nature

Article

Learning from prepandemic data to forecast viral escape

https://doi.org/10.1038/s41586-023-06617-0

Received: 20 July 2022

Accepted: 6 September 2023

Published online: 11 October 2023

Nicole N. Thadani[1,6], Sarah Gurev[1,2,6], Pascal Notin[3,6], Noor Youssef[1], Nathan J. Rollins[1,5], Daniel Ritter[1], Chris Sander[1,4], Yarin Gal[3] & Debora S. Marks[1,4]✉

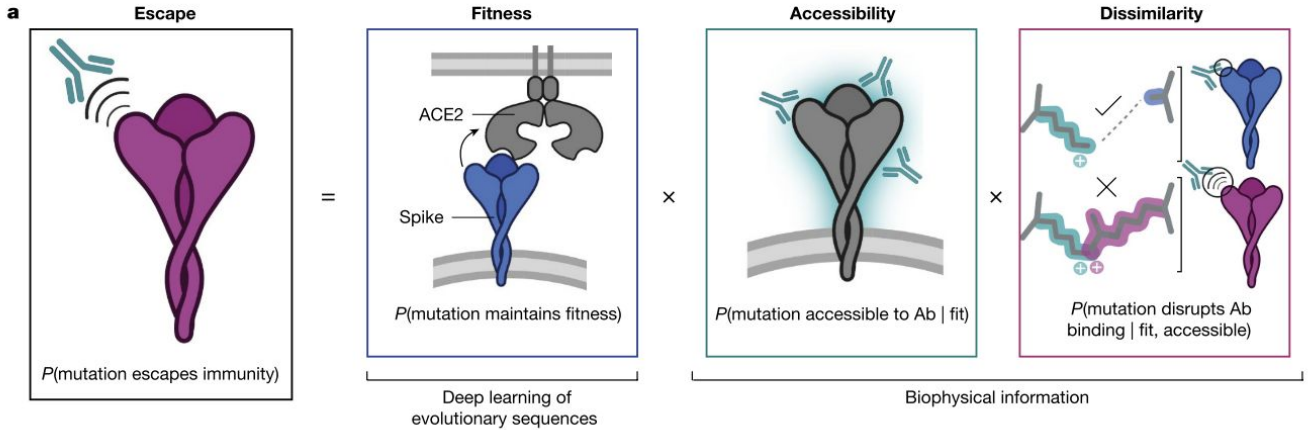
MUTATIONS EVADING THE IMMUNE RESPONSE

- main goal of **pandemic preparedness**:
  - **which mutations** can evade host immune response
  - **how likely** it is that they will occur
- general approach
  - experiments that require **host polyclonal antibodies**
  - computational methods that require **enormous sequencing data** to reliable estimate strain prevalences
- essentially, **a bunch of people need to get infected** before we are able to tell anything useful about the virus

→ **EVEscape**

- uses **historical sequences** with **biophysical** and **structural** information
- **applicable before surveillance sequencing, DMS and 3D structure information of antibody complexes**

EVESCAPE: GENERAL CONCEPT



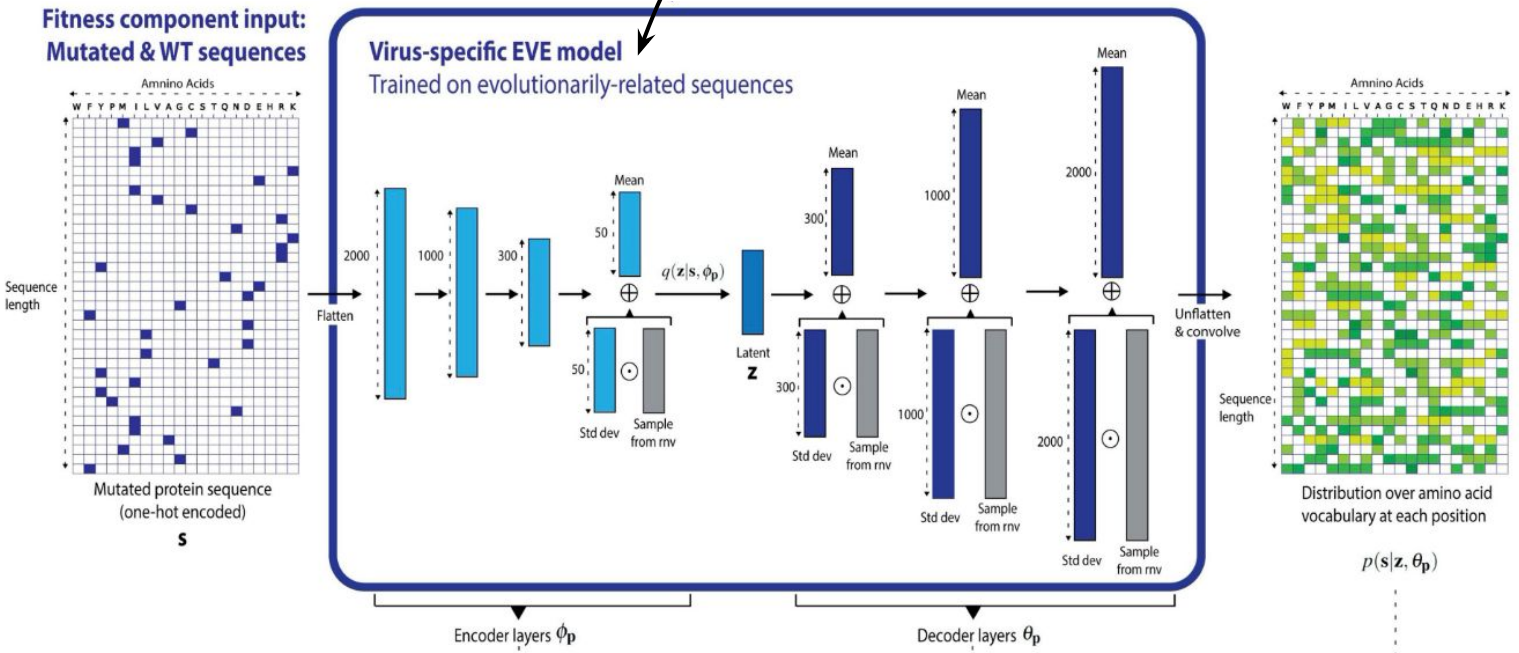
probability that a mutation can escape immunity:

- probability that it maintains viral fitness
- probability that it is located in an accessible region by antibodies (given it maintains fitness)
- probability that it disrupts antibody binding (given it maintains fitness and is accessible)

} product

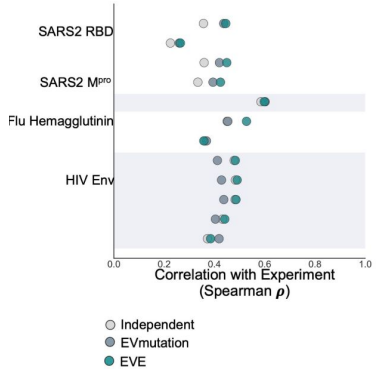
MODEL: FITNESS

a Bayesian variational autoencoder



training data: prepandemic coronavirus sequences

fitness estimates correlate with DMS results



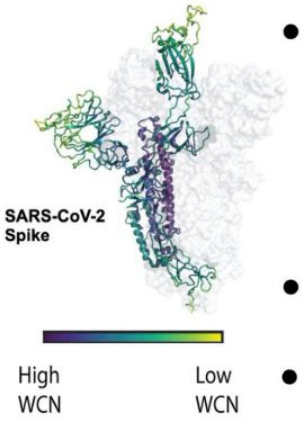
$$ELBO(\mathbf{s}) = N.\mathbb{E}_{p(\mathbf{s})}\left[\mathbb{E}_{q(\theta_\mathbf{p}),q(\mathbf{z}|\mathbf{s})}\left(\log p(\mathbf{s}|\mathbf{z},\theta_\mathbf{p})\right) - D_{KL}(q(\mathbf{z}|\mathbf{s},\phi_\mathbf{p})||p(\mathbf{z}))\right] - D_{KL}(q(\theta_\mathbf{p})||p(\theta_\mathbf{p}))$$

~ lower bound for the log-likelihood

$$P(\mathbf{s}\text{ is fit}) = f(\log\frac{p(\mathbf{s}|\theta_\mathbf{p})}{p(\mathbf{w}|\theta_\mathbf{p})}) \sim f(ELBO(\mathbf{s}) - ELBO(\mathbf{w}))$$

MODEL: ACCESSIBILITY & DISSIMILARITY

Accessibility component input:
Protein structure(s)



- Weighted contact number for residue i in conformer c:

$$WCN_i^{(c)} = \sum_{j \neq i} \frac{1}{\left(r_{ij}^{(c)}\right)^2}$$

where r_{ij} is the distance between the geometric centers of the residue i and residue j side chains

- The negative $WCN_i^{(c)}$ captures surface accessibility and protrusion from the core structure
- If there are multiple conformers, we take the maximum of negative $WCN_i^{(c)}$ values across conformers
- P(Mutation accessible to Ab | fit) $= f\left(\max_c\{-WCN_i^{(c)}\}\right)$

Dissimilarity component input:
Mutated & WT amino acid(s)

- Change in **hydrophobicity** at residue i (Eisenberg-weiss scale) $\Delta_h^{(i)}$
- Change in **charge** at residue i (at physiological pH) $\Delta_c^{(i)}$
- P(Mutation disrupts Ab binding | fit, accessible)

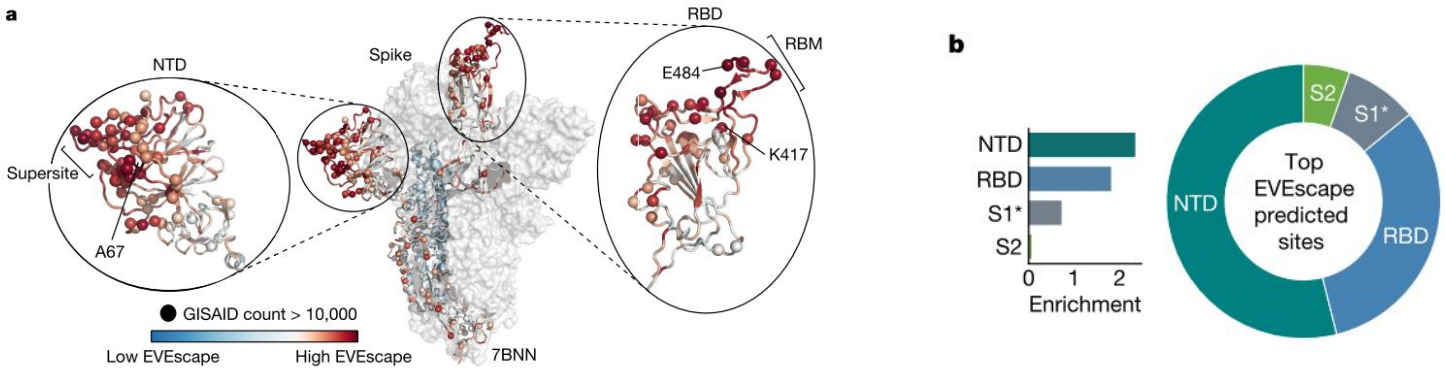
$$= f\left(\sigma(\Delta_h^{(i)}) + \sigma(\Delta_c^{(i)})\right)$$

where $\sigma(.)$ applies standard scaling

weighted contact number ~ large for central residues, low for surface ones → use **negative WCN**

dissimilarity ~ how much physical properties (**charge, hydrophobicity**) change

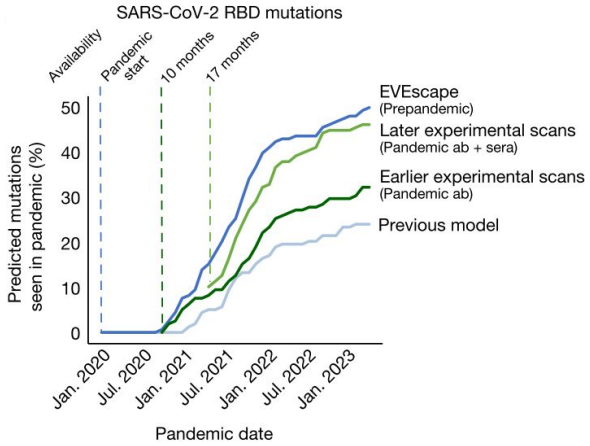
RESULTS: ANTIGENIC REGIONS

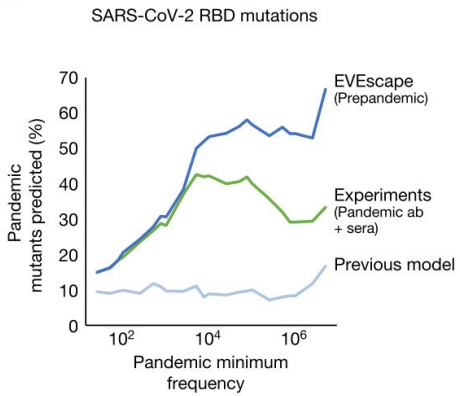


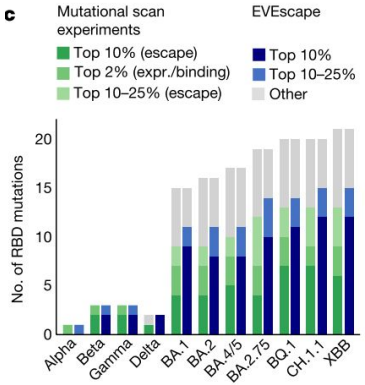
EVEscape **identifies antigenic regions** without prior knowledge on antibody binding regions or epitopes

→ useful in early vaccine development

RESULTS: PREDICTED MUTATIONS





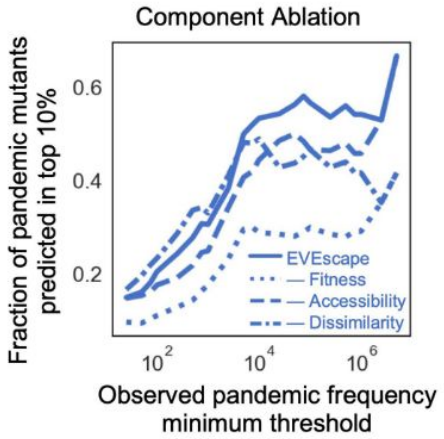


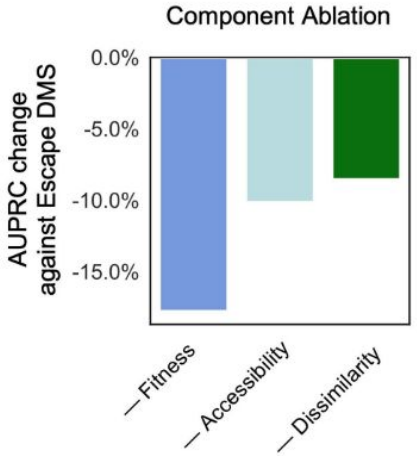
- **EVEscape-predicted mutations emerge** during the course of the pandemic
- **most high-frequency mutations were predicted by EVEscape**
- VOC defining mutations have high EVEscape scores

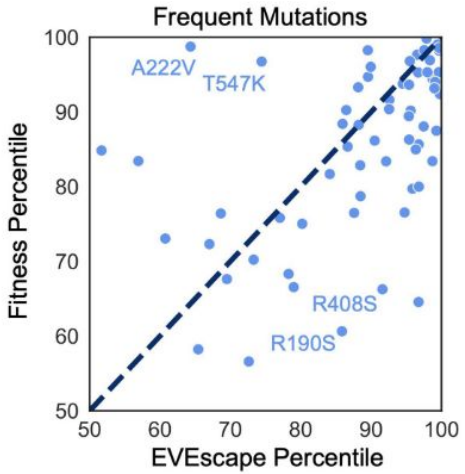
→ escape sites can be predicted before the pandemic and can forewarn waning therapy effectiveness

RESULTS: PREDICTED MUTATIONS

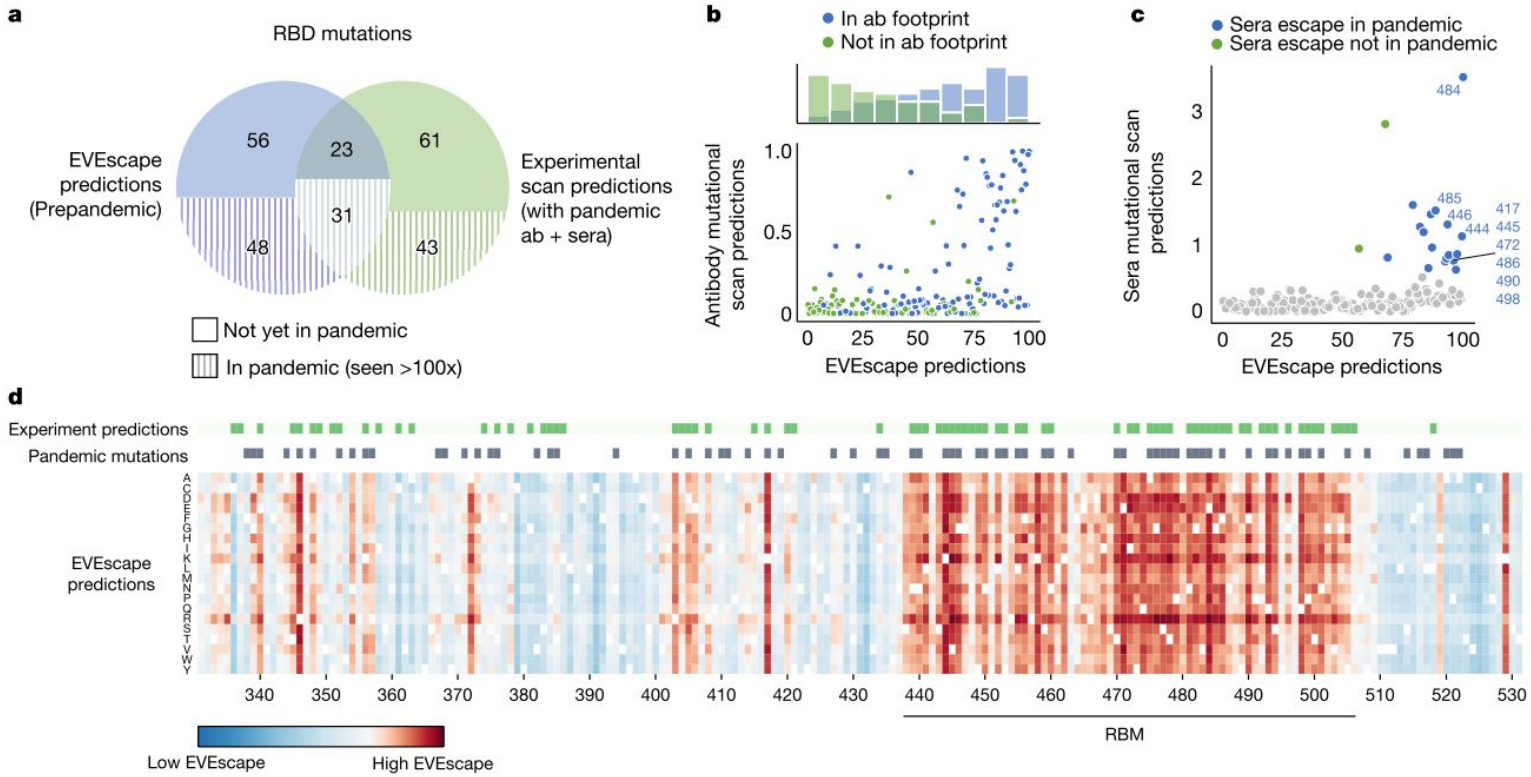
- EVE alone (**fitness alone) is better at predicting mutations with low frequency** that are assumably irrelevant in terms of immune escape but maintain fitness
- **all parts of the model play an important role in predictions**





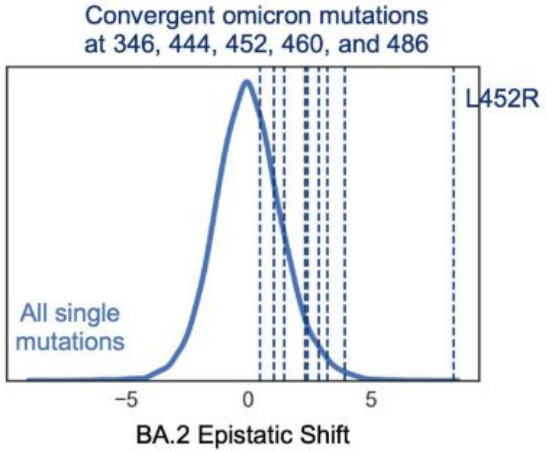


RESULTS: EVESCAPE VS. DMS



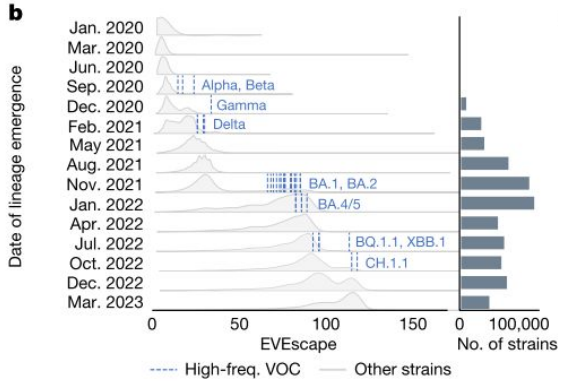
MODULAR DESIGN

- model components can be switched to accommodate specific tasks
  - TranceptEVE → indels
  - including glycosylation in the dissimilarity component for HIV
  - retraining fitness component on pandemic data
    → epistatic shifts

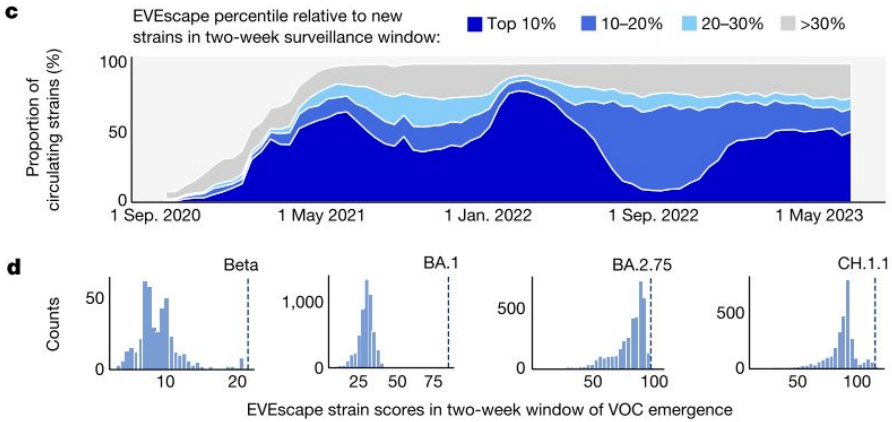


- comparing EVE scores on a Wuhan full Spike model and on an omicron (BA.2) full Spike model
- **BA.2 epistatic shift**: the Wuhan linear regression residual for a model fit to the two sets of EVE scores for all single mutations to full Spike

STRAIN FORECASTING



distribution of non-VOC EVEscape strain scores (aggregated EVEscape scores for unique combinations of mutations)

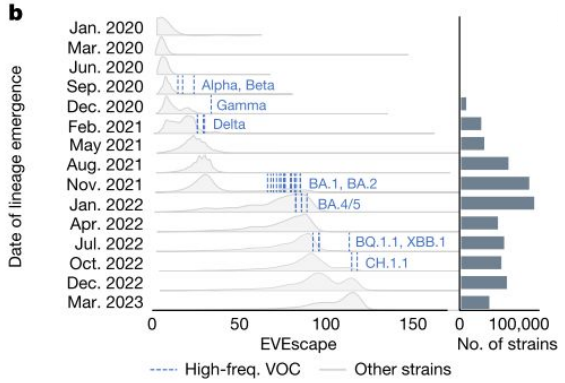


more than 40% of circulating strains on average fall into the top decile of EVEscape strain scores

VOCs are among the highest scoring strains in the two-week windows of their emergence

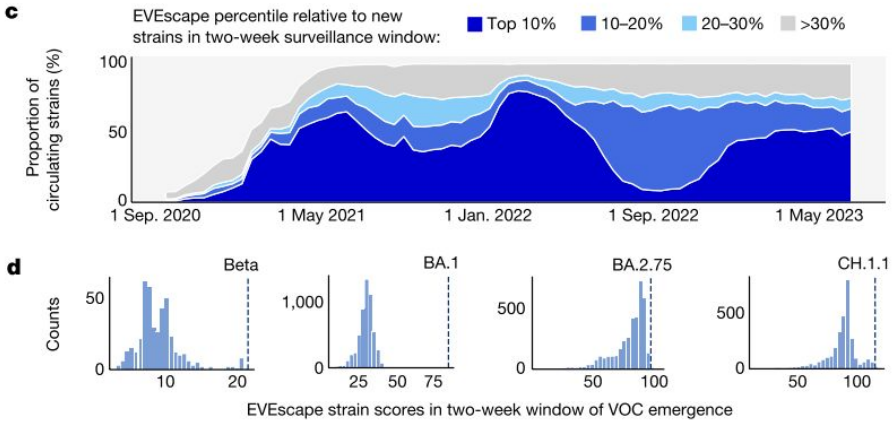
→ **VOCs can be forecasted after a single observation**!

STRAIN FORECASTING



distribution of non-VOC EVEscape strain scores (aggregated EVEscape scores for unique combinations of mutations)

These are all based on prepandemic data!

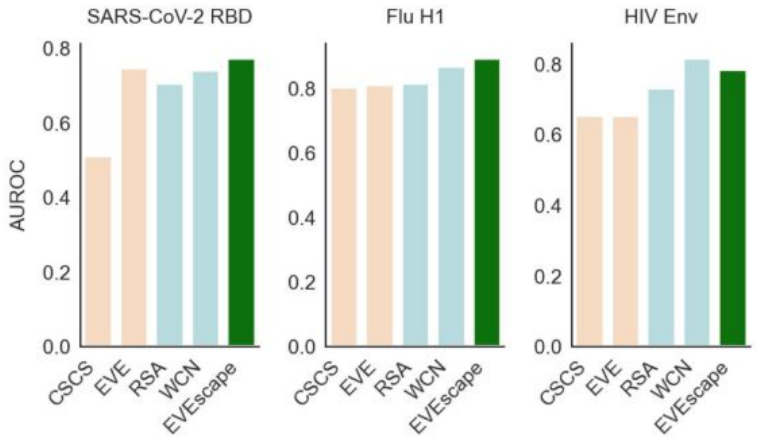


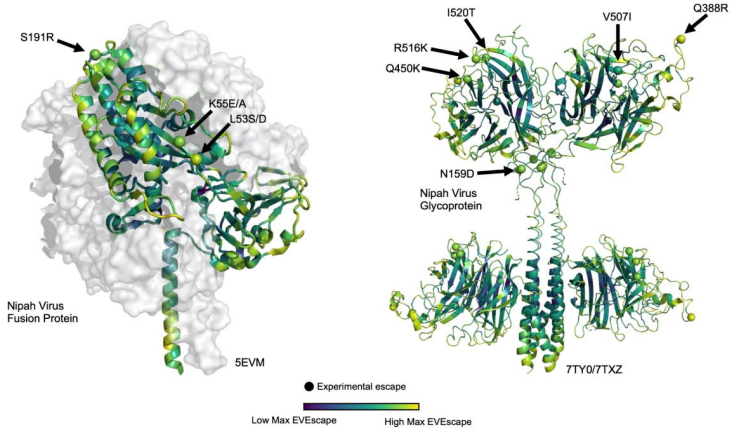
more than 40% of circulating strains on average fall into the top decile of EVEscape strain scores

VOCs are among the highest scoring strains in the two-week windows of their emergence

→ **VOCs can be forecasted after a single observation**!

GENERALIZABILITY: OTHER VIRUSES





At the beginning:

especially **useful for non-pandemic viruses** (Nipah, Lassa)

During the pandemy:

active surveillance of emerging strains

evescape.org