UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

ANA CAROLINA DAMASCENO SANCHES

Aprendizagem de Máquinas na Identificação de Resíduos-Chave de Variantes na Interação das Proteínas ACE2 com Spike do SARS-CoV-2

RIBEIRÃO PRETO - SP, BRASIL
2022

Este trabalho foi apresentado e aprovado pela Comissão Coordenadora do Curso deInformática Biomédica em 15/03/2023.

ANA CAROLINA DAMASCENO SANCHES

Aprendizagem de Máquinas na Identificação de Resíduos-Chave de Variantes na Interação das Proteínas ACE2 com Spike do SARS-CoV-2

Monografia apresentada à Faculdade de Medicina de Ribeirão Preto como parte dos requisitos para conclusão do curso de graduação em Informática Biomédica.

Orientador(a): Prof.ª Dr.ª Silvana Giuliatti
Coorientador(a): Me. Levy Bueno Alves

RIBEIRÃO PRETO - SP, BRASIL
2022

AGRADECIMENTOS

Ao Serviço de Tecnologia da Informação (STI) da USP agradeço pelo uso dos clusters, que foi fundamental para o desenvolvimento desse projeto.

O presente trabalho foi realizado com apoio do processo nº 2022/02782-5, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade do(s) autor(es) e não necessariamente refletem a visão da FAPESP.

RESUMO

Desde de 2020 a população mundial vem enfrentando a pandemia de COVID-19, doença infecciosa causada pela Síndrome Respiratória Aguda Grave do Coronavírus-2 (SARS-CoV-2), que, até o momento, causou mais de 6,6 milhões de vítimas. O SARS-CoV-2 possui genoma de RNA, o que o torna mais propenso a mutações que um vírus de DNA, fato observável pelo surgimento de diferentes variantes. Tais mutações podem implicar no aumento da afinidade do vírus com o receptor ACE2, melhorar a suscetibilidade à infecção, impactar no diagnóstico e, não só reduzir a eficácia de tratamentos como também reduzir a neutralização por anticorpos. Além disso, características do hospedeiro, como fatores de risco e genéticos, também podem afetar a suscetibilidade e desenvolvimento da doença, e como a ACE2 possui interação direta com a proteína S, polimorfismos presentes nessa podem estar associados com a suscetibilidade ao SARS-CoV-2. Entretanto, como essas mutações e polimorfismos, encontrados em diferentes populações, contribuem para melhorar a estabilidade e afinidade de interação entre o complexo SARS-CoV-2-ACE2 não é totalmente compreendido. Métodos comumente usados para compreender este processo, como Dinâmicas Moleculares, geram grandes quantidades de dados, consequentemente interpretar e extrair informações desses métodos não é um processo simples. Métodos de aprendizado de máquinas são usados em análises de grande quantidade de dados, pois reduzem a dimensionalidade do problema, assim, auxiliam na identificação de diferenças significantes entre as trajetórias resultantes de dinâmicas moleculares. Portanto, propôs-se usar os classificadores Multilayer Perceptron e Random Forest a fim de identificar regiões ou resíduos críticos na interação entre variantes do vírus e hospedeiro que possam impactar significativamente a funcionalidade das proteínas ACE2 do hospedeiro e Spike do SARS-CoV-2. Ainda que estes classificadores sofreram overfitting, esses foram capazes de identificar pares contendo resíduos chaves na interação entre proteínas. Contudo, é necessário um melhor ajuste dos parâmetros e avaliação estrutural dos pares identificados.

Palavras-chave: COVID-19. Bioinformática estrutural. Simulações moleculares.

ABSTRACT

Since 2020, the world's population has been facing the COVID-19 pandemic, an infectious desease caused by Severe Acute Respiratory Syndrome coronavirus-2 (SARS-CoV-2), which, to the present day, caused more than 6,6 million victims. The SARS-CoV-2 has a RNA genome, which makes it more prone to mutations than a DNA virus, notable by the different variants that emerged. Such mutations may lead to an increase of binding affinity of the virus and its receptor ACE2, improving infection susceptibility, impact diagnosis, and not only reduce the effectiveness of treatments but also reduce the antibody neutralization. Futhermore, host characteristics, like risk and genetic factors, may also affect susceptibility and the desease development, and as ACE2 has a direct interaction with the S protein, polymorphisms in this protein may be associated with SARS-CoV-2 susceptibility. However, how this mutation and polymorphisms, found in different populations, contribute to improve stability and affinity between the complex SARS-CoV-2-ACE2 is not fully understood. Methods commonly used to understand this process, for exemple Molecular Dinamics, generate large amounts of data, consequently interpreting and extracting information from these methods is not a simple process. Machine learning methods are used in analyzes of big data, because it reduces the dimensionality of the problem, thus, it helps to identify significant differences between the trajectories resulting from molecular dynamics. Therefore, it was proposed the use of the classifiers Multilayer Perceptron and Random Forest in order to identify critical regions or residues between virus and host variants that could significantly impact the functionality of the host ACE2 and Spike of SARS-CoV-2 proteins. Although these classifiers were overfitted, these were able to identify pairs containing key residues in the interaction between proteins. However, a better adjustment of the parameters and structural evaluation of the identified pairs is necessary.

Keywords: COVID-19. Structural bioinformatics. Molecular simulations.

LISTA DE QUADROS

SUMÁRIO

1. INTRODUÇÃO

1.1. COVID-19

Em 11 de março de 2020, a Organização Mundial da Saúde (OMS) caracterizou como pandemia a COVID-19 (WHO DIRECTOR GENERAL'S SPEECHES, 2020), doença infecciosa causada pela Síndrome Respiratória Aguda Grave do Coronavírus-2 (do inglês *Severe Acute Respiratory Syndrome coronavirus-2 - SARS-CoV-2*). Até o presente momento, novembro de 2022, já foram confirmados mais de 630 milhões de casos confirmados, incluindo mais de 6,6 milhões de mortes globalmente. Só no Brasil, são mais de 35 milhões de casos com quase 690 mil mortes (WORLD HEALTH ORGANIZATION, 2022).

A COVID-19 é uma doença respiratória, transmitida pelas células epiteliais do pulmão por meio de aerossóis, que pode acarretar desde uma pneumonia viral leve até Síndrome do Desconforto Respiratório Agudo (SDRA), e em casos ainda mais graves levando a falência múltipla dos órgãos (CHOUDHARY et al., 2020).

1.2. SARS-CoV-2

SARS-CoV-2 pertence a família de Coronaviridae (CoVs), que são divididos em quatro gêneros: alfas-CoVs, beta-CoVs, gamma-CoVs e delta-CoVs. Desses, o SARS-CoV-2 pertence ao gênero beta-CoVs, que assim como os alfas-CoVs, em sua grande maioria, originam-se em morcegos e infectam outros mamíferos. Fazem parte também desse gênero os altamente patogênicos coronavírus humanos Síndrome Respiratória Aguda Grave do Coronavírus (do inglês *Severe Acute Respiratory Syndrome coronavirus* - SARS-CoV) e Síndrome respiratória do Oriente Médio (do inglês *Middle East Respiratory Syndrome* - MERS), com taxas de letalidade de 10.9% e 34.3%, respectivamente (SINGH et al., 2021).

Sendo um vírus de RNA de fita simples, de sentido positivo, o SARS-CoV-2 possui um genoma de aproximadamente 30 mil nucleotídeos, que codifica duas grandes poliproteínas (pp1a e pp1ab) que são clivadas em 16 proteínas não estruturais (NSPs), essenciais para síntese de RNA viral, e outras funções. Além disso, codifica quatro proteínas estruturais, as quais são Spike(S), Envelope(E), Membrana(M) e proteínas do nucleocapsídeo(N), necessárias para entrada e

montagem do vírus. E nove proteínas acessórias que devem atuar na neutralização da imunidade do hospedeiro durante a infecção (PENG et al., 2021).

O ponto inicial da infecção viral é a entrada do vírus na célula, dessa forma é um dos processos mais importantes, sendo o alvo no desenvolvimento de vacinas e medicamentos. E esse processo com o vírus SARS-CoV-2 ocorre mediante a interação de sua proteína estrutural Spike com a proteína humana, presente na membrana da célula, enzima conversora de angiotensina-II (do inglês *angiotensin-converting enzyme 2*- ACE2) (PENG et al., 2021).

1.2.1. Spike (S)

Há, em média, de 30 a 60 trímeros da proteína Spike sobressaindo do envelope do vírus (PENG et al., 2021), com o aspecto de coroa característico da família CoVs. O monômero da S é uma proteína de fusão tipo I, com 1273 aminoácidos, sendo a maior maquinaria de fusão viral identificada até o presente momento (HUANG et al., 2020; PENG et al., 2021).

Esse pode ser dividido em um peptídeo sinal (1-13 amino ácidos) localizado no N-terminal e duas subunidades: subunidade 1 (S1) localizada entre os resíduos de aminoácidos 14-685, sendo responsável pela ligação com o receptor, e a subunidade 2 (S2) presente entre os resíduos 686-1273 aa, que realiza a fusão com a membrana da célula hospedeira (HUANG et al., 2020).

Na S1 tem-se a região N-terminal (14-305 aa) e o domínio de interação com receptor (319-541 aa – do inglês *Receptor Binding Domain* - RBD). Já a S2 possui o peptídeo de fusão (788-806 aa - do inglês *Fusion Peptide* - FP), as sequências de repetição heptapeptídeo (do inglês *Heptapeptide Repeat* - HR) HR1 (912-984 aa) e HR2 (1163-1213 aa), o domínio transmembranar (1213-1237 aa - do inglês *TransMembrane* - TM) e o domínio citoplasmático (1237-1273 aa - do inglês *cytoplasm domain* - CT) (HUANG et al., 2020).

As proteínas S atuam como precursor inativo no organismo, de forma que, durante a infecção, essas proteínas são ativadas e clivadas em suas subunidades S1 e S2, por proteases do hospedeiro, sendo uma etapa indispensável para a fusão da membrana do vírus com a membrana da célula-alvo (HUANG et al., 2020).

1.2.2. Variabilidade do Vírus

Vírus que possuem genoma de RNA, como o SARS-CoV-2, em geral, sofrem mutações mais rápido e com uma taxa maior do que vírus de DNA, logo o SARS-CoV-2 tem alta probabilidade de sofrer mutações e melhor se adaptar ao ambiente. Tais mutações podem ocorrer em diversas regiões do genoma, sendo essas sinônimas ou não-sinônimas, e podem implicar no aumento na afinidade do vírus com o receptor, na melhora a suscetibilidade a infecção, impactar diagnóstico e, não só reduzir a eficácia de tratamentos como também reduzir a neutralização por anticorpos produzidos por infecções anteriores ou resultantes da vacinação (ZEPEDA-CERVANTES et al., 2022).

Tal fato pode ser observado pelo surgimento das variantes e as ondas causadas por elas ao longo desses quase 3 anos. Então, para rastrear tais variantes a Organização Mundial da Saúde e colaboradores propuseram a caracterização específica de Variantes de Interesse (do inglês *Variants of Interest* - VOIs), Variantes de Preocupação (do inglês *Variants of Concern* – VOCs), e Variantes sob Monitoramento (do inglês *Variants under Monitoring* – VUMs) (WHO, 2022).

As Variantes de Preocupação, VOCs, são aquelas que aumentam a transmissibilidade, causam alteração negativa na epidemiologia da COVID-19, o aumento na virulência, a alteração na apresentação clínica, ou diminuição da eficácia das medidas sociais e saúde, diagnóstico, terapias ou vacinas disponíveis. A atual VOC em circulação é a Omicron (B.1.1.529 - vários países), anteriores a essa, classificada como VOC, tem-se a Alfa (B.1.1.7 - Estados Unidos), Beta (B.1.351 - África do Sul), Gamma (P.1 - Brasil) e Delta (B.1.617.2 - Índia) (WHO, 2022).

A variante Alfa possui 23 mutações (14 não sinônimas; 6 sinônimas) e 3 deleções, sendo as mais significantes a H69-V70del, N501Y e P681H presentes na proteína S. Já a variante Beta apresenta 12 mutações não sinônimas e uma deleção, as quais dez dessas, incluindo a deleção, encontram-se na S (L18F, D80A, D215G, LAL 242-244 DEL, R246I, K417N, E484K, N501Y, D614G e A701V). Na variante Gamma há 17 mutações não sinônimas, 4 mutações sinônimas e 1 deleção, e em sua proteína S encontram-se 12 mutações (L18F, T20N, P26S, D138Y, R190S, K417T, E484T, N501Y, D614G, H655Y, T1027I e.V1176F). Quanto à variante Delta, há três sublinhagens, B.1.617.1, B.1.617.2 e B.1.617.3, e são 17 mutações das

quais as seguintes 4 são preocupantes: L452R, T478K, D614G, P681R. A ômicron apresenta mais de 18 mil mutações em seu genoma, sendo que na região codificadora são mais de 2,9 mil deleções e mutações sinônimas (mais de 2 mil), e não sinônimas (mais de 11 mil). Na proteína S, há 30 mutações (ARAF et al., 2022; MOHAMMADI; SHAYESTEHPOUR; MIRZAEI, 2021).

As Variantes de Interesse são aquelas com mudanças genéticas previstas ou conhecidas por afetar as características dos vírus (transmissibilidade, gravidade, escape imunológico, escape diagnóstico ou terapêutico) e causar transmissão comunitária significativa ou múltiplos grupos, em diversos países com prevalência relativa crescente com aumento no número de casos ao longo do tempo, ou demais impactos epidemiológicos que sugerem um risco a saúde pública mundial. Atualmente, não há VOIs circulando, mas oito variantes já foram classificadas como tal, por exemplo a Zeta (P.2), cujo a amostra mais antiga registrada é no Brasil (WHO, 2022).

Neste momento também não há Variantes em Monitoramento (VOM), mas 15 variante foram consideradas VOMs, por serem variantes que possuíam alterações genéticas suspeitas de afetar as características do vírus que poderia indicar um risco futuro, mas as evidências quanto ao impacto fenotípico ou epidemiológicos ainda não são suficientes, exigindo monitoramento e avaliação enquanto aguarda novas informações (WHO, 2022).

1.3. Proteínas Alvo: ACE2

A COVID-19 afeta principalmente indivíduos com comorbidades e/ou algum tipo de imunossupressão. Além de que algumas pessoas desenvolvem a forma severa da COVID-19, enquanto outros são assintomáticos. Assim, as características do hospedeiro, como fatores de risco e genéticos, também podem afetar a suscetibilidade e desenvolvimento da doença (CHOUDHARY et al., 2020; ZEPEDA-CERVANTES et al., 2022).

A principal proteína receptora do SARS-CoV-2, a ACE2, atua como carboxipeptidase simples que cliva polipeptídeos do sistema renina/angiotensina, possuindo papel essencial na função cardíaca, sendo expresso em vários tecidos e órgãos, o que sugere a potencial capacidade de infecção sistêmica em pacientes com COVID-19 (PENG et al, 2021). Variantes dessa proteína já foram associadas com hipertensão e outras doenças cardiovascular, e como a proteína possui

interação direta com a proteína S, tais polimorfismos podem estar associados com a suscetibilidade ao SARS-CoV-2 (CHOUDHARY et al., 2020; ZEPEDA-CERVANTES et al., 2022).

Durante a interação com a proteína ACE2, a proteína viral S sofre alterações conformacionais na qual muda do estado de pré-fusão para a fusão propriamente dita, por meio da clivagem por uma proteína como Catepsina B e L, no compartimento endossomal, ou , em especial, pela serino-protease transmembrana 2 (do inglês, *transmembrane protease serine 2* - TMPRSS2), que tem papel fundamental na patogênese e propagação viral, além de que seus inibidores bloqueiam a entrada do vírus (ZEPEDA-CERVANTES et al., 2022).

Após a interação com a ACE2 e a clivagem pela TMPRSS2, a fusão então é feita e o RNA viral entra na célula hospedeira e sequestra sua maquinaria para sintetizar as poliproteínas, que são posteriormente clivadas nas proteínas não estruturais, as proteínas estruturais e o RNA, que depois de montados são transportados para fora da célula por exocitose (PRAJAPAT et al., 2020).

1.4 Aprendizagem de Máquina e Dinâmica Molecular

As mutações do sofridas pelo SARS-CoV-2 observadas nas suas variantes assim como os polimorfismos observados em uma das principais estruturas responsáveis pela entrada do vírus no hospedeiro, levantam questões,como por exemplo, se variabilidade genética do vírus e do hospedeiro poderiam explicar os diferentes graus de severidade nos casos de infecção, e como essas variabilidades genéticas afetam a interação entre SARS-CoV-2 e ACE2. Além disso, levanta questionamento em relação à possibilidade dos polimorfismos de ACE2 afetarem a eficácia dos tratamentos. Para mais, o Brasil é o quinto país com maior número de casos e o segundo em número de mortes por COVID-19 (WORLD HEALTH ORGANIZATION, 2022), tendo em vista que há na população brasileira uma grande diferença entre as proporções ancestrais, a interação entre vírus-hospedeiro em indivíduos brasileiros apresenta comportamento diferente quando comparada a outras populações?

Portanto, as variantes de SARS-Cov-2 associadas com os polimorfismos encontrados na ACE2 em diferentes populações podem ser os principais fatores para o entendimento da COVID-19. Entretanto, como essas mutações e polimorfismos contribuem para melhorar a estabilidade e afinidade de interação

entre os complexos SARS-CoV-2-ACE2 não é totalmente compreendido. A análise de modos normais dos movimentos conformacionais das estruturas, assim como dinâmica molecular são exemplos de abordagens empregadas na tentativa de alcançar total compreensão do processo.

Para avaliar o efeito de uma mutação ou comparar mudanças estruturais entre complexos, as simulações de dinâmica molecular proporcionam um conhecimento único. Uma vez que, são capazes de prever como cada átomo em um sistema se move por um dado período de tempo, sendo capazes de capturar vários processos biomoleculares, tal como a resposta da molécula a perturbações, mutações, adição ou remoção de ligantes etc (HOLLINGSWORTH; DROR, 2018). Todavia, geram grandes quantidades de dados de milhares de átomos a cada intervalo de tempo (FLEETWOOD et al., 2020). Assim, interpretar e extrair informações dessas trajetórias não é um processo simples.

Métodos de aprendizado de máquinas são usados em análises de grande quantidade de dados, pois reduzem a dimensionalidade do problema (FLEETWOOD et al., 2020). Por isso, essas abordagens auxiliam na identificação de diferenças significantes entre as diversas trajetórias obtidas durante a simulação de dinâmica molecular, mesmo que essas diferenças sejam tênues.

Portanto, propõe-se nesse projeto usar trajetórias resultantes de simulações de dinâmica molecular e abordagens de aprendizado de máquina a fim de revelar as diferenças em linhagens de SARS-CoV-2 na região de interação com a ACE2.

1.5. Objetivo

Identificar, por meio de aprendizagem de máquinas, regiões ou resíduos críticos na interação entre variantes do vírus e hospedeiro que possam impactar significativamente a funcionalidade das proteínas ACE2 do hospedeiro e Spike do SARS-CoV-2.

2.MATERIAL E MÉTODOS

2.1. Trajetórias das Dinâmicas Moleculares

Para esse projeto foram utilizadas as trajetórias, de 100 ns cada, dos complexos formados pela ACE2, sem mutações, com a região RBD presente na subunidade S1 da proteína Spike sem mutações, e das variantes mais recentes Delta, Omicron, e Zeta, por estarem presentes na população brasileira. Todos esses complexos, bem como suas trajetórias, foram fornecidos pela mestranda Ana Luísa Rodrigues de Ávila.

2.2. Aprendizagem de Máquina

Simulações convencionais de Dinâmicas Moleculares podem chegar a gerar terabytes de dados sem qualquer pré-processamento, pois os sistemas podem ter uma alta dimensionalidade, visto que as interações podem ser entre dezenas e centenas de milhares de átomos, que ao final precisam ser condensados para possibilitar interpretação humana das informações obtidas. Este problema com dimensionalidade pode ser resolvido com uso de métodos de aprendizado de máquina (FLEETWOOD et al., 2020). Dessa forma, baseado nas abordagens usadas por Fleetwood e colaboradores (2020), foram usados os métodos supervisionados *Multilayer Perceptron* (MLP) e *Random Forest* (RF) com os dados das trajetórias para identificar resíduos que mais contribuem para a diferença no comportamento dinâmico entre os complexos. O uso de ambos, fornece um resultado mais robusto.

2.2.1 Pré-processamento

As características de entrada para tais algoritmos consiste no inverso das distâncias entre os resíduos da ACE2 e S1 da Spike. Para isso, foi utilizado a biblioteca Mdtraj (MCGIBBON et al., 2015), desenvolvida em python (https://www.python.org/), para obtenção das distâncias entre os resíduos, sendo mantidas só aqueles pares que em algum frame possuía distância igual ou menor que 15 Å, distância suficiente para englobar todos os contatos na região de interação, e essas então foram normalizadas.

Ao final desse processo tem-se os pares como características e os frames como os valores de entrada para os algoritmos de aprendizado de máquina. Nesse

processo também foram usadas as bibliotecas pandas (https://pandas.pydata.org/) e numpy (https://numpy.org/) para manipulação dos dados, ambas também desenvolvidas em python.

2.2.2. Multilayer Perceptron

Perceptron de múltiplas camadas (do inglês *Multilayer Perceptron* - MLP) consiste em uma Rede Neural Artificial feedforward (do inglês *feedforward Artificial Neural Network*) em que suas camadas são totalmente conectadas (FLEETWOOD et al., 2020). Como função de ativação dos neurônios foi utilizado *Rectified Linear Unit* (ReLU) (GLOROT; BORDES; BENGIO, 2011), por ser mais eficiente que outras funções como a sigmóide, e para a otimização dos pesos será usado o Adam (KINGMA; BA, 2015). Quanto às camadas, foi usado dois tipos de redes, uma cinco camadas com cem neurônios em cada e outra em que testou-se o uso de uma distribuição decrescente dos neurônios, tendo assim oito camadas com 100, 75, 50, 40, 30, 20, 10 e 5 neurônios respectivamente. Na implementação da rede usou-se a biblioteca python Scikit-learn (https://scikit-learn.org/stable/index.html).

Para treinamento e teste dessas redes, depois de obter os valores inversos das distâncias, foi construída uma matriz de correlação para obtenção do primeiro perfil. Outros 4 perfis foram obtidos por meio de bootstrap e valor de threshold de 0.9.

2.2.3. LRP-0

Assim como outros algoritmos de aprendizagem de máquina, o MLP é criticado por ser semelhante a uma caixa preta, uma vez que prejudica o entendimento humano de seus resultados (FLEETWOOD et al., 2020). Portanto, para extrair as características importantes para classificação, foi aplicado o *Layer-Wise Relevance Propagation* (LRP). O LRP consiste em uma técnica que redistribui o que foi recebido pelo neurônio para camada anterior, mantendo a quantidade (MONTAVON et al., 2019). O LRP possui diversos tipos, e nesse trabalho foi usado sua regra básica LRP-0:

$$R_j = \sum_k \frac{a_j w_{jk}}{\Sigma_{0,j} a_j w_{jk}} R_k \qquad \text{(eq. 1)}$$

Em que o a_j é o valor assumido pelo neurônio na camada j, w_{jk} são os pesos que conectam a camada superior k a camada inferior j e R_k consiste na relevância da camada superior que vai ser propagada para a camada inferior (FLEETWOOD et al., 2020).

2.2.4. *Random Forest*

O *Random Forest* consiste em um algoritmo cuja predição é resultado da média de um conjunto de várias árvores de decisão, sendo que cada árvore é ajustada com uma subamostra do conjunto de dados (FLEETWOOD et al., 2020). Para esse modelo foi usado o coeficiente de impureza Gini e cem árvores de decisão. Para computar a importância da característica para um certo estado, foi usado o método um contra todos (do inglês, *one-versus-the-rest*), em que é usado um *Random Forest* para cada estado, gerando um classificador binário para cada uma das classes, em que depois é computado a importância de Gini para essas. Uma vez os dados invertidos e normalizados estes já podem ser usados como entrada para o modelo, que também foi implementado usando a biblioteca Scikit-learn (https://scikit-learn.org/stable/index.html),

3.RESULTADOS E DISCUSSÃO

3.1 MLP

Ao final da seleção dos pares com pelo menos 15 A° em pelo menos um frame e do bootstrap, obteve-se 5 perfis com 1828, 1907, 1925, 1909 e 1934 pares, respectivamente. Todos com 40 mil frames. Então cada perfil foi usado para o treinamento dos dois tipos de redes, o que resultou em 10 redes com acurácia de 100% e perdas menores que 0.01, como pode-se observar no Quadro 1 e 2. Demais métricas presentes no reporte de classificação (do inglês *Classification Report*) podem ser observadas em apêndices 1 e 2.

Quadro 1 - Acurácia e Perda dos MLPs com cinco camadas de 100 neurônios.

Perfil (Pares por Frame)	Acurácia	Perda
1828 x 40 mil	1.00	0.0011
1907 x 40 mil	1.00	0.0028
1925 x 40 mil	1.00	0.0005
1909 x 40 mil	1.00	0.0008
1934 x 40 mil	1.00	0.0027

Fonte: Elaborado pelo autor

Quadro 2 - Acurácia e Perda dos MLPs com oito camadas com distribuição decrescente de neurônios.

Perfil (Pares por Frame)	Acurácia	Perda
1828 x 40 mil	1.00	0.0022
1907 x 40 mil	1.00	0.0014
1925 x 40 mil	1.00	0.0036
1909 x 40 mil	1.00	0.0008
1934 x 40 mil	1.00	0.0017

Fonte: Elaborado pelo autor

Dado essas métricas, de acurácia alta e perda muito pequena das redes, é provável que tenha ocorrido overfitting, ou seja, ao invés de aprender com os dados os algoritmos se ajustaram a esses. Logo, é necessário mais testes, fazendo o uso

de diferentes parâmetros, para encontrar os mais adequados para esses algoritmos. Posto isso, ainda assim é interessante avaliar os pares de resíduos indicados ao final por esse métodos.

Nos quadros 3 e 4 pode-se observar os 5 pares mais importantes para a classificação de cada complexo, resultante do LRP-0 para as redes. Nos apêndices de 4 a 7 pode-se observar os gráficos de relevância para a rede com cinco camadas, e em apêndices 8 a 11 os gráficos da rede com oito camadas.

Quadro 3 - Pares as maiores relevância para cada complexo, resultado LRP-0 para rede com cinco camadas de cem neurônios.

Complexos	**Pares (ACE2, SPIKE) e Importância**
ACE2-SPIKE(Selvagem)	(SER106, ASN487) = 8.991 (GLU22, SER477) = 7.337 (SER19, GLY485) = 6.243 (ASN338, GLN498) = 5.638 (SER19, ASP467) = 5.390
ACE2-SPIKE(Delta)	(LEU45, ASN450) = 4.736 (THR324, GLU406) = 4.588 (LEU391, LYS417) = 4.210 (GLN42, SER443) = 4.022 (LYS31, SER494) = 3.886
ACE2-SPIKE(Ômicron)	(GLU312, GLY504) = 1.142 (PHE72, TYR501) = 1.092 (GLY352, TYR451) = 1.002 (GLN325, GLY447) = 0.950 (GLU329, SER438) = 0.874
ACE2-SPIKE(P2)	(SER19, ASN477) = 1.298 (SER19, PRO479) = 0.908 (SER105, TYR489) = 0.781 (ANS338, THR500) = 0.640 (SER106, PHE486) = 0.595

Fonte: Elaborado pelo próprio autor.

Quadro 4 - Pares as maiores relevância para cada complexo, resultado LRP-0 para rede com oito camadas com distribuição decrescente de neurônios..

Complexos	Pares (ACE2, SPIKE) e Importância
ACE2-SPIKE(Selvagem)	(SER106, GLY485) = 4.777 (VAL107, PHE486) = 4.745 (GLN89, SER477) = 4.663 (SER19, PRO479) = 3.608 (ALA71, GLU484) = 2.778
ACE2-SPIKE(Delta)	(ASP30, GLU484) = 3.583 (GLN24, LYS417) = 1.667 (GLY352, ARG408) = 1.467 (ALA65, SER443) = 1.111 (ASN33, GLN498) = 0.965
ACE2-SPIKE(Ômicron)	(GLU329, SER438) = 2.434 (GLN42, SER349) = 2.094 (TYR381, GLY502) = 1.759 (GLY352, ASN448) = 1.656 (GLY354, GLY504) = 1.501
ACE2-SPIKE(P2)	(PRO321, ARG403) = 6.209 (SER19, ASN477) = 5.333 (SER19, PRO479) = 4.838 (GLN325, SER371) = 4.581 (GLU37, THR415) = 4.220

Fonte: Elaborado pelo próprio autor.

Vale de nota que a mudança nos parâmetros, neste caso a mudança no número de camadas, resultou na mesma acurácia e valores de perda semelhantes, mas em um diferente conjunto de pares considerados os mais relevantes para a classificação dos complexos. Isso é decorrente da diferença do número e distribuição de neurônios, visto que isso afeta a quantidade e valores dos pesos, que implicam em valores de neurônios diferentes e, portanto, em um resultado do LRP-0 diferente.

Analisando os pares indicados nota-se a presença dos resíduos GLN24, GLN42, LEU45, GLN325, GLU329 e GLY354, localizados na proteína ACE2, que são considerados resíduos chave na interação com a proteína S (SURYAMOHAN et

al., 2021). Além desses, um dos resíduos-chave da ACE2 que apareceu mais de uma vez dentre os pares é o SER19, inclusive mutação nesse resíduo (S19P) foi identificado como sendo um dos polimorfismos que aumentam a interação proteína ACE2/S, o que difere dos resíduos ASN33 , PHE72 e GLY352 cujas mutações N33I, F72V e G352V respectivamente, diminuem tal interação (SURYAMOHAN et al., 2021).

Ademais, outros resíduos-chave que apareceram nos pares são: o resíduo ASP30, que faz ponte salina com o resíduo LYS407 da S, mas aqui aparece como par do resíduo GLU484, sendo necessária uma investigação estrutural; o resíduo LYS31, em que a mutação K31R é prevista como sendo uma das mutações que reduzem a interação com a proteína S; e o resíduo GLU37, que coordena contatos polares como os resíduos TYR505 e GLY502 do RBD da S (SURYAMOHAN et al., 2021), contudo aqui faz par com o resíduo THR415, enquanto o resíduo GLY502 faz com o TYR381 da ACE2.

Quanto aos resíduos da proteína Spike, aparecem nos pares os resíduos críticos PHE486, SER494 e TYR501 para interação com a ACE2, sendo que último a mutação N501Y, que está presente em múltiplas VOCs, e garante uma maior afinidade de ligação com a ACE2 dado interação mais forte com seus resíduos TYR41 e LYS353 (SINGH et al., 2021), entretanto nesses resultados aparece em par com o resíduo PHE72.

No que se refere ao resíduo LYS417, a mutação K417N aumenta a transmissibilidade do vírus, enquanto a mutação no resíduo SER477 (S477N) aumenta afinidade de ligação e no resíduo GLU484 a mutação E484K está associado a resistência a anticorpos (SINGH et al., 2021). Há também os resíduos SER371 e GLN498, em que suas respectivas mutações S371L e Q498R estão associadas a uma maior afinidade de ligação entre a variante Ômicron e a proteína ACE2 (ARAF et al., 2022).

Essa diferença na formação de pares pode indicar que selecionar apenas os pares que em um dos frames tenha distância menor que 15 Å seja um parâmetro abrangente, tendo em vista que os complexos se movem ao longo da trajetória e por um breve período podem estar próximos o suficiente para ter tal distância. Desta maneira, é necessário que os pares sejam visualizados na estrutura do complexo para melhor compreensão.

3.2 RF

Para o *Random Forest* foram usados 3201 pares, também com 40 mil frames, que, como foi usado o bootstrap implementado neste classificador, foi dividido em amostras para o treinamento das cem árvores de decisão que integraram esse modelo. E como o um-contra-todos foi usado, obteve-se ao final 4 classificadores *Random Forest*, um para cada classe, em que obteve-se uma acurácia de 100%, o que também indica overfitting. O *Classification Report* pode ser observado em apêndice 3.

No quadro 5 pode-se observar os cinco pares mais importantes pelo Random Forest, sendo resultado da importância de Gini dos classificadores. Nos apêndices 5 a 8 tem-se o gráfico para cada complexo.

Quanto à importância dada pela impureza de Gini no Random Forest, os pares são formados por alguns dos resíduos já mencionados no MLP, mas também pelos resíduos da Spike TYR505, cuja mutação pode aumentar a transmissão (SINGH et al., 2021). Pelo resíduo ARG498, que é uma das mutações presentes na variante Ômicron. responsável pela maior afinidade desta com a proteína ACE2 (ARAF et al., 2022). Em relação a ACE2, o resíduo SER19 foi recorrente entre os pares, esse sendo um resíduo-chave, assim como os resíduos LYS353 e THR27, que também apareceram entre os pares (SURYAMOHAN et al., 2021).

Quadro 5 - Pares com maior importância de Gini

Complexos	Pares (ACE2, SPIKE) e Importância
ACE2-SPIKE(Selvagem)	(SER19, VAL483) = 0.027 (SER19, CYS488) = 0.026 (SER19, CYS480) = 0.016 (SER44, TYR505) = 0.0143 (SER19, GLN474) = 0.0143
ACE2-SPIKE(Delta)	(ALA36, ASN501) = 0.021 (GLY66, ASN501) = 0.020 (ALA342, THR500) = 0.019 (ASN103, TYR505) = 0.015 (LYS68, ASN501) = 0.013
ACE2-SPIKE(Ômicron)	(ALA25, ASN417) = 0.031 (GLN24, ASN417) = 0.031 (ILE21, ASN417) = 0.031 (LYS353, ARG498) = 0.029 (THR27, ASN417) = 0.028
ACE2-SPIKE(P2)	(SER106, LYS484) = 0.256 (SER19, CYS480) = 0.023 (SER105, ASN487) = 0.022 (GLY104, ASN487) = 0.020 (SER105, LYS484) = 0.020

Fonte: Elaborado pelo próprio autor.

.

5.CONCLUSÃO

Mediante ao exposto ao longo desse trabalho, pode-se concluir que as técnicas de aprendizado de máquina apresentadas, MLP e RF, e seus respectivos métodos de extração de características LRP-0 e Importância de Gini, possuem capacidade de apresentar resultados interessantes quanto a resíduos e pares dos complexos abordado. Contudo, ainda é necessário mais testes em relação aos parâmetros de ambos modelos, a fim de adequá-los para que não ocorra o overfitting. Além disso, vale ressaltar que, uma vez ajustado os parâmetros e obtenção dos valores adequados, faz-se necessário também a análise na estrutura dos pares indicados como os mais importantes para classificação, o que pode levar a insights quanto ao impacto dessas na interação do complexo ACE2/S.

REFERÊNCIAS

ARAF, Yusha; AKTER, Fariya; TANG, Yan dong; et al. Omicron variant of SARS-CoV-2: Genomics, transmissibility, and responses to current COVID-19 vaccines. Journal of Medical Virology, v. 94, n. 5, p. 1825–1832, 2022. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/35023191/>. Acesso em: 7 abr. 2022.

CHOUDHARY, Sarita; SREENIVASULU, Karli; MITRA, Prasenjit; et al. Role of genetic variants and gene expression in the susceptibility and severity of COVID-19. Annals of Laboratory Medicine, v. 41, n. 2, p. 129–138, 2020. Disponível em: </pmc/articles/PMC7591285/>. Acesso em: 9 abr. 2022.

FLEETWOOD, Oliver; KASIMOVA, Marina A.; WESTERLUND, Annie M.; et al. Molecular Insights from Conformational Ensembles via Machine Learning. Biophysical Journal, v. 118, n. 3, p. 765–780, 2020. Disponível em: </pmc/articles/PMC7002924/>. Acesso em: 9 abr. 2022.

GLOROT, Xavier; BORDES, Antoine; BENGIO, Yoshua. Deep sparse rectifier neural networks. In: Journal of Machine Learning Research. [s.l.: s.n.], 2011, v. 15, p. 315–323.

HOLLINGSWORTH, Scott A.; DROR, Ron O. Molecular Dynamics Simulation for All. Neuron, v. 99, n. 6, p. 1129–1143, 2018.

HUANG, Yuan; YANG, Chan; XU, Xin feng; et al. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacologica Sinica, v. 41, n. 9, p. 1141–1149, 2020. Disponível em: <https://www.nature.com/articles/s41401-020-0485-4>. Acesso em: 1 dez. 2022.

KINGMA, Diederik P.; BA, Jimmy Lei. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. [s.l.]: International Conference on Learning Representations, ICLR, 2015. Disponível em: <https://arxiv.org/abs/1412.6980v9>. Acesso em: 8 dez. 2022.

MCGIBBON, Robert T.; BEAUCHAMP, Kyle A.; HARRIGAN, Matthew P.; et al. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophysical Journal, v. 109, n. 8, p. 1528–1532, 2015.

MOHAMMADI, Mehrdad; SHAYESTEHPOUR, Mohammad; MIRZAEI, Hamed. The impact of spike mutated variants of SARS-CoV2 [Alpha, Beta, Gamma, Delta, and Lambda] on the efficacy of subunit recombinant vaccines. Brazilian Journal of Infectious Diseases, v. 25, n. 4, p. 101606, 2021. Disponível em: </pmc/articles/PMC8367756/>. Acesso em: 9 abr. 2022.

MONTAVON, Grégoire; BINDER, Alexander; LAPUSCHKIN, Sebastian; et al. Layer-Wise Relevance Propagation: An Overview. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). [s.l.: s.n.], 2019, v. 11700 LNCS, p. 193–209.

PENG, Ruchao; WU, Lian Ao; WANG, Qingling; et al. Cell entry by SARS-CoV-2. Trends in Biochemical Sciences, v. 46, n. 10, p. 848–860, 2021. Disponível em: <https://doi.org/10.1016/j.tibs.2021.06.001>. Acesso em: 2 set. 2021.

PRAJAPAT, Manisha; SARMA, Phulen; SHEKHAR, Nishant; et al. Update on the target structures of SARS-CoV-2: A systematic review. Indian Journal of Pharmacology, v. 52, n. 2, p. 142–149, 2020. Disponível em: </pmc/articles/PMC7282679/>. Acesso em: 9 abr. 2022.

SINGH, Jalen; PANDIT, Pranav; MCARTHUR, Andrew G.; et al. Evolutionary trajectory of SARS-CoV-2 and emerging variants. Virology Journal, v. 18, n. 1, p. 166, 2021. Disponível em: </pmc/articles/PMC8361246/>. Acesso em: 3 set. 2021.

SURYAMOHAN, Kushal; DIWANJI, Devan; STAWISKI, Eric W.; et al. Human ACE2 receptor polymorphisms and altered susceptibility to SARS-CoV-2. Communications Biology, v. 4, n. 1, p. 1–11, 2021. Disponível em: <https://www.nature.com/articles/s42003-021-02030-3>. Acesso em: 9 abr. 2022.

WHO. Tracking SARS-CoV-2 variants. Who. Disponível em: <https://www.who.int/activities/tracking-SARS-CoV-2-variants>. Acesso em: 3 dez. 2022.

WHO DIRECTOR GENERAL'S SPEECHES. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. WHO Director General's speeches, n. March, p. 4, 2020. Disponível em: <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>. Acesso em: 28 nov. 2022.

WORLD HEALTH ORGANIZATION. WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. World Health Organization, p. 1–5, 2021. Disponível em: <https://covid19.who.int/>. Acesso em: 12 dez. 2022.

ZEPEDA-CERVANTES, Jesús; MARTÍNEZ-FLORES, Daniel; RAMÍREZ-JARQUÍN, Josué Orlando; et al. Implications of the Immune Polymorphisms of the Host and the Genetic Variability of SARS-CoV-2 in the Development of COVID-19. Viruses, v. 14, n. 1, 2022. Disponível em: </pmc/articles/PMC8778858/>. Acesso em: 9 abr. 2022.

APÊNDICE

Apêndice 1 - *Classification Report* MLP com cinco camadas de cem neurônios.

Perfil	*Classification Report*
1828 x 40 mil	precision recall f1-score support ACE2-SPIKE(P2) 1.00 1.00 1.00 2003 ACE2-SPIKE(WT) 1.00 1.00 1.00 2018 ACE2-SPIKE(delta1) 1.00 1.00 1.00 2000 ACE2-SPIKE(omicron1) 1.00 1.00 1.00 1980 accuracy 1.00 8001 macro avg 1.00 1.00 1.00 8001 weighted avg 1.00 1.00 1.00 8001 Perda: 0.0011579129333462378
1907 x 40 mil	precision recall f1-score support ACE2-SPIKE(P2) 1.00 1.00 1.00 2003 ACE2-SPIKE(WT) 1.00 1.00 1.00 2018 ACE2-SPIKE(delta1) 1.00 1.00 1.00 2000 ACE2-SPIKE(omicron1) 1.00 1.00 1.00 1980 accuracy 1.00 8001 macro avg 1.00 1.00 1.00 8001 weighted avg 1.00 1.00 1.00 8001 Perda: 0.0028196864327025587
1925 x 40 mil	precision recall f1-score support ACE2-SPIKE(P2) 1.00 1.00 1.00 2003 ACE2-SPIKE(WT) 1.00 1.00 1.00 2018 ACE2-SPIKE(delta1) 1.00 1.00 1.00 2000 ACE2-SPIKE(omicron1) 1.00 1.00 1.00 1980 accuracy 1.00 8001 macro avg 1.00 1.00 1.00 8001 weighted avg 1.00 1.00 1.00 8001 Perda: 0.000468362446264934
1909 x 40 mil	precision recall f1-score support ACE2-SPIKE(P2) 1.00 1.00 1.00 2003 ACE2-SPIKE(WT) 1.00 1.00 1.00 2018 ACE2-SPIKE(delta1) 1.00 1.00 1.00 2000 ACE2-SPIKE(omicron1) 1.00 1.00 1.00 1980 accuracy 1.00 8001 macro avg 1.00 1.00 1.00 8001 weighted avg 1.00 1.00 1.00 8001 Perda: 0.0007678636780219338

1934 x 40 mil	precision recall f1-score support ACE2-SPIKE(P2) 1.00 1.00 1.00 2003 ACE2-SPIKE(WT) 1.00 1.00 1.00 2018 ACE2-SPIKE(delta1) 1.00 1.00 1.00 2000 ACE2-SPIKE(omicron1) 1.00 1.00 1.00 1980 accuracy 1.00 8001 macro avg 1.00 1.00 1.00 8001 weighted avg 1.00 1.00 1.00 8001 Perda: 0.0027129004542402733

Fonte: Elaborado pelo autor.

Apêndice 2 - *Classification Report* MLP com distribuição decrescente de neurônios em oito camadas.

Perfil	***Classification Report***
1828 x 40 mil	precision recall f1-score support ACE2-SPIKE(P2) 1.00 1.00 1.00 2003 ACE2-SPIKE(WT) 1.00 1.00 1.00 2018 ACE2-SPIKE(delta1) 1.00 1.00 1.00 2000 ACE2-SPIKE(omicron1) 1.00 1.00 1.00 1980 accuracy 1.00 8001 macro avg 1.00 1.00 1.00 8001 weighted avg 1.00 1.00 1.00 8001 Perda: 0.0021944622069453297
1907 x 40 mil	precision recall f1-score support ACE2-SPIKE(P2) 1.00 1.00 1.00 2003 ACE2-SPIKE(WT) 1.00 1.00 1.00 2018 ACE2-SPIKE(delta1) 1.00 1.00 1.00 2000 ACE2-SPIKE(omicron1) 1.00 1.00 1.00 1980 accuracy 1.00 8001 macro avg 1.00 1.00 1.00 8001 weighted avg 1.00 1.00 1.00 8001 Perda: 0.001368776453017395
1925 x 40 mil	precision recall f1-score support ACE2-SPIKE(P2) 1.00 1.00 1.00 2003 ACE2-SPIKE(WT) 1.00 1.00 1.00 2018 ACE2-SPIKE(delta1) 1.00 1.00 1.00 2000 ACE2-SPIKE(omicron1) 1.00 1.00 1.00 1980 accuracy 1.00 8001

	macro avg 1.00 1.00 1.00 8001 weighted avg 1.00 1.00 1.00 8001 Perda: 0.003572032391250793
1909 x 40 mil	precision recall f1-score support ACE2-SPIKE(P2) 1.00 1.00 1.00 2003 ACE2-SPIKE(WT) 1.00 1.00 1.00 2018 ACE2-SPIKE(delta1) 1.00 1.00 1.00 2000 ACE2-SPIKE(omicron1) 1.00 1.00 1.00 1980 accuracy 1.00 8001 macro avg 1.00 1.00 1.00 8001 weighted avg 1.00 1.00 1.00 8001 Perda: 0.0008118542414011292
1934 x 40 mil	precision recall f1-score support ACE2-SPIKE(P2) 1.00 1.00 1.00 2003 ACE2-SPIKE(WT) 1.00 1.00 1.00 2018 ACE2-SPIKE(delta1) 1.00 1.00 1.00 2000 ACE2-SPIKE(omicron1) 1.00 1.00 1.00 1980 accuracy 1.00 8001 macro avg 1.00 1.00 1.00 8001 weighted avg 1.00 1.00 1.00 8001 Perda: 0.0017346427187478351

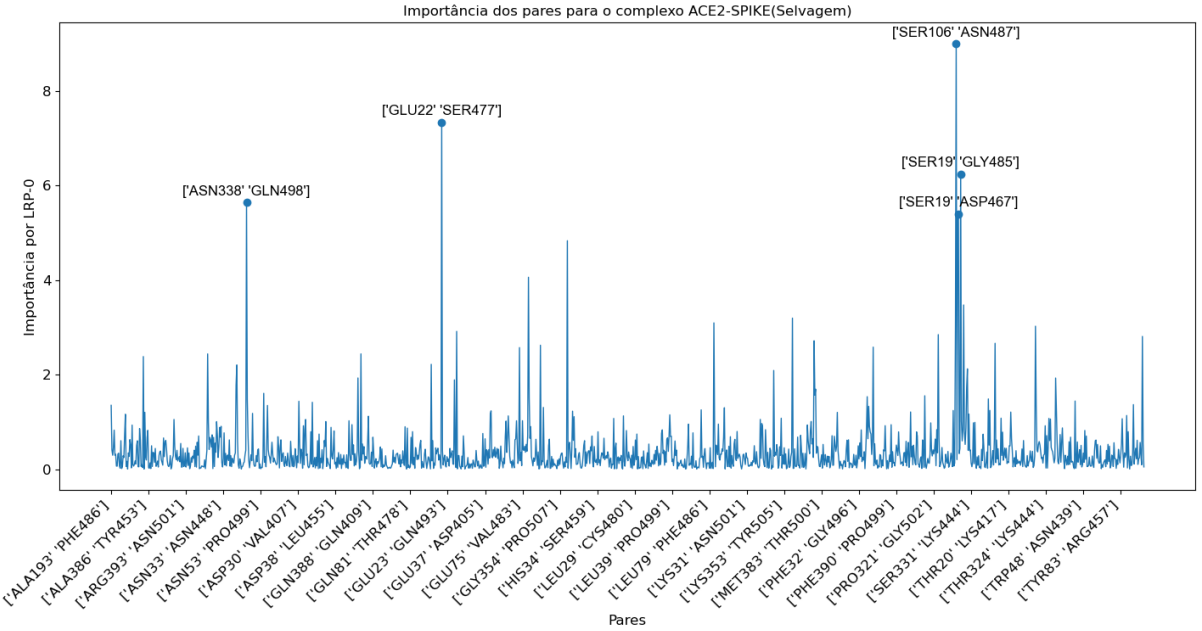
Fonte: Elaborado pelo autor.

Apêndice 3 - *Classification Report* do *Random Forest*

	precision	recall	f1-score	support
ACE2-SPIKE(P2)	1.00	1.00	1.00	2003
ACE2-SPIKE(WT)	1.00	1.00	1.00	2018
ACE2-SPIKE(delta1)	1.00	1.00	1.00	2000
ACE2-SPIKE(omicron1)	1.00	1.00	1.00	1980
accuracy			1.00	8001
macro avg	1.00	1.00	1.00	8001
weighted avg	1.00	1.00	1.00	8001

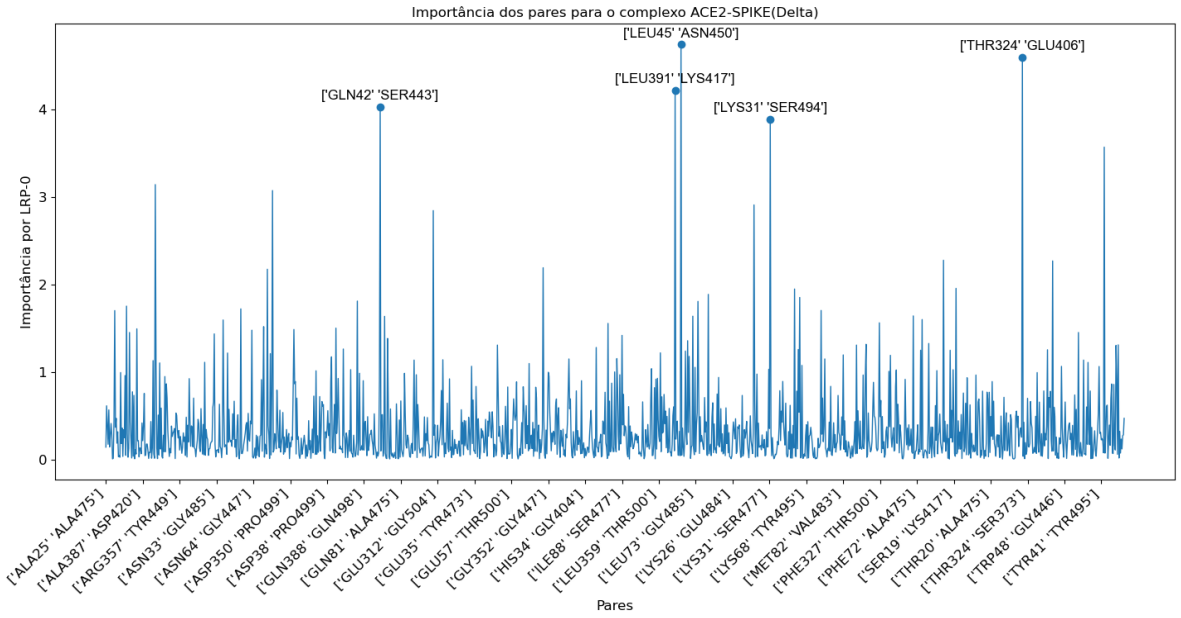
Fonte: Elaborado pelo autor.

Apêndice 4 - Gráfico dos resultados do LRP-0 da rede com cinco camadas para o complexo ACE2-SPIKE(Selvagem)



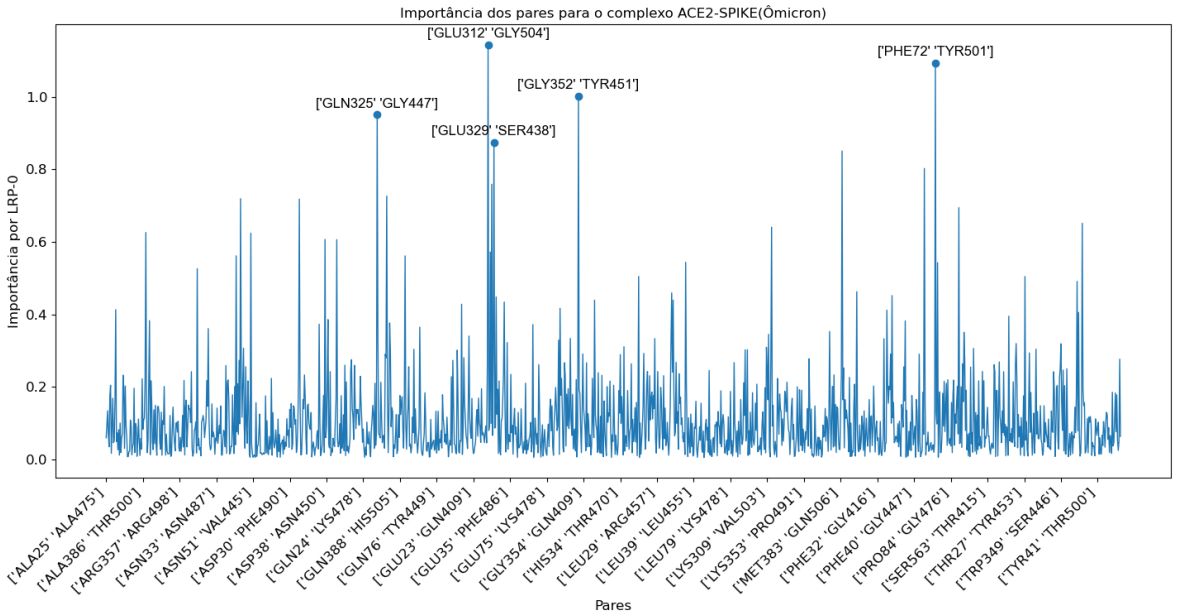
Fonte: Elaborado pelo autor.

Apêndice 5 - Gráfico dos resultados do LRP-0 da rede com cinco camadas para o complexo ACE2-SPIKE(Delta)



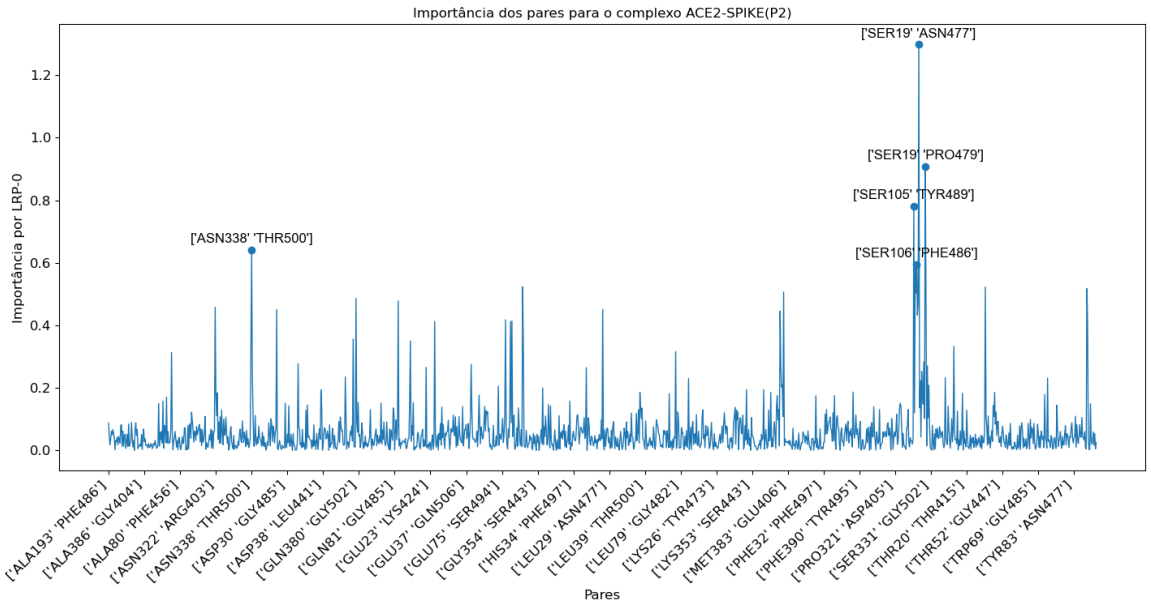
Fonte: Elaborado pelo autor.

Apêndice 6 - Gráfico dos resultados do LRP-0 da rede com cinco camadas para o complexo ACE2-SPIKE(Ômicron)



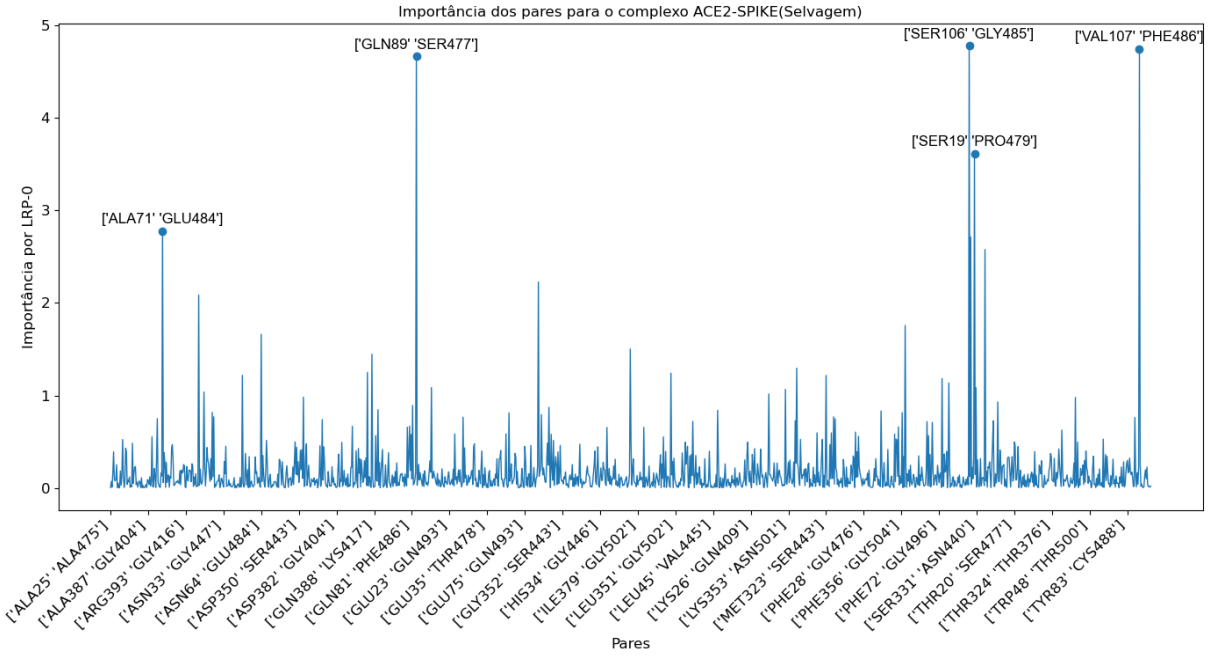
Fonte: Elaborado pelo autor.

Apêndice 7 - Gráfico dos resultados do LRP-0 da rede com cinco camadas para o complexo ACE2-SPIKE(P2)



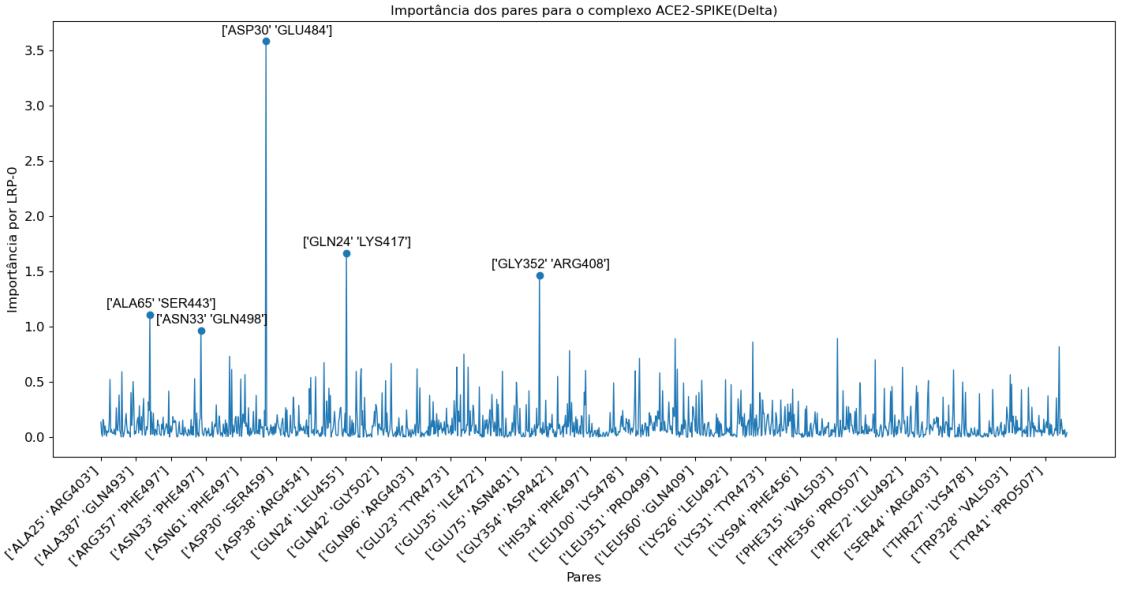
Fonte: Elaborado pelo autor.

Apêndice 8 - Gráfico dos resultados do LRP-0 da rede com distribuição decrescente de neurônios em oito camadas para o complexo ACE2-SPIKE(Selvagem)



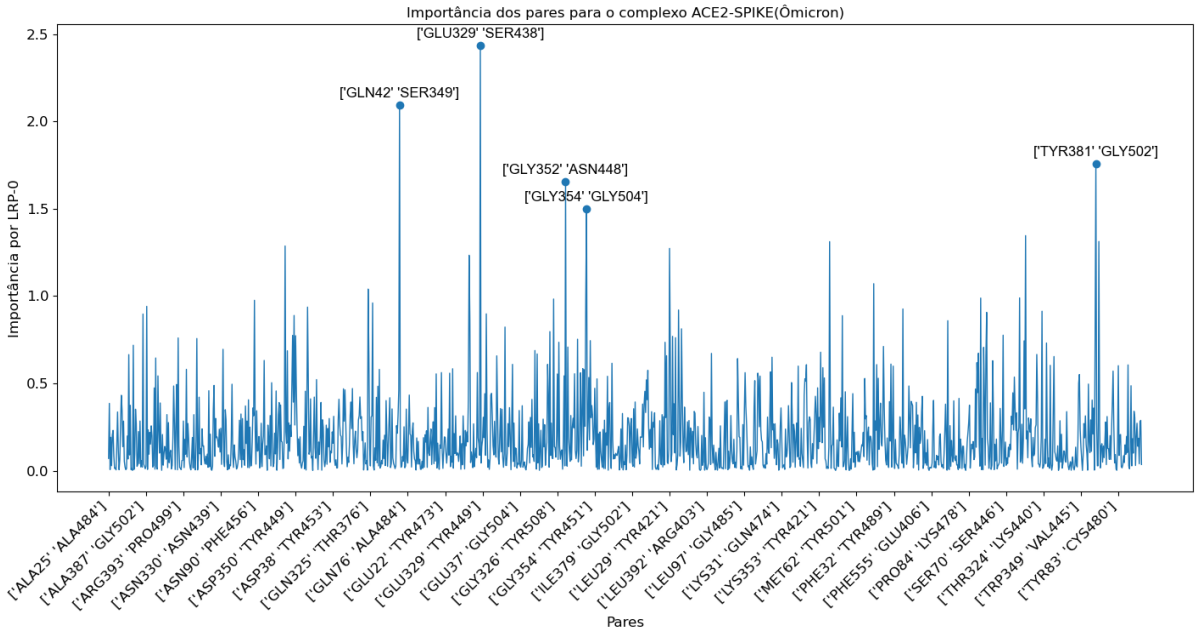
Fonte: Elaborado pelo autor.

Apêndice 9 - Gráfico dos resultados do LRP-0 da rede com distribuição decrescente de neurônios em oito camadas para o complexo ACE2-SPIKE(Delta)



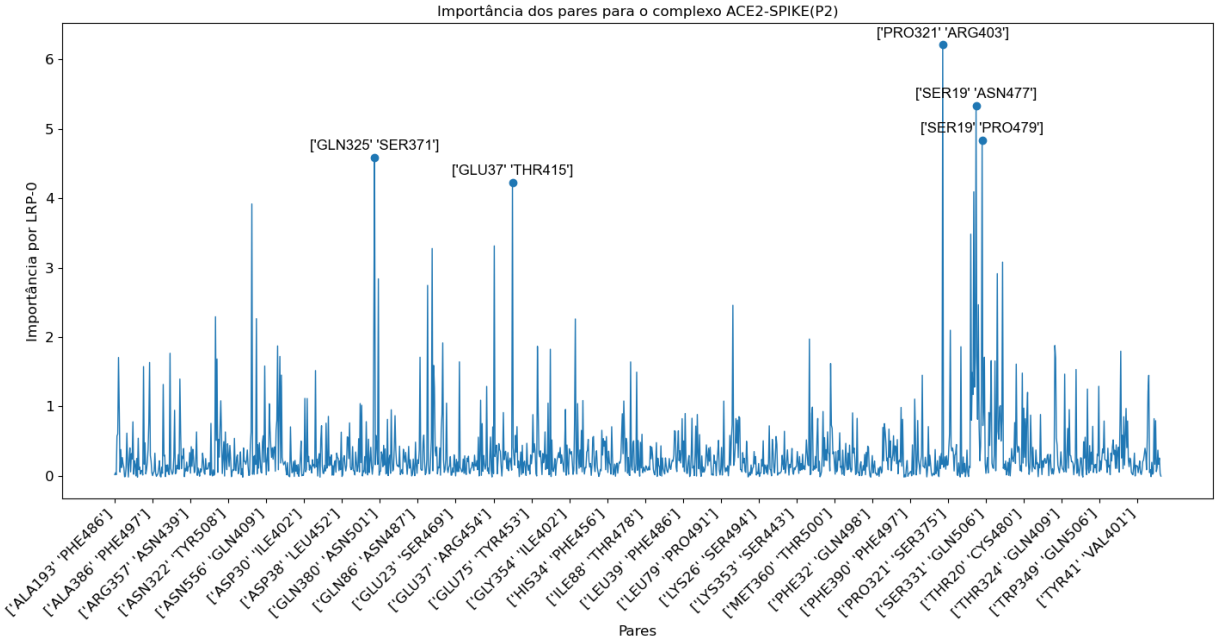
Fonte: Elaborado pelo autor.

Apêndice 10 - Gráfico dos resultados do LRP-0 da rede com distribuição decrescente de neurônios em oito camadas para o complexo ACE2-SPIKE(Ômicron)



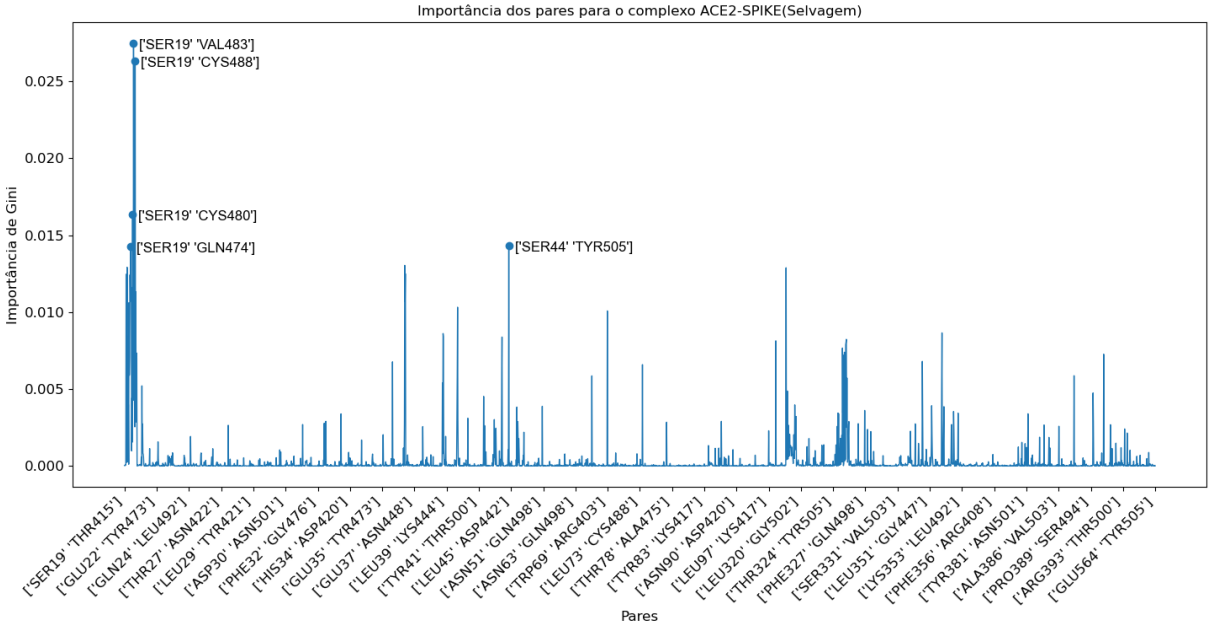
Fonte: Elaborado pelo autor.

Apêndice 11 - Gráfico dos resultados do LRP-0 da rede com distribuição decrescente de neurônios em oito camadas para o complexo ACE2-SPIKE(P2).



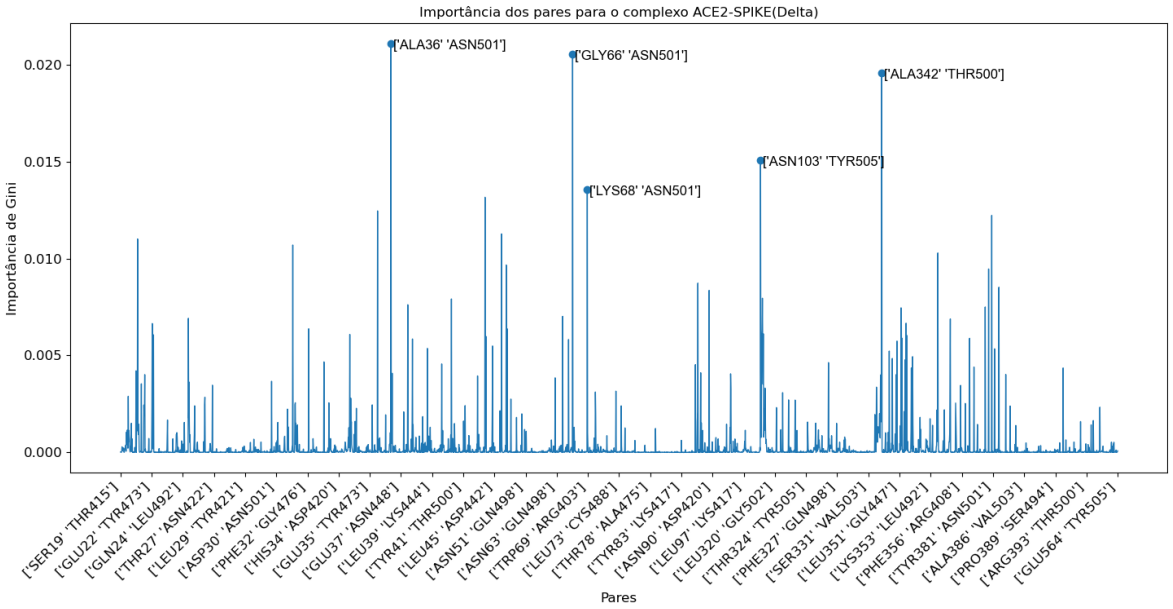
Fonte: Elaborado pelo autor.

Apêndice 12 - Gráfico dos resultados da Importância de Gini do *Random Forest* para o complexo ACE2-SPIKE(Selvagem).



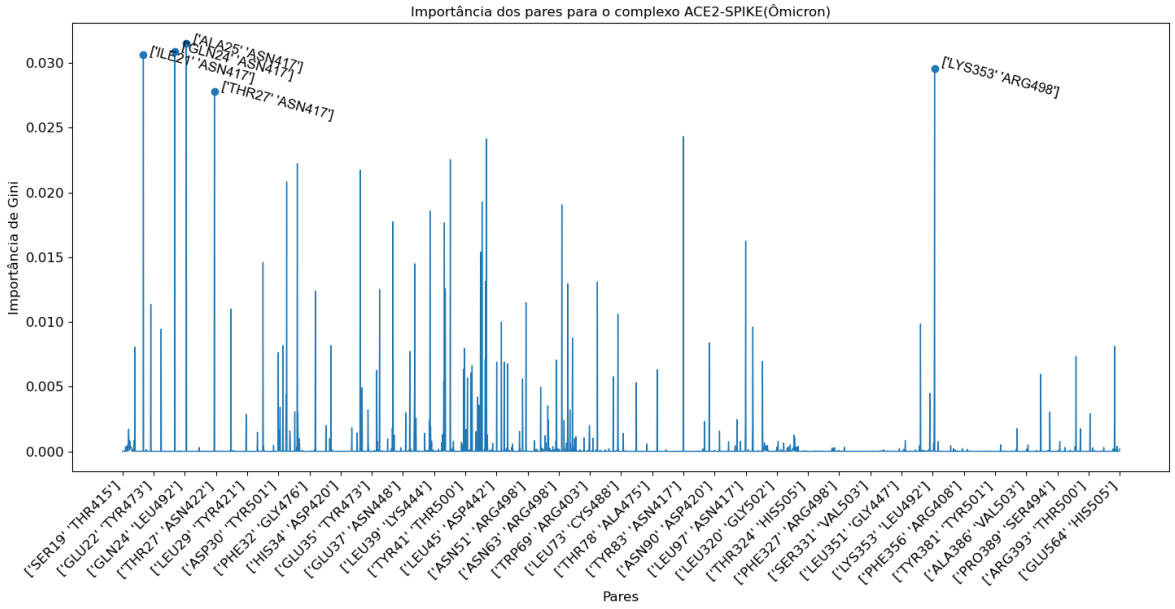
Fonte: Elaborado pelo autor.

Apêndice 13 - Gráfico dos resultados da Importância de Gini do *Random Forest* para o complexo ACE2-SPIKE(Delta).



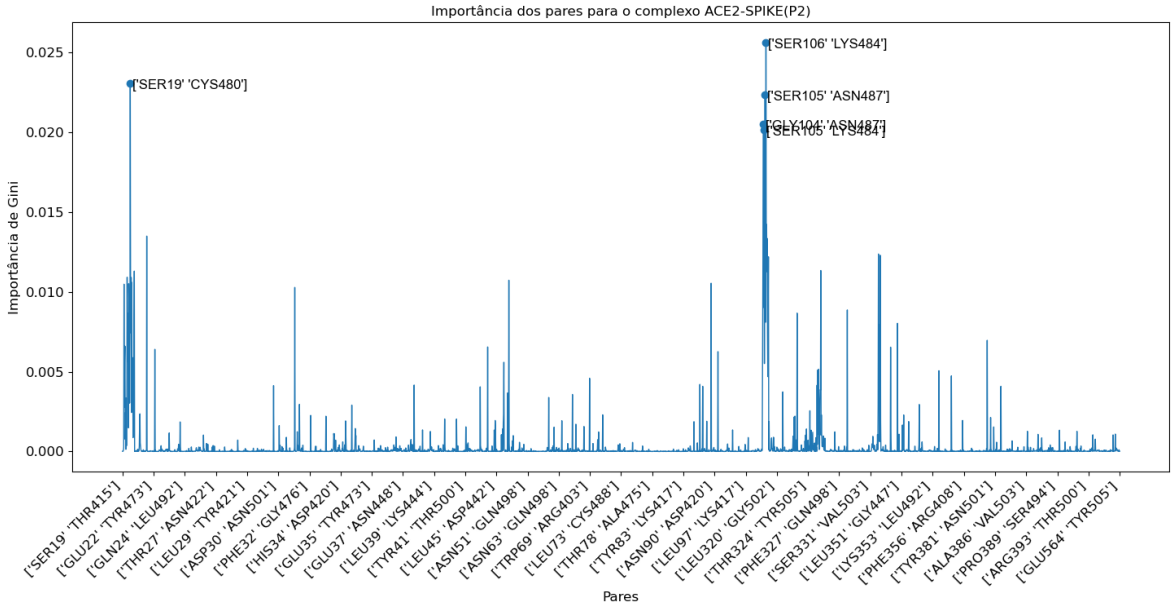
Fonte: Elaborado pelo autor.

Apêndice 14 - Gráfico dos resultados da Importância de Gini do *Random Forest* para o complexo ACE2-SPIKE(Ômicron).



Fonte: Elaborado pelo autor.

Apêndice 15 - Gráfico dos resultados da Importância de Gini do *Random Forest* para o complexo ACE2-SPIKE(P2).



Observação: os pares que se sobrepõem são (GLY104, ASN487) e (SER105, LYS484).

Fonte: Elaborado pelo autor.