Title: Computation of Antigenicity Predicts SARS-CoV-2 Vaccine Breakthrough Variants

Authors:

Ye-fan Hu[1,3,†], Jing-chu Hu[2,†], Hua-rui Gong[1,†], Antoine Danchin[1,5], Ren Sun[1], Hin Chu[4], Ivan Fan-Ngai Hung[3], Kwok Yung Yuen[4], Kelvin Kai-Wang To[4,], Bao-zhong Zhang[2,*], Thomas Yau[3,*], Jian-Dong Huang[1,2,*]

Affiliations:

[1] School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, University of Hong Kong, 3/F, Laboratory Block, 21 Sassoon Road, Hong Kong, China

[2] CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

[3] Department of Medicine, Li Ka Shing Faculty of Medicine, University of Hong Kong, 4/F Professional Block, Queen Mary Hospital, 102 Pokfulam Road, Hong Kong, China

[4] Department of Microbiology, Li Ka Shing Faculty of Medicine, University of Hong Kong, 19/F T Block, Queen Mary Hospital, 102 Pokfulam Road, Hong Kong, China

[5] Kodikos Labs / Stellate Therapeutics, Institut Cochin, 24 rue du Faubourg Saint-Jacques, 75014 Paris, France

[†] These authors contributed equally to this work.

* Corresponding authors. B.Z.Z. (bz.zhang3@siat.ac.cn), T.Y. (tyaucc@hku.hk), J.D.H. (jdhuang@hku.hk)

Abstract

It has been reported that multiple SARS-CoV-2 variants of concerns (VOCs) including B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma), and B.1.617.2 (Delta) can reduce neutralisation by antibodies, resulting in vaccine breakthrough infections. Virus-antiserum neutralisation assays are typically performed to monitor potential vaccine breakthrough strains. However, such experimental-based methods are slow and cannot instantly validate whether newly emerging variants can break through current vaccines or therapeutic antibodies. To address this, we sought to establish a computational model to predict the antigenicity of SARS-CoV-2 variants by sequence alone and in real time. In this study, we firstly identified the relationship between the antigenic difference transformed from the amino acid sequence and the antigenic distance from the neutralisation titres. Based on this correlation, we obtained a computational model for the receptor binding domain (RBD) of the spike protein to predict the fold decrease in virus-antiserum neutralisation titres with high accuracy (~0.79). Our predicted results were comparable with experimental neutralisation titres of variants, including B.1.1.7 (Alpha), B.1.351 (Beta), B.1.617.2 (Delta), B.1.429 (Epsilon), P.1 (Gamma), B.1.526 (Iota), B.1.617.1 (Kappa), and C.37 (Lambda), as well as SARS-CoV. Here, we firstly predicted the fold of decrease of B.1.1.529 (Omicron) as 17.4-fold less susceptible to neutralisation. We visualised all 1521 SARS-CoV-2 lineages to indicate variants including B.1.621 (Mu), B.1.630, B.1.633, B.1.649, and C.1.2, which can induce vaccine breakthrough infections in addition to reported VOCs B.1.351 (Beta), P.1 (Gamma), B.1.617.2 (Delta), and B.1.1.529 (Omicron). Our study offers a quick approach to predict the antigenicity of SARS-CoV-2 variants as soon as they emerge. Furthermore, this approach can facilitate future vaccine updates to cover all major variants. An online version can be accessed at http://jdlab.online .

Up to January 2022, there have been several SARS-CoV-2 variants including B.1.1.7 (Alpha) [1-5], B.1.351 (Beta) [2,3,6,7], P.1 (Gamma) [1,2,8], and B.1.617.2 (Delta) [9,10] that are experimentally tested to lead vaccine breakthrough infections, thus they have been designated as variants of concerns (VOCs) by the world health organization (WHO). There is a concern that other untested emerging variants may lead to vaccine breakthrough infections [11-16]. The most recent case is the validation of B.1.1.529 (Omicron). The current virological and epidemiological techniques took several weeks to validate whether the variant is capable of reducing the efficacy of current vaccines [17,18] or therapeutic antibodies [18,19], even though their viral sequences have been shared in real time via the Global Initiative for Sharing All Influenza Data (GISAID) [20]. The speed of validation of vaccine breakthrough variants can hardly catch up with the fast-emerging rate of new variants. Thus, it is crucial to develop new approaches for identifying the next potential vaccine breakthrough variant as soon as it is reported.

Here, we established a computational approach for predicting the antigenicity of SARS-CoV-2 variants from viral sequences alone, with the aim to accelerate the identification of potential vaccine breakthrough variants. Our approach is founded on the concept of antigenic mapping, also named antigenic cartography. This method has been used to monitor vaccine breakthrough variants of influenza virus using haemagglutination inhibition (HI) assay data [21,22], dengue virus [23] and SARS-CoV-2 circulating strains [24] using pairwise antisera data. In antigenic mapping, the antigenic distance is calculated from the fold change of the neutralisation titre between the reference virus and its variant, to measure the change of antigenicity between two variants. A computational approach for predicting antigenic distances to indicate vaccine breakthrough variants could theoretically provide much more rapid results once the variant sequence is reported. Past studies proposed a linear relationship between amino acid changes in antigenic sites and neutralisation fold decrease [25-29]. Computational prediction approaches based on such a relationship could also provide reliable estimates of neutralisation titres for existing antiserum against the vaccine breakthrough variants with similar accuracy to experiment-based approaches used in previous studies [25-29]. However, these predictions were optimised for influenza virus instead of SARS-CoV-2. For example, the neutralisation titre decrease of any SARS-CoV-2 variant should be less than that of SARS-CoV comparing to the ancestral

strain of SARS-CoV-2, because the cross protection between the SARS-CoV-2 variant and the ancestral strain is stronger than that between SARS-CoV and SARS-CoV-2. Thus, it is difficult to use a linear relationship to predict the decrease in neutralisation titre which saturates with the increase in the mutation numbers of variants. A SARS-CoV-2 optimised model for predicting antigenicity is urgently needed.

In this study, we established a computational sequence-based method to predict the antigenicity of SARS-CoV-2 variants to reveal potential vaccine breakthrough variants. This method can also predict the neutralisation titre of VOCs in comparison to the ancestral strain of SARS-CoV-2. Our predicted results were comparable with experimental neutralisation titres of VOCs, including B.1.1.7 (Alpha), B.1.351 (Beta), B.1.617.2 (Delta), B.1.429 (Epsilon), P.1 (Gamma), B.1.526 (Iota), B.1.617.1 (Kappa), and C.37 (Lambda), as well as SARS-CoV. Here, we predicted that B.1.1.529 (Omicron) is 17.4-fold less susceptible to neutralisation, which is consistent with reported decrease folds ranging from 10 to 40 [17,18].

A computational model for predicting antigenicity of SARS-CoV-2 variants

To predict the antigenicity of SARS-CoV-2 variants, we firstly integrated the reported conformational or linear epitopes **(Fig. S1 &Table S1)** on the SARS-CoV-2 Spike protein **(Fig. 1a)** with the reported experimental virus-antiserum neutralisation titres against SARS-CoV-2 variants including B.1.1.7 [1-5], B.1.351 [2,3,6,7], and P.1 [1,2,8] **(Table S2a)**. Considering the distinct assays used in the different studies, we standardised the neutralisation titres of each variant to the titre of the ancestral strain of SARS-CoV-2 (lineage A) using the same assay in each study on a log 2 scale, and thus we got observed antigenic distance (H_{ab}) from neutralisation titres **(Fig. 1b).** For the antigenic difference (D_{ab}), we used Poisson distance to represent the difference between two amino acid sequences **(Fig. 1b)**. By comparing the observed antigenic distance with the antigenic difference, we found a relationship between observed antigenic distance and the antigenic difference: $H_{ab}=T_{max}\cdot D_{ab}/(D_{50}+D_{ab})$, where T_{max} is the maximal fold of decrease and D_{50} is the antigenic difference which may lead to neutralisation decrease at the 50% level of the maximal decrease (the fold change between SARS-CoV-2 and SARS-CoV). This relationship described that the decrease of neutralisation titre increases with the accumulation of amino acid changes, and then

reaches at the maximal decrease **(Figs. 1c-d).** Based on this correlation, we obtained a computational model using the receptor binding domain (RBD) of the spike protein to predict the fold decrease in virus-antiserum neutralisation titres with higher accuracy **(~0.79, the calculation of accuracy in Methods)** compared with other fragments of spike (entire spike, N terminal domain plus RBD, or S1, **Fig. 1d**). With repeated 5-fold or 10-fold cross validation **(Fig. 1d)**, we found that prediction using RBD is relatively robust in terms of root-mean-square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2) and accuracy.

To further validate our model, we predicted the fold decreases in neutralisation titres (comparing to the ancestral of SARS-CoV-2) of multiple variants including B.1.1.7 (Alpha), B.1.351 (Beta), B.1.617.2 (Delta), B.1.429 (Epsilon), P.1 (Gamma), B.1.526 (Iota), B.1.617.1 (Kappa), and C.37 (Lambda), as well as SARS-CoV and WIV1-CoV using datasets without the variant that we aimed to validate. Previous studies have reported that VOCs can elicit vaccine breakthrough infections, which correlated with fold decreases in the neutralisation titres from experimental assays was disclosed **(Table S2)**. Our predicted results were highly consistent with the neutralisation assay results **(Fig. 1e)**. We also predicted the fold of decrease in neutralisation titre of the most recent VOC, B.1.1.529 (Omicron). Considering 15 mutations in the spike of B.1.1.529 (Omicron), the variant is estimated to have a 17.44-fold (95% confidence interval: 13.7, 22.2) decrease in neutralisation titre (shown as a blue point in **Fig. 1c**). , The predicted result is consistent with reported decrease folds ranging from 10 to 40 [17,18]. This result alarmed the risk of vaccine breakthrough or re-infection of B.1.1.529 (Omicron) due to the dramatic decrease in neutralization.

The prediction of potential vaccine breakthrough strains

To predict the next potential SARS-CoV-2 vaccine breakthrough variants, we visualised the antigenicity of all available SARS-CoV-2 variants as an indicator of their vaccine breakthrough potential. We firstly selected all 1521 lineage variants using PANGO [30] updated on December 6, 2021 (**Table S3**) to predict their antigenicity. Then we calculated the pairwise distances of different variants. For visualising these results, we captured two principal components from the high-dimensional data of antigenic distance [25]. We used all spike amino acid sequences to plot the ‘genetic map’ of SARS-CoV-2 to represent the genetic difference among

different variants **(Fig. 2a-b)**. We then plotted the 'antigenic map' using the predicted antigenic distances **(Fig. 2c-d,** online versions available at http://jdlab.online).

Based on the relationship between neutralisation titre fold change and protective efficacy [31], it was convenient to set up some 'cut-offs' in the current vaccine coverage. We included phase 3 and real-world results of vaccine efficacy or effectiveness, as well as neutralisation titre data from phase 1 and 2 studies **(Table S4-5)**. Thus, we got the relationship between neutralisation titre and protective efficacy against a symptomatic COVID-19 **(Fig. 2e)**. A 3.93-fold decrease in neutralisation titres induced by VOCs that can dampened the efficacy of some vaccines to lower than 50%. In this way, one cut-off of 1.98 arbitrary units (A.U.) represented a 3.93-fold decrease in the neutralisation titre (shown as a pink circle in **Fig. 2c-d**). All variants outside this cut-off have the potential to be vaccine breakthrough variants. By comparing the "genetic map" and antigenic map, we can set up the border of antigenic map. Although there are >200 mutations in the SARS-CoV and WIV1-CoV spike (**Fig 2a**), the antigenic distance is around 4.9 A.U. which mean ~ 30-fold decrease in the neutralisation titre (shown as a dark red circle in **Fig. 2c-d**).

To reveal the distribution of variant, we plotted the density of variants on the 'genetic map' and antigenic map due to overlapping dots. In the genetic map, hotspots are located at lineage A (>10%) and B.1 (>40%) mainly, as well as AY.* and P.1 (**Fig. 2b**). While in the antigenic map, hotspots are placed at lineage A (>40%) mainly, together with AY.* (**Fig. 2d**). Although most variants were shown to be close to the ancestral strain (**Figs. 2b&d**), multiple variants were found to decrease neutralisation titres significantly (**Fig. 2c**). In addition to reported VOCs including B.1.351 (Beta, containing sub-lineages like B.1.351.2 and B.1.351.5) [2,3,6,7], P.1 (Gamma, containing sub-lineages like P.1.11 and P.1.3) [1,2,8], B.1.617.2 (Delta, containing sub-lineages AY.*) [9] , and B.1.621 (Mu, containing sub-lineage B.1.621.1), B.1.1.529 (Omicron) showed over 3.93-fold decrease in the neutralisation titre. Other variants B.1.630, B.1.633, B.1.649, and C.1.2 also have the potential to be vaccine breakthrough variants with more than 3.93-fold decrease **(Fig. 2c)**. Besides the pandemic of B.1.617.2 (Delta) [9] and the outbreak of B.1.1.529 (Omicron), multiple variants should be investigated immediately as they have the potential to become tomorrow's VOCs.

Discussion

Predicting neutralisation responses against all SARS-CoV-2 variants based on sequences alone is vital for selecting the next vaccine seeds for the development of effective COVID-19 vaccines. We established a computational approach to predict neutralisation titres and validated these predictions using experimental data. Our computational approach could potentially provide the first hints of whether a newly identified variant can break through vaccines just by its sequence information, which would greatly shorten the time for the crucial early warning of emerging vaccine breakthrough strains.

In the prediction of the antigenicity of SARS-CoV-2 variants, we proposed that the limit of neutralisation titre decrease is set by SARS-CoV **(Fig. 1)**. In recent studies, SARS-CoV is ~ 36-fold less susceptible to neutralisation comparing to the ancestral strain of SARS-CoV-2. Based on this result, a non-linear curve was established to describe the relationship between the observed antigenic distance and the antigenic difference. We further performed calculation using different fragments of the Spike protein **(Fig. 1d)**. Among the Spike protein and the RBD, NTD-RBD, and S1 fragments, we found the prediction using amino acid sequences of RBD was able to estimate the neutralisation titre more accurately than the others (**Figs. 1d**). Thus, we used the RBD-based computations to determine the neutralisation titres.

A major concern of our computation of the neutralisation titre is that the data is based on diverse neutralisation assays of serum samples from both patients and vaccinees against both live virus and pseudovirus **(Table S2).** Although the results were consistent qualitatively, the variation of fold change is too large to be ignored **(Fig. 1e)**. Considering the variation in the real world, we set up values 2-fold or less than the experimental values as the criteria based on previous studies [28]. It is better to establish a convenient and standardised neutralisation pipeline in the future, like the haemagglutination inhibition (HI) assay for influenza virus. Such a pipeline can allow the precise estimation of neutralisation titres. Together with estimating the association of neutralisation with protection, it will help to develop next generation vaccines.

It is crucial to update vaccines to cover all vaccine breakthrough strains that have significant amino acid and glycosylation changes to prevent further infectious outbreaks. However, not all predicted SARS-CoV-2 vaccine breakthrough variants will have the chance to cause an outbreak due to their changed viral fitness [32] or by

pure luck. Based on previous studies of influenza viruses, it is possible for variants to have alterations that change the antigenicity, but fail to cause outbreaks in the wider population [33]. Considering immune escape elicited by variants, updating current vaccine seeds with new variants should extend the vaccine coverage. As SARS-CoV-2 showed different variant directions in the antigenic map **(Fig. 2)**, the use of multiple virus seeds based on the different directions might be appropriate to cover all major variants in the long term. Our method could help in the selection of SARS-CoV-2 variants for updating vaccines.

Methods

Antigenic footprint

We collected 149 confirmed conformational epitopes with protein structures released in the Protein Data Bank (PDB) (https://www.rcsb.org/) or annotated epitope footprints and 76 linear epitopes published in the literature **(Table S1)**. We plotted the footprint of all Spike protein epitopes from the aforementioned 225 epitopes using R-3.6.6.

Antigenic distances from neutralisation data

We calculated antigenic distances from the neutralisation data based on previous publications[26]. For virus variant *a*, reference virus *b*, and antiserum β (referencing virus *b*), we defined the antigenic distance of variant *a* to reference virus *b* in terms of the standardised log titre as $H_{ab}=\log_2 T_{a\beta} - \log_2 T_{b\beta}$, where $T_{b\beta}$ is the titre of antiserum β against virus *b*, and $T_{a\beta}$ is the titre of antiserum β against virus *a* [26]. Merged data with reference virus lineage A (the ancestral strain of SARS-CoV-2) were collected from several publications **(Table S2)**.

Genetic and antigenic difference calculation

We selected 1521 SARS-CoV-2 lineages using PANGO (v.3.1.15) updated on December 6, 2021 (https://cov-lineages.org/). Spike protein amino acid sequences of these lineages were obtained from GISAID, using the earliest collected for each lineage **(Table S3)**. All sequences with neutralisation titres were also included **(Table S3)**. For genetic distances, we used Molecular Evolutionary Genetics Analysis (MEGA) X to calculate the pairwise distances among Spike protein amino acid sequences in the SARS-CoV-2 variants using a Poisson model. For antigenic distance, we used an information theory-based approach *p-all-epitope* [27,28] to measure the pairwise distances among amino acid sequences of the antigenic footprint ('antigenic positions'). The distance is based on the number of different amino acids n_d between two *n*-mer viral sequences of variants *a* and *b*. Under the assumption that the number of amino acid substitutions per site follows a Poisson distribution, we can then calculate the distance between *a* and *b* as $D_{ab}=-\ln(1-n_d/n)$.

Modelling and performance measurement

A model considering the maximal neutralisation tire decrease was applied to examine the antigenic distance from the neutralisation data H_{ab} and our computed results D_{ab} as $H_{ab}=T_{max}\cdot D_{ab}/(D_{50}+D_{ab})$, where T_{max} is the maximal decrease and D_{50} is the antigenic difference which may lead to neutralisation decrease at the 50% level of the maximal decrease. The predicted neutralisation titre is then given as $P_{ab}\approx\hat{H}_{ab}=T_{max}\cdot D_{ab}/(D_{50}+D_{ab})$. Root-mean-square error (RMSE), mean absolute error (MAE), and coefficient of determination (R-squared R^2) were used to measure the performance of the linear correlation.

Reproducibility was determined by pairwise sequences and neutralisation titres. Neutralisation titre data were converted into variables by calculating the relative difference in the neutralisation titres between reference virus and variant against the antiserum. Accuracy was the percentage of correctly predicted neutralisation titres using amino acid sequences. Based on previous studies [28], computational values 2-fold or less than the experimental values were considered to be similar (correct) and those more than 2-fold lower were considered dissimilar (error). Here, 10-time repeated 5-fold and 10-fold cross validation were applied in terms of root-mean-square error (RMSE), mean absolute error (MAE), coefficient of determination (R-squared R^2), and accuracy.

Genetic and antigenic maps

After calculating genetic and antigenic distances, we used classical multidimensional scaling (CMDS) to display the data as a plot using R-3.6.6. We set up SARS-CoV-2 lineage A as the origin and scaled the data in two and three dimensions. We then acquired the genetic and antigenic maps of SARS-CoV-2 lineages. An online version can be obtained at http://jdlab.online .

Logistic model

Following past studies[31], we used a logistic model in R-3.6.6 to describe the relationship between antigenic distance (neutralization level) and protective efficacy/effectiveness: $E=1/(1+\exp(-k(H-H_{50})))$. E is the protective efficacy/effectiveness at a specific neutralization level H. H is the mean of neutralisation titres in vaccinees divided by corresponding mean of titres in convalescent patients, which is the antigenic distance to convalescent patients in log 2.

282 H_{50} is the antigenic distance at which an individual will have a 50% protective
283 efficacy/effectiveness.

References

1. Garcia-Beltran WF, Lam EC, Denis KS, et al. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* 2021.
2. Chen RE, Zhang X, Case JB, et al. Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. *Nature Medicine* 2021.
3. Wang P, Nair MS, Liu L, et al. Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7. *Nature* 2021.
4. Xie X, Liu Y, Liu J, et al. Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera. *Nature Medicine* 2021.
5. Muik A, Wallisch A-K, Sänger B, et al. Neutralization of SARS-CoV-2 lineage B.1.1.7 pseudovirus by BNT162b2 vaccine–elicited human sera. *Science* 2021: eabg6105.
6. Supasa P, Zhou D, Dejnirattisai W, et al. Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. *Cell* 2021.
7. Zhou D, Dejnirattisai W, Supasa P, et al. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine induced sera. *Cell* 2021.
8. Liu Y, Liu J, Xia H, et al. Neutralizing Activity of BNT162b2-Elicited Serum. *New England Journal of Medicine* 2021.
9. Edara V-V, Lai L, Sahoo MK, et al. Infection and vaccine-induced neutralizing antibody responses to the SARS-CoV-2 B.1.617.1 variant. *bioRxiv* 2021: 2021.05.09.443299.
10. Wall EC, Wu M, Harvey R, et al. Neutralising antibody activity against SARS-CoV-2 VOCs B.1.617.2 and B.1.351 by BNT162b2 vaccination. *The Lancet* 2021; **397**(10292): 2331-3.
11. Hacisuleyman E, Hale C, Saito Y, et al. Vaccine Breakthrough Infections with SARS-CoV-2 Variants. *New England Journal of Medicine* 2021.
12. COVID-19 breakthrough case investigations and reporting. 2021.
13. Kustin T, Harel N, Finkel U, et al. Evidence for increased breakthrough rates of SARS-CoV-2 variants of concern in BNT162b2 mRNA vaccinated individuals. *medRxiv* 2021: 2021.04.06.21254882.
14. Jacobson KB, Pinsky BA, Rath MEM, et al. Post-vaccination SARS-CoV-2 infections and incidence of the B.1.427/B.1.429 variant among healthcare personnel at a northern California academic medical center. *medRxiv* 2021: 2021.04.14.21255431.
15. Abu-Raddad LJ, Chemaitelly H, Butt AA. Effectiveness of the BNT162b2 Covid-19 Vaccine against the B.1.1.7 and B.1.351 Variants. *New England Journal of Medicine* 2021.
16. Cavanaugh AM, Fortier S, Lewis P, et al. COVID-19 Outbreak Associated with a SARS-CoV-2 R.1 Lineage Variant in a Skilled Nursing Facility After Vaccination Program - Kentucky, March 2021. *MMWR Morb Mortal Wkly Rep* 2021; **70**(17): 639-43.
17. Carreño JM, Alshammary H, Tcheou J, et al. Activity of convalescent and vaccine serum against SARS-CoV-2 Omicron. *Nature* 2021.
18. Liu L, Iketani S, Guo Y, et al. Striking Antibody Evasion Manifested by the Omicron Variant of SARS-CoV-2. *Nature* 2021.
19. Cao Y, Wang J, Jian F, et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* 2021.
20. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* 2017; **1**(1): 33-46.
21. Smith DJ, Lapedes AS, de Jong JC, et al. Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science* 2004; **305**(5682): 371-6.
22. Koel BF, Burke DF, Bestebroer TM, et al. Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution. *Science* 2013; **342**(6161): 976-9.
23. Katzelnick LC, Fonville JM, Gromowski GD, et al. Dengue viruses cluster antigenically but not as discrete serotypes. *Science* 2015; **349**(6254): 1338-43.
24. Liu C, Ginn HM, Dejnirattisai W, et al. Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell* 2021; **184**(16): 4220-36.e13.
25. Sitaras I. Antigenic Cartography: Overview and Current Developments. In:

Spackman E, ed. Animal Influenza Virus: Methods and Protocols. New York, NY: Springer US; 2020: 61-8.

26. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proceedings of the National Academy of Sciences* 2016; **113**(12): E1701-E9.

27. Pan K, Subieta KC, Deem MW. A novel sequence-based antigenic distance measure for H1N1, with application to vaccine effectiveness and the selection of vaccine strains. *Protein Engineering, Design and Selection* 2011; **24**(3): 291-9.

28. Anderson CS, McCall PR, Stern HA, Yang H, Topham DJ. Antigenic cartography of H1N1 influenza viruses using sequence-based antigenic distance calculation. *BMC Bioinformatics* 2018; **19**(1): 51.

29. Gupta V, Earl DJ, Deem MW. Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine* 2006; **24**(18): 3881-8.

30. O'Toole Á, Scher E, Underwood A, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution* 2021; **7**(2).

31. Khoury DS, Cromer D, Reynaldi A, et al. Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection. *Nature Medicine* 2021.

32. Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. *Science* 2021; **371**(6526): 284-8.

33. Plante JA, Liu Y, Liu J, et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* 2020.

34. Woo H, Park S-J, Choi YK, et al. Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. *The Journal of Physical Chemistry B* 2020; **124**(33): 7128-37.

Acknowledgements

We acknowledge the authors and originating and submitting laboratories of the sequences from GISAID's EpiCoV Database on which this research is based (GISAID acknowledgments are in Table S3).

Funding: The work was supported by grants from the Health and Medical Research Fund, the Food and Health Bureau, The Government of the Hong Kong Special Administrative Region (COVID190117, COVID1903010) and Guangdong Science and Technology Department (2020B1212030004) to J.H. J.H. thanks the L & T Charitable Foundation and the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2019BT02Y198) for their support.

Author contributions: Y.F.H., B.Z.Z., T.Y., H.C., K.K.W.T., and J.D.H. designed the study. Y.F.H. and J.C.H. analysed the sequences from GISAID and performed the computations and built the online tool. I.F.N.H. and K.K.W.T. recruited the patients and volunteers and collected samples from patients and volunteers. performed the neutralisation assay. Y.F.H., A.D., R.S., K.Y.Y., K.T., H.R.G, and J.D.H. analysed

379 the results. Y.F.H., A.D., T.Y., and J.D.H. wrote the initial draft, and all authors
380 edited the final version.

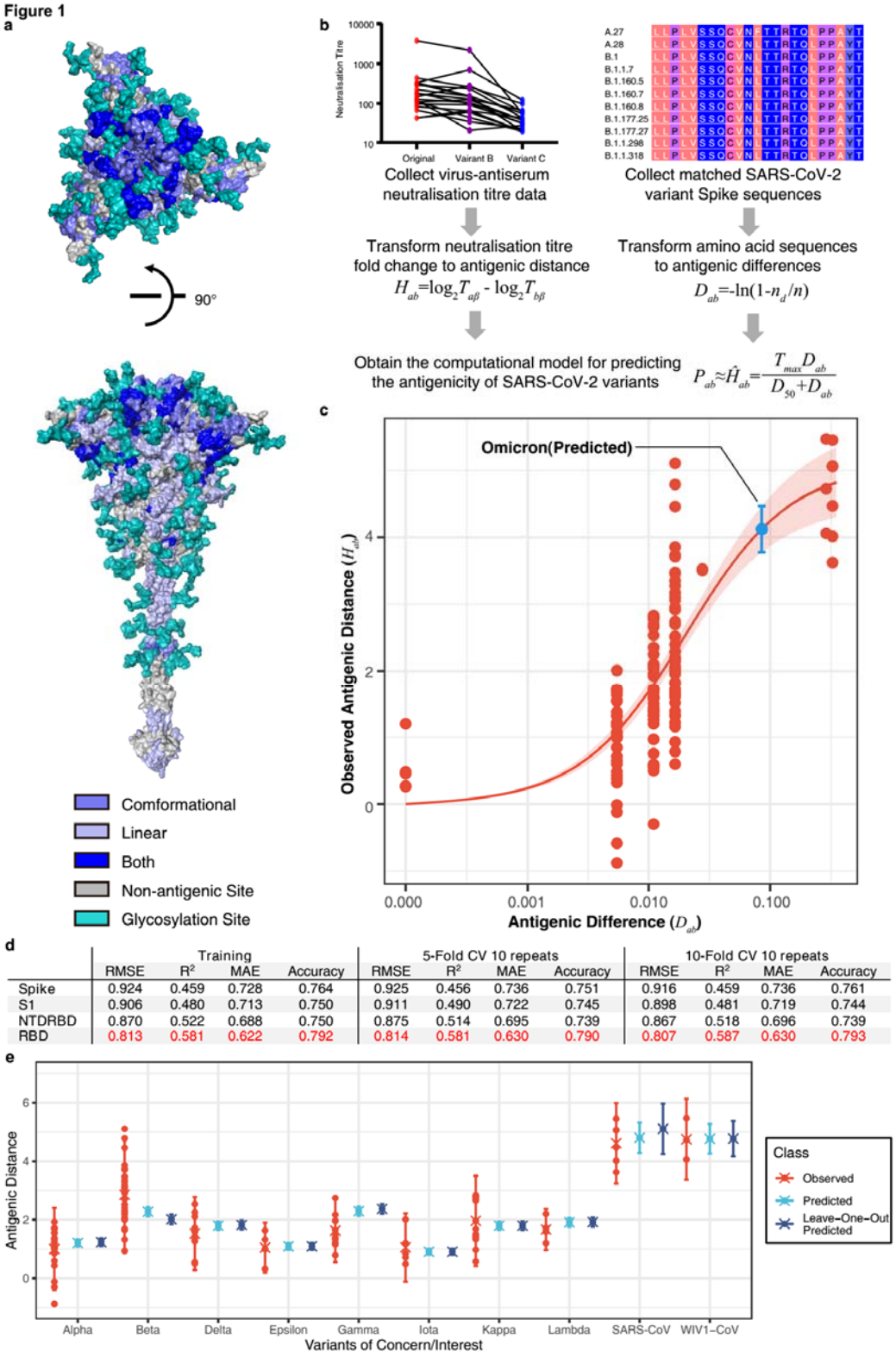
381 **Competing interests:** All authors declare no competing interests.

382 **Data and materials availability**: All sequence data listed in TableS3 are from
383 GISAID's EpiCoV Database.

384 **Supplementary Materials**

385 **Figures S1 to S2**

386 **Tables S1 to S5**



	Training				5-Fold CV 10 repeats				10-Fold CV 10 repeats			
	RMSE	R^2	MAE	Accuracy	RMSE	R^2	MAE	Accuracy	RMSE	R^2	MAE	Accuracy
Spike	0.924	0.459	0.728	0.764	0.925	0.456	0.736	0.751	0.916	0.459	0.736	0.761
S1	0.906	0.480	0.713	0.750	0.911	0.490	0.722	0.745	0.898	0.481	0.719	0.744
NTDRBD	0.870	0.522	0.688	0.750	0.875	0.514	0.695	0.739	0.867	0.518	0.696	0.739
RBD	0.813	0.581	0.622	0.792	0.814	0.581	0.630	0.790	0.807	0.587	0.630	0.793

Fig. 1| Sequence-based prediction of antigenic distance. (a) The top view and the side view of antigenic sites on the full-length Spike protein [34]. The conformational epitopes are coloured in slate and linear epitopes in light blue. Some antigenic

positions in both conformational epitopes and linear epitopes are coloured in blue. All glycosylation sites are in teal. **(b)** A flowchart of the process to establish the sequence-based computational model of SARS-CoV-2 antigenicity. The antigenic distance of variant *a* to reference virus *b* from neutralisation titre was defined as $H_{ab}=\log_2 T_{a\beta}$ - $\log_2 T_{b\beta,}$ where β, $T_{a\beta}$, and $T_{b\beta}$ denote antiserum (referencing virus *b*), the titre of antiserum β against virus *b*, and the titre of antiserum β against virus *a* [26]. The antigenic distance of variant *a* to reference virus *b* from amino acid sequences was defined as $D_{ab}=-\ln(1-n_d/n)$, where n_d is the number of amino acid substitutions between variant *a* and reference virus *b*, *n* is the number of antigenic sites. Then, we proposed a relationship between observed antigenic distance and the antigenic difference: $H_{ab}=T_{max}\cdot D_{ab}/(D_{50}+D_{ab})$, where T_{max} is the maximal fold of decrease and D_{50} is the antigenic difference which may lead to neutralisation decrease at the 50% level of the maximal decrease. **(c)** The relationship between the antigenic difference and the observed antigenic distance. The predicted antigenic distance of B.1.1.529 (Omicron) is marked in cyan. **(d)** The performance of the model in different fragments of the spike protein in terms of root-mean-square error (RMSE), mean absolute error (MAE), coefficient of determination (R-squared R^2), and accuracy. **(e)** Predicted versus observed antigenic distances of variants of concern. Here, The observed antigenic distances as fold decreases in the neutralisation titres of variants of concern versus the original strain on a log 2 scale. Each point shows the mean of antigenic distances in each assay. Predicted antigenic distances are based on the prediction in (c). Leave-one-out predicted antigenic distances are predicted based on the datasets without the variant that we aim to compare.

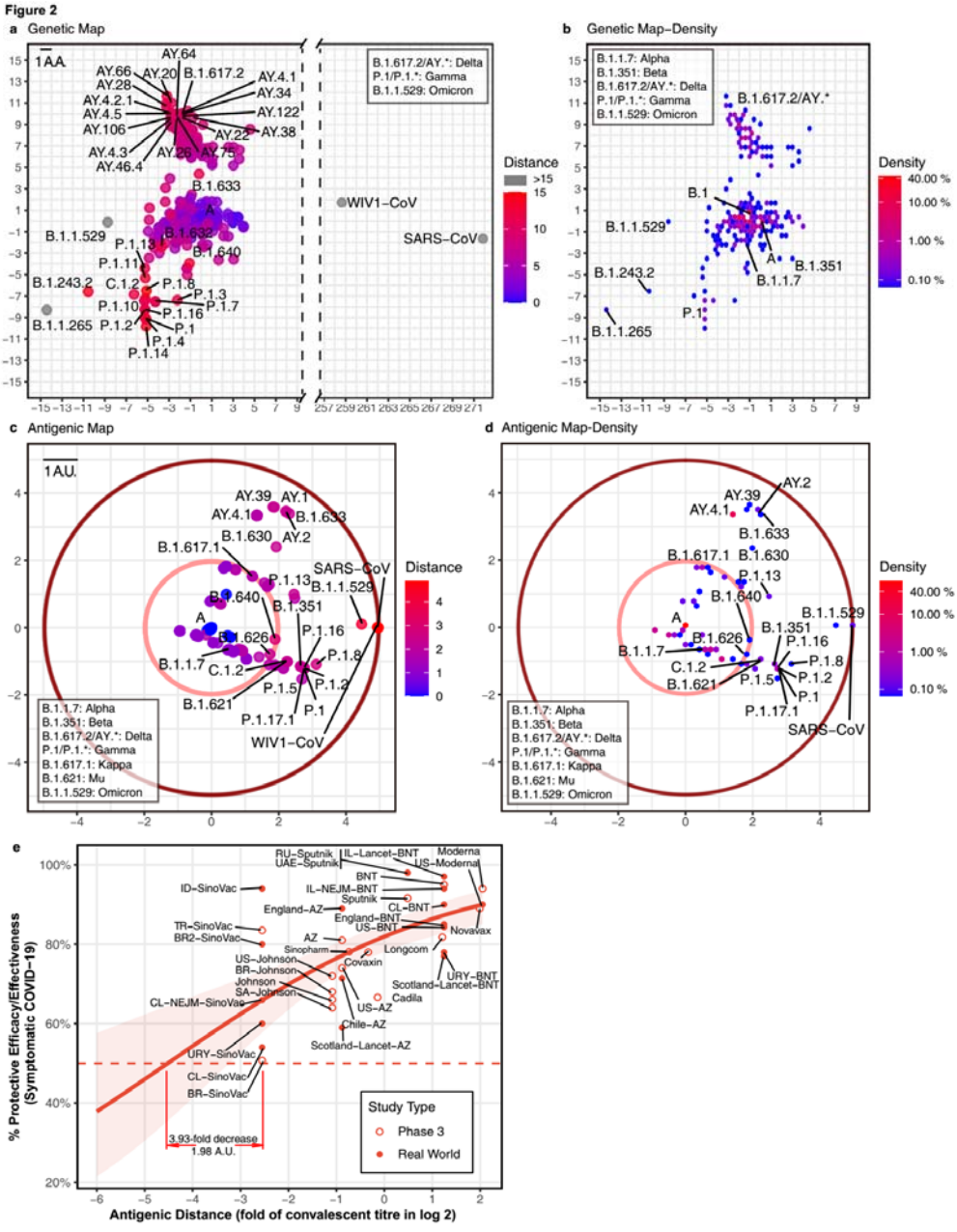
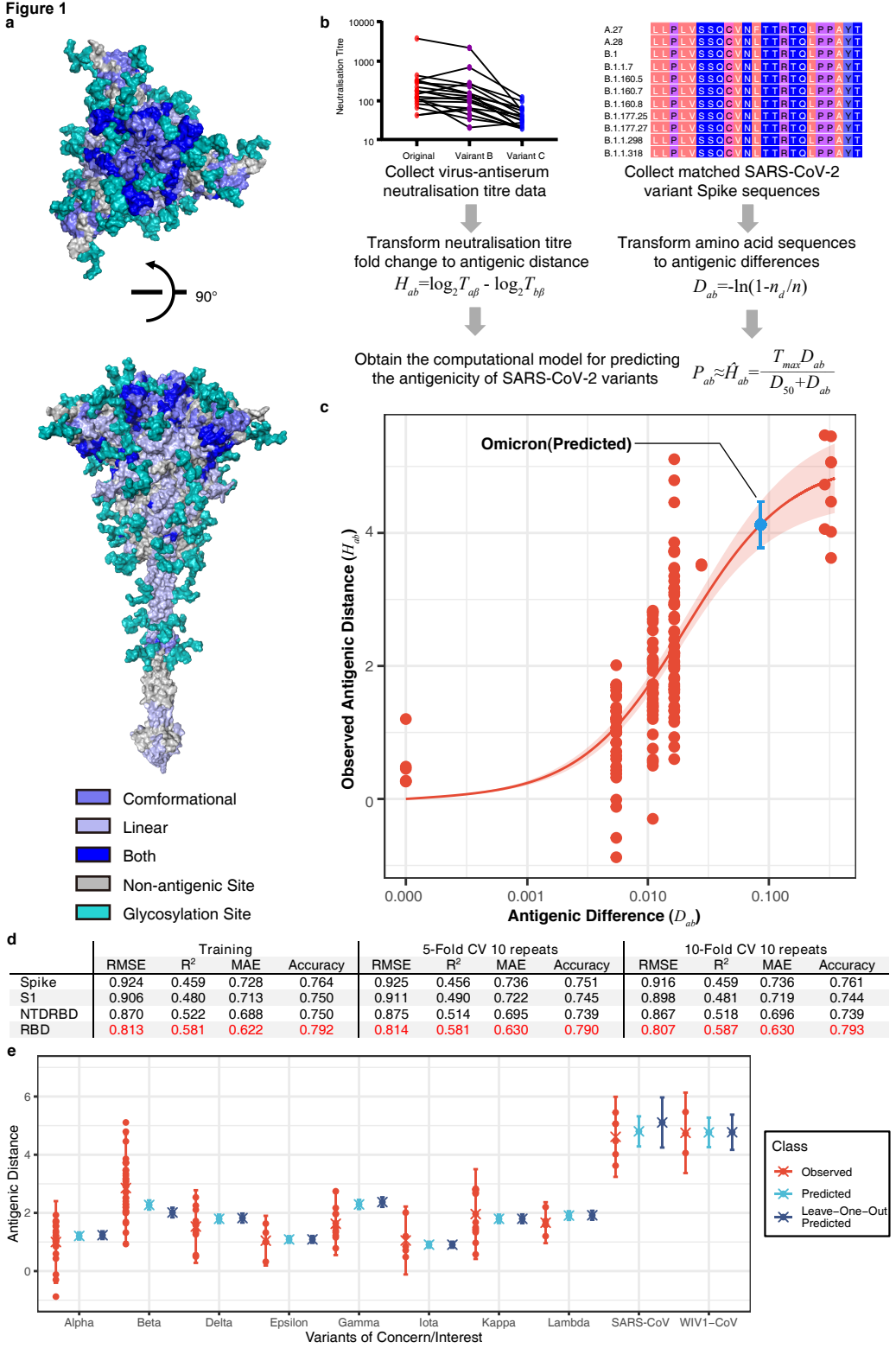


Fig. 2| Genetic and antigenic mapping of SARS-CoV-2 variants. (a) Genetic map of SARS-CoV-2 variant strains shows amino acid mutation numbers of spike proteins, and **(b)** the density of genetic map shows distribution of variants. The vertical and horizontal axes represent the measured relative genetic distances (1 amino acid/1 A.A. = 1 amino acid difference). **(c)** Antigenic map of SARS-CoV-2 variant strains shows the antigenic distance between variants, and **(d)** the density of antigenic map shows distribution of variants. Variants outside the pink circle are vaccine breakthrough candidates. The red circle suggested the border of antigenic map. The antigenic

distance is based on RBD amino acid sequences. The vertical and horizontal axes represent the measured relative antigenic distances (1 arbitrary unit/1 A.U. = 1-fold decrease in the neutralisation titre on a log 2 scale). Colours show the antigenic distance to the SARS-CoV-2 original strain (lineage A). **(e)** Relationship between antigenic distance (mean of neutralisation titres in vaccinees divided by corresponding mean of titres in convalescent patients in log 2) and protection from SARS-CoV-2 infection. The reported mean neutralization level from phase 1 or 2 studies **(Table S4)** and the protective efficacy or effectiveness from phase 3 trials or real-world studies **(Table S5)** for different vaccines. The red line indicates the logistic model, and the red shading indicates the 95% confidence interval of the model. Here, we mark the basis of setting up the cut-off of 3.93-fold decrease (1.98 A.U.).

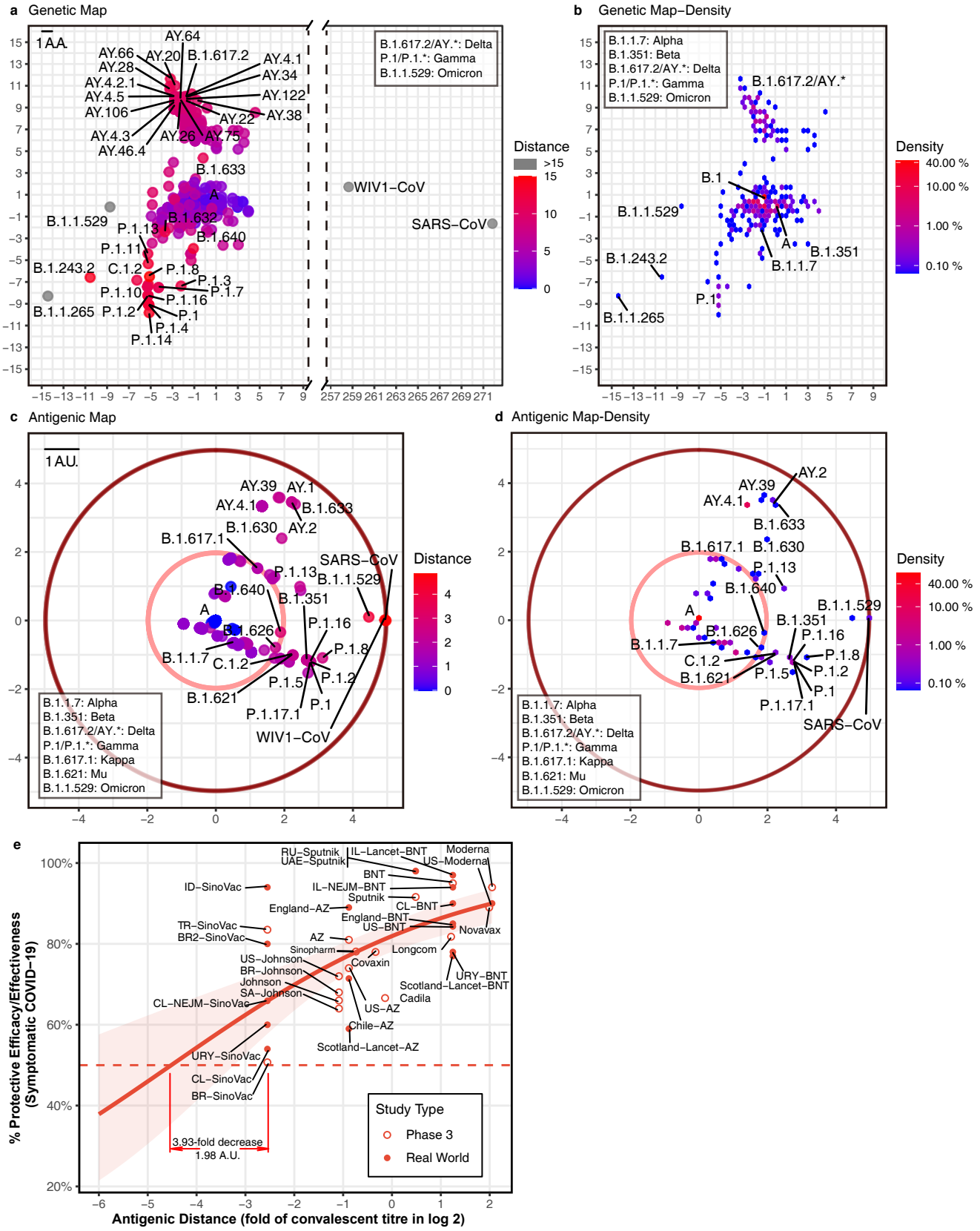
Figure 1



d

	Training				5-Fold CV 10 repeats				10-Fold CV 10 repeats			
	RMSE	R^2	MAE	Accuracy	RMSE	R^2	MAE	Accuracy	RMSE	R^2	MAE	Accuracy
Spike	0.924	0.459	0.728	0.764	0.925	0.456	0.736	0.751	0.916	0.459	0.736	0.761
S1	0.906	0.480	0.713	0.750	0.911	0.490	0.722	0.745	0.898	0.481	0.719	0.744
NTDRBD	0.870	0.522	0.688	0.750	0.875	0.514	0.695	0.739	0.867	0.518	0.696	0.739
RBD	0.813	0.581	0.622	0.792	0.814	0.581	0.630	0.790	0.807	0.587	0.630	0.793

Figure 2


a Genetic Map
1 A.A.
B.1.617.2/AY.*: Delta
P.1/P.1.*: Gamma
B.1.1.529: Omicron
AY.64
AY.66
AY.20
B.1.617.2
AY.4.1
AY.28
AY.34
AY.4.2.1
AY.122
AY.4.5
AY.106
AY.22
AY.38
AY.4.3
AY.46.4
AY.26
AY.75
B.1.633
A
WIV1−CoV
B.1.1.529
B.1.632
SARS−CoV
P.1.13
B.1.640
P.1.11
B.1.243.2
C.1.2
P.1.8
P.1.3
P.1.10
P.1.7
P.1.16
B.1.1.265
P.1.2
P.1
P.1.4
P.1.14
Distance
>15
15
10
5
0
b Genetic Map−Density
B.1.1.7: Alpha
B.1.351: Beta
B.1.617.2/AY.*: Delta
P.1/P.1.*: Gamma
B.1.1.529: Omicron
B.1.617.2/AY.*
B.1
B.1.1.529
A
B.1.351
B.1.243.2
B.1.1.7
P.1
B.1.1.265
Density
40.00 %
10.00 %
1.00 %
0.10 %
c Antigenic Map
1 A.U.
AY.39
AY.1
AY.4.1
B.1.633
B.1.630
AY.2
B.1.617.1
SARS−CoV
P.1.13
B.1.1.529
B.1.640
B.1.351
A
P.1.16
B.1.626
P.1.8
B.1.1.7
C.1.2
P.1.2
P.1.5
B.1.621
P.1
P.1.17.1
WIV1−CoV
B.1.1.7: Alpha
B.1.351: Beta
B.1.617.2/AY.*: Delta
P.1/P.1.*: Gamma
B.1.617.1: Kappa
B.1.621: Mu
B.1.1.529: Omicron
Distance
4
3
2
1
0
d Antigenic Map-Density
AY.2
AY.39
AY.4.1
B.1.633
B.1.617.1
B.1.630
P.1.13
B.1.640
B.1.1.529
A
B.1.351
B.1.626
P.1.16
B.1.1.7
C.1.2
P.1.8
B.1.621
P.1.5
P.1.2
P.1
P.1.17.1
SARS−CoV
B.1.1.7: Alpha
B.1.351: Beta
B.1.617.2/AY.*: Delta
P.1/P.1.*: Gamma
B.1.617.1: Kappa
B.1.621: Mu
B.1.1.529: Omicron
Density
40.00 %
10.00 %
1.00 %
0.10 %
e
% Protective Efficacy/Effectiveness (Symptomatic COVID−19)
Antigenic Distance (fold of convalescent titre in log 2)
RU−Sputnik
UAE−Sputnik
IL−Lancet−BNT
Moderna
US−Moderna
BNT
ID−SinoVac
IL−NEJM−BNT
Sputnik
CL−BNT
England−AZ
England−BNT
TR−SinoVac
US−BNT
Novavax
BR2−SinoVac
AZ
Sinopharm
Longcom
Covaxin
US−Johnson
BR−Johnson
Johnson
URY−BNT
Scotland−Lancet−BNT
SA−Johnson
Cadila
CL−NEJM−SinoVac
US−AZ
Chile−AZ
Scotland−Lancet−AZ
URY−SinoVac
CL−SinoVac
BR−SinoVac
3.93-fold decrease
1.98 A.U.
Study Type
Phase 3
Real World