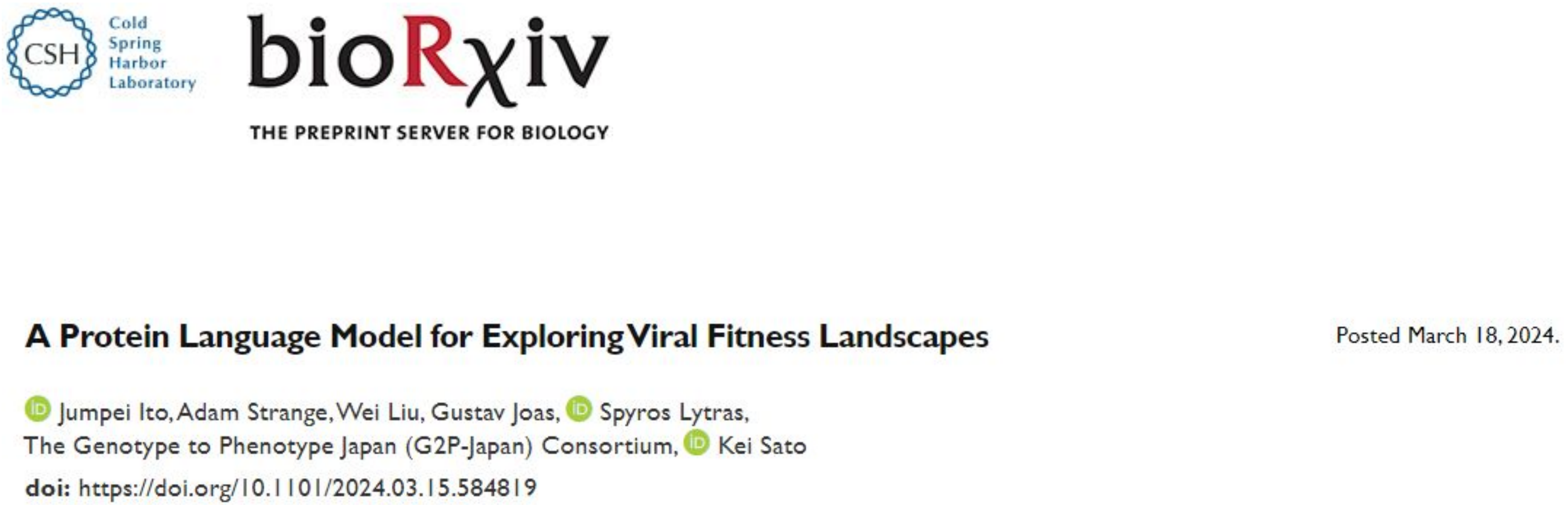


Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

A Protein Language Model for Exploring Viral Fitness Landscapes

Posted March 18, 2024.

Jumpei Ito, Adam Strange, Wei Liu, Gustav Joas, Spyros Lytras,
The Genotype to Phenotype Japan (G2P-Japan) Consortium, Kei Sato

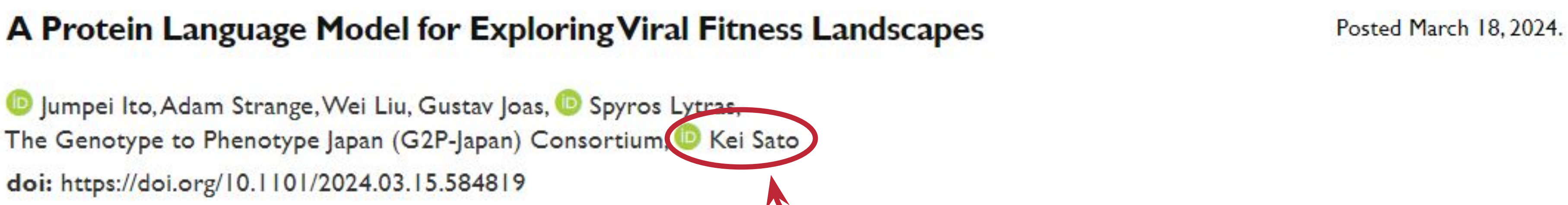
doi: https://doi.org/10.1101/2024.03.15.584819

https://github.com/TheSatoLab/CoVFit

CSH Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY



A Protein Language Model for Exploring Viral Fitness Landscapes

Posted March 18, 2024.

Jumpei Ito, Adam Strange, Wei Liu, Gustav Joas, Spyros Lytras,
The Genotype to Phenotype Japan (G2P-Japan) Consortium, Kei Sato

doi: https://doi.org/10.1101/2024.03.15.584819

Number of **(IF > 50)** papers **since 2022**:

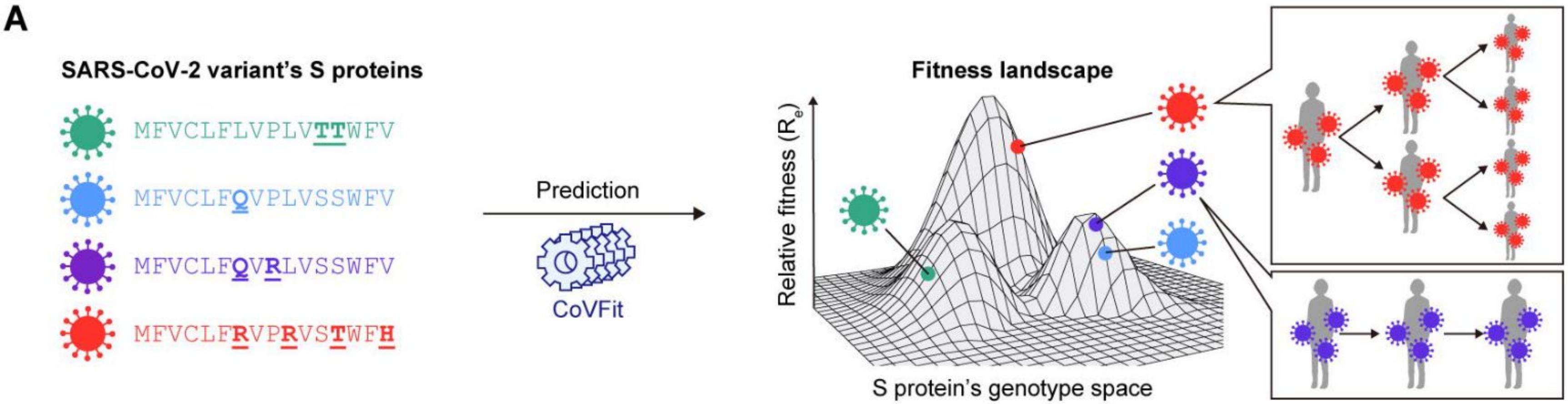
- Nature: 3
- Cell: 3
- The Lancet Infectious Diseases: 7

https://github.com/TheSatoLab/CoVFit

MAIN GOAL

- **predicting SARS-CoV-2 fitness** from S protein **AA sequence**

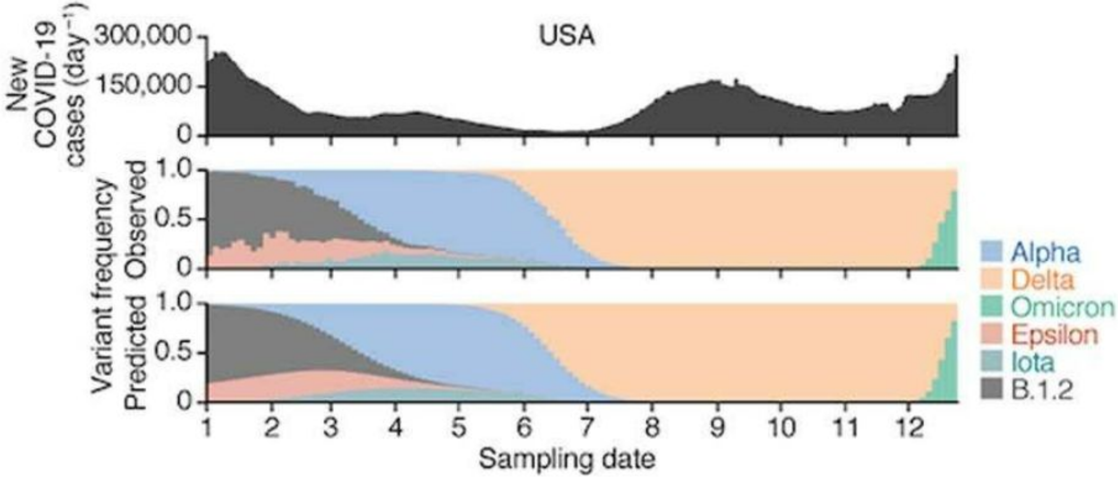
  fitness defined for whole S protein sequences and **not for single mutations (epistasis)**



TARGET: FITNESS

- **fitness ~ effective reproduction number** (in each country)
  - based on **count data** obtained from GISAID for each **haplotype**

    (S protein haplotypes are defined by a unique set of AA mutations; they do not necessarily correspond to traditional lineage definitions)



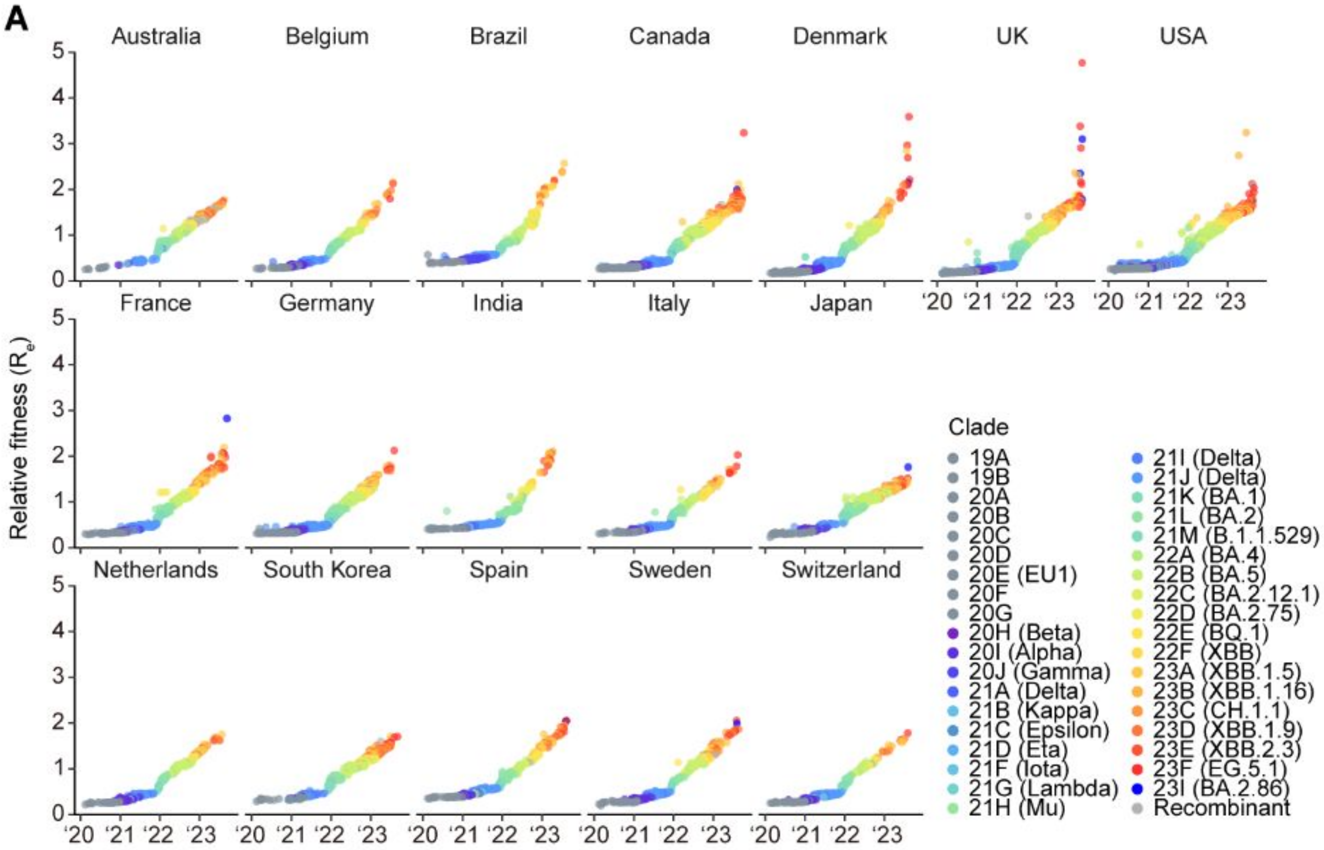
- they assume that count data follows a **multinomial distribution**, with **time-dependent probabilities** of each category (lineage)
- the time dependent probabilities follow the form:

$$p_l(t) = \mathrm{softmax}\{b_{0,l} + b_{1,l}t\}$$

- **relative effective reproduction number**:

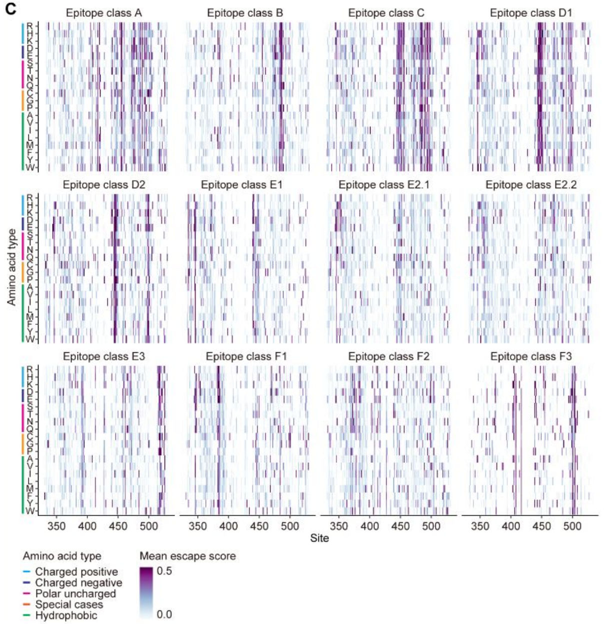
$$R_{e,l} = e^{b_{1,l}\tau}$$

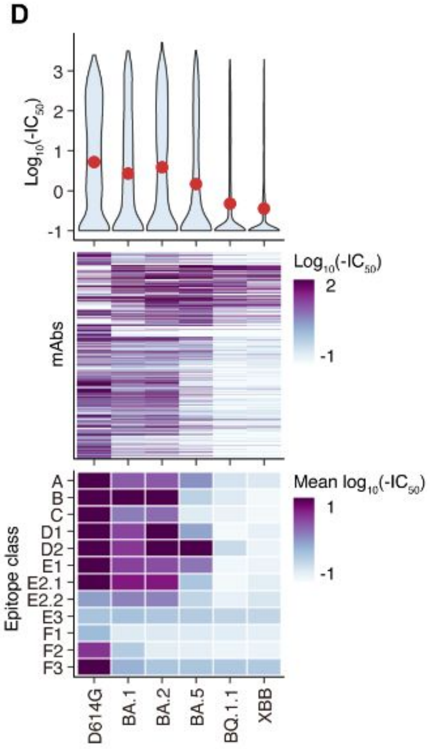
TARGET: FITNESS



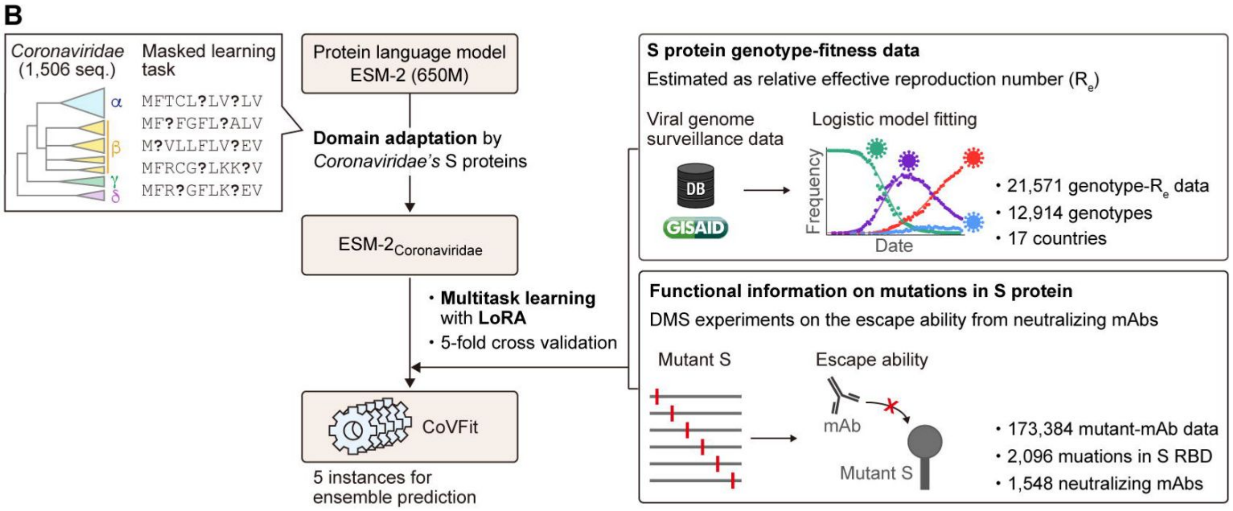
TARGET: ESCAPE SCORES

- **mAb escape scores**
  - based on **DMS experiments**
  - they do not directly influence fitness and are not a "real" target, they are added to the model for **better generalisation**





MODEL



B
Coronaviridae (1,506 seq.)
Masked learning task
α
β
γ
δ
MFTCL?LV?LV
MF?FGFL?ALV
M?VLLFLV?EV
MFRCG?LKK?V
MFR?GFLK?EV
Protein language model ESM-2 (650M)
Domain adaptation by Coronaviridae's S proteins
ESM-2Coronaviridae
• Multitask learning with LoRA
• 5-fold cross validation
CoVFit
5 instances for ensemble prediction
S protein genotype-fitness data
Estimated as relative effective reproduction number (Re)
Viral genome surveillance data
DB
GISAID
Logistic model fitting
Frequency
Date
• 21,571 genotype-Re data
• 12,914 genotypes
• 17 countries
Functional information on mutations in S protein
DMS experiments on the escape ability from neutralizing mAbs
Mutant S
Escape ability
mAb
Mutant S
• 173,384 mutant-mAb data
• 2,096 muations in S RBD
• 1,548 neutralizing mAbs

INPUT:

first 1024 AAs of the S protein, tokenized

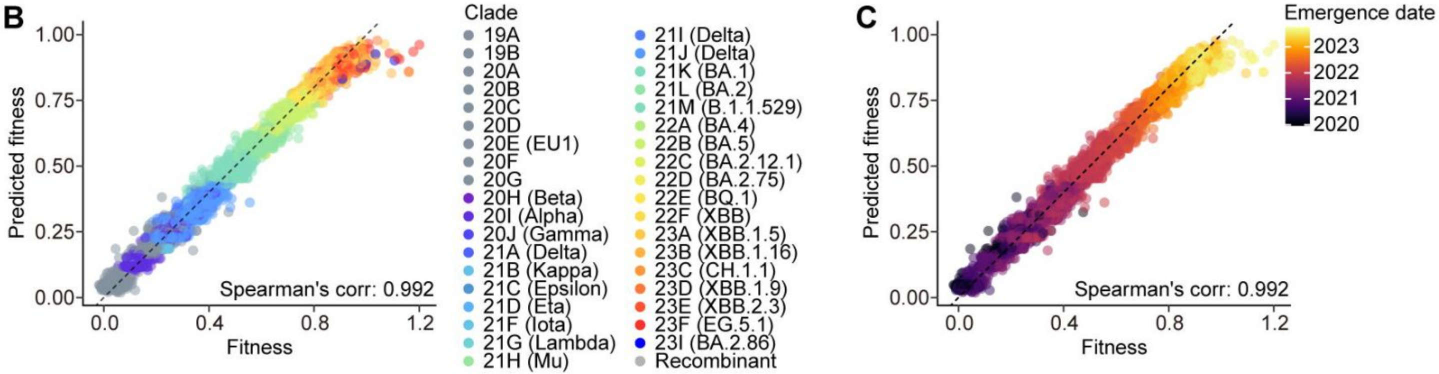
↓

OUTPUT:

17 (country-wise) **fitness scores +** 1548 mAb **escape scores**

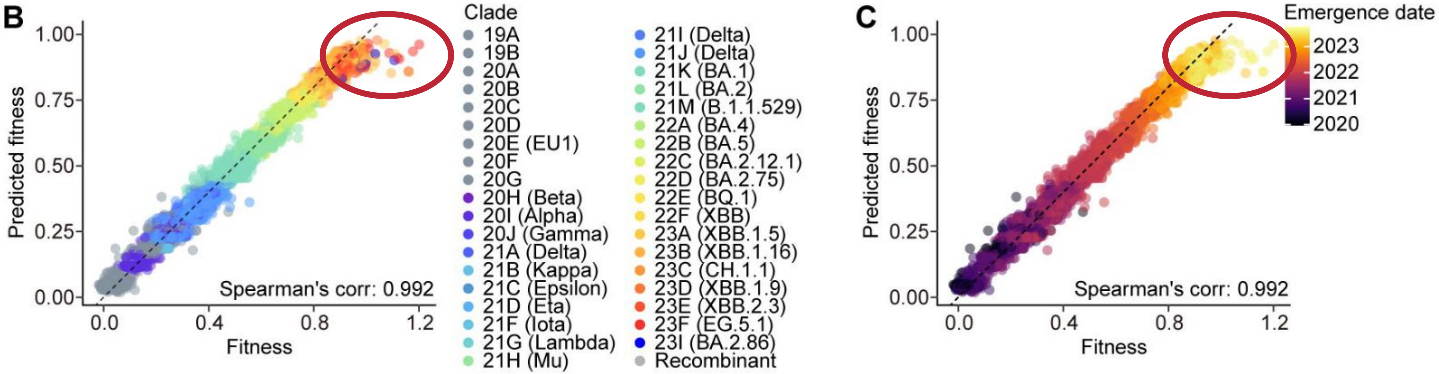
1. domain adaptation: fine-tuning ESM-2 for masked language modelling with historic Coronavirus S protein sequences **AND SARS-CoV-2 sequences prior to August 31, 2022 → ESM-2$_{\text{Coronaviridae}}$**
2. fine-tune ESM-2$_{\text{Coronaviridae}}$ with surveillance data (AA sequences) and **multi-task targets for generalisability** (R_e and mAb escape scores)

BASIC PERFORMANCE: FITNESS



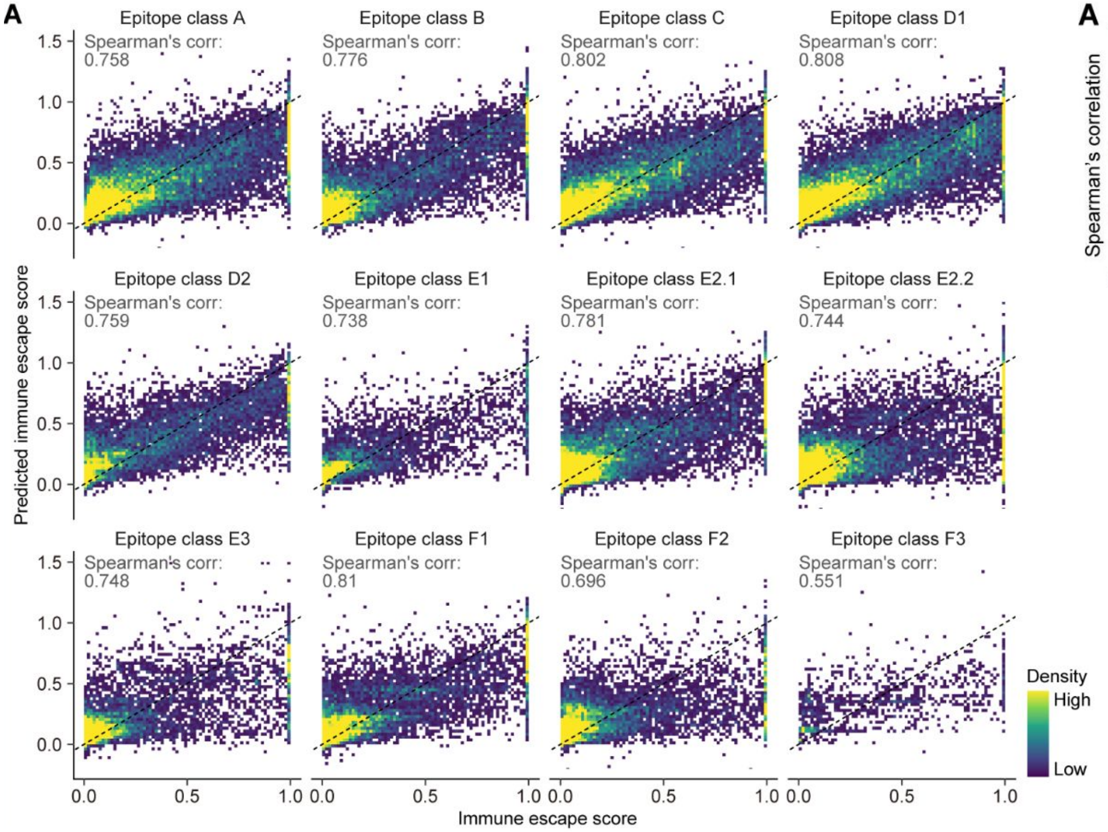
- good predictions for older sequences
- systematic offset for newer data, even with random train-test splits

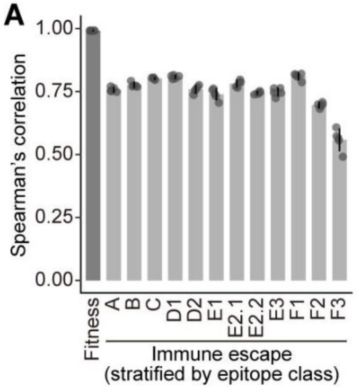
BASIC PERFORMANCE: FITNESS



- good predictions for older sequences
- systematic offset for newer data, even with random train-test splits

BASIC PERFORMANCE: IMMUNE ESCAPE

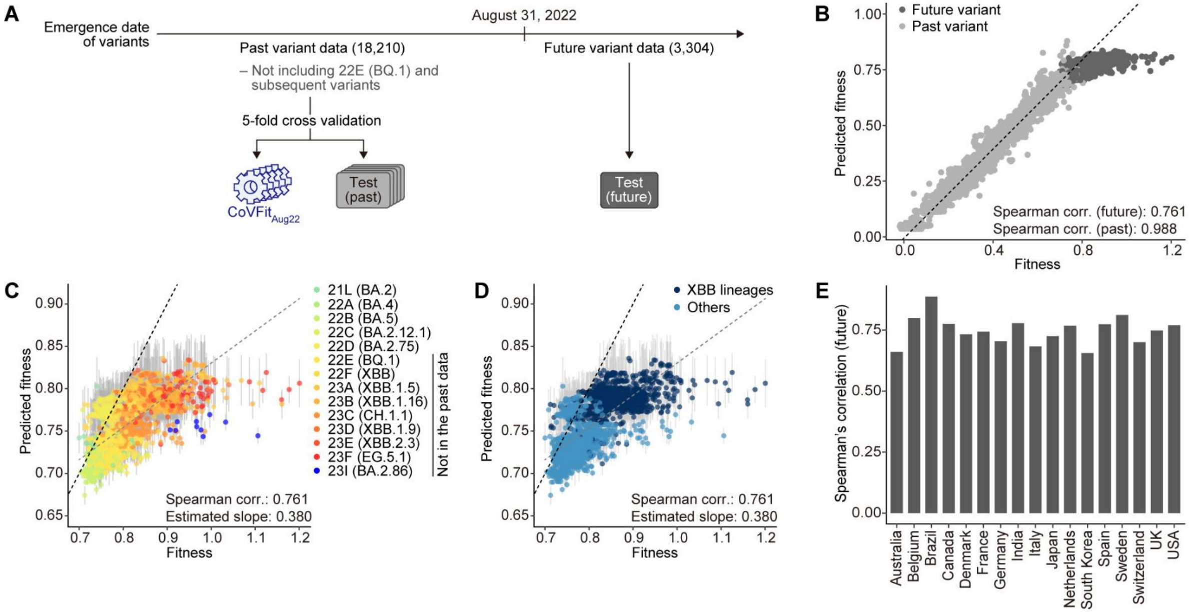




- positive correlations, but low accuracy
- not a main target, kept only for generalisability

REAL PERFORMANCE: FITNESS

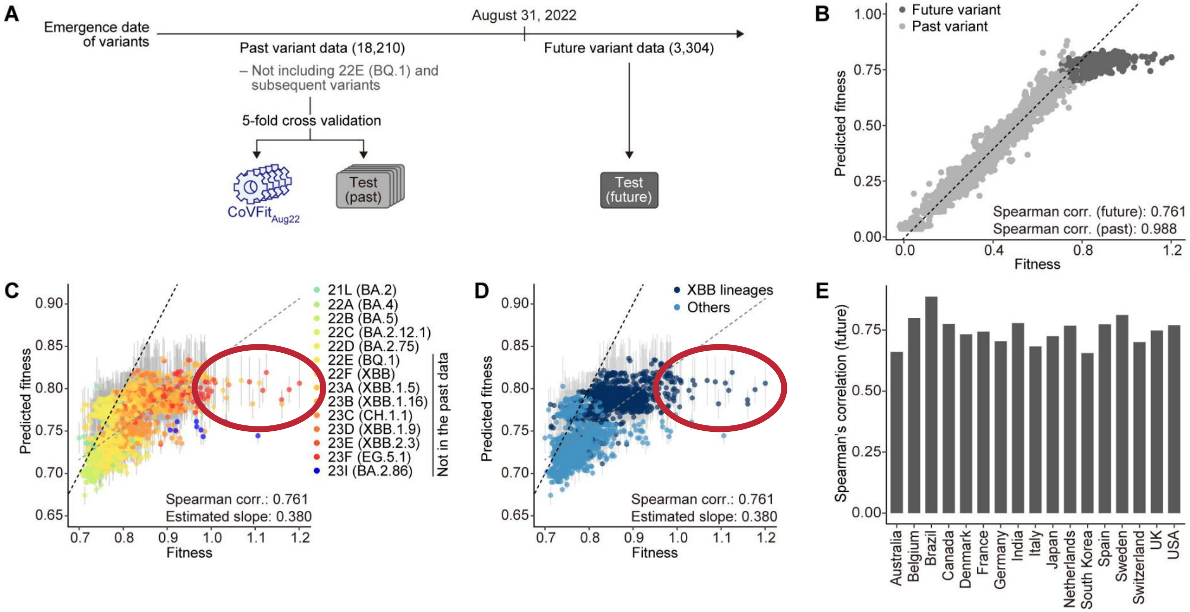
- with a **past vs. future split** for training and validation



- systematically underestimates fitness of future variants by a factor of ~ 0.38 and even lower for variants with very high fitness

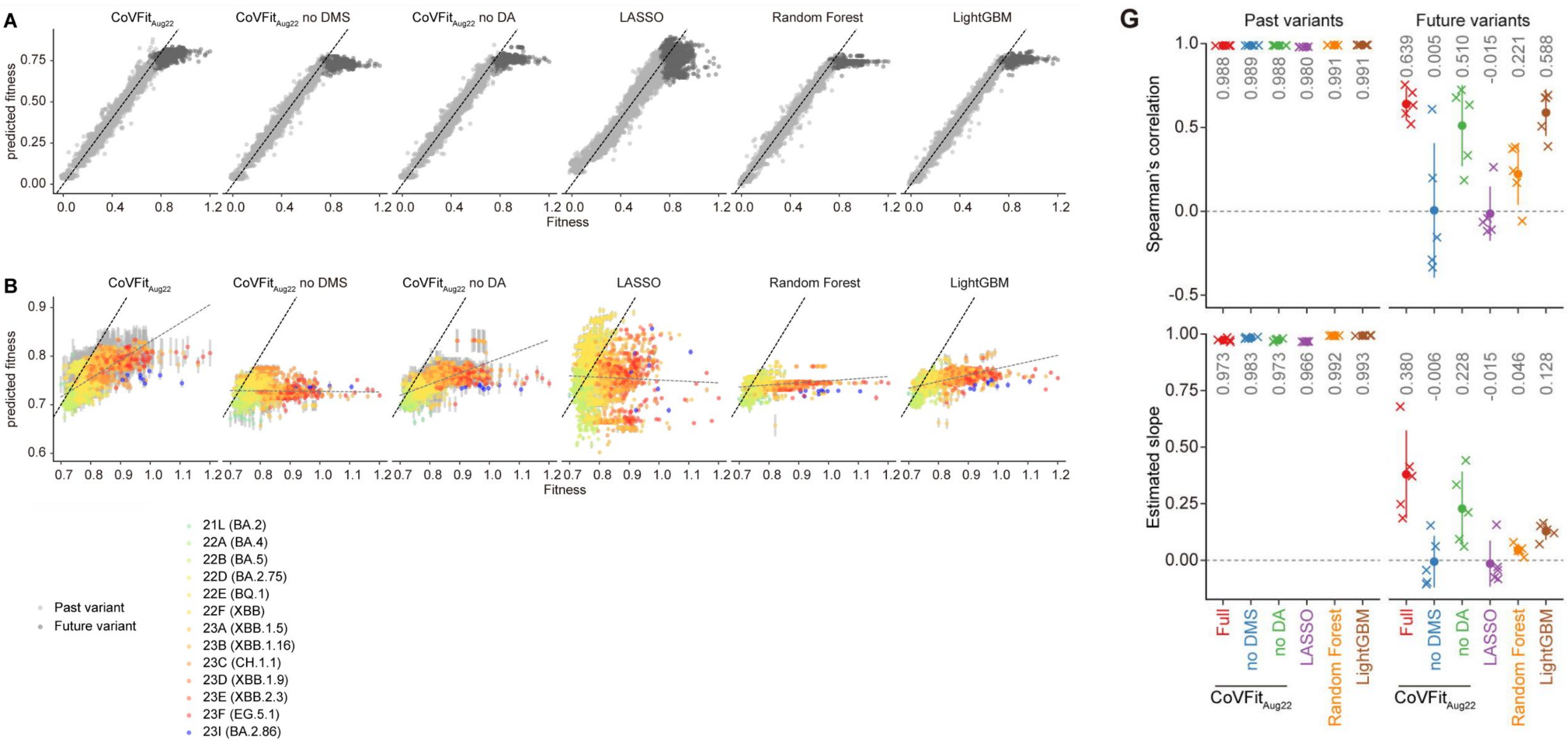
REAL PERFORMANCE: FITNESS

- with a **past vs. future split** for training and validation

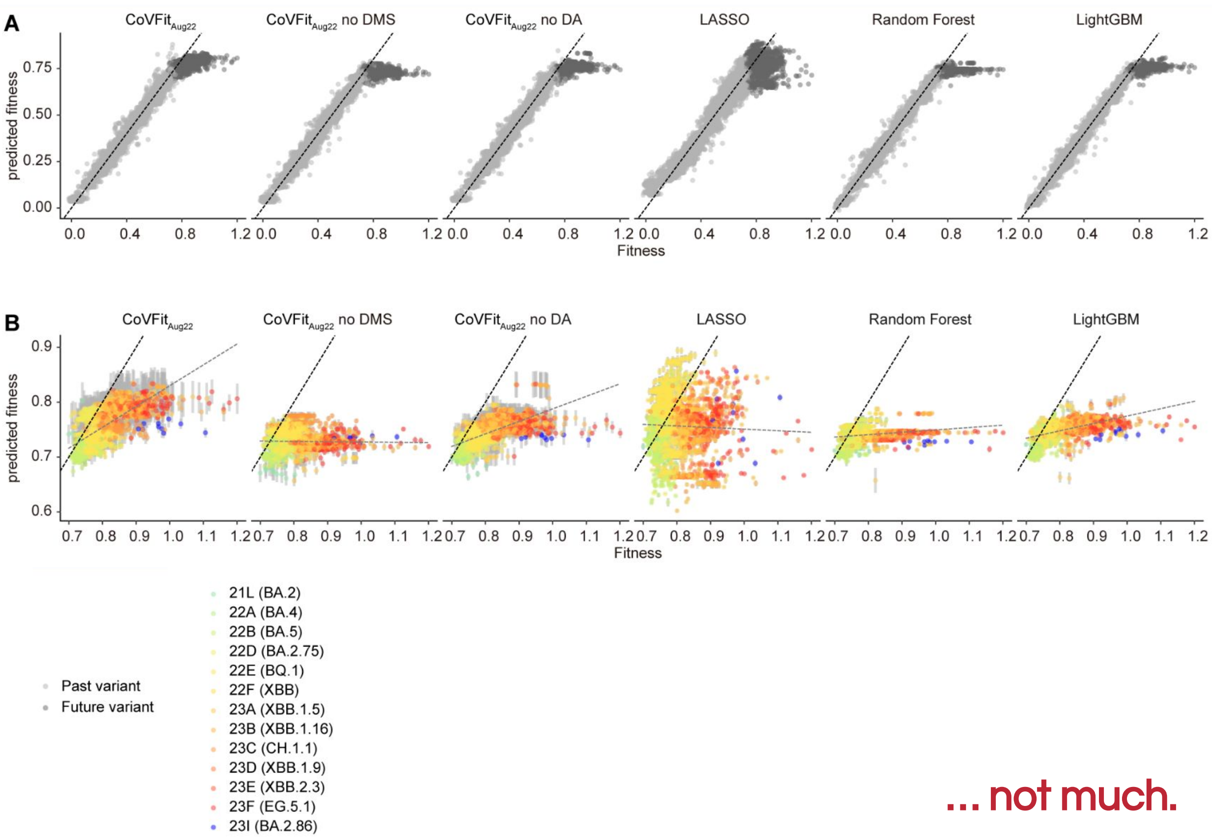


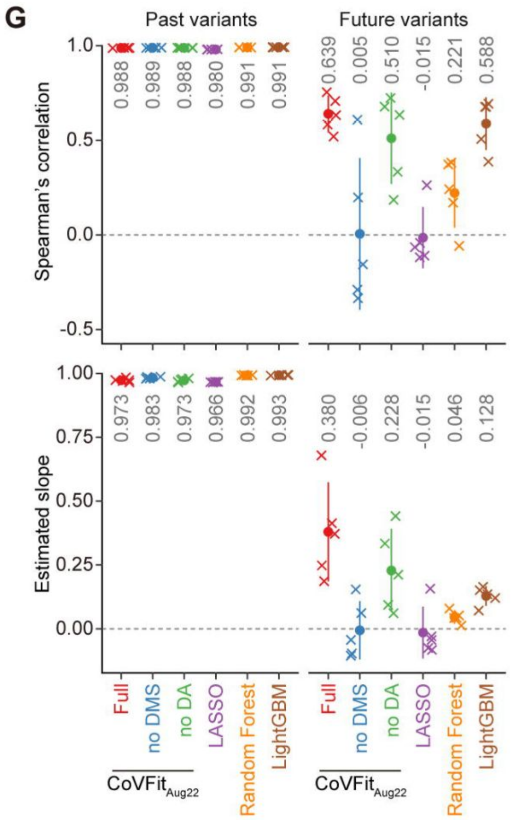
- systematically underestimates fitness of future variants by a factor of ~ 0.38 and even lower for variants with very high fitness

IS IT BETTER THAN SIMPLER MODELS?



IS IT BETTER THAN SIMPLER MODELS?

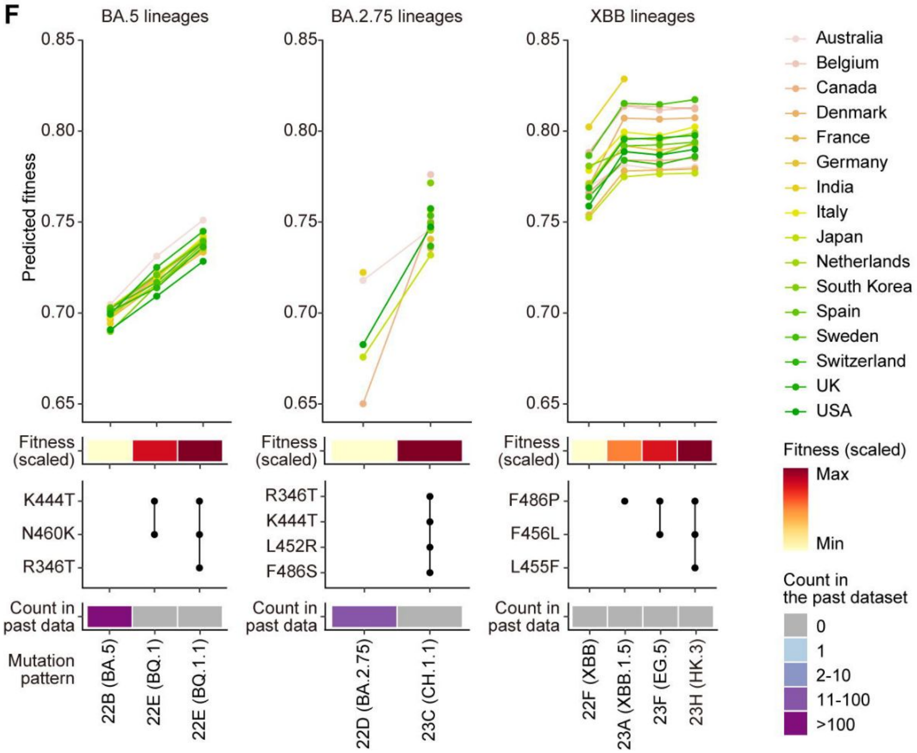




... not much.

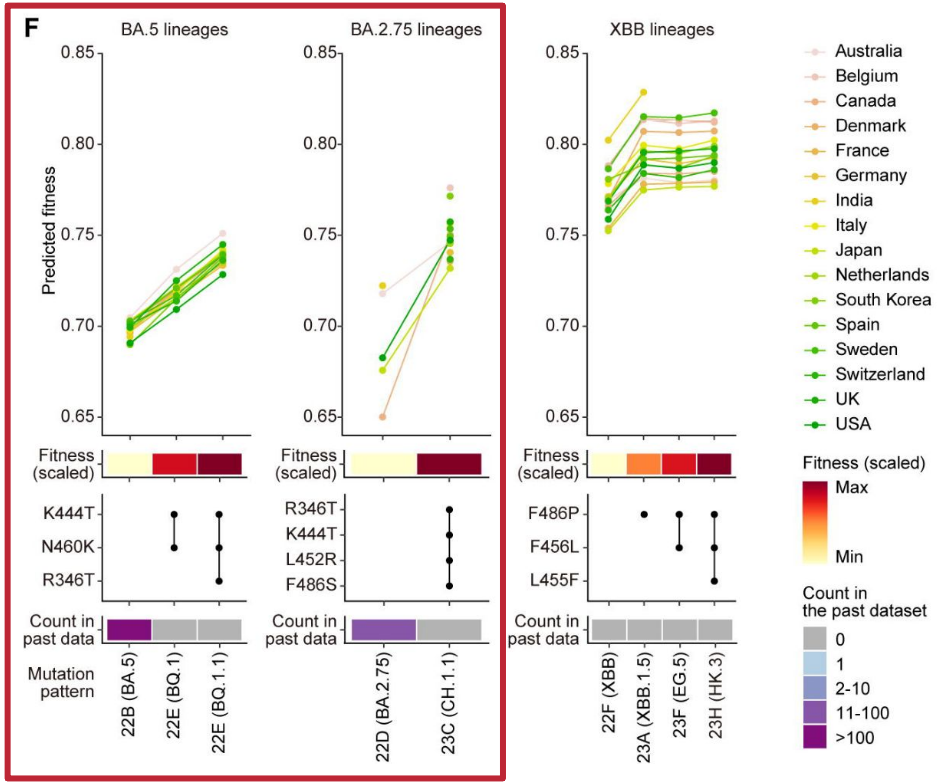
SO WHAT CAN IT DO?

- it can predict the **fact of fitness gain** (in luckier setups and not the actual amount)



SO WHAT CAN IT DO?

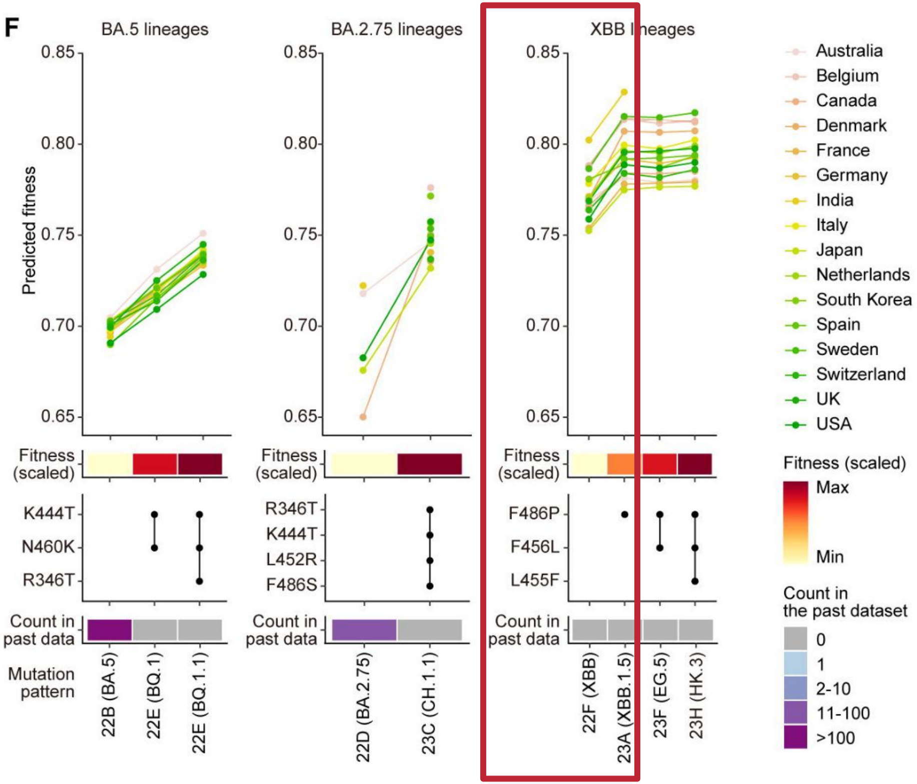
- it can predict the **fact of fitness gain** (in luckier setups and not the actual amount)



- **if the ancestral sequence is present** in the training set, it can predict that descendants with additional mutations have higher fitness

SO WHAT CAN IT DO?

- it can predict the **fact of fitness gain** (in luckier setups and not the actual amount)

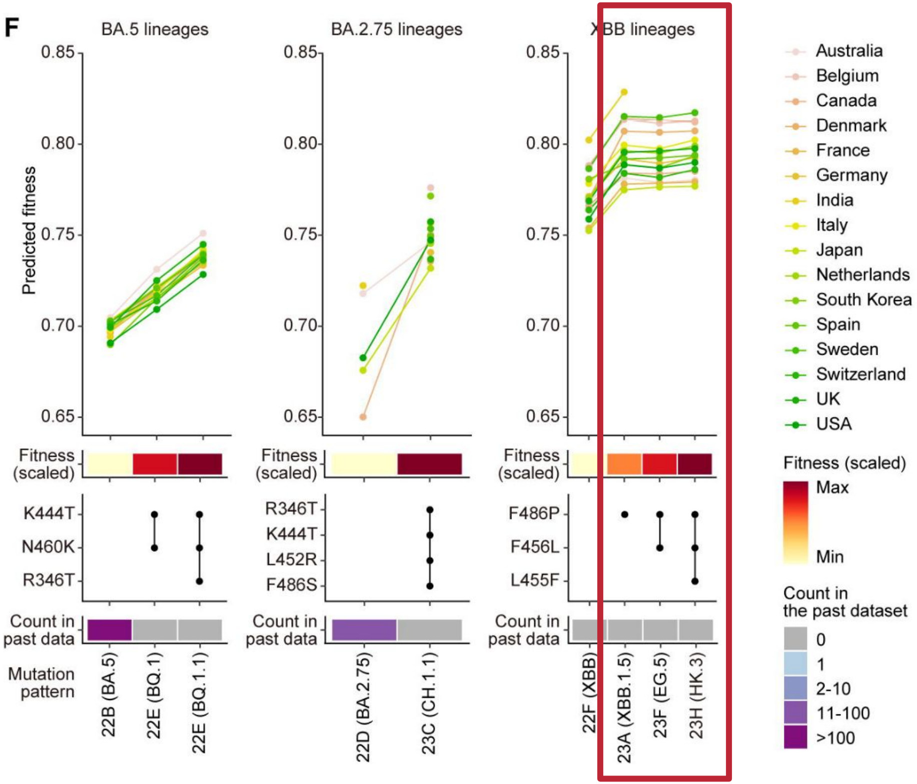


- ***if the ancestral sequence is present*** in the training set, it can predict that descendants with additional mutations have higher fitness

- even if the ancestral strain is missing from the training set, it can predict that adding a single mutation is beneficial

SO WHAT CAN IT DO?

- it can predict the **fact of fitness gain** (in luckier setups and not the actual amount)



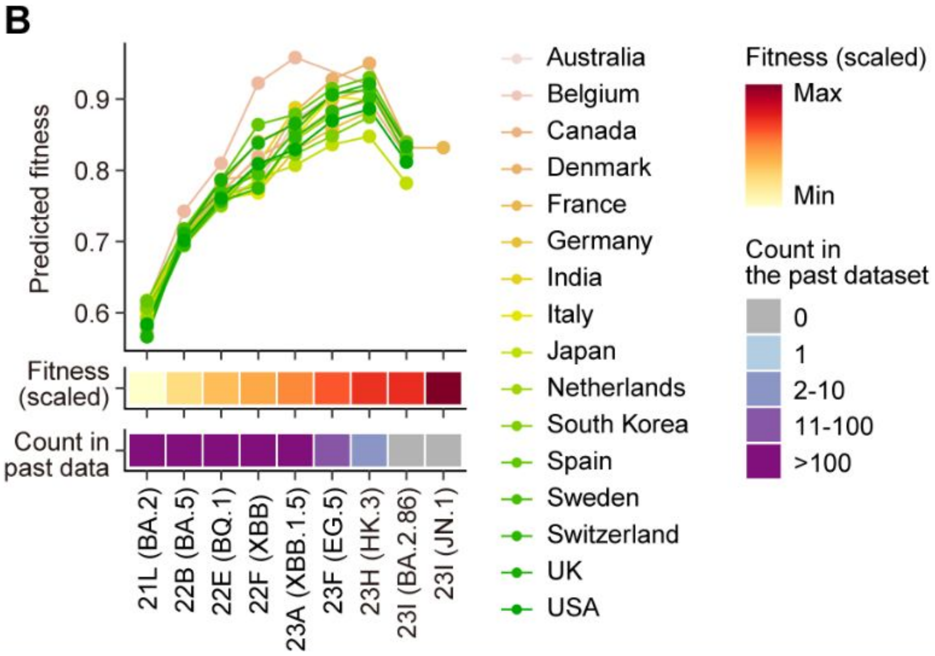
- **if the ancestral sequence is present** in the training set, it can predict that descendants with additional mutations have higher fitness

- even if the ancestral strain is missing from the training set, it can predict that adding a single mutation is beneficial

- ***it fails to predict the additional benefit of further mutations***

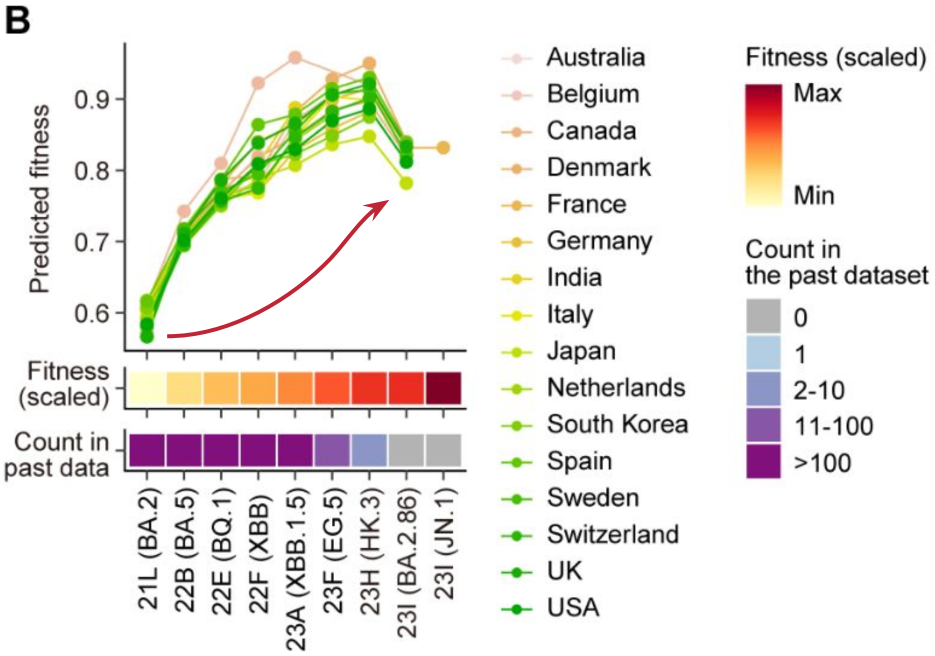
SO WHAT CAN IT DO?

- it can predict the **fact of fitness gain** (in luckier setups and not the actual amount)



SO WHAT CAN IT DO?

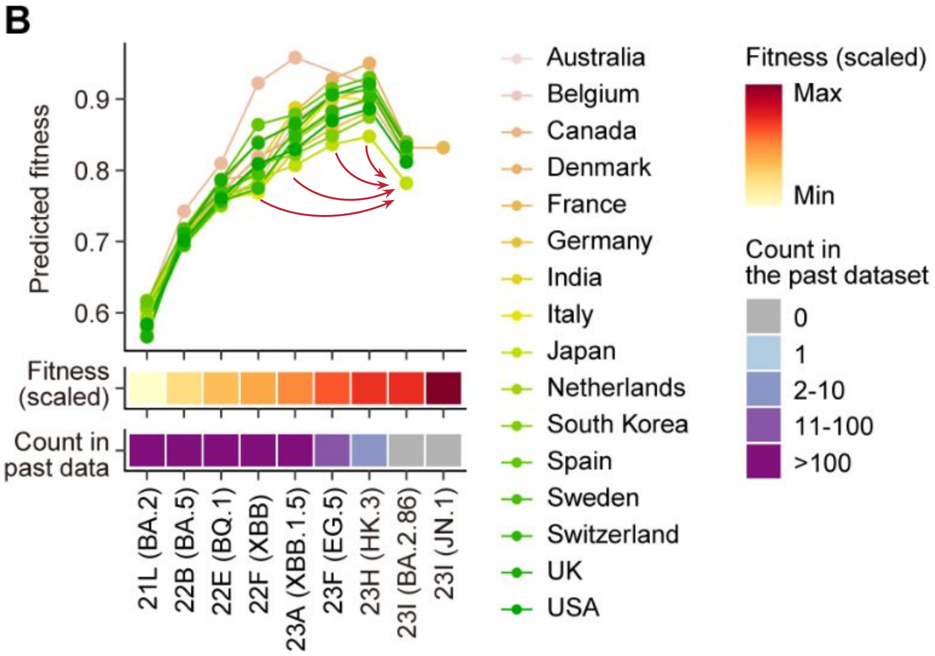
- it can predict the **fact of fitness gain** (in luckier setups and not the actual amount)



- it can predict that **acquiring a total of 30 new mutations in one step is beneficial** (BA.2 → BA.2.86)

SO WHAT CAN IT DO?

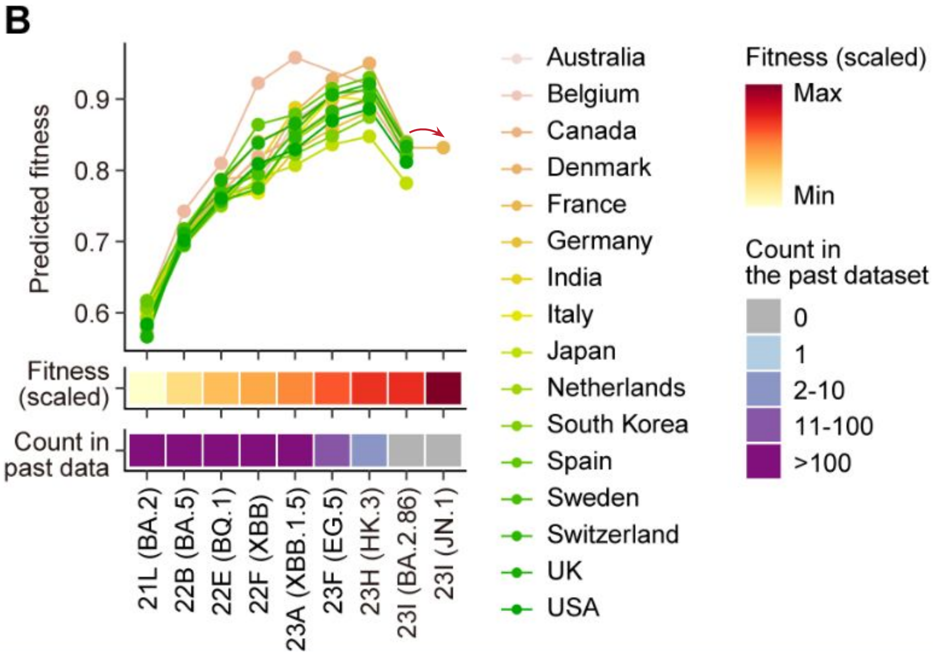
- it can predict the **fact of fitness gain** (in luckier setups and not the actual amount)



- it can predict that **acquiring a total of 30 new mutations in one step is beneficial** (BA.2 → BA.2.86)
- it fails to predict the fitness gain compared to non-ancestral sequences

SO WHAT CAN IT DO?

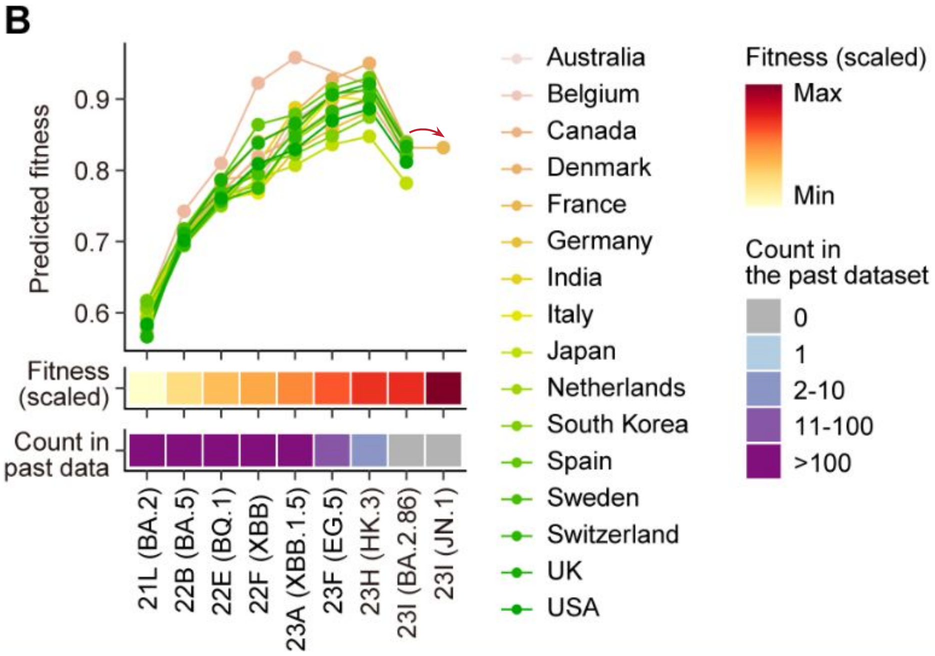
- it can predict the **fact of fitness gain** (in luckier setups and not the actual amount)



- it can predict that **acquiring a total of 30 new mutations in one step is beneficial** (BA.2 → BA.2.86)
- it fails to predict the fitness gain compared to non-ancestral sequences
- it fails to predict the additional fitness benefit of further descendants (BA.2.86 → JN.1)

SO WHAT CAN IT DO?

- it can predict the **fact of fitness gain** (in luckier setups and not the actual amount)

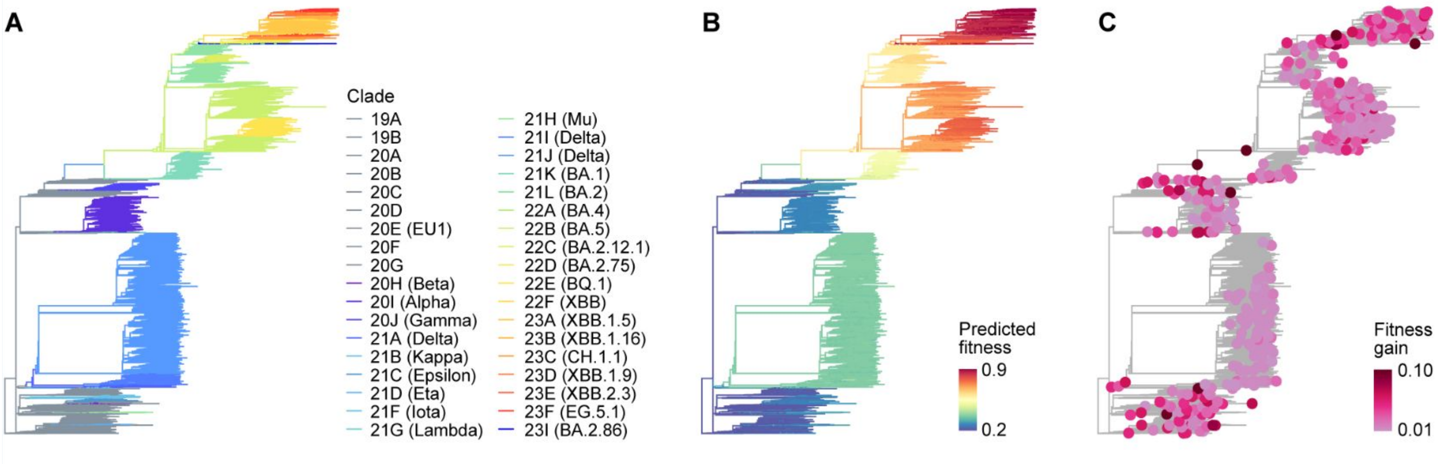


- it can predict that **acquiring a total of 30 new mutations in one step is beneficial** (BA.2 → BA.2.86)
- it fails to predict the fitness gain compared to non-ancestral sequences
- it fails to predict the additional fitness benefit of further descendants (BA.2.86 → JN.1)

These are special cases and no information is present about the potential benefit/disadvantage of other (not seen) mutations.

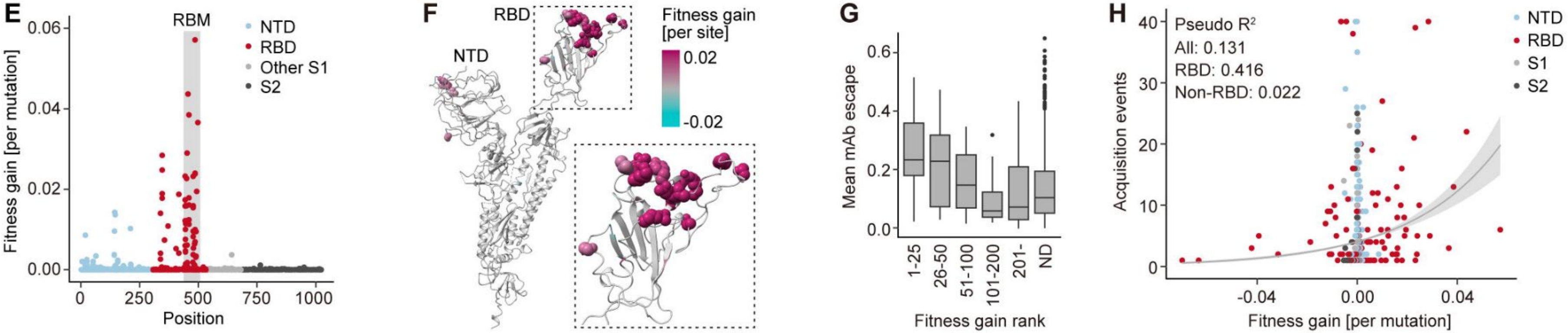
SANITY CHECKS

- they used the model **trained on the whole pandemic surveillance dataset** (no past vs. future split) → more of a sanity check than an actual "prediction" or forecast
- **predicted fitness** for all branches of the (reconstructed) phylogenetic tree
- checked for **significant fitness elevation** between branches and parent nodes
- more than half of the branches with a significant fitness elevation were **within the Omicron lineage**



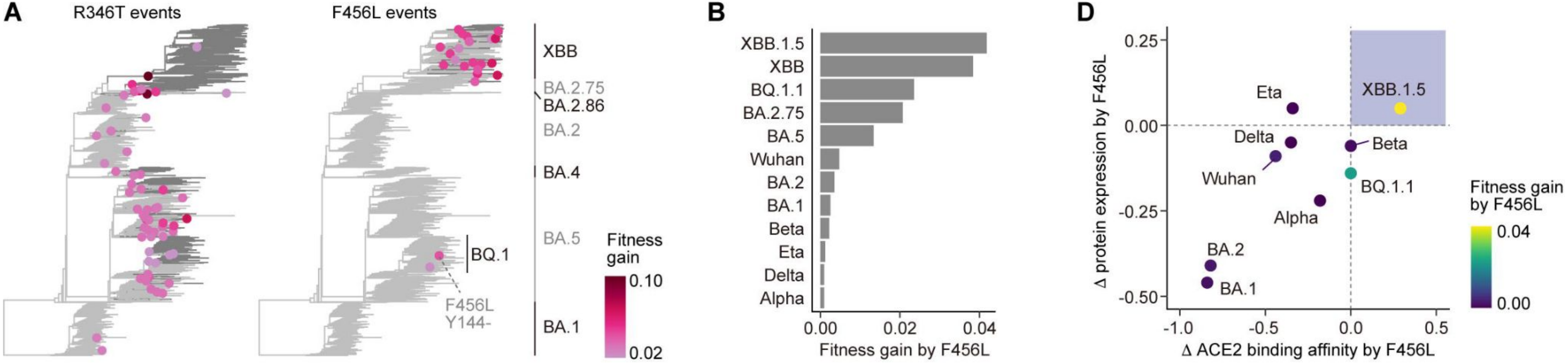
SANITY CHECKS

- **mutations with high associated fitness gain** are predominantly found in the **RBD** of the S protein, particularly in its **receptor binding motif**
- they also enhance the virus’s ability to **evade humoral immunity**
- **acquired multiple times** in a convergent manner throughout Omicron’s evolution



CAPTURING EPISTASIS

- fitness effects of a specific mutation on **various S protein backbones**



- some mutations occur independently multiple times throughout evolution (e.g. R346T)
- some mutations occur only for specific variants (e.g. F456L)
- **different backbones have different fitness gains when a single AA is changed**

COMPARISON WITH OTHER APPROACHES

	CoVFit	Bloom & Neher	EVEscape
dataset	prepandemic & surveillance data	surveillance data	**prepandemic data**
target	rel. repr. number estimated from count data (sequence-wise)	fitness based on four-fold degenerate sites (mutation-wise)	immune escape (fitness + accessibility + dissimilarity)
genomic region	S protein 1-1024 AA	whole genome	S protein
epistasis	**inherently included**	not considered	can be included when retraining on pandemic data
future predictions	widely inaccurate	—	**surprisingly good**