as ferramentas foram capazes de diferenciar ligantes reais de ligantes falsos (*decoys*).

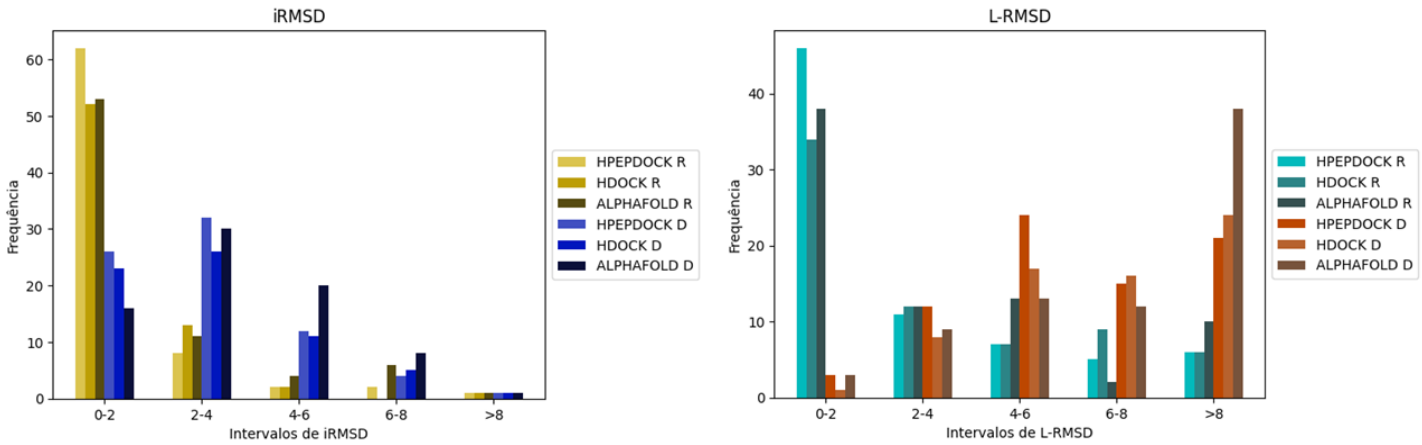


Figura 12. Frequência de complexos proteína-peptídeo modelados por intervalos de iRMSD e L-RMSD. O gráfico à direita representa a frequência de complexos nos intervalos de L-RMS. As barras em tons de azul representam os complexos reais (R) e as barras em tons de laranja representam os complexos falsos (*decoys*, D). O gráfico à esquerda representa a frequência de complexos nos intervalos de iRMS. As barras em tons de amarelo representam os complexos reais (R) e as barras em tons de azul representam os complexos falsos (*decoys*, D). Em ambos os gráficos observamos que os complexos reais se encontram em intervalos de valores menores de iRMS e L-RMS.

fonte: próprio autor

Além do RMSD, é importante avaliarmos os contatos entre os resíduos da interface que se mantiveram da estrutura original no modelo predito, ou seja, calcular a fração de contatos nativos (Fnat). O Fnat é a razão entre o número de contatos nativos preservados na interface do modelo e o número de contatos nativos totais existentes na interface original. Ele varia de 0 a 1, e quanto mais próximo de 1, mais contatos nativos foram mantidos entre as estruturas comparadas.

Segundo a Figura 14, a Fnat dos modelos de reais (barras em tons verdes) foi maior comparada com a Fnat dos modelos falsos (barras em tons roxos). Esse resultado corrobora com os resultados anteriores, mostrando que mais contatos foram mantidos do complexo experimental naqueles modelos dos complexos reais quando comparados com os complexos com *decoys*.

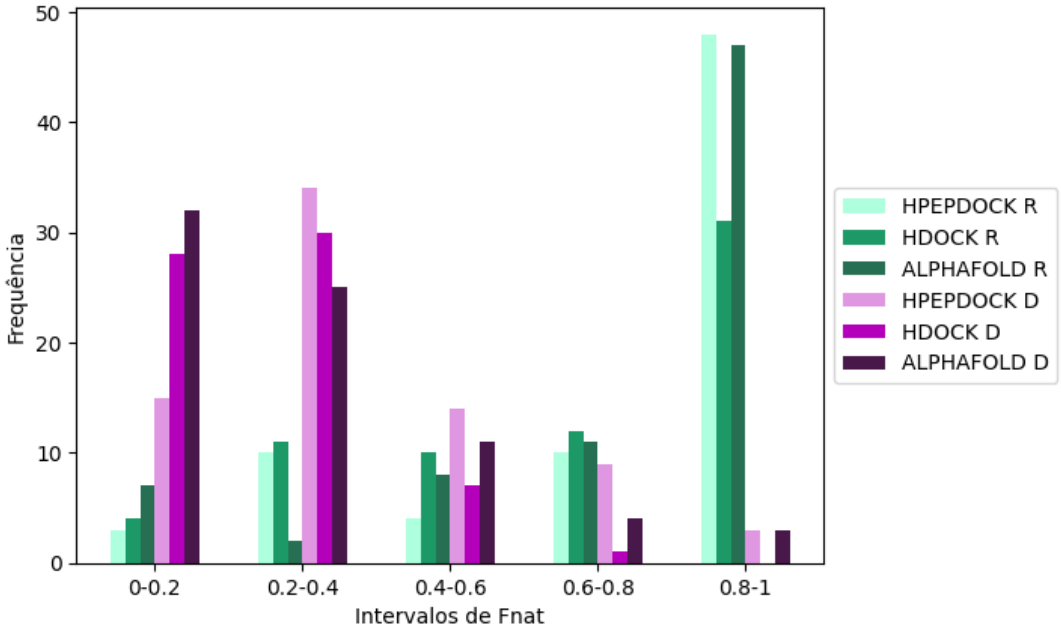
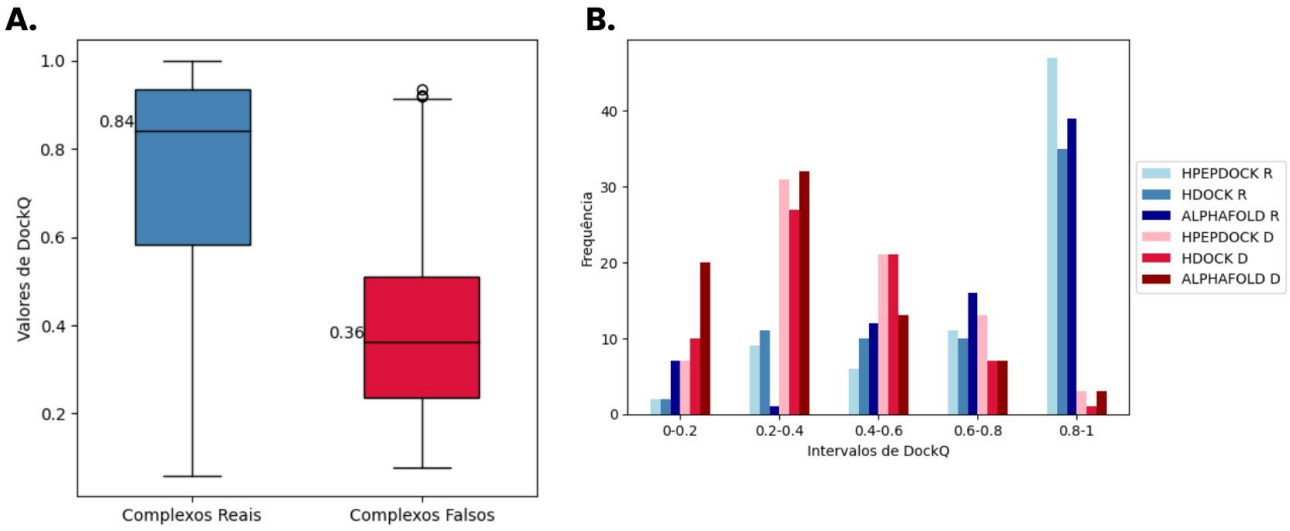


Figura 13. Frequência de complexos proteína-peptídeo modelados por intervalos de Fnat. As barras em tons de verde representam os complexos reais, e as em tons de roxo, os complexos falsos. Observamos que os complexos reais apresentaram valores maiores de Fnat, do que os complexos falsos.

fonte: próprio autor

4.1.2 DockQ

O DockQ é uma métrica que varia de 0 a 1, usado para avaliar complexos proteína-proteína, e leva em consideração uma média entre os parâmetros já avaliados acima (iRMSD, L-RMSD e Fnat). Ele considera um modelo incorreto quando o score varia de 0 a 0.23, aceitável de 0.23 a 0.49, qualidade média de 0.49 a 0.8, e alta qualidade de 0.8 até 1. (BASU; WALLNER, 2016)



C.

	Complexos reais (R)		Complexos falsos (D)	
	Média	Mediana	Média	Mediana
HPEPDOCK	0.79	0.935	0.44	0.398
HDOCK	0.70	0.817	0.37	0.374
AlphaFold Multimer	0.72	0.824	0.34	0.319

Figura 14. Comparação dos valores de DockQ entre os complexos reais e falsos. A. Boxplot dos valores de DockQ representando os complexos reais (em azul) e falsos (em vermelho). A barra dentro de cada retângulo representa a mediana dos valores de DockQ de cada conjunto. **B.** Histograma da frequência de complexos proteína-peptídeo modelados por intervalos de DockQ score. As barras em tons de azul representam os complexos reais e as barras em tons de vermelho representam os complexos falsos. **C.** Tabela comparativa entre as médias e medianas dos valores de DockQ para os complexos reais e falsos, de acordo com cada ferramenta.

fonte: próprio autor

A Figura 15A apresenta um gráfico *boxplot* comparando os valores de DockQ dos complexos reais (em azul) com os falsos (em vermelho). A mediana do conjunto dos complexos reais é 0.84 enquanto dos complexos falsos é de 0.36. Podemos observar que há uma diferença visual em ambos os gráficos, mostrando que os valores de DockQ para os complexos reais são maiores do que para os complexos falsos. Essa conclusão também é evidenciada na Figura 15B, que mostra a frequência de complexos proteína-peptídeo modelados por intervalos de DockQ score. Observamos que as barras em tons de azul (representam os complexos reais) se encontram mais concentradas à direita do gráfico, representam maiores valores de DockQ. Enquanto isso, as barras de tons de vermelho/rosa, se encontram em faixas mais variadas de valores de DockQ. Destaca-se assim, a capacidade das três ferramentas em melhor modelar/ancorar os peptídeos reais comparados aos *decoys*.

A Figura 15 C apresenta uma tabela com as médias e medianas de valor de *DockQ* para cada um dos experimentos. A média dos valores de DockQ é estatisticamente maior para os complexos reais do que para os complexos falsos, com um p-valor <0.05. Dentre todos os complexos reais e falsos, 56% dos complexos reais obtiveram valores de DockQ considerados de alta qualidade (>0.8), contra somente 3% dos complexos falsos, evidenciando que as ferramentas conseguiram discriminar os peptídeos, melhor ancorando ao sítio aqueles reais.

4.1.3 Funções de pontuação

As funções de pontuação das ferramentas utilizadas também foram avaliadas. De acordo com Allen et al., 2015, uma falha na pontuação ocorre quando uma pose correta é amostrada no experimento de *docking*, mas não é classificada como a melhor pose. Baseado nessa definição, calculamos de forma empírica, o sucesso de pontuação.

O sucesso de pontuação foi calculado sendo a razão : **Nc/N**, sendo **Nc** = número de complexos cujo modelo melhor ranqueado pela ferramenta também obteve o melhor valor de DockQ e **N** = número total de complexos. A Tabela 5 apresenta os valores de sucesso de pontuação para cada ferramenta, separando pelos conjuntos de dados dos complexos reais e falsos , além de uma taxa geral por ferramenta, abrangendo ambos os complexos.

Tabela 5. Valores de sucesso de pontuação referente a cada ferramenta.

	Sucesso de pontuação		
	Complexos reais (R)	Complexos falsos (D)	Geral
HPEPDOCK	0.52	0.04	0.28
HDOCK	0.55	0.06	0.31
AlphaFold Multimer	0.16	0.13	0.14

O HDOCK foi o programa que obteve maior taxa de sucesso em ranquear as melhores poses, com uma taxa de sucesso de pontuação de 31%. Ou seja, das 134 proteínas avaliadas (68 experimentais e 66 *decoys)*, 42 delas obtiveram o melhor valor de DockQ no primeiro modelo dos 100 gerados pela ferramenta. Já o *AlphaFold* obteve a menor taxa de sucesso entre todos os programas, de 14%. É importante ressaltar, que o *AlphaFold* é a única ferramenta dentre as três avaliadas que não possui função de pontuação própria, e ranqueia os modelos gerados somente pelo valor de plDDT, cujo objetivo é estimar a qualidade do modelo e não a afinidade

dos complexos. Esse pode ser o motivo da menor taxa de sucesso em relação ao *score*, uma vez que as outras ferramentas (HDOCK e HPEPDOCK) são específicas para *docking* e possuem funções de pontuação mais robustas. Esse resultado corrobora com AGRAWAL et al., 2019 em que as melhores poses dos ligantes não foram as melhores ranqueadas pelas ferramentas de *docking* proteína-peptideo. Isso mostra a necessidade do desenvolvimento de função de pontuação geral que funcione para vários algoritmos de *docking* molecular e que possam ranquear com melhor precisão os modelos gerados.

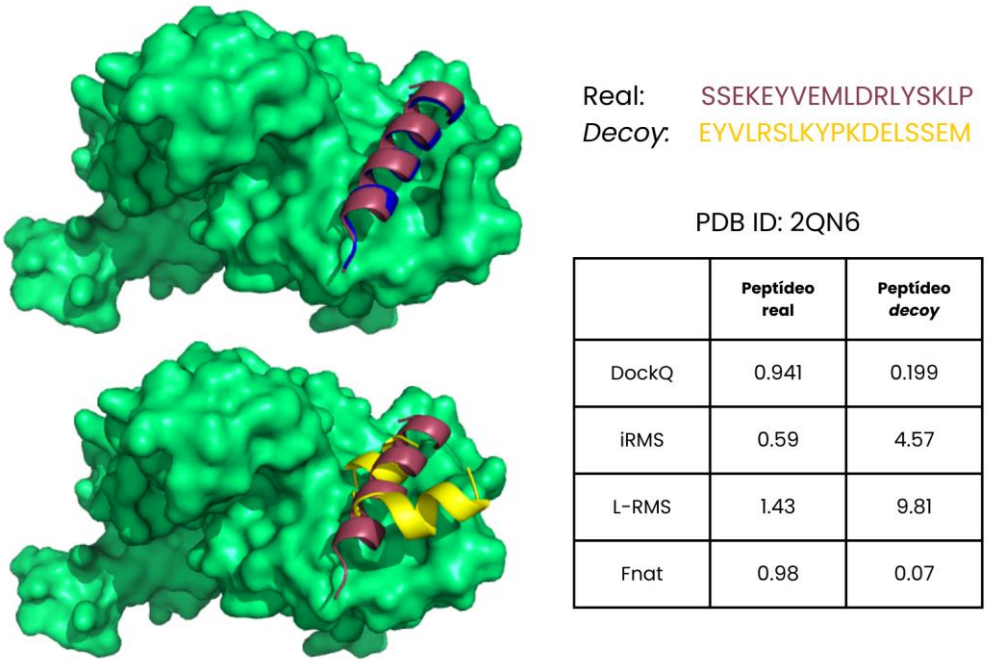
Nas próximas seções (Exemplo 1 até Exemplo 4), apresentaremos exemplos de complexos que tiveram diferentes desempenhos ao serem ancorados e modelados pelas ferramentas. Serão apresentados complexos cujos modelos reais ficaram bem modelados e os falsos não (Exemplo 1), complexos cujos ambos os modelos ficaram bem (Exemplo 2) e mal modelados (Exemplo 3), e um exemplo de um complexo que entrou em uma exceção da classificação do DockQ (Exemplo 4).

4.1.4 Exemplo 1: complexo real bem modelado e complexo falso mal modelado

O complexo proteína-peptídeo de PDB ID 2QN6 se trata da subunidade gama do fator de iniciação da tradução 2 (proteína, 414 resíduos), complexado com a subunidade beta (peptídeo, 18 resíduos).

Essa estrutura é um dos exemplos cuja modelagem feita pelo *AlphaFold Multimer* da estrutura real foi considerada de alta qualidade (DockQ = 0.941) mas a modelagem do complexo com peptídeo *decoy* obteve baixa qualidade (DockQ = 0.199). Dentre todos os complexos modelados por todas as ferramentas, a maior parte deles obteve melhor resultado (maior DockQ) para os complexos reais do que para os complexos falsos. Esse resultado era o esperado, uma vez que os peptídeos *decoys* foram gerados com o intuito de não interagirem tão bem com o sítio de ligação da proteína quanto os peptídeos originais.

A Figura 16 mostra imagem da proteína 2QN6 (em verde) ligado ao peptídeo resolvido experimentalmente do complexo (em rosa), o peptídeo real resultado da modelagem pelo *AlphaFold Multimer* (em azul) e o peptídeo *decoy* modelado também pela mesma ferramenta (em amarelo). O peptídeo real modelado se encontra em pose quase idêntica ao peptídeo experimental, enquanto o peptídeo *decoy* apresenta pose mais distante, comprovado pelas métricas de iRMS e LRMS maiores.



	Peptídeo real	Peptídeo *decoy*
DockQ	0.941	0.199
iRMS	0.59	4.57
L-RMS	1.43	9.81
Fnat	0.98	0.07

Figura 15. Exemplo de complexo real bem modelado e complexo falso mal modelado pelo *AlphaFold Multimer*. Receptor (verde) e peptídeos: i.experimental (rosa), ii.peptídeo real modelado (azul) e iii.peptídeo *decoy* modelado (amarelo). Ambos os modelos foram feitos pelo *AlphaFold Multimer*. Ao lado direito das imagens, é mostrada uma tabela apresentando os valores das métricas utilizadas para avaliação dos modelos para os complexos reais e falsos (DockQ, iRMS, L-RMS e Fnat). Os valores das métricas comprovam a melhor modelagem do complexo real comparado ao complexo falso. Acima da tabela, é apresentada a sequência do peptídeo real e do peptídeo *decoy*. Ambos possuem o mesmo tamanho e mesmos resíduos, porém permutados.

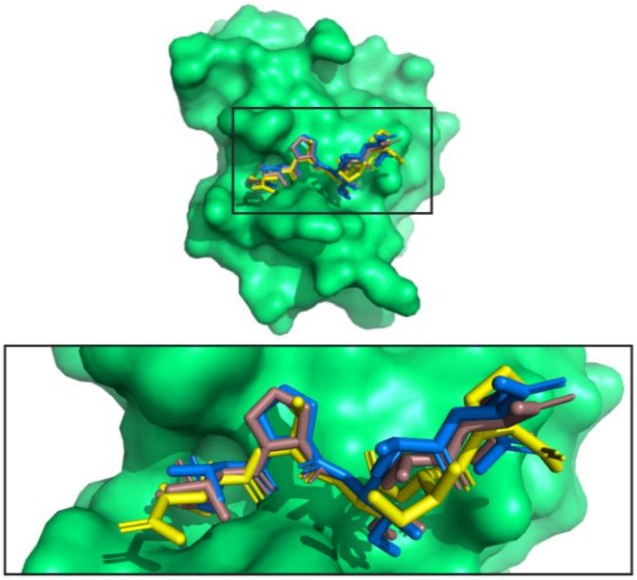
fonte: próprio autor

4.1.5 Exemplo 2: complexos real e falso bem modelados

O complexo de PDB ID 3D9U exemplifica um caso em que tanto o peptídeo real quanto o *decoy* foram bem modelados pela ferramenta HPEPDOCK. O complexo se trata do domínio BIR3 da proteína inibidora de apoptose 1 (cIAP1) complexado com o peptídeo SMAC, que funciona como regulador da atividade dessa proteína.

Conforme ilustrado na Figura 17, o peptídeo experimental (em rosa), o real resultado do *redocking* (em azul) e o peptídeo *decoy* (em amarelo) estão bem posicionados no sítio de ligação. As métricas utilizadas para avaliar os modelos também demonstram modelos de qualidade (altos DockQ e Fnat, e baixos iRMS e L-RMS). Uma hipótese para esse fato seria o tamanho do peptídeo. Observamos que os peptídeos *decoys* menores (de 5 a 7 resíduos) possuíram melhores resultados quando comparado com os maiores (acima de 7 resíduos). Nesse caso, o peptídeo possui apenas 6 resíduos, e mesmo criando um *decoy* permutando sua sequência de aminoácidos, não houve tanta diferença na sequência original, como há em casos de peptídeos maiores. Devido aos menores graus de liberdade conformacionais, peptídeos

menores apresentam menos combinações de ângulos torcionais para serem explorados pela ferramenta de *docking*, aumentando a probabilidade de encontrar uma conformação ideal que interaja de maneira semelhante ao peptídeo experimental com os resíduos da proteína.



Real: AVPIAQ
Decoy: AQVAIP

PDB ID: 3D9U

	Peptídeo real	Peptídeo *decoy*
DockQ	0.951	0.920
iRMS	0.24	0.40
L-RMS	0.69	1.15
Fnat	0.89	0.85

Figura 16. Exemplo de peptídeo real e falso bem ancorados à proteína pelo HPEPDOCK. Receptor (verde) e peptídeos: i.experimental (rosa), ii.peptídeo real ancorado (azul) e iii.peptídeo *decoy* ancorado (amarelo). Ambos os *dockings* foram feitos pelo HPEPDOCK. Ao lado direito das imagens, é mostrada uma tabela apresentando os valores das métricas utilizadas para avaliação dos modelos para os complexos reais e falsos (DockQ, iRMS, L-RMS e Fnat). Os valores das métricas comprovam uma boa ancoragem dos peptídeos reais e *decoys* pela ferramenta. Acima da tabela, é apresentada a sequência do peptídeo real e do peptídeo *decoy*. Ambos possuem o mesmo tamanho e mesmos resíduos, porém permutados.

fonte: próprio autor

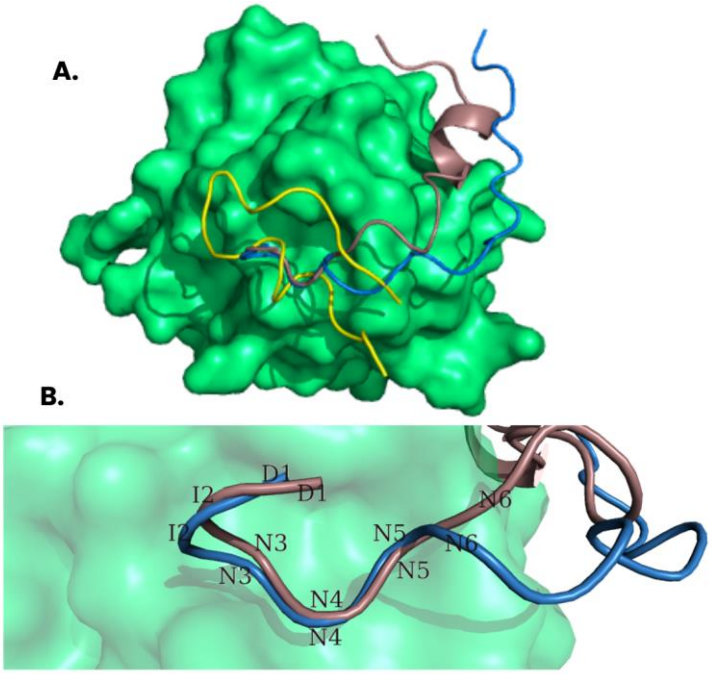
4.1.6 Exemplo 3: complexo real e falso mal modelados

O complexo 2IHS modelado pelo *AlphaFold Multimer* obteve resultados de DockQ considerados aceitável para o peptídeo real (0.490) e incorreto para o peptídeo *decoy* (0.129).

O complexo se trata do domínio B30.2/SPRY de uma proteína SOCS box (SSB) de Drosophila, complexada com um peptídeo derivado de helicase de RNA. Neste caso específico, esse domínio está relacionado ao desenvolvimento do oócito em Drosophila.

A Figura 18A apresenta o receptor (em verde), e os peptídeos: em rosa, o peptídeo experimental, em azul o peptídeo real modelado, e em amarelo, o peptídeo *decoy*. Podemos observar que o peptídeo *decoy* foi mal modelado em toda a extensão do peptídeo, ocupando somente parte do sítio de ligação real, o que explica o maior valor de iRMS e L-RMS. Por consequência, os contatos nativos se perderam, levando ao menor valor de Fnat.

Segundo WOO et al., 2006, a região N-terminal do peptídeo (DINNNN) é a principal responsável pela ligação de alta afinidade entre ele e a proteína em questão. Essa região do peptídeo real (em azul) foi bem modelada quando comparada ao experimental (em rosa), tendo todos os contatos mantidos, representados na Figura 18B. Somente a partir do resíduo N6, o peptídeo foi mal modelado, perdendo a estrutura de hélice na sua região C-terminal. Esse fato explica o valor aceitável de DockQ para o complexo real, uma vez que somente parte do peptídeo foi bem modelada.



Real: DINNNNNIVEDVERKREFYI
Decoy: RINDINEVFNNKERVDNEYI

PDB ID: 2IHS

	Peptídeo real	Peptídeo *decoy*
DockQ	0.490	0.129
iRMS	2.53	8.13
L-RMS	5.65	18.1
Fnat	0.52	0.17

Figura 17. Exemplo de complexos real e falso mal modelados pelo *AlphaFold Multimer*. A. Receptor (verde) e peptídeos: i.experimental (rosa), ii.peptídeo real modelado (azul) e iii.peptídeo *decoy* modelado (amarelo). **B.** Evidenciando os resíduos da região N terminal dos peptídeos experimental (rosa) e real modelado (azul). Essa região foi bem modelada comparada com o restante do peptídeo.Ao lado direito das imagens, é mostrada uma tabela apresentando os valores das métricas utilizadas para avaliação dos modelos para os complexos reais e falsos (DockQ, iRMS, L-RMS e Fnat). Os valores das métricas comprovam a modelagem ruim de ambos os complexos. Acima da tabela, é apresentada a sequência do peptídeo real e do peptídeo *decoy*. Ambos possuem o mesmo tamanho e mesmos resíduos, porém permutados.

fonte: próprio autor

4.1.7 Exemplo 4: Modelagem do complexo PDB ID *2WHX pelo AlphaFold Multimer*

De acordo com BASU et al., 2016, um modelo com L-RMS relativamente alto (>10Å) ainda pode ser classificado como 'aceitável' se apresentar um Fnat elevado (>0,50) e um iRMS baixo (<3,0Å). Esse foi o caso do complexo de PDB ID 2WHX. Essa é uma proteína do vírus da dengue (NS3) que possui uma parte de serino-protease (resíduos 1 a 168) que se liga ao peptídeo, NS2B, uma parte *linker* (resíduos 169 a 179) e uma parte de helicase (resíduos 180 a

618). (Figura 19) (LUO et al., 2010). Quando modelada pelo *AlphaFold Multimer*, obteve-se o L-RMS = 30,73 (alto), mas seu iRMS = 1,07 (baixo), o Fnat = 0,95 (alto), e DockQ = 0,561 (considerado modelo de qualidade média).

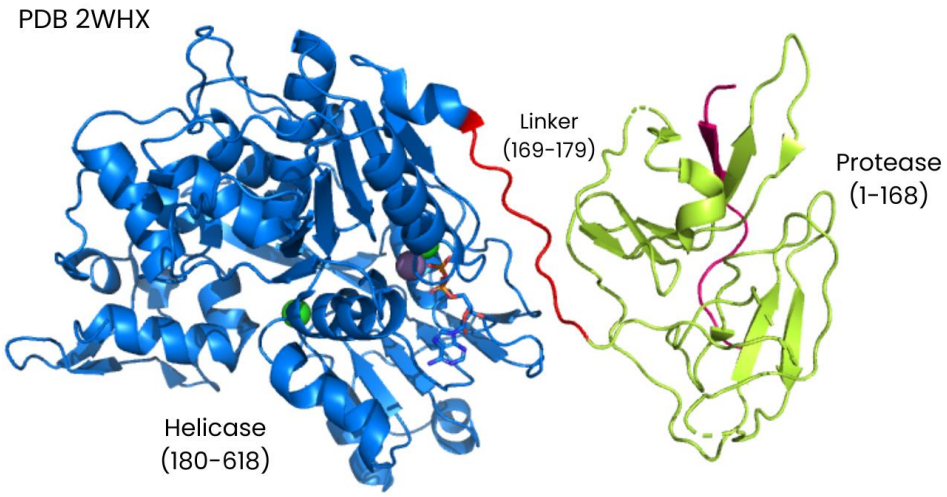


Figura 18. Representação da proteína NS3 do vírus da dengue complexada com o peptídeo NS2B. PDB ID: 2WHX. Na figura, observamos a proteína, que possui três regiões. Uma serino-protease do resíduo 1 ao 168 (em verde) onde se encontra o peptídeo NS3 ligado (em roxo), uma região linker do resíduo 169 ao 179 (em vermelho), e uma região de helicase do resíduo 180 ao 618 (em azul).

fonte: próprio autor

Segundo (LUO et al., 2010), a proteína adquire duas conformações alternativas na região da protease, que ocorrem sob condições de cristalização semelhantes, devido à flexibilidade da região *linker*. Como as sequências de ambas as proteínas são idênticas, o *AlphaFold Multimer* modelou o complexo na **conformação I** (representada no PDB pelo ID 2VBC), (figura 20A) enquanto a estrutura utilizada neste trabalho, presente na base de dados PepPro (PDB ID 2WHX), apresenta a **conformação II** (Figura 20C). Por esse motivo, o valor do L-RMS ficou elevado, uma vez que o peptídeo foi modelado distante da posição original. No entanto, o iRMS permaneceu baixo, já que a interface proteína-peptídeo se manteve quase inalterada, assim como os contatos.

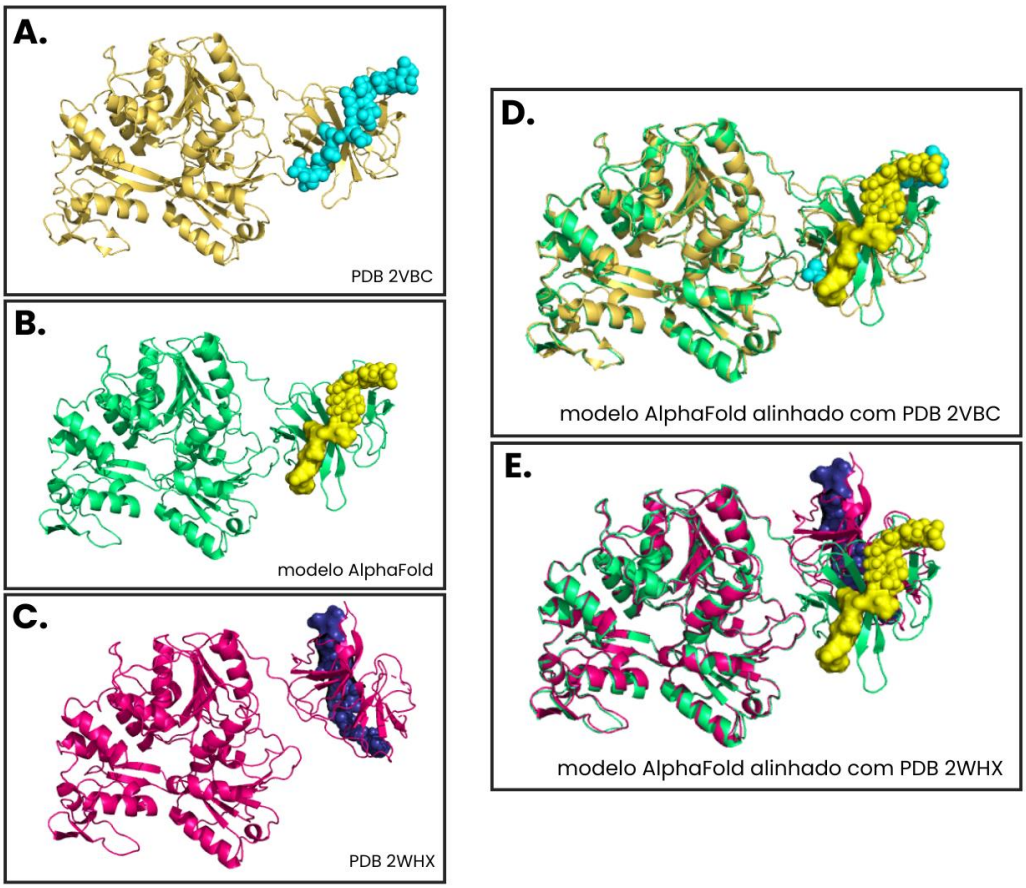


Figura 19. Comparação entre estruturas 2WHX, 2VBC e modelo predito pelo *AlphaFold Multimer*. As figuras A e C representam os complexos 2VBC e 2WHX, respectivamente. Ambos os complexos são referentes à proteína NS3, do vírus da dengue, complexado com o peptídeo NS2B, portanto suas sequências são idênticas. A diferença entre eles está na conformação da região da protease (onde o peptídeo interage), onde a figura A mostra a conformação I e a figura C, a conformação II. A figura B apresenta o modelo criado pelo *Alphafold Multimer* utilizando a sequência da proteína e do peptídeo. As figuras D e E representam alinhamentos entre a estrutura modelada pelo *AlphaFold Multimer* e as estruturas experimentais 2VBC (figura D) e 2WHX (figura E), mostrando que o modelo é mais parecido com a conformação I (PDB ID 2VBC) do que com a II (PDB ID 2WHX).

fonte: próprio autor

O objetivo proposto nesta seção foi comparar três ferramentas de *docking* e modelagem molecular na tarefa de ancorar/modelar complexos proteína-peptídeos reais e falsos.

As três ferramentas utilizadas neste estudo cumpriram com a função de ancorar/modelar bem complexos proteína-peptídeo reais e não tão bem complexos falsos, como era o esperado. A ferramenta HDOCK, embora se destaque na tarefa de classificar as melhores poses e demonstre eficácia na distinção entre ligantes reais e falsos, requer um uso cauteloso no *docking* proteína-peptídeo. Especialmente em peptídeos menores, há o risco da modelagem do peptídeo cortar as suas extremidades, resultando em uma redução do tamanho real do peptídeo.

A diversidade da base de dados utilizada, PepPro, foi importante pois assim tivemos exemplos de complexos bem e mal modelados pelas três ferramentas. Não houve, em nenhum complexo modelado por nenhuma ferramenta, caso em que o complexo falso fosse melhor modelado do que o complexo real, indicando o sucesso tanto na modelagem pelas ferramentas quanto na geração dos peptídeos *decoys*. Do mesmo modo, não houve nenhum exemplo de complexo que tenha sido mal modelado nas três ferramentas, mas houve complexos bem modelados nas três, ou em pelo menos, duas delas. Isso enfatiza a conhecida importância de utilizar várias ferramentas com abordagens diferentes para a realização de *docking* proteína-peptídeo.

4.2 Parte II - Comparação de modelagem de peptídeos nas formas *holo* e *apo*

Na segunda parte deste trabalho comparamos a modelagem de peptídeos nas formas *holo* e *apo*, utilizando o *AlphaFold 2*. A forma *holo* refere-se aos peptídeos modelados juntamente com os receptores, enquanto a forma *apo* se refere aos peptídeos modelados individualmente. Já se sabe que a tarefa de modelar e resolver estruturas de peptídeos na forma *apo* é desafiadora, pois os peptídeos tendem a ter estrutura quando ligados aos receptores devido à redução de sua liberdade conformacional.

A Figura 21 apresenta um histograma onde podemos comparar o RMSD normalizado (RMSD/número de resíduos do peptídeo) das estruturas modeladas nas forma *apo* e *holo*. A média do valor dos RMSD normalizado é de 0.214 e a mediana é de 0.142. Dessa forma, percebemos que grande parte dos peptídeos obtiveram modelagem pelo *AlphaFold2* parecida em ambas as formas, apresentando pequenas diferenças estruturais. Alguns exemplos de estruturas mais específicas serão apresentados a seguir.

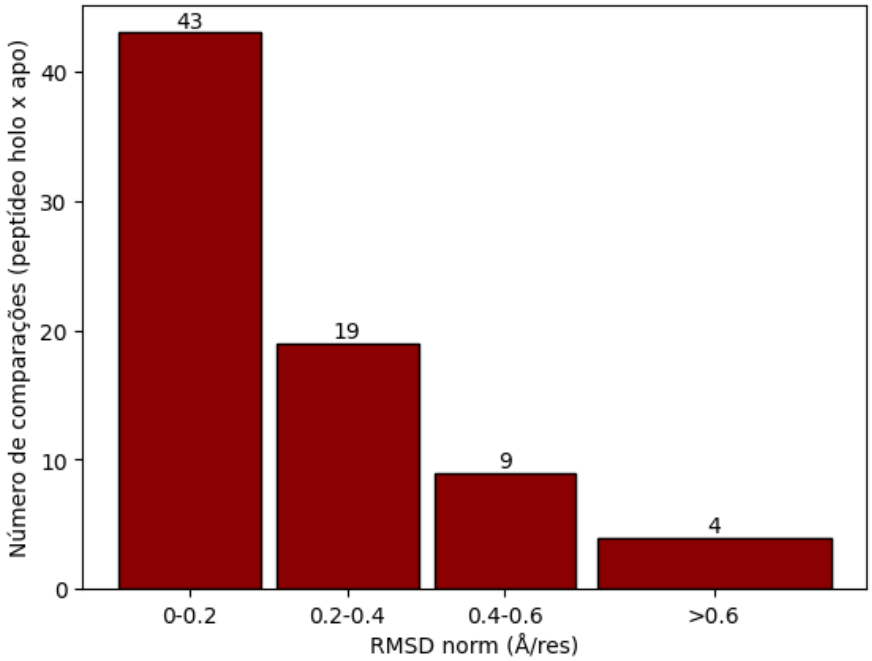


Figura 20. Número de comparações de peptídeos holo e apo por intervalo de RMSD normalizado. Comparando as estruturas dos peptídeos modelados de forma *holo* e *apo*. Observamos que a maior parte dos dados se encontra à esquerda do gráfico, indicando menores valores de RMSD entre peptídeos *holo* e *apo*.

fonte: próprio autor

Dois peptídeos serão utilizados como exemplos de estruturas com **(i)** modelagens parecidas nas formas *holo* e *apo*, e **(ii)** modelagens distintas nas formas *holo* e *apo*.

Como exemplo **(i),** temos o complexo 2NM1, em que o RMSD normalizado entre a estrutura *holo* e *apo* do peptídeo foi de 0.04, mostrando que ambas as estruturas tiveram modelagem semelhante pelo *AlphaFold 2*. O complexo de PDB ID 2NM1 se trata do domínio de ligação da neurotoxina botulínica sorotipo B (proteína, cadeia A), complexada com o domínio luminal do seu receptor, Sinaptotagmina-2 (peptídeo, cadeia B). As sinaptotagminas constituem uma família de receptores relacionados à liberação de neurotransmissores dependentes de cálcio, e portanto, são importantes na transmissão da sinapse em junções neuromusculares (JIN et al., 2006). A Figura 22A ilustra a proteína (em rosa) e o peptídeo (em azul) formando um complexo. Na Figura 22B, observamos o peptídeo modelado da forma holo (em azul) e apo (em laranja). A cadeia principal foi praticamente mantida em ambas as modelagens, com algumas mudanças somente em algumas posições das cadeias laterais, corroborando com o valor baixo de RMSD normalizado entre as duas estruturas (0.04).

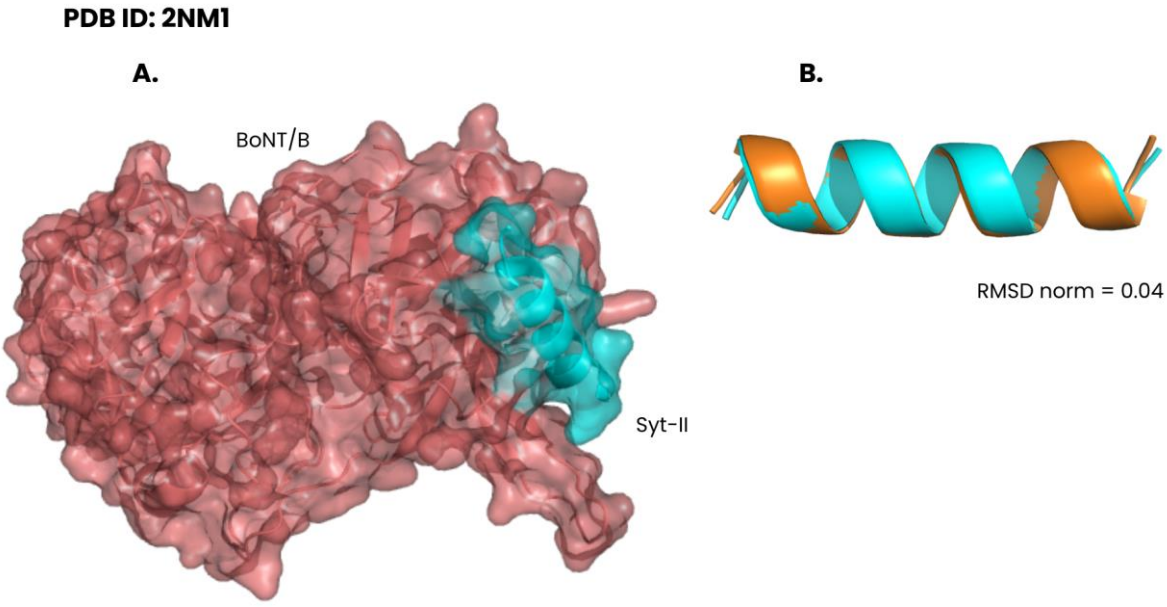


Figura 21. Comparação entre o peptídeo do complexo PDB ID 2NM1 nas formas *holo* e *apo*. A. Complexo PDB ID 2NM1 ilustrando, em rosa, o domínio de ligação ao receptor da neurotoxina botulínica sorotipo B (BoNT/B) e, em azul, o domínio luminal da proteína de membrana Sinaptotagmina 2 (Syt-II). **B.** Alinhamento entre o modelo *holo* (em azul) e *apo* (em laranja) do domínio da proteína Syt-II.

fonte: próprio autor

Por outro lado, temos como exemplo de estrutura com **(ii)** modelagens distintas nas formas *holo* e *apo*, o peptídeo pertencente ao complexo de PDB ID 4XOE. Esse complexo se trata de uma proteína receptora (FimH) e de um peptídeo (DsG) relacionados com a adesão de bactérias patogênicas *Escherichia coli* ao epitélio (SAUER et al., 2016). Na estrutura resolvida experimentalmente, o DsG (peptídeo) possui estrutura secundária de folha beta e interage com duas folhas beta da proteína, se posicionando entre elas. A figura 23A ilustra cada resíduo do peptídeo e suas ligações de hidrogênio com os resíduos da proteína, que constituem as duas folhas betas (em amarelo) no entorno do peptídeo. Na figura 23C, observamos o peptídeo modelado nas formas *holo* (em vermelho) e *apo* (em verde). Quando modelado da forma *apo*, o peptídeo se dobra sobre si mesmo, gerando duas folhas betas que passam a interagir entre si, uma vez que não há outra proteína no entorno para interagir com ele. Dessa forma, o RMSD normalizado das duas estruturas é de 0.760, mostrando que houve diferença nas estruturas modeladas de ambas as formas.

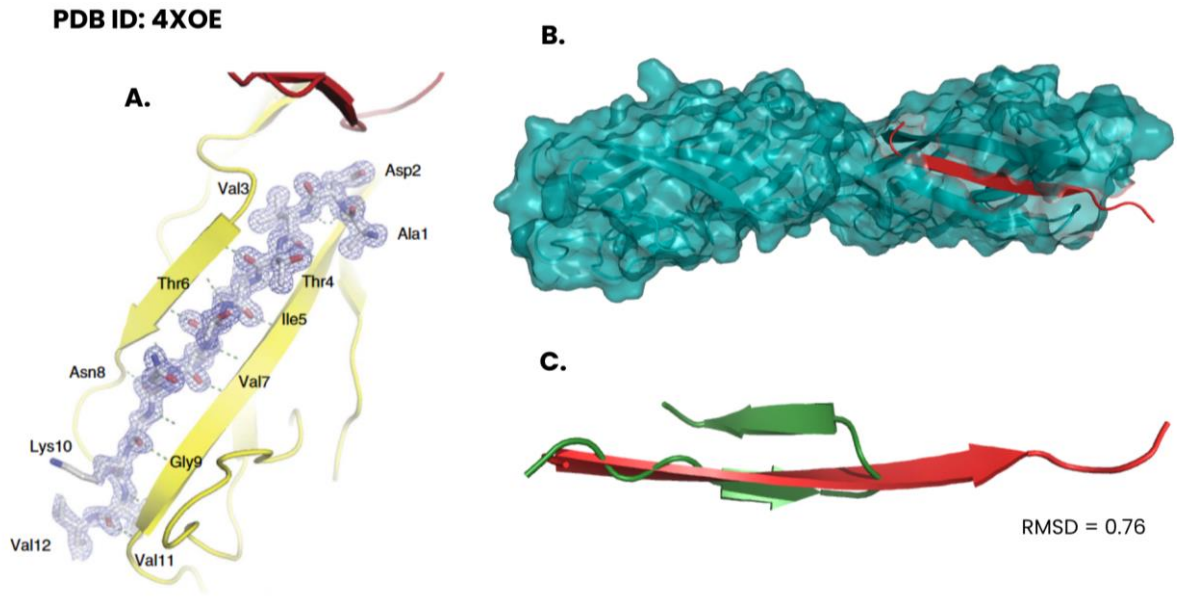


Figura 22. Comparação entre o peptídeo do complexo PDB ID 4XOE nas formas *holo* e *apo*. A.Peptídeo DsG, representado em azul, e duas folhas beta da proteína FimH, em amarelo. Os traços representam as ligações de hidrogênio entre os resíduos do peptídeo e da proteína. **B.** Representação da proteína FimH, em azul e peptídeo DsG, em vermelho. **C.** Alinhamento entre o modelo *holo* (em vermelho) e *apo* (em verde) do peptídeo DsG.

fonte: Adaptado de Sauer et al., 2016.

O objetivo desta seção foi comparar a modelagem de peptídeos, utilizando o *AlphaFold 2*, nas formas *holo* e *apo*. Apesar de ser uma ferramenta desenvolvida para a modelagem de proteínas e de complexos, desejávamos testá-la também para a modelagem de peptídeos com até 30 resíduos. O *AlphaFold 2* modelou os peptídeos de formas parecidas, uma vez que 43 dos 75 peptídeos testados apresentaram um RMSD normalizado inferior a 0.2. Esse resultado está em consonância com MCDONALD et al., 2023 que comparou o *AlphaFold 2* para modelagem de peptídeos *apo* com outras como RoseTTAFold e PEPFOLD-3, evidenciando que o *AlphaFold 2* obteve um desempenho superior. Nesse mesmo trabalho, os autores compararam os modelos do *AlphaFold 2* com estruturas de peptídeos resolvidas experimentalmente por ressonância magnética nuclear (NMR) e concluíram que a métrica pLDDT não conseguiu ranquear os modelos com tanta precisão. No presente trabalho, utilizamos exclusivamente o modelo mais bem ranqueado pelo pLDDT para realizar as comparações. É possível que a testagem de outros modelos, especialmente para os peptídeos que apresentaram menor precisão na modelagem, resulte em melhorias nos resultados.

4.3 Parte III -Modelos de predição de complexos reais e falsos

Nesta seção do trabalho buscamos desenvolver modelos de aprendizado de máquina para predição de complexos proteína-peptídeo reais e falsos.

4.3.1 Teste de parâmetro: Interfaces

Para avaliar o melhor valor a ser usado como *cutoff* para calcular as interfaces dos complexos, foram construídos modelos com assinaturas das interfaces com distâncias de 4, 5, 6, 6.5, 7 e 8 Å. Com base na Figura 24, é possível perceber que o melhor valor de cutoff para a interface foi o valor de 6Å, com um F1-score de 0.707, corroborando com os achados de (PIRES et al., 2011). Esse valor foi utilizado para todos os experimentos subsequentes.

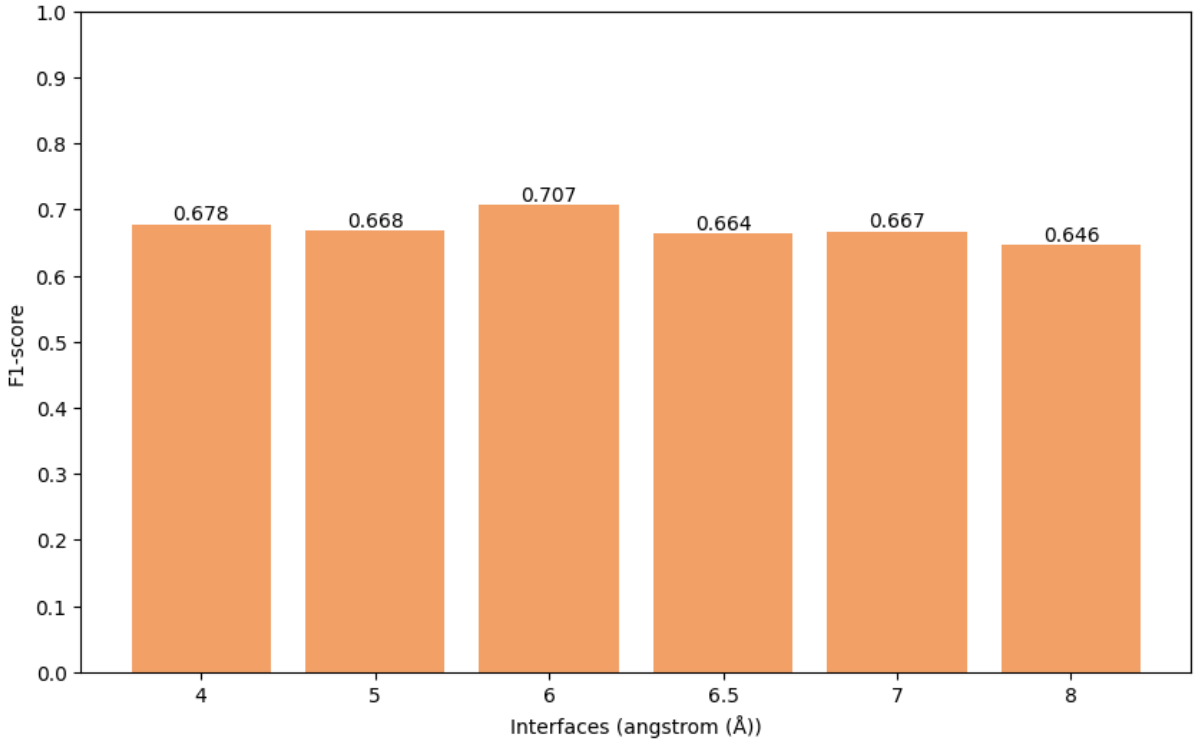


Figura 23. Valores de F1-score dos modelos de aprendizado de máquina para cada valor testado de interface.

fonte: próprio autor

4.3.2 Teste de parâmetros: Assinaturas estruturais

Comparou-se as assinaturas *CSM* e *aCSM_all* em relação ao poder de predição de complexos proteína-peptídeo reais e falsos. A Tabela 6 apresenta os melhores modelos gerados para cada teste. Dentre os testes realizados utilizando a assinatura *CSM*, o que obteve melhor F1-score foi o Teste 1, utilizando os parâmetros *default* da assinatura: *limite de corte* = 30 e *passo de corte* = 0.2.

Os melhores modelos foram obtidos utilizando a assinatura *aCSM_all,* sendo o Teste 5 o que obteve maior valor de F1-score = 0.741, usando os parâmetros *limite de corte* = 20 e *passo de corte* = 0.2. A assinatura *aCSM_all*, diferentemente da CSM, leva em consideração a natureza dos átomos que estão em contato e consequentemente, o tipo de interação que está sendo feita entre eles. Essa característica faz com que a assinatura seja mais robusta e mais completa, obtendo melhores resultados em tarefas de predições relacionadas a estruturas de proteínas (PIRES et al., 2011, 2013; RODRIGUES et al., 2022).

Tabela 6. Resultados dos experimentos com os cálculos de assinaturas CSM e *aCSM_all.*

	Parâmetros			Resultados
Teste	Assinatura	*limite de corte*	*passo de corte*	F1-score
1	CSM	30	0.2	0.710
2	CSM	20	0.2	0.698
3	CSM	10	0.1	0.687
4	aCSM_all	30	0.2	0.711
5	aCSM_all	20	0.2	0.741
6	aCSM_all	10	0.1	0.735

4.3.3 Modelos de aprendizado de máquina

Ainda com o objetivo de predizer quando um complexo proteína-peptídeo é real ou não, testou-se diferentes paradigmas de aprendizado de máquina usando as assinaturas estruturais dos complexos. A Figura 25 apresenta os resultados de F1-score para os melhores modelos gerados com cada paradigma de aprendizado de máquina, incluindo KNN, Redes Neurais, *Gradient Boosting*, Regressão Linear, SVM, *Random Forest*, *Naive Bayes* e Árvore de Decisão. Os melhores valores de *F1-score* (0,741 e 0.735) foram alcançados com o uso de Redes Neurais e Regressão Logística, respectivamente. Em contrapartida, o pior desempenho foi obtido pelo KNN, com *F1-score* igual a 0,587. Esses resultados contrastam com os de estudos anteriores, como os de (PIRES et al., 2011, 2013), nos quais o *KNN e Gradient Boosting* se destacaram como os melhores algoritmos de aprendizado de máquina para a predição de função e classificação estrutural de proteínas utilizando assinaturas estruturais.

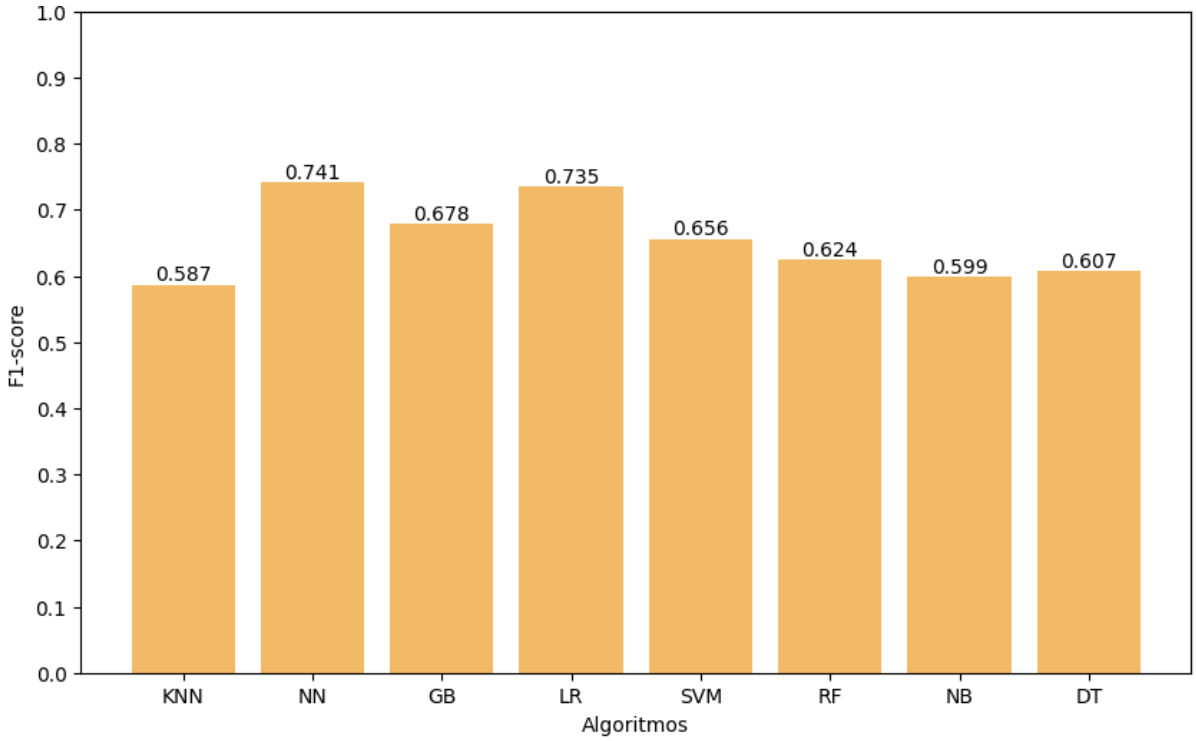


Figura 24. Valores de F1-score para cada modelo gerado a partir de algoritmos diferentes de aprendizado de máquina. KNN: K-Nearest Neighbors, NN: Neural Networks, GB: Gradient Boosting, LR: Linear Regression, SVM: Support Vector Machine, RF: Random Forest, NB: Naive Bayes, DT: Decision Tree.

fonte: próprio autor

4.3.4 Avaliação do classificador

Dentre todos os modelos gerados para classificação de complexos reais e falsos, o que obteve maior sucesso foram as Redes Neurais, com F1-score igual a 0,741. A Figura 26 apresenta o gráfico da curva ROC e matriz de confusão relativas a ele, e a Tabela 7 as métricas de avaliação do modelo. Dentre os 441 complexos proteína-peptídeo, o modelo classificou corretamente 327, sendo deles 172 complexos reais (representados pela letra R) e 155 complexos falsos (representados pela letra D). Além disso, o modelo classificou incorretamente 53 complexos reais e 61 falsos.

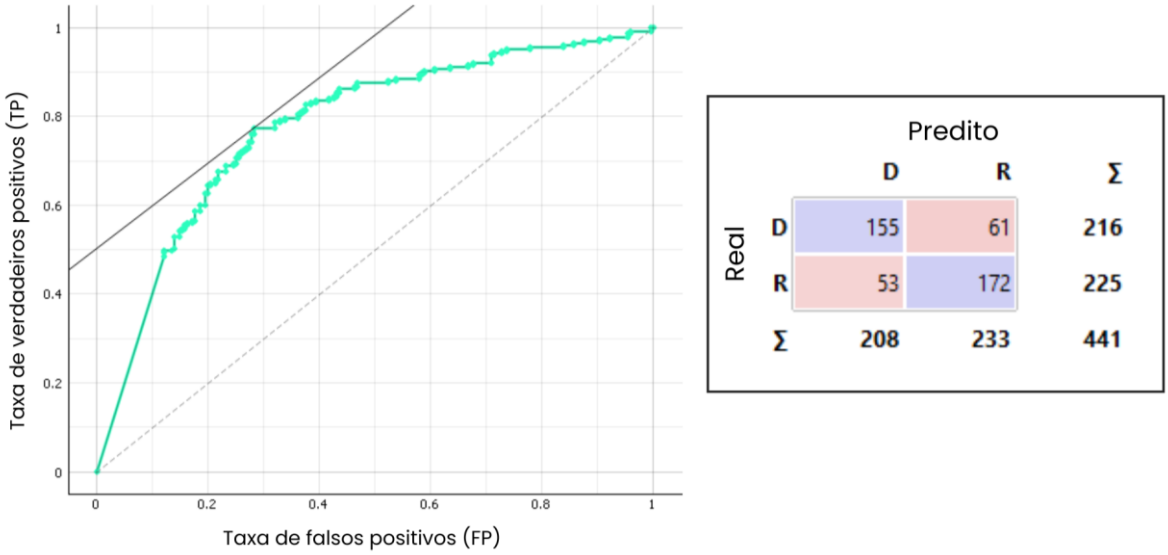


Figura 25. Curva ROC e matriz de confusão referentes ao melhor modelo de classificação de complexos reais (R) e falsos (D).

fonte: próprio autor

Tabela 7. Métricas de avaliação do modelo de classificação de complexos reais e falsos..

AUC	CA	F1-score	Precisão	Revocação	Especificidade	MCC
0,779	0,741	0,741	0,742	0,741	0,741	0,483

A Tabela S3 apresenta as classificações equivocadas de cada complexo feitas pelo modelo baseado em Redes Neurais. Acreditávamos que existiria uma relação entre o valor de DockQ e sua classe predita. Em outras palavras, esperávamos que complexos com valores de DockQ mais altos (entre 0,6 e 1) fossem classificados pelo modelo de aprendizado supervisionado como complexos reais, e complexos com valores de DockQ mais baixos (abaixo de 0.6) fossem classificados como falsos. O cálculo do aCSM leva em consideração (i). a distância entre os átomos da proteína e do peptídeo e (ii). os contatos que estão sendo realizados entre eles. A distância (i) se relaciona com as métricas de iRMS e L-RMS e os contatos (ii), com a métrica Fnat, ambos avaliados no cálculo de DockQ. Entretanto, não foi esse o observado: não houve uma relação entre a qualidade da modelagem do complexo (valor de DockQ) e a classificação como complexo real ou falso. Isso se comprova pelo valor das médias dos DockQ entre os complexos preditos como reais (média = 0.672) e falsos (média = 0.444) quando comparado com as médias de DockQ dos complexos com seus rótulos verdadeiros: reais (média = 0.738) e falsos (média = 0.388).

O complexo com o PDB ID 1U00 teve o maior número de classificações equivocadas. Entre os 6 modelos (3 complexos reais e 3 falsos), 4 foram classificados incorretamente. O complexo em questão se trata do domínio de ligação da proteína chaperona HscA interagindo com o peptídeo ELPPVKIHC (CUPP-VICKERY et al., 2004). Embora tenha sido classificado incorretamente, observa-se que o peptídeo foi bem ancorado à proteína pela ferramenta HPEPDOCK (Figura 27A). A interface proteína-peptídeo foi praticamente preservada, com um valor de i-RMS de 0,33 e DockQ de 0,966 em comparação ao peptídeo cristalizado. Por outro lado, o melhor modelo do peptídeo real gerado pelo *AlphaFold Multimer* (Figura 27C) alcançou um valor de DockQ de 0,608, sendo considerado um modelo de qualidade aceitável. Já em relação aos *decoys*, o modelo também os classificou incorretamente. Eles obtiveram valores de DockQ = 0,664 (Figura 27B) e 0,583 (Figura 27D).

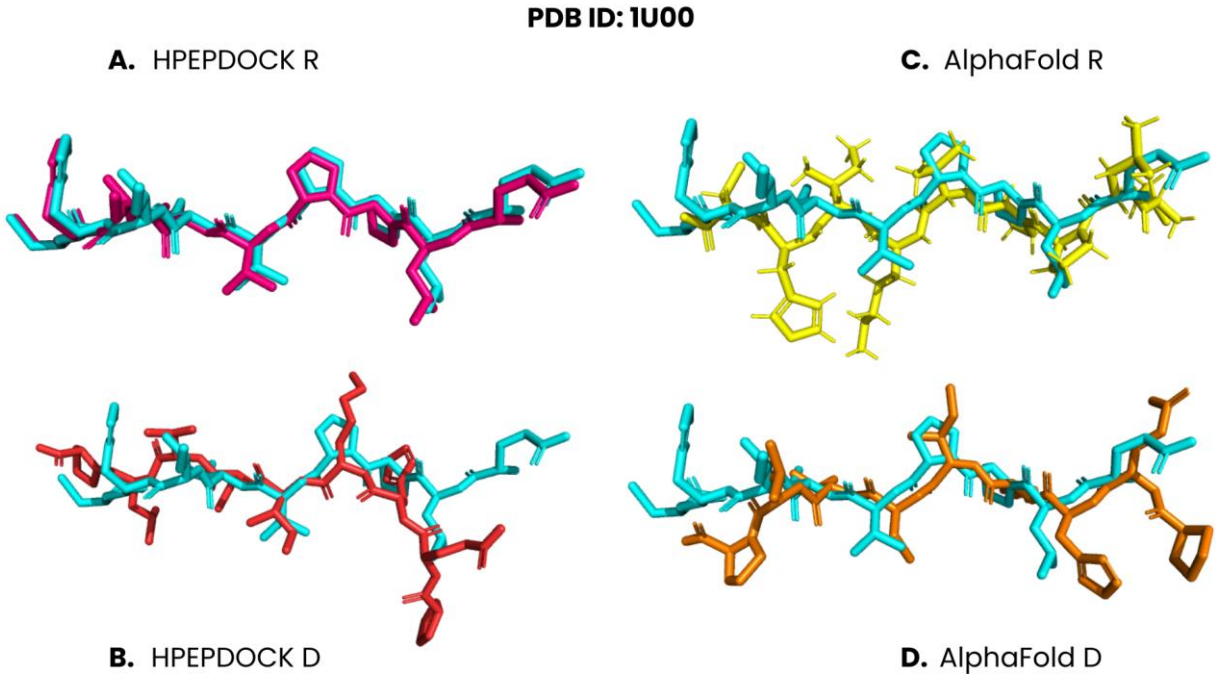


Figura 26. Complexos PDB 1U00 classificados incorretamente pelo modelo de aprendizado de máquina. A. Redocking usando a ferramenta HPEPDOCK. Em azul, o peptídeo cristalizado, e em rosa, o modelo. **B.** Redocking do peptídeo *decoy* usando a ferramenta HPEPDOCK. Em azul, o peptídeo cristalizado, e em vermelho, o modelo. **C.** Modelagem do peptídeo real usando a ferramenta *AlphaFold Multimer*. Em azul, o peptídeo cristalizado, e em amarelo, o modelo. **D.** Modelagem do peptídeo *decoy* usando a ferramenta *AlphaFold Multimer*. Em azul, o peptídeo cristalizado, e em laranja, o modelo.

fonte: próprio autor

O melhor modelo gerado para predição de complexos proteína-peptídeo reais e falsos foi então baseado em Redes Neurais, com parâmetros *Neurons in hidden layers:* 500, *activation*: "ReLu", *solver*: "L-Bfgs-B", *regularization* α = 0, *maximal number of iterations:* 100 e *replicable training.* Esse modelo foi treinado e testado utilizando as assinaturas do tipo *aCSM_all*, com parâmetros *limite de corte* = 20 e passo de corte = 0.2 e considerando os átomos da interface como os átomos do receptor que tivessem a até no máximo 6Å de distância dos átomos do peptídeo.

Em ZHAO et al., 2011, os pesquisadores utilizam informações estatísticas das interfaces de complexos proteína-proteína para classificar complexos reais e falsos. O modelo de classificação supervisionado utilizando o algoritmo SVM obteve acurácia de 0.78. Outro trabalho relacionado (JANDOVA et al., 2021) busca distinguir *decoys* de estruturas nativas proteína-proteína com base em sua estabilidade durante simulação de dinâmica molecular, atingindo acurácia de 0.85 usando o algoritmo *Random Forest*. Apesar de obterem resultados ligeiramente melhores, não foram testados para predição de complexos proteína-peptídeo.

A predição de complexos nativos proteína-proteína e, principalmente proteína-peptídeo, continua sendo um desafio significativo na bioinformática estrutural. O desenvolvimento de modelos capazes de realizar essa tarefa mostra-se promissor, especialmente como uma ferramenta complementar para aprimorar as funções de pontuação de ferramentas de *docking*.

4.4 Parte IV -Modelos de predição de interação proteína-peptídeo

Apresentaremos nesta parte, os resultados referentes aos modelos de predição de interação proteína-peptídeo, considerando proteína e peptídeo separados, ou seja, sem a realização do *docking*. Para atingir esse objetivo, concatenamos as assinaturas das interfaces das proteínas com a assinatura dos peptídeos reais e *decoys*, na forma *apo,* modelados pelo *AlphaFold Multimer*. Então, utilizamos essas assinaturas para o treinamento de modelos de aprendizado de máquina para classificação dos peptídeos como ligantes e não-ligantes.

4.4.1 Teste de parâmetros: assinaturas estruturais

Testamos os parâmetros da assinatura *aCSM_all* para a predição de interação proteína-peptídeo. Segundo a Tabela 8, os melhores parâmetros da assinatura para as interfaces da proteína e peptídeos são *limite de corte* = 20 e *passo de corte* = 0.2. Esses valores foram utilizados como parâmetros para os experimentos subsequentes.

Tabela 8. Resultados dos experimentos com os parâmetros das assinaturas *aCSM_all.*

	Proteína			Peptídeo			Resultados
Teste	Assinatura	*limite de corte*	*passo de corte*	Assinatura	*limite de corte*	*passo de corte*	F1-score
1	aCSM	30	0.2	CSM	30	0.2	0.714
2	aCSM	20	0.2	CSM	20	0.2	0.719
3	aCSM	10	0.1	CSM	10	0.1	0.655

4.4.2 Modelos de aprendizado de máquina

Desenvolvemos modelos supervisionados para classificação de peptídeos ligantes e não-ligantes. Os algoritmos testados foram KNN, Redes Neurais, *Gradient Boosting*, Regressão Linear, SVM, *Random Forest*, *Naive Bayes* e Árvore de Decisão. O melhor desempenho foi relativo ao algoritmo *Gradient Boosting*, obtendo *F1-score* igual a 0,930 e o pior desempenho, foi relativo ao KNN, obtendo *F1-score* igual a 0,526, segundo a Figura 28.

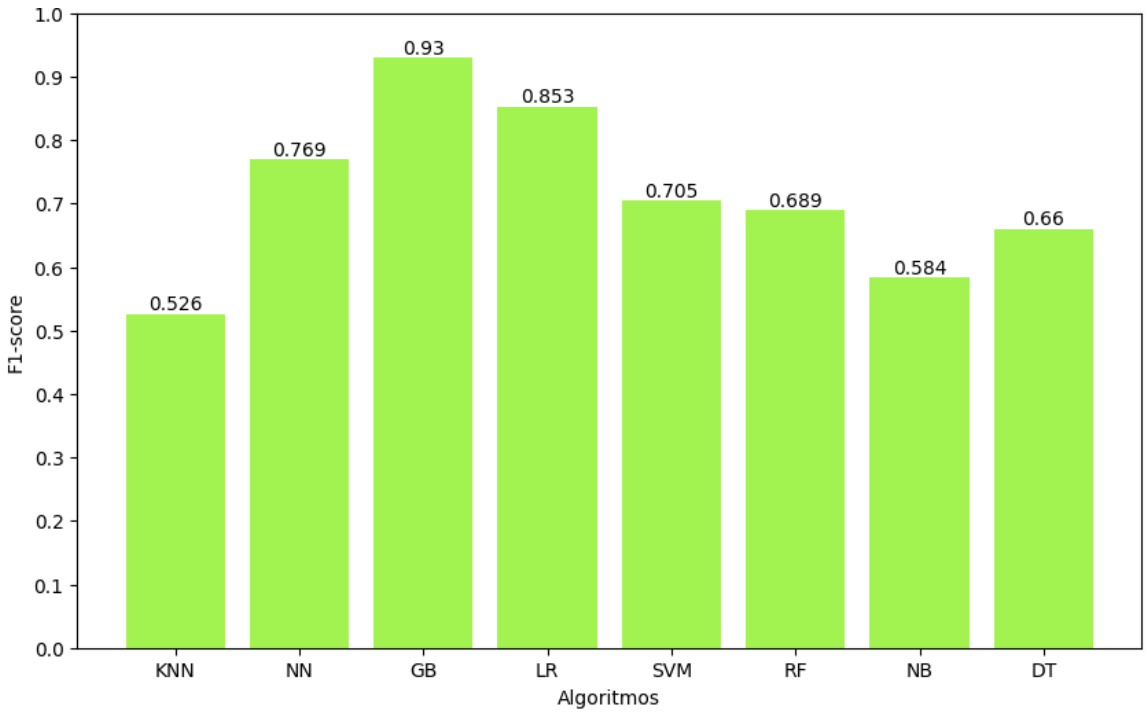


Figura 27. Valores de F1-score para cada modelo gerado a partir de algoritmos diferentes de aprendizado de máquina para predição de ligação proteína-peptídeo. KNN: K-Nearest Neighbors, NN: Neural Networks, GB: Gradient Boosting, LR: Linear Regression, SVM: Support Vector Machine, RF: Random Forest, NB: Naive Bayes, DT: Decision Tree.

fonte: próprio autor

4.4.3 Avaliação do classificador

A Figura 29 apresenta o gráfico da curva ROC para as predições de ligantes (R) e a matriz de confusão relativas ao melhor modelo desenvolvido (usando o *Gradient Boosting*). A Tabela 9 apresenta as métricas de avaliação desse modelo. Dentre os 441 complexos proteína-peptídeo, o modelo classificou corretamente 410, sendo deles 212 complexos com peptídeos ligantes (representados pela letra R) e 198 complexos com peptídeos não-ligantes (representados pela letra D). O modelo classificou incorretamente somente 13 peptídeos ligantes e 18 peptídeos *decoys*. Na Tabela S4 estão presentes os erros de classificação do modelo.

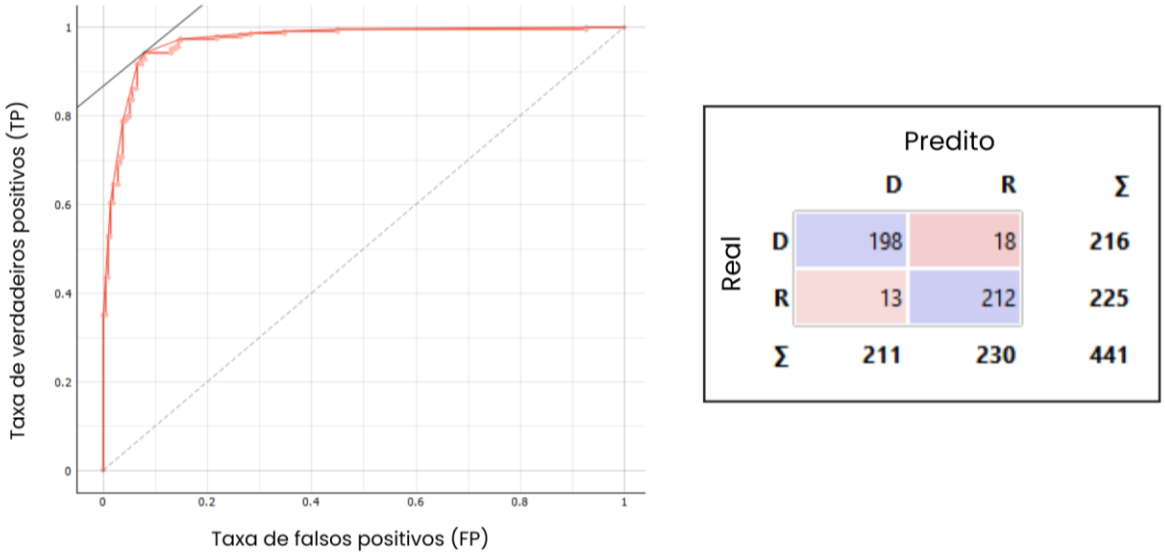


Figura 28. Curva ROC e matriz de confusão referentes ao melhor modelo de classificação de peptídeos ligantes e não ligantes, usando o algoritmo Gradient Boosting.

fonte: próprio autor

Tabela 9. Métricas de avaliação do modelo de classificação de peptídeos ligantes e não ligantes.

AUC	CA	F1-score	Precisão	Revocação	Especificidade	MCC
0,967	0,930	0,930	0,930	0,930	0,929	0,860

Com o objetivo de facilitar a triagem virtual de peptídeos, desenvolvemos um classificador que pudesse predizer se um peptídeo teria capacidade de se ligar ou não no sítio de uma proteína. O melhor modelo, obtendo AUC = 0,967, foi baseado no algoritmos *Gradient*

Boosting, com parâmetros *number of trees: 200, learning rate "0.100", replicable training, growth control for limit depth of individual trees: 3, "do not split subsets smaller than 2*" e *fraction of training instances*: 1.00. Esse modelo foi treinado e testado utilizando as assinaturas do tipo *aCSM_all*, com parâmetros limite de corte = 20 e passo de corte = 0.2 e considerando os átomos da interface como os átomos do receptor que tivessem a até no máximo 6Å de distância dos átomos do peptídeo.

Apesar de os peptídeos muitas vezes adquirirem suas estruturas somente ao interagir com proteínas e outros ligantes, utilizamos suas assinaturas na forma *apo*, pois nossa proposta inicial era predizer a ligação sem a necessidade de realizar *docking*. Acreditamos que a assinatura estrutural dos peptídeos, mesmo que não na sua conformação correta quando ligado, pode trazer informações sobre as prováveis interações entre seus resíduos, baseado nos tipos de resíduo e na distância na qual se encontram.

Yin e colaboradores (YIN; MI; SHUKLA, 2024), realizaram uma revisão da literatura citando modelos de aprendizado de máquina existentes para a predição de interação proteína-peptídeo. A grande maioria, que obteve as melhores métricas (AUC, MCC, F1-score, entre outras) foi baseada em aprendizado profundo. O *InterPepRank*, por exemplo, é baseado em Redes Convolucionais em Grafos (GCN), e obteve AUC = 0.86 na predição de sítios de ligação de peptídeos, um problema diferente do que estamos abordando neste trabalho.

O trabalho de Lei e colaboradores (LEI et al., 2021) também buscou desenvolver um método utilizando aprendizado de máquina que facilitasse a triagem virtual de peptídeos, além de predizer resíduos importantes para a ligação proteína-peptídeo, diminuindo o tempo de processamento. Apesar de alcançarem acurácias acima de 0.8, os atributos utilizados para treinamento do modelo são baseados nas sequências das proteínas e peptídeos.

Baranwal e colaboradores criaram o *Struct2Graph* (BARANWAL et al., 2022), uma arquitetura de redes neurais que utilizam assinaturas estruturais baseadas em grafos para predizer interações proteína-proteína. Nessa estratégia, usam GNCs para obter os atributos geométricos relevantes do par das proteínas usadas como entrada, concatenam esses atributos e então, alimentam uma Rede Neural *feedforward* (FNN) que possui como saída a probabilidade de aquele complexo pertencer a duas classes: ligante ou não ligante. O melhor modelo gerado atingiu AUC = 0,991 e F1 = 0,975. Assim como o *Struct2Graph*, no presente trabalho utilizamos as assinaturas estruturais de proteínas. Esses métodos tendem a gerar modelos de aprendizado de máquina mais robustos do que aqueles que utilizam somente informações sobre as sequências, uma vez que elas são menos informativas do que as estruturas. O *Struct2Graph* se compara diretamente com o modelo proposto nesta dissertação (parte IV), embora utilize

abordagem para predizer interações proteína-proteína. Apesar de os modelos gerados pelo *Struct2Graph* obterem métricas ligeiramente melhores (*Struct2Graph* com AUC = 0,991 contra AUC = 0,967 do nosso modelo), eles utilizam dados estruturais de 117 mil complexos proteína-proteína provenientes de duas bases de dados: IntAct (ORCHARD et al., 2014) e STRING (SZKLARCZYK et al., 2019), enquanto no presente trabalho obtivemos resultado semelhante utilizando base de dados menor. Como perspectiva, pretendemos treinar e testar nosso modelo usando assinaturas dos complexos presentes em outras bases de dados, como a Propedia, bem como utilizar modelos de redes neurais profundas, para buscarmos alcançar maiores acurácias.

5. Conclusões

Neste trabalho, investigamos o papel das assinaturas estruturais na interação entre proteínas e peptídeos utilizando aprendizado de máquina. Dividimos o trabalho em quatro partes, cada uma responsável por atingir um objetivo específico.

Na primeira parte, avaliamos três ferramentas de *docking* e modelagem molecular para predição de interação proteína-peptídeo, utilizando o DockQ como métrica para avaliar os modelos gerados. Os resultados mostraram que, de maneira geral, os complexos foram bem modelados, de acordo com os valores obtidos de DockQ. Planejamos, como próximo passo, também calcular e avaliar o PAE de todos os modelos para complementar nossa análise, uma vez que essa métrica indica qualidade relativa às posições das cadeias modeladas. Apesar de não haver nenhum caso em que o peptídeo *decoy* fora melhor modelado do que o peptídeo original, houve casos em que eles ainda assim foram bem modelados e parecem ter afinidades semelhantes pelo sítio da proteína. Por fim, identificamos a necessidade de aprimoramento das funções de pontuação de *docking* proteína-peptídeo, o que se mostra uma tarefa ainda desafiadora para o estudo de interações proteína-peptídeo.

Na parte II do trabalho, buscamos comparar estruturas de peptídeos preditas pelo *AlphaFold 2* nas formas *holo* e *apo.* A maior parte dos peptídeos foram modelados de modo similar ligado e desligado da proteína, obtendo valores de RMSD normalizados abaixo de 0.2.

A parte III do trabalho consistiu em criar modelos para predição de complexos reais e falsos. Como dito anteriormente, esse tipo de modelo e análise pode ser de grande utilidade para aprimorar funções de pontuação de *docking*. O fato de não obtermos modelos tão acurados (acurácia máxima de 0.741) demonstra a dificuldade ainda existente em diferenciar complexos reais de falsos, mesmo utilizando técnicas atuais como o uso das assinaturas estruturais para representar proteínas.

Por fim, na parte IV deste trabalho, utilizamos assinaturas dos peptídeos na forma *apo*, concatenadas com assinaturas das proteínas, para treinar um modelo de predição de ligação. Idealmente, deveríamos usar as assinaturas dos peptídeos na conformação ligada à proteína (*holo*), pois isso refletiria a estrutura exata que eles assumem quando ligados. No entanto, como nosso objetivo foi avaliar a capacidade do modelo em prever a ligação sem realizar o *docking*. Por esse motivo, utilizamos as assinaturas dos peptídeos na forma *apo*, já que não disporíamos de suas estruturas na forma *holo*. Assim, quanto mais similares forem as estruturas dos peptídeos nas formas *holo* e *apo*, mais semelhantes serão suas assinaturas, aumentando a confiança em nosso modelo de predição de ligação. Os modelos gerados obtiveram alta acurácia e se mostraram eficientes para a classificação proposta. Em trabalhos futuros, pretendemos realizar um estudo de caso para testar o modelo em exemplo real de triagem de peptídeos.

Embora nosso trabalho tenha mostrado resultados promissores, reconhecemos algumas limitações, como a necessidade de expandir as análises e avaliar outras ferramentas de *docking* proteína-peptídeo, bem como a necessidade da utilização de base de dados maior, como a Propedia, para treinamento e teste dos modelos desenvolvidos. Além disso, necessitamos ainda de realizar estudo de caso para teste dos modelos com exemplos reais de triagem virtual de peptídeos;

Por fim, nosso trabalho oferece uma contribuição para a área de Bioinformática Estrutural, proporcionando uma base para estudos futuros que busquem entender um pouco mais sobre as interações proteína-peptídeo. Acreditamos que nosso trabalho poderá ser aplicado em diversas áreas para a busca de fármacos baseados em peptídeos, destacando a relevância e o potencial impacto de nossa pesquisa.

6. Perspectivas

Como perspectivas futuras para este trabalho, pretendemos refinar os modelos de aprendizado de máquina utilizando os dados de estruturas de complexos proteína-peptídeo da base de dados Propedia 2, que hoje contém informações sobre mais de 49 mil complexos. Aumentar a quantidade de dados para treinamento dos modelos potencialmente aumentará a precisão e a robustez das predições de interação proteína-peptídeo. Além disso, planejamos realizar estudos de caso de triagem de peptídeos para avaliar a eficácia do modelo na predição de ligações proteína-peptídeo em situações práticas. Por fim, visamos criar um *pipeline* que integre o modelo de aprendizado de máquina de predição de ligação à Propedia, facilitando o uso por

outros pesquisadores e ampliando a aplicabilidade e o impacto de nossas pesquisas na comunidade científica.

7. Referências bibliográficas

AGRAWAL, P. et al. Benchmarking of different molecular docking methods for protein-peptide docking. **BMC Bioinformatics**, v. 19, n. S13, p. 426, fev. 2019.

ALAM, N. et al. High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. **PLOS Computational Biology**, v. 13, n. 12, p. e1005905, 27 dez. 2017.

ALLEN, W. J. et al. DOCK 6: Impact of new features and current docking performance. **Journal of Computational Chemistry**, v. 36, n. 15, p. 1132–1156, 5 jun. 2015.

ALQURAISHI, M. AlphaFold at CASP13. **Bioinformatics**, v. 35, n. 22, p. 4862–4865, 15 nov. 2019.

BARANWAL, M. et al. Struct2Graph: a graph attention network for structure based predictions of protein–protein interactions. **BMC Bioinformatics**, v. 23, n. 1, p. 370, 10 set. 2022.

BARASHKOVA, A. S.; RYAZANTSEV, D. Y.; ROGOZHIN, E. A. Rational Design of Plant Hairpin-like Peptide EcAMP1: Structural–Functional Correlations to Reveal Antibacterial and Antifungal Activity. **Molecules**, v. 27, n. 11, p. 3554, 31 maio 2022.

BASU, S.; WALLNER, B. DockQ: A Quality Measure for Protein-Protein Docking Models. **PLOS ONE**, v. 11, n. 8, p. e0161879, 25 ago. 2016.

BRÄNDÉN, C.-I.; TOOZE, J. **Introduction to protein structure**. 2nd ed ed. New York: Garland Pub, 1999.

CHENG, J. et al. BERTMHC: improved MHC–peptide class II interaction prediction with transformer and multiple instance learning. **Bioinformatics**, v. 37, n. 22, p. 4172–4179, 18 nov. 2021.

CIEMNY, M. et al. Protein–peptide docking: opportunities and challenges. **Drug Discovery Today**, v. 23, n. 8, p. 1530–1537, ago. 2018.

CUPP-VICKERY, J. R. et al. Crystal structure of the molecular chaperone HscA substrate binding domain complexed with the IscU recognition peptide ELPPVKIHC. **Journal of Molecular Biology**, v. 342, n. 4, p. 1265–1278, 24 set. 2004.

DE VRIES, S. J. et al. The pepATTRACT web server for blind, large-scale peptide-protein docking. **Nucleic Acids Research**, v. 45, n. W1, p. W361–W364, 3 jul. 2017.

DEMŠAR, J. et al. Orange: From Experimental Machine Learning to Interactive Data Mining. Em: BOULICAUT, J.-F. et al. (Eds.). **Knowledge Discovery in Databases: PKDD 2004**. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. v. 3202p. 537–539.

EVANS, R. et al. **Protein complex prediction with AlphaFold-Multimer**. , 4 out. 2021. Disponível em: <http://biorxiv.org/lookup/doi/10.1101/2021.10.04.463034>. Acesso em: 24 abr. 2024

FEI, Y. et al. LTPConstraint: a transfer learning based end-to-end method for RNA secondary structure prediction. **BMC Bioinformatics**, v. 23, n. 1, p. 354, 23 ago. 2022.

GAO, H.-R.; GAO, H.-Y. Cardiovascular functions of central corticotropin-releasing factor related peptides system. **Neuropeptides**, v. 75, p. 18–24, jun. 2019.

GEORGE, C.; BYUN, A.; HOWARD-THOMPSON, A. New Injectable Agents for the Treatment of Type 2 Diabetes Part 2—Glucagon-Like Peptide-1 (GLP-1) Agonists. **The American Journal of Medicine**, v. 131, n. 11, p. 1304–1306, 1 nov. 2018.

HELLINGER, R.; GRUBER, C. W. Peptide-based protease inhibitors from plants. **Drug Discovery Today**, v. 24, n. 9, p. 1877–1889, set. 2019.

HUANG, N.; SHOICHET, B. K.; IRWIN, J. J. Benchmarking Sets for Molecular Docking. **Journal of Medicinal Chemistry**, v. 49, n. 23, p. 6789–6801, 1 nov. 2006.

HUANG, S.-Y.; ZOU, X. An iterative knowledge-based scoring function for protein-protein recognition. **Proteins**, v. 72, n. 2, p. 557–579, ago. 2008.
HUGO VERLI, ET AL. **Bioinformática: da biologia à flexibilidade molecular**. [s.l.] SBBq, 2014.
JANDOVA, Z.; VARGIU, A. V.; BONVIN, A. M. J. J. Native or Non-Native Protein–Protein Docking Models? Molecular Dynamics to the Rescue. **Journal of Chemical Theory and Computation**, v. 17, n. 9, p. 5944–5954, 14 set. 2021.
JANIN, J. et al. CAPRI: A Critical Assessment of PRedicted Interactions. **Proteins: Structure, Function, and Bioinformatics**, v. 52, n. 1, p. 2–9, 2003.
JIN, R. et al. Botulinum neurotoxin B recognizes its protein receptor with high affinity and specificity. **Nature**, v. 444, n. 7122, p. 1092–1095, dez. 2006.
JOHANSSON-ÅKHE, I.; WALLNER, B. Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. **Frontiers in Bioinformatics**, v. 2, p. 959160, 26 set. 2022.
JUMPER, J. et al. Highly accurate protein structure prediction with AlphaFold. **Nature**, v. 596, n. 7873, p. 583–589, 26 ago. 2021.
KUMAR, N.; SRIVASTAVA, R. Deep learning in structural bioinformatics: current applications and future perspectives. **Briefings in Bioinformatics**, v. 25, n. 3, p. bbae042, 27 mar. 2024.
KUMAR, R. et al. Protein Sub-Nuclear Localization Prediction Using SVM and Pfam Domain Information. **PLoS ONE**, v. 9, n. 6, p. e98345, 4 jun. 2014.
KUNKEL, D. et al. Efficacy of the glucagon-like peptide-1 agonist exenatide in the treatment of short bowel syndrome. **Neurogastroenterology & Motility**, v. 23, n. 8, p. 739-e328, 2011.
LEHNINGER, A. L.; NELSON, D. L.; COX, M. M. **Lehninger principles of biochemistry**. 7th ed ed. New York: W.H. Freeman, 2017.
LEI, Y. et al. A deep-learning framework for multi-level peptide–protein interaction prediction. **Nature Communications**, v. 12, n. 1, p. 5465, 15 set. 2021.
LENSINK, M. F.; VELANKAR, S.; WODAK, S. J. Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. **Proteins**, v. 85, n. 3, p. 359–377, mar. 2017.
LI, H. et al. Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. **Molecular Informatics**, v. 34, n. 2–3, p. 115–126, fev. 2015.
LI, Y. et al. PredAmyl-MLP: Prediction of Amyloid Proteins Using Multilayer Perceptron. **Computational and Mathematical Methods in Medicine**, v. 2020, p. 1–12, 20 nov. 2020.
LONDON, N. et al. Rosetta FlexPepDock web server—high resolution modeling of peptide–protein interactions. **Nucleic Acids Research**, v. 39, n. Web Server issue, p. W249–W253, 1 jul. 2011.
LONDON, N.; MOVSHOVITZ-ATTIAS, D.; SCHUELER-FURMAN, O. The Structural Basis of Peptide-Protein Binding Strategies. **Structure**, v. 18, n. 2, p. 188–199, fev. 2010.
LUO, D. et al. Flexibility between the Protease and Helicase Domains of the Dengue Virus NS3 Protein Conferred by the Linker Region and Its Functional Implications. **Journal of Biological Chemistry**, v. 285, n. 24, p. 18817–18827, jun. 2010.
MA, Z.; ZOU, X. MDock: A Suite for Molecular Inverse Docking and Target Prediction. **Methods in Molecular Biology (Clifton, N.J.)**, v. 2266, p. 313–322, 2021.
MACIEJEWSKI, M. W. et al. NMRbox: A Resource for Biomolecular NMR Computation. **Biophysical Journal**, v. 112, n. 8, p. 1529–1534, abr. 2017.
MANCINI, A. L. et al. STING Contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. **Bioinformatics**, v. 20, n. 13, p. 2145–2147, 2004.
MARIANO, D. Métricas de avaliação em machine learning: acurácia, sensibilidade, precisão,

especificidade e F-score. Em: MARIANO, D. et al. (Eds.). **BIOINFO - Revista Brasileira de Bioinformática e Biologia Computacional**. 1. ed. [s.l.] Alfahelix, 2021.
MARIANO, D. C. B. et al. A Computational Method to Propose Mutations in Enzymes Based on Structural Signature Variation (SSV). **International Journal of Molecular Sciences**, v. 20, n. 2, p. 333, jan. 2019.
MARTINS, P. et al. Propedia v2.3: A novel representation approach for the peptide-protein interaction database using graph-based structural signatures. **Frontiers in Bioinformatics**, v. 3, p. 1103103, 16 fev. 2023.
MARTINS, P. M. et al. Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm. **BMC Bioinformatics**, v. 22, n. 1, p. 1, dez. 2021.
MAUPETIT, J.; DERREUMAUX, P.; TUFFERY, P. PEP-FOLD: an online resource for de novo peptide structure prediction. **Nucleic Acids Research**, v. 37, n. Web Server, p. W498–W503, 1 jul. 2009.
MCDONALD, E. F. et al. Benchmarking AlphaFold2 on peptide structure prediction. **Structure**, v. 31, n. 1, p. 111- 119.e2, jan. 2023.
MELO, R. C. et al. Finding protein-protein interaction patterns by contact map matching. 2007.
MIN, S.; LEE, B.; YOON, S. Deep learning in bioinformatics. **Briefings in Bioinformatics**, p. bbw068, 29 jul. 2016.
NESHICH, G. et al. STING Millennium: A web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. **Nucleic acids research**, v. 31, n. 13, p. 3386–3392, 2003.
ORCHARD, S. et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. **Nucleic Acids Research**, v. 42, n. D1, p. D358–D363, jan. 2014.
PIMENTEL, V. et al. VTR: A Web Tool for Identifying Analogous Contacts on Protein Structures and Their Complexes. **Frontiers in Bioinformatics**, v. 1, p. 730350, 8 nov. 2021.
PIRES, D. E. et al. Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. **BMC Genomics**, v. 12, n. S4, p. S12, dez. 2011.
PIRES, D. E. V. et al. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. **Bioinformatics**, v. 29, n. 7, p. 855–861, 1 abr. 2013.
RAJPOOT, S. et al. A Novel Therapeutic Peptide Blocks SARS-CoV-2 Spike Protein Binding with Host Cell ACE2 Receptor. **Drugs in R&D**, v. 21, n. 3, p. 273–283, set. 2021.
REY, J. et al. PEP-FOLD4: a pH-dependent force field for peptide structure prediction in aqueous solution. **Nucleic Acids Research**, v. 51, n. W1, p. W432–W437, 5 jul. 2023.
RODRIGUES, C. H. M. et al. CSM-peptides: A computational approach to rapid identification of therapeutic peptides. **Protein Science**, v. 31, n. 10, p. e4442, out. 2022.
SANTANA, C. A. et al. GRaSP: a graph-based residue neighborhood strategy to predict binding sites. **Bioinformatics**, v. 36, n. Supplement_2, p. i726–i734, 30 dez. 2020.
SAUER, M. M. et al. Catch-bond mechanism of the bacterial adhesin FimH. **Nature Communications**, v. 7, n. 1, p. 10738, 7 mar. 2016.
SILVA, M. F. M. et al. Proteingo: Motivation, user experience, and learning of molecular interactions in biological complexes. **Entertainment Computing**, v. 29, p. 31–42, mar. 2019.
SOBOLEV, V. et al. Automated analysis of interatomic contacts in proteins. **Bioinformatics (Oxford, England)**, v. 15, n. 4, p. 327–332, 1999.
SU, X. et al. Protein- and Peptide-Based Virus Inactivators: Inactivating Viruses Before Their Entry Into Cells. **Frontiers in Microbiology**, v. 11, 25 maio 2020.
SZKLARCZYK, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. **Nucleic**

Acids Research, v. 47, n. Database issue, p. D607–D613, 8 jan. 2019.
TEMML, V.; KUTIL, Z. Structure-based molecular modeling in SAR analysis and lead optimization. **Computational and Structural Biotechnology Journal**, v. 19, p. 1431–1444, 2021.
TORRISI, M.; POLLASTRI, G.; LE, Q. Deep learning methods in protein structure prediction. **Computational and Structural Biotechnology Journal**, v. 18, p. 1301–1310, 2020.
VARGA, J. K.; SCHUELER-FURMAN, O. Who Binds Better? Let Alphafold2 Decide! **Angewandte Chemie International Edition**, v. 62, n. 28, p. e202303526, 10 jul. 2023.
VERDONK, M. L. et al. Improved protein-ligand docking using GOLD. **Proteins**, v. 52, n. 4, p. 609–623, 1 set. 2003.
WANG, L. et al. Therapeutic peptides: current applications and future directions. **Signal Transduction and Targeted Therapy**, v. 7, n. 1, p. 48, 14 fev. 2022.
WEBB, B.; SALI, A. Comparative Protein Structure Modeling Using MODELLER. **Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]**, v. 54, p. 5.6.1-5.6.37, 20 jun. 2016.
WEN, Z. et al. PepBDB: a comprehensive structural database of biological peptide–protein interactions. **Bioinformatics**, v. 35, n. 1, p. 175–177, 1 jan. 2019.
WENG, G. et al. Comprehensive Evaluation of Fourteen Docking Programs on Protein–Peptide Complexes. **Journal of Chemical Theory and Computation**, v. 16, n. 6, p. 3959–3969, 9 jun. 2020.
WOO, J.-S. et al. Structural Basis for Protein Recognition by B30.2/SPRY Domains. **Molecular Cell**, v. 24, n. 6, p. 967–976, dez. 2006.
XU, X.; ZOU, X. PepPro: A Nonredundant Structure Data Set for Benchmarking Peptide–Protein Computational Docking. **Journal of Computational Chemistry**, v. 41, n. 4, p. 362–369, 5 fev. 2020.
YAN, Y. et al. Addressing recent docking challenges: A hybrid strategy to integrate template-based and free protein-protein docking. **Proteins: Structure, Function, and Bioinformatics**, v. 85, n. 3, p. 497–512, mar. 2017a.
YAN, Y. et al. HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. **Nucleic Acids Research**, v. 45, n. W1, p. W365–W373, 3 jul. 2017b.
YAN, Y.; ZHANG, D.; HUANG, S.-Y. Efficient conformational ensemble generation of protein-bound peptides. **Journal of Cheminformatics**, v. 9, n. 1, p. 59, 22 nov. 2017.
YASUO, N.; SEKIJIMA, M. Improved Method of Structure-Based Virtual Screening via Interaction-Energy-Based Learning. **Journal of Chemical Information and Modeling**, v. 59, n. 3, p. 1050–1061, 25 mar. 2019.
YIN, S.; MI, X.; SHUKLA, D. Leveraging machine learning models for peptide–protein interaction prediction. **RSC Chemical Biology**, p. 10.1039.D3CB00208J, 2024.
ZHAO, N. et al. Feature-based classification of native and non-native protein–protein interactions: Comparing supervised and semi-supervised learning approaches. **PROTEOMICS**, v. 11, n. 22, p. 4321–4330, nov. 2011.
ZHOU, P. et al. HPEPDOCK: a web server for blind peptide–protein docking based on a hierarchical algorithm. **Nucleic Acids Research**, v. 46, n. W1, p. W443–W450, 2 jul. 2018.

8. Apêndices

Tabela S1. Características dos complexos proteína-peptídeo selecionados da base de dados PepPro. Adaptado de XU et al., 2020

PDB ID	Peptideo			Proteína	
	Sequência	**N° Res**	**ES**	**Nome**	**N° Res**
1avf_Q:J	AVVKVPLKKFKSIRETMKEKGLLGEF	26	HE	Gastricsin	323
1d4t_B:A	KSLTIYAQVQK	11	bE	T cell signal transduction molecule SAP	104
1dkd_E:A	SWMTTPWGFLHP	12	sE	GroEL	146
1gux_E:B	DLYCYEQLN	9	C	Retinoblastoma protein	141
1hc9_C:A	WRYYESSLLPYPD	13	bE	α-bungarotoxin	74
1j2x_B:A	SEADEMAKALEAELNDLM	18	H	Transcription initiation factor IIF	67
1lb6_B:A	KQEPQEIDF	9	bE	TNF receptor-associated factor 6	155
1oj5_B:A	LPPTEQDLTKLLLE	14	H	Steroid receptor coactivator 1A	105
1skg_B:A	VAFRS	5	C	Phospholipase A2	121
1sqk_B:A	DLPKVAENLKSQLEGFNQDKLKNAS	25	pH	Actin, α skeletal muscle	360
1t0j_C:B	GAQQLEEDLKGYLDWITQAE	20	H	Voltage-gated calcium channel subunit β2a	187
1u00_P:A	ELPPVKIHC	9	bE	Chaperone protein hscA	227
1wkw_B:A	PGGTRIIYDRKFLMECRNSP	20	pH	Eukaryotic translation initiation factor 4E	191
2b1j_C:A	MGDSILSQAEIDALLN	16	H	Chemotaxis protein cheY	128
2bba_P:A	TNYLFSPNGPIARAW	15	pH	Ephrin type-B receptor 4	185
2c5k_P:T	KSLRVSSLNKDRRLLLREFYNLEN	24	H	T-snare affecting a late golgi compartment protein 1	89

2g30_P:A	DDGLDEAFSRLAQSRT	16	H	AP-2 complex subunit β-1	233
2hpl_B:A	DDLYG	5	C	PNGase	100
2ihs_D:B	DINNNNNIVEDVERKREFYI	20	pH	Gustavus, CG2944-PF	195
2nm1_B:A	EDMFAKLKDKFFNEINK	17	H	Botulinum neurotoxin type B	430
2okr_C:A	IKIKKIEDASNPLLLKRRKKARAL	24	pH	Mitogen-activated protein kinase 14	339
2peh_C:A	KRKSRWDETP	10	C	Splicing factor 45	104
2puy_E:B	ARTKQTARKS	10	bE	PHD finger protein 21A	60
2pv1_B:A	WEYIPNV	7	C	Chaperone surA	103
2qn6_C:A	SSEKEYVEMLDRLYSKLP	18	H	Translation initiation factor 2 γ subunit	393
2qos_A:C	LRYDSTAERLY	11	sE	Complement protein C8 γ	173
2qxv_B:A	KTMFSSNRQKILERTETLNQEWKQRRIQPV	30	H	Embryonic ectoderm development	352
2r9q_Y:B	VEVPLAGAV	9	C	2'-deoxycytidine 5'-triphosphate deaminase	342
2rl0_E:D	GQVTTESNLVEFDEESTK	18	bE	Fibronectin	89
2whx_C:A	ADLSLEKAANVQWD	14	bE	Serine protease / NTPase / Helicase NS3	582
2x0x_D:A	YLVGQIDSEVDTDDLSNFQL	20	HE	Ribonucleotide-diphosphate Reductase 1 subunit α	728
2xs1_B:A	SREKPYKEVTEDLLHLNSLF	20	H	Programmed cell death 6-interacting protein	697
2xum_S:A	HLEVVKLLLEHGADVDAQDK	20	pH	Hypoxia-inducible factor 1-α inhibitor	349
2ybf_B:A	SKYRKKHKSEFQLLVDQARKGYKKIAG	27	H	Ubiquitin-conjugating enzyme E2 B	150
3at0_B:A	GSWNSGSSGTGSTGNQ	16	bE	Clumping factor B	317

3awr_C:A	GPRLSRLLSSAGC	13	H	Mitochondrial import receptor subunit TOM20 homolog	73
3d9u_B:A	AVPIAQ	6	bE	Baculoviral IAP repeat-containing protein 2	92
3dab_B:A	SQETFSDLWKLLPEN	15	H	Mdm4 protein	88
3dy0_B:A	RSQRLVFNRPFLMFIVDNNILFLGKVNRP	29	sE	N-terminus Plasma serine protease inhibitor	328
3hbv_Z:P	AKASQAA	7	bE	Secreted protease C	380
3ik5_B:A	AYQQGQNQLYNELNLGRR	18	H	Protein Nef	119
3kj0_B:A	GSGGRPEIWYAQELRRIGDEFNAYYAR	27	H	Induced myeloid leukemia cell differentiation protein Mcl-1	157
3kut_C:A	SNLNPNAAEFVPGVKYG	17	C	Polyadenylate-binding protein 1	84
3l81_B:A	TYKFFEQ	7	bE	AP-4 complex subunit mu-1	250
3lu9_C:B	ATNATLDPRSFLLRNPNDKYEPFWE	25	C	Prothrombin	251
3n3x_B:A	SDDDMG	6	C	Ribosome inactivating protein	246
3njf_B:A	PQIINRPQN	9	bE	Peptidase	112
3o37_E:A	ARTKQTARKS	10	bE	Transcription intermediary factor 1-α	175
3plv_C:A	SLSIEETNELRASLGLKLIPP	21	pH	Ubiquitin-like modifier HUB1	80
3r7g_B:A	KSLYKIKPRHDSGIKAKISMKT	22	HE	Protein spire homolog 1	154
3ro2_B:A	RNSFYMGTCQDEPEQLDDWNRIAELQQR	28	pH	G-protein-signaling modulator 2	328
3ryb_B:A	SLSQSLSQS	9	bE	Oligopeptide-binding protein oppA	563
3so6_Q:A	NSINFDNPVYQKTT	14	bE	LDL receptor adaptor protein	137
3ukx_C:B	GSRRRRRRKRKREWDDDDDPPKKRRRLD	28	C	Importin subunit α-2	426
4aom_T:A	KNIPSLLRVQAHIRKKMV	18	H	Myosin A tail domain interacting protein	143

4dj9_B:A	ETQVVLINAVKDVAKALGDLISATKAAAG	29	H	Vinculin	242
4ext_B:C	RTANILKPLMSPPSREEIMATLL	23	HE	Mitotic spindle assembly checkpoint protein MAD2B	198
4hh6_Z:A	KKWDSVYASLFEKINLKK	18	H	Putative type VI secretion protein	157
4htp_C:A	DVPQWEVFFKR	11	pH	DNA ligase 4	221
4j1v_G:A	ALPAWARPDYNPPLVESWRR	20	pH	MOB kinase activator 1A	166
4k0u_B:A	RTFRQVQSSISDFYD	15	H	Lipoprotein OutS	95
4m5s_B:A	GERTIPITRE	10	bE	α-crystallin B	87
4oni_C:A	QGAASRPAILYALLSSSLK	19	H	Human nuclear receptor LRH1	241
4q5u_C:A	ARKEVIRNKIRAIGKMARVFSVLR	24	H	Calmodulin	145
4qqi_X:A	KAFVHMPTLPNLDFHKT	17	C	Ankyrin repeat family A protein 2	176
4uwx_D:B	TPRSARLERMAQALALQAGSP	21	H	Protein diaphanous homolog 1	230
4x3h_B:A	RIPSYRYRY	9	bE	Activity-regulated cytoskeleton-associated protein	79
4xoe_B:A	ADVTITVNGKVVAK	14	bE	FimH protein	279
4yl6_B:A	MDEQEALNSIMNDLVALQMNRR	22	H	Malcavernin	88
4yz6_B:A	ELPIARRASLHRFLEKRKDRVT	22	H	Transcription factor MYC3	175
5crw_B:A	GKTKEGVLYVG	11	C	Protein disulfide-isomerase	242
5epp_B:A	SPEEMRRQRLHRFDS	15	H	Transitional endoplasmic reticulum ATPase	170
5f67_C:A	GPGSRGKSTVTGRMISGWL	19	sE	Inactivation-no-after-potential D protein	98
5fzt_B:A	PELDDILYHVKGMQRIVNQWSEK	23	H	TALIN-1	306

5gtu_B:A	GQQDLMINNPLSQDEG SLWNKFFQDKE	27	pH	Vacuolar protein sorting-associated protein 29	186

PDB ID: cadeia do peptídeo:cadeia da proteína

ES: estrutura secundária do peptídeo (H: helix, pH: partial Helix, C:coil, HE: helix e B-strand, bE:binding stand, sE: self-folding strand)

Tabela S2. Sequências dos peptídeos originais e dos peptídeos *decoys.*

PDB	**Peptideo Original**	**Peptídeo *Decoy***
1avf	AVVKVPLKKFKSIRETMKEKGL	VIKKFVKLETPLVAMKKKRESG
1d4t	KSLTIYAQVQK	IVQKQATLKSY
1dkd	WMTTPWGFLHP	HTWTWFMLPGP
1gux	DLYCYEQLN	ELDLYCQYN
1hc9	WRYYESSLLPYPD	YELPYSRWPDYLS
1j2x	SEADEMAKALEAELNDLM	DAENEEAEASLKLADMLM
1lb6	KQEPQEIDF	EKIPQQEFD
1oj5	LPPTEQDLTKLLLE	TPKLLQELDTLEPL
1skg	VAFRS	RVSAF
1sqk	DLPKVAENLKSQLEGFNQDKLKNAS	KKKLFAGKESALQNLDSDELQNVNP
1t0j	QQLEEDLKGYLDWITQ	QDIKWLETLEQGDQLY
1u00	ELPPVKIHC	PEHKICVLP
1wkw	PGGTRIIYDRKFLMECRNSP	NLYTIMRFRDIEGSRKPGPC
2b1j	GDSILSQAEIDALLN	ADISDQNLGLILASE
2bba	NYLFSPNGPIARAW	PSYNWNPFAIGARL
2c5k	KSLRVSSLNKDRRLLLREFYNL	KLELYFVLLSRSLRDSRRLKNN
2g30	DDGLDEAFSRLAQSRT	QFDDAAEGSLTRDSRL
2hpl	DDLYG	GLYDD
2ihs	DINNNNNIVEDVERKREFYI	RINDINEVFNNKERVDNEYI
2nm1	EDMFAKLKDKFFNEINK	KFELDFMANKFKKIDNE

2okr	IKIKKIEDASNPLLLKRRKKARAL	LKIARNAIKAPDKKLRRILSKELK
2peh	KRKSRWDETP	KKTWRRESDP
2puy	ARTKQTARKS	KQRTTAASRK
2pv1	WEYIPNV	IYWVPEN
2qn6	SSEKEYVEMLDRLYSKLP	EYVLRSLKYPKDELSSEM
2qos	LRYDSTAERLY	AYRESLRDTYL
2qxv	TMFSSNRQKILERTETLNQEWKQRRIQPV	RVTIKQQLRTREELQNSMEWFNTKSQPIR
2r9q	VEVPLAGAV	GPVVEAALV
2rl0	QVTTESNLVEFDEEST	TEVVTFNELSESDQET
2whx	ADLSLEKAANVQWD	KVENLADWQSAADL
2x0x	QIDSEVDTDDLSNFQL	DTLVDLNQEFDQDISS
2xs1	EKPYKEVTEDLLHLN	NTYKDLHLEEKEPLV
2xum	EVVKLLLEHGADVDAQDK	KELEKVGLLDQVDHVDAA
2ybf	KYRKKHKSEFQLLVDQARKGY	KQAKYKGRQHSVLKLFKRYED
3at0	WNSGSSGTGSTG	SSNWSTGGGGTS
3awr	GPRLSRLLSSAG	GLSLRSRPLAGS
3d9u	AVPIAQ	AQVAIP
3dab	ETFSDLWKLLPE	WDTLPSEFLEKL
3dy0	SQRLVFNRPFLMFIVDNNILFLGKVNRP	DGVRFPRKILPVVFRSLNLINQNNMFFL
3hbv	AKASQAA	AAKQSAA
3ik5	AYQQGQNQLYNELNLG	NENYLGYLQQLQANGQ
3kj0	GRPEIWYAQELRRIGDEFNAYYAR	GYAYGNIFARDQWAYERPELIERR
3kut	NLNPNAAEFVPGVKYG	PYKFGAPGVNVNLENA
3l81	TYKFFEQ	KFTEFQY
3lu9	ATNATLDPRSFLLRNPNDKYEPFWE	KWTALNFRNEYETRNSLPDLDPAPF

3n3x	SDDDMG	DSDDMG
3njf	PQIINRP	IPRIPQN
3o37	ARTKQTARKS	KAKRAQTSRT
3plv	SLSIEETNELRASLGLKLIPP	NEIELSLPSESPTLIKGARLL
3r7g	YKIKPRHDSGIKAKISMKT	KKRIMGDIYSIKAKPKHST
3ro2	FYMGTCQDEPEQLDDWNRIAEL	TQIDQEFLMCLWYPNGDERADE
3ryb	SLSQSLSQS	SLLSSQSSQ
3so6	SINFDNPVYQKTT	QVDSPNTNKYFIT
3ukx	RRRKRKREWDDDDDPPKKRRRLD	KDRWRRDKRRDRRKPDEKPDDLR
4aom	KNIPSLLRVQAHIRKKMV	LNRKKIHSRPIVQKVLMA
4dj9	VVLINAVKDVAKALGDLISATK	LAVTKINASALIVVADDLVGKK
4ext	RTANILKPLMSPPSREEIMATLL	ESRPLMSAMITLPPKLILEATRN
4hh6	WDSVYASLFEKINL	SDEAWFKINYVLSL
4htp	DVPQWEVFFK	FEKPFQVWVD
4j1v	LPAWARPDYNPPLVE	LLPEVPWNDYRAAPP
4k0u	RTFRQVQSSISDFYD	STSDFVSQRIYDQFR
4m5s	GERTIPITRE	IPTEGERTIR
4oni	GAASRPAILYALLSSS	ASASILLPGYRLSSAA
4q5u	ARKEVIRNKIRAIGKMARVFSVLR	IIKKGRVEASRIVRNRAKVFRLAM
4qqi	AFVHMPTLPNLDFHKT	TNVDAMLFTKFPPHHL
4uwx	TPRSARLERMAQALAL	LQMPESLATALRAARR
4x3h	RIPSYRYRY	YRYIRRYSP
4xoe	ADVTITVNGKVVAK	KGDTNATVAVVVIK
4yl6	DEQEALNSIMNDLVALQM	IVEDLAQNENLMMSADQL
4yz6	LPIARRASLHRFLEKRKD	RIRFPKLRLKDSARAHEL
5crw	GKTKEGVLYVG	EKYGGLTVVGK

5epp	SPEEMRRQRLHRFDS	RMRSEERFSQPRLDH
5f67	GPGSRGKSTVTGRMISGWL	PGRSKWRMSVGGTSTGGIL
5fzt	PELDDILYHVKGMQRIVNQWSEK	NQMPDSIWEIKYEVHLDGQKVLR
5gtu	QDLMINNPLSQDEGSLWNKFFQDK	GQDPQNLSKWFDQLENINKDLSFM

Tabela S3. Erros de predição de complexos reais (R) e falsos (D) pelo modelo de Redes Neurais.

	Complexo	**Classe real**	**Classe predita**
1	1avf - AlphaFold	D	R
2	1d4t - AlphaFold	D	R
3	1d4t - AlphaFold	R	D
4	1dkd - AlphaFold	D	R
5	1dkd - AlphaFold	R	D
6	1gux - HDOCK	R	D
7	1gux - HDOCK	D	R
8	1gux - HPEPDOCK	D	R
9	1hc9 - AlphaFold	R	D
10	1hc9 - HDOCK	D	R
11	1j2x - AlphaFold	R	D
12	1j2x - AlphaFold	D	R
13	1lb6 - AlphaFold	D	R
14	1oj5 - AlphaFold	D	R
15	1oj5 - HPEPDOCK	R	D
16	1skg - AlphaFold	R	D
17	1skg - AlphaFold	D	R

18	1skg - HDOCK	R	D
19	1skg - HPEPDOCK	D	R
20	1t0j - AlphaFold	D	R
21	1t0j - HPEPDOCK	D	R
22	1u00 - AlphaFold	R	D
23	1u00 - AlphaFold	D	R
24	1u00 - HPEPDOCK	D	R
25	1u00 - HPEPDOCK	R	D
26	1wkw - AlphaFold	D	R
27	1wkw - HPEPDOCK	D	R
28	2b1j - AlphaFold	R	D
29	2b1j - HPEPDOCK	R	D
30	2bba - AlphaFold	D	R
31	2bba - HDOCK	D	R
32	2c5k - HPEPDOCK	R	D
33	2g30 - HPEPDOCK	D	R
34	2hpl - HPEPDOCK	D	R
35	2ihs - AlphaFold	R	D
36	2okr - HPEPDOCK	D	R
37	2okr - HPEPDOCK	R	D
38	2puy - HPEPDOCK	D	R
39	2pv1 - HDOCK	R	D
40	2qn6 - HPEPDOCK	D	R
41	2qos - AlphaFold	D	R

42	2qos - AlphaFold	R	D
43	2qxv - HDOCK	D	R
44	2qxv - HPEPDOCK	R	D
45	2r9q - AlphaFold	D	R
46	2rl0 - HDOCK	R	D
47	2rl0 - HDOCK	D	R
48	2rl0 - HPEPDOCK	R	D
49	2whx - AlphaFold	D	R
50	2x0x - AlphaFold	R	D
51	2xs1 - HPEPDOCK	D	R
52	2xum - AlphaFold	R	D
53	2ybf - AlphaFold	R	D
54	3at0 - AlphaFold	D	R
55	3at0 - HPEPDOCK	D	R
56	3d9u - AlphaFold	D	R
57	3d9u - AlphaFold	R	D
58	3d9u - HDOCK	R	D
59	3d9u - HPEPDOCK	D	R
60	3dy0 - HDOCK	D	R
61	3hbv - AlphaFold	R	D
62	3ik5 - AlphaFold	D	R
63	3ik5 - HPEPDOCK	D	R
64	3kj0 - AlphaFold	D	R
65	3n3x - AlphaFold	R	D

66	3n3x - HDOCK	R	D
67	3n3x - HPEPDOCK	R	D
68	3o37 - AlphaFold	R	D
69	3o37 - HDOCK	R	D
70	3o37 - HPEPDOCK	R	D
71	3plv - AlphaFold	R	D
72	3plv - HPEPDOCK	R	D
73	3r7g - AlphaFold	R	D
74	3r7g - HPEPDOCK	R	D
75	3ro2 - HPEPDOCK	R	D
76	3ryb - AlphaFold	D	R
77	3ryb - AlphaFold	R	D
78	3so6 - AlphaFold	R	D
79	3so6 - HPEPDOCK	R	D
80	3ukx - AlphaFold	D	R
81	4aom - AlphaFold	R	D
82	4dj9 - AlphaFold	D	R
83	4dj9 - AlphaFold	R	D
84	4ext - HPEPDOCK	D	R
85	4ext - HPEPDOCK	R	D
86	4htp - AlphaFold	R	D
87	4j1v - AlphaFold	D	R
88	4j1v - HPEPDOCK	R	D
89	4k0u - AlphaFold	D	R

90	4k0u - AlphaFold	R	D
91	4m5s - HPEPDOCK	D	R
92	4m5s - HPEPDOCK	R	D
93	4oni - HDOCK	D	R
94	4oni - HPEPDOCK	R	D
95	4q5u - AlphaFold	D	R
96	4qqi - AlphaFold	D	R
97	4qqi - HPEPDOCK	D	R
98	4uwx - AlphaFold	D	R
99	4uwx - HPEPDOCK	D	R
100	4x3h - HPEPDOCK	D	R
101	4xoe - AlphaFold	D	R
102	4xoe - HPEPDOCK	D	R
103	4yz6 - AlphaFold	R	D
104	4yz6 - HPEPDOCK	D	R
105	5crw - HPEPDOCK	R	D
106	5crw - HPEPDOCK	D	R
107	5epp - AlphaFold	D	R
108	5epp - HPEPDOCK	R	D
109	5f67 - AlphaFold	D	R
110	5f67 - HPEPDOCK	R	D
111	5fzt - AlphaFold	D	R
112	5fzt - HPEPDOCK	R	D
113	5fzt - HPEPDOCK	D	R

114	5gtu - HPEPDOCK	D	R

Tabela S4. Erros de predição de peptídeos ligantes (R) e não ligantes (D) pelo modelo usando Gradient Boosting.

	Complexo	**Classe Real**	**Classe Predita**
1	4m5s - AlphaFold	D	R
2	4ext - HPEPDOCK	D	R
3	3d9u - HPEPDOCK	D	R
4	1avf - AlphaFold	D	R
5	3d9u - AlphaFold	D	R
6	4ext - AlphaFold	D	R
7	4xoe - AlphaFold	D	R
8	2hpl - HPEPDOCK	D	R
9	3so6 - AlphaFold	D	R
10	1oj5 - AlphaFold	D	R
11	3hbv - HPEPDOCK	D	R
12	3hbv - AlphaFold	D	R
13	3ryb - AlphaFold	D	R
14	3n3j - HPEPDOCK	D	R
15	4dj9 - HPEPDOCK	D	R
16	4dj9 - AlphaFold	D	R
17	2b1j - AlphaFold	D	R
18	2whx - AlphaFold	D	R
19	1lb6 - HDOCK	R	D
20	4j1v - HPEPDOCK	R	D

21	3ro2 - HPEPDOCK	R	D
22	1u00 - AlphaFold	R	D
23	2ihs - AlphaFold	R	D
24	3hbv - AlphaFold	R	D
25	1wkw - HPEPDOCK	R	D
26	2rl0 - HPEPDOCK	R	D
27	3plv - HPEPDOCK	R	D
28	2b1j - AlphaFold	R	D
29	4m5s - HPEPDOCK	R	D
30	1d4t - AlphaFold	R	D
31	3o37 - HPEPDOCK	R	D