UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MATHEUS DO VALE ALMEIDA

APLICAÇÃO DE AUTOCODIFICADOR VARIACIONAL PARA AGRUPAMENTO DE SEQUÊNCIA DE ANTICORPOS

FORTALEZA

2024

MATHEUS DO VALE ALMEIDA

APLICAÇÃO DE AUTOCODIFICADOR VARIACIONAL PARA AGRUPAMENTO DE SEQUÊNCIA DE ANTICORPOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientador: Prof. Dr. César Lincoln Cavalcante Mattos.
Coorientador: Dr. Geraldo Rodrigues Sartori.

FORTALEZA

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

A449a Almeida, Matheus do Vale.
Aplicação de autocoficador variacional para agrupamento de sequência de anticorpos / Matheus do Vale Almeida. – 2024.
36 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Computação, Fortaleza, 2024.
Orientação: Prof. Dr. César Lincoln Cavalcante Mattos.
Coorientação: Prof. Dr. Geraldo Rodrigues Sartori.

1. Aprendizagem de máquina. 2. Anticorpos. 3. Aprendizado profundo. 4. Agrupamento. I. Título.
CDD 005

MATHEUS DO VALE ALMEIDA

APLICAÇÃO DE AUTOCODIFICADOR VARIACIONAL PARA AGRUPAMENTO DE SEQUÊNCIA DE ANTICORPOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em: xx/xx/xxxx.

BANCA EXAMINADORA

Prof. Dr. César Lincoln Cavalcante Mattos (Orientador)
Universidade Federal do Ceará (UFC)

Dr. Geraldo Rodrigues Sartori (Coorientador)
Fundação Oswaldo Cruz – Ceara (FIOCRUZ-CE)

Prof. Me. Diego da Silva de Almeida
Fundação Oswaldo Cruz – Ceara (FIOCRUZ-CE)

Prof. Me. Francisco Flávio de Assunção Rabelo
Universidade Federal do Ceará (UFC)

À minha família, por sua capacidade de acreditar em mim e investir em mim. Mãe, seu cuidado e dedicação foi que deram, em alguns momentos, a esperança para seguir.

AGRADECIMENTOS

Aos meus orientadores, Prof. Dr. César Lincoln Cavalcante e Dr. Geraldo Rodrigues Sartori, pela excelente orientação, apoio constante e pela oportunidade de contribuírem de forma decisiva na construção deste trabalho.

Ao laboratório da Fiocruz e ao Grupo de Pesquisa, por me acolherem e pela oportunidade de participar de um ambiente científico tão estimulante. Em especial, ao Prof. Dr. João Hermínio Martins da Silva, por todo o apoio e incentivo para minha permanência no grupo.

Aos professores membros da banca examinadora, Diego da Silva de Almeida e Francisco Flávio de Assunção Rabelo, pela generosidade de seu tempo e pelas valiosas colaborações e sugestões, que enriqueceram este projeto.

À minha mãe e familiares, meus verdadeiros pilares, por todo o suporte, seja financeiro ou emocional, que me permitiu continuar minha jornada acadêmica.

Aos colegas e amigos de turma, pela troca de ideias, reflexões, críticas e sugestões, que contribuíram para o meu aprendizado ao longo do curso. Em especial, aos amigos Silvio e Nadiana, pela companhia e pelos ótimos momentos vividos durante toda a graduação.

Agradeço à Universidade Federal do Ceará por proporcionar uma educação de excelência e aos professores do curso, cujos ensinamentos e oportunidades foram fundamentais para a realização deste trabalho e para o meu crescimento acadêmico. Por fim, deixo meu agradecimento ao Programa Institucional de Bolsas de Iniciação Científica e Tecnológica (BICT) da Funcap, pelo suporte financeiro essencial ao desenvolvimento deste projeto.

RESUMO

A descoberta de novos tratamentos baseados em anticorpos é uma área de crescente importância na biotecnologia, especialmente em imunoterapias para doenças como câncer e doenças autoimunes. No entanto, a enorme diversidade de estrutura e sequência dos anticorpos apresenta um desafio significativo na identificação de candidatos terapêuticos promissores. Este trabalho tem como objetivo acelerar esse processo por meio de técnicas de aprendizado de máquina, com foco no agrupamento de anticorpos com base em suas características estruturais. Ao utilizar modelos avançados de aprendizado profundo, como Autocodificadores Variacionais, e métodos de agrupamento, propomos uma abordagem que busca identificar padrões latentes e agrupamentos naturais em grandes volumes de dados de anticorpos. Essa estratégia pode facilitar a triagem e priorização de candidatos ao ruinir e agrupar características compartilhadas entre anticorpos. A expectativa é que este trabalho contribua para acelerar a pesquisa e o desenvolvimento de imunoterapias, permitindo uma seleção mais eficiente de candidatos promissores, com base em suas propriedades estruturais.

Palavras-chave: aprendizagem de máquina; anticorpos; aprendizado profundo; agrupamento.

ABSTRACT

The discovery of new antibody-based treatments is an area of growing importance in biotechnology, particularly in immunotherapies for diseases such as cancer and autoimmune disorders. However, the vast diversity in the structure and sequence of antibodies presents a significant challenge in identifying promising therapeutic candidates. This work aims to optimize this process through machine learning techniques, focusing on the clustering of antibodies based on their structural characteristics. By utilizing advanced deep learning models, such as Variational Autoencoders, and clustering methods, we propose an approach that seeks to identify latent patterns and natural clusters within large volumes of antibody data. This strategy can facilitate the screening and prioritization of candidates by gathering and grouping shared characteristics among antibodies. The expectation is that this work will help accelerate research and the development of immunotherapies, enabling a more efficient selection of promising candidates based on their structural properties

Keywords: machine learning; antibody; deep learning; clustering.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANARCI	*Antibody Numbering And Repertoire Classification Interface*
CDR	Regiões Determinantes de Complementariedade
GAE	*Graph Autoencoder*
IA	Inteligência Artificial
LSTM	*Long Short-Term Memory*
ML	*Machine Learning*
OAS	Oligoclonal Antibody Sequences
PCA	Análise de Componentes Principais
PDB	*Protein Data Bank*
UMAP	*Uniform Manifold Approximation and Projection*
VAE	Autocodificadores Variacionais

LISTA DE SÍMBOLOS

μ	Média
σ^2	Variância

SUMÁRIO

1 INTRODUÇÃO

A descoberta de novos tratamentos para doenças continua a ser um dos maiores desafios na área terapêutica. A identificação e o desenvolvimento de terapias eficazes exigem a análise de dados biológicos complexos. No caso de anticorpos, a análise se torna ainda mais desafiadora devido à diversidade de estrutura e funcional dessas biomoléculas e sua interação altamente específica com antígenos, o que demanda técnicas avançadas para entender suas variações e potenciais terapêuticos.

Os anticorpos, com sua capacidade única de reconhecer e se ligar de forma específica a alvos que apresentam epítopos em antígenos, desempenham um papel crucial na resposta imunológica. Esses alvos são identificados pelos anticorpos de maneira altamente seletiva, permitindo que o sistema imunológico neutralize ou elimine as ameaças de forma precisa. Por conta dessa especificidade, os anticorpos possuem um potencial significativo como agentes terapêuticos e são amplamente utilizados em imunoterapias para uma variedade de doenças, como o câncer e doenças autoimunes (JANEWAY *et al.*, 2001). No entanto, a complexidade estrutural e a diversidade das sequências de anticorpos tornam a tarefa de identificar candidatos promissores um grande desafio (FERNANDEZ-QUINTERO *et al.*, 2023)

Nesse contexto, o agrupamento de anticorpos emerge como uma abordagem poderosa para acelerar o processo de descoberta de novos fármacos. O agrupamento, uma técnica que agrupa dados com base em similaridades, vai além da simples redução de redundância, ele permite a identificação de padrões complexos dentro de grandes conjuntos de dados biológicos (HUANG *et al.*, 2022). Essa análise aprofundada facilita a compreensão das interações entre diferentes anticorpos e antígenos, levando a *insights* sobre a funcionalidade e a especificidade dos anticorpos. Ao identificar grupos de anticorpos com propriedades desejáveis, o agrupamento não apenas acelera a seleção para o desenvolvimento terapêutico, mas também potencializa a exploração de novas estratégias de *design* e criação de fármacos, contribuindo significativamente para a eficiência da pesquisa em biomedicina (LI *et al.*, 2001).

A aplicação de técnicas de *Machine Learning* (ML) tem demonstrado um enorme potencial na análise de anticorpos, especialmente com o uso de algoritmos baseados em aprendizado profundo, que são capazes de lidar com grandes volumes de dados e descobrir padrões complexos que escapam aos métodos tradicionais. Modelos como os Autocodificadores Variacionais (VAE) podem ser utilizados para aprender representações latentes de dados biológicos, facilitando a identificação de padrões ocultos e grupos naturais em sequências de anticorpos.

Embora os VAE tenham sido amplamente estudados em diversas áreas, como a geração de imagens e processamento de linguagem natural (KINGMA; WELLING, 2013), sua aplicação à análise de anticorpos ainda está em fase de desenvolvimento. No entanto, essa abordagem oferece o potencial de proporcionar avanços valiosos para o desenvolvimento de novas terapias e otimização de tratamentos existentes.

Neste trabalho, exploraremos a aplicação de VAE para agrupamento de anticorpos, com o objetivo de otimizar a descoberta de novos tratamentos. Utilizando dados de sequência das Regiões Determinantes de Complementariedade (CDR) e a região do paratopo, buscamos identificar agrupamentos que possam revelar novas oportunidades para o desenvolvimento de terapias eficazes. Ao identificar padrões e agrupamentos naturais dentro das sequências, pretendemos facilitar a triagem de anticorpos com propriedades desejáveis, potencializando a eficácia no desenvolvimento de imunoterapias. Essa abordagem não só contribui para um entendimento mais profundo das interações entre anticorpos e antígenos, mas também oferece um caminho para a descoberta de novos candidatos terapêuticos que podem ser mais explorados em pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo estabelece a base teórica essencial para a compreensão deste trabalho. Inicialmente, são discutidos os princípios fundamentais da estrutura dos anticorpos, incluindo suas características estruturais e funcionais. Por último, o capítulo foca em técnicas e ferramentas de aprendizado de máquina. Em particular, são discutidos os modelos de *autoencoders*.

2.1 Anticorpos

Os anticorpos são proteínas produzidas pelo sistema imunológico em resposta a antígenos, que são substâncias estranhas ao organismo (JANEWAY *et al.*, 2001).Um patógeno é qualquer organismo ou agente capaz de causar doenças, e, ao invadir o corpo, seus componentes são identificados como antígenos, desencadeando a produção de anticorpos. Dessa forma, os anticorpos desempenham um papel fundamental na defesa do organismo, reconhecendo e se ligando a esses antígenos, neutralizando-os ou marcando-os para destruição por outras células do sistema imunológico.

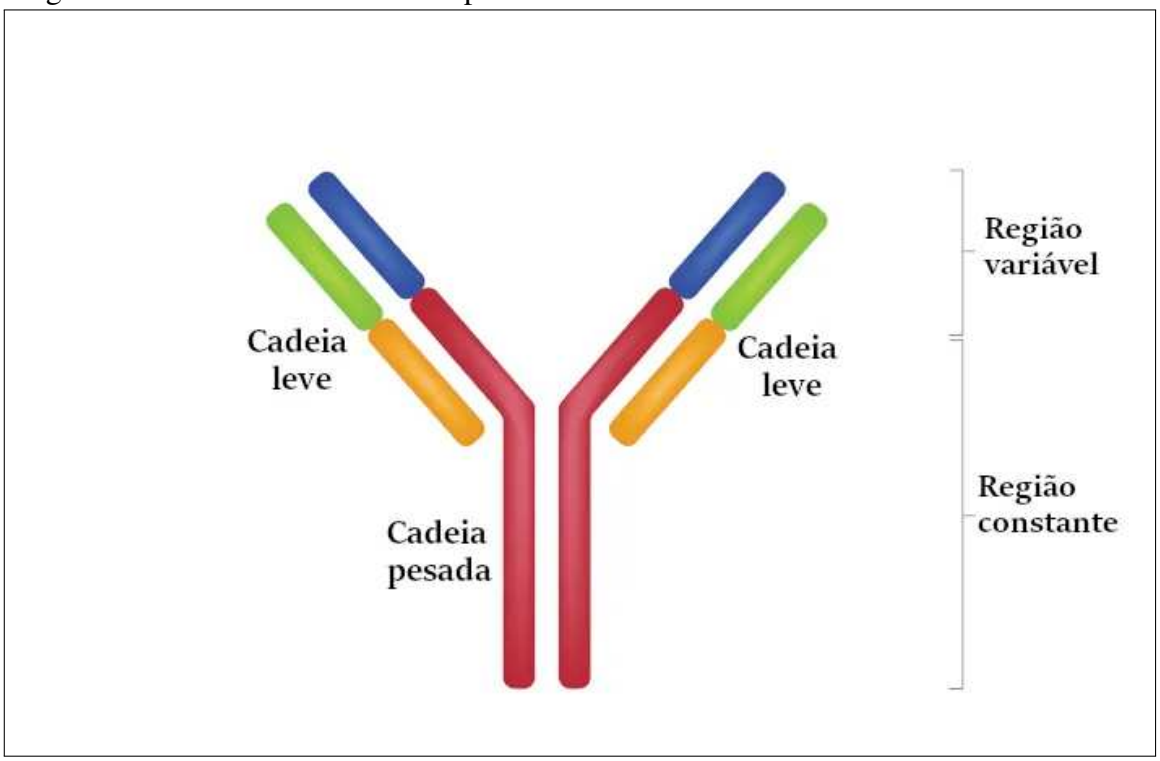
Estruturalmente, os anticorpos são compostos por duas cadeias pesadas e duas cadeias leves, cada uma contendo um domínio variável e um domínio constante. O domínio variável é o responsável pela especificidade de ligação ao antígeno, enquanto o domínio constante determina a classe do anticorpo, influenciando sua função na resposta imune, como neutralização, opsonização e ativação do sistema complemento. A imagem 1 mostra a estrutura convencional de um anticorpo.

Dentro do domínio variável, encontram-se as regiões hipervariáveis, também conhecidas como CDR, que são as principais responsáveis pela alta especificidade de reconhecimento nos anticorpos. Essas regiões são altamente variáveis em sequência e estrutura, permitindo que os anticorpos se liguem a uma vasta gama de epítopos, regiões específicas dos antígenos com as quais interagem (JANEWAY *et al.*, 2001). A diversidade das CDRs é um dos fatores chave que permitem ao sistema imunológico reconhecer e neutralizar patógenos com grande eficiência.

O local específico no anticorpo que se liga ao antígeno é conhecido como paratopo. Essa região é uma estrutura complementar ao epítopo, e sua interação é essencial para a eficácia da resposta imune. Essa interação é altamente específica, permitindo que cada anticorpo se ligue a um antígeno particular, o que é crucial na defesa contra infecções e doenças.

A análise dessa diversidade é central para a pesquisa biotecnológica moderna, sendo

Figura 1 – Estrutura de um anticorpo.



Fonte: (SANTOS, s.d.)

possível estudar essas sequências detalhadamente com o auxílio de técnicas computacionais e bancos de dados estruturais, como aqueles que utilizam os formatos *Protein Data Bank* (PDB) e FASTA, oferecendo novas oportunidades para o desenvolvimento de terapias inovadoras.

2.2 Formatos PDB e FASTA

2.2.1 Formato PDB

O PDB é um repositório global onde são depositadas informações sobre estruturas tridimensionais experimentais de proteínas e outras macromoléculas biológicas. Esse repositório é essencial para a pesquisa científica, pois fornece um formato de arquivo padronizado, com a extensão *.pdb*, amplamente utilizado para armazenar e compartilhar essas informações cruciais. O formato PDB oferece uma maneira organizada e acessível de representar as estruturas biológicas, permitindo que pesquisadores ao redor do mundo colaborem e compartilhem dados. Cada arquivo PDB contém coordenadas atômicas que representam a posição dos átomos de uma ou mais moléculas, como também inclui informações adicionais sobre o arranjo desses átomos e suas ligações químicas (BANK, s.d.).

Os arquivos PDB são organizados em linhas com campos específicos, cada uma

representando diferentes aspectos da estrutura. A figura 2 mostra um trecho do arquivo. As linhas geralmente começam com um prefixo que indica o tipo de informação, como ATOM para coordenadas atômicas, que facilita a leitura e interpretação dos dados.

Figura 2 – Trecho de um arquivo no formato PDB

```
ATOM   7143  O   ARG H  18      55.167 120.391  92.780  1.00 89.93           O
ATOM   7144  CB  ARG H  18      52.047 120.813  91.877  1.00 83.62           C
ATOM   7145  CG  ARG H  18      50.545 120.947  92.057  1.00 84.59           C
ATOM   7146  CD  ARG H  18      49.880 119.599  92.292  1.00 84.93           C
ATOM   7147  NE  ARG H  18      48.505 119.574  91.806  1.00 88.40           N
ATOM   7148  CZ  ARG H  18      48.160 119.442  90.532  1.00 88.26           C
ATOM   7149  NH1 ARG H  18      49.069 119.325  89.577  1.00 86.26           N
ATOM   7150  NH2 ARG H  18      46.870 119.428  90.208  1.00 89.34           N
ATOM   7151  N   LYS H  19      53.948 118.494  92.866  1.00 91.97           N
ATOM   7152  CA  LYS H  19      55.072 117.580  92.683  1.00 94.47           C
ATOM   7153  C   LYS H  19      54.883 116.874  91.344  1.00 92.11           C
ATOM   7154  O   LYS H  19      54.188 115.857  91.260  1.00 91.72           O
ATOM   7155  CB  LYS H  19      55.153 116.592  93.846  1.00 93.40           C
ATOM   7156  CG  LYS H  19      56.081 115.395  93.632  1.00 87.75           C
ATOM   7157  CD  LYS H  19      57.494 115.681  94.119  1.00 87.54           C
ATOM   7158  CE  LYS H  19      58.185 114.411  94.585  1.00 91.47           C
ATOM   7159  NZ  LYS H  19      58.043 113.308  93.597  1.00 90.99           N
ATOM   7160  N   LEU H  20      55.502 117.416  90.299  1.00 85.17           N
ATOM   7161  CA  LEU H  20      55.452 116.795  88.984  1.00 87.08           C
```

Fonte: elaborada pelo autor.
Nota: As linhas começam com "ATOM", indicando dados sobre átomos específicos, com detalhes como número do átomo, nome do átomo e do resíduo, cadeia, e coordenadas (x, y, z).

2.2.2 Formato FASTA

O formato FASTA é um tipo de arquivo amplamente utilizado para armazenar sequências de nucleotídeos ou aminoácidos, facilitando o intercâmbio de dados biológicos entre diferentes plataformas e ferramentas de análise. Enquanto o formato PDB é essencial para a representação tridimensional das macromoléculas, o formato FASTA foca na descrição linear das sequências, sendo fundamental em bioinformática e em bancos de dados de sequências.

Cada entrada em um arquivo FASTA é composta por um cabeçalho e uma ou mais linhas de sequência. Essa estrutura simples permite que múltiplas sequências sejam armazenadas no mesmo arquivo, cada uma iniciando com seu respectivo cabeçalho. A terminação de cada sequência é geralmente definida pela quebra de linha, ou seja, cada sequência pode ocupar uma ou mais linhas, e a sequência continua até o próximo cabeçalho ou até o final do arquivo. A figura mostra um exemplo de um arquivo no formato FASTA.

Figura 3 – Trecho de um arquivo no formato FASTA

```
>00001bd259ac44ec3b445fe0ae91bf14_H
EVQLVQSGAEVKKPGESLRISCKGSGYSFTNYWISWVRQMPGKGLEWMGRINPSDSHTNYSPSFQGHVTISADKSISTAYLQWSSLKASDTAMY
YCSRPHYYGSGADYWGQGTLVTVSS

>00001bd259ac44ec3b445fe0ae91bf14_L
EIVMTQSPATLSVSPGERATLSCRASQSVGSNLAWYQQKPGQAPSLLIYGASTRATGFPGRFSGSGSGTEFTLTISSLQSEDSAVYYCLQYNNW
YTFGQGTKLEIK

>00002e606050cc0351b3c831c67c6cef_H
EVQLLESGGGLVQPGGSLRLSCAASGFTFNSYAMSWVRQAPGKGLEWVATISGSGGTIYYADSVKGRFTISRDSSKNTLYLQMNSLRAEDTAVY
YCAKDGSPREWLVNEFWGQGTLVTVSS

>00002e606050cc0351b3c831c67c6cef_L
EIVLTQSPGTLSLSPGERATLSCRASQSVSSSYLAWYQQKPGQAPRLLIYGASSRATGIPDRFSGSGSGTDFTLTISRLEPEDFAVYYCQQYGS
APRTFGQGTKVEIK

>00005f40fa39f872f0a4760df98a371f_H
KVQLVESGGGLVQPGGSLRLSCAASGFTFSNYAMTWVRQAPGKGLEWVSSISGSGQSTYYADSVKGRFTISRDNSKNTLYLQMSSLRAEDTAIY
YCASAYSIGWYYFDYWGQGTLVTVSS

>00005f40fa39f872f0a4760df98a371f_L
EIVLAQSPGTLSLSPGERATLSCRASQSVSSSYLAWYQQKPGQAPRLLIYGASSRTTGIPDRFSGSGSGTDFTLTVSRLEPEDFAVYYCQQYGS
SPVTFGQGTKLEIK
```

Fonte: elaborada pelo autor.
Nota: O cabeçalho começa com um caractere de maior (>), seguido por um identificador e, opcionalmente, uma descrição adicional da sequência.

2.3 Ferramentas para análise de anticorpos

2.3.1 ANARCI

O *Antibody Numbering And Repertoire Classification Interface* (ANARCI) é uma ferramenta amplamente utilizada para a numeração e classificação de anticorpos. Ele atribui números aos resíduos em anticorpos com base na sua posição em estruturas de referência, permitindo a comparação e análise de diferentes anticorpos (DUNBAR; DEANE, 2016). Esta ferramenta facilita a interpretação de sequências de anticorpos e é fundamental para a análise estrutural e funcional.

Existem diversos tipos de numerações estabelecidos na literatura, como a numeração de Chothia (CHOTHIA; LESK, 1992) e a numeração de Kabat (KABAT *et al.*, 1991), cada uma com suas próprias convenções e utilidades. A numeração correta é crucial, pois desempenha um papel importante na identificação das regiões hipervariáveis. Ao numerar esses resíduos, os pesquisadores podem identificar rapidamente as CDRs e analisar suas variações entre diferentes anticorpos, contribuindo para a compreensão da diversidade e funcionalidade dos anticorpos.

2.3.2 Parapred

O Parapred é um modelo de aprendizado profundo que prediz as regiões de parátopos de anticorpos utilizando uma combinação de redes neurais. Ele processa sequências de aminoácidos das CDRs, representando cada resíduo com codificações *one-hot* e características adicionais. O modelo utiliza redes neurais convolucionais (CNN, do inglês *Convolutional Neural Network*) para capturar padrões locais, redes neurais recorrentes (RNN, do inglês *Recurrent Neural Network*) para aprender dependências de longo prazo, e perceptrons multicamadas (MLP, do inglês *Multilayer Perceptron*) para a classificação final, identificando quais resíduos são ligantes ou não ao antígeno (LIBERIS *et al.*, 2018).

2.4 Aprendizagem de Máquina Profundo

Aprendizado de máquina profundo (DL, do ingles Deep Learning) é uma subárea da Inteligência Artificial (IA) que ganhou destaque nas últimas décadas devido aos avanços em *hardware* e à crescente disponibilidade de grandes volumes de dados. Nos últimos anos, a área de biotecnologia vivenciou uma explosão no uso de DL, especialmente em aplicações como a descoberta de fármacos, a genômica e a medicina personalizada. Essa evolução foi impulsionada pelo aumento da capacidade computacional e pela disponibilidade de conjuntos de dados massivos, permitindo modelos mais complexos e precisos. Estudos recentes demonstraram que algoritmos de DL podem identificar padrões em dados biológicos que seriam quase impossíveis de detectar com abordagens tradicionais, revolucionando assim a pesquisa biomédica (JUMPER *et al.*, 2021).

Enquanto o ML tradicional utiliza algoritmos que aprendem a partir de dados estruturados, muitas vezes exigindo a definição manual de características, o DL se destaca por permitir o aprendizado automático a partir de grandes volumes de dados (IBM, 2024). Embora frequentemente associado a dados não estruturados, como imagens, áudio e texto, o DL também é amplamente utilizado em dados estruturados. Um exemplo claro disso são as redes neurais em grafos, que foram projetadas para lidar com dados representados por grafos, permitindo capturar relações complexas entre nós e suas conexões. Essas redes são amplamente aplicadas em áreas como análise de redes sociais, predição de propriedades moleculares e análise de interações entre proteínas (ZHOU *et al.*, 2021). Além disso, outros modelos de DL, como RNN e *Transformers*, também processam dados estruturados, como séries temporais ou sequências de DNA. Portanto,

a versatilidade do DL permite que ele seja aplicado tanto a dados não estruturados quanto estruturados, oferecendo soluções eficientes para uma variedade de problemas complexos.

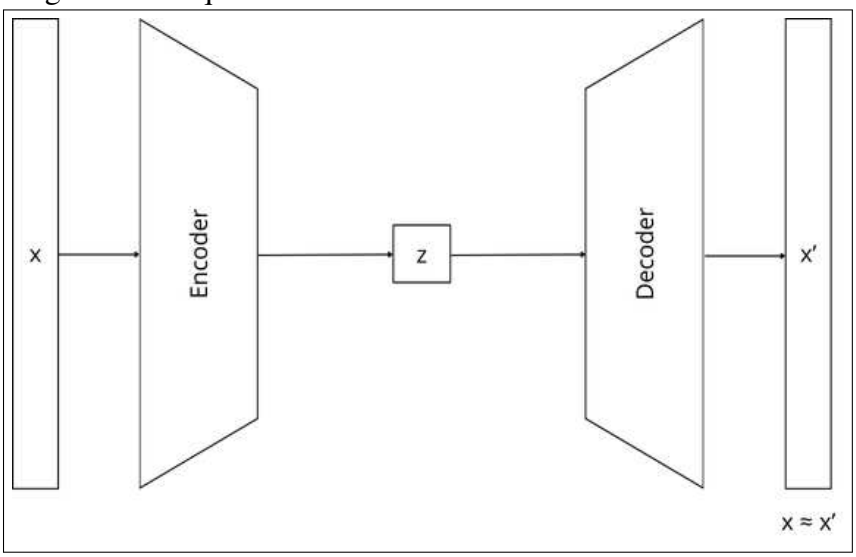
Essas redes são compostas por múltiplas camadas de neurônios artificiais, e, à medida que a rede se torna mais complexa, adicionando mais camadas, sua capacidade de entender e processar informações se torna mais avançada. Isso permite que a rede resolva problemas mais desafiadores, como reconhecer a fala e classificar imagens. Embora o DL tenha revolucionado diversas indústrias, como a de saúde, ele ainda apresenta desafios, como a necessidade de grandes quantidades de dados para o treinameto e alto poder computacional (SHARIFANI; AMINI, 2023).

2.5 Autocodificador Variacional

O VAE é uma abordagem popular dentro da área de modelos gráficos probabilísticos, especificamente quando há variáveis latentes contínuas, variáveis que não podem ser observadas diretamente, mas que influenciam as variáveis que observamos. O objetivo do VAE é estimar uma função de limite inferior, utilizando uma função objetivo estocástica para otimização, frequentemente baseada em maximização de verossimilhança ou máximo *a posteriori* para os parâmetros globais. O princípio fundamental do VAE é aprender uma aproximação variacional para a distribuição posterior, facilitando a inferência em grandes conjuntos de dados, e permitindo o uso eficiente de variáveis latentes para a geração de novos exemplos, o que é útil para tarefas como geração de imagens e amostragem (KINGMA; WELLING, 2013).

No VAE, os dados passam primeiro pelo *encoder*, que transforma as entradas originais em uma distribuição probabilística no espaço latente. Essa distribuição geralmente é assumida como gaussiana, caracterizada por uma média (μ) e variância (σ^2). Após essa transformação, amostras são extraídas dessa distribuição latente, que são então passadas para o *decoder*. O *decoder* realiza o processo inverso: a partir das amostras do espaço latente, ele tenta reconstruir os dados originais. A Figura 4 fornece uma ilustração da arquitetura do modelo. Esse processo permite ao VAE aprender representações latentes compactas que preservam as características mais relevantes dos dados, o que é particularmente útil em tarefas de geração e reconstrução (DOERSCH, 2021). A eficiência desse modelo é medida pela capacidade do *decoder* em reproduzir os dados de entrada com alta precisão, enquanto as amostras latentes continuam a ser representações probabilísticas aproximadas dos dados reais.

Figura 4 – Arquitetura de um Autoencoder Variacional.



Fonte: elaborada pelo autor.
Nota: Aqui, *x* representa a entrada original na qual o modelo codifica *x* em uma representação latente *z* e, em seguida, decodifica essa representação para gerar uma nova amostra *x'*. A amostra *x'* busca se aproximar da entrada original x.

2.6 *K-Means*

O algoritmo *K-Means* é amplamente utilizado para realizar tarefas de agrupamento em várias áreas da ciência de dados. No contexto de modelos de aprendizado não supervisionado, como o VAE, o *K-Means* pode ser aplicado diretamente no espaço latente aprendido. O espaço latente de um VAE representa uma versão comprimida dos dados originais, onde as principais características dos dados são capturadas em uma representação de menor dimensionalidade. Ao aplicar o *K-Means* nesse espaço, é possível identificar agrupamentos naturais presentes nos dados, o que pode revelar padrões subjacentes que não seriam imediatamente visíveis nos dados originais de alta dimensionalidade.

O *K-Means* particiona o espaço latente em k grupos, minimizando a variância intracluster e maximizando a separação entre clusters. Cada ponto no espaço latente é associado ao centroide mais próximo, e os centróides são recalculados iterativamente até que a convergência seja alcançada. Esta técnica é particularmente útil para o estudo de características de dados de natureza complexa, como imagens, sequências biológicas e interações moleculares, fornecendo *insights* sobre as relações entre amostras no espaço latente aprendido (IKOTUN *et al.*, 2023).

3 TRABALHOS RELACIONADOS

Este capítulo oferece uma visão geral dos trabalhos relacionados a agrupamento de sequência de anticorpos. O objetivo é apresentar técnicas utilizadas para a análise e agrupamento dessas moléculas. Ao examinar esses trabalhos similares, buscamos contextualizar a importância do agrupamento no avanço da imunologia e no desenvolvimento de terapias baseadas em anticorpos.

3.1 *Many-against-Many Searching*

O *Many-against-Many searching* (*MMseqs2*) é uma suíte de *software* de código aberto projetada para a busca e agrupamento de grandes conjuntos de sequências (CHOMICZ *et al.*, 2024). Com suporte para sistemas operacionais como Linux, Mac OS e Windows, o *MMseqs2* é altamente escalável, permitindo sua execução em múltiplos núcleos e servidores.

A arquitetura do *MMseqs2* consiste em dois módulos principais: pré-filtragem e alinhamento. O módulo de pré-filtragem realiza um cálculo rápido e sensível de similaridades entre as sequências de um banco de dados de consulta e um banco de dados alvo, utilizando uma técnica de correspondência de *k-mers* consecutivos. Ambos os módulos são paralelizados, permitindo que o software utilize eficientemente todos os núcleos disponíveis, o que aumenta significativamente a eficiência de tempo e processamento (STEINEGGER; SöDING, 2017).

Além de buscas rápidas e precisas, o *MMseqs2* é especialmente eficiente em tarefas de agrupamento. Ele agrupa sequências semelhantes em clusters com base em um grafo de similaridade gerado pelos módulos de pré-filtragem ou alinhamento. Um dos diferenciais do *MMseqs2* é a capacidade de atualizar clusters com novas sequências, mantendo identificadores de cluster estáveis, o que evita a necessidade de reprocessar o conjunto de dados completo. Essa funcionalidade é fundamental para a manutenção de bancos de dados dinâmicos e em constante atualização.

Ao comparar a abordagem do *MMseqs2* com a utilizada em nosso trabalho, que emprega um VAE para explorar um espaço latente considerando também a região paratopo dos anticorpos, notamos diferenças. Enquanto o *MMseqs2* se concentra na similaridade de sequencias, nossa metodologia busca capturar informações adicionais que podem influenciar a formação de clusters, levando a uma representação mais rica e informativa.

Devido à sua combinação de sensibilidade, velocidade e escalabilidade, o *MMseqs2*

é amplamente utilizado em projetos que envolvem a análise de grandes volumes de dados biológicos, como estudos de metagenômica, nos quais se busca atribuir clados taxonômicos e clusters funcionais a dados que outros *softwares* poderiam não ser capazes de mapear. Assim, o *MMseqs2* se destaca como uma ferramenta poderosa e eficiente para a análise de sequências, superando várias limitações de métodos mais tradicionais (KEMPEN *et al.*, 2024).

3.2 Predição de Similaridade de Paratopos de Anticorpos

No contexto de predição de parátopos e agrupamento de anticorpos, (GHANBARPOUR; AL., 2023) apresentam uma abordagem computacional que visa predizer os resíduos do parátopo e gerar representações para medir similaridades entre anticorpos. Os autores destacam a importância de ferramentas *in silico* para o processo de descoberta de novos anticorpos, especialmente para superar os desafios impostos por alvos complexos. Essa abordagem é relevante para a aceleração do processo de descoberta, reduzindo custos e aumentando a diversidade de epítopos na seleção de anticorpos. Além disso, os autores apontam que métodos tradicionais, como a análise baseada unicamente na similaridade de sequência, não são suficientes para agrupar anticorpos com a mesma especificidade de epítopo, devido à variabilidade estrutural nas regiões de ligação.

Este trabalho se alinha a esse objetivo de agrupamento computacional ao focar na utilização de métodos de aprendizado de máquina para agrupamento de anticorpos com base em características de sequência. Ambos os trabalhos compartilham a premissa de que a similaridade de parátopo pode ser um indicador mais preciso de agrupamento funcional do que a similaridade de sequência.

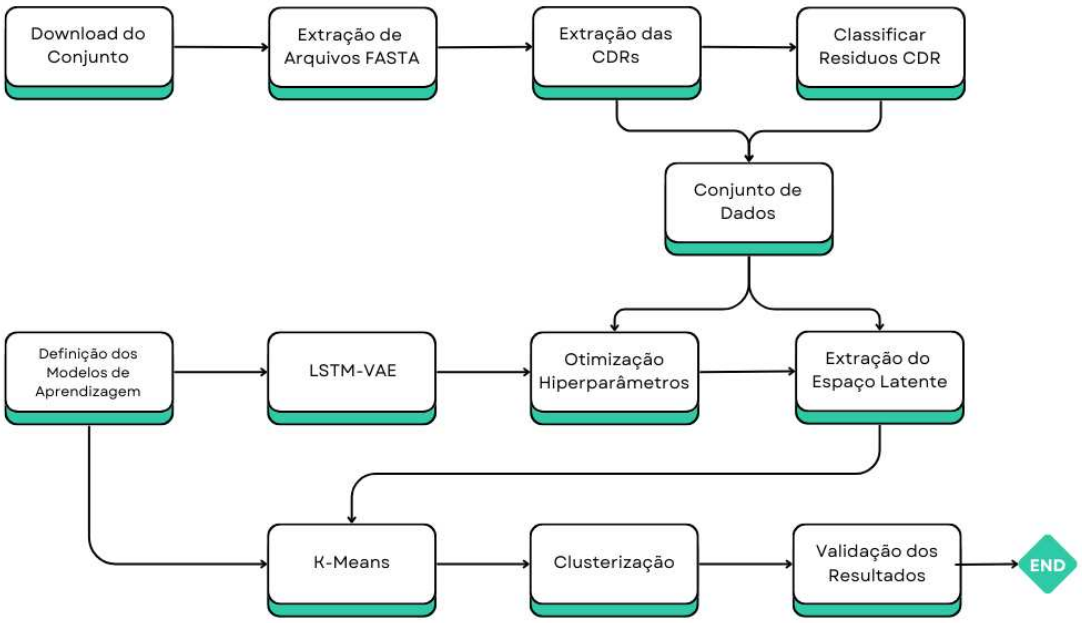
No entanto, enquanto o trabalho relacionado se concentra em desenvolver um modelo preditivo para agrupar anticorpos sem conhecimento prévio do antígeno, este trabalho explora a combinação de técnicas de agrupamento e validação através de comparações experimentais com o *MMseqs2*, que utiliza alinhamento de sequências para validar a formação dos clusters.

Em resumo, o trabalho apresentado por (GHANBARPOUR; AL., 2023) e esta pesquisa compartilham a meta de melhorar a classificação e descoberta de anticorpos utilizando abordagens computacionais inovadoras, porém divergem nos métodos e na ênfase das análises.

4 METODOLOGIA

Este capítulo descreve as etapas seguidas para a seleção e o pré-processamento dos dados, além da definição, treinamento, otimização de hiperparâmetros e validação do modelo utilizado neste trabalho. A Figura 5 apresenta uma sumarização em diagrama desse capítulo.

Figura 5 – Diagrama metodológico.



Fonte: elaborada pelo autor.

4.1 Seleção do Conjunto de Dados

Para o conjunto de dados, foram obtidas aproximadamente 1,3 milhões de estruturas de anticorpos modelados, provenientes do Oligoclonal Antibody Sequences (OAS) a partir de um repositório disponível online no GitHub Graylab (2023). Esses anticorpos foram gerados a partir de amostras de sangue de quatro doadores não relacionados, e, para garantir a qualidade dos dados, foi realizada uma filtragem rigorosa das sequências (JAFFE *et al.*, 2022).

O banco de dados é composto por arquivos no formato PDB e FASTA, com as estruturas de anticorpos sendo modeladas por técnicas de predição baseadas em aprendizado de máquina. Essa abordagem de modelagem permite a construção de representações tridimensionais

de anticorpos a partir de suas sequências de aminoácidos, proporcionando informações estruturais valiosas. Todavia, devido à simplicidade dos arquivos FASTA, foi decidido utilizá-los nas etapas subsequentes do projeto, já que oferecem uma maneira mais eficiente de trabalhar com sequências de aminoácidos.

Inicialmente, a escolha desta base de dados se deu pela sua quantidade significativa de dados disponíveis, o que contribui para a generalização dos modelos de aprendizado de máquina aplicados nas etapas subsequentes deste trabalho.

No entanto, por conta de limitações de *hardware* enfrentadas nas etapas finais do projeto, o conjunto de dados precisou ser reduzido para 15.000 amostras. A redução foi realizada de maneira aleatória, que, apesar de não ser a abordagem mais ideal para preservar a integridade dos dados, tornou-se necessária devido ao tempo limitado disponível para a conclusão do projeto. Dessa forma, 80% dos dados foram utilizados para o treino e teste dos modelos, enquanto os 20% restantes foram reservados para o conjunto de validação, assegurando a validação adequada durante o aprendizado.

4.2 Pré-processamento dos dados

Para o processamento dos dados, utilizamos uma série de *scripts* em Python que foram responsáveis por preparar os dados para o treinamento dos modelos. O primeiro passo envolveu a extração dos arquivos FASTA do conjunto de dados.

Em seguida, foi realizada a extração das CDRs das cadeias leves e pesadas utilizando a ferramenta Anarci. A numeração de Chothia foi escolhida para este estudo especificamente porque ela é compatível com a ferramenta Parapred, que utiliza essa numeração para prever paratopo, com a inclusão de dois resíduos adicionais em cada extremidade das CDRs (LIBERIS *et al.*, 2018). A Figura 6 mostra uma visualização destacando no anticorpo a região que foi extraída para a criação do conjunto.

No processo, utilizamos um limiar de 67% de probabilidade, como definido pelo Parapred, para definir quais resíduos são classificados como parte do paratopo. Isso significa que qualquer resíduo que tenha uma probabilidade maior ou igual a 67% de pertencer ao paratopo foi marcado de forma binária como pertencente a essa região. Essa abordagem permite uma definição mais precisa das áreas de interação relevantes para o reconhecimento de antígenos, mantendo um equilíbrio entre a sensibilidade do modelo e a precisão da seleção dos resíduos.

A Tabela 1 ilustra o formato dos dados contidos nesse arquivo, mostrando as sequên-

Figura 6 – Visualização de anticorpo, destacado em laranja as regiões CDRs.

Fonte: elaborada pelo autor.

cias de resíduos das CDRs juntamente com as informações binárias relacionadas à região paratopo.

Tabela 1 – Conjunto de dados

cdrh1 ph1	cdrh2 ph2	cdrh3 ph3
GSGYSFTNYWISW 0000001111001	GSGYSFTNYWISW 1010110100	SRPHYYGSGADYWG 00111111000000
GSGYSFTNYWISW 0000000111001	TISGSGGTIY 1011111101	AKDGSPREWLVNEFWG 0000111111000000
GSGYSFTNYWISW 0000000111001	SISGSGQSTY 0011111101	ASAYSIGWYYFDYWG 000110111100000

Fonte: elaborada pelo autor.
Nota: ph1, ph2 e ph3 representam as probabilidades de cada resíduo nas respectivas CDRs (cdrh1, cdrh2, cdrh3) em pertencer à região paratopo.

Devido à natureza das sequências de resíduos, não foi possível inserir os dados diretamente no modelo. Para contornar, utilizou-se a codificação *One-Hot* para representar os vinte tipos de resíduos possíveis. Essa codificação mapeia cada resíduo para um vetor, no qual apenas um elemento assume o valor de um, enquanto os demais elementos permanecem com o valor zero. Além disso, foi utilizado o *padding* para garantir que todas as CDRs tivessem o mesmo comprimento. O *padding* consiste em adicionar um caractere especial no final das sequências menores, preenchendo-as até que atinjam o tamanho das maiores. Isso assegura

que todas as sequências possam ser processadas de maneira uniforme pelo modelo, sem alterar o significado biológico dos dados, uma vez que o caractere de *padding* é ignorado durante o treinamento.

4.3 Definição e treinamento do modelo

Inicialmente, pretendíamos utilizar um modelo baseado em *Graph Autoencoder* (GAE) devido à sua capacidade de lidar com dados estruturados em grafos. No entanto, a arquitetura do GAE não permitia a realização de testes com múltiplos grafos,pois não foi projetada para aceitar esse tipo de entrada.Em função dessa limitação, optamos por utilizar um VAE baseado em *Long Short-Term Memory* (LSTM) (LSTMVAE), sendo realizada uma conversão para dados tabulares. No qual, combina redes neurais recorrentes com autocodificadores variacionais para capturar a dinâmica temporal, em nosso caso, as sequências de resíduos dos anticorpos, e a estrutura latente dos dados, oferecendo uma maneira mais eficaz de trabalhar com as características sequenciais presentes no conjunto de dados.

4.4 Hiperparâmetros e Treinamento

Para otimizar o desempenho do LSTMVAE, foram definidos e ajustados os seguintes hiperparâmetros:

- **Taxa de Aprendizado (*Learning Rate*)**: Controla a magnitude das atualizações dos pesos do modelo durante o treinamento.
- **Tamanho do Lote (*Batch Size*)**: Número de amostras processadas antes de atualizar os pesos.
- **Tamanho da Camada Oculta (*Hidden Size*)**: Número de neurônios na camada oculta da LSTM.
- **Tamanho Latente (*Latent Size*)**: Dimensionalidade do espaço latente.

4.4.1 Otimização de Hiperparâmetros

A otimização dos hiperparâmetros foi realizada utilizando o *Optuna*, uma biblioteca de otimização de hiperparâmetros. O processo incluiu as seguintes etapas:

- **Definição do Objetivo**: A função objetivo foi definida para minimizar a perda de validação do modelo. A perda de validação (*eval_loss*) é calculada após cada época de treinamento e

é utilizada para avaliar o desempenho do modelo em dados não vistos.

– **Execução dos Ensaios**: Vários ensaios foram executados, com diferentes combinações de hiperparâmetros. Cada ensaio treinou o modelo com a combinação de hiperparâmetros especificada e avaliou o desempenho. A tabela 2 mostra a distribuição dos hiperparâmetros.
– **Seleção do Melhor Modelo**: O *Optuna* selecionou a combinação de hiperparâmetros que resultou na menor perda de validação. As combinações de hiperparâmetros foram ajustadas iterativamente para melhorar a performance do modelo.

Tabela 2 – Hiperparâmetros do LSTMVAE

Hiperparâmetro	Valor
Taxa de Aprendizado (*learning_rate*)	1.00×10^{-3} a 1.00×10^{-1}
Tamanho do *Batch* (*batch_size*)	64, 128, 256
Tamanho da Camada Oculta (*hidden_size*)	128, 256, 512
Tamanho Latente (*latent_size*)	2, 4, 10, 16, 32, 64

Fonte: elaborada pelo autor.

4.5 Avaliação e Resultados

Após o treinamento do modelo, utilizamos o conjunto de teste, composto por dados que o modelo ainda não havia visto, para avaliar o desempenho do LSTMVAE. Para isso, geramos o espaço latente das amostras do conjunto de teste e, em seguida, aplicamos a técnica *Uniform Manifold Approximation and Projection* (UMAP) para reduzir as dimensões desse espaço latente para duas componentes principais, facilitando a visualização dos dados.

O UMAP é um método de redução de dimensionalidade que preserva as relações de proximidade entre os dados no espaço de alta dimensão, sendo especialmente útil para visualização e análise de agrupamentos em espaços latentes. Diferente de outras técnicas, como Análise de Componentes Principais (PCA), o UMAP é capaz de capturar tanto a estrutura global quanto local dos dados, o que o torna adequado para a análise de dados complexos e não lineares, como os gerados por modelos baseados em autoencoders variacionais (MCINNES *et al.*, 2020).

4.6 Agrupamento com K-Means

Uma vez que o espaço latente é aprendido pelo VAE, aplicamos o algoritmo K-Means para realizar a agrupamento dos dados nesse espaço comprimido. O processo de agrupamento é

útil para identificar grupos naturais de amostras que compartilham características semelhantes, facilitando a análise de padrões e a categorização dos dados.

O algoritmo *K-Means* é aplicado a essas representações latentes para identificar padrões e segmentar os dados em grupos distintos. A escolha do número k de clusters é crucial para a eficácia do *K-Means*. O valor de k determina quantos grupos distintos o algoritmo irá formar, e essa escolha pode ter um impacto significativo nos resultados.

Para otimizar a seleção do número ideal de clusters, técnicas como *GridSearchCV* podem ser usadas em conjunto com o *K-Means*. O *GridSearchCV* é um método automatizado que realiza a busca exaustiva através de uma grade de parâmetros para encontrar a combinação que resulta na melhor desempenho do modelo. No contexto do *K-Means*, isso envolve testar diferentes valores de k e avaliar o desempenho do modelo para cada valor de k. Através de métricas de avaliação, como o índice de Davies-Bouldin, é possível determinar qual valor de k proporciona a melhor segmentação dos dados.

4.7 Validação dos resultados

Para validar os grupos obtidos com o algoritmo *K-Means*, utilizamos o *MMseqs2*, que é uma ferramenta de agrupamento baseada em similaridade de sequências para identificar pares de sequências agrupados.

A validação consiste em verificar se os pares de sequências identificados pelo *MMseqs2* estão corretamente agrupados dentro dos clusters definidos pelo *K-Means*. Para isso, comparamos os pares de sequências do *MMseqs2* e checamos se ambas pertencem ao mesmo grupo seguindo o *K-Means*. Se um par de sequências estiver no mesmo grupo em ambas as classificações, contamos essa correspondência como uma validação bem-sucedida.

Finalmente, calculamos a porcentagem de pares de sequências que foram corretamente agrupados dentro de cada grupo. Esse cálculo nos permite avaliar a coesão dos grupos formados pelo *K-Means*, comparando-os com os grupos independentes fornecidos pelo *MMseqs2*.

5 RESULTADOS

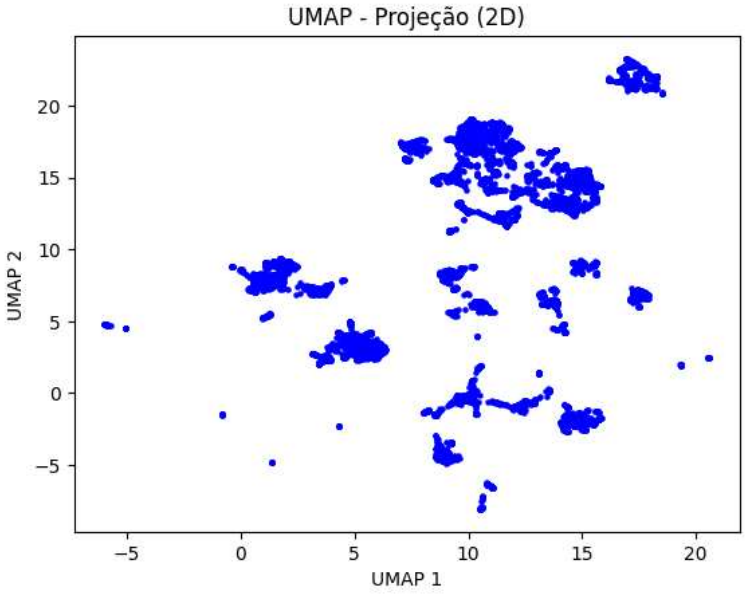
Durante a otimização dos hiperparâmetros do LSTMVAE, que envolveu 50 ensaios, os melhores parâmetros encontrados estão listados na Tabela 3. Esses parâmetros foram usados para converter os dados para uma representação latente e para a projeção dos dados com o UMAP, resultando na visualização das duas principais componentes do espaço latente, conforme mostrado na Figura 7. Esta projeção revelou a distribuição dos dados e a formação dos grupos, permitindo uma validação visual da separação deles.

Tabela 3 – Hiperparâmetros encontrados

Hiperparâmetro	Valor
Taxa de Aprendizado (*learning_rate*)	7.51×10^{-4}
Tamanho do *Batch* (*batch_size*)	64
Tamanho da Camada Oculta (*hidden_size*)	256
Tamanho Latente (*latent_size*)	10
Melhor Perda	0.578

Fonte: elaborada pelo autor.

Figura 7 – Projeção dos dados em duas dimensões.

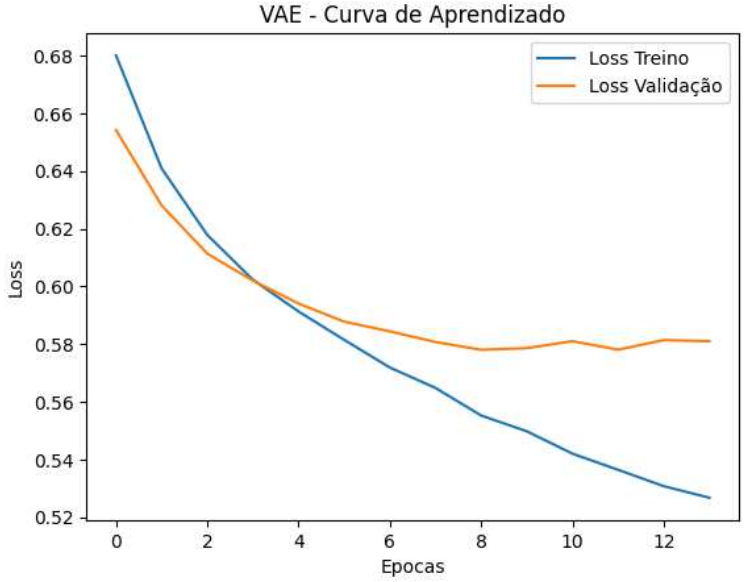


Fonte: elaborada pelo autor.

Os resultados de perca do treinamento do modelo são ilustrados na Figura 8, que mostra a curva de aprendizado ao longo de 14 épocas. Durante esse processo, utilizamos um conjunto de validação para monitorar o desempenho e detectar sinais de *overfitting*. Para evitar que o modelo superajustasse aos dados de treinamento, foi utilizado a técnica de parada

prematura (*early stopping*), para interromper o treinamento. Essa estratégia permitiu manter a generalização do modelo, evitando o ajuste excessivo aos dados de treino e melhorando o desempenho geral no conjunto de validação.

Figura 8 – Curva de aprendizado do LSTMVAE.



Fonte: elaborada pelo autor.

Durante a execução do *GridSearch* para o *K-means*, utilizando o espaço latente gerado a partir do conjunto de teste, foi determinado que o número ideal de *clusters* era 20. Essa escolha foi baseada na análise do índice de Davies-Bouldin, que avaliou a compactação e separação dos grupos, na qual obtemos uma pontuação aproximada de *1.46*. O índice DB, que mede a qualidade do agrupamento ao comparar a distância média entre grupos e a dispersão dentro dos grupos, indicou que 20 grupos proporcionavam uma divisão equilibrada e representativa dos dados no espaço latente.

A comparação entre os 20 grupos obtidos e os agrupamentos do *MMseqs2* indicou que, em média, os grupos apresentaram uma correspondência de aproximadamente 82,55% na similaridade dos pares, como mostrado na Tabela 4. Esse valor sugere que a correspondência entre os grupos do *K-means* e os agrupamentos do *MMseqs2* é razoável, mas não perfeita. É importante ressaltar que o *MMseqs2* realiza agrupamentos com base no alinhamento de sequência, enquanto o método de agrupamento deste trabalho considera as sequências das CDRs e a região paratopo dos anticorpos. Essa abordagem adicional pode ter contribuído para a diferença nas estatísticas de correspondência, já que o *MMseqs2* não leva em conta essa dimensão extra de informação. Portanto, a análise sugere que a inclusão da região paratopo oferece uma perspectiva

mais diferenciada, que pode não ser completamente capturada pelos métodos de alinhamento de sequência utilizados pelo *MMseqs2*.

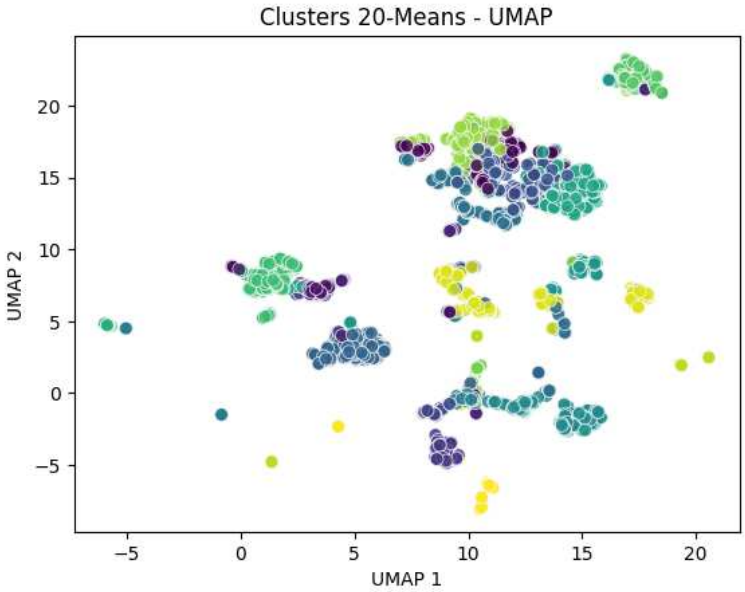
Tabela 4 – Dados dos Clusters

Cluster	Tamanho	Pares agrupados pelo *MMseqs2*	Percentual similaridade
0	306	208	67.97%
1	205	171	83.41%
2	70	60	85.71%
3	196	191	97.45%
4	138	76	55.07%
5	391	153	39.13%
6	602	591	98.17%
7	447	353	78.97%
8	161	146	90.68%
9	448	257	57.37%
10	194	181	93.30%
11	647	562	86.86%
12	45	45	100.00%
13	497	443	89.13%
14	319	306	95.92%
15	62	62	100.00%
16	573	394	68.76%
17	66	51	77.27%
18	509	470	92.34%
19	124	116	93.55%

Fonte: elaborada pelo autor.

Por fim, a Figura 9 exibe a distribuição dos grupos no plano. Apesar de a visualização ser limitada pela alta quantidade de grupos, ela ainda contribui para uma melhor compreensão da estrutura dos grupos gerados pelo modelo.

Figura 9 – Projeção dos dados em duas dimensões.



Fonte: elaborada pelo autor.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou uma nova abordagem para a análise e agrupamento de anticorpos utilizando técnicas de aprendizado de máquina. Através da aplicação de VAE, conseguimos mapear os dados de sequência de anticorpos em um espaço latente, o que possibilitou em um agrupamento por meio do algoritmo *K-Means*. Com isso, foi possível identificar grupos de anticorpos com características semelhantes, proporcionando uma base sólida para futuras investigações sobre suas similaridades.

Os resultados alcançados demonstraram que a integração de técnicas de aprendizado profundo com métodos de agrupamento se mostrou eficiente para organizar grandes volumes de dados de anticorpos, permitindo a identificação de padrões e características importantes. Essa abordagem contribui para a otimização do processo de descoberta de novos tratamentos terapêuticos, especialmente no campo das imunoterapias, ao utilizar técnicas modernas de aprendizado de máquina em um problema de alta relevância biológica. A aplicação de técnicas de clusterização possibilitou a classificação de anticorpos com base nas suas propriedades das sequências de CDRs, acelerando assim a triagem de potenciais terapias e o desenvolvimento de tratamentos direcionados.

Apesar dos avanços obtidos, há ainda várias direções promissoras para o aprimoramento deste trabalho. Uma das principais áreas de interesse para trabalhos futuros é a utilização de arquivos no formato PDB ao invés de FASTA. Diferente dos arquivos de sequência simples, os PDB contêm informações tridimensionais sobre a estrutura dos anticorpos, o que oferece uma visão mais detalhada das suas propriedades físico-químicas e de interação. Especificamente, esses arquivos descrevem a posição exata dos átomos que compõem a proteína no espaço tridimensional, permitindo a visualização e análise de aspectos como dobras, interações de superfície e a proximidade entre resíduos que influenciam diretamente sua funcionalidade e ligação com antígenos. Dessa forma, o uso de dados estruturais pode fornecer insights mais robustos para o entendimento das interações moleculares e a melhoria de modelos preditivos.

Outro ponto a ser explorado é a representação das estruturas de anticorpos por meio de grafos, uma vez que essa abordagem permite capturar diretamente a relação espacial e estrutural das proteínas. A matriz de adjacência, junto com as *features* dos nós, seria usada para representar as interações entre os resíduos e, assim, fornecer um modelo mais realista e detalhado da estrutura do anticorpo.

Além das abordagens já mencionadas, é importante explorar outras formas de validar

e verificar a qualidade dos dados e dos resultados obtidos. Uma alternativa promissora é o uso de ferramentas especializadas como o CLAP (*Classification of Proteins*). O CLAP é um servidor web que realiza a classificação automática de proteínas, com foco especial em proteínas multi-domínios (CHOMICZ *et al.*, 2024). Essa ferramenta pode ser especialmente útil para a validação dos grupos gerados neste trabalho, pois oferece uma abordagem adicional para o agrupamento com base na região de paratopo, complementando a análise realizada com os Autocodificadores Variacionais e o *K-Means*.

Além disso, é de interesse a utilização de todo o conjunto de dados disponível, que compreende mais de 1,3 milhões de amostras. Essa abordagem se justifica pela possibilidade de fornecer ao modelo uma maior diversidade de amostras, capturando variações que podem não estar presentes em um subconjunto menor. Dessa forma, um conjunto de dados mais extenso tem o potencial de melhorar a robustez e a generalização do modelo, tornando as descobertas aplicáveis a um número maior de casos.

Apesar das vantagens potenciais de utilizar todo o conjunto de dados disponível é importante reconhecer que o aumento indiscriminado da quantidade de dados nem sempre resulta em melhorias automáticas. Em cenários futuros, será necessário garantir que esse crescimento venha acompanhado de um processo rigoroso de filtragem, para evitar a inclusão de amostras redundantes ou de baixa qualidade, que poderiam comprometer a eficiência do treinamento.

Com o uso de técnicas como o aprendizado em grafos e dados estruturais mais completos, é possível que no futuro este trabalho contribua de maneira ainda mais significativa para a otimização de tratamentos imunoterapêuticos e a compreensão das interações entre anticorpos e antígenos.

REFERÊNCIAS

BANK, R. P. D. **Guide to Understanding PDB Data**. s.d. Disponível em: https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/pdb-overview. Acesso em: 27 mai. 2024.

CHOMICZ, D.; KOńCZAK, J.; WRóBEL, S.; SATłAWA, T.; DUDZIC, P.; JANUSZ, B.; TARKOWSKI, M.; DESZYńSKI, P.; GAWłOWSKI, T.; KOSTYN, A.; ORłOWSKI, M.; KLAUS, T.; SCHULTE, L.; MARTIN, K.; COMEAU, S. R.; KRAWCZYK, K. Benchmarking antibody clustering methods using sequence, structural, and machine learning similarity measures for antibody discovery applications. **Frontiers in Molecular Biosciences**, v. 11, 2024. ISSN 2296-889X. Disponível em: https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2024.1352508.

CHOTHIA, C.; LESK, A. M. Canonical structures for antibody varieties. **Journal of Molecular Biology**, v. 196, n. 4, p. 901–917, 1992.

DOERSCH, C. **Tutorial on Variational Autoencoders**. 2021. Disponível em: https://arxiv.org/abs/1606.05908.

DUNBAR, J.; DEANE, C. M. Anarci: Antigen receptor numbering and receptor classification. **Bioinformatics**, Oxford University Press, v. 32, n. 2, p. 298–300, 2016.

FERNANDEZ-QUINTERO, M. L.; POMARICI, N. D.; FISCHER, A.-L. M.; HOERSCHINGER, V. J.; KROELL, K. B.; RICCABONA, J. R.; KAMENIK, A. S.; LOEFFLER, J. R.; FERGUSON, J. A.; PERRETT, H. R.; LIEDL, K. R.; HAN, J.; WARD, A. B. Structure and dynamics guiding design of antibody therapeutics and vaccines. **Antibodies**, v. 12, n. 4, 2023. ISSN 2073-4468. Disponível em: https://www.mdpi.com/2073-4468/12/4/67.

GHANBARPOUR, A.; AL. et. Structure-free antibody paratope similarity prediction for in silico epitope binning via protein language models. **iScience**, v. 26, n. 2, p. 106036, 2023.

Graylab. **IgFold: Antibody Structure Dataset**. 2023. Acesso em: 10 mai. 2024. Disponível em: https://github.com/Graylab/IgFold?tab=readme-ov-file.

HUANG, Y.; ZHANG, Z.; ZHOU, Y. Abagintpre: A deep learning method for predicting antibody-antigen interactions based on sequence information. **Frontiers in Immunology**, v. 13, p. 1053617, 12 2022.

IBM. **O que é deep learning?** 2024. Disponível em: https://www.ibm.com/br-pt/topics/deep-learning. Acesso em: 08 jun. 2024.

IKOTUN, A. M.; EZUGWU, A. E.; ABUALIGAH, L.; ABUHAIJA, B.; HEMING, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. **Information Sciences**, v. 622, p. 178–210, 2023. ISSN 0020-0255. Disponível em: https://www.sciencedirect.com/science/article/pii/S0020025522014633.

JAFFE, D. B.; SHAHI, P.; ADAMS, B. A. *et al.* Functional antibodies exhibit light chain coherence. **Nature**, Nature Publishing Group, v. 611, n. 7934, p. 352–357, 2022.

JANEWAY, C. A. J.; TRAVERS, P.; WALPORT, M. *et al.* **Immunobiology: The Immune System in Health and Disease**. 5th. ed. New York: Garland Science, 2001. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK27160/.

JUMPER, J. *et al.* Highly accurate protein structure prediction with alphafold. **Nature**, Nature Publishing Group, v. 596, p. 583–589, 2021.

KABAT, E. A.; WU, T. T.; REID, W. **Sequences of Proteins of Immunological Interest**. [*S. l.*]: National Institutes of Health, 1991.

KEMPEN, M. van; KIM, S.; TUMESCHEIT, C. *et al.* Fast and accurate protein structure search with foldseek. **Nature Biotechnology**, v. 42, p. 243–246, 2024.

KINGMA, D. P.; WELLING, M. Auto-encoding variational bayes. **arXiv preprint arXiv:1312.6114**, 2013. Fixes a typo in the abstract, no other changes. Disponível em: https://doi.org/10.48550/arXiv.1312.6114.

LI, W.; JAROSZEWSKI, L.; GODZIK, A. Clustering of highly homologous sequences to reduce the size of large protein databases. **Bioinformatics**, Oxford University Press, v. 17, n. 3, p. 282–283, 2001.

LIBERIS, E.; VELIčKOVIć, P.; SORMANNI, P.; VENDRUSCOLO, M.; LIò, P. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. **Bioinformatics**, Oxford University Press, v. 34, n. 17, p. 2944–2950, 2018. Disponível em: https://doi.org/10.1093/bioinformatics/bty305.

MCINNES, L.; HEALY, J.; MELVILLE, J. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. 2020. Disponível em: https://arxiv.org/abs/1802.03426.

SANTOS, V. Sardinha dos. **Anticorpos**. s.d. Mundo Educação. Disponível em: https://mundoeducacao.uol.com.br/biologia/anticorpos.htm. Acesso em: 04 set. 2024.

SHARIFANI, K.; AMINI, M. Machine learning and deep learning: A review of methods and applications. **World Information Technology and Engineering Journal**, v. 10, n. 07, p. 3897–3904, 2023. Disponível em: https://ssrn.com/abstract=4458723.

STEINEGGER, M.; SöDING, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. **Nature Biotechnology**, v. 35, n. 11, p. 1026–1028, 2017. Disponível em: https://doi.org/10.1038/nbt.3988.

ZHOU, J.; CUI, G.; HU, S.; ZHANG, Z.; YANG, C.; LIU, Z.; WANG, L.; LI, C.; SUN, M. **Graph Neural Networks: A Review of Methods and Applications**. 2021. Disponível em: https://arxiv.org/abs/1812.08434.