UNIVERSIDADE FEDERAL DE PELOTAS

Programa de Pós-Graduação em Biotecnologia

Tese

Desenvolvimento de uma pipeline para vacinologia reversa baseada em arquitetura *transformer*

Amanda Munari Guimarães

Pelotas, 2025

Amanda Munari Guimarães

Desenvolvimento de uma pipeline para vacinologia reversa baseada em arquitetura *transformer*

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Doutora em Ciências (área do Conhecimento: Bioinformática).

Orientador: Dr. Frederico Schmitt Kremer
Coorientador: Dr. Luciano da Silva Pinto

Pelotas, 2025

BANCA EXAMINADORA

Prof. Dr. Frederico Schmitt Kremer

Prof. Dr. Luciano da Silva Pinto

Prof. Dr. Ulisses Brisolara Correa

Prof. Dr. Alan McBride

Dr. Rafael Woloski

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

G963d Guimarães, Amanda Munari

Desenvolvimento de uma pipeline para vacinologia reversa baseada em arquitetura *transformer* [recurso eletrônico] / Amanda Munari Guimarães ; Frederico Schmitt Kremer, orientador. — Pelotas, 2025.
94 f.

Tese (Doutorado) — Programa de Pós-Graduação em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2025.

1. Bioinformática. 2. Proteoma. 3. Processamento de linguagem natural. 4. Bactérias. 5. Patógenos. I. Kremer, Frederico Schmitt, orient. II. Título.

CDD 615.372

Elaborada por Gabriela Machado Lopes CRB: 10/1842

Para minha filha Esther, com imensa gratidão e amor.

Dedico.

Agradecimentos

Primeiramente, agradeço à Universidade Federal de Pelotas, a qual possibilitou o meu aperfeiçoamento intelectual e científico, através do contato com os professores, alunos, técnicos e funcionários. Minha história com a universidade e o meio acadêmico iniciou em 2014 e se encerra após 11 anos, tendo passado pela graduação, mestrado e doutorado. Serei sempre muito grata.

Aos meus pais, por me darem a oportunidade de viver, por terem me dado amor e carinho, por terem permitido que eu crescesse com conforto, boas oportunidades e rodeada de pessoas maravilhosas. Sou eternamente grata à minha mãe pelas noites mal dormidas, pelas madrugadas nos plantões, pelos abraços nos momentos mais críticos, pelas comemorações nos momentos de alegria e por sempre apoiar minhas escolhas, fazendo o possível, e às vezes o impossível, para eu conquistar meus sonhos. Por ter me ensinado a ser responsável, ter caráter, ser uma pessoa que sempre busca o melhor em cada oportunidade.

Ao Lucas Farias, meu marido, que desde o início esteve ao meu lado me dando suporte, força e me incentivando a sempre buscar a melhor jornada. Obrigada por tornar a vida mais leve, alegre e mais fácil. Nossa união trouxe tranquilidade e muito propósito de vida.

À minha filha, Esther, que me trás tanta alegria e força, além de ter me proporcionado um incentivo imenso apenas por existir.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código do Financiamento 001.

“É justo que muito custe o que muito vale”.

Santa Teresa d’Ávila

Resumo

GUIMARÃES, Amanda Munari. Desenvolvimento de uma pipeline para vacinologia reversa baseada em arquitetura *transformer*. Orientador Frederico Schmitt Kremer. 2025. 92f. Tese (Doutorado em Biotecnologia) - Programa de Pós Graduação em Biotecnologia, Universidade Federal de Pelotas, Pelotas, 2025.

A vacinologia reversa, associada à bioinformática, permite a predição de proteínas imunogênicas a partir de dados genômicos, viabilizando o desenvolvimento de produtos biotecnológicos. A predição *in silico* de alvos vacinais envolve modelos baseados em processamento de linguagem natural (NLP), aplicados na análise estrutural e funcional de proteínas, mineração de textos biomédicos e geração de dados sintéticos. O estudo apresenta o ReVarcine, uma ferramenta projetada para prever características proteicas relevantes na imunogenicidade. Seu processo inclui pré-processamento de sequências dos bancos UniProt e SwissProt, remoção de entradas incompletas ou redundantes e tokenização baseada em Keras para representação numérica. As sequências foram divididas em conjuntos de treinamento e teste, diferenciando bactérias Gram-positivas e Gram-negativas. Foram desenvolvidos oito modelos de aprendizado profundo para prever peptídeos sinal, sites de clivagem, localização subcelular e estruturas secundárias como barris beta e alfa-hélices. As arquiteturas incluem camadas de incorporação, transformadores, LSTM e camadas densas, integradas em um pipeline modular. Esse design garante escalabilidade e adaptabilidade para diferentes organismos e necessidades analíticas. A avaliação de desempenho utilizou métricas como MCC e F1 score, comparando o ReVarcine com ferramentas consolidadas, incluindo SignalP, PSORT e PSIPRED. Os resultados indicam que o ReVarcine frequentemente supera preditores tradicionais na análise proteômica, destacando-se pela versatilidade e precisão. O ReVarcine inova ao unificar múltiplas tarefas preditivas, reduzindo a dependência de diversas ferramentas especializadas. Isso otimiza o tempo computacional e minimiza erros no processamento de dados entre plataformas. Sua abordagem integrada estabelece um novo padrão para imuno informática, tornando-o uma ferramenta promissora para suprir as crescentes demandas da vacinologia reversa.

Palavras-chave: bioinformática, proteoma, processamento de linguagem natural, bactérias, patógenos

Abstract

GUIMARÃES, Amanda Munari. Development of a pipeline for reverse vaccinology based on transformer architecture. Advisor Frederico Schmitt Kremer. 2025. 92f. Thesis (PhD in Biotechnology) - Postgraduate Program in Biotechnology, Federal University of Pelotas, Pelotas, 2025.

Reverse vaccinology, combined with bioinformatics, enables the prediction of immunogenic proteins from genomic data, facilitating the development of biotechnological products. The in silico prediction of vaccine targets involves models based on natural language processing (NLP), applied to the structural and functional analysis of proteins, biomedical text mining, and synthetic data generation. This study presents ReVarcine, a tool designed to predict protein characteristics relevant to immunogenicity. Its process includes preprocessing sequences from the UniProt and SwissProt databases, removing incomplete or redundant entries, and using Keras-based tokenization for numerical representation. The sequences were divided into training and testing sets, distinguishing Gram-positive and Gram-negative bacteria. Eight deep learning models were developed to predict signal peptides, cleavage sites, subcellular localization, and secondary structures such as beta barrels and alpha helices. The architectures include embedding layers, transformers, LSTM, and dense layers, integrated into a modular pipeline. This design ensures scalability and adaptability for different organisms and analytical needs. Performance evaluation used metrics such as MCC and F1 score, comparing ReVarcine with established tools, including SignalP, PSORT, and PSIPRED. The results indicate that ReVarcine often outperforms traditional predictors in proteomic analysis, standing out for its versatility and accuracy. ReVarcine innovates by unifying multiple predictive tasks, reducing reliance on various specialized tools. This optimizes computational time and minimizes errors in data processing across platforms. Its integrated approach sets a new standard for immunoinformatics, making it a promising tool to meet the growing demands of reverse vaccinology.

Keywords: bioinformatics, reverse vaccinology, proteome, transformers, natural language processing

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas e siglas

AAs Aminoácidos
AM Aprendizado de Máquina
ANN *Artificial Neural Networks*
CMV Citomegalovírus
CNN Redes Neurais Convolutivas
DL *Deep Learning*
DNN *Deep Neural Networks*
EBV Vírus Epstein-Barr
HIV Vírus da Imunodeficiência Humana
HMM *Hidden Markov Model*
HSV Vírus Herpes Simplex
IA Inteligência Artificial
ML *Machine Learning*
NCBI *National Center for Biotechnology Information*
NLP *Natural Language Processing*
OMS Organização Mundial da Saúde
RSV Vírus Epstein-Barr
TB Tuberculose
VR Vacinologia Reversa
WSG Sequenciamento Completo de Genomas

Sumário

1. Introdução Geral

As vacinas previnem milhões de mortes anualmente, bem como reduzem a mortalidade e morbidade da população vacinada. Elas podem ser definidas como um produto biológico que contém antígenos oriundos de patógenos. Esses antígenos devem, de forma segura, induzir uma resposta imune protetora ao hospedeiro contra futuras exposições ao patógeno em questão (POLLARD & BIJKER, 2021). O desenvolvimento de vacinas tradicionais foi amplamente baseado nos princípios de Pasteur. Essa abordagem conta com o isolamento, inativação e injeção do agente causador da doença. No entanto, com tecnologias de sequenciamento de nova geração e o crescente número de sequências dos genomas dos microorganismos patogênicos disponíveis em bancos de dados como NCBI (AGARWALA et al., 2018), GenBank (CLARK et al., 2016), EuPathDB (WARRENFELTZ et al., 2018) e *Virus Pathogen Database and Analysis Resource* (ViPR) (PICKETT et al., 2012), combinados com os avanços na patogênese e imunologia, permitiu o nascimento de uma nova era da pesquisa e do desenvolvimento de vacinas. Além da inovação na produção de vacinas, foi possível reduzir o custo financeiro e o tempo de identificação de novos candidatos de antígenos. Essa estratégia foi denominada de Vacinologia Reversa (VR) (ADU-BOBIE et al., 2003; LILJEROOS et al., 2015; RAPPUOLI, 2000).

A VR agregada à bioinformática tem como objetivo principal rastrear *in silico* o repertório proteico e gênico do patógeno e determinar quais são as proteínas capazes de induzir uma resposta imune no hospedeiro. Rino Rappuoli foi o primeiro a usar o termo vacinologia reversa (RAPPUOLI, 2001). Esse conceito foi validado com *Neisseria meningitides* serogrupo B (MenB), bactéria Gram-negativa, responsável por causar meningite meningocócica globalmente (KELLY; RAPPUOLI, 2005). A partir desses avanços na genômica, proteômica e imunoinformática, ferramentas dessas abordagens vêm sendo consecutivamente aplicadas no diagnóstico, na terapêutica e no desenvolvimento de vacinas.

Sabe-se que as proteínas de membrana, secretadas ou extracelulares são mais facilmente acessíveis ao anticorpo do que as proteínas intracelulares e, portanto, representam candidatos a vacinas ideais. Por isso, torna-se crucial determinar a localização subcelular, especificamente identificar proteínas de membrana ou proteínas exportadas pela célula. A identificação dessas proteínas também se baseiam em características como domínios transmembranares, peptídeo sinal, motivos de ancoragem da membrana, como por exemplo, estruturas de barril-beta e hélice transmembrânicas, atividade de adesina, presença de domínios conservados e funções moleculares associadas a fatores de virulência (BRENNAN & SHAHIN, 1996; ROOT, 2006; SERRUTO & RAPPUOLI, 2006).

A abordagem de VR está resumida na Figura 1. O primeiro passo é processar as sequências genômicas individualmente e passá-las por vários filtros computacionais de seleção (HE; RAPPUOLI; DE-GROOT, 2010). Dessa forma, o conjunto de proteínas pode ser processado com programas de software dedicados a predizer sua possível localização celular, presença de peptídeo sinal, posição de corte do peptídeo sinal, presença de estruturas específicas como a de arrojados transmembranares de barril-beta, alfa-hélice transmembrana, dentre outras características. Com base na presença dessas qualidades, é feita uma predição de quais são os melhores candidatos vacinais, para que possam, então, ser realizados os testes *in vivo*.

De uma forma geral, a predição das estruturas que são fundamentais na etapa de elencar os alvos *in silico*, ocorre através de modelos de *Machine Learning* (ML). O Estado da Arte para predição dessas estruturas conta com SignalP (PETERSEN et al., 2011) e Phobius (KÄLL; KROGH; & SONNHAMMER, 2007) para predição de peptídeo sinal. O SignalP é baseado no método em rede neural profunda, combinado com classificação de campo aleatório condicional e aprendizagem de transferência otimizada para previsão de peptídeo sinal. Já a ferramenta Phobius é baseada em Modelos Ocultos de Markov (*Hidden Markov Model*, HMM). A ferramenta

Cello (YU et al., 2006) e Psort (YU et al, 2010) são utilizadas para predição de localização subcelular. O algoritmo do Cello é baseado em classificação binária com regressão logística, enquanto o Psort é baseado em *Bayesian Network*. Para predição de estruturas de barril beta em proteínas, tem-se as ferramentas HHomp (REMMERT et al, 2009) e BOMP (BERVEN et al, 2004). Enquanto que a ferramenta HHomp se baseia em HMM, a ferramenta TMHMM (SONNHAMMER et al., 1998) consiste em uma ferramenta útil para predição da estrutura de hélice transmembrânica em proteínas, a qual utiliza em seu algoritmo HMM.

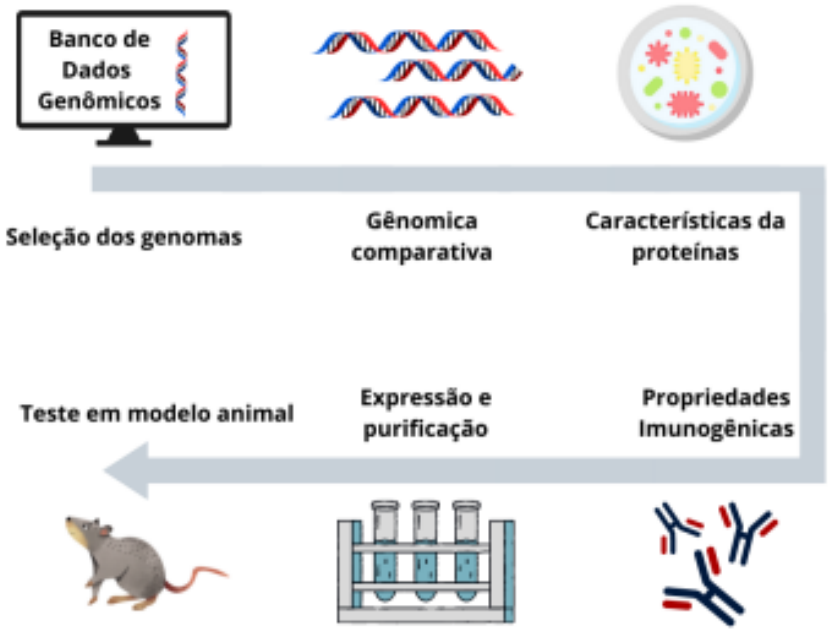
Além dessas ferramentas Estado da Arte, têm-se disponíveis diversas pipelines que integram ferramentas de predição de estruturas associadas à VR. Por exemplo, a ferramenta PanRV (NAZ et al, 2019) combina as abordagens de pangenoma e vacinologia reversa através da integração de ferramentas de bioinformática como Prokka (SEEMANN, 2014), Roay (PAGE et al., 2015), Blast+ (CAMACHO et al., 2009), PSORTb (YU et al., 2010), HMMTOP (SIMON, 2001), ABCPred (SAHA & RAGHAVA, 2006), Propred (SINGH & RAGHAVA, 2001), Vaxijen (DOYTCHINOVA & FLOWER, 2007) , e bancos de dados como DEG (LUO et al., 2013), VFDB (CHEN et al., 2012), MvirDB (ZHOU et al., 2006), RefSeq (PRUITT, TATUSOVA, & MAGLOTT, 2007), UniProt-SwissProt. Existe, também, a *pipeline* ReVac (D'MELLO et al., 2019), a qual combina várias ferramentas independentes para previsões como PSORTb, LipoP (ORVIS et al., 2010), TMHMM (KENT, 2002), SignalP, SPAAN (AASBURNER et al., 2000), Surface HMMs (GERTZ et al., 2006), OrthoMCL (LI et al., 2003), dentre outras. Esse tipo de *pipeline* depende de múltiplas ferramentas e/ou bancos de dados diferentes, cada qual com suas respectivas dependências para serem executados apropriadamente, dessa forma a configuração, execução e portabilidade destas *pipelines* se tornam um processo complexo e trabalhoso.

Nesse aspecto, desenvolver uma ferramenta que tenha modelos de ML próprios, além de ser um processo inovador, conta com menor complexidade de execução. Nesse contexto, o ML possui uma ferramenta

poderosa conhecida como redes neurais artificiais (*Artificial Neural Networks,* ANN). As ANN podem ser definidas como o mapeamento não linear matemático que simula as redes neurais biológicas através dos neurônios que emitem impulsos elétricos em resposta a estímulos (JAIN; MAO; MOHIUDDUN, 1996). Essa arquitetura complexa inspirada na biologia conta com modelos lógico-matemáticos de neurônios artificiais que simulam o comportamento e as funções dos neurônios biológicos. O modelo clássico de ANN é composto por 3 camadas, sendo a primeira camada denominada de camada de entrada (input layer), a segunda é a camada oculta (hidden layer) e a terceira é a camada de saída (output layer) (BASHEER; HAJMEER, 2000; GERSHENSON, 2003). Redes neurais que apresentam estruturas mais complexas com um maior número de camadas e um grande número de neurônios, são chamadas de redes neurais profundas (deep neural networks, DNN). As redes profundas são responsáveis por diversos avanços relacionados à visão computacional, processamento de linguagem natural como textos, reconhecimento de imagens e de áudios (DENG et al., 2012; SZE et al., 2017).

Considerando o contexto da VR em conjunto dos modelos de *Deep Learning*, para os propósitos deste estudo, as tarefas serão vistas dentro do escopo de *Natural Language Processing* (NLP), visto que os dados que serão processados são sequências de aminoácidos, logo textos (CHOWDHARY, 2020). As sequências de proteínas podem ser representadas naturalmente como sequências de letras. O alfabeto proteico consiste de 20 aminoácidos comuns (AAs). Além disso, proteínas são compostas de elementos modulares exibindo pequenas variações que podem ser rearranjadas e reunidas de forma hierárquica. Por esta analogia, motivos e domínios de proteínas, que são a construção funcional básica blocos de proteínas, são semelhantes a palavras, frases e sentenças em linguagem humana (BRANDES et al., 2021). Nesse sentido, NLP aliada a bioinformática são ferramentas importantes pois permitem desenvolver preditores de estrutura tridimensional de proteínas, função biológica, localização subcelular, interação entre moléculas dentre outras estruturas.

Figura 1 - Esquema sobre a estratégia de Vacinologia Reversa. A imagem abrange os principais passos da Vacinologia Reversa, desde a obtenção dos dados genômicos até os testes *in vivo*.



Fonte: Imagem gerada com Canva (canva.com)

Nesse contexto, o presente trabalho visa o desenvolvimento de uma pipeline para Vacinologia Reversa baseada em aprendizagem de máquina profunda denominada ReVarcine. Esta ferramenta permitirá integrar a metodologia de aprendizado de máquina profundo aliado às etapas de predição das estruturas necessárias para a elencação de alvos vacinas, como predição de peptídeo sinal, hélices transmembrânicas, barril-beta, localização sub-celular.

2. Revisão Bibliográfica

2.1 Vacinas

A etimologia da palavra vacina remonta ao termo latino vaccinus, que significa "relativo à vaca", derivado de vacca, palavra latina para "vaca". Essa origem está intimamente conectada à descoberta pioneira de Edward Jenner no século XVIII, que observou que a exposição ao vírus da varíola bovina (*cowpox*) conferia imunidade contra a varíola humana, uma doença mortal na época. Jenner utilizou material de lesões da varíola bovina para inocular indivíduos, um processo denominado "vacinação", em alusão às vacas, consolidando assim o termo vacina como um pilar da imunização moderna (HOLLAND, 2010). Em 1700, o agricultor Benjamin Jesty, em colaboração com Jenner, observou que as vacas leiteiras não eram afetadas pela varíola, inferindo que a varíola bovina poderia protegê-las. Jesty inoculou sua própria família com esporos da varíola (PEAD & JESTY, 2009), enquanto Jenner posteriormente conduziu ensaios clínicos no século XVIII e divulgou seus resultados ao mundo, contribuindo para a erradicação da varíola (JENNER, 1798; BAZIN, 2011).

Após Jenner, Louis Pasteur descobriu a atenuação de bactérias pela exposição a condições adversas, utilizando o organismo causador da cólera das galinhas, agora conhecido como *Pasteurella multocida* (PASTER, 1880). Pasteur acreditava que a resistência se devia ao esgotamento de um elemento crucial para o crescimento do organismo. Embora ele não compreendesse completamente o mecanismo das vacinas, seus resultados práticos foram notáveis. Na última década do século XIX, o desenvolvimento de vacinas começou a ser realizado por pesquisadores da Grã-Bretanha, Alemanha, Estados Unidos e do laboratório de Pasteur na França. Durante os últimos anos do século XIX e início do século XX, foram produzidas e testadas vacinas de células inteiras inativadas contra a febre tifoide (PFEIFFER & KOLLE, 1896), cólera (KOLLE, 1896) e peste (HAFFKINE, 1896).

Nesse contexto, a vacina é uma preparação biológica que pode conter um patógeno vivos atenuados, microorganismos mortos e inativados, componentes microbianos purificados, conjugados de proteína transportadora de polissacarídeo ou proteínas recombinantes. Os dois tipos básicos de vacinas, baseada no microrganismo em um forma atenuada, mas viva ou no microrganismo morto e inativado, estavam entre as primeiras vacinas geradas. Um terceiro tipo de vacinas, feito a partir dos toxoides diftérico e tetânico, foi desenvolvido na década de 1920 e representa um produto muito mais sofisticado. Independente do tipo de vacina o que se espera de comportamento é que ao ser administrado ao hospedeiro, estimula uma resposta protetora das células do sistema imunológico (PLOTKIN et al., 2012). Os antígenos presentes na vacina fazem com que o sistema imunológico do corpo reconheça o agente como estranho, induzindo respostas imunes específicas (imunogenicidade). As vacinas podem ser profiláticas (para prevenir ou reduzir os efeitos de uma infecção futura por qualquer patógeno natural) ou terapêuticas (como vacinas contra o câncer) (LEVINE e SZTEIN, 2004).

Epidemiologicamente a implementação direcionada de vacinas acarretou na diminuição da morbidade e mortalidade por doenças infecciosas que antes eram flagelos e encargos económicos (como sarampo, poliomielite, difteria, invasivo Haemophilus influenzae tipo b e infecções pneumocócicas) (ORENSTEIN e AHMED, 2017). Segundo a Organização Mundial da Saúde (OMS), a vacinação pode prevenir entre dois e três milhões de mortes anualmente em todo o mundo (WHO). Além disso, as imunizações permitiram a erradicação da varíola e estão agora perto de erradicar a poliomielite (Comissão Regional Africana para a Certificação de Poliomielite, 2020; FENNER et al., 1988). Bem como, as vacinas possuem um impacto econômico significativo quando o assunto é redução de custos hospitalares com tratamentos e hospitalização. Essa redução de custos em 2017 foi superior a 500 bilhões de dólares (ORENSTEIN e AHMED, 2017; OZAWA et al., 2016).

Embora as vacinas tenham um impacto grandioso na saúde pública, ainda existem muitas doenças infecciosas para as quais não foram desenvolvidas vacinas eficazes. Para os patógenos humanos como a imunodeficiência humana vírus (HIV), tuberculose (TB), vírus sincicial respiratório (RSV), citomegalovírus (CMV), vírus herpes simplex (HSV) e vírus Epstein-Barr (EBV) até agora não tiveram sucesso. O HIV, por exemplo, causou 39 milhões de mortes em todo o mundo e mais de 36 milhões de pessoas convivem com o vírus (PANDEY & GALVANI, 2019). Da mesma forma, a tuberculose causa 1.6 milhões de mortes anuais (SINGH et al., 2020).

O surgimento de novos patógenos, como doenças respiratórias agudas graves, vírus da síndrome e a disseminação agressiva de patógenos reconhecidos, como vírus do Nilo Ocidental também estimula programas de desenvolvimento de vacinas. O combate aos surtos exige o rápido desenvolvimento de vacinas que normalmente não era possível com os métodos tradicionais das plataformas de vacinas. Estes desafios suscitaram um intenso interesse no desenvolvimento de novas tecnologias de vacinas.

2.2 Vacinologia reversa

Até recentemente, o desenvolvimento de vacinas dependia de métodos bioquímicos, imunológicos e microbiológicos. Embora eficazes e seguras para muitas doenças infecciosas, essas vacinas apresentam algumas desvantagens, como a necessidade de cultivar o patógeno em condições adequadas, o risco de reativação ou contaminação do produto final, a baixa imunogenicidade para alguns antígenos e a incapacidade de se adaptar rapidamente a novos patógenos emergentes (RAPPUOLI, 2000).

A vacinologia reversa (VR) é uma tecnologia que utiliza métodos computacionais para analisar o proteoma de um patógeno e identificar alvos vacinais. O termo "reverso" refere-se à abordagem inovadora de iniciar a descoberta de vacinas a partir de dados genômicos analisados por

computador, em vez de um organismo vivo. Essa técnica foca em antígenos que provavelmente induzirão respostas protetoras e nas regiões precisas desses antígenos, os epítopos, reconhecidos pelo sistema imunológico (POLAND, 2009). A VR permite identificar potenciais alvos de vacinas de maneira mais rápida e eficiente, sendo especialmente útil para patógenos difíceis de cultivar em laboratório (LM, 2010; DONATI & RAPPUOLI, 2013).

Essa abordagem ganhou destaque no final da década de 1990, após o sequenciamento completo do genoma da bactéria Neisseria meningitidis (meningococo), conforme revisado por MASIGNANI et al, 1990. Este avanço resultou na criação e aprovação da Bexsero, a primeira vacina contra cepas B de meningite meningocócica. Desde então, a VR tem sido utilizada para identificar candidatos a antígenos para diversos patógenos, incluindo os vírus da hepatite C, influenza e Zika.

A vacinologia reversa tornou-se possível graças ao sequenciamento completo de genomas (WGS) e aos avanços na biologia populacional, como revisado por Maiden, um pioneiro nas ferramentas epidemiológicas moleculares cruciais para a concepção de vacinas para doenças infecciosas e não infecciosas. Recentemente, Bianconi et al. aplicaram com sucesso a abordagem clássica de VR para *Pseudomonas aeruginosa*. Um dos principais avanços relacionados à vacinologia reversa é a incorporação de novos conceitos e tecnologias para facilitar a concepção de vacinas, incluindo contribuições da proteômica, imunologia, biologia estrutural, biologia de sistemas e modelagem matemática. Assim, a VR hoje abrange muito mais do que a inovação na descoberta de antígenos, representando uma mudança significativa na direção e ação na pesquisa de vacinas.

Organismos patogênicos são compostos por milhares de proteínas. O objetivo central da VR é reduzir esse número, deixando apenas os candidatos mais promissores para investigação laboratorial. Esse objetivo é alcançado prevendo ou coletando características de proteínas que apoiem ou se opunham à candidatura usando programas de bioinformática ou acessando bancos de dados biológicos, respectivamente

A abordagem de VR permite identificar sistematicamente todos os potenciais antígenos de um patógeno, tornando, teoricamente, possível desenvolver um agente vacinal seguro e eficaz contra qualquer doença infecciosa. Existem alguns protocolos que podem ser aplicados dentro do conceito de vacinologia reversa

2.3 Abordagens bioinformáticas para análise genômica de procariotos

Sequenciamento é o processo que determina a ordem de uma sequência nucleotídica do DNA ou RNA de um organismo. Essa técnica teve início em 1975 com Sanger e Coulson (SANGER & COULSON, 1975). No entanto, foi com o advento dos sequenciamentos de nova geração que houve um aumento expressivo no número de dados genômicos. Com o grande fluxo de dados gerados tanto pelo sequenciamento, como pelas análises genômicas, surgiu a necessidade de tratar esses dados e armazená-los de forma virtual, ou seja, armazenar dados biológicos utilizando ferramentas de informática. Nesse contexto, a bioinformática, atualmente, é centro da interpretação dos dados genômicos, uma vez que possibilitou a criação de banco de dados e ferramentas que aperfeiçoam o processo de análises. Juntamente com a bioinformática, os bioinformatas exercem papel crucial na evolução da pesquisa, pois a análise dos dados gerados é provavelmente o próximo gargalo na pesquisa biológica (LUCACIU et al., 2019).

As abordagens da bioinformática e da biologia molecular permitem a prospecção *in silico* de informações, a partir de dados genômicos, possibilitando a exploração dos recursos genéticos com a finalidade de obter produtos biotecnológicos como antibióticos, agentes terapêuticos, vacinas, enzimas, polímeros, além de propiciar melhores prognósticos e prevenções de doenças emergentes (MCLNERNEY et al., 2008).

Nesse sentido a caracterização de organismos procariotos é de extrema relevância possuindo, atualmente, diferentes focos, por exemplo, entendimento da genética e fisiologia, eventos evolutivos, bioquímica, desenvolvimento de novos compostos biotecnológicos e patogênese de doenças (HANDELSMAN, 2004; WOYKE et al., 2010). A genômica de microrganismo conta com um imenso banco, na base de dados *Nucleotide* do NCBI (www.ncbi.nlm.nih.gov) constam, aproximadamente, 52.291.489 entradas de genomas para bactérias. Já no banco GOLD (www.genomesonline.org), atualmente, estão registrados cerca de 136.778 projetos genomas para Procariotos. Essa grande quantidade de dados possibilitou o surgimento de novas linhas de pesquisas, tais como a genômica funcional e proteômica, as quais permitem a análise comparativa dos genomas bacterianos de forma a identificar mecanismos adaptativos, aspectos evolutivos e produtos biotecnológicos, bem como permitiu o nascimento de uma nova era da pesquisa e do desenvolvimento de vacinas.

2.4 Ferramentas bioinformática para análise de proteínas e antígenos para vacinologia reversa

A análise de proteínas e antígenos desempenha um papel central na vacinologia reversa, permitindo a identificação de novos alvos vacinais de forma rápida e eficiente. Esta abordagem bioinformática explora os genomas de patógenos, com foco em proteínas imunogênicas, e elimina a necessidade de cultivar micro-organismos em laboratório, acelerando o processo de desenvolvimento de vacinas (RAPPUOLI, 2000).

Em vacinologia reversa, o processo começa com a análise de sequências genômicas do patógeno para identificar proteínas-alvo com alta probabilidade de desencadear uma resposta imunológica protetora. Uma etapa essencial é selecionar proteínas de superfície ou secretadas, já que essas estruturas são diretamente acessíveis ao sistema imunológico do

hospedeiro (PIZZA et al., 2000). Diversas ferramentas bioinformáticas são usadas nesta triagem, permitindo análises precisas e baseadas em dados genômicos, tais como PSORTb (YU et al., 2010), SIganlP, LipoP (RAHMAN et al., 2008), Blast, Ortholog Finder e ProtParam.

Além dessas ferramentas, outras abordagens bioinformáticas avançadas, como a modelagem estrutural realizada por softwares como PyMOL e AlphaFold, permitem compreender a estrutura tridimensional das proteínas. A avaliação da estrutura é essencial para predições mais robustas sobre a antigenicidade e o desenho racional de epítopos.

Estas ferramentas, em conjunto, oferecem uma abordagem sistemática e altamente eficiente para identificar e validar proteínas-alvo para vacinas. A integração dos dados de localização subcelular (YU et al., 2010), presença de peptídeos sinal (ALMAGRO ARMENTEROS *et al., 2019*) e características estruturais permite uma seleção precisa, reduzindo os custos e aumentando a taxa de sucesso no desenvolvimento de vacinas.

2.4.1 SignalP

O SignalP é uma ferramenta amplamente utilizada em bioinformática para prever peptídeos sinais em proteínas, ou seja, identificar regiões em proteínas que indicam se elas são destinadas a ser secretadas ou transportadas para fora da célula. Esses peptídeos sinalizadores geralmente estão presentes na extremidade N-terminal das proteínas, e sua identificação é crucial em pesquisas de biologia molecular e biotecnologia. O SignalP foi desenvolvido para oferecer uma maneira confiável de prever a existência e a localização desses peptídeos sinal com base em sequências de proteínas. A versão mais recente, SignalP 6.0, apresenta avanços consideráveis na precisão da predição, utilizando redes neurais profundas e métodos de aprendizado de máquina para obter melhores resultados em diversos organismos (BENDTSEN et al., 2004).

A versão original do SignalP foi descrita por NIELSEN et al. (1997) e, ao longo dos anos, passou por várias atualizações para melhorar a sensibilidade e especificidade na predição de peptídeos sinal. Essas versões mais recentes, como o SignalP 5.0 e 6.0, incorporam modelos de aprendizado de máquina e outros avanços em métodos computacionais, o que permite resultados mais precisos e abrangentes para diferentes tipos de organismos e sequências de proteínas (BENDTSEN et al., 2016; HUANG & RASMUSSEN, 2023).

2.4.2 PSORT

A ferramenta PSORT é uma das mais antigas e bem estabelecidas ferramentas de bioinformática para predição de localização subcelular de proteínas em organismos procariotos e eucariotos. Desenvolvida inicialmente por NAKAI e KANEHISA (1991), ela utiliza uma abordagem baseada em dados experimentais e informações de sequência para prever em qual compartimento celular uma proteína é localizada, como núcleo, citoplasma, membranas, entre outros.

PSORT realiza a predição com base em características das sequências de proteínas, como sinais de direcionamento, composição de aminoácidos e motivos específicos associados à localização subcelular. Ao longo do tempo, a ferramenta evoluiu e deu origem a variações mais especializadas, como o PSORTb, otimizado para bactérias, e o WoLF PSORT, mais adequado para proteínas de organismos eucariotos.

A versão PSORTb foi desenvolvida especificamente para bactérias, oferecendo alta precisão, particularmente em organismos gram-positivos e gram-negativos. É uma das ferramentas mais amplamente usadas para microbiomas e estudos em microbiologia (YU et al. 2010). Já a versão WoLF é otimizada para predição em eucariotos, que usa um sistema baseado em aprendizado de máquina para melhorar a classificação (HORTON & NAKAI, 2007).

PSORT é amplamente utilizada em estudos proteômicos para associar proteínas a compartimentos celulares específicos. Isso tem relevância não apenas para caracterização funcional de proteínas, mas também para entender mecanismos celulares e para estudos relacionados ao desenvolvimento de novos medicamentos ou biotecnologias.

2.4.3 PSIPRED

A ferramenta PSIPRED é uma das mais populares e eficazes para a predição de estruturas secundárias de proteínas, ajudando na análise da organização tridimensional de proteínas a partir de suas sequências de aminoácidos. PSIPRED foi inicialmente desenvolvida por MCGUFFIN et al. (2000) e utiliza um modelo baseado em redes neurais para prever as estruturas secundárias, como hélices alfa, barril beta e regiões não estruturadas, a partir de uma sequência de proteínas.

A versão inicial, PSIPRED 1.0, fez uso de uma abordagem simples baseada em redes neurais, e desde então, foi significativamente aprimorada em versões posteriores, como o PSIPRED 3.0. Este método foi baseado em alinhamentos de sequências e perfis gerados pela ferramenta PSI-BLAST para melhorar a precisão das previsões. PSIPRED 3.0, que é a versão mais recente, combina dois tipos de redes neurais, uma para predição da estrutura de três estados (hélice, barril e desordenada) e outra para predição de pares de resíduos, melhorando a precisão na previsão de estruturas de proteínas e aumentando a capacidade de lidar com proteínas complexas e desordenadas (JONES, 2007).

2.5 Aprendizado de máquina

O aprendizado de máquina (AM) é uma subárea da inteligência artificial que permite criar sistemas capazes de aprender e adaptar seu comportamento a partir de dados.. Essa abordagem tem sido amplamente

empregada em diversas áreas, como medicina, tecnologia e finanças, devido à sua capacidade de extrair padrões significativos de grandes volumes de dados e realizar previsões com alta acurácia. Avanços recentes na capacidade computacional e na disponibilidade de dados têm acelerado a adoção de técnicas de aprendizado de máquina, consolidando seu papel na sociedade moderna (ZHOU et al., 2022).

O aprendizado de máquina é comumente classificado em três principais categorias: aprendizado supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, algoritmos são treinados com conjuntos de dados rotulados, permitindo que prevejam resultados específicos em novos dados. Por outro lado, no aprendizado não supervisionado, os algoritmos buscam identificar padrões ou agrupamentos de forma autônoma a partir de dados não rotulados. Já o aprendizado por reforço baseia-se em um agente que toma decisões sequenciais em um ambiente, aprendendo a maximizar recompensas acumuladas ao longo do tempo por meio de tentativa e erro (RUSSELL & NORVING, 2021).

Apesar dos avanços, o aprendizado de máquina enfrenta desafios significativos, como a necessidade de garantir a qualidade e representatividade dos dados de entrada. Modelos de aprendizado de máquina são altamente dependentes de dados confiáveis; informações incompletas ou enviesadas podem comprometer sua eficácia. Adicionalmente, a interpretabilidade dos modelos permanece uma preocupação crescente, especialmente em aplicações críticas, como na área da saúde, onde decisões precisam ser explicáveis para profissionais e pacientes. Investimentos em pesquisa para desenvolver algoritmos mais transparentes e robustos são essenciais para ampliar a confiança no uso de aprendizado de máquina (CARVALHO et al., 2019).

2.5.1 Aprendizado profundo

Nos últimos anos, impulsionado pelo rápido aumento de dados e recursos computacionais, o aprendizado profundo (*deep learning*) ganhou espaço e alcançou avanços relevantes na ciência biológica. Essencialmente, o aprendizado profundo é uma parte do campo de aprendizado de máquina, um subcampo da inteligência artificial (IA) que se preocupa com o aprendizado de representações de dados usando métodos computacionais.

Em algoritmos tradicionais de aprendizado de máquina, é necessário selecionar manualmente as características e o classificador, enquanto em um algoritmo de aprendizado profundo, as características são extraídas automaticamente pelo próprio algoritmo, aprendendo com seus próprios erros. É essa extração automática de características que distingue o aprendizado profundo do aprendizado de máquina.

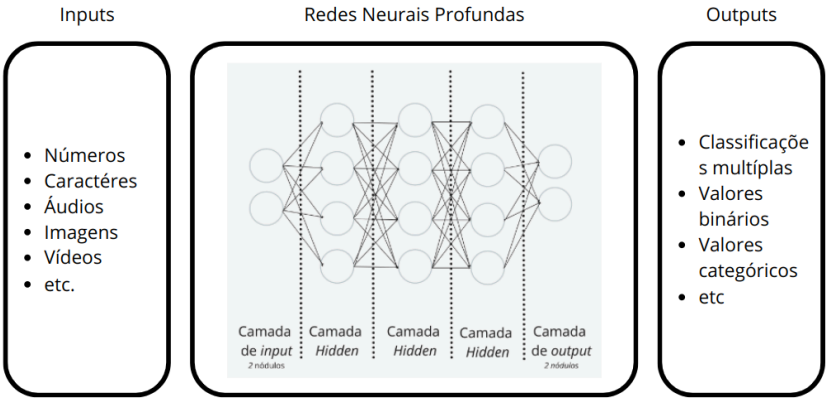
A origem das Redes Neurais Profundas (DNNs) remonta a 1943, quando McCulloch e Pitts propuseram a primeira rede neural artificial (MCULLOCH & PITTS, 1943). Somente em 1985, HINTON et al. propuseram o algoritmo de retropropagação, que estimulou enormemente o desenvolvimento do campo (ACKLEY, HINTON & SEJNOWSKI, 1985). Quase ao mesmo tempo, foi proposto o "*neocogitron*", que inspirou as Redes Neurais Convolucionais (CNNs), as Redes Neurais Recorrentes (RNNs) e as DNNs (FUKUSHIMA & MIYAKE, 1982; JORDAN, 1997; LECUN et al., 1998). Contudo, devido às limitações do hardware da época, esses modelos eram difíceis de usar para lidar com grandes volumes de dados, o que novamente freou o avanço do aprendizado profundo.

Nos últimos anos, duas aplicações impressionantes de aprendizado profundo ganharam destaque global e chocaram o mundo. Uma delas é o AlphaGo, que derrotou os campeões mundiais de Go usando aprendizado profundo com o suporte de recursos de hardware abundantes (DEEPMIND, 2022). A outra é o AlphaFold, que resolveu o desafio de 50 anos do problema de dobramento de proteínas. Essas conquistas estimularam ainda

mais o rápido desenvolvimento do aprendizado profundo em diversos domínios. Atualmente, com os avanços nas Unidades de Processamento Gráfico (GPUs) e na Computação de Alto Desempenho (HPC), o aprendizado profundo tornou-se uma das ferramentas mais eficientes, apresentando desempenho notável em praticamente todos os domínios.

Geralmente, a arquitetura de uma DNN consiste em múltiplas camadas de neurônios, incluindo uma camada de entrada, uma camada de saída e uma ou várias camadas ocultas (Figura 2). Cada neurônio está conectado a outro neurônio para transmitir informações. A entrada para uma DNN pode ser números, caracteres, áudios, imagens, etc., que são divididos em bits de dados binários que um computador pode processar. A saída pode ser valores contínuos, valores binários ou valores categóricos, dependendo das tarefas.

Figura 2 - **Modelo clássico de Redes Neurais Artificiais em comparação ao Modelo de *Deep Learning*.** Está sendo representado um modelo de Rede Neural baseada em *Deep Learning*, a qual possui uma estrutura de camadas mais complexas, sendo a primeira a camada de entrada (*input layer*), a segunda, terceira e quarta camadas ocultas (*hidden layer*) e a quinta é a camada de saída (*output layer*).



Fonte: Adaptado de JAIN; MAO; MOHIUDDUN, 1996 e CICHY; KAISER, 2019

2.5.2 Camadas ocultas em modelos *deep learning*

Os modelos de *deep learning* são compostos por múltiplas camadas ocultas, cada uma desempenhando um papel fundamental na extração e transformação de características a partir dos dados de entrada. Essas camadas são projetadas para aprender representações cada vez mais abstratas e complexas, permitindo que os modelos capturem padrões e dependências nos dados. A escolha das camadas ocultas e sua configuração dependem do tipo de problema que se deseja resolver. A seguir, detalharemos algumas das principais camadas utilizadas em deep learning e seus funcionamentos.

A camada de *embedding* é amplamente utilizada no NLP para transformar palavras ou símbolos discretos em representações vetoriais densas. Essa transformação é necessária, pois os modelos de deep learning operam melhor com dados numéricos e contínuos. O *embedding* permite que palavras semanticamente semelhantes possuam representações próximas no espaço vetorial, auxiliando a aprendizagem do modelo. Diferentes técnicas de embedding podem ser utilizadas, incluindo o Word2Vec, que emprega abordagens Continuous Bag of Words (CBOW) e Skip-gram para aprender representações contextuais de palavras; o GloVe, que se baseia em matrizes de coocorrência para capturar relações semânticas; e embeddings treinados diretamente nas camadas de redes neurais profundas (MIKOLOV et al., 2013). A vantagem de usar *embeddings* é a capacidade de capturar relações semânticas e sintáticas dentro dos textos, o que melhora significativamente o desempenho de modelos de *deep learning* aplicados a tarefas como análise de sentimento, tradução automática e respostas automáticas a perguntas.

As camadas convolucionais são fundamentais para o processamento de imagens, embora também possam ser aplicadas a séries temporais e textos. Elas utilizam filtros que deslizam sobre os dados de entrada para extrair padrões locais, reduzindo a dimensionalidade e capturando características relevantes. Essa abordagem é inspirada no funcionamento

do córtex visual humano (LECUN et al., 1998). As CNNs são compostas por múltiplas camadas convolucionais, seguidas por camadas de *pooling* e normalização, que ajudam a reduzir a complexidade do modelo e melhorar a capacidade de generalização. Dentro da estrutura de CNNs, temos diferentes tipos de camadas, incluindo: camadas de Convolução, as quais são responsáveis pela aplicação de filtros que extraem características essenciais da imagem; camadas de *Pooling* que são usadas para reduzir a dimensionalidade das representações e manter apenas as informações mais relevantes; camadas de Normalização que auxiliam na padronização dos dados para otimizar a aprendizagem da rede; camadas *Fully Connected*, as quais são utilizadas nas últimas etapas para a classificação final. CNNs são extremamente eficientes em tarefas de visão computacional, sendo utilizadas em aplicações como reconhecimento facial, diagnóstico médico por imagens e classificação de objetos em fotos.

As redes neurais recorrentes (RNNs) são projetadas para lidar com dados sequenciais, permitindo que o modelo retenha informações passadas para a previsão de elementos futuros. No entanto, RNNs tradicionais sofrem com o problema do desaparecimento do gradiente, tornando difícil a aprendizagem de dependências de longo prazo. Para mitigar essa limitação, foram desenvolvidas variantes como *Long Short-Term Memory* (LSTM) (HOCHREITER & SCHIDHUBER, 1997) e *Gated Recurrent Units* (GRU) (CHO et al., 2014), que utilizam mecanismos de portas para controlar o fluxo de informação e permitir a retenção de dependências de longo prazo.

A camada LSTM inclui três portas principais: porta de entrada, porta de esquecimento e porta de saída. A porta de entrada decide quais novas informações serão armazenadas na memória da célula. A porta de esquecimento controla quais informações antigas devem ser descartadas. Já a porta de saída determina a saída final do nó. As camadas GRU são uma simplificação das LSTMs, combinando as portas de entrada e esquecimento em uma única estrutura, reduzindo a complexidade computacional sem perder eficácia.

As arquiteturas baseadas em Transformers revolucionaram o campo do NLP e estão sendo aplicadas a diversas outras áreas. Estas redes utilizam mecanismos de autoatenção, permitindo que o modelo pondere a importância de diferentes partes da entrada, independentemente da posição sequencial. O modelo Transformer foi introduzido por VASWANI et al. (2017) e é a base de modelos modernos como BERT (DEVLIN et al., 2018) e GPT (BROWN et al., 2020).

2.5.3 Plataformas e recursos de aprendizado profundo

Atualmente, as duas plataformas de código aberto mais renomadas para aprendizado profundo de ponta a ponta são o TensorFlow (ABADI et al., 2016) e o PyTorch (PASZE et al., 2019). Ambas oferecem ecossistemas abrangentes e flexíveis de ferramentas, bibliotecas e recursos comunitários que permitem que engenheiros e pesquisadores desenvolvam e implantem facilmente aplicações impulsionadas por aprendizado profundo.

TensorFlow (https://www.tensorflow.org/, acessado em 27 de novembro de 2024) foi desenvolvido por pesquisadores e engenheiros do Google e lançado em 2015. Trata-se de uma biblioteca de matemática simbólica, ideal para programação de fluxo de dados em uma ampla variedade de tarefas. Ele oferece múltiplos níveis de abstração para a construção e o treinamento de Redes Neurais Profundas (DNNs). Além disso, adotou o Keras (https://keras.io/, acessado em 27 de novembro de 2024), uma API funcional que estende o TensorFlow, permitindo que os usuários codifiquem facilmente seções funcionais de alto nível. O TensorFlow também fornece funcionalidades específicas, como pipelines, estimadores e execução imediata, suportando diversas topologias com diferentes combinações de entradas, saídas e camadas.

O PyTorch (https://pytorch.org/, acessado em 27 de novembro de 2024) é baseado no Torch e é relativamente mais novo em comparação ao TensorFlow. Foi desenvolvido por pesquisadores do Facebook e lançado em

2017. É conhecido por sua simplicidade, facilidade de uso, flexibilidade, uso eficiente de memória e gráficos computacionais dinâmicos. Devido ao seu poder computacional e à sensação de programação nativa, o PyTorch está se destacando como favorito. Além disso, possui uma grande comunidade de desenvolvedores e pesquisadores que construíram ferramentas e bibliotecas robustas para expandir suas capacidades. Algumas bibliotecas populares incluem GPyTorch, BoTorch e Allen NLP.

2.5.4 Processamento de linguagem natural

Com os avanços das tecnologias de sequenciamento de genomas, a quantidade de dados continua a crescer. Essas tecnologias oferecem aos pesquisadores o desafio de extrair informações estatísticas e biológicas significativas de dados de alta dimensionalidade. As propriedades matemáticas e estatísticas da alta dimensionalidade muitas vezes são pouco compreendidas ou ignoradas na modelagem e análise de dados (HASTIE et al., 2009; VAN DER MAATEN et al., 2008). A capacidade de incorporar essas sequências textuais em um espaço vetorial é um passo importante para o desenvolvimento de métodos de análise eficazes (MIKOLOV et al., 2013). Uma abordagem possível é usar métodos de incorporação existentes do mundo do processamento de linguagem natural (NLP) e adaptá-los para sequências genômicas (ASGARI; MOFRAD, 2015; YANG et al., 2018).

O NLP é um campo crucial e amplamente pesquisado. Trata-se de um subcampo dos deep neural networks (DNN) que visa capacitar os computadores a compreenderem texto e linguagem falada de maneira semelhante à forma como os humanos fazem. Devido às ambiguidades da linguagem humana, o NLP representa um problema bastante desafiador. Algumas das tarefas populares envolvidas incluem reconhecimento de fala (GRAVES et al., 2013), análise de sentimentos (PANG; LEE, 2008), tradução automática (BAHDANAU et al., 2014) e resposta a perguntas (RAJPURKAR et al., 2016), que serão apresentadas a seguir.

No contexto biológico, NLP pode ser aplicado em diferentes contextos, tais como predição de estruturas e funções de proteínas, mineração de textos biomédicos, predição de alvos vacinais e geração de dados sintéticos. No campo de predição de estruturas e funções de proteínas a representação de sequências proteicas como uma linguagem tem permitido o uso de modelos baseados em transformadores, como *ProtBERT* e *ESM (Evolutionary Scale Models)*, para prever estruturas secundárias, domínios funcionais e mutações patogênicas (ELNAGGAR, et al., 2020). Para a mineração de textos NLP é usado para extrair informações de grandes volumes de literatura científica. Modelos como *BioBERT* foram desenvolvidos para resolver tarefas como reconhecimento de entidades biomédicas (*e.g.*, genes, proteínas) e análise de relações causais (como a associação gene-doença) (LEE et al.,2020).

Considerando o contexto da VR em conjunto dos modelos de *Deep Learning*, para os propósitos deste estudo, as tarefas serão vistas dentro do escopo de NLP, visto que os dados que serão processados são sequências de aminoácidos, logo textos (CHOWDHARY, 2020). As sequências de proteínas podem ser representadas naturalmente como sequências de letras. O alfabeto proteico consiste de 20 aminoácidos comuns (AAs). Além disso, proteínas são compostas de elementos modulares exibindo pequenas variações que podem ser rearranjadas e reunidas de forma hierárquica. Por esta analogia, motivos e domínios de proteínas, que são a construção funcional básica blocos de proteínas, são semelhantes a palavras, frases e sentenças em linguagem humana (BRANDES et al., 2021). Nesse sentido, NLP aliada a bioinformática são ferramentas importantes pois permitem desenvolver preditores de estrutura tridimensional de proteínas, função biológica, localização subcelular, interação entre moléculas dentre outras estruturas.

2.5.5 Métricas de avaliação de desempenho de modelos de *deep learning*

A avaliação de modelos de *deep learning* é fundamental para determinar sua eficácia em diferentes tarefas, sendo comumente utilizada uma variedade de métricas quantitativas. Entre as métricas mais empregadas para classificação, destacam-se a acurácia, precisão, recall e a métrica F1-score, que combinam aspectos da taxa de verdadeiros positivos e falsos positivos para uma avaliação equilibrada do desempenho do modelo (GÉRON, 2019). Para problemas de regressão, métricas como erro médio absoluto (MAE), erro quadrático médio (MSE) e raiz do erro quadrático médio (RMSE) são amplamente utilizadas, permitindo quantificar a diferença entre os valores previstos e reais (GOODFELLOW; BENGIO; COURVILLE, 2016). Além disso, métricas específicas, como a AUC-ROC para classificação binária e a métrica IoU (Intersection over Union) para segmentação de imagens, são essenciais para avaliar modelos conforme o contexto da aplicação. A escolha da métrica apropriada deve ser feita considerando a natureza dos dados e o objetivo do modelo, garantindo uma avaliação confiável e comparável entre diferentes abordagens.

A acurácia é uma das métricas mais utilizadas para avaliar o desempenho de modelos de *deep learning*, especialmente em problemas de classificação. Ela é definida como a proporção de previsões corretas em relação ao total de previsões realizadas, sendo calculada conforme a Equação 1:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}$$

Onde TP (*True Positives*) representa os verdadeiros positivos, TN (*True Negatives*) os verdadeiros negativos, FP (*False Positives*) os falsos positivos e FN (*False Negatives*) os falsos negativos (GÉRON, 2019).

Embora seja uma métrica intuitiva e de fácil interpretação, a acurácia pode ser enganosa em cenários de classes desbalanceadas, onde um modelo pode obter alta acurácia apenas por prever a classe majoritária, sem de fato aprender a distinguir as classes corretamente. Para contornar essa limitação, métricas complementares, como precisão, recall e F1-score, são frequentemente empregadas na avaliação de modelos de deep learning (GOODFELLOW; BENGIO; COURVILLE, 2016). Assim, a escolha da acurácia como métrica principal deve ser feita com cautela, considerando a distribuição dos dados e a relevância de outras métricas para o problema específico.

A precisão é uma métrica amplamente utilizada para avaliar modelos de deep learning, especialmente em problemas de classificação binária, e é particularmente útil em cenários onde o custo de falsos positivos é alto. Ela é definida como a razão entre o número de verdadeiros positivos (TP) e a soma dos verdadeiros positivos (TP) e falsos positivos (FP), conforme a Equação 2:

$$\text{Precisão} = \frac{TP}{TP + FP}$$

A precisão reflete a capacidade do modelo em não rotular erroneamente as instâncias negativas como positivas, sendo especialmente útil em contextos em que a classificação incorreta de uma instância como positiva pode resultar em sérias consequências. Contudo, em problemas com classes desbalanceadas, a precisão por si só pode ser uma métrica insuficiente, já que um modelo pode apresentar alta precisão ao classificar a maioria das instâncias como negativas, ignorando as positivas. Nesse sentido, é comum utilizar a precisão em conjunto com outras métricas, como recall e F1-score, para obter uma avaliação mais robusta do desempenho do modelo (GOODFELLOW; BENGIO; COURVILLE, 2016).

O recall, também conhecido como sensibilidade ou taxa de verdadeiros positivos, é uma métrica crucial na avaliação de modelos de deep learning, especialmente em problemas onde a captura de todos os exemplos positivos é mais importante do que a precisão. O recall é definido como a razão entre o número de verdadeiros positivos (TP) e a soma dos verdadeiros positivos (TP) e falsos negativos (FN), conforme a Equação 3:

$$\text{Recall} = \frac{TP}{TP + FN}$$

O recall é especialmente relevante em cenários onde é mais importante identificar o maior número possível de casos positivos, como em diagnósticos médicos, onde a falha em detectar uma doença pode ter consequências graves. Contudo, um modelo com alto recall pode ter uma taxa de falsos positivos elevada, o que pode ser mitigado pelo uso de métricas complementares, como a precisão ou a F1-score, que combinam precisão e recall em uma única métrica (GOODFELLOW; BENGIO; COURVILLE, 2016). Dessa forma, a utilização do recall deve ser contextualizada de acordo com o impacto de falsos negativos e a natureza do problema.

A métrica F1-score é uma medida combinada de precisão e recall, sendo especialmente útil em problemas de deep learning com classes desbalanceadas, onde a avaliação isolada de precisão ou recall pode não refletir adequadamente o desempenho do modelo. O F1-score é definido como a média harmônica entre a precisão (P) e o recall (R), conforme a Equação 4:

$$\text{F1-score} = 2 \cdot \frac{P \cdot R}{P + R}$$

O F1-score busca equilibrar a necessidade de precisão e *recall*, sendo uma métrica especialmente relevante em cenários onde tanto os falsos positivos quanto os falsos negativos têm um impacto significativo. A vantagem do F1-score em relação a métricas isoladas é que ele penaliza mais severamente os modelos que apresentam desequilíbrio entre precisão e recall, oferecendo uma visão mais equilibrada do desempenho do modelo (GOODFELLOW; BENGIO; COURVILLE, 2016). Assim, o F1-score é amplamente utilizado em tarefas de classificação em que se busca um compromisso entre a identificação de exemplos positivos e a minimização de falsos positivos.

A métrica de *Matthews Correlation Coefficient* (MCC) é uma métrica robusta para avaliar modelos de deep learning, especialmente em cenários de classes desbalanceadas, pois considera todos os elementos da matriz de confusão (verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos). O MCC é definido pela Equação 5:

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

O MCC varia entre -1 e 1, sendo que um valor de 1 indica uma previsão perfeita, 0 indica uma previsão aleatória e -1 indica uma previsão totalmente errada. Uma das principais vantagens do MCC em relação a outras métricas, como acurácia e F1-score, é que ele leva em consideração a distribuição das classes, oferecendo uma avaliação equilibrada do desempenho do modelo, mesmo quando as classes são altamente desbalanceadas (GOODFELLOW; BENGIO; COURVILLE, 2016). Por essa razão, o MCC é uma métrica preferida em situações onde a precisão da classificação para ambas as classes é de interesse, e a simples contagem de acertos não é suficiente para refletir o verdadeiro desempenho do modelo.

A métrica *Intersection over Union* (IoU) é amplamente utilizada na avaliação de modelos de deep learning, especialmente em tarefas de segmentação de imagens, como na detecção de objetos. Ela calcula a sobreposição entre a área predita e a área real (*ground truth*), oferecendo uma medida de quão bem o modelo consegue prever as regiões de interesse. O IoU é definido pela Equação 6:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|}$$

Nesta equação, A representa a área predita pelo modelo, B é a área real, ∩\cap∩ é a interseção das duas áreas, e ∪\cup∪ é a união entre elas (GÉRON, 2019). O IoU varia de 0 a 1, onde 0 indica nenhuma sobreposição e 1 indica uma correspondência perfeita entre a área predita e a área real. A principal vantagem do IoU é sua capacidade de penalizar previsões imprecisas e sobreposições parciais, tornando-a uma métrica robusta, especialmente em contextos de segmentação de objetos em imagens. No entanto, o IoU pode ser sensível a pequenas variações nas bordas das áreas preditas, o que pode afetar a avaliação de modelos com desempenho próximo ao ideal. Assim, é comum utilizar o IoU em combinação com outras métricas, como a acurácia ou o F1-score, para uma avaliação mais abrangente do desempenho do modelo em tarefas de segmentação (GOODFELLOW; BENGIO; COURVILLE, 2016).

3. Hipótese e Objetivo

3.1 Hipótese

O desenvolvimento de uma pipeline unificada, fundamentada em técnicas avançadas de aprendizado de máquina profundo, é capaz de superar o desempenho das atuais ferramentas bioinformáticas aplicadas à vacinologia reversa, proporcionando uma identificação mais precisa, eficiente e robusta de candidatos vacinais.

3.2 Objetivo Geral

Desenvolver uma *pipeline* de modelos de aprendizado profundo, capaz de predizer e identificar potenciais candidatos vacinais para bactérias, por vacinologia reversa.

3.3 Objetivo Específico

Para identificação e seleção de candidatos vacinais em potencial para bactérias, foram estabelecidos os seguintes objetivos específicos:

- Obter dois modelos de aprendizado profundo para predição de peptídeo sinal para bactérias Gram-positivas e outro para Gram-negativas com resultados das métricas de *recall* e precisão igual ou superior às ferramentas SignalP;
- Obter um modelo de aprendizado profundo para predição de estruturas barril betas para bactérias Gram-negativas, com resultados das métricas de acurácia, *recall* e precisão igual ou superior às ferramentas PSIPRED;
- Obter um modelo de aprendizado profundo para predição de proteína de hélice transmembrânica para bactérias Gram-positivas, com resultados das métricas de AUC ROC e IOU igual ou superior à ferramenta PSIPRED;
- Obter dois modelos de aprendizado profundo para predição de

localização subcelular para bactérias Gram-positivas e outro para Gram-negativas, com resultados das métricas de *F1-score* igual ou superior às ferramentas Cello;

- Integrar os 8 modelos de aprendizado profundo em uma estrutura de *pipeline*;
- Disponibilizar a *pipeline* desenvolvida para uso público.

4. Capítulo 1 - Teste e avaliação de arquitetura de modelos

4.1 Introdução

As vacinas desempenham um papel crucial na prevenção de milhões de mortes anuais, contribuindo significativamente para a redução da mortalidade e morbidade na população imunizada. Esses imunizantes são definidos como produtos biológicos que contêm antígenos derivados de patógenos, projetados para induzir uma resposta imunológica protetora de forma segura, protegendo o hospedeiro contra futuras infecções pelo mesmo agente (POLLARD & BIKHER, 2021). Com os avanços nas tecnologias de sequenciamento de nova geração e o crescente volume de genomas de microrganismos patogênicos disponíveis em bancos de dados tornou-se possível inaugurar uma nova era na pesquisa e desenvolvimento de vacinas. Além de revolucionar a produção de imunizantes, essa abordagem possibilitou a redução de custos e do tempo necessário para a identificação de novos antígenos candidatos. Esse avanço levou ao surgimento da Vacinologia Reversa (VR) (ADU-BOBIE et al., 2003; LILJEROOS et al., 2015; RAPPUOLI, 2000).

A VR, quando aliada à bioinformática, tem como principal objetivo a análise in silico do repertório proteico e gênico de patógenos, permitindo a identificação de proteínas com potencial para induzir resposta imune no hospedeiro. O termo vacinologia reversa foi inicialmente introduzido por Rino Rappuoli (RAPPUOLI, 2001) e validado a partir de estudos com Neisseria meningitidis do sorogrupo B (MenB), uma bactéria Gram-negativa responsável por casos de meningite meningocócica em nível global (KELLY; RAPPUOLI, 2005). Com a evolução das abordagens genômicas, proteômicas e imunoinformáticas, essas técnicas vêm sendo continuamente aplicadas em diagnóstico, terapia e desenvolvimento de vacinas.

A seleção de antígenos vacinais idealmente considera proteínas de membrana e proteínas extracelulares, uma vez que são mais acessíveis ao sistema imunológico do hospedeiro em comparação com proteínas intracelulares. Dessa forma, a determinação da localização subcelular é um fator crítico, especialmente na identificação de proteínas de membrana ou exportadas pela célula. Esse processo de identificação leva em conta características estruturais como domínios transmembranares, peptídeo sinal, motivos de ancoragem à membrana (como estruturas de barril beta e hélices transmembrânicas), atividade de adesina, presença de domínios conservados e funções moleculares associadas à virulência

do patógeno (BRENNAN & SHAHIN, 1996; ROOT, 2006; SERRUTO & RAPPUOLIi, 2006).

Diante desse cenário, o desenvolvimento de uma ferramenta baseada em ML que possua modelos próprios representa uma solução inovadora e de menor complexidade operacional. O aprendizado de máquina conta com um método particularmente poderoso, conhecido como redes neurais artificiais (Artificial Neural Networks, ANN). Essas redes consistem em modelos matemáticos que simulam o funcionamento dos neurônios biológicos, processando informações de maneira hierárquica (JAIN; MAO; MOHIUDDUN, 1996). O modelo clássico de ANN é composto por três camadas: entrada (input layer), oculta (hidden layer) e saída (output layer) (BASHEER; HAJMEER, 2000; GERSHENSON, 2003). Modelos mais sofisticados, que incluem múltiplas camadas e maior número de neurônios, são denominados redes neurais profundas (deep neural networks, DNN) e são responsáveis por avanços significativos em visão computacional, reconhecimento de padrões e processamento de linguagem natural (DENG et al., 2012; SZE et al., 2017).

No contexto da VR e do aprendizado profundo, este estudo considera as tarefas dentro da abordagem de Processamento de Linguagem Natural (Natural Language Processing, NLP), uma vez que as sequências de aminoácidos podem ser representadas como sequências textuais (CHOWSHARY, 2020). A estrutura modular das proteínas é análoga à composição da linguagem, onde motivos e domínios funcionam como palavras e frases (BRANDES et al., 2021). Assim, a integração de NLP e bioinformática é essencial para desenvolver preditores estruturais e funcionais para proteínas.

Com isso, este trabalho propõe o desenvolvimento da ferramenta ReVarcine, uma pipeline de VR baseada em aprendizado profundo, voltada para a identificação de antígenos vacinais potenciais por meio da predição de características estruturais essenciais.

4.2 Metodologia

Revisão Bibliográfica

A primeira tarefa foi a realização de uma pesquisa exploratória sobre o uso de *Deep Learning* para classificação de texto. Essa etapa foi necessária para proporcionar a compreensão do campo de estudo, possibilitando assim a definição dos objetivos de pesquisa e a obtenção dos modelos. Após a coleta e leitura dos artigos foi possível realizar a segunda tarefa da fase exploratória, onde foi realizada a definição dos recursos necessários para o desenvolvimento do projeto. Nessa etapa, foram verificados quais recursos deveriam ser construídos na fase de coleta de dados e desenvolvimento, apresentada na próxima seção.

Coleta de dados e pré-processamento

O primeiro conjunto de dados usado neste trabalho foi gerado para treinar e testar o método do ReVarcine. Os dados foram extraídos do UniProt/SwissProt submetidos até Janeiro de 2021, a fim de obter proteínas de bactérias Gram-positivas e Gram-negativas pertencentes a Filos de microorganismos patogênicos (Tabela 1). A extração ocorreu de forma automatizada através de um script em Bash. Será gerado um novo conjunto de dados de referência para comparar diferentes abordagens na detecção de peptídeo sinal e previsão do local de clivagem, localização subcelular, estrutura de beta barril transmembranar e alfa-hélice transmembranar. Para a validação do método, selecionaremos proteínas do Uniprot liberadas após Agosto de 2022. Isso permitirá excluir qualquer proteína já incluída no conjunto de dados usado para o treinamento.

Tabela 1 - Tabela dos Filos das bactérias, seus respectivos códigos taxonômicos e classificações de Gram.

Filo	Código Taxonômico	Tipo de membrana [a]
Actinobacteria	201174	Monoderme
Firmicutes	1239	Monoderme
Proteobacteria	1224	Diderme
Aquificae	200783	Diderme
Planctomycetes	203682	Diderme
Verrucomicrobia	74201	Diderme
Chlamydiae	204428	Diderme
Fibrobacteres	65842	Diderme
Cyanobacteria	1117	Diderme
Bacteroidetes	976	Diderme
Chlorobi	1090	Diderme
Spirochetes Kmlps6-30	135386	Diderme
Acidobacteria	57723	Diderme
Tenericutes	544448	Diderme

a = monoderme compreende os filos de bactérias classificados como Gram-positivas, enquanto diderme os classificados como Gram-negativos, mesmo para filos que não coram com o método de Gram.

Treinamento dos modelos

Neste trabalho, aplicamos Redes Neurais Profundas à análise de sequência de proteínas. Nesse caso, o domínio de entrada é um sinal unidimensional, onde cada posição em uma sequência é representada por um vetor multicanal (ou seja, multidimensional) que codifica o tipo de resíduo em cada posição de um proteína, um canal para cada tipo de resíduo. Os métodos foram desenvolvidos utilizando a biblioteca Keras em Python.

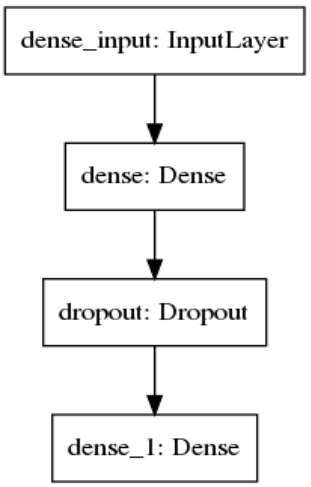
A previsão de peptídeo-sinal é uma tarefa relevante da classificação de proteínas onde o objetivo é detectar a presença/ausência do sinal sequência no terminal N da proteína. Para isso, usou-se a metodologia de classificação binária através da métrica *binary cross-entropy*, a qual refere-se a classe de métrica de entropia cruzada a ser usada quando houver apenas duas classes de rótulos 0 e 1, sendo 0 a classe de proteínas sem peptídeo, e 1 a classe de proteínas com peptídeo. A entropia cruzada binária compara cada uma das probabilidades previstas com a saída real da classe, que pode ser 0 ou 1. Em seguida, calcula a

pontuação que penaliza as probabilidades com base na distância do valor esperado. Isso significa quão perto ou longe do valor real.

A Figura 3a resume a arquitetura do método definido neste trabalho para a previsão peptídeo sinal para bactérias com uma membrana, compreendendo dois módulos básicos: a extração de recursos e a classificação. A base de dados utilizada possui 93.292 proteínas, sendo 2.433 da classe de proteínas com peptídeo sinal, e 90.859 da classe sem peptídeo sinal. O dataset para teste foi determinado para usar 25% da base de dados. Para balancear os dados, usou-se o método de undersampling.

Já a Figura 3b resume a arquitetura do método definido para bactérias com duas membranas. A base de dados utilizada possui 229.228 proteínas, sendo 7.381 da classe de proteínas com peptídeo sinal, e 221.899 da classe sem peptídeo sinal. O dataset para teste foi determinado para usar 25% da base de dados. Para balancear os dados, usou-se o método de undersampling.

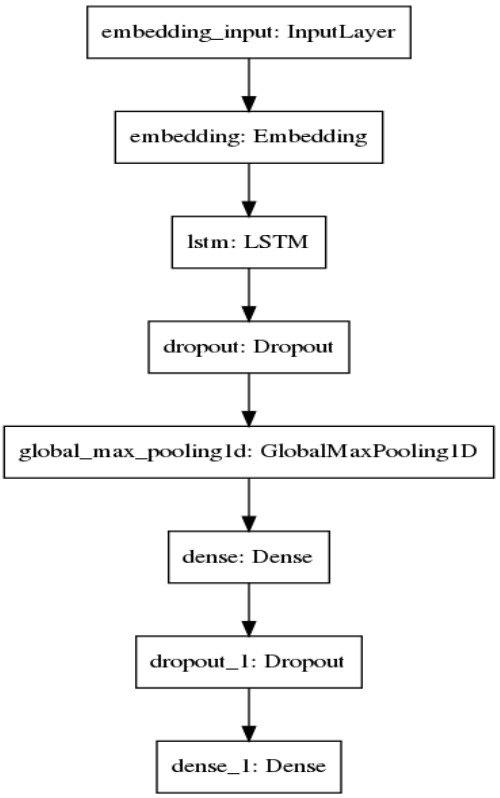
Figure 3 - Arquitetura do método para detectar peptídeo sinal em Gram-positivas e Gram-negativas, tendo como entrada sequência de proteína.



Quando um peptídeo sinal foi detectado com o método descrito acima, a sequência da proteína passa para o segundo estágio de previsão que identifica o

local de clivagem do peptídeo sinal. Nesse caso, cada local de clivagem foi considerado uma classe, nesse caso o método utilizado foi o modelo de classificação por esparsidade através da métrica de por *sparse categorical cross-entropy*. Esse tipo de modelo permite resolver problemas de classificação multiclasse que possuem vetores de recurso com alvos inteiros. A Figura 4 resume a arquitetura do método definido neste trabalho para a previsão local de clivagem do peptídeo sinal para bactérias com duas membranas. A base de dados utilizada possui 7.381 proteínas, todas com peptídeo sinal. O dataset para teste foi determinado para usar 25% da base de dados. Não foi realizado nenhum tipo de balanceamento dos dados.

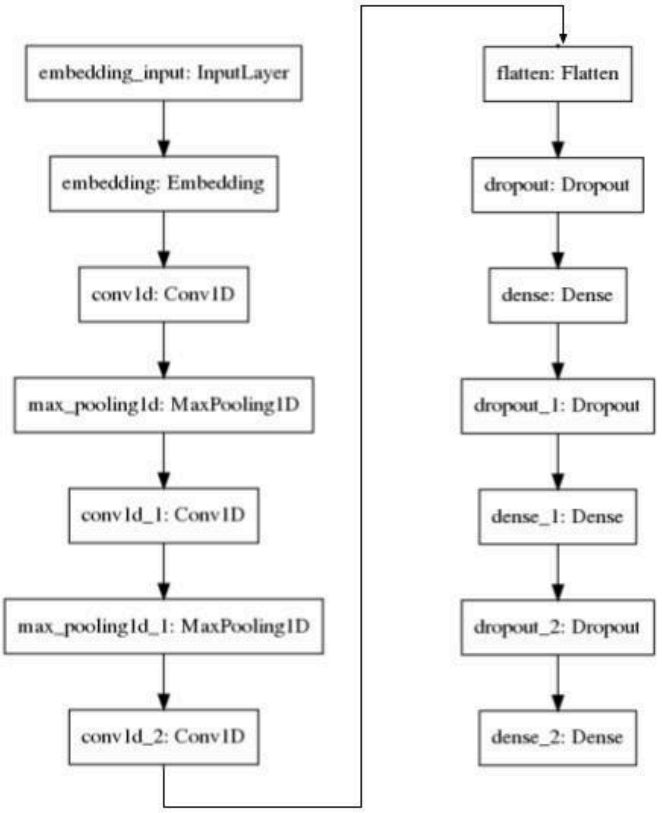
Figure 4 - Arquitetura do método para determinar o local de clivagem do peptídeo sinal, tendo como entrada sequência de proteína.



Já para o método para predição da localização subcelular para bactérias com duas membranas utilizou-se a metodologia de *categorical cross-entropy.* Essa

tarefa consiste em uma classificação multiclasse em que uma sequência de proteína só pode pertencer a uma das muitas categorias possíveis de localização subcelular, e o modelo deve decidir qual delas. A Figura 5 resume a arquitetura do método definido neste trabalho para a previsão da localização subcelular. A base de dados utilizada possui 90.177 proteínas, com as seguintes categorias: periplasmática, citoplasmática, fora da membrana e entre membranas. O dataset para teste foi determinado para usar 25% da base de dados. Não foi realizado nenhum tipo de balanceamento dos dados.

Figure 5 - Arquitetura do método para determinar a localização subcelular, tendo como entrada sequência de proteína.



Avaliação

O desempenho de previsão de detecção de peptídeo sinal foi medido usando o coeficiente de correlação de Matthews (MCC), no qual tanto as previsões verdadeiras quanto as falsas positivas e negativas são contadas em nível de sequência. O método de posição de clivagem do peptídeo sinal foi avaliado através da métrica Intersection over Union (IoU), também conhecido como Índice Jaccard, é uma das métricas bastante usadas em problemas de segmentação. A IoU é a área de sobreposição entre a segmentação predita e a verdadeira, dividida pela área de união entre a segmentação predita e a verdadeira. Além da IoU, utilizou-se a métrica de acurácia. Para o método de localização subcelular, utilizou-se a métrica de acurácia.

Validação e comparação do método com resultados do Estado da Arte

O método de predição para peptídeo sinal foi comparado ao Estado da Arte SignalP através das métricas Matthew score e acurácia. O método para posição de clivagem do peptídeo sinal foi avaliado através da métrica de IoU comparado ao software SignalP.

Implementação do método de peptídeo sinal e seu local de clivagem

A etapa final é a implantação do método desenvolvido como um serviço da web para facilitar o acesso à comunidade de pesquisa. Para isso, utilizou-se o Streamlit (https://streamlit.io/), o qual é uma biblioteca Python que pode ser usada para criar aplicativos da Web interativos usando código simples. Para dispor de forma gratuita a interface web do método, o código foi disponibilizado em um repositório público no GitHub. O aplicativo da web aceita como *input* a sequência de aminoácidos na forma de string e retorna se há ou não peptídeo sinal na sequência, e caso tenha ele retorna a sua respectiva posição de clivagem.

4.3 Resultados

Método de predição de peptídeo sinal

O método para predição de peptídeo sinal para bactérias com uma membrana apresentou em sua primeira versão um MCC de 0.55. As perda ao longo das épocas (Figura 6a) demonstra que a loss final foi de 0.08. Além disso, apresentou uma acurácia de validação de 0.97 (Figura 7a).

Já o método para bactérias com duas membranas, apresentou um MCC de 0.61. Na Figura 6b a perda ao longo das épocas teve um valor de 0.63. Apresentou, também, uma acurácia de validação de 0.80 (Figura 7b).

Figura 6 - Perdas ao longo das épocas no método de peptídeo sinal em bactérias. (A) método para bactérias com uma membrana. (B) método para bactéria com duas membranas.

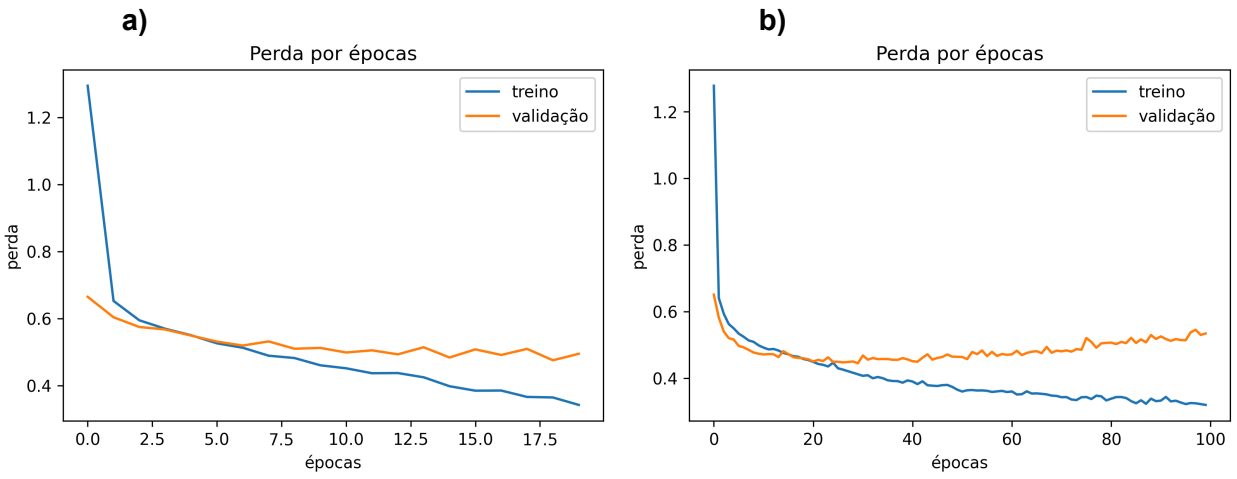
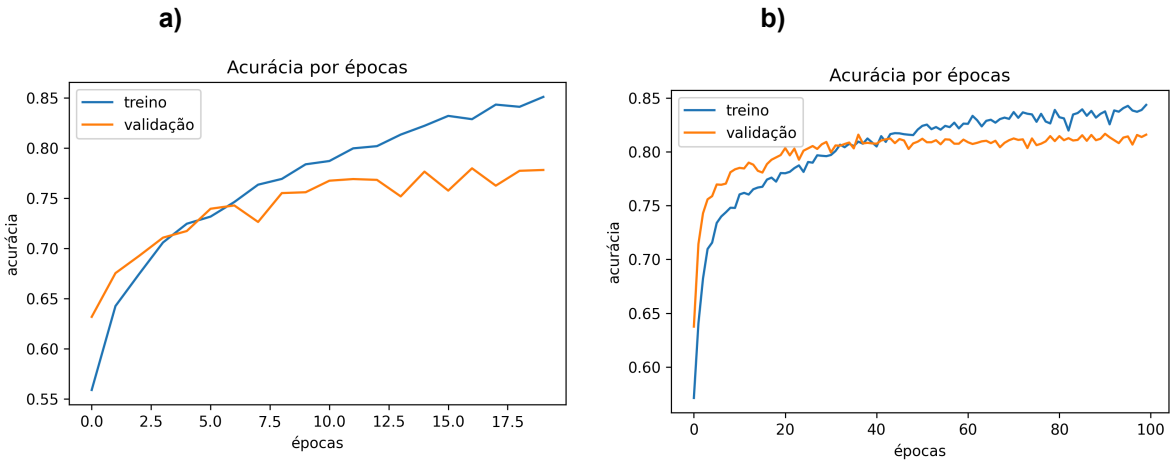


Figura 7 - Acurácia ao longo das épocas no método de peptídeo sinal em bactérias. (A) método para bactérias com uma membrana. (B) método para bactéria com duas membranas.



Método para posição de clivagem do peptídeo sinal

O método para posição de clivagem do peptídeo sinal, em sua primeira versão, apresentou IoU de a 0.97, loss de 0.89 (Figura 8) e acurácia de validação de 0.80 (Figura 9).

Figura 8 - Perdas ao longo das épocas no método de posição de clivagem do peptídeo sinal em bactérias com duas membranas.

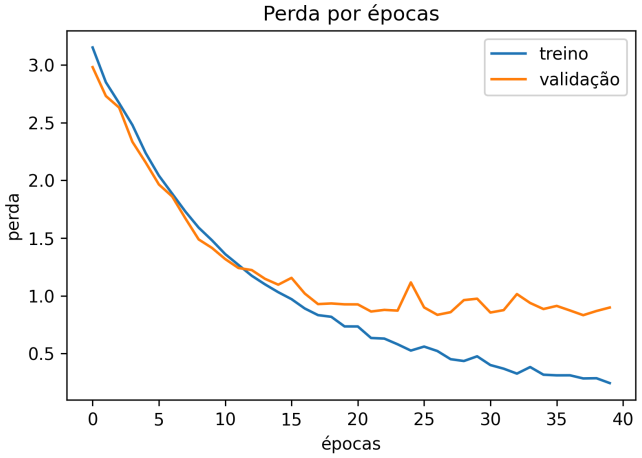
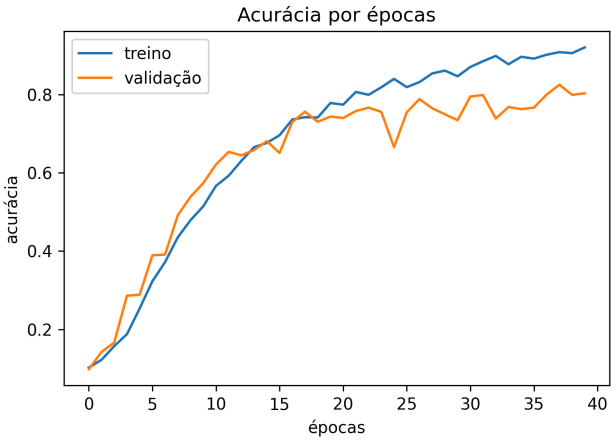


Figura 9 - Acurácia ao longo das épocas do método de posição de clivagem do peptídeo sinal em bactérias com duas membranas.



Método para localização subcelular

O método para classificação de localização subcelular apresentou uma loss de 0.19 e (Figura 10) acurácia de 0.94 (Figura 11).

Figura 10 - Perdas ao longo das épocas no método de localização subcelular para bactérias com duas membranas.

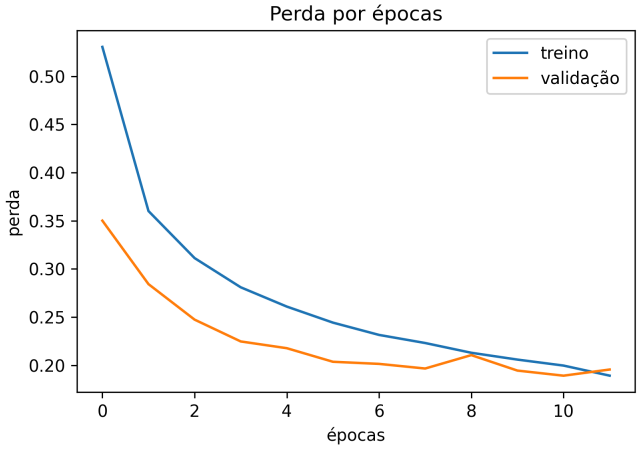
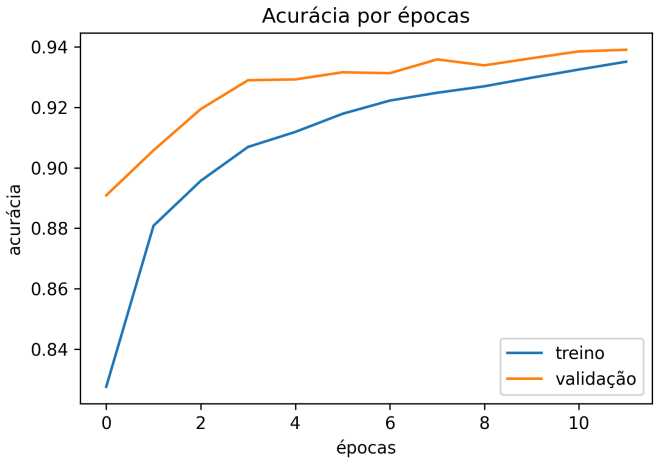


Figura 11 - Acurácia ao longo das épocas no método de localização subcelular para bactérias com duas membranas.



Validação e comparação do método com resultados do Estado da Arte

As métricas obtidas na validação e comparação do algoritmo SignalP, Estado da Arte na predição de peptídeo Sinal, com o método desenvolvido neste trabalho estão resumidas na Tabela 2.

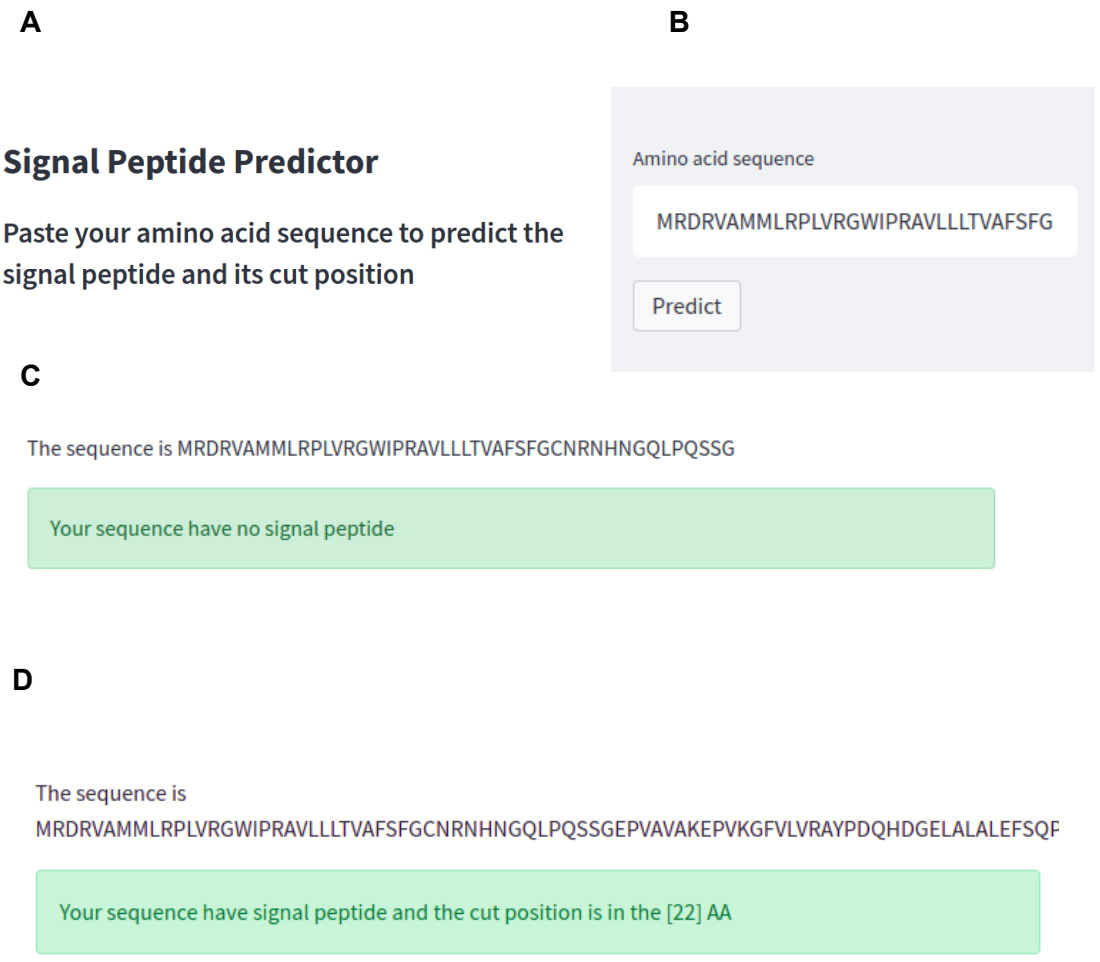
Tabela 2 - Tabela comparativa das métricas f1 score, Matthew score e acurácia do algoritmo SignalP e do método de peptídeo sinal ReVarcine.

Métrica	SignalP	Revarcine
f1 score	0.96	0.82
Matthew score	0.93	0.64
Acurácia	0.96	0.87

Implementação do modelo de posição de corte do peptídeo

A página inicial do servidor web do método de posição de clivagem do peptídeo sinal é mostrada na Figura 12a, enquanto a Figura 12b destaca o processo de submissão da sequência de aminoácidos. A Figura 12c ilustra a página de resultados mostrando qual é a posição de clivagem, ou seja, qual a posição do último resíduo do peptídeo sinal.

Figura 12 - Servidor Web do método de posição de clivagem do peptídeo sinal. (a) Página inicial do servidor Web; (b) local de submissão da sequência de aminoácidos; (c) resultado do método para proteínas sem peptídeo sinal; (d) resultado do método para proteínas com peptídeo sinal e seu respectivo local de clivagem.



5. Capítulo 2 - Artigo 1

Artigo submetido à revista *Protein Science*

Title: ReVarcine: novel transformer-based models for signal peptide, subcellular location, alpha helix and beta-barrel prediction in bacteria genomes

Amanda Munari Guimarães[1] ; Gratchela Rodrigues Dutra[2] ; Jacques Ben Bilombe[3] ; Luciano da Silva Pinto[1] ; Frederico Schmitt Kremer [1#]

[1] *Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, Rio Grande do Sul, Brazil*

[2] *Instituto de Biologia Departamento de Microbiologia e Parasitologia, Universidade Federal de Pelotas, Pelotas, Rio Grande do Sul, Brazil*

[3] *Departamento de Eletrônica, Centro Federal de Educação e Tecnologia Celso Suckow da Fonseca, Rio de Janeiro, Rio de Janeiro, Brazil*

Correspondence

Frederico Schmitt Kremer, *Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, Rio Grande do Sul, Brazil*

Email: fred.s.kremer@gmail.com

Abstract

Reverse vaccinology approach is an *in silico* methodology in order to identify antigens that are good vaccine candidates in a simple, safety and inexpensive way, with reduced time and effort. This strategy is based on bioinformatics tools, which predict important protein structural features, such as: integral transmembrane β-barrel arrangements; transmembrane alpha-helices; signal peptides; and secreted proteins. In this context, specific tools have been developed for the prediction of critical structural features, however despite significant progress, challenges persist due to the lack of integration in existing methods and the limited robustness and generalization of deep learning models with sparse data. To address these gaps, we introduce ReVarcine, a transformer-based deep neural network designed to automate the identification of signal peptides, subcellular localization, β-barrels, and alpha helices in bacteria, while prioritizing vaccine targets, advancing immunoinformatics and next-generation vaccine development. ReVarcine integrates predictions of signal peptides, subcellular localization, and structural features into a single automated workflow, generating detailed reports and prioritizing vaccine targets. Benchmarks against SignalP, PSORT, and PSIPRED demonstrate its superior predictive capabilities across diverse proteomes. By addressing key limitations in immunoinformatics, ReVarcine sets a new standard for computational tools in immunology and vaccinology, with potential for significant contributions through ongoing refinement and experimental validation.

Key-words: reverse vaccinology, vaccine, microbiology, bioinformatics

1. Introduction

The etymology of the word "vaccine" traces back to the Latin term *vaccinus*, which means "relating to the cow," derived from *vacca*, the Latin word for "cow." This origin is closely linked to the pioneering discovery of Edward Jenner in the 18th century, who observed that exposure to the cowpox virus provided immunity against smallpox, a deadly disease at the time. Jenner used material from cowpox lesions to inoculate individuals, a process known as "vaccination," referencing cows, thus solidifying the term vaccine as a cornerstone of modern immunization [1]. In the 1700s, farmer Benjamin Jesty, in collaboration with Jenner, noted that dairy cows were not affected by smallpox, inferring that cowpox might protect them. Jesty inoculated his own family with smallpox spores [2], while Jenner later conducted clinical trials in the 18th century and published his results worldwide, contributing to the eradication of smallpox [3, 4].

The development of traditional vaccines was largely based on the principles of Pasteur. This approach involves isolating, inactivating, and injecting the causative agent of the disease. However, with next-generation sequencing technologies and the growing number of pathogen genome sequences available in databases like NCBI [5], GenBank [6], EuPathDB [7], and the Virus Pathogen Database and Analysis Resource (ViPR) [8], combined with advances in pathogenesis and immunology, a new era in vaccine research and development has emerged.

In recent years, the ever-increasing and refinements in quality of pathogenic bacteria genomes with advances in bioinformatics allowed the development and establishment of the process known as Reverse Vaccinology (RV) [9]. RV approach is an *in silico* methodology in order to identify antigens that are good vaccine candidates in a simple, safety and inexpensive way, with reduced time and effort. This strategy is based on bioinformatics tools, which predict important protein structural features, such as: integral transmembrane β-barrel arrangements; transmembrane alpha-helices; signal peptides; and secreted proteins [10, 11]. This provided

a rigorous analysis and the identification of vaccine candidates for subsequent evaluation in the laboratory.

Signal Peptides (SPs) are short peptides in the N-terminal of proteins, and are found in secreted and transmembrane (TM) proteins in Bacteria machine, as well as in proteins inside organelles in eukaryotic cells. [12]. In Gram-negative bacteria, the protein could be retained in the periplasm, or be inserted into the outer membrane as a β-barrel transmembrane protein [13]. In Gram-positive bacteria, the protein could be attached to the cell wall [14]. Comparative sequence analysis showed that SPs have a typical length of 20-30 residues, but in some cases may be up to 60 residues long [15].

Protein's function is directly associated with its subcellular location, the prediction of these classification plays an important role in RV. In particular, proteins that are exposed on the surfaces of pathogens attract the attention of the host immune system, and are mostly antigenic and responsible for host pathogenesis, therefore the most suitable vaccine candidates. The secreted proteins are likely toxins or other virulence factors.

With the advent of artificial intelligence, deep learning has established itself as a promising tool in areas such as computational biology and immunoinformatics. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models have been widely applied in the prediction of biological features, such as protein secondary structures and antigenic epitopes (CHEN et al., 2019; XIAO et al., 2022). These models stand out for their ability to capture complex patterns in biological data, often outperforming traditional methods based on statistics or conventional machine learning.

In the context of reverse vaccinology, specific tools have been developed for the prediction of critical structural features. Signal peptide prediction models, such as SignalP, are widely used to identify secreted proteins, which often act as vaccine targets [18]. Similarly, subcellular localization prediction, performed by tools such as PSORT, allows prioritizing proteins relevant to host-pathogen interaction. Finally, advanced secondary

structure prediction techniques, such as beta-barrel and alpha helices, have received increasing attention due to their relevance in antigenic stability and functionality (ZHANG et al., 2020).

Although these advances represent significant progress, gaps still remain. Many of the current methods lack integration, making it difficult to create comprehensive, automated pipelines for reverse vaccinology. Furthermore, issues related to the robustness and generalization of deep learning models, particularly in scenarios with limited data, remain challenging [20]. In this sense, we present ReVarcine, a method based on a deep neural network, combined with the transformer architecture, whose main objective is to provide an efficient computational tool that automates the identification of signal peptides, subcellular localization, β-barrel and alpha helix in bacteria and prioritization of vaccine targets, contributing to significant advances in the field of immunoinformatics and in the development of new generation vaccines.

2. Results

The performance of ReVarcine models in signal peptide prediction, subcellular localization, and secondary structure tasks highlights its superior capabilities, driven by innovative integrations of advanced deep learning architectures, such as transformers and recurrent networks, alongside optimized preprocessing and modular model design. For signal peptide detection, ReVarcine achieved notable results, including an overall MCC of 0.937 for monodermal bacteria and 0.969 for didermal bacteria (Table 1). Intersection over Union (IoU) metrics were equally impressive, reaching 0.954 and 0.970 for monodermal and didermal bacteria, respectively (Table 1). Furthermore, the model demonstrated precision scores of 0.935 and 0.912, accuracy levels of 0.953 and 0.967, and recall values of 0.732 and 0.790 for monoderms and diderms, respectively (Table 2). These outcomes underscore the model's ability to detect and delineate signal peptides with superior precision, a performance facilitated by its use of transformer blocks and LSTM layers, which enable the capture of complex sequential dependencies. The advanced architecture and fine-tuned hyperparameters of ReVarcine demonstrate its substantial advantages over traditional approaches such as SignalP.

The performance of ReVarcine models in subcellular localization tasks demonstrates its superior capabilities, driven by the integration of advanced deep learning architectures, including transformers and recurrent networks, as well as optimized preprocessing and modular model design. In terms of accuracy, ReVarcine outperformed the benchmark PSORT, achieving 0.953 for monodermal bacteria and 0.967 for didermal bacteria, compared to PSORT's 0.741 and 0.698, respectively. For precision, ReVarcine also showed substantial improvement, with values of 0.912 for monodermal and 0.935 for didermal bacteria, while PSORT scored 0.326 and 0.272, respectively. Additionally, the model demonstrated recall scores of 0.732 for monoderms and 0.790 for didermal bacteria, surpassing PSORT's 0.486 and 0.467 (Table 2). These results highlight ReVarcine's ability to predict subcellular localizations with greater accuracy and reliability, emphasizing its

advantages over traditional tools like PSORT. This superior performance is enabled by its use of advanced architectures such as transformer blocks and recurrent layers, which capture complex dependencies within the data.

The performance of ReVarcine models in beta barrel tasks underscores its advanced capabilities. In terms of F1 score, ReVarcine outperformed the benchmark PSIPRED, achieving 0.872 for didermal bacteria, compared to PSORT's 0.791 (Table 3). For the alpha helix task, ReVarcine demonstrated further enhancement, with an F1 score of 0.826 for didermal bacteria, while PSIPRED attained 0.819 (Table 4). These results highlight ReVarcine's exceptional ability to predict protein structures with greater accuracy, making it a robust tool for structural bioinformatics applications and potential vaccine design.

The superior results obtained by ReVarcine in comparison to traditional tools can be attributed to several factors that ensure its efficacy and predictive accuracy, such as innovative deep learning architectures and the modular integration of models. While benchmark tools rely on more static or limited approaches, ReVarcine adopts advanced deep learning architectures, including transformers and recurrent neural networks. The use of complex models and integrated pipelines allows ReVarcine to more effectively model the intricate dependencies within the data, contributing to its superior performance, personalized training with large datasets, and the inclusion of multiple evaluation metrics.

Another key factor distinguishing ReVarcine is its use of large databases, such as UniProt, for training its models. Training with robust datasets provided ReVarcine with greater generalization capacity and more accurate predictions, which is essential in reverse vaccinology contexts, where the nuances of biological data play a crucial role. Additionally, ReVarcine incorporates several evaluation metrics, including MCC (Matthews Correlation Coefficient), IoU (Intersection over Union), precision, and recall. Considering these multiple metrics enables a deeper analysis of the model's performance, capturing nuances that traditional tools cannot adequately assess, resulting in a more comprehensive and precise evaluation.

ReVarcine’s modular approach, which allows for the integration of predictions from various stages of the process (such as signal peptide prediction, subcellular localization, and beta-barrel and alpha-helix structure prediction), is also a significant advantage. This integration results in more detailed and comprehensive analyses, which contribute to the optimization of the model's final performance and its greater ability to deliver accurate and reliable predictions. Combined, these factors enable ReVarcine to stand out among other tools, providing superior performance and a more robust predictive capacity.

3. Discussion

The field of immunoinformatics has rapidly evolved, offering computational methods that augment traditional immunology approaches by enabling the prediction, analysis, and modeling of immune responses [21]. However, several gaps remain in existing methodologies, particularly concerning the seamless integration of automated pipelines for tasks such as antigen discovery, epitope prediction, and vaccine design [22].

ReVarcine was specifically designed to address these challenges by offering a unified and automated framework for the prediction of signal peptides, subcellular localization, and structural motifs like beta-barrels and alpha-helices. Traditional tools, such as SignalP, PSORT, and PSIPRED, often excel at individual predictive tasks but lack the interoperability required for comprehensive immunoinformatics analyses [23, 24, 25]. ReVarcine fills this critical gap by integrating these predictive capabilities into a single pipeline, significantly reducing computational overhead and enhancing accuracy across tasks [26]. This integrated approach provides a holistic framework for the *in silico* identification of vaccine targets [27, 28].

One major advantage of ReVarcine is its potential to streamline vaccine design workflows. Automated pipelines like ReVarcine align with the growing need for rapid and scalable methods, particularly in the context of emerging infectious diseases, where time-efficient solutions are paramount [29, 30]. Its ability to integrate multiple predictions within a single framework minimizes manual data transfer between tools, thereby reducing the likelihood of errors and improving data consistency.

The comprehensive evaluation metrics, including F1 score, accuracy, precision, recall, and Matthews correlation coefficient (MCC), demonstrate the robustness of Revarcine across all evaluated domains. These superior metrics are indicative of a broader trend in immunoinformatics: the adoption of deep learning and artificial intelligence to enhance prediction accuracy and pipeline integration [31, 32]. By leveraging neural network architectures, ReVarcine outperforms classical machine learning approaches traditionally

employed in tools like SignalP, PSORT and PSIPRED, offering better generalization and improved adaptability to diverse datasets. Moreover, ReVarcine provides researchers with actionable insights that are directly applicable to reverse vaccinology. Its accurate predictions of subcellular localization, combined with secondary structure analysis, enable the identification of targetable epitopes more efficiently than standalone tools. This comprehensive functionality underscores its transformative potential in vaccine design, helping researchers accelerate the journey from genome to vaccine [33].

Another key application of ReVarcine lies in its role as part of preclinical antigen screening pipelines. The tool accelerates the identification of viable candidates by reducing dependency on time-intensive laboratory analyses, allowing researchers to focus on experimentally validating a narrowed list of the most promising targets [34]. This capability is especially critical in the context of rapid vaccine development, where timelines are constrained, such as during pandemics or outbreaks of resistant pathogens.

Importantly, Revarcine’s versatility and precision underscore its potential to bridge gaps in current computational tools by providing a unified solution for multiple predictive tasks. By streamlining the prediction process, it minimizes the dependency on multiple specialized tools, reducing computational time and errors arising from cross-platform data handling,it sets a new standard for tools designed to meet the ever-growing demands of computational immunology and vaccinology. [35].

4. Conclusions

The ReVarcine pipeline was developed to integrate the models, allowing predictions of signal peptide, subcellular localization, and structural features to be combined into a single workflow. Each module of the pipeline was designed to generate detailed reports and prioritize vaccine targets. The benchmark performed with the SignalP, PSORT and PSIPRED tools indicates that it can predict structures across the proteome of diverse organisms, often better than expert predictors. ReVarcine exemplifies the type of innovation required to address longstanding limitations in immunoinformatics. By combining advanced predictive capabilities with automation and integration, it sets a new standard for tools designed to meet the ever-growing demands of computational immunology and vaccinology. With continued refinement, experimental validation, and expanded data integration, ReVarcine is poised to make substantial contributions to the fields of immunoinformatics and vaccinology.

5. Materials and methods

The manuscript primarily serves as a tool description. Therefore, we provide only a brief overview here, with additional details—including all intermediate steps, and commands in the Supporting Information and the online Tutorial.

5.1 Data collection and preprocessing

The protein sequences used in this study were obtained from the UniProt and Swissprot database, recognized for its comprehensiveness and high-quality curation. Only complete sequences annotated with the features required for the prediction tasks, such as the presence of signal peptides, subcellular localization, and structural features including transmembrane β-barrels and alpha helices, were selected. In the filtering step, incomplete or redundant sequences were excluded. Furthermore, the datasets of Gram-positive bacteria were separated from Gram-negative, since specific models were developed for each of them. For the subcellular localization dataset, an extra classification selection step was applied, in which the classifications of cell inner membrane, cell outer membrane, cytoplasm, periplasm and secreted were selected. In the encoding step, the amino acid sequences were processed to ensure that deep learning models could interpret and learn patterns from the data. To do this, a tokenization method based on the Keras library was used, which is suitable for dealing with categorical sequences, such as amino acids. This process involved tokenization, in which each sequence was transformed into a numerical representation suitable for computational analysis. The tokenizer was configured with a maximum limit for the vocabulary. This parameter defines the maximum number of unique tokens that were considered during the process, prioritizing the most frequent tokens in the training sequences. The fit_on_texts method was applied to the training sequences contained in the sequence column of the dataframe. This step creates a vocabulary, where

each token (amino acid or subsequence) is mapped to a numerical index based on its frequency. The amino acid sequences were converted to lists of numerical indexes using the texts_to_sequences method. In this step, each token in the original sequence is replaced by the corresponding index defined in the generated vocabulary. Unrecognized tokens or those outside the vocabulary boundary are discarded. To ensure compatibility with deep learning models, all sequences were standardized to a fixed length defined by the sequence size. Sequences shorter than the fixed length were padded with zeros at the end, while longer sequences were truncated to the defined length, keeping the first elements. This process ensures that all sequences have the same length without changing the order of the tokens. The operation was performed using the pad_sequences function from the Keras library. In addition, the data was divided into training (75%) and test (25%) sets, ensuring a balance between the classes in each set. The dataset for identifying signal peptides in Gram-negative and Gram-positive bacterias; predicting peptide cleavage sites for both bacterial classifications; performing multi-class classification of subcellular localization for both types of bacteria; and detecting beta-barrel and alpha-helix structures in Gram-negative bacteria is summarized in Table 5. The choice of the number of inputs in the training set aimed to guarantee statistical robustness for model adjustment, while the number in the test set ensured a reliable assessment of performance in real situations. Additional tests were conducted to verify the representativeness of the generated subsets.

5.2 Development of deep learning models

Eight models were developed, each tailored to a specific task: identifying signal peptides in Gram-positive and Gram-negative bacteria (Supplementary material 1); predicting peptide cleavage sites for both bacterial classifications (Supplementary material 2); performing multi-class classification of subcellular localization for both types of bacteria (Supplementary material 3); and detecting beta-barrel (Supplementary

material 4) and alpha-helix structures in Gram-negative bacteria (Supplementary material 5). For the signal peptide prediction task, a deep learning model was developed combining different layers to capture important features of protein sequences. The model uses an architecture based on embeddings, transformer blocks and recurrent networks to deal with complex sequence patterns. The model was implemented using the Keras library, with the following main layers: embedding layer, transformer block, Long Short-Term Memory (LSTM) layer, dense layers, dropout. In addition, binary cross-entropy was used, appropriate for binary classification problems. The Adam algorithm was chosen due to its efficiency in adjusting weights in deep networks. Peptide cleavage site model was designed for multi-class classification tasks, leveraging advanced deep learning techniques such as transformer blocks, LSTM networks, and pooling mechanisms. In this framework, the cleavage site was identified as the amino acid position marking the start of the cleavage, with each position within a sequence treated as a distinct category. The model was designed to predict the subcellular localization of proteins based on their sequences, using a hybrid approach that combines convolutional layers (Conv1D), batch normalization (Batch Normalization), transformers and dropout regularization. For both the binary classification model of the presence of beta barrel and alpha helix, the model was designed with the combination of embedding layers, Transformer blocks, LSTM and dense layers, to capture complex relationships and extract relevant information from the input sequences. The flowchart of each model architecture was generated using Python and demonstrates each layer constructed (Supplementary material 1-5).

5.3 Integrated models pipeline

ReVarcine is structured to integrate and execute multiple deep learning models in sequence, enabling the analysis of protein data in a scalable manner that is adaptable to the specific characteristics of the sequences (Figure 1). Model integration occurs in three main steps: pre-processing, model execution, and post-processing. In the pre-processing step, the input sequences are provided in a .fasta file and are processed,

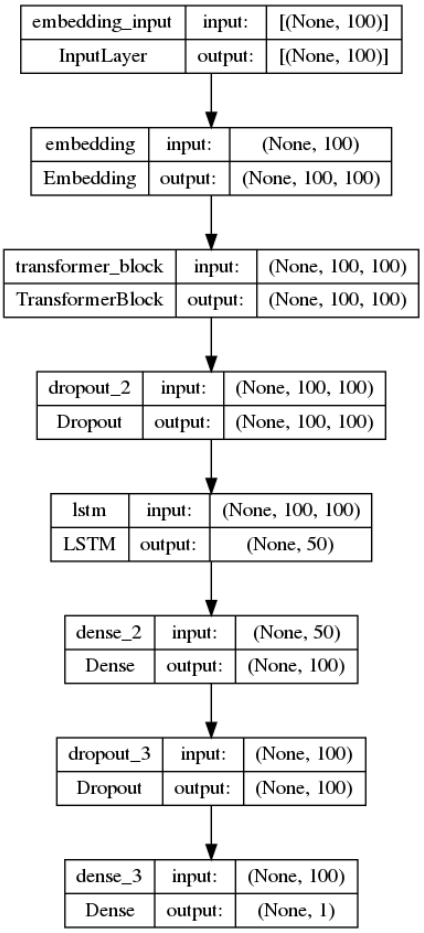
generating the result of a dataframe containing the sequences prepared to be used by the models, ensuring consistency in format and attributes. Depending on the classification of the bacteria as Gram-positive (G+) or Gram-negative (G-), a specific set of models is selected (Table 6). The models are executed in a modular and sequential manner, each model is associated with a specific class (G+ or G-) with methods that encapsulate its logic. The selection of models occurs based on the parameters provided by the user, which determine which analyses will be performed. For each sequence, the selected models are applied. If a model is successfully called, its results are stored for later processing. After the models are run, the raw results are processed to combine and organize the data into a useful format for prioritizing vaccine targets. This step synthesizes the results of the different analyses and provides a consolidated report.

5.4 Evaluation metrics

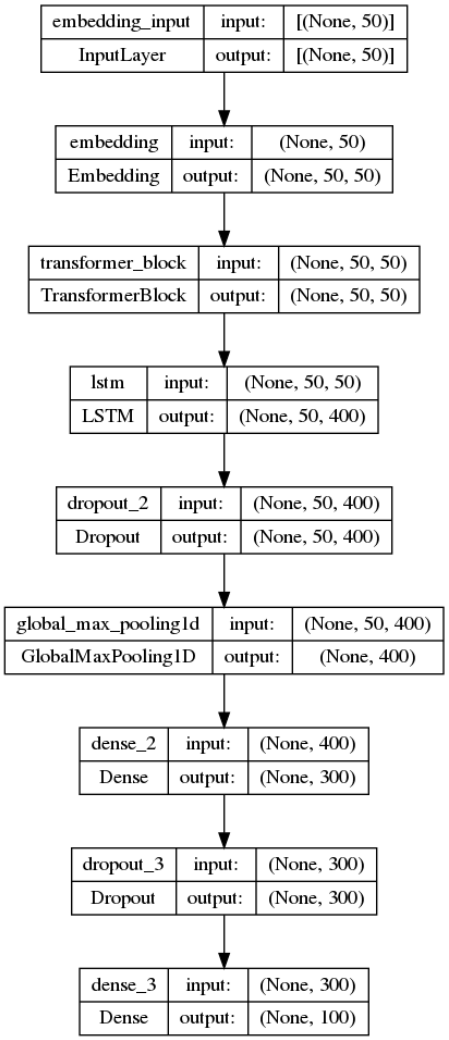
The prediction performance of the SP detection models was measured using the Matthews correlation coefficient (MCC), in which true and false positive and negative predictions are counted at the sequence level. We also used intersection over union (IoU), which in segmentation analyses evaluates the overlap between the predicted and true masks to identify how well the model segmented the sequence. To evaluate the subcellular localization models, we used precision and recall to evaluate the predictions, where precision is defined as the fraction of the location predictions that are correct, and recall is the fraction of true locations that are predicted as the correct subcellular location type and have the correct classification assigned. Also MCC and F1 score for beta barrel and alpha helix. F1 score emphasizes the balance between correctly identified positive instances and the minimization of false results by prioritizing the harmonic mean rather than a simple arithmetic mean, the F1 score effectively penalizes extreme disparities between precision and recall, ensuring that neither metric is ignored. Consequently, it has become a standard evaluation criterion in binary classification problems.

Furthermore, we evaluated each ReVarcine model against the tools SignalP for signal peptide, PSORT for subcellular localization, and PSIPRED for beta-barrel and alpha-helix structure, which are available both as web servers and for CLI use.

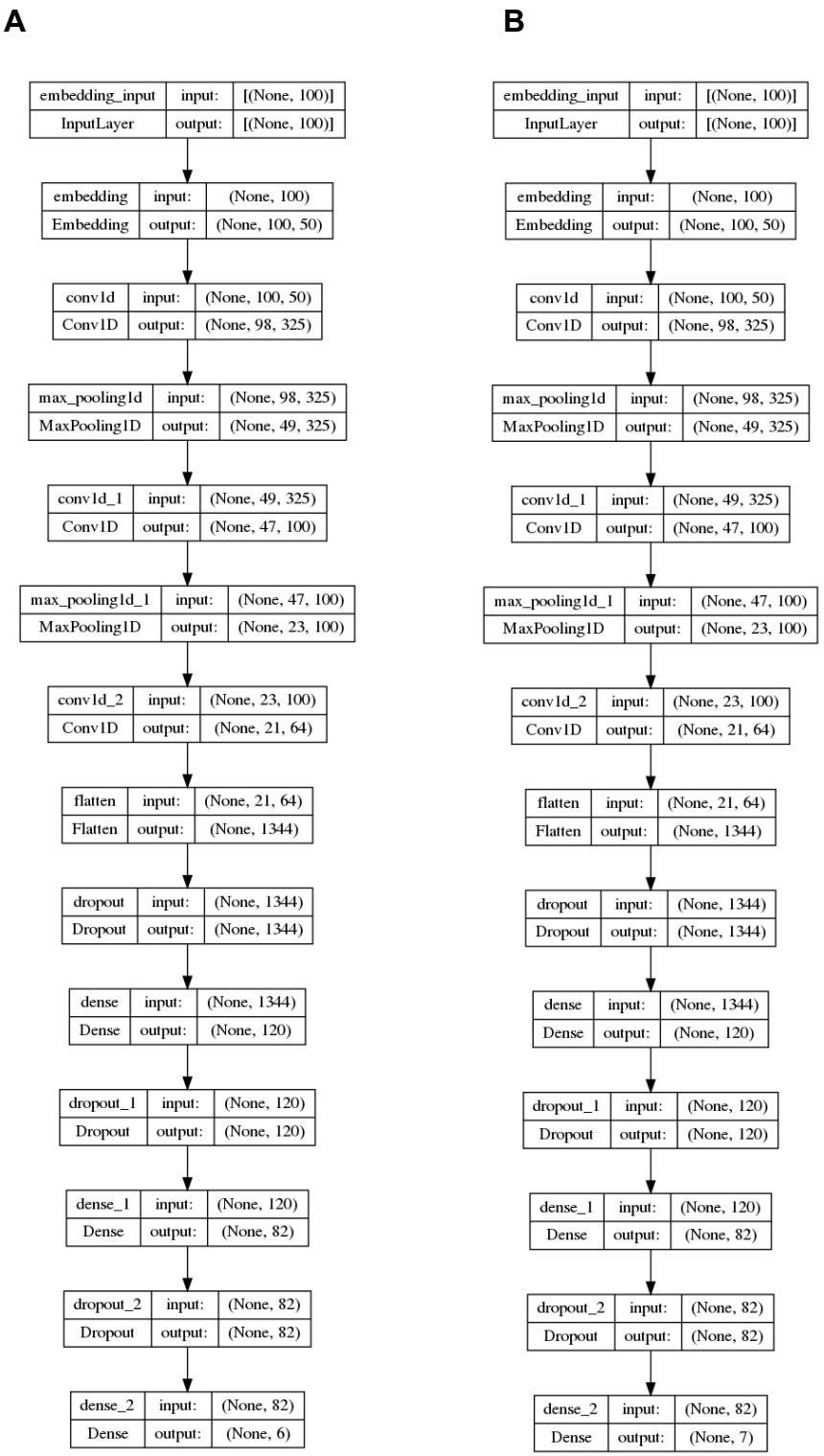
6. Supplementary material



Supplementary material 1. Signal peptide deep learning model architecture for Gram-positive and Gram-negative bacterias. The network starts with an Input Layer that accepts sequences of shape (None, 100), followed by an Embedding Layer that maps the input into a representation of shape (None, 100, 100). A Transformer Block processes these embeddings, keeping the same shape (None, 100, 100). A Dropout Layer is applied for regularization. Next, an LSTM Layer extracts sequential features and reduces the dimensionality to (None, 50). This is followed by a Dense Layer that expands the representation to (None, 100), another Dropout Layer for regularization, and a final Dense Output Layer that maps the processed features to a single output of shape (None, 1).

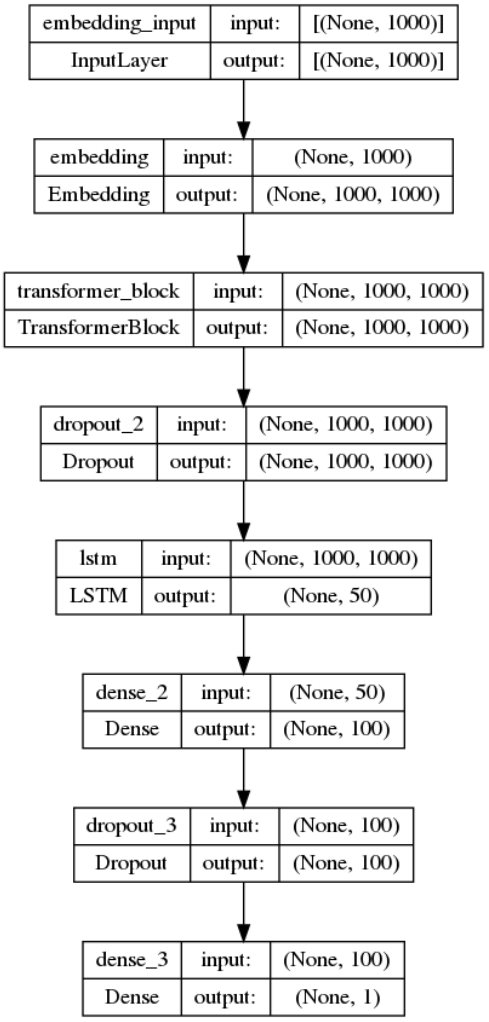


Supplementary material 2. Signal peptide cut position deep learning model architecture for Gram-positive and Gram-negative bacterias. The network begins with an Input Layer that accepts sequences of shape (None, 50), followed by an Embedding Layer, which maps the input to a higher-dimensional space (None, 50, 50). A Transformer Block processes these embeddings, maintaining the shape (None, 50, 50). Next, an LSTM Layer extracts sequential features, expanding the representation to (None, 50, 400). A Dropout Layer is applied for regularization, followed by a Global Max Pooling Layer, which reduces the dimensionality to (None, 400). A Dense Layer maps the representation to (None, 300), followed by another Dropout Layer. Finally, a Dense Output Layer produces the output representation of shape (None, 100).

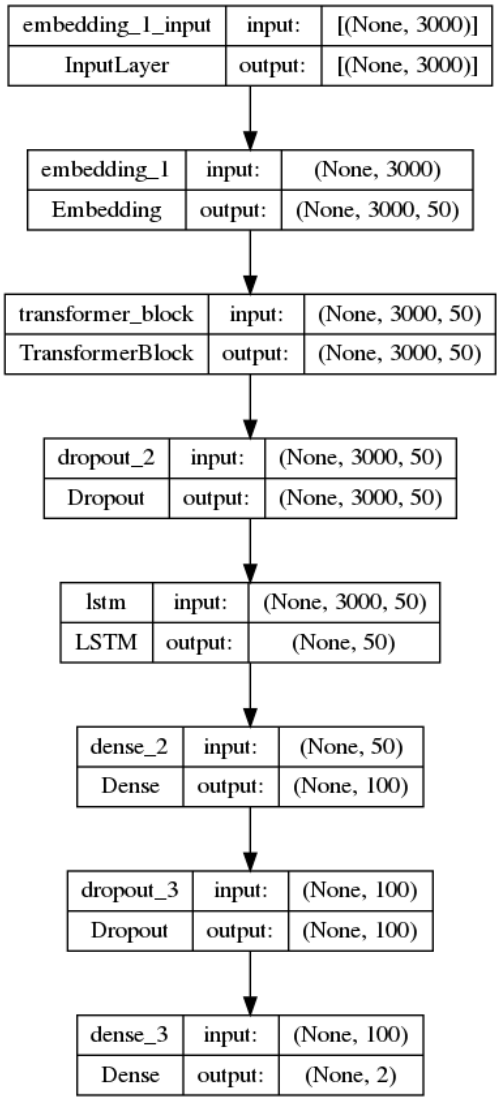


Supplementary material 3. Subcellular location deep learning model architecture for Gram-positive and Gram-negative bacterias. **A** - Gram-negative networks, the model starts with an Input Layer accepting sequences of shape (None, 100), followed by an Embedding Layer that transforms the input into a representation of shape (None, 100, 50). Three Conv1D Layers with kernel operations extract hierarchical features, each followed by

MaxPooling1D Layers to reduce dimensionality. These layers produce progressively smaller feature maps, culminating in a shape of (None, 21, 64). A Flatten Layer reshapes the output into a 1D vector of size (None, 1344). The network includes two fully connected Dense Layers of sizes 120 and 82, interspersed with Dropout Layers for regularization. The final Dense Output Layer maps the representation to a vector of size (None, 6), suitable for multi-class classification or regression tasks. **B** - Gram-positive networks, the network starts with an Input Layer that processes sequences of shape (None, 100). An Embedding Layer projects the input into a representation of shape (None, 100, 50). Three Conv1D Layers extract hierarchical features with output shapes (None, 98, 325), (None, 47, 100), and (None, 21, 64), each followed by corresponding MaxPooling1D Layers for dimensionality reduction. A Flatten Layer reshapes the final feature map into a 1D vector of size (None, 1344). The network includes two fully connected Dense Layers of sizes 120 and 82, interspersed with Dropout Layers for regularization. The final Dense Output Layer produces an output vector of shape (None, 7), suitable for a seven-class classification task.



Supplementary material 4. Beta barrel deep learning model architecture for Gram-negative bacterias. The network begins with an Input Layer accepting sequences of shape (None, 1000) and an Embedding Layer that projects them into a higher-dimensional space (None, 1000, 1000). A Transformer Block processes these embeddings, followed by a Dropout Layer for regularization. An LSTM Layer extracts sequential features, reducing the representation to (None, 50). This is followed by a Dense Layer expanding the features to (None, 100) and another Dropout Layer. The final Dense Output Layer maps the processed features to a single output (None, 1), suitable for the task at hand.



Supplementary material 5. Alpha helix deep learning model architecture for Gram-negative bacterias. The network begins with an Input Layer accepting sequences of shape (None, 3000), followed by an Embedding Layer that converts the input into a representation of shape (None, 3000, 50). A Transformer Block processes these embeddings, maintaining the shape (None, 3000, 50). A Dropout Layer is applied for regularization. Next, an LSTM Layer extracts sequential features and reduces the dimensionality to (None, 50). This is followed by a Dense Layer expanding the features to (None, 100) and another Dropout Layer for regularization. Finally, a Dense Output Layer maps the features to a two-dimensional vector (None, 2), suitable for binary classification.

7. Acknowledgments

We thank the financial support from the Capes Institute.

8. Data availability statement

Revarcine is freely available at https://github.com/omixlab/revarcine-pipeline.

9. References

1. Holland M. *Vaccines: A Biography*. Springer; 2010.
2. Pead, P.J. Benjamin Jesty: Dorset's Vaccination Pioneer. Timefile Books, Chichester, 2009.
3. Jenner, E. An Inquiry into the Causes and Effects of the Variolae Vaccinae. Low, London, 1798.
4. Bazin, H. Vaccination: A History. From Lady Montagu to Genetic Engineering. John Libbey Eurotext, Montrouge, 2011.
5. Agarwala, R., et al. (2018). NCBI: National Center for Biotechnology Information. Available at: https://www.ncbi.nlm.nih.gov
6. Clark, K., et al. GenBank: Nucleotide sequence database. Nucleic Acids Research, 44(D1), 2016.
7. Warrenfeltz, S., et al. (2018). EuPathDB: Bioinformatics tools for the study of pathogens. Available at: https://eupathdb.org/eupathdb/
8. Pickett, B. E., et al. *ViPR: The Virus Pathogen Database and Analysis Resource*. BMC Bioinformatics, 13, 510, 2012.
9. Rappuoli R. Reverse vaccinology. *Curr Opin Microbiol*. 2000;3:445–50.
10. Heinson AI, Woelk CH, Newell M-L. The promise of reverse vaccinology. *Int Health*. 2015;7:85–9.
11. Bowman BN, McAdam PR, Vivona S, Zhang JX, Luong T, Belew RK, et al. Improving reverse vaccinology with a machine learning approach. *Vaccine*. 2011;29:8156–64.
12. Owji H, Nezafat N, Negahdaripour M, Hajiebrahimi A, Ghasemi Y. A CHENcomprehensive review of signal peptides: Structure, roles, and applications. *Eur J Cell Biol*. 2018;97:422–41.
13. Duong F, Eichler J, Price A, Leonard MR, Wickner W. Biogenesis of the Gram-Negative Bacterial Envelope. *Cell*. 1997;91:567–73.
14. Mazmanian SK, Liu G, Ton-That H, Schneewind O. Staphylococcus aureus Sortase, an enzyme that anchors surface proteins to the cell wall. *Science*. 1999;285:760–3.
15. Hiss JA, Schneider G. Architecture, function and prediction of long signal peptides. *Brief Bioinform*. 2009;10:569–78.
16. Almagro JC, et al. Signal peptide prediction using machine learning approaches. *Bioinformatics Advances*. 2019;35(4):782-790.
17. Krogan, N. J., et al. (2021). *Deep Learning Approaches for Reverse Vaccinology*. Nature Reviews Microbiology, 19(7), 491-504.
18. Flower DR. Bioinformatics for vaccines and immunotherapeutics. *Bioinformatics*. 2003;19(2):239–245.
19. Liao Y, et al. Applications of machine learning in biomedical research. *Trends Pharmacol Sci*. 2020;41(3):172–186.
20. Petersen TN, et al. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*. 2011;8(10):785–786.
21. Nakai K, Kanehisa M. Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins: Struct Func Genet*.

1991;11(2):95-110.
22. Bryson K, et al. Protein structure prediction with PSIPRED. *Proteins: Struct Func Bioinform*. 2005;61(S7):184-192.
23. Goodfellow I, Bengio Y, Courville A. *Deep learning*. MIT Press; 2016.
24. Smith A, et al. Integrative approaches in computational vaccinology: A review. *J Comput Biol*. 2021;28(10):1055-1072.
25. Jones DT, Lee D. Advances in computational tools for reverse vaccinology. *Curr Opin Immunol*. 2022;75:45-53.
26. Petrovsky N, Brusic V. Computational predictive models for T-cell epitopes. *Nat Rev Immunol*. 2002;2(9):707-718.
27. Flower DR, et al. Reverse vaccinology in the age of big data. *Immunology*. 2021;162(2):154-169.
28. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
29. Li W, et al. A comprehensive review of AI in vaccine design. *Comput Biol Chem*. 2022;100:107717.
30. Soria-Guerra RE, et al. An overview of bioinformatics tools for epitope prediction. *Drug Discov Today*. 2015;20(4):403–410.
31. Vita R, et al. The immune epitope database (IEDB). *Nucleic Acids Res*. 2019;47(D1):D339–D343.
32. Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge: MIT Press; 2016.

10. Tables

Table 1 - Signal peptide prediction performance evaluation between ReVarcine and SignalP.

Metric	Membrane type [a]	ReVarcine	SignalP
Matthew correlation coefficient	monoderm	0.937	0.896
	diderm	0.969	0.928
Intersection Over Union	monoderm	0.954	0.946
	diderm	0.970	0.937

a = monoderm comprises the phyla of bacteria classified as Gram-positive, while diderm those classified as Gram-negative, even for phyla that do not stain with the Gram method.

Table 2 - Subcellular location prediction performance evaluation between ReVarcine and PSORT.

Metric	Membrane type [a]	ReVarcine	PSORT
Acurácia	monoderm	0.953	0.741
	diderm	0.967	0.698
Precisão	monoderm	0.912	0.326
	diderm	0.935	0.272
Recall	monoderm	0.732	0.486
	diderm	0.790	0.467

a = monoderm comprises the phyla of bacteria classified as Gram-positive, while diderm those classified as Gram-negative, even for phyla that do not stain with the Gram method.

Table 3 - Beta barrel prediction performance evaluation between ReVarcine and PSIPRED.

Metric	Membrane type [a]	ReVarcine	PSIPRED
F1 score	diderm	0.872	0.791

a = diderm those classified as Gram-negative, even for phyla that do not stain with the Gram method.

Table 4 - Alpha helix prediction performance evaluation between ReVarcine and PSIPRED.

Metric	Membrane type [a]	ReVarcine	PSIPRED
F1 score	diderm	0.826	0.819

a = diderm those classified as Gram-negative, even for phyla that do not stain with the Gram method.

Table 5 - Total protein sequences for train and test data for the signal peptide, signal peptide cut position, subcellular location, beta barrel and alpha helix models, respectively.

Membrane type	Predicted structure model	Total protein sequence train data	Total protein sequence test data
Monoderm / Gram-positive	Signal peptide	3483	1161
	Subcellular location	25497	8499
	Signal peptide cut position	3588	1196
Diderm/ Gram-negative	Signal peptide	11072	3690
	Signal peptide cut position	6588	2196
	Subcellular location	20661	6887
	Alpha helix	14203	4734
	Beta barrel	13586	4528

Table 6 - ReVarcine specific set of models for each classification of bacteria.

Membrane type	ReVarcine model name	Predicted structure
Monoderm	spp	Signal peptide
	scp	Subcellular location
	cgp	Signal peptide cut position
Diderm	spgn	Signal peptide
	cgn	Signal peptide cut position
	scn	Subcellular location
	ahn	Alpha helix
	bbn	Beta barrel

11. Figure

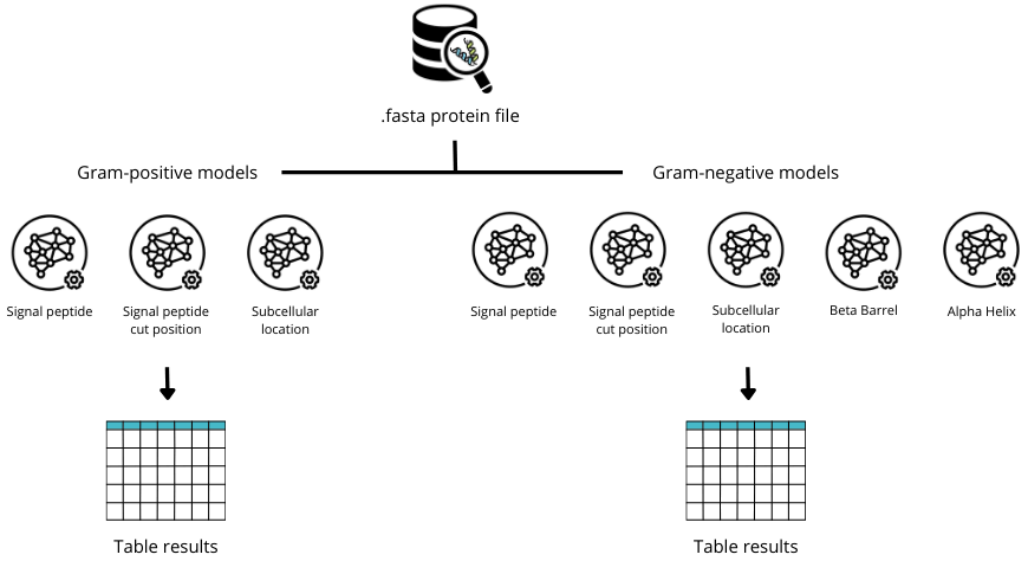


Figure 1. ReVarcine pipeline flowchart.The first step requires input data in .fasta format, containing one or more amino acid sequences. If the input sequences are from Gram-positive bacteria, users can choose among three prediction models: Signal Peptide, Cut Position, and Subcellular Location. For sequences from Gram-negative bacteria, five prediction models are available: Signal Peptide, Cut Position, Subcellular Location, Beta Barrel, and Alpha Helix. The output, regardless of whether the input is from Gram-positive or Gram-negative bacteria, is presented in a tabular format, listing the selected models and their respective results.

6. Conclusões Gerais

No contexto da vacinologia reversa, as proteínas alvos precisam apresentar certas características para serem candidatas ideais a antígenos vacinais. Essas características estão relacionadas a sua localização celular, as proteínas da superfície celular ou secretadas são preferidas, pois são mais acessíveis ao sistema imunológico; proteínas com estruturas tridimensionais como barril beta e alfa hélice bem definidas e domínios conservados podem ser mais imunogênicas e estáveis. No que se refere à localização celular ferramentas como PSORT e SignalP são usadas para prever localização subcelular e peptídeos sinal. Já no cenário de estruturas tridimensionais ferramentas como Alphafold e PSIPRED ajudam na predição estrutural.

Apesar desses avanços, ainda existem lacunas, visto que muitos métodos atuais carecem de integração, dificultando a criação de pipelines abrangentes e automatizados para vacinologia reversa. Além disso, a robustez e a capacidade de generalização de modelos de deep learning, especialmente em cenários com dados limitados, permanecem desafios. Nesse contexto, ReVarcine, uma ferramenta baseada em redes neurais profundas e arquitetura transformer, tem como objetivo automatizar a identificação de peptídeos sinal, localização subcelular, estruturas β-barril e hélices alfa em bactérias, além de priorizar alvos vacinais, contribuindo para avanços significativos na imunoinformática e no desenvolvimento de vacinas de nova geração.

O modelo ReVarcine apresentou desempenho superior em diversas tarefas de predição relacionadas à vacinologia reversa, incluindo predição de peptídeo sinal, localização subcelular e estrutura secundária de proteínas. Para a predição de peptídeos sinal, ReVarcine obteve métricas como MCC, IoU, precisão e a acurácia superiores ao SignalP. Em tarefas de localização subcelular, ReVarcine obteve acurácia muito acima do PSORT. Além disso, em predições de estruturas secundárias, como barril beta e alfa-hélice, o modelo se mostrou superior ao PSIPRED.

Esses resultados superiores refletem uma tendência crescente em imunoinformática: a adoção de deep learning e inteligência artificial para aprimorar a precisão das predições e a integração de pipelines. Ao utilizar arquiteturas de redes neurais, o ReVarcine supera abordagens tradicionais de machine learning, como SignalP, PSORT e PSIPRED, oferecendo melhor generalização e adaptabilidade a diversos conjuntos de dados.

Além disso, o ReVarcine proporciona insights valiosos para a vacinologia reversa, com previsões precisas de localização subcelular e análise de estruturas secundárias, facilitando a identificação de epítopos alvo de forma mais eficiente do que ferramentas isoladas. Sua funcionalidade abrangente reforça seu potencial transformador no design de vacinas, acelerando a transição do genoma para a vacina.

Outra aplicação crucial do ReVarcine é em pipelines de triagem de antígenos pré-clínicos. O modelo acelera a identificação de candidatos viáveis, reduzindo a dependência de análises laboratoriais demoradas, o que é vital para o desenvolvimento rápido de vacinas, especialmente em contextos de pandemias ou surtos de patógenos resistentes. Sua versatilidade e precisão estabelecem um novo padrão em ferramentas computacionais, oferecendo uma solução unificada que otimiza o processo de predição, reduzindo o tempo computacional e os erros provenientes do uso de várias plataformas. Isso destaca seu papel essencial nas demandas crescentes de imunologia computacional e vacinologia.

7. Referências

ABADI, M. et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. **arXiv**, 2016.

ACKLEY, D. H.; HINTON, G. E.; SEJNOWSKI, T. J. A learning algorithm for Boltzmann machines. **Cognitive Science**, v. 9, p. 147–169, 1985.

AFRICA REGIONAL COMMISSION FOR THE CERTIFICATION OF POLIOMYELITIS ERADICATION. Certifying the interruption of wild poliovirus transmission in the WHO African region on the turbulent journey to a polio-free world. *The Lancet Global Health*, v. 8, p. e1345–e1351, 2020.

ALLEN, E. E.; BANFIELD, J. F. Community genomics in microbial ecology and evolution. **Nature Reviews Microbiology**, v. 3, n. 6, p. 489–498, 2005.

ALMAGRO ARMENTEROS, J. J.; TSIRIGOS, K. D.; SØNDERBY, C. K.; et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, [s.l.], v. 37, p. 420–423, 2019.

ASGARI, E.; MOFRAD, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. **PLOS One**, v. 10, n. 11, 2015.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In: Proceedings of the International Conference on Learning Representations (ICLR), 2014.

BAZIN, H. Vaccination: A History. From Lady Montagu to Genetic Engineering. John Libbey Eurotext, Montrouge, 2011.

BENDTSEN, J. D. et al. Improved prediction of signal peptides: SignalP 3.0. **Journal of Molecular Biology**, v. 340, n. 4, p. 783–795, 2004.

BENDTSEN, J. D. et al. Prediction of signal peptides using deep learning. **Journal of Molecular Biology**, v. 428, n. 22, p. 3583–3593, 2016.

BERVEN, F. S. et al. BOMP: a program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. **Nucleic Acids Research**, v. 32, n. suppl_2, p. W394-W399, 2004.

BENDTSEN, J.D.; NIELSEN, H.; VON HEIJNE, G.; BRUNAK, S. Improved prediction of signal peptides: SignalP 3.0. **Journal of Molecular Biology**, v. 340, n. 4, p. 783–795, 2004.

BIACONI, R.M. et al. Genome-Based Approach Delivers Vaccine Candidates Against *Pseudomonas aeruginosa*. **Frontiers in Immunology**, v. 9, p. 3021, 2018.

BIANCONI, I. et al. Genome-based approach delivers vaccine candidates against *Pseudomonas aeruginosa*. **Frontiers in Immunology**, v. 9, p. 3021, 2019.

BOJANOWSKI, P. et al. Enriching word vectors with subword information. **Facebook AI Research**, 2017.

BRADBURY, J. et al. Pytorch: An imperative style, high-performance deep learning library. **Advances in Neural Information Processing Systems**, v. 32, p. 1–12, 2019.

CARVALHO, D.V.; PEREIRA, E.M.; CARDOSO, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. **Artificial Intelligence Review**, v. 51, n. 2, p. 245–275, 2019.

DAHL, G. E. Deep learning approaches to problems in speech recognition, computational chemistry, and natural language text processing. University of Toronto, 2015.

DEEPMIND. **AlphaGo**. Disponível em: https://www.deepmind.com/research/highlighted-research/alphago. Acesso em: 2 nov. 2022.

DENG, L. Three classes of deep learning architectures and their applications: a tutorial survey. **APSIPA Transactions on Signal and Information Processing**, v. 57, p. 58, 2012.

DONATI, C.; RAPPUOLI, R. Reverse vaccinology in the 21st century: improvements over the original design. **Annals of the New York Academy of Sciences**, v. 1285, p. 115–132, 2013.

ELNAGGAR, A. et al. ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and transfer learning. **Bioinformatics**, 2020.

FENNER, F. et al. Smallpox vaccine and vaccination in the intensified smallpox eradication programme. World Health Organization, Geneva, 1988.

FENNER, F. et al. Smallpox and Its Eradication, Volume 6. World Health Organization, 1988.

FLOWER, D. R. et al. Computer-aided selection of candidate vaccine antigens. **BMC Immunome Research**, v. 6, 2010.

FUKUSHIMA, K.; MIYAKE, S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: *Competition and Cooperation in Neural Nets*. Berlin/Heidelberg: Springer, 1982. p. 267–285.

GRAVES, A.; MOHAMED, A.-R.; HINTON, G. Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, Vancouver. Proceedings […]. IEEE, 2013, p. 6645-6649.

GRAVES, J. et al. A review of deep learning methods for antibodies. **Antibodies**, v. 9, n. 2, p. 12, 2020.

GROOT, A. S. Immunomics: discovering new targets for vaccines and therapeutics. **Drug Discovery Today**, v. 11, n. 5-6, p. 203–209, 2006.

HAFFKINE, W.M. Remarks on the plague prophylactic fluid. **British Medical Journal**, v. 1, p. 1461–1462, 1897.

HANDELSMAN, J. Metagenomics: Application of genomics to uncultured microorganisms. **Microbiology and Molecular Biology Reviews**, v. 68, p. 669–685, 2004.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. New York: Springer, 2009.

HOLLAND, M. **Vaccines: A Biography**. Springer, 2010.

HUANG, P.; RASMUSSEN, M. D. SignalP 6.0: A robust, deep-learning-based tool for signal peptide prediction. **Bioinformatics**, v. 39, n. 5, p. 1537-1539, 2023.

JANG, B.; KIM, I.; KIM, J. W. Word2vec convolutional neural networks for classification of news articles and tweets. **PLOS One**, v. 14, n. 8, 2019.

JENNER, E. An Inquiry into the Causes and Effects of the Variolae Vaccinae. Low, London, 1798.

JONES, D. T. PSIPRED: A simple and accurate method for predicting protein secondary structure from amino acid sequences. **Journal of Molecular Biology**, v. 319, n. 3, p. 599-610, 2007.

JORDAN, M. I. Serial order: A parallel distributed processing approach. In: *Advances in Psychology*. Amsterdam: Elsevier, 1997. v. 121, p. 471–495.

KÄLL, L.; KROGH, A.; SONNHAMMER, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. **Nucleic Acids Research**, v. 35, p. W429–W432, 2007.

KOLLE, W. Zur aktiven Immunisierung der Menschen gegen Cholera. **Zentralblatt für Bakteriologie und Parasitenkunde**, v. 19, p. 97–104, 1896.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, p. 2278–2324, 1998.

LEE, J. et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. **Bioinformatics**, 2020.

LEVINE, M.M.; SZTEIN, M.B. Vaccine development strategies for improving immunization: the role of modern immunology. **Nature Immunology**, v. 5, p. 460–464, 2004.

L.M. L. New strategies for vaccine development. **SPCV**, v. 2, p. e4, 2010.

LU, D.R. et al. Identifying functional anti-*Staphylococcus aureus* antibodies by sequencing antibody repertoires of patient plasmablasts. **Clinical Immunology**, v. 152, p. 77–89, 2014.

MAIDEN, M.C.J. et al. The Impact of Nucleotide Sequence Analysis on Meningococcal Vaccine Development and Assessment. **Frontiers in Immunology**, v. 9, p. 3151, 2018.

MASIGNANI, V. et al. Vaccination against *Neisseria meningitidis* using three variants of the lipoprotein GNA1870. **Journal of Experimental Medicine**, v. 197, n. 6, p. 789-799, 2003.

MCINERNEY, M. J. et al. Physiology, ecology, phylogeny, and genomics of microorganisms capable of syntrophic metabolism. **Annals of the New York Academy of Sciences**, v. 1125, n. 1, p. 58–72, 2008.

McCULLOCH, W. S.; PITTS, W. A. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, v. 5, p. 115–133, 1943.

MCGUFFIN, L. J.; BRYSON, K.; JONES, D. T. The PSIPRED protein structure prediction server. **Bioinformatics**, v. 16, n. 4, p. 404-405, 2000.

MIJOLOV, T. et al. Efficient Estimation of Word Representations in Vector Space. In: Proceedings of the International Conference on Learning Representations (ICLR), 2013.

NIELSEN, H. et al. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. **Protein Engineering**, v. 10, n. 1, p. 1-6, 1997.

ORENSTEIN, W.A.; AHMED, R. Simply put: Vaccination saves lives. **Proceedings of the National Academy of Sciences**, v. 114, p. 4031–4033, 2017.

OZAWA, S. et al. Return on investment from childhood immunization in low- and middle-income countries, 2011-2020. **Health Affairs**, v. 35, p. 199–207, 2016.

PANDEY, A.; GALVANI, A.P. The global burden of HIV and prospects for control. **Lancet HIV**, v. 6, p. e809–e811, 2019.

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. **Foundations and Trends in Information Retrieval**, v. 2, n. 1–2, p. 1-135, 2008.

PASTEUR, L. De l'atténuation du virus du choléra des poules. **Comptes Rendus de l'Académie des Sciences**, v. 91, p. 673–680, 1880.

PEAD, P.J. Benjamin Jesty: Dorset's Vaccination Pioneer. Timefile Books, Chichester, 2009.

PETERSEN, T. et al. SignalP 4.0: discriminating signal peptides from transmembrane regions. **Nature Methods**, v. 8, p. 785–786, 2011.

PFEIFFER, R.; KOLLE, W. Experimentelle Untersuchungen zur Frage der Schutzimpfung des Menschen gegen Typhus Abdominalis. **Deutsche Medizinische Wochenschrift**, v. 22, p. 735–737, 1896.

PIZZA, M.; SCARLATO, V.; MASIGNANI, V.; et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science*, [s.l.], v. 287, n. 5459, p. 1816–1820, 2000.

PLOTKIN, S. A.; ORENSTEIN, W. A.; OFFIT, P. A. Vaccines. **Expert Review of Vaccines**, v. 11, n. 10, p. 1225-1243, 2012.

POLAND, G.A.; OVSYANNIKOVA, I.G.; JACOBSON, R.M. Application of pharmacogenomics to vaccines. **Pharmacogenomics**, v. 10, n. 5, p. 837–852, 2009.

POOLMAN, J. T. et al. The history of pneumococcal conjugate vaccine development: dose selection. **Expert Review of Vaccines**, v. 12, n. 12, p. 1379-1394, 2013.

RAHMAN, S. A.; SAIT, M.; SUNDARAM, S. A. LipoP: Predictions of lipoprotein signal peptides in gram-negative bacteria. *Protein Science*, [s.l.], v. 17, n. 3, p. 371–374, 2008.

RAJPURKAR, P. et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, Austin. Proceedings […]. 2016, p. 2383-2392.

RAPPUOLI, R. Reverse vaccinology. **Current Opinion in Microbiology**, v. 3, n. 5, p. 445–450, 2000.

RAPPUOLI, R. Vaccines: science, health, longevity, and wealth. **Proceedings of the National Academy of Sciences**, v. 111, p. 12282, 2014.

REMMERT, M. et al. HHomp - prediction and classification of outer membrane proteins. **Nucleic Acids Research**, v. 37, p. W446, 2009.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. Pearson, 2021.

SANCHEZ-TRINCADO, J.L. et al. Fundamentals and methods for T- and B-cell epitope prediction. **Journal of Immunological Research**, v. 2017, p. 2680160, 2017.

SERRUTO, D. et al. Biotechnology and vaccines: application of functional genomics to *Neisseria meningitidis* and other bacterial pathogens. **Journal of Biotechnology**, v. 113, 2004.

SINGH, R. et al. Recent updates on drug resistance in *Mycobacterium tuberculosis*. **Journal of Applied Microbiology**, v. 128, p. 1547–1567, 2020.

SONNHAMMER, E. L. L. et al. A hidden Markov model for predicting

transmembrane helices in protein sequences. In: *Proceedings of the ISMB*, 1998. p. 175–182.

VAN DER MAATEN, L.; HINTON, G. Visualizing Data using t-SNE. **Journal of Machine Learning Research**, v. 9, p. 2579-2605, 2008.

VIRAL GENOMICS et al. Recent Advances in the Development of SARS-CoV-2 Vaccine Platforms. **Viruses**, v. 15, p. 2130, 2023.

WILLIAMS, A.E. et al. Innate imprinting by the modified heat-labile toxin of *Escherichia coli* (LTK63) provides generic protection against lung infectious disease. **Journal of Immunology**, v. 173, p. 7435–7443, 2004.

WIZEMANN, T. M. et al. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. **Infection and Immunity**, v. 69, n. 3, p. 1593–1598, 2001.

WORLD HEALTH ORGANIZATION (WHO). Ten threats to global health in 2019. Available from: https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019. Accessed February 06, 2024.

WOYKE, T. et al. One Bacterial Cell, One Complete Genome N. Ahmed (ed.). **PLoS ONE**, v. 5, n. 4, p. E10314, 2010.

YANG, K.K.; ZHANG, G.; SAILER, Z.R. Machine learning and protein engineering. **Current Opinion in Structural Biology**, v. 48, p. 168-176, 2018.

YAO, X. et al. Predictive Modelling in Healthcare: Advances in AI Techniques. **Journal of Medical Informatics**, v. 10, n. 4, p. 34–47, 2023.

YU, C. S.; CHEN, Y. C.; LU, C. H. Cello: A tool for predicting cellular localization of proteins. **Computational Biology and Chemistry**, v. 30, n. 1, p. 49-59, 2006.

YU, C.S. et al. Prediction of protein subcellular localization. **Proteins**, v. 64, p. 643–651, 2006.

YU, N.Y. et al. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. **Bioinformatics**, v. 26, n. 13, p. 1608–1615, 2010.

YU, N. Y. et al. PSORTdb—an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. **Nucleic Acids Research**, v. 39, n. suppl_1, p. D241-D244, 2010.

XUE, L. C.; CHEN, Y. C.; LU, M.; YU, C. S. Cello 2.0: A tool for predicting subcellular localization of proteins using sequence-derived features. **Bioinformatics**, v. 37, n. 2, p. 280-287, 2021.

ZHOU, K. et al. Machine Learning in Manufacturing: Current Applications and Future Directions. **International Journal of Production Research**, v. 60, n. 15, p. 4620–4645, 2022.

ZIMMERMANN, N. et al. Human isotype-dependent inhibitory antibody responses against *Mycobacterium tuberculosis*. **EMBO Molecular Medicine**, v. 8, p. 1325–1339, 2016.