

Abordagem Baseada em Embeddings e Transformers para Vigilância Genômica do SARS-CoV-2

Alessandro S. Silva*, Gabriel B. M. Câmara*[†] e Marcelo A. C. Fernandes*^{†‡}

*InovAI Lab, nPITI/IMD, UFRN, Natal, RN, Brasil.

[†]Centro Multiusuário de Bioinformática (BioME) - IMD, UFRN, Natal, RN, Brasil.

[‡]Departamento de Engenharia da Computação e Automação (DCA), UFRN, Natal, RN, Brasil.

Email: asdseng@gmail.com, gmottaz@gmail.com, mfernandes@dca.ufrn.br

Resumo—A pandemia de COVID-19 evidenciou a importância da vigilância genômica para o monitoramento da evolução do SARS-CoV-2 e a identificação de variantes de preocupação. Este trabalho propõe uma abordagem inovadora baseada em técnicas de *Natural Language Processing* (PLN), especificamente *embeddings* e *Transformers*, para representação vetorial e análise das substituições de aminoácidos presentes em sequências virais. Utilizando dados provenientes do *Global Initiative on Sharing Avian Influenza Data* (GISAID), foram extraídas e processadas aproximadamente 548 mil amostras com substituições únicas. Cada amostra foi tratada como um "documento", onde as substituições representam "palavras", permitindo a aplicação de modelos do tipo all-MiniLM-L6-v2 para geração de vetores de *embedding* de 384 dimensões. Esses vetores foram utilizados em análises de agrupamento, visualização e similaridade genética entre variantes. Os resultados obtidos demonstraram a eficácia da abordagem proposta, com a formação de clusters coerentes que refletem padrões específicos de variantes virais, como Delta, Alpha e Omicron. A estrutura vetorial gerada a partir das substituições de aminoácidos mostrou-se sensível às diferenças mutacionais entre essas variantes, permitindo sua segmentação em agrupamentos distintos no espaço latente.

Index Terms—SARS-CoV-2, vigilância genômica, Transformers, embeddings, variantes

I. INTRODUÇÃO

A emergência e rápida propagação do SARS-CoV-2 evidenciaram a necessidade de novas estratégias para monitorar sua evolução genômica e identificar precocemente variantes com potencial epidemiológico elevado [1]. À medida que milhões de sequências genéticas se tornaram disponíveis por meio de plataformas como o *Global Initiative on Sharing Avian Influenza Data* (GISAID) [2], a comunidade científica voltou-se para métodos computacionais escaláveis e capazes de lidar com a alta dimensionalidade e dinâmica das mutações virais. Dessa forma, técnicas de aprendizado profundo, especialmente aquelas inspiradas em *Natural Language Processing* (PLN), têm-se destacado por sua capacidade de extrair representações vetoriais ricas e sensíveis ao contexto de sequências mutacionais [3].

Trabalhos como [4], propõem uma analogia conceitual, tratando mutações como "palavras" e genomas como "documentos" em um grande "corpus" de evolução viral. Utilizando

modelos como Word2Vec e *Dynamic Topic Modeling* (DTM), esses autores conseguiram rastrear assinaturas mutacionais ao longo do tempo e associá-las a linhagens específicas e regiões geográficas distintas. Aspecto também abordado por [5], que propuseram um modelo baseado em atenção interpretável aplicado à sequência da proteína Spike para prever desfechos clínicos com base no padrão.

Outros pesquisadores avançaram na direção da predição de mutações. Por exemplo, no trabalho apresentado em [6] os autores desenvolveram o modelo chamado "TEMPO", uma arquitetura baseada em *transformers* que integra a estrutura de árvores filogenéticas com sequências evolutivas do vírus. Ao incorporar *embeddings* treinados sobre sequências temporais extraídas de trajetórias filogenéticas, o modelo "TEMPO" foi capaz de prever mutações futuras com desempenho superior a diversos métodos existentes, não apenas para o SARS-CoV-2, mas também para vírus como H1N1, H3N2 e H5N19. Essa generalização demonstra a robustez de arquiteturas *Transformers* no contexto de previsão evolutiva viral. Enquanto esses estudos focam principalmente na região da proteína spike, o trabalho apresentado em [7] amplia esse escopo ao propor o modelo *Strainflow*, que aplica *embeddings* não supervisionados sobre códons da spike, usando técnicas como FastTex e Word2Vec. Os autores utilizaram a entropia das representações latentes para prever surtos de casos com até dois meses de antecedência, além de correlacionar dimensões do *embedding* com mutações específicas de variantes de preocupação, como Delta e Omicron.

Outras pesquisas adotaram arquiteturas mais baseadas em aprendizagem profunda para classificar sequências genômicas. No trabalho apresentado em [8], os autores propuseram uma arquitetura híbrida baseada em múltiplos *Transformers* combinados com *Convolutional Neural Network* (CNNs) e *Long Short-Term Memory* (LSTM), treinada sobre diferentes representações numéricas de DNA, como *k*-mers e *one-hot encoding*, alcançando alta acurácia na classificação de variantes virais. Seguindo o mesmo caminho, o trabalho apresentado em [9] propôs modelos híbridos com atenção, combinando CNNs e LSTM bidirecionais (BLSTM) para classificar genomas com base em sua sequência completa. Os modelos atingiram uma

acurácia em torno 98% na diferenciação entre SARS-CoV-2 e outros vírus, demonstrando o poder de arquiteturas profundas para classificação. De forma semelhante, os trabalhos apresentados em [10]–[14] exploram diferentes formas de representações e classificação de sequências usando CNNs, alcançando acurácia superior a 98% na distinção entre SARS-CoV-2 e outros vírus da família *Coronaviridae*, reforçando a eficácia de abordagens baseadas em aprendizado profundo para a classificação genômica.

Apesar do valor dessas contribuições, a maior parte das abordagens da literatura permanece fortemente ancorada em sequências de nucleotídeos e ou sequências de aminoácidos. Poucos trabalhos têm priorizado a representação direta das substituições de aminoácidos, também chamados *AASubstitutions*, com entrada de dados. A representação em *AASubstitutions* possui um aspecto importante para entender os efeitos funcionais das mutações, especialmente quando se trata da diversidade observada em nível proteico. Baseado nisto, este trabalho tem como proposta a utilização do formato *AASubstitutions*, como entrada em modelos de linguagem baseados em *Transformers*. Aqui, cada amostra viral do dataset será representada como um “documento”, composto por um conjunto de substituições de aminoácidos, sem necessidade de alinhamento ou referência genômica explícita. Esses documentos são então transformados para um espaço latente via *embeddings* alta dimensionalidade por meio do modelo chamado all-MiniLM-L6-v2 [15].

Os resultados revelaram que os *embeddings* gerados a partir do formato em *AASubstitutions* foram capazes de capturar padrões estruturais característicos das variantes virais do SARS-CoV-2. A aplicação de algoritmos não supervisionados de agrupamento sobre os vetores resultou em *clusters* bem definidos, nos quais variantes como Delta, Omicron, Gamma e Alpha mostraram-se fortemente concentradas em grupos específicos. A variante Delta, por exemplo, apresentou grande dispersão em dois *clusters*, refletindo sua diversidade mutacional, enquanto a variante Omicron formou um grupo coeso e isolado, compatível com seu perfil genômico singular e altamente mutado. Essas observações estão em consonância com estudos apresentados em [16], [17], que destacam a complexidade estrutural da Omicron como determinante para sua evasão imune e segregação genômica. Da mesma forma, a predominância da Delta e sua sobreposição com outras regiões no espaço de projeção reforçam os estudos apresentados em [18], [19], que descrevem sua rápida disseminação global e variabilidade adaptativa. Assim, os agrupamentos obtidos nesta pesquisa não refletem apenas coerência interna, mas também correspondem a características biológicas descritas na literatura, evidenciando o potencial dos modelos baseados em *Transformers* para discriminar com precisão perfis mutacionais complexos sem a necessidade de alinhamento de sequência.

II. METODOLOGIA

A metodologia proposta baseia-se na transformação do formato de substituições de aminoácidos, *AASubstitutions*, presentes em sequências genômicas do SARS-CoV-2 em

representações vetoriais densas (*embeddings*), geradas por modelos baseados em arquitetura do tipo *Transformers*. Essa representação possibilita a aplicação de técnicas baseadas em aprendizado de máquina, análise vetorial e construção de redes de similaridade para fins de vigilância genômica. O processo metodológico é estruturado nas seguintes etapas: coleta e pré-processamento de dados, remoção de outliers associados as substituições de aminoácidos, geração dos *embeddings* e análise dos vetores resultantes.

A. Coleta e Pré-processamento dos Dados

A Figura 1 detalha as etapas iniciais do fluxo metodológico, abrangendo desde a coleta de dados até o processo de curadoria das substituições de aminoácidos. Os dados foram obtidos do repositório GISAID [2], a partir do conjunto de metadados coletados em 28 de março de 2024. O *dataset* original continha 16.630.814 sequências (ou amostras), contendo vários campos entre eles o campo sobre *AASubstitutions*, informação crucial para este trabalho.

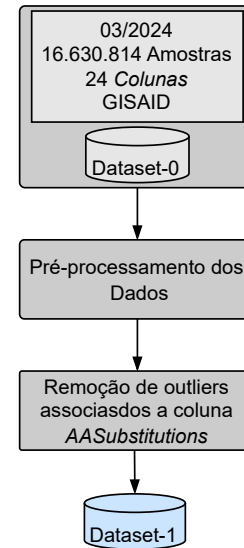


Figura 1: Etapas iniciais da metodologia, da coleta no GISAID à remoção de outliers das substituições de aminoácidos para formação do *Dataset-1*.

Com o objetivo de garantir a qualidade das análises genômicas, foi realizado o pré-processamento que envolveu a filtragem das sequências marcadas como completas e com alta cobertura. Sequência classificadas como completa diz respeito a genomas cuja extensão cobre integralmente a sequência de referência, enquanto a cobertura genômica indica que cada base foi sequenciada múltiplas vezes, o que reduz a taxa de erro e assegura maior confiabilidade nas inferências sobre mutações [20]. Além disso, foram selecionadas apenas amostras classificadas como Variantes de Preocupação (VOC). A Organização Mundial da Saúde (OMS) estabelece três categorias principais para as variantes do SARS-CoV-2: variantes

sob monitoramento (VUM), variantes de interesse (VOI), além da VOC. As VOCs são aquelas que apresentam evidências consistentes de aumento na transmissibilidade, na gravidade da doença ou impacto na eficácia de diagnósticos, vacinas e tratamentos, sendo, portanto, as mais relevantes do ponto de vista de saúde pública [21]. Após o pré-processamento, o *dataset* foi reduzido para 936.638 amostras.

Como mostrado na Figura 1, procedeu-se à exclusão de *outliers* com base na quantidade de substituições de aminoácidos por amostra. Foram removidas as amostras que apresentavam valores abaixo do percentil 0,5% ou acima do percentil 99,5%, com o objetivo de eliminar extremos que pudessem comprometer a qualidade das representações vetoriais. O conjunto final, aqui representado pelo *dataset-1*, foi utilizado na geração dos *embeddings* e é composto por 927.064 amostras. A Figura 2 apresenta a distribuição do número de substituições de aminoácidos por amostra antes e após a remoção dos *outliers*. Observa-se uma redução significativa na dispersão dos dados, promovendo maior homogeneidade entre as amostras mantidas. Inicialmente, o número de substituições variava de 1 a 3093, com desvio padrão de aproximadamente 21. Após a filtragem, os valores passaram a variar entre 22 e 71 substituições, com desvio padrão reduzido para cerca de 8,8. A média e a mediana ficaram em 35 e 33 substituições, respectivamente. Essa redução na variabilidade confere maior consistência às análises subsequentes, mitigando efeitos de ruído ou viés durante a modelagem vetorial.

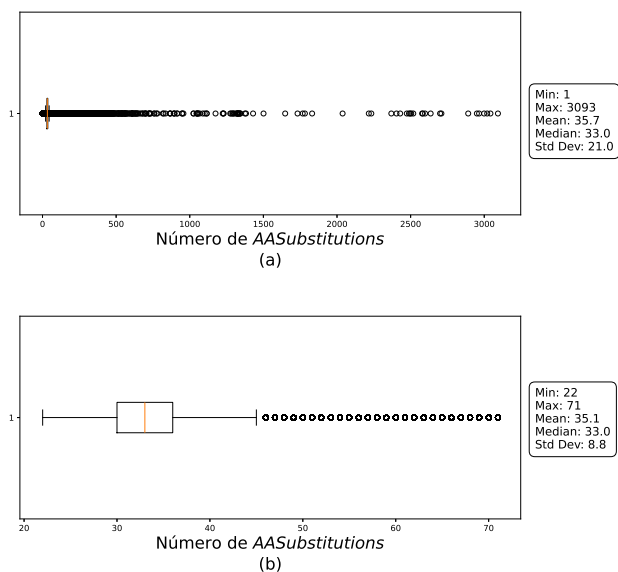


Figura 2: Representação da distribuição do numero de AA-Substitutions (a) com outliers e (b) sem outliers.

B. AA Substitution

A coluna *AA Substitutions* contém, para cada n -ésima amostra genômica, a lista de substituições de aminoácidos identificadas em relação à sequência de referência do SARS-CoV-2. Cada substituição é representada em uma notação padronizada que segue a estrutura *Gene:Substituição*, onde o

nome da proteína ou região genômica (por exemplo, *Spike*, *NSP12*, *ORF3a*) é seguido por um código de substituição no formato *X123Y*. Neste código, *X* representa o aminoácido original, *123* é a posição da mutação na cadeia da proteína, e *Y* representa o novo aminoácido substituído. Um exemplo clássico é *Spike:D614G*, que indica uma substituição do aminoácido ácido aspártico (D) por glicina (G) na posição 614 da proteína *Spike*.

Cada n -ésima amostra pode conter múltiplas substituições separadas por espaço ou vírgula (dependendo do formato exportado), sendo interpretadas como um conjunto de eventos mutacionais associados a uma única sequência viral. Essa estrutura permite tratar cada amostra como um “documento” composto por “palavras”, onde cada palavra representa uma substituição distinta. Essa analogia viabiliza a aplicação de técnicas PLN, como *embeddings* e *Transformers*, para gerar representações vetoriais densas das sequências com base na coocorrência e nos padrões semânticos entre as substituições. Para ilustrar a estrutura e diversidade dessas substituições, as Tabelas I e II apresentam exemplos concretos extraídos da coluna *AA Substitutions*.

A Tabela II mostra amostras representativas da variante Omicron provenientes de diferentes continentes. Nela, observa-se a recorrência de mutações como *Spike:N501Y*, *E484A*, *Q498R*, entre outras, frequentemente associadas ao escape imunológico e aumento da afinidade ao receptor ACE2 [22]. A diversidade das mutações entre regiões geográficas (por exemplo, presença de *NSP3:P1103S* na América do Sul ou *NSP13:R392C* na Ásia) destaca a capacidade do vírus de acumular adaptações regionais mesmo dentro da mesma linhagem.

A Tabela I, por sua vez, reúne amostras da variante Delta também estratificadas por continente. As substituições como *Spike:L452R*, *P681R*, *T478K* e *D614G* aparecem como uma das principais assinaturas da linhagem, refletindo sua alta transmissibilidade. No entanto, variantes geograficamente específicas também surgem, como *NSP3:I273V* na América do Sul e *NSP2:D511N* na África, evidenciando a plasticidade adaptativa da linhagem Delta, isto é, sua capacidade de acumular mutações específicas em diferentes regiões geográficas em resposta a pressões seletivas locais, favorecendo sua persistência e disseminação global [23].

Apesar da organização padronizada dos dados, não é trivial separar diretamente os padrões mutacionais de diferentes variantes com base apenas na inspeção dessas listas de substituições. A grande quantidade de mutações por amostra, a variabilidade na ordem e frequência das substituições, bem como a sobreposição de mutações entre variantes distintas, dificultam a extração de regularidades com métodos convencionais. No entanto, ao utilizar modelos de linguagem baseados em *Transformers*, torna-se possível capturar relações contextuais não lineares por meio de suas camadas de atenção [24].

Tabela I: Exemplo de amostras da variante Omicron extraídas do campo *AASubstitutions*.

Continente	<i>AASubstitutions</i>
Europa	(NSP5_P132H,Spike_H69del,Spike_T95I,Spike_A67V,Spike_S373P,Spike_Q493R,Spike_N969K,Spike_H655Y,Spike_N856K,N_R203K,Spike_G142D,NSP3_A1892T,Spike_Q954H,N_P13L,NSP3_L1266I,Spike_N501Y,N_R32del,M_Q19E,Spike_N440K,NSP4_T492I,Spike_N679K,Spike_N764K,Spike_L212I,Spike_Y505H,Spike_R346K,NSP6_I189V,Spike_T547K,M_D3G,Spike_D796Y,N_G204R,Spike_T478K,Spike_V143del,M_A63T,Spike_G496S,Spike_K417N,Spike_S371L,Spike_G339D,NSP3_S1265del,NSP14_I42V,Spike_P681H,Spike_Y144del,Spike_ins214EPE,N_S33del,Spike_S375F,Spike_Q498R,Spike_G446S,Spike_S477N,N_E31del,NSP3_K38R,Spike_N211del,Spike_E484A,E_T9I,Spike_V70del,Spike_L981F,NSP12_P323L,Spike_D614G,Spike_Y145del)
Africa	(NSP5_P132H,Spike_H69del,Spike_T95I,NSP6_A88V,Spike_A67V,NSP3_G489S,NSP4_T327I,NSP6_S106del,Spike_Q493R,Spike_N969K,Spike_H655Y,N_R203K,Spike_G142D,Spike_Q954H,N_P13L,Spike_N501Y,N_R32del,NS3_T223I,NSP14_I42V,Spike_P681H,Spike_Y144del,M_Q19E,N_S33del,NSP1_S135R,NSP4_T492I,Spike_Q498R,Spike_N679K,Spike_N764K,Spike_S477N,NSP6_G107del,Spike_Y505H,N_E31del,NSP6_F108del,Spike_D796Y,Spike_E484A,N_G204R,Spike_T478K,E_T9I,N_S413R,Spike_V70del,Spike_V143del,M_A63T,NSP12_P323L,Spike_Y145del,Spike_D614G)
Asia	(NSP5_P132H,NSP3_G489S,NSP4_T327I,Spike_S373P,Spike_Q493R,Spike_N969K,Spike_H655Y,N_R203K,Spike_V213G,Spike_G142D,Spike_Q954H,N_P13L,Spike_N501Y,NS3_T223I,Spike_T19I,M_Q19E,Spike_N440K,NSP4_T492I,Spike_N679K,Spike_N764K,Spike_Y505H,Spike_D796Y,N_G204R,Spike_T478K,N_S413R,M_A63T,Spike_S371F,NSP13_R392C,Spike_K417N,Spike_T376A,Spike_G339D,Spike_R408S,NSP4_L438F,NSP14_I42V,Spike_P681H,NSP4_L264F,NSP3_T24I,NSP1_S135R,Spike_S375F,Spike_D405N,Spike_Q498R,Spike_S477N,NSP15_T112I,Spike_E484A,E_T9I,NSP12_P323L,Spike_D614G)
America do Norte	(NSP5_P132H,Spike_H69del,Spike_T95I,Spike_A67V,NS3_S165F,Spike_S373P,Spike_Q493R,Spike_N969K,Spike_H655Y,Spike_N856K,N_R203K,Spike_G142D,NSP3_A1892T,Spike_Q954H,NS3_L106F,N_P13L,Spike_N501Y,N_R32del,M_Q19E,Spike_N440K,NSP4_T492I,NSP6_L105del,Spike_N679K,Spike_N764K,Spike_L212I,NSP6_G107del,Spike_Y505H,NSP6_I189V,Spike_T547K,M_D3G,Spike_D796Y,N_G204R,N_D343G,Spike_T478K,Spike_V143del,M_A63T,Spike_G496S,Spike_K417N,NSP6_S106del,Spike_S371L,Spike_G339D,NSP3_S1265del,NSP14_I42V,Spike_P681H,Spike_Y144del,Spike_ins214EPE,N_S33del,Spike_S375F,NSP2_A385V,Spike_Q498R,Spike_G446S,Spike_S477N,N_E31del,NSP3_K38R,Spike_N211del,Spike_E484A,E_T9I,Spike_V70del,Spike_L981F,NSP12_P323L,Spike_D614G,Spike_Y145del)
America do Sul	(NSP5_P132H,Spike_T95I,Spike_S373P,Spike_Q493R,Spike_N969K,Spike_H655Y,Spike_N856K,N_R203K,Spike_G142D,NSP3_A1892T,NSP3_P1103S,Spike_Q954H,N_P13L,NSP3_L1266I,Spike_N501Y,M_Q19E,Spike_N440K,NSP4_T492I,Spike_N679K,Spike_N764K,Spike_Y505H,NSP6_I189V,Spike_T547K,M_D3G,Spike_D796Y,N_G204R,Spike_T478K,Spike_V143del,M_A63T,Spike_G496S,Spike_ins215PEX,Spike_K417N,Spike_S371L,Spike_G339D,NSP3_S1265del,NSP14_I42V,Spike_P681H,Spike_Y144del,Spike_S375F,Spike_Q498R,Spike_G446S,Spike_D215E,Spike_S477N,NSP3_K38R,Spike_E484A,E_T9I,Spike_L981F,NSP12_P323L,Spike_D614G,Spike_Y145del)

C. Geração de Embeddings

Cada n -ésima amostra do *dataset-1* (ou “documento”) pode ser representada por N_n palavras que representam substituições de aminoácidos, como descrito na seção acima. Assim, cada “documento” pode ser expresso como

$$\mathbf{s}_n = [s_{n,1} \quad \dots \quad s_{n,i} \quad \dots \quad s_{n,N_n}] \quad (1)$$

onde $s_{n,i}$ representa a i -ésima substituição (ou palavra) da n -ésima amostra e N_n é o número de substituições para cada n -ésima amostra viral. Para converter as substituições das amostras virais em vetores numéricos, foi utilizado o modelo pré-treinado all-MiniLM-L6-v2 em cada n -ésima amostra $\mathbf{s}_n$ gerando um novo vetor denso de 384 dimensões que pode ser expresso como

$$\mathbf{e}_n = [e_{n,1} \quad e_{n,2} \quad \dots \quad e_{n,384}] \quad (2)$$

onde $\mathbf{e}_n$ representa o *embedding* da n -ésima amostra. É importante destacar que, conforme ilustrado na Figura 3 foi criado um subconjunto chamado aqui de *dataset-2* a partir do *dataset-1*, contendo exclusivamente a coluna *AASubstitutions*. Em seguida, foram removidas todas as entradas duplicadas com base nessa informação, resultando em um total de 550.088 amostras distintas. Dessa forma, o modelo all-MiniLM-L6-v2 foi utilizado para gerar *embeddings* únicos, o que permitiu preservar a diversidade das sequências virais representadas, ao mesmo tempo em que otimizou o uso de recursos computacionais, evitando o processamento redundante de amostras com perfis mutacionais idênticos.

D. Análises Baseadas em Embeddings

Conforme ilustrado na Figura 3, após a etapa de geração dos *embeddings* por meio do modelo de *Transfor-*

Tabela II: Exemplo de amostras da variante Delta extraídas do campo *AASubstitutions*.

Continente	<i>AASubstitutions</i>
Europa	(N_G215C,N_D63G,N_R203M,NSP12_G671S,Spike_G142D,NS3_S26L,NSP3_R1341Q,Spike_P681R,Spike_R158del,NSP3_P1228L,NS6_K42N,NS7a_V82A,NSP3_A488S,Spike_F157del,NSP4_T492I,NSP14_A394V,Spike_T19R,NS7a_T120I,N_D377Y,M_I82T,Spike_D950N,NS7b_T40I,NSP13_P77L,Spike_E156G,NSP3_P1469S,Spike_T478K,NSP6_T77A,NSP12_K807E,NSP6_A2V,NSP12_P323L,Spike_D614G,NSP4_V167L,Spike_L452R)
Africa	(N_G215C,N_D63G,N_R203M,NSP12_G671S,Spike_G142D,NS3_S26L,Spike_P681R,Spike_R158del,NSP3_P1228L,NS7a_V82A,NSP2_D511N,NSP3_A488S,Spike_F157del,NSP4_T492I,NSP14_A394V,Spike_T19R,NS7a_T120I,N_D377Y,M_I82T,Spike_D950N,NS7b_T40I,NSP13_P77L,Spike_E156G,NSP3_P1469S,Spike_T478K,NSP6_T77A,NS8_G66S,NSP12_P323L,Spike_D614G,NSP4_V167L,Spike_L452R,NSP13_E365D)
Asia	(N_G215C,Spike_T951I,N_D63G,N_R203M,NSP12_G671S,NSP15_Q19H,NS3_S26L,NS3_T221K,Spike_P681R,Spike_R158del,NSP3_P1228L,Spike_D950B,NS7a_V82A,NSP3_A488S,Spike_F157del,NSP4_T492I,Spike_T19R,NSP14_A394V,NS7a_T120I,N_D377Y,M_I82T,NSP3_A231V,NS7b_T40I,NSP13_P77L,Spike_E156G,NSP14_N67S,NSP3_P1469S,Spike_T478K,NSP6_T77A,NSP12_P323L,NS3_G49V,Spike_D614G,NSP4_V167L,Spike_L452R,Spike_P26L)
América do Norte	(N_G215C,N_D63G,N_R203M,NSP12_G671S,NSP13_I334V,Spike_G142D,NS3_S26L,Spike_P681R,Spike_R158del,NSP3_P1159L,NSP3_P1228L,NS7a_V82A,NSP3_A488S,Spike_F157del,NS8_P36S,NSP4_T492I,NSP14_A394V,Spike_T19R,NS7a_T120I,N_D377Y,M_I82T,Spike_D950N,NS7b_T40I,NSP13_P77L,Spike_E156G,NSP3_P1469S,NSP3_S1877N,Spike_T478K,NSP6_T77A,NSP12_P323L,Spike_D614G,NSP4_V167L,Spike_L452R)
América do Sul	(N_G215C,N_D63G,N_R203M,NSP8_T145I,NSP12_G671S,NS3_S26L,Spike_P681R,Spike_R158del,NSP3_P1228L,NS7a_V82A,NSP3_A488S,Spike_F157del,NSP4_T492I,NSP14_A394V,Spike_T19R,NS7a_T120I,N_D377Y,M_I82T,Spike_D950N,NS7b_T40I,NSP13_P77L,Spike_E156G,NSP3_P1469S,Spike_T478K,NSP6_T77A,NSP12_P323L,Spike_D614G,Spike_L452R,NSP4_V167L,NSP3_I273V)

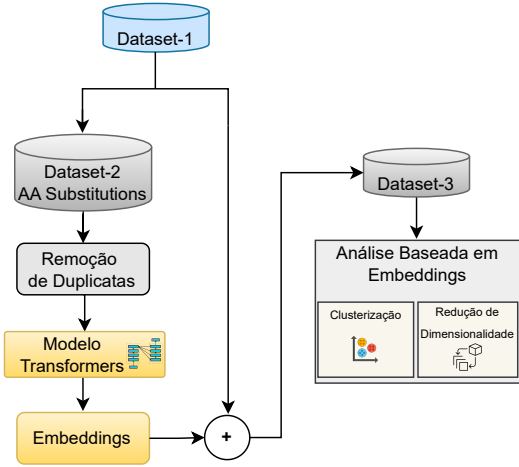


Figura 3: Etapas finais da metodologia, com integração dos *embeddings* ao *dataset-1* e geração do *dataset-3* para análise vetorial.

mer all-MiniLM-L6-v2, as representações vetoriais das substituições de aminoácidos foram incorporadas ao conjunto de dados original (*Dataset-1*), resultando em um novo conjunto denominado *Dataset-3*. Este novo *dataset* agrega os metadados epidemiológicos e clínicos, e também os veto-

res contextuais que capturam relações semânticas entre as mutações de cada amostra viral. Com o objetivo de investigar padrões estruturais entre as amostras virais representadas no espaço vetorial, foram aplicadas duas técnicas amplamente reconhecidas na literatura: o algoritmo de agrupamento *K-means* para realizar os agrupamentos (clustering) e a técnica de redução de dimensionalidade *t-distributed Stochastic Neighbor Embedding* (t-SNE) para visualizar os agrupamentos.

Conforme ilustrado na Figura 3, após a geração dos *embeddings* por meio do modelo *Transformer* all-MiniLM-L6-v2, as representações vetoriais das substituições de aminoácidos foram integradas ao conjunto de dados original (*Dataset-1*), originando um novo conjunto denominado *Dataset-3*. Esse novo *dataset* agrega, de forma sinérgica, os metadados clínicos e epidemiológicos às representações contextuais das mutações, que capturam relações semânticas entre as alterações observadas em cada amostra viral. A fim de investigar padrões estruturais latentes no espaço vetorial, aplicaram-se duas técnicas amplamente consolidadas na literatura: o algoritmo de agrupamento *K-means*, para a identificação de agrupamentos, e a técnica de redução de dimensionalidade *t-distributed Stochastic Neighbor Embedding* (t-SNE), utilizada para projeção e visualização das amostras em duas dimensões, permitindo a interpretação da organização dos grupos formados de maneira clara e intuitiva [25].

Neste trabalho, o *K-means*, foi aplicado diretamente sobre os *embeddings* gerados, previamente normalizados por meio

da padronização Z-score. Foram testados valores de K variando de 4 a 10 *clusters*, e fundamentada pela análise do método do cotovelo e do índice de *Silhouette* foi definido que a melhor configuração seria com quatro grupos. Essa configuração foi posteriormente validada visualmente por meio da projeção *t*-SNE. A técnica *t*-SNE foi empregada para viabilizar a visualização dos dados de alta dimensionalidade (384 dimensões) em um espaço bidimensional [26]. Para garantir que a métrica de comparação entre os *embeddings* refletisse semelhanças estruturais relevantes, foi adotada a normalização L2 e a distância do cosseno como métrica. A projeção obtida com *t*-SNE foi importante para revelar a separabilidade dos *clusters*, identificar sobreposições e visualizar a concentração de amostras por variante.

III. RESULTADOS

A metodologia baseada em *embeddings* de substituições de aminoácidos, demonstrou-se eficaz na identificação de padrões estruturais relevantes entre variantes do SARS-CoV-2. A projeção bidimensional via *t*-SNE permitiu visualizar esses agrupamentos de forma clara e interpretável. Conforme ilustrado na Figura 4, cada *cluster* ocupa uma região distinta no espaço vetorial, indicando que os perfis mutacionais codificados pelos *embeddings* resultaram em separações consistentes entre grupos.

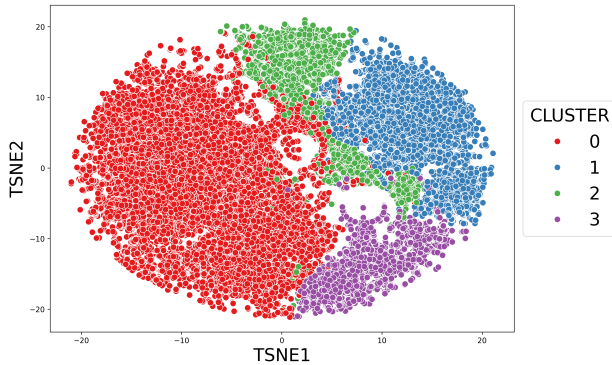


Figura 4: Projeção *t*-SNE dos *embeddings* com destaque para os *clusters* identificados.

A Figura 5 apresenta a distribuição das variantes virais sobre a mesma projeção *t*-SNE utilizada para visualização dos *clusters*. Nota-se uma correspondência direta entre regiões ocupadas por determinados grupos e as variantes associadas às amostras, indicando que o modelo de representação vetorial utilizado é sensível às diferenças genéticas entre linhagens virais. Essa tendência é confirmada através da Figura 6, que ilustra por meio de gráfico de barras a distribuição das variantes em cada *cluster*. A Tabela III complementa a Figura 6 trazendo um resumo dos valores absolutos e percentuais do número de amostras de cada variante dentro de cada *cluster*.

Os dados revelam que o *cluster* 0 é composto quase exclusivamente por amostras da variante Delta, que responde por 99,79% das observações desse grupo, evidenciando uma forte

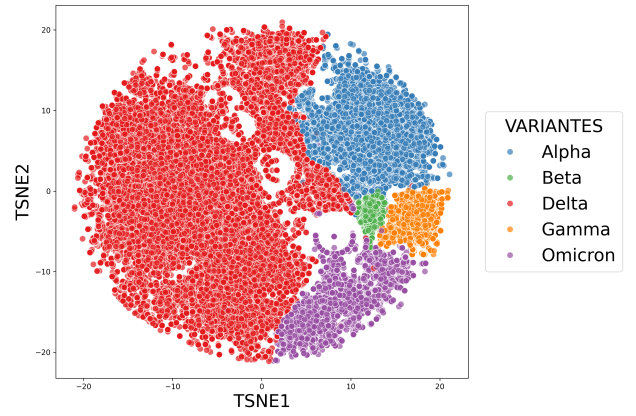


Figura 5: Projeção *t*-SNE destacando a distribuição das variantes virais.

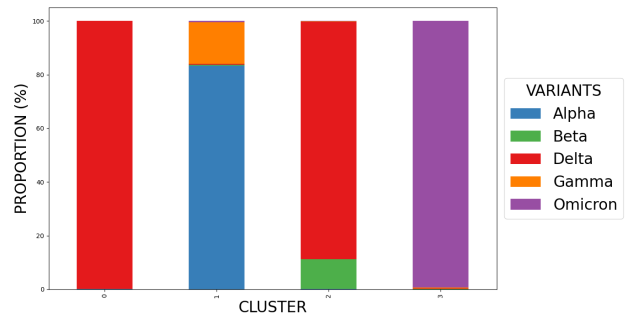


Figura 6: Frequência relativa das variantes em cada *cluster* identificado.

homogeneidade interna. Já o *cluster* 1 apresenta um perfil diferente, sendo dominado pelas variantes Alpha, que representa 83,32% das amostras, e Gamma, com 15,61%, o que indica um agrupamento característico de variantes com perfis mais semelhantes entre si, mas ainda distintos da Delta. O *cluster* 2 também é amplamente formado por amostras da variante Delta, com 88,58%, mas traz uma contribuição importante da variante Beta, que representa 11,02% das amostras, sugerindo possíveis semelhanças estruturais entre essas duas linhagens no espaço de *embeddings*. Por fim, o *cluster* 3 mostra-se fortemente caracterizado pela variante Omicron, que responde por mais de 99,0% das amostras nesse grupo, demonstrando um padrão altamente específico dessa linhagem em relação às demais.

IV. ANÁLISE DOS RESULTADOS

A análise dos agrupamentos obtidos por meio dos *embeddings* das substituições de aminoácidos revelou padrões consistentes com a biologia molecular das variantes virais do SARS-CoV-2. A formação de *clusters* com alta homogeneidade interna e boa separação intergrupo, especialmente para as variantes Delta e Omicron, reforça a capacidade do modelo baseado em *Transformers* de capturar representações semânticas com relevância biológica.

Tabela III: Distribuição absoluta e percentual das variantes do SARS-CoV-2 por *cluster* identificado.

Clusters	Número de amostras				
	Alpha	Beta	Delta	Gamma	Omicron
0	955 (0,17%)	33 (0,006%)	550,665 (99,79%)	10 (0,0018%)	178 (0,032%)
1	137.028 (83,32%)	631 (0,38%)	469 (0,28%)	25.694 (15,61%)	741 (0,45%)
2	181 (0,18%)	11.128 (11,02%)	89.425 (88,58%)	77 (0,076%)	154 (0,15%)
3	46 (0,042%)	139 (0,127%)	111 (0,101%)	456 (0,416%)	108.943 (99,31%)

A predominância da variante Delta nos clusters 0 e 2 está associada à sua assinatura mutacional, que inclui mutações como L452R e P681R. Estas alterações foram descritas como críticas para o aumento da afinidade de ligação ao receptor ACE2 e para o escape parcial à neutralização por anticorpos. Trabalhos como os de [18] demonstraram que essas mutações contribuíram para a rápida disseminação da Delta em países como a Índia, durante a segunda onda da pandemia.

A variante Alpha, por sua vez, mostrou-se concentrada no cluster 1, comportamento coerente com sua assinatura mutacional característica, incluindo N501Y, P681H e del69-70, que conferem maior infectividade e capacidade de evasão imunológica. Estudos como os de [27] destacam que essas mutações contribuíram para a dominância da Alpha no Reino Unido durante o início de 2021, especialmente em ambientes escolares.

A variante Gamma, identificada no Brasil, apresenta mutações como K417T, E484K e N501Y no domínio de ligação ao receptor (RBD), aumentando sua afinidade pelo ACE2 e sua evasão de anticorpos neutralizantes [28]. A homogeneidade dos *embeddings* gerados para essa linhagem, observada na clusterização, indica que o modelo capturou adequadamente o padrão estrutural e funcional associado a esse perfil mutacional.

A variante Beta, embora com menor expressão nos dados, concentrou-se parcialmente no cluster 2 e compartilha com Gamma e Alpha a presença da mutação E484K, o que pode explicar sua proximidade vetorial com essas linhagens. A convergência de mutações no RBD entre variantes distintas sugere a atuação de pressões seletivas comuns ao longo da evolução viral, o que também foi destacado por [29].

O Cluster 3 concentra quase exclusivamente amostras da variante Ômicron, como já mencionado. Essa variante, identificada inicialmente em novembro de 2021 na África do Sul, se destaca por um número excepcionalmente alto de mutações na proteína spike, totalizando mais de 30 alterações. Dentre essas, diversas ocorrem no domínio de ligação ao receptor (RBD), como K417N, H655Y, E484A, N969K, Q498R, N501Y e Y505H, que possuem implicações significativas tanto na transmissibilidade quanto na capacidade de escape imunológico do vírus, como relatado por [30].

Os resultados obtidos confirmam que os *embeddings* derivados de substituições de aminoácidos permitem a distinção eficaz entre variantes virais, mesmo na ausência de alinhamento de sequências. A estrutura latente construída se mostrou coerente com os conhecimentos atuais da biologia do SARS-CoV-2, como demonstrado por comparações com estudos filogenéticos e epidemiológicos prévios. Essa convergência

entre evidência empírica e literatura reforça a robustez da metodologia adotada e abre caminho para aplicações mais amplas em vigilância genômica assistida por inteligência artificial.

V. CONCLUSÃO

Este estudo propôs uma abordagem baseada em representações vetoriais para análise de substituições de aminoácidos, *AASubstitutions*, em amostras virais do SARS-CoV-2, com o objetivo de identificar agrupamentos consistentes de variantes sem a necessidade de alinhamento de sequências. A aplicação de modelos pré-treinados do tipo *Transformer*, neste trabalho sendo utilizado o modelo *all-MiniLM-L6-v2*, possibilitou a conversão de perfis mutacionais em *embeddings* de alta dimensão, que foram posteriormente analisados por técnicas de aprendizado não supervisionado, como *K-means* e *t-SNE*.

Os resultados demonstraram que os *embeddings* gerados preservam informações estruturais e funcionais relevantes, sendo capazes de diferenciar eficazmente as principais variantes de preocupação. A distribuição das amostras nos *clusters* revelou padrões coerentes com as assinaturas mutacionais conhecidas de variantes como Alpha, Beta, Gamma, Delta e Omicron, destacando a sensibilidade do modelo às características genéticas específicas de cada linhagem. Em especial, a variante Omicron apresentou forte segregação no espaço vetorial, refletindo seu extenso conjunto de mutações na proteína Spike.

A metodologia proposta demonstrou ser uma alternativa robusta e escalável para vigilância genômica baseada em dados de mutações, com potencial aplicação em sistemas de monitoramento em tempo real, apoio à epidemiologia molecular e integração em arquiteturas do tipo *Retrieval-Augmented Generation* (RAG). Ao dispensar etapas custosas de alinhamento e explorar diretamente o padrão das substituições, o método amplia as possibilidades de análise genômica de larga escala, especialmente em contextos com grandes volumes de dados heterogêneos. Como trabalhos futuros, pretende-se incorporar metadados clínicos e epidemiológicos às análises, expandir o conjunto de variantes incluídas, e explorar o uso dos *embeddings* gerados como base vetorial para sistemas de busca semântica e recomendação em vigilância viral assistida por inteligência artificial.

AGRADECIMENTOS

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte e financiamento.

REFERÊNCIAS

- [1] L. Ran, X. Tan, and Y. Zhang, "Precise community-based public health management: Crucial experience responding to covid-19 in wuhan, china," *Risk Management and Healthcare Policy*, pp. 171–178, 2022.
- [2] S. Khare, C. Gurry, L. Freitas, M. B. Schultz, G. Bach, A. Diallo, N. Akite, J. Ho, R. T. Lee, N. K. Yeo, C. Team, and S. Maurer-Stroh, "Gisaid's role in pandemic response," *China CDC Weekly*, vol. 3, no. 49, pp. 1049–1051, 2021. [Online]. Available: <https://doi.org/10.46234/ccdcw2021.255>
- [3] M. Zvyagin, A. Brace, K. Hippe, Y. Deng, B. Zhang, C. Orozco Bohorquez, A. Clyde, B. Kale, D. Perez-Rivera, H. Ma *et al.*, "Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*. ACM, 2022.
- [4] R. Singh, S. Nagpal, N. K. Pinna, and S. S. Mande, "Tracking mutational semantics of sars-cov-2 genomes," *Scientific Reports*, vol. 12, no. 1, p. 15704, 2022. [Online]. Available: <https://doi.org/10.1038/s41598-022-20000-5>
- [5] B. A. Sokhansanj, Z. Zhao, and G. L. Rosen, "Interpretable and predictive deep neural network modeling of the sars-cov-2 spike protein sequence to predict covid-19 disease severity," *Biology*, vol. 11, no. 12, 2022. [Online]. Available: <https://www.mdpi.com/2079-7737/11/12/1786>
- [6] B. Zhou, H. Zhou, X. Zhang, X. Xu, Y. Chai, Z. Zheng, A. C. Kot, and Z. Zhou, "Tempo: A transformer-based mutation prediction framework for sars-cov-2 evolution," *Computers in Biology and Medicine*, vol. 152, p. 106264, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0010482522009726>
- [7] S. Nagpal, R. Pal, Ashima, A. Tyagi, S. Tripathi, A. Nagori, S. Ahmad, H. P. Mishra, R. Malhotra, R. Kutum, and T. Sethi, "Genomic surveillance of covid-19 variants with language models and machine learning," *Frontiers in Genetics*, vol. 13, p. 858252, 2022.
- [8] A. Qayyum, A. Benzinou, O. Saidani, F. Alhayan, M. A. Khan, A. Masood, and M. Mazher, "Assessment and classification of covid-19 dna sequence using pairwise features concatenation from multi-transformer and deep features with machine learning models," *SLAS Technology*, vol. 29, p. 100147, 2024.
- [9] A. M. Mutawa, "Attention-based hybrid deep learning models for classifying covid-19 genome sequences," *AI*, vol. 6, no. 1, p. 4, 2025.
- [10] G. B. M. Câmara, M. G. F. Coutinho, L. M. D. d. Silva, W. V. D. N. Gadelha, M. F. Torquato, R. d. M. Barbosa, and M. A. C. Fernandes, "Convolutional neural network applied to sars-cov-2 sequence classification," *Sensors*, vol. 22, no. 15, p. 5730, 2022. [Online]. Available: <https://www.mdpi.com/1424-8220/22/15/5730>
- [11] K. S. Azevedo, L. C. de Souza, M. G. F. Coutinho, R. d. M. Barbosa, and M. A. C. Fernandes, "Deepvirusclassifier: a deep learning tool for classifying sars-cov-2 based on viral subtypes within the coronavirusidae family," *BMC Bioinformatics*, vol. 25, no. 231, 2024. [Online]. Available: <https://doi.org/10.1186/s12859-024-05754-1>
- [12] L. C. de Souza, K. S. Azevedo, J. G. de Souza, R. d. M. Barbosa, and M. A. C. Fernandes, "New proposal of viral genome representation applied in the classification of sars-cov-2 with deep learning," *BMC Bioinformatics*, vol. 24, no. 92, 2023. [Online]. Available: <https://doi.org/10.1186/s12859-023-05188-1>
- [13] M. G. F. Coutinho, G. B. M. Câmara, R. d. M. Barbosa, and M. A. C. Fernandes, "Sars-cov-2 virus classification based on stacked sparse autoencoder," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 284–298, 2023. [Online]. Available: <https://doi.org/10.1016/j.csbj.2022.12.007>
- [14] P. K. Vesapogu and B. R. Surampudi, "A comparative study of machine learning approaches for the detection of sars-cov-2 and its variants," in *Procedia Computer Science*, vol. 235. Elsevier, 2024, pp. 1190–1201, presented at the International Conference on Machine Learning and Data Engineering (ICMLDE 2023). [Online]. Available: <https://doi.org/10.1016/j.procs.2024.04.113>
- [15] Hugging Face, "all-minilm-l6-v2," 2025, accessed: 2025-04-20. [Online]. Available: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>
- [16] M. Bhattacharya, A. R. Sharma, K. Dhama, G. Agoramoorthy, and C. Chakraborty, "Omicron variant (b.1.1.529) of sars-cov-2: understanding mutations in the genome, s-glycoprotein, and antibody-binding regions," *GeroScience*, vol. 44, no. 2, pp. 619–637, 2022. [Online]. Available: <https://doi.org/10.1007/s11357-022-00532-4>
- [17] Y. Fan, X. Li, L. Zhang, S. Wan, L. Zhang, and F. Zhou, "Sars-cov-2 omicron variant: recent progress and future perspectives," *Signal Transduction and Targeted Therapy*, vol. 7, no. 1, p. 141, 2022. [Online]. Available: <https://doi.org/10.1038/s41392-022-00997-x>
- [18] S. Cheria, V. Potdar, S. Jadhav, P. Yadav, N. Gupta, M. Das, P. Rakshit, S. Singh, P. Abraham, and S. Panda, "Sars-cov-2 spike mutations, l452r, t478k, e484q and p681r, in the second wave of covid-19 in maharashtra, india," *Microorganisms*, vol. 9, no. 7, p. 1542, 2021. [Online]. Available: <https://www.mdpi.com/2076-2607/9/7/1542>
- [19] D. Tian, Y. Sun, J. Zhou, and Q. Ye, "The global epidemic of the sars-cov-2 delta variant, key spike mutations and immune escape," *Frontiers in Immunology*, vol. Volume 12 - 2021, 2021.
- [20] D. Sims, I. Sudbery, N. E. Illott, A. Heger, and C. P. Ponting, "Sequencing depth and coverage: key considerations in genomic analyses," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 121–132, Feb. 2014. [Online]. Available: <https://doi.org/10.1038/nrg3642>
- [21] World Health Organization. (2025) Tracking sars-cov-2 variants. World Health Organization. Accessed: 2025-04-02. [Online]. Available: <https://www.who.int/activities/tracking-SARS-CoV-2-variants>
- [22] Y. Cao, J. Wang, F. Jian, T. Xiao, W. Song, A. Yisimayi, W. Huang, Q. Li, P. Wang, R. An, J. Wang, Y. Wang, X. Niu, S. Yang, H. Liang, H. Sun, T. Li, Y. Yu, Q. Cui, S. Liu, X. Yang, S. Du, Z. Zhang, X. Hao, F. Shao, R. Jin, X. Wang, J. Xiao, Y. Wang, and X. S. Xie, "Omicron escapes the majority of existing sars-cov-2 neutralizing antibodies," *Nature*, vol. 602, no. 7898, pp. 657–663, 2022, published February 2022. [Online]. Available: <https://doi.org/10.1038/s41586-021-04385-3>
- [23] Y. Zhan, H. Yin, and J.-Y. Yin, "B.1.617.2 (delta) variant of sars-cov-2: features, transmission and potential strategies," *International Journal of Biological Sciences*, vol. 18, no. 5, pp. 1844–1851, 2022, published online 2022-02-14, eCollection 2022. [Online]. Available: <https://doi.org/10.7150/ijbs.66881>
- [24] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives, "Msa transformer," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8844–8856. [Online]. Available: <https://proceedings.mlr.press/v139/rao21a.html>
- [25] A. Rykov, R. C. De Amorim, V. Makarenkov, and B. Mirkin, "Inertia-based indices to determine the number of clusters in k-means: An experimental evaluation," *IEEE Access*, vol. 12, pp. 11761–11773, 2024. [Online]. Available: <https://doi.org/10.1109/ACCESS.2024.3350791>
- [26] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: <http://jmlr.org/papers/v9/vandermaaten08a.html>
- [27] S. N. Ladhani, G. Ireland, F. Baawuah, J. Beckmann, I. O. Okike, S. Ahmad, J. Garstang, A. J. Brent, B. Brent, F. Aiano, Z. Amin-Chowdhury, M. Kall, R. Borrow, E. Linley, M. Zambon, J. Poh, L. Warrenner, A. Lackenby, J. Ellis, G. Amirthalingam, K. E. Brown, and M. E. Ramsay, "Emergence of sars-cov-2 alpha (b.1.1.7) variant, infection rates, antibody seroconversion and seroprevalence rates in secondary school students and staff: Active prospective surveillance, december 2020 to march 2021, england," *Journal of Infection*, vol. 83, no. 5, pp. 573–580, 2021. [Online]. Available: <https://doi.org/10.1016/j.jinf.2021.08.019>
- [28] N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. d. S. Candido, S. Mishra, M. A. E. Crispim, F. C. Sales, I. Hawryluk, J. T. McCrone *et al.*, "Genomics and epidemiology of the p.1 sars-cov-2 lineage in manaus, brazil," *Science*, vol. 372, no. 6544, pp. 815–821, 2021. [Online]. Available: <https://www.science.org/doi/10.1126/science.abb2644>
- [29] H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, J. E. San, N. Msomi *et al.*, "Detection of a sars-cov-2 variant of concern in south africa," *Nature*, vol. 592, no. 7854, pp. 438–443, 2021. [Online]. Available: <https://www.nature.com/articles/s41586-021-03402-9>
- [30] M. Bhattacharya, A. R. Sharma, K. Dhama, G. Agoramoorthy, and C. Chakraborty, "Omicron variant (b.1.1.529) of sars-cov-2: understanding mutations in the genome, s-glycoprotein, and antibody-binding regions," *GeroScience*, vol. 44, no. 2, pp. 619–637, 2022. [Online]. Available: <https://doi.org/10.1007/s11357-022-00532-4>