UNIVERSIDADE FEDERAL DO RIO GRANDE - FURG

CIÊNCIAS DA SAÚDE

CURSO DE MESTRADO EM CIÊNCIAS DA SAÚDE

Dissertação de Mestrado

Avaliação comparativa de Ferramentas para Predição Estrutural de proteínas com Mutação Pontual: estudo de caso em *Mycobacterium tuberculosis*

Veridiana Piva Richter

Dissertação de Mestrado apresentada ao Ciências da Saúde da Universidade Federal do Rio Grande - FURG, como requisito parcial para a obtenção do grau de Mestre em Ciências da Saúde

Orientador: Profa. Dra. Karina dos Santos Machado

Rio Grande, 2026

AGRADECIMENTOS

Agradeço pelo amparo da minha família e pela dedicação e ajuda do meu namorado, por todo o suporte que vocês me deram durante a realização do meu mestrado. Sem vocês, eu não estaria concluindo este sonho. Sou grata aos meus professores e à minha orientadora, por terem me guiado ao longo dessa trajetória, aprendi muito ao seu lado. Desejo manifestar meu reconhecimento aos meus colegas do Laboratório de Biologia Computacional (COMBI-Lab), principalmente ao mestrando Abdirasak Mohamed Osman, que me ajudou na automatização e criação de códigos para a realização deste trabalho. Gostaria de reconhecer o apoio do Programa de Pós-Graduação em Ciências da Saúde da FURG e também o auxílio da bolsa da CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, sem esse auxílio financeiro, eu não conseguiria me dedicar exclusivamente à conclusão desta pesquisa. Muito obrigada!

RESUMO

RICHTER, Veridiana Piva. **Avaliação comparativa de Ferramentas para Predição Estrutural de proteínas com Mutação Pontual: estudo de caso em *Mycobacterium tuberculosis***. 2026. 115 f. Dissertação (Mestrado) – Ciências da Saúde. Universidade Federal do Rio Grande - FURG, Rio Grande.

As proteínas são compostas por uma ou mais cadeias longas de aminoácidos dobradas de forma específica, desempenhando um papel crucial na estrutura, função e regulação dos processos celulares e vias metabólicas. Elas podem sofrer mutações, como as do tipo *missense*, caracterizadas pela substituição de um par de bases no DNA, e consequente a troca de um aminoácido por outro na sequência da proteína, o que pode alterar propriedades como conformação, estabilidade, flexibilidade e resistência a medicamentos. A tuberculose é uma doença infecciosa causada pelo *Mycobacterium tuberculosis*, uma bactéria que atinge principalmente os pulmões e quando o tratamento é realizado inadequadamente, podem surgir cepas resistentes aos medicamentos, dificultando o controle da infecção e aumentando a mortalidade. Por isso, o diagnóstico precoce e o uso correto de antibióticos são fundamentais para prevenir essas situações. Nesse cenário, a bioinformática estrutural surge como campo fundamental, reunindo repositórios, algoritmos e ferramentas para explorar, analisar, prever e simular estruturas e interações. Os avanços na área, somados ao aumento da capacidade computacional, aos novos modelos de inteligência artificial e ao crescimento dos bancos de dados estruturais, têm permitido o desenvolvimento de modelos preditivos de estruturas tridimensionais de proteínas altamente confiáveis. Como análises experimentais demandam tempo e recursos elevados, muitas estruturas permanecem indeterminadas, exigindo métodos computacionais in silico para geração de modelos tridimensionais. Nesse contexto, um dos grandes desafios é a predição de estruturas de proteínas com mutações pontuais. Nosso objetivo foi avaliar diferentes ferramentas de predição de estruturas tridimensionais de proteínas com mutações pontuais *missense* e analisar diferentes métricas de validação a fim de descobrir qual das ferramentas performa melhor nesses casos. Este trabalho utiliza os algoritmos de predição estrutural — ColabFold, trRosetta, Swiss-Model, AlphaFold3, I-TASSER, Modeller e Phyre2 — para gerar modelos tridimensionais de proteínas com mutações *missense*. Para análise da qualidade dos modelos gerados, são utilizados validadores como MolProbity, SAVES (Verify3D e ERRAT), VoroMQA, QMEAN e QMEANDisCo, que oferecem métricas para identificar quais ferramentas foram as mais adequadas para predição estrutural com mutações. O estudo foca em proteínas-alvo relevantes no contexto da tuberculose, com mutações associadas à resistência a medicamentos. Os resultados obtidos com esse estudo mostram que AlphaFold3 e trRosetta se destacaram como as mais eficazes, com consistência e um bom desempenho na maioria dos casos, seguido pelo SWISS-MODEL, que apresentou desempenho estável. O ColabFold teve resultados medianos, porém consistentes, enquanto

as ferramentas I-TASSER e Phyre2 mostraram variabilidade elevada e menor confiabilidade. Por fim, o MODELLER teve desempenho inferior em todas as métricas. A análise com o algoritmo de *Borda Count* reforça esses resultados encontrados e oferece uma base sólida para a escolha criteriosa de ferramentas em estudos de modelagem estrutural voltados à mutações pontuais *missense* associadas à resistência a fármacos no tratamento da tuberculose.

Palavras-chave: Proteínas, mutações, predição, estrutura.

ABSTRACT

RICHTER, Veridiana Piva. **Comparative Evaluation of Tools for Structural Prediction of Proteins with Point Mutations: A Case Study in *Mycobacterium tuberculosis***. 2026. 115 f. Dissertação (Mestrado) – Ciências da Saúde. Universidade Federal do Rio Grande - FURG, Rio Grande.

Proteins are composed of one or more long chains of amino acids folded in a specific way, playing a crucial role in the structure, function, and regulation of cellular processes and metabolic pathways. They can undergo mutations, such as missense mutations, characterized by the substitution of a base pair in the DNA and, consequently, the replacement of one amino acid by another in the protein sequence, which can alter properties such as conformation, stability, flexibility, and drug resistance. Tuberculosis is an infectious disease caused by Mycobacterium tuberculosis, a bacterium that primarily affects the lungs, and when treatment is carried out inadequately, drug-resistant strains may emerge, making infection control difficult and increasing mortality. Therefore, early diagnosis and the correct use of antibiotics are essential to prevent these situations.

In this scenario, structural bioinformatics emerges as a fundamental field, bringing together repositories, algorithms, and tools to explore, analyze, predict, and simulate structures and interactions. Advances in the area, combined with increased computational capacity, new artificial intelligence models, and the growth of structural databases, have enabled the development of highly reliable predictive models of protein three-dimensional structures. As experimental analyses demand considerable time and resources, many structures remain undetermined, requiring in silico computational methods to generate three-dimensional models.

In this context, one of the major challenges is the prediction of protein structures with point mutations. Our objective was to evaluate different tools for predicting the three-dimensional structures of proteins with missense point mutations and to analyze different validation metrics in order to determine which of the tools performs best in these cases. This work uses structural prediction algorithms — ColabFold, trRosetta, Swiss-Model, AlphaFold3, I-TASSER, Modeller, and Phyre2 — to generate three-dimensional models of proteins with missense mutations.

To analyze the quality of the generated models, validation tools such as MolProbity, SAVES (Verify3D and ERRAT), VoroMQA, QMEAN, and QMEANDisCo are used, which provide metrics to identify which tools were most suitable for structural prediction with mutations. The study focuses on target proteins relevant in the context of tuberculosis, with mutations associated with drug resistance. The results obtained from this study show that AlphaFold3 and trRosetta stood out as the most effective, with consistency and good performance in most cases, followed by SWISS-MODEL, which showed stable performance. ColabFold had average but

consistent results, while the tools I-TASSER and Phyre2 showed high variability and lower reliability. Finally, MODELLER had inferior performance in all metrics. The analysis with the Borda Count algorithm reinforces these findings and provides a solid basis for the careful selection of tools in structural modeling studies focused on missense point mutations associated with drug resistance in tuberculosis treatment.

Keywords: proteins, mutations, prediction, structure.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BAAR	Bacilos álcool-ácido resistentes
BD	Banco de Dados
CASP	Critical Assessment of Structure Prediction
CC	Concentração crítica
Combi-Lab	Laboratório de Biologia Computacional
DP	Desvio Padrão
EBI	*European Bioinformatics Institute*
EMBL	*European Molecular Biology Laboratory*
FM	*Free modeling*
FURG	Universidade Federal do Rio Grande
GMQE	*Global Model Quality Estimate*
LDDT	*Local Distance Difference Test*
MSA	*Alinhamentos Múltiplos de Sequências*
MTBC	*Complexo Mycobacterium tuberculosis*
NCBI	*National Center for Biotechnology Information*
OMS	Organização Mundial da Saúde
PDB	*Protein Data Bank*
pLDDT	*Predicted local distance difference test*
PPV	*Valor Preditivo Positivo*
pTM	*Predicted template modelling*
mRNA	RNA Mensageiro
rRNA	*RNA Ribossômico*
RIF	Rifampicina
SAVeS	*Structure Analysis and Verification Serve*
SMTL	*SWISS-MODEL Template Library*
SNP	Polimorfismo de nucleotídeo único

TB	Tuberculose
TBM	*Template-based modeling*
TB-MDR	Tuberculose Multirresistente
tRNA	RNA Transportador
WT	Wild-type
wwPDB	*Worldwide Protein Data Bank*

SUMÁRIO

1 INTRODUÇÃO

O estudo realizado em 1970 provou que a estrutura terciária de uma proteína depende da sua sequência de aminoácidos [8]. Desde que ocorreu essa descoberta, a predição da estrutura de proteínas tornou-se alvo de pesquisas, inclusive de competições na área, como o *Critical Assessment of Structure Prediction* (CASP) [1], que acontece a cada dois anos desde 1994 e estabelece o atual cenário da predição da estrutura de proteínas, identificando quais foram os avanços feitos e onde ainda existem lacunas necessitando de esforços e melhorias [111].

Há atualmente muitas ferramentas para predição de estruturas de proteínas como por exemplo o Phyre2 [105], Swiss-Model [81], I-Tasser [218], MODELLER [202], entre outras. Essas ferramentas têm por objetivo reduzir a lacuna existente entre a quantidade de sequências de proteínas que temos disponíveis nos bancos de dados públicos e o número de estruturas determinadas experimentalmente [154]. Mais recentemente surgiram os métodos de predição de estrutura de proteínas que utilizam aprendizagem profunda como o Alphafold2 [101], Alphafold3 [2] e trRosetta [66] que fornecem com precisão estruturas muito próximas das obtidas experimentalmente [9].

As mutações pontuais ocorrem em proteínas de duas formas, mutações *missense* que alteram um único aminoácido em uma sequência protéica [149] e as mutações sem sentido, quando o códon de um aminoácido sofre mutação para formar um códon de parada [14].

Visando um melhor entendimento sobre o impacto dessas mutações na estrutura e a relação da mesma com a resistência da proteína a fármacos. É fundamental termos um conhecimento mais aprofundado sobre as estruturas das proteínas e suas mutações, principalmente as pontuais.

A tuberculose é uma doença infecciosa crônica que não possuía cura eficaz até a primeira metade do século XX. O maior entendimento sobre a doença e a descoberta de antibióticos eficazes aceleraram a recuperação dos pacientes e reduziram a prevalência e a transmissão dos casos. No entanto, o surgimento de cepas resistentes a medicamentos no final da década de 1980 levou a OMS a declarar a tuberculose como uma emergência de saúde global em 1993 e que perdura até hoje [127].

O *M. tuberculosis* tem um genoma que se adapta rapidamente ao estresse causado pelo

[1] https://predictioncenter.org/

medicamento, aumentando a resistência aos fármacos anti-tuberculose, o que está frequentemente associado à mutações moleculares. As mutações no genoma do *M. tuberculosis* levam a alterações no alvo de ação do medicamento, afetando a sua ação. Hoje, os medicamentos antituberculose de primeira e segunda linha continuam a ser a base do tratamento da TB, no entanto, pode desenvolver-se resistência medicamentosa se não forem utilizados adequadamente [87].

1.1 Objetivo Geral

Avaliar diferentes ferramentas de predição de estrutura de proteínas para a modelagem de proteínas com mutações pontuais *missense*, comparando diferentes métricas a fim de descobrir quais ferramentas são mais apropriadas para a modelagem desse tipo de proteína.

1.2 Objetivos específicos

- Automatizar o uso de ferramentas de predição e validação para um estudo de caso da TB;
- Gerar e avaliar modelos de estruturas de proteínas com mutações pontuais através de ferramentas de predição;
- Disponibilizar um banco de dados contendo as estruturas com mutação relacionadas a TB em um *website* para acesso aos resultados das validações dos modelos obtidos.

2 REFERENCIAL TEÓRICO

Esta seção tem como objetivo apresentar os principais conceitos que fundamentam o nosso trabalho de dissertação, oferecendo uma base teórica para a análise dos dados e a discussão dos resultados.

2.1 Dogma central

O dogma central da biologia molecular, conforme apresentado na Figura 1, é composto principalmente pelos passos de transcrição e tradução, e descreve como a informação genética contida no DNA é usada para produzir proteínas [83].

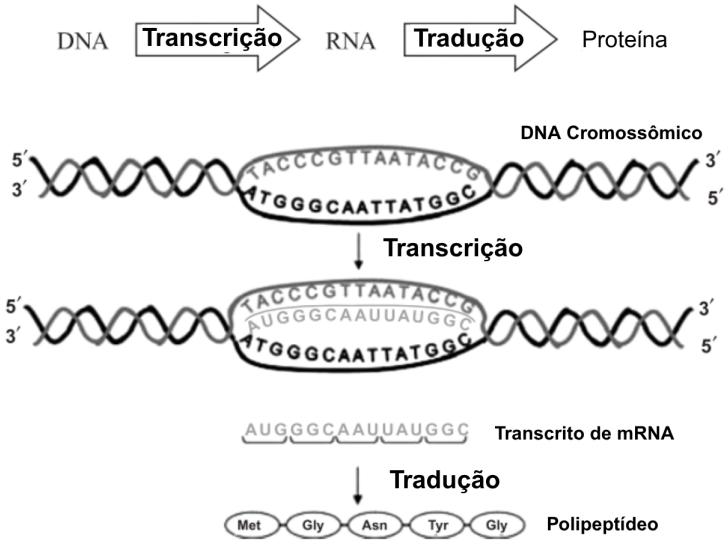

Figura 1: Dogma central da biologia molecular [83].

2.2 Replicação

A replicação é o processo pelo qual uma molécula de DNA dupla-fita é copiada para originar duas moléculas de DNA idênticas, sendo uma etapa imprescindível para a divisão celular e para a transmissão da informação genética. Este mecanismo é caracterizado como semiconservativo, o que significa que cada nova molécula de DNA formada contém uma fita original, conservada da molécula parental, e uma fita recém-sintetizada [91, 73].

O processo envolve uma complexa maquinaria multiproteica e se inicia nas origens de replicação, onde a dupla-fita de DNA é desenrolada e separada por enzimas como as helicases, expondo as fitas simples que atuarão como moldes. A partir de um primer de RNA, DNA polimerases atuam adicionando desoxirribonucleotídeos complementares, sintetizando as novas cadeias invariavelmente no sentido 5' para 3'. Devido à natureza antiparalela das fitas de DNA, a síntese ocorre de forma contínua na fita líder e de forma descontínua na fita tardia, com a formação de fragmentos de Okazaki. A alta fidelidade desse processo é essencial para a estabilidade do genoma, embora a introdução de variações ocasionais contribua para a diversidade genética e a evolução [59, 63, 176].

2.3 Transcrição

A transcrição é a primeira etapa da expressão gênica e consiste na síntese de uma molécula de RNA a partir de um molde de DNA. Esse processo ocorre no núcleo da célula e envolve a leitura de um trecho de uma das fitas do DNA, utilizado como molde para gerar uma molécula de RNA complementar e antiparalela. Para que a transcrição ocorra, são necessários alguns elementos essenciais: a fita molde de DNA, os nucleotídeos que vão compor o RNA e um conjunto de proteínas, entre elas a RNA polimerase — a enzima responsável por catalisar a formação da nova molécula de RNA [25].

2.4 Tradução

Tradução é a síntese de um polipeptídio com base em uma molécula de mRNA. É o processo pelo qual uma molécula de mRNA é decodificada nos ribossomos para especificar a síntese de um polímero formado pela união de várias unidades monoméricas de aminoácidos ligados em série (um polipeptídio). A tradução é encerrada quando o códon de parada é lido por uma proteína conhecida como fator de liberação. As subunidades ribossômicas são então separadas [107]. Os polipeptídios são o segundo maior componente dos organismos vivos, ficando atrás apenas das moléculas de água em termos de massa total [44].

Grande parte de nosso genoma não é codificada, representando sequências não codificadoras dos próprios genes, sequências repetitivas não funcionais de DNA, pseudogenes e fragmentos de genes que permaneceram no genoma ao longo do processo evolutivo. Os três principais tipos de RNA celulares produzidos durante a transcrição (RNA mensageiro (mRNA), RNA ri-

bossômico (rRNA) e RNA transportador (tRNA)) atuam em conjunto na tradução para produzir um polipeptídio [83].

Na tradução, a molécula de mRNA é a responsável por direcionar a síntese de uma cadeia polipeptídica, cuja sequência de aminoácidos é determinada pela sequência de nucleotídios presente no mRNA, que, por conseguinte, é derivada da sequência de nucleotídios contida no seu molde de DNA. Por outro lado, as moléculas de tRNA são as responsáveis por reconhecer a sequência de nucleotídios presente no mRNA e correlacioná-la com aminoácidos correspondentes. Por último, todo o processo acontece nos ribossomos, que são formados por cerca de 3 a 5 moléculas de rRNA e 50 a 90 proteínas diferentes, funcionando como grandes fábricas moleculares de proteínas. Os ribossomos são capazes de posicionar corretamente os tRNA com o mRNA, catalisando as ligações entre os aminoácidos que são adicionados ao polipeptídio nascente [64]. As regras que regem o processo de leitura dos nucleotídios na molécula de mRNA e a decodificação dos nucleotídios em aminoácidos compõem o que chamamos de código genético [69].

A informação contida no código genético corresponde ao conjunto de regras que determina como a tradução de uma molécula de mRNA é traduzida em uma proteína. A primeira característica do código genético é que ele é lido em trincas consecutivas, ou seja, cada sequência de três nucleotídeos, chamada de códon, será responsável por especificar um aminoácido [83].

A Figura2 retirada e adaptada do livro Química Orgânica. V.2, de Francis A. Carey[37], mostra o código genético para traduzir cada trinca de nucleotídeos no mRNA em um aminoácido ou indicar o fim da produção de uma nova proteína.

Primeira posição	Segunda posição U	Segunda posição C	Segunda posição A	Segunda posição G	Terceira posição
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Parada	UGA Parada*	A
	UUG Leu	UCG Ser	UAG Parada†	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Sr	U
	AUC Ile	ACC Thr	AAC Asn	AGC Sr	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Figura 2: O código genético (códons de RNA mensageiro) [37].

2.5 Transcrição reversa

A conversão do RNA para o DNA, chamada de transcrição reversa, acontece por causa da atuação da enzima transcriptase reversa que tem grandes aplicações em ensaios de biologia molecular [203]. Em células humanas, o DNA é usado como molde para criar o mRNA. Os

retrovírus, por outro lado, codificam e carregam dentro de seus vírions a enzima transcriptase reversa, que é uma DNA polimerase dependente de RNA. A transcriptase reversa é capaz de transcrever reversamente o genoma de RNA fita simples (ssRNA) em uma fita linear de DNA complementar de fita dupla (cDNA), que é então integrada a um cromossomo da célula hospedeira [122].

2.6 Aminoácidos

Existe um total de 20 aminoácidos que compõem as proteínas, sendo que nove são essenciais e onze são não essenciais. Os seres humanos não têm a capacidade de sintetizar os nove aminoácidos essenciais, portanto, é necessário obtê-los por meio da alimentação [41]. Também são encontrados na natureza mais de 700 aminoácidos não proteicos [132].

Além desses 20 aminoácidos, codificados pelos 61 códons padrão de três bases, existem outros dois que são encontrados em alguns organismos, sendo eles a selenocisteína (sel) vigésimo primeiro aminoácido e a pirrolisina (pyr) vigésimo segundo aminoácido [27]. O selênio é incorporado às proteínas na forma de selenocisteína, em uma versão da cisteína na qual o enxofre é substituído pelo selênio [182]. No caso da pirrolisina, trata-se de uma lisina modificada, em que um anel de pirolina está ligado à extremidade da sua cadeia lateral [130]. Selenocisteína e pirrolisina são inseridas durante a síntese proteica em códons de parada especiais que foram realocados (UGA para a selenocisteína e UAG para a pirrolisina) [52, 51].

A composição molecular de um aminoácido típico geralmente contém um átomo de carbono central (C) que está ligado a um grupo amino (NH2), um átomo de hidrogênio (H), um grupo carboxila (COOH) e uma cadeia lateral (R), conforme mostrado na Figura 3 [97]. Os aminoácidos são classificados como básicos, ácidos, aromáticos, alifáticos ou contendo enxofre, com base na composição e nas propriedades das suas cadeias laterais (R) [167].

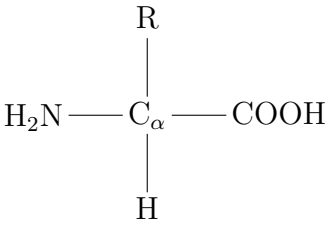

Figura 3: Estrutura geral de um aminoácido

Os aminoácidos são os constituintes fundamentais das proteínas do nosso corpo e servem como substratos para a síntese proteica [49]. A simples disposição dessas letras, que representam a sequência de aminoácidos, contém informações valiosas, que podem ser desvendadas por meio do uso de ferramentas de bioinformática [144].

Com a finalidade de padronizar o uso, a nomenclatura dos aminoácidos também é organizada em abreviações de uma ou três letras. Os nomes dos 20 aminoácidos essenciais que compõem as proteínas, bem como suas abreviaturas estão descritos na Tabela 1.

Tabela 1: 20 Aminoácidos Padrão

Aminoácidos	**Abreviação de 3 letras**	**Abreviação de 1 letra**
Alanina	Ala	A
Isoleucina	Ile	I
Leucina	Leu	L
Metionina	Met	M
Fenilalanina	Phe	F
Valina	Val	V
Prolina	Pro	P
Glicina	Gly	G
Lisina (+)	Lys	K
Histidina (+)	His	H
Arginina (+)	Arg	R
Aspartato (-)	Asp	D
Glutamato (-)	Glu	E
Glutamina	Gln	Q
Asparagina	Asn	N
Serina	Ser	S
Treonina	Thr	T
Tirosina	Tyr	Y
Cisteína	Cys	C
Triptofano	Trp	W

2.7 Proteínas

As proteínas são formados pela ligação do grupo α-carboxila de um aminoácido ao grupo α-amino de outro aminoácido através de uma ligação peptídica [16]. Cada proteína possui uma sequência primária de aminoácidos única que leva a uma estrutura tridimensional específica, o que, por sua vez, influencia diretamente sua função[144].

As proteínas são as macromoléculas biológicas mais abundantes e apresentam enorme diversidade de funções biológicas. A proteína é a expressão da informação genética, a executora de vários tipos de funções biológicas e a sustentadora das atividades metabólicas nos organismos [139]. Em geral, as proteínas que ocorrem naturalmente são compostas pelos 20 aminoácidos mais comuns, cada um com cadeias laterais diferentes. Cada tipo de aminoácido possui propriedades únicas que afetam a estrutura e a função das proteínas, dependendo de suas cadeias laterais [194].

2.7.1 Níveis de informação estrutural

A estrutura de uma proteína é organizada em quatro níveis hierárquicos de complexidade: primária, secundária, terciária e quaternária, conforme mostra um exemplo na figura 4.

[1] `https://www.wwpdb.org/`

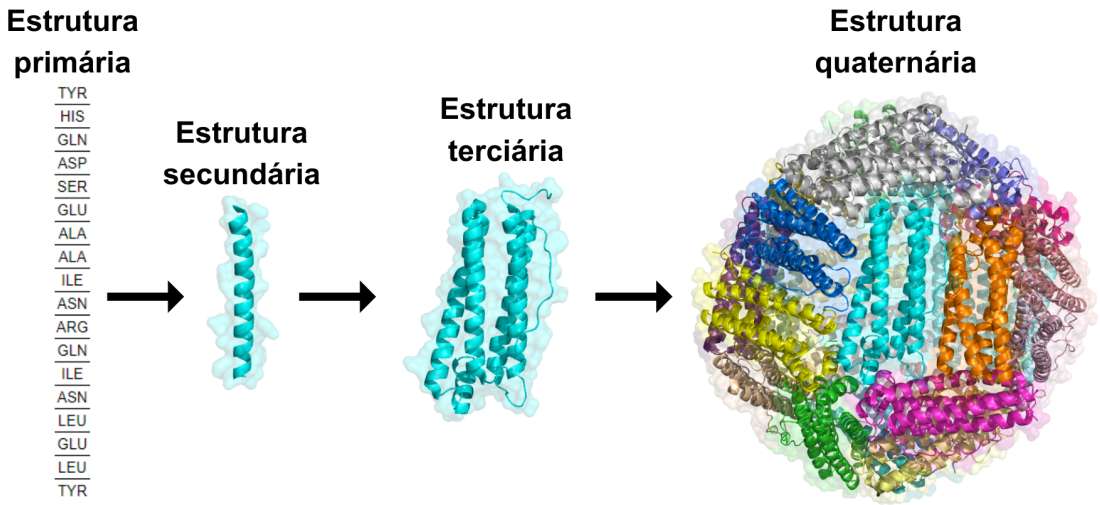

Figura 4: Diferentes níveis de estrutura nas proteínas.

2.7.1.1 Primária

Estrutura primária refere-se à sequência de aminoácidos. Os aminoácidos são unidos por ligações peptídicas para formar cadeias polipeptídicas [39].

O *National Center for Biotechnology Information* (NCBI) Protein[1] é um banco de dados de proteínas que possui 1.338.769.287 sequências de proteínas e 223.775 estruturas determinadas experimentalmente em 21 de agosto de 2024 [173].

2.7.1.2 Secundária

A estrutura secundária descreve a organização espacial dos resíduos de aminoácidos que estão próximos na sequência, as cadeias polipeptídicas podem se enovelar em estruturas regulares que constituem elementos da estrutura secundária, como a *alfa-hélice*, a *folha-beta*, voltas e alças [157]. Prever a estrutura secundária a partir da sequência de aminoácidos continua sendo um desafio significativo [180].

2.7.1.3 Terciária

Na estrutura terciária, as proteínas podem se enovelar em estruturas globulares ou fibrosas. Ela descreve a organização espacial dos resíduos de aminoácidos que estão distantes uns dos outros na sequência, bem como o padrão das pontes dissulfeto. A estrutura terciária é a forma tridimensional geral de um polipeptídio inteiro [21]. Nessa forma a proteína já pode exercer sua função.

[1] https://www.ncbi.nlm.nih.gov/protein

2.7.1.4 Quaternária

A estrutura quaternária diz respeito à organização espacial das subunidades e à natureza de suas interações, nela as cadeias polipeptídicas podem se associar para formar estruturas compostas por múltiplas subunidades [33].

A taxa crescente do número de sequências depositadas nos bancos de dados é muito maior do que a quantidade de estruturas proteicas conhecidas determinadas experimentalmente, conforme consta na página oficial de estatísticas do *Protein Data Bank*[2] (PDB). Ele reúne dados obtidos por cristalografia de raios-X, espectroscopia de ressonância magnética nuclear (RMN), microscopia eletrônica e modelagem integrativa/híbrida submetidos por cientistas do mundo todo [195].

O PDB é o único repositório global de dados experimentais de estruturas tridimensionais de macromoléculas biológicas. Desde 2003, o PDB é gerenciado pelo consórcio *Worldwide Protein Data Bank* (wwPDB) [1], um consórcio internacional que supervisiona de forma colaborativa a submissão, validação, biocuração e disseminação em acesso aberto dos dados de estruturas 3D de macromoléculas [31].

No acesso realizado dia 23 de julho de 2025, constava como número total de entradas disponíveis no PDB 239.521 estruturas e 9.842 estruturas disponibilizadas somente no ano de 2025 até o presente momento [159]. Além dele, o banco de dados *Protein nr* também é extenso e está crescendo rapidamente, possuindo mais de 595 milhões de sequências e 234 bilhões de resíduos. Portanto, precisamos da bioinformática para criar ferramentas computacionais de predição com o intuito de reduzir essa lacuna [95].

2.8 Mutações

Uma mutação é uma alteração na sequência de nucleotídeos, que pode surgir devido a erros nos processos celulares internos, como a replicação do DNA, ou influências externas, como exposição à radiação ionizante ou à luz UV. As mutações ocorrem de diversas maneiras, desde alterações em um único nucleotídeo até movimentações de elementos transponíveis dentro ou entre genomas, além de modificações no número e na estrutura dos cromossomos [80].

As mutações podem alterar a forma ou a composição da interface de interação da proteína, levando à perda ou ao ganho de novos parceiros de interação. As mutações também podem desestabilizar uma proteína, causando alterações na conformação, solubilidade ou outros atributos que determinam a sua função [185]. Uma única alteração em um aminoácido pode ter diversos efeitos, incluindo uma modificação no processo de dobramento, alterações na química de um sítio ativo ou perturbações em redes de ligações iônicas e de hidrogênio, resultando em uma alteração na estabilidade global e na dinâmica da estrutura dobrada [17].

As mutações podem ser classificadas de acordo com a sua causa como: **Mutações espontâneas** que podem se originar naturalmente durante processos celulares, podendo ocorrer

[2] https://www.rcsb.org/

durante a replicação do DNA, a recombinação ou o reparo; **Mutações induzidas** que são causadas por agentes mutagênicos físicos (radiação ultravioleta, por exemplo), químicos (como pesticidas, produtos químicos) ou biológicos [197].

As mutações espontâneas ou induzidas por pressão seletiva, causada por exemplo pelo uso indevido de medicamentos como antibióticos, pode ocasionar a resistência a fármacos que ameaça a longevidade das pessoas, pois restringe as opções de tratamento dos pacientes, sendo um importante problema de saúde pública mundial. Qualquer organismo capaz de evoluir e se diversificar pode desenvolver resistência sob pressão seletiva [113]. Embora as mutações sejam as principais responsáveis pela resistência aos medicamentos, a resistência pode resultar de uma interação complexa de vários fatores [119].

As mutações genéticas podem ser classificadas também dependendo de como alteram o DNA podendo modificar propriedades importantes das proteínas, como conformação, estabilidade, flexibilidade e resistência a medicamentos [197]:

- **Mutações pontuais** causam alterações de um único nucleotídeo do DNA, que podem levar à produção de proteínas não funcionais. Substituições que trocam um aminoácido por outro podem levar a:
  - **Mutações de troca de sentido** (em inglês, *missense*): são desafiadoras e difíceis de compreender, pois alteram apenas um único aminoácido em uma sequência proteica, podendo ter efeitos imperceptíveis, que não variam, até a perda total da função [149].
  - **Mutações sem sentido** (em inglês, *nonsense*): ocorre quando o códon de um aminoácido sofre mutação para formar um códon de parada, então o ribossomo para a leitura da sequência e o resto da proteína não é produzido [14].A cadeia polipeptídica encurtada não consegue se dobrar adequadamente, sendo geralmente detectada pela célula e degradada [53].
- **Mutações de inserção ou deleção**: são adições ou perdas de pares de nucleotídeos em um gene e podem causar mutações com alteração da fase de leitura de uma mensagem genética.
- **Mutações de duplicação:** um segmento de DNA é duplicado e na maioria dos casos a segunda cópia geralmente permanece localizada próxima ou após a cópia original [146].
- **Mutações de translocação:** um segmento de DNA é transferido de seu local original para outra posição na mesma molécula de DNA ou em outra [53].

Neste trabalho o foco é em mutações pontuais missense, principalmente as relacionadas a resistência a fármacos. Nesse contexto, muitas vezes as estruturas das proteínas com mutações pontuais não estão disponíveis em BD de estrutura de proteínas, como o PDB e o

EBI-Alphafold. Assim, é necessário encontrar formas de obter modelos dessas estruturas com mutações para uso em estudos de busca de novos medicamentos e com o intuito de prever o impacto dessa mutação na estrutura das proteínas, sendo necessário o uso de ferramentas de bioinformática capazes de predizer com qualidade e garantia a estrutura tridimensional de uma proteína, sendo dada a sequência com uma mutação pontual.

2.9 Bioinformática

A bioinformática é uma disciplina que estuda os processos e ferramentas necessários para a representação e análise de dados biológicos no nível molecular, utilizando como fonte de dados a sequência de DNA, RNA (genômica) e as proteínas (proteômica), estruturas tridimensionais das moléculas, dados sobre redes biológicas e suas interrelações (metabolômica), dados de expressão gênica, entre outros [123, 54].

Os experimentos das áreas das ômicas podem gerar muitos terabytes de informações e a interpretação desses dados utiliza substancialmente métodos computacionais. Sendo assim, bioinformática corresponde a métodos computacionais de extração de informação útil a partir de conjuntos de dados complexos gerados a partir de experimentos das ômicas (genômica, trascriptômica, proteômica e metabolômica) [13].

Entre as muitas áreas interdisciplinares da bioinformática, a bioinformática estrutural obteve avanços surpreendentes nos últimos anos e está relacionada à análise e predição da estrutura tridimensional de biomacromoléculas. Ela ajuda nossa compreensão sobre os principais processos celulares através da análise de inúmeros bancos de dados, algoritmos e ferramentas, que armazenam, categorizam e interpretam a mensagem biológica [108].

Seus interesses de pesquisa concentraram-se principalmente na análise e previsão das estruturas tridimensionais e funções relacionadas de macromoléculas biológicas, como proteínas, RNA e DNA. Entretanto, ela está cada vez mais diversificada e suas aplicações estão se expandindo para mais campos. Porém, vale ressaltar que os métodos computacionais não se opõem aos experimentais, mas sim os complementam e são incorporados a eles, impulsionando o desenvolvimento de técnicas e promovendo o avanço de métodos mais novos e a serem utilizados [212].

Uma das áreas de atuação da bioinformática estrutural possibilita a realização de pesquisas para a previsão dos efeitos de mutações em proteínas através do uso de várias ferramentas computacionais que utilizam diferentes abordagens para análise [150]. Essa abordagem envolve ferramentas para a predição da estrutura das proteínas que têm mutação. Nesse trabalho, o foco será nas ferramentas de predição de estruturas, aplicadas à predição de estruturas com mutações pontuais.

2.9.1 Bancos de dados de estruturas de proteínas

O principal banco de dados (BD) de estruturas terciárias e quaternárias de proteínas, determinadas experimentalmente, é o *Protein Data Bank*[3][18]. O PDB surgiu em 1971 para armazenar e permitir o compartilhamento padronizado de dados sobre estruturas tridimensionais de moléculas biológicas, incluindo proteínas, ácidos nucleicos e complexos macromoleculares.

O EBI-Alphafold [4] [101] é uma base de dados desenvolvida em parceria com o Instituto Europeu de Bioinformática do Laboratório Europeu de Biologia Molecular (EMBL-EBI, do inglês *European Molecular Biology Laboratory-European Bioinformatics Institute*). Nesse BD, é encontrado de forma gratuita e aberta mais de 200 milhões de predições de estruturas de proteínas que foram geradas pelo AlphaFold (algoritmo baseado em inteligência artificial (IA) desenvolvido pelo Google DeepMind, capaz de prever estruturas de proteínas com alta precisão e velocidade através de métodos computacionais). A DeepMind estima que o banco de dados de estruturas de proteínas do AlphaFold já foi utilizado por mais de 2 milhões de pesquisadores, economizando cumulativamente até 1 bilhão de anos de pesquisa. Em reconhecimento ao impacto do desenvolvimento do AlphaFold2, John Jumper e Demis Hassabis foram agraciados com o Prêmio Nobel de Química de 2024 [70].

2.10 Tuberculose

A tuberculose (TB) é frequentemente considerada uma doença do passado em países desenvolvidos e com alta renda [135]. Atualmente, o desenvolvimento de antibióticos, juntamente com melhorias nos cuidados em saúde e nas condições de vida, contribuiu significativamente para a redução do número de casos [86]. No entanto, em regiões de baixa renda, a TB continua sendo disseminada e é considerada uma das principais causas de morte por doenças infecciosas em todo o mundo [58]. O aumento estimado na incidência de TB entre 2021 e 2023 é amplamente atribuído às interrupções no diagnóstico e no tratamento durante a pandemia de COVID-19, quando o número de novas notificações de casos de TB diminuiu. Infelizmente, acredita-se que essas reduções tenham levado ao aumento do número de pessoas com TB não diagnosticada e não tratada [209]. O Brasil está classificado entre os 30 países com maior carga de tuberculose no mundo, com uma incidência de 36,3 casos por 100.000 habitantes e uma taxa de mortalidade de 2,3 óbitos por 100.000 habitantes em 2022 [191].

Ela é causada por uma bactéria chamada *Mycobacterium tuberculosis*, também conhecida como bacilo de Koch. A doença geralmente afeta os pulmões, mas também pode atingir outras partes do corpo, ou várias partes ao mesmo tempo, e, se não for tratada adequadamente, pode ser fatal. No entanto, nem todas as pessoas infectadas com TB desenvolvem a doença — existem duas condições: infecção latente e TB ativa. Sem tratamento, pessoas com infecção latente, forma inativa, podem desenvolver a forma ativa da doença a qualquer momento e ficar

[3] https://www.rcsb.org/
[4] https://alphafold.ebi.ac.uk/

doentes [77].

Os principais sintomas da TB são: tosse por três semanas ou mais; febre vespertina; sudorese noturna; e perda de peso. A transmissão ocorre pela via respiratória, por meio da liberação de aerossóis produzidos ao tossir, falar ou espirrar por uma pessoa com tuberculose ativa e não tratada. Outras pessoas podem se infectar ao inalar essas partículas. É importante enfatizar que a TB não é transmitida por objetos compartilhados [136].

O diagnóstico da tuberculose pode ser estabelecido por métodos microbiológicos, clínico-radiológicos ou histopatológicos. A baciloscopia de escarro é a técnica mais amplamente utilizada, devido ao seu baixo custo e facilidade de implementação. Ela consiste na coloração do material biológico pelo método de Ziehl-Neelsen, que permite a visualização direta do *M. tuberculosis*, identificado como bacilos álcool-ácido resistentes (BAAR). No entanto, os testes moleculares oferecem maior sensibilidade e especificidade, facilitando e melhorando a precisão diagnóstica. A cultura para *Mycobacterium tuberculosis* continua sendo o padrão-ouro. Além desses métodos, a radiografia de tórax é o exame de imagem mais frequentemente solicitado para avaliar o comprometimento pulmonar em pacientes com suspeita de tuberculose [65]. No entanto, um diagnóstico rápido e preciso da TB é essencial, já que os métodos atuais ainda apresentam limitações [128].

O tratamento medicamentoso da tuberculose dura pelo menos seis meses e tem três objetivos principais: eliminar rapidamente os bacilos, evitar a seleção de cepas resistentes aos medicamentos e eliminar os bacilos persistentes para prevenir recaídas. Desde 2009, o Brasil adotou e recomenda um esquema padrão composto por quatro medicamentos para o tratamento dos casos de TB: rifampicina, isoniazida, pirazinamida e etambutol, utilizados no tratamento da tuberculose sensível a medicamentos em adultos e adolescentes ($\geq$ 10 anos de idade). A TB é curável quando o tratamento é realizado de forma adequada. Nas primeiras semanas de tratamento, o paciente geralmente começa a se sentir melhor, por isso os profissionais de saúde devem orientá-lo e incentivá-lo a seguir o tratamento até o fim, independentemente do desaparecimento dos sintomas. É importante lembrar que o tratamento irregular pode agravar a doença e levar ao desenvolvimento de *M. tuberculosis* resistente a medicamentos [36].

Os medicamentos de primeira linha são fundamentais no tratamento de casos novos de tuberculose e geralmente constituem a base dos esquemas terapêuticos padrão. Eles são frequentemente combinados para aumentar a eficácia do tratamento e reduzir o risco de desenvolvimento de resistência aos medicamentos [188]. Sendo eles:

- **Isoniazida (INH)**;
- **Rifampicina (RIF)**;
- **Pirazinamida (PZA)**;
- **Etambutol (EMB)**

Já os medicamentos de segunda linha para tuberculose são fármacos utilizados no tratamento do *M. tuberculosis* resistente a medicamentos [207]. Sendo eles:

- **Grupo A**
  - Levofloxacino (LFX);
  - Moxifloxacino (MFX);
  - Bedaquilina (BDQ);
  - Linezolida (LZD)
- **Grupo B**
  - Clofazimina (CFZ)
- **Grupo C**
  - Delamanida (DLM);
  - Amicacina (AMK);
  - Estreptomicina (STM);
  - Etionamida (ETO);
  - Protionamida (PTO)

Atualmente, a única vacina licenciada contra a tuberculose (TB) é o Bacilo de Calmette-Guérin (BCG), que é normalmente administrada a recém-nascidos e protege de forma eficaz contra as formas graves da doença. No entanto, sua eficácia contra a tuberculose pulmonar, a forma predominante da doença, é limitada, com uma redução significativa observada cerca de 10 anos após a vacinação infantil, existindo a necessidade de criar uma nova vacina contra a TB com eficácia duradoura [115].

Para combater a tuberculose em nível global, é necessário avançar no desenvolvimento de uma vacina eficaz, implementar intervenções de controle coordenadas, enfrentar os determinantes socioeconômicos e manter altos níveis de apoio político para transformar os avanços científicos em ações concretas [141]. Além disso, a colaboração eficaz entre os serviços de saúde, os sistemas de apoio social e outros setores da comunidade é fundamental para enfrentar esse desafio [153]. Esses pontos são essenciais para alcançar as metas estabelecidas no plano nacional de eliminação da tuberculose até 2030 [206].

2.10.1 Resistência a medicamentos

No cenário atual, as opções de tratamento para pacientes com doenças infecciosas vêm crescendo rapidamente, com novos esquemas totalmente orais e mais curtos representando um avanço. Porém, todo esse progresso está ameaçado pelo aumento da resistência aos fármacos.

Novos compostos químicos estão entrando em ensaios clínicos, aumentando as esperanças de esquemas de tratamento totalmente novos que possam superar as crescentes taxas de resistência aos remédios convencionais [93].

A resistência farmacológica no *M. tuberculosis* surge como resultado de alterações genéticas no genoma bacteriano, que impactam a eficácia dos alvos farmacológicos ou a funcionalidade de enzimas auxiliares. Os polimorfismos de nucleotídeo único (SNPs) representam o tipo mais comum de variação genética detectada. A detecção dessas pequenas alterações na sequência do DNA pode ser facilmente realizada por meio do processo de amplificação, oferecendo um método rápido e preciso para avaliar a resistência [118].

A maioria dos mecanismos de resistência do *M. tuberculosis* a medicamentos conhecidos está relacionada a mutações em genes envolvidos na resistência, incluindo genes que codificam reguladores transcricionais [61].

A prevalência da tuberculose multirresistente (TB-MDR), que refere-se à forma de TB que apresenta resistência tanto à rifampicina quanto à isoniazida, está aumentando em escala global [57]. Estima-se que o *M. tuberculosis* resistente a medicamentos seja responsável por 13% de todas as mortes atribuíveis à resistência antimicrobiana em todo o mundo, sendo impulsionada tanto pela aquisição contínua de resistência quanto pela transmissão de pessoa para pessoa [74]. Garantir o diagnóstico precoce e o acesso oportuno a tratamentos eficazes é crucial para mitigar a disseminação da infecção e impedir o desenvolvimento de resistência aos medicamentos decorrente de terapias incorretas.

2.10.2 Catálogo de mutações do *Mycobacterium tuberculosis* e sua associação com a resistência a medicamentos

O "Catálogo de Mutações do *Mycobacterium tuberculosis* associado à resistência a medicamentos - segunda edição [5]" é um catálogo elaborado com o objetivo de trazer mutações genéticas conhecidas por influenciar a resistência do *M. tuberculosis* à diferentes fármacos. Disponibilizado pela Organização Mundial da Saúde (OMS) junto com instituições de pesquisa internacionais, ele reúne dados provenientes de estudos genômicos e fenotípicos, permitindo uma correlação confiável entre variantes genéticas e perfis de resistência. Esse catálogo tem sido essencial tanto para a vigilância molecular da resistência quanto para a interpretação clínica de testes genéticos, contribuindo para diagnósticos mais precisos e para o direcionamento de terapias mais eficazes [208].

Esse catálogo apresenta tabelas detalhadas que organizam as variantes genéticas de acordo com diversos parâmetros relevantes para a predição de resistência a medicamentos. A seguir, na Figura 5 e na Figura 6 são apresentadadas uma explicação das principais colunas dessas tabelas, com foco na estrutura das colunas, nos critérios utilizados para classificar a confiabilidade das mutações e no significado estatístico dos dados mostrados.

As duas primeiras colunas da tabela referem-se ao medicamento analisado (Drug) e à vari-

[5] https://www.who.int/publications/i/item/9789240082410

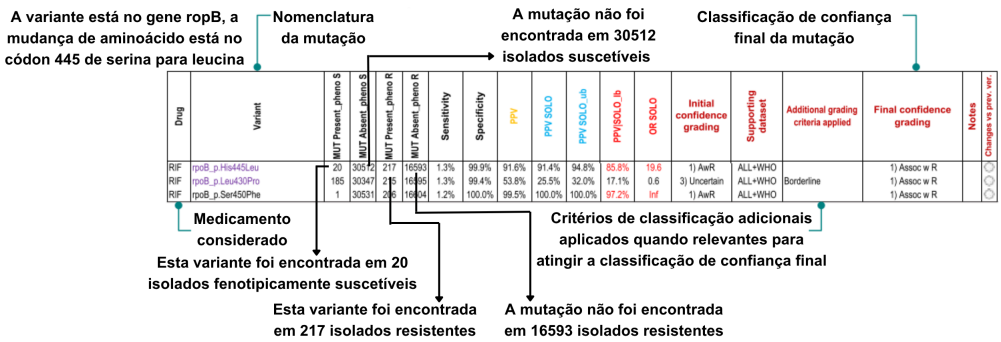

Figura 5: Catálogo de mutações do *Mycobacterium tuberculosis* e sua associação com a resistência a medicamentos.

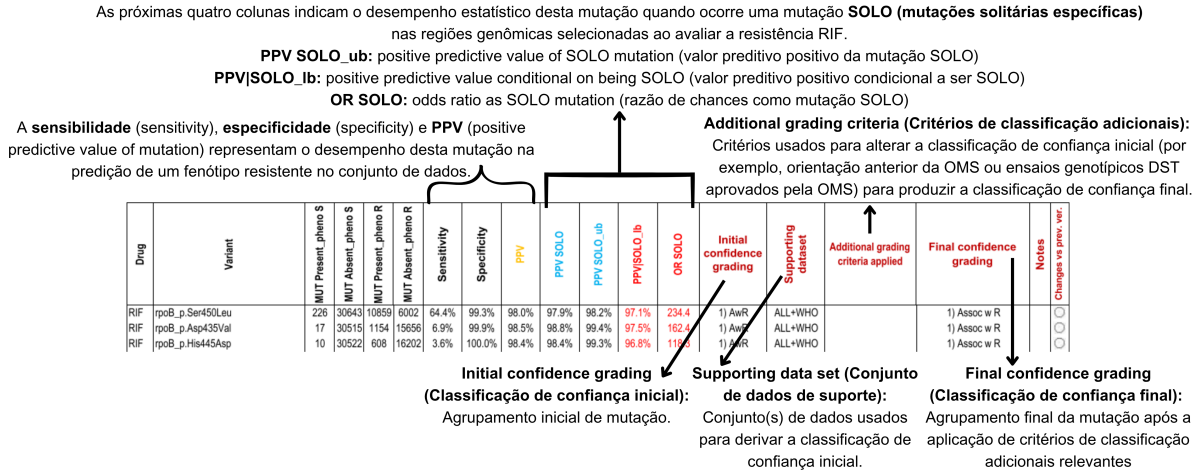

Figura 6: Continuação da explicação sobre o Catálogo de mutações do *Mycobacterium tuberculosis* e sua associação com a resistência a medicamentos.

ante genética identificada (Variant). A coluna do medicamento indica qual fármaco foi considerado na análise de resistência, já a coluna da variante apresenta a mutação específica, descrita com base no gene afetado e na alteração de aminoácido correspondente

As quatro colunas seguintes — `MUT_Present_pheno_S`, `MUT_Present_pheno_R`, `MUT_Absent_pheno_S` e `MUT_Absent_pheno_R` — mostram quantos isolados fenotipicamente suscetíveis (pheno S) ou resistentes (pheno R) apresentaram ou não a mutação em questão.

No exemplo da Figura 5, o medicamento considerado foi a rifampicina (RIF). A variante (mutação) está no gene rpoB, a mudança de aminoácido está no aminoácido 445 (numeração do Complexo *Mycobacterium tuberculosis* (MTBC), do inglês *Mycobacterium Tuberculosis* Complex) e a mudança é de serina (Ser ou S) para leucina (Leu ou L) [200]. Esta variante foi encontrada em 20 isolados fenotipicamente suscetíveis e em 217 isolados resistentes. A mutação não foi encontrada em 30.512 isolados suscetíveis e em 16.593 isolados resistentes.

A seguir temos as colunas referentes a sensibilidade, especificidade e valor preditivo po-

sitivo (PPV, do inglês positive predictive value) representam o desempenho desta mutação na predição de um fenótipo resistente no conjunto de dados. Essas métricas ajudam a avaliar se a mutação é confiável para prever corretamente a resistência do fenótipo. Abaixo, segue uma breve explicação dos principais conceitos utilizados para o cálculo dessas métricas.

- **VP = Verdadeiro Positivo** Casos em que o método acertou ao prever positivo (e realmente era positivo) [198].
- **FN = Falso Negativo:** Casos em que o método errou ao prever negativo, mas na verdade era positivo [62].
- **VN = Verdadeiro Negativo:** Casos em que o método acertou ao prever negativo (e realmente era negativo) [84].
- **FP = Falso Positivo:** Casos em que o método errou ao prever positivo, mas na verdade era negativo [172].
- **Sensibilidade:** Avalia a capacidade do método de detectar com sucesso resultados classificados como positivos [116].

$$\text{sensibilidade} = \frac{VP}{VP + FN} \tag{1}$$

- **Especificidade:** Avalia a capacidade do método de detectar resultados negativos [138].

$$\text{especificidade} = \frac{VN}{VN + FP} \tag{2}$$

- **Precisão:** Avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos [50]:

$$\text{precisão} = \frac{VP}{VP + FP} \tag{3}$$

- **PPV (Positive Predictive Value):** Valor preditivo positivo. Probabilidade de que um caso que foi predito como resistente realmente seja resistente. É uma medida de precisão das previsões positivas [174].

As próximas indicam o desempenho estatístico da mutação quando ocorre uma mutação SOLO nas regiões genômicas selecionadas ao avaliar a resistência a RIF. Os valores fornecidos são o PPV do ponto médio o lb e o ub correspondentes e a razão de chances para a mutação SOLO (OR SOLO), do inglês odds ratio as SOLO mutation. Essas métricas refletem como a mutação se comporta em situações onde ocorre sozinha, ajudando a entender melhor a influência isolada dessa mutação na resistência à rifampicina.

- **PPV SOLO_ub:** Mede a chance de quando uma mutação solo é observada, ela realmente esteja associada à resistência.
- **PPV SOLO_lb:** Calcula a precisão das previsões de resistência somente quando a mutação é analisada em um contexto isolado, sem outras mutações interferindo.
- **OR SOLO:** Mede a associação entre a presença da mutação solo e a resistência à rifampicina.
  - Um valor de OR maior que 1 sugere que a presença da mutação solo está associada a uma maior chance de resistência.
  - Enquanto valores abaixo de 1 sugerem uma menor chance de resistência em comparação com a ausência da mutação.

Uma vez identificadas, por meio do algoritmo SOLO, as variantes associadas e não associadas a fenótipos de resistência, e geradas as estatísticas de associação relevantes, foi aplicado um conjunto de limiares estatísticos consensuais e regras adicionais de classificação para graduar a confiança e ranquear as mutações observadas no *M. tuberculosis* complex (MTBC). Os critérios de classificação foram aplicados de forma igual a todas as mutações para todos os medicamentos.

As variantes presentes no catálogo foram estratificadas em um de cinco grupos de acordo com a quantidade e a qualidade das evidências disponíveis para suportar a associação estatisticamente. A Tabela 2 mostra como são divididos os grupos e qual a interpretação associada a cada um deles.

Tabela 2: Classificação de grupos de mutação com base em sua associação com resistência.

Grupo	**Interpretação**
Grupo 1	Associado à resistência (Assoc c R)
Grupo 2	Associado à resistência – provisório (Assoc c R – provisório)
Grupo 3	Significado incerto
Grupo 4	Não associado à resistência – provisório (Não assoc c R – provisório)
Grupo 5	Não associado à resistência (Não assoc c R)

Abaixo se encontra um diagrama, Figura 7, explicando como funciona essa classificação nesses diferentes grupos.

As variantes dos Grupos 1 e 2 devem ser interpretadas como marcadores de resistência fenotípica clinicamente relevante, ou seja, mutações associadas à resistência fenotípica em uma concentração crítica (CC) reconhecida pela OMS. As variantes dos Grupos 4 e 5 não são marcadores de resistência. Já o papel das mutações do Grupo 3 permanece incerto a partir das evidências disponíveis.

Para entrar em um dos cinco grupos é preciso cumprir os seguintes critérios para classificação:

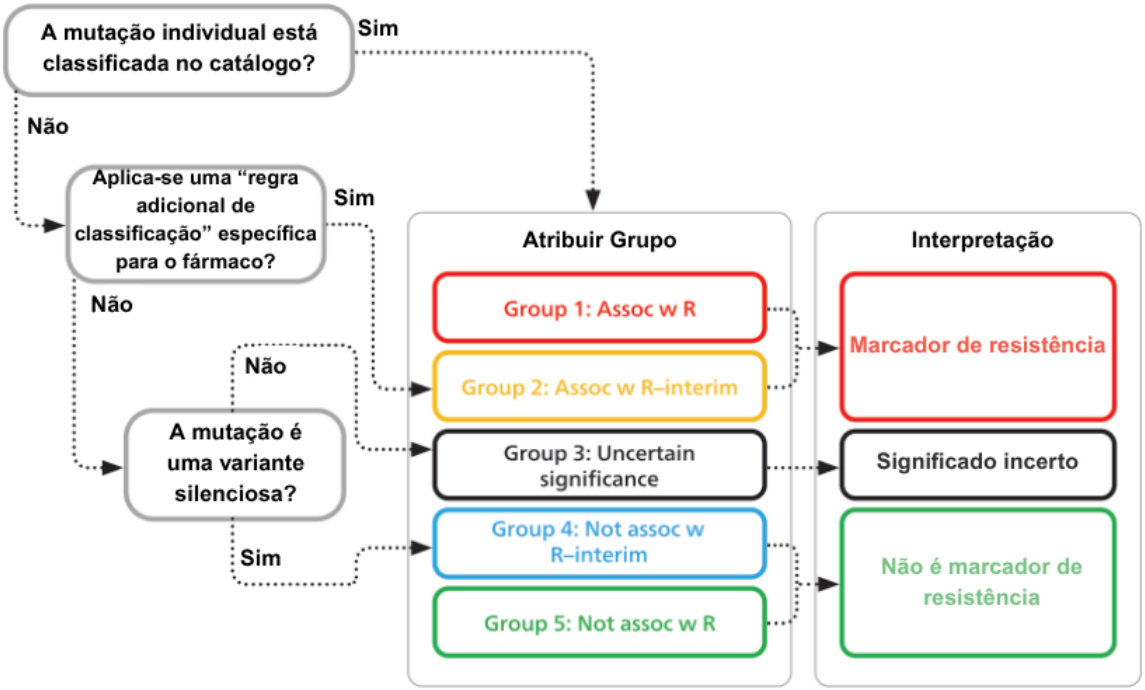

Figura 7: Instruções para uso do catálogo.

- **Grupo 1:** Esse grupo inclui mutações que atendem a cinco critérios, sendo eles:

  1. Soma de isolados resistentes e suscetíveis com mutação SOLO (mutação isolada): Deve ser ≥ 5.
  2. Limite inferior (lb) de 95% do valor preditivo positivo condicional (PPV SOLO_lb) deve ser $\geq 25\%$.
  3. Razão de chances (OR) maior que 1: Isso sugere uma associação positiva com a resistência.
  4. OR SOLO > 1 (se o critério 4 for atendido).
  5. Significância estatística de OR SOLO: Avaliada com o teste exato de Fisher com correção para descoberta falsa.

- **Grupo 2:** Refere-se a mutações no gene pncA que atendem aos critérios "relaxados" de associação com resistência.

  1. Isolados resistentes com mutação SOLO: Deve ser ≥ 2.
  2. PPV $\geq 50\%$.

- **Grupo 3:** Todas as mutações que não atenderam aos critérios de inclusão nos Grupos 1, 2, 4 ou 5.

- **Grupo 4:** Mutações silenciosas não classificadas como neutras nas etapas a e b do algoritmo de neutralidade. Mutações que atenderam aos critérios "relaxados" para *pncA*:

1. PPV SOLO $< 40\%$
2. Limite superior do intervalo de confiança de 95% do PPV SOLO $< 75\%$

- **Grupo 5:** Mutações neutras que foram mascaradas antes da aplicação do algoritmo.

Como as mutações individuais associadas à resistência à PZA são encontradas com menor frequência e estão distribuídas de forma mais ampla entre genes como o pncA, em comparação com outras mutações de resistência, foi necessária uma consideração especial. Assim, aplicamos critérios de classificação "relaxados", com limiares menos rigorosos, para identificar mutações adicionais, porém infrequentes, associadas e não associadas à resistência à PZA apenas no gene pncA.

2.11 Predição de estruturas tridimensionais de proteínas

Existem inúmeras estruturas proteicas que ainda não foram determinadas experimentalmente. Na ausência de informações experimentais, a abordagem computacional para a predição de estruturas tridimensionais de uma proteína a partir de sua sequência de aminoácidos tem sido utilizada [145].

Em 1972, Christian Anfinsen recebeu o Prêmio Nobel de Química ao propor que a estrutura de uma proteína em seu ambiente fisiológico padrão é determinada pela sequência de aminoácidos que a compõe. Essa ideia ficou conhecida como o dogma de Anfinsen, que sugeria a possibilidade de prever a estrutura de uma proteína a partir de sua sequência de aminoácidos [7]. Entretanto, na década de 1960, Cyrus Levinthal demonstrou que há uma grande quantidade de conformações possíveis que uma cadeia proteica poderia assumir. Esse conceito ficou conhecido como paradoxo de Levinthal e foram essas descobertas que estimularam a busca por um método capaz de identificar com precisão a estrutura nativa de uma proteína, apenas com base em sua sequência [133].

Novos estudos estabeleceram que o enovelamento de proteínas pode ser descrito por uma paisagem energética em forma de funil, resultado da evolução de sequências proteicas enoveláveis de acordo com o princípio da frustração mínima, o que permite que as proteínas se enovelem rapidamente em suas conformações nativas biologicamente funcionais [89, 143, 75]. Para uma família de proteínas com um determinado enovelamento funcional, o princípio da frustração mínima sugere que, independentemente da sequência, todas as proteínas dentro dessa família devem se enovelar com taxas semelhantes [196, 205].

Na próxima seção serão descritos os principais métodos de predição de estruturas de proteínas e as respetivas ferramentas.

2.11.1 Métodos de predição de estruturas de proteínas

Os métodos de predição de estrutura de proteínas são tradicionalmente divididos em duas categorias principais: modelagem baseada em molde (*template-based modeling* – TBM) e mo-

delagem livre (*free modeling* – FM). A TBM depende da disponibilidade de estruturas de proteínas homólogas em bancos de dados para modelar a proteína-alvo com base na similaridade estrutural, o que a torna relativamente precisa e computacionalmente eficiente. Em contraste, a FM é utilizada quando não há estruturas homólogas disponíveis, representando um desafio muito maior devido à ausência de moldes evolutivos [211].

Por mais de uma década, métodos baseados em fragmentos, como o Rosetta e o I-TASSER — ambos combinando pequenos fragmentos estruturais com funções estatísticas de energia — foram as abordagens líderes em FM nas competições CASP [215]. No entanto, o surgimento de métodos baseados em aprendizado profundo, especialmente o AlphaFold2 e, mais recentemente, o AlphaFold3, marcou uma mudança transformadora na área. Esses modelos melhoraram drasticamente a precisão das predições, mesmo em casos difíceis de FM, utilizando estruturas de aprendizado de ponta a ponta que não dependem mais do pareamento com moldes nem da montagem por fragmentos.

2.11.2 Métodos experimentais para resolução da estrutura tridimensional de proteínas

Os principais métodos experimentais para resolver as estruturas de proteínas em resolução atômica são a cristalografia por difração de raios-X, a espectroscopia de ressonância magnética nuclear (RMN) e a criomicroscopia eletrônica (cryo-EM) [175]. Apesar de seus avanços indiscutíveis, cada um possui limitações específicas, conforme será discutido abaixo. Devido às dificuldades na determinação experimental das estruturas tridimensionais de proteínas, os métodos computacionais para predição de estruturas proteicas têm ganhado mais popularidade [179].

2.11.3 Cristalografia por Difração de Raios-X

A cristalografia por difração de raios-X é uma das técnicas mais utilizadas para determinar a estrutura tridimensional de macromoléculas biológicas, como proteínas, ácidos nucleicos ou partículas virais. O poder da cristalografia de raios-X é demonstrado pelas estruturas de alta resolução, que nos permitem localizar as cadeias laterais das proteínas em nível atômico, ajudando a elucidar a função e a dinâmica das proteínas [102].

A cristalografia pode fornecer respostas confiáveis para muitas questões relacionadas à estrutura, desde o dobramento global até detalhes atômicos de ligações. A estrutura tridimensional oferece informações detalhadas sobre a posição dos átomos, interações atômicas específicas, além de indícios sobre a flexibilidade da molécula. Ela também pode fornecer *insights* sobre os centros de sítios ativos e os mecanismos de reação de enzimas, mudanças conformacionais que ocorrem após a ligação de ligantes, efeitos de mutações pontuais na dobra da proteína e suas repercussões na função [181, 24]. Contudo, seu maior desafio é produzir quantidade suficiente de proteína e obter cristais com boa difração [134].

A selenometionina (SeMet) é um análogo da metionina em que o enxofre é substituído por selênio. Na cristalografia de raios X, a SeMet é amplamente utilizada para resolver o problema

de fase, pois o átomo de selênio possui mais elétrons, gerando sinal suficiente para técnicas de faseamento experimental, como dispersão anômala de comprimento de onda único (SAD) e dispersão anômala de múltiplos comprimentos de onda (MAD). A SeMet consolidou-se como uma ferramenta essencial para a determinação estrutural de proteínas, impulsionando avanços na cristalografia macromolecular [142, 140, 193, 156].

A determinação estrutural por essa técnica começa com a formação de cristais da macromolécula de interesse cuja estrutura se deseja determinar. Em seguida, um feixe de raios-X é direcionado ao cristal. Os raios-X interagem com as nuvens eletrônicas dos átomos no cristal, e o arranjo atômico regular e repetitivo dá origem a um padrão complexo de feixes difratados, que são registrados por um detector como pontos (padrão de difração). Esse padrão contém informações sobre as posições de todos os átomos no cristal. No entanto, são necessários cálculos matemáticos para gerar um mapa de densidade eletrônica. Idealmente, os picos no mapa de densidade eletrônica correspondem às posições dos átomos na molécula. Esse mapa é interpretado por meio da construção de um modelo atômico da molécula. Esse modelo é refinado, até que se obtenha um modelo final de alta qualidade [29, 184].

O problema das fases é a questão central no campo da cristalografia, ao calcular mapas de densidade eletrônica a partir de dados de difração. Com o desenvolvimento da predição de estruturas de proteínas, o método de substituição molecular, que se baseia no uso de modelos semelhantes para o cálculo inicial das fases, ganhou mais importância [121, 181]. A substituição molecular é um método amplamente empregado na cristalografia de raios X, mas seu sucesso depende da similaridade estrutural entre o modelo de busca e a proteína previamente cristalizada [12, 79, 104].

2.11.4 Ressonância Nuclear Magnética

A ressonância magnética nuclear (RMN) é, atualmente, um método consolidado em diversas áreas científicas, como física, química, biologia e medicina. A espectroscopia de RMN é uma ferramenta poderosa para interessados em determinar a estrutura tridimensional de biomoléculas, dinâmica e interações de macromoléculas biológicas [103, 32].

Em contraste com a cristalografia de raios X, a RMN não requer amostras cristalinas para a medição. Embora a RMN não seja uma ferramenta ideal para determinar a estrutura de proteínas com tamanho superior a 100 kDa, ela é excelente para estudar a dinâmica das interações proteína-proteína e proteína-ligante [164]. A principal limitação da RMN continua sendo o tamanho das proteínas, por isso a maioria das estruturas depositadas no PDB obtidas por RMN corresponde a proteínas menores ou domínios proteicos isolados [131].

2.11.5 Criomicroscopia eletrônica

A criomicroscopia eletrônica (Crio-EM) é um método que permite visualizar estruturas de proteínas em detalhes atômicos, oferecendo boa informação. Provavelmente, a área de pesquisa em que ela exerce o maior impacto é na caracterização estrutural de proteínas de membrana

[11]. O número de estruturas determinadas por crio-EM está crescendo rapidamente, e espera-se que em breve esse número ultrapasse o da cristalografia de raios-X [120]. Uma limitação importante, no entanto, é que a cryo-EM é aplicável apenas a complexos macromoleculares grandes, deixando muitas proteínas importantes de menor tamanho fora do seu alcance [40].

Essa técnica envolve o congelamento rápido de soluções contendo proteínas ou outras biomoléculas, e em seguida é feita sua exposição a feixes de elétrons para gerar imagens microscópicas de moléculas individuais. Então, essas imagens são utilizadas para reconstruir a forma ou estrutura tridimensional da molécula. O desenvolvimento tecnológico, tanto na captura de imagens quanto nos softwares de processamento, tornou possível obter reconstruções tridimensionais de moléculas [38]. A cryo-EM não requer a cristalização das proteínas e conta com microscópios avançados, além de softwares para transformar as imagens capturadas em estruturas moleculares mais nítidas [34, 43].

2.11.6 Modelagem por homologia

As técnicas de modelagem por homologia são empregadas para criar modelos estruturais tridimensionais de uma proteína alvo a partir de sua sequência de aminoácidos, utilizando como referência uma proteína similar, homóloga, (modelo) cuja estrutura já foi determinada experimentalmente e está depositada em bancos de dados com estrutura tridimensional conhecida.

Geralmente, considera-se um mínimo de 25% de identidade de sequência para que uma sequência seja adequada à modelagem por homologia. As regiões desalinhadas, ou regiões de lacunas (gaps), devem ser modeladas por meio de modelos preditivos, uma vez que ainda não há moldes conhecidos disponíveis [126].

O processo de modelagem por homologia envolve as etapas principais, conforme ilustrado na Figura 8 [106]:

1. Identificação de proteínas evolutivamente relacionadas com estruturas resolvidas experimentalmente que podem ser usadas como modelo(s) para a proteína alvo de interesse.

2. Mapeamento de resíduos correspondentes da sequência alvo e das estrutura(s) modelo por meio de métodos de alinhamento de sequência e ajuste manual.

3. Construção do modelo tridimensional com base no alinhamento.

4. Avaliação da qualidade do modelo resultante. Este procedimento pode ser iterado (refeito) até que um modelo satisfatório seja obtido.

Existem três abordagens principais utilizadas na modelagem por homologia [98]:

Montagem por Corpo Rígido (*Rigid-Body Assembly*): Neste método, as regiões alinhadas entre a proteína-alvo e o molde são tratadas como corpos rígidos. As coordenadas da cadeia principal (*backbone*) da estrutura do molde são copiadas diretamente para construir o núcleo conservado do modelo. A estrutura final é montada combinando essas regiões conservadas com

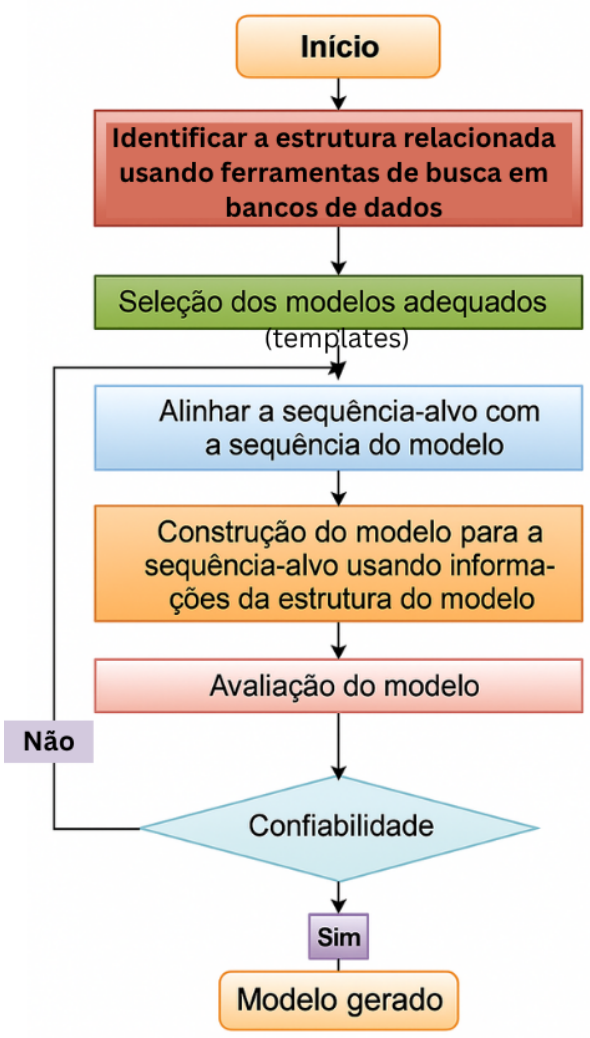

Figura 8: Modelagem por homologia.

alças (loops) e cadeias laterais variáveis. Ferramentas como o *SWISS-MODEL* utilizam essa abordagem [161].

Montagem Baseada em Fragmentos (*Fragment-Based Assembly*): Essa técnica utiliza um subconjunto de coordenadas atômicas — tipicamente os átomos de ($C\alpha$) — como posições-guia. O modelo é construído por meio da busca e do encaixe dos fragmentos estruturais mais adequados a partir de um banco de dados de estruturas conhecidas de proteínas. Essa abordagem oferece maior flexibilidade para modelar variações locais de conformação [92].

Modelagem Baseada em Restrições (*Restraint-Based Modeling*): Nesse método, restrições espaciais são geradas a partir das regiões alinhadas do molde e combinadas com restrições estereoquímicas para a sequência-alvo. O processo de modelagem busca minimizar as violações dessas restrições. O modelo final é aquele que apresenta menor desvio em relação às restrições definidas. O software *MODELLER* é uma implementação amplamente utilizada dessa estratégia [202].

2.11.7 Modelagem *Ab-initio*

Quando não se encontra uma estrutura modelo adequada, é possível utilizar métodos de previsão de estrutura *de novo* (ou *ab-initio*) para criar modelos tridimensionais de proteínas sem a necessidade de ter uma estrutura homóloga como referência [88]. Na modelagem *ab-initio*, geralmente é realizada uma busca conformacional guiada por uma função de energia projetada, gerando várias conformações possíveis, das quais são selecionados os modelos finais [168]. No entanto, é um método computacionalmente complexo e desafiador.

Para ser bem-sucedida, essa modelagem depende de três fatores principais [165]:

1. Uma função de energia precisa, que assegure que a estrutura nativa da proteína seja o estado mais estável termodinamicamente entre todas as possíveis conformações;

2. Um método de busca eficiente, para identificar rapidamente os estados de baixa energia;

3. Uma estratégia eficaz para selecionar modelos quase nativos a partir do conjunto de estruturas chamariz.

Resumidamente, os métodos de predição *ab-initio* visam prever a estrutura tridimensional de uma proteína apenas a partir da informação da sequência primária de aminoácidos, sem o uso de estruturas previamente conhecidas como molde, e utilizando princípios físicos, termodinâmicos e potenciais estatísticos, para tentar identificar a estrutura nativa de uma proteína [100].

Métodos *ab-initio* tradicionais, como os implementados no Rosetta, utilizam montagem por fragmentos combinada com funções de pontuação baseadas em energia para explorar o espaço conformacional. No entanto, são computacionalmente intensivos e geralmente limitados a proteínas pequenas ou de tamanho moderado [199].

2.11.8 Modelagem *Threading*

O *threading*, também conhecido como reconhecimento de dobramento (*fold recognition*), é um método de predição estrutural que identifica dobramentos estruturalmente compatíveis para uma sequência-alvo, mesmo quando a identidade de sequência é baixa ou inexistente. Ele avalia a compatibilidade da sequência-alvo com moldes estruturais existentes por meio de funções de pontuação baseadas em contatos entre resíduos, acessibilidade ao solvente e alinhamento de estruturas secundárias, diferentemente do que ocorre na modelagem por homologia, que depende de uma similaridade significativa entre as sequências [4, 23].

Muitos métodos de modelagem por *threading* buscam superar as abordagens baseadas em sequências ao combinar informações homólogas com diversos tipos de dados estruturais [155]. O *threading* se baseia na ideia de que o número total de dobras distintas na natureza é limitado a algumas milhares. Esse método tenta alinhar a nova sequência à melhor opção entre uma seleção de estruturas de estados nativos possíveis [48]. Uma dificuldade associada ao *threading*

é que, devido a restrições estéricas, pode não ser possível encaixar uma sequência em uma parte de uma estrutura nativa de uma sequência diferente [42].

Os métodos de *threading* têm três aplicações principais na previsão da estrutura de proteínas [56]:

1. Identificar templates de estruturas de proteínas apropriadas para modelar uma proteína alvo;

2. Identificar sequências de proteínas que adotam uma dobra de proteína conhecida;

3. Avaliar a qualidade de um modelo de proteína.

O *threading* atua como uma ponte conceitual e metodológica entre as abordagens de modelagem baseada em molde e modelagem livre. Pois, uma vez identificado um dobramento compatível, a estrutura alinhada é refinada de maneira semelhante à modelagem por homologia.

2.11.9 Modelagem baseada em Inteligência Artificial (IA)

Ferramentas como o ChatGPT[6], Google Gemini[7], Llama[8], DeepSeek[9] e o DALL-E[10] ganharam destaque nos últimos anos, impulsionadas pelos avanços significativos na área de IA. O interesse por essas tecnologias cresce à medida que suas aplicações se expandem, despertando tanto entusiasmo pelas oportunidades quanto preocupações em relação aos impactos sociais. Apesar da atenção recente, o uso da IA para promover descobertas em diferentes áreas da ciência já vem sendo desenvolvido há bastante tempo [160].

Nos últimos anos, tem se popularizado os métodos de predição de estrutura de proteínas que utilizam inteligência artificial, mais especificamente algoritmos de aprendizagem profunda, que tem sido capazes de predizer estruturas com precisão, se aproximando consideravelmente das estruturas obtidas experimentalmente [20, 82]. Softwares de modelagem 3D de proteínas baseados em IA revolucionaram a biologia estrutural, prevendo, na maioria dos casos, estruturas proteicas com um alto nível de confiança [166].

Em geral, os métodos baseados em IA preveem estruturas proteicas aprendendo a partir de muitas sequências e modelos de proteínas conhecidos. No entanto, a precisão da predição é limitada em outras proteínas, como aquelas com poucos homólogos [213]. Estudos recentes estão explorando estratégias de *transfer learning* e *meta-learning* para aprimorar a capacidade de generalização entre diferentes estruturas de proteínas, permitindo a construção de modelos de predição mais eficientes e versáteis. Além disso, a integração de abordagens computacionais

[6]https://chatgpt.com
[7]https://gemini.google.com/app
[8]https://www.llama.com/
[9]https://www.deepseek.com/en
[10]https://openai.com/index/dall-e-3/

e experimentais, aliado com o uso de técnicas interdisciplinares, vêm promovendo avanços significativos e soluções na área de predição de estruturas de proteínas [45].

O surgimento de modelos baseados em aprendizado profundo revolucionou a modelagem. Ferramentas como o AlphaFold2 e o AlphaFold3 utilizam redes neurais treinadas para prever distâncias entre resíduos, orientações e restrições estruturais diretamente a partir dos dados da sequência. Esses modelos alcançaram uma precisão sem precedentes na predição estrutural, mesmo para alvos sem qualquer molde homólogo conhecido, redefinindo efetivamente as capacidades dos métodos *ab-initio* [45].

Um aspecto crucial dos métodos de predição de estrutura de proteínas baseados em IA é a viabilidade prática em relação aos recursos computacionais. O treinamento desses modelos envolve o processamento de grandes volumes de dados biológicos e a otimização de milhões ou bilhões de parâmetros, o que demanda um poder computacional elevado, geralmente fornecido por clusters com múltiplas GPUs de alto desempenho. Isso implica em um consumo significativo de energia e custos operacionais elevados. Além disso, embora a etapa de inferência seja menos exigente do que o treinamento, ela ainda requer recursos substanciais, especialmente ao lidar com proteínas grandes ou em grande quantidade. Essa forte dependência de infraestrutura computacional avançada representa uma barreira importante para o acesso de muitos pesquisadores, laboratórios e instituições ao redor do mundo que não dispõem desses recursos [213, 112].

2.11.10 Predição de estruturas de proteínas com mutações pontuais

A maior disponibilidade de dados de mutação de alta qualidade e os avanços nas abordagens computacionais apoiaram e permitiram o desenvolvimento de várias ferramentas computacionais com o objetivo de compreender como as mutações afetam o enovelamento e a estabilidade das proteínas. Sabe-se que quando a identidade do modelo alvo para modelagem de homologia cai abaixo de 40%, ocorre uma deterioração de desempenho para métodos baseados em estrutura [152]. Além disso, sabemos que a estabilidade da proteína está intimamente ligada à funcionalidade da proteína, assim, a perda da estabilidade da proteína devido a mutações acaba resultando na redução da sua funcionalidade [19]. O AlphaFold2 proporcionou um enorme avanço no campo da biologia estrutural. No entanto, ele possui uma limitação, sendo incapaz de prever os efeitos estruturais de uma sequência de entrada contendo mutações pontuais. Provavelmente isso ocorre, pois não existe um banco de dados para armazenar as mutações e treiná-lo [30].

As redes neurais são extremamente sensíveis à quantidade de dados disponíveis no conjunto de treinamento e a escassez de dados experimentais faz com que a eficácia do aprendizado profundo na tarefa de previsão de $\Delta\Delta G$ seja comprometida. Um estudo não conseguiu identificar uma forma de aplicar os avanços do AlphaFold para solucionar a tarefa de previsão de energia $\Delta\Delta G$ após mutações [151].

2.12 Ferramentas para predição tridimensional de proteínas

2.12.1 trRosetta

Figura 9: Página inicial para utilização via web da ferramenta de predição trRosetta.

O trRosetta[11] (transform-restrained Rosetta) é uma plataforma baseada em aprendizado profundo disponível na web para previsão rápida e precisa da estrutura de proteínas [96]. A entrada da sequência de aminoácidos da proteína, será pesquisada no banco de dados de sequência para gerar um alinhamento de sequência múltipla. Depois disso, a rede neural profunda é usada para prever as geometrias entre resíduos, incluindo distância e orientações. As geometrias previstas são então transformadas em restrições para orientar a previsão da estrutura com base na minimização direta de energia, que é implementada sob a estrutura Rosetta. Em geral, leva aproximadamente 1 hora para prever a estrutura final de uma proteína com aproximadamente 300 aminoácidos [66].

A utilização mais comum do trRosetta é a predição de modelos estruturais para alvos específicos, sendo a modelagem *ab-initio* um dos seus principais recursos. Contudo, para aprimorar a precisão em alvos com estruturas homólogas detectáveis, o servidor inclui automaticamente modelos homólogos. Ao combinar modelagem *ab-initio* com modelagem baseada em modelos, ele funciona bem para uma ampla variedade de alvos.

O servidor web, mostrado na Figura 9 oferece uma interface acessível para a geração de modelos estruturais a partir de sequências de entrada. Ele automatiza a geração de alinhamentos múltiplos de sequências (MSAs), a inferência por redes neurais profundas e a construção da estrutura, com a opção de aplicar um refinamento adicional utilizando o DeepAccNet — um modelo de aprendizado profundo que estima o erro por resíduo e orienta a correção estrutural [85]. Essa integração do aprendizado profundo com o refinamento baseado em física permite

[11] https://yanglab.nankai.edu.cn/trRosetta/

uma previsão estrutural precisa e eficiente, como demonstrado nos testes de referência do CASP [96].

2.12.2 ColabFold - AlphaFold2

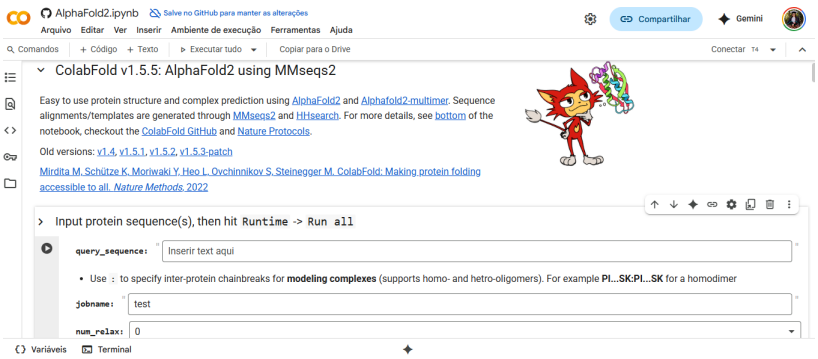

Figura 10: Página inicial para utilização via web da ferramenta de predição ColabFold.

O ColabFold [12] utiliza aprendizado de máquina que integra conhecimento físico e biológico sobre a estrutura das proteínas, ao incorporar alinhamentos de múltiplas sequências no desenvolvimento de seu algoritmo de aprendizado profundo, fornecendo estimativas precisas da confiabilidade de cada resíduo da proteína. Além disso, ele incorpora novas arquiteturas de redes neurais e procedimentos de treinamento que consideram as restrições evolutivas, físicas e geométricas das estruturas proteicas [101]

Ele elimina a necessidade de grandes recursos computacionais ao executar inteiramente na nuvem, tornando a tecnologia de predição de estruturas proteicas amplamente acessível à comunidade científica [137]. Ele é um software de código aberto, rápido e fácil de usar para a previsão de estruturas de proteínas, para uso como Jupyter Notebook dentro do Google Colaboratory, conforme mostra a Figura 10. O ColabFold oferece previsão acelerada de estruturas, combinando a precisão de modelagem do AlphaFold2 [101] com a busca eficiente de sequências do MMseqs2, que é aproximadamente 50 vezes mais rápida [137].

As previsões do ColabFold são classificadas pelo teste de diferença de distância local prevista (pLDDT, do inglês predicted local distance difference test), onde ‘rank_001’ representa a estrutura com maior confiança. Para complexos proteicos, as previsões são classificadas por pontuação de modelagem por molde prevista (pTM, do inglês predicted template modelling). O ColabFold também fornece um gráfico mostrando a pontuação pLDDT para cada posição de aminoácido. Já os PAEs (protein alignment error) de todos os modelos são armazenados como um arquivo png. O arquivo MSA (formato A3M) contém as sequências usadas pelo ColabFold na previsão, podendo ser inspecionado usando um visualizador de alinhamento. A cobertura e diversidade da MSA podem ser examinadas visualizando o gráfico de cobertura (_coverage.png) que mostra o número de sequências no MSA e sua identidade com posições específicas de aminoácidos.

[12] https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

Embora as previsões geradas pelo AlphaFold2 sejam frequentemente precisas, é crucial reconhecer que algumas partes dessas previsões são incompatíveis com os dados experimentais. Em comparação com os modelos depositados no PDB, as previsões do AlphaFold2 não são superiores, pois ele é incapaz de considerar a presença de ligantes, íons, modificações covalentes e condições ambientais, sendo irreal esperar que represente corretamente e completamente os detalhes das estruturas proteicas dependentes desses fatores [192].

2.12.3 Alphafold3

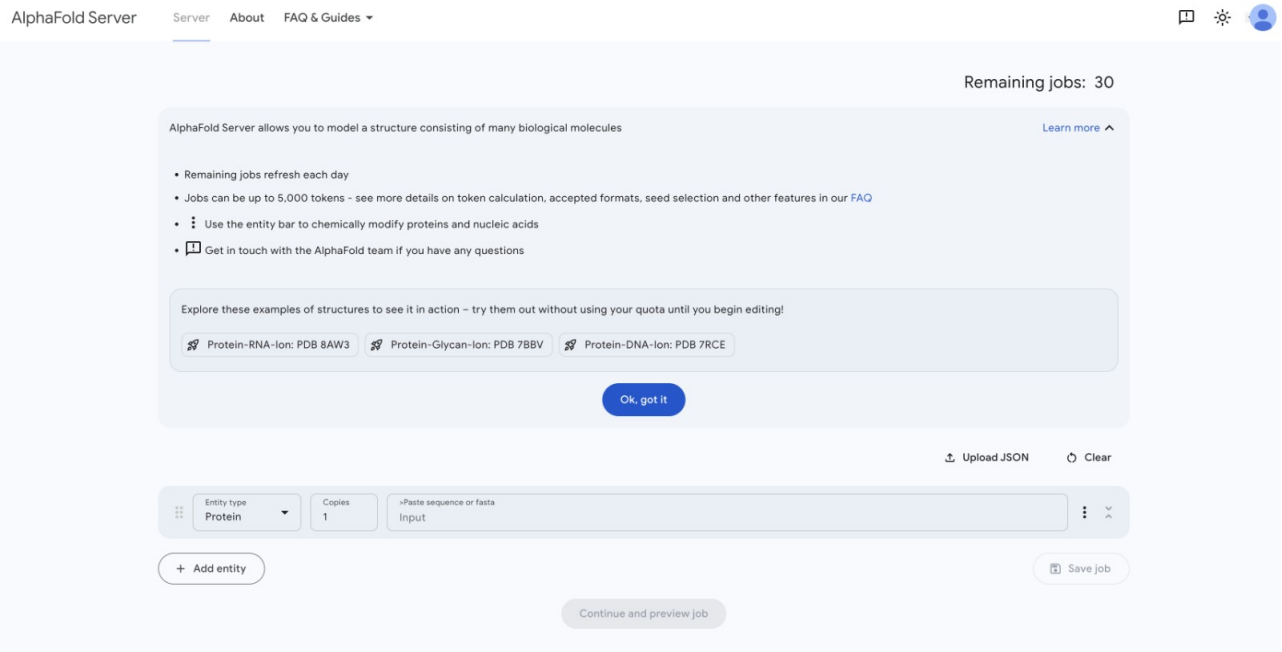

Figura 11: Página inicial para utilização via web da ferramenta de predição AlphaFold3.

O Alphafold3 [13], Figura 11, lançado em maio de 2024 por Abramson *et al.* [1], expandiu suas capacidades muito além das do AlphaFold2. Ele está disponível gratuitamente por meio de um servidor web para usuários de pesquisa não comercial, que pode gerar previsões de estrutura biomolecular altamente precisas, onde podemos submeter 30 trabalhos por dia sem a necessidade de codificação especializada ou possuir uma infraestrutura computacional avançada. Ele melhora a precisão da modelagem do complexo proteico e pode prever muitas biomoléculas, como proteínas, DNA, RNA, bem como moléculas pequenas (ligantes), íons e também modelar modificações químicas [110]. Sua capacidade de modelar simultaneamente diversos tipos de interações torna-o uma ferramenta essencial para a biologia estrutural, descoberta de fármacos e design de macromoléculas [60].

Ele é baseado em uma nova arquitetura fundamentada em difusão, combinada com um módulo no estilo transformer chamado *"Pairformer"*, o AlphaFold3 é capaz de prever conjuntamente, em nível atômico, a estrutura de complexos que incluem proteínas, ácidos nucleicos,

[13] https://alphafoldserver.com/

ligantes de pequenas moléculas, íons e modificações pós-traducionais — tudo dentro de um sistema de aprendizado profundo unificado. Isso representa um salto notável em relação aos métodos anteriores, que exigiam ferramentas especializadas ou separadas para lidar com diferentes tipos de interações biomoleculares.

2.12.4 OmegaFold

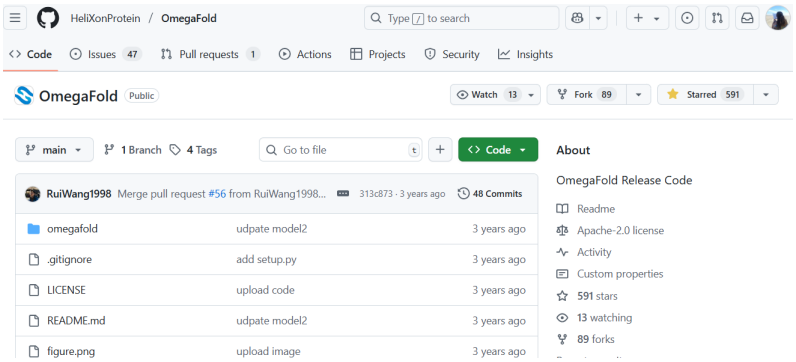

Figura 12: Página do GitHub com orientações para utilização da ferramenta de predição OmegaFold.

O OmegaFold [14] é um método baseado em aprendizado profundo que permite a predição de estruturas proteicas em alta resolução diretamente a partir de uma única sequência primária de aminoácidos, sem a necessidade de alinhamentos múltiplos de sequências (MSAs) ou perfis evolutivos. Introduzido por Wu e Ding em 2022 [210], o OmegaFold combina um modelo de linguagem proteica com uma arquitetura transformer inspirada em geometria, alcançando uma acurácia comparável à do AlphaFold2 ou do RoseTTAFold na ausência de dados de MSA. Sua principal vantagem está na capacidade de modelar de forma confiável proteínas que não possuem homólogos — como proteínas órfãs, anticorpos de rápida evolução ou sequências projetadas de forma de novo — contextos nos quais métodos baseados em informação evolutiva tendem a ter desempenho inferior.

Em estudos de benchmarking realizados com conjuntos de dados como CAMEO e CASP15, o OmegaFold apresentou desempenho robusto, frequentemente rivalizando com métodos de última geração mesmo utilizando apenas uma única sequência como entrada, embora fique ligeiramente atrás dos modelos mais avançados baseados em MSA em termos de precisão final. Ele oferece vantagens consideráveis em velocidade, sendo capaz de prever a estrutura de uma proteína com 500 resíduos em apenas alguns segundos — ordens de grandeza mais rápido do que o AlphaFold2 em condições semelhantes. Sua arquitetura de modelo, com aproximadamente 124 milhões de parâmetros, alcança um equilíbrio entre eficiência e precisão, tornando o OmegaFold acessível a pesquisadores que não dispõem de grandes recursos computacionais.

[14] https://github.com/HeliXonProtein/OmegaFold

Existem diversas publicações que tratam da comparação entre ferramentas de predição de estruturas proteicas, mas geralmente focam principalmente no AlphaFold2 e em ferramentas semelhantes, tornando limitado o número de estudos utilizando o OmegaFold [90, 162].

2.12.5 I-TASSER

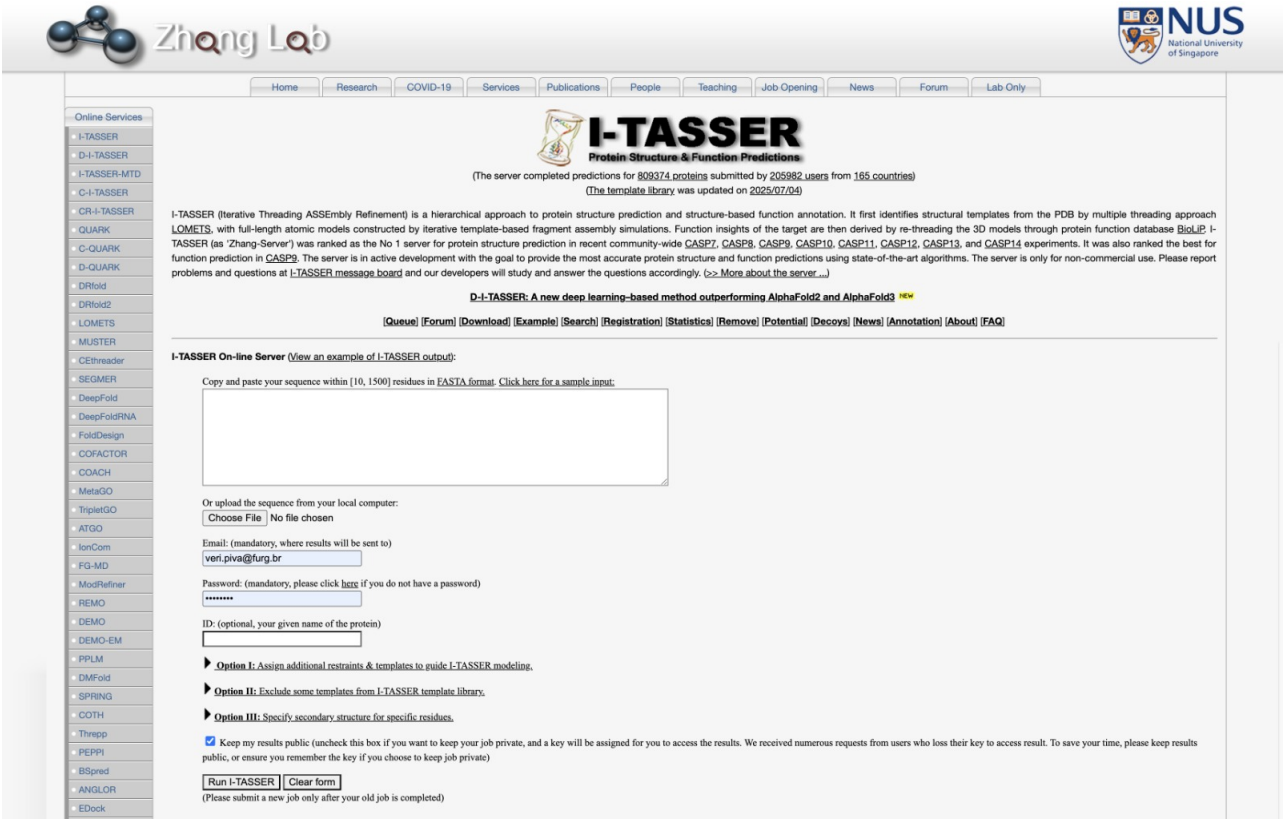

Figura 13: Página inicial para utilização via web da ferramenta de predição I-TASSER.

O I-TASSER [15], Figura 13, é um método hierárquico para predição de estruturas proteicas que refina iterativamente modelos gerados a partir do algoritmo TASSER original. O processo começa com a identificação de modelos homólogos e análogos por meio de alinhamento perfil–perfil. Com base nos escores Z obtidos nesses alinhamentos, as sequências de consulta são classificadas em categorias de modelagem fácil, intermediária ou difícil. A predição da estrutura secundária é realizada utilizando modelos estatísticos específicos de contexto, como o CAS, e a cadeia proteica é dividida em regiões off-lattice (alinhadas a templates) e lattice-based (modeladas *ab-initio*), equilibrando precisão e flexibilidade [218]. A robustez do pipeline do I-TASSER pode ser atribuída às simulações compostas de montagem por fragmentos, que combinam estruturas derivadas tanto de dobramento *ab-initio* quanto do refinamento de templates obtidos por *threading* [214].

A montagem da estrutura terciária no I-TASSER baseia-se no paradigma do ROSETTA,

[15] https://zhanggroup.org/I-TASSER/

mas utiliza fragmentos maiores (com mais de 20 resíduos) e uma busca de Monte Carlo com troca de réplicas mais eficiente. Como resultado, o tempo computacional é reduzido, requerendo cerca de 5 horas de CPU por alvo, em comparação com os aproximadamente 150 dias de CPU exigidos pelo ROSETTA [217].

A principal inovação do I-TASSER está em seu processo de refinamento iterativo. Após a primeira etapa de montagem da estrutura, o método agrupa os modelos decoy usando o SPICKER e extrai restrições de consenso a partir das conformações de baixa energia. Essas restrições são então utilizadas em uma segunda fase de modelagem, juntamente com novos análogos estruturais identificados pelo TM-align, para refinar ainda mais a precisão do modelo. Esse processo melhora significativamente a qualidade estereoquímica: em um conjunto de referência com 200 proteínas, o I-TASSER reduziu o número médio de colisões estéricas de aproximadamente 79 para apenas 0,8 por modelo. Em comparação, o MODELLER produziu uma média de 16,7 colisões sob as mesmas condições [216]. Esses refinamentos permitem que o I-TASSER atinja desempenho de ponta, especialmente em alvos desafiadores que carecem de homólogos claros.

2.12.6 MODELLER

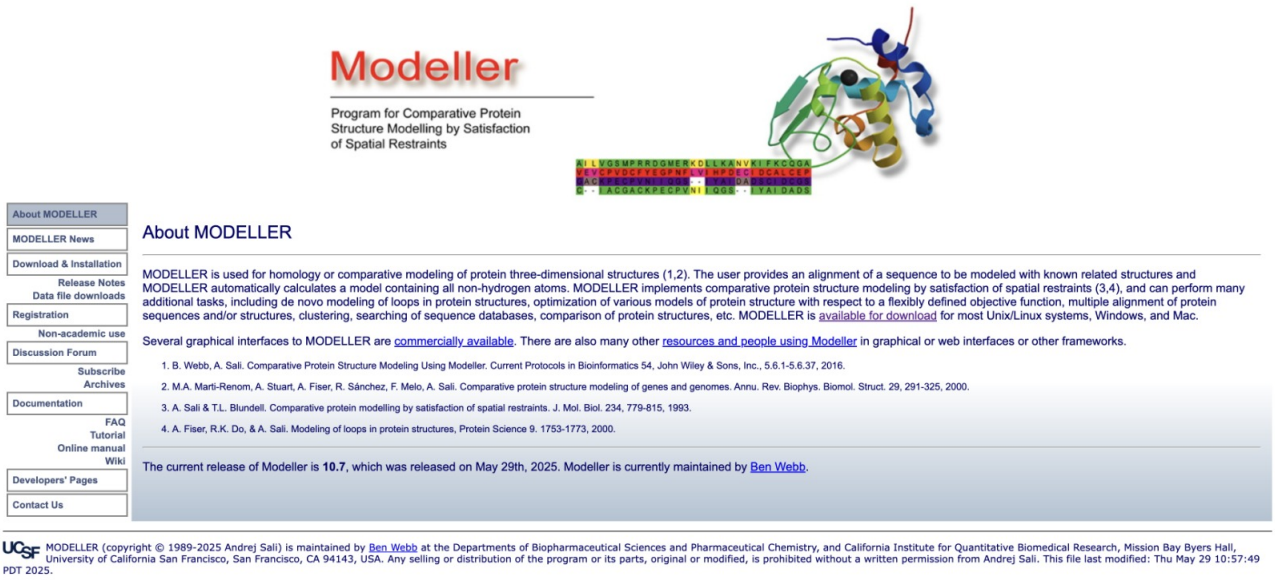

Figura 14: Página inicial com informações sobre a ferramenta de predição Modeller.

O MODELLER [16], Figura14, é um método de modelagem comparativa de estruturas proteicas que constrói modelos atômicos tridimensionais de uma sequência-alvo com base em seu alinhamento a uma ou mais estruturas de modelos homólogos. O método opera segundo o princípio da satisfação de restrições espaciais, no qual características como distâncias interatômicas, comprimentos de ligações, ângulos de ligação e ângulos diedros são representadas

[16] https://salilab.org/modeller/

como restrições probabilísticas derivadas dos templates [169]. Essas restrições são expressas por funções de densidade de probabilidade, combinando parâmetros estereoquímicos físicos de campos de força da mecânica molecular (por exemplo, CHARMM) com informações estatísticas sobre estruturas proteicas [76]. Interações não ligadas e efeitos de solvatação são modelados por meio de potenciais estatísticos de força média [72]. O modelo final é gerado otimizando uma função objetivo que representa a probabilidade de satisfazer todas as restrições espaciais, utilizando técnicas numéricas de otimização como gradientes conjugados, dinâmica molecular e *simulated annealing* [202].

Uma das principais vantagens do MODELLER é sua flexibilidade em incorporar uma ampla variedade de informações estruturais. Além das restrições baseadas em modelos "templates", o programa pode aceitar dados experimentais, como distâncias derivadas de RMN, dados de cross-linking ou mapas de crio-EM. Ele também oferece suporte a restrições definidas pelo usuário, permitindo modelagem integrativa ou híbrida. Para regiões de alça (loops) ou inserções não cobertas pelo "template", o MODELLER utiliza uma busca conformacional baseada em fragmentos, seguida de uma pontuação baseada em energia para modelar com precisão essas regiões variáveis [76]. O software é amplamente utilizado em bioinformática estrutural devido ao seu grau de automação, interface personalizável em Python e desempenho consistente em testes comparativos e tarefas de predição estrutural [72].

2.12.7 Phyre2

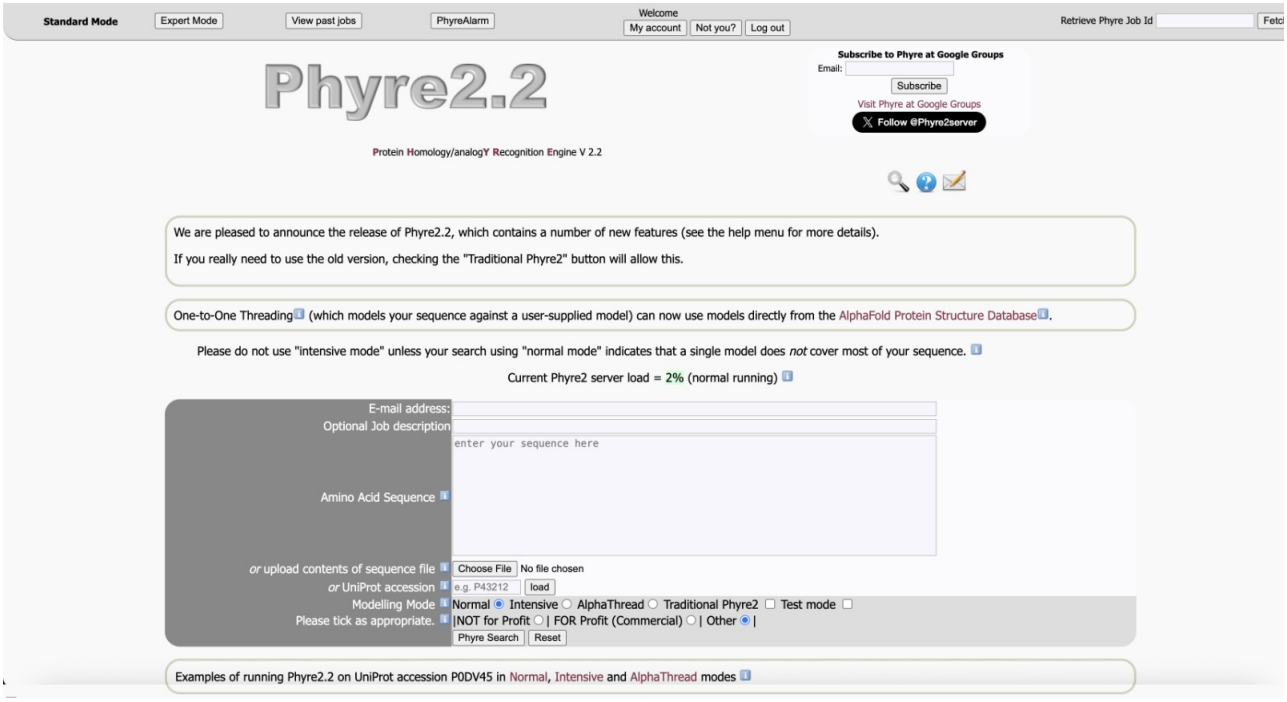

Figura 15: Página inicial para utilização via web da ferramenta de predição Phyre2.

O Phyre2 [17] [105] possui uma interface web simples para o usuário, conforme mostra a Figura15, ele utiliza métodos avançados de detecção de homologia para construir modelos 3D, prever sítios de ligação de ligantes e analisar o efeito de variantes de aminoácidos na sequência proteica fornecida. Uma previsão da estrutura é retornada entre 30 minutos e 2 horas após a submissão na ferramenta. O Phyre2 é uma plataforma amplamente utilizada, construída sobre um cluster de computação compartilhado com aproximadamente 300 núcleos de CPU [105].

Ao invés dele depender de um único algoritmo, o Phyre2 integra diversas ferramentas, possibilitando uma análise versátil com foco principal na modelagem da estrutura tridimensional de sequências individuais de proteínas. O pipeline de predição é dividido em quatro etapas:

1. Primeiro, sequências homólogas são detectadas usando o HHblits [163], e a estrutura secundária é predita com o PSIPRED [99].

2. Em seguida, o HHsearch compara as características preditas com uma biblioteca de HMMs estruturais para identificar modelos adequados [183].

3. Na terceira etapa, inserções e deleções são modeladas usando uma biblioteca de fragmentos e ajustadas por meio do método de descida cíclica das coordenadas (CCD); se necessário, uma etapa opcional *ab-initio* com o Poing é utilizada [94].

4. Por fim, o protocolo R3 é aplicado para otimizar o posicionamento das cadeias laterais.

Uma melhoria significativa ocorreu com a introdução do Phyre2.2 [158], que integra modelos da base de dados AlphaFold, aumentando consideravelmente a precisão das predições para proteínas sem homólogos conhecidos. Além disso, o servidor apresentou uma interface redesenhada para facilitar a identificação de domínios proteicos e uma nova biblioteca abrangente de modelos que mapeia quase todas as sequências UniProt encontradas no Banco de Dados de Proteínas (PDB). Essas atualizações não apenas aprimoram o desempenho da modelagem, mas também tornam a plataforma mais acessível e informativa para qualquer perfil de usuários.

Entretanto, o Phyre2 apresenta duas limitações principais:

1. Ausênsia de homologia entre a sequência fornecida pelo usuário e uma sequência de estrutura conhecida, torna a modelagem impossível ou não confiável.

2. Prever os efeitos estruturais das mutações pontuais. Ele tem funcionalidade para prever o efeito fenotípico de uma mutação pontual, mas é incapaz de determinar com precisão, além da posição estimada de uma cadeia lateral, o efeito estrutural mais amplo de uma mutação pontual.

Nos resultados dos modelos, a coluna confiança (Confidence) não representa a precisão esperada do modelo, mas sim a probabilidade (de 0 a 100) de que a correspondência entre sua sequência e este modelo seja uma homologia verdadeira. Mais de 90% de confiança indica que a proteína é modelada com alta precisão (2–4 Å RMSD da estrutura nativa verdadeira) [189].

[17] http://www.sbg.bio.ic.ac.uk/ phyre2/html/page.cgi?id=index

2.12.8 SWISS-MODEL

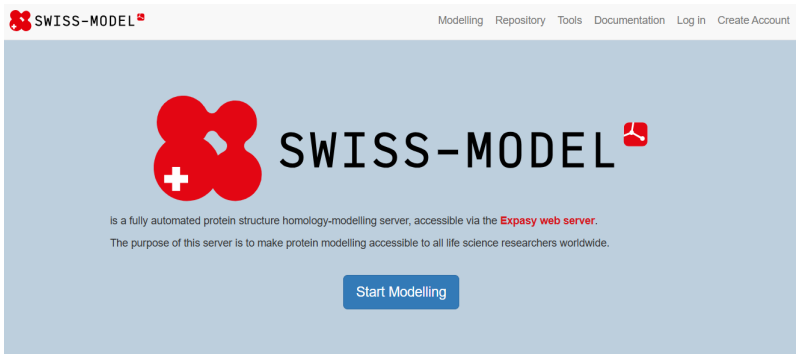

Figura 16: Página inicial para utilização via web da ferramenta de predição SWISS-MODEL.

O SWISS-MODEL[18] [81]é uma plataforma de modelagem acessível *on-line* e foi o primeiro servidor de modelagem de homologia de proteínas totalmente automatizado, sendo continuamente desenvolvido e melhorado [10]. Sua página inicial é apresentada na Figura 16, ele utiliza o mecanismo ProMod3, desenvolvido dentro do framework OpenStructure [187], para identificar modelos estruturais a partir da SWISS-MODEL Template Library (SMTL), com base no alinhamento de sequências. Para regiões que não possuem modelos adequados, o ProMod3 aplica uma amostragem conformacional baseada em Monte Carlo para gerar estruturas plausíveis [201]. As cadeias laterais dos resíduos não conservados são modeladas utilizando uma biblioteca de rotâmeros dependente da cadeia principal (*backbone*), otimizadas por meio do algoritmo *TreePack*, baseado em grafos, que minimiza a função de energia do SCWRL4 [109]. Os modelos finais são refinados usando rotinas de minimização de energia implementadas na biblioteca OpenMM [67], com a parametrização do campo de força CHARMM22/CMAP para maior precisão [125].

Resumidamente, para modelar uma proteína, ele busca em uma biblioteca de estruturas proteicas experimentais para identificar modelos apropriados. Em seguida, é gerado um modelo tridimensional da proteína alvo com base no alinhamento entre a sequência da proteína alvo e a estrutura modelo. Ferramentas de avaliação da qualidade do modelo são utilizadas para estimar a confiabilidade dos modelos obtidos. Normalmente, o processamento leva menos de 1 hora, porém, esse tempo não inclui o período necessário para visualização e interpretação do modelo [26].

Além disso, a interface web do SWISS-MODEL é de fácil acessibilidade, permitindo que os usuários gerenciem múltiplos projetos de modelagem por meio de espaços de trabalho personalizados e oferece três modos de modelagem — automatizado, por alinhamento e por projeto — adaptados a diferentes níveis de experiência dos usuários e à complexidade das modelagens [201, 81]. A plataforma também possui uma ampla documentação e tutoriais, tornando-a adequada para qualquer tipo de usuário, tanto para iniciantes quanto para os mais experientes.

[18]https://swissmodel.expasy.org/

Essas escolhas de design, combinadas com a integração contínua de métodos computacionais de ponta, consolidaram o SWISS-MODEL como um recurso fundamental na bioinformática estrutural [186].

2.13 Métricas para validação da predição de estruturas tridimensionais

O uso combinado das ferramentas de validação permite uma certificação e avaliação robusta, permitindo a análise da integridade estrutural e a identificação de possíveis imprecisões, garantindo uma validação completa e precisa das estruturas proteicas mutantes, contribuindo para a confiabilidade geral dos modelos tridimensionais. Essa validação das estruturas tridimensionais de proteínas pode ser realizada utilizando as ferramentas MolProbity, SAVES (Verify3D e ERRAT), VoroMQA, QMEAN e QMEANDisCo. Há diversos artigos utilizando estas mesmas ferramentas para realizar as validações das estruturas tridimensionais [6, 170, 22, 3]. A seguir essas ferramentas são descritas.

2.13.1 MolProbity

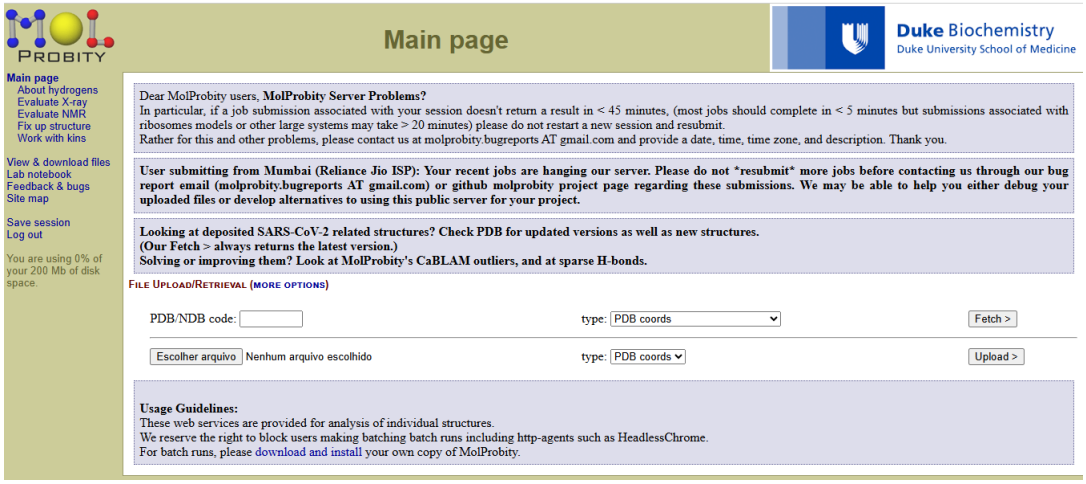

Figura 17: Página inicial para utilização via web da ferramenta de validação MolProbity.

O MolProbity[19] [204], Figura17 , é um kit de ferramentas gratuitas disponíveis para acesso na web desenvolvido pelo Laboratório Richardson, que serve como uma plataforma abrangente para análise estrutural, oferecendo funcionalidades como identificação de conflitos estéricos, validação de estrutura geométrica, análise de gráficos Ramachandran e avaliação de interações de ligações de hidrogênio.

A utilização do MolProbity começa com o usuário carregando um arquivo de coordenadas a partir do seu computador ou buscando um dos bancos de dados no formato PDB ou no formato mmCIF. Os arquivos de coordenadas enviados são processados para fornecer ao usuário um retorno sobre o conteúdo interpretado. Na página principal, há opções de edição, como remover

[19]http://molprobity.biochem.duke.edu/

cópias extras de cadeias, e opções relacionadas ao arquivo, como escolher manter os átomos de hidrogênio (H) do arquivo de entrada em vez de permitir que o programa os otimize. No entanto, a primeira etapa da validação é quase sempre a adição de átomos de hidrogênio, o que é necessário para aproveitar a análise de contatos atômicos completos (all-atom contact analysis). Normalmente, é preciso adicionar e otimizar os átomos de hidrogênio para a maioria das estruturas antes que a validação completa possa ser realizada.

O resultado é apresentado na forma de uma tabela-resumo codificada nas cores verde, amarelo, vermelho, seguida de detalhes apresentados em gráficos e tabelas. As análises incluem: contatos atômicos completos (*all-atom contact analysis*), rotâmeros de cadeias laterais (*side-chain rotamers*), critérios de conformação da cadeia principal baseados no gráfico de Ramachandran, escore MolProbity (MolProbity score), geometria covalente, entre outros [46].

O MolProbity Score é um número que representa as estatísticas centrais de qualidade da proteína. Ele funciona como um indicador único e combinado da qualidade geral do modelo. Esse escore é baseado em uma função ponderada que leva em conta colisões atômicas, resíduos favorecidos no gráfico de Ramachandran e desvios de rotâmeros. O valor é escalonado e normalizado de forma que se aproxime da resolução na qual o modelo seria considerado de qualidade média. Um escore MolProbity mais baixo indica melhor qualidade estrutural [47, 46].

Para o MolProbity score e para o clashscore, a tabela com os resultados finais exibe o percentil correspondente em relação a estruturas de resolução similar. Assim, o MolProbity score e seu percentil oferecem uma maneira rápida e geral para que os usuários finais comparem diferentes entradas da mesma molécula em diferentes resoluções.

2.13.2 SAVES

O SAVeS (*Structure Analysis and Verification Server*)[20], Figura 18, é um metaservidor que fornece uma plataforma comum para realizar avaliação de qualidade com cinco ferramentas diferentes (PROCHECK, WHATCHECK, PROVE, ERRAT e Verify3D) [177]. Ele fornece uma interface de usuário interativa na qual é necessário dar a entrada de uma estrutura de proteína de interesse no formato de arquivo PDB e ele produzirá os resultados dessas ferramentas, Figura 19.

2.13.3 Verify3D

O Verify3D [21] determina a compatibilidade de um modelo atômico (3D) com sua própria sequência de aminoácidos (1D), atribuindo uma classe estrutural com base em sua localização e ambiente (alfa, beta, alça, polar, apolar) e comparando os resultados com estruturas de referência [28, 68].

Ele é um método representativo que compara o ambiente estrutural de um modelo proteico em análise com os perfis estruturais esperados para a proteína, derivados de estruturas

[20] https://saves.mbi.ucla.edu/
[21] https://www.doe-mbi.ucla.edu/verify3d/

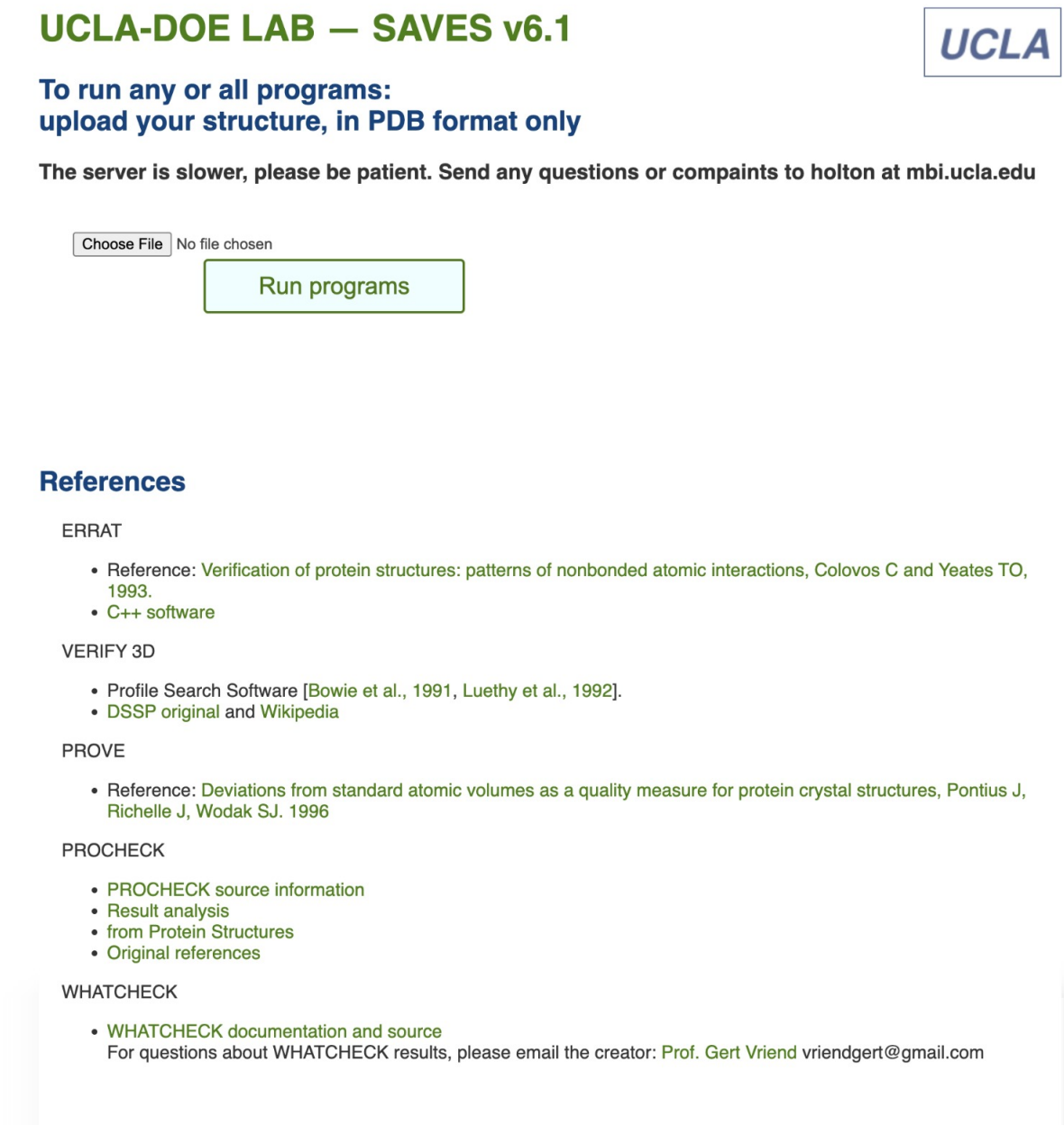

Figura 18: Página inicial para utilização via web da ferramenta de validação SAVES.

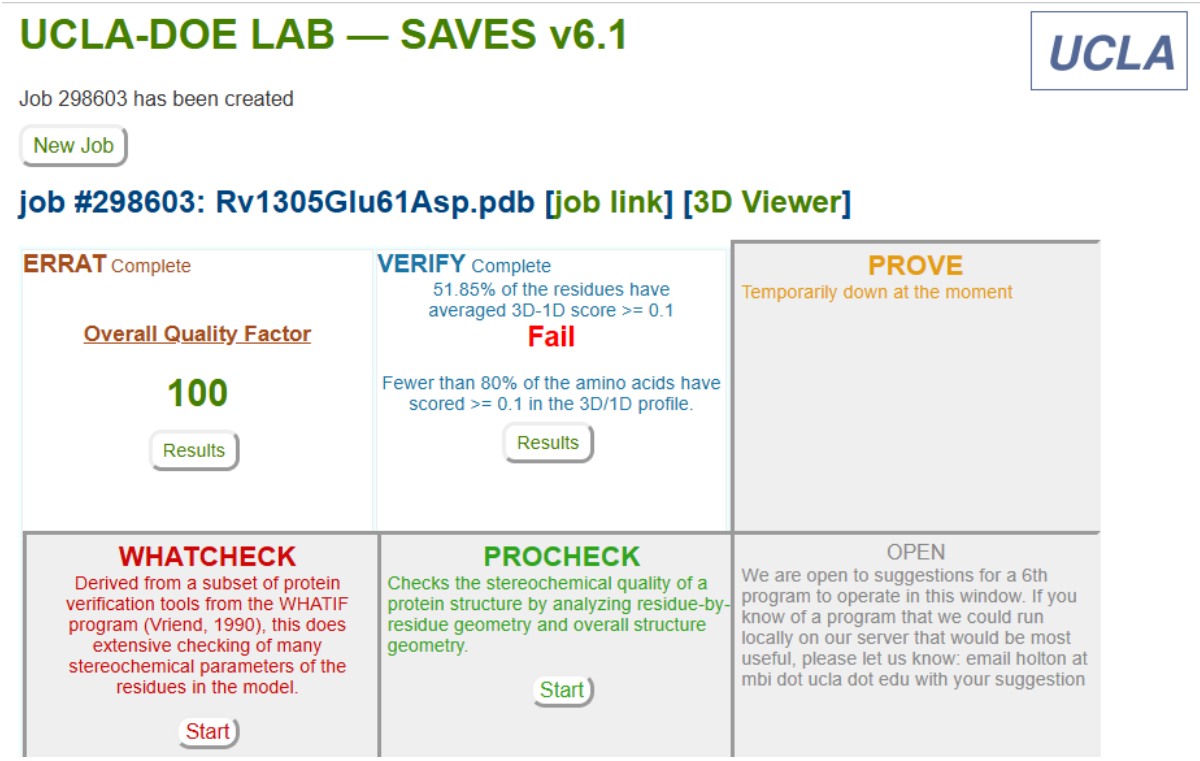

Figura 19: Exemplo de como visualizamos os resultados nas ferramentas ERRAT e Verify3D.

nativas, a fim de predizer a qualidade do modelo. As informações utilizadas para gerar esses perfis estatísticos incluem a estrutura secundária, a acessibilidade ao solvente e a polaridade dos resíduos [171, 35].

O Verify3D gera uma pontuação de perfil 3D para cada resíduo no modelo da proteína. Uma pontuação alta indica que o resíduo é compatível com seu ambiente, enquanto uma pontuação baixa sugere possíveis problemas no modelo [124]. Por exemplo, um modelo de proteína com pontuação Verify3D acima de 0,2 é geralmente considerado um bom modelo, enquanto uma pontuação abaixo de 0,1 pode indicar problemas significativos [117]. Ele utiliza os seguintes critérios como avaliação:

- **Passou (Pass):** Quando o resultado for equivalente a $>=$80% dos aminoácidos com pontuação $>=$0,2 no perfil 3D-1D, o que significa que a estrutura passou com boa qualidade.

- **Falhou (Fail):** Já, quando $< 80\%$ dos aminoácidos obtiveram pontuação $>= 0,2$ no perfil 3D-1D para má qualidade, correspondendo a uma estrutura que apresentou erro e qualidade inferior.

2.13.4 ERRAT

ERRAT [22] é um algoritmo de verificação de estrutura de proteína para avaliar o progresso da construção e refinamento de modelos cristalográficos. Ele pode detectar regiões incorretas de estruturas proteicas de acordo com erros que levam a distribuições aleatórias de átomos, que podem ser distinguidas de distribuições corretas [55, 190].

Ele avalia a estabilidade e a confiança estatística das conformações dos resíduos com base em estruturas de referência [55]. O ERRAT fornece um valor de confiança de erro para a estrutura. Estruturas resolvidas experimentalmente tendem a apresentar valores acima de 90%. As pontuações do ERRAT, significam que quanto mais altas, maior é a qualidade da estrutura. Normalmente, para estruturas de alta resolução, esse resultado fica em torno de 95% ou superior.

Ele também analisa as estatísticas das interações não covalentes entre diferentes tipos de átomos e plota o valor da função de erro em relação à posição de uma janela deslizante de 9 resíduos, calculada por meio da comparação com estatísticas de estruturas altamente refinadas [114].

2.13.5 VoroMQA

Figura 20: Página inicial para utilização via web da ferramenta de validação VoroMQA.

O servidor VoroMQA [23], Figura 20 aplica uma abordagem de avaliação da qualidade de modelos baseada na tesselagem de Voronoi de esferas atômicas, conhecida como “Avaliação da Qualidade de Modelos baseada em Diagramas de Voronoi” (Voronoi diagram-based Model Quality Assessment). Esse método estima a qualidade de estruturas proteicas integrando potenciais estatísticos com informações geométricas derivadas da divisão do espaço atômico. Ele avalia tanto as interações diretas entre átomos quanto os efeitos indiretos das interações com o solvente, utilizando as áreas de superfície de contato. A saída inclui pontuações de qualidade

[22] https://www.doe-mbi.ucla.edu/errat/
[23] https://bioinformatics.lt/wtsam/voromqa

nos níveis atômico, de resíduos e da estrutura como um todo [147].

A maioria das estruturas determinadas experimentalmente com alta qualidade apresenta pontuações VoroMQA superiores a 0,4. Além disso, praticamente nenhuma estrutura nativa possui pontuação VoroMQA inferior a 0,3. Assim, o modelo é classificado como provavelmente ruim se a pontuação for inferior a 0,3; se for superior a 0,4, o modelo é considerado provavelmente bom; se estiver entre 0,3 e 0,4, a qualidade do modelo não pode ser classificada de forma confiável como boa ou ruim utilizando apenas o VoroMQA. Vale salientar que a mesma estrutura proteica sempre irá gerar a mesma pontuação no VoroMQA, uma característica considerada muito importante para a reprodutibilidade do teste [148].

Como entrada, o servidor web VoroMQA aceita uma ou mais estruturas (modelos) nos formatos PDB ou mmCIF. As estruturas fornecidas podem conter múltiplas cadeias, e montagens biológicas também são aceitas. O servidor oferece uma interface de fácil utilização, permitindo aos usuários realizar uma análise detalhada dos resultados de pontuação. Como saída, o servidor fornece pontuações globais, pontuações locais (por resíduo) e informações contextuais adicionais sobre estrutura secundária e acessibilidade ao solvente. Além disso, quando é solicitada a análise de interações entre cadeias, o servidor também fornece pontuações específicas da interface, energias estimadas de interface e pontuações localizadas para os resíduos envolvidos nos contatos entre cadeias [24].

O software VoroMQA para Linux e macOS está disponível também no GitHub [25].

2.13.6 QMEAN

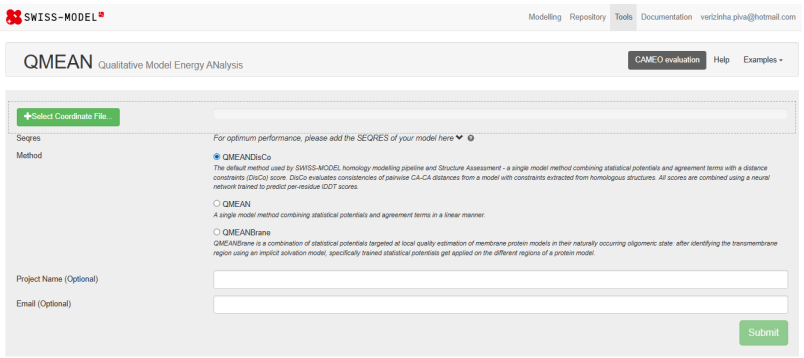

Figura 21: Página inicial para utilização via web da ferramenta de validação QMEANDisCo e QMEAN.

O QMEAN (Quantitative Model Energy Analysis) [26], Figura21, é uma função de pontuação composta que permite obter estimativas absolutas de qualidade tanto globais (para toda a estrutura) quanto locais (por resíduo), com base em um único modelo. Existem dois valores de pontuação global: QMEAN4 e QMEAN6. O QMEAN4 é uma combinação linear de quatro ter-

[24] https://bioinformatics.lt/wtsam/voromqa/help
[25] https://github.com/kliment-olechnovic/voronota
[26] https://swissmodel.expasy.org/qmean

mos de potenciais estatísticos. O QMEAN6, além desses quatro, utiliza dois termos adicionais que avaliam a consistência das características estruturais com predições baseadas na sequência. Ambas as pontuações globais estão originalmente na faixa de [0,1], sendo 1 indicativo de boa qualidade. Por padrão, esses valores são transformados em Z-scores para compará-los com o que seria esperado de estruturas obtidas por cristalografia de raios X de alta resolução — e é isso que é exibido nas páginas de resultado. Caso prefira os valores brutos, eles podem ser obtidos nos arquivos disponíveis para download. As pontuações locais são combinações lineares dos quatro termos estatísticos e dos termos de concordância, calculadas individualmente para cada resíduo. Também estão na faixa de [0,1], sendo 1 indicativo de boa qualidade [15].

O formato de entrada necessário para executar o QMEAN e o QMEANDisCo é o mesmo. Pode-se enviar um modelo no formato PDB ou arquivos compactados no formato .tar.gz contendo múltiplos modelos em PDB que compartilham a mesma sequência de referência (SEQRES) [27].

2.13.7 QMEANDisCo

O QMEANDisCo é uma função de pontuação composta que permite obter estimativas absolutas de qualidade tanto globais (para toda a estrutura) quanto locais (por resíduo), com base em um único modelo. Ele utiliza os mesmos termos individuais do QMEAN como base. A principal melhoria é a inclusão de um novo termo que prevê estimativas locais de qualidade por resíduo, avaliando a concordância entre distâncias par-a-par de resíduos e conjuntos de restrições de distância (DisCo) extraídas de estruturas homólogas ao modelo avaliado.

Os homólogos são identificados com o HHblits. Caso nenhum homólogo seja encontrado, as pontuações DisCo não são utilizadas. Como os resultados tendem a ser razoáveis na ausência dessas pontuações, uma mensagem de aviso é exibida nesses casos. Todos os termos são combinados por meio de redes neurais treinadas para prever a pontuação LDDT por resíduo, em uma escala de [0,1] [129]. A pontuação global do QMEANDisCo é a média das pontuações por resíduo, e a estimativa de erro fornecida é baseada nas pontuações globais do QMEANDisCo calculadas para um grande conjunto de modelos, representando a raiz do erro quadrático médio (ou seja, o desvio padrão) entre a pontuação global do QMEANDisCo e o LDDT (considerado o valor de referência) [201]. Como a confiabilidade da predição depende fortemente do tamanho do modelo, a estimativa de erro fornecida é calculada com base em modelos de tamanho semelhante ao da entrada [186].

[27] https://swissmodel.expasy.org/qmean/help

3 METODOLOGIA

Este capítulo tem como objetivo apresentar em detalhes os procedimentos adotados para a execução deste trabalho, incluindo os processos, a limpeza de dados, as automações e as ferramentas utilizadas. O primeiro passo consiste na obtenção dos arquivos FASTA das proteínas wild-type e mutantes de interesse. Nesse trabalho, devido ao estudo de caso aplicado a TB, nossa fonte de dados foi a tabela com a lista de mutações presente no 'Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance - Second edition' Com essas sequências em mãos, o próximo passo envolve a modelagem, utilizando diferentes ferramentas de modelagem tridimensional de proteínas, utilizando as sequências resultantes da etapa anterior com as mutações. Finalmente, com os modelos PDBs gerados, o último estágio consiste em aplicar esses arquivos em ferramentas de validação, com o objetivo de extrair métricas que possam auxiliar na escolha do melhor gerador de modelos preditivos para proteínas com mutações pontuais.

A Figura 23 mostra um detalhamento da metodologia proposta para essa dissertação, onde foram brevemente descritas as etapas de pré-processamento, predição dos modelos e validação dos modelos. A Figura 22 mostra um fluxograma da metodologia desenvolvida na nossa pesquisa.

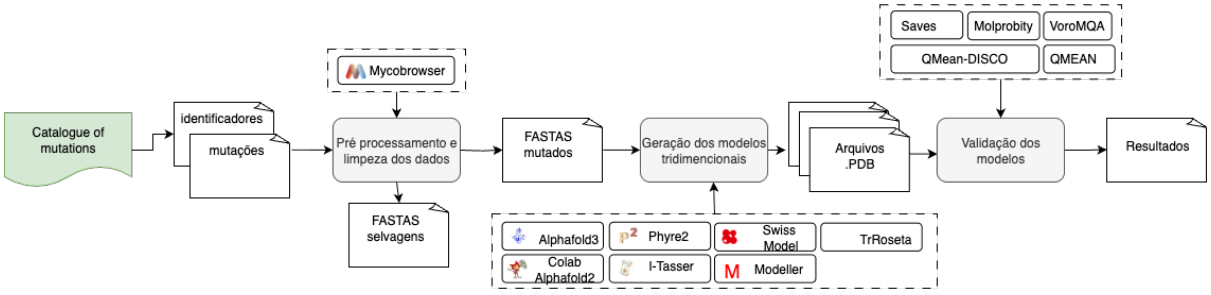

Figura 22: Fluxograma da metodologia realizada no trabalho.

O objetivo principal proposto neste projeto é identificar, com base em diversas métricas provenientes de vários validadores, qual é a ferramenta mais eficaz para a predição de proteínas com mutação pontual *missense*, recomendando assim para os usuários qual a mais indicada

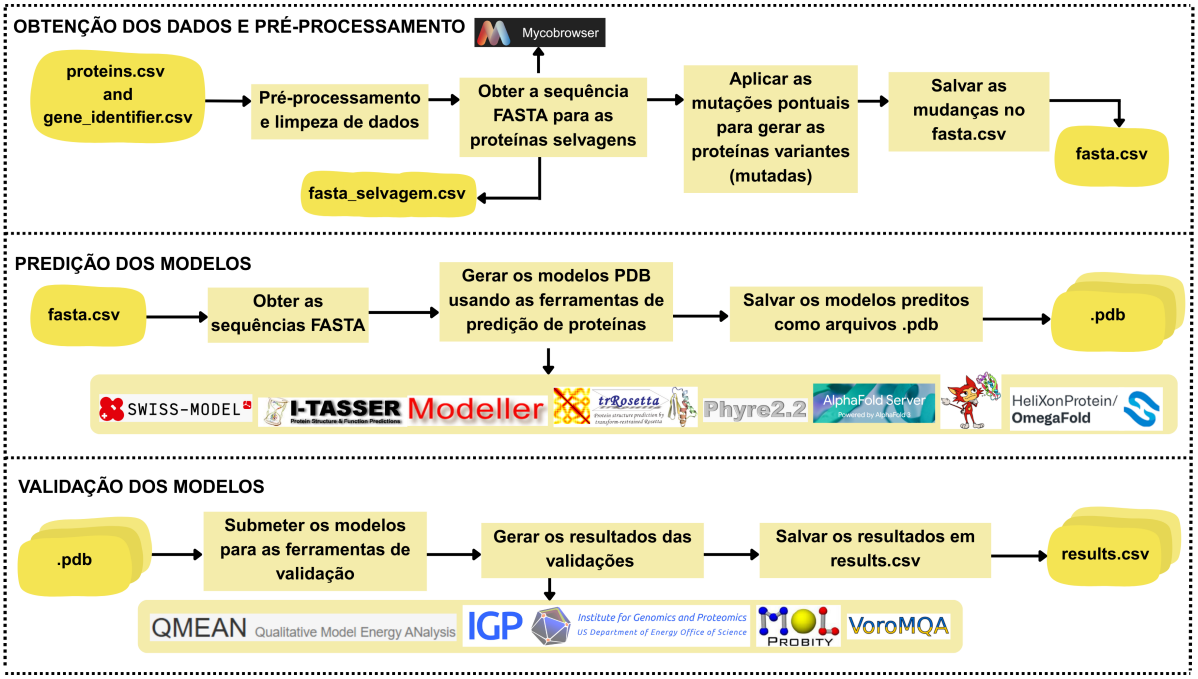

Figura 23: Representação da metodologia realizada no trabalho.

para esse caso, gerando os modelos mais próximos da estrutura real, reduzindo assim a necessidade de realização de métodos experimentais. Além disso, desenvolvemos um site com todas as informações sobre as proteínas analisadas, suas sequências FASTA e os resultados obtidos, promovendo e disseminando o conhecimento adquirido ao longo da pesquisa, seguindo a ideologia de ciência disponível para todos (*open science*).

3.1 Obtenção dos dados e pré-processamento

3.1.1 Dados das mutações

Foi utilizada a tabela com a lista de mutações presente no 'Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance - Second edition', publicado pela Organização Mundial da Saúde (OMS)[1] [208], conforme mostra a Figura 24. Neste livro, encontramos um endereço do repositório no GitHub [2], onde foi disponibilizada uma planilha contendo milhares de mutações de diversos tipos e diferentes informações sobre cada uma delas. Nessa planilha, cada linha representa uma mutação diferente associada a uma droga de tratamento da TB e contém dezenas de atributos diferentes caracterizando essa mutação. É importante mencionar que, para aplicar essa metodologia em outros estudos de caso, a etapa de obtenção e pré-processamento seria adaptada, mantendo as demais etapas conforme proposto.

[1] https://www.who.int/home
[2] https://github.com/GTB-tbsequencing/mutation-catalogue-2023

drug	gene	mutation	variant	tier	effect	genomic position	algorithm_pass	Present_SOLO_SR	Present_SOLO_R	Present_SOLO_S	Present_R	Present_S	Absent_R	Absent_S	Sens	Sens_lb	Sens_ub	Spec	Spec_lb	Spec_ub	PPV	PPV_lb	PPV_ub
Amikacin	bacA	c.102G	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.1044	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	91	1112	2369	20866	3,70%	2,99%	4,52%	94,94%	94,64%	95,23%	7,56%	6,13%	9,21%
Amikacin	bacA	c.105C	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.1065	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	3	2460	21975	0,00%	0,00%	0,15%	99,99%	99,96%	100,00%	0,00%	0,00%	70,76%
Amikacin	bacA	c.1080	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	2	2460	21976	0,00%	0,00%	0,15%	99,99%	99,97%	100,00%	0,00%	0,00%	84,19%
Amikacin	bacA	c.1140	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.1170	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.1194	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.1210	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.1212	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.1212	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.1242	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.129T	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.1317	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	1	0	2459	21978	0,04%	0,00%	0,23%	100,00%	99,98%	100,00%	100,00%	2,50%	100,00%
Amikacin	bacA	c.1323	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.1348	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.135G	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.1371	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	1	2	2459	21976	0,04%	0,00%	0,23%	99,99%	99,97%	100,00%	33,33%	0,84%	90,57%
Amikacin	bacA	c.1374	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%
Amikacin	bacA	c.138C	bacA_c	2	synonymous_variant	(see "Genomi	NA	NA	NA	NA	0	1	2460	21977	0,00%	0,00%	0,15%	100,00%	99,97%	100,00%	0,00%	0,00%	97,50%

Figura 24: Representação de como é a tabela disponibilizada no GitHub e que utilizamos para realizar o estudo de caso proposto para validar a metodologia.

3.1.2 Limpeza dos dados

Uma vez obtidos esses dados, o primeiro passo foi aplicar um processo de limpeza e formatação na tabela, conforme representado na Figura 25. Essa etapa foi essencial para garantir que o conjunto de dados ficasse livre de inconsistências, informações irrelevantes, duplicações e assegurar que os dados estivessem prontos para prosseguir para a análise, ou seja, transformamos os dados para que eles tivessem um formato útil e eficiente.

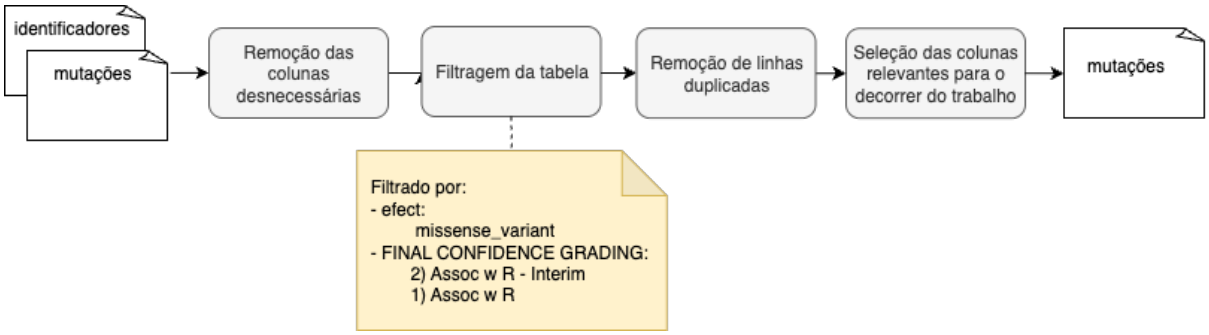

Figura 25: Esquema ilustrativo das etapas realizadas no pré-processamento dos dados para o estudo de caso para a TB a partir dos dados da tabela do "Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance - Second edition".

A fim de reduzir o escopo e trabalhar apenas com os dados necessários, o primeiro passo foi fazer uma seleção dos atributos (colunas), com o objetivo de filtrar apenas as mutações do tipo pontuais *missense* associadas à resistência a fármacos utilizados no tratamento da TB. Além disso, foi feito um mapeamento para cada um dos 65 genes únicos com o seu identificador, informação essa que também estava presente no mesmo livro de onde obtivemos a tabela inicial.

Desse modo, as colunas mantidas foram:

- **gene;**
- **variant;**
- **effect;**
- **FINAL CONFIDENCE GRADING;**
- **identifier;** Nova coluna que foi adicionada para incluir o identificador do gene;

Com isso, reduzimos um total de 114 colunas para apenas cinco.

Possuindo um dataset menor, o nosso próximo passo foi filtrar pelas mutações relevantes para o estudo de caso, ou seja, as que na coluna "effect" possuíam o valor "missense_variant", indicando que são mutações *missense*, que é um tipo de mutação pontual.

Já na coluna "FINAL CONFIDENCE GRADING", nós filtramos pelas linhas que continham os grupos "1 Assoc w R" e "2 Assoc w R -Interim", selecionando as mutações associadas à resistência contra as drogas utilizadas para tratar a tuberculose. Ao finalizarmos esse processo, tivemos uma redução de 48.152 instâncias (linhas) para 411 instâncias. A Figura 26 mostra como ficou a aparência do nosso dataset com os filtros já aplicados.

	gene	variant	effect	FINAL CONFIDENCE GRADING	identifier
2198	atpE	atpE_p.Ala63Pro	missense_variant	2) Assoc w R - Interim	Rv1305
2201	atpE	atpE_p.Asp28Ala	missense_variant	2) Assoc w R - Interim	Rv1305
2202	atpE	atpE_p.Asp28Gly	missense_variant	2) Assoc w R - Interim	Rv1305
2203	atpE	atpE_p.Asp28Val	missense_variant	2) Assoc w R - Interim	Rv1305
2204	atpE	atpE_p.Glu61Asp	missense_variant	2) Assoc w R - Interim	Rv1305

Figura 26: Representação de como ficou a aparência do nosso dataset após a filtragem.

O próximo processo aplicado aos dados foi a remoção de dados duplicados. Como na tabela original haviam sido encontradas as mesmas mutações, porém com resistência a diferentes fármacos, nós ainda tínhamos informações repetidas que foram necessárias remover. Isso causou uma redução para 384 mutações pontuais *missense* únicas.

Por fim, o último passo relacionado à limpeza dos dados foi remover as colunas 'effect' e 'FINAL CONFIDENCE GRADING', uma vez que os valores presentes nelas já haviam sido utilizados na filtragem e não seriam mais úteis para o resto do nosso trabalho. Assim, nossa base de dados inicial passou de 48.152 linhas e 114 colunas para 384 e 3, respectivamente, contendo 15 genes diferentes.

3.2 Obter as sequências das proteínas de tipo selvagem e mutantes

Com os dados limpos, o próximo passo foi conseguir obter as sequências FASTAs para cada um dos genes. Na Figura 27 segue um diagrama com o passo a passo executado a fim de

obter essas sequências.

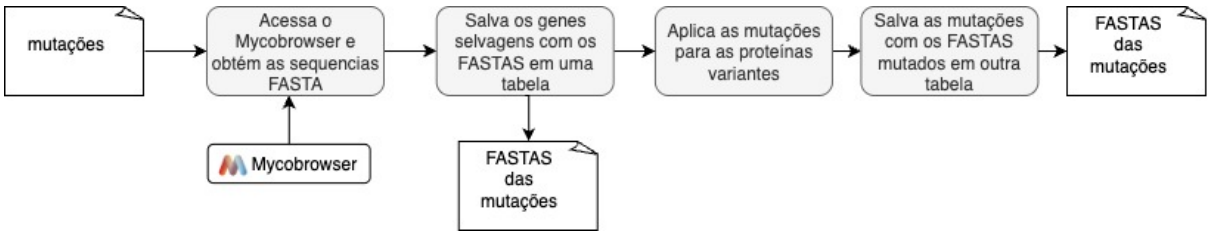

Figura 27: Fluxograma de como foi realizado a obtenção das sequências FASTAs selvagens (wild-type) e mutadas.

O primeiro passo foi buscar a sequência específica para cada proteína, para isso, foi desenvolvido um *script* para realizar o "web scrapping" que acessava o site do Mycobrowser [3], buscava por cada um dos genes, acessava sua sequência FASTA e salvava seu conteúdo. Ao finalizar essa etapa, para maior clareza, foi criada uma outra tabela apenas para as proteínas selvagens, removendo as mutações, mantendo apenas o gene, o identificador, a url do Mycobrowser, e a sequência FASTA. A tabela resultante ficou com 3 colunas - gene, mycobrowser_url e fasta - de 15 genes, conforme mostrado na Figura 28.

	gene	mycobrowser_url	fasta
0	atpE	https://mycobrowser.epfl.ch/genes/Rv1305	MDPTIAAGALIGGGLIMAGGAIGAGIGDGVAGNALISGVARQPEAQ...
1	Rv0678	https://mycobrowser.epfl.ch/genes/Rv0678	VSVNDGVDQMGAEPDIMEFVEQMGGYFESRSLTRLAGRLLGWLLVC...
2	tlyA	https://mycobrowser.epfl.ch/genes/Rv1694	VARRARVDAELVRRGLARSRQQAAELIGAGKVRIDGLPAVKPATAV...
3	ddn	https://mycobrowser.epfl.ch/genes/Rv3547	MPKSPPRFLNSPLSDFFIKWMSRINTWMYRRNDGEGLGGTFQKIPV...
4	embB	https://mycobrowser.epfl.ch/genes/Rv3795	MTQCASRRKSTPNRAILGAFASARGTRWVATIAGLIGFVLSVATPL...

Figura 28: Visualização de como ficou nosso dataset após a obtenção dos FASTAs selvagens através do site Mycobrowser.

Uma vez que o trabalho é voltado às proteínas mutantes e resistentes a medicamentos, além de obter as sequências FASTAs selvagens das proteínas, precisamos também das sequências com as mutações aplicadas. As sequências com as mutações foram geradas alterando o aminoácido pelo aminoácido substituído da mutação por meio de um *script* em *Python*. Para isso, utilizamos nossa primeira tabela resultante do pré-processamento do Catálogo de Mutações para pegar a mutação (variant), extrair a posição na sequência que deve ser alterada e o aminoácido que deve ser substituído, e aplicamos essa alteração nas sequências FASTAs das proteínas selvagens, salvando a sequência FASTA mudada na tabela de mutações. A aplicação dessas etapas de pré-processamento resultou em uma tabela contendo 384 mutações, cada uma relacionada a um gene e a uma sequência FASTA mutada, como pode ser observado na Figura 29.

[3] https://mycobrowser.epfl.ch/

	gene	variant	fasta
0	atpE	atpE_p.Ala63Pro	MDPTIAAGALIGGGLIMAGGAIGAGIGDGVAGNALISGVARQPEAQ...
1	atpE	atpE_p.Asp28Ala	MDPTIAAGALIGGGLIMAGGAIGAGIGAGVAGNALISGVARQPEAQ...
2	atpE	atpE_p.Asp28Gly	MDPTIAAGALIGGGLIMAGGAIGAGIGGGVAGNALISGVARQPEAQ...
3	atpE	atpE_p.Asp28Val	MDPTIAAGALIGGGLIMAGGAIGAGIGVGVAGNALISGVARQPEAQ...
4	atpE	atpE_p.Glu61Asp	MDPTIAAGALIGGGLIMAGGAIGAGIGDGVAGNALISGVARQPEAQ...

Figura 29: Visualização de como ficou nosso dataset após a realizar a substituição do aminoácido da sequência selvagem pelo aminoácido da mutação pontual *missense*.

3.3 Modelar as estruturas 3D das proteínas com mutações pontuais utilizando diferentes algoritmos/ferramentas

Em posse das sequências FASTAs mutadas, seguimos para a etapa de predição da modelagem tridimensional. Para modelar as estruturas 3D das estruturas com mutações pontuais, foram utilizadas as seguintes ferramentas:

- **ColabFold;**
- **Alphafold3;**
- **SWISS-MODEL;**
- **Phyre2;**
- **I-TASSER;**
- **trRosetta;**
- **MODELLER;**
- **OmegaFold;**

Como cada ferramenta funciona de maneira diferente, recebendo como entrada diferentes formatos e possuindo diferentes limitações. Foi necessário realizar uma abordagem única para submeter as sequências para cada uma das ferramentas.

As automações das ferramentas foram desenvolvidas com a ajuda de colaboradores do grupo Combi-Lab [4]. Todos os códigos estão disponíveis para acesso através do GitHub [5].

[4] https://github.com/combilab-furg
[5] https://github.com/combilab-furg/Automation-prediction-tools

3.3.1 ColabFold

Essa é uma ferramenta web, desenvolvida utilizando o Google Colab, na qual a interface gráfica e o próprio código ficam disponíveis no "notebook" em *Python*, tornando possível criar uma cópia do arquivo e alterar conforme necessário. Ele é capaz de predizer apenas uma estrutura por vez, então, para atingir nosso objetivo e executar para todas as nossas sequências, adaptamos o código presente diretamente no Google Colab para receber nossa tabela contendo as sequências e percorrer cada uma das linhas dela de forma automática, submetendo a sequência FASTA. Uma vez que a execução terminava e a predição havia sido concluída, era possível realizar o download do arquivo .zip contendo todos os arquivos, incluindo o .pdb da estrutura gerada.

A imagem abaixo, Figura 30, mostra como é a interface do formulário, onde quem estiver usando a ferramenta deve adicionar a sequência e os demais parâmetros necessários.

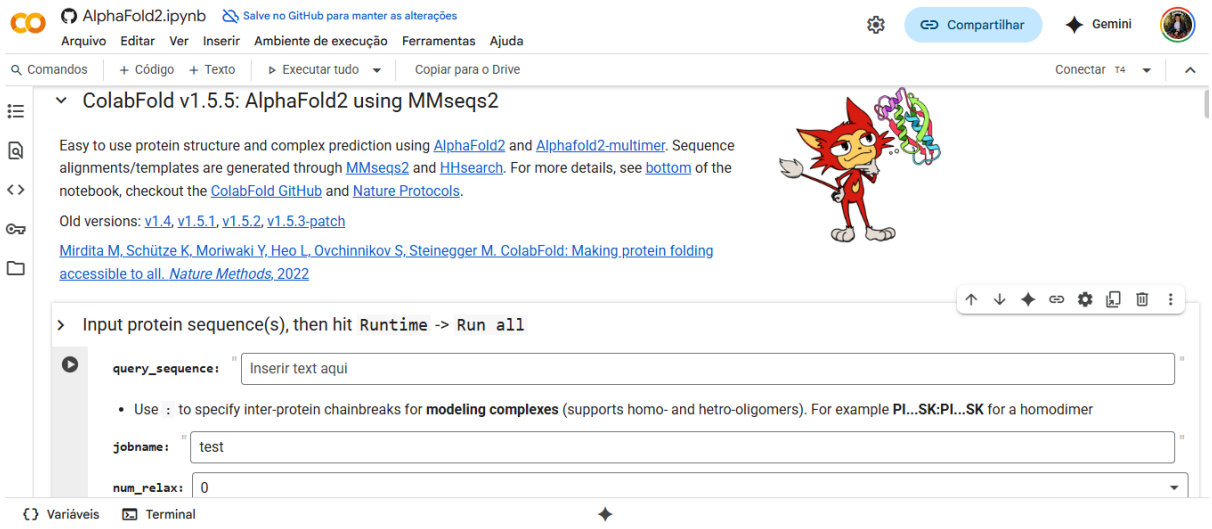

Figura 30: Interface do ColabFold, mostrando o formulário para adicionar a sequencia e preencher outros parâmetros

Para essa ferramenta, optamos por gerar os modelos pelo próprio Google Colab, aproveitando a infraestrutura oferecida e alterando o código localmente. O primeiro passo para automação dessa ferramenta foi compreender o código e os parâmetros disponíveis no *Python* notebook do ColabFold. Criamos uma cópia do Colab Notebook para podermos editar e abrimos cada uma das células de código para compreender como era feito o uso dessa ferramenta e o que precisava ser editado. A próxima imagem, Figura 31, mostra um exemplo de código disponível, no qual aplicamos nossas alterações.

A primeira alteração feita no código foi uma limpeza das variáveis e funções que não faziam sentido para nosso contexto. Como a ferramenta utiliza diversos parâmetros para diferentes funcionalidades e tipos de proteínas, tem muitos trechos de código que não seriam utilizados para gerar o modelo. Alguns condicionais, por exemplo, onde já sabíamos o caminho que seria

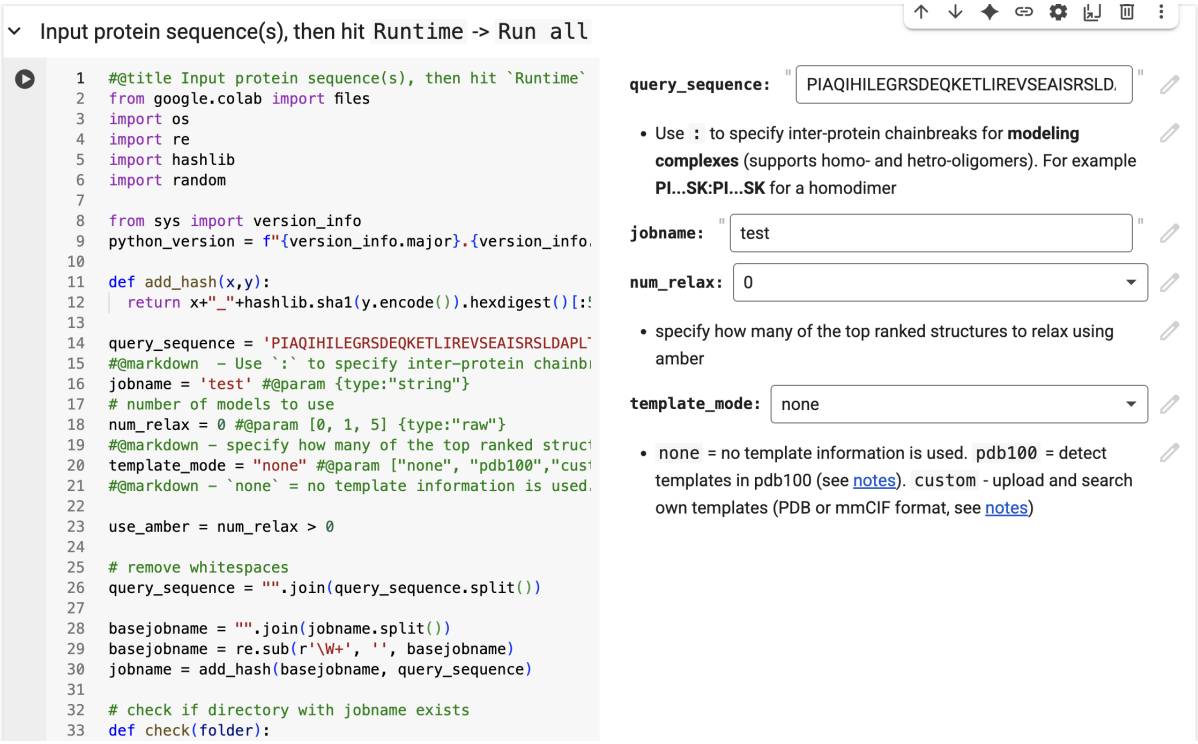

Figura 31: Exemplo de código disponível no Notebook Colab pelo ColabFold

adotado, puderam ser simplificados ou removidos por completo, tornando o código mais simples e direto. Outro ponto, ainda relacionado com a limpeza, foi transformar todas as variáveis que não seriam alteradas em valores constantes, deixando como variável apenas o que seria nossa entrada, ou seja, os valores das sequências fastas.

Como o código da ferramenta está formatado em diversas células distintas, não seria possível executar tudo uma única vez, os códigos estavam espalhados ao longo do Colab Notebook. Para tornar essa automação de rodar apenas uma vez e a partir de uma única célula, a ideia foi criar uma função extra, responsável por executar as demais funções ao invés de chamar cada uma separadamente. Ela recebe como parâmetro uma sequência FASTA e sua execução é o equivalente a executar cada célula de código individualmente de forma sequencial, até o término da execução.

A próxima parte a ser alterada foi utilizar a biblioteca do Pandas para acessar a tabela resultante da etapa do projeto, o pré-processamento e limpeza dos dados, para termos acesso às sequências FASTA que queremos usar como entrada para essa ferramenta. Em posse dessas sequências, bastou criar um laço e iterar sobre cada linha da tabela, extrair o valor da sequência e submeter para o modelo.

Quando o modelo termina sua execução, tem uma função responsável por realizar o download dos resultados em um arquivo .zip.

Abaixo segue um passo a passo de cada ação tomada para automatizar essa ferramenta:

1. Remover variáveis e funções desnecessárias.

2. Transformar variáveis em constantes.
3. Criar uma função responsável por chamar todas as demais de forma sequencial.
4. Adicionar tabela ao Google Drive e carregar ela.
5. Criar um laço para iterar sobre cada elemento da tabela, lendo a sequência FASTA.
6. Chamar a função para submeter cada sequência FASTA para o modelo.
7. Extrair dos resultados apenas os arquivos relevantes, .PDB, e agregar tudo dentro de uma única pasta para facilitar o download.

O código alterado resultou em uma parte inicial onde são definidas as constantes, seguido pelas funções responsáveis pela geração do modelo, uma função responsável por orquestrar tudo isso e fazer o download dos resultados, e uma função para filtrar apenas os arquivos dos modelos .pdb e concatenar tudo em uma única pasta. A Figura 32 mostra um diagrama de como o código funciona.

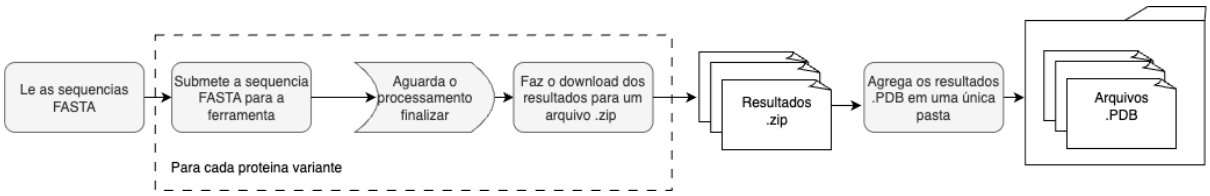

Figura 32: Diagrama do funcionamento do ColabFold após alterações

3.3.2 Alphafold3

O AlphaFold3 é uma ferramenta web disponibilizada para predizer estruturas tridimensionais de proteínas. Ele suporta a submissão de até 30 sequências por dia, mas, anteriormente, era possível submeter apenas 20 por dia. A fim de automatizar esse processo, foi utilizada a automatização dessa ferramenta disponível em um repositório do GitHub [6], que foi desenvolvida em parceria com os alunos de pós-graduação do Combi-Lab, no laboratório de bioinformática localizado no campus da FURG.

Ela aceita múltiplas entradas a partir de uma tabela com o gene, a mutação e a sequência FASTA, o que já está de acordo com a tabela resultante do pré-processamento. Um outro ponto importante de salientar, é que essa mesma automação também é responsável por formatar o arquivo resultante do AlphaFold3, que originalmente possui o formato .cif, mas é convertido para .PDB, que é utilizado nessa pesquisa.

Essa ferramenta de automação possui uma interface intuitiva e fácil de usar. Utilizamos essa interface para submeter nossa lista de sequências FASTA de uma única vez e obter os modelos

[6] https://github.com/combilab-furg/AutoFold

já prontos. Abaixo segue uma imagem, Figura 33, da interface da nossa ferramenta que busca na tabela.

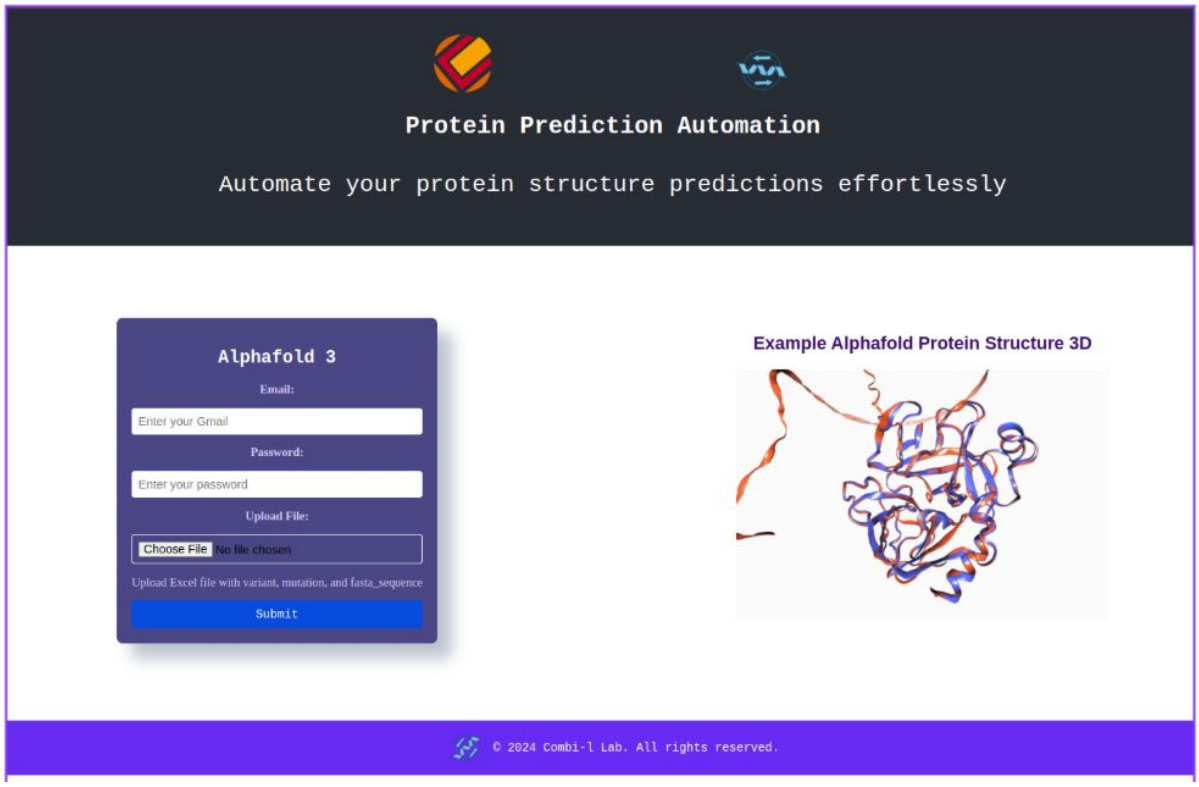

Figura 33: Interface da ferramenta desenvolvida para utilização do AlphaFold3

3.3.3 SWISS-MODEL

O SWISS-MODEL é outra ferramenta web e também possui a limitação de conseguir executar uma submissão única, sem o processamento em lotes. Para contornar isso, utilizamos as APIs [7]disponibilizadas pela própria ferramenta para submeter cada sequência da nossa tabela individualmente e obter um identificador para o processamento, e com esse identificador, baixar os respectivos resultados.

Começamos carregando nossa tabela utilizando novamente a biblioteca do Pandas, e iteramos sobre cada uma das linhas do dataframe, obtendo o valor da sequência FASTA.

Feito isso, fizemos uma requisição para a API disponibilizada pelo próprio SWISS-MODEL para iniciar a execução da ferramenta e começar a geração do modelo e então, uma vez que o processo inteiro ocorre de forma síncrona para essa ferramenta, aguardamos o processamento terminar, e uma vez que temos um modelo gerado, a API retorna um identificador para a execução em questão.

Então, com esse identificador, realizamos uma segunda requisição, para uma outra API, para fazer o download do modelo. Todos esses downloads então são salvos e todos os arquivos

[7]https://swissmodel.expasy.org/api-docs/

.PDB extraídos para uma pasta contendo apenas os arquivos dos modelos.

Abaixo, Figura 34, segue um diagrama simplificado de como funciona a execução da automação do SWISS-MODEL.

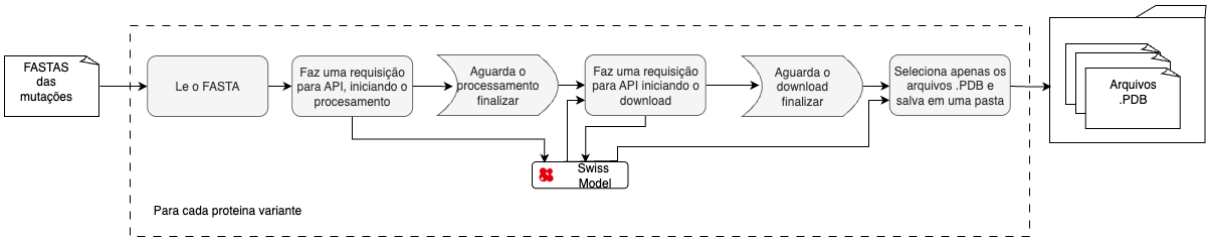

Figura 34: Diagrama do funcionamento da automação da ferramenta SWISS-MODEL

3.3.4 Phyre2

A única ferramenta onde a execução foi feita de forma manual. Também foi utilizada sua versão web e, apesar de possuir uma limitação, ela é de 100 proteínas simultâneas, ou seja, para que fossem gerados os modelos tridimensionais de todas as mutações, foi necessário rodar 4 vezes a ferramenta. A saída também vem no formato de PDB e foi feito um pequeno script para filtrar e separar apenas os arquivos desejados, ou seja, todos aqueles no formato .PDB.

3.3.5 I-TASSER

Apesar de possuir uma versão web para essa ferramenta, ela é muito limitada quando comparada ao software, que foi a versão utilizada para esse trabalho. A versão completa do I-TASSER não possui limitações quanto à quantidade de execuções, esse limite é dado pela capacidade da máquina na qual ele está sendo executado.

No caso dessa pesquisa, foi criado um script para executar apenas 20 paralelamente. Como entrada, é utilizada uma pasta com arquivos contendo as sequências FASTA de cada gene, que será percorrida e processada, gerando o modelo PDB. Para isso, utilizamos os FASTAS presentes na tabela e salvamos em um diretório específico utilizado pela ferramenta.

Abaixo iremos listar a configuração dos parâmetros que utilizamos para rodar o I-TASSER pelo terminal:

- **pkgdir /home/instala/I-TASSER5.2** É a pasta (diretório) do pacote I-TASSER, significa o caminho do pacote I-TASSER. Devemos ter todos os programas neste diretório. O padrão é adivinhar pela localização do script runI-TASSER.pl.
- **libdir /home/instala/ITLIB** É a pasta para arquivos (diretório) da biblioteca I-TASSER, significa o caminho das bibliotecas de modelo.
- **ntemp 1** Número de modelos principais exibidos para cada programa de *threading*; o padrão é 20, com intervalo de [1,50].

- **nmodel 1** Número de modelos finais gerados pelo I-TASSER; o valor padrão é 5, com intervalo de [1,10].

- **seqname fasta_executar** Esse nome deve ser diferente para diferentes destinos para que você possa executar vários trabalhos ao mesmo tempo. Significa o nome exclusivo da sua sequência de consulta.

- **datadir /home/instala/I-TASSER5.2/fastas** Este é o diretório onde sua sequência de entrada "seq.fasta" está localizada. Quando você executa vários jobs, diferentes alvos precisam ser colocados em pastas diferentes.

- **light true** Esta opção significa executar o I-TASSER no modo rápido (cada simulação é executada por padrão por 5 horas no máximo). True ou false, (padrão: false) esta opção executa simulações rápidas.

- **hours 6** Especifique o número máximo de horas de simulações (padrão=5 quando -light=true).

- **traj false** true or false, (padrão: true) vai depositar os arquivos de trajetória.

- **outdir /home/instala/I-TASSER5.2/modelos_gerados** É onde os resultados finais devem ser salvos (o valor padrão é definido como data_dir).

3.3.6 trRosetta

O trRosetta funciona de forma semelhante ao AlphaFold3, contudo, para automatizar a submissão das sequências FASTA foi necessário desenvolver uma ferramenta específica para percorrer a tabela, submetendo cada sequência para a ferramenta, gerando e realizando o download do PDB para cada uma das mutações. Para criar essa ferramenta foram utilizadas bibliotecas do Selenium, uma ferramenta de automação de navegadores compatível com diversas linguagens de programação, como Python. Dessa forma foi possível interagir com a interface web e fazer tudo de forma automática. O primeiro passo foi iterar sobre cada uma das linhas lendo a sequência FASTA e submetendo para o trRosetta. Nessa parte há um limite de apenas 50 projetos rodando simultaneamente por usuário, então apenas 50 eram executados por vez, enquanto os processamentos que já estavam em andamento finalizarem. Enquanto eles estavam sendo executados, uma outra função era responsável por acessar a página de resultados e verificar se o resultado já estava disponível para download e, caso positivo, baixava o arquivo .PDB.

Explicando melhor o passo a passo executado na automação do trRosetta, começamos novamente lendo a tabela com a biblioteca Pandas e acessando o site com a biblioteca Selenium.

Como a ferramenta possui uma limitação na quantidade de execuções em paralelo, verificamos esse limite, e iteramos sobre cada linha, enviando o fasta para a ferramenta, clicamos de

forma automática nos locais necessários para começar a execução, até que o limite seja atingido. Quando o processo de gerar o modelo é iniciado, extraímos a url de onde os resultados serão exibidos e salvamos.

Como essa ferramenta processa de forma assíncrona, ou seja, a gente precisa iniciar a execução e depois verificar de tempo em tempo quando está pronta, adicionamos, após essa etapa onde iniciamos o processamento, uma etapa onde utilizamos as URLs salvas de onde estarão os resultados para verificar os modelos que já foram gerados, e se gerados, remover um da lista de processos em execução, abrindo espaço para outro começar, e salvando o modelo em uma pasta com os .PDBs resultantes.

Como essa ferramenta possui uma limitação e funciona de forma assíncrona, seu funcionamento acabou acontecendo por etapas, onde de tempos em tempos executávamos o programa, ele verificava o limite disponível, iniciava o processamento até atingir o limite, e depois verificava se alguns dos que já tinham sido iniciados anteriormente estavam prontos.

Abaixo podemos ver um diagrama, Figura 35, com o funcionamento mais detalhado da automação do trRosetta.

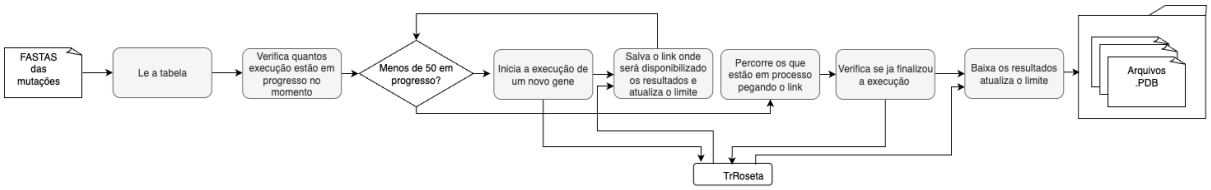

Figura 35: Diagrama do funcionamento da automação da ferramenta trRosetta

3.3.7 MODELLER

Assim como o Alphafold3, para gerar os modelos para cada uma das proteínas mutadas, também foram utilizados os scripts de uma automatização do uso do Modeller disponibilizada em um repositório do GitHub desenvolvidos no Combi-Lab [8]. Essa ferramenta recebe como entrada uma tabela contendo gene, mutação e fasta e utiliza uma imagem docker com o MODELLER instalado para gerar todos os arquivos necessários, fazer as predições e disponibilizar os modelos .PDBs resultantes.

A tabela utilizada como entrada é bem similar à utilizada pela ferramenta de automação do AlphaFold3, contendo uma coluna para gene, uma para variante e uma com a sequência fasta. Os resultados ficam disponíveis em uma pasta “results” contendo todos os arquivos .PDBs gerados.

[8] https://github.com/rvlampert/modeller

3.4 Validação das estruturas preditas computacionalmente

Terminando a etapa de modelagem, temos à nossa disposição os arquivos .PDB para cada uma das mutações geradas por cada uma das ferramentas usadas para predizer a estrutura tridimensional das proteínas. Assim, o próximo passo é submeter esses arquivos resultantes para diferentes ferramentas de validação - Saves (Errat, Verify3D), MolProbity, QMEAN-DISCO, QMEAN e VoroMQA - a fim de obter informações relevantes que possam auxiliar na comparação das ferramentas de modelagem. Aqui, diferente da etapa anterior de modelagem, as ferramentas seguiram todas o mesmo padrão, como são todas ferramentas web, com a mesma entrada sendo o arquivo .PDB e diferentes métricas como resultado, foi possível seguir o mesmo fluxograma para cada uma delas.

Sabendo o caminho do arquivo para cada um dos modelos de cada ferramenta, o próximo passo é submeter as ferramentas de validação, e para isso foram utilizadas novamente as bibliotecas do Selenium para acessar cada uma das ferramentas, submetendo o arquivo referente àquela determinada variante e ferramenta de modelagem, clicando nos botões de forma automática e configurando o que fosse necessário até que as métricas resultantes fossem exibidas e pudessem ser extraídas para uma nova tabela de resultados.

Abaixo, na Figura 36, segue um diagrama explicando como foi feita a automação das ferramentas de validação.

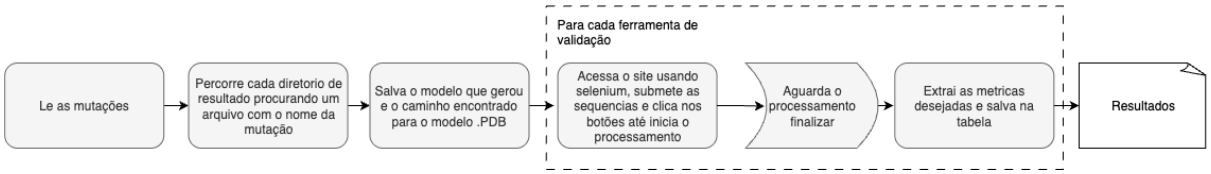

Figura 36: Diagrama do funcionamento da automação das ferramentas

O resultado final de todo esse processamento foi uma tabela unificada, com 3.072 linhas (8 ferramentas * 384 variantes), sendo uma para cada modelo gerado por cada ferramenta de predição, e colunas, incluindo o gene, a variante, o fasta, o modelo e diversas métricas obtidas através das seis ferramentas de validação utilizadas, como mostra a figura abaixo 37

	gene	variant	fasta	model	errat	verify	poor_rotamers	poor_rotamers_percentage	favored_rotamers	favored_rotamers_percentage	...	cb_deviations_percentage	bad_bonds	ba
0	atpE	atpE_Ala63Pro	MDPTIAAGALIGGGLIMAGGAIGAGIGDGVAGNALISGVARQPEAQ...	swiss_model	100.0	29.63% >= 0.1	0.0	0.00%	52.0	98.11%	...	0.00%	0 / 582	
1	atpE	atpE_Ala63Pro	MDPTIAAGALIGGGLIMAGGAIGAGIGDGVAGNALISGVARQPEAQ...	colab_alphafold2	100.0	48.15% >= 0.1	2.0	3.77%	50.0	94.34%	...	0.00%	20 / 581	
2	atpE	atpE_Ala63Pro	MDPTIAAGALIGGGLIMAGGAIGAGIGDGVAGNALISGVARQPEAQ...	modeller	100.0	28.40% >= 0.1	1.0	1.89%	50.0	94.34%	...	1.49%	0 / 582	
3	atpE	atpE_Ala63Pro	MDPTIAAGALIGGGLIMAGGAIGAGIGDGVAGNALISGVARQPEAQ...	phyre2	100.0	24.69% >= 0.1	0.0	0.00%	53.0	100.00%	...	0.00%	3 / 581	
4	atpE	atpE_Ala63Pro	MDPTIAAGALIGGGLIMAGGAIGAGIGDGVAGNALISGVARQPEAQ...	i_tasser	100	17.28% >= 0.1	0.0	0.00%	53.0	100.00%	...	0.00%	46 / 581	

Figura 37: Visualização da tabela com os resultados das ferramentas de validação para cada um dos modelos preditos com mutação pontual *missense*.

3.5 Website com a base de dados de mutações pontuais associadas a resistência

Um outro objetivo desse trabalho é a disponibilização de forma fácil aos resultados e estruturas geradas e obtidas ao longo dessa pesquisa. Assim, a solução adotada foi a criação de um *website* para fazer o depósito das estruturas com mutações e das suas respectivas métricas de validação para cada modelo gerado.

Devido ao fato da criação desse website necessitar de conhecimentos mais aprofundados em diversas tecnologias específicas de desenvolvimento e linguagens de programação voltadas para web, a construção dele foi feita em parceria com os alunos de pós-graduação do Combi-Lab.

4 RESULTADOS

Nesta sessão serão apresentados os resultados do estudo de caso para *M. tuberculosis* e as validações a partir das ferramentas de validação.

4.1 Pré-processamento

O trabalho foi realizado considerando proteínas do *M. tuberculosis* e mutações pontuais *missense* associadas à resistência aos principais fármacos utilizados no tratamento da TB. A Tabela 3 mostra como ficou nosso dataset após o pré-processamento. Podemos observar que alguns genes possuem uma maior quantidade de mutações em comparação a outros e que o tamanho das sequências varia da menor, com 81 aminoácidos, até a maior com 1.172 aminoácidos.

Identifier	Gene	Número de Aminoácidos	Número de Mutações
Rv1305	atpE	81	6
Rv0682	rpsL	124	4
Rv3547	ddn	151	1
Rv0678	Rv0678	165	8
Rv2043c	pncA	186	204
Rv0701	rplC	217	1
Rv3919c	gid	224	9
Rv1694	tlyA	268	2
Rv1484	inhA	269	1
Rv3854c	ethA	489	9
Rv0005	gyrB	675	8
Rv1908c	katG	740	5
Rv0006	gyrA	838	10
Rv3795	embB	1098	13
Rv0667	rpoB	1172	103

Tabela 3: Dados de genes com número de aminoácidos e mutações

4.2 Modelagem das estruturas

Neste trabalho, nós tivemos um total de 3.072 estruturas modeladas (sete ferramentas de predição de estrutura 3D X 384 sequências FASTA com mutações).

4.3 Avaliação dos modelos obtidos por ferramenta de validação

Nessa sessão organizamos os resultados por cada ferramenta de validação utilizada neste trabalho.

4.3.1 ERRAT

Os escores ERRAT indicam que, quanto maior o valor, melhor a qualidade da estrutura. Tipicamente, para estruturas de alta resolução, esse escore é em torno de 95% ou superior.

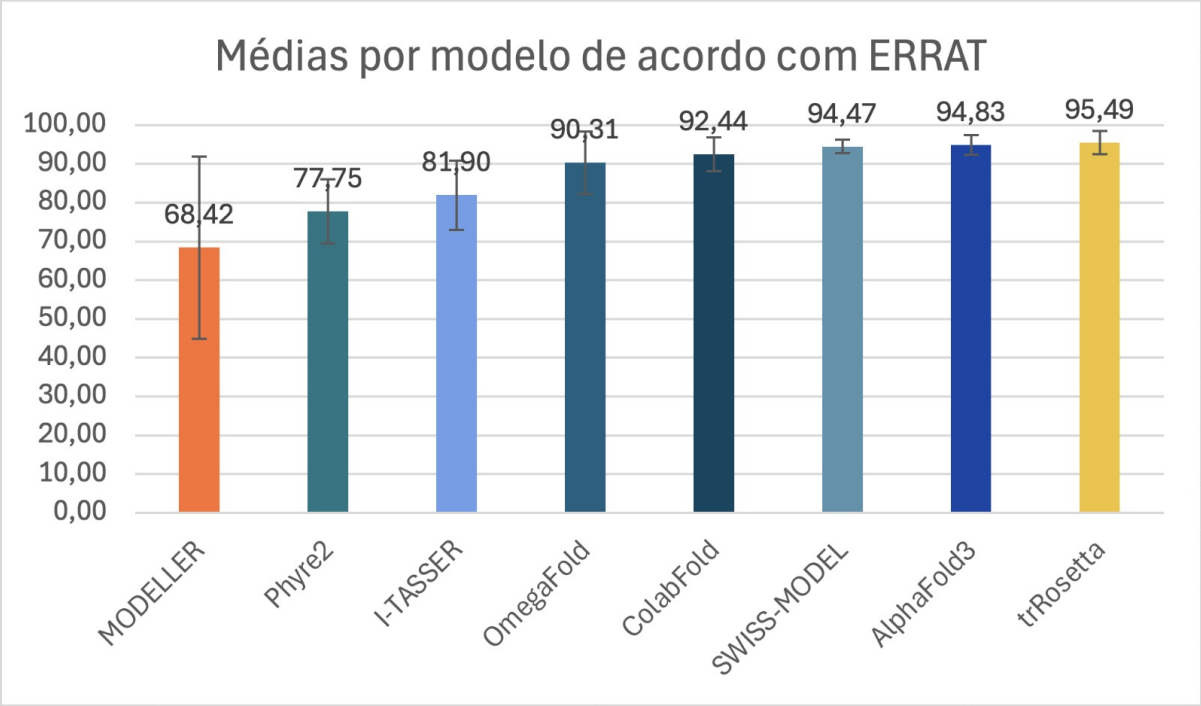

Figura 38: Gráfico de barras mostrando no eixo X as ferramentas de predição de estruturas 3D e no eixo Y os valores médios para cada uma delas, segundo a medida de qualidade do ERRAT. As barras incluem os respectivos desvios padrão para cada ferramenta.

Analisando o gráfico de médias, Figura 38, temos acima do 90 o OmegaFold, o ColabFold, o SWISS-MODEL, AlphaFold3 e trRoseta com 92.44, 94.47, 94.83 e 95.49, respectivamente. O último colocado em relação à qualidade dos modelos gerados foi o MODELLER com 68.42, seguido pelo Phyre2 com 77.75, depois o I-TASSER com 81.90.

Ao analisarmos o percentual de modelos rejeitados pela ferramenta no gráfico da Figura 39, podemos ver que com menos de 50% de rejeição temos o AlphaFold3 e o I-TASSER com 44.79% e 40.10%, respectivamente. A seguir, veio o ColabFold e o SWISS-MODEL, com

58.59% e 78.13%. Com aproximadamente 95% de rejeição, 95.05% temos o trRosetta. As duas ferramentas que performaram pior e se mantiveram nas duas últimas posições foram, o Phyre2 e o MODELLER, com 95.83% e 98.43% respectivamente.

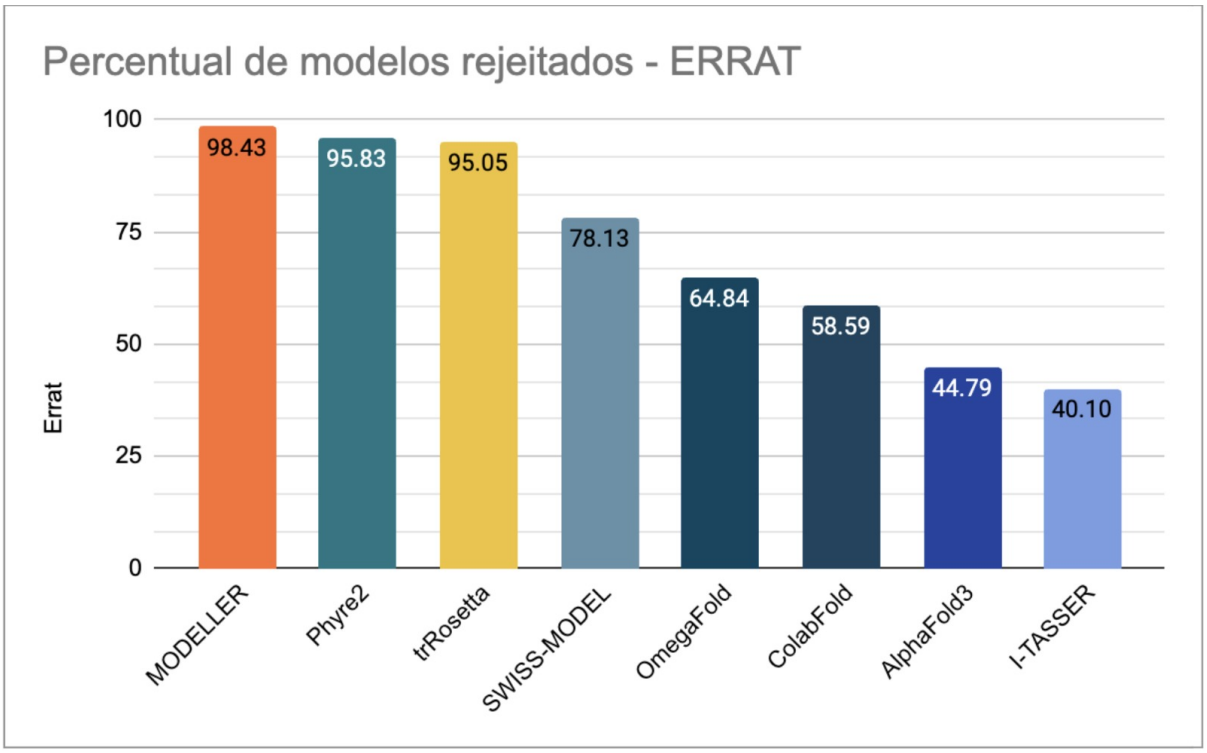

Figura 39: Gráfico de barras mostrando no eixo X as ferramentas de predição e no eixo Y os percentuais de estruturas rejeitadas em cada uma delas, segundo o ERRAT.

4.3.2 VERIFY3D

A próxima métrica utilizada para avaliar os modelos gerados foi o VERIFY3D. Ela representa uma porcentagem, então seus valores podem variar no gráfico de zero até cem. Uma pontuação Verify3D acima de 0,2 é geralmente considerada indicativa de um bom modelo, enquanto uma pontuação abaixo de 0,1 pode indicar problemas significativos com o modelo (modelos rejeitados).

- **Falhou (Fail):** Menos de 80% dos aminoácidos obtiveram pontuação $\geq 0,1$ no perfil 3D/1D.
- **Passou (Pass):** Mais de 80% dos aminoácidos obtiveram pontuação $\geq 0,1$ no perfil 3D/1D.

Podemos ver no gráfico da Figura 40 um resultado muito similar quando comparado ao ERRAT para o melhor e os piores modelos, contudo, as posições intermediárias trocaram. De acordo com o VERIFY3D, as duas melhores ferramentas foram trRosetta e AlphaFold3 com

75.39 e 73.98, respectivamente. O modelo que pior performou foi novamente o MODELLER, com uma média de 58.86, seguida pelas ferramentas SWISSMODEL, Phyre2, ColabFold e I-TASSER, todos com médias próximas a 71, sendo elas 71.02, 71.32, 71.93 e 71.96, respectivamente.

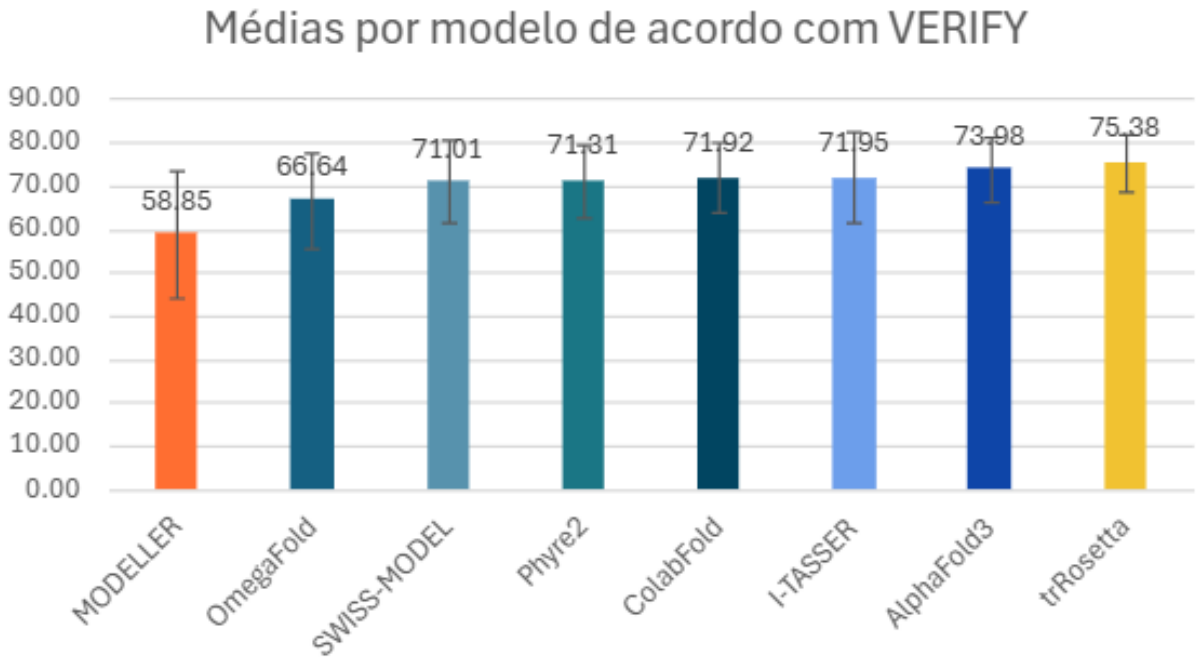

Figura 40: Gráfico de barras mostrando no eixo X as ferramentas de predição de estruturas 3D e no eixo Y os valores da média para cada uma delas, segundo a medida de qualidade de validação do VERIFY3D. As barras incluem os respectivos desvios padrão para cada ferramenta.

Para o VERIFY3D, temos na Figura 41 o gráfico do percentual de modelos que apresentaram falha segundo sua métrica. Novamente, as ferramentas que foram pior avaliadas permaneceram nas últimas posições, enquanto aquela que ficou melhor colocada mostrou-se ter uma queda quando comparada à sua posição no gráfico de médias.

Os modelos que melhor desempenharam, de acordo com essa métrica, foram o AlphaFold3 com 83.07% dos modelos falhando e o ColabFold, onde 85.68% dos modelos apresentaram falha. Podemos ver no gráfico três modelos onde mais de 95% dos seus modelos retornaram falha, o MODELLER com 99.74%, o Phyre2 com 97.40% e o SWISS-MODEL com 95.83%, seguidos pelos modelos trRosetta e I-TASSER, ambos com aproximadamente 89%, sendo 89.58 e 89.84 respectivamente.

4.3.3 MolProbity

O terceiro modelo de validação estudado foi o MolProbity score, ele trazia diversas métricas diferentes. Nesse estudo, optamos por focar apenas no valor MolProbity retornado pela própria ferramenta.

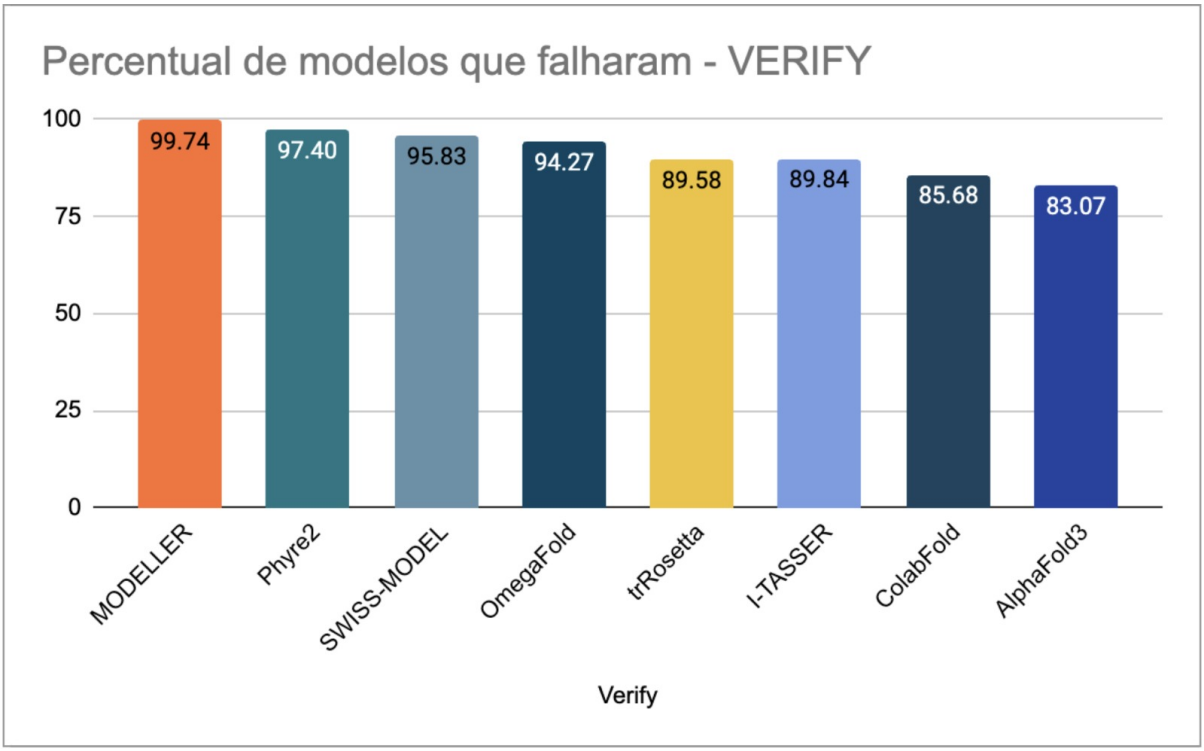

Figura 41: Gráfico de barras mostrando no eixo X as ferramentas de predição e no eixo Y os percentuais de estruturas que falharam em cada uma delas, segundo o VERIFY3D.

A fim de comparar os modelos, geramos o gráfico da Figura 42, com as médias obtidas em cada uma das ferramentas preditivas. Diferente das demais, para essa métrica, pontuações mais baixas indicam maior qualidade estrutural.

Nas três primeiras colocações temos o SWISS-MODEL, o trRosetta e o AlphaFold3, com 0.93, 1.08 e 1.38 de médias, respectivamente. Em quinto lugar temos o ColabFold com aproximadamente 2.05 de média. Novamente o Phyre2 ficou em sexto lugar, com uma média de 2.47. Mais uma vez o MODELLER ficou em último lugar quando comparado aos demais, com uma média de 3.02.

4.3.4 VoroMQA

A próxima métrica a ser estudada é o VoroMQA, seus valores variam de 0 a 1 e quanto mais próximo do 1, melhor o modelo se saiu na tarefa de predizer as proteínas tridimensionais. Para relembrar, a interpretação do escore global do VoroMQA de um modelo estrutural, pode-se utilizar a seguinte regra:

- **Modelo provavelmente é bom:** Se o escore for maior que 0,4.
- **Modelo provavelmente é ruim:** Se o escore for menor que 0,3.
- **Modelo não pode ser classificado com confiabilidade como bom ou ruim apenas com base no VoroMQA:** Se o escore estiver entre 0,3 e 0,4.

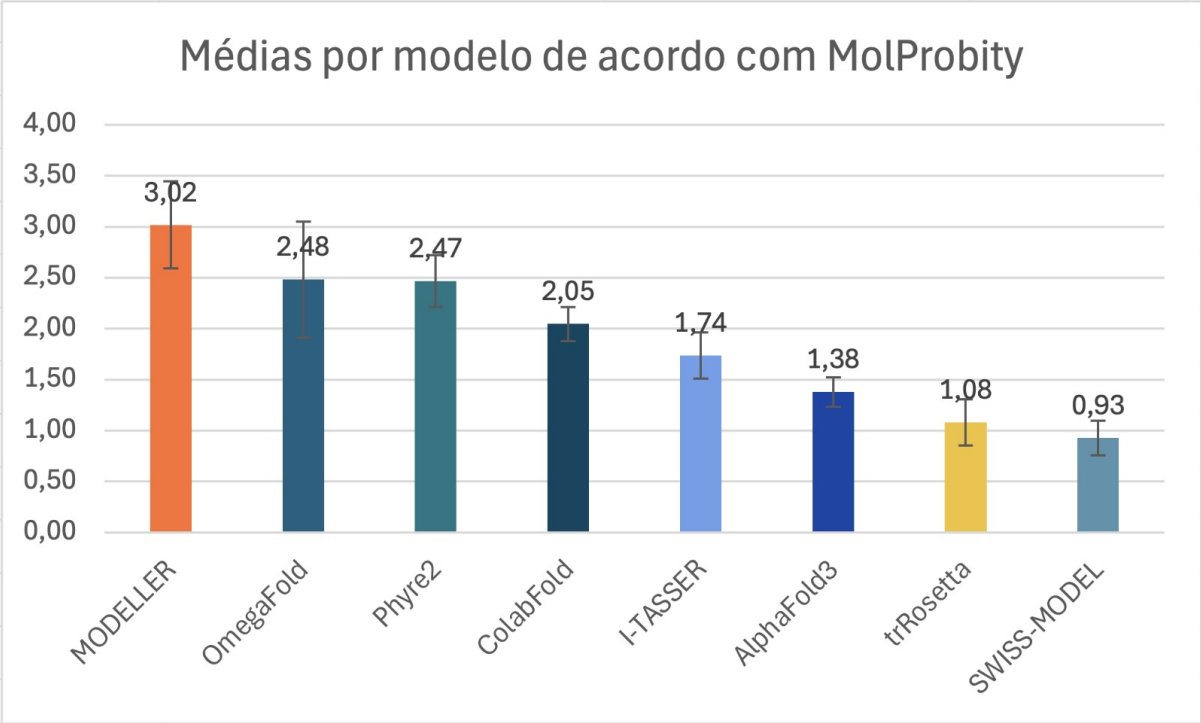

Figura 42: Gráfico de barras mostrando no eixo X as ferramentas de predição de estruturas 3D e no eixo Y os valores das médias para cada uma delas, segundo a validação do MolProbity score. As barras incluem os respectivos desvios padrão para cada ferramenta.

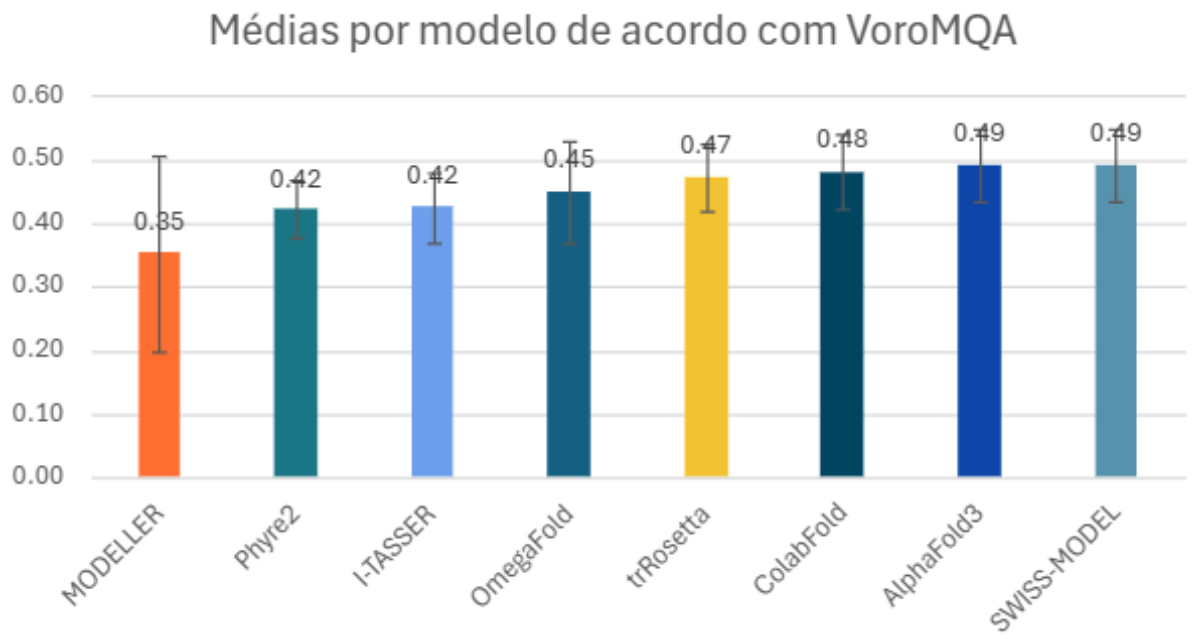

Figura 43: Gráfico de barras mostrando no eixo X as ferramentas de predição de estruturas 3D e no eixo Y os valores das médias para cada uma delas, segundo a validação do VoroMQA. As barras incluem os respectivos desvios padrão para cada ferramenta.

Podemos ver no gráfico da Figura 43, que os valores variaram pouco, ficando todos entre 0.3 e 0.5. Em primeiro lugar, houve um empate, tanto o SWISS-MODEL quanto o AlphaFold3 tiveram resultados iguais e atingiram a média de 0.49. Após eles temos o ColabFold e o trRosetta com médias também muito próximas, 0.47 e 0.48, respectivamente. Por fim, as ferramentas de modelagem que obtiveram as piores médias foi o MODELLER com 0.35, seguidos pelo Phyre2 e I-TASSER, que, de acordo com essa métrica, performaram de maneira similar, ambos obtendo o empate com a média de 0.42.

A ferramenta classifica eles em 3 categorias baseadas no valor retornado, alta qualidade, baixa qualidade e uma categoria intermediária, considerado inconclusiva. Geramos então, na Figura 44, um gráfico comparando quanto por cento dos modelos foram considerados com qualidade alta pelo validador.

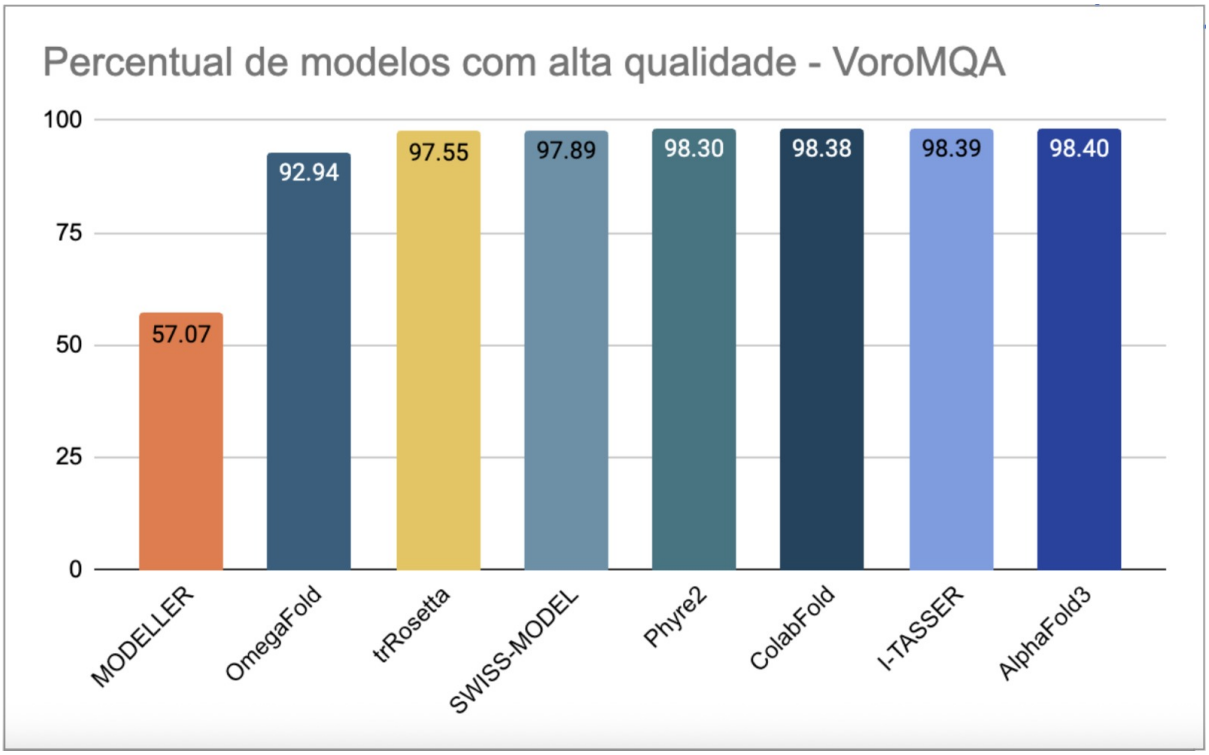

Figura 44: Gráfico de barras mostrando no eixo X as ferramentas de predição de estruturas 3D e no eixo Y o percentual de modelos com alta qualidade para cada uma delas, segundo a validação do VoroMQA.

Essa métrica foi a que apresentou o maior contraste entre os resultados do MODELLER e os demais. Enquanto o MODELLER teve apenas 57.07% dos modelos considerados de alta qualidade, todos os demais tiveram mais de 97% de modelos nessa categoria. Os resultados foram bem semelhantes entre os demais, sendo 97.55 e 97.89 as médias do trRosetta e SWISS-MODEL respectivamente, seguido por outras três médias muito próximas, com uma diferença de um décimo entre os percentuais. O ColabFold, o I-TASSER e o AlphaFold obtiveram 98.38, 98.39 e 98.40, respectivamente.

4.3.5 QMEAN

Por fim, a última ferramenta a ser explorada é o QMEAN, que traz também duas métricas diferentes, QMEAN e QMEANDisCo. Começaremos a análise pelo QMEAN gerando o gráfico da Figura 45 com as médias obtidas para cada uma das ferramentas preditivas.

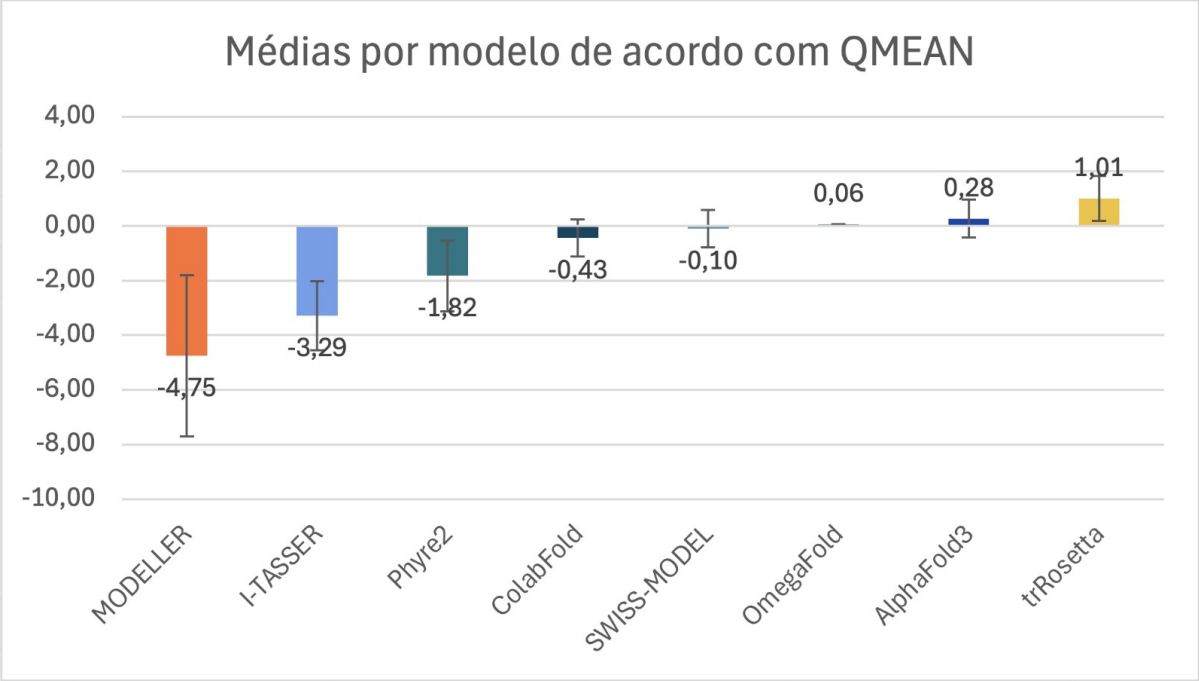

Figura 45: Gráfico de barras mostrando no eixo X as ferramentas de predição de estruturas 3D e no eixo Y os valores das médias para cada uma delas, segundo a validação do QMEAN. As barras incluem os respectivos desvios padrão para cada ferramenta.

Essa é a única métrica onde os valores podem ser negativos, já que os valores que aparecem nas páginas de resultado são transformados em escores Z e, quanto mais negativos, menos preciso o modelo gerado.

Com valores positivos temos o AlphaFold3 com 0.28 e em primeiro lugar o trRosetta com a média de 1.01. Podemos ver que cinco dos sete modelos tiveram sua média de valores negativas, sendo o MODELLER a média mais negativa de todas, com -4,75, seguida pelo I-TASSER com -3.29, e pelo Phyre2 com -1.82. E fechando a lista das ferramentas com médias negativas, temos respectivamente, ColabFold e SWISS-MODEL com -0.43 e -0.10.

4.3.6 QMEAN-DisCo

Por último, analisamos o gráfico da Figura 46, resultante das médias do QMEAN-DisCo. Aqui, os valores podem variar em um intervalo de zero a um, sendo que quanto maior o valor, mais precisa foi a predição. Apesar de serem geradas pela mesma ferramenta e de possuírem o mesmo nome, o gráfico obtido para essa métrica se assemelha ao gráfico com as médias do QMEAN apenas nos dois modelos que performaram pior, sendo novamente o MODELLER e o

I-TASSER, com 0.63 e 0.80. As duas próximas ferramentas obtiveram a mesma média de 0.83, que são o trRosetta e o Phyre2. Seguido pelo ColabFold com 0.87. Por fim, como melhores ferramentas, temos o AlphaFold3 e o SWISS-MODEL, ambas com média de 0.88.

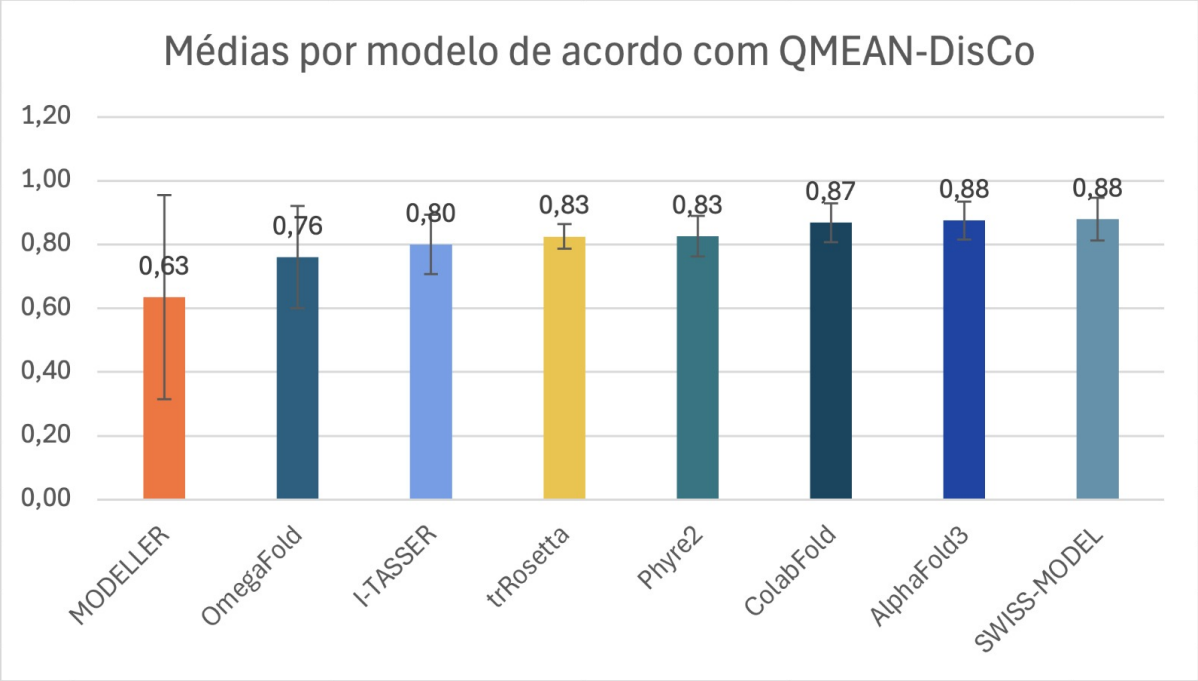

Figura 46: Gráfico de barras mostrando no eixo X as ferramentas de predição de estruturas 3D e no eixo Y os valores das médias para cada uma delas, segundo a validação do QMEANDisCo. As barras incluem os respectivos desvios padrão para cada ferramenta.

Em posse desses resultados, agora temos dados suficientes para analisar e chegar a uma resposta mais assertiva sobre qual modelo performa melhor a tarefa de predizer estruturas tridimensionais para proteínas que sofreram mutações pontuais.

4.4 Avaliação dos modelos obtidos (interpretação por gene)

Nesta sessão serão expostos os resultados obtidos, com enfoque em cada gene individualmente. Importante salientar que esses resultados podem apresentar um viés devido à quantidade de mutações para cada gene, que muda drasticamente, tendo genes com uma única mutação (que é o caso do ddn, inhA e rplC), até genes com 204 mutações (pncA).

4.4.1 ERRAT

O primeiro validador utilizado para analisar as ferramentas de predição foi o ERRAT. Geramos uma tabela, disponível no Anexo 7, com as médias e o desvio padrão obtidos por cada ferramenta em relação a cada um dos genes, ordenados por ordem de tamanho, do menor FASTA até o maior.

Ferramentas como o SWISS-MODEL, Phyre2, e trRosetta frequentemente produzem pontuações altas, indicando uma boa qualidade estrutural em suas previsões para a maioria dos genes analisados; por exemplo, trRosetta teve pontuações superiores a 90 para vários genes, incluindo *rplC* e *tlyA*. Por outro lado, o MODELLER apresentou pontuações mais baixas em diversos genes, como *rpsL* e *embB*, o que sugere menos confiança na precisão de suas estruturas preditas. AlphaFold3 também tende a gerar pontuações altas, semelhante ao SWISS-MODEL, trRosetta, e Phyre2, conforme observado em genes como *atpE*, *pncA*, e *Rv0678*. ColabFold, também apresenta bom desempenho ao longo da tabela, com pontuações variando entre médias e altas. Vimos que todas as ferramentas obtiveram um valor muito próximo de 100 para o gene atpE, no qual a menor média foi a do AlphaFold3, de 98.17.

4.4.2 VERIFY3D

A próxima ferramenta analisada foi o VERIFY3D. Ao observar a tabela, no Anexo 8, podemos perceber que, com exceção da ferramenta do MODELLER, todos os validadores tiveram valores bem similares. Ferramentas como SWISS-MODEL e trRosetta frequentemente mostram pontuações mais altas. Por exemplo, trRosetta obteve pontuações acima de 76 para diversos genes, como *rpsL* e *gyrB*. ColabFold, também demonstra um bom desempenho com pontuações relativamente altas, como observado no gene *katG* com uma pontuação de 91,68 ± 1.49. Genes específicos como *katG* atingem as mais altas pontuações em várias ferramentas, especialmente SWISS-MODEL, Phyre2, e trRosetta. Além disso, o gene *ethA* apresenta variação nas pontuações, onde métodos como I-TASSER e trRosetta registram pontuações acima de 85.

4.4.3 MolProbity

Quando analisamos a tabela do MolProbity por gene, no Anexo 9, podemos perceber que assim como a tabela do VERIFY3D, alguns modelos performaram melhor que outros, sabendo que o valor do MolProbity, quanto mais próximo a zero melhor, vemos que os modelos SWISS-MODEL, AlphaFold3 e trRosetta performaram melhor para todos os genes, enquanto os demais acabaram tendo uma performance não tão boa. Quanto à estabilidade, podemos notar que o tamanho não influenciou e não existe nenhum valor na tabela que indique que algum modelo performa melhor ou pior levando em consideração a quantidade de aminoácidos que o gene possui.

4.4.4 VoroMQA

A próxima tabela, no Anexo 10, foi referente às médias por gene para o validador VoroMQA. Ele possuiu um comportamento semelhante ao do VERIFY3D e novamente o gene no qual as ferramentas foram menos assertivas predizendo o modelo foi o atpE e novamente a exceção foi o MODELLER, cujo o gene que obteve os piores resultados foi o rpsL.

4.4.5 QMEAN-DisCo

A penúltima métrica analisada foi o QMEAN-DisCo, podemos ver na tabela, no anexo 11, que para a maioria das ferramentas ele performou de maneira similar, o mais destoante foi o MODELLER, onde a variação entre os melhores e piores foi superior às demais. Mais uma vez podemos ver que o tamanho do gene não exerceu influência na qualidade dos modelos gerados.

4.4.6 QMEAN

Por fim, realizamos a análise baseada no QMEAN, disponível na tabela do Anexo 12. Assim como o ERRAT, vimos que alguns modelos tiveram uma estabilidade em seus resultados maior do que outros. As ferramentas Phyre2, I-TASSER e MODELLER, além de possuírem as menores médias, foram as que tiveram maior variação entre os melhores e piores resultados, enquanto a média das demais tiveram um comportamento mais linear independente do gene.

4.5 Website com a base de dados de mutações pontuais associadas a resistência

Além desses resultados, também é objetivo desse trabalho disponibilizá-los de forma fácil para que mais pesquisadores possam usá-los. Para isso, foi criado um *website* onde tudo que foi gerado, assim como todos os resultados obtidos ficarão disponíveis.

O site está disponível para acesso [1]. A Figura 47 mostra como é a interface gráfica que o usuário tem acesso. A Figura 48 mostra a barra de seleção onde podemos filtrar pelas ferramentas de predição que gostaríamos de analisar os modelos gerados. Nele, o usuário pode consultar e visualizar as estruturas, utilizar filtros para buscar e fazer download das informações desejadas. Posteriormente, podemos adicionar mais funcionalidades ao site, pois o objetivo é mantê-lo em funcionamento e atualizado.

Isso é útil para a comunidade acadêmica que estuda tanto mutações da TB, quanto ferramentas de predição de estruturas terciárias. Além disso, nós não encontramos nenhuma plataforma com este tipo de informações de forma gratuita e aberta ao público em geral. Esse é o primeiro site com o depósito de uma grande quantidade de estruturas terciárias do *M. tuberculosis* com mutações pontuais *missense* preditas com o uso de ferramentas *in-sílico* de bioinformática estrutural e utilizando diferentes métricas de validação.

[1] https://combilab-furg.github.io/gene-data/

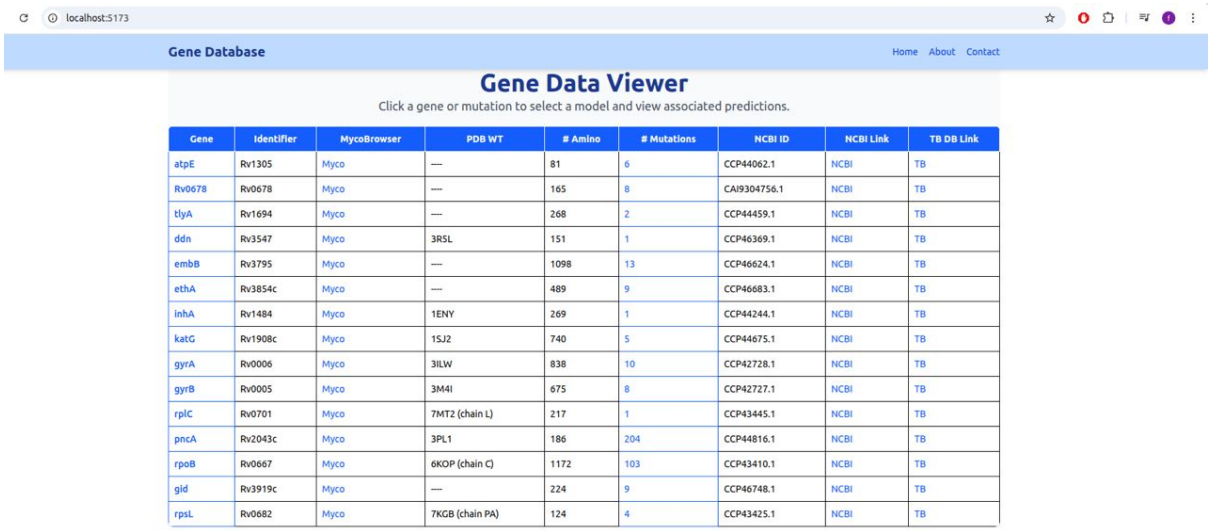

Figura 47: Página inicial do site desenvolvido para divulgar os resultados obtidos neste trabalho.

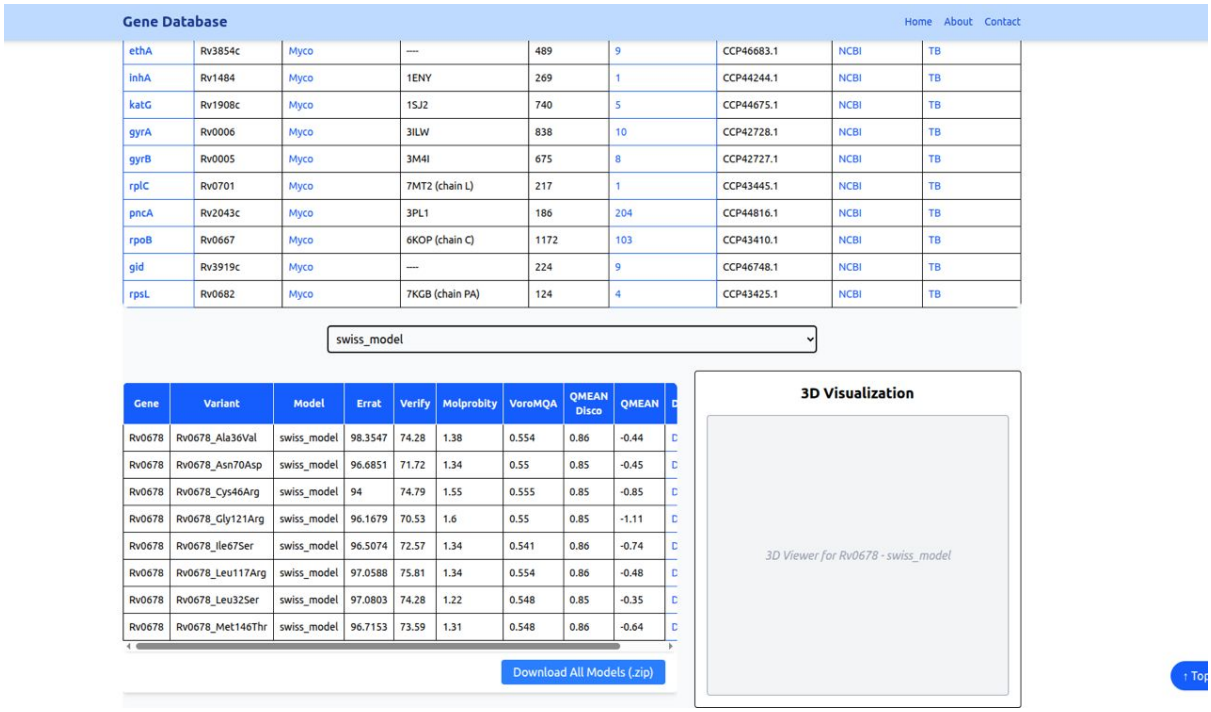

Figura 48: Continuação do site que foi desenvolvido pelo Combi-Lab.

5 DISCUSSÃO

Esses dados obtidos através das ferramentas de validação permitem traçar um panorama da qualidade dos modelos gerados por diferentes algoritmos de predição estrutural. A análise das médias, desvios padrão e percentuais de aprovação por métrica revela padrões importantes sobre o desempenho relativo de cada abordagem.

Tabela 4: Valor da média e desvio padrão para cada uma das ferramentas de modelagem estrutural. Os valores destacados em verde mostram as ferramentas que obtiveram os melhores resultados para cada um dos validadores utilizados. Já os valores em vermelho, mostram quais foram as ferramentas que obtiveram os piores resultados.

Métrica	**SWISS--MODEL**	**Colab_Fold**	**Modeller**	**Phyre2**	**I-TASSER**	**trRosetta**	**AlphaFold3**	**OmegaFold**
ERRAT	94.47 ± 1.73	92.44 ± 4.31	**68.42 ± 23.46**	77.75 ± 8.29	81.90 ± 8.98	**95.49 ± 3.00**	94.83 ± 2.58	90.31 ± 7.99
VERIFY3D	71.02 ± 9.53	71.93 ± 8.03	**58.86 ± 14.44**	71.32 ± 8.47	71.96 ± 10.37	**75.39 ± 6.65**	73.98 ± 7.45	66.64 ± 10.92
MolProbity	**0.93 ± 0.17**	2.05 ± 0.17	**3.02 ± 0.43**	2.47 ± 0.26	2.74 ± 0.23	1.08 ± 0.23	1.38 ± 0.15	2.48 ± 0.57
VoroMQA	**0.49 ± 0.06**	0.48 ± 0.06	**0.35 ± 0.15**	0.42 ± 0.05	0.42 ± 0.05	0.47 ± 0.05	**0.49 ± 0.06**	0.45 ± 0.08
QMEAN-DisCo	**0.88 ± 0.07**	0.87 ± 0.06	**0.63 ± 0.32**	0.83 ± 0.06	0.80 ± 0.09	0.83 ± 0.04	**0.88 ± 0.06**	0.76 ± 0.16
QMEAN	-0.10 ± 0.68	-0.43 ± 0.69	**-4.75 ± 2.95**	-1.82 ± 1.29	-3.29 ± 1.26	**1.01 ± 0.83**	0.28 ± 0.70	0.06 ± 0.0

5.1 Modeller

O MODELLER destacou-se negativamente na maioria das métricas avaliadas. Tanto em ERRAT quanto em VERIFY3D, ele apresentou médias mais baixas do que os demais modelos. Este desempenho foi confirmado pelos altos percentuais de rejeição, acima de 95%. Se considerarmos o validador Molprobity, o MODELLER novamente teve o pior resultado, reforçando a tendência de baixa qualidade estrutural. O mesmo acontece ao considerarmos os resultados do VoroMQA, onde teve o menor valor médio. Para QMEAN e QMEANDisCo não foi diferente, o MODELLER mostrou abaixo da média das demais.

Em resumo, o MODELLER apresentou consistentemente o pior desempenho. Seus modelos foram mais frequentemente rejeitados, com maior variabilidade e menor fidelidade estrutural.

5.2 AlphaFold3 e trRosetta

As ferramentas AlphaFold3 e trRosetta mostraram desempenhos superiores de forma consistente em basicamente todas as métricas. No ERRAT e VERIFY, obtiveram as maiores médias. O trRosetta foi a ferramenta que obteve os melhores valores em ambas as validações. No MolProbity, apesar de nenhum deles ser o melhor, o AlphaFold3 e o trRosetta ficaram em segundo e terceiro lugar, mostrando uma consistência nos seus modelos, ficando atrás apenas do SWISS-MODEL. Nas métricas QMEAN e QMEAN-DisCo, o trRosetta se destacou, sendo um dos únicos com valor positivo para o QMEAN (AlphaFold3 também ficou positivo). Ambas foram as únicas ferramentas que obtiveram valores positivos. Para essa métrica, o trRosetta mais uma vez foi o que obteve os melhores valores. O AlphaFold3 e o SWISS-MODEL também se posicionaram bem nessas métricas, sendo os melhores colocados em relação ao QMEAN-DisCo.

Desse modo, é possível afirmar que tanto AlphaFold3 quanto trRosetta demonstram ser boas ferramentas para predição estrutural de proteínas com mutação pontual, considerando o estudo de caso apresentado, com consistência, qualidade e fidelidade. O trRosetta lidera em mais métricas, enquanto AlphaFold3 mostra maior consistência.

5.3 SWISS-MODEL

Essa ferramenta apresentou um bom desempenho intermediário de forma geral. O SWISS-MODEL teve bons valores para MolProbity, além do QMEAN-DisCo e VoroMQA, se equiparando ao AlphaFold3. O SWISS-MODEL, porém, apresentou um desempenho levemente inferior no QMEAN, ficando na fronteira entre os bons e os medianos. Embora não supere os melhores modelos, no modo geral, ele mostrou ser uma ferramenta confiável.

5.4 ColabFold

Essa ferramenta obteve resultados razoáveis no geral, ela não se destacou como melhor nem pior em nenhum dos modelos de validação, o que indica uma consistência em seus resultados. Apesar de ter resultados inferiores considerando o valor médio da métrica VERIFY3D quando comparado com o AlphaFold3 e o trRosetta, ela apresentou uma performance sólida nas demais métricas.

5.5 I-TASSER e Phyre2

O desempenho do I-TASSER e do Phyre2 foram os mais inconsistentes. Apesar de apresentarem valores médios bons para algumas das validações, ambas as ferramentas mostraram valores médios ruins nas métricas com Molprobity e VoroMQA, ficando apenas na frente do MODELLER. Em relação ao QMEAN, seus valores foram negativos, tendo o I-TASSER o se-

gundo pior desempenho. No QMEAN-DisCo, ambos ficaram no meio da tabela. No geral, o I-TASSER e o Phyre2 apresentam desempenho aceitável, mas com variabilidade alta e falta de consistência entre métricas.

5.6 *Borda count*

Por fim, com a finalidade de buscar uma resposta mais precisa sobre qual ferramenta de predição tridimensional possui os melhores resultados em relação a proteínas que sofreram mutação pontual, é necessário realizar uma concatenação desses resultados. Isso foi realizado utilizando um algoritmo de concatenação de ranqueamento chamado "*Borda Count*".

A ideia desse método é combinar ou somar as posições (ranks) de cada ferramenta e, em seguida, ordenar os rótulos das classes de acordo com suas novas pontuações combinadas [71, 5, 78]. Resumidamente, cada posição do ranqueamento possui um peso, e depois somamos os pesos obtidos em cada um dos validadores, gerando um único ranqueamento final, conforme é representado na Figura 49 [178].

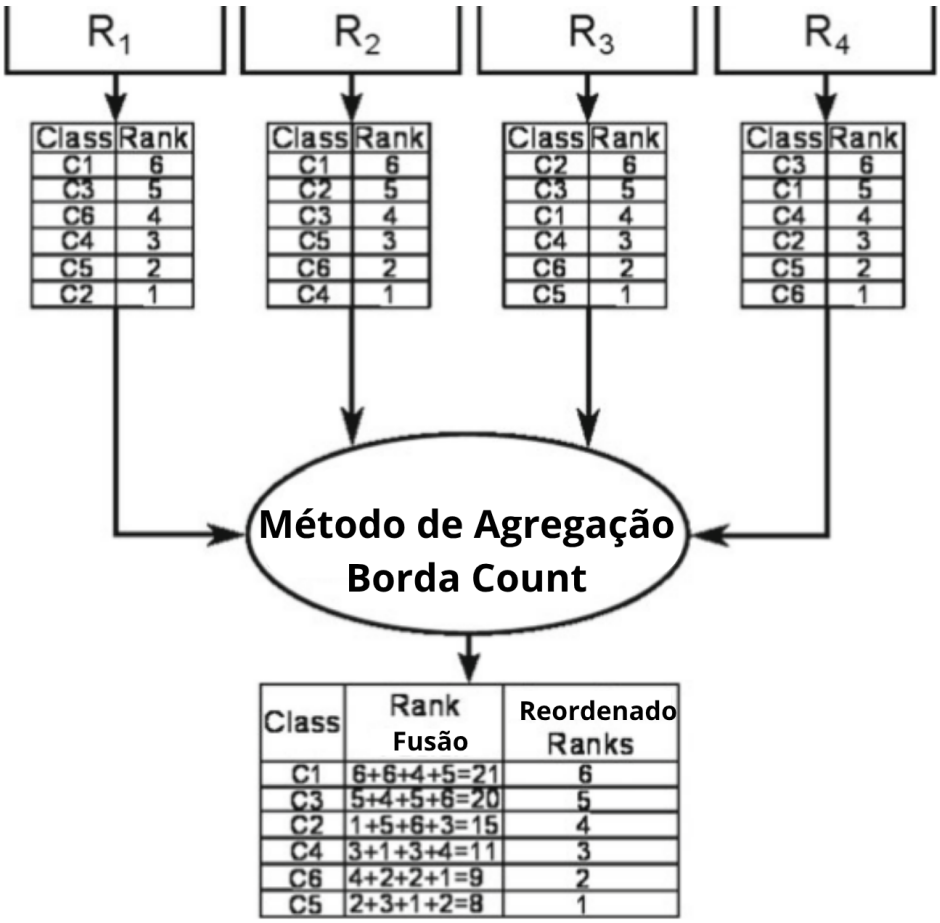

Figura 49: Um exemplo do método de agregação de rankings *Borda count*.

Tabela 5: Posição e valor das ferramentas de modelagem estrutural para cada métrica avaliada.

Modelo	**Tipo**	**ERRAT**	**VERIFY**	**MolProbity**	**VoroMQA**	**QMEAN-DISCO**	**QMEAN**
SWISS-MODEL	Posição	3	6	1	1	7	4
	Valor	5	2	7	7	7	4
ColabFold	Posição	4	4	4	3	3	5
	Valor	4	4	4	3	3	5
Modeller	Posição	8	8	8	8	8	3
	Valor	0	0	0	0	0	8
Phyre2	Posição	7	5	5	7	4	2
	Valor	1	3	3	1	4	2
I-TASSER	Posição	6	3	7	6	6	7
	Valor	2	5	1	2	2	1
trRosetta	Posição	1	1	2	4	5	1
	Valor	7	7	6	4	3	7
AlphaFold3	Posição	2	2	3	2	2	2
	Valor	6	6	5	6	6	6
OmegaFold	Posição	5	7	6	5	7	3
	Valor	3	1	2	3	1	5

Como possuímos sete modelos diferentes, os valores variam entre um e sete, correspondendo a sua posição. Por exemplo, uma ferramenta que chegou em quinto lugar o valor 5. Optamos por utilizar apenas o ranqueamento das médias, uma vez que é a única métrica comum para todas as ferramentas.

O ranqueamento final, aplicando o algoritmo e *Borda Count* está descrito na Tabela 6.

Tabela 6: Soma total dos valores atribuídos a cada modelo de predição estrutural. Sendo que valores mais baixos indicam melhor posição no ranking (melhor desempenho segundo o *borda count*) e valores mais altos significam piores classificações no ranking.

Modelo	**Valor somado**
AlphaFold3	13
trRosetta	14
SWISS-MODEL	16
ColabFold	23
OmegaFold	33
I-TASSER	35
Phyre2	34
Modeller	48

Esse ranqueamento corrobora com a discussão feita anteriormente, onde as ferramentas AlphaFold3 e trRosetta obtiveram os melhores valores e tiveram um desempenho melhor, seguido por SWISS-MODEL, ColabFold, depois I-TASSER, Phyre2, e por fim o MODELLER,

sendo a pior performance referente a tarefa de predizer modelos tridimensionais com mutações pontuais.

Sendo assim, diante da análise detalhada das métricas de validação e da aplicação do algoritmo de *Borda Count* para consolidar os resultados, é possível afirmar que as ferramentas AlphaFold3 e trRosetta são as mais indicadas para a predição estrutural de proteínas com mutações pontuais associadas à resistência a fármacos. Ambas demonstraram alta consistência, qualidade e fidelidade estrutural. O SWISS-MODEL também se mostrou uma opção viável e estável, enquanto o ColabFold apresentou um desempenho equilibrado, embora sem grande destaque. Já o I-TASSER e o Phyre2 revelaram inconsistências importantes, e o MODELLER, por sua vez, obteve os piores resultados na maioria das métricas avaliadas, sendo desaconselhado para esse tipo de aplicação. Esses resultados fornecem uma base sólida para decisões futuras na escolha da ferramenta de modelagem mais apropriada para estudos estruturais envolvendo mutações pontuais.

6 CONCLUSÃO

Este trabalho teve como objetivo principal estudar qual a melhor ferramenta de predição de estrutura tridimensional para modelar mutações pontuais do tipo *missense*. Com o estudo de caso foram modeladas 384 proteínas com mutações pontuais associadas à resistência aos principais fármacos utilizados no tratamento da TB de 15 diferentes genes do *Mycobacterium tuberculosis*. Tivemos um total de 3.072 modelos tridimensionais, utilizando 8 ferramentas de predição. Foi possível realizar uma análise detalhada sobre a qualidade das estruturas geradas utilizando 5 ferramentas de validação, permitindo um estudo mais aprofundado sobre como essas ferramentas performam quando o gene em questão sofreu uma mutação pontual *missense* e comparar esses preditores de modelos para descobrir qual é o mais indicado para essa situação.

Ao longo do estudo, observou-se uma variação na qualidade dos modelos entre as diferentes ferramentas de predição. O MODELLER apresentou desempenho consistentemente, porém inferior em todas as métricas avaliadas. Por outro lado, ferramentas como AlphaFold3 e trRosetta se destacaram, apresentando médias superiores e uma boa consistência entre as diferentes métricas. O SWISS-MODEL, embora tenha oscilado em algumas categorias, mostrou-se uma alternativa estável e confiável. O ColabFold obteve um desempenho mediano, mas consistente, enquanto I-TASSER e Phyre2 oscilaram bastante, apresentando bons resultados em alguns aspectos e não tão bons em outros.

A aplicação do algoritmo *Borda Count* permitiu concatenar esses resultados provenientes das diversas métricas em um único ranqueamento, reforçando as observações feitas durante a análise individual. AlphaFold3 e trRosetta se consolidaram como as ferramentas mais indicadas para a tarefa de modelagem estrutural de proteínas mutantes da TB, seguidas por SWISS-MODEL, com desempenho também satisfatório. Já o MODELLER ficou claramente atrás dos demais, sendo desaconselhado para esse tipo de análise.

Outro importante resultado deste trabalho foi o desenvolvimento do website Gene Database, que disponibiliza todos os dados gerados. Esse site visa facilitar o acesso aos resultados encontrados nessa pesquisa e às estruturas tridimensionais analisadas e foi feito com intuito de servir como repositório para futuros estudos. A interface permite consultas filtradas por gene, ferramenta de predição, validador, tipo de mutação e outros campos, o que facilita bastante a busca por algum parâmetro específico, tornando uma ferramenta ainda mais fácil e intuitiva.

Além dos resultados e análises realizadas nessa pesquisa, este trabalho ressalta a importância da escolha de ferramentas *in-sílico* em estudos envolvendo predições estruturais. Ainda que existam diversas soluções disponíveis, nem todas entregam a mesma qualidade.

Como a metodologia proposta tem a etapa de coleta, integração e pré-processamento independente, ao atualizar somente essa etapa é possível o uso para outros estudos de caso. Assim, como trabalhos futuros, pretendemos ampliar a quantidade de sequências a serem utilizadas nas ferramentas de predição, pois tendo em mente um dataset e um pré-processamento diferentes, podemos utilizar as mesmas ferramentas de predição e validação automatizadas para realizar novas análises e chegar a novos resultados. Além disso, pretende-se expandir o escopo do banco de dados incluindo genes relacionados a outras doenças infecciosas e genéticas, não se restringindo apenas à TB. A ampliação pode também considerar outros tipos de mutações, como inserções, deleções e mutações *nonsense*.

Adicionalmente, novas ferramentas de predição e validação poderão ser incorporadas à medida que forem desenvolvidas, bem como aprimoramentos na interface do site, permitindo, por exemplo, visualização interativa das estruturas, integração com bancos de dados externos e sugestões automatizadas baseadas no perfil da mutação. Com essas melhorias, o Gene Database poderá se tornar uma plataforma de referência para estudos estruturais de proteínas mutantes, incentivando a reprodutibilidade e a colaboração entre diferentes grupos de pesquisa.

Este trabalho, portanto, além de oferecer uma análise crítica e robusta das ferramentas atuais de predição estrutural, inaugura uma base pública acessível para consulta e exploração científica, contribuindo de forma significativa para a bioinformática estrutural e o estudo de resistência antimicrobiana.

A TABELAS DAS AVALIAÇÕES POR GENE

A.1 ERRAT

Gene	SWISS-MODEL	ColabFold	MODELLER	Phyre2	I-TASSER	trRosetta	AlphaFold3	OmegaFold
atpE	99.54 ± 1.12	100.00 ± 0.00	99.77 ± 0.56	100.00 ± 0.00	100.00 ± 0.00	99.09 ± 2.24	98.17 ± 2.83	100.00 ± 0.00
rpsL	86.28 ± 0.07	74.44 ± 2.92	29.14 ± 8.34	75.98 ± 1.72	75.31 ± 8.45	89.97 ± 5.01	96.09 ± 0.65	72.81 ± 3.15
ddn	97.90	86.01	60.47	89.22	77.62	90.65	94.24	89.36 ± 0.00
Rv0678	96.57 ± 1.22	86.46 ± 7.19	59.55 ± 7.98	99.91 ± 0.26	90.56 ± 5.17	99.68 ± 0.49	99.41 ± 0.43	99.36 ± 0.85
pncA	94.67 ± 1.26	95.39 ± 1.46	87.37 ± 3.96	76.44 ± 4.83	87.71 ± 5.21	96.06 ± 3.16	96.21 ± 1.21	94.98 ± 2.72
rplC	97.60	71.74	41.57	82.98	67.02	91.49	97.01	81.86
gid	91.90 ± 2.26	90.81 ± 1.59	89.08 ± 3.28	92.96 ± 1.26	73.68 ± 4.06	94.47 ± 1.26	93.62 ± 2.34	97.49 ± 0.78
tlyA	97.27 ± 0.00	96.68 ± 1.38	26.65 ± 1.46	70.97 ± 0.57	80.19 ± 0.82	98.83 ± 0.57	96.09 ± 0.54	98.07 ± 0.00
inhA	94.99	93.49	83.52	98.46	93.10	95.38	98.08	100.00
ethA	92.04 ± 0.25	86.57 ± 1.34	32.36 ± 3.22	55.53 ± 3.31	62.62 ± 4.69	91.98 ± 1.63	89.21 ± 1.97	87.73 ± 0.87
gyrB	97.43 ± 0.10	92.37 ± 0.99	55.78 ± 3.65	74.11 ± 0.99	72.47 ± 3.40	97.67 ± 0.64	96.89 ± 0.36	94.35 ± 0.62
katG	96.04 ± 0.69	89.26 ± 0.32	60.49 ± 4.40	90.78 ± 0.10	73.93 ± 2.05	92.49 ± 6.58	95.63 ± 0.12	81.68 ± 2.80
gyrA	94.82 ± 0.60	91.34 ± 0.53	52.71 ± 2.36	89.12 ± 0.26	77.88 ± 3.37	94.89 ± 0.99	96.33 ± 1.01	94.27 ± 0.19
embB	95.43 ± 0.28	86.18 ± 1.79	29.07 ± 2.63	58.80 ± 0.75	66.89 ± 2.76	92.55 ± 3.70	95.07 ± 0.70	65.01 ± 2.05
rpoB	93.79 ± 0.20	89.19 ± 0.43	41.11 ± 2.19	78.74 ± 0.26	74.84 ± 3.39	94.86 ± 0.96	91.63 ± 0.74	82.84 ± 1.85

Tabela 7: Pontuações ERRAT para diferentes genes em diversas ferramentas de predição de estruturas tridimensionais.

A.2 Verify3D

Gene	SWISS-MODEL	ColabFold	MODELLER	Phyre2	I-TASSER	trRosetta	AlphaFold3	OmegaFold
atpE	35.19 ± 6.52	39.71 ± 11.08	31.28 ± 9.58	22.64 ± 5.10	26.75 ± 10.98	44.24 ± 6.82	33.95 ± 14.11	44.86 ± 3.88
rpsL	68.75 ± 1.21	71.57 ± 2.02	24.19 ± 4.49	75.21 ± 0.41	66.13 ± 6.07	76.01 ± 1.53	67.54 ± 5.75	54.64 ± 3.92
ddn	80.13	80.13	64.90	65.22	91.39	84.11	79.47	80.13
Rv0678	73.45 ± 1.73	69.24 ± 4.89	76.36 ± 4.41	63.02 ± 2.54	75.08 ± 7.79	73.03 ± 4.88	69.70 ± 5.10	67.27 ± 4.17
pncA	66.54 ± 6.81	67.74 ± 2.52	69.55 ± 2.86	69.63 ± 2.13	72.78 ± 3.64	73.54 ± 3.70	72.49 ± 3.10	69.72 ± 3.39
rplC	84.79	74.65	62.21	73.71	83.41	76.50	80.18	71.43
gid	75.32 ± 2.02	77.78 ± 2.56	76.29 ± 2.78	69.37 ± 2.16	71.53 ± 7.54	79.76 ± 3.42	77.93 ± 2.97	80.21 ± 2.18
tlyA	64.18 ± 1.06	72.58 ± 0.26	33.77 ± 0.79	63.02 ± 0.00	70.15 ± 7.38	69.78 ± 0.53	70.15 ± 6.86	68.10 ± 0.56
inhA	67.63	62.45	51.30	73.88	78.81	71.00	58.74	59.85
ethA	87.84 ± 0.84	88.64 ± 0.68	59.83 ± 5.05	80.74 ± 5.73	85.77 ± 3.39	90.73 ± 1.76	89.14 ± 2.40	88.89 ± 1.53
gyrB	76.69 ± 0.52	76.24 ± 0.89	46.26 ± 1.85	76.84 ± 0.30	73.67 ± 1.23	79.96 ± 1.57	75.66 ± 1.33	79.04 ± 1.16
katG	94.69 ± 0.36	91.68 ± 1.49	70.27 ± 1.15	92.47 ± 0.00	91.65 ± 1.14	92.65 ± 2.10	93.65 ± 1.12	89.27 ± 0.63
gyrA	77.42 ± 0.82	73.51 ± 1.02	41.65 ± 2.37	68.47 ± 0.13	71.01 ± 1.55	79.36 ± 2.02	75.94 ± 0.52	71.41 ± 0.66
embB	64.08 ± 0.32	63.58 ± 1.08	42.18 ± 2.18	57.66 ± 0.50	36.21 ± 2.03	66.15 ± 1.68	65.95 ± 1.53	24.64 ± 1.96
rpoB	78.65 ± 0.18	79.93 ± 0.81	41.99 ± 3.34	78.05 ± 0.42	74.97 ± 4.70	78.93 ± 1.70	78.03 ± 1.95	61.72 ± 4.12

Tabela 8: Pontuações VERIFY3D para cada gene em diferentes ferramentas de modelagem de proteínas.

A.3 MolProbity

Gene	SWISS-MODEL	ColabFold	MODELLER	Phyre2	I-TASSER	trRosetta	AlphaFold3	OmegaFold
atpE	0.61 ± 0.00	2.51 ± 0.17	2.23 ± 0.12	2.02 ± 0.08	1.84 ± 0.06	0.74 ± 0.26	0.94 ± 0.36	2.19 ± 0.11
rpsL	0.79 ± 0.06	2.22 ± 0.12	3.18 ± 0.39	2.08 ± 0.01	2.44 ± 0.07	1.08 ± 0.19	1.31 ± 0.13	3.14 ± 0.10
ddn	0.84	1.94	2.63	1.99	2.62	0.65	1.20	2.58
Rv0678	1.39 ± 0.13	2.24 ± 0.09	3.54 ± 0.32	1.83 ± 0.07	2.30 ± 0.14	0.75 ± 0.18	1.09 ± 0.16	2.00 ± 0.12
pncA	1.00 ± 0.14	1.92 ± 0.07	2.70 ± 0.13	2.61 ± 0.11	2.68 ± 0.12	1.16 ± 0.19	1.42 ± 0.12	2.10 ± 0.21
rplC	0.77	2.10	3.43	2.22	2.51	0.57	1.09	3.05
gid	0.63 ± 0.14	2.40 ± 0.05	2.46 ± 0.18	1.84 ± 0.08	3.21 ± 0.13	1.01 ± 0.19	1.70 ± 0.12	2.16 ± 0.04
tlyA	0.77 ± 0.04	2.18 ± 0.01	3.57 ± 0.18	2.74 ± 0.01	2.73 ± 0.04	0.70 ± 0.14	1.18 ± 0.00	1.76 ± 0.02
inhA	0.78	2.34	3.04	1.98	2.34	1.14	1.21	1.78
ethA	0.63 ± 0.07	2.09 ± 0.03	3.54 ± 0.11	3.09 ± 0.04	3.14 ± 0.08	0.90 ± 0.06	1.38 ± 0.08	1.87 ± 0.05
gyrB	1.00 ± 0.04	2.19 ± 0.03	3.28 ± 0.15	2.58 ± 0.01	2.76 ± 0.03	0.88 ± 0.06	1.26 ± 0.08	2.22 ± 0.03
katG	1.14 ± 0.17	2.26 ± 0.04	3.31 ± 0.08	2.02 ± 0.00	2.84 ± 0.08	1.31 ± 0.68	1.48 ± 0.06	3.25 ± 0.07
gyrA	0.80 ± 0.04	2.34 ± 0.03	3.25 ± 0.11	2.06 ± 0.00	2.69 ± 0.10	0.81 ± 0.08	1.39 ± 0.05	2.27 ± 0.02
embB	0.69 ± 0.01	2.38 ± 0.05	3.65 ± 0.09	2.65 ± 0.02	3.02 ± 0.10	1.18 ± 0.45	1.30 ± 0.15	3.76 ± 0.06
rpoB	0.85 ± 0.03	2.11 ± 0.02	3.50 ± 0.06	2.32 ± 0.02	2.86 ± 0.16	1.03 ± 0.12	1.35 ± 0.06	3.21 ± 0.04

Tabela 9: Pontuações MolProbity para cada gene em diferentes ferramentas de modelagem.

A.4 VoroMQA

Gene	SWISS-MODEL	ColabFold	MODELLER	Phyre2	I-TASSER	trRosetta	AlphaFold3	OmegaFold
atpE	0.17 ± 0.01	0.15 ± 0.01	0.15 ± 0.01	0.14 ± 0.01	0.14 ± 0.01	0.15 ± 0.01	0.14 ± 0.01	0.15 ± 0.01
rpsL	0.40 ± 0.00	0.38 ± 0.00	0.09 ± 0.01	0.39 ± 0.00	0.36 ± 0.02	0.40 ± 0.01	0.40 ± 0.00	0.36 ± 0.01
ddn	0.44	0.44	0.41	0.48	0.45	0.46	0.44	0.43
Rv0678	0.55 ± 0.00	0.37 ± 0.01	0.36 ± 0.02	0.37 ± 0.01	0.39 ± 0.02	0.40 ± 0.01	0.39 ± 0.01	0.39 ± 0.01
pncA	0.51 ± 0.04	0.51 ± 0.01	0.48 ± 0.01	0.44 ± 0.01	0.46 ± 0.01	0.50 ± 0.01	0.52 ± 0.01	0.51 ± 0.01
rplC	0.40	0.39	0.17	0.36	0.38	0.39	0.40	0.40
gid	0.57 ± 0.01	0.56 ± 0.00	0.54 ± 0.01	0.55 ± 0.00	0.42 ± 0.01	0.53 ± 0.01	0.55 ± 0.00	0.56 ± 0.01
tlyA	0.50 ± 0.00	0.51 ± 0.00	0.14 ± 0.00	0.39 ± 0.00	0.40 ± 0.02	0.50 ± 0.00	0.51 ± 0.00	0.51 ± 0.00
inhA	0.58	0.47	0.38	0.46	0.45	0.48	0.48	0.48
ethA	0.56 ± 0.00	0.55 ± 0.00	0.22 ± 0.01	0.38 ± 0.01	0.39 ± 0.01	0.54 ± 0.01	0.56 ± 0.00	0.56 ± 0.00
gyrB	0.48 ± 0.00	0.47 ± 0.00	0.20 ± 0.01	0.41 ± 0.00	0.38 ± 0.01	0.46 ± 0.00	0.49 ± 0.00	0.48 ± 0.00
katG	0.49 ± 0.00	0.46 ± 0.00	0.21 ± 0.00	0.45 ± 0.00	0.40 ± 0.01	0.44 ± 0.01	0.46 ± 0.00	0.43 ± 0.01
gyrA	0.50 ± 0.00	0.48 ± 0.00	0.21 ± 0.00	0.45 ± 0.00	0.39 ± 0.01	0.49 ± 0.00	0.49 ± 0.00	0.49 ± 0.00
embB	0.49 ± 0.00	0.48 ± 0.00	0.23 ± 0.01	0.42 ± 0.00	0.41 ± 0.01	0.47 ± 0.01	0.50 ± 0.00	0.23 ± 0.01
rpoB	0.46 ± 0.00	0.44 ± 0.00	0.16 ± 0.00	0.40 ± 0.00	0.38 ± 0.01	0.44 ± 0.00	0.46 ± 0.00	0.36 ± 0.01

Tabela 10: Pontuações VoroMQA para cada gene em diferentes ferramentas de predição de estruturas tridimensionáis de proteínas.

A.5 QMEAN-DisCo

Gene	SWISS-MODEL	ColabFold	MODELLER	Phyre2	I-TASSER	trRosetta	AlphaFold3	OmegaFold
atpE	0.76 ± 0.01	0.74 ± 0.01	0.75 ± 0.02	0.76 ± 0.02	0.72 ± 0.01	0.74 ± 0.01	0.69 ± 0.02	0.75 ± 0.01
rpsL	0.73 ± 0.00	0.72 ± 0.00	0.24 ± 0.03	0.68 ± 0.01	0.69 ± 0.01	0.73 ± 0.01	0.73 ± 0.00	0.58 ± 0.01
ddn	0.78	0.78	0.67	0.83	0.70	0.76	0.78	0.72
Rv0678	0.86 ± 0.01	0.80 ± 0.00	0.66 ± 0.04	0.85 ± 0.00	0.80 ± 0.01	0.78 ± 0.01	0.81 ± 0.01	0.80 ± 0.00
pncA	0.92 ± 0.05	0.92 ± 0.01	0.91 ± 0.01	0.85 ± 0.03	0.88 ± 0.01	0.85 ± 0.01	0.92 ± 0.01	0.88 ± 0.02
rplC	0.85	0.84	0.35	0.75	0.77	0.81	0.83	0.66
gid	0.90 ± 0.01	0.86 ± 0.00	0.87 ± 0.01	0.89 ± 0.00	0.64 ± 0.01	0.82 ± 0.01	0.87 ± 0.01	0.84 ± 0.00
tlyA	0.83 ± 0.00	0.81 ± 0.01	0.27 ± 0.01	0.77 ± 0.01	0.66 ± 0.01	0.81 ± 0.01	0.83 ± 0.00	0.82 ± 0.00
inhA	0.90	0.86	0.79	0.90	0.87	0.83	0.88	0.88
ethA	0.65 ± 0.00	0.64 ± 0.01	0.23 ± 0.01	0.51 ± 0.01	0.52 ± 0.01	0.65 ± 0.01	0.64 ± 0.01	0.64 ± 0.00
gyrB	0.82 ± 0.00	0.80 ± 0.00	0.39 ± 0.00	0.75 ± 0.00	0.63 ± 0.00	0.79 ± 0.00	0.80 ± 0.00	0.77 ± 0.01
katG	0.88 ± 0.00	0.85 ± 0.00	0.34 ± 0.01	0.87 ± 0.00	0.77 ± 0.01	0.83 ± 0.02	0.87 ± 0.00	0.65 ± 0.00
gyrA	0.83 ± 0.00	0.81 ± 0.00	0.34 ± 0.01	0.87 ± 0.00	0.65 ± 0.01	0.80 ± 0.00	0.83 ± 0.00	0.80 ± 0.00
embB	0.81 ± 0.01	0.78 ± 0.00	0.23 ± 0.01	0.73 ± 0.01	0.69 ± 0.01	0.77 ± 0.02	0.80 ± 0.00	0.28 ± 0.02
rpoB	0.86 ± 0.00	0.84 ± 0.00	0.23 ± 0.01	0.82 ± 0.01	0.75 ± 0.02	0.82 ± 0.01	0.86 ± 0.00	0.58 ± 0.02

Tabela 11: Pontuações QMEAN-DisCo para cada gene em diferentes ferramentas de predição tridimensional.

A.6 QMEAN

Gene	SWISS-MODEL	ColabFold	MODELLER	Phyre2	I-TASSER	trRosetta	AlphaFold3	OmegaFold
atpE	-2.88 ± 0.13	-3.67 ± 0.24	-3.45 ± 0.38	-3.98 ± 0.06	-3.39 ± 0.44	-2.43 ± 0.11	-3.51 ± 0.17	0.10 ± 0.00
rpsL	-0.56 ± 0.07	-0.51 ± 0.28	-6.31 ± 0.84	-2.03 ± 0.07	-2.32 ± 0.45	0.69 ± 0.36	-0.12 ± 0.18	0.10 ± 0.00
ddn	0.33	1.15	-1.93	-0.70	-1.68	1.61	1.06	0.07
Rv0678	-0.63 ± 0.26	-1.81 ± 0.25	-4.52 ± 0.80	-0.28 ± 0.12	-1.66 ± 0.31	0.51 ± 0.33	-0.81 ± 0.39	0.07 ± 0.00
pncA	-0.44 ± 0.31	-0.22 ± 0.16	-2.39 ± 0.34	-2.25 ± 0.45	-2.74 ± 0.51	0.71 ± 0.41	0.10 ± 0.17	0.06 ± 0.00
rplC	0.66	0.45	-7.82	-0.40	-2.08	1.16	0.72	0.06
gid	0.59 ± 0.25	-1.03 ± 0.31	-0.48 ± 0.35	0.77 ± 0.20	-5.88 ± 0.94	0.87 ± 0.42	-0.51 ± 0.19	0.06 ± 0.00
tlyA	-0.22 ± 0.01	-0.40 ± 0.01	-10.47 ± 0.08	-3.62 ± 0.09	-4.13 ± 0.66	0.74 ± 0.13	0.27 ± 0.15	0.05 ± 0.00
inhA	1.13	-0.18	-0.99	0.76	-1.02	0.26	0.22	0.05
ethA	1.42 ± 0.14	1.69 ± 0.11	-9.66 ± 0.54	-6.88 ± 0.53	-8.29 ± 0.67	2.60 ± 0.20	1.60 ± 0.11	0.05 ± 0.00
gyrB	-0.19 ± 0.03	-1.15 ± 0.15	-6.43 ± 0.36	-1.88 ± 0.19	-4.95 ± 0.23	1.51 ± 0.15	0.61 ± 0.13	0.05 ± 0.00
katG	-0.02 ± 0.31	-1.01 ± 0.26	-7.10 ± 0.42	0.14 ± 0.04	-3.19 ± 0.48	0.49 ± 1.02	-0.03 ± 0.10	0.05 ± 0.00
gyrA	-0.10 ± 0.07	-1.75 ± 0.12	-6.78 ± 0.36	0.51 ± 0.03	-3.51 ± 0.25	1.66 ± 0.20	0.15 ± 0.12	0.05 ± 0.00
embB	-0.56 ± 0.06	-1.55 ± 0.07	-9.13 ± 0.27	-3.52 ± 0.07	-5.28 ± 0.30	0.03 ± 0.95	-0.32 ± 0.09	0.05 ± 0.00
rpoB	0.63 ± 0.05	-0.37 ± 0.08	-8.31 ± 0.38	-0.86 ± 0.09	-3.50 ± 0.90	1.79 ± 0.35	0.95 ± 0.08	0.05 ± 0.00

Tabela 12: Pontuações QMEAN para cada gene em diferentes ferramentas de predição de estruturas tridimensionais.

REFERÊNCIAS

[1] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. (2024a). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500.

[2] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. (2024b). Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3.

[3] Adebiyi, M. and O. Olugbara, O. (2021). Binding site identification of COVID-19 main protease 3D structure by homology modeling. *Indonesian Journal of Electrical Engineering and Computer Science*, 21(3):1713.

[4] Agnihotry, S., Pathak, R. K., Singh, D. B., Tiwari, A., and Hussain, I. (2022). Protein structure prediction. In *Bioinformatics*, pages 177–188. Elsevier.

[5] Ahmed, A., Saeed, F., Salim, N., and Abdo, A. (2014). Condorcet and borda count fusion method for ligand-based virtual screening. *Journal of Cheminformatics*, 6(1):1–10.

[6] Alhumaid, N. K. and Tawfik, E. A. (2024). Reliability of AlphaFold2 Models in Virtual Drug Screening: A Focus on Selected Class A GPCRs. *International Journal of Molecular Sciences*, 25(18):10139.

[7] Anfinsen, C. B. (1972). The formation and stabilization of protein structure. *Biochemical Journal*, 128(4):737–749.

[8] Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230.

[9] Antunes, D. (2025). Alphafold e a obsolescência dos biólogos computacionais (sqn). https://sbi.org.br/sblogi/alphafold-e-a-obsolescencia-dos-biologos-computacionais-sqn/. Acesso em: 01 set. 2025.

[10] Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201.

[11] Azinas, S. and Carroni, M. (2023). Cryo-EM uniqueness in structure determination of macromolecular complexes: A selected structural anthology. *Current Opinion in Structural Biology*, 81:102621.

[12] Barbarin-Bocahu, I. and Graille, M. (2022). The x-ray crystallography phase problem solved thanks to alphafold and rosettafold models: a case-study report. *Acta Crystallographica Section D Structural Biology*, 78:517–531.

[13] Baynes, J. and Dominiczak, M. (2019). *Bioquímica Médica*, volume 5 ed. Elsevier Health Sciences, Rio de Janeiro.

[14] Benhabiles, H., Jia, J., and Lejeune, F. (2016). General Aspects Related to Nonsense Mutations. In *Nonsense Mutation Correction in Human Diseases*, pages 1–76. Elsevier.

[15] Benkert, P., Biasini, M., and Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, 27(3):343–350.

[16] Berg, Jeremy M.; Tymoczko, J. L., Gatto Jr., G. J., and Stryer, L. (2021). *Bioquímica*, volume 9 ed. Grupo GEN, Rio de Janeiro.

[17] Bergendahl, L. T., Gerasimavicius, L., Miles, J., Macdonald, L., Wells, J. N., Welburn, J. P. I., and Marsh, J. A. (2019). The role of protein complexes in human genetic disease. *Protein Science*, 28(8):1400–1411.

[18] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

[19] Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., and Tawfik, D. S. (2006). Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, 444(7121):929–932.

[20] Bertoline, L. M., Lima, A. N., Krieger, J. E., and Teixeira, S. K. (2023). Before and after AlphaFold2: An overview of protein structure prediction.

[21] Bhagavan, N. and Ha, C.-E. (2011). Three-Dimensional Structure of Proteins. In *Essentials of Medical Biochemistry*, pages 29–38. Elsevier.

[22] Bhattacharya, A., Wunderlich, Z., Monleon, D., Tejero, R., and Montelione, G. T. (2008). Assessing model accuracy using the homology modeling automatically software. *Proteins: Structure, Function, and Bioinformatics*, 70(1):105–118.

[23] Bhattacharya, S., Roche, R., Shuvo, M. H., Moussad, B., and Bhattacharya, D. (2023). Contact-Assisted Threading in Low-Homology Protein Modeling. *Methods in Molecular Biology*, 2627:41–59.

[24] Bijak, V., Szczygiel, M., Lenkiewicz, J., Gucwa, M., Cooper, D. R., Murzyn, K., and Minor, W. (2023). The current role and evolution of X-ray crystallography in drug discovery and development.

[25] Blanco, A. and Blanco, G. (2022). The genetic information (I). In *Medical Biochemistry*, number I, pages 501–534. Elsevier.

[26] Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., and Schwede, T. (2009). Protein structure homology modeling using SWISS-MODEL workspace. *Nature Protocols*, 4(1):1–13.

[27] BORGES-OSÓRIO, M. R. L. and ROBINSON, W. M. (2013). *Genética humana*. Art-Med, Porto Alegre, 3 edition. E-book. p.751. ISBN 9788565852906.

[28] Bowie, J. U., Lüthy, R., and Eisenberg, D. (1991). A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science*, 253(5016):164–170.

[29] Brito, J. A. and Archer, M. (2013). X-ray Crystallography. In *Practical Approaches to Biological Inorganic Chemistry*, pages 217–255. Elsevier.

[30] Buel, G. R. and Walters, K. J. (2022). Can AlphaFold2 predict the impact of missense mutations on structure? *Nature Structural & Molecular Biology*, 29(1):1–2.

[31] Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Costanzo, L. D., Christie, C., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranovic, V., Guzenko, D., Hudson, B. P., Liang, Y., Lowe, R., Peisach, E., Periskova, I., Randle, C., Rose, A., Sekharan, M., Shao, C., Tao, Y.-P., Valasatava, Y., Voigt, M., Westbrook, J., Young, J., Zardecki, C., Zhuravleva, M., Kurisu, G., Nakamura, H., Kengaku, Y., Cho, H., Sato, J., Kim, J. Y., Ikegawa, Y., Nakagawa, A., Yamashita, R., Kudou, T., Bekker, G.-J., Suzuki, H., Iwata, T., Yokochi, M., Kobayashi, N., Fujiwara, T., Velankar, S., Kleywegt, G. J., Anyango, S., Armstrong, D. R., Berrisford, J. M., Conroy, M. J., Dana, J. M., Deshpande, M., Gane, P., Gáborová, R., Gupta, D., Gutmanas, A., Koča, J., Mak, L., Mir, S., Mukhopadhyay, A., Nadzirin, N., Nair, S., Patwardhan, A., Paysan-Lafosse, T., Pravda, L.,

Salih, O., Sehnal, D., Varadi, M., Vařeková, R., Markley, J. L., Hoch, J. C., Romero, P. R., Baskaran, K., Maziuk, D., Ulrich, E. L., Wedell, J. R., Yao, H., Livny, M., and Ioannidis, Y. E. (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, 47(D1):D520–D528.

[32] Burley, S. K., Berman, H. M., Duarte, J. M., Feng, Z., Flatt, J. W., Hudson, B. P., Lowe, R., Peisach, E., Piehl, D. W., Rose, Y., Sali, A., Sekharan, M., Shao, C., Vallat, B., Voigt, M., Westbrook, J. D., Young, J. Y., and Zardecki, C. (2022). Protein Data Bank: A Comprehensive Review of 3D Structure Holdings and Worldwide Utilization by Researchers, Educators, and Students. *Biomolecules*, 12(10):1425.

[33] Buxbaum, E. (2015). *Fundamentals of Protein Structure and Function*. Springer International Publishing, Cham.

[34] Callaway, E. (2020). Revolutionary cryo-EM is taking over structural biology. *Nature*, 578(7794):201–201.

[35] Cao, R., Wang, Z., Wang, Y., and Cheng, J. (2014). Smoq: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC Bioinformatics*, 15:120.

[36] Cardoso, A. P., Rabello, E., de Q. Mello, F. C., et al. (2021). *Diagnóstico e tratamento em pneumologia*. Manole, Barueri, 2 edition. E-book, p. 201. ISBN 9786555764383. Disponível em: `https://integrada.minhabiblioteca.com.br/reader/books/9786555764383/`. Acesso em: 02 jul. 2025.

[37] CAREY, F. A. (2011). *Química orgânica. V.2*. AMGH, Porto Alegre, 7 edition. E-book. p.1212. ISBN 9788580550542.

[38] Carroni, M. and Saibil, H. R. (2016). Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods*, 95(2016):78–85.

[39] Casem, M. L. (2016). Proteins. In *Case Studies in Cell Biology*, pages 23–71. Elsevier.

[40] Castells-Graells, R., Meador, K., Arbing, M. A., Sawaya, M. R., Gee, M., Cascio, D., Gleave, E., Debreczeni, J. É., Breed, J., Leopold, K., Patel, A., Jahagirdar, D., Lyons, B., Subramaniam, S., Phillips, C., and Yeates, T. O. (2023). Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold. *Proceedings of the National Academy of Sciences*, 120(37):1–8.

[41] Chandel, N. S. (2021). Amino Acid Metabolism. *Cold Spring Harbor Perspectives in Biology*, 13(4):a040584.

[42] Chang, I., Cieplak, M., Dima, R. I., Maritan, A., and Banavar, J. R. (2001). Protein threading by learning. *Proceedings of the National Academy of Sciences*, 98(25):14350–14355.

[43] Chari, A. and Stark, H. (2023). Prospects and Limitations of High-Resolution Single-Particle Cryo-Electron Microscopy. *Annual Review of Biophysics*, 52(1):391–411.

[44] Chatterjee, S. and Yadav, S. (2019). The Origin of Prebiotic Information System in the Peptide/RNA World: A Simulation Model of the Evolution of Translation and the Genetic Code. *Life*, 9(1):25.

[45] Chen, L., Li, Q., Nasif, K. F. A., Xie, Y., Deng, B., Niu, S., Pouriyeh, S., Dai, Z., Chen, J., and Xie, C. Y. (2024). AI-Driven Deep Learning Techniques in Protein Structure Prediction. *International Journal of Molecular Sciences*, 25(15):1–21.

[46] Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010). MolProbity : all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D Biological Crystallography*, 66(1):12–21.

[47] Chen, V. B., Wedell, J. R., Wenger, R. K., Ulrich, E. L., and Markley, J. L. (2015). MolProbity for the masses–of data. *Journal of Biomolecular NMR*, 63(1):77–83.

[48] Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, 357(6379):543–544.

[49] Church, D. D., Hirsch, K. R., Park, S., Kim, I.-Y., Gwin, J. A., Pasiakos, S. M., Wolfe, R. R., and Ferrando, A. A. (2020). Essential Amino Acids and Protein Synthesis: Insights into Maximizing the Muscle and Whole-Body Response to Feeding. *Nutrients*, 12(12):3717.

[50] Ciarleglio, A. (2021). Measures of variability and precision in statistics: appreciating, untangling and applying concepts. *BJPsych Advances*, 27(2):137–139.

[51] Clark, D. and Pazdernik, N. (2012). *Molecular Biology*. Academic Cell.

[52] Clark, D. and Pazdernik, N. (2015). *Biotechnology*. Academic Cell.

[53] Clark, D. P., Pazdernik, N. J., and McGehee, M. R. (2019). Mutations and Repair. In *Molecular Biology*, pages 832–879. Elsevier.

[54] Colicchio, T. (2020). *Introdução à informática em saúde: Fundamentos, aplicações e lições aprendidas com a informatização do sistema de saúde americano*. Artmed Editora, Porto Alegre.

[55] Colovos, C. and Yeates, T. O. (1993). Verification of protein structures: Patterns of non-bonded atomic interactions. *Protein Science*, 2(9):1511–1519.

[56] Cymerman, I. A., Feder, M., PawŁowski, M., Kurowski, M. A., and Bujnicki, J. M. (2008). Computational methods for protein structure prediction and fold recognition. In *Computational Methods for Protein Structure Prediction and Fold Recognition*, volume 1, pages 1–21.

[57] Datta, D., Jamwal, S., Jyoti, N., Patnaik, S., and Kumar, D. (2024). Actionable mechanisms of drug tolerance and resistance in Mycobacterium tuberculosis. *The FEBS Journal*, 291(20):4433–4452.

[58] Dattani, S., Spooner, F., Ritchie, H., and Roser, M. (2023). Tuberculosis. *Our World in Data*. https://ourworldindata.org/tuberculosis.

[59] Davis, T. H. (2004). Meselson and stahl: The art of dna replication. *Proceedings of the National Academy of Sciences*, 101:17895–17896.

[60] Desai, D., Kantliwala, S. V., Vybhavi, J., Ravi, R., Patel, H., and Patel, J. (2024). Review of AlphaFold 3: Transformative Advances in Drug Design and Therapeutics. *Cureus*.

[61] Dheda, K., Mirzayev, F., Cirillo, D. M., Udwadia, Z., Dooley, K. E., Chang, K.-C., Omar, S. V., Reuter, A., Perumal, T., Horsburgh, C. R., Murray, M., and Lange, C. (2024). Multidrug-resistant tuberculosis. *Nature Reviews Disease Primers*, 10(1):22.

[62] Diez, P. (2018). Introduction. In *Smart Wheelchairs and Brain-Computer Interfaces*, pages 1–21. Elsevier, second edi edition.

[63] Ditse, Z., Lamers, M. H., and Warner, D. F. (2017). Dna replication in mycobacterium tuberculosis. *Microbiology Spectrum*, 5.

[64] Doherty, J. and Guo, M. (2016). Transfer RNA. In *Encyclopedia of Cell Biology*, volume 1, pages 309–340. Elsevier.

[65] dos Santos de Lemos, A. and Lins, R. S. (2023). *Doenças infecciosas na emergência: diagnóstico e tratamento*. Manole, Barueri. E-book, p. 283. ISBN 9786555763232. Available at: `https://integrada.minhabiblioteca.com.br/reader/books/9786555763232/`. Accessed: July 2, 2025.

[66] Du, Z., Su, H., Wang, W., Ye, L., Wei, H., Peng, Z., Anishchenko, I., Baker, D., and Yang, J. (2021). The trRosetta server for fast and accurate protein structure prediction. *Nature Protocols*, 16(12):5634–5651.

[67] Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., and Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):e1005659.

[68] Eisenberg, D., Lüthy, R., and Bowie, J. U. (1997). [20] VERIFY3D: Assessment of protein models with three-dimensional profiles. In *Methods in enzymology*, volume 277, pages 396–404.

[69] Ellenbroek, B. and Youn, J. (2016). The Genetic Basis of Behavior. In *Gene-Environment Interactions in Psychiatry*, volume 15, pages 19–46. Elsevier.

[70] Ellmen, I., Raybould, M. I. J., and Deane, C. M. (2025). The protein universe in 3D. *Nature Chemical Biology*, 21(1):27–28.

[71] Emerson, P. (2013). The original Borda count and partial voting. *Social Choice and Welfare*, 40(2):353–358.

[72] Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M., Eramian, D., Shen, M.-y., Pieper, U., and Sali, A. (2006). Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics*, 15(1):1–30.

[73] Fajri, N. and Petryk, N. (2024). Monitoring and quantifying replication fork dynamics with high-throughput methods. *Communications Biology*, 7:729.

[74] Farhat, M., Cox, H., Ghanem, M., Denkinger, C. M., Rodrigues, C., Abd El Aziz, M. S., Enkh-Amgalan, H., Vambe, D., Ugarte-Gil, C., Furin, J., and Pai, M. (2024). Drug-resistant tuberculosis: a persistent global health concern. *Nature Reviews Microbiology*, 22(10):617–635.

[75] Finkelstein, A. V., Bogatyreva, N. S., Ivankov, D. N., and Garbuzynskiy, S. O. (2022). Protein folding problem: enigma, paradox, solution.

[76] Fiser, A., Do, R. K. G., and Šali, A. (2000). Modeling of loops in protein structures. *Protein Science*, 9(9):1753–1773.

[77] for Disease Control, C. and (CDC), P. About Tuberculosis — Tuberculosis (TB) — CDC.

[78] Fox, N. B. and Bruyns, B. (2025). An Evaluation of Borda Count Variations Using Ranked Choice Voting Data. *arXiv preprint arXiv:2501.00618*, (2024):1–19.

[79] Graille, M., Sacquin-Mora, S., and Taly, A. (2023). Best practices of using ai-based models in crystallography and their impact in structural biology.

[80] Griffiths, Anthony J F.; Doebley, J. and Peichel, C. e. a. (2022). *Introdução à Genética*, volume 12 ed. Grupo GEN, Rio de Janeiro.

[81] Guex, N., Peitsch, M. C., and Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *ELECTROPHORESIS*, 30(S1):162–173.

[82] Hameduh, T., Haddad, Y., Adam, V., and Heger, Z. (2020). Homology modeling in the time of collective and artificial intelligence. *Computational and Structural Biotechnology Journal*, 18:3494–3506.

[83] Helix, T. D. (2009). Biomolecular Principles: Nucleic Acids. In *Biomedical Engineering*, pages 82–140. Cambridge University Press.

[84] Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., and Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1):5979.

[85] Hiranuma, N., Park, H., Baek, M., Anishchanka, I., Dauparas, J., and Baker, D. (2020). Improved protein structure refinement guided by deep learning based accuracy estimation.

[86] Holloway-Kew, K. and Henneberg, M. (2023). Dynamics of tuberculosis infection in various populations during the 19th and 20th century: The impact of conservative and pharmaceutical treatments. *Tuberculosis*, 143(S):102389.

[87] Hou, K., Jabeen, R., Sun, L., and Wei, J. (2024). How do Mutations of Mycobacterium Genes Cause Drug Resistance in Tuberculosis? *Current Pharmaceutical Biotechnology*, 25(6):724–736.

[88] Huang, E. S., Samudrala, R., and Park, B. H. (2000). Scoring Functions for ab initio Protein Structure Prediction. In *Protein Structure Prediction*, volume 143, pages 223–245. Humana Press, New Jersey.

[89] Hunter, P. (2006). Into the fold. *EMBO reports*, 7:249–252.

[90] Hýskova, A., Maršálková, E., and Šimeček, P. (2025). Balancing Speed and Precision in Protein Folding: A Comparison of AlphaFold2, ESMFold, and OmegaFold.

[91] Ille, A. M., Lamont, H., and Mathews, M. B. (2022). The central dogma revisited: Insights from protein synthesis, crispr, and beyond. *WIREs RNA*, 13.

[92] Jahnke, W. (2007). Fragment-Based Approaches. In *Comprehensive Medicinal Chemistry II*, pages 939–957. Elsevier.

[93] Janssen, S., Murphy, M., Upton, C., Allwood, B., and Diacon, A. H. (2025). Tuberculosis: An Update for the Clinician. *Respirology*, 30(3):196–205.

[94] Jefferys, B. R., Kelley, L. A., and Sternberg, M. J. (2010). Protein Folding Requires Crowd Control in a Simulated Cell. *Journal of Molecular Biology*, 397(5):1329–1338.

[95] Jiang, R., Zhang, X., and Zhang, M. Q. (2013). *Basics of Bioinformatics*, volume 9783642389. Springer Berlin Heidelberg, Berlin, Heidelberg.

[96] Jianyi, Y., Anishchenkob, I., Hahnbeom, P., Pengd, Z., Ovchinnikove, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3).

[97] Jing, X., Dong, Q., Hong, D., and Lu, R. (2020). Amino Acid Encoding Methods for Protein Sequences: A Comprehensive Review and Assessment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6):1918–1931.

[98] Jisna, V. A. and Jayaraj, P. B. (2021). Protein Structure Prediction: Conventional and Deep Learning Perspectives. *Protein Journal*, 40(4):522–544.

[99] Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202.

[100] Jothi, A. (2012). Principles, Challenges and Advances in ab initio Protein Structure Prediction. *Protein & Peptide Letters*, 19(11):1194–1204.

[101] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

[102] Kadakeri, S., Arul, M. R., Bordett, R., Duraisamy, N., Naik, H., and Rudraiah, S. (2020). Protein synthesis and characterization. In *Artificial Protein and Peptide Nanofibers*, pages 121–161. Elsevier.

[103] Kaliva, M. and Vamvakaki, M. (2020). Nanomaterials characterization. In *Polymer Science and Nanotechnology*, pages 401–433. Elsevier.

[104] Keegan, R. M., Simpkin, A. J., and Rigden, D. J. (2024). The success rate of processed predicted models in molecular replacement: implications for experimental phasing in the alphafold era. *Acta Crystallographica Section D Structural Biology*, 80:766–779.

[105] Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6):845–858.

[106] Khan, F. I., Wei, D.-q., Gu, K.-r., Hassan, M. I., and Tabrez, S. (2016). Current updates on computer aided protein modeling and designing. *International Journal of Biological Macromolecules*, 85:48–62.

[107] Klug, W. S., Cummings, M. R., Spencer, C. A., et al. (2010). *Conceitos de Genética*. ArtMed, Porto Alegre, 9 edition. E-book. Acesso em: 25 mai. 2025.

[108] Koča, J., Svobodová Vařeková, R., Pravda, L., Berka, K., Geidl, S., Sehnal, D., and Otyepka, M. (2016). *Structural Bioinformatics Tools for Drug Design*. SpringerBriefs in Biochemistry and Molecular Biology. Springer International Publishing, Cham.

[109] Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics*, 77(4):778–795.

[110] Krokidis, M. G., Koumadorakis, D. E., Lazaros, K., Ivantsik, O., Exarchos, T. P., Vrahatis, A. G., Kotsiantis, S., and Vlamos, P. (2025). AlphaFold3: An Overview of Applications and Performance Insights. *International Journal of Molecular Sciences*, 26(8):3671.

[111] Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2021). Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins*, 89(12):1607–1617.

[112] Kumar, H. and Kim, P. (2024). Artificial intelligence in fusion protein three-dimensional structure prediction: Review and perspective. *Clinical and Translational Medicine*, 14(8).

[113] Kurt Yilmaz, N. and Schiffer, C. A. (2021). Introduction: Drug Resistance. *Chemical Reviews*, 121(6):3235–3237.

[114] Laboratory, U.-D. (2025). Saves v6.0: Structure analysis and verification server. `https://saves.mbi.ucla.edu/`. Accessed: 2025-07-17.

[115] Lawrence, A. (2024). Bacillus Calmette-Guérin (BCG) Revaccination and Protection Against Tuberculosis: A Systematic Review. *Cureus*, 16(3).

[116] Lazarus, R. M. (1999). Definition of sensitivity and specificity. *The American Journal of Clinical Nutrition*, 69(1):158.

[117] Lee, S. (2025). Verify3D: The Ultimate Tool for Protein Model Validation.

[118] Leo, S., Narasimhan, M., Rathinam, S., and Banerjee, A. (2024). Biomarkers in diagnosing and therapeutic monitoring of tuberculosis: a review. *Annals of Medicine*, 56(1):–.

[119] Li, R. and Sperling, A. (2013). Drug Resistance. In *Brenner's Encyclopedia of Genetics*, volume 1, pages 418–420. Elsevier.

[120] Li, S., Terashi, G., Zhang, Z., and Kihara, D. (2025). Advancing structure modeling from cryo-EM maps with deep learning. *Biochemical Society Transactions*, 53(01):259–265.

[121] Li, Z., Fan, H., and Ding, W. (2024). Solving protein structures by combining structure prediction, molecular replacement and direct-methods-aided model completion. *IUCrJ*, 11:152–167.

[122] Louten, J. (2016). Virus Replication. In *Essential Human Virology*, pages 49–70. Elsevier.

[123] Luscombe, N. M., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics? an introduction and overview. *Yearbook of medical informatics*, 10(01):83–100.

[124] Lüthy, R., Bowie, J. U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, 356(6364):83–85.

[125] Mackerell, A. D., Feig, M., and Brooks, C. L. (2004). Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry*, 25(11):1400–1415.

[126] Maia, R. T. (2024). Protein structure prediction by computational homology modeling: a brief explanation. *International Journal of Molecular Biology Open Access*, 7(1):118–120.

[127] Mancuso, G., Midiri, A., De Gaetano, S., Ponzo, E., and Biondo, C. (2023). Tackling Drug-Resistant Tuberculosis: New Challenges from the Old Pathogen Mycobacterium tuberculosis. *Microorganisms*, 11(9):2277.

[128] Mao, X., Wang, J., Xu, J., Xu, P., Hu, H., Li, L., Zhang, Z., and Song, Y. (2025). Current diagnosing strategies for Mycobacterium tuberculosis and its drug resistance: a review. *Journal of Applied Microbiology*, 136(5).

[129] Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728.

[130] Marintchev, A. (2012). *Fidelity and Quality Control in Gene Expression*. Number v. 86 in Advances in Protein Chemistry and Structural Biology. Elsevier Science.

[131] Marion, D. (2013). An Introduction to Biological NMR Spectroscopy. *Molecular Cellular Proteomics*, 12(11):3006–3025.

[132] MCMURRY, J. (2016). *Química Orgânica - Combo: Tradução da 9ª edição norte-americana*. +A Educação - Cengage Learning Brasil, Porto Alegre, 3 edition. E-book. p.1060. ISBN 9788522125876.

[133] Merz, K. M. and Le Grand, S. M. (1994). *The Protein Folding Problem and Tertiary Structure Prediction*, volume 4. Birkhäuser Boston, Boston, MA.

[134] Middaugh, C. R. and Pearlman, R. (1999). Proteins as Drugs: Analysis, Formulation and Delivery. In *Handbook of Experimental Pharmacology*, volume 199, pages 33–58.

[135] Millet, J.-P., Moreno, A., Fina, L., del Baño, L., Orcau, A., de Olalla, P. G., and Caylà, J. A. (2013). Factors that influence current tuberculosis epidemiology. *European Spine Journal*, 22(S4):539–548.

[136] Ministério da Saúde (2025). Tuberculose — Ministério da Saúde.

[137] Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682.

[138] Monaghan, T. F., Rahman, S. N., Agudelo, C. W., Wein, A. J., Lazar, J. M., Everaert, K., and Dmochowski, R. R. (2021). Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. *Medicina*, 57(5):503.

[139] Morris, R., Black, K. A., and Stollar, E. J. (2022). Uncovering protein function: from classification to complexes. *Essays in Biochemistry*, 66(3):255–285.

[140] Mundlapati, V. R., Sahoo, D. K., Ghosh, S., Purame, U. K., Pandey, S., Acharya, R., Pal, N., Tiwari, P., and Biswal, H. S. (2017). Spectroscopic Evidences for Strong Hydrogen Bonds with Selenomethionine in Proteins. *The Journal of Physical Chemistry Letters*, 8(4):794–800.

[141] Naidoo, K. and Perumal, R. (2023). Advances in tuberculosis control during the past decade. *The Lancet Respiratory Medicine*, 11(4):311–313.

[142] Nam, K. H. (2023). Radiation Damage on Selenomethionine-Substituted Single-Domain Substrate-Binding Protein. *Crystals*, 13(12):1620.

[143] Nassar, R., Dignon, G. L., Razban, R. M., and Dill, K. A. (2021). The protein folding problem: The role of theory. *Journal of Molecular Biology*, 433:167126.

[144] Nelson, David L.; Cox, M. M. e. A. A. H. (2022). *Princípios de Bioquímica de Lehninger*, volume 1. Artmed, Porto Alegre.

[145] Ó'Fágáin, C. (2017). Protein Stability: Enhancement and Measurement. In Walls, D. and Loughran, S. T., editors, *Protein Chromatography*, volume 1485 of *Methods in Molecular Biology*, pages 101–129. Springer New York, New York, NY.

[146] Ohno, S. (1970). *Evolution by Gene Duplication*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[147] Olechnovič, K. and Venclovas, Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins: Structure, Function, and Bioinformatics*, 85(6):1131–1145.

[148] Olechnovič, K. and Venclovas, Č. (2019). VoroMQA web server for assessing three-dimensional structures of proteins and protein complexes. *Nucleic Acids Research*, 47(W1):W437–W442.

[149] Ozturk, K. and Carter, H. (2022). Predicting functional consequences of mutations using molecular interaction network features. *Human Genetics*, 141(6):1195–1210.

[150] Paiva, V. d. A., Gomes, I. d. S., Monteiro, C. R., Mendonça, M. V., Martins, P. M., Santana, C. A., Gonçalves-Almeida, V., Izidoro, S. C., de Melo-Minardi, R. C., and Silveira, S. d. A. (2022). Protein structural bioinformatics: An overview. *Computers in Biology and Medicine*, 147(June).

[151] Pak, M. A., Markhieva, K. A., Novikova, M. S., Petrov, D. S., Vorobyev, I. S., Maksimova, E. S., Kondrashov, F. A., and Ivankov, D. N. (2023). Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLOS ONE*, 18(3):e0282689.

[152] Pan, Q., Nguyen, T. B., Ascher, D. B., and Pires, D. E. V. (2022). Systematic evaluation of computational tools to predict the effects of mutations on protein stability in the absence of experimental structures. *Briefings in Bioinformatics*, 23(2):1–13.

[153] Pavinati, G., de Lima, L. V., Bernardo, P. H. P., Dias, J. R., Reis-Santos, B., and Magnabosco, G. T. (2024). A critical analysis of the decreasing trends in tuberculosis cure indicators in Brazil, 2001–2022. *Jornal Brasileiro de Pneumologia*, 50(2):1–11.

[154] Pearce, R. and Zhang, Y. (2021). Toward the solution of the protein structure prediction problem. *Journal of Biological Chemistry*, 297(1):100870.

[155] Peng, J. and Xu, J. (2010). Low-homology protein threading. *Bioinformatics*, 26(12):i294–i300.

[156] Pereira, P. J. B., Royant, A., Panjikar, S., and de Sanctis, D. (2013). In-house UV radiation-damage-induced phasing of selenomethionine-labeled protein structures. *Journal of Structural Biology*, 181(2):89–94.

[157] Petsko, G. and Ringe, D. (2004). *Protein Structure and Function*. Primers in biology. New Science Press.

[158] Powell, H. R., Islam, S. A., David, A., and Sternberg, M. J. (2025). Phyre2.2: A Community Resource for Template-based Protein Structure Prediction. *Journal of Molecular Biology*.

[159] Protein Data Bank (2025). PDB Statistics: Overall Growth of Released Structures Per Year.

[160] Qiu, X., Li, H., Ver Steeg, G., and Godzik, A. (2024). Advances in AI for Protein Structure Prediction: Implications for Cancer Drug Discovery and Development. *Biomolecules*, 14(3):339.

[161] Qu, X., Swanson, R., Day, R., and Tsai, J. (2009). A Guide to Template Based Structure Prediction. *Current Protein Peptide Science*, 10(3):270–285.

[162] Raisinghani, N., Alshahrani, M., Gupta, G., Tian, H., Xiao, S., Tao, P., and Verkhivker, G. (2024). Probing Functional Allosteric States and Conformational Ensembles of the Allosteric Protein Kinase States and Mutants: Atomistic Modeling and Comparative Analysis of AlphaFold2, OmegaFold, and AlphaFlow Approaches and Adaptations. *The Journal of Physical Chemistry B*, 128(45):11088–11107.

[163] Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2):173–175.

[164] Renaud, J.-P. (2020). *Structural biology in drug discovery : methods, techniques, and practices*. John Wiley Sons.

[165] Rigden, D. J. (2017). *From Protein Structure to Function with Bioinformatics*. Springer Netherlands, Dordrecht.

[166] Robert, X., Guillon, C., and Gouet, P. (2025). FoldScript: a web server for the efficient analysis of AI-generated 3D protein models. *Nucleic Acids Research*, 53(W1):W277–W282.

[167] Rodwell, V., Bender, D., Botham, K., Kennelly, P., and Weil, P. (2021). *Bioquímica Ilustrada de Harper*, volume 31 ed. McGraw Hill Brasil, Porto Alegre.

[168] Roterman-Konieczna, I. (2012). A short description of other selected ab initio methods for protein structure prediction. In *Protein Folding in Silico*, pages 165–189. Elsevier.

[169] Šali, A. and Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, 234(3):779–815.

[170] Sandy, S. and Winanda, E. (2022). Homology Modeling Odorant-binding Protein-1 (OBP1) Anopheles Farauti Protein Target for Mosquito Repellent. *Biomedical and Pharmacology Journal*, 15(3):1759–1768.

[171] Sasin, J. M. and Bujnicki, J. M. (2004). Colorado3d, a web server for the visual analysis of protein structures. *Nucleic Acids Research*, 32:W586–W589.

[172] Sathyanarayanan, S. (2024). Confusion Matrix-Based Performance Evaluation Metrics. *African Journal of Biomedical Research*, pages 4023–4031.

[173] Sayers, E. W., Beck, J., Bolton, E. E., Brister, J. R., Chan, J., Connor, R., Feldgarden, M., Fine, A. M., Funk, K., Hoffman, J., Kannan, S., Kelly, C., Klimke, W., Kim, S., Lathrop, S., Marchler-Bauer, A., Murphy, T. D., O'Sullivan, C., Schmieder, E., Skripchenko, Y., Stine, A., Thibaud-Nissen, F., Wang, J., Ye, J., Zellers, E., Schneider, V. A., and Pruitt, K. D. (2025). Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Research*, 53(D1):D20–D29.

[174] Schelde, A. B. and Kornholt, J. (2021). Validation studies in epidemiologic research: estimation of the positive predictive value. *Journal of Clinical Epidemiology*, 137:262–264.

[175] Schneider, M., Belsom, A., and Rappsilber, J. (2018). Protein Tertiary Structure by Crosslinking/Mass Spectrometry. *Trends in Biochemical Sciences*, 43(3):157–169.

[176] Seo, K. and Ichihashi, N. (2023). Investigation of compatibility between dna replication, transcription, and translation for in vitro central dogma. *ACS Synthetic Biology*, 12:1813–1822.

[177] Shanker, A. (2018). *Bioinformatics: Sequences, Structures, Phylogeny*. Springer Singapore, Singapore.

[178] Sharif, M. M., Tharwat, A., Hassanien, A. E., Hefny, H. A., and Schaefer, G. (2016). Enzyme Function Classification Based on Borda Count Ranking Aggregation Method. *Studies in Big Data*, 19:75–85.

[179] Sharma, N., Chauhan, P., and Shakya, M. (2025). Analysis for a 3D structure of a protein through molecular modelling and simulation. *Bulletin of Biomathematics*, 3(1):21–36.

[180] Smolarczyk, T., Roterman-Konieczna, I., and Stapor, K. (2020). Protein Secondary Structure Prediction: A Review of Progress and Directions. *Current Bioinformatics*, 15(2):90–107.

[181] Smyth, M. S. and Martin, J. H. J. (2000). x ray crystallography. *Molecular Pathology*, 53:8–14.

[182] Snyder, L. and Snyder, L. (2024). *Bacterial Genetics and Genomics*. CRC Press.

[183] Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics*, 21(7):951–960.

[184] Srivastava, A. A., Nagai, T., Srivastava, A. A., Miyashita, O., and Tama, F. (2018). Role of Computational Methods in Going beyond X-ray Crystallography to Explore Protein Structure and Dynamics. *International Journal of Molecular Sciences*, 19(11):3401.

[185] Strokach, A., Corbi-Verge, C., Teyra, J., and Kim, P. M. (2019). Predicting the Effect of Mutations on Protein Folding and Protein-Protein Interactions. In *Methods in Molecular Biology*, volume 1851, pages 1–17.

[186] Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., and Schwede, T. (2020). QMEANDisCo—distance constraints applied on model quality estimation. *Bioinformatics*, 36(6):1765–1771.

[187] Studer, G., Tauriello, G., Bienert, S., Biasini, M., Johner, N., and Schwede, T. (2021). ProMod3—A versatile homology modelling toolbox. *PLOS Computational Biology*, 17(1):e1008667.

[188] Sun, Q., Li, S., Gao, M., and Pang, Y. (2024). Therapeutic Strategies for Tuberculosis: Progress and Lessons Learned. *Biomedical and environmental sciences : BES*, 37(11):1310–1323.

[189] Syed Ibrahim, K., Gurusubramanian, G., Zothansanga, Yadav, R. P., Senthil Kumar, N., Pandian, S. K., Borah, P., and Mohan, S. (2017a). *Bioinformatics - A Student's Companion*. Springer Singapore, Singapore.

[190] Syed Ibrahim, K., Gurusubramanian, G., Zothansanga, Yadav, R. P., Senthil Kumar, N., Pandian, S. K., Borah, P., and Mohan, S. (2017b). *Bioinformatics - A Student's Companion*. Springer Singapore, Singapore.

[191] Tavares, R. B. V., Berra, T. Z., Alves, Y. M., Popolin, M. A. P., Ramos, A. C. V., Tártaro, A. F., de Souza, C. F., and Arcêncio, R. A. (2024). Unsuccessful tuberculosis treatment outcomes across Brazil's geographical landscape before and during the COVID-19 pandemic: are we truly advancing toward the sustainable development/end TB goal? *Infectious Diseases of Poverty*, 13(1):17.

[192] Terwilliger, T. C., Liebschner, D., Croll, T. I., Williams, C. J., McCoy, A. J., Poon, B. K., Afonine, P. V., Oeffner, R. D., Richardson, J. S., Read, R. J., and Adams, P. D.

(2024). AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods*, 21(1):110–116.

[193] Thompson, R. A., Spring, A. M., Sheng, J., Huang, Z., and Germann, M. W. (2015). The importance of fitting in: conformational preference of selenium 2 modifications in nucleosides and helical structures. *Journal of Biomolecular Structure and Dynamics*, 33(2):289–297.

[194] Tomii, K. (2019). Protein Properties. In *Encyclopedia of Bioinformatics and Computational Biology*, volume 1-3, pages 28–33. Elsevier.

[195] Tong, J. C. and Ranganathan, S. (2013). Scientific publications and databases. In *Computer-Aided Vaccine Design*, pages 21–46. Elsevier.

[196] Tzul, F. O., Vasilchuk, D., and Makhatadze, G. I. (2017). Evidence for the principle of minimal frustration in the evolution of protein folding landscapes. *Proceedings of the National Academy of Sciences*, 114(9):E1627–E16322.

[197] Urry, Lisa A.; Cain, M. L., Wasserman, S. A., and et al. (2022). *Biologia de Campbell*, volume 12 ed. Grupo A, Porto Alegre.

[198] van Ravenzwaaij, D. and Ioannidis, J. P. A. (2019). True and false positive rates for different criteria of evaluating statistical evidence from clinical trials. *BMC Medical Research Methodology*, 19(1):218.

[199] Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology*, 13(1):e1005324.

[200] Wanger, A., Chavez, V., Huang, R. S., Wahed, A., Actor, J. K., and Dasgupta, A. (2017). Overview of Bacteria. In *Microbiology and Molecular Diagnosis in Pathology*, pages 75–117. Elsevier.

[201] Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., and Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1):W296–W303.

[202] Webb, B. and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, 54(1):1–55.

[203] WILCZYNSKI, S. P. (2009). Molecular Biology. In *Modern Surgical Pathology*, volume 1, pages 85–120. Elsevier.

[204] Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S., and Richardson, D. C. (2018). MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science*, 27(1):293–315.

[205] Wolynes, P. G. (2015). Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie*, 119:218–230.

[206] World Health Organization (2015). *Implementing the End TB Strategy: The Essentials*. World Health Organization, Geneva, Switzerland. WHO/HTM/TB/2015.31.

[207] World Health Organization (2020). *WHO consolidated guidelines on tuberculosis. Module 4: treatment - drug-resistant tuberculosis treatment*. World Health Organization, Geneva.

[208] World Health Organization (2021). *Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance*, volume 2. Geneva.

[209] World Health Organization (2024). *WHO: operational handbook on tuberculosis. Module 1: prevention - tuberculosis preventive treatment, second edition*. Geneva.

[210] Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., and Peng, J. (2022). High-resolution de novo structure prediction from primary sequence.

[211] Wuyun, Q., Chen, Y., Shen, Y., Cao, Y., Hu, G., Cui, W., Gao, J., and Zheng, W. (2024). Recent Progress of Protein Tertiary Structure Prediction.

[212] Xu, Q., Dai, H., Zhao, T., and Wei, D. (2015). Introduction to Structural Bioinformatics. In *Advances in Experimental Medicine and Biology*, volume 827, pages 1–7.

[213] Zhan, Z.-H., Hong, J., Li, J.-Y., Wang, C., He, L., Xu, Z., and Zhang, J. (2025). Artificial intelligence-based methods for protein structure prediction: a survey. *Artificial Intelligence Review*, 58(10):328.

[214] Zhang, W., Yang, J., He, B., Walker, S. E., Zhang, H., Govindarajoo, B., Virtanen, J., Xue, Z., Shen, H. B., and Zhang, Y. (2016). Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11. *Proteins: Structure, Function and Bioinformatics*, 84(S1):76–86.

[215] Zhang, Y., Zheng, W., and Li, Y. (2021). Recent progress in protein structure prediction and its impact on structural biology. *Journal of Protein Chemistry*, 40:1–15.

[216] Zheng, W., Wuyun, Q., Li, Y., Liu, Q., Zhou, X., Peng, C., Zhu, Y., Freddolino, L., and Zhang, Y. (2025). Deep-learning-based single-domain and multidomain protein structure prediction with D-I-TASSER. *Nature Biotechnology*.

[217] Zheng, W., Zhang, C., Li, Y., Pearce, R., Bell, E. W., and Zhang, Y. (2021). Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Reports Methods*, 1(3):100014.

[218] Zhou, X., Zheng, W., Li, Y., Pearce, R., Zhang, C., Bell, E. W., Zhang, G., and Zhang, Y. (2022). I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. *Nature Protocols*, 17(10):2326–2353.