Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes

Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, Martin Weigt

Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes

Juan Rodriguez-Rivas[a,1], Giancarlo Croce[b,c,1], Maureen Muscat[a], and Martin Weigt[a,2]

[a]CNRS, Institut de Biologie Paris Seine, Laboratory of Computational and Quantitative Biology, Sorbonne Université, 75005 Paris, France; [b]Department of Oncology, Ludwig Institute for Cancer Research Lausanne, University of Lausanne, 1011 Lausanne, Switzerland; and [c]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Edited by John Barton, Physics and Astronomy, University of California, Riverside, CA; received July 16, 2021; accepted December 13, 2021 by Editorial Board Member Mehran Kardar

The emergence of new variants of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a major concern given their potential impact on the transmissibility and pathogenicity of the virus as well as the efficacy of therapeutic interventions. Here, we predict the mutability of all positions in SARS-CoV-2 protein domains to forecast the appearance of unseen variants. Using sequence data from other coronaviruses, preexisting to SARS-CoV-2, we build statistical models that not only capture amino acid conservation but also more complex patterns resulting from epistasis. We show that these models are notably superior to conservation profiles in estimating the already observable SARS-CoV-2 variability. In the receptor binding domain of the spike protein, we observe that the predicted mutability correlates well with experimental measures of protein stability and that both are reliable mutability predictors (receiver operating characteristic areas under the curve ∼0.8). Most interestingly, we observe an increasing agreement between our model and the observed variability as more data become available over time, proving the anticipatory capacity of our model. When combined with data concerning the immune response, our approach identifies positions where current variants of concern are highly overrepresented. These results could assist studies on viral evolution and future viral outbreaks and, in particular, guide the exploration and anticipation of potentially harmful future SARS-CoV-2 variants.

SARS-CoV-2 | mutability | data-driven models | epistasis | direct coupling analysis

The emergence of variants of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a major global health concern. Mutations observed in circulating variants of interest (VOIs) or variants of concern (VOCs) have been associated with increased transmissibility (1, 2), reduced efficacy of antibody treatments (3, 4), and lower antibody neutralization (5). It is currently under investigation how far circulating or future mutants can escape the human immune response induced by vaccination or previous infection (6).

Since the beginning of the COVID-19 pandemic, genomic surveillance of SARS-CoV-2 strains has played a pivotal role in tracking new mutations as they appear and expand. Viral sequences sampled from infected individuals from various parts of the world have been continuously deposited in the GISAID database (https://www.gisaid.org/) (7), and—as of May 2021—more than 1,500,000 genomes are available. Genome-wide analysis of circulating strains (8, 9) has shown that mutated positions are heterogeneously distributed across SARS-CoV-2 proteins: While the vast majority of positions have remained largely invariant to date, a restricted set is accumulating diversity. According to Nextstrain (10) global analysis (May 2021, 3,883 genomes), no mutational event has occurred for 58% of the entire proteome, while only 14% has experienced more than two events. In particular, the protein cores tend to be less variable as mutations in the core usually have a deleterious effect on the stability of the protein (11, 12). In contrast, the exposed regions of the spike protein have accumulated a large number of mutations resulting in variants with increased affinity with the human ACE2 receptor (13, 14) and transmissibility (1, 2) and reduced antibody neutralization (5). Each residue of the SARS-CoV-2 proteome is subjected to different selective pressures which affect the evolution of the virus, thus constraining the variability of SARS-CoV-2 sequences. This suggests that statistical patterns in sequences could be used to distinguish mutable from constrained positions.

In recent years, data-driven models trained on sequence data of patients affected by HIV have been used in this spirit. They identify regions subject to strong selective constraints and therefore less likely to variate (15, 16), guiding the immunogen design of therapeutic strategies being effective against current and future HIV strains (17, 18). Such approaches are trained on large amounts of HIV sequence data, resulting from decades of study and high rates of intrapatient evolution (19). One of the most important lessons of these studies is the importance of epistasis, i.e., the dependence of mutational effects on other preexisting mutations: Epistatic models outperform significantly simpler nonepistatic modeling approaches based on independent conservation patterns of individual residue positions.

Significance

During the COVID pandemic, new severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants emerge and spread, some being of major concern due to their increased infectivity or capacity to reduce vaccine efficiency. Anticipating mutations, which might give rise to new variants, would be of great interest. We construct sequence models predicting how mutable SARS-CoV-2 positions are, using a single SARS-CoV-2 sequence and databases of other coronaviruses. Predictions are tested against available mutagenesis data and the observed variability of SARS-CoV-2 proteins. Interestingly, predictions agree increasingly with observations, as more SARS-CoV-2 sequences become available. Combining predictions with immunological data, we find an overrepresentation of mutations in current variants of concern. The approach may become relevant for potential outbreaks of future viral diseases.

Author contributions: M.W. designed research; J.R.-R., G.C., and M.M. performed research; J.R.-R. and G.C. contributed new reagents/analytic tools; J.R.-R., G.C., M.M., and M.W. analyzed data; and J.R.-R., G.C., and M.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. J.B. is a guest editor invited by the Editorial Board.

[1]J.R.-R. and G.C. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: martin.weigt@sorbonne-universite.fr.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2113118119/-/DCSupplemental.

Published January 12, 2022.

The techniques used for studying HIV sequences are not directly applicable to SARS-CoV-2, for which less than 2 y of data are available, and intrapatient evolution is more limited due to the typically short duration of SARS-CoV-2 infections. We therefore use a strategy which requires only a single SARS-CoV-2 genome to be known; this genome serves as a reference to build alignments of homologous but diverged sequences, which in our case belong mostly to other coronaviruses. These sequences are used to train statistical sequence models, which in turn can be used to predict the mutability of each position in the proteins of the reference SARS-CoV-2. The advantage of this approach is that predictions rely exclusively on data available very early in the outbreak, and predictions can be tested while more data accumulate.

Current approaches predict the mutability along the SARS-CoV-2 proteome using conservation profiles built from multiple sequence alignments (MSA) of SARS-CoV-2 and other related coronaviruses (20, 21). The resulting models (hereinafter independent models, or IND) have few parameters and can be trained using limited sets of data. Unfortunately, they also have only limited predictive power as they assume that positions within a protein evolve independently from each other, disregarding that residues can affect each other's evolution via epistatic interactions.

In this work, we construct unsupervised probabilistic models to predict SARS-CoV-2 mutable and constrained positions. We base our approach on the direct coupling analysis (DCA) (22) that overcomes the aforementioned limitations by explicitly including pairwise epistatic terms in our modeling. The DCA models are trained using families of homologous sequences, broadly collected from all known coronavirus genomes, allowing us to model the general selective pressures acting on the family of coronaviruses. While the use of other coronaviruses substantially enlarges the datasets, making data-driven modeling more robust, we may, however, partially lose information about host-specific constraints like the interaction with host-cell receptors (e.g., ACE2 for SARS-CoV-2) or with the host's immune system.

While models are learned from diverged homologs, the prediction of mutable sites requires a SARS-CoV-2 genome (in our case the Wuhan-Hu-1 strain) to be used as reference: Our models assign a mutability score to each position in each SARS-CoV-2 protein. This score reflects the constraints acting on a position when mutating away from the reference strain. Other SARS-CoV-2 genomes are only required to test our predictions: We assess the predictive power of our approach and of IND models by validating the predictions with the mutations actually observed in SARS-CoV-2 proteomes deposited in GISAID (7). We carry out a detailed study of the receptor binding domain (RBD) as it plays a pivotal role in viral attachment, fusion, and entry and is the primary target for antibody therapies and vaccine development (23, 24). For this specific domain, additional deep mutational scanning (DMS) data are available measuring how amino acid mutations of RBD affect protein expression (a proxy of protein stability) and binding to the human ACE2 receptor (25), allowing us to investigate more deeply their relationship with the DCA mutability score and with the observed variability across SARS-CoV-2 variants.

Most observed mutations are neutral and do not affect the virus phenotype (26); however, mutations occurring in SARS-CoV-2 immunogenic regions, i.e., targets of human B and T cells, may allow the virus to evade the immune response induced by vaccination or previous infection. By combining our DCA-mutability scores with data from the Immune Epitope Database [IEDB (27)], we identify a restricted set of positions in the RBD that are expected to be both mutable and highly immunogenic. Interestingly, we observe that most circulating SARS-CoV-2 VOCs or VOIs have mutations in a subset of those positions. This combined approach also suggests novel positions that are more likely to mutate in the future and whose mutations could induce a reduction in immune response. In this sense, our predictions may help the rational design of new immunogenic or therapeutic strategies, such as monoclonal antibodies or vaccines, to become more efficient against potential future SARS-CoV-2 strains by targeting less-mutable positions.

Data are highly dynamic during the ongoing pandemic. A new variant, Omicron (B.1.1.529), has recently emerged and was rapidly declared a VOC during the final revision of this paper. Due to the great interest in characterizing this variant, we have included a new analysis of its RBD mutations. As compared to all preexisting variants, Omicron increases even the number of mutable and immunogenic positions.

Also, beyond the case of SARS-CoV-2, our study can be seen as a proof-of-concept study. Since we need only a single reference genome to search for distant homologs and make predictions, the approach can be applied very early in any potential viral outbreak in the future. Such predictions may be particularly valuable in situations where observational data on newly emerging pathogens are missing.

Results

According to the Pfam protein-domain family database (28), the SARS-CoV-2 proteome (isolate Wuhan-Hu-1) contains 39 protein domains (see *SI Appendix*, Table S1) covering 81% (7,860 out of 9,748 residues) of the entire proteome. For each of these domains, we predict the mutability using both the epistatic DCA and the independent IND models following the general scheme illustrated in Fig. 1 and detailed in *Materials and Methods*:

- For each protein (domain), we extract MSA of homologous sequences from public sequence databases. These sequences, which belong almost exclusively to other Coronaviridae, diverged during up to $\sim 10^3$ to 10^8 y from their common ancestors with SARS-CoV-2 (29). They are used to train IND and DCA models. These models are applied to the protein sequences of the SARS-CoV-2 reference strain Wuhan-Hu-1 to predict the mutability of each site. Note that only a single SARS-CoV-2 sequence is needed in this step.
- We validate the models using DMS data measuring protein expression, which are currently available only for the RBD of the SARS-CoV-2 spike protein. To this aim, we compare experimentally measured mutational effects with model-based predictions.
- We use SARS-CoV-2 sequences extracted from GISAID to estimate the empirical variability among circulating SARS-CoV-2 strains and to test our predictions. Note that these data are independent from the data used in the first step.

In both approaches, IND and DCA, we thus use the MSA of distant homologs to learn a family-specific sequence landscape $E(a_1,\ldots,a_L)$, or "statistical energy," which provides low values to good (functional) and high values to bad (nonfunctional) sequences. In this context, $(a_1,\ldots,a_L)$ stands for an aligned sequence, i.e., the entries may be any of the 20 natural amino acids or an alignment gap. Any variant containing one or more mutations with respect to the reference sequence in Wuhan-Hu-1 can now be characterized by the statistical-energy change $\Delta E = E(\mathit{reference}) - E(\mathit{variant})$ assigning positive values to variants predicted to be beneficial and negative values to variants predicted to be deleterious. To obtain a position-specific (but not amino acid-specific) mutability score, we average ΔE over all amino acid changes in this position reachable by a single nucleotide mutation; see *Materials and Methods* for the precise definition of $E(a_1,\ldots,a_L)$ and the derived mutability scores.

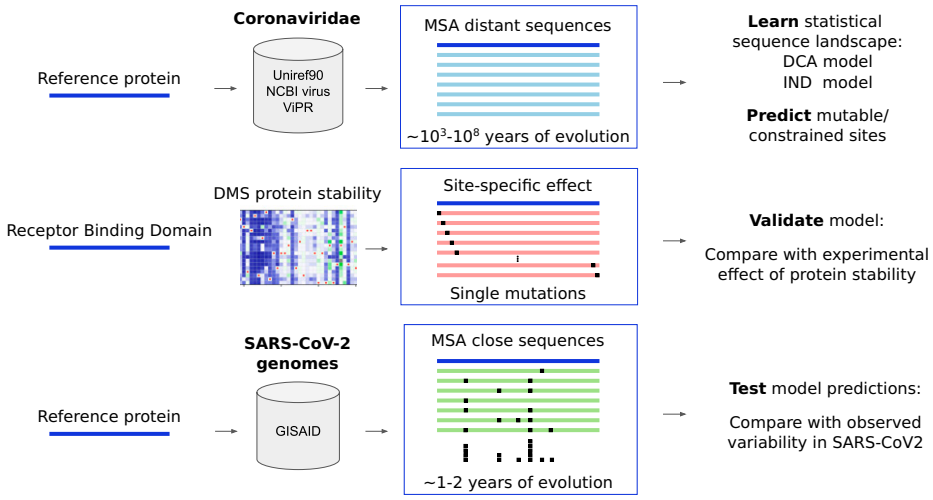

Fig. 1. Scheme of the protocol and data used in the study. The DCA (epistatic) and IND (independent) models are trained with MSA of diverse sequences coming from large sequence databases. For the RBD, we add results of DMS experiments for protein expression (a proxy for stability) (25) which are used as a first independent validation of the models. Model-based predictions are tested against the observed variability, which is derived from SARS-CoV-2 genomes available in GISAID. The estimate on the years of evolution in MSAs of distant sequences provided here (29) is indicative; it can vary strongly between distinct protein domains of SARS-CoV-2.

For testing these mutability scores, we use the same 39 protein domains for extracting a second MSA with variants of SARS-CoV-2 from the GISAID database. To minimize frequency biases due to the extremely heterogeneous sequencing efforts in different countries, we decided to remove redundant amino acid sequences and keep each distinct sequence only once. The position-specific observed variability is now defined as the number of distinct sequences in the resulting MSA having a mutation in the position under consideration, when compared to the Wuhan-Hu-1 reference amino acid sequences (see *Materials and Methods* for more details).

In the case of the RBD, these data (predicted mutability and observed variability) are complemented with the experimental measures for protein expression, used as a proxy for protein stability by Starr et al. (25). This type of data currently exists only for the RBD; we therefore decided to use the RBD for extensive validation of our predictions and to report predictions for the other 38 domains only at the end of *Results*.

Data-Driven Sequence Landscapes Predict Mutability and Mutational Effects in the RBD. To assess our prediction framework, we proceed with the following steps. First, we show that the mutational effect predictions of single-site amino acid mutations are correlated with the protein expression changes measured in the aforementioned deep mutational scan. Second, we show that the RBD variants available in the GISAID database are significantly more neutral than a randomized sequence library of the same sequence divergence from the Wuhan-Hu-1 reference, the latter showing an accumulation of deleterious mutations. These two observations allow us to use predicted close-to-neutrality of a position as an indicator of its high mutability, while positions predicted to have mostly deleterious mutations are expected to be of low mutability. We use extensive comparisons with the sequence variability derived from GISAID to test this hypothesis.

As a first test, we compare the agreement of the computationally predicted mutational effects with the experimental protein expression. Taking into consideration the region of the RBD domain defined by the Pfam profile *bCoV_S1_RBD* (PF09408, aligned length $L = 178$), we compare our predictions with the experimentally measured single-site mutations. We focus on the position-specific mutability, obtained by averaging both predictions and experiments over all accessible amino acid changes in a position (*Materials and Methods*). The DCA model is well correlated with experimental expression (Fig. 2*A*; Spearman's $\rho = 0.54$), clearly superior to the IND model (*SI Appendix*, Fig. S1*A*; $\rho = 0.32$). We also check that individual position and amino acid-specific predictions follow a similar trend (*SI Appendix*, Fig. S1*B*; $\rho = 0.49$ DCA and *SI Appendix*, Fig. S1*C*; $\rho = 0.29$ IND). In brief, we observe that the protein expression and the epistatic model are well correlated, notably better than for the IND model. We note that the model predicts some mutations to be deleterious which are neutral in the expression experiments. Two reasons are possible: 1) In limited datasets of functional sequences, undersampled neutral variants may appear deleterious and 2) mutations without effect on expression may be deleterious for other phenotypes contributing to protein fitness. This observation agrees with what has been observed across other protein families (30–32) where phenotypes better describing fitness also correlate better to sequence-based predictions.

Our DCA and IND models are built from MSAs of diverged species, where we explicitly remove sequences similar to the Wuhan-Hu-1 reference (*Materials and Methods*) to avoid overlaps with the GISAID sequences used to estimate the local SARS-CoV-2 observed variability. We compare how fit, according to the DCA model, the natural sequences are compared to sequences having the same number of random mutations. This can be achieved by comparing the statistical-energy differences $\Delta E = E(reference) - E(variant)$ between the two sets of sequences. Fig. 2*B* shows that the natural SARS-CoV-2 variants are significantly better according to the model than randomly mutated sequences, i.e., the naturally occurring mutations are, according to the model, significantly more neutral than the predominantly deleterious random mutations, as is to be expected by evolution under purifying selection. This finding indicates the capacity of DCA trained on diverged homologs to capture local constraints acting on the evolution of SARS-CoV-2 proteins.

The combination of these two observations is key for our work: The epistatic model is able to capture mutational effects, and the SARS-CoV-2 variants, which emerged over the last months, are significantly more neutral than random mutations. Can we use this to predict possible new variants of SARS-CoV-2 by identifying positions with favorable mutability scores? To test this idea, we compare the currently observable SARS-CoV-2 variability with the one predicted using the model-based mutability

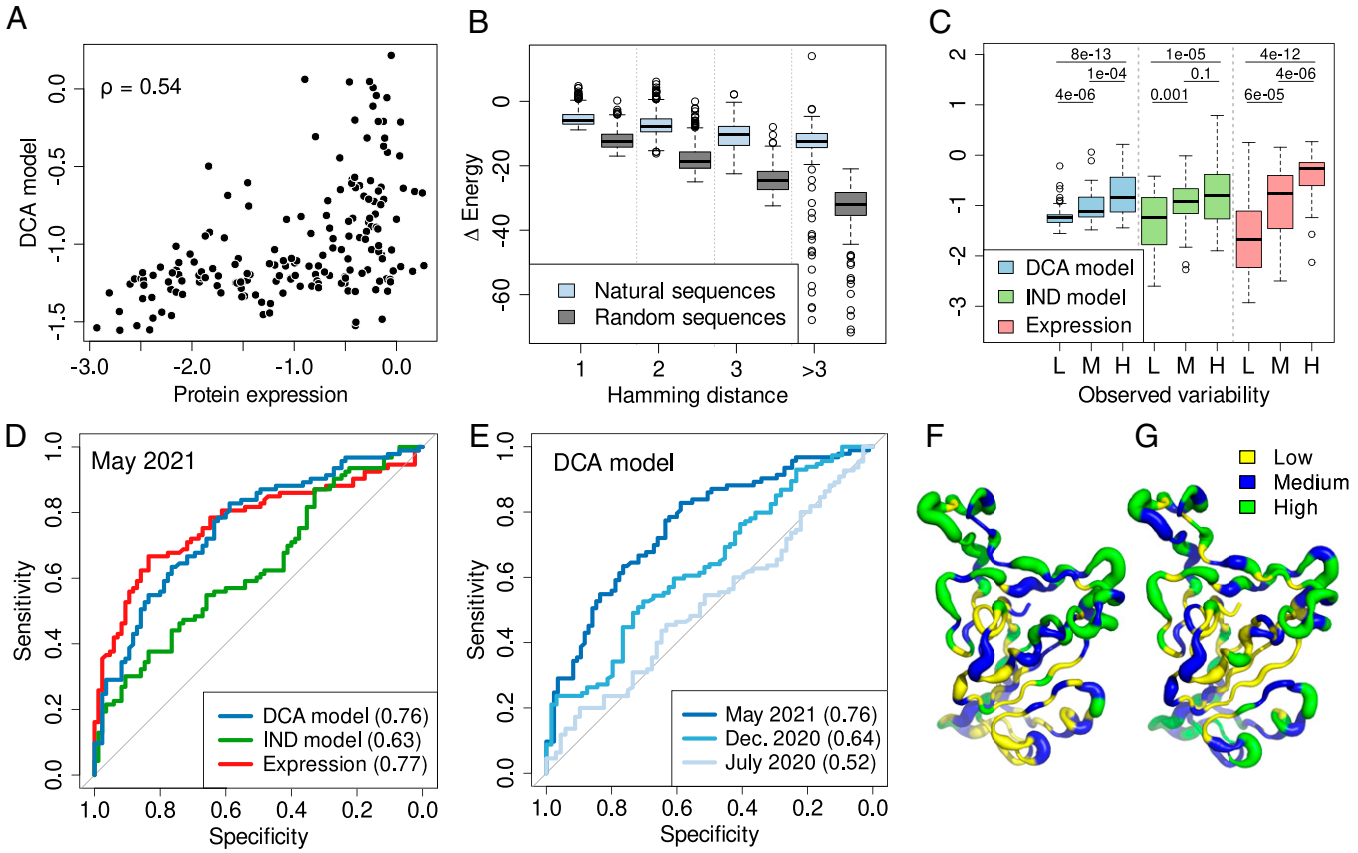

Fig. 2. (*A*) DCA-predicted mutational scores for the 178 positions of the RBD as a function of the experimental protein expression. (*B*) ΔEnergy of GISAID sequences and random sequences compared to the reference RBD sequence (from isolate Wuhan-Hu-1) indicating how well they fit the DCA model as a function of their Hamming distance. (*C*) Distributions of scores from the three predictors for positions with low (L, cutoff: <9, $n = 61$), medium (M, [9,16], $n = 57$) and high (H, >16, $n = 60$) observed variability in GISAID. The *P* values are obtained with the Wilcoxon signed-rank test. L, M, and H in the *x* axis correspond to low, medium, and high observed variability, respectively. (*D*) ROC curves for the DCA model for positions with low (cutoff: ≤12, $n = 93$) vs. high (>12, $n = 85$) observed variability for the three predictors. (*E*) ROC curves for positions with low versus high observed variability, where the observed variability is quantified with the SARS-CoV-2 genomes available at July 2020, December 2020, and May 2021 (*SI Appendix*, Fig. S2*A*), i.e., with increasing accuracy. (*F* and *G*) RBD 3D structure [Protein Data Bank ID code: 6M0J (33)] colored according to three levels of mutational scores from the DCA model (*F*) and the protein expression (*G*). Lower mutational scores are shown in yellow, medium in blue, and higher in green. The width (wider for higher variability) in both panels corresponds to three levels of the observed variability (same cutoffs as in *C*). In all ROC curves, the AUC of the ROC is shown in the legend. Cutoffs to define positions with low and high variability in the ROC analyses were chosen to split into balanced subsets with the most similar number of observations possible in each subset (*SI Appendix*, Fig. S2*A*), i.e., the median is used as the cutoff. ROC curves for the other combinations of predictors and dates are shown in *SI Appendix*, Fig. S3.

score and with the mutations expected by the experimental protein expression. As mentioned before and illustrated in Fig. 1, we operationally assess the observable variability by the number of distinct GISAID sequences having a variant amino acid (compared to the reference Wuhan-Hu-1) in the specific position under study. We observe that the DCA model and expression are similarly correlated with variability (Spearman's $\rho = 0.61$ and 0.6, respectively), while the correlation is weaker for the IND model ($\rho = 0.34$). This trend can also be observed in Fig. 2*C* by looking at the distribution of mutational scores after grouping positions by their observed variability; they grow accordingly with the variability (note that the scores are not comparable between methods, but *P* values indicate a higher significance for DCA and expression than for IND).

We analyze in detail the performance of these different measures as a predictor of SARS-CoV-2 mutability through receiver operating characteristic (ROC) curves and the resulting areas under the curve (AUC), which range from 0.5 for random to 1.0 for perfect predictions. We perform the ROC analysis using a variability cutoff of 12, because it splits the set of positions into two balanced subsets of positions with low (≤12, $n = 93$) or high (>12, $n = 85$) variability. The DCA model and protein expression (AUC 0.76 and 0.77) show a remarkable performance in distinguishing positions with low or high variability, clearly outperforming the IND model (AUC 0.63). This result is not dependent on the particular cutoff chosen, as a similar trend is observed for a large range of variability cutoffs (*SI Appendix*, Fig. S2*B*). The protein expression and the DCA model perform similarly (see *SI Appendix*, Fig. S2*B*; averaged ROC AUC of 0.83 and 0.81, respectively), followed by the IND model (0.66). The experiment-based predictor performs comparatively better at high variability while the sequence-based ones are better at low variability (*SI Appendix*, Fig. S2*B*), which is probably related to the fact that highly conserved positions are usually very relevant to the function of the protein (34). In conclusion, we observe that the different measures have a substantial predictive power of the mutability, although the IND model is worse compared to the others. The performance of the DCA models is surprisingly competitive, with a performance similar to the experimental measurements, with an advantage for lower mutabilities and a slight disadvantage for higher mutabilities (cf. Fig. 2*B*).

The performance measured in the previous analysis is not only dependent on the intrinsic predictive power of each method but also on how well the ground truth is defined, in this case the variability estimated from GISAID data. Although only a fraction of all possible variants emerged in the short time since SARS-CoV-2 appeared, the observed variability has greatly evolved over time due to the great effort of sequencing SARS-CoV-2 genomes (*SI Appendix*, Fig. S2*A*). As an example, the number of RBD positions without observed variants has shrunk from 58 out of 178 in July 2020 to only 3 in May 2021. Interestingly, we observe a great increase in predictive performance when evaluated against more recent and richer variant libraries considered in the estimation of variability. As shown in Fig. 2*E*, the performance has increased from an AUC of only 0.52 considering sequences collected until July 2020 up to 0.76 with sequences until May 2021 (at each time point, the median variability is used to partition data into low vs. high variability; see *SI Appendix*, Fig. S2*B* and *Materials and Methods* for details). This improvement is not only evident from the ROC analysis but also by looking at the correlation between the DCA model and the variability with Spearman correlations of $\rho = 0.13$, 0.35, and 0.61 in July 2020, December 2020, and May 2021, respectively (cf. *SI Appendix*, Fig. S4). This result indicates that the remarkable increase of SARS-CoV-2 genomes has led to a much better estimation of the variability. More importantly, it shows that our computational model is able to anticipate future variability; it suggests that the performance could be even higher with more data, leading to an even better estimation of the variability. We have complemented this analysis also for mutability predictions done using the IND score or protein expression (*SI Appendix*, Fig. S3); all show the tendency to improve with the increase of SARS-CoV-2 data available in the GISAID database. Only the IND score-based analysis shows no clear trend between December 2020 and May 2021.

Recently deep learning has been used to improve mutational predictions in the case of large training MSA, but the accuracy for viral proteins remains rather limited (35). To explore this issue, we have used DeepSequence (35). Its predictions correlate well with the DCA model predictions ($\rho = 0.6$; *SI Appendix*, Fig. S5*C*), but the correlations with protein expression (0.31; *SI Appendix*, Fig. S5*A*) and observed variability (0.35; *SI Appendix*, Fig. S5*D*) are smaller than those of the DCA model ($\rho = 0.54$ and 0.61), congruent with the prior observations for other viral proteins.

To conclude the comparison of computational predictions, observed variability, and experimental DMS data for the RBD, we explore how these are distributed within the three-dimensional RBD structure (Fig. 2 *F* and *G*). Apart from an overall agreement between the three quantities, there is a clear trend for lower values in the core of the RBD and higher in the exposed parts of the structure, probably related to the greater impact of mutations on the stability in the core and the selective pressure for immune escape in the surface.

As a summary of this section, we conclude that the epistatic DCA prediction for the position-specific mutability of RBD positions in SARS-CoV-2 is highly informative about the observable variability across the increasing number of sequenced SARS-CoV-2 variants. The increased accuracy when compared to the most recent versions of GISAID proves the anticipatory power of our approach: The positions that are predicted as mutable by our approach are more likely to be associated with future SARS-CoV-2 variants.

Combining Mutability Predictions with Immune Response Frequencies Identifies Mutations Present in Several SARS-CoV-2 VOCs. Nonsynonymous mutations of SARS-CoV-2 occurring in immunogenic regions can cause the virus to (partially) escape the human B and T immune response induced by vaccination or previous infection. B and T cells target specific regions of the viral proteome, known as B/T cell epitopes. Epitope mutations can reduce the ability of the immune cells to recognize and bind epitopes and thus the effectiveness of the immune response.

Antibody-escaping mutations are already present in circulating variants (4, 36). On 18 May 2021 the World Health Organization's weekly epidemiological report on COVID-19 (available at https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19—18-may-2021) classified six SARS-CoV-2 variants as VOIs and four—posing an increased risk to global public health—as VOCs. Both VOIs and VOCs are likely to affect transmission, diagnostics, therapeutics, or immune escape (37). Within the RBD domain, only seven positions of VOI and VOC strains are mutated with respect to the Wuhan-Hu-1 reference strain (Fig. 3*A*).

The IEDB (27) collects experimentally validated B and T cell epitopes. Most of the SARS-CoV-2 epitopes are localized on the spike protein: a total of 913 epitopes, 459 B cell and 463 T cell epitopes, as of 16 May 2021.

Each position of the spike protein is associated with a site response frequency (RF) (*Materials and Methods*). The RF is calculated as the number of positively responding subjects relative to the total number of those tested, averaged over all epitopes mapped to that position. Nonsynonymous mutations in position with high RF typically modify multiple epitopes with the risk of negatively affecting the human immune response.

In Fig. 3*B* we plot the IEDB RF versus the DCA mutability score for each position of the SARS-CoV-2 RBD domain. Interestingly, only a restricted set of positions has high DCA and RF scores at the same time (the upper right corner of Fig. 3*B*), and four of them are observed in circulatining VOCs and VOIs, including the well-known positions N501 and E484. *SI Appendix*, Fig. S6 shows analogous plots of the RF vs. the IND score or the expression data; the enrichment of VOC/VOI mutations becomes less pronounced as compared to the DCA score. These results highlight the potential for our approach to identify positions that are likely to mutate (high DCA score) and whose mutations may cause immune escape (high IEDB RF). The first 20 predictions, sorted according to the DCA mutability score, are given in Table 1. Note that nine predictions have a high RF (>0.3, highlighted in bold), and several of them are not yet part of current VOCs and VOIs. Our results suggest them as potentially dangerous positions likely to mutate in future SARS-CoV-2 strains.

As the virus is constantly changing through mutation, other circulating SARS-CoV-2 VOCs/VOIs have emerged during the redaction of the paper. Also, more epitopes have been tested and immunological data are rapidly accumulating, and statistically more reliable IEDB RFs are now available. In December 2021 we repeated the same analysis using updated IEDB data (downloaded 22 November 2021) and the five current VOCs. The results reported in *SI Appendix*, Fig. S7 confirm the enrichment of dangerous mutations in the upper right corner, which is particularly pronounced for the newly emerged Omicron (B.1.1.529) variant. Indeed, of the 14 RBD mutations present in Omicron, 6 (K417, N440, E484, Q493, Q498, N501) are in the first top 20 DCA predictions (Table 1). Remarkably, mutations in positions N440, Q493, and Q498 occur for the first time in Omicron; they are not shared by other VOIs and VOCs.

We repeated our analysis distinguishing between T and B cell epitopes (*SI Appendix*, Fig. S8). While for B cell epitopes it is still possible to clearly identify a subset of positions with high DCA and RF scores, this is not the case for T cell epitopes. This is expected as B cell antibodies directly bind the pathogen, while T cell epitope must be presented by the human leukocyte antigens (HLAs)—one of the most polymorphic genes in the human genome—and have a much larger sequence variability. Limited T cell data makes it arduous to obtain a statistically

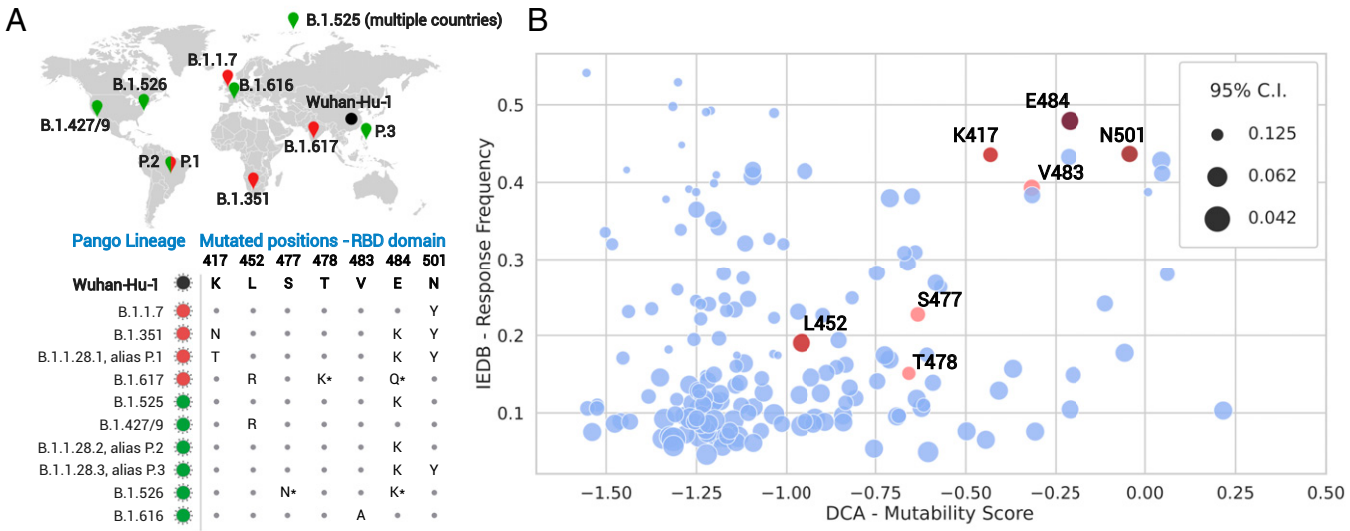

Fig. 3. (*A*) SARS-CoV-2 strains classified in May 2021 as VOCs (red, now also named Alpha [B.1.1.7], Beta [B.1.351], Gamma [P.1], and Delta [B.1.617.2]) and VOIs (green). The figure shows the corresponding amino acid mutations with respect to the Wuhan-Hu-1 reference in the RBD domain and the geographical area where they were first detected. The B.1.617 lineage is divided into three sublineages; the E484Q and T478K (with asterisks) mutations are not shared by all sublineages. The same is true for E484K and S477N in the B.1.526 lineage. (*B*) The IEDB RF and the DCA mutability score for each position of the RBD domain. The upper right corner contains potentially dangerous positions, as they are predicted to be mutable (high DCA mutability score) and are shared by multiple positively responding epitopes (high IEDB RF). Mutated positions observed in VOCs and VOIs strains are depicted in red, and darker shades correspond to the most frequent mutations. The size of each point is inversely proportional to the IEDB 95% CI [size ∼1/(upper bound − lower bound)], thus larger points correspond to more statistically reliable IEDB RF.

reliable T cell IEDB RF, even after restricting the analysis to a subset of T cell epitopes shared by a large fraction of the population (*SI Appendix*, Fig. S9) (39).

Mutability Predictions Are Extendable to All SARS-CoV-2 Proteins. Thanks to the wide availability of sequence data as compared to experimental data, a key advantage of our data-driven modeling approach is the possibility to obtain predictions for all the protein domains in the SARS-CoV-2 proteome. We extend the analysis to all 39 protein domains covering 81% of the entire proteome (8,037 out of 9,748 positions). First, we observed that both the mutational scores from DCA and IND models systematically grow with the observed variability (Fig. 4*A*), with a more pronounced change for the DCA model reflected by smaller *P* values. The performance is influenced by both the quality of the models, which depends on the available sequence data, and the definition of the ground truth, i.e., the observed variability. In accordance with previous findings for the RBD, we observe that the performance of the DCA model improves as more data are used to estimate the variability; the ROC AUC goes from 0.59 in July 2020 to 0.71 in May 2021. The increased performance observed in the RBD can be

Table 1. The first 20 predictions, sorted according to the DCA mutability score, with the corresponding IEDB RF and the VOIs and VOIs in which the position has mutated

Position	AA Wuhan-Hu-1	DCA mutability score	IEDB RF (95% CI)	Pango lineage (ref. 38)
519	H	0.22	0.10 (0.08:0.14)	
403	R	0.06	0.28 (0.24:0.32)	
490	**F**	**0.05**	**0.41 (0.38:0.45)**	
493	**Q**	**0.04**	**0.43 (0.40:0.46)**	
372	**A**	**0.01**	**0.39 (0.32:0.46)**	
501	**N**	**−0.04**	**0.44 (0.40:0.47)**	**B.1.1.7; B.1.351; P.1; P.3**
445	V	−0.06	0.18 (0.15:0.21)	
498	Q	−0.11	0.24 (0.21:0.28)	
441	L	−0.20	0.15 (0.12:0.19)	
440	N	−0.21	0.10 (0.08:0.14)	
484	**E**	**−0.21**	**0.48 (0.45:0.51)**	**B.1.351; P.1; B.1.617; B.1.525; P.2; P.3**
486	**F**	**−0.21**	**0.43 (0.40:0.47)**	
443	S	−0.31	0.08 (0.05:0.11)	
494	**S**	**−0.32**	**0.38 (0.35:0.42)**	
483	**V**	**−0.32**	**0.39 (0.36:0.43)**	**B.1.616**
460	N	−0.37	0.16 (0.13:0.19)	
444	K	−0.41	0.13 (0.10:0.16)	
417	**K**	**−0.43**	**0.44 (0.40:0.48)**	**B.1.351; P.1**
439	N	−0.44	0.07 (0.04:0.10)	
402	I	−0.50	0.08 (0.05:0.11)	

Positions with IEDB RF above 0.3 are shown in bold.

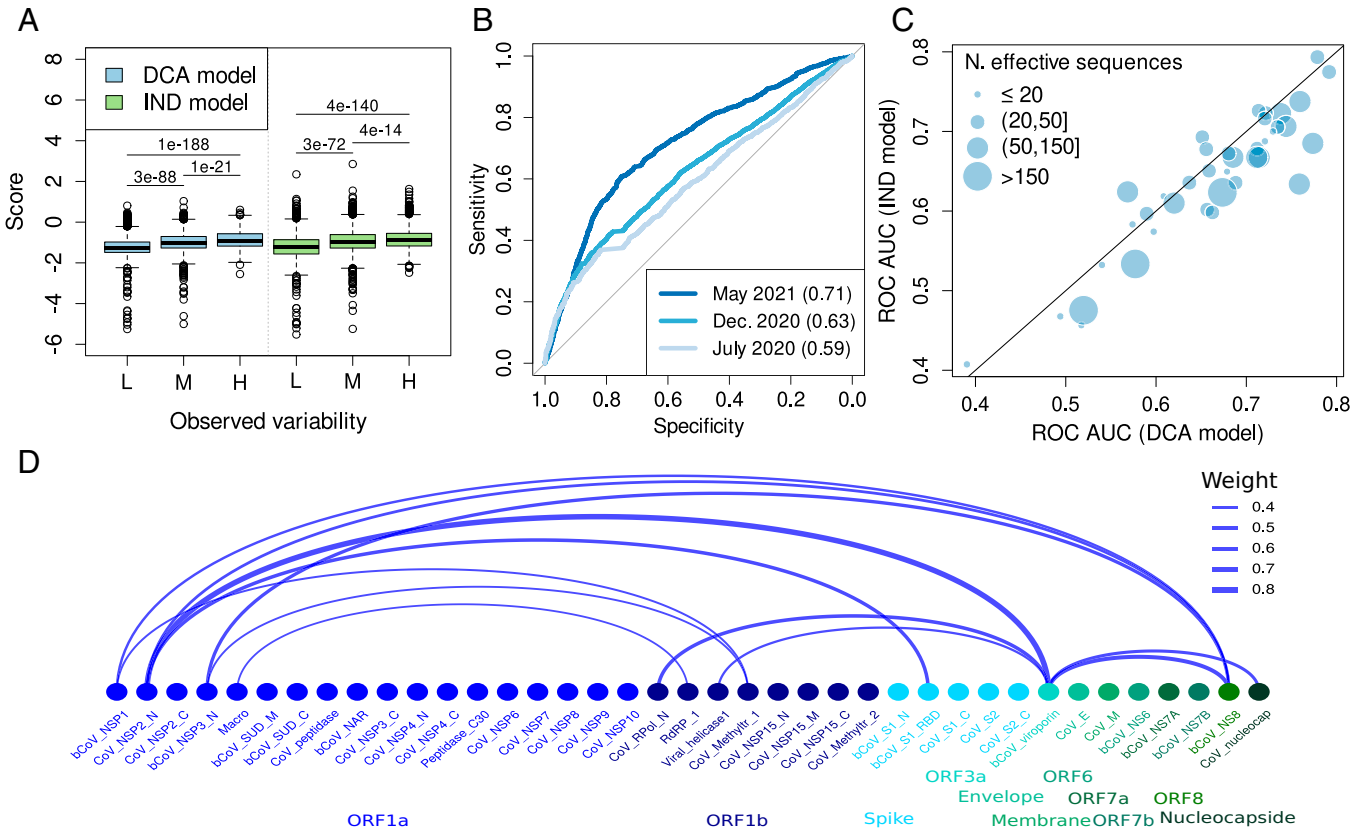

Fig. 4. (*A*) Distribution of DCA and IND scores as a function of the variability (L, low <7, n = 2,757 positions; M, medium = [7,15], n = 2,647; H, high >15, n = 2,554) for the entire SARS-CoV-2 proteome (*P* values from the Wilcoxon signed-rank test). L, M, and H in the x axis correspond to low, medium, and high observed variability, respectively. (*B*) ROC curve for the classification provided by the DCA model for positions with low (≤3, n = 4,873 in December 2020) or high (>3, n = 3,085 in December 2020) variability, where the variability is estimated from data until May 2021, December 2020, or July 2020. (*C*) Comparison of ROC AUC obtained by the DCA and IND models for the 39 domains in the proteome. The variability cutoff for each domain is chosen to give rise to two balanced subsets of positions. (*D*) The nodes represent the Pfam domains in the proteome with a link between pairs of domains when they have at least one relatively strong epistatic coupling. The width of the link is proportional to the strength of the signal, or weight, which comes from the strongest coupling among all the interdomain pairs of positions. Protein domains codified within the same open reading frame (ORF) share the same color.

partially attributed to a better estimation of the variability, as it is one of the most variable regions of the SARS-CoV-2 proteome (40). Another important factor is the impact of the available sequence data (*SI Appendix*, Table S1). To take into consideration this factor, we split the 39 protein domains into two sets with at least 50 or fewer than 50 effective sequences (i.e., nonredundant at 80% identity; see *Materials and Methods*). As expected, the performance is greater when more sequence data are available to build the models, as shown in *SI Appendix*, Fig. S10. A systematic comparison between the models reveals that the DCA model is better than the IND in most cases, especially in those with a higher number of effective sequences (Fig. 4*C*).

Beyond the RBD domain, other protein domains have an important role in triggering an immune response in humans. We extend the analysis of the previous section (combining immunological data with DCA predictions) to all the protein domains of the SARS-CoV-2 proteome. The results are available on the GitHub page.

Our predictions are based on individual Pfam protein domains. While we argued that epistasis is a crucial ingredient to our models, we currently do not include epistasis between distinct domains. The main reason is that multidomain studies risk again limiting available sequence data. To get a first impression of the potential role of epistasis between distinct protein domains in SARS-CoV-2 evolution, we have used DCA to detect epistatic couplings between all 741 pairs of the 39 present domains (*SI Appendix*, *SI Text*). As is shown in Fig. 4*D*, we find a sparse network of only 12 potentially coupled domain pairs, out of the 601 pairs providing sufficient data for our analysis (*SI Appendix*, *SI Text*). While the sparsity of this network makes it unlikely that our mutability predictions suffer from our domain-centric modeling approach, our results suggest the existence of interdomain and interprotein epistasis in SARS-CoV-2. This conclusion is coherent with that of ref. 41, which differently from our analysis is based entirely on an analysis of the SARS-CoV-2 genomes deposited in GISAID. It is worth mentioning that the strongest epistatic coupling found in our analysis is between the domain Cov_NSP2_N and bCov_viroporin (*SI Appendix*, Table S3), which is also highlighted in ref. 41. However, a biological interpretation of these findings is not obvious due to the limited availability of experimental information about potential physical or functional interactions between SARS-CoV-2 proteins.

Discussion

In this work, we propose to use statistical models to predict the mutability of individual positions in SARS-CoV-2 proteins. The models are based on MSAs coming from various coronaviruses. The inclusion of epistasis in the DCA-based modeling framework allows us to capture local evolutionary constraints specific to the SARS-CoV-2 sequence background. Using several tests

for the RBD of the spike protein, for which the most extensive experimental datasets are available, we were able to establish that our computational predictions are able to anticipate position mutated in variants of SARS-CoV-2 from sequence alignments not containing SARS-CoV-2 sequences. This fact is particularly evident in Fig. 2*E*, which shows that more recent and thus richer releases of the GISAID database of SARS-CoV-2 genomes follow more accurately our model predictions. The inclusion of epistasis in the modeling was found to be essential to improve the quality of the mutability predictions.

The combination of our predictions with available immune response frequencies allows for identifying a relatively small group of 9 positions out of the 178 positions in the RBD which are highly mutable and have a high potential for immune escape. Interestingly, four out of these nine positions are mutated in the current VOCs or VOIs. The other five positions are predicted to potentially give similar advantages for emerging SARS-CoV-2 variants. In fact, a new variant was declared a VOI in June 2021 (lineage C.37, or Lambda). This variant has two RBD mutations in positions L452 and F490. While the first is shared with other lineages (even if substituted to another amino acid), the second one was not part of any previous VOI or VOC, but it is the third predicted position in terms of mutability and the first with high RF (cf. Table 1). Even more recently, in November 2021, the variant Omicron (B.1.1.529) emerged and was declared a VOC; it shows new mutation in positions Q493, Q498, and N440, which were not mutated in preexisting VOIs and VOCs but take ranks 4, 8, and 10 in Table 1 (cf. also *SI Appendix*, Fig. S7). Our approach therefore highlights the importance of monitoring these positions, which could also be taken into account when exploring potential therapeutic or vaccine targets.

This can be illustrated by the following example. A monoclonal antibody was recently isolated which has neutralizing activity against all SARS-CoV-2 VOCs identified to date (42). The antibody targets a region of about 600 $Å^2$ of the spike protein surface centered in residue F486. This residue is predicted to be quite mutable (rank 12 in Table 1, next to E484), and mutations might have immunological relevance as indicated by a high IEDB RF—so a mutation in F486 emerging in a new variant might decrease the neutralizing capacities of the antibody. However, this residue is also in contact with the ACE2 human receptor, and thus a mutation might also decrease the affinity with the host protein, resulting in an evolutionary disadvantage for the virus. While our model, trained on long-term evolutionary data, does not contain specific knowledge about the RDB–ACE2 interaction, it suggests that positions like F486 should be carefully studied with complementary structural and experimental approaches and considered when designing antibodies effective against novel strains.

Being based on readily available sequence data is one of the advantages of our approach over more labor-intensive experimental approaches like the DMS data, such as the effect of mutations on RBD expression and binding to the human ACE2 receptor, allowing us to provide useful predictions of mutability for most of the SARS-CoV-2 proteome. It also has its limitations, most importantly its dependence on the availability of sufficiently large and diverged sequence ensembles. In fact, we observe that a greater number of sequences usually increases the performance of the approach (Fig. 4*C* and *SI Appendix*, Fig. S10). However, it is important to note that the inclusion of more divergent sequences might not always be the best strategy as the model might capture constraints that are not relevant for the specific SARS-CoV-2 context. This trade-off will be explored in future work.

Our approach can be extended in several ways. One is to include how different domains might constrain the variability of other domains. However, according to our analysis in the previous section, interdomain epistasis seems to play only a minor role, even if more sequence data might be needed to better estimate the influence of interdomain or interprotein epistasis. Another is to model constraints due to specific virus–host interaction, which is currently out of our scope, as we do not consider host sequences in the MSAs. Indeed, we observe the correlation of experimental binding to ACE2 and our predictions (Pearson's $r = 0.27$) can be fully explained through the protein expression (Pearson's r partial correlation controlled by expression $= -0.02$). In an attempt to explore this issue, we built coalignments of RBDs with homologs of ACE2 present in the hosts of other coronaviruses. Since the binding mechanism between RBD and ACE2 homologs is present only in sarbecoviruses, the resulting coalignment of RBD and ACE2 homologs contains only seven effective sequences (nonredundant sequences at 80% identity; cf. *Materials and Methods*), a number insufficient to capture the complex virus–host interactions.

Predicting evolution is an undoubtedly daunting task (43, 44). While there is little, if any, hope to predict specific future evolutionary events, we have shown that data-driven approaches capturing statistical patterns in sequence data can effectively identify more general evolutionary trends, such as which positions are more likely to mutate and represent a concern to current therapeutic interventions. In this sense, our work is a step forward to a more precise characterization of the SARS-CoV-2 evolution fueled by a huge worldwide effort of research and monitoring of the virus, whose evolution is unfolding in almost real time at an unprecedented level of detail.

While the main application of our work is the insights provided on SARS-CoV-2, our study can also be seen as a proof of concept. In the case of emergence of a new viral pathogen, a single sequenced genome can be used as the reference to first extract families of homologous sequences from public databases, which allow for learning the statistical models needed for mutability predictions. These predictions can therefore be done in very early stages of a possible outbreak, before large amounts of observational or experimental data become available, forecast future variability, and thereby help to direct our attention to as-yet-unobserved mutations.

Materials and Methods

Sequence Data. Sequence data in FASTA format were downloaded from the following databases: GISAID (ref. 7, release 16 May 2021), Uniref90 (ref. 45, release December 2020), ViPR (ref. 46, downloaded in September 2020), NCBI viral genomes (ref. 47, downloaded in September 2020), and MERS coronavirus database (ref. 48, downloaded in September 2020). The amino acid sequence of isolate Wuhan-Hu-1 was used as the reference proteome (GenBank accession no. MN908947). Protein domains were detected using the HMMER suite (ref. 49, version 3.1b2) and the HMM profiles from Pfam.

A global database including distant species was built by combining Uniref90, ViPR, NCBI viral genomes, and the MERS coronavirus database and used to train the DCA and IND models. We built MSAs by running jackhmmer with five iterations and starting both with the full-length reference protein sequence (except for the ORF1ab) and with the trimmed domain sequence (see *SI Appendix, SI Text* for more details). For each domain, we selected the MSA with more nonredundant sequences between the two resulting MSAs for further analysis, which increases the amount of available sequence data. As quality controls, all sequences including nonstandard amino acids were removed as well as repeated sequences or sequences covering less than 80% of the reference; predictions are robust when modifying this threshold (cf. *SI Appendix*, Fig. S11*A*). To separate training from test data, all sequences closer than 90% sequence identity to the Wuhan-Hu-1 reference were filtered out (i.e., all SARS-CoV-2 sequences, including close relatives in nonhuman hosts). The exclusion of SARS-CoV-2 reference sequences has a negligible influence on the predictions, e.g., the spearman's correlation on the RBD of the DCA scores with protein expression with and without the reference sequence is the same ($\rho = 0.54$) as well as in the case of the observed variability ($\rho = 0.61$).

For the GISAID database, a MSA for each domain sequence was built with only one iteration in jackhmmer as the GISAID sequences are very similar to

the reference sequences. We applied the same quality controls as before but kept sequences closer than 90% sequence identity and removing sequences corresponding to a nonhuman host. The July and December 2020 subsets of sequences were collected until the 16th of the corresponding month. The alignments of GISAID sequences were used exclusively for testing our predictions.

The random sequences in Fig. 2*B* are generated by randomly selecting a position and variant following a uniform distribution. For each GISAID sequence, a random sequence is produced with the same number of mutations to the reference.

Statistical Models. For each protein domain in the reference proteome we built an independent-site model (IND) and an epistatic model (DCA) using the previously described global MSA containing a diverged set of species but no SARS-CoV-2 sequences.

Independent or sequence profile model. Assuming statistical independence of positions, a simple probabilistic model $P_{IND}(a_1, a_2, \ldots, a_L)$, where $(a_1, a_2, \ldots, a_L)$ represents an aligned sequence of amino acids (with the gaps "—" to account for insertions or deletions) of length L, for a protein family is defined by

$$P_{IND}(a_1, \ldots, a_L) = \prod_{i=1}^{L} f_i(a_i) \ .$$

The factors $f_i(a)$ equal the empirical frequencies of amino acid a in column $i = 1, \ldots, L$ of the global MSA (with L columns). Therefore, the probability that any sequence of length L belongs to the protein family is factorized into the individual position-specific contributions of each of its amino acids. Similar to ref. 30, the effect of an amino acid mutation $a_i \to b$ can be computed as

$$\begin{aligned}\Delta E_{IND}(i, b) &= \log P_{IND}(a_1, \ldots, a_i, \ldots, a_L) - \log P_{IND}(a_1, \ldots, \mathrm{b}, \ldots, a_L) \\ &= \log f_i(a_i) - \log f_i(b).\end{aligned}$$

In contrast to previous work (30), positive values correspond to beneficial mutations while negatives correspond to deleterious mutations. Therefore, this value can be more naturally interpreted as a proxy of the selective pressure acting across coronaviruses.

DCA or epistatic model. It is possible to overcome the assumption of independence between positions by introducing two-site coupling terms as done in DCA models:

$$P_{DCA}(a_1, \ldots, a_L) = \frac{1}{Z} \exp\left(\sum_{1 \le i \le L} h_i(a_i) + \sum_{1 \le i < j \le L} J_{ij}(a_i, a_j) \right),$$

where Z is a normalization constant. The inference of model parameters is a computationally hard task and a number of approximations have been proposed (50–53). In this work, we rely on the widely used asymmetric plmDCA approach (52), which provides one of the best trade-offs between computational cost and performance. Following standard practice (50), we apply a sampling correction by counting the number of sequences with higher than 80% identity and reweighting them (results are robust to the specific value of this parameter; see *SI Appendix*, Fig. S11*B*). The number of effective sequences refers to the number of sequences that are not redundant at 80% sequence identity. As before, the effect of a single mutation $a_i \to b$ can be computed as the difference between a wild-type sequence and single-mutant sequence:

$$\Delta E_{DCA}(i, b) = \log P_{DCA}(a_1, \ldots, a_i, \ldots, a_L) - \log P_{DCA}(a_1, \ldots, b, \ldots, a_L).$$

As we focus on the mutability of each position in the SARS-CoV-2 proteome, for each of the models IND and DCA we derive a single mutational score $S_{IND/DCA}$ for each position i as

$$S_{IND/DCA}(i) = \frac{1}{q} \sum_{k=1}^{q} \Delta E_{IND/DCA}(i, b_k),$$

where $\Delta E(i, b_k)$ is the effect of the k th single mutations ($a_i \to b_k$) in position i. We restrict the set of amino acids to the ones reachable by a single nucleotide missense mutation from the corresponding codon in the Wuhan-Hu-1 reference genome (i.e., the alphabet size q depends on the specific codon used in position i). To make the quantification more interpretable and comparable between distinct domains, we divide the mutational score by the average score considering all the positions in the domain:

$$MS_{IND/DCA}(i) = S_{IND/DCA}(i) / \textstyle\sum_{j=1}^{L} S_{IND/DCA}(j).$$

This final mutational score is positive for beneficial mutations and negative for deleterious mutations. Values close to 0 can be interpreted as neutral, values in the range $(-1, 0)$ as better than average, and lower than -1 as worse than average and more deleterious.

Estimating Variability of SARS-CoV-2 Sequences from GISAID Data. The variability of each position was estimated by counting the number of sequences that have a different amino acid in the corresponding position compared to the reference. Only nonidentical sequences were considered to avoid the strong sequencing bias due to the highly diverse number of genomes sequenced in different countries. This corresponds to the standard reweighting procedure used in DCA but at a 100% similarity threshold adapted to the high sequence similarities between SARS-CoV-2 strains. Results are robust with respect to this procedure, since the empirical variability estimated without any reweighting shows, in the RBD, a Pearson correlation of 0.89 (Spearman correlation of 0.92) to the reweighted estimates.

DMS Data. The DMS data measuring protein expression and the binding to ACE2 obtained by Starr et al. (25) was collected from https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS trimmed to the RBD alignment (which contains 178 Pfam positions instead of the 201 in the experiment) and merged into our framework.

IEDB data. B and T cell epitope data were collected from the IEDB webserver by selecting Organism SARS-CoV-2 (ID: 2697049, SARS2) and restricting to B and T cells assay. For each protein of the SARS-CoV-2 proteome, a list of experimentally validated epitopes is provided. Following the definition of https://help.iedb.org/hc/en-us/articles/114094147751, it is possible to introduce an RF for each position i of the proteome. RF is defined as the number of positively responding subjects relative to the total number of those tested, averaged over all epitopes mapped to that position. Large values thus correspond to positions of high potential for immune escape.

The IEDB website only reports the upper and lower bounds of the 95% CI for the RF score—and not the RF score itself—to correct for the sample size. In our analysis we compute the mean RF score from the IEDB epitope data and use the 95% confidence upper and lower bounds provided by the IEDB to compute the confidence interval (CI = upper bound − lower bound) for each position.

Performance Evaluation. All ROC analyses were performed in R (ref. 54, version 3.6.3) using the package pROC (ref. 55, version 1.16.2). Controls and cases were defined by a variability cutoff parameter. Cutoffs for the variability, which define the subset of positions with low or high variability, were chosen to split into balanced subsets with the most similar number of observations possible in each subset, i.e., using the median. Positions with higher variability than the cutoff are considered positives.

Data Availability. To ensure reproducibility and access to our results we provide at https://giancarlocroce.github.io/DCA_SARS-CoV-2/ the data generated in the course of this research and a Jupyter notebook to reproduce key figures and guide data analysis. This notebook will also contain data updated as compared to the datasets used in this article. The code to generate the predictions for the IND and DCA models is available at https://github.com/juan-rodriguez-rivas/covmut. All other study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. We thank Erik Aurell, Matteo Bisardi, and David Gfeller for interesting discussions, David Gfeller in particular for his help with the immunologic data, and Richard Neher for his valuable feedback on our manuscript. Our work was partially funded by the Faculty of Science and Engineering of Sorbonne University in the context of the call SU-COVID19-FSI, by the EU H2020 Research and Innovation Programme MSCA-RISE-2016 under Grant Agreement 734439 InferNet and by the H2020 Marie Sklodowska Curie Individual Fellowship (H2020-MSCA-IF-2020), No. 101027973 to G.C.

1. N. G. Davies *et al.*; CMMID COVID-19 Working Group; COVID-19 Genomics UK (COG-UK) Consortium, Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, eabg3055 (2021).
2. L. Zhang *et al.*, SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat. Commun.* **11**, 6013 (2020).
3. M. Hoffmann *et al.*, SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. *Cell* **184**, 2384–2393.e12 (2021).
4. W. F. Garcia-Beltran *et al.*, Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* **184**, 2372–2383.e9 (2021).
5. D. Planas *et al.*, Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596**, 276–280 (2021).
6. R. K. Gupta, Will SARS-CoV-2 variants of concern affect the promise of vaccines? *Nat. Rev. Immunol.* **21**, 340–341 (2021).

7. S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **1**, 33–46 (2017).
8. B. Dearlove *et al.*, A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 23652–23662 (2020).
9. L. van Dorp *et al.*, Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020).
10. J. Hadfield *et al.*, Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
11. B. W. Matthews, Structural and genetic analysis of protein stability. *Annu. Rev. Biochem.* **62**, 139–160 (1993).
12. D. Shortle, W. E. Stites, A. K. Meeker, Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry* **29**, 8033–8041 (1990).
13. F. Ali, A. Kasry, M. Amin, The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutant. *Med. Drug Discov.* **10**, 100086 (2021).
14. B. Luan, H. Wang, T. Huynh, Enhanced binding of the N501Y-mutated SARS-CoV-2 spike protein to the human ACE2 receptor: Insights from molecular dynamics simulations. *FEBS Lett.* **595**, 1454–1461 (2021).
15. V. Dahirel *et al.*, Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11530–11535 (2011).
16. A. L. Ferguson *et al.*, Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* **38**, 606–617 (2013).
17. R. H. Y. Louie, K. J. Kaczorowski, J. P. Barton, A. K. Chakraborty, M. R. McKay, Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E564–E573 (2018).
18. D. K. Murakowski *et al.*, Adenovirus-vectored vaccine containing multidimensionally conserved parts of the HIV proteome is immunogenic in rhesus macaques. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2022496118 (2021).
19. E. Callaway, The coronavirus is mutating – Does it matter? *Nature* **585**, 174–177 (2020).
20. S. F. Ahmed, A. A. Quadeer, M. R. McKay, COVIDep: A web-based platform for real-time reporting of vaccine target recommendations for SARS-CoV-2. *Nat. Protoc.* **15**, 2141–2142 (2020).
21. M. Yarmarkovich, J. M. Warrington, A. Farrel, J. M. Maris, Identification of SARS-CoV-2 vaccine epitopes predicted to induce long-term population-scale immunity. *Cell Rep. Med.* **1**, 100036 (2020).
22. S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, M. Weigt, Inverse statistical physics of protein sequences: A key issues review. *Rep. Prog. Phys.* **81**, 032601 (2018).
23. W. Tai *et al.*, Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: Implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cell. Mol. Immunol.* **17**, 613–620 (2020).
24. L. Premkumar *et al.*, The receptor binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in SARS-CoV-2 patients. *Sci. Immunol.* **5**, eabc8413 (2020).
25. T. N. Starr *et al.*, Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310.e20 (2020).
26. H. Banoun, Evolution of SARS-CoV-2: Review of mutations, role of the host immune system. *Nephron* **145**, 392–403 (2021).
27. R. Vita *et al.*, The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* **47** (D1), D339–D343 (2019).
28. J. Mistry *et al.*, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49** (D1), D412–D419 (2021).
29. J. O. Wertheim, D. K. W. Chu, J. S. M. Peiris, S. L. Kosakovsky Pond, L. L. M. Poon, A case for the ancient origin of coronaviruses. *J. Virol.* **87**, 7039–7045 (2013).
30. M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, M. Weigt, Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.* **33**, 268–280 (2016).
31. T. A. Hopf *et al.*, Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
32. J. Trinquier, G. Uguzzoni, A. Pagnani, F. Zamponi, M. Weigt, Efficient generative modeling of protein sequences using simple autoregressive models. *Nat. Commun.* **12**, 5800 (2021).
33. J. Lan *et al.*, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
34. R. V. Eck, M. O. Dayhoff, Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences. *Science* **152**, 363–366 (1966).
35. A. J. Riesselman, J. B. Ingraham, D. S. Marks, Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
36. A. J. Greaney *et al.*, Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463–476.e6 (2021).
37. F. Konings *et al.*, SARS-CoV-2 variants of interest and concern naming scheme conducive for global discourse. *Nat. Microbiol.* **6**, 821–823 (2021).
38. A. Rambaut *et al.*, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
39. A. A. Quadeer, S. F. Ahmed, M. R. McKay, Landscape of epitopes targeted by T cells in 852 individuals recovered from COVID-19: Meta-analysis, immunoprevalence, and web platform. *Cell Rep. Med.* **2**, 100312 (2021).
40. S. Vilar, D. G. Isom, One year of SARS-CoV-2: How much has the virus changed? *Biology (Basel)* **10**, 91 (2021).
41. H.-L. Zeng, V. Dichio, E. Rodríguez Horta, K. Thorell, E. Aurell, Global analysis of more than 50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 31519–31526 (2020).
42. C. Fenwick *et al.*, A highly potent antibody effective against SARS-CoV-2 variants of concern. *Cell Rep.* **37**, 109814 (2021).
43. P. Nosil *et al.*, Natural selection and the predictability of evolution in Timema stick insects. *Science* **359**, 765–770 (2018).
44. M. Lässig, V. Mustonen, A. M. Walczak, Predicting evolution. *Nat. Ecol. Evol.* **1**, 77 (2017).
45. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu; UniProt Consortium, UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
46. S. Phadke, S. Macherla, R. H. Scheuermann, "Database and analytical resources for viral research community" in *Encyclopedia of Virology*, D. Bamford, M. Zuckerman, Eds. (Elsevier, 2021), pp. 141–152.
47. J. R. Brister, D. Ako-Adjei, Y. Bao, O. Blinkova, NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–D577 (2015).
48. E. L. Hatcher *et al.*, Virus variation resource – Improved response to emergent viral outbreaks. *Nucleic Acids Res.* **45** (D1), D482–D490 (2017).
49. S. R. Eddy, Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
50. F. Morcos *et al.*, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).
51. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009).
52. M. Ekeberg, T. Hartonen, E. Aurell, Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.* **276**, 341–356 (2014).
53. D. T. Jones, D. W. A. Buchan, D. Cozzetto, M. Pontil, PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).
54. R Core Team, R: A language and environment for statistical computing (Version 3.1, R Foundation for Statistical Computing, Vienna, 2020).
55. X. Robin *et al.*, pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).