A DBSCAN-Based Approach for Evaluating Protein Clusters Using Embeddings Generated from k-mer Images with Vision Transformer

Giovanna A. P. Soares[1], Hannah I. S. Marques[1], Raquel de M. Barbosa[3] Marcelo A. C. Fernandes[1,2,4]

[1]InovAI Lab, nPITI/IMD, UFRN, 59.078-900, Natal, RN, Brazil

[2]Bioinformatics Multidisciplinary Environment (BioME), IMD, UFRN, Natal, Brazil

[3]Faculty of Pharmacy, Granada University, Granada, Spain

[4]Department of Computer Engineering and Automation, UFRN, Natal, Brazil

giovanna.assuncao.705@ufrn.edu.br, hannahisabele1516@gmail.com

rbarbosa@ugr.es, mfernandes@dca.ufrn.br

Abstract. *This work presents an approach for analyzing protein clusters from vector embeddings generated by k-mer images processed with a Vision Transformer (ViT) model. The approach, independent of alignment, enables the use of density-based clustering methods such as DBSCAN, applied to the UniRef100 and UniRef90 datasets. Two metrics were proposed: contamination, which measures the purity of clusters with respect to the original labels, and dispersion, which quantifies the fragmentation of a label into multiple groups. The results show that UniRef100 had low contamination, while UniRef90 had higher dispersion. The methodology allows the identification of substructures and internal divisions within labels and provides measures for curation, refinement, and functional annotation of biological databases.*

Resumo. *Este trabalho apresenta uma abordagem para análise de agrupamentos de proteínas a partir de embeddings vetoriais gerados por imagens de k-mers processadas por um modelo Vision Transformer (ViT). A abordagem, independente de alinhamento, permite o uso de métodos de agrupamento baseados em densidade, como o DBSCAN, aplicado aos conjuntos UniRef100 e UniRef90. Foram propostas duas métricas: contaminação, que mede a pureza dos clusters em relação aos rótulos originais, e espalhamento, que quantifica a fragmentação de um rótulo em múltiplos grupos. Os resultados mostram que o UniRef100 apresentou baixa contaminação, enquanto o UniRef90 apresentou maior dispersão. A metodologia possibilita identificar subestruturas e divisões internas em rótulos e fornecer medidas para curadoria, refinamento e anotação funcional de bases biológicas.*

1. Introdução

Com o crescimento acelerado das bases de dados biológicas, em especial aquelas relacionadas a sequências de proteínas, surge a necessidade de métodos mais eficientes e interpretáveis para organização, curadoria e análise funcional desses dados.

Conjuntos como o UniProt e suas variantes derivadas, como o UniRef100 e UniRef90 [Consortium 2023], fornecem agrupamentos de proteínas com base em critérios de identidade, mas ainda enfrentam limitações no que diz respeito à coesão interna e à representação da diversidade funcional presente dentro de cada grupo. Embora métodos tradicionais baseados em alinhamento e identidade de sequência sejam amplamente utilizados, esses nem sempre são capazes de capturar relações mais sutis de similaridade estrutural ou funcional. Nesse cenário, abordagens baseadas em aprendizado profundo e representação vetorial têm ganhado destaque por sua capacidade de representar sequências em espaços latentes ricos em informação semântica. Em particular, a conversão de sequências em imagens de k-mers [Câmara et al. 2022, De Souza et al. 2022, Coutinho et al. 2023] e o uso de modelos de visão computacional como o *Vision Transformer* (ViT) [Dosovitskiy et al. 2020] têm se mostrado promissores para extrair *embeddings* discriminativos, abrindo caminho para novas formas de análise e agrupamento em dados biológicos complexos.

Vários trabalhos recentes demonstraram o potencial de representar sequências genéticas em forma de imagens a partir de contagens de k-mers, permitindo a aplicação de técnicas de *deep learning* para tarefas de classificação. Em [Câmara et al. 2022], foi proposta uma arquitetura baseada em redes convolucionais (CNN) para classificar sequências de SARS-CoV-2 com elevada acurácia, utilizando imagens construídas a partir da frequência de k-mers de genomas virais. A proposta mostrou-se robusta mesmo sem alinhamentos, revelando que a estrutura visual das sequências carrega padrões discriminativos relevantes. De forma complementar, o trabalho de [De Souza et al. 2022] aplicou essa representação em tarefas de predição de interações droga-alvo, reforçando a viabilidade da abordagem para diferentes domínios bioinformáticos. Em [Coutinho et al. 2023], *autoencoders* empilhados (SSAE) foram utilizados para extrair representações latentes de imagens k-mer derivadas de sequências virais, ampliando as aplicações possíveis desse paradigma. Esses estudos consolidam uma base metodológica que este trabalho expande ao utilizar imagens k-mers processadas por ViT, explorando uma arquitetura alternativa para geração de *embeddings* vetoriais ricos em informação biológica estrutural.

O ViT é uma arquitetura de redes neurais baseada em mecanismos de atenção originalmente desenvolvida para tarefas de visão computacional, como classificação de imagens e detecção de objetos. Diferentemente de arquiteturas convolucionais tradicionais, o ViT divide a imagem de entrada em pequenos blocos (*patches*), que são linearizados e tratados como *tokens* de uma sequência, permitindo o uso direto dos mecanismos de autoatenção do *Transformer*. Essa abordagem possibilita que o modelo aprenda relações de longo alcance entre regiões da imagem de forma mais direta e interpretável [Dosovitskiy et al. 2020, Yin et al. 2024, Yang et al. 2024, Ma et al. 2023]. No contexto deste trabalho, o ViT é utilizado para extrair embeddings vetoriais de imagens construídas a partir de sequências de proteínas, codificadas em representações de k-mers. A capacidade do ViT de capturar padrões estruturais distribuídos espacialmente o torna particularmente adequado para representar variações locais e globais em dados biológicos complexos, sendo uma escolha promissora para tarefas que envolvem agrupamento e caracterização de proteínas.

O DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) é um algoritmo de agrupamento amplamente utilizado por sua capacidade de identificar grupos

em dados com formatos arbitrários e presença de ruído. Ao contrário de métodos tradicionais como k-means, o DBSCAN não exige que o número de clusters seja previamente definido, o que o torna especialmente útil em cenários exploratórios e em dados de alta complexidade. Sua abordagem baseada em densidade permite detectar automaticamente regiões densas no espaço de características e separar pontos esparsos como ruído, oferecendo uma segmentação mais flexível e robusta. Em contextos nos quais os dados são representados em espaços vetoriais latentes, como *embeddings* derivados de imagens ou sequências, o DBSCAN se destaca por conseguir identificar agrupamentos com diferentes formas, tamanhos e densidades. Essa flexibilidade torna o algoritmo particularmente atrativo para aplicações envolvendo dados não estruturados ou provenientes de representações geradas por redes neurais profundas, como os utilizados neste trabalho [Wang et al. 2022, Karim et al. 2021, Singh et al. 2022, Kulkarni and Burhanpurwala 2024].

Assim, este trabalho propõe uma abordagem baseada em *embeddings* vetoriais extraídos de imagens de k-mers processadas por um ViT, combinada com o algoritmo DBSCAN para a análise de agrupamentos em dados proteicos. Ao invés de depender de alinhamentos ou medidas de identidade direta, a proposta explora representações latentes que preservam padrões estruturais e funcionais, permitindo a aplicação de técnicas de agrupamento baseadas em densidade. Para avaliar a qualidade dos agrupamentos gerados, são utilizadas duas métricas complementares: a contaminação, que quantifica a pureza de cada cluster em relação aos rótulos de origem, e o espalhamento, que mede a dispersão de um mesmo rótulo ao longo de múltiplos clusters. A aplicação dessa metodologia aos conjuntos UniRef100 e UniRef90 demonstra que o modelo é capaz de revelar subestruturas internas e heterogeneidades não detectadas por métodos tradicionais, oferecendo uma nova perspectiva para a organização e curadoria de grandes bancos de proteínas.

2. Metodologia

2.1. Dataset

Neste trabalho foram utilizados dois conjuntos derivados da base UniRef: o UniRef100 e o UniRef90. Ambos agrupam sequências de proteínas a partir do UniProtKB com base em critérios de similaridade. O UniRef100 contém todas as sequências únicas, enquanto o UniRef90 agrupa sequências com pelo menos 90% de identidade. Essa diferença impacta diretamente na coesão dos grupos, sendo o UniRef90 naturalmente mais permissivo e, portanto, mais heterogêneo. Foram selecionados manualmente 10 grupos (ou *clusters*) de cada conjunto, buscando diversidade funcional e taxonômica. A Tabela 1 apresenta os *clusters* escolhidos do UniRef100, com informações sobre o identificador do *cluster*, o nome do gene ou proteína de referência, o número de aminoácidos (AA), a espécie ou grupo predominante e o número de sequências por *cluster*. De maneira semelhante, a Tabela 2 resume os *clusters* utilizados do UniRef90, também detalhando a diversidade de organismos associados a cada grupo.

No caso do UniRef90, observa-se que vários *clusters* apresentam múltiplos organismos ou espécies relacionadas (como diferentes sorotipos virais ou subgrupos taxonômicos), o que contribui para o aumento da dispersão observada nas análises posteriores. Essas características tornam o UniRef90 particularmente interessante para avaliar a capacidade do modelo em identificar padrões de agrupamento mesmo diante de maior variabilidade biológica.

Tabela 1. Distribuição detalhada dos grupos (*clusters*) UniRef100.

Uniref Cluster ID	Gene/Proteína	Nº AA (Ref.)	Espécie ou grupo predominante	Nº Sequências
A0A0F7YU55	Matrix protein 1	252	*Influenza A virus*	971
A0A0U0T1S6	PPE family protein	73	*Mycobacterium orygis 112400015*	1
			Mycobacterium tuberculosis	1
			Missing	830
A0A0Z0R4A7	Surface protein G	125	*Escherichia coli*	1
			Leptospira borgpetersenii serovar Ballum	1
			Staphylococcus aureus	6
			Missing	992
A0A1P8KIE2	Genome polyprotein	3391	*Dengue virus*	36
			Dengue virus type 1	3
			Dengue virus type 2	834
A0A3G1IDJ8	Major capsid protein	540	*Norovirus GII*	618
			Norovirus GII.17	241
			Outros Norovirus GII.17 (isolados individuais)	75
A0A517E530	Gag protein (Fragment)	123	*Human immunodeficiency virus type 1*	799
A0A6M6AQ24	Pol protein (Fragment)	1009	*Human immunodeficiency virus type 1*	887
Q5MXE2	p72 protein (Fragment)	138	*African swine fever virus*	827
Q67953	Large envelope protein	445	*Hepatitis B virus*	994
W6JI24	Cytochrome c oxidase subunit 1	512	Diversas espécies de crustáceos (*Acantholobulus*)	956

2.2. *k*-mers e ViT

A Figura 1 ilustra o fluxo de processamento adotado para a geração dos *embeddings* utilizados neste trabalho. A partir de uma sequência de aminoácidos, aplica-se uma contagem de k-mers por meio do módulo Seq2MC, resultando em uma matriz de frequência que representa os padrões locais da sequência. Essa matriz é então convertida em uma imagem por meio do método MC2Image, conforme proposto em [Câmara et al. 2022, De Souza et al. 2022, Coutinho et al. 2023], produzindo uma representação visual que preserva informações estruturais da sequência original. A dimensão da imagem gerada depende do valor de k, sendo normalmente projetada como uma matriz quadrada de dimensão $\left\lceil \sqrt{21^k} \right\rceil \times \left\lceil \sqrt{21^k} \right\rceil$, garantindo consistência de tamanho mesmo entre sequências de comprimentos variados. As imagens são processadas por um modelo ViT B/16, inicialmente treinado no *ImageNet-21k* e posteriormente refinado no *ImageNet-1k*, com entradas redimensionadas para 384×384 *pixels*. Esse modelo extrai *embeddings* vetoriais que capturam características estruturais, funcionais e relações de similaridade entre as sequências. Diferentemente das abordagens tradicionais baseadas em alinhamento, essa estratégia permite representar sequências com diferentes tamanhos de forma unificada, viabilizando o uso de métricas vetoriais e técnicas de agrupamento baseadas em densidade. Além disso, o uso de imagens e modelos visuais oferece vantagens computacionais importantes, como paralelização e compatibilidade com aceleradores (GPU/TPU), sendo também altamente escalável para grandes bancos de dados.

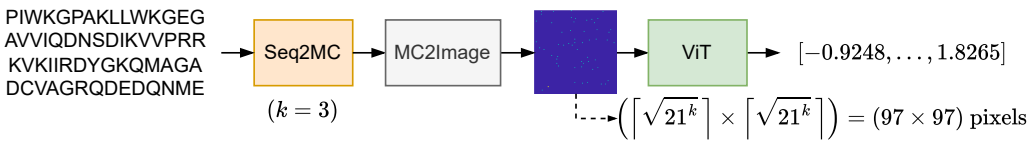


Figura 1. Fluxo de transformação de uma sequência de aminoácidos em *embedding* vetorial via representação por imagem de k-mers e processamento por *Vision Transformer* (ViT). Exemplo da proteína HV1 Pol protein (Fragment) da espécie *Human immunodeficiency virus type 1*.

Tabela 2. Distribuição detalhada dos grupos (*clusters*) UniRef90.

Uniref Cluster ID	Gene/Proteína	Nº AA (Ref.)	Espécie ou Isolado predominante	Nº Sequências
A0A0E3X5I8	Capsid protein	540	*Norovirus GII*	277
			Norovirus GII.17	220
			Outros Norovirus (isolados individuais)	restante
A0A0U0T1S6	PPE family protein	73	Missing	990
			Mycobacterium orygis 112400015	1
			Mycobacterium tuberculosis	1
A0A0Z0R4A7	Surface protein G	125	Missing	891
			Staphylococcus aureus	6
			Staphylococcus schweitzeri	1
			Escherichia coli	1
			Leptospira borgpetersenii serovar Ballum	1
A0A1D8B9X1	Cytochrome c oxidase subunit 1	511	*Acantholobulus bermudensis*	3
			Outros Acantholobulus	restante
A0A290XZ71	Major capsid protein p72	133	*African swine fever virus*	900
A0A517E5J2	Gag protein (Fragment)	123	*Human immunodeficiency virus type 1*	925
P03141	Large envelope protein	400	*Hepatitis B virus*	895
			Outros isolados de Hepatitis B virus	restante
P05777	Matrix protein 1	252	*Influenza A virus*	807
			Outros isolados de Influenza A virus	restante
P29990	Genome polyprotein	3391	*Dengue virus type 2*	740
			Dengue virus	132
			Outros isolados Dengue virus	restante
Q8UTD3	Pol protein (Fragment)	1007	*Human immunodeficiency virus type 1*	925

2.3. Agrupamento com o DBSCAN

Para os conjuntos UniRef90 e UniRef100, foi realizada uma análise sistemática utilizando o algoritmo DBSCAN, com 6 experimentos variando o parâmetro P, que representa o número mínimo de pontos na vizinhança necessários para definir uma região como densa, no intervalo de 2 a 7. Para cada n-ésimo experimento com valor P^n, foi estimado um valor específico de ε^n com base na análise da curva da j-distância. Essa variável ε^n define a distância máxima entre dois pontos para que sejam considerados vizinhos no contexto do DBSCAN. A curva da v-distância é construída ordenando, de forma crescente, as distâncias entre cada ponto e seu v-ésimo vizinho mais próximo, sendo $v = P - 1$. A forma da curva permite visualizar a transição entre regiões densas e esparsas nos dados, e o ponto de inflexão, conhecido como cotovelo, é interpretado como um limiar natural de densidade. O valor de ε correspondente a esse cotovelo foi adotado como parâmetro ideal para a aplicação do DBSCAN, por representar um equilíbrio entre a sensibilidade à densidade local e a separação entre grupos.

A seguir, para cada n-ésimo experimento com a combinação de parâmetros (ε^n, P^n), o algoritmo DBSCAN foi aplicado sobre os embeddings das amostras. A fim de evitar ambiguidade entre os agrupamentos de referência definidos nos conjuntos UniRef100 (ver Tabela 1) e UniRef90 (ver Tabela 2) e os agrupamentos gerados pelo DBSCAN, adotamos neste artigo a convenção de chamar os grupos de origem dos dados de rótulos, reservando o termo grupos (ou *clusters*) exclusivamente para os grupos encontrados pelo algoritmo de DBSCAN.

Em cada n-ésimo experimento, foram calculadas quatro métricas principais: (i) o valor médio do índice de silhueta ($\bar{s}^n$), que quantifica a coesão dos agrupamentos realizados pelo DBSCAN; (ii) a contaminação global média, que avalia a pureza geral dos agrupamentos do DBSCAN ($\bar{\delta}^n$) em relação aos rótulos do UniRef90 e UniRef100; (iii) o número total de *clusters* identificados (C^n), excluindo ruído; e (iv) o número de amostras

classificadas como ruído (R^n), associadas ao rótulo -1. O cálculo da contaminação foi realizado da seguinte forma. Para cada i-ésimo *cluster* do n-ésimo experimento, representado aqui por $\mathbf{c}_i^n = [c_{i,1}^n, \ldots, c_{i,j}^n, \ldots, c_{i,N_i^n}^n]$, foram extraídos os rótulos dominantes das N_i^n amostras presentes no *cluster*, $c_{i,j}^n$. O rótulo dominante foi definido como aquele com maior frequência absoluta dentro do *cluster*, $\mathbf{c}_i^n$. A contaminação do i-ésimo *cluster* do n-ésimo experimento com o r-ésimo rótulo dominante, $w_{i,r}^n$, pode ser expressa como

$$w_{i,r}^n = 1 - \frac{M_{i,r}^n}{N_i^n} \tag{1}$$

em que $M_{i,r}^n$ representa o número de amostras com o rótulo dominante r em $\mathbf{c}_i^n$ do i-ésimo *cluster* do n-ésimo experimento, onde $M_i^n \leq N_i^n$. Quanto mais próximo de zero o valor da contaminação, maior a pureza do *cluster*, ou seja, tende a ter apenas um rótulo r dominante.

Com base nas contaminações individuais de cada i-ésimo *cluster* do n-ésimo experimento, $w_{i,r}^n$, foi então calculada a contaminação global média, $\bar{\delta}^n$, definida como a média ponderada das contaminações de todos os *cluster* não ruidosos, ponderada pelo número de amostras em cada *cluster* e expressa como

$$\bar{\delta}^n = \frac{1}{\sum_{i=1}^{C^n} N_i^n} \sum_{i=1}^{C^n} N_i^n w_{i,r}^n. \tag{2}$$

Essa métrica reflete o grau médio de impureza dos agrupamentos, penalizando *clusters* mistos de forma proporcional ao seu tamanho.

As métricas obtidas foram utilizadas para avaliar a qualidade dos agrupamentos e selecionar os parâmetros mais adequados. Em seguida, a contaminação de cada rótulo de referência foi analisada em detalhe, e as divisões internas de seus agrupamentos foram cruzadas com informações biológicas, como diversidade de organismos ou funções associadas, para investigar a coerência biológica dos agrupamentos detectados. Do ponto de vista biológico, a contaminação de um *cluster* representa o grau de heterogeneidade em relação à classificação funcional ou taxonômica das amostras agrupadas. Um valor baixo de contaminação sugere que o DBSCAN foi capaz de isolar, de forma consistente, amostras pertencentes a um mesmo grupo funcional ou evolutivo, indicando que os embeddings preservam bem a estrutura semântica dos dados biológicos. Por outro lado, altos níveis de contaminação podem indicar sobreposição entre diferentes categorias biológicas, refletindo possíveis limitações do modelo de representação, regiões funcionais ambíguas ou relações filogenéticas próximas entre os grupos misturados. Nesses casos, a contaminação pode apontar regiões do espaço de embedding onde os limites taxonômicos se tornam difusos, ou ainda revelar casos biologicamente plausíveis de transição funcional ou convergência evolutiva.

Além da métrica de contaminação, foi introduzida uma segunda métrica denominada espalhamento que está associada aos rótulos. O espalhamento de um r-ésimo rótulo associado ao n-ésimo experimento, γ_r^n, tem como objetivo quantificar o grau de fragmentação de um rótulo verdadeiro ao longo dos *clusters* gerados pelo DBSCAN. Essa métrica reflete em que medida as amostras associadas a um mesmo rótulo estão distribuídas em diferentes agrupamentos, independentemente de o rótulo ser dominante

em cada *cluster*. O espalhamento, γ_r^n, pode ser expresso como

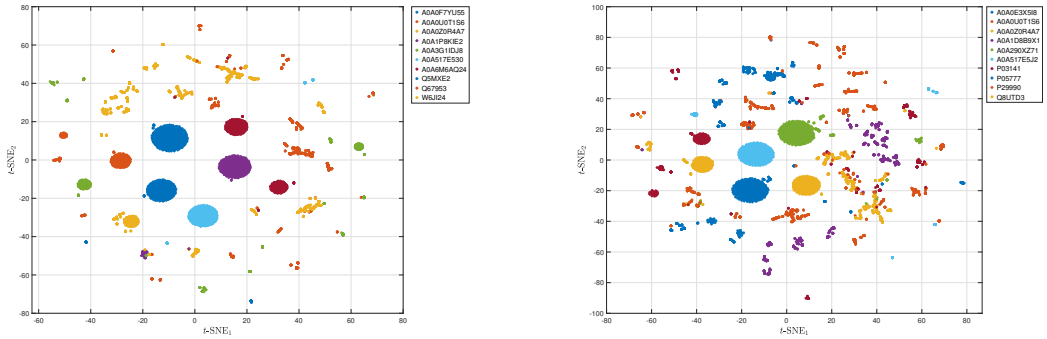
$$\gamma_r^n = \frac{C_r^n - 1}{C^n - 1} \tag{3}$$

onde C_r^n o número de *clusters* nos quais o rótulo r aparece em pelo menos uma amostra. Essa métrica assume valor 0 quando todas as amostras do rótulo r estão concentradas em um único cluster, e tende a 1 quando o rótulo aparece disperso em praticamente todos os *clusters*. Dessa forma, o espalhamento serve como uma medida relativa da coesão estrutural de um rótulo dentro da organização de agrupamentos produzida pelo DBSCAN.

A métrica de espalhamento fornece uma visão complementar, indicando o quanto um determinado grupo de amostras associadas a um mesmo rótulo de referência encontra-se disperso ao longo dos diferentes agrupamentos formados. Um espalhamento baixo sugere que as amostras do grupo são internamente coesas e bem delimitadas no espaço de *embedding*, o que é desejável para rótulos associados a famílias taxonômicas bem definidas ou funções altamente conservadas. Em contraste, um espalhamento elevado pode indicar heterogeneidade interna no grupo biológico, presença de subgrupos funcionais, variação estrutural significativa ou ainda efeito de ruído nos dados. Do ponto de vista biológico, essa dispersão pode revelar diversidade não capturada por rótulos simplificados, sinalizando a existência de subtipos, variantes ou características específicas que poderiam ser relevantes para refinamentos taxonômicos ou funcionais posteriores.

3. Resultados e Análises

As Figuras 2(a) e 2(b) apresentam a projeção bidimensional dos *embeddings* extraídos por meio do modelo ViT aplicados às imagens construídas com base em k-mers (ver Figura 1). As representações foram obtidas utilizando o algoritmo t-SNE e correspondem, respectivamente, aos agrupamentos dos *clusters* UniRef100 e UniRef90, chamados neste trabalho de rótulos.



(a) UniRef100 (b) UniRef90

Figura 2. Projeções t-SNE dos *embeddings* ViT gerados a partir das imagens k-mers dos rótulos (*clusters* UniRef).

A análise das projeções obtidas por t-SNE indica que os *embeddings* extraídos pelo modelo ViT são capazes de representar características discriminativas dos dados. Na Figura 2(a) (UniRef100), observa-se uma segmentação bem definida entre os diferentes

rótulos (*clusters* UniRef), com agrupamentos compactos e minimamente sobrepostos, sugerindo que a estrutura de similaridade entre as proteínas foi preservada mesmo após a conversão em imagens. Esse comportamento evidencia a capacidade do ViT em capturar informações relevantes para diferenciação biológica das sequências representadas. Na Figura 2(b) (UniRef90), embora a formação de agrupamentos também seja evidente, há maior dispersão e sobreposição entre alguns rótulos (*clusters* UniRef). Tal característica é condizente com a natureza mais permissiva do UniRef90, que agrupa sequências com até 90% de identidade. Ainda assim, o modelo manteve uma organização latente nas distribuições, o que reforça a robustez da técnica de *embedding* baseada em visão computacional aplicada a sequências proteicas.

As Tabelas 3 e 4 apresentam os resultados da aplicação do algoritmo DBSCAN aos conjuntos UniRef100 e UniRef90, respectivamente. Cada linha representa um experimento realizado com uma configuração específica de parâmetros e resume as principais métricas obtidas: número de *clusters* identificados, C^n, quantidade de amostras classificadas como ruído, R^n, índice médio de silhueta, s^n, e contaminação global média, $\bar{\delta}^n$. Esses indicadores permitem avaliar, de forma comparativa, a qualidade dos agrupamentos sob diferentes configurações de densidade, considerando aspectos como coesão interna, separabilidade e pureza em relação aos rótulos originais. Os resultados mostram contrastes marcantes entre os dois conjuntos, refletindo diferenças estruturais que serão discutidas nas análises a seguir.

Tabela 3. Resultados do DBSCAN para o conjunto UniRef100 com variação de p e ε.

n-ésimo experimento	$(p^n; \varepsilon^n)$	C^n	R^n	s^n	$\bar{\delta}^n$
1	$(2; 2{,}2533)$	185	163	0,80438	0,00000
2	$(3; 2{,}9945)$	91	173	0,80834	0,00034
3	$(4; 3{,}3987)$	65	191	0,80371	0,00056
4	$(5; 3{,}5136)$	52	226	0,81025	0,00057
5	$(6; 3{,}7297)$	47	234	0,81614	0,00057
6	$(7; 4{,}0618)$	42	227	0,81849	0,00057

Tabela 4. Resultados do DBSCAN para o conjunto UniRef90 com variação de p e ε.

n-ésimo experimento	$(p^n; \varepsilon^n)$	C^n	R^n	s^n	$\bar{\delta}^n$
1	$(2; 3{,}2017)$	217	268	0,72438	0,00067
2	$(3; 3{,}9374)$	135	307	0,72251	0,00079
3	$(4; 4{,}0399)$	108	384	0,72719	0,00079
4	$(5; 4{,}2766)$	96	399	0,72548	0,00080
5	$(6; 5{,}0551)$	76	388	0,69790	0,01840
6	$(7; 5{,}6205)$	64	382	0,65672	0,05460

Essas métricas permitem comparar a coesão e a pureza dos agrupamentos gerados sob diferentes configurações de densidade. Enquanto s^n avalia a separabilidade e a compacidade interna dos *clusters*, a métrica $\bar{\delta}^n$ quantifica a proporção média de contaminação observada nos *clusters* com base nos rótulos de referência. O número de amostras rotuladas como -1 (R_k) indica a quantidade de pontos considerados ruído pelo algoritmo.

A análise comparativa entre os experimentos visa identificar o melhor equilíbrio entre densidade, pureza e cobertura dos dados.

A análise comparativa dos resultados obtidos para os conjuntos UniRef100 e UniRef90 revela diferenças significativas na estrutura dos agrupamentos identificados pelo DBSCAN. No caso do UniRef100 (ver Tabela 3), observou-se uma alta coesão nos agrupamentos, refletida pelos elevados valores do índice de silhueta (s^n), todos acima de $0{,}80$, e por níveis de contaminação global média praticamente nulos ($\bar{\delta}^n \approx 0$). Isso indica que os embeddings gerados para as amostras do UniRef100 formam grupos bem separados e com alta pureza em relação aos rótulos de referência. Além disso, o número de amostras rotuladas como ruído (R_k) permaneceu relativamente estável, mesmo com variações nos parâmetros.

Por outro lado, os resultados obtidos para o UniRef90 (Ver Tabela 4) indicam uma estrutura de agrupamento menos definida. Embora os valores de s^n se mantenham em torno de $0{,}72$ nos experimentos com $p^n \leq 5$, observa-se um aumento gradual na contaminação global média à medida que p_k cresce, atingindo $\bar{\delta}^n = 0{,}05460$ com $p^n = 7$. Esse comportamento sugere maior sobreposição entre os grupos no espaço de *embeddings*, possivelmente refletindo a maior heterogeneidade intrínseca do UniRef90, que agrupa sequências com até 90% de identidade. O número de amostras classificadas como ruído (R_k) também foi consistentemente mais alto em comparação ao UniRef100, reforçando a hipótese de maior dispersão das amostras neste conjunto. Em conjunto, esses resultados evidenciam que o UniRef100 apresenta uma estrutura de agrupamento mais coesa e biologicamente consistente, enquanto o UniRef90 demanda uma análise mais cuidadosa devido à sua complexidade estrutural.

As Tabelas 5 e 6 apresentam um resumo estatístico das métricas de contaminação, $w^n_{i,r}$ e espalhamento γ^n_r por rótulo, considerando todos 6 experimentos com as diferentes configurações de p e ε para os conjuntos UniRef100 e UniRef90, respectivamente. Para cada rótulo, são reportados os valores médios ($\overline{w}_r$ e $\overline{\gamma}_r$), medianas ($\widetilde{w}_r$ e $\widetilde{\gamma}_r$) e desvios padrão (σ_{w_r} e σ_{γ_r}) das métricas de contaminação e espalhamento. Esses resultados permitem avaliar a estabilidade estrutural de cada grupo ao longo dos diferentes experimentos, fornecendo evidências sobre a coesão e a pureza dos agrupamentos. Rótulos com contaminação próxima de zero e espalhamento reduzido indicam maior consistência nos agrupamentos, enquanto valores elevados em ambas as métricas sugerem maior heterogeneidade e possível fragmentação funcional ou taxonômica.

Os resultados obtidos para o UniRef100 (ver Tabela 5) revelam um padrão altamente coeso e consistente nos agrupamentos formados. Para todos os rótulos avaliados, a contaminação média ($\overline{w}_r$) foi igual a zero, com desvios padrão nulos, indicando que os *clusters* compostos por amostras desses rótulos não foram contaminados por outros rótulos em nenhuma das configurações testadas. Além disso, os valores médios de espalhamento ($\overline{\gamma}_r$) variaram em sua maioria entre $0{,}01$ e $0{,}42$, com destaque para o rótulo `A0A0Z0R4A7`, que apresentou o maior espalhamento, refletindo certa dispersão ao longo dos *clusters*. Por outro lado, rótulos como `A0A0F7YU55` e `Q5MXE2` apresentaram espalhamento nulo, indicando que suas amostras estiveram consistentemente agrupadas em um único cluster ao longo de todos os experimentos. Esses resultados sugerem que os *embeddings* gerados para o UniRef100 são suficientemente expressivos para separar os grupos biológicos com elevada pureza e estabilidade estrutural.

Tabela 5. Estatísticas por rótulo para o UniRef100: média, mediana e desvio padrão da contaminação e do espalhamento.

Rótulo	$\overline{w}_r$	$\widetilde{w}_r$	σ_{w_r}	$\overline{\gamma}_r$	$\widetilde{\gamma}_r$	σ_{γ_r}
A0A0Z0R4A7	0,00000	0,00000	0,00000	0,41989	0,39721	0,09271
A0A0U0T1S6	0,00222	0,00265	0,00149	0,11744	0,10371	0,03421
Q67953	0,00000	0,00000	0,00000	0,10614	0,10570	0,03143
A0A3G1IDJ8	0,00000	0,00000	0,00000	0,10203	0,10347	0,02998
W6JI24	0,00000	0,00000	0,00000	0,06066	0,06114	0,00946
A0A6M6AQ24	0,00000	0,00000	0,00000	0,02795	0,02307	0,01046
A0A1P8KIE2	0,00000	0,00000	0,00000	0,01813	0,01796	0,00476
A0A517E530	0,00000	0,00000	0,00000	0,01632	0,01762	0,00709
A0A0F7YU55	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
Q5MXE2	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000

Tabela 6. Estatísticas por rótulo para o UniRef90: média, mediana e desvio padrão da contaminação e do espalhamento.

Rótulo	$\overline{w}_r$	$\widetilde{w}_r$	σ_{w_r}	$\overline{\gamma}_r$	$\widetilde{\gamma}_r$	σ_{γ_r}
P29990	0,00000	0,00000	0,00000	0,31474	0,31761	0,00944
A0A0Z0R4A7	0,00000	0,00000	0,00000	0,17197	0,18352	0,03701
P03141	0,00000	0,00000	0,00000	0,12451	0,12075	0,01319
A0A0E3X5I8	0,00000	0,00000	0,00000	0,08129	0,07470	0,01803
A0A1D8B9X1	0,00000	0,00000	0,00000	0,08095	0,07470	0,01035
P05777	0,00000	0,00000	0,00000	0,06279	0,06333	0,00416
A0A0U0T1S6	0,04264	0,00271	0,07043	0,03420	0,02663	0,02781
A0A517E5J2	0,00000	0,00000	0,00000	0,03213	0,02981	0,00994
Q8UTD3	0,00000	0,00000	0,00000	0,02576	0,02545	0,01052
A0A290XZ71	0,00000	0,00000	0,00000	0,01314	0,01361	0,00770

A análise do UniRef90 (ver Tabela 6) revela uma maior diversidade nos padrões de agrupamento, com evidências de maior heterogeneidade em comparação ao UniRef100. Embora a maioria dos rótulos tenha apresentado contaminação média nula, o rótulo `A0A0U0T1S6` se destacou com $\overline{w}_r = 0{,}04264$ e um desvio padrão relativamente elevado ($\sigma_{w_r} = 0{,}07043$), indicando variações relevantes na pureza de seus agrupamentos ao longo das configurações. Em relação ao espalhamento, rótulos como `P29990` e `A0A0Z0R4A7` apresentaram os maiores valores médios de $\overline{\gamma}_r$ (0,31 e 0,17, respectivamente), sugerindo uma tendência desses grupos a se fragmentarem em múltiplos *clusters*. Já rótulos como `A0A290XZ71` e `Q8UTD3` apresentaram baixos valores de espalhamento, com pouca variação, refletindo agrupamentos mais coesos. Esses resultados reforçam a maior complexidade estrutural do UniRef90, cuja construção permite até 90% de identidade entre sequências, resultando em grupos biologicamente mais amplos e internamente heterogêneos.

Os valores de espalhamento obtidos também revelam a presença de subgrupos internos mesmo entre rótulos que, a princípio, seriam considerados homogêneos. No caso do UniRef100 (ver Tabela 1), o rótulo `A0A0Z0R4A7` apresentou o maior espalhamento médio ($\overline{\gamma}_r = 0{,}42$), sugerindo uma fragmentação acentuada ao longo dos *clusters*. Essa

observação é compatível com a diversidade taxonômica listada para esse grupo, que inclui diferentes espécies bacterianas e registros ausentes (missing), indicando que o agrupamento original já incorporava variações biológicas relevantes. Por outro lado, rótulos como `A0A0F7YU55` e `Q5MXE2` apresentaram espalhamento nulo, mantendo suas amostras sempre coesas em um único cluster ao longo de todos os experimentos. Esses casos refletem maior uniformidade taxonômica e reforçam a capacidade do modelo em preservar estruturas altamente conservadas.

No UniRef90 (Tabela 2), observa-se ainda maior evidência de subestruturas internas. Rótulos como `P29990` e `A0A0Z0R4A7` apresentaram os maiores espalhamentos médios ($\overline{\gamma}_r = 0{,}31$ e $0{,}17$, respectivamente), indicando que suas amostras foram sistematicamente divididas em múltiplos *clusters*. Essa dispersão é coerente com a composição taxonômica mais heterogênea desses grupos, os quais incluem diferentes sorotipos virais, subgrupos taxonômicos e espécies distintas, como evidenciado nas descrições da Tabela 2. Tais padrões de espalhamento reforçam que a metodologia proposta permite detectar subgrupos latentes dentro de rótulos amplos, oferecendo suporte para refinamentos funcionais ou taxonômicos em bases de dados biológicas.

4. Conclusões

Este trabalho apresentou uma abordagem para análise de agrupamentos de proteínas a partir de *embeddings* vetoriais extraídos de imagens de k-mers processadas por um modelo ViT. A combinação dessa representação com o algoritmo DBSCAN permitiu investigar, de forma não supervisionada, a coesão e a fragmentação de grupos proteicos nos conjuntos UniRef100 e UniRef90. Para isso, foram propostas duas métricas complementares: contaminação, que mede a pureza dos *clusters* em relação aos rótulos de referência, e espalhamento, que quantifica o grau de dispersão de cada rótulo ao longo dos agrupamentos formados. Os resultados demonstraram que o UniRef100 apresenta uma organização estrutural mais coesa e consistente, com baixa contaminação e baixos níveis de espalhamento, indicando que os *embeddings* preservaram bem a estrutura semântica dos grupos biológicos. Em contrapartida, o UniRef90 evidenciou maior heterogeneidade interna, refletida por maiores valores de espalhamento em diversos rótulos (especialmente naqueles com composição taxonômica mais ampla), o que revela a existência de subgrupos funcionais ou estruturais latentes. A metodologia proposta mostrou-se eficaz na caracterização quantitativa da qualidade dos agrupamentos e na identificação de padrões internos não evidenciados por abordagens tradicionais. Esses resultados indicam que a integração de representações visuais com modelos de aprendizado profundo e técnicas baseadas em densidade oferece uma nova perspectiva para o refinamento, curadoria e anotação de bases de dados proteicos em larga escala.

Agradecimentos

Este trabalho foi financiado pelo CNPq, pelo Sistema Único de Saúde (SUS) e pelo Ministério da Saúde (MS), por meio do projeto nº 444306/2023-4 — *PharmaGenoNet: Uma Plataforma Integrada de Farmacogenômica Populacional e Deep Learning para Predição de Interações Fármaco-Alvo*.

Referências

Consortium, T. U. (2023). Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531.

Coutinho, M. G. F., Câmara, G. B. M., Barbosa, R. d. M., and Fernandes, M. A. C. (2023). Sars-cov-2 virus classification based on stacked sparse autoencoder. *Computational and Structural Biotechnology Journal*, 21:284–298.

Câmara, G. B. M., Coutinho, M. G. F., Silva, L. M. D. d., Gadelha, W. V. d. N., Torquato, M. F., Barbosa, R. d. M., and Fernandes, M. A. C. (2022). Convolutional neural network applied to sars-cov-2 sequence classification. *Sensors*, 22(15):5730.

De Souza, J. G., Fernandes, M. A., and de Melo Barbosa, R. (2022). A novel deep neural network technique for drug–target interaction. *Pharmaceutics*, 14(3):625.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., and Decker, S. (2021). Deep learning-based clustering approaches for bioinformatics. *Briefings in bioinformatics*, 22(1):393–415.

Kulkarni, O. and Burhanpurwala, A. (2024). A survey of advancements in dbscan clustering algorithms for big data. In *2024 3rd International conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC)*, pages 106–111. IEEE.

Ma, S., Gao, X., Jiang, L., and Xu, R. (2023). A review of visual transformer research. In *International Conference on Image, Vision and Intelligent Systems*, pages 349–356. Springer.

Singh, H. V., Girdhar, A., and Dahiya, S. (2022). A literature survey based on dbscan algorithms. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 751–758. IEEE.

Wang, J., Li, Z., and Zhang, J. (2022). Visualizing the knowledge structure and evolution of bioinformatics. *BMC bioinformatics*, 23(Suppl 8):404.

Yang, Q., Bai, Y., Liu, F., and Zhang, W. (2024). Integrated visual transformer and flash attention for lip-to-speech generation gan. *Scientific Reports*, 14(1):4525.

Yin, Y., Tang, Z., and Weng, H. (2024). Application of visual transformer in renal image analysis. *BioMedical Engineering OnLine*, 23(1):27.