scientific reports

Check for updates

OPEN

Variant-driven early warning via unsupervised machine learning analysis of spike protein mutations for COVID-19

Adele de Hoffer[1,2,14], Shahram Vatani[3,4,14], Corentin Cot[5,14], Giacomo Cacciapaglia[3,4]✉, Maria Luisa Chiusano[6,7], Andrea Cimarelli[8], Francesco Conventi[9,10], Antonio Giannini[11], Stefan Hohenegger[3,4] & Francesco Sannino[2,9,12,13]✉

Never before such a vast amount of data, including genome sequencing, has been collected for any viral pandemic than for the current case of COVID-19. This offers the possibility to trace the virus evolution and to assess the role mutations play in its spread within the population, in real time. To this end, we focused on the Spike protein for its central role in mediating viral outbreak and replication in host cells. Employing the Levenshtein distance on the Spike protein sequences, we designed a machine learning algorithm yielding a temporal clustering of the available dataset. From this, we were able to identify and define emerging persistent variants that are in agreement with known evidences. Our novel algorithm allowed us to define persistent variants as chains that remain stable over time and to highlight emerging variants of epidemiological interest as branching events that occur over time. Hence, we determined the relationship and temporal connection between variants of interest and the ensuing passage to dominance of the current variants of concern. Remarkably, the analysis and the relevant tools introduced in our work serve as an early warning for the emergence of new persistent variants once the associated cluster reaches 1% of the time-binned sequence data. We validated our approach and its effectiveness on the onset of the Alpha variant of concern. We further predict that the recently identified lineage AY.4.2 ('Delta plus') is causing a new emerging variant. Comparing our findings with the epidemiological data we demonstrated that each new wave is dominated by a new emerging variant, thus confirming the hypothesis of the existence of a strong correlation between the birth of variants and the pandemic multi-wave temporal pattern. The above allows us to introduce the epidemiology of variants that we described via the Mutation epidemiological Renormalisation Group framework.

It is of primary importance to understand the diffusion of a virus and the establishment of its variants, to further understand the infection mechanisms and fight the associated disease, especially in view of an efficient

[1]Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy. [2]Scuola Superiore Meridionale, Largo S. Marcellino 10, 80138 Naples, Italy. [3]Institut de Physique des 2 Infinis (IP2I), UMR5822, CNRS/IN2P3, 69622 Villeurbanne, France. [4]Université de Lyon, Université Claude Bernard Lyon 1, 69001 Lyon, France. [5]Laboratoire de Physique des 2 Infinis Irène Joliot Curie (UMR 9012), CNRS/IN2P3, 15 Rue Georges Clemenceau, 91400 Orsay, France. [6]Department of Agricultural Sciences, Università degli Studi di Napoli Federico II, 80055 Portici, Italy. [7]Department of Research Infrastructures for Marine Biological Resources (RIMAR), Stazione Zoologica "Anton Dohrn", 80121 Naples, Italy. [8]Centre International de Recherche en Infectiologie (CIRI), Inserm, U1111, CNRS, UMR5308, ENS de Lyon, Univ Lyon, Université Claude Bernard Lyon 1, 46 Allée d'Italie, 69007 Lyon, France. [9]INFN Sezione di Napoli, Complesso Universitario di Monte S. Angelo Edificio 6, Via Cintia, 80126 Naples, Italy. [10]Dipartimento di Ingegneria, Università degli studi di Napoli Parthenope, Centro Direzionale di Napoli, Isola C 4, lato Sud, 80143 Naples, Italy. [11]University of Science and Technology of China (USTC), No.96, JinZhai Road, Baohe District, Hefei 230026, Anhui, China. [12]Dipartimento di Fisica E. Pancini, Università di Napoli Federico II, Complesso Universitario di Monte S. Angelo Edificio 6, Via Cintia, 80126 Naples, Italy. [13]CP3-Origins and the Danish Institute for Advanced Study, University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark. [14]These authors contributed equally: Adele de Hoffer, Shahram Vatani and Corentin Cot. ✉email: g.cacciapaglia@ipnl.in2p3.fr; sannino@cp3.sdu.dk

vaccination campaign. This task could not be efficiently approached in past extended pandemics caused by infectious diseases, like for instance the "Spanish" Influenza of 1918–1919[1], mainly due to the paucity of the available data. The current COVID-19 crisis is, on the contrary, revolutionising our understanding of pandemics because of the efficient collection of a large amount of data (e.g. genome sequencing, epidemiology, etc) in real time, which allows for a timely identification of viral variants that successfully radiate throughout the world. Among the mutations that characterise SARS-CoV-2 variants, those that can be traced along the spike protein (S) sequence are major, although not unique, drivers of viral spread in the human population for the role that this protein plays in mediating the virus entrance into target cells as well as for its role in mediating escape from antibody responses. For the above reasons, in this work we apply and validate the hypothesis that mutations on the SARS coronavirus Spike protein are sufficient to identify the emergence of new variants of epidemiological relevance, which can have dominant diffusion within the infected population. In this work, a *mutation* is a single change in the amino acid sequence of the Spike protein (substitution, addition, deletion), while a *Spike variant*, or simply *variant*, is a unique sequence of amino acids in the Spike protein that appears in clusters. Like other coronaviruses, the SARS-CoV-2 has relatively low mutation rates[2], nevertheless the current COVID-19 pandemic has seen the emergence of several epidemiologically relevant variants. Efficient nucleotide sequencing has allowed to track sequence mutations along the genome of SARS-CoV-2, and to identify dangerous variants[3,4] that appeared to increase the infectivity compared to the initial form that was sequenced from the outbreak in Wuhan, China[5] (GenBank: MN908947.3). Since the second half of 2020, variants of concern (VoCs) and of interest (VoIs) have been identified in various regions of the world. For instance, following the naming scheme of the WHO[6] (Pango lineage[7], GISAID[8,9]): The Alpha VoC (B.1.1.7, GRY), first identified in September 2020 in the UK[10,11]; the Beta VoC (B.1.351, GH/501Y.V2) first found in South Africa in May 2020[12]; the Gamma VoC (P.1, GR/501Y.V3) first detected in Brazil in November 2020[13], which has been spreading in Manaus notwithstanding the high rate of previous infections; the Delta VoC (B.1.617.2, G/478K.V1) identified in India in October 2020; and the Epsilon VoI (B.1.427+429, GH/452R.V1) found in California in March 2020[14]. An exhaustive list can be found on the WHO website (www.who.int). Considering the Alpha VoC as an example, it has been possible to study its infectious power in lab experiments, finding a higher rate of transmission by 67–75%, compared to the previous ones[11]. The transmission advantage has been confirmed by epidemiological data in the UK[15,16]. Most analyses of the epidemiological data are done applying the time-honoured compartmental models of the SIR type[17–19], appropriately extended by including more compartments[20]. The main drawback in this approach is the large number of parameters, which need to be fixed by hand or extracted from the data. In this work, we bypassed this bottleneck by using a simplified and effective approach based on theoretical physics methods, the epidemic Renormalisation Group (eRG) framework[21–23], combined with information directly extracted from the Spike protein sequence via a simple Machine Learning approach. This novel method allowed us to analyse, at the same time, the variability of the SARS-CoV-2 Spike protein in multiple countries and regions of the world, and thus provide a direct comparison of the epidemiological impact of the different Spike variants. A theoretical analysis of the variants within the eRG framework is presented in a companion publication[24].

In the present work, we analysed the protein sequence data for the UK nations downloaded from the GISAID repository[8,9]. We implemented a simple Machine Learning (ML) algorithm based on the Levenshtein measure (LM)[25,26] in order to cluster protein sequences based on their distance in terms of number of amino acid substitutions (i.e. the number of amino acid mutations needed to transform one sequence into the other, or vice-versa). The clusters have been defined by setting cutoffs on the Ward distance between branches of a proximity tree, built by the use of a standard hierarchical clustering algorithm. We applied the clustering algorithm on the data binned in temporal units, specifically here identified by months or weeks. Each time unit may include more clusters according to the cutoffs set on the Ward distance and, within each cluster, the dominant variant is the most frequent in terms of identical sequences over the total number of sequences in the cluster. We then developed an algorithm that links clusters appearing in consecutive time units and creates chains of clusters that share the same dominant Spike variant. Empirically we determined that chains that persist for longer than three time units identify emerging variants. In order to reconstruct the origin of each emerging variant, we associated the initial cluster of each chain with a cluster from the previous time unit that maximises the overlap in their sequence content. Hence, a branching relation emerges in our procedure. For clarity, in this work we use the following definitions for variants and mutations:

1. A *dominant variant* in a cluster is the Spike variant that is most frequently appearing in the cluster. Note that the chains are created by linking consecutive clusters when they possess the same dominant variant.
2. An *emerging variant* is defined as an established chain that contains more than three consecutive clusters, defined using our linkage algorithm. This criterium is established empirically from the results of the chain reconstruction.

It should be noted that some of the emerging variants defined by our procedure can be associated to VoCs and VoIs, as defined by the WHO, as they share the same characteristic Spike mutations. Unlike phylogenetic analyses plus Bayesian inference methods[27,28], the proposed ML method is unsupervised and does not rely on any statistical model (nor any prior) to describe the growth and evolution of a lineage. Furthermore, the algorithm is data driven and does not require Monte Carlo simulations for calibration. The time evolution emerges as a natural consequence of the hierarchical clustering algorithm procedure, and the emerging variant (defined as a stable time-ordered sequence of clusters) is only subsequently compared to VoCs and VoIs defined by WHO, which we take as an example. The method is computationally very efficient (a few hours for the entire UK dataset on 144 core Intel(R) Xeon(R) Gold 5220 CPU at 2.20GHz CPU with 500 Gb RAM) and allows to quickly identify new potentially dangerous variants and to reconstruct relevant spike mutation dynamics within stable chains.

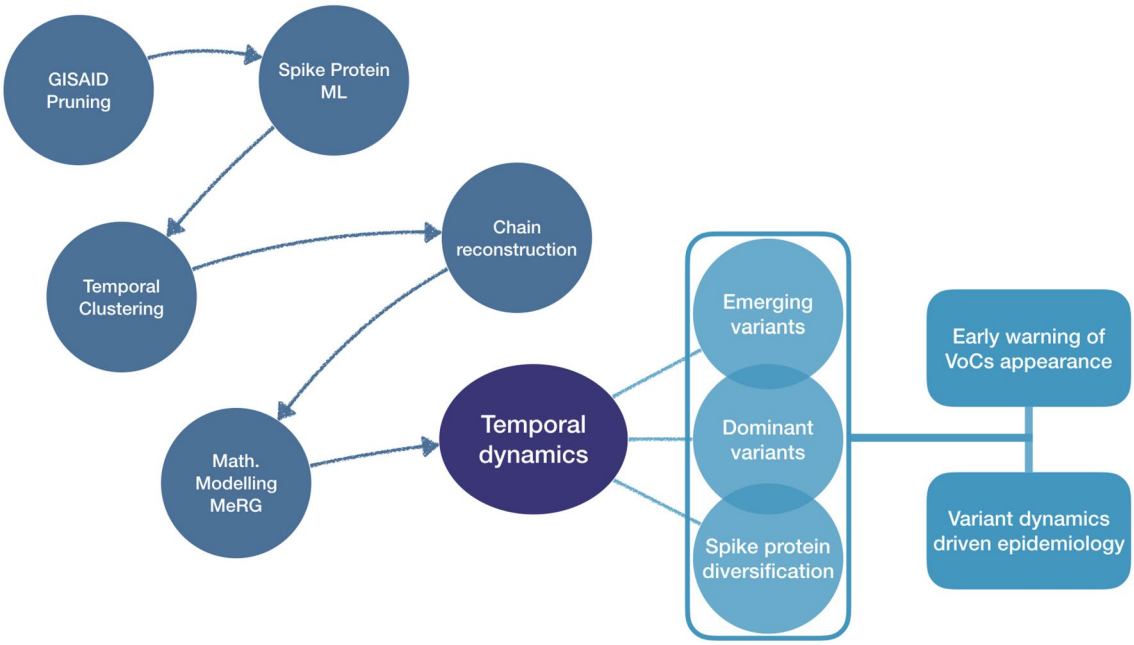

Figure 1. Methodology and main outcomes. Schematic representation of the work-flow we follow in this work. The figure is read from left to right, with the blue circles summarising the main steps of our investigation, with leads to a reconstruction of the temporal dynamics of the diffusion of the variants (dark blue oval). The arrows show their logical sequence. Intermediate results are assembled in the vertical box. In the light blue boxes on the right, we summarise the two main applications of the results: the early warning tool for the appearance of relevant variants and the reconstruction of the variant epidemiological dynamics.

The procedure above has been independently repeated for each geographical region in our study. We validated our results by showing that our approach identifies the Alpha VoC, independently, in all the distinct UK regions we studied. Once the dominant variants were identified, we analysed their temporal spreading within the affected population. Given that only a small fraction of the infected individuals have their viral charge sampled and sequenced, we estimated the number of people infected by each variant by multiplying the number of positive tests by the rate of occurrence of each variant in the sequencing data. This rough approximation allows us to reliably extract the temporal evolution of each variant in the population. Note that each infected individual is, in practice, associated to the variant that is most frequently reappearing in their viral charge, following the practice of the sequence reporting. Thus, the data we use track the time development of the dominance of each variant at the individual level.

To analyse the time evolution of the individuals infected by each variant, we employed the economical eRG approach[21] that allows to organise the pandemic waves according to temporal symmetry principles similar to those found in high energy physics[29,30]. The approach has been extensively tested[23,31], shown to be equivalent to traditional SIR compartmental models with time-dependent parameters[32], and, last but not least, summarised in a comprehensive review alongside other approaches[33]. The economy of the model rests in the fact that, once the overall number of infected individuals is fixed, the diffusion rate of the virus is captured by a single parameter γ that measures the speed at which the virus spreads in the population. This value can be extracted by fitting the number of new daily infections or the cumulated number of infections. The value of γ contains not only the infectivity of the virus variant, but also the effect of pharmaceutical and non-pharmaceutical interventions, as well as the response of the population. It has been established[21,32] that a constant γ correctly describes the time-evolution over the time-scale of a single wave: this is due to the fact that social effects, like the decrease in people's mobility, affect the diffusion of the virus with a delay of a few weeks[34], while vaccinations administered during a wave have minor impact[31,35] . This is a special property of the eRG approach, to be contrasted to time-honoured compartmental models, of the SIR type, where a time-dependent reproduction number $R_0(t)$[32] is necessary to capture the correct effects. The effects of non-pharmaceutical interventions have been studied in detail within more traditional mathematical modelling[36–39].

A visual summary of the methodology followed by our analysis with its main outcomes is shown in Fig. 1, while more details are reported in the Supplementary material.

The main goal of this work is to understand the viral dynamics that characterises wave patterns stemming from infectious diseases like COVID-19. The eRG approach additionally offers a natural mathematical understanding in terms of the dynamical flow of the system[40,41]. Importantly, by employing ML analysis to genomic data, we discovered that each pandemic wave is driven by a single emerging and persisting variant. The findings demonstrate that the variant dynamics is one of the main engines behind the emergence of wave patterns for COVID-19. This result can be used as a template for similar infectious diseases. As direct consequence of our study we propose a novel evolutionary model for the interpretation of the virus diffusion that is mutation driven.

Results

Spike protein sequences have been extracted from the GISAID repository on a country-specific basis and the date-stamp associated to each sequence has been used to obtain a temporal dimension of viral variants appearance. Note that each genome sequence in the GISAID data collection corresponds to the most frequent Spike variant occurring in a single infected individual. The pruned dataset (see Supplementary material) has been clustered per month to obtain groups of distinct variants. Firstly, we computed the LM between each pair of sequences, thus counting by an unweighted approach the minimal number of amino acid substitutions, deletions and insertions needed to transform one sequence into the other, and vice-versa. Secondly, the algorithm constructed a tree of proximity by pairing sequences that are the closest to each other into a branch. To combine branches that contain more than one sequence, we used Ward's method, after having checked that other choices do not significantly affect the results (more details in the Supplementary material). The tree is completed when all sequences are grouped into a single branch. To define the clusters, we considered a cutoff in the distance so that branches whose Ward distance is larger than the cutoff are considered as separate clusters. We applied the same cutoff to all branches. The clustering procedure is applied to sequencing data binned in time, where the duration of each bin depends on the available data per day during the period of interest. The procedure depends on two parameters, the cutoff and a threshold in the size of each cluster, which can be tuned by optimising the dataset coverage and number of clusters, without any prior knowledge of the variants. Henceforth, emerging variants are defined as persisting chains of clusters, sharing the same dominant Spike variant.

As England has the largest available sequencing sample, with 646.697 sequences as of the end of August 2021, we mainly focused on this dataset. This minimises statistical and sampling bias errors. After pruning, 461.122 sequences were retained, out of which we identified 13.887 distinct ones.

Emerging variants as time-ordered cluster chains. The time evolution and emergence of SARS-CoV-2 variants can be studied by applying our ML algorithm to the Spike sequence data binned in time, by calendar month. Hence, we have divided the sequence dataset for England following the date tag in the GISAID repository. For each month, we run the ML algorithm on the pruned data to define clusters, retaining only the ones comprising at least 1% of the monthly dataset. The cutoff on the Ward distance r_W between branches, as well as the 1% threshold above, were chosen to optimise the coverage of the dataset (i.e. we required that the defined clusters cover at least 90% of the data) while keeping the number of clusters below 10. The optimisation analysis, presented in the Supplementary material, showed that the optimal range for the branch cutoff is $r_W \in [50, 200]$. After this, we compared the clusters in consecutive months to link those with a "strong similarity", i.e. those that share the same dominant sequence (strong links). More details on this procedure and its validation can be found in the Supplementary material. The linkage algorithm we employed allowed to define "chains of clusters" that we associate to emerging variants. The results are shown in Fig. 2 for two choices of the Ward distance: $r_W = 100$ in the left and $r_W = 200$ in the right plots. For the two choices, we identified 6 and 4 cluster chains, respectively, that last more than 3 months. In the middle and bottom rows of Fig. 2 we show the new monthly infections (per 100k inhabitants) and the frequencies of the cluster chains, which we identify as emerging variants in the following. In this respect, the results for $r_W = 200$ in Fig. 2B can be directly compared to the VoCs identified by the WHO: Comparing the frequencies of occurrence in Fig. 2D, we see that v2 can be associated to the Alpha VoC, while v3 matches the epidemiological data for the Delta VoC. We also checked that the dominant Spike variant for the two chains presents the mutations characteristic of the two VoCs: N501Y, D614Y and P681H for the Alpha VoC; L452R, T478K, D614G and P681R for the Delta VoC. These results have been corroborated by a cluster analysis of the global dataset, without time binning, and by a similar analysis for the data of Wales and Scotland, as shown in the Supplementary material.

The chain analysis, however, allowed us to better probe the time evolution and emergence of the persisting variants. To do so, for the clusters at the beginning of each chain, we defined a branching link with the cluster in the previous month. These connections are shown as grey diagonal links in the top plots of Fig. 2. From the case $r_W = 200$ in Fig. 2B, we clearly see that v1, which is responsible for the second wave, branched off from v0 in October 2020. Similarly, v2, which corresponds to the Alpha VoC, also branched off from v0 a month later. The Delta VoC v3, instead, developed from v1 from February to May 2021, via two intermediate clusters, 20 and 22. Finally we see the emergence of a branch, 20–23–26, which died off being dominated by the Delta VoC starting with cluster 25. By lowering the cutoff that defines clusters, see Fig. 2A for $r_W = 100$, one can see how v1 splits in two distinct, but closely related, chains, as well as the Delta VoC v3. The Delta VoC is now seen as branching off from v1b. The closeness of the clusters splitting from v1 and v3 is confirmed by comparing the dominant Spike sequences, showing that v1b differs from v1a only by the mutation L18F, while v3b differs from v3a only by the mutation T95I. In particular, cluster 43 emerged in August 2021 and its dominant variant bears the Y145H and A222V mutations that identify the AY.4.2 lineage ('Delta plus' variant)[42], which has been classified by Pango[7] at the beginning of September. Out of the many new lineages that have been recently isolated, only this one is highlighted by our ML analysis. As such, and with the caveat that our analysis includes only data up to August 2021, the ability of this novel variant to give raise to a stable chain in the near future deserves close attention.

These results firstly show that the phylogenetic relation between variants emerges from our simple ML algorithm applied exclusively to the Spike protein sequence. Furthermore, we see a distinctive pattern relating the emergence of a persistent variant and the exponential increase in infections that ignites a new pandemic wave. A new wave only emerges when a new variant is generated, which has the virological strength to overcome the old ones. This is seen very clearly with v2 (or Alpha VoC) which spins off from v0 closely to v1 and takes over by generating a third wave. We also see the emergence of short-lived variants that do not have the power to start a new wave and therefore die off without infecting a sizeable number of individuals. All short-lived chains have less than two clusters, hence we define a minimum length of three for persisting chains.

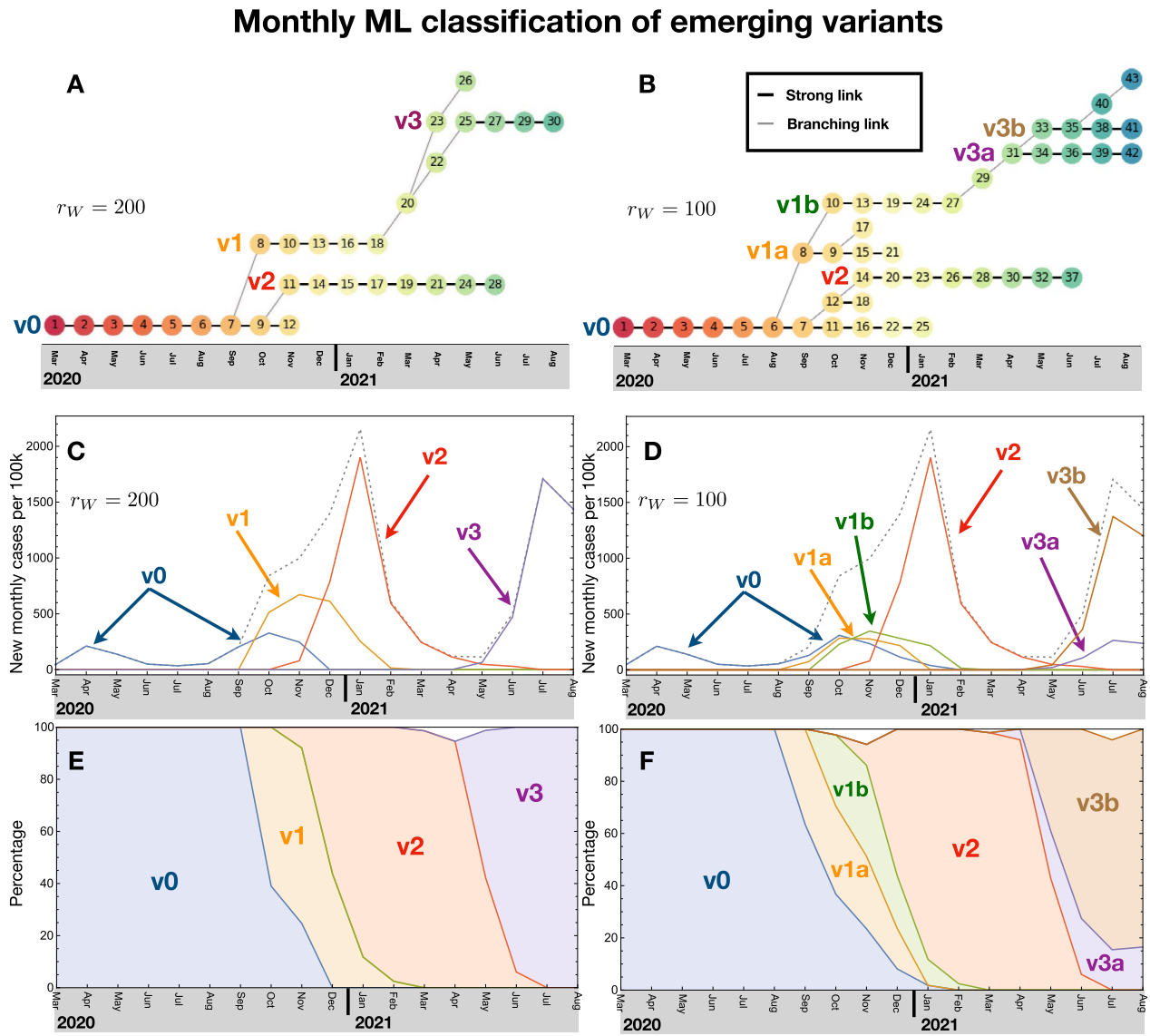

Figure 2. Monthly ML analysis and chain variants. The clusters are linked to form chains, which are then identified with emerging variants, as shown in the top plots (**A**,**B**). In the middle (**C**,**D**) and bottom (**E**,**F**) plots we show the number of monthly infected per 100k inhabitants and percentage of occurrence for each emerging variant. The shaded regions represent the relative percentage of the variants present at each time. The left plots (**A**,**C** and **E**) correspond to a cutoff in the Ward distance of $r_W = 200$ while the right ones (**B**,**D** and **F**) to $r_W = 100$. Note that the chains v2 and v3 for $r_W = 200$ can be associated to the Alpha and Delta VoC, respectively.

Performance of the ML algorithm as an early warning tool for emerging variants. The results for time-ordered chains with monthly clustering (Fig. 2) demonstrate that our ML algorithm is able to efficiently identify the emergence of new variants that have strong impact on the epidemiological evolution of the disease. Hence, it can be used as an early warning tool for new potentially dangerous variants that may become VoCs. To test the performance of this tool, we validated the procedure on the emergence of the Alpha VoC. The first Alpha VoC case has been found on the 20th of September, 2020. Following the monthly analysis, we identified a cluster dominated by the Alpha VoC in November 2020, at the beginning of chain v2. Once the first cluster is identified, one would need to add the data for the following months to confirm its persistence (as the process is additive, the cluster definitions in the previous months are not affected). We saw empirically from Fig. 2 that persisting chains contain at least three clusters, hence an emerging variant could be defined only 2 months later (January 2021 for the Alpha VoC).

To improve the performance of the ML algorithm in terms of prediction, we reanalysed the same data using a weekly—rather than monthly-based binning (Table 1). The cutoff on the Ward distance needs to be adjusted for the weekly analysis, thus leading to a choice that differs slightly from that of Fig. 2. Using a cutoff $r_W = 95$ (instead of 100) as well as the 1% threshold, the Alpha VoC chain was identified as branching off in week 44, at

Cal. week	Date (Mon)	Total seq.	% of Alpha VoC	No. of clusters	
38	14 Sept.	1948	0.05	–	First detection
39	21 Sept.	3394	0.06	–	
40	28 Sept.	2203	0.09	5	
41	5 Oct.	3891	0.1	5	
42	12 Oct.	4598	0.07	5	
43	19 Oct.	5921	0.4	2	
44	26 Oct.	4557	1.5	1	First warning (weekly)
45	2 Nov.	7589	3	1	
46	9 Nov.	7200	7	1	
47	16 Nov.	4669	12	1	Emerging persistent variant (weekly)
48	23 Nov.	2343	12	1	
49	30Nov.	1971	21	1	First warning (monthly)
50	7 Dec.	6382	38	1	
51	14 Dec.	8059	50	1	
52	21 Dec.	4864	53	1	
53	28 Dec.	7766	65	1	WHO classification as VoC

Table 1. Early warning for the Alpha VoC. Results of the ML analysis applied to weekly binned data after the first detection of the Alpha VoC in England. The columns contain the 2020 calendar week number, with the initial date (Monday), the total number of sequences in the dataset for each week and the percentage of Alpha VoC sequences identified a posteriori in the data. The fifth column contains the minimal number of clusters in the ML output that allows to isolate the Alpha VoC cases. When this indicator is equal to 1, an early warning can be issued (week 44). After 3 weeks of the cluster persisting, we identify it with an emerging variant (week 47). This date can be compared to the WHO classification decision (29 Dec., week 53).

a moment when this variant represented 1.5% of the total weekly sequences, as illustrated by the top diagram in Fig. 3A. By week 46 (stable chain of at least three clusters) this analysis would have been able to identify the Alpha variant as VoC by the 9th of November. We next analysed the same data by lowering both cutoff and threshold to determine whether this could influence the performance of the ML algorithm. Using less stringent parameters it is possible to define separate clusters containing the Alpha VoC sequences earlier than week 44 (Fig. 3A lower two panels, highlighted by a red arrow). However, in these cases the analysis yields to an increase in the number of clusters and chains, which makes it difficult to identify unequivocally new emergent variants. To quantify the performance, therefore, we counted how many additional chains appear when an Alpha VoC chain can be isolated at each week, as shown in the fifth column of Table 1. For all weeks after 44 included, it suffices to generate one new cluster besides the two that contain the dominant Spike variant of v0 and v1, while for week 43 at least two new clusters are necessary. The inverse of the number of clusters defined above quantifies a "confidence indicator" (C.I.) for the early warning, and it is plotted in Fig. 3B as a function of the Alpha VoC percentage in each week dataset. The C.I. can be interpreted as the probability of identification of the correct emerging variant via the first cluster. The result shows that probabilities above 50% require the presence of the new variant in at least 1% of the sequences. In Fig. 3C, we show the Alpha VoC percentage as function of time in weeks. An early warning can be issued as soon as the new variant surpasses 1% of the data, leading to an early warning for the Alpha VoC in week 44, i.e. 6 weeks after the first detection. If the chain persists for 3 more weeks, an emerging variant is identified in week 47, hence 6 weeks before the official classification as a VoC by WHO. It is possible to reduce the time-scale below a week by increasing the sequencing. To obtain a result with statistical uncertainty below 10%, for instance, a few thousand sequences in each time bin would be required. For the Alpha VoC data, this is achieved for weekly binning, as shown in Table 1 (the statistical error is shown as a band in Fig. 3C).

To test if these conclusions are general, we performed the same weekly analysis at the onset of the Delta variant, which started spreading in the UK in May 2021. For this time period, the number of weekly sequences available on GISAID amounts to a few thousands, hence offering a situation similar to that of the Alpha VoC onset. This allows for a fair comparison between the two analyses. We found that, in both cases, the VoC can be uniquely identified by setting the r_W cutoff around 100 and a threshold of 1%, confirming that the parameters of the clustering algorithm do not depend on the specificities of the variant. We also computed the percentage of the VoC Spike variant in the time bin where we obtain a C.I. of 0.5, i.e. a 50% probability of identifying the emerging variant: the result is $1.0 \pm 0.5\%$ for Alpha and $3.0 \pm 0.5\%$ for Delta, where the error is statistical. The larger value for Delta is due to higher transmissibility of this VoC with respect to the Alpha (see Supplementary material for more details). Henceforth, we can conclude that a reliable identification of a new emerging variant can be obtained if such variant reaches a few % of the time-binned sequences, where larger fractions are needed for rapidly increasing variants.

At the moment of submission of this work, there is an increasing coverage in the media about the possible concerns due to the raise of a new variant in the UK (AY.4.2 Pango lineage). Our monthly analysis already identified this lineage as a dominant Spike variant in cluster 43 (Fig. 2), branching off from the main Delta chain in August 2021 (date at which data acquisition was stopped for the purposes of this work). To determine whether

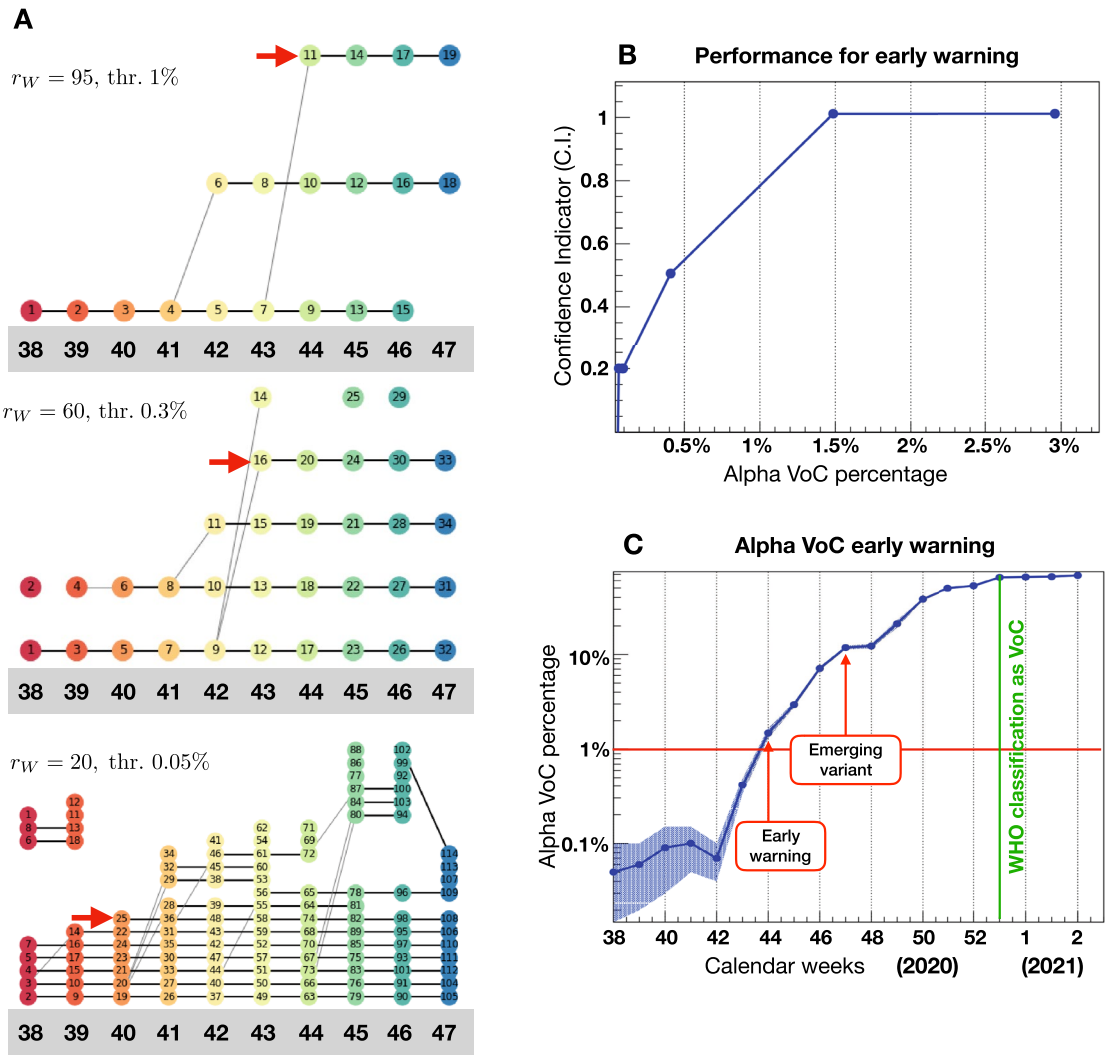

Figure 3. Early warning performance. (**A**) Cluster chains from week 38 to week 47 obtained at different working points. The chain containing the Alpha VoC Spike variant is highlighted by the red arrow. (**B**) Confidence indicator for the early warning performance as a function of the percentage of the new variant in the time-binned data. (**C**) Data for the Alpha VoC, indicating the early warning and definition of the emerging variant, using our ML on a weekly binning, compared to the time of the WHO classification as VoC.

this novel variant could be considered of concern, we carried out our ML analysis with a weekly binning. Despite the fact that this analysis was carried out with a suboptimal $r_w = 100$ (due to the urgency of the situation analyses to determine the optimal r_w are ongoing), our weekly analysis clearly indicates that AY.4.2 Pango lineage has formed a stable chain of 3 clusters by the19th of September, 2021. This analysis thus indicates that this variant is truly establishing in the UK as a variant of concern.

Features of Spike mutation dynamics within stable chains. This method allows for a temporal analysis of the accumulation of viral Spike variants at different resolution, according to the chosen r_w. The $r_w = 100$ analysis, for example, highlights how the Alpha variant (v2 stable chain comprised between clusters 14 and 37) arouse from cluster 7 (v0 stable chain of clusters 1–25) through an ephemeral intermediate cluster (cluster 12). The mapping of mutations that appeared in both dominant and subdominant variants in each cluster over time on the cryo-EM trimeric structure of the Spike protein[43] is of interest (Fig. 4A,B). In addition to the mutations fixed in the dominant Spike variant, mutations in subdominant ones (detected at equal/above 1% of the entire pool of sequences in each cluster, and marked by cyan dots in the 3D spike models in Fig. 4C) accumulate with increasing frequency along each chain. This is not surprising per se and it likely reflects high viral replication rates over time at the population level. While so far mutations that become fixed in the dominant variant appear and reach dominance within 1 month, the fixation of the H69/V70 deletion seems to have undergone a transient sampling state with the N439K mutation in the RBM. Viruses bearing the N439K Spike mutation have been characterised ex vivo and the mutation had been described to allow for antibody-mediated immunity escape while not affecting viral fitness[44]. Given that this mutation rapidly vanished (non-productive chain 12–18), it is clear that such mutation exhibited a defect in viral fitness that cannot be recapitulated with ex vivo studies.

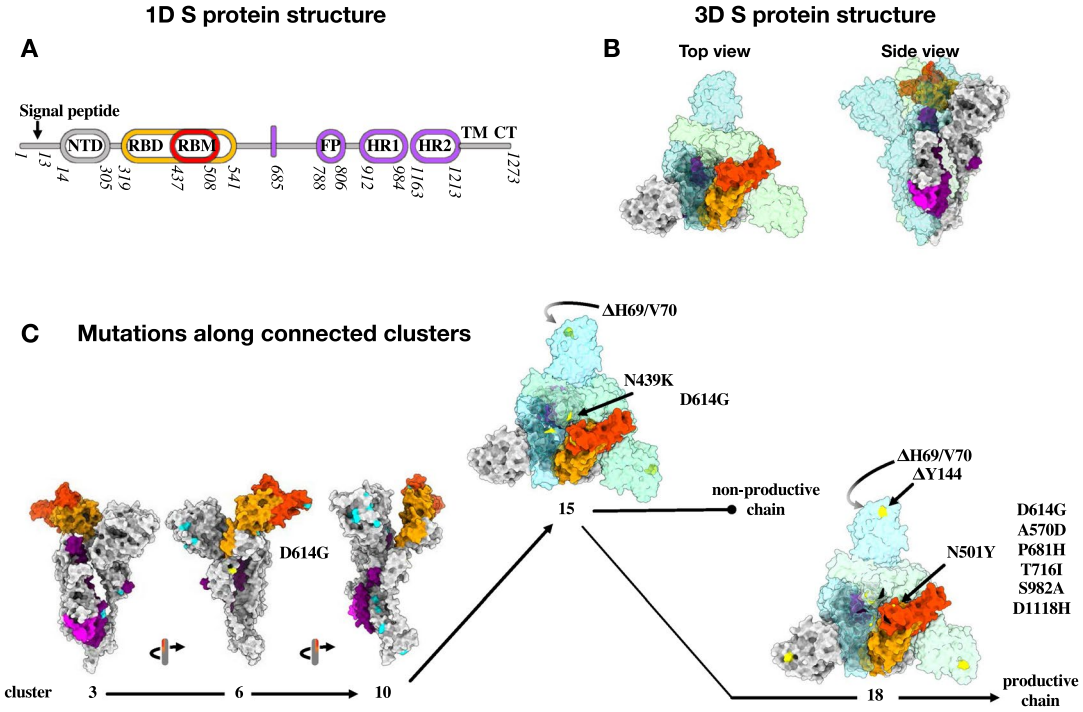

Figure 4. Temporal analysis of mutations arising in the Spike protein during the genesis of the Alpha VoC. (**A**) Schematic representation of the SARS-CoV-2 Spike protein (S): N terminal domain (NTD); receptor binding domain and motif (RBD and RBM, respectively); fusion peptide (FP); heptad repeat 1 and 2 (HR1 and HR2, respectively); transmembrane domain (TM) and cytoplasmic tail (CT). (**B**) Top and side view of the trimeric Spike protein in its closed conformation[43] (PDB: 6ZGI). For simplicity, the colour codes of the different domains are provided for a single chain of the trimer. (**C**) Accumulation of the different mutations in the different clusters leading to the establishment of the Alpha VoC: domains colour codes as in (**A**) for a single S; yellow indicates mutations fixed in the dominant variant and cyan indicates mutations appearing in subdominant ones. In cluster 14 ($r_W = 100$), the original position of the N439K mutation that was lost to the profit of the near N501Y mutation is marked in black.

However, the N439K mutation may still have served an important role in the emergence of the subsequent dominant variant by allowing sufficient time for variants bearing the H69/V70 deletion to combine with more advantageous mutations (N501Y, DY144 etc).

In Fig. 5 we also show the pattern of diversification occurring along the Spike protein in terms of single amino acid substitutions in the chains reconstructed with the cutoff $r_W = 100$, i.e. v0, v1a, v1b, v2, v3a and v3b. The heatmaps in Fig. 5A associated to each main chain show the position and the number of amino acid changes with respect to the previous cluster in the chain (see Supplementary material for more details). The picture shows where and how often a change occurs within the chain along time and offers an overview of these dynamics. Our time-ordered analysis along each chain allows to distinguish the extent of Spike mutations within stable chains. For instance, along v0 one can see higher variability in clusters 6 and 7, corresponding to when v1a/b and v2 branched off (see Fig. 2), and in the last two clusters before the variant disappeared from the data. Furthermore, we observe that v2 has a lower degree of variability compared to the other 5 chains. This could reflect the fact that v2 did not lead to new stable chains, while the other ones did, including those associated to the Delta VoC. This preliminary overview surely deserves further investigations, while it is here reported to highlight the additional level of details that a temporal variant analysis of the Spike protein offers.

As a final remark, our analysis also permits to observe the most conserved or variable regions per chain (regions were changes are minor or not occurring and regions with frequent substitutions). Interestingly, the comparison among the plots also shows that, although there are main hot spots of mutations that are detectable along the spike protein and along all the chains, each chain has a typical pattern of substitution. Considered altogether (see Fig. 5B), this analysis can allow us to identify regions of the Spike protein that also provide hints for more efficient targeting in monitoring or pharmaceutical interventions.

Epidemiological data and MeRG. The results of our ML analysis firmly suggest that there is a strong relation between the genesis of a new emerging variant and the onset of a new wave, with exponential increase in the number of infections, in the epidemiological data. In a companion article[24] we developed a framework that can be used to describe the evolution of each variant. The model is based on the eRG approach by including mutations (MeRG).

The MeRG framework models the time evolution of the cumulated number of infected by each variant in terms of a logistic function (sigmoid), solution of the eRG equation, and given by:

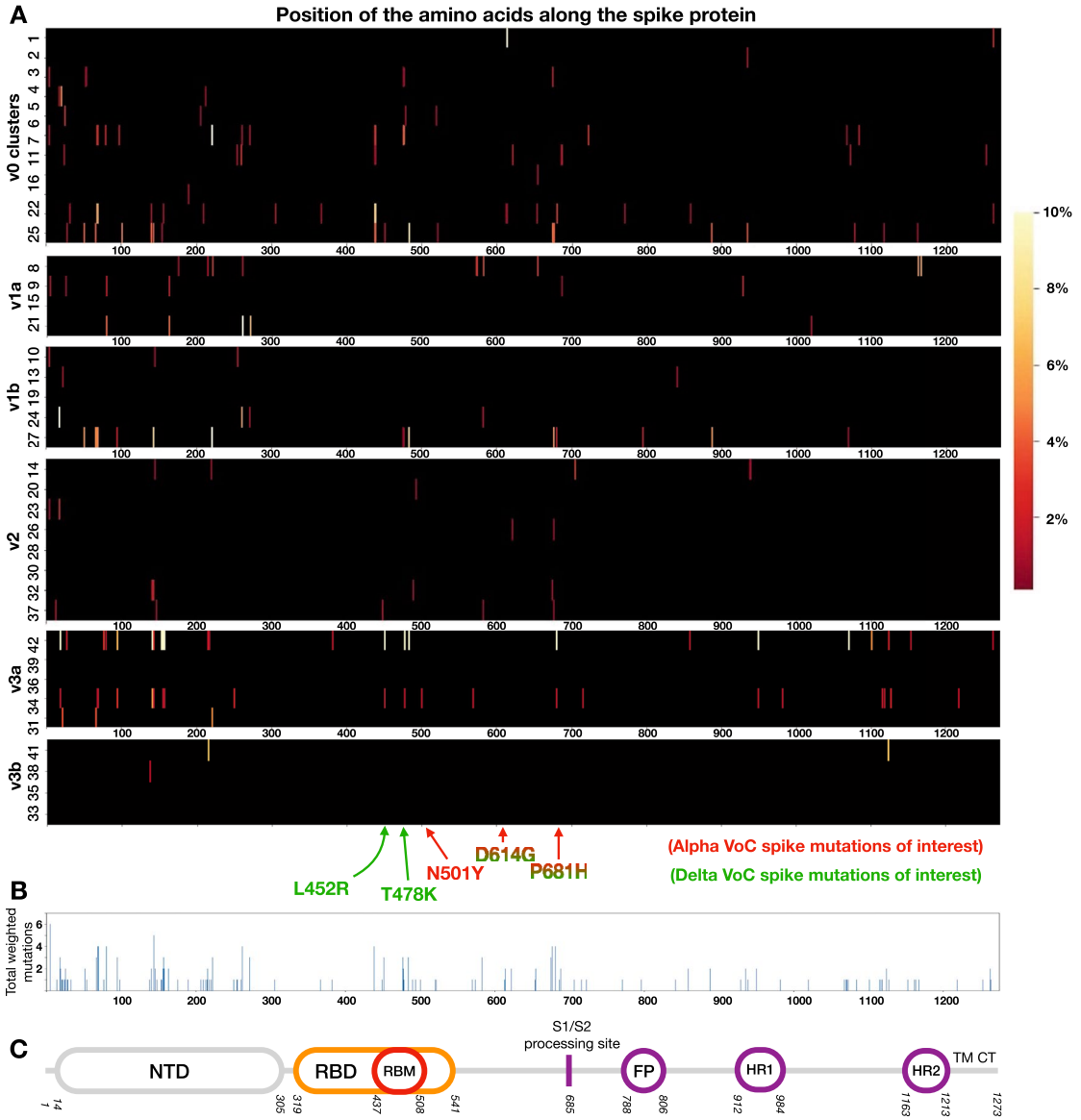

Figure 5. Patterns of spike protein diversification of the emerging variants. Heatmaps indicate for each cluster in a chain (**A**) the location and number of amino acid substitutions counted per all variants in a cluster when compared to the previous cluster and normalised accordingly (see Supplementary material). Subdominant variants in each cluster are retained if their frequency is above 1% in the cluster. (**B**) recapitulates the sum of all the values from all chains in A, while (**C**) shows the S protein structure for reference.

$$\mathscr{I}_c(t) = A\frac{e^{\gamma(t-t_0)}}{1+e^{\gamma(t-t_0)}}, \tag{1}$$

where $\mathscr{I}_c$ is the cumulative number of infected, γ is the infection rate (in inverse days) and A is the total affected individuals after the wave (per 100.000 inhabitants). These two parameters control the size and duration of the wave. The parameter t_0 controls the timing of the wave, and is of no concern in this study. We recall that the parameter γ encodes the effective diffusion speed of the variant, including not only its intrinsic viral power but also the effect of pharmaceutical measures (like vaccinations) and social distancing measures. Nevertheless, it is possible to compare the value of these parameters between different variants. If the diffusion occurs under similar social conditions, this represents a measure of the ability of the new variant to spread and infect new individuals.

Hence, we used the logistic function above to fit the epidemiological data, after distributing the new daily infected to each variant proportionally to the variant frequency observed in the sequencing data. This procedure yields a reliable estimate of the diffusion of each variant. For this purpose, we used the full dataset from GISAID for the whole UK, using the VoC classification embedded in the GISAID data. As shown before, this classification is equivalent to the result of our ML approach. The result is shown in the top row of Fig. 6, where we show the number of sequences (left plot), the new number of infections per variant and the result of the MeRG fit (middle) and the frequency of the VoCs (right). Note that the total numbers are plotted in blue, while the VoCs

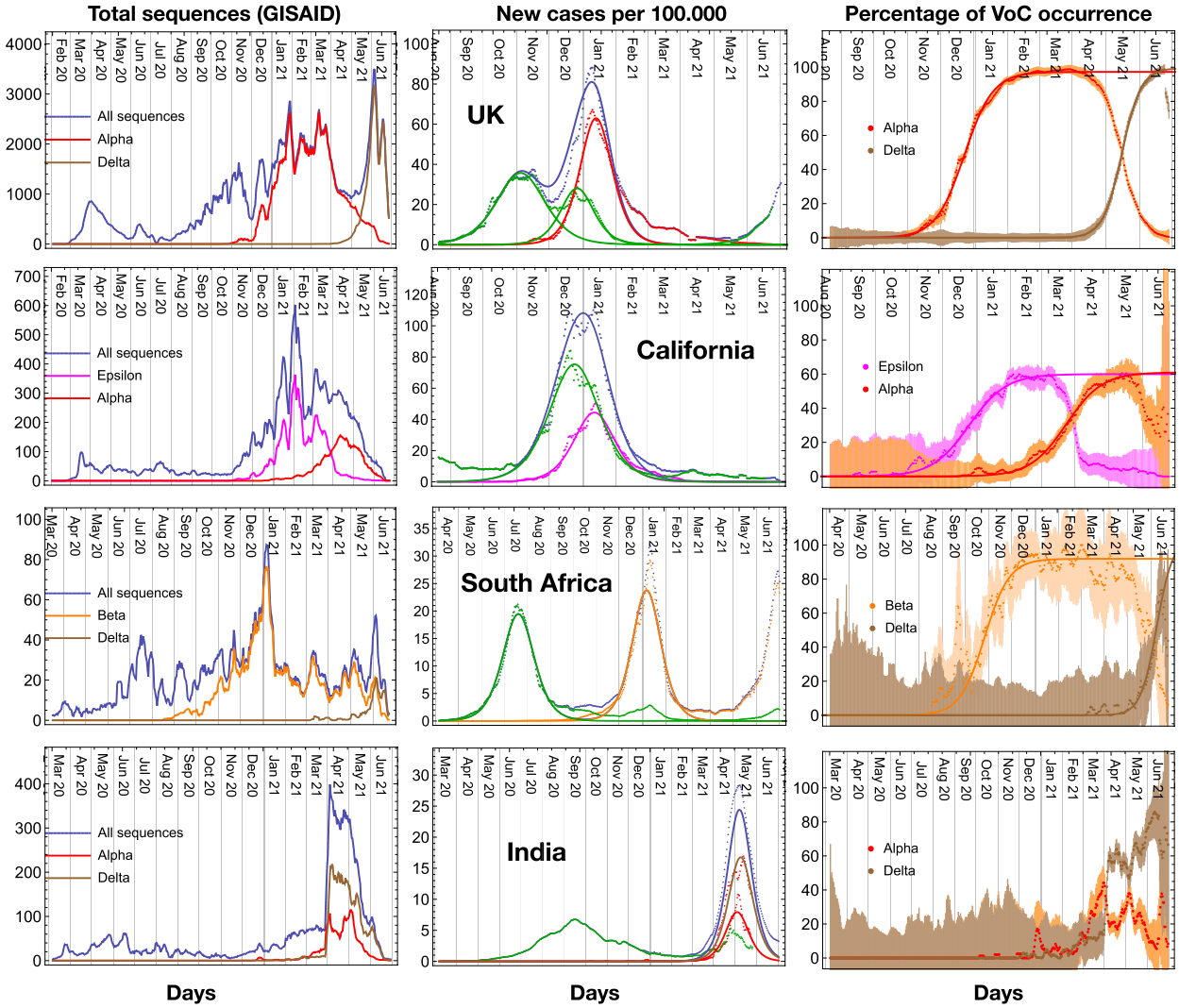

Figure 6. MeRG model for epidemiological data of variants. Results of the MeRG fitting of the number of infected associated to each relevant variant. Each row corresponds to a geographical region. In the left column we show the total number of sequencing available on GISAID (in colour the ones associated to the relevant VoC or VoI); the middle column shows the number of new daily infected (per 100.000 inhabitants); the right column shows the percentage of each VoC or VoI in the sequencing data. All plots show daily rates, with data smoothened over a period of 7 days. In the middle plots, the data are shown by dots, where blue corresponds to the total and the colours show the number of infected associated to each variant. The solid lines show the result of the fits to the MeRG model (note that only for the UK we fit the "standard variant"—in green-with two logistic functions). In the left plots, the error derives from the expected statistical variation on the number of daily sequences (after smoothening). For all the plots, the classification in variants derived from the GISAID data.

in colours. We considered the epidemiological data from the most recent waves, which developed between September 2020 and February 2021. The green curve in the middle plot shows that, after the first peak at the beginning of November, a second smaller peak developed. We describe the two with two independent sigmoids. The second sigmoid is subtracted from the data when fitting for the Alpha VoC data. The parameters from the fit are reported in Table 2. As the social conditions during this period did not change substantially, it is meaningful to compare the γ parameters for the Alpha and Delta VoC with the other ones (in green). We observed a marked increase in the transmissibility, by 49% for Alpha, which is compatible with laboratory tests. Interestingly, the frequency percentage for the VoCs, shown in the left plot, can also be fitted very accurately with a logistic function in Eq. (1) as long as only one VoC dominates. The results are also reported in Table 2. The fit parameter $\gamma_{\%}$ is a measure of how more infectious is the new VoC with respect to the previously dominant one. This plot also shows very effectively the switch between the two variants, occurring in May 2021.

We repeated the same analysis for South Africa, California and India, which show very good fits notwithstanding the more limited sequencing statistics available on GISAID. This is clearly shown in the left plots, where we report the statistical uncertainly at 65% confidence level, due to the available sequencing. The results, shown in Fig. 6 and Table 2, demonstrate that the MeRG framework provides an excellent modelling of the data.

Region	Standard variant		Variant of concern			Transmissibility	VoC percentage	
	A	γ	A_{VoC}	γ_{VoC}	VoC/VoI	Increase	$A_{\%}$	$\gamma_{\%}$
UK	2140 (12)	0.0668 (5)	2530 (10)	0.0994 (7)	Alpha	49%	97.3 (3)%	0.076 (1)
			–	–	Delta	–	99 (1)%	0.115 (2)
South Africa	1104 (2)	0.0705 (4)	1161 (2)	0.0904 (5)	Beta	28%	91.9 (8)%	0.061 (4)
			–	–	Delta	–	96 (6)%	0.090 (7)
India	717 (3)	0.0358 (4)	497.8 (8)	0.0858 (3)	Alpha	140%	–	–
			908 (5)	0.0747 (6)	Delta	109%	–	–
California	4773 (7)	0.0620 (3)	2250 (5)	0.0758 (5)	Epsilon	22%	59.9 (6)%	0.059 (4)
			–	–	Alpha	–	61.0 (6)%	0.0610 (2)

Table 2. MeRG fit parameters. Parameters from the fit of the VoC/VoI for the UK, South Africa, California and India, also shown in Fig. 6. The fit follows the MeRG model, according to which each variant can be fitted by an independent logistic function. For the UK, the "standard variant" fit corresponds to the first peak, in October–November 2020. The transmissibility increase is computed by comparing the gamma of the VoC with that of the standard variant in the same country. For the new variants that have not reached the peak of diffusion, it is not possible to extract reliable values for the eRG parameters.

Discussion

We presented a ML algorithm that allows to identify, classify and track epidemiologically relevant variants of SARS-CoV-2. It is based on the Levenshtein distance of the Spike protein sequences and is unbiased in the sense that it requires no prior knowledge of any of the variants' properties. For each time bin, the algorithm first produces an independent clustering of the Spike protein sequences. It then links clusters in subsequent bins with a common dominant Spike variant, thereby creating chains of clusters depicting temporal waves of variants. The results for England empirically showed that the a chain persisting at leat 3 consecutive clusters is a strong indication for an increased viral fitness of its dominant variant. This criterion allowed to identify emerging variants that pose a significant epidemiological threat. We validated the method with both monthly and weekly time binning.

We applied the algorithm to the sequencing data from England, which offers the largest dataset on the GISAID open-source genome repository. Among the emerging variants, the officially recognised VoCs (Alpha and Delta) were clearly identified and isolated. Similar results for Wales and Scotland (despite a more limited number of available sequences) confirmed the effectiveness of the algorithm, while comparison of our approach (that uses data of the Spike protein sequences only) to other informed methods based on the complete genome validated the algorithm. Furthermore, the temporal organisation of clusters into chains served as a tool not only to monitor the genetic evolution of the Spike protein but also to help shed light on its mechanisms. On the one hand, from the temporal chain analysis branching relations arouse from which we can reconstruct the evolutionary diversification that leads to the establishing of emergent variants. On the other hand, within a single chain, the analysis of mutations of subdominant Spike variants permitted to distinguish regions of the sequence with a high frequency of mutations from those in which no amino acid substitutions take place over time.

Using the relative percentage of each variant in the sequencing dataset to estimate the number of individuals infected by each variant, we correlated our temporal chain analysis with epidemiological data. We discovered that each new wave of the COVID-19 pandemic in England (and similarly in Scotland and Wales) was driven and dominated by a new emerging variant. This observation corroborates the hypothesis that there exists a strong and direct causal relation between the emergence of a new variant and the onset of a new epidemic wave. We modelled the cumulative number of infected individuals by use of the MeRG framework that we proposed in a companion manuscript[24]. We also used epidemiological data from the whole UK, California, India and South Africa to confirm the validity of the model.

Finally, in view of potential future waves of COVID-19, we tested the viability and performance of our ML algorithm as an early warning tool to detect the emergence of a new, epidemiologically dangerous, variant. We demonstrated that the Alpha VoC could be established as the dominant variant of an emerging persistent chain 9 weeks after its first detection in the England data set. This precedes its classification as a VoC by the WHO by 6 weeks. We showed, more generally, that an early warning for the emergence of a new persistent variant can be issued once its associated cluster reaches 1% of the time-binned sequence data. Interestingly, the Spike protein was a reliable and sufficient reference to meet a successful goal. Despite being preliminary in light of the need to better adjust the clustering cutoff, our analysis of the most recent data in the UK stresses the emergence of the Pango lineage AY.4.2 into a stable chain and thus of a true variant of concern. Our analysis indicates that an early warning could have been issued as of the 19th of September. Thus, our ML analysis tool and these early warning indications could be used by policy makers to implement immediate actions that globally limit the spread of this and of other variants that will emerge in the future.

Limitations. This study was performed on the sequencing data from a single region, England. This is justified by the fact that the sequencing dataset associated to England on the GISAID open-source genome repository is by far the largest and most uniform compared to other countries/regions. Biases relative to specific data-taking practices un England may induce biases in the analysis. To validate the results, however, we have also analysed the data for Wales and Scotland, as presented in the Supplementary material. We chose the other two nations of

the Great Britain island because they have a very similar epidemiological history compared to England, thus we would expect comparable outcomes. As such, by comparing the results we would test the reliability of the ML procedure alone. In fact, the results for Wales and Scotland, while less significant with respect to statistics, show the same patterns we obtained for England.

As an early warning tool, our ML approach is triggered when a new cluster chain branches off. This procedure is sensitive to the working point chosen for the clustering algorithm, and it has a certain chance to produce a false positive. However, the extensive use of this tool on future data and on sequences from other countries/regions will allow to estimate more reliably the false positive probability.

Conclusions

The results of our ML analysis have profound impact, both scientifically and epidemiologically: They provide new insights that are crucial for the development of new strategies to study how SARS-CoV-2 variants emerge and to predict the evolutionary pattern as well as the characteristics of future mutations of the Spike protein. We provide a tool that allows for an efficient and unbiased identification of emerging variants, for the tracking of the evolution and diversification of their Spike proteins and, most importantly, for an early warning system to identify epidemiological threats for the population.

The concrete results presented in this work can be viewed as a relevant step in the development of alternative strategies aimed at understanding the genesis of variants in epidemic or pandemic infectious diseases. The temporal dynamics of variants, in fact, allows to study the branching off of new relevant variants, and to track the evolutionary pattern of amino acid substitutions in the Spike protein highlighting persistent variants and structural trends, in terms of hotspots of mutations and/or of conserved regions. Further studies are necessary to fully exploit this information.

Our work furthermore underlines the importance of sufficient genomic data in order to both scientifically understand and track the temporal evolution of viral diseases, but also to issue sufficiently early warnings of epidemiologically dangerous variants so that decision makers can take efficient preventative measures. Our approach can be applied to other viral diseases, like influenza, provided that sufficient sequencing data is available.

Data availability

All raw data used in this work are obtained from open-source repositories: GISAID (https://www.gisaid.org/) for the sequencing; Ourworldindata.org (https://ourworldindata.org/) and the UK Coronavirus Dashboard (https://coronavirus.data.gov.uk/details/cases) for the epidemiological data. The Machine Learning code is available at https://github.com/AdeledeHoffer/ML-Covid.

Received: 24 November 2021; Accepted: 28 April 2022
Published online: 03 June 2022

References

1. Taubenberger, J. K. & Morens, D. M. 1918 influenza: The mother of all pandemics. *Rev. Biomed.* **17**(1), 69–79 (2006).
2. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J. Virol.* **84**, 9733–9748 (2010).
3. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116–121. https://doi.org/10.1038/s41586-020-2895-3 (2021).
4. Korber, B. *et al.* Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827 (2020).
5. Wu, A. *et al.* Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe* **27**(3), 325–328. https://doi.org/10.1016/j.chom.2020.02.001 (2020).
6. Konings, F. *et al.* SARS-CoV-2 variants of interest and concern naming scheme conducive for global discourse. *Nat. Microbiol.* **6**, 821–823. https://doi.org/10.1038/s41564-021-00932-w (2021).
7. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407. https://doi.org/10.1038/s41564-020-0770-5 (2020).
8. Elbe, S. & Buckland-Merret, G. Data, disease and diplomacy: Gisaid's innovative contribution to global health. *Glob. Chall.* **1**, 33–46. https://doi.org/10.1002/gch2.1018 (2017).
9. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data—From vision to reality. *EuroSurveillance* **22**(13), 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 (2017).
10. Rambaud, A. *et al.* Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. In *COVID-19 Genomics Consortium UK (CoG-UK) Report*. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (2020).
11. Mahase, E. Covid-19: What have we learnt about the new variant in the UK?. *BMJ* **371**, 1–2. https://doi.org/10.1136/bmj.m4944 (2020).
12. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv*https://doi.org/10.1101/2020.12.21.20248640 *(2020).*
13. Sabino, E. C. *et al.* Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *Lancet* **397**, 452–455. https://doi.org/10.1016/S0140-6736(21)00183-5 (2021).
14. Pater, A. A. *et al.* Emergence and evolution of a prevalent new SARS-CoV-2 variant in the United States. *bioRxiv*https://doi.org/10.1101/2021.01.11.426287 *(2021).*
15. Rasigade, J.-P. *et al.* A viral perspective on worldwide non-pharmaceutical interventions against COVID-19. *medRxiv*https://doi.org/10.1101/2020.08.24.20180927 *(2020).*
16. Volz, E. *et al.* Transmission of SARS-CoV-2 lineage in B.1.1.7 England: Insights from linking epidemiological and genetic data. *medRxiv*https://doi.org/10.1101/2020.12.30.20249034 *(2021).*
17. Kermack, W. O., McKendrick, A. & Walker, G. T. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. A* **115**, 700–721 (1927).
18. Perc, M. *et al.* Statistical physics of human cooperation. *Phys. Rep.* **687**, 1–51 (2017).
19. Wang, Z., Andrews, M. A., Wu, Z.-X., Wang, L. & Bauch, C. T. Coupled disease-behavior dynamics on complex networks: A review. *Phys. Life Rev.* **15**, 1–29 (2015).

20. Giordano, G. *et al.* Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy. *Nat. Med.*https://doi.org/10.1038/s41591-021-01334-5 *(2021).*
21. Della Morte, M., Orlando, D. & Sannino, F. Renormalization group approach to pandemics: The COVID-19 case. *Front. Phys.* **8**, 144. https://doi.org/10.3389/fphy.2020.00144 (2020).
22. Cacciapaglia, G. & Sannino, F. Interplay of social distancing and border restrictions for pandemics (COVID-19) via the epidemic Renormalisation Group framework. *Sci. Rep.* **10**, 15828. https://doi.org/10.1038/s41598-020-72175-4 (2020). arxiv:2005.04956.
23. Cacciapaglia, G., Cot, C. & Sannino, F. Second wave COVID-19 pandemics in Europe: A temporal playbook. *Sci. Rep.* **10**, 15514. https://doi.org/10.1038/s41598-020-72611-5 (2020). arxiv:2007.13100.
24. Cacciapaglia, G. *et al.* Epidemiological theory of virus variants. *Physica A Stat. Mech. Appl.* **596**, 127071. https://doi.org/10.1016/j.physa.2022.127071 (2022). arxiv:2106.14982.
25. Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Dokl. Akad. Nauk* **163**, 845–848 (1965).
26. Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Cybern. Control Theory* **10**, 707–710 (1966).
27. Bouckaert, R. *et al.* BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**(4), e1003537. https://doi.org/10.1371/journal.pcbi.1003537 (2014).
28. Obermeyer, F. H. *et al.* Analysis of 2.1 million SARS-CoV-2 genomes identifies mutations associated with transmissibility. https://doi.org/10.1101/2021.09.07.21263228 (2021).
29. Wilson, K. G. Renormalization group and critical phenomena. 1. Renormalization group and the Kadanoff scaling picture. *Phys. Rev. B* **4**, 3174–3183. https://doi.org/10.1103/PhysRevB.4.3174 (1971).
30. Wilson, K. G. Renormalization group and critical phenomena. 2. Phase space cell analysis of critical behavior. *Phys. Rev. B* **4**, 3184–3205. https://doi.org/10.1103/PhysRevB.4.3184 (1971).
31. Cacciapaglia, G., Cot, C., Islind, A. S., Óskarsdóttir, M. & Sannino, F. Impact of us vaccination strategy on COVID-19 wave dynamics. *Sci. Rep.* **11**, 10960. https://doi.org/10.1038/s41598-021-90539-2 (2021). arxiv:2012.12004.
32. Della Morte, M. & Sannino, F. Renormalization group approach to pandemics as a time-dependent sir model. *Front. Phys.* **8**, 583. https://doi.org/10.3389/fphy.2020.591876 (2021).
33. Cacciapaglia, G. *et al.* The field theoretical ABC of epidemic dynamics (2021). arxiv:2101.11399.
34. Cacciapaglia, G., Cot, C. & Sannino, F. Mining google and apple mobility data: Temporal anatomy for COVID-19 social distancing. *Sci. Rep.* **11**, 4150. https://doi.org/10.1038/s41598-021-83441-4 (2020). arxiv:2008.02117.
35. Cacciapaglia, G., Hohenegger, S. & Sannino, F. Effective mathematical modelling of health passes during a pandemic. *Sci. Rep.* **12**, 6989. https://doi.org/10.1038/s41598-022-10663-5 (2022).
36. Brauner, J. M. *et al.* Inferring the effectiveness of government interventions against COVID-19. *Science* **371**(6531), eabd9338. https://doi.org/10.1126/science.abd9338 (2021).
37. Sharma, M. *et al.* Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe. *Nat. Commun.* **12**(1), 5820. https://doi.org/10.1038/s41467-021-26013-4 (2021).
38. Li, Y. *et al.* The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: A modelling study across 131 countries. *Lancet Infect. Dis.* **21**(2), 193–202. https://doi.org/10.1016/S1473-3099(20)30785-4 (2021).
39. Liu, Y., Morgenstern, C., Kelly, J., Lowe, R. & Jit, M. The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. *BMC Med.* **19**(1), 40. https://doi.org/10.1186/s12916-020-01872-8 (2021).
40. Cacciapaglia, G., Cot, C. & Sannino, F. Multiwave pandemic dynamics explained: How to tame the next wave of infectious diseases. *Sci. Rep.* **11**, 6638. https://doi.org/10.1038/s41598-021-85875-2 (2021). arxiv:2011.12846.
41. Cacciapaglia, G. & Sannino, F. Evidence for complex fixed points in pandemic data. *Front. Appl. Math. Stat.* **7**, 659580. https://doi.org/10.3389/fams.2021.659580 (2021). arxiv:2009.08861.
42. Latif, A. A. *et al.* AY.4.2 Lineage Report. *outbreak.info*. https://outbreak.info/situation-reports?pango=AY.4.2 (2021).
43. Wrobel, A. G. *et al.* SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat. Struct. Mol. Biol.* **27**(8), 763–767. https://doi.org/10.1038/s41594-020-0468-7 (2020).
44. Thomson, E. C. *et al.* Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* **184**(4), 1171–1187. https://doi.org/10.1016/j.cell.2021.01.037 (2021).

Acknowledgements

We acknowledge with gratitude the authors, originating and submitting laboratories of the genetic sequence and metadata made available through GISAID. A full listing of all authors and laboratories is available on the GISAID website.

Author contributions

This work has been designed and performed conjointly and equally by all the authors. In particular: A.d.H., S.V., A.G. and F.C. have developed the Machine Learning algorithm and analysed the spike protein sequencing data; A.C. and M.L.C. contributed in the analysis of the spike protein diversification; C.C. has collected and analysed the clade and variant of concern data from GISAID; G.C., C.C., S.H. and F.S. have developed the theoretical framework. All authors have equally contributed to the writing of the text.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-12442-8.

Correspondence and requests for materials should be addressed to G.C. or F.S.