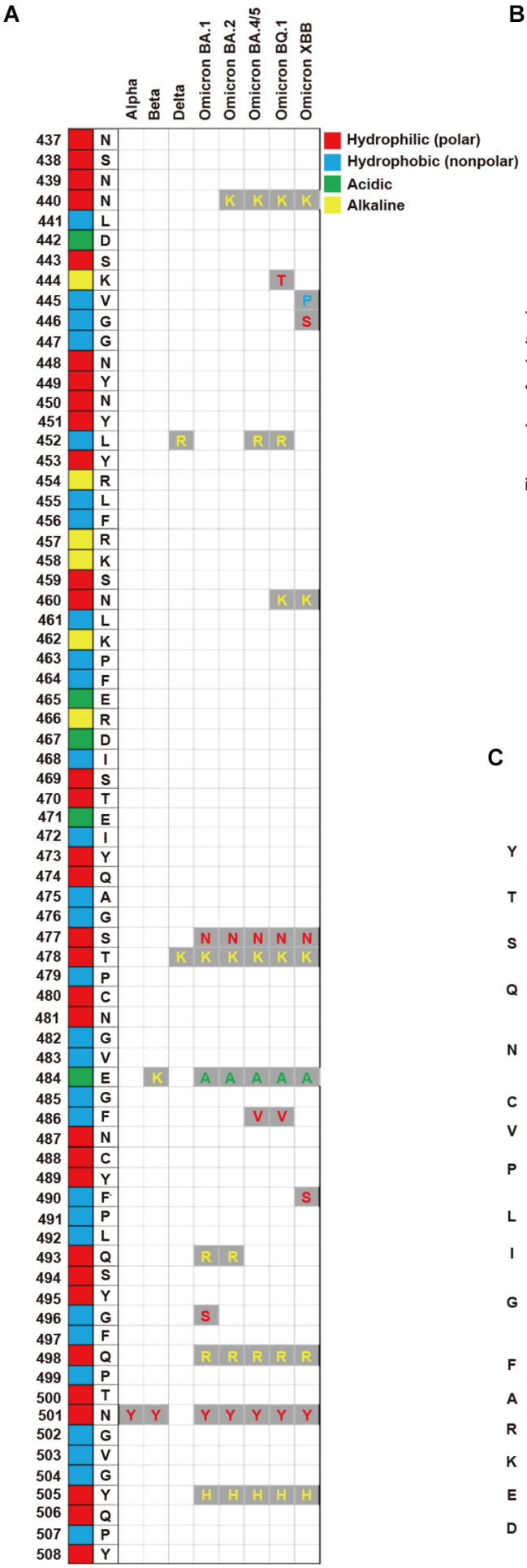
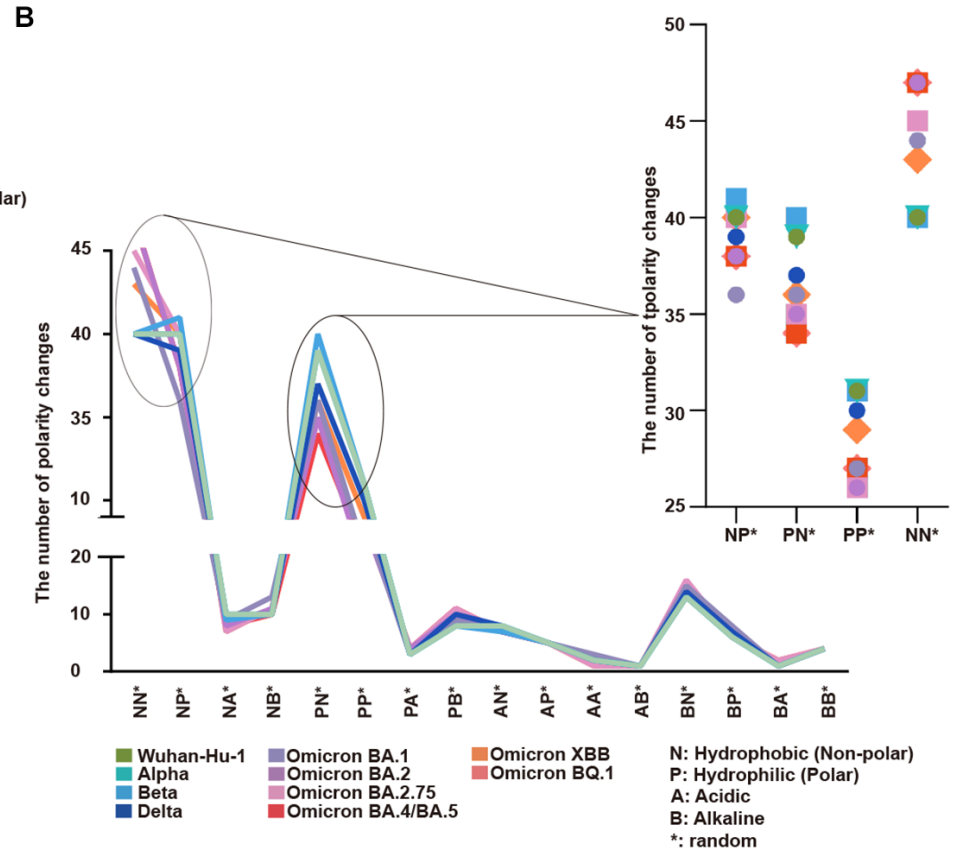
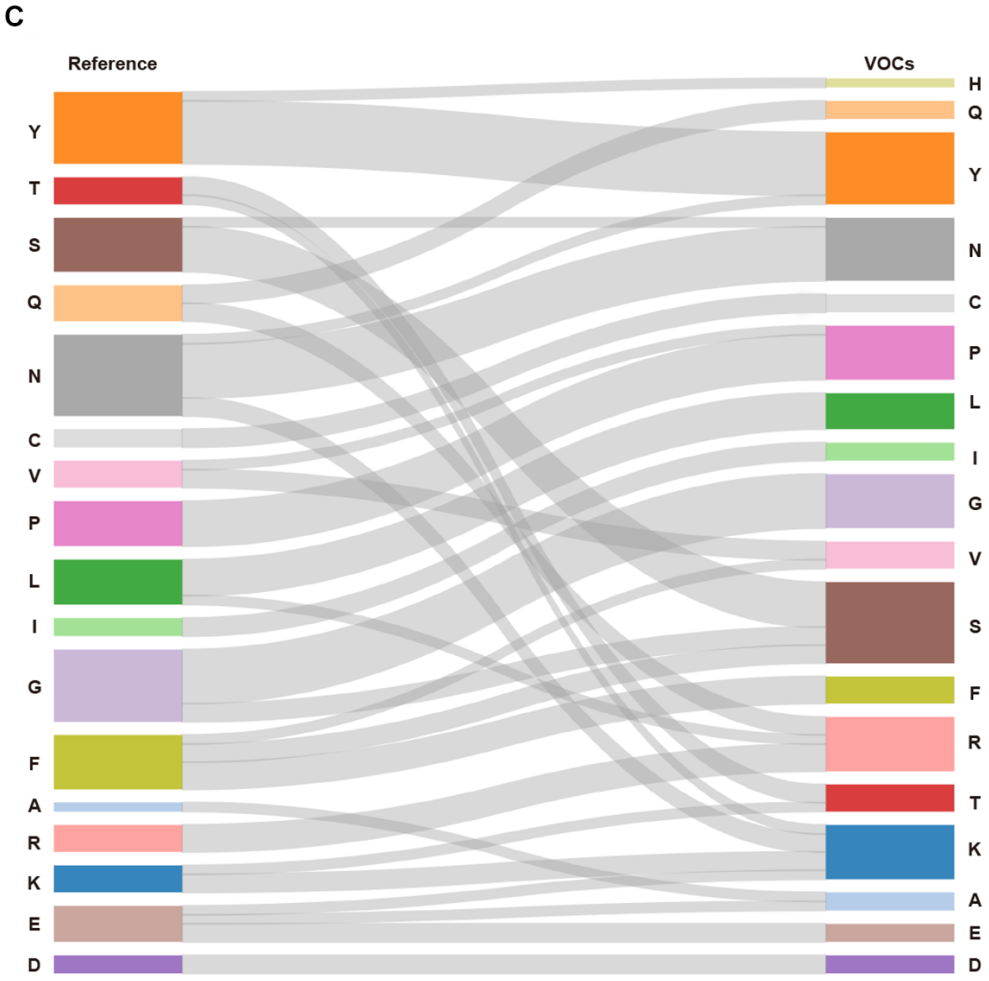
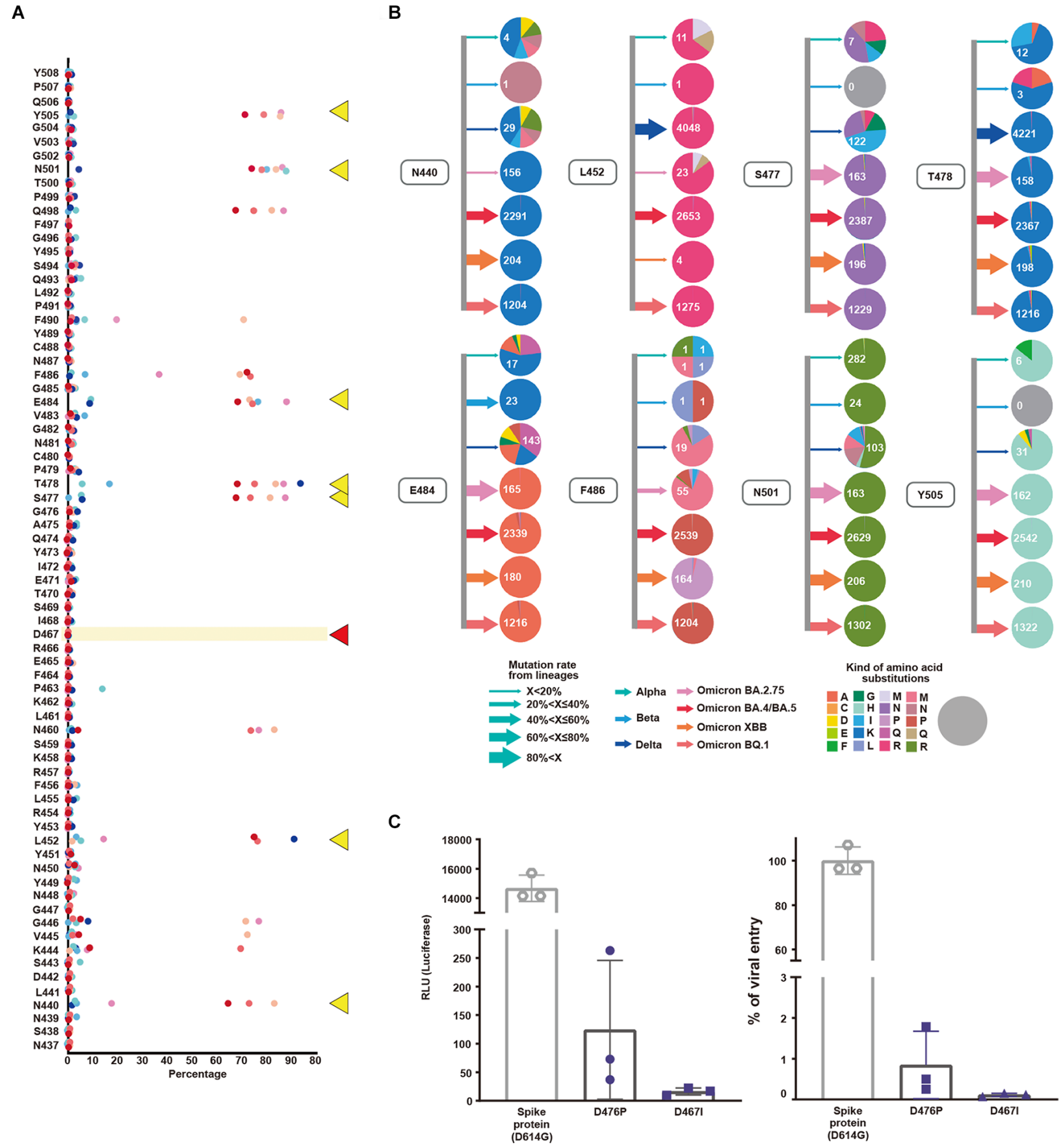
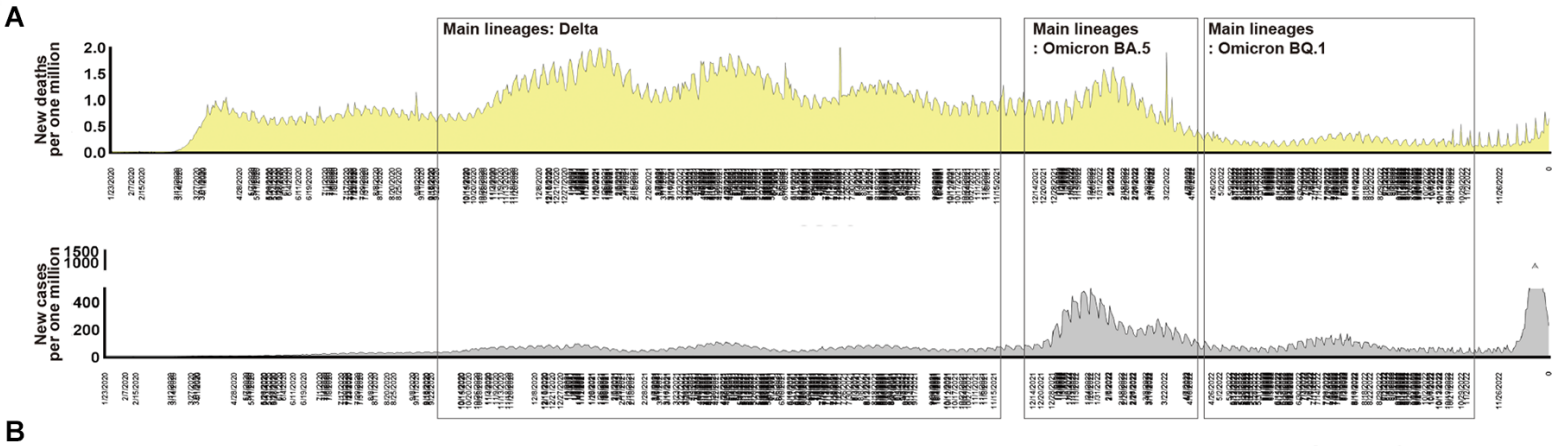
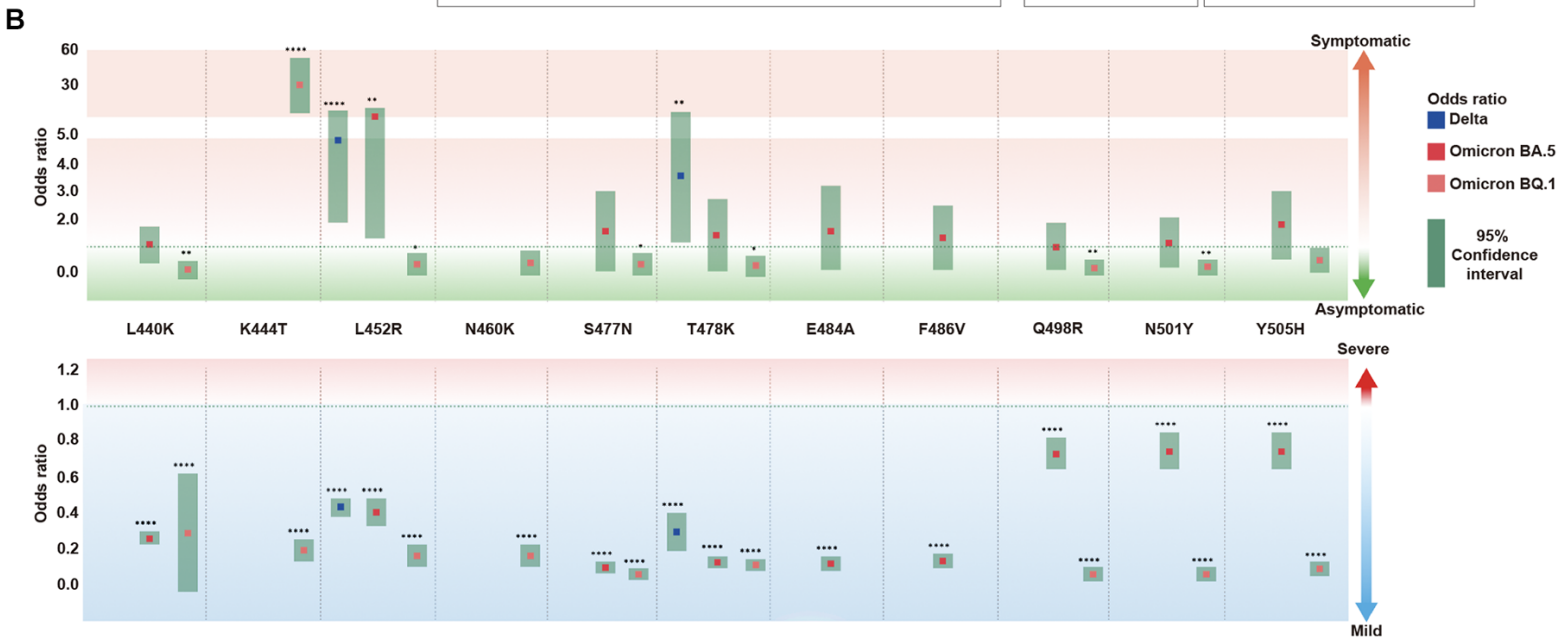

A
Alpha
Beta
Delta
Omicron BA.1
Omicron BA.2
Omicron BA.4/5
Omicron BQ.1
Omicron XBB
Hydrophilic (polar)
Hydrophobic (nonpolar)
Acidic
Alkaline

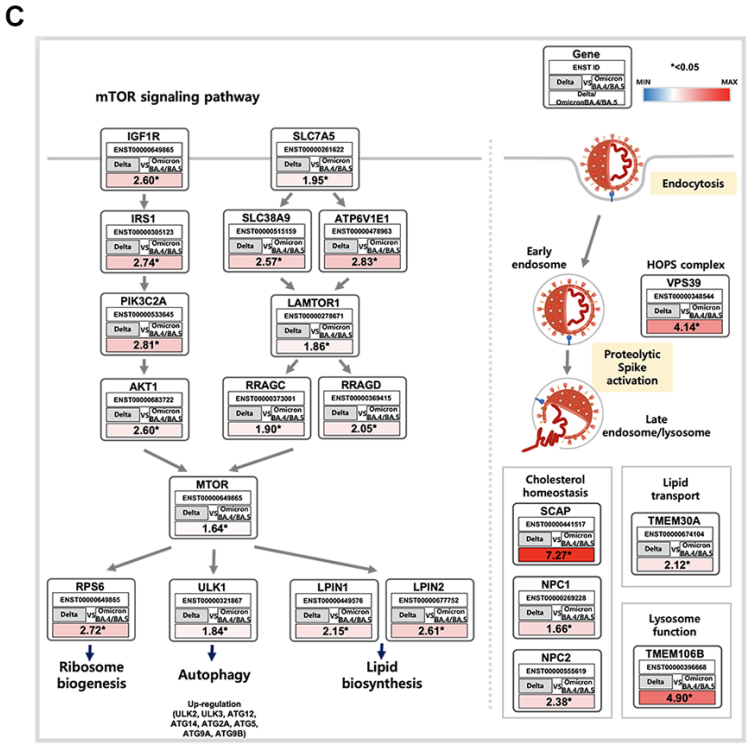

B
The number of polarity changes
The number of tpolarity changes
NP*
PN*
PP*
NN*
NN*
NP*
NA*
NB*
PN*
PP*
PA*
PB*
AN*
AP*
AA*
AB*
BN*
BP*
BA*
BB*
Wuhan-Hu-1
Alpha
Beta
Delta
Omicron BA.1
Omicron BA.2
Omicron BA.2.75
Omicron BA.4/BA.5
Omicron XBB
Omicron BQ.1
N: Hydrophobic (Non-polar)
P: Hydrophilic (Polar)
A: Acidic
B: Alkaline
*: random

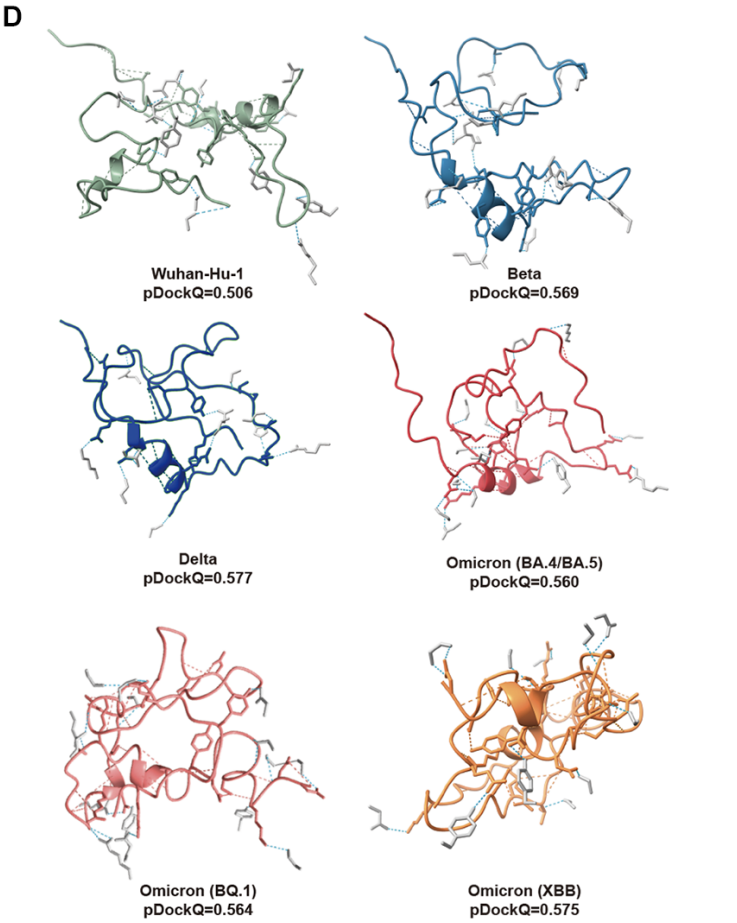

C
Reference
VOCs


A
Y508
P507
Q506
Y505
G504
V503
G502
N501
T500
P499
Q498
F497
G496
Y495
S494
Q493
L492
P491
F490
Y489
C488
N487
F486
G485
E484
V483
G482
N481
C480
P479
T478
S477
G476
A475
Q474
Y473
I472
E471
T470
S469
I468
D467
R466
E465
F464
P463
K462
L461
N460
S459
K458
R457
F456
L455
R454
Y453
L452
Y451
N450
Y449
N448
G447
G446
V445
K444
S443
D442
L441
N440
N439
S438
N437
0 10 20 30 40 50 60 70 80 90 80
Percentage
B
N440
4
1
29
156
2291
204
1204
L452
11
1
4048
23
2653
4
1275
S477
7
0
122
163
2387
196
1229
T478
12
3
4221
158
2367
198
1216
E484
17
23
143
165
2339
180
1216
F486
1 1
1 1
1 1
19
55
2539
164
1204
N501
282
24
103
163
2629
206
1302
Y505
6
0
31
162
2542
210
1322
Mutation rate from lineages
X<20%
20%<X≤40%
40%<X≤60%
60%<X≤80%
80%<X
Alpha
Beta
Delta
Omicron BA.2.75
Omicron BA.4/BA.5
Omicron XBB
Omicron BQ.1
Kind of amino acid substitutions
A C D E F G H I K L M N P Q R M N P Q R
C
RLU (Luciferase)
18000
16000
14000
300
250
200
150
100
50
0
Spike protein (D614G)
D476P
D467I
% of viral entry
100
80
60
3
2
1
0
Spike protein (D614G)
D476P
D467I


A
Main lineages: Delta
Main lineages : Omicron BA.5
Main lineages : Omicron BQ.1
New deaths per one million
New cases per one million


B
Odds ratio
Symptomatic
Asymptomatic
Odds ratio
Delta
Omicron BA.5
Omicron BQ.1
95% Confidence interval
L440K
K444T
L452R
N460K
S477N
T478K
E484A
F486V
Q498R
N501Y
Y505H
Severe
Mild


C
mTOR signaling pathway
Gene
ENST ID
Delta vs Omicron BA.4/BA.5
Delta vs OmicronBA.4/BA.5
*<0.05
MIN
MAX
IGF1R
2.60*
SLC7A5
1.95*
IRS1
2.74*
SLC38A9
2.57*
ATP6V1E1
2.83*
PIK3C2A
2.81*
LAMTOR1
1.86*
AKT1
2.60*
RRAGC
1.90*
RRAGD
2.05*
MTOR
1.64*
RPS6
2.72*
ULK1
1.84*
LPIN1
2.15*
LPIN2
2.61*
Ribosome biogenesis
Autophagy
Lipid biosynthesis
Up-regulation (ULK2, ULK3, ATG12, ATG14, ATG2A, ATG5, ATG9A, ATG9B)
Endocytosis
Early endosome
HOPS complex
VPS39
4.14*
Proteolytic Spike activation
Late endosome/lysosome
Cholesterol homeostasis
SCAP
7.27*
NPC1
1.66*
NPC2
2.38*
Lipid transport
TMEM30A
2.12*
Lysosome function
TMEM106B
4.90*


D
Wuhan-Hu-1
pDockQ=0.506
Beta
pDockQ=0.569
Delta
pDockQ=0.577
Omicron (BA.4/BA.5)
pDockQ=0.560
Omicron (BQ.1)
pDockQ=0.564
Omicron (XBB)
pDockQ=0.575

A

INPUT

SARS-CoV-2 receptor binding motif (RBM, 72 AA)
NSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY

SCPS (Sub-Clustering of Protein Structure)

① Protein structure prediction (AlphaFold)

pLDDT — Median
BA.5 73.33
Delta 69.99
Alpha 55.91
Wuhan-Hu-1 46.53

② Clustering (pLDDT & PAE)

T1 T2 T3 T4 T5 T6 T4

Analysis

$$SCPS_i = \begin{cases} 0.9 \text{ if } SCPS_i \in \{T2, T4\} \\ 0.1 \text{ if } SCPS_i \in \{T1, T3, T5, T6\} \end{cases}$$

PCS (Polaritvy Change Score)

Hydrophobic, Hydrophilic, Alkaline, Acidic

Wuhan-Hu-1 → Lineages

Analysis

① Nucleotide

$$PCS_i(w,m) = \begin{cases} 0.9 & \text{if } PCS_i(w_i) \in PP^* \wedge PCS_i(m_i) \notin PP^* \\ 0.5 & \text{if } PCS_i(w_i) \notin PP^* \wedge PCS_i(m_i) \in PP^* \\ 0.1 & \text{if otherwise} \end{cases}$$

Wuhan-Hu-1 n=17
Alpha n=17
Delta n=16
Omicron BA5 n=14

MR (Mutation Rate)

① Nucleotide Rate — Wild, Mut

② Codon — 1st 2nd 3th: 28% 49% 22%

Mutation rate

Analysis

$$MR_i(w,m) = \frac{mr(w_i) - mr(m_i)}{mr(w_i)}$$

$$mr(x) = \prod_{j=1} N_j(x) \times C_j(x)$$

MR: Wuhan-Hu-1 0.2, Alpha 3.3, Delta 11.2, Omicron BA.5

BPES (Biochemical Properties Eigen Score)

Amino acid property — PC1, PC2, H Residue, pH, Hydrophobic

Analysis

$$BPES_i(w,m) = \frac{bpes(w_i) - bpes(m_i)}{bpes(w_i)}$$

$$bpes(x) = \|R(x)\| \frac{R(x) \cdot P(x)}{\|R\| \cdot \|P(x)\|} + \|P(x)\| + \|H(x)\| \frac{H(x) \cdot P(x)}{\|H\| \cdot \|P(x)\|}$$

BPES: Wuhan-Hu-1 0, Alpha -1.31, Delta 27.2, Omicron BA.5 46.2

APESS

Wuhan-hu-1 / Delta / Omicron BA.5 — bpes, mr, pcs, scps

$$APESS(w,m) = \sum_{i=1}^{n} SCPS_i(w) \times PCS_i(w,m) \times MR_i(w,m) \times BPES_i(w,m)$$

Density(scaled): Beta 0.006, Alpha -0.001, Delta 1.640, XBB 1.967, BA.2 2.053, BA.1 1.950, BQ.1 2.080, BA.2.75 1.927, BA.5 2.154

B

Omicron BQ.1 2.080 — BQ.1.3(+S477K) 4.384

Omicron XBB 1.967 — XBB1.5 (+S477K, +F486P) 4.287

Omicron BA.5 2.154 — BA.5.1, BA.5.2.1, BA.5.2.34, BA.5.6 (+S477K) 4.458

Omicron BA.2.75 1.927 — BA.2.75.5 (+S477K) 4.318

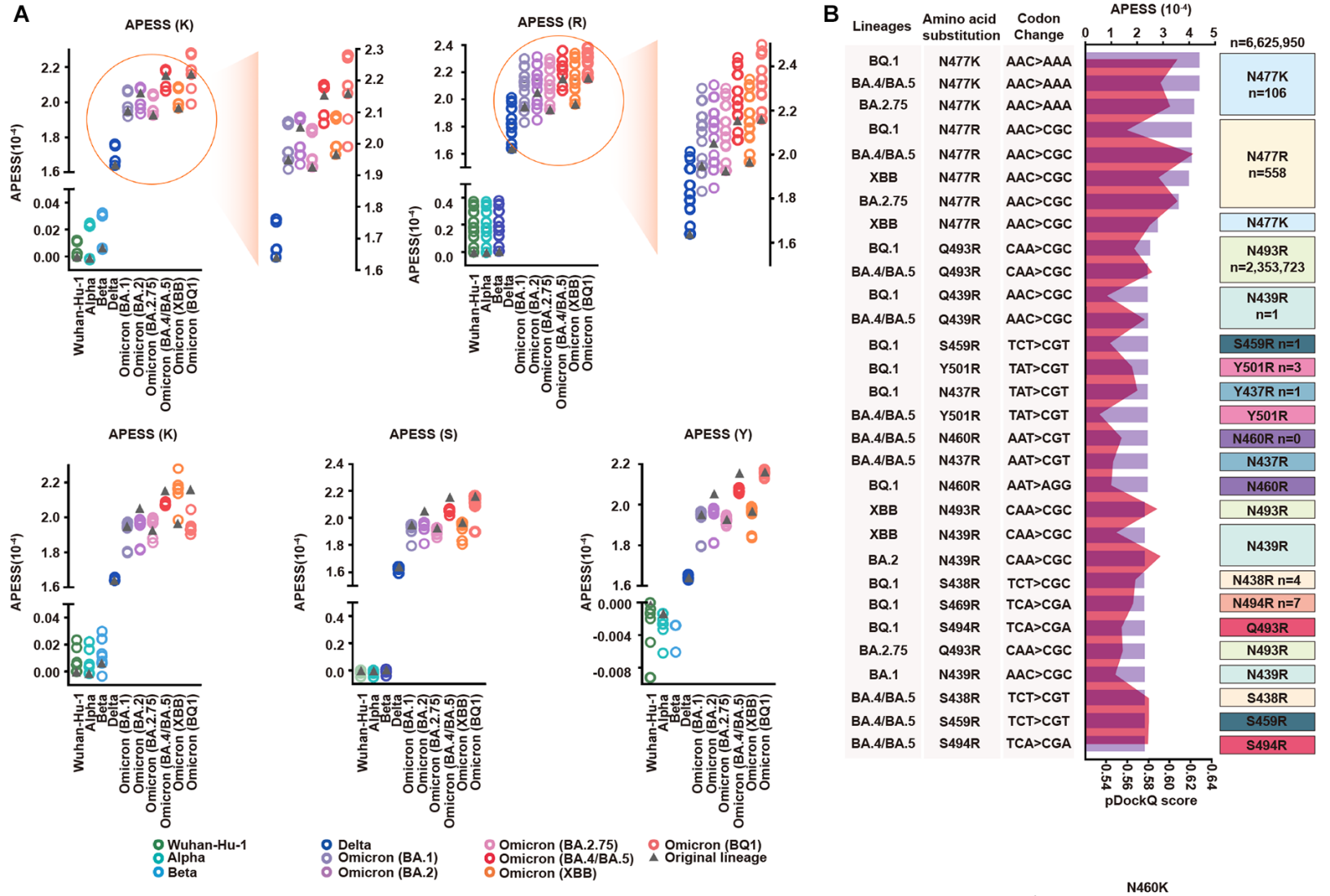
Delta 1.640 — AY.95 (+S477R, +Y508H) 3.64

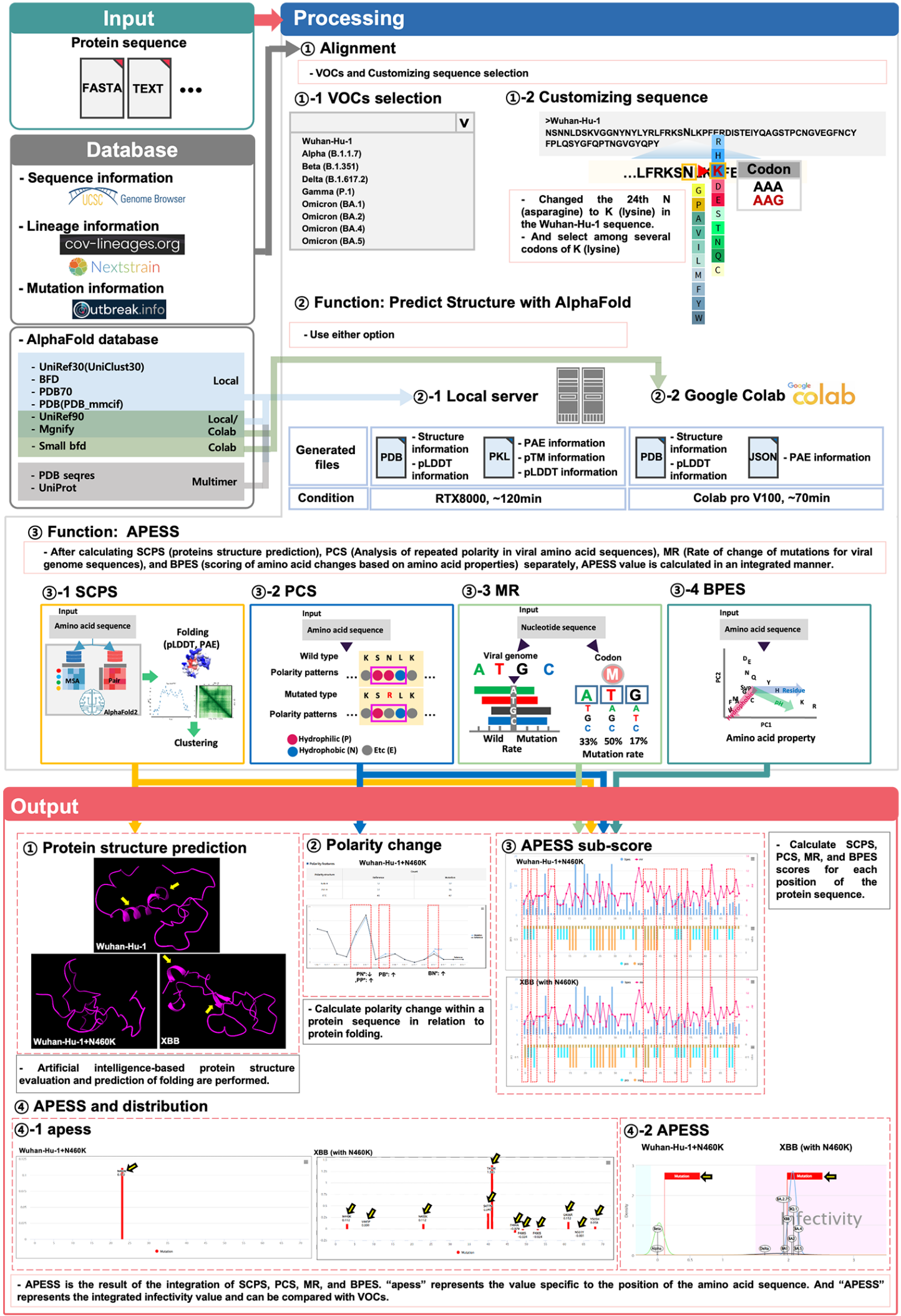
Beta 0.006 — B.1.351 (+L452R, +T478K)

Alpha -0.001 — B.1.1.7 (+N440K, +L452Q, +S477N, +T478K, +E484A, +Q493R, +Q498R, +Y505H)

Alpha, Beta, Delta, Omicron BA.2.75, Omicron BA.5, Omicron XBB, Omicron BQ.1, main lineage

APESS


A
APESS (K)
APESS (R)
APESS (S)
APESS (Y)
APESS(10⁻⁴)
Wuhan-Hu-1
Alpha
Beta
Delta
Omicron (BA.1)
Omicron (BA.2)
Omicron (BA.2.75)
Omicron (BA.4/BA.5)
Omicron (XBB)
Omicron (BQ1)
Wuhan-Hu-1
Alpha
Beta
Delta
Omicron (BA.1)
Omicron (BA.2)
Omicron (BA.2.75)
Omicron (BA.4/BA.5)
Omicron (XBB)
Omicron (BQ1)
Original lineage
B
Lineages
Amino acid substitution
Codon Change
APESS (10⁻⁴)
n=6,625,950
BQ.1 N477K AAC>AAA
BA.4/BA.5 N477K AAC>AAA
BA.2.75 N477K AAC>AAA
BQ.1 N477R AAC>CGC
BA.4/BA.5 N477R AAC>CGC
XBB N477R AAC>CGC
BA.2.75 N477R AAC>CGC
XBB N477R AAC>CGC
BQ.1 Q493R CAA>CGC
BA.4/BA.5 Q493R CAA>CGC
BQ.1 Q439R AAC>CGC
BA.4/BA.5 Q439R AAC>CGC
BQ.1 S459R TCT>CGT
BQ.1 Y501R TAT>CGT
BQ.1 N437R TAT>CGT
BA.4/BA.5 Y501R TAT>CGT
BA.4/BA.5 N460R AAT>CGT
BA.4/BA.5 N437R AAT>CGT
BQ.1 N460R AAT>AGG
XBB N493R CAA>CGC
XBB N439R CAA>CGC
BA.2 N439R CAA>CGC
BQ.1 S438R TCT>CGC
BQ.1 S469R TCA>CGA
BQ.1 S494R TCA>CGA
BA.2.75 Q493R CAA>CGC
BA.1 N439R AAC>CGC
BA.4/BA.5 S438R TCT>CGT
BA.4/BA.5 S459R TCT>CGT
BA.4/BA.5 S494R TCA>CGA
N477K n=106
N477R n=558
N477K
N493R n=2,353,723
N439R n=1
S459R n=1
Y501R n=3
Y437R n=1
Y501R
N460R n=0
N437R
N460R
N493R
N439R
N438R n=4
N494R n=7
Q493R
N493R
N439R
S438R
S459R
S494R
pDockQ score
C
Probability
N440K
V445P
G446S
N460K
S477N
T478K
E484A
F486P
F490S
Q498R
N501Y
Y505H
L452R
E484K
Q493R
G496S
F486V
K444T
S494R
T500R
Q493K
T470K
N439K
N460R
Before omicron
Omicron BA.2
Omicron BA.4
Omicron BA.5
Omicron XBB
Omicron BQ.1
N460K
Omicron XBB
Omicron BQ.1
Omicron BA.5
Omicron BA.4
Omicron BA.2
Before omicron
Q493R
Omicron XBB
Omicron BQ.1
Omicron BA.5
Omicron BA.4
Omicron BA.2
Before omicron
D
RLU (Lucifierase)
% of viral entry
Spike protein (D614G)
N437R
N460K


Input
Protein sequence
FASTA
TEXT
Database
- Sequence information
UCSC Genome Browser
- Lineage information
cov-lineages.org
Nextstrain
- Mutation information
outbreak.info
- AlphaFold database
- UniRef30(UniClust30)
- BFD
- PDB70
- PDB(PDB_mmcif)
Local
- UniRef90
- Mgnify
Local/ Colab
- Small bfd
Colab
- PDB seqres
- UniProt
Multimer
Processing
① Alignment
- VOCs and Customizing sequence selection
①-1 VOCs selection
Wuhan-Hu-1
Alpha (B.1.1.7)
Beta (B.1.351)
Delta (B.1.617.2)
Gamma (P.1)
Omicron (BA.1)
Omicron (BA.2)
Omicron (BA.4)
Omicron (BA.5)
①-2 Customizing sequence
>Wuhan-Hu-1
NSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCY
FPLQSYGFQPTNGVGYQPY
...LFRKSNLKFE
Codon
AAA
AAG
- Changed the 24th N (asparagine) to K (lysine) in the Wuhan-Hu-1 sequence.
- And select among several codons of K (lysine)
② Function: Predict Structure with AlphaFold
- Use either option
②-1 Local server
②-2 Google Colab
Generated files
PDB
- Structure information
- pLDDT information
PKL
- PAE information
- pTM information
- pLDDT information
PDB
- Structure information
- pLDDT information
JSON
- PAE information
Condition
RTX8000, ~120min
Colab pro V100, ~70min
③ Function: APESS
- After calculating SCPS (proteins structure prediction), PCS (Analysis of repeated polarity in viral amino acid sequences), MR (Rate of change of mutations for viral genome sequences), and BPES (scoring of amino acid changes based on amino acid properties) separately, APESS value is calculated in an integrated manner.
③-1 SCPS
Input
Amino acid sequence
MSA
Pair
AlphaFold2
Folding (pLDDT, PAE)
Clustering
③-2 PCS
Input
Amino acid sequence
Wild type
Polarity patterns
Mutated type
Polarity patterns
Hydrophilic (P)
Hydrophobic (N)
Etc (E)
③-3 MR
Input
Nucleotide sequence
Viral genome
Codon
Wild
Mutation
Rate
33% 50% 17%
Mutation rate
③-4 BPES
Input
Amino acid sequence
Amino acid property
Output
① Protein structure prediction
Wuhan-Hu-1
Wuhan-Hu-1+N460K
XBB
- Artificial intelligence-based protein structure evaluation and prediction of folding are performed.
② Polarity change
Wuhan-Hu-1+N460K
- Calculate polarity change within a protein sequence in relation to protein folding.
③ APESS sub-score
Wuhan-Hu-1+N460K
XBB (with N460K)
- Calculate SCPS, PCS, MR, and BPES scores for each position of the protein sequence.
④ APESS and distribution
④-1 apess
Wuhan-Hu-1+N460K
XBB (with N460K)
④-2 APESS
Wuhan-Hu-1+N460K
XBB (with N460K)
Infectivity
- APESS is the result of the integration of SCPS, PCS, MR, and BPES. "apess" represents the value specific to the position of the amino acid sequence. And "APESS" represents the integrated infectivity value and can be compared with VOCs.

AIVE: accurate predictions of SARS-CoV-2 infectivity from comprehensive analysis

Jongkeun Park[1,*], Won Jong Choi[1,*], Do Young Seong[1,*], Seung Pil Jeong[1], Ju Young Lee[1], Hyo Jeong Park[1], Dae Sun Chung[1], Ki Jong Yi[2], Uijin Kim[3], Ga-Yeon Yoon[3], Hyeran Kim[4,5], Taehoon Kim[4,5], Sooyeon Go[6], Eun Jeong Min[7], Hyun-Soo Cho[3], Nam-Hyeok Cho[4,5,8], and Dongwan Hong[1,9,10,11#]

Affiliations:

1. Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea
2. Graduate School of Medical Science and Engineering, Korea Advanced Institute and Technology (KAIST), Daejeon 34141, Republic of Korea
3. Department of Systems Biology, College of Life Science and Biotechnology, Yonsei University, Seoul 03722, Republic of Korea
4. Department of Microbiology and Immunology, Seoul National University College of Medicine, Seoul 03080, Republic of Korea
5. Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul 03080, Republic of Korea
6. School of Chemical and Biological Engineering, Seoul National University, Seoul 08826, Republic of Korea
7. Department of Medical Life Sciences, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea
8. Seoul National University Bundang Hospital, Seongnam, Gyeonggi-do 13620, Republic of Korea

9. Precision Medicine Research Center, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea
10. Cancer Evolution Research Center, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea
11. CMC Institute for Basic Medical Science, the Catholic Medical Center of The Catholic University of Korea, Seoul 06591, Republic of Korea

* Co-first authors with equal contribution

Correspondence

Address for correspondence
Dongwan Hong, Ph.D.
Professor
College of Medicine, The Catholic University of Korea
222 Banpo-daero Seocho-gu, Seoul 06591, Republic of Korea
E-mail: dwhong@catholic.c.kr
Tel: +82-2-3147-8424
Fax: +82-2-2258-7749

Abstract

This study presents an innovative research model utilizing big data science and protein structure prediction AI software. An unprecedented amount of SARS-CoV-2 data has been accumulated compared with previous infectious diseases, enabling insights into its evolutionary process and more thorough analyses. We identified amino acid substitutions ranging from hydrophilic to hydrophobic, or positively charged amino acids in the RBM region. An increased frequency of amino acid substitutions to lysine (K) and arginine (R) was detected in Variants of Concern (VOCs) and viral sequencing data. As the virus evolved to Omicron, commonly occurring mutations became fixed components of the new viral sequence. Furthermore, in specific positions, only one type of amino acid substitution and a notable absence of mutations at D467 was detected across viral sequences in VOCs. The binding affinity with the ACE2 receptor increased for later lineages. We developed APESS, a mathematical model evaluating infectivity based on biochemical and mutational properties calculated from a protein prediction of AlphaFold. We validated discoveries of features found through APESS. Infectivity was evaluated in silico using real-world viral sequences and in vitro viral entry assays. Using Machine Learning, we predicted mutations that had the potential to become more prominent. APESS and characteristics we discovered are featured in AIVE, a web-based system, accessible at https://ai-ve.org. AIVE provides an infectivity measurement of mutations entered by users which is available on fast APESS calculations and visualization of results without GPU installation. We established a clear link between specific viral properties and increased infectivity. Comprehensive analysis and specialized AIVE reporting enhance our understanding of SARS-CoV-2 and enable more accurate predictions of infectivity.

Keywords: AIVE, APESS, Protein Prediction, Infectivity, and SARS-CoV-2

Introduction

The importance of big data science has led to the accumulation of huge amounts of data, recently, a generative AI model with large language model (LLM) using various types of big data has further been emerging across various fields, including healthcare, omics research, and industry[1,2]. In response to the continuous emergence of new SARS-CoV-2 variants, our study introduces an innovative model and platform that integrate big data analysis, protein structure prediction using AlphaFold, and AI learning to predict highly transmissible novel SARS-CoV-2 variants. The validation of predictions made using AI is crucial, so we present the proofs through in-silico and in-vitro experiments for highly infectious variants.

The COVID-19 pandemic-related deaths have decreased, but the infectivity of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) persists. This may be because of neutralizing antibodies from previous infections or numerous mutations in the virus, from the Alpha variant to current Omicron sublineages[3,4].

Genomic databases, such as Nextstrain and the Global Initiative on Sharing All Influenza Data (GISAID), Our World in Data, containing epidemiological data, have been instrumental in collecting valuable data during the COVID-19 pandemic[5-7]. Reports on the current global efforts of genomic surveillance strategies and sequencing show varied levels of data accumulation[8]. Nonetheless, the increasing amount of available SARS-CoV-2 data has made various analytical approaches possible.

SARS-CoV-2 mutations can alter gene function. For example, the SARS-CoV-2 furin gene plays an important role in the cleavage of the spike protein, and mutations in the gene significantly affect SARS-CoV-2 fusion with the host cell membrane[9,10]. Mutations in the genes have a direct impact on its protein structure, influencing the pathway used by the virus to infect host cell[11]. Mutations in the spike protein leads to differences in the protein structure and binding affinity to ACE2, the receptor through which the virus penetrates the host cell[12,13].

Deep learning methods predict mutations and protein structures, and their application in research has led to enhanced genomic analysis with large datasets[14,15]. This approach has proved valuable in predicting disease-related mutations and identifying genes of interest[16].

Recently, deep learning has been utilized to analyze mutations in SARS-CoV-2[17,18]. Prediction models, such as AlphaFold2, have been used to assess protein 3D structures[19]. This has led to the detection of structural changes caused by mutations in Delta (B.1.617.2) and Omicron (B.1.1.529) [20]. Alphafold2 has also been used to analyze protein genotypes and phenotypes in the RBD region[21].

Prior approaches may not be intuitive without expertise in the field. Also, structure-based predictions of mutations cannot fully account for all aspects of SARS-CoV-2 infection. In this study, we analyzed SARS-CoV-2 mutations using artificial intelligence (AI) methods and leveraged large datasets to make various predictions. We discovered amino acid polarity changes and substitutions then evaluated the infectivity from significant mutations in the receptor binding motif (RBM) region. Hydrophobic properties in the amino acid sequence affects protein folding[22]. Coronavirus hydrophobicity has significant effects on amino acid properties and protein folding[23]. Therefore, we specifically examined the effect of protein structures from hydrophobicity to hydrophilic and alkaline properties.

Compared with prior approaches, our evaluation involved a comprehensive analysis of epidemiological, and genomic data, capitalizing on the availability of large SARS-CoV-2 datasets. We extracted properties from the VOCs and each sublineage at the nucleotide and amino acid levels. We developed an evaluation model called Amino acid Property Eigen Selection Score (APESS) to calculate SARS-CoV-2 sequences. Various methods were utilized to validate our findings and we used machine learning to make further predictions.

Finally, we present our findings and evaluation models through Artificial Intelligence Analytics Toolkit for predicting virus mutations in protEin (AIVE), a web-based platform that integrates SARS-CoV-2 data, visualizes APESS score distribution, offers 3D protein predictions, and increases accessibility for researchers. Overall, our research provides a model for pandemic preparedness and the study of infectious diseases.

Results

Discovery of significant properties in the amino acid sequence

We identified consecutive hydrophilic amino acids within the SARS-CoV-2 spike protein in the following lineages: Wuhan-Hu-1, Alpha, Beta, and Omicron (BA.1, BA.2, BA.2.75, BA.4/BA.5, XBB, BQ.1). A series of hydrophilic amino acid regions were observed in the RBM region (amino acid sequence 437–508). In specific SARS-CoV-2 lineages, amino acid substitutions were observed in these regions (Fig. 1A and Supplementary Fig. 1). For example, the Delta and Omicron lineages contained a substitution of T478K of alkaline amino acids.

Each position showed differences in the number of polarity changes; specifically, changes from hydrophilic to hydrophobic or positively charged amino acids. For most positions, we did not detect significant differences in polarity between the lineages investigated (Supplementary Tables 1-3).

As the virus progressed from the Wuhan-Hu-1 (40 in NN* and 31 in PP*) to Omicron (average 45.5 in NN* and average 27 in PP*), the number of polarity changes increased in NN* while decreasing in PP*. For these positions, polarity changes occurred in the lineages as they evolved to the Omicron variant (Fig. 1B). More polarity changes occurred in Delta, XXB, Omicron BQ.1, and Omicron BA.4/BA.5 than in the other VOCs.

We investigated the amino acid substitutions in the VOCs and compared them with the Wuhan-Hu-1 sequence as the reference. We discovered a two-fold increase in amino acid substitutions to lysine (K) and arginine (R) in later SARS-CoV-2 lineages. Meanwhile, glutamine (Q), phenylalanine (F), and glutamic acid (E) levels decreased by half in the VOCs. For phenylalanine (F), hydrophobic residue mutations occurred in areas irrelevant to the regions of consecutive hydrophilic amino acids (Fig. 1C and Supplementary Table 4 and Supplementary Fig. 2).

Various coronaviruses, including MERS-CoV, SARS-CoV-1, and SARS-CoV-2, did not show significant differences in polarity across positions (Supplementary Fig. 3A).

We investigated the polarity (hydrophilic, hydrophobic, alkaline, and acidic) of amino acids in Wuhan-Hu-1, Alpha, Beta, Omicron (BA.1, BA.2, BA.2.75, BA.4/BA/5, XBB, BQ.1), and the number of amino acids in the RBM sequence. Most amino acids were either hydrophilic or hydrophobic across lineages, with a slight increase in alkaline amino acid levels in the Omicron variants than in the others (Supplementary Fig. 3B).

We investigated the nucleotide sequences of the SARS-CoV-2 spike protein gene. We analyzed the transitions and transversions of 7,335,614 samples from the GISAID viral sequence. The rate of change from G to U was higher than that of U to G, and the rate from C to U was higher than that of U to C. Both C to G and G to C rates were extremely low (Fig. 2C and Supplementary Fig. 3C and Supplementary Table 5) [24,25].

Identification of properties in amino acid substitutions

We studied each mutation in the RBM region of lineages and sublineages using verified mutation data: Alpha, Beta, Delta, Omicron (BA.2.75), Omicron (BA.4/BA.5), Omicron (XBB), and Omicron (BQ.1).

From the viral sequences submitted to GISAID, we extracted data (n = 7,335,614) of the sublineages defined on the outbreak.info platform. The most noteworthy mutations were N440, L452, S477, T478, E484, F486, N501, and Y505 (Fig. 2A). Only seven individuals possessed mutations at D467 out of the 7,335,614 investigated (Supplementary Fig. 4A).

Importantly, the molecular work results and amino acid substitutions from the viral sequences were the same. Almost no mutations were found at the D467 position of the RBM sequence. In vitro experiments using a luciferase assay and viral entry experiments demonstrated that infectivity associated with D467 mutagenesis (D467P and D467I) was lower than the wild type (spike protein D614G) (Fig. 2C and Supplementary Fig. 5).

We investigated the mutation rates and amino acid substitutions in VOCs and variants under monitoring (VUMs). In the earlier Alpha, Beta, and Delta lineages, various amino acid changes occurred at multiple locations. In later lineages, we observed a reduction in the diversity of amino acid changes due to substitutions (Fig. 2B). L452R and T478K were the most numerous non-synonymous mutations while more synonymous mutations developed.

Differences in molecular cell levels that affect infectivity and severity in the evolution of SARS-CoV-2

We investigated the molecular cell level differences that affect the severity and infectivity of SARS-CoV-2 in its evolutionary process. From 1/23/2020 to 12/31/2022, we examined the number of infections and deaths using data from OWID and compared the periods where lineages occurred. Based on epidemiological data, we established three main periods: the major outbreak period of Delta, Omicron BA.5, and Omicron BQ. During the Delta period, there was an increase in the ratio of deaths to infections, for the Omicron BA.5 period, both infections and deaths increased, and finally in the Omicron BQ period, both decreased (Fig. 3A). We attributed these trends to mutations in the major lineages and further investigation of sublineages through viral sequence information revealed stabilization per lineage (Supplementary Tables 6 and 7).

We investigated the epidemiological relationship between symptoms and disease severity in patients by examining mutations in the RBM region for major SARS-CoV-2 lineages and their sublineages. We found that these mutations became fixed over time (Fig. 3B and Supplementary Tables 8 and 9). In the Delta variant, a significant increase in symptomatic infections occurred for mutations L452R (odds ratio 4.346, 95% CI 2.378-8.191) and T478K (odds ratio 3.116, 95% CI 1.595-6.292). For most patients with Omicron BA.5 and BQ.1, mutations were asymptomatic. However, among patients with Omicron BQ.1 mutations, only those with K447T were symptomatic. The Delta, Omicron BA.5, and Omicron BA.1 variants more frequently produced milder outcomes with more mutations. Consequently, we showed that as SARS-CoV-2 evolves to Omicron, the number of asymptomatic patients with mild outcomes increased, with infection symptoms becoming less severe.

Since we cannot determine the severity and presence of symptoms in SARS-CoV-2 from mutations alone, we analyzed gene expression within host cells of infected patients. We utilized GSE235262 from the NCBI GEO database to analyze the gene expression of controls (uninfected), Alpha (B.1.1.7), Delta (B.1.617.2), and Omicron (B.1.529, BA.2, BA.4, and BA.5) variants during waves of COVID-19. In particular, we focused on the expression differences between Alpha (B.1.1.7) and Omicron (BA.4/BA.5), and between Delta (B.1.617.2) and Omicron (BA.4/BA.5), conducting enrichment analyses on genes that showed significant differences

(Supplementary Table 10). Compared to Omicron (BA.4/BA.5), most genes in Delta (B.1.617.2) were up regulated in molecular and virus infection pathways ($p<0.05$). Notably, in the mTOR pathway, there was a correlation with receptor genes leading to the PI3K-AKT pathway, lysosome, ribosome biogenesis, autophagy, and lipid biosynthesis, In the virus infection pathway, endocytosis and metabolic pathways were related (Fig. 3C). Compared to Omicron (BA.4/BA.5), Delta is more associated with various pathways within the host cell, leading to inflammation, replication, and severity.

We believe that Omicron's characteristics are not solely due to its molecular cell level features within the host cell but also stem from its evolution, affecting its binding affinity to the host's ACE2. Therefore, we utilized pDockQ and HADDOCK to predict the binding affinity between SARS-CoV-2 and the ACE2 receptor affected by mutations[26,27]. Based on the pDockQ results, with Wuhan-Hu-1 as the standard, the SARS-CoV-2 variants showed strong affinity in descending order of Delta (0.577), Omicron XBB (0.575), Beta (0.569), Omicron BQ.1 (0.564), and Omicron BA.4/BA.5 (0.560) (Fig. 3D and Supplementary Fig. 6). Compared to the binding affinity of Wuhan-Hu-1 (0.506), the lineages with higher binding affinity being associated with infectivity (Supplementary Tables 11-13 and Supplementary Fig. 7).

The results showed the following descending order of binding affinities for the ACE2 receptor: Delta, XBB, Beta, BQ.1, BA.4/BA.5, BA.1/BA.2, Alpha, and Wuhan-Hu-1. While Beta had a high affinity score, it has had a low occurrence rate since its emergence, and it did not receive a high score from our evaluation model.

To check for interspecies infection in humans, bats, and pangolins, we measured the bond affinity between the virus and host. We also measured the bond affinity of SARS-CoV-1. The bond strength of SARS-CoV-2 to the Homo sapiens ACE2 receptor is weaker than that of SARS-CoV-1[28]. Although the bond strength with the same host is higher, it weakens with a different host. This suggests that viruses that evolve within a human host are likely to be the most infectious to humans. As SARS-CoV-2 evolved to Omicron, the binding affinity increased for the virus.

Overall, the Delta variant was found to be more associated with the host molecular pathway and severity. Meanwhile, the Omicron variant showed a higher interaction between the ACE2 and RBM region. The Omicron variant is more infectious.

Development of APESS, an evaluation scoring model and the evaluation of lineages

We discovered consecutive amino acids along with amino acid substitutions to lysine (K) and arginine (R). Based on these findings, we developed APESS, an evaluation model, to analyze viral sequences. APESS was calculated from four previous calculations: sub-clustering of protein structure (SCPS), polarity change score (PCS), mutation rate (MR), and biochemical properties eigen score (BPES) (Fig. 4A and Supplementary Tables 14-17 and Supplementary Fig. 8). APESS evaluates the infectivity of the SARS-CoV-2 lineage. Lower APESS scores indicated that the lineage is less infectious, whereas higher APESS scores indicated that it may be more infectious.

We calculated the APESS scores of 7,335,614 viral sequences encompassing VOCs and their sublineages (Fig. 4B). The Alpha and Beta variants received lower APESS scores, whereas the Delta and Omicron variants (BA.5, BQ.1, and XBB) received higher scores.

We generated an APESS distribution graph for the VOC sublineages. Significant variants, including Delta and Omicron, showed APESS values higher than 1.62, indicating high infectivity (Fig. 4A and Supplementary Fig. 9). The APESS scores of the sublineages were similar to those of the VOCs. Some sublineages had lower APESS scores because of the wide variety of amino acid substitutions in the RBM region (Fig. 4B). The sublineages with the highest APESS scores possessed amino acid substitutions to lysine (K) and arginine (R) at S477.

Based on cumulative prevalence results from investigations on 21 March 2023 (outbreak.info), 2% cumulative prevalence of the S477K mutation was confirmed in the Omicron variants BA.5.1 and BA.5.2. Severe infections were associated with amino acid substitutions to lysine (K) and arginine (R) at the S477 position.

Utilizing a gaussian mixture model (GMM), we divided the APESS scores of 30,000 randomly selected sublineages into four components. Four groups were identified, G1 (Alpha and Beta: a centroid at 0.0289), G2 (BA.1, BA.2.75, and XBB: a centroid at 1.9474), G3 (Delta: a centroid at 1.6383). G4 (BA.2, BA.4/5 and BQ.1: a centroid at 2.0802). We also examined the distribution for all 7,335,614 lineages and found that the results were the same as the 30,000 sampled

sublineages. We observed that most of Alpha and Beta were included in the component with a centroid of 0.0289, while the component with a centroid of 2.0627 included various Omicron variants.

Creation of candidate lineages and protein structure predictions

Candidate lineages were created by performing amino acid substitutions at lysine (K) and arginine (R) to evaluate their association with infectivity. APESS scores of these sequences were higher than of the SARS-CoV-2 variants. In particular, S477K and S477R amino acid substitutions were linked to increased infectivity. In contrast, amino acid substitutions of asparagine (N), serine (S), tyrosine (Y), and glycine (G) resulted in APESS scores similar to or lower than of the SARS-CoV-2 variants (Fig. 5A).

Of the mutated sequences, amino acid substitutions of K and R showed the highest APESS and pDockQ scores (Fig. 5B).

We used deep learning and machine learning methods to determine the probability of amino acid substitutions at specific locations in the lineages before and after Omicron variant emergence. Models such as random forest, light-gradient boosting machine (LightGBM), extreme Gradient Boosting (XGBoost), and ensemble methods were used, and consistent results were obtained (Supplementary Fig. 10 and Supplementary Tables 18-22). Specifically, for N460K, we found that the probability increased as the lineage progressed to Omicron. For Q493R, the probability was relatively constant for Omicron lineages (Fig. 5C and Supplementary Table 23).

To validate the binding kinetics of Q493R and N460K variant RBDs to human ACE2 at the macromolecular level, we performed surface plasmon resonance (SPR). In line with our computational prediction, both variants show approximately a three-fold increase in binding affinity compared to the wild-type RBD (Supplementary Figs. 11 and 12).

We further verified the infectivity of the predicted mutations using luciferase and viral entry assays (Supplementary Fig. 5). N437R mutation led to a two-fold increase in luciferase activity

and viral entry compared to the wild-type (spike protein D614G). In contrast, N460K and Q493R mutations led to a decrease in luciferase activity and viral entry compared to the wild type (spike protein D614G) (Fig. 5D and Supplementary Fig. 11). However, viral entry of N460K was 10-fold higher than that of Q493R. N460K mutations occurred at a low rate but had a higher probability of infection compared to other mutations.

Therefore, our results imply that the maintenance of the proliferative rate of virus was due to new mutations, high docking scores, and an increased presence of Omicron variants (Supplementary Figs. 13-22).

Based on our findings, we developed AIVE (https://ai-ve.org/), a web-based platform enabling rapid and precise prediction of protein structure and calculation of SARS-CoV-2 infectivity. AIVE facilitates the analysis of virus sequences entered by users and provides visualization and analysis report, allowing it to be used in environments without GPU installation (Fig. 6). For example, we used a customized sequence wherein N460K substitution was introduced into the Wuhan-Hu-1 sequence (Wuhan-Hu-1+N460K) as an input. The output generated the following four results:

First, protein structure prediction results showed that Wuhan-Hu-1+N460K had increased folding compared to Wuhan-Hu-1. Meanwhile, XBB, which contains N460K substitution, demonstrated a more stabilized protein structure compared to Wuhan-Hu-1+N460K due to the alpha-helix in the 493-495 region and the beta-sheet in the 465-467 and 505-507 regions (Panel ① of Fig. 6. Second, for polarity changes, there was a decrease in consecutive hydrophilic amino acids in Wuhan-Hu-1+N460K due to the mutation (Panel ② of Fig. 6). Third, calculations of the APESS sub-scores showed a difference in scores for XBB compared to Wuhan Hu-1+N460K owing to mutations (Panel ③ of Fig. 6). Fourth, the APESS distribution was displayed as "apess", which reflects the score at each position in the amino acid sequence. The apess score at N460K for both Wuhan-Hu-1+N460K and XBB (with N460K) was 0.112 (Panel ④-1 of Fig. 6). Furthermore, the distribution of the total "APESS" score, which comprehensively evaluates each position of the sequencing results, was used to determine the infectivity sections. Usage guidelines of AIVE can be found on supplementary information (Supplementary Fig. 23).

An apess score at N460K for XBB (with N460K) was 1.967. XBB (with N460K) mutation was located in the infectivity region with a statistically significant association (> 95%, pink color),

whereas APESS score increased for Wuhan Hu-1+N460K compared to Wuhan Hu-1 but was not included in the high infectivity section containing XBB (Panel ④-2 of Fig. 6). Therefore, our research findings enable rapid protein structure prediction and APESS via meticulous and systematic structuring of data. Visualization of these analyses and the evaluation of infectivity are available on AIVE.

Discussion

In this study, we aimed to comprehensively analyze the viral sequence and protein structure of SARS-CoV-2 and investigated its association with epidemiological data. Specifically, we analyzed the impact of these factors on infection and predicted the occurrence of new mutations. Our approach involved a multistep analysis. First, we identified specific amino acid substitutions within the viral genome, focusing on their potential impact on SARS-CoV-2 protein structure. Second, we explored the impact of these structural changes on the interaction between SARS-CoV-2 and the host. Third, we assessed the potential effects of these changes on virus infectivity. This systematic analysis allowed us to gain useful insights into the behavior and evolution of the virus.

Hydrophobic protein structure plays an important role in protein stability and folding[29,30]. This affects structural changes from hydrophilic to hydrophobic or positively charged residues. Flaviviruses, including Zika virus, Japanese encephalitis virus, and Dengue virus, are enveloped viruses that use envelope proteins to infiltrate the host. Infection is facilitated by E-protein folding, which causes the virus to fuse with its host[31]. Recently, SARS-CoV-2 treatments targeted the folding of the nonstructural protein NSP3 of the virus[32]. Research on avian coronaviruses has shown that the introduction of a hydrophobic domain into the infectious bronchitis virus E protein affects the infectivity of the virus[33]. Amino acid substitutions such as L411F, F472S, D510S, and I529T have been reported in MERS-CoV[34,35]. In addition to D510S, consecutive hydrophilic amino acids were observed in MERS-CoV, which we believe contributed to its low infectivity (Supplementary Fig. 24).

Our analysis primarily focused on the RBM (amino acids 437–508) region of the spike protein of SARS-CoV-2 that directly interacts with the ACE2 receptor, allowing the virus to infiltrate host

cells and discovered mutations in VOCs and VUMs. In the Alpha, Beta, and Delta lineages, specific amino acid substitutions occurred in various locations. For Omicron, less diversity was observed for amino acid substitutions (Figs. 2B and 3A).

Amino acid substitutions became fixed as SARS-CoV-2 lineages progressed. Examination of epidemiological data revealed four distinct periods. During the first and third periods, there was less diversity in amino acid substitutions, leading to a decrease in both infections and deaths (Fig. 3A). In contrast, during the second period, there was more diversity in amino acid substitutions for the sublineages, leading to an increased number of cases and deaths. Thus, we established an association between the fixation of SARS-CoV-2 amino acid substitutions and infectivity.

The N437R mutation led to increased viral infectivity and received a high APESS score but was barely observed in patients. We detected only three instances of N437R amino acid substitutions, all with the AAT codon, where two or more codon positions must undergo alterations. Such changes were only in V445P, a mutation found in XBB and its sublineages, but it occurred at a miniscule rate of 0.14%. Although not yet prominent, N437R is expected to be associated with high infectivity and remains a primary candidate for close monitoring in future.

Frequency prediction of N460K through machine learning showed a 9-fold increase compared to BA.1. And the mutation showed decreased viral entry in vitro and low binding affinity in silico. The N460K mutation affects docking but combination with other mutations could increase infectivity. Ito et al. confirmed that N460K mutation forms a hydrogen bond to N-linked glycan on N90 of human ACE2[36].

The D467 position plays a key role in a salt-bridge interaction within the Delta variant[37]. Due to this and mutations at the position being extremely rare suggests that D467 contributes to the structural stability of the virus. Specifically, amino acid substitutions D467P and D467I disrupt the salt bridge formed by R457 and R454 (Supplementary Fig. 4). Salt bridges play an important role in thermostabilization[38] and represent electrostatic interactions between positively charged and negatively charged residues. Alterations in positively charged residues can disrupt these salt-bridge interactions with negatively charged residues of SARS-CoV-2. Overall, mutations at this site impair the ability of SARS-CoV-2 to effectively infect the host. In silico results showed

amino acid substitutions at D467 caused the alpha helix to present in the wild type to change into a linear structure (Supplementary Fig. 25).

XBB predicted by ML exhibited a decrease in frequency, displaying a low slope and signifying slow disappearance (Fig. 5C). Protein measurement proved to be difficult in vitro. Notably, there was an approximate 2-fold increase in binding affinity against the wild type in SPR analysis. For the Q493R mutation itself, the docking is low but combined with other mutations, it showed increased docking. We believe that the mutation itself negatively affects the spike protein structure but is important in binding with the host overall. Recent structural research which identified RBD-ACE2 complex structures of BQ1.1 (N460K) and B1.1.529 (Q493R) supports our results[39].

Our research shows that consecutive hydrophobic amino acids in the RBM region and specific amino acid substitutions affect not only infection but also protein structure. We assessed the infectivity of SARS-CoV-2 lineages using protein structure prediction. We observed that the emergence of various mutations corresponded to changes in binding affinity. As SARS-CoV-2 lineages progressed to Omicron, their binding affinity became stronger, resulting in increased infectivity, notably in the Delta and Omicron variants. Conversely, the Beta variant showed high binding affinity but not high infectivity. Our assumption is that for consecutive amino acids and amino acid substitutions, lysine (K) and arginine (R) have less weight; therefore, Beta is unlikely to develop into a significant variant. We speculate that Beta variant emerged early in the pandemic, potentially limiting its ability to infect a large population, while widespread vaccination efforts may have contributed to decreased infectivity.

During the evolutionary process of SARS-CoV-2, the increase in infectivity and severity of variants from Alpha to Delta was attributed to abnormalities in the molecular signaling systems of the host. In contrast, Omicron showed lower severity but increased infectivity. We believe this is because of structural changes that allow the variant to bind well with the ACE2 receptor (Fig. 3 and Supplementary Fig. 7). In addition, epidemiological data reveals a decrease in deaths, reproduction rate is maintained, and Omicron has high docking which confirms the high infectivity. We created mutagenesis sequences created from the backbone of Wuhan-Hu-1 and VOCs with amino acid substitutions to lysine (K) and arginine (R). Sequences from the backbone of Omicron XBB were predicted to have high infectivity and still liable to occur. We can therefore expect consistent infections caused by SARS-CoV-2 going forward.

We comprehensively evaluated genomic and epidemiological data. This multifaceted approach, along with the usage of APESS, machine learning, in vitro assays, and AIVE ensured a more comprehensive understanding of SARS-CoV-2 behavior and evolution. In addition, by applying GMM to the APESS score of a new sample, we can predict its membership within a specific component (Fig. 4A). This prediction allows us to make informed conjectures about its potential strains and infectivity.

Based on our findings, we predicted ~440 mutations with a probability of occurring in the near future. One year after generating the predicted datasets, BA.2.86, EG.5.1, and HK.3 were included in the datasets were reported to the highest frequencies with 43%, 28%, and 12% at Nextstrain (https://nextstrain.org/ncov/gisaid/global/6m; 21 December 2023). APESS values of them also were high at 2.051, 1.982, and 1.986, respectively. These mutations were provided in AIVE and can be considered as important targets to prevent a new wave of infections.

AIVE features these analyses and can be easily utilized without expert knowledge of its algorithm and the analysis features can be used without GPU setup or the need for software library installations. Furthermore, AIVE offers a user-friendly interface with the flexibility to input and append experiment data for analysis. Additionally, ongoing and past analyses can be accessed after processing, providing many options for management. The time required for analysis in the local server or Google Colab can be verified. The local server run on RTX8000 takes 50 minutes, offers a large database, and is free to use (Fig. 6). After completing the job, user can examine and manage the analysis results permanently with their account generated on AIVE.

Prior research of SARS-CoV-2 included analysis of epidemiological data, molecular work, and AI applications. We adopted a comprehensive approach, utilizing real-world data and multi-faceted validation. Our research reveals that SARS-CoV-2 increased in infectivity over time, illuminating significant trends in viral infections. Our discoveries, evaluation model, and the AIVE platform will serve as the foundation for preparedness against further developments in future pandemics.

Materials and Methods

Data collection

We used the World Health Organization tracking of SARS-CoV-2 variants to verify the currently circulating VOCs. In the tracking SARS-CoV-2 variants page, variants are divided into currently circulating variants of concern (VOCs), currently circulating variants of interest (VOIs), and currently circulating variants under monitoring (VUMs). In our study, the primary currently circulating VOCs were Omicron, including BA.1, BA.2, BA.3, BA.4, and BA.5 (https://www.who.int/activities/tracking-SARS-CoV-2-variants).

To obtain the precise lineage and mutation data, cov-lineages.org (cov-lineages.org) and outbreak.info (outbreak.info) databases were used. For the sublineage analysis of the viral sequences collected through GISAID, we used accurately defined sublineage information from the outbreak.info platform.

Using WHO VOCs as standards, Alpha (B.1.1.7), Beta (B.1.351), Delta (B.1.617.2), and Omicron (BA.1, BA.2, and BA.4/BA.5) were selected as the main lineages. From WHO's currently circulating VUMs as of 17 May 2023, we included Omicron BA.2.75 nicknamed "Centaurus," BQ.1, and XBB.

We downloaded the mutation data of the currently reported 7,335,614 viral sequences from GISAID and utilized the data for the following purposes: first, the epidemiological data were categorized based on the patient status column: symptomatic or asymptomatic, and severity: mild or severe (including hospitalized and deceased). Second, for VOCs, including sublineages, we categorized lineage, date, and mutations in the RBM region. Third, to utilize the learning data, we categorized clades, lineages, and mutations in the RBM region. Our epidemiological data on COVID-19 used our world's data (OWID, https://ourworldindata.org/). For confirmed cases and deaths, we used 7-day rolling average worldwide data. For vaccine doses, cumulative vaccine doses were used.

Detection of amino acid features from SARS-CoV-2 sequences

We defined four properties: hydrophilicity (polar, P), hydrophobicity (non-polar, N), alkalinity (basic, B), and acidity (A). We divided two consecutive positions of amino acids with these properties into NN*, NP*, NA*, NB*, PN*, PP*, PA*, PB*, AN*, AP*, AA*, AB*, BN*, BP*, BA*, and BB*, with the most common combination being NN*, NP*, PN*, and PP*.

We verified consecutive amino acids in the SARS-CoV-2 viral sequence spike protein. First, we categorized alanine (A), valine (V), leucine (L), glycine (G), isoleucine (I), methionine (M), tryptophan (W), phenylalanine (F), and proline (P) as hydrophobic amino acids. Second, we categorized serine (S), cysteine (C), asparagine (N), glutamine (Q), threonine (T), and tyrosine (Y) as hydrophilic amino acids. Third, lysine (K), arginine (R), and histidine (H) were categorized as alkaline amino acids (B: positively charged). Finally, aspartic acid (D) and glutamic acid were categorized as acidic amino acids.

We performed the following analysis to observe the polarity changes due to specific amino acid substitutions. Based on polarity and considering the possibility of mutations occurring, we classified each adjacent amino acid for each mutation and divided them into 64 patterns of three amino acids grouped together. We determined the pattern changes caused by the mutations. The analysis methods are available publicly in the web-based platform AIVE (https://ai–ve.org) and GitHub (https://github.com/Honglab-Research/AIVE).

Protein 3D structure evaluation

We utilized Alphafold2 to measure protein structure folding and used pDockQ and HADDock to measure the binding affinity between SARS-CoV-2 and the ACE2 receptor[40]. We used AlphaFold2 to predict changes in the 3D structure of the RBM at amino acid positions 437–508 in the SARS-CoV-2 viral sequence. We performed predictions for Wuhan-Hu-1, Alpha, Beta, Delta, and Omicron (BA.1, BA.2, BA2.75, BA.4, BA.5, XBB, and BQ.1).

Through UniProt (P0DTC2 · SPIKE_SARS2), we used the viral sequence in the RBM region ‘NSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY.’ We performed an analysis using mutation data per lineage (https://www.uniprot.org/uniprotkb/P0DTC2/entry).

We used AlphaFold2 and through structure predictions, we observed structural similarity, distance and folding between amino acids. The predicted local distance difference test (pLDDT), Predicted Aligned Error (PAE), and AlphaFold2 results were used for analysis. Additionally, protein data bank (PDB) files and another AlphaFold2 result, were used to visualize the 3D structures using ChimeraX[41]. We measured the binding affinity between the human ACE2 receptor and the SARS-CoV-2 RBM region. We obtained the human ACE2 receptor sequence from UniProt (UniProt: Q9BYF1).

To increase the confidence of binding affinity analysis, first, we used pDockQ to score binding affinity. Second, we used HADDOCK 2.2 for analysis and comprehensively evaluated the HADDOCK score, root mean square deviation (RMSD), van der Waals energy, electrostatic energy, and desolvation energy (Supplementary Fig. 7). AIVE can predict the structural features of SARS-CoV-2 in different regions and the binding affinity between the human ACE2 receptor and the RBM region of SARS-CoV-2.

Analysis of the association between epidemiology and mutations

From 7,335,614 viral sequences sourced from GISAID, we identified mutations commonly discovered in VOCs and VUMs: L440K, K444T, L452R, N460K, S477N, T478K, E484A, F486V, Q498R, and Y501H. We measured the odds ratios for mutations commonly discovered in VOCs and VUMs: L440K, K444T, L452R, N460K, S477N, T478K, E484A, F486V, Q498R, and Y501H, and epidemiological indicators (symptomatic/asymptomatic and mild/severe). These analyses were performed using the traditional statistical method of logistic regression.

In the patient status column of the data file, we used asymptomatic and symptomatic rows. Deceased, hospitalized, and severe patients were categorized as severe, while outpatients, not hospitalized, and mild were categorized as mild. R version 4.1.1 and python 3.9.13 were used in our analysis.

Evaluation of SARS-CoV-2 mutation infectivity through evaluation modeling

We developed an evaluation model to comprehensively evaluate SARS-CoV-2 at the RNA, amino acid, and protein structure levels. SCPS, PCS, MR, and BPES were designed and calculated for each variant.

Sub-Clustering of Protein Structure (SCPS)

To better understand the molecular structure of SARS-CoV-2 lineages, we used the PAE, pLDDT, pdb, and hydrogen bonds that caused abnormal physical and chemical properties of the compounds to proceed with clustering.

Using clustering, the SARS-CoV-2 lineages were divided into six clusters and the main variants were mostly observed in the first and third clusters. For the cluster distribution for each lineage, “cluster 2” showed the largest difference in folding compared to the Wuhan-Hu-1 variant, and the lowest difference in folding compared to the Delta and Omicron (BA.4/BA.5) variants. Meanwhile, “cluster 4” contained the lineages with the most extreme folding showed a relaxation of folding in the main variants.

Based on this analysis, protein structure prediction-based clustering was defined as follows:

$$\mathrm{SCPS}_i = \begin{cases} 0.9 \; if \; SCPS_i \in \{\mathrm{T2}, \mathrm{T4}\} \\ 0.1 \; if \; SCPS_i \in \{\mathrm{T1}, \mathrm{T3}, \mathrm{T5}, \mathrm{T6}\} \end{cases}$$

where T1 is cluster 1 (437–449), T2 is cluster 2 (450-460), T3 is cluster 3 (461-471), T4 is cluster 4 (472-479, 501-508), T5 is cluster 5 (480-489), and T6 is cluster 6 (490-500), which are positioned in the RBM of the RBD.

Omicron and Delta variants take up 88.3% of the 7,335,614 viral sequences from GISAID. Therefore, we applied a weight of 0.9 to T2 and T4 clusters and 0.1 to the remaining clusters.

Polarity Change Score (PCS)

To calculate the amino acid substitutions in the RBM sequence, we counted each lineage: Wuhan-Hu-1, Alpha, Beta, Delta, and Omicron (BA.1, BA.2, BA2.75, BA.4, BA.5, BQ.1, and XBB). For the counting method, we started from the beginning of the RBM, sliding three positions after each count until the final 72nd position was reached.

$$PCS_i\,(w, m) = \begin{cases} 0.9\ if\ \ PCS_i\,(w_i) \in PP^* \wedge PCS_i\,(m_i) \notin PP^* \\ 0.5\ if\ \ PCS_i\,(w_i) \notin PP^* \wedge PCS_i\,(m_i) \in PP^* \\ 0.1\ if\, otherwise \end{cases}$$

$$PP^* = [(a_1, a_2, a_3) \mid a_1, a_2 \in P,\ a_3 \notin P\},\ \text{P is hydrophilic set and } a_1 \text{ is an amino acid}$$

Three consecutive amino acids with two hydrophilic amino acids in Wuhan-Hu-1 changes due to mutation (0.9), when the inverse occurs (0.5), and for mutations unrelated to consecutive hydrophilic amino acids (0.1).

Mutation Rate (MR)

Based on the base trans rate and codon compositional bias, we investigated the ratio of A, T, G, and C in SARS-CoV-2 spike protein nucleotides and calculated the rate of change of bases depending on the mutations occurring. On average, the SARS-CoV-2 virus comprises 30.2% A, 19.9% G, 32.4% U, 17.6% C. For 17 lineages {Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), Zeta (P.2), Epsilon (B.1.427), Epsilon (B.1.429), Eta (B.1.525), Lambda (C.37), Mu (B.1.621), Omicron (BA.1), Omicron (BA.2), Omicron (BA.4), Omicron (BA.5), Omicron (BA.2.75), Omicron (BQ.1), and Omicron (XBB)}, we checked the differences between the codon positions where mutations occurred. We set the 1st codon at 28%, 2nd codon at 49%, and the 3rd codon at 22% to create an amino acid mutation property equation based on biological properties.

$$MR_i\,(w, m) = \frac{mr\,(w_i) - mr\,(m_i)}{mr\,(w_i)}\quad,\quad \mathrm{mr}\,(x) = \prod_{j=1}^{3} \mathrm{N}_j(x) \times \mathrm{C}_j(x)$$

For each codon position, MR_i (w,m) calculates the frequency of mutation (m) against the reference (w). mr(x) evaluates the frequency change for each nucleotide in a codon. $N_j(x)$ is the frequency for j-the nucleotide. $C_j(x)$ is the frequency for j-the codon position: C_1 (1st codon 28%), C_2 (2nd codon 49%), and C_3 (3rd codon 22%).

The total mutation rate interval is [0.0072, 0.0201] for 64 codons. 16 amino acids (G, E, A, V, R, K, N, T, I, Q, H, L, W, C, Y, and S) are amino acid substitutions that occur in SARS-CoV-2. The mutation rate interval for these amino acid substitutions is [0.0115, 0.0137].

Biochemical Properties Eigen Score (BPES)

We used amino acid properties such as residue, pH, and hydrophobic properties using principal component analysis (PCA). From VOCs, we analyzed the 72 amino acids included in the RBM region. Regarding amino acid properties, amino acids with residues C, D, E, K, R, and Y were in the order R:12.48, K:10.53, and Y:10.07. pH was in the order of R:10.76, K:9.74, and H:7.59. The hydrophobic amino acids were in the order of R: -4.5 and K: -3.9, where K and R showed similarly high scores. K and R received high scores for all the three amino acid properties (Supplementary Table 24). Based on the three biochemical properties of amino acids (residue, pH, and hydrophobicity) and the PCA clustering, we defined the eigen score as follows.

$$BPES_i\,(w, m) = \frac{bpes\,(w_i) - bpes\,(m_i)}{bpes\,(w_i)}$$

$$bpes\,(x) = \|R(x)\|\,\frac{R(x)\,\cdot\,P(x)}{\|R\|\,\cdot\,\|p(x)\|} + \|P(x)\| + \|H(x)\|\,\frac{H(x)\,\cdot\,P(x)}{\|H\|\,\cdot\,\|p(x)\|}$$

For each amino acid property, $BPES_i$ (w,m) calculates the value of the mutation (m) against the reference (w). For bpes(x), R is the eigen vector of a residue score, H is the eigen vector of a hydrophobic score, P is the eigen vector of pH score, || || is norm of the vector, x is the amino acid, and · is the dot product.

Amino acid Property Eigen Selection Score (APESS)

$$APESS\ (w,m)=\sum_{i=1}^{n} SCPS_i(w) \times PCS_i(w,m) \times MR_i(w,m) \times BPES_i(w,m)$$

APESS is the comprehensive value of SCPS, PCS, MR, and BPES. n is the number of positions, w means wild, and m means mutated.

The Kernel Density Estimator (kdeplot) from Python's seaborn package was used to create a distribution function for APESS scores calculated across randomly sampled 30,000 SARS-CoV-2 sublineages sourced from GISAID. For all 7,335,614 lineages, we also calculated the scores and studied the distribution. This distribution function revealed that the SARS-CoV-2 lineages could be differentiated and characterized based on their APESS scores (Fig. 5A). Based on this observation, we applied Gaussian mixture models (GMM) to identify unique components of APESS score. We divided the range of the APESS score into two regions: the lower region for scores less than or equal to 3.2 and the upper region for scores greater than 3.2. To select the optimal number of components for GMM in each region, we used Bayesian Information Criterion (BIC), resulting in 4 and 5 as the optimal number of components in the lower and upper regions, respectively. We fit GMM for each region to construct prediction models with the corresponding optimal number of components.

Analysis of gene regulation from SARS-CoV-2

We analyzed the differences in gene expression for various SARS-CoV-2 lineages when they infect the host. We downloaded GSE235262 from the Gene Expression Omnibus (GEO) to obtain data on human gene expression. Comparisons were made between controls (uninfected people), Alpha (B.1.1.7), Delta (B.1.617.2), Omicron (B.1.529), Omicron (BA.2), and Omicron (BA.4/BA.5 and its sublineages). Only the genes that showed significant expression levels (log2 (TPM+1)) were selected. We used enrichR (https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/cpz1.90) for the enrichment analysis of these selected genes then visualized the comparison between Delta and Omicron (BA.4/BA.5).

SARS-CoV-2 variant prediction and machine learning model

The model predicted the probability of mutations in the RBM region of Alpha, Beta, Delta, and Omicron (BA.1, BA.2, BA.2.75, BA.4, BA.5, BQ.1, and XBB) variants and the 24 mutagenesis variants generated using machine learning. Eighteen variants were used for probability prediction. These were N440K, V445P, G446S, N460K, S477N, T478K, E484A, F486P, F490S, Q498R, N501Y, Y505H, L452R, E484K, Q493R, G496S, F486V, and K444T). Additionally, a total of 24 VOCs (S477R, S477K, N439R, Y501R, N437R, S438R, S459R, S469R, S494R, T470R, T500R, Q493K, Q493R, T470K, T500K, S469K, S494K, Y501K, N437K, N439K, N460R, N460K, S438K, and S459K) excluding overlapping variants and variants lacking expression in data were also used for probability prediction (Supplementary Fig. 26).

The prediction was carried out using data collected from 2019-12-23 to 2023-05-17 (1,242 days), and then with six different types of data, excluding specific clades:

- The first group consists of clades before 21M (Omicron B.1.1.529) including 19A, 19B, 20A, 20B, 20C, 20D, 20E, 20F, 20G, 20H, 20I, 20J, 21A, 21B, 21C, 21D, 21E, 21F, 21G, 21H, 21I, and 21J.

- The second group consisted of clades before 21K (Omicron BA.1) and 21L (Omicron ~BA.2), including the prior group and the addition of 21M, 21K, and 21L.

- The third group consisted of clades before 22D (Omicron BA.2.75) and 22C (Omicron BA.2.12.1), including the prior group and the addition of 22C and 22D.

- The fourth group consists of clades before 22A (Omicron BA.4) and 22B (Omicron, BA.5), including the prior group and the addition of 22A and 22B.

- The fifth group consisted of clades before 22E (Omicron BQ.1), including the prior group and the addition of 22E.

- The sixth group consisted of clades before 22F (Omicron XBB), 23A (Omicron XBB 1.5), and 23 B (Omicron XBB), including the first group and the addition of 22F, 23A, and 23 B.

Machine learning determined the clade information and presence of mutations in 24 variants. The possibility of each variant is extracted through learning. Machine learning was performed using a LightGBM, which is a highly efficient gradient boosting decision tree, XGBoost, which is a scalable tree boosting system, Random Forest, which is a random forest guided tour, and an ensemble model that combined the three previous models. To calculate the probability of mutation occurrences, our custom script and following libraries were used in Python 3.9.13. Data processing employed scikit-learn for scaling, data splitting, and accuracy assessment. In the case of machine learning, LightGBM and XGBoost utilized the 'lightgbm' and 'xgboost' libraries, respectively, to build their models. Meanwhile, the Random Forest and ensemble models utilized scikit-learn's subfunctions to create their models. The deep learning model used TensorFlow and its lower-level library, Keras, for model development."

The Flow of Data Preprocessing and Machine Learning

We retrieved information from SARS-CoV-2 samples collected on GISAID from December 23, 2019, to May 17, 2023, spanning 1,242 days. We filtered the data using various columns, including date, Nextstrain clade, species, substitution, AA substitution, deletion, and insertion, among others.

First, for species data, we extracted the data where the host was either "homo" or "homo sapiens". Then, we removed the samples with date values that did not follow the 'YYYY-MM-DD' format. Based on the insertion, deletion, and AA substitution columns, we removed the samples where deletions, insertions, or nonsense mutations occurred. Following this, we removed columns other than date, clade, and lineage, and then eliminated samples with missing values (NaN).

For the extracted samples, we organized the amino acid mutation information from the AA substitution column for the 437-508 region of the S protein. We created columns indicating the presence (1) or absence (0) of 40 target mutations. We removed mutations not found in the GISAID data, leaving the columns representing the presence or absence of 24 mutations (N440K, V445P, G446S, N460K, S477N, T478K, E484A, F486P, F490S, Q498R, N501Y, Y505H, L452R, E484K, Q493R, G496S, F486V, K444T, S494R, T500R, Q493K, T470K, N439K, and N460R).

To prepare the filtered data for training, we transformed the values for the dates into integers, representing the number of days that had passed since the initial collection date. We then normalized these values to fall within the range of [0-1]. Using one-hot encoding, we preprocessed the Nextstrain clade information. From the preprocessed data, we created six datasets, each based on different clade groups. Finally, we used the random state function to generate the training and testing datasets used in the learning process.

The training data is used for learning as the input for five different models created. The learning model then predicts the likelihood of the occurrence of 24 target mutations (Supplementary Fig. 27).

Cell culture

HEK293T and 293T-hACE2 cells were maintained in Dulbecco's Modified Eagle Medium (DMEM, Welgene, South Korea) supplemented with 10% fetal bovine serum (FBS, Gibco) and 1% penicillin-streptomycin (Gibco) at 37 ℃ in a humidified 5% CO_2. A HEK293T cell line stably

expressing human ACE2, 293T-hACE2, was generated using a lentivirus-mediated gene transduction system under the antibiotic pressure of hygromycin.

Molecular cloning and plasmid construction

The human codon-optimized full-length SARS-CoV-2 spike (Wuhan-Hu-1 strain) gene was obtained from Sino Biological Inc. (Beijing, China). The last 19 amino acids of the SARS-CoV-2 spike were truncated to increase protein expression, and a flag tag was inserted at the C-terminal end. Spike gene mutation constructs, including D614G, N437R, N460K, D467P, D467I, and Q493R, were generated using a PlatinumTM SuperFi II PCR Master Mix (Invitrogen, Waltham, MA, USA).

Spike protein expression test

Viral spike genes were cloned into pcDNA3.1(+), and plasmid DNA was transfected into HEK293T cells using the PEI reagent (Sigma). After 48 h, the cells were lysed with a 1% NP-40 lysis buffer (150 mM NaCl, 1% NP-40, 50 mM Tris-HCl), and the extracted proteins were analyzed by SDS-PAGE and immunoblotting. A mouse anti-FLAG antibody (Sigma) was used to detect spike protein expression.

Pseudotyped virus and viral entry

The pseudotyped virus was generated from HEK293T cells (Invitrogen) by co-transfection with human immunodeficiency virus backbone plasmids expressing firefly luciferase as described previously. We used packaging plasmids (pLP1, pLP2, and pLP/VSV-G; all from Invitrogen) and pLVX-Luc-IRES-ZsGreen1_ Cat. 632187 (Luc stands for luciferase, and IRES stands for internal ribosomal entry site). For S protein pseudotyping, the full-length cDNA of the S gene (Sino Biological Inc.) was cloned into pcDNA3 and used for transfection instead of pLP/VSV-G.

A plasmid carrying the gene encoding the indicated mutation in the S protein was generated using a PlatinumTM SuperFi ll PCR Master Mix (Invitrogen, Waltham, MA, USA) based on the wild-type construct, and the point mutation was confirmed by sequencing. Viral supernatants were harvested 48 h after transfection and normalized using the Lenti-X reverse transcription-quantitative PCR (qRT-PCR) titration kit _ Cat. 631235 according to the manufacturer's protocol. Infected 293T-hACE2 cells were lysed 48 h after infection and the efficiency of viral entry was measured by comparing the luciferase activity of pseudotyped viruses bearing the wildtype or mutant S protein. A VSV-G-pseudotyped lentivirus was used as a positive control. Relative luciferase activity in the cell lysates was measured using a luciferase assay kit (Promega).

Cloning, expression, and purification of ACE2, SARS-CoV-2 RBD, and RBD variants

The SARS-CoV-2 spike protein RBD, and its mutations (N460K and Q493R), along with the N-terminal peptidase domain of human ACE2 was expressed using the Bac-to-Bac® baculovirus system (Invitrogen). The RBD (Residues Arg319-Phe541) and the N-terminal peptidase domain of human ACE2 (Residues Ser19-Asp615) with an N-terminal gp67 signal peptide were cloned upstream of the cleavable deca-histidine tag and FLAG tag. RBD variants were introduced into the wild-type RBD construct using PCR-based site-directed mutagenesis. The resulting constructs were expressed in Spodoptera frugiperda (Sf9) insect cells and secreted into the medium after being cultured at 27°C for 72 h.

Subsequently, ACE2, RBD, and its variants were isolated from the filtered supernatant using a HisTrap Excel column (Cytiva, Marlbourough, MA, USA). The C-terminal tags were cleaved using an in-house-generated HRV3C protease. The protease and cleaved C-terminal tags were removed using Ni-NTA resin (Takara, 635662), and the proteins were further purified by size-exclusion chromatography on a HiLoad 26/600 Superdex 200 pg (Cytiva). Purification was performed in a buffer containing 20 mM Tris-HCl (pH 7.5), 150 mM NaCl, and 1 mM DTT. To maintain stability during storage, each protein sample was frozen at -80°C and supplemented with 10% glycerol. For surface plasmon resonance (SPR), proteins were thawed from frozen storage immediately before the experiment.

Binding affinity measurement

The SPR experiments were performed at room temperature using BiaCore T200 with a series of S CM5 sensor chips (Cytiva). The surfaces of the sample and reference flow cells were activated using a 1:1 mixture of NHS (N-hydroxysuccinimide) and EDC (3-(N,N-dimethylamino) propyl-N-ethylcarbodiimide). The reference flow cell was left blank. All the surfaces were blocked with ethanolamine (pH 8.0). HBS-EP + (Cytiva) was used as running buffer.

For the binding affinity assay, the purified N-terminal domain of ACE2 was diluted in 10 mM sodium acetate buffer (pH 4.0) and immobilized on the chip at approximately 800 response units. The SARS-CoV-2 RBD and RBD mutations along the gradient were observed on the chip surface. After each cycle, the sensor surface was regenerated using 10 mM glycine-HCl pH 2.5). The data were fitted to a 1:1 interaction steady-state binding model using BIAevaluation 3.1 software. Curve fitting was performed via nonlinear regression using a one-site-specific binding equation in GraphPad Prism version 8.4.0 (GraphPad Software, Boston, MA, USA).

AIVE platform

The AIVE platform (https://ai-ve.org/) provides real-time protein structure predictions and APESS score calculations. AIVE reports the analysis results and mutation data from known sequences, randomly sampled SARS-CoV-2 sequences, or custom sequences provided by users. Based on the mutations in the SARS-CoV-2 lineages, Alphafold2 was utilized to analyze the protein structures. Using the Wuhan-Hu-1 sequence as a reference, we performed a comparative analysis of the main VOCs: Alpha, Beta, Delta, and Omicron. Reports are generated using visualization tools that provide Alphafold2 outputs such as pLDDT and PAE. Monomer folding and docking were visualized and scored. Differences caused by mutations are shown in the input amino acid sequences based on polarity and protein properties. APESS can predict the infectivity of sequences through the distribution of values obtained through the GMM prediction model and using previously known lineages. The results of the user's APESS sequence scores can be visualized. AIVE runs on a high-performance computing (HPC) system with 3 RTX8700 GPUs, 96 CPUs, and 256 GB main memory.

References

1 Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat Med* **29**, 1930-1940 (2023). https://doi.org/10.1038/s41591-023-02448-8

2 Yang, X. *et al.* A large language model for electronic health records. *NPJ Digit Med* **5**, 194 (2022). https://doi.org/10.1038/s41746-022-00742-2

3 Tsai, K. C., Lee, Y. C. & Tseng, T. S. Comprehensive Deep Mutational Scanning Reveals the Immune-Escaping Hotspots of SARS-CoV-2 Receptor-Binding Domain Targeting Neutralizing Antibodies. *Front Microbiol* **12**, 698365 (2021). https://doi.org/10.3389/fmicb.2021.698365

4 Cao, Y. *et al.* Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657-663 (2022). https://doi.org/10.1038/s41586-021-04385-3

5 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22** (2017). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

6 Mathieu, E. *et al.* A global database of COVID-19 vaccinations. *Nat Hum Behav* **5**, 947-953 (2021). https://doi.org/10.1038/s41562-021-01122-8

7 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018). https://doi.org/10.1093/bioinformatics/bty407

8 Chen, Z. *et al.* Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nat Genet* **54**, 499-507 (2022). https://doi.org/10.1038/s41588-022-01033-y

9 Peacock, T. P. *et al.* The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. *Nat Microbiol* **6**, 899-909 (2021). https://doi.org/10.1038/s41564-021-00908-w

10 Johnson, B. A. *et al.* Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature* **591**, 293-299 (2021). https://doi.org/10.1038/s41586-021-03237-4

11 Bouhaddou, M. *et al.* SARS-CoV-2 variants evolve convergent strategies to remodel the host response. *Cell* **186**, 4597-4614 e4526 (2023). https://doi.org/10.1016/j.cell.2023.08.026

12 Ali, F., Kasry, A. & Amin, M. The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutant. *Med Drug Discov* **10**, 100086 (2021). https://doi.org/10.1016/j.medidd.2021.100086

13 Seyran, M. *et al.* The structural basis of accelerated host cell entry by SARS-CoV-2dagger. *FEBS J* **288**, 5010-5020 (2021). https://doi.org/10.1111/febs.15651

14 Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710 (2020). https://doi.org/10.1038/s41586-019-1923-7

15 Theodosiou, A. A. & Read, R. C. Artificial intelligence, machine learning and deep learning: Potential resources for the infection clinician. *J Infect* **87**, 287-294 (2023). https://doi.org/10.1016/j.jinf.2023.07.006

16 Yang, Z. *et al.* A deep learning framework identifies dimensional representations of Alzheimer's Disease from brain structure. *Nat Commun* **12**, 7065 (2021). https://doi.org/10.1038/s41467-021-26703-z

17 Berman, D. S., Howser, C., Mehoke, T., Ernlund, A. W. & Evans, J. D. MutaGAN: A sequence-to-sequence GAN framework to predict mutations of evolving protein populations. *Virus Evol* **9**, vead022 (2023). https://doi.org/10.1093/ve/vead022

18 Zhou, B. *et al.* TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution. *Comput Biol Med* **152**, 106264 (2023). https://doi.org/10.1016/j.compbiomed.2022.106264

19 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org/10.1038/s41586-021-03819-2

20 Bhowmick, S. *et al.* In Silico Protein Folding Prediction of COVID-19 Mutations and Variants. *Biomolecules* **12** (2022). https://doi.org/10.3390/biom12111665

21 Kilim, O., Mentes, A., Pal, B., Csabai, I. & Gellert, A. SARS-CoV-2 receptor-binding domain deep mutational AlphaFold2 structures. *Sci Data* **10**, 134 (2023). https://doi.org/10.1038/s41597-023-02035-z

22 Lins, L. & Brasseur, R. The hydrophobic effect in protein folding. *The FASEB Journal* **9**, 535-540 (1995). https://doi.org/https://doi.org/10.1096/fasebj.9.7.7737462

23 Shekhawat, U. & Roy Chowdhury, A. Computational and comparative investigation of hydrophobic profile of spike protein of SARS-CoV-2 and SARS-CoV. *Journal of Biological Physics* **48**, 399-414 (2022). https://doi.org/10.1007/s10867-022-09615-x

24 Panchin, A. Y. & Panchin, Y. V. Excessive G-U transversions in novel allele variants in SARS-CoV-2 genomes. *PeerJ* **8**, e9648 (2020). https://doi.org/10.7717/peerj.9648

25 Yi, K. *et al.* Mutational spectrum of SARS-CoV-2 during the global pandemic. *Exp Mol Med* **53**, 1229-1237 (2021). https://doi.org/10.1038/s12276-021-00658-z

26 van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* **428**, 720-725 (2016). https://doi.org/10.1016/j.jmb.2015.09.014

27 Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* **13**, 1265 (2022). https://doi.org/10.1038/s41467-022-28865-w

28 Cao, W. *et al.* Biomechanical characterization of SARS-CoV-2 spike RBD and human ACE2 protein-protein interaction. *Biophys J* **120**, 1011-1019 (2021). https://doi.org/10.1016/j.bpj.2021.02.007

29 Pace, C. N. *et al.* Contribution of hydrophobic interactions to protein stability. *Journal of molecular biology* **408**, 514-528 (2011). https://doi.org/10.1016/j.jmb.2011.02.053

30 Islam, M. M., Kobayashi, K., Kidokoro, S.-I. & Kuroda, Y. Hydrophobic surface residues can stabilize a protein through improved water–protein interactions. *The FEBS Journal* **286**, 4122-4134 (2019). https://doi.org/https://doi.org/10.1111/febs.14941

31 Hu, T., Wu, Z., Wu, S., Chen, S. & Cheng, A. The key amino acids of E protein involved in early flavivirus infection: viral entry. *Virol J* **18**, 136 (2021). https://doi.org/10.1186/s12985-021-01611-2

32 Bergasa-Caceres, F. & Rabitz, H. A. Interdiction of Protein Folding for Therapeutic Drug Development in SARS CoV-2. *J Phys Chem B* **124**, 8201-8208 (2020). https://doi.org/10.1021/acs.jpcb.0c03716

33 Ruch, T. R. & Machamer, C. E. The hydrophobic domain of infectious bronchitis virus E protein alters the host secretory pathway and is important for release of infectious virus. *J Virol* **85**, 675-685 (2011). https://doi.org/10.1128/JVI.01570-10

34 Wong, L. R. *et al.* Middle East respiratory syndrome coronavirus Spike protein variants exhibit geographic differences in virulence. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org/10.1073/pnas.2102983118

35 Kleine-Weber, H. *et al.* Mutations in the Spike Protein of Middle East Respiratory Syndrome Coronavirus Transmitted in Korea Increase Resistance to Antibody-Mediated Neutralization. *J Virol* **93** (2019). https://doi.org/10.1128/JVI.01381-18

36 Ito, J. *et al.* Convergent evolution of SARS-CoV-2 Omicron subvariants leading to the emergence of BQ.1.1 variant. *Nat Commun* **14**, 2671 (2023). https://doi.org/10.1038/s41467-023-38188-z

37 Baral, P. *et al.* Mutation-induced changes in the receptor-binding interface of the SARS-CoV-2 Delta variant B.1.617.2 and implications for immune evasion. *Biochem Biophys Res Commun* **574**, 14-19 (2021). https://doi.org/10.1016/j.bbrc.2021.08.036

38 Ban, X. *et al.* Evolutionary Stability of Salt Bridges Hints Its Contribution to Stability of Proteins. *Comput Struct Biotechnol J* **17**, 895-903 (2019). https://doi.org/10.1016/j.csbj.2019.06.022

39 Han, P. *et al.* Receptor binding and complex structures of human ACE2 to spike RBD from omicron and delta SARS-CoV-2. *Cell* **185**, 630-640 e610 (2022). https://doi.org/10.1016/j.cell.2022.01.001

40 Khan, J. M. & Ranganathan, S. pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immunome Res* **6 Suppl 1**, S2 (2010). https://doi.org/10.1186/1745-7580-6-S1-S2

41 Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* **30**, 70-82 (2021). https://doi.org/https://doi.org/10.1002/pro.3943

Acknowledgements

We thank the Global Science experimental Data hub Center (GSDC) and the Korea Research Environment Open NETwork (KREONET) service for data computing and network provided by the Korea Institute of Science and Technology Information (KISTI).

Funding

This work was supported in part by grants from the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (NRF-2021M3H9A2097227, NRF: 2021M3A9I2080490, NRF-2022R1A2C3008162, and RS-2023-00220840), the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (RS-2023-00265923), and the Basic Medical Science Facilitation Program through the Catholic Medical Center of the Catholic University of Korea funded by the Catholic Education Foundation.

Authors' contributions

D.H conceived and designed these studies. W.J.C, D.Y.S, J.Y.L, H.J.P, D.S.C, S.P.J, K.J.Y and J.P gathered SARS-CoV-2 data and performed bioinformatic analyses. S.P.J, J.P, W.J.C, D.Y.S, E.J.M, and D.H built mathematics models and performed statistical validation. J.P, W.J.C, D.S.C, D.Y.S, (relevant team) analyzed protein 3D structures and utilized machine learning for predictions. W.J.C, D.Y.S, J.P and D.H designed and built the AIVE system. U.K, G.Y.Y, H.K, T.K, S.G, H.S.C and N.H.C performed experiments in the experiment of luciferase and estimation of virus viability. D.Y.S, J.P and D.H wrote the manuscript with input from all other authors.

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Data Availability

All data supporting the findings reported here are available in the paper and Supplementary Information. The raw and processed data, custom scripts, and codes used in this study are deposited in the GitHub repository (https://github.com/Honglab-Research/AIVE, https://github.com/Honglab-Research/AIVE-prediction) and AIVE (https://ai-ve.org).

Figure Legends

Figure 1. Analysis of protein properties discovered in the SARS-CoV-2 amino acid sequence

(A) The SARS-CoV-2 amino acid sequence between positions 437 and 508 in the Receptor Binding Motif (RBM) is displayed with the corresponding amino acids in the original positions. Amino acid substitutions are shown for Alpha, Beta, Delta, Omicron BA.1, Omicron BA.2, Omicron BA.4/BA.5, Omicron BQ.1, and Omicron XBB. Hydrophilic (polar) amino acids are displayed in red, hydrophobic (non-polar) in blue, acidic in green, and alkaline (positively charged) in yellow.

(B) The number of polarity changes [N: hydrophobic (nonpolar), P: hydrophilic (polar), A: acidic, and B: alkaline (basic)] in the Receptor Binding Domain (RBD) region is displayed. Wuhan-Hu-1, Alpha, Beta, Delta, Omicron (BA.1), Omicron (BA.2), Omicron (BA.2.75), Omicron (BA.4/BA.5), Omicron (XBB), and Omicron (BQ.1) are presented on the graph with each lineage color-coded. The polarity change count for NN*, NP*, PN*, and PP* are shown in more detail. PN* (Wuhan-Hu-1: 39, Alpha: 39, Beta: 40, Delta: 37, Omicron BA.1: 36, Omicron BA.2: 35, Omicron BA.2.75: 35, Omicron BA.4/BA.5: 34, Omicron XBB: 36, Omicron BQ.1: 34) PP* (Wuhan-Hu-1: 31, Alpha: 31, Beta: 31, Delta: 30, Omicron BA.1: 27, Omicron BA.2: 26, Omicron BA.2.75: 26, Omicron BA.4/BA.5: 27, Omicron XBB: 29, Omicron BQ.1: 27). Overall, polarity decreased from PN* to PP* across all SARS-CoV-2 lineages.

(C) The amino acid substitutions in the RBM region from the reference to VOCs are displayed. The seventeen amino acids in the reference list are tyrosine (Y), threonine (T), serine (S), glutamine (Q), asparagine (N), cysteine (C), valine (V), proline (P), leucine (L), isoleucine (I), glycine (G), phenylalanine (F), alanine (A), arginine (R), lysine (K), glutamic acid (E), and aspartic acid (D). For VOCs, the amino acid substitutions are indicated by gray lines. There was a more than two-fold increase in lysine (K) and arginine (R) in VOCs compared with the reference.

Figure 2. Evaluation and validation of amino acid substitutions in the SARS-CoV-2 RBM region

(A) Mutations, their occurrence rates as percentages, and the original amino acid at the position are shown. Alpha, Beta, Delta, Omicron (BA.2.75), Omicron (BA.4/BA.5), Omicron (XBB), and Omicron (BQ.1) lineages are displayed with corresponding colors. N440, L452, S477, T478, E484, F486, N501, and Y505 are indicated by yellow triangles while D467 is indicated by a red triangle.

(B) For positions N440, L452, S477, T478, E484, F486, N501, and Y505, lineages and amino acid substitutions are displayed. Arrows indicate the mutation rate where width corresponds with the percentage and the colors indicate lineages Alpha, Beta, Delta, Omicron (BA.2.75), Omicron (BA.4/BA.5), Omicron (XBB), and Omicron (BQ.1). The colors in the pie chart indicate amino acids. The mutation rate of Alpha is lower than 20%. At the 501st position in the RBM region, amino acid substitutions from asparagine (N) to tyrosine (Y) (n = 282) occurred along with other substitutions. For the Beta variant, the 484th position showed a mutation rate over 60% with E484K. The Delta variant showed a mutation rate of 60% at the 452nd position. L452R and T478K amino acid substitutions along with various mutations were observed. In Omicron, the mutation rate for S477, T478, E484, N501, and Y505 was over 40%. The amino acid substitutions were S477K, T478K, E484A, N501Y, and Y505H. We calculated the mutation rates for the following positions: T478K (99.95%), Q498R (94.14%), N501Y (99.52%), and Y505H (97.66%).

(C) The effect of the D467 amino acid substitution on viral infection was evaluated in vitro via luciferase and viral entry assays. Mutagenesis at D467 to hydrophobic amino acids proline (P) and isoleucine (I) was performed. There was a significant decrease in the RLU and viral entry percentages for both D467P and D467I (<0.0001).

Figure 3. Association between SARS-CoV-2 mutations and epidemiological data

(A) For each SARS-CoV-2 lineage, the position and distribution of amino acid substitutions were analyzed alongside epidemiological data. The first period, from November 2020 to December 2021, was characterized by low infections and high deaths. Delta was the most prominent during this period, coinciding with worldwide vaccinations. Amino acid substitutions L452R and T478K were observed at the highest frequency while T478K and L452R were independently observed at the highest frequencies. The second period, from January 2022 to April 2022, saw an increase in the new cases and deaths. BA.5 was prominent during this period with various amino acid substitutions observed. The third period, from May 2022 to November 2022, showed a significant decrease in infections and deaths. Omicron, specifically BQ.1 was prominent during this period and worldwide vaccination rates decreased.

(B) From the viral sequences of the patients, the association between the primary mutations of Delta, BA.5, and BQ.1, and epidemiological data (symptoms and severity) was analyzed. Odds ratios are displayed for L440K, K444T, L452R, N460K, S477N, T478K, E484A, F486V, Q498R, N501Y, and Y505H. L452R was an indicator of symptomaticity in Delta and BA.5. K477T was associated with symptomaticity in BQ.1. All mutations were associated with mildness. The 95% confidence intervals are shown for Delta, BA.5, and BQ.1.

(C) We analyzed the expression of Delta compared to BA.4/BA.5 using the GSE235262 dataset. In Delta, the mTOR pathway was observed to regulate ribosome biogenesis, autophagy, and lipid biosynthesis while also playing a role in viral infection pathways. The expression values were calculated as Delta [log2(TPM+1)] / Omicron BA.4+BA.5 [log2(TPM+1)], with ENST Ensembl Transcript IDs, and * indicating a significance level of $p<0.05$.

(D) The folding structures and pDockQ scores (0.506, 0.569, 0.577, 0.560, 0.564, and 0.575 for Wuhan-Hu-1, Beta, Delta, BA.4/BA.5, BQ.1, and XBB respectively) were shown.

Figure 4. APESS: a comprehensive evaluation model of SARS-CoV-2 mutations

(A) Amino acid Property Eigen Selection Score (APESS), an evaluation model based on the properties discovered in the RBM and the infectivity of SARS-CoV-2, was developed. A 72-amino acid-long RBM sequence of SARS-CoV-2 was used to comprehensively evaluate the sub-clustering of protein structure (SCPS), polarity change score (PCS), mutation rate (MR), and biochemical properties eigen score (BPES). Through comprehensive analysis of each position, the infectivity of the input sequence could be evaluated against preexisting lineages.

(B) The APESS scores were calculated for SARS-CoV-2 lineages Alpha, Beta, Delta, and Omicron (BA.2.75, BA.5, XBB, BQ.1), and the data were obtained for sublineages from viral sequences. The original lineages are displayed with a gray triangle and their APESS scores, whereas the sublineages are color-coded differently. The S477K substitution resulted in the highest APESS score.

Figure 5. Multifaceted evaluation of SARS-CoV-2: evaluation model, machine learning, and in vitro assay

(A) Mutagenesis sequences containing consecutive hydrophilic amino acids were evaluated with APESS. They were based on Wuhan-Hu-1, Alpha, Beta, Delta, BA.1, BA.2, BA.2.75, BA.4/BA.5, XBB, and BQ.1, as indicated by colors and gray triangles. APESS values for the K, R, N, S, and Y mutated sequences of the lineages are displayed. Mutagenesis of lysine (K) and arginine (R) in Omicron sublineages resulted in increased APESS scores, whereas mutagenesis of asparagine (N), serine (S), and tyrosine (Y) resulted in decreased APESS scores. Specific regions for K and R are magnified to show the distribution of the APESS scores of the mutagenesis sequences in more detail.

(B) To predict mutations with high infectivity using APESS, mutagenesis was performed using Wuhan-Hu-1, Alpha, Beta, Delta, BA.1, BA.2, BA.2.75, BA.4/BA.5, XBB, and BQ.1 as the backbone. The presence of these amino acid substitutions was verified using the viral sequence data from GISAID. For each lineage, the amino acid substitutions resulted in 280 mutagenic sequences. Thirty sequences with the highest APESS and pDockQ scores are displayed. N460R and S469R have not been observed naturally, whereas N439R, S459R, N437R, Y501R, S438R, and S494R have been observed in ten people or less.

(C) For mutations occurring in lineages and mutations evaluated through APESS, AI learning models (Random Forest, LightGBM, XGBoost, Ensemble, and deep learning) were used to investigate the probability. For N460K, there was a 9-fold increase in the probability of XBB compared to prior Omicron lineages. Q493R is not found in XBB but still has a high probability of occurrence.

(D) The effects of N437 and Q493 amino acid substitutions on viral infection were evaluated in vitro using luciferase and viral entry assays. There was a significant increase in the Relative Light Units and viral entry percentage for N437R, vice versa for Q493R.

Figure 6. Prediction of potential SARS-CoV-2 mutations through integrated evaluation and prediction

(Input). This figure consists of three steps: ‘Input’, ‘Processing’, and ‘Output’. Users can select a custom sequence from the entire SARS-CoV-2 sequence, choose VOCs, or create customized sequences for analysis. Depending on the user's system environment, analysis can be done through local prediction (server) or Google Colab.

(Processing) Three types of analyses are performed. First, the protein 3D structure prediction is analyzed. This includes protein 3D structure, pLDDT, and PAE. Second, the infectivity is evaluated using APESS (2.12). For each position, the structural difference graph for BPES, MR, PCS, and SCPS is visualized. The APESS distribution is visualized for known VOCs and created variants. Third, polarity changes are visualized in sequences.

(Output) Four results comparing Wuhan-Hu-1+N460K and XBB (with N460K) are output and visualized. First, through protein structure prediction, secondary structures can be confirmed in XBB (with N460K) compared to Wuhan-Hu-1+N460K (Yellow arrow). Second, comparison of polarity changes through the mutation of Wuhan-Hu-1+N460K (red dotted line) is done. Third, in XBB (with N460K), which has more mutations than Wuhan-Hu-1+N460K, the difference in values at each position in the protein sequence of SCPS, PCS, MR, and BPES is displayed (red dotted line). Fourth, the distribution of APESS, which represents the comprehensive value of SCPS, PCS, MR, and BPES, is shown. “apess” indicates the score for each position in the customized protein sequence. In the case of XBB (with N460K), which has more mutations than Wuhan-Hu-1+N460K, an apess score distribution of XBB (with N460K) has values from -0.079 to 1.385 is shown (Yellow arrow). X-axis presents position in RBM (72aa) and Y-axis presents APESS score, respectively (④-1). APESS is the summed value of each position and can evaluate infectivity. A region including XBB (with N460K) shows infectivity due to many mutations, and also shows an increase in APESS score due to the N460K mutation. X-axis presents APESS score and Y-axis presents a density, respectively (④-2).

AIVE comprehensively evaluates viral infectivity, protein structure, amino acid substitutions, and polarity changes in preexisting and potential SARS-CoV-2 sequences.