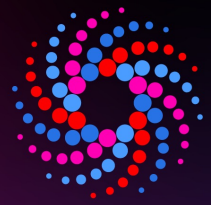
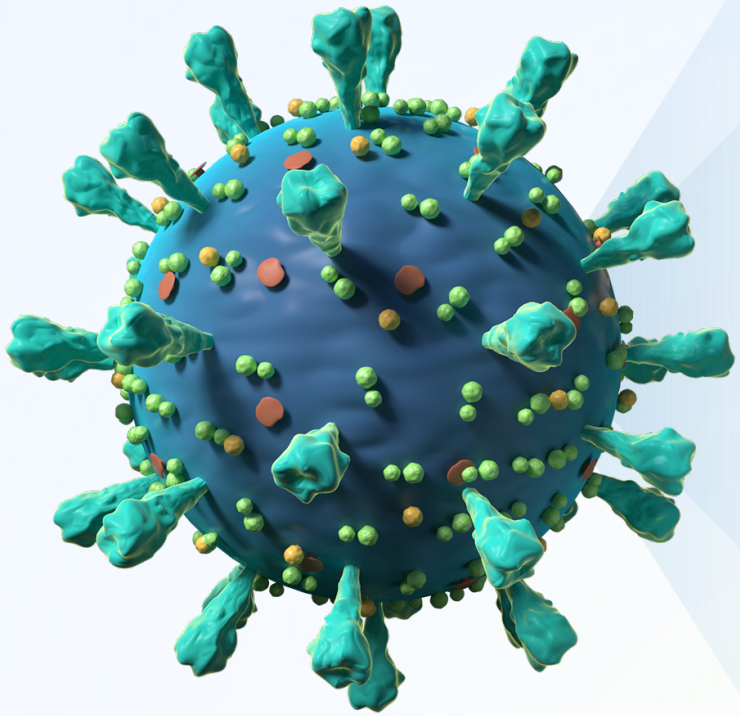


SC22
Dallas, TX | hpc accelerates.

Running Ahead of Evolution

AI based simulation for predicting future high-risk SARS-CoV-2 variants

Peng Cheng Laboratory AI Team



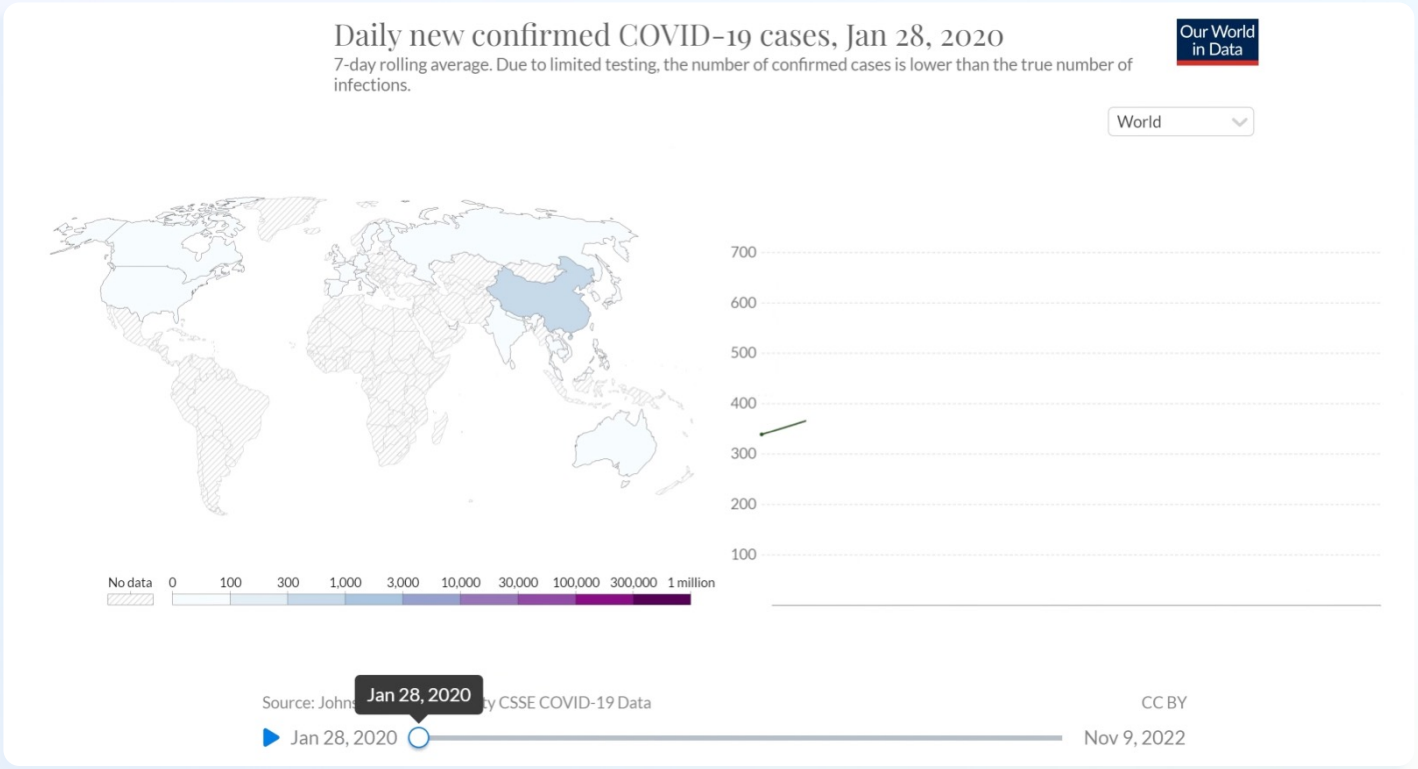
Background
& Highlight

Technical challenges
& Innovations

Science results
& Key performance

Summary
& Prospect

COVID-19: A world-wide pandemic



Never-ending story

Timeline of outbreak

2021.01

3 months

2021.04

4 months

2021.08

5 months

2022.01

6 months

2022.07

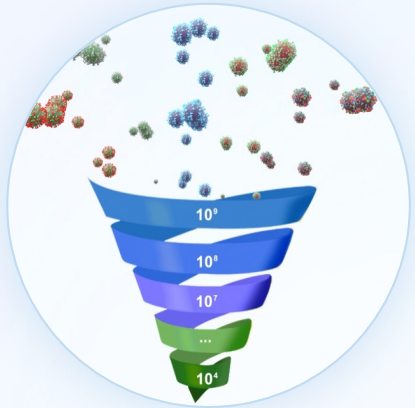
? months

? ? ?

Running ahead of evolution

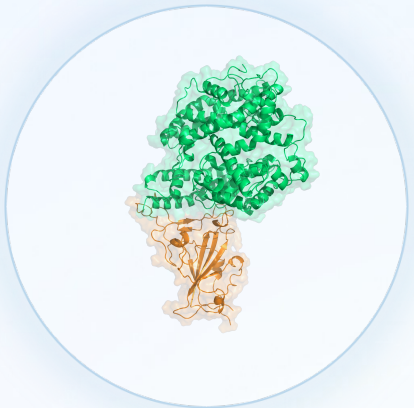
Stage 1

In silico mutation simulation



Stage 2

Molecular dynamics simulation



Stage 3

Wet-lab experiments

Stage 4

Drug design

Pandemic control

SARS-CoV-2 is the most sequenced virus, EVER.

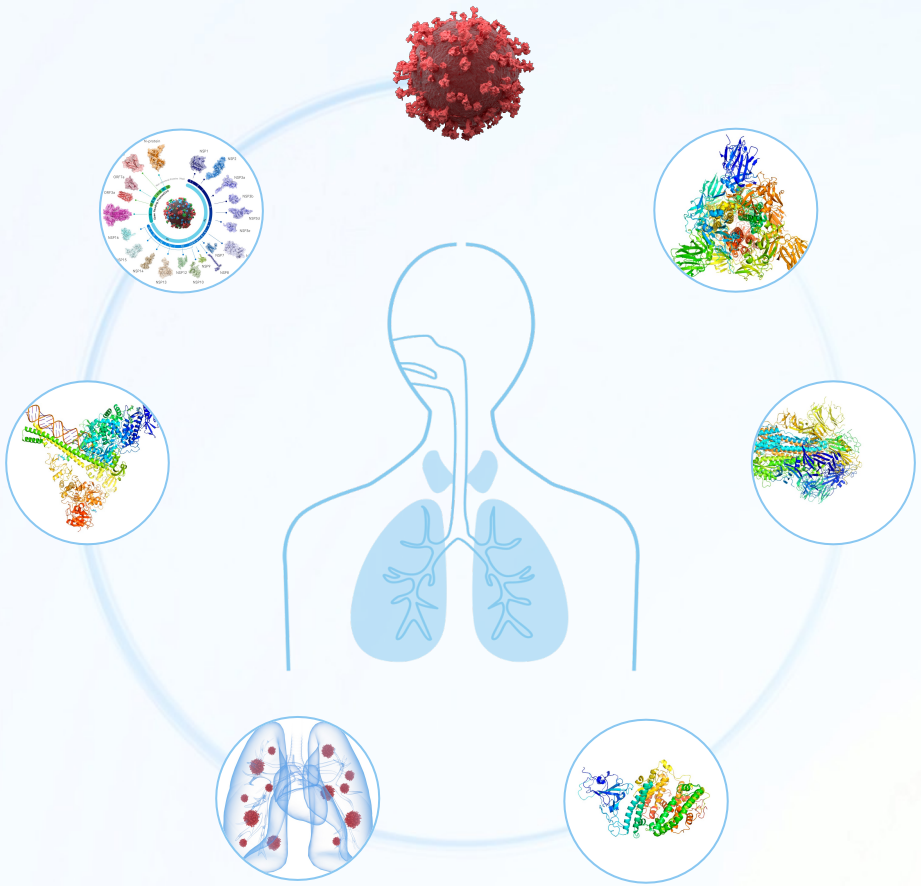
hCoV-19 data sharing via GISAID

13,890,985

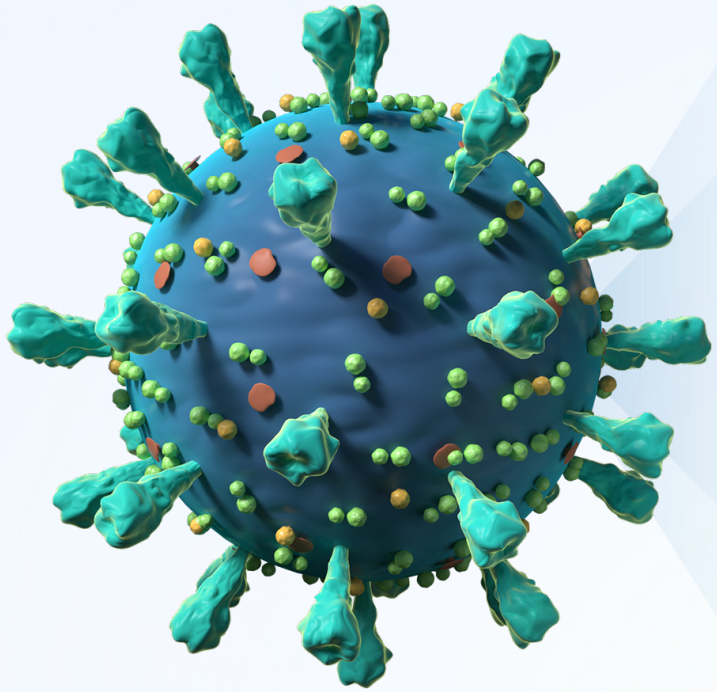
Genome sequence submissions

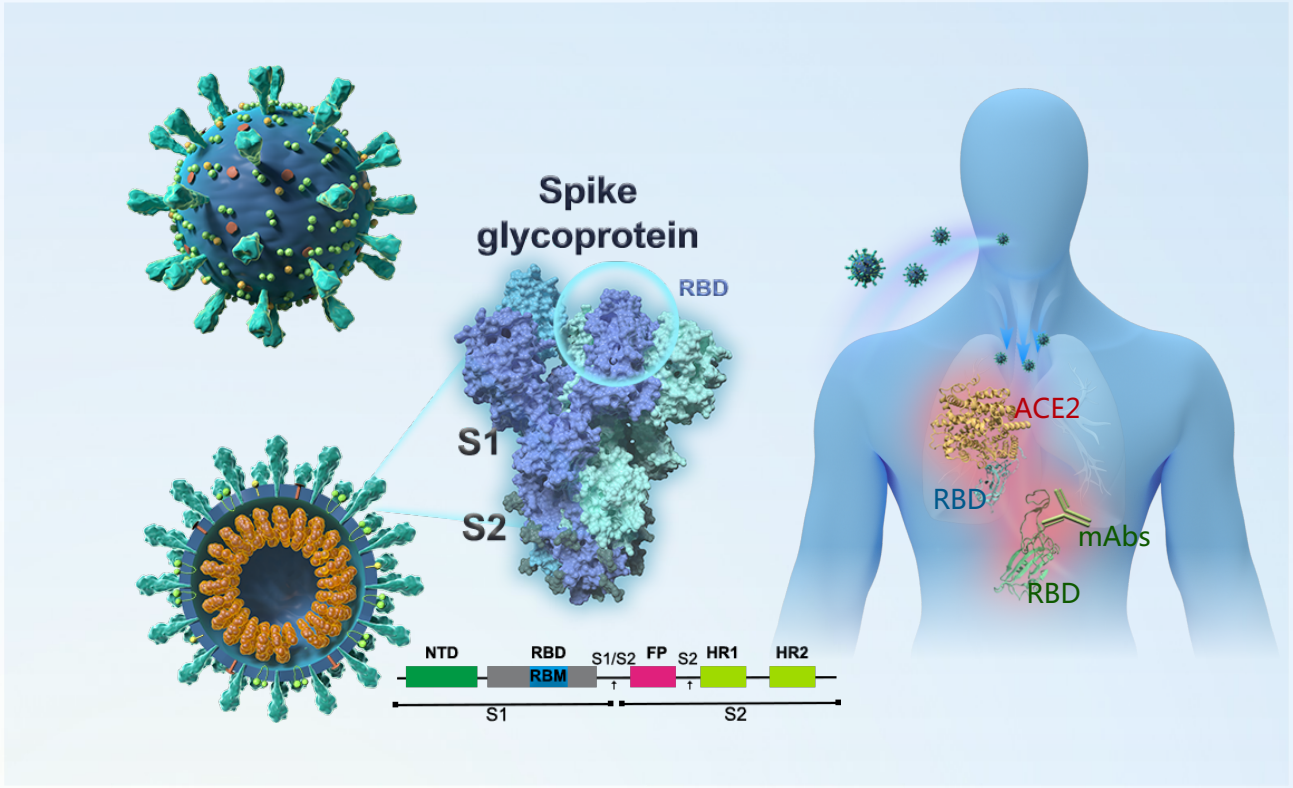
Can the viral mutational landscape and trends be inferred from these 13.8 million viral genome sequences?

The infection process of SARS-CoV-2



The most mutated region of SARS-CoV-2





The RBD region of the Spike protein is an area of concern because it has a high mutation rate, which can significantly affect binding to hACE2 and antibodies.

Highlight: SARS-CoV-2 RBD mutation simulation and variants prediction

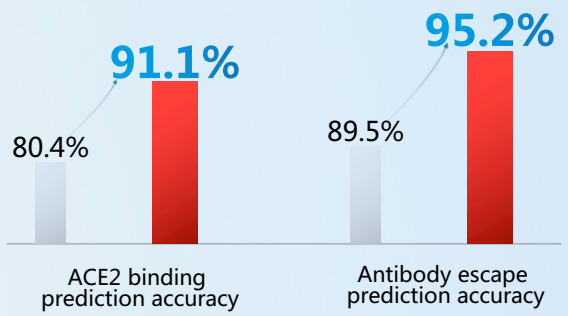
The first work of SARS-CoV-2 RBD mutation simulation									
Module	Generation				Screening			Feedback	
Details	Massive sequence processing	Pretraining	Fine-tuning	Inference	ACE2 binding	Antibody escape	Transmissibility	Comprehensive analysis	Recursively fine-tuning
Our work	✓	✓	✓	✓	✓	✓	✓	✓	✓
Nicola De Maio et al, 2022	✓	✗	✗	✓	✗	✗	✗	✓	✗
Vladimir Shchur et al, 2022	✓	✗	✗	✓	✗	✗	✗	✓	✗
Jiahua Chen et al, 2020	✗	✓	✗	✗	✓	✗	✗	✗	✗
Joseph M.Taft et al, 2021	✗	✓	✗	✗	✓	✓	✗	✓	✗

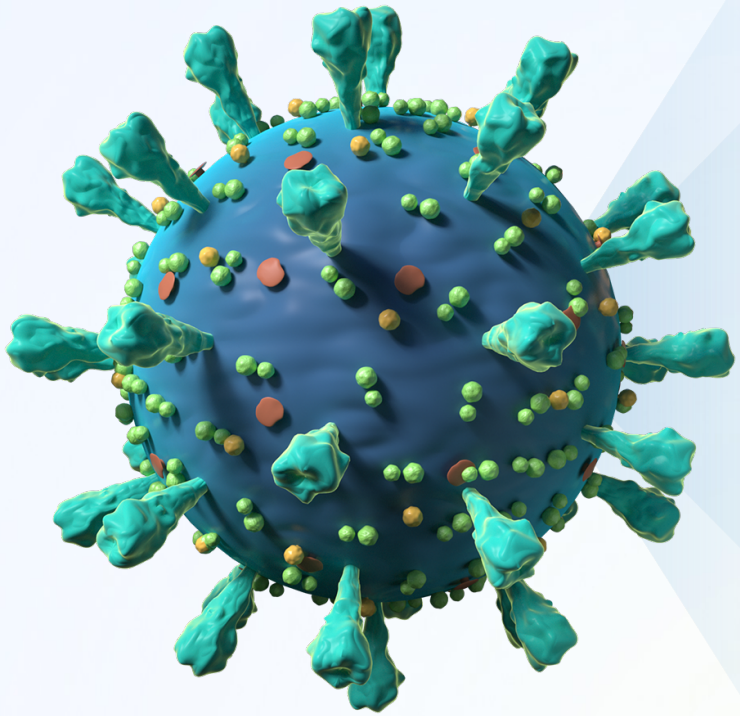
Alpha	BF.7
Beta	BQ.1
Gamma	BF.14
Delta	BA.2.75.2
Omicron BA.5	BA.4.6

100% success rate for the prediction of VOCs, except Omicron

Most of VUMs are predicted accurately

Algorithm Achieved SOTA





Background
& Highlight

Technical challenges
& Innovations

Science results
& Key performance

Summary
& Prospect

High performance prediction of high-dimensional mutational space

Number of Mutations = (amino acid types)$^{\text{sequence length}}$ = 20^{201} = 3.213876×10^{261}

The length of RBD domain is 201 (331-531)

There are 20 amino acids

We need high-throughput variants generation and screen!

Innovation 1: mutation simulation

Simulate the natural process

"NITNLCP…CGPPKST "

"NITDLCP…CAPKKST "

…

"KITNLKP…CYPKKFT "

Millions of RBDs

Process

Algorithm Innovation

10^9

10^8

10^7

…

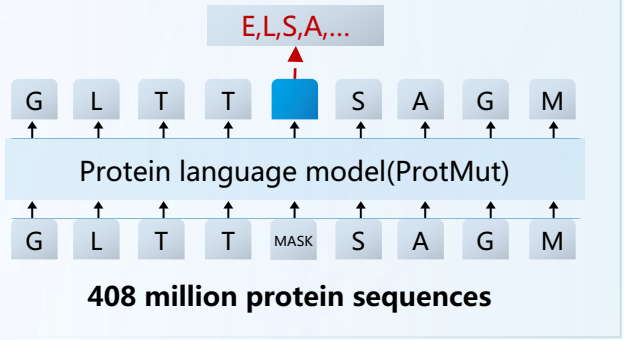
10^4

High-throughput screening

Fine-tune

Generate

AI-HPC Innovation



408 million protein sequences

ProtMut

Input

RBD sequences mutation probability

Innovation 1: mutation simulation

Reduction of mutational space

Learning of virus RBD sequence

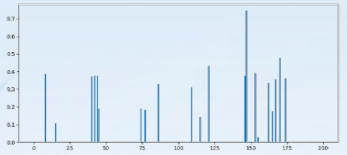
Screening

Recursively fine-tuning

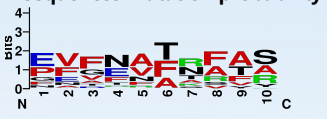
...

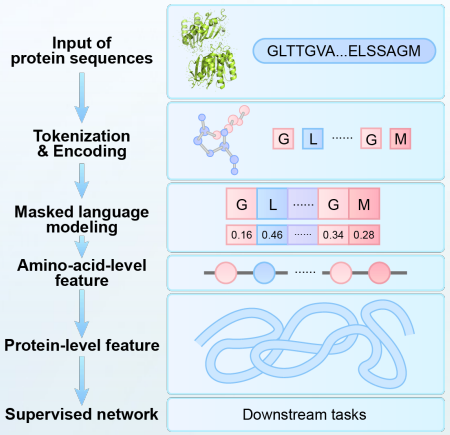
reduced mutational space

Mutation probability distribution of viruses in the real world



RBD sequences mutation probability





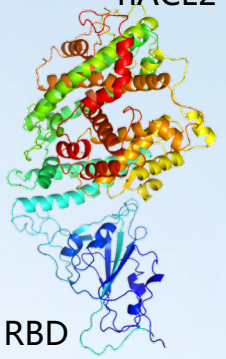
Mutation generator based on natural proteins and SARS-CoV-2 sequences

Simulation of evolution selection of viruses in the real world

High-throughput screening

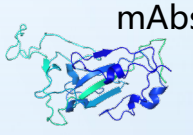
ACE2 binding

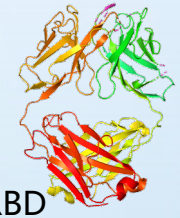
hACE2



RBD

mAbs



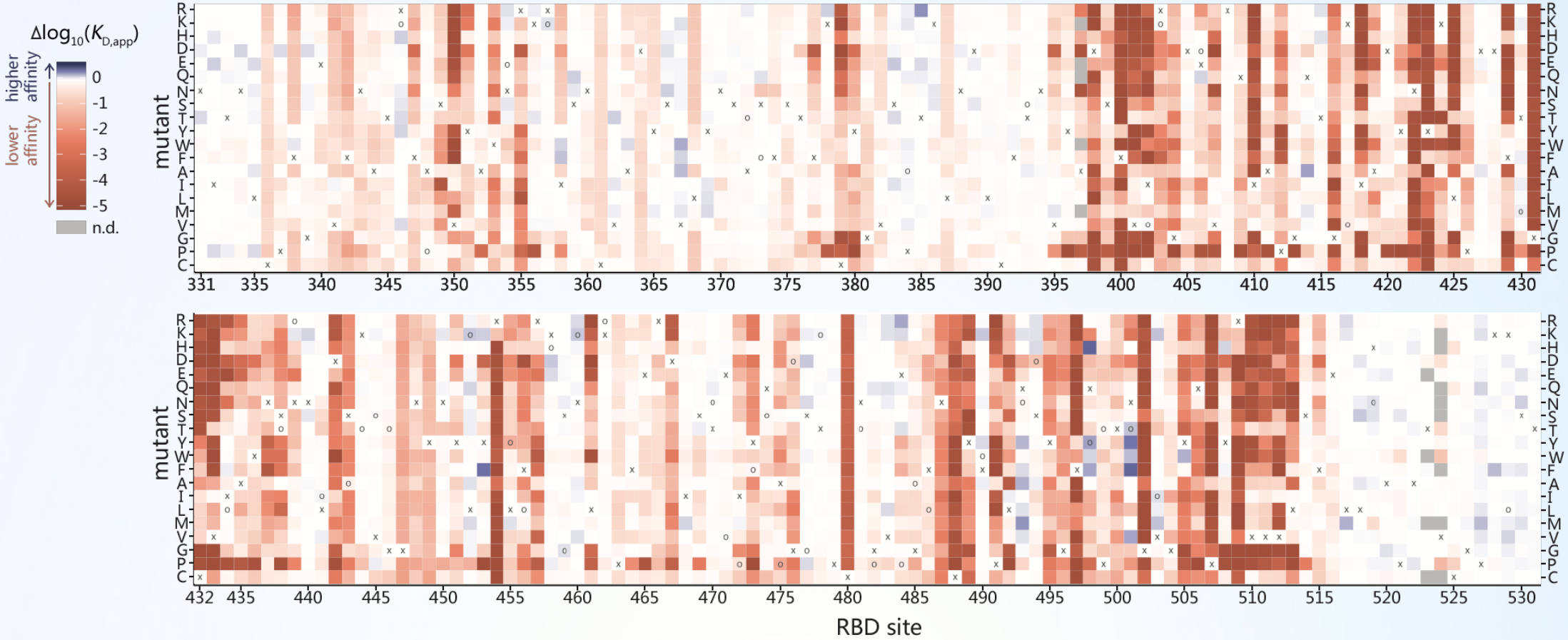


RBD

Antibody escape

Innovation 2: science-driven prediction

Mutation effects on **binding**

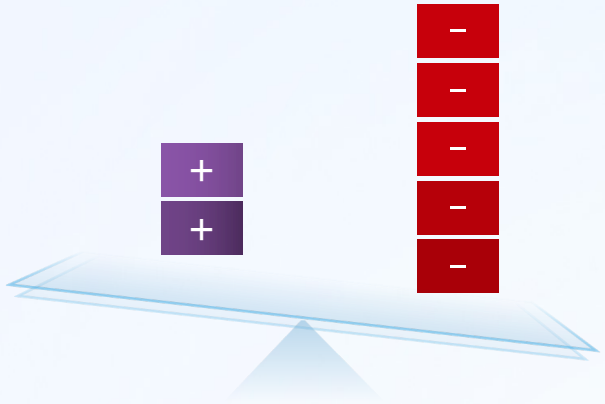


Deep mutational scanning

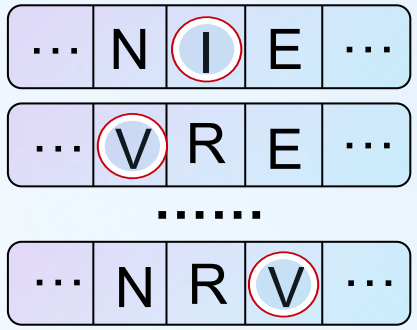
Starr T N, Greaney A J, Hilton S K, et al. Cell, 2020, 182(5): 1295-1310. e20.

Innovation 2: science-driven prediction

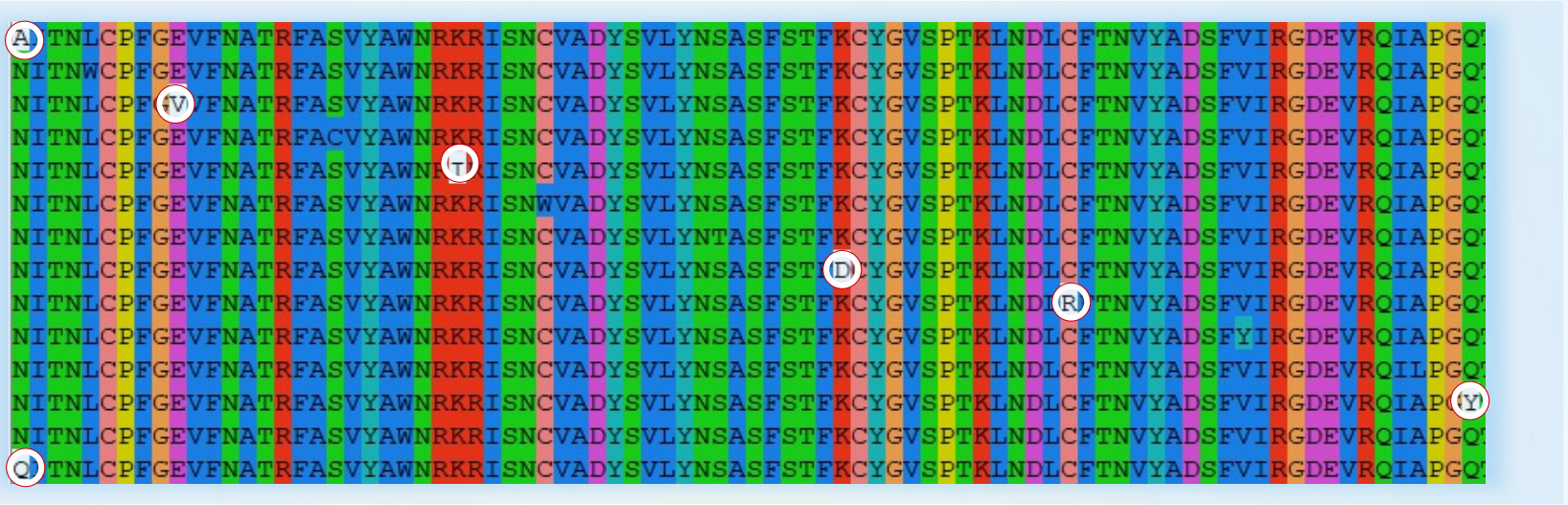
Data characteristics 1:
serious category imbalance



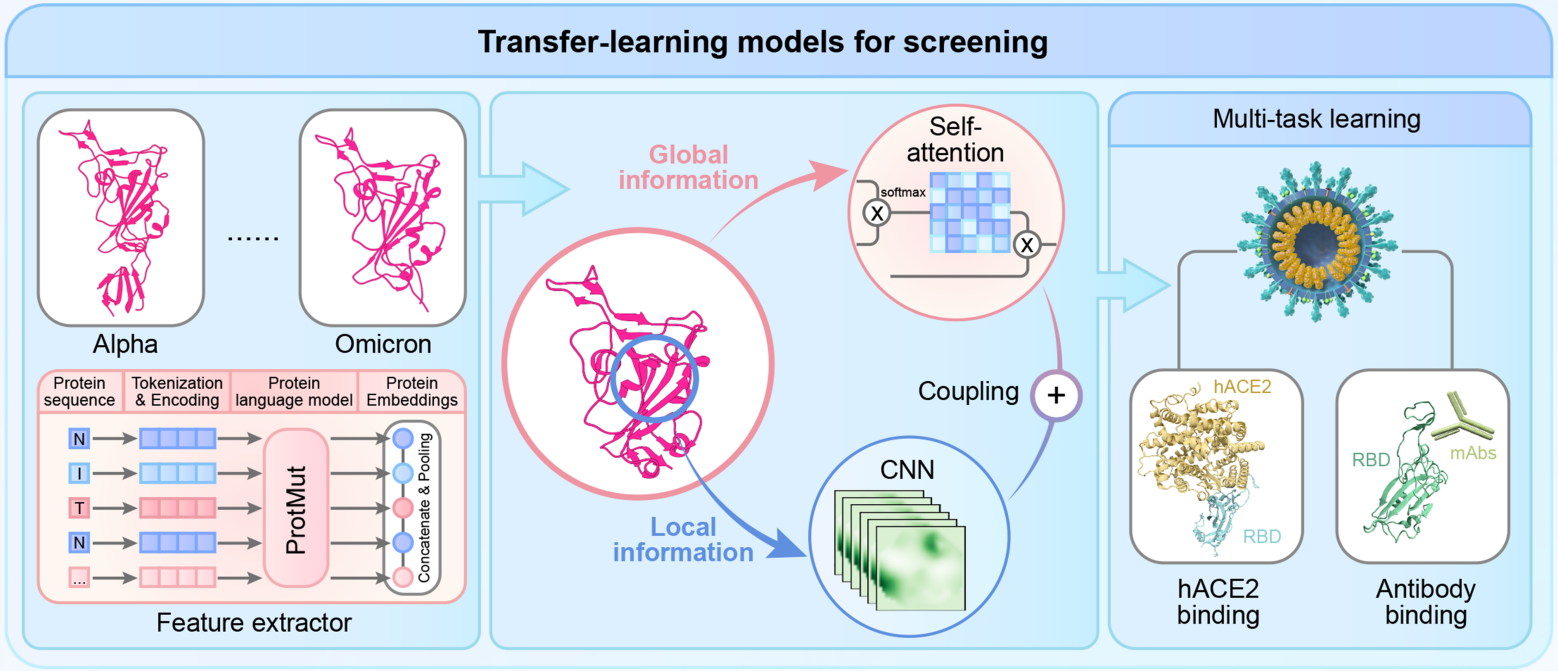
Data characteristics 2:
very similar between sequences

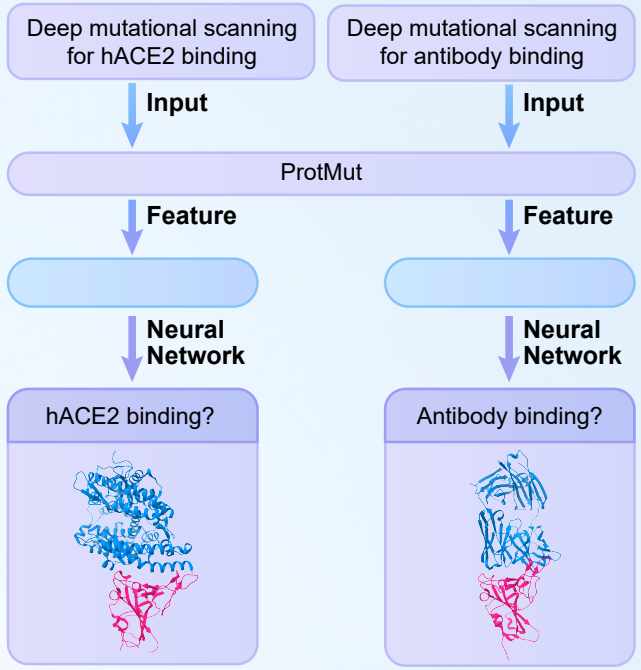


Sequences of DMS



Innovation 2: science-driven prediction

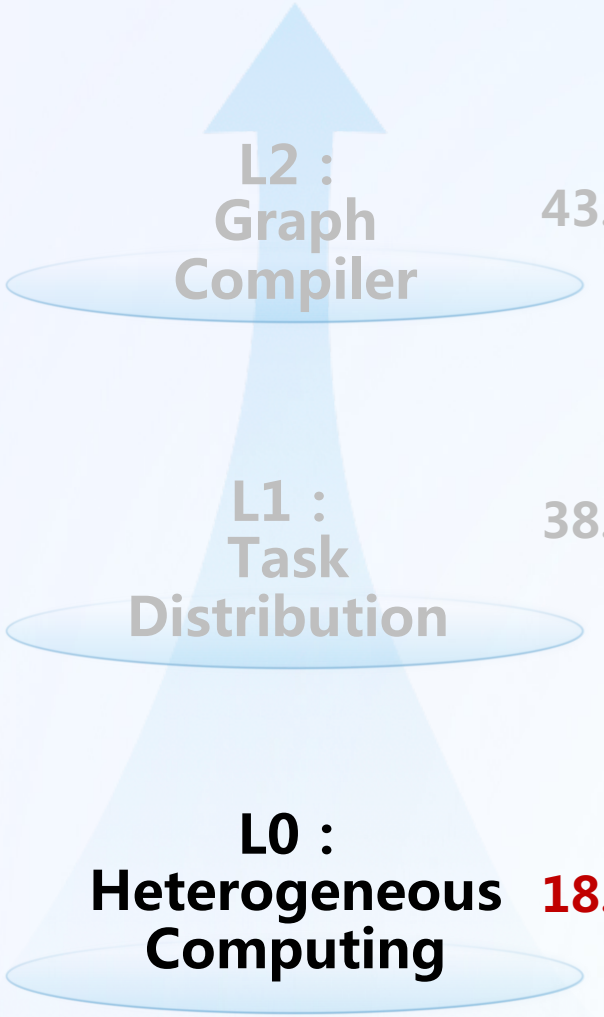




Variants	RBD-31	RBD-32	RBD-33	RBD-34	RBD-312	RBD-36	.....	RBD-71
Consistent with wet-lab experiments?	True	True	True	True	True	True	True	True

Zahradník J, Marciano S, Shemesh M, et al. Nature microbiology, 2021, 6(9): 1188-1198.

Innovation 3: multi-level optimization strategy

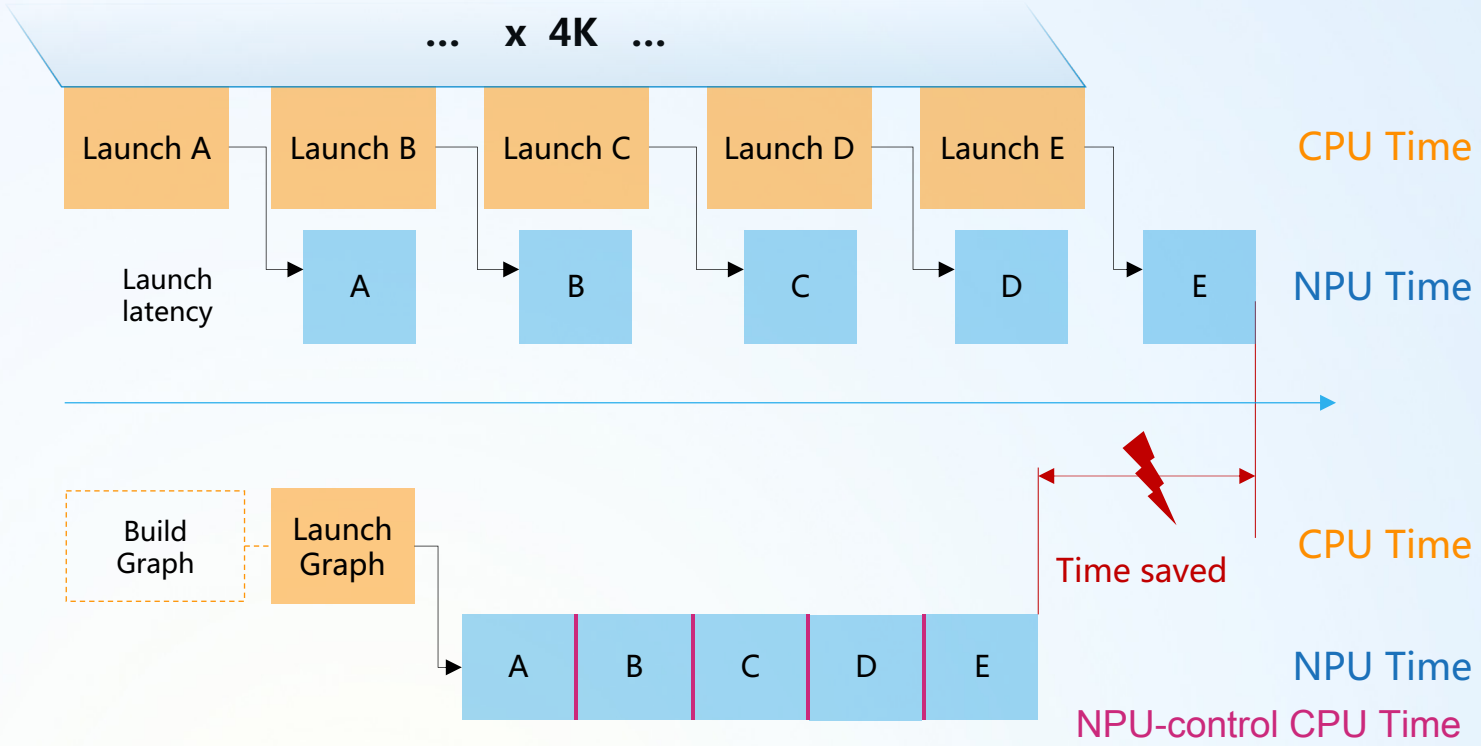


43.08%

38.69%

18.23%

Challenge : total 4k operators , cost 82ms\step.



Optimization : using the special NPU-control cpu to run full sink mode
Improvment : 82 ms/step , up 18.23%

Innovation 3: multi-level optimization strategy

L2 : Graph Compiler

43.08%

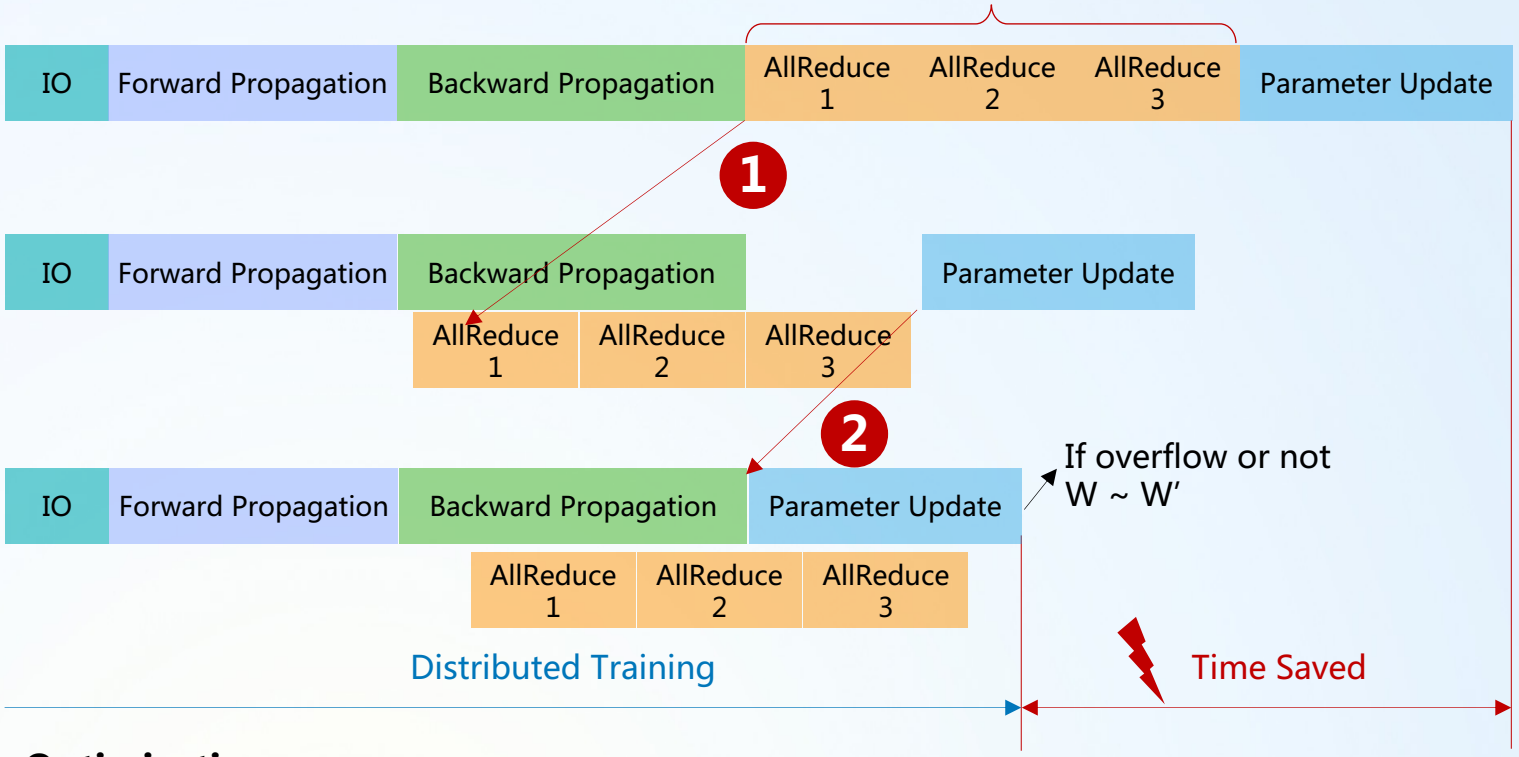
L1 : Task Distribution

38.69%

L0 : Heterogeneous Computing

18.23%

Challenge : In large scale distributed system , communication cost the most.



Optimization :
① Computational Communication overlap
② Parameter Update Communication overlap
Improvment : 174 ms/step , up 38.69%

Innovation 3: multi-level optimization strategy

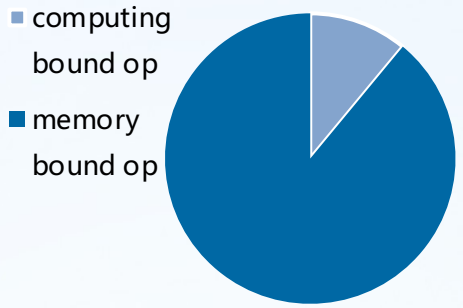
L2 : Graph Compiler — 43.08% ↑

L1 : Task Distribution — 38.69% ↑

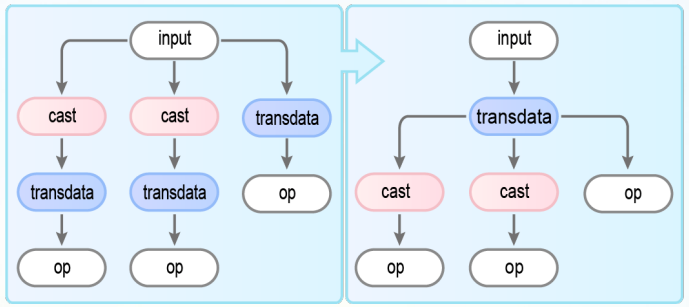
L0 : Heterogeneous Computing — 18.23% ↑

Challenge : large nums of memory bound operators lead to extra time overhead

Cube task OP nums distribution



Operator fusion

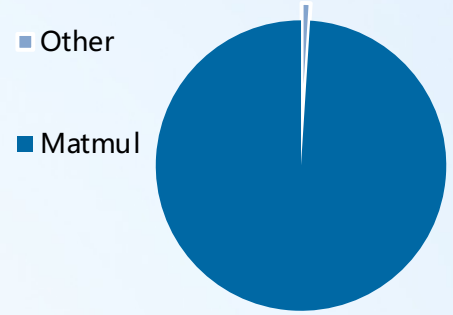


Optimization :
- layerNorm fusion
- matmul +add
- allreduce+gradients

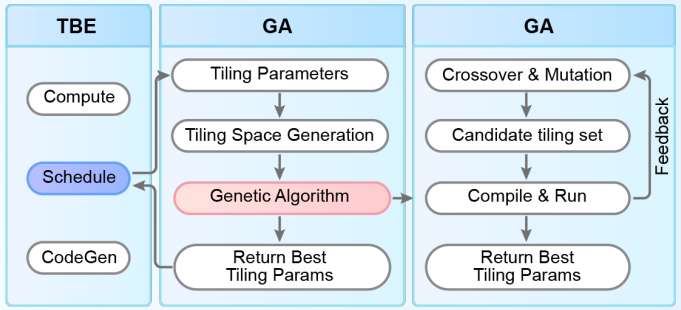
Improvment : up 30%

Challenge : Matmul computation >99% and time consumption>40%

Matmul ratio of computation

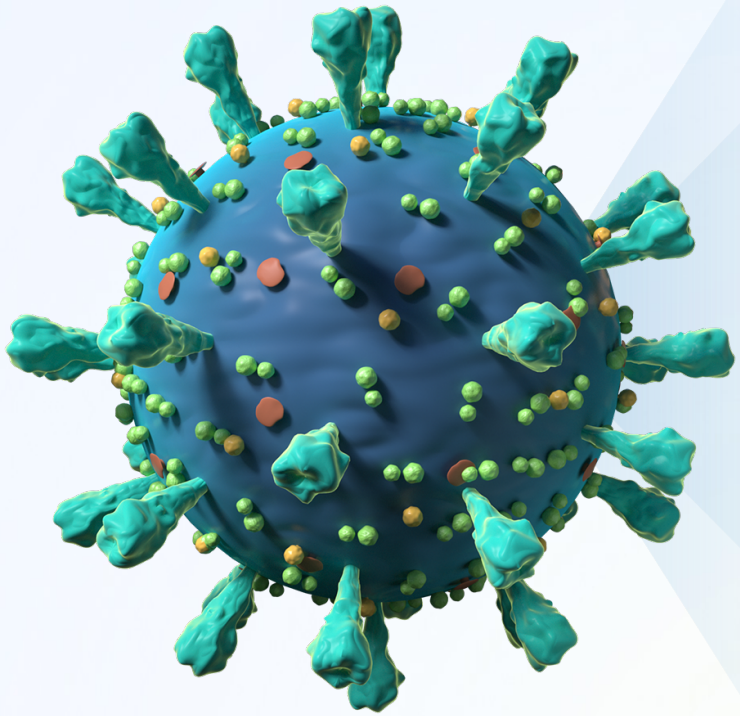


Operator auto-tuning



Optimization : auto-tuning , using GA\RL algorithm

Improvment : up 13%



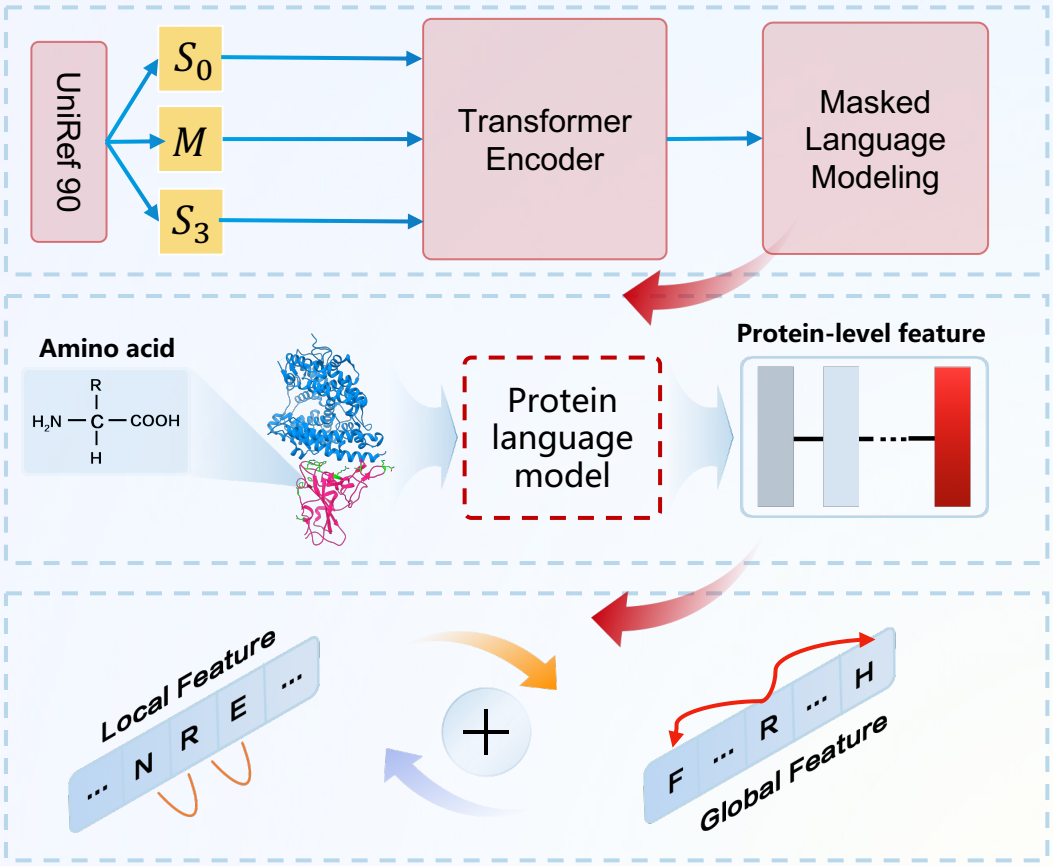
Background
& Highlight

Technical challenges
& Innovations

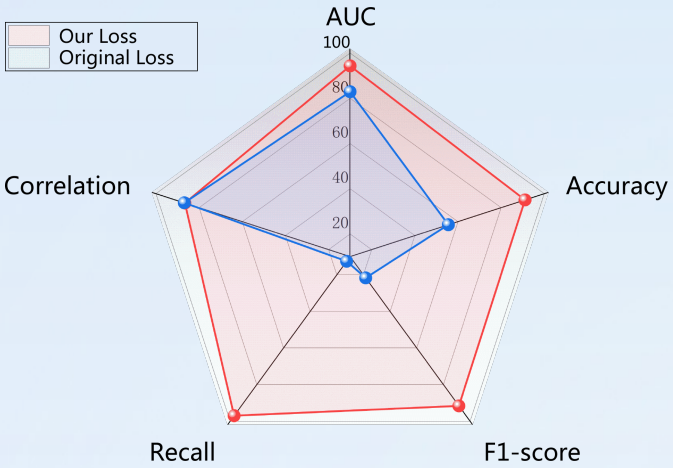
Science results
& Key performance

Summary
& Prospect

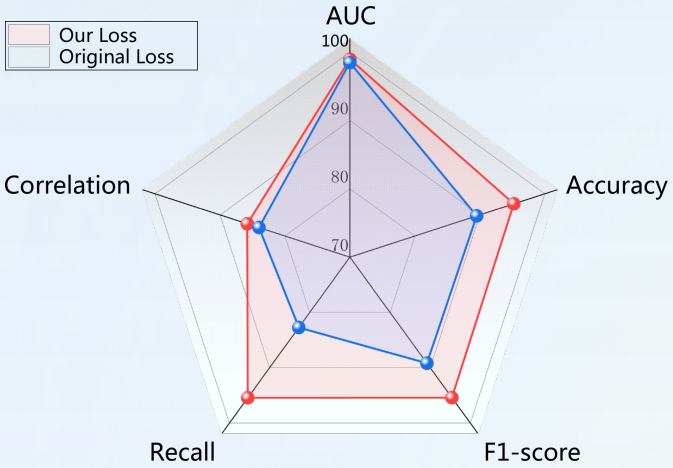
Reveal the latent virus RBD sequence information



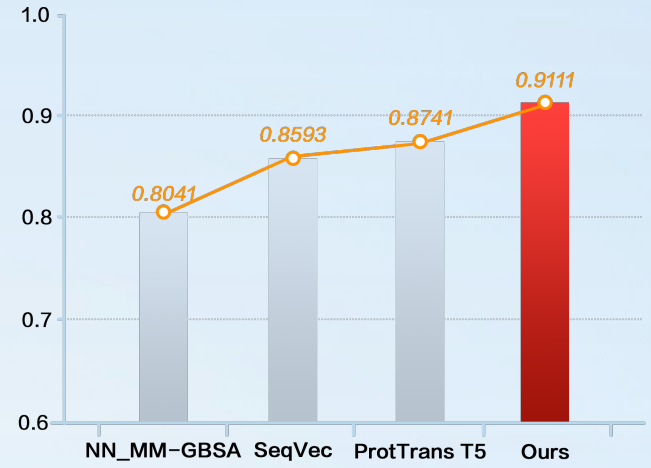
Algorithm achieved SOTA



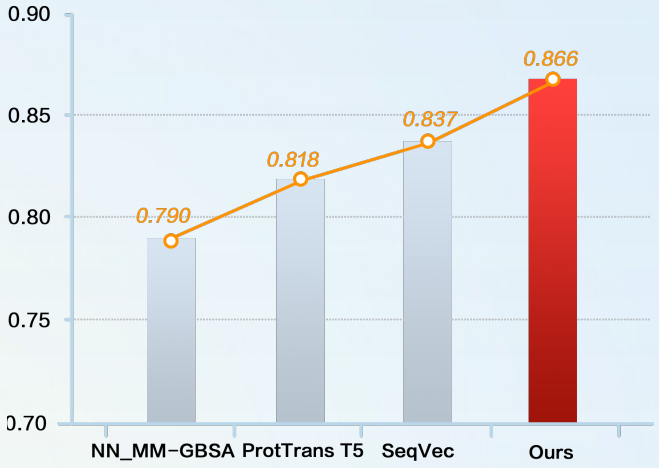
ACE2 binding



Antibody escape



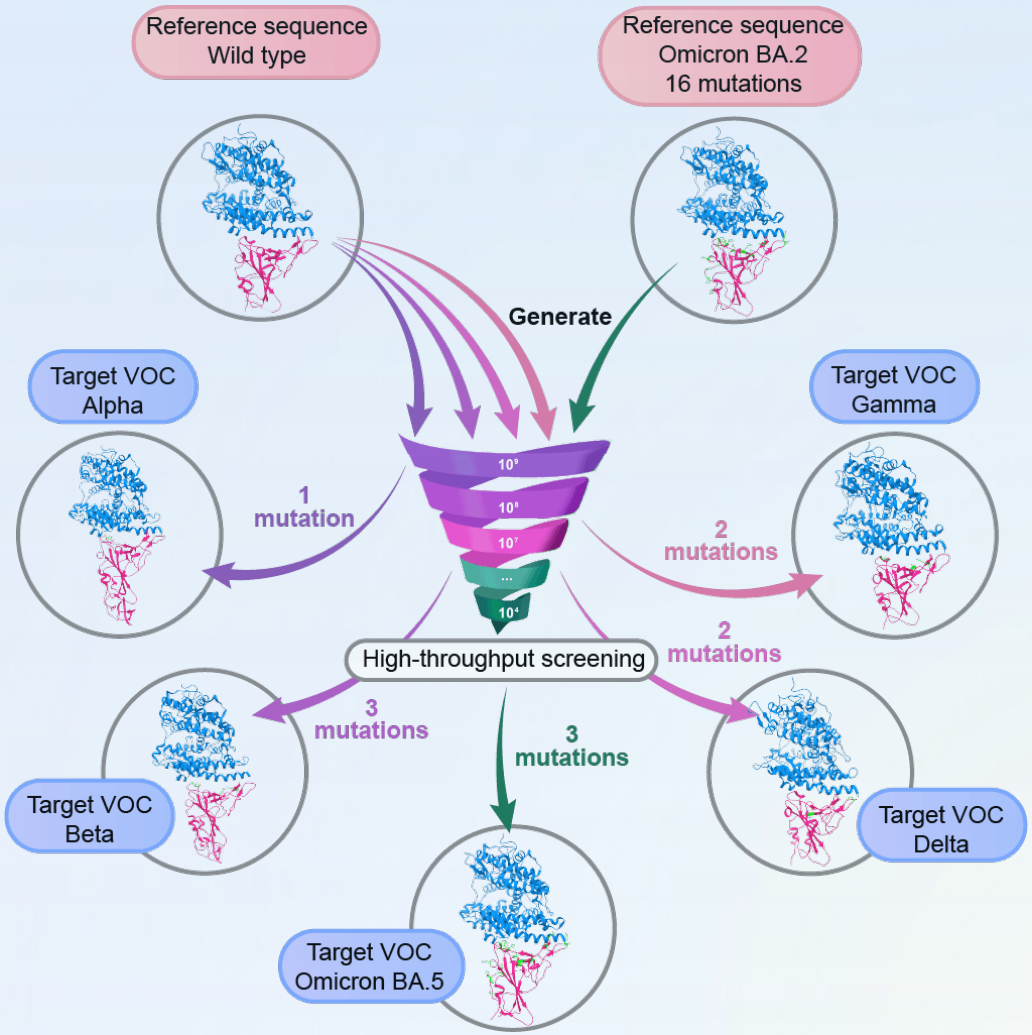
Accuracy



Correlation

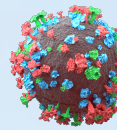
Prediction of VOCs and VUMs

Two stage validation of VOCs

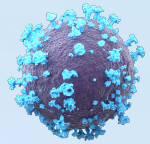


- **Validation before Omicron**
  - **WT ------Alpha Complete**
  - **WT ------Beta Complete**
  - **WT ------Gamma Complete**
  - **WT ------Delta Complete**
- **Validation after Omicron**
  - **Omicron BA.2 ------BA.5 Complete**

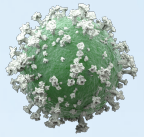
Validation of VUMs



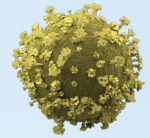
BF.7 ✓



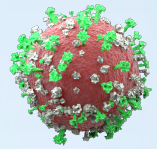
BF.14 ✓



BQ.1 ✓



BA.2.75.2 ✓



BA.4.6 ✓

High throughput generating & screening

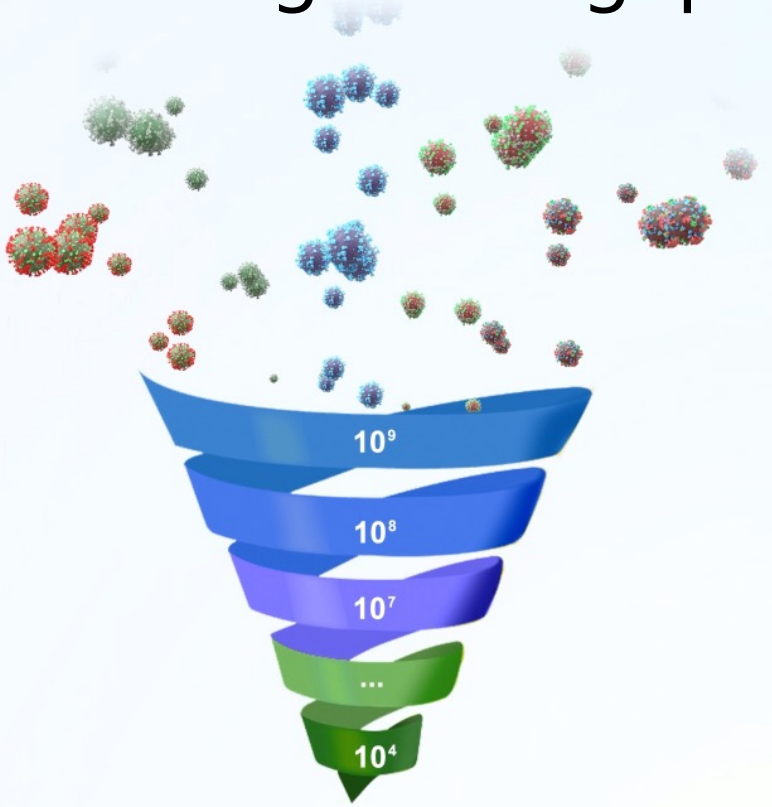


Table1:High throughput screening of various variants

Variant	1st screening*	2st screening**
Alpha	39.8%	2.0%
Beta	13.3%	51.3%
Gamma	45.2%	33.8%
Delta	46.7%	19.1%
Omicron BA.5	90.4%	80.2%

*Proportion after hACE2 binding screening

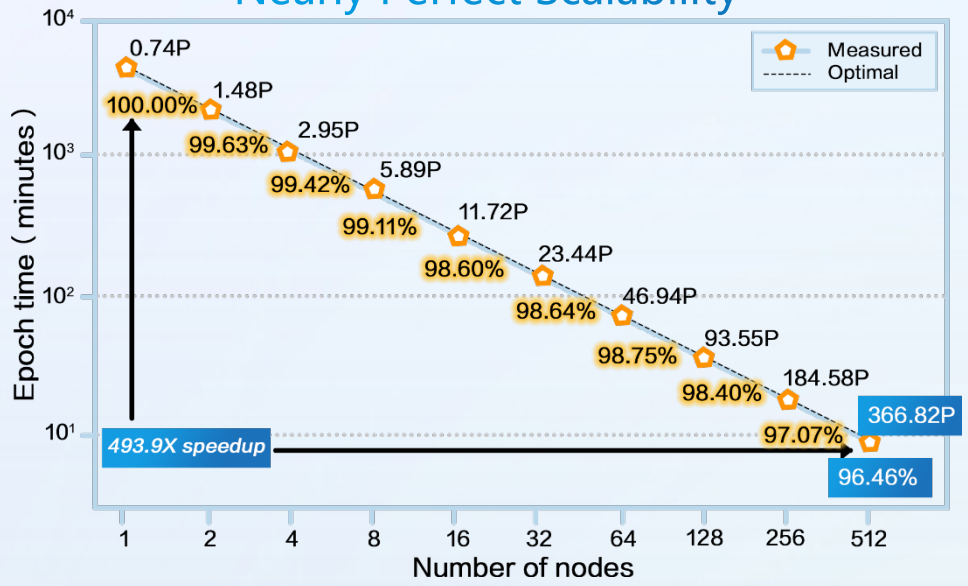
**Proportion after antibody binding screening

Results on full-scale system

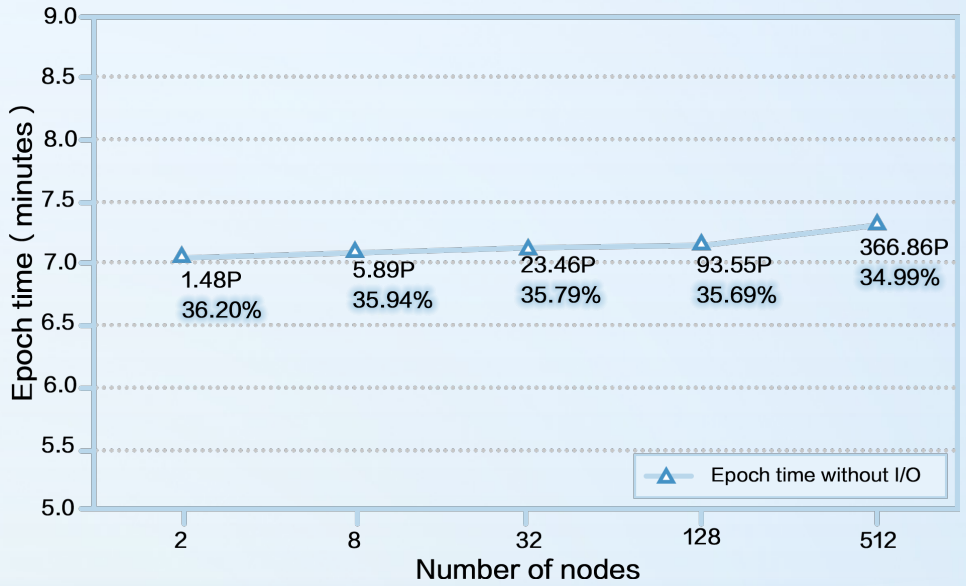
VOCs	Number of generation	Time of generation & screening (h)
Alpha	10^6	0.00024
Beta	10^{11}	24
Gamma	10^{10}	2.4
Delta	10^8	0.024
Omicron BA.5	10^9	0.24

Strong scaling & Weak scaling

Nearly Perfect Scalability



Stable Utilization

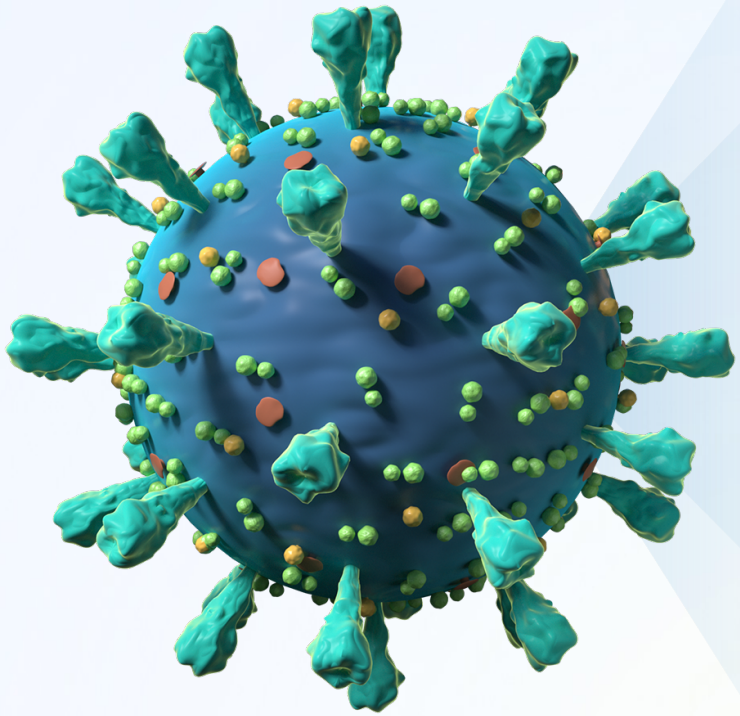


- 4096 Neural Processing Units (NPUs) and 2048 Kunpeng 920 Processors, provides 1 EFLOPS.
- 512 Nodes , each node has 4 Processors and 8 NPUs.
- Each NPU with 32GB memory provides 256 TFLOPS.
- Every Processor is connected via 100G RoCE.

Key performance

- **63.4** zettaflops*
- **366.8** PFLOPS on 4096 NPUs
- **96.5%** scalability
- **493.9X** speedup
- **34.9%** utilization (training)

* Measured in mixed-precision



Background
& Highlight

Technical challenges
& Innovations

Science results
& Key performance

Summary
& Prospect

63.4 zettaflops in exchange for almost 5 months

Scientific Results	HPC Performance
First simulation of SARS-CoV-2 RBD	**Achieving over 19.8 zettaflops**
Algorithm achieved SOTA	**366.8 PFLOPS on 4096 NPUs**
Almost all VOCs are predicted accurately	**96.5% scalability**
Most of VUMs are predicted accurately	**493.9X speedup**

The period between outbreaks ≈ 5 months

***In silico* simulation time on full-scale system ≈ 2 days**