# Ultrafast topological data analysis reveals pandemic-scale dynamics of convergent evolution

Michael Bleher[1*†], Lukas Hahn[1†], Maximilian Neumann[1,2*†], Zachary Ardern[7], Juan Ángel Patiño-Galindo [4], Mathieu Carrière[5], Ulrich Bauer[6], Raúl Rabadán[3*], Andreas Ott[1*]

[1]Institute for Mathematics, Heidelberg University, Heidelberg, Germany

[2]Institute for Biological Interfaces 5, Microbial Genetics & Biotechnology, Karlsruhe Institute of Technology, Karlsruhe, Germany

[3]Program for Mathematical Genomics, Department of Systems Biology, Columbia University, New York, NY, USA

[4]Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[5]DataShape, Centre Inria d'Université Côte d'Azur, Biot, France

[6]TUM Department of Mathematics and Munich Data Science Institute, Munich, Germany

[7]Wellcome Trust Sanger Institute, Hinxton, United Kingdom

[†]These authors contributed equally to this work.

[*]Corresponding authors:
mbleher@mathi.uni-heidelberg.de (M.B.)
mn.neumann@kit.edu (M.N.)
rr2579@cumc.columbia.edu (R.R.)
aott@mathi.uni-heidelberg.de (A.O.)

## Abstract

Genome variants which re-occur independently across evolutionary lineages are key molecular signatures of adaptation. Inferring the dynamics of such genetic changes from pandemic-scale genomic datasets is now possible, which opens up unprecedented insight into evolutionary processes. However, existing approaches depend on the construction of accurate phylogenetic trees, which remains challenging at scale. Here we present EVOtRec, an organism-agnostic, fast and scalable Topological Data Analysis approach that enables the inference of convergently evolving genomic variants over time directly from topological patterns in the dataset, without requiring the construction of a phylogenetic tree. Using data from both simulations and published experiments, we show that EVOtRec can robustly identify variants under positive selection and performs orders of magnitude faster than state-of-the-art phylogeny-based approaches, with comparable results. We apply EVOtRec to three large viral genome datasets: SARS-CoV-2, influenza virus A subtype H5N1 and HIV-1. We identify key convergent genome variants and demonstrate how EVOtRec facilitates the real-time tracking of high fitness variants in large datasets with millions of genomes, including effects modulated by varying genomic backgrounds. We envision our Topological Data Analysis approach as a new framework for efficient comparative genomics.

## Introduction

Convergent evolution in genomes is a biological signal which often corresponds to variants of high fitness [1]. Recurrently arising variants are of critical importance for instance in the case of microbial pathogens, where adaptive strategies can even be shared across diverse species [2–4]. Identifying these genome variants and mapping their evolutionary dynamics is central to understanding pathogen biology and can inform public health responses. New access to pandemic-scale datasets of millions of genomes creates new opportunities for studying this kind of molecular evolution at an unprecedented scale but also new computational challenges [5].

Experimental approaches consisting of mutating many genome positions *in vivo* and testing the effect on certain phenotypes, are limited by the vast number of possible variations, and also, for instance, the lack of tractable laboratory assays for most viral proteins [6] and technical limitations in implementing unbiased assays at scale [7]. Computational methods are well-suited to massive-scale analyses, but face their own challenges [8]. These challenges have been addressed by several new approaches, including machine learning approaches [9–11], clade-growth approaches [12–14], as well as optimized phylogenetic approaches [6, 15–18]. Phylogenetic approaches have shown good concordance with experimental fitness measurements [6]. High fitness variants can be inferred via genomic recurrence [19]—if a mutation confers an evolutionary advantage, we expect it to re-occur in independent lineages of the phylogeny and its frequency should increase with time.

However, phylogenetic approaches rely on the construction of accurate phylogenetic trees, which is challenging when the number of genomes exceeds tens of thousands [5]. Common strategies to overcome this difficulty include downsampling [20] and stepwise addition of new sequences onto an existing phylogenetic tree [21–23]. However, downsampling can reduce accuracy through loss of information, and stepwise addition can be used for very large genomic datasets but results in less accurate tree topologies. An additional conceptual challenge for phylogenetic approaches is that homoplasies and recombination can confuse the construction of phylogenetic trees [24–26]. Such merging of distinct "vertical" lineages can arise from several evolutionary scenarios. If a genome of an organism acquires genetic material from a different

genome through horizontal gene transfer or recombination, we will observe that parts of the newly generated genome resemble the "vertical" parent, while others resemble the genome of the "horizontal" organism that exported that material [27–30]. Homoplasies are generated when the same mutation occurs independently twice and is retained in both lineages, making the two sampled strains more similar than expected [31].

Here we present EVOtRec (https://github.com/ottamj/evotrec), an organism-agnostic, fast and scalable approach based on persistent homology, a method from Topological Data Analysis [32], that overcomes limitations of phylogenetic approaches by quantifying the dynamics of convergent evolution solely through its topological footprint in the genomic dataset, without the need to reconstruct a phylogenetic tree. We analyze the sensitivity and robustness of our method with simulated evolutionary scenarios and apply it to three pandemic-scale viral genome datasets: coronavirus SARS-CoV-2, avian influenza virus A subtype H5N1 and human immunodeficiency virus HIV-1. We assess EVOtRec's computational performance in relation to leading phylogenetic tools, and also assess the different tools' ability to identify fitness-increasing genomic variants detected in previously published experimental data. Using our new approach, we track potentially fitness-increasing effects of individual genomic variants over time and investigate hotspots of convergent evolution dynamics across the genome, modulated by varying genomic backgrounds.

## Results

### Topological recurrence analysis with EVOtRec quantifies convergent evolution

Chan *et al.* [33] initiated the use of persistent homology to extract global evolutionary features from genome datasets. This method detects topological cycles in the dataset which correspond to apparent reticulate events (implying homoplasy or recombination) in the phylogeny. Persistent homology captures all these events, and also the scale of the events. All information is compiled into a stable and unbiased descriptor known as a *persistence barcode* (Figure 1 and Supplementary Figure 1). Bars in the barcode starting at small genetic distance scale typically correspond to homoplasies, while bars starting at larger scale are indicative of recombination events involving the exchange of larger portions of genetic material [33–35].

We use persistent homology to introduce a novel index of recurrence that quantifies convergent evolution directly through topological cycles in the genomic dataset, without the need for computationally expensive reconstruction of a phylogenetic tree. *SNV cycles* in the genomic dataset are topological cycles that consist of a series of genomes that approximates all evolutionary steps as faithfully as possible in terms of single nucleotide variations (SNV). We define the *topological recurrence index (tRI)* of a specific SNV as the total number of SNV cycles containing this SNV (Figure 1 and Methods). The rationale behind this definition is that mutations observed in an SNV cycle occur independently twice in different lines of descent within this cycle, which implies that the mutation may confer some evolutionary advantage, given that the genomes observed result from processes including filtering by natural selection [31, 36]. A proportion of tRI signals may also be due to other factors, and the nonuniformity of mutational spectra [37] can bias tRI signals towards particular nucleotide mutation types (Supplementary Figure 8). To compensate for these effects, we implemented several tRI corrections in the topological recurrence analysis (Methods). Most importantly, we normalize the tRI by its mean per nucleotide mutation type across four-fold degenerate sites [6, 37, 38].

The topological recurrence analysis can be carried out for sliding windows across the genome (windows of a defined length or single genomic regions), based on sub-alignments of appropriately truncated genome sequences (Figure 4, Supplementary Figure 9 and Methods) [35]. In

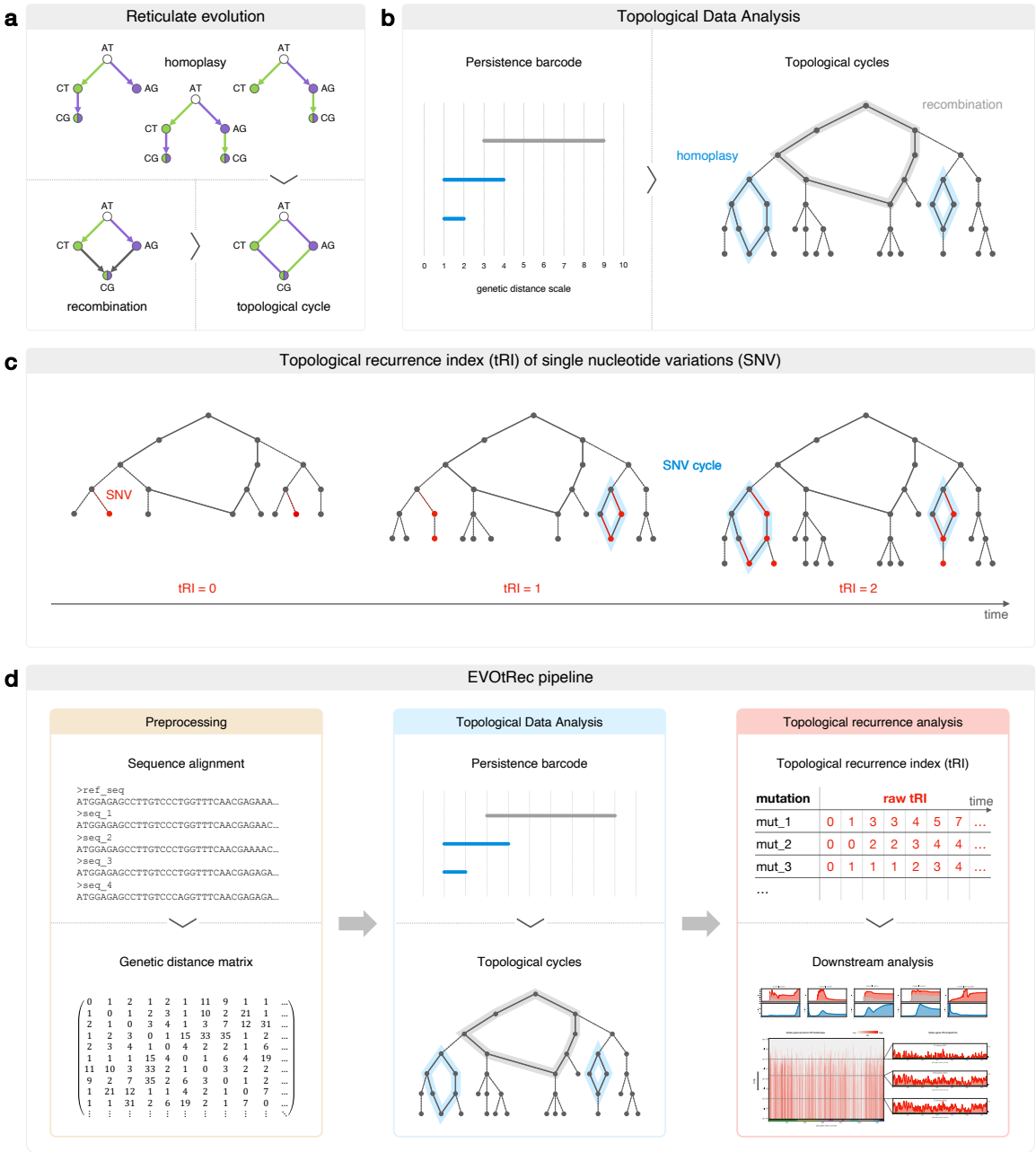



**Figure 1. Topological recurrence analysis with EVOtRec quantifies convergent evolution.** **(a)** Possible evolutionary histories for a dinucleotide sequence. The coloring of the edges corresponds to a specific mutation, while the coloring of the nodes indicates the presence of the genomic variant resulting from this mutation. Reticulate events (implying homoplasy or recombination) lead to the presence of four alleles for which there is no single consistent phylogeny. On the genomic level, genetically identical individuals cannot be distinguished and incompatible phylogenies are represented by a topological cycle in the corresponding phylogenetic network. **(b)** Topological Data Analysis detects reticulate events solely by means of their topological footprint in the genomic dataset, without the need to construct a possibly ambiguous phylogenetic tree (Methods). On the technical level, this is achieved by computing a persistence barcode. Each bar in the barcode corresponds to a topological cycle in the reticulate phylogeny, with bars starting at small (large) genetic distance scale typically corresponding to homoplasies (recombination events). **(c)** SNV cycles are topological cycles in which sequences corresponding to adjacent nodes differ by single nucleotide variations (SNV) only, i.e., the genetic distance (edge length) between any two adjacent nodes is 1. These cycles correspond to bars starting at genetic distance $= 1$ in the persistence barcode. The raw topological recurrence index (raw tRI) of a specific SNV is the total number of SNV cycles which contain an edge corresponding to this SNV (Methods). Intuitively, mutations contributing to tRI signals are more likely to happen if the mutation confers some evolutionary advantage, because they appear independently in different lines of descent. In the displayed phylogeny, the SNV colored in red has a raw tRI of 2 as it is acquired in precisely two SNV cycles. For genome sequence alignments which incorporate temporal information, the time-dependent tRI can resolve topological signals of recurrence at different time steps (Methods). **(d)** The EVOtRec (topological Recurrence in EVOlution) pipeline (Methods) takes as input a sequence alignment (including temporal information) with a fixed reference sequence and returns (time series) tRI data for every SNV in the alignment.

this case, mutations outside the window are ignored, which typically leads to the creation of more topological cycles and stronger tRI signals, revealing a more detailed picture of convergent evolution in that region of the genome.

We implemented our method in the EVOtRec (topological Recurrence in EVOlution) pipeline (https://github.com/ottamj/evotrec). This pipeline takes as input a sequence alignment with a fixed reference sequence and returns a list of SNVs with corresponding tRI scores (Figure 1). EVOtRec is designed for the efficient and scalable analysis of very large genomic datasets, and is based on highly optimized algorithms for the rapid computation of the persistent homology of genomic datasets. Hammingdist [39] is used to compute genetic distances, and Ripser [40] is used for the subsequent computation of the persistence barcode and the localization of topological cycles (Figure 1, Supplementary Figure 1 and Methods). For sequence alignments that integrate temporal information, such as the collection date of genomic isolates, EVOtRec applies the MuRiT algorithm [41] to enable a time series analysis of convergent evolution based on time-dependent tRI (Figure 1 and Methods). EVOtRec processes genetic distances in a tailored sparse matrix format which keeps track of short distances only, thereby greatly reducing the complexity of persistent homology computations (Supplementary Figure 1 and Methods). In this way, EVOtRec serves as a useful tool for the real-time monitoring of emerging potentially adaptive genomic variants in ongoing evolutionary processes and for the mapping of convergently evolving genomic variants over time.

### Simulations show that topological recurrence analysis is sensitive to positive selection and robust to topological noise and stochastic sequencing errors

To validate the reliability of the tRI as a robust indicator of positive selection, we employed simulations of genome evolution and statistical tests. First, topological cycles can arise randomly in neutral evolution, causing some background noise in tRI signals. However, from neutral evolutionary scenarios simulated with SANTA-SIM [42], we inferred that for SARS-CoV-2 this effect is very small and amounts to less than 5% of topological/SNV cycles (Supplementary Figure 5) and less than 0.5% of total tRI sum (Figure 2 and Methods).

Second, while consensus genomes obtained with current sequencing technologies have very high sequencing accuracy of almost 100% [6, 43, 44], sequencing errors can still affect phylogenetic analyses [24, 26, 45]. To mitigate the impact of sequencing errors on tRI, we applied several filtering rules in the preprocessing of sequence alignments to ensure that only high-quality genomes were included in the topological recurrence analysis (Methods). Furthermore, because of the combinatorial structure of SNV cycles (each SNV must occur independently twice in SNV cycles; Figure 1) it is unlikely that random nucleotide substitutions give rise to SNV cycles, and we therefore do not expect that stochastic sequencing errors drive tRI much. To support this argument, we added simulated stochastic sequencing errors to a synthetic alignment generated under neutral evolution with SANTA-SIM, and to a large real SARS-CoV-2 whole genome alignment (Methods). We observed only a minimal tRI signal increase in the simulated alignment and none at all in the SARS-CoV-2 alignment (Figure 2 and Supplementary Figure 6).

A Monte Carlo permutation test showed that the frequency distribution of raw tRI signals differs significantly from a uniform distribution (Kolmogorov–Smirnov test, $p < 0.05$; Supplementary Figure 7). We propose a conservative permutation-based tRI significance level to account for the effects of background noise and sequencing errors (Methods).

On the other hand, systematic sequencing errors may confound results of the topological recurrence analysis—a phenomenon that is also present in phylogenetic inference [21, 24, 26, 45, 46]. To contain this effect, we recommend that in the downstream analysis of tRI signals,

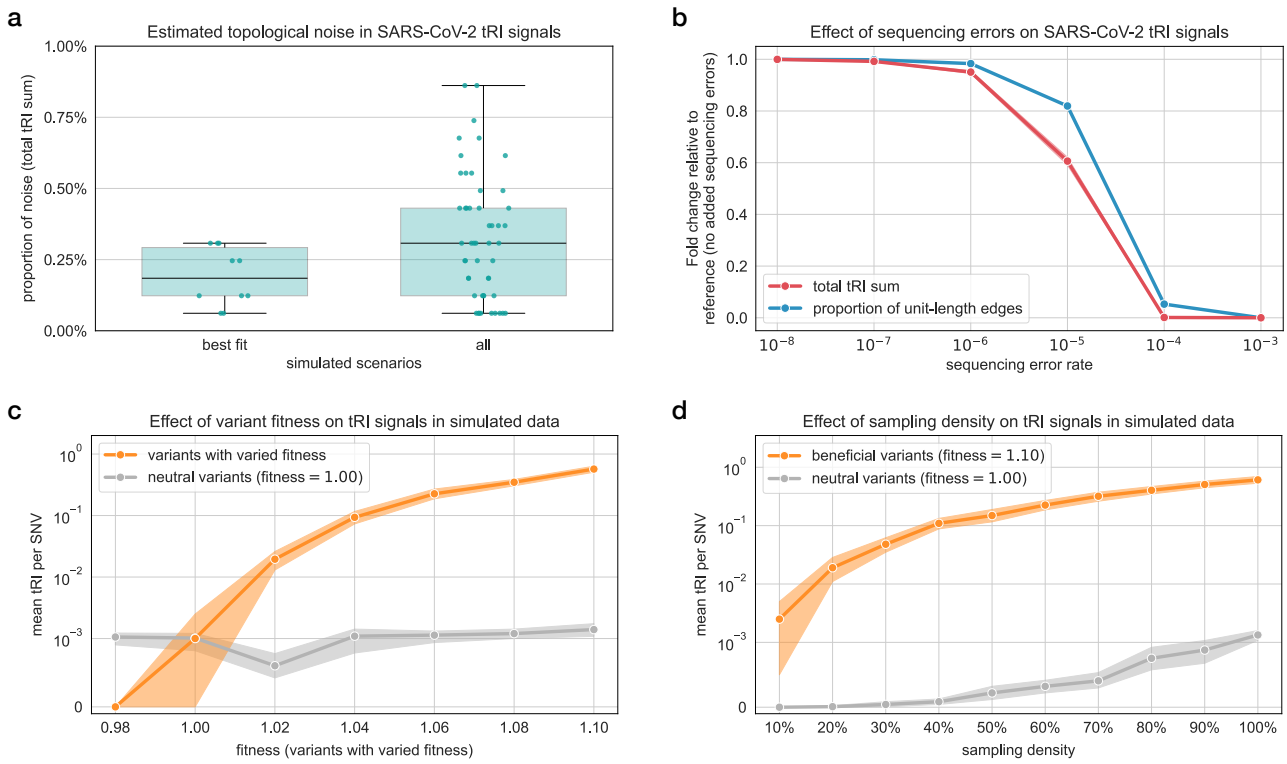



**Figure 2. Simulations show that topological recurrence analysis is sensitive to positive selection and robust to topological noise and stochastic sequencing errors. (a)** Box plot with overlaid strip plot showing the proportion of background topological noise in tRI (total tRI sum across all variants) of a SARS-CoV-2 whole genome alignment with 100k sequences, estimated from five different simulated neutral evolutionary scenarios (for details see Supplementary Figure 5 and Methods). **(b)** We assessed how stochastic sequencing errors affect SARS-CoV-2 tRI signals by adding simulated stochastic sequencing errors to a real SARS-CoV-2 whole genome alignment with 200k sequences, for varying error rates ranging from $10^{-8}$ to $10^{-3}$ (for details see Supplementary Figure 6 and Methods). The line plot (mean ± 95% confidence band) shows the relative change in the total tRI sum across all variants and in the proportion of unit-length edges among all possible edges in the genomic dataset. Here, relative change refers to the alignment with added sequencing errors compared against the reference alignment without added errors. We found that the total tRI sum did not increase after adding sequencing errors but instead decreased to zero at higher error rates. This phenomenon is explained by the fact that higher sequencing error rates increase genomic distances, thus progressively reducing the proportion of unit-length edges which are the building blocks of SNV cycles (Figure 1). **(c-d)** Dependence of tRI signals on variations in variant fitness and sampling density in simulated synthetic alignments (for details see Supplementary Figure 3 and Methods). (c) Line plot (mean ± 95% confidence band) showing the dependence of tRI on variations in variant fitness. We increased the fitness parameter stepwise from 0.98 to 1.10 for 1.7% of mutations (variants with varied fitness), while keeping it fixed at 1.00 for all other mutations (neutral variants). (d) Line plot (mean ± 95% confidence band) showing the dependence of tRI on variations in sampling density. We increased the sampling density stepwise from 10% to 100%, with fitness fixed at 1.10 for beneficial variants and at 1.00 for neutral variants. We note that tRI signals of neutral variants occurring in panels (c) and (d) arise from background topological noise in neutral evolution as in (a).

masking schemes [45, 47] are applied to exclude problematic sites.

Finally, we used synthetic evolutionary scenarios generated with SANTA-SIM to examine the sensitivity of the tRI for positive selection (Figure 2, Supplementary Figure 3 and Methods). We found a very strong monotonic correlation (Spearman's $\rho = 1.0$, $p < 1e{-}100$) between fitness and tRI for variants with varied fitness. The simulations moreover show that tRI is sensitive to beneficial variants, whereas deleterious variants produce almost no tRI signal. Next, we quantified the dependence on sampling density of the effect size (Cohen's $d$) for the difference in tRI signals between beneficial and neutral variants. We found a large effect size ($1.31 \leq d \leq 4.42$) for sampling densities ranging from 20% to 100%, and a medium effect size ($d = 0.64$) for a sampling density of 10%. Overall, these findings demonstrate that tRI is sufficiently sensitive to detect positive selection, and that this sensitivity is robust to variations in sampling density, topological noise and stochastic sequencing errors.

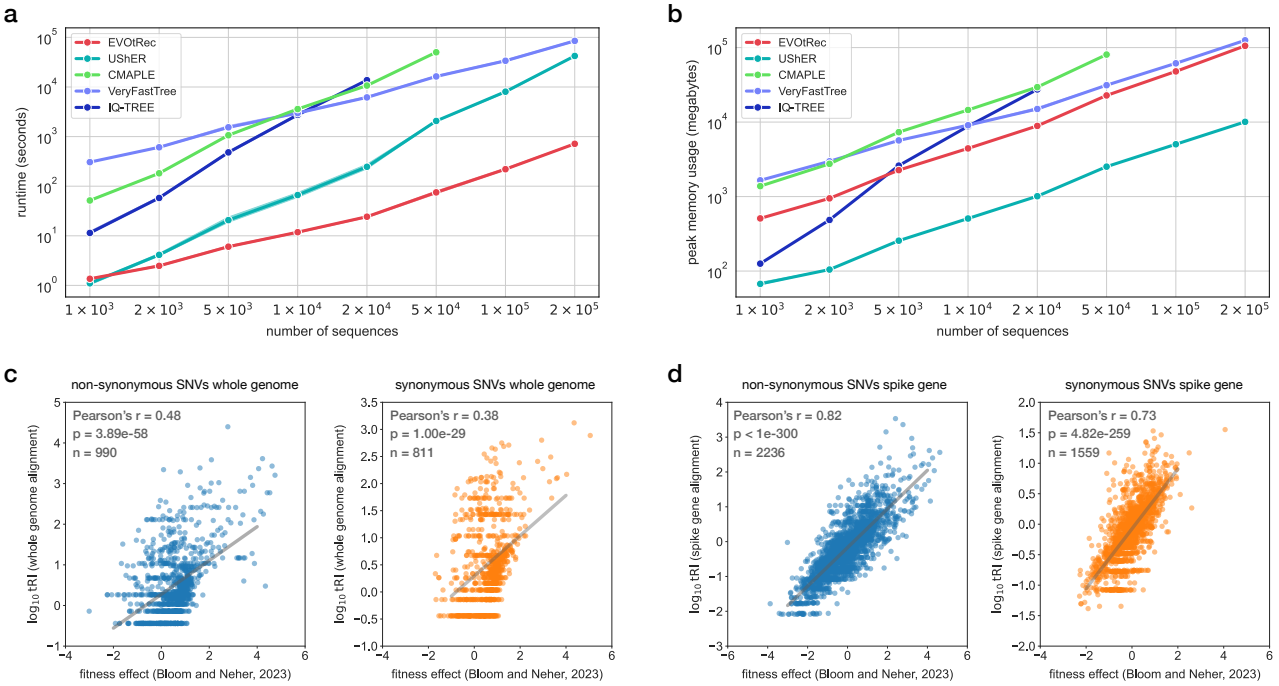



**Figure 3. EVOtRec performs faster than phylogenetic tools with congruent results on fitness-increasing effects. (a, b)** Line plots (mean ± 95% confidence band) showing runtime and peak memory usage (RAM) of EVOtRec vs. the following phylogenetic inference tools: IQ-TREE [48], VeryFastTree [49] and CMAPLE [50] (de novo analysis), and UShER [21] (phylogenetic sample placement). For the benchmarking we used SARS-CoV-2 genome alignments with 1k up to 200k distinct sequences, obtained by subsampling the SARS-CoV-2 whole genome alignment (Supplementary Table 1 and Methods). **(c, d)** Correlation between log-transformed positive tRI and phylogeny-based fitness effect estimates [6] for SARS-CoV-2 SNVs on the whole genome (c) and the spike gene (d) (Methods).

## EVOtRec performs faster than phylogenetic tools with congruent results on fitness-increasing effects

We compared EVOtRec with state-of-the-art phylogeny-based approaches for the assessment of fitness-increasing effects of genomic variants. In a first step, we used SARS-CoV-2 whole genome alignments to benchmark the runtime and memory usage of EVOtRec against four tools for phylogenetic tree reconstruction: IQ-TREE [48], VeryFastTree [49] and CMAPLE [50] for de novo analysis, and UShER [21] for phylogenetic sample placement (Figure 3, Supplementary Table 1 and Methods). We found that for large alignments with more than 20k genomes, EVOtRec performed at least two orders of magnitude faster than IQ-TREE, VeryFastTree and CMAPLE, and more than one order of magnitude faster than UShER. For example, EVOtRec completed the de novo analysis of the 200k genomes alignment in 11min 55s, yielding >100-fold speedup over the fastest de novo analysis tool (VeryFastTree, 23h 37min) and >50-fold speedup over the fastest sample placement tool (UShER, 11h 44min) (Supplementary Table 1). Further, EVOtRec required more memory (RAM) than tools for phylogenetic sample placement (UShER), but less memory than tools for de novo phylogenetic inference (IQ-TREE, VeryFastTree and CMAPLE), with a peak memory usage of 106 GB for the 200k genomes alignment (UShER, 10 GB and VeryFastTree, 125 GB) (Supplementary Table 1). EVOtRec gains computational efficiency by employing only short genetic distances, which make up less than 1% of distances in large SARS-CoV-2 whole genome alignments. This reduces the time (memory) consumption for the computation of persistence barcodes with Ripser [40] by 5 (3) orders of magnitude compared to the standard persistent homology computation (Supplementary Figure 1).

Next, we compared tRI signals with phylogenetic mutation count-based estimates of fitness

effects. Using simulated sequence alignments generated with SANTA-SIM, we first analyzed the Spearman correlation between raw tRI profiles and phylogenetic homoplasy counts computed with IQ-TREE and TreeTime [51] (Methods). We observed a moderate-to-strong correlation with $\rho = 0.59$ for beneficial variants, but no correlation for neutral variants ($\rho = 0.04$), with several neutral variants exhibiting no tRI signal but having positive homoplasy count (Supplementary Figure 4).

Moreover, mutation counts in the UShER phylogenetic tree have previously been used [6] to estimate fitness effects of SARS-CoV-2 mutations. We found a moderate-to-strong Pearson correlation between log-transformed tRI from our analysis and these fitness effect estimates [52] for the whole genome ($0.38 \leq r \leq 0.48$) and spike gene ($0.73 \leq r \leq 0.82$) (Figure 3 and Methods). Divergence between tRI and phylogeny-based fitness effect estimates likely reflects the fact that tRI is sensitive to beneficial mutations, while fitness effect estimates have been reported to be better for estimating effects of deleterious or nearly neutral mutations in comparison with clade-growth methods [6, 37].

These findings highlight that our approach, based on topological recurrence, differs from traditional recurrence analyses on global phylogenetic trees in that it analyzes the independent occurrence of mutations within topological cycles locally in the genomic dataset, without constructing a phylogenetic tree (Figure 1).

### EVOtRec reveals the pandemic-scale dynamics of SARS-CoV-2 convergent evolution

Convergent mutations play a key role in SARS-CoV-2 evolution and are indicative of potential adaptation [53–57]. To demonstrate EVOtRec's ability to map the pandemic-scale dynamics of convergent evolution in time-resolved genomic datasets, we performed a time series topological recurrence analysis of SARS-CoV-2 genome data shared via GISAID, the global data science initiative [58], collected between January 2020 and February 2024 and covering the entire COVID-19 pandemic. The raw dataset contains 14,956,592 whole genomes along with temporal information given by collection date. We analyzed two high quality nucleotide sequence alignments which we compiled from this dataset, one containing whole genomes and another one containing spike genes (Supplementary Table 2 and Methods). Restricting the analysis to the spike gene alignment resulted in stronger tRI signals, and provides a more detailed picture of convergence in the spike gene (Supplementary Figure 9 and Methods).

We used EVOtRec to study the convergent evolution of single nucleotide variations by means of *tRI/prevalence time series charts* showing time-dependent tRI vs. time-dependent prevalence at daily resolution (Methods). For example, for the beneficial spike gene amino acid changes Q52H [59, 60], E484K [61, 62], N501Y [63–65] and P681R [66] we observed an increase in tRI over a certain period of time, which is accompanied, or followed, by an increase also in prevalence (Figure 4). These charts can be used for the real-time tracking of convergently evolving genomic variants and to identify time intervals in which a specific mutation potentially conferred a selective advantage. We found that on the last day of each time series (29/30 January 2024), tRI and prevalence were moderately correlated with Pearson's $r = 0.24$ for the whole genome alignment ($r = 0.25$ for the spike gene alignment) (Supplementary Figure 10). This is consistent with positive selection acting as a confounding factor, where evolutionary advantages lead to an increase in both tRI and prevalence. For some mutations we observed a different pattern—a typical example is the highly recurrent mutation L3606F [8, 26, 45, 53] on the ORF1a gene, for which tRI and prevalence showed diverging trends (Figure 4), suggesting that this mutation is neutral or deleterious.

We recovered the known convergently evolving spike gene amino acid sites R346, K444, L452, F456, A475, N460 and F486. Amino acid changes at these sites have been shown to recurrently

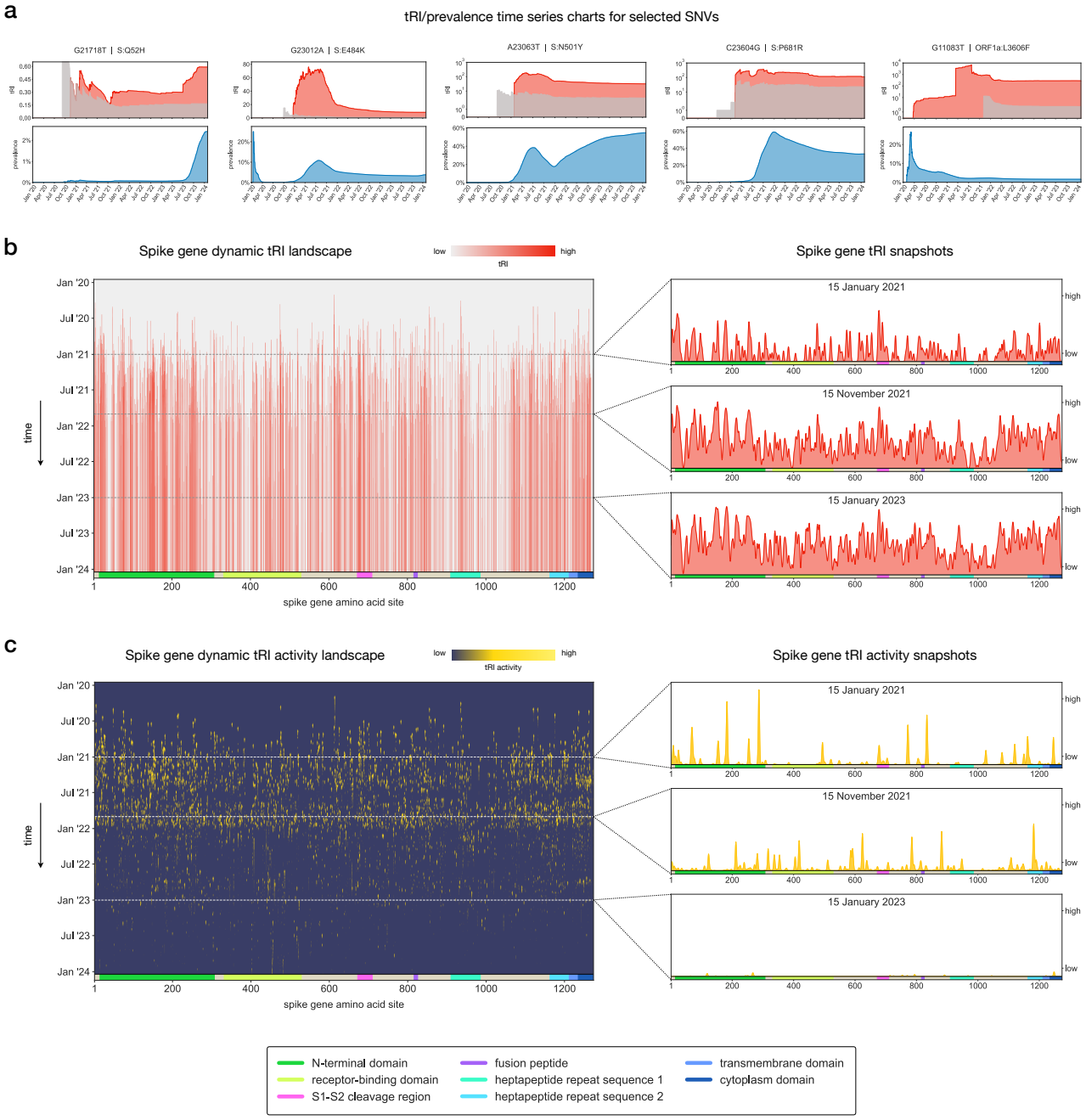



**Figure 4. EVOtRec reveals the pandemic-scale dynamics of SARS-CoV-2 convergent evolution.** We performed a time series topological recurrence analysis of the evolution of SARS-CoV-2 during the COVID-19 pandemic [57] with EVOtRec, analyzing the whole genome and spike gene alignments based on data from GISAID [58] collected between January 2020 and February 2024 (Supplementary Table 2 and Methods). **(a)** Time series charts showing tRI (red) and prevalence (blue) at daily resolution. The area shaded in grey marks the tRI significance level (i.e. the red tRI signal is significant whenever it is above the grey area). Charts for the fitness-increasing spike amino acid changes Q52H [59, 60], E484K [61, 62], N501Y [63–65] and P681R [66] obtained from the spike gene alignment, and for the highly recurrent mutation L3606F [8, 26, 45, 53] on the ORF1a gene obtained from the whole genome alignment. **(b, c)** Dynamic fitness landscapes mapping the spike gene convergent evolution throughout the entire COVID-19 pandemic from January 2020 until January 2024 (left), and snapshots (horizontal slices) showing fitness landscapes at fixed time points in the early, middle and late phase of the pandemic (right). Dynamic tRI landscapes (b) comprehensively capture time-dependent tRI signals across all amino acid sites on the spike gene. Vertical line segments (red) indicate the strength of tRI signals over time at a specific site. Dynamic tRI activity landscapes (c) capture changes in the intensity of convergent evolution by mapping positive growth rates of tRI signals in (b). Vertical line segments (yellow) indicate periods of time during which tRI signals at a specific amino acid site were increasing.

occur in several Omicron sub-lineages [67–69], and we detected significant tRI signal at each of these sites after the emergence of the Omicron variant in late 2021 (Supplementary Figure 11).

While tRI/prevalence time series charts map the convergent evolution of individual SNVs,

a more comprehensive picture can be obtained by simultaneously comparing time-dependent tRI signals across all amino acid sites on the genome. For this, we integrated EVOtRec tRI time series data into a *dynamic tRI landscape* which captures tRI signals across the entire spike gene for the period from January 2020 until February 2024 at daily resolution (Figure 4 and Methods). Landscapes of tRI signals at specific time points can then be extracted by taking *snapshots* (horizontal slices) in the dynamic tRI landscape. To localize hotspots of convergent evolution both in time and genome position, we further integrated EVOtRec time series data into a *dynamic tRI activity landscape* that visualizes positive tRI growth rates across all amino acid sites on the spike gene in a two-dimensional heatmap (Figure 4 and Methods). Applying Gaussian smoothing can help to reduce image detail and enhance hotspots of convergent evolution (Figure 5). In summary, landscapes of tRI signals enable the rapid and convenient identification of genomic regions shaped by convergent evolution over time. The interpretation should be supported at the level of individual SNVs by more detailed tRI/prevalence charts and, if available, additional experimental data.

### Topological recurrence correlates with adaptation

For the biological validation of our method, we compared EVOtRec tRI signals with published experimental and epidemiological data characterizing adaptation for three different viruses: SARS-CoV-2, influenza A subtype H5N1 and HIV-1. First, we demonstrated EVOtRec's ability to capture the convergent evolution in pandemic-scale SARS-CoV-2 data by comparing spike gene tRI signals with published deep mutational scanning (DMS) measurements for the wild-type and the Omicron BA.1, BA.2 and XBB.1.5 sub-lineages. We found that the Pearson correlation between time-dependent tRI obtained from the spike gene alignment and ACE2-binding affinity for the wild-type receptor-binding domain (RBD) [70] was rising during the first 10 months of the COVID-19 pandemic (Figure 5) when the wild type was the predominant circulating variant in an immunologically naive host population [71], before it began to decline again as immune escape became an increasingly important evolutionary factor [65, 71] (Methods). We moreover investigated the performance of tRI in identifying fitness-increasing amino acid changes (i.e. positive DMS effect) for the BA.1, BA.2 and XBB.1.5 sub-lineages. We found a statistically significant association between binarized tRI (tRI $> 0$ vs. tRI $= 0$) obtained from the Omicron spike gene alignment and binarized DMS effect (DMS effect $> 0$ vs. DMS effect $\leq 0$) for antibody escape in BA.1 [72] and spike-mediated entry in BA.2 [59] (Fisher's exact test, $p < 0.01$; Figure 5 and Methods). For XBB.1.5, a logistic model predicting positive tRI signals as a function of the full three-dimensional XBB.1.5 DMS data [59] suggests a correlation with ACE2-binding ($\beta_1 = 0.53$), spike mediated entry ($\beta_2 = 1.44$) and human sera escape ($\beta_3 = 1.02$) (Supplementary Figure 12 and Methods). Findings from this logistic model are consistent with results from a similar phylogeny-based ordinary least squares regression analysis using published fitness effect estimates [6, 17] (Supplementary Figure 12). The strength of correlations between tRI and DMS measurements is likely limited by differences in mutational and selective factors in virus transmission in the wild versus selective experiments in the laboratory. Mutational factors include host RNA editing processes [37, 79–83] and hyper-mutable sites [45, 81]. Selective factors include transmissability [57], as well as epistatic interactions across the whole genome or varying genomic backgrounds [37, 84–86]. Aside from biological differences, systematic sequencing errors can also cause apparent genomic recurrence [26, 45], increasing tRI at some sites.

We further used these experimental DMS data to compare the performance of tRI with phylogeny-based fitness effect estimates [6, 17] in identifying fitness-increasing amino acid changes (i.e. positive DMS effect) for three different SARS-CoV-2 phenotypes (Supplemen-

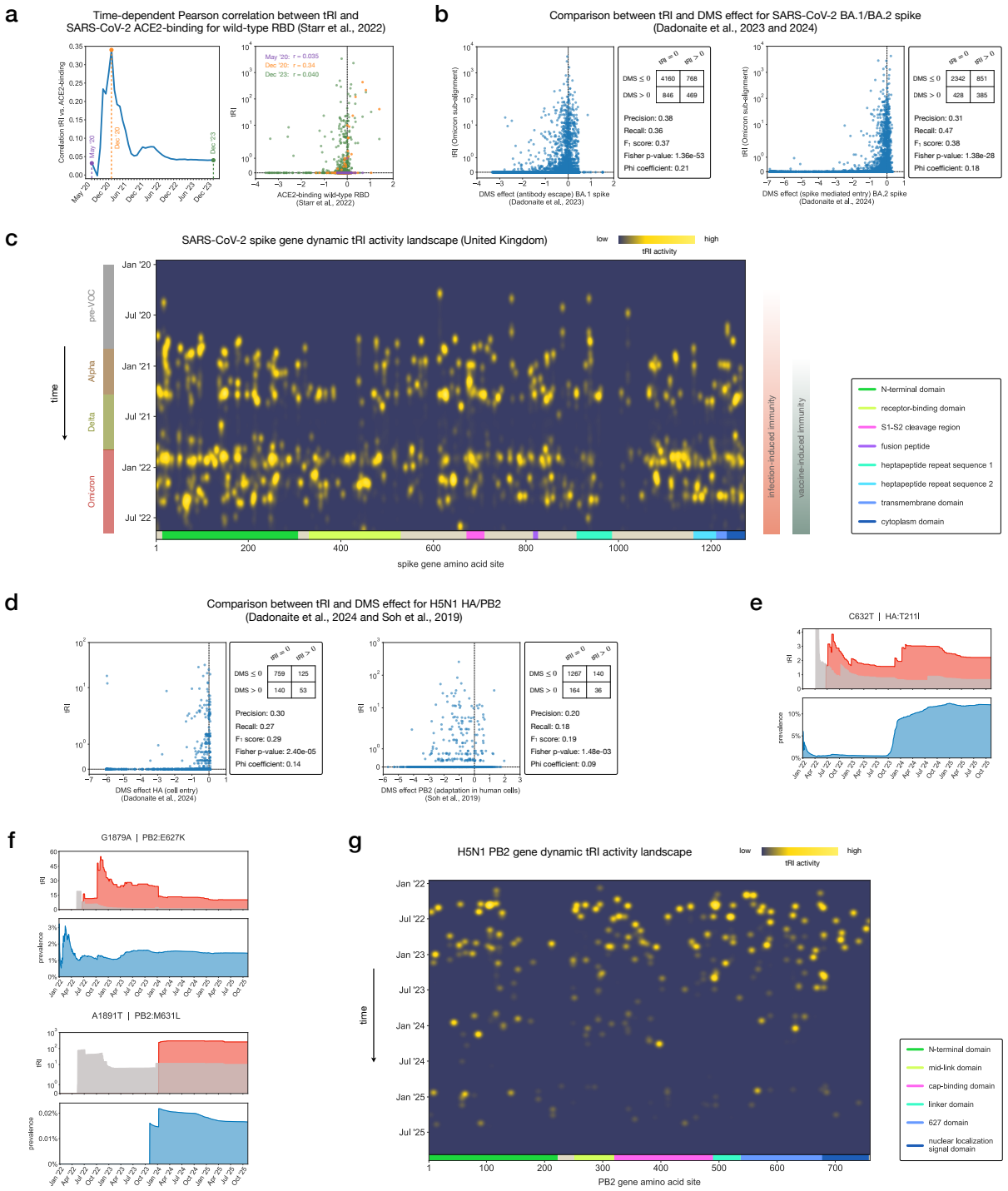



**Figure 5. Topological recurrence correlates with adaptation.** **(a)** Time-dependent Pearson correlation between tRI and ACE2-binding affinity for amino acid changes in the SARS-CoV-2 wild-type RBD (Methods). Experimental data taken from yeast display deep mutational scanning (DMS) measurements [70]. **(b)** Comparison between tRI and published pseudovirus DMS measurements (antibody escape and cell entry) [59, 72] of amino acid changes in the SARS-CoV-2 Omicron spike gene alignment for the BA.1 and BA.2 sub-lineages. Statistical analyses are based on contingency tables for binarized tRI/DMS effects (Methods). **(c)** SARS-CoV-2 spike gene dynamic tRI activity landscape for the United Kingdom covering the period from January 2020 until August 2022 (Methods). Hotspots of tRI activity reflect changes in genomic background (emergence of dominant genomic variants) and in host immunity status (infection and vaccination) [71]. **(d)** Comparison between tRI and published pseudovirus DMS measurements (cell entry and adaptation in human cells) [73, 74] of amino acid changes in the H5N1 HA and PB2 gene alignments. Statistical analyses as in (b). **(e-f)** Time series charts showing tRI (red; area shaded in grey marks tRI significance level) and prevalence (blue) at daily resolution for the mammalian-adaptive mutations HA:T211I (T199I in H3 numbering) [75] (e) and PB2:E627K [76, 77] and PB2:M631L [78] (f). **(g)** H5N1 PB2 gene dynamic tRI activity landscape covering the period from January 2021 until November 2025 (Methods). Hotspots of tRI activity reflect changes in genomic background (emergence of genotypes B3.13 in late 2023/early 2024 and D1.1 in late 2024/early 2025).

tary Figure 12). We found that both methods yield comparable $F_1$ score and precision, while the tRI tends to exhibit higher recall. Notably, our unbiased, organism- and site-agnostic approach achieves predictive performance on par with the phylogeny-based fitness effect estimates in [17] which invoke SARS-CoV-2 site-specific mutation rates and sequence context.

Moreover, we observed that at least 84% of all amino acid changes that had fixed in the Alpha, Beta, Gamma, Delta or Omicron variants also exhibited significant tRI, while among all amino acid changes only 15% had this property (Supplementary Figure 7).

To assess EVOtRec's ability to map the time-dependence of convergent evolution in the context of changing genomic backgrounds, we investigated to what extent the dynamic tRI landscape of the SARS-CoV-2 spike gene was modulated by an increase in host immunity and the emergence of dominant virus variants over the course of the COVID-19 pandemic in the United Kingdom [57, 71, 87, 88] between January 2020 and August 2022 (Figure 5, Supplementary Table 2 and Methods). We observed a significant rise in tRI activity in the second half of 2020, with topological signals accumulating in the S1-S2 cleavage region, which is a determinant of SARS-CoV-2 transmissibility [71, 89–91], in regions in the RBD which are known to enhance ACE2-binding, such as for example site N501 [63–65, 92], and in the N-terminal domain which allosterically modulates S1-S2 cleavage and the conformation of the RBD for ACE2-binding [59, 93]. This is in line with the observation that adaptation during the first year of the pandemic was mainly enhancing infectiousness and transmissibility in an immunologically naive host population, as seen in the Alpha and Delta variants [71, 94]. Immune escape became an evolutionary factor early in the year 2021, when infection-induced population immunity was on the rise on the Alpha and Delta genomic backgrounds [55, 65, 71]. Correspondingly, we recorded a further increase in tRI signals in the N-terminal domain and the RBD, which contain epitopes for antibody binding [89, 95–99] such as site E484 [61, 62], and in the S1-S2 cleavage region [93], as for example at site P681 [66] (Figure 5). After a longer period with almost no tRI activity, we observed another significant rise in activity towards the end of the year 2021, with distinct hotspots in the N-terminal domain, the RBD, the S1-S2 cleavage region, the fusion peptide and other regions in the S2 subunit, which play a key role in immune escape [59, 67, 86, 95–98, 100–103] (Figure 5). This is congruent with the observation that during this later phase of the pandemic, immune evasion became the dominant factor of positive selection in the presence of widespread population immunity, as seen in the convergent evolution of the Omicron variant and its sub-lineages [67–69, 71, 103, 104].

These findings indicate that dynamic tRI activity landscapes can capture shifts in variant fitness resulting from changes in genomic background and interactions with genetic changes at other sites in the genome [55, 65, 105] (Figure 5). It is a particular feature of our method that, since genomes in SNV cycles share the same genomic background (Figure 1), tRI automatically keeps track of shifting genomic backgrounds without the need to analyze sub-alignments corresponding to different phylogenetic clades (Supplementary Figure 9).

Finally, we examined EVOtRec's ability to identify fitness-increasing genomic variants in the evolution of influenza A subtype H5N1 and HIV-1. Since the sampling rate in the corresponding alignments is lower than in SARS-CoV-2, we augmented the data using ancestral sequence reconstruction (Supplementary Figure 2 and Methods). For H5N1, we performed time series topological recurrence analyses of the hemagglutinin gene (HA) and the polymerase basic 2 gene (PB2) alignments belonging to the currently circulating highly pathogenic avian influenza H5N1 strain 2.3.4.4b (Supplementary Table 3 and Methods). Using tRI data from the last day of each time series (7/8 November 2025), we found a statistically significant association between binarized tRI (tRI $>$ 0 vs. tRI $=$ 0) and binarized DMS effect (DMS effect $>$ 0 vs. DMS effect $\leq$ 0) for pseudovirus entry into 293T cells for the HA gene [73] and viral

| mutation | tRI | indication of adaptation | references |
|---|---|---|---|
| F317I | 244.0 | phylogeny-based analysis | Yoshida et al., 2011 |
| N302K | 244.0 | associated with antibody resistance | Witvrouw et al., 2005 |
| A281D | 214.5 | associated with antibody resistance | Watkins et al., 1993; Yoshimura et al., 2010; Galvez et al., 2024 |
| W17R | 188.1 | potentially associated with antibody resistance | Upadhyay et al., 2018 |
| F353I | 183.0 | deep mutational scanning | Haddox et al., 2016 |

**Table 1. Topological recurrence correlates with adaptation.** Comparison of tRI with published data on fitness increasing mutations in HIV-1. The table shows the top five amino acid changes on the HIV-1 envelope gene (Env) with strongest tRI signal. These five substitutions have been associated with (potential) adaptation in the literature (see Supplementary Table 5 for a more comprehensive list). The tRI data was generated with EVOtRec from the Env gene alignment (Supplementary Table 4 and Methods), and the indication of adaptation is taken from [106–112].

adaptation in human A549 cells for the PB2 gene [74] (Fisher's exact test, $p < 0.01$; Figure 5 and Methods). The observed association between tRI and DMS measurements is less strong for H5N1 than for SARS-CoV-2, which may be explained by the fact that the percentage of mutations in the HA and PB2 alignments among all mutations that have been tested in the DMS experiments is lower than in SARS-CoV-2 (Supplementary Figure 13). We detected statistically significant tRI signals for key signatures of mammalian adaptation [113–116], including the amino acid substitutions T211I (T199I in H3 numbering) [75] in the HA gene, and E249G [117], A588T [118], E627K [76, 77], M631L [78, 119], D701N [77, 120] and D740N [78] in the PB2 gene (Figure 5 and Supplementary Figure 11). Moreover, the dynamic tRI activity landscape of the PB2 gene (Figure 5 and Methods) reveals patterns of topological signals likely associated with changes in genomic background following spillover events from wild birds to dairy cattle in North America [78, 116, 121]. For example, we observed an increase in tRI for the key adaptive mutations M631L [78] in the B3.13 genotype in late 2023/early 2024, and D701N [122] in the D1.1 genotype in early 2025. We further report HA and PB2 amino acid changes currently under convergent evolution (Supplementary Figure 14).

For HIV-1, we performed the tRI analysis for the envelope gene (Env), which is essential for host cell binding and entry, and has the capability to escape host immunity through its rapid evolution [106, 123] (Supplementary Table 4 and Methods). We found that among the top 40 amino acid changes on the Env gene with strongest tRI signals, at least 70% have been associated with potential positive selection in the literature (Table 1 and Supplementary Table 5).

## Discussion

In this study, we introduce a method from Topological Data Analysis that can rapidly identify the presence of recurrent genomic variants in pandemic-scale sequence alignments. Our approach uses persistent homology to quantify convergent evolution in terms of global topological patterns in genomic datasets, without relying on the challenging prior reconstruction of an accurate phylogenetic tree. We implemented the method in the EVOtRec pipeline, which in benchmark tests performed orders of magnitude faster than state-of-the-art tree inference tools like IQ-TREE [48], VeryFastTree [49], UShER [21] and CMAPLE [50], while results were congruent with phylogeny-based fitness effect estimates [6, 17]. The pipeline is distance-based

and organism-agnostic, and enables an unbiased and scalable analysis of millions of genomes, which facilitates leveraging the ever-increasing wealth of genomic sequencing data.

Our method is designed for the rapid analysis of the dynamics of convergent evolution in very large time-resolved genomic datasets. To demonstrate its utility and versatility, we applied the EVOtRec pipeline to pandemic-scale viral genome datasets from SARS-CoV-2, influenza A subtype H5N1 and HIV-1, representing three different virus families. The SARS-CoV-2 dataset we analyzed contained more than 14 million genomes, covering the entire COVID-19 pandemic. We showed that EVOtRec's speed and scalability enable the real-time monitoring of emerging potentially adaptive mutations in ongoing evolutionary processes over time. In addition, the method reveals hotspots of convergent evolution across a two-dimensional dynamic fitness landscape by mapping genomic recurrence simultaneously in both time and genome position. The method is also well-suited to the retrospective analysis of convergent evolution in historical genomic data. We validated our approach with simulated data and published experimental data (deep mutational scanning) across the three viruses, confirming that EVOtRec can robustly identify fitness-increasing genomic variants. Notably, we demonstrated that our method matches or exceeds the predictive power of the most biologically-informed phylogeny-based approaches [6, 17].

Care should be taken in the interpretation of EVOtRec tRI signals. Our method resolves homoplasies only at small genetic distance scales, does not distinguish between forward and backward mutations, and does in its present form not include insertions and deletions. While we applied several filtering rules in the preprocessing of sequence data and our analyses indicate that EVOtRec is robust to stochastic sequencing errors, systematic sequencing errors may still confound tRI signals. Moreover, we observed that mutations identified as topologically recurrent need not necessarily be adaptive (e.g. at hyper-mutable sites). We therefore recommend that in the downstream analysis, additional biological information, such as masking schemes, is used to identify problematic sites. Finally, the computational performance of EVOtRec may depend on the sampling rate, the ordering of sequences in the alignment and the topological structure of the genomic dataset.

Future developments in this method for viral comparative genomics could include normalizing the tRI by more detailed or site-specific mutational spectra, and investigating the utility of our approach for finding sites under purifying selection or subject to epistatic interactions. The method can motivate and guide experimental studies by reducing the biologically relevant combinatorial complexity of possible genetic variants to explore in the laboratory. Promising applications beyond viral genomics could support studies of bacterial evolution and somatic mutation processes in cancer. Analyzing evolution over greater genomic distances will come with new challenges, but the general method presented here is highly adaptable to new data types including amino acid alignments from bacterial genes or somatic mutation profiles in cancer genomes. The growing size of comparative genomics demands innovative approaches, and we propose Topological Data Analysis as a method that opens the way for new data-driven insights into molecular evolution at large scale.

## Methods

### Topological Data Analysis of genome evolution

Topological Data Analysis [32] originated from the mathematical field of *algebraic topology* and studies the shape of datasets by extracting topological structures and patterns. Such topological structures have associated dimensions: structures of dimension zero can be thought of as the connected components, and structures of dimension one are essentially the loops, or topological

cycles, of the dataset. Structures of higher dimensions can also be defined, but are also more difficult to interpret. Here we are interested in reticulate evolutionary processes, thus we choose to focus on topological structures in dimension one, since topological cycles can be interpreted as signals of divergence from phylogenetic trees (Figure 1a).

Datasets often come as point clouds: in our setting, each point corresponds to a virus genome sample, and lies in a high-dimensional space where each nucleotide of the genome is a dimension [35]. A common way to extract the phylogenetic network from this point cloud simply amounts to connecting samples as soon as their genetic distance is less than a given threshold $r > 0$. This results in a (neighborhood) graph, whose set of cycles provides candidates for the topological structures in dimension one of the true underlying network. However, a main limitation of this approach comes from the fact that relevant topological structure typically appears at multiple scales.

The most common way to handle this issue in Topological Data Analysis is to actually compute and track the cycles for all possible values of $r$ ranging from 0 to $+\infty$. As $r$ increases, some cycles can appear, and some already existing cycles can disappear, or get filled in. The whole point of Topological Data Analysis is to record, for each cycle, its radius of appearance, or birth time, and radius of disappearance, or death time. This is called the *persistent homology* of the point cloud [32, 35]. The construction, based on a scale parameter $r$, can be summarized as follows (Supplementary Figure 1a). The input is a distance matrix describing the dataset, considered as a finite metric space. First, consider the *geometric graph* at scale $r$, whose vertices are the data points, with any two points connected by an edge whenever their distance is at most $r$. Generalizing this construction, the *Vietoris–Rips complex* at scale $r$ connects any subset of the data points with a simplex (an edge, a triangle, a tetrahedron, or a higher-dimensional generalization thereof) whenever all pairwise distances in the subset are at most $r$. A Vietoris–Rips complex is a particular type of *simplicial complex*, a higher-dimensional generalization of graphs which is of crucial interest in algebraic topology, in particular in homology theory. The family of Vietoris–Rips complexes for all parameters $r$ is called the *Vietoris–Rips filtration*. It provides a multiscale method to extract cycles of various sizes, and to encode them in a so-called *persistence barcode*: each bar, or interval, in this barcode corresponds to a *topological cycle* representing a topological feature (a reticulate evolutionary process in the present case), and the bar endpoints correspond to its radii of birth and death (the maximum genetic distance between consecutive samples forming the cycle, and, roughly, the maximum pairwise genetic distance between samples forming the cycle).

Each bar indicates the presence of a reticulate event, implying that the evolutionary history cannot be fully explained in terms of a single phylogenetic tree [33–35]. The mathematical background of this phenomenon is a classical theorem due to Rips, which asserts that trees have trivial persistent Vietoris–Rips homology [33, 124]. The corresponding cycle in the associated reticulate phylogeny can then be localized in the sequence alignment by tracing it back to the isolates that constitute the reticulate event. Moreover, the length of the bar represents the cycle size. In our case, this corresponds to the length of the reticulate evolutionary process, which allows to distinguish, for instance, between homoplasies and recombinations (Figure 1b).

### The EVOtRec pipeline

The EVOtRec pipeline uses persistent homology to quantify convergent evolution in genomic datasets. For time-resolved datasets, EVOtRec can map the time-dependence of these signals. A basic version of the EVOtRec pipeline that computes raw tRI is implemented in the Python script evotrec.py (`https://github.com/ottamj/evotrec`). We applied the pipeline to SARS-CoV-2, H5N1 and HIV-1 datasets, proceeding in three main steps (Figure 1d). We first outline

these steps and provide more details in subsequent sections.

**1. Preprocessing.** We downloaded and prepared a sequence alignment with a fixed reference sequence. We used Hammingdist [39] to compute genetic distances in the sequence alignment. For the time series analysis, we further applied the MuRiT algorithm [41] to rescale genetic distances depending on temporal information. To improve efficiency, genetic distances were stored in a specifically designed sparse matrix format which keeps track of short genetic distances only.

**2. Topological Data Analysis.** We used Ripser [40] to compute the persistence barcode of the genetic distance matrix and to localize topological cycles in the genomic dataset. Ripser outputs a list of topological cycles that correspond to the bars in the barcode. The persistent homology computation is compatible with the EVOtRec sparse distance matrix format and was restricted to topological cycles in which all edges have short length (with respect to genetic distance).

**3. Topological recurrence analysis.** We performed a topological recurrence analysis on the Ripser output. SNV cycles were extracted and evaluated for the computation of the raw topological recurrence index (raw tRI). Time points of topological signals were recovered from the time-rescaled length of edges in SNV cycles, and corrections were applied to the raw tRI to obtain final tRI values. Subsequently, in the downstream analysis time series tRI data were visualized using tRI charts and two-dimensional dynamic tRI landscapes.

## Acquisition and preparation of sequence alignments

### Coronavirus

We used SARS-CoV-2 genome data provided by the GISAID EpiCoV Database [58] and accessible at https://doi.org/10.55876/gis8.240314pc (alignment msa_0222.fasta downloaded on 24 February 2024). This alignment comprises 14,956,592 SARS-CoV-2 nucleotide sequences aligned to the reference sequence hCoV-19/Wuhan/WIV04/2019. We generated four sub-alignments by applying the following operations:

(1) Sequences were truncated to specific genomic regions.
(2) High-quality genomes (complete, no ambiguous nucleotide sites, complete collection date) were selected.
(3) Only sequences that occurred in at least three identical copies in the alignment were retained, and among identical sequences, only the earliest (by collection date) was retained.
(4) Sequences in the alignment were arranged in time-reversed order (by collection date).

**Whole genome alignment:** 283,687 distinct whole genomes; length: 51,395nt; country: any; lineage: any; collection date: between December 2019 and February 2024.

**Spike gene alignment:** 158,190 distinct spike genes; length: 8,237nt; country: any; lineage: any; collection date: between December 2019 and February 2024.

**Omicron spike gene alignment:** 72,309 distinct spike genes; length: 8,237nt; country: any; lineages: Omicron BA.1, BA.2 and XBB.1.5; collection date: between November 2021 and February 2024.

**United Kingdom spike gene alignment:** 25,713 distinct spike genes; length: 8,237nt; country: United Kingdom; lineages: any; collection date: between December 2019 and August 2022.

For more details see Supplementary Information Section 1.1.

### Avian influenza virus

We used two avian influenza A subtype H5N1 (strain 2.3.4.4b) genome datasets, one for the HA gene and another one for the PB2 gene, provided by the GISAID EpiFlu Database [58] and accessible at https://doi.org/10.55876/gis8.251116eh (HA gene) and https://doi.org/10.55876/gis8.251116bc (PB2 gene). We downloaded all available nucleotide sequences (complete, with complete collection date not before 30 December 2021) as of 16 November 2025. The HA dataset comprises 24,141 sequences and the PB2 dataset comprises 23,237 sequences. We aligned each of these datasets to the reference sequence A/American_Wigeon/South_Carolina/22-000345-001/2021 (EPI_ISL_18133029) using MAFFT v7.526 [125]. We generated two alignments by applying the following operations:

(1) High-quality genomes (no ambiguous nucleotide sites) were selected.
(2) Data augmentation (ancestral sequence reconstruction) was applied.
(3) Duplicate sequences were removed, and among identical sequences, only the earliest (by collection date) was retained.
(4) Sequences in the alignment were arranged in time-reversed order (by collection date).

**HA gene alignment:** 17,223 distinct HA genes; length: 2,571nt; country: any; lineage: 2.3.4.4b; collection date: between December 2021 and November 2025.

**PB2 gene alignment:** 18,139 distinct PB2 genes; length: 2,357nt; country: any; lineage: 2.3.4.4b; collection date: between December 2021 and November 2025.

For more details see Supplementary Information Section 1.2.

### Human immunodeficiency virus

We used HIV-1 (strain HXB2) genome data provided by the European Nucleotide Archive (ENA) [126] (for accession numbers see Supplementary Table 4). These data comprise 30,964 envelope (Env) gene nucleotide sequences with genome coverage of at least 78% (downloaded on 1 March 2025), which we aligned to the HIV1/HTLV-III/LAV reference genome (ENA accession number K03455.1) using MAFFT v7.526 [125]. We generated an alignment by applying the following operations:

(1) Genomes with no ambiguous nucleotide sites were selected.
(2) Data augmentation (ancestral sequence reconstruction) was applied.
(3) Duplicate sequences were removed.

**Env gene alignment:** 30,088 distinct Env genes; length: 36,265nt; country: any; lineage: HXB2; collection date: any.

For more details see Supplementary Information Section 1.3.

### Distance matrices

We used the Python package Hammingdist v1.4.0 [39] to compute the *genetic distance matrix* of each sequence alignment. For any pair of sequences, this matrix encodes the *Hamming distance* between the two sequences, which is the number of positions at which the nucleotides in the two aligned sequences differ. Our convention in this work is that insertions and deletions (dashes "-" in aligned sequences) do not contribute to the genetic distance. To gain efficiency, the genetic distance matrix was stored as a sparse matrix in coordinate format (COO). In addition, since we were only interested in SNV cycles in persistent homology, it was sufficient to keep track of short distances only; all entries $> 3$ in the distance matrix were therefore reset to the

value 3. Correspondingly, Hammingdist was called with the arguments `max_distance=4` and `threshold=3`. This reduced the time and memory usage required for storing genetic distances (Supplementary Figure 1b-c). For the SARS-CoV-2 and H5N1 alignments, we applied the MuRiT algorithm [41] to incorporate temporal information into the distance matrix. Every entry of 1 in the sparse distance matrix was replaced by the number of days between the date of the reference sequence and the most recent date in the pair of sequences corresponding to this entry in the distance matrix, and every entry of 2 (3) was replaced by the number of days between the date of the reference sequence and the most recent date occurring in the sequence alignment, plus 1 (plus 2).

### Persistent homology, persistence barcodes and topological cycles

We used Ripser [127] to compute the persistent homology (persistence barcode and topological cycles) with $\mathbb{F}_2$ coefficients of each sequence alignment from its sparse genetic distance matrix. Ripser is a state-of-the-art software for the computation of persistent homology based on the topological construction of Vietoris–Rips complexes (Supplementary Figure 1a). For the computation of the persistence barcode, Ripser resorts to various optimizations, which are crucial when handling datasets of the size considered in this work. Notably, Ripser computes persistent cohomology, which is not based on cycles but instead on cocycles, often described intuitively as *cuts* that tear open a hole. In order to obtain topological cycles representing the evolutionary features in persistent homology, we used a custom version of Ripser that subsequently carries out a second computation, this time based on cycles instead of cocycles. This custom version of Ripser is available via commit f545200 on branch `tight-representative-cycles` of the GitHub repository `https://github.com/Ripser/ripser`. While a naive computation based on homology would be prohibitively expensive, the previous computation of the persistence barcode based on cocycles makes the subsequent computation of representative cycles feasible.

The homological features (i.e. bars in the persistence barcode) identified by persistent homology admit different cycle representatives in the dataset. In order to obtain topological cycles that fit tightly to the data points, the custom version of Ripser employs a method called *exhaustive reduction* [128, 129], which can be roughly summarized as follows: Whenever a representative cycle contains an edge that also appears in another cycle as the longest edge, a tighter representative can be obtained by replacing the edge with the remaining edges from the other cycles, which all have shorter length. The ambiguity of cycle representatives and bases has almost no effect on tRI, as we show in Supplementary Information Section 2.5.

We passed the sparse genetic distance matrix to Ripser with arguments `--dim 1 --format sparse --threshold 2`. To enable time series analysis of sequence alignments, the fixed threshold of 2 was replaced by the number of days between the date of the reference sequence and the most recent sequence date in the alignment, plus 1. Ripser returned a list of topological cycles, the edges of which are given by pairs of sequences. Our choice of the `threshold` argument ensured that Ripser only returned those topological cycles in which all edges are of length 1, i.e. SNV cycles.

We observed that Ripser performs faster if sequence alignments are sorted according to collection date in descending (i.e. reverse chronological) order [130].

### Topological recurrence analysis

*SNV cycles* are topological cycles all of whose edges correspond to SNVs (single nucleotide variations) in the sequence alignment. Every edge in an SNV cycle has Hamming length = 1, and the endpoints of an edge correspond to a pair of sequences that differ at exactly one

nucleotide site at which there is no dash "-". Then for every SNV in the sequence alignment (defined with respect to the reference sequence), the *raw topological recurrence index (raw tRI)* of this SNV is defined as the total number of SNV cycles in the sequence alignment that contain an edge corresponding to the SNV (Figure 1c). The idea behind this definition is that the presence of a specific mutation in SNV cycles is indicative of convergent evolution in a twofold sense: (i) the mutation occurs independently twice in different lines of descent within each cycle, and furthermore, (ii) if the mutation is contained in different SNV cycles then it also occurs independently in different lines of descent within the phylogeny.

To compute the raw tRI of a given SNV, we analyzed the edges in the SNV cycles that had previously been found with Ripser. Edges corresponding to the given SNV were counted under the conditions that (i) in each cycle, at most one such edge is counted, and (ii) no edge is counted twice. This analysis was restricted to SNVs satisfying the following properties: (i) one of the two nucleotides involved in the SNV is the same as in the reference sequence at that site, and (ii) the other two nucleotides in the codon containing the SNV are the same as in the reference sequence. These two conditions ensure that the corresponding SAAV is uniquely determined by the SNV with respect to the reference sequence.

In the case of the SARS-CoV-2 and H5N1 alignments which include collection dates, we obtained time series raw tRI values by recovering the temporal information from the edges of SNV cycles as follows. The length of each edge in an SNV cycle encodes the most recent date of the two sequences defining that edge, given as the number of days counted from the date of the reference sequence. This date was then assigned to the tRI signal contributed by the edge.

The final tRI was computed by applying two specific correction steps: First, raw tRI values were adjusted by subtracting 1 from every signal. The motivation for this first correction step is the observation that a mutation present in precisely one SNV cycle in the dataset could have occurred just once in the evolutionary process (Figure 1a). Only the presence of multiple SNV cycles containing a specific mutation is therefore indicative of the independent occurrence of that mutation.

Second, we normalized the adjusted raw tRI by its mean per nucleotide mutation type across four-fold degenerate sites, following the approach taken in [6]. The motivation for this second correction step is the nonuniformity of mutational spectra [37]. We first determined all nucleotide sites on the genome that are four-fold degenerate, i.e. any of the four possible nucleotide variations at that site do not change the corresponding amino acid. Then for each nucleotide mutation type, the mean adjusted raw tRI was computed across all four-fold degenerate SNVs in the sequence alignment (defined with respect to the reference sequence) that matched the given mutation type (Supplementary Figure 8). To obtain the final tRI of an SNV, the raw tRI of that SNV was rescaled by dividing by this mean raw tRI associated with the mutation type of the given SNV. If this mean raw tRI happened to be zero, the rescaling was not well-defined and no rescaling was applied. In the case of time series raw tRI data, the rescaling was applied at every time step. This may lead to spurious tRI signals at the onset of positive rescaling factors.

To focus the topological recurrence analysis on single genes or specific genomic regions (e.g. the SARS-CoV-2 spike gene), we restricted the analysis to sliding windows on the genome. For this, we first created a sub-alignment of genome sequences appropriately truncated to fit the sliding window, and then performed a separate topological recurrence analysis for this particular alignment. This process typically leads to stronger tRI signals and more detailed tRI profiles (Supplementary Figure 9a).

We note that the time series topological recurrence analysis may exhibit an initial warm-up period, during which SNV cycles are formed with some delay. This phenomenon can lead to

spurious tRI growth rates and should be taken into account when interpreting early signals in tRI activity landscapes.

### Downstream analysis

We analyzed time series tRI signals for the SARS-CoV-2 whole genome and spike gene alignments (Supplementary Table 2), both for individual mutations and across genes on the genome. For individual SNVs, we generated *tRI/prevalence time series charts* showing time-dependent tRI, together with the conservative tRI significance level, and time-dependent prevalence of the SNV (Figure 4a). In these charts, the horizontal axis represents the time range at daily resolution. The time-dependent prevalence was computed as the quotient of the number of sequences carrying the SNV by the total number of sequences in the alignment, up to a given point in time.

We generated a *dynamic tRI landscape* from SARS-CoV-2 spike gene time series tRI data to comprehensively map time-dependent tRI signals across this gene (Figure 4b). This tRI landscape is a two-dimensional heat map, where the horizontal axis represents the amino acid positions on the gene (from left to right), the vertical axis represents the time range (from top to bottom) and the heat parameter is the log-transformed tRI. Vertical lines in this heat map represent time-dependent tRI signals at a specific amino acid site. For this, we aggregated tRI signals of non-synonymous SNVs into tRI signals of amino acid sites by summing up tRI signals of all SNVs that give rise to SAAVs at the same amino acid site. We further derived a *dynamic tRI activity landscape* from this tRI landscape, which is a two-dimensional heat map where the heat parameter is the positive part of the tRI growth rate (the first discrete differences of the exponentially weighted moving average of the log-transformed tRI) in the vertical direction. We generated dynamic tRI activity landscapes of the SARS-CoV-2 spike gene (Figure 4c) and UK spike gene (Figure 5c) alignments and for the H5N1 PB2 gene alignment (Figure 5g). For better visualization and to enhance hotspots of convergent evolution, we applied two-dimensional Gaussian smoothing with standard deviation 0.4 (3) in the horizontal (vertical) direction in the spike gene analysis, and with standard deviation 4 (10) in the horizontal (vertical) direction in the UK spike gene and the PB2 gene analyses. For the SARS-CoV-2 spike gene, we took horizontal slices in dynamic tRI (activity) landscapes to extract *snapshots* showing landscapes of tRI (tRI growth rates) across the spike gene at specific points in time. For better visualization we applied one-dimensional Gaussian smoothing with standard deviation 2.

### Statistical analyses

Topological cycles can arise randomly in neutral evolution, causing some background noise in tRI (Figure 2a and Supplementary Figure 5). To quantitatively estimate the effect of this background noise in SARS-CoV-2 whole genome tRI signals, we simulated neutral evolutionary scenarios using a Wright-Fisher forward model of viral evolution using SANTA-SIM v1.0 [42] with (i) fixed parameters: genome length ($30,000$ nt), number of generations ($N = 10,000$), number of sequences sampled from the population per time step ($n = 15$), recombination rate ($\rho = 0$), transition bias (1.0), and (ii) variable parameters: mutation rate per site per generation $\mu$, population size $p$, carrying population $c$, population growth rate per generation $g$. We considered five scenarios: In scenarios I-III we varied the mutation rate $\mu$ under the assumption of fixed population size $p$, while in scenarios IV and V we investigated the effects of logistic growth $g$ of the population with carrying population $c$. For more details see Supplementary Information Section 2.1.

We analyzed the effect of stochastic sequencing errors on tRI by simulating stochastic se-

quencing errors with varying error rates in two reference alignments (Figure 2b and Supplementary Figure 6). As reference alignments, we used a synthetic alignment simulated under neutral evolution with SANTA-SIM v1.0 [42], and a specially tailored real SARS-CoV-2 200k whole genomes alignment. We swept sequencing error rates from $10^{-8}$ to $10^{-3}$ on a logarithmic scale. For more details see Supplementary Information Section 2.2.

We applied a Kolmogorov–Smirnov test to compare the frequency distributions of raw tRI signals in the SARS-CoV-2 whole genome and the spike gene alignments with the frequency distributions obtained from a Monte Carlo permutation test under the null hypothesis that raw tRI signals are distributed uniformly across the genome. For more details see Supplementary Information Section 2.3.

These findings led us to introduce a conservative tRI significance level, by means of a permutation test under the null hypothesis of uniformly distributed raw tRI signals, with $p = 0.05$ as the threshold for statistically significant signals. For more details see Supplementary Information Section 2.4.

### Validation on simulated data

We analyzed the sensitivity of the tRI for positive selection by varying fitness and sampling density in synthetic nucleotide sequence alignments generated with SANTA-SIM v1.0 [42]. For 1.7% of single nucleotide mutations, we varied the fitness parameter between 0.98 and 1.10 in steps of 0.02 (variants with varied fitness), while all other mutations were assigned a fitness parameter of 1.00 (neutral variants). Fitness values in this range are typical for single nucleotide substitutions in RNA viruses [131]. For each scenario, we computed the mean tRI per SNV separately across all variants with varied fitness and across all neutral variants. We then determined the Spearman correlation between fitness and mean tRI signals across variants with varied fitness. Next, we kept the fitness of variants with varied fitness fixed at 1.10 (beneficial variants) and randomly selected $10\%, 20\%, \ldots, 100\%$ of sequences in simulated alignments. For each scenario, we computed the effect size (Cohen's $d$) for the difference between mean tRI across beneficial variants and mean tRI across neutral variants. For more details see Supplementary Information Section 3.1.

We compared raw tRI signals with phylogenetic homoplasy counts using synthetic nucleotide sequence alignments generated with SANTA-SIM v1.0 [42]. For 7.1% of single nucleotide mutations, we increased the fitness parameter to 1.3 (beneficial variants), while all other mutations were assigned a neutral fitness parameter of 1.0 (neutral variants). For each simulated alignment, we (i) removed duplicate sequences and computed raw tRI signals with EVOtRec, and (ii) reconstructed phylogenetic trees with IQ-TREE v3.0.1 [48] and subsequently computed homoplasy counts with TreeTime v0.11.4 [51]. We then determined the Spearman correlation between raw tRI profiles and phylogenetic homoplasy counts, separately for profiles across neutral variants and for profiles across beneficial variants. For more details see Supplementary Information Section 3.2.

### Validation on experimental data

We validated tRI signals of amino acid changes in the SARS-CoV-2 (Omicron) spike gene alignments (Supplementary Table 2) and in the H5N1 HA and PB2 gene alignments (Supplementary Table 3) on published deep mutational scanning (DMS) data. For this, we aggregated tRI signals of non-synonymous single nucleotide variants (SNV) into tRI signals of single amino acid variants (SAAV) by summing up the tRI of all SNVs that give rise to the same SAAV. We computed the time-dependent Pearson correlation between SARS-CoV-2 spike gene tRI and

DMS effect [70]. For more details see Supplementary Information Section 4.1.

Moreover, we compared tRI signals in the SARS-CoV-2 Omicron spike gene alignment with DMS effects for the BA.1 [72], BA.2 [59] and XBB.1.5 [59] sub-lineages (Figure 5b and Supplementary Figure 12a-d), based on $2\times2$ contingency tables for binarized tRI (tRI $> 0$ vs. tRI $= 0$) and binarized DMS effect (DMS effect $> 0$ vs. DMS effect $\leq 0$). In addition, for XBB.1.5 we computed a logistic regression model predicting positive tRI as a function of the three XBB.1.5 DMS effects. We repeated all analyses with tRI replaced by published phylogeny-based fitness effect estimates from [6, 17]. In this case, for XBB.1.5 we computed an ordinary least squares regression model predicting positive fitness effect as a function of the three XBB.1.5 DMS effects. For more details see Supplementary Information Section 4.2.

Similarly, we compared tRI signals obtained from the H5N1 HA and PB2 gene alignments with DMS effects for the HA gene [73] and PB2 gene [74] (Figure 5d), based on $2\times2$ contingency tables for binarized tRI (tRI $> 0$ vs. tRI $= 0$) and binarized DMS effect (DMS effect $> 0$ vs. DMS effect $\leq 0$). For more details see Supplementary Information Section 4.3.

### Comparison with phylogeny-based fitness effect estimates

We compared EVOtRec tRI signals with published phylogeny-based fitness effect estimates [6] for SNVs on the SARS-CoV-2 whole genome and spike gene (Figure 3c-d). We log-transformed positive tRI scores of the SARS-CoV-2 whole genome and spike gene alignments (Supplementary Table 2). Fitness effect estimates were taken from the column "delta_fitness" in [52]. We subsequently computed the Pearson correlation between log-transformed positive tRI and fitness effect estimates, separately for non-synonymous and synonymous SNVs on the whole genome (resp. spike gene) for which both tRI and fitness effect estimate data were available.

### Performance analysis and benchmarking

We analyzed the runtime and memory usage of EVOtRec in comparison with state-of-the-art tools for the reconstruction of phylogenetic trees (Figure 3a-b). We applied the following tools to eight SARS-CoV-2 whole genome alignments with up to 200k distinct sequences, so long as the runtime did not exceed 24 hours: The Python script evotrec.py v1.0, available via `https://github.com/ottamj/evotrec`; IQ-TREE v3.0.1 [48]; VeryFastTree v4.0.5 [49]; CMAPLE v1.1.0 [50] and UShER v0.6.6 [21]. To obtain the sequence alignments, we sorted the SARS-CoV-2 whole genome alignment according to collection date in ascending order and selected the first $n$ sequences, for $n = 1\times10^3, 2\times10^3, 5\times10^3, 1\times10^4, 2\times10^4, 5\times10^4, 1\times10^5, 2\times10^5$. We analyzed the runtime and peak memory usage (RAM), computing mean $\pm$ 95% confidence interval (CI) over 3 independent runs for each tool and each alignment. All computations were performed on a server with Intel Xeon Gold 6230R processors (2.10GHz) and 52 cores. For more details see Supplementary Information Section 5.

### Data availability

The findings of this study are based (i) on genome data of 14,956,592 SARS-CoV-2 nucleotide sequences available from the GISAID EpiCoV Database as of 24 February 2024, via `https://doi.org/10.55876/gis8.240314pc`, (ii) on genome data of 24,141 H5N1 HA gene nucleotide sequences available from the GISAID EpiFlu Database as of 16 November 2025, via `https://doi.org/10.55876/gis8.251116eh`, (iii) on genome data of 23,237 H5N1 PB2 gene nucleotide sequences available from the GISAID EpiFlu Database as of 16 November 2025, via `https://doi.org/10.55876/gis8.251116bc`, and (iv) on genome data of 30,964 HIV-1

Env gene nucleotide sequences available from the European Nucleotide Archive (ENA) as of 1 March 2025, via https://www.ebi.ac.uk/ena/browser/home using the accession numbers provided in Supplementary Table 4. Experimental data (deep mutational scanning) of the SARS-CoV-2 wild-type RBD were taken from the published study [70] and are available from [132]. Experimental data (deep mutational scanning) of the SARS-CoV-2 Omicron spike gene were taken from the published studies [59, 72] and are available from [133] for BA.1, from [134] for BA.2 and from [135] for XBB.1.5. Data on SARS-CoV-2 fitness effect estimates were taken from the published studies [6, 17] and are available from [52, 136] and [137–139]. Experimental data (deep mutational scanning) of the avian influenza A subtype H5N1 HA and PB2 genes were taken from the published studies [73, 74] and are available from [140, 141]. Results of the topological recurrence analysis (tRI and prevalence data) for SARS-CoV-2, H5N1 and HIV-1 are available via https://doi.org/10.5281/zenodo.18332026.

## Code availability

Code of the EVOtRec pipeline is available via https://github.com/ottamj/evotrec. Code of Hammingdist and Ripser is available via https://github.com/ssciwr/hammingdist and https://github.com/Ripser/ripser/tree/tight-representative-cycles. All other code is available from the corresponding authors upon request.

## Acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative and the European Nucleotide Archive (ENA), on which this research is based. This work was supported by the de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE (Forschungszentrum Jülich and W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014, W-de.NBI-016, W-de.NBI-022). The authors further acknowledge support from the Interdisciplinary Center for Scientific Computing at Heidelberg University and the Scientific Software Center of Heidelberg University. This research was supported by the DFG Collaborative Research Center SFB/TRR 109 "Discretization in Geometry and Dynamics". M.B. was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster). L.H. thanks the Evangelisches Studienwerk Villigst for their support. A.O. and M.N. were supported by the Vector Stiftung. M.N. was supported by the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health". A.O. acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 281869850 (RTG 2229). This research was funded by the Federal Ministry of Education and Research (BMBF) and the Baden-Württemberg Ministry of Science as part of the Excellence Strategy of the German Federal and State Governments. Work conducted by Z.A. at the Wellcome Sanger Institute was supported by core funding from the Wellcome Trust (Grant No. 220540/Z/20/A).

## Author contributions

M.B., L.H., M.N., Z.A., J.P.-G., R.R., A.O. designed the study; A.O. managed and supervised the project; M.B., L.H., A.O. curated data; M.B., L.H., M.N, Z.A., M.C., J.P.-G., A.O. performed computational analyses; M.B., L.H., M.N., U.B., A.O. developed and implemented

software; M.B., L.H., A.O. acquired computing resources; M.B., L.H., M.N., Z.A., U.B., R.R., A.O. wrote the manuscript; M.B., L.H., M.N., A.O. designed and implemented the EVOtRec pipeline; M.N. designed the MuRiT algorithm for time series topological recurrence analysis; A.O. designed and implemented tRI/prevalence charts and dynamic tRI landscapes; M.B., M.N., A.O. implemented the MuRiT algorithm in the EVOtRec pipeline; U.B. designed and implemented exhaustive reduction for cycle localization in Ripser; M.B designed and implemented analyses of simulated evolutionary scenarios; M.B., A.O. performed simulations; A.O. designed, implemented and performed benchmarking and comparison with phylogenetic methods; M.B, A.O. performed analyses of simulated evolutionary scenarios; M.N, A.O. designed stochastic sequencing error analysis; A.O. performed stochastic sequencing error analysis; M.N. designed and implemented ancestral sequence reconstruction for alignments with low sampling rate; M.N. designed and implemented study of effect of ambiguity of cycle representatives; A.O. analyzed topological signals at four-fold degenerate sites; A.O. performed time series topological recurrence analyses of genomic datasets; M.N., A.O. designed comparison with experimental data; A.O. performed comparison with experimental data; M.B., L.H., M.N., Z.A., U.B., A.O. designed and created figures; all authors contributed to the final version of the paper.

## Competing interests

R.R. is a founder of Genotwin, he is member of the Scientific Advisory Board of AimedBio and consults for Arquimea Research. The other authors declare no competing interests.

## Supplementary Material

- Supplementary Tables
- Supplementary Figures
- Supplementary Information

**Correspondence and requests for materials** should be addressed to M.B., M.N., R.R. or A.O.

## References

1. Sackton, T. B. & Clark, N. Convergent evolution in the genomics era: new insights and directions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374,** 20190102 (2019). doi:10.1098/rstb.2019.0102.
2. Gatt, Y. E. & Margalit, H. Common Adaptive Strategies Underlie Within-Host Evolution of Bacterial Pathogens. *Molecular Biology and Evolution* **38,** 1101–1121 (2020). doi:10.1093/molbev/msaa278.
3. Gutierrez, B., Escalera-Zamudio, M. & Pybus, O. G. Parallel molecular evolution and adaptation in viruses. *Current Opinion in Virology* **34,** 90–96 (2019). doi:10.1016/j.coviro.2018.12.006.
4. Koonin, E. V., Dolja, V. V. & Krupovic, M. The logic of virus evolution. *Cell Host & Microbe* **30,** 917–929 (2022). doi:10.1016/j.chom.2022.06.008.
5. Hodcroft, E. B., De Maio, N., Lanfear, R., *et al.* Want to Track Pandemic Variants Faster? Fix the Bioinformatics Bottleneck. *Nature* **591,** 30–33 (2021). doi:10.1038/d41586-021-00525-x.
6. Bloom, J. D. & Neher, R. A. Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evolution* **9** (2023). doi:10.1093/ve/vead055.

7. Wei, H. & Li, X. Deep mutational scanning: A versatile tool in systematically mapping genotypes to phenotypes. *Frontiers in Genetics* **14** (2023). doi:10.3389/fgene.2023.1087267.

8. Morel, B., Barbera, P., Czech, L., *et al.* Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. *Molecular Biology and Evolution* (ed Malik, H.) msaa314 (2020). doi:10.1093/molbev/msaa314.

9. Torada, L., Lorenzon, L., Beddis, A., *et al.* ImaGene: a convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics* **20** (2019). doi:10.1186/s12859-019-2927-x.

10. Thadani, N. N., Gurev, S., Notin, P., *et al.* Learning from prepandemic data to forecast viral escape. *Nature* **622,** 818–825 (2023). doi:10.1038/s41586-023-06617-0.

11. West, C., Walker, C. R., Arasti, S., *et al.* Detecting interspecific positive selection using convolutional neural networks. *Molecular Biology and Evolution* (ed Ouangraoua, A.) (2025). doi:10.1093/molbev/msaf154.

12. Obermeyer, F., Jankowiak, M., Barkas, N., *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376,** 1327–1332 (2022). doi:10.1126/science.abm1208.

13. Maher, M. C., Bartha, I., Weaver, S., *et al.* Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Science Translational Medicine* **14** (2022). doi:10.1126/scitranslmed.abk3445.

14. Lee, B., Quadeer, A. A., Sohail, M. S., *et al.* Inferring effects of mutations on SARS-CoV-2 transmission from genomic surveillance data. *Nature Communications* **16** (2025). doi:10.1038/s41467-024-55593-0.

15. Zhao, L., Wymant, C., Blanquart, F., *et al.* Phylogenetic estimation of the viral fitness landscape of HIV-1 set-point viral load. *Virus Evolution* **8** (2022). doi:10.1093/ve/veac022.

16. Meijers, M., Ruchnewitz, D., Eberhardt, J., *et al.* Population immunity predicts evolutionary trajectories of SARS-CoV-2. *Cell* **186,** 5151–5164.e13 (2023). doi:10.1016/j.cell.2023.09.022.

17. Haddox, H. K., Angehrn, G., Sesta, L., *et al.* The mutation rate of SARS-CoV-2 is highly variable between sites and is influenced by sequence context, genomic region, and RNA structure. *Nucleic Acids Research* **53** (2025). doi:10.1093/nar/gkaf503.

18. Lefrancq, N., Duret, L., Bouchez, V., *et al.* Learning the fitness dynamics of pathogens from phylogenies. *Nature* **637,** 683–690 (2025). doi:10.1038/s41586-024-08309-9.

19. Maeso, I., Roy, S. W. & Irimia, M. Widespread Recurrent Evolution of Genomic Features. *Genome Biology and Evolution* **4,** 486–500 (2012). doi:10.1093/gbe/evs022.

20. Hadfield, J., Megill, C., Bell, S. M., *et al.* Nextstrain: Real-Time Tracking of Pathogen Evolution. *Bioinformatics* **34,** 4121–4123 (2018). doi:10.1093/bioinformatics/bty407.

21. Turakhia, Y., Thornlow, B., Hinrichs, A. S., *et al.* Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics* **53,** 809–816 (2021). doi:10.1038/s41588-021-00862-7.

22. De Maio, N., Kalaghatgi, P., Turakhia, Y., *et al.* Maximum likelihood pandemic-scale phylogenetics. *Nature Genetics* **55,** 746–752 (2023). doi:10.1038/s41588-023-01368-0.

23. Ye, C., Thornlow, B., Hinrichs, A., *et al.* matOptimize: a parallel tree optimization method enables online phylogenetics for SARS-CoV-2. *Bioinformatics* **38,** 3734–3740 (2022). doi:10.1093/bioinformatics/btac401.

24. Steenwyk, J. L., Li, Y., Zhou, X., *et al.* Incongruence in the phylogenomics era. *Nature Reviews Genetics* **24,** 834–850 (2023). doi:10.1038/s41576-023-00620-x.

25. Kapli, P., Flouri, T. & Telford, M. J. Systematic errors in phylogenetic trees. *Current Biology* **31,** R59–R64 (2021). doi:10.1016/j.cub.2020.11.043.

26. Turakhia, Y., De Maio, N., Thornlow, B., *et al.* Stability of SARS-CoV-2 Phylogenies. *PLOS Genetics* **16,** e1009175 (2020). doi:10.1371/journal.pgen.1009175.

27. Bjornson, S., Verbruggen, H., Upham, N. S. & Steenwyk, J. L. Reticulate evolution: Detection and utility in the phylogenomics era. *Molecular Phylogenetics and Evolution* **201,** 108197 (2024). doi:10.1016/j.ympev.2024.108197.

28. Stott, C. M. & Bobay, L.-M. Impact of homologous recombination on core genome phylogenies. *BMC Genomics* **21** (2020). doi:10.1186/s12864-020-07262-x.

29. Posada, D. & Crandall, K. A. The Effect of Recombination on the Accuracy of Phylogeny Estimation. *Journal of Molecular Evolution* **54,** 396–402 (2002). doi:10.1007/s00239-001-0034-9.

30. Schierup, M. H. & Hein, J. Consequences of Recombination on Traditional Phylogenetic Analysis. *Genetics* **156,** 879–891 (2000). doi:10.1093/genetics/156.2.879.

31. Crispell, J., Balaz, D. & Gordon, S. V. HomoplasyFinder: A Simple Tool to Identify Homoplasies on a Phylogeny. *Microbial Genomics* **5** (2019). doi:10.1099/mgen.0.000245.

32. Carlsson, G. Topology and data. *Bulletin of the American Mathematical Society* **46,** 255–308 (2009). doi:10.1090/s0273-0979-09-01249-x. http://dx.doi.org/10.1090/S0273-0979-09-01249-X.

33. Chan, J. M., Carlsson, G. & Rabadan, R. Topology of Viral Evolution. *Proceedings of the National Academy of Sciences* **110,** 18566–18571 (2013). doi:10.1073/pnas.1313480110.

34. Cámara, P. G., Levine, A. J. & Rabadán, R. Inference of Ancestral Recombination Graphs through Topological Data Analysis. *PLOS Computational Biology* **12** (ed Pond, S. L. K.) e1005071 (2016). doi:10.1371/journal.pcbi.1005071.

35. Rabadán, R. & Blumberg, A. J. *Topological Data Analysis for Genomics and Evolution* doi:10.1017/9781316671665 (Cambridge University Press, 2019).

36. Stayton, C. T. The definition, recognition, and interpretation of convergent evolution, and two new measures for quantifying and assessing the significance of convergence. *Evolution* **69,** 2140–2153 (2015). doi:10.1111/evo.12729. http://dx.doi.org/10.1111/evo.12729.

37. Bloom, J. D., Beichman, A. C., Neher, R. A. & Harris, K. Evolution of the SARS-CoV-2 Mutational Spectrum. *Molecular Biology and Evolution* **40** (ed Hepp, C.) (2023). doi:10.1093/molbev/msad085.

38. De Maio, N., Walker, C. R., Turakhia, Y., *et al.* Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. *Genome Biology and Evolution* **13** (2021). doi:10.1093/gbe/evab087.

39. Keegan, L. & Kempf, D. *Hammingdist: A Fast Tool to Calculate Hamming Distances* version 1.4.0. 2025. https://github.com/ssciwr/hammingdist visited on 2025-11-01.

40. Bauer, U. Ripser: efficient computation of Vietoris-Rips persistence barcodes. *Journal of Applied and Computational Topology* (2021). doi:10.1007/s41468-021-00071-5.

41. Neumann, M., Bleher, M., Hahn, L., *et al.* MuRiT: efficient computation of pathwise persistence barcodes in multi-filtered flag complexes via Vietoris-Rips transformations. *arXiv* (2022). doi:10.48550/arXiv.2207.03394.

42. Jariani, A., Warth, C., Deforche, K., *et al.* SANTA-SIM: Simulating Viral Sequence Evolution Dynamics under Selection and Recombination. *Virus Evolution* **5** (2019). doi:10.1093/ve/vez003.

43. Bull, R. A., Adikari, T. N., Ferguson, J. M., *et al.* Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nature Communications* **11** (2020). doi:10.1038/s41467-020-20075-6.

44. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics* **3** (2021). doi:10.1093/nargab/lqab019.

45. De Maio, N., Walker, C., Borges, R., *et al. Issues with SARS-CoV-2 Sequencing Data - SARS-CoV-2 Coronavirus / nCoV-2019 Genomic Epidemiology* Virological. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473 visited on 2024-03-24.

46. Hunt, M., Hinrichs, A. S., Anderson, D., *et al.* Addressing pandemic-wide systematic errors in the SARS-CoV-2 phylogeny. *bioRxiv* (2024). doi:10.1101/2024.04.29.591666.

47. Turakhia, Y., De Maio, N., Thornlow, B., *et al. Problematic Sites SARS-CoV-2* Github. https://github.com/W-L/ProblematicSites_SARS-CoV2/blob/master/problematic_sites_sarsCov2.vcf visited on 2024-08-31.

48. Wong, T., Ly-Trong, N., Ren, H., *et al.* IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models. *EcoEvoRxiv* (2025). doi:10.32942/x2p62n.

49. Piñeiro, C. & Pichel, J. C. Efficient phylogenetic tree inference for massive taxonomic datasets: harnessing the power of a server to analyze 1 million taxa. *GigaScience* **13** (2024). doi:10.1093/gigascience/giae055.

50. Ly-Trong, N., Bielow, C., De Maio, N. & Minh, B. Q. CMAPLE: Efficient Phylogenetic Inference in the Pandemic Era. *Molecular Biology and Evolution* **41** (ed Rzhetsky, A.) (2024). doi:10.1093/molbev/msae134.

51. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-Likelihood Phylodynamic Analysis. *Virus Evolution* **4** (2018). doi:10.1093/ve/vex042.

52. Bloom, J. D. & Neher, R. A. *Fitness effects of SARS-CoV-2 amino-acid mutations estimated from observed versus expected mutation counts* Github. https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/results_gisaid_2024-04-24/nt_fitness/ntmut_fitness_all.csv visited on 2024-06-30.

53. Van Dorp, L., Acman, M., Richard, D., *et al.* Emergence of Genomic Diversity and Recurrent Mutations in SARS-CoV-2. *Infection, Genetics and Evolution* **83,** 104351 (2020). doi:10.1016/j.meegid.2020.104351.

54. Rochman, N. D., Wolf, Y. I., Faure, G., *et al.* Ongoing Global and Regional Adaptive Evolution of SARS-CoV-2. *Proceedings of the National Academy of Sciences* **118,** e2104241118 (2021). doi:10.1073/pnas.2104241118.

55. Zahradník, J., Nunvar, J. & Schreiber, G. Perspectives: SARS-CoV-2 Spike Convergent Evolution as a Guide to Explore Adaptive Advantage. *Frontiers in Cellular and Infection Microbiology* **12** (2022). doi:10.3389/fcimb.2022.748948.

56. Kistler, K. E., Huddleston, J. & Bedford, T. Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. *Cell Host & Microbe* **30,** 545–555.e4 (2022). doi:10.1016/j.chom.2022.03.018.

57. Markov, P. V., Ghafari, M., Beer, M., *et al.* The evolution of SARS-CoV-2. *Nature Reviews Microbiology* **21,** 361–379 (2023). doi:10.1038/s41579-023-00878-2.

58. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1,** 33–46 (2017). doi:10.1002/gch2.1018.

59. Dadonaite, B., Brown, J., McMahon, T. E., *et al.* Spike deep mutational scanning helps predict success of SARS-CoV-2 clades. *Nature* **631,** 617–626 (2024). doi:10.1038/s41586-024-07636-1.

60. Roemer, C. & Neher, R. *SARS-CoV-2 variant reports (2023-07-26)* GitHub. https://github.com/neherlab/SARS-CoV-2_variant-reports/blob/main/reports/variant_report_2023-07-26.md visited on 2024-06-26.

61. Collier, D. A., De Marco, A., Ferreira, I. A. T. M., *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* **593,** 136–141 (2021). doi:10.1038/s41586-021-03412-7.

62. Liu, Z., VanBlargan, L. A., Bloyet, L.-M., *et al.* Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host & Microbe* **29,** 477–488.e4 (2021). doi:10.1016/j.chom.2021.01.014.

63. Liu, H., Zhang, Q., Wei, P., *et al.* The basis of a more contagious 501Y.V1 variant of SARS-CoV-2. *Cell Research* **31,** 720–722 (2021). doi:10.1038/s41422-021-00496-8.

64. Liu, Y., Liu, J., Plante, K. S., *et al.* The N501Y spike substitution enhances SARS-CoV-2 infection and transmission. *Nature* **602,** 294–299 (2021). doi:10.1038/s41586-021-04245-0.

65. Martin, D. P., Weaver, S., Tegally, H., *et al.* The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184,** 5189–5200.e7 (2021). doi:10.1016/j.cell.2021.09.003.

66. Liu, Y., Liu, J., Johnson, B. A., *et al.* Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant. *Cell Reports* **39,** 110829 (2022). doi:10.1016/j.celrep.2022.110829.

67. Cao, Y., Jian, F., Wang, J., *et al.* Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. *Nature* (2022). doi:10.1038/s41586-022-05644-7.

68. Ito, J., Suzuki, R., Uriu, K., *et al.* Convergent evolution of SARS-CoV-2 Omicron subvariants leading to the emergence of BQ.1.1 variant. *Nature Communications* **14** (2023). doi:10.1038/s41467-023-38188-z.

69. Jian, F., Wang, J., Yisimayi, A., *et al.* Evolving antibody response to SARS-CoV-2 antigenic shift from XBB to JN.1. *Nature* (2024). doi:10.1038/s41586-024-08315-x. http://dx.doi.org/10.1038/s41586-024-08315-x.

70. Starr, T. N., Greaney, A. J., Hannon, W. W., *et al.* Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science* **377,** 420–424 (2022). doi:10.1126/science.abo7896.

71. Carabelli, A. M., Peacock, T. P., Thorne, L. G., *et al.* SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nature Reviews Microbiology* (2023). doi:10.1038/s41579-022-00841-7.

72. Dadonaite, B., Crawford, K. H., Radford, C. E., *et al.* A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. *Cell* **186,** 1263–1278.e20 (2023). doi:10.1016/j.cell.2023.02.001.

73. Dadonaite, B., Ahn, J. J., Ort, J. T., *et al.* Deep mutational scanning of H5 hemagglutinin to inform influenza virus surveillance. *PLOS Biology* **22,** e3002916 (2024). doi:10.1371/journal.pbio.3002916.

74. Soh, Y. S., Moncla, L. H., Eguia, R., *et al.* Comprehensive mapping of adaptation of the avian influenza polymerase protein PB2 to humans. *eLife* **8** (2019). doi:10.7554/elife.45079.

75. Good, M. R., Fernández-Quintero, M. L., Ji, W., *et al.* A single mutation in dairy cow-associated H5N1 viruses increases receptor binding breadth. *Nature Communications* **15** (2024). doi:10.1038/s41467-024-54934-3.

76. Carrique, L., Fan, H., Walker, A. P., *et al.* Host ANP32A mediates the assembly of the influenza virus replicase. *Nature* **587,** 638–643 (2020). doi:10.1038/s41586-020-2927-z.

77. Peacock, T. P., Sheppard, C. M., Lister, M. G., *et al.* Mammalian ANP32A and ANP32B Proteins Drive Differential Polymerase Adaptations in Avian Influenza Virus. *Journal of Virology* **97** (ed Jung, J. U.) (2023). doi:10.1128/jvi.00213-23.

78. Dholakia, V., Quantrill, J. L., Richardson, S. A. S., *et al.* Polymerase mutations underlie early adaptation of H5N1 influenza virus to dairy cattle and other mammals. *Nature Communications* (2026). doi:10.1038/s41467-026-68306-6.

79. Giorgio, S. D., Martignano, F., Torcia, M. G., *et al.* Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Science Advances* **6** (2020). doi:10.1126/sciadv.abb5813.

80. Graudenzi, A., Maspero, D., Angaroni, F., *et al.* Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. *iScience* **24,** 102116 (2021). doi:10.1016/j.isci.2021.102116.

81. Simmonds, P. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **5** (ed Schwemmle, M.) (2020). doi:10.1128/msphere.00408-20.

82. Ratcliff, J. & Simmonds, P. Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. *Virology* **556,** 62–72 (2021). doi:10.1016/j.virol.2020.12.018.

83. Zhao, L., Hall, M., de Cesare, M., *et al.* The mutational spectrum of SARS-CoV-2 genomic and antigenomic RNA. *Proceedings of the Royal Society B: Biological Sciences* **289** (2022). doi:10.1098/rspb.2022.1747.

84. Moulana, A., Dupic, T., Phillips, A. M., *et al.* Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA.1. *Nature Communications* **13** (2022). doi:10.1038/s41467-022-34506-z.

85. Moulana, A., Dupic, T., Phillips, A. M., *et al.* The landscape of antibody binding affinity in SARS-CoV-2 Omicron BA.1 evolution. *eLife* **12** (2023). doi:10.7554/elife.83442.

86. Javanmardi, K., Segall-Shapiro, T. H., Chou, C.-W., *et al.* Antibody escape and cryptic cross-domain stabilization in the SARS-CoV-2 Omicron spike protein. *Cell Host & Microbe* **30,** 1242–1254.e6 (2022). doi:10.1016/j.chom.2022.07.016.

87. Vöhringer, H. S., Sanderson, T., Sinnott, M., *et al.* Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature* **600,** 506–511 (2021). doi:10.1038/s41586-021-04069-y.

88. Du Plessis, L., McCrone, J. T., Zarebski, A. E., *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371,** 708–712 (2021). doi:10.1126/science.abf2946.

89. Huang, Y., Yang, C., Xu, X.-f., *et al.* Structural and Functional Properties of SARS-CoV-2 Spike Protein: Potential Antivirus Drug Development for COVID-19. *Acta Pharmacologica Sinica* **41,** 1141–1149 (2020). doi:10.1038/s41401-020-0485-4.

90. Johnson, B. A., Xie, X., Bailey, A. L., *et al.* Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature* **591,** 293–299 (2021). doi:10.1038/s41586-021-03237-4.

91. Peacock, T. P., Goldhill, D. H., Zhou, J., *et al.* The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. *Nature Microbiology* **6,** 899–909 (2021). doi:10.1038/s41564-021-00908-w.

92. Zahradník, J., Marciano, S., Shemesh, M., *et al.* SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nature Microbiology* **6,** 1188–1198 (2021). doi:10.1038/s41564-021-00954-4.

93. Meng, B., Datir, R., Choi, J., *et al.* SARS-CoV-2 spike N-terminal domain modulates TMPRSS2-dependent viral entry and fusogenicity. *Cell Reports* **40,** 111220 (2022). doi:10.1016/j.celrep.2022.111220.

94. Volz, E., Mishra, S., Chand, M., *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593,** 266–269 (2021). doi:10.1038/s41586-021-03470-x.

95. Harvey, W. T., Carabelli, A. M., Jackson, B., *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* **19,** 409–424 (2021). doi:10.1038/s41579-021-00573-0.

96. Piccoli, L., Park, Y.-J., Tortorici, M. A., *et al.* Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell* **183,** 1024–1042.e21 (2020). doi:10.1016/j.cell.2020.09.037.

97. Dejnirattisai, W., Zhou, D., Ginn, H. M., *et al.* The antigenic anatomy of SARS-CoV-2 receptor binding domain. *Cell* **184,** 2183–2200.e22 (2021). doi:10.1016/j.cell.2021.02.032.

98. Chen, Y., Zhao, X., Zhou, H., *et al.* Broadly neutralizing antibodies to SARS-CoV-2 and other human coronaviruses. *Nature Reviews Immunology* **23,** 189–199 (2022). doi:10.1038/s41577-022-00784-3.

99. Klinakis, A., Cournia, Z. & Rampias, T. N-terminal domain mutations of the spike protein are structurally implicated in epitope recognition in emerging SARS-CoV-2 strains. *Computational and Structural Biotechnology Journal* **19,** 5556–5567 (2021). doi:10.1016/j.csbj.2021.10.004.

100. Barnes, C. O., Jette, C. A., Abernathy, M. E., *et al.* SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588,** 682–687 (2020). doi:10.1038/s41586-020-2852-1.

101. Lei, R., Qing, E., Odle, A., *et al.* Functional and antigenic characterization of SARS-CoV-2 spike fusion peptide by deep mutational scanning. *Nature Communications* **15** (2024). doi:10.1038/s41467-024-48104-8.

102. Jian, F., Feng, L., Yang, S., *et al.* Convergent evolution of SARS-CoV-2 XBB lineages on receptor-binding domain 455–456 synergistically enhances antibody evasion and ACE2 binding. *PLOS Pathogens* **19** (ed Wan, X.-F.) e1011868 (2023). doi:10.1371/journal.ppat.1011868.

103. Roemer, C., Sheward, D. J., Hisner, R., *et al.* SARS-CoV-2 evolution in the Omicron era. *Nature Microbiology* **8,** 1952–1959 (2023). doi:10.1038/s41564-023-01504-w.

104. Willett, B. J., Grove, J., MacLean, O. A., *et al.* SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. *Nature Microbiology* **7,** 1161–1179 (2022). doi:10.1038/s41564-022-01143-7.

105. Johnson, M. S., Reddy, G. & Desai, M. M. Epistasis and evolution: recent advances and an outlook for prediction. *BMC Biology* **21** (2023). doi:10.1186/s12915-023-01585-3.

106. Yoshida, I., Sugiura, W., Shibata, J., *et al.* Change of Positive Selection Pressure on HIV-1 Envelope Gene Inferred by Early and Recent Samples. *PLoS ONE* **6** (ed Liberles, D.) e18630 (2011). doi:10.1371/journal.pone.0018630.

107. Witvrouw, M., Fikkert, V., Hantson, A., *et al.* Resistance of Human Immunodeficiency Virus Type 1 to the High-Mannose Binding Agents Cyanovirin N and Concanavalin A. *Journal of Virology* **79,** 7777–7784 (2005). doi:10.1128/jvi.79.12.7777-7784.2005.

108. Watkins, B. A., Reitz, M. S., Wilson, C. A., *et al.* Immune escape by human immunodeficiency virus type 1 from neutralizing antibodies: evidence for multiple pathways. *Journal of Virology* **67,** 7493–7500 (1993). doi:10.1128/jvi.67.12.7493-7500.1993.

109. Yoshimura, K., Harada, S., Shibata, J., *et al.* Enhanced Exposure of Human Immunodeficiency Virus Type 1 Primary Isolate Neutralization Epitopes through Binding of CD4 Mimetic Compounds. *Journal of Virology* **84,** 7558–7568 (2010). doi:10.1128/jvi.00227-10.

110. Galvez, N. M., Cao, Y., Nitido, A. D., *et al.* HIV broadly neutralizing antibody escapability drives the therapeutic efficacy of vectored immunotherapy. *bioRxiv* (2024). doi:10.1101/2024.07.11.603156.

111. Upadhyay, C., Feyznezhad, R., Yang, W., *et al.* Alterations of HIV-1 envelope phenotype and antibody-mediated neutralization by signal peptide mutations. *PLOS Pathogens* **14,** e1006812 (2018). doi:10.1371/journal.ppat.1006812.

112. Haddox, H. K., Dingens, A. S. & Bloom, J. D. Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. *PLOS Pathogens* **12** (ed Swanstrom, R.) e1006114 (2016). doi:10.1371/journal.ppat.1006114.

113. Mostafa, A., Naguib, M. M., Nogales, A., *et al.* Avian influenza A (H5N1) virus in dairy cattle: origin, evolution, and cross-species transmission. *mBio* **15** (ed Yount, J.) (2024). doi:10.1128/mbio.02542-24.

114. Peacock, T. P., Moncla, L., Dudas, G., *et al.* The global H5N1 influenza panzootic in mammals. *Nature* **637,** 304–313 (2024). doi:10.1038/s41586-024-08054-z.

115. Capelastegui, F. & Goldhill, D. H. H5N1 2.3.4.4b: a review of mammalian adaptations and risk of pandemic emergence. *Journal of General Virology* **106** (2025). doi:10.1099/jgv.0.002109.

116. Nguyen, T.-Q., Hutter, C. R., Markin, A., *et al.* Emergence and interstate spread of highly pathogenic avian influenza A(H5N1) in dairy cattle in the United States. *Science* **388** (2025). doi:10.1126/science.adq0900.

117. Yamaji, R., Yamada, S., Le, M. Q., *et al.* Identification of PB2 Mutations Responsible for the Efficient Replication of H5N1 Influenza Viruses in Human Lung Epithelial Cells. *Journal of Virology* **89** (ed Dermody, T. S.) 3947–3956 (2015). doi:10.1128/jvi.03328-14.

118. Fan, S., Hatta, M., Kim, J. H., *et al.* Novel residues in avian influenza virus PB2 protein affect virulence in mammalian hosts. *Nature Communications* **5** (2014). doi:10.1038/ncomms6021.

119. Gu, C., Maemura, T., Guan, L., *et al.* A human isolate of bovine H5N1 is transmissible and lethal in animal models. *Nature* **636,** 711–718 (2024). doi:10.1038/s41586-024-08254-7.

120. Gao, Y., Zhang, Y., Shinya, K., *et al.* Identification of Amino Acids in HA and PB2 Critical for the Transmission of H5N1 Avian Influenza Viruses in a Mammalian Host. *PLoS Pathogens* **5** (ed Sherry, B.) e1000709 (2009). doi:10.1371/journal.ppat.1000709.

121. Krammer, F., Hermann, E. & Rasmussen, A. L. Highly pathogenic avian influenza H5N1: history, current situation, and outlook. *Journal of Virology* **99** (ed Favoreel, H. W.) (2025). doi:10.1128/jvi.02209-24.

122. Pekar, J. E., Peacock, T. P., Moncla, L., *et al. Timing and molecular characterisation of the transmission to cattle of H5N1 influenza A virus genotype D1.1, clade 2.3.4.4b* Virological. https://virological.org/t/timing-and-molecular-characterisation-of-the-transmission-to-cattle-of-h5n1-influenza-a-virus-genotype-d1-1-clade-2-3-4-4b/991 visited on 2025-11-26.

123. Arrildt, K. T., Joseph, S. B. & Swanstrom, R. The HIV-1 Env Protein: A Coat of Many Colors. *Current HIV/AIDS Reports* **9,** 52–63 (2012). doi:10.1007/s11904-011-0107-3.

124. Gromov, M. in *Essays in Group Theory* (ed Gersten, S. M.) 75–263 (Springer, New York, NY, 1987). doi:10.1007/978-1-4613-9586-7_3.

125. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30,** 3059–3066 (2002). doi:10.1093/nar/gkf436.

126. Leinonen, R., Akhtar, R., Birney, E., *et al.* The European Nucleotide Archive. *Nucleic Acids Research* **39,** D28–D31 (2010). doi:10.1093/nar/gkq967.

127. Bauer, U. *Ripser* Github. https://github.com/Ripser/ripser/tree/tight-representative-cycles visited on 2024-01-15.

128. Edelsbrunner, H. & Ölsböck, K. Holes and Dependences in an Ordered Complex. *Computer Aided Geometric Design* **73,** 1–15 (2019). doi:10.1016/j.cagd.2019.06.003.

129. Zomorodian, A. & Carlsson, G. Computing Persistent Homology. *Discrete & Computational Geometry* **33,** 249–274 (2005). doi:10.1007/s00454-004-1146-y.

130. Bauer, U. & Roll, F. *Gromov Hyperbolicity, Geodesic Defect, and Apparent Pairs in Vietoris-Rips Filtrations* in (Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022). doi:10.4230/LIPICS.SOCG.2022.15.

131. Sanjuán, R., Moya, A. & Elena, S. F. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences* **101,** 8396–8401 (2004). doi:10.1073/pnas.0400146101.

132. Starr, T. N., Greaney, A. J., Stewart, C. M., *et al. Deep mutational scanning of the SARS-CoV-2 RBD in variant backgrounds* Github. https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS_variants/blob/main/results/final_variant_scores/final_variant_scores.csv visited on 2024-08-23.

133. Dadonaite, B., Crawford, K. H., Radford, C. E., *et al. Deep mutational scanning of SARS-CoV-2 Omicron BA.1 spike using a barcoded lentiviral platform* Github. https://github.com/dms-vep/SARS-CoV-2_Omicron_BA.1_spike_DMS_mAbs/blob/main/results/muteffects_functional/muteffects_observed.csv visited on 2024-08-23.

134. Dadonaite, B., Brown, J., McMahon, T. E., *et al. Deep mutational scanning of SARS-CoV-2 BA.2 spike* Github. https://github.com/dms-vep/SARS-CoV-2_Omicron_BA.2_spike_ACE2_binding/blob/main/results/summaries/summary.csv visited on 2024-08-23.

135. Dadonaite, B., Brown, J., McMahon, T. E., *et al. Deep mutational scanning of SARS-CoV-2 XBB.1.5 spike* Github. https://github.com/dms-vep/SARS-CoV-2_XBB.1.5_spike_DMS/blob/main/results/summaries/summary.csv visited on 2024-08-23.

136. Bloom, J. D. & Neher, R. A. *Fitness effects of SARS-CoV-2 amino-acid mutations estimated from observed versus expected mutation counts* Github. https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/results_gisaid_2024-04-24/aa_fitness/aamut_fitness_by_clade.csv visited on 2025-06-15.

137. Haddox, H. K., Angehrn, G., Sesta, L., *et al. Fitness effects of nucleotide and amino acid mutations updated with novel estimates of neutral rates* Github. https://github.com/neherlab/SARS2-mut-fitness-v2/blob/main/results/aamut_fitness/21K_aamut_fitness.csv visited on 2025-06-15.

138. Haddox, H. K., Angehrn, G., Sesta, L., *et al. Fitness effects of nucleotide and amino acid mutations updated with novel estimates of neutral rates* Github. https://github.com/neherlab/SARS2-mut-fitness-v2/blob/main/results/aamut_fitness/BA.2_aamut_fitness.csv visited on 2025-06-15.

139. Haddox, H. K., Angehrn, G., Sesta, L., *et al. Fitness effects of nucleotide and amino acid mutations updated with novel estimates of neutral rates* Github. https://github.com/neherlab/SARS2-mut-fitness-v2/blob/main/results/aamut_fitness/XBB_aamut_fitness.csv visited on 2025-06-15.

140. Dadonaite, B., Ahn, J. J., Ort, J. T., *et al. Deep mutational scanning of H5N1 influenza haemagglutinin using a barcoded lentiviral platform* Github. https://github.com/dms-vep/Flu_H5_American-Wigeon_South-Carolina_2021-H5N1_DMS/blob/main/results/summaries/phenotypes.csv visited on 2025-03-01.

141. Soh, Y. S., Moncla, L. H., Eguia, R., *et al.* *Deep mutational scanning of avian influenza PB2 to comprehensively map adaptation to the human host* Github. https://github.com/jbloomlab/PB2-DMS/blob/master/results/diffsel/summary_prefs_effects_diffsel.csv visited on 2025-03-01.

142. Férir, G., Gordts, S. & Schols, D. HIV-1 and Its Resistance to Peptidic Carbohydrate-Binding Agents (CBAs): An Overview. *Molecules* **19,** 21085–21112 (2014). doi:10.3390/molecules191221085.

143. Mahalanabis, M., Jayaraman, P., Miura, T., *et al.* Continuous Viral Escape and Selection by Autologous Neutralizing Antibodies in Drug-Naïve Human Immunodeficiency Virus Controllers. *Journal of Virology* **83,** 662–672 (2009). doi:10.1128/jvi.01328-08.

144. Crooks, E. T., Almanza, F., D'Addabbo, A., *et al.* Engineering well-expressed, V2-immunofocusing HIV-1 envelope glycoprotein membrane trimers for use in heterologous prime-boost vaccine regimens. *PLOS Pathogens* **17** (ed Swanstrom, R.) e1009807 (2021). doi:10.1371/journal.ppat.1009807.

145. Li, H., Xu, C.-F., Blais, S., *et al.* Proximal Glycans Outside of the Epitopes Regulate the Presentation of HIV-1 Envelope gp120 Helper Epitopes. *The Journal of Immunology* **182,** 6369–6378 (2009). doi:10.4049/jimmunol.0804287.

146. Anastassopoulou, C. G., Ketas, T. J., Klasse, P. J. & Moore, J. P. Resistance to CCR5 inhibitors caused by sequence changes in the fusion peptide of HIV-1 gp41. *Proceedings of the National Academy of Sciences* **106,** 5318–5323 (2009). doi:10.1073/pnas.0811713106.

147. Anastassopoulou, C. G., Ketas, T. J., Sanders, R. W., *et al.* Effects of sequence changes in the HIV-1 gp41 fusion peptide on CCR5 inhibitor resistance. *Virology* **428,** 86–97 (2012). doi:10.1016/j.virol.2012.03.008.

148. Mielke, D., Bandawe, G., Zheng, J., *et al.* ADCC-mediating non-neutralizing antibodies can exert immune pressure in early HIV-1 infection. *PLOS Pathogens* **17** (ed Silvestri, G.) e1010046 (2021). doi:10.1371/journal.ppat.1010046.

149. Caskey, M., Schoofs, T., Gruell, H., *et al.* Antibody 10-1074 suppresses viremia in HIV-1-infected individuals. *Nature Medicine* **23,** 185–191 (2017). doi:10.1038/nm.4268.

150. MacLeod, D. T., Choi, N. M., Briney, B., *et al.* Early Antibody Lineage Diversification and Independent Limb Maturation Lead to Broad HIV-1 Neutralization Targeting the Env High-Mannose Patch. *Immunity* **44,** 1215–1226 (2016). doi:10.1016/j.immuni.2016.04.016.

151. Cale, E. M., Gorman, J., Radakovich, N. A., *et al.* Virus-like Particles Identify an HIV V1V2 Apex-Binding Neutralizing Antibody that Lacks a Protruding Loop. *Immunity* **46,** 777–791.e10 (2017). doi:10.1016/j.immuni.2017.04.011.

152. Schoofs, T., Barnes, C. O., Suh-Toma, N., *et al.* Broad and Potent Neutralizing Antibodies Recognize the Silent Face of the HIV Envelope. *Immunity* **50,** 1513–1529.e9 (2019). doi:10.1016/j.immuni.2019.04.014.

153. Wibmer, C. K., Bhiman, J. N., Gray, E. S., *et al.* Viral Escape from HIV-1 Neutralizing Antibodies Drives Increased Plasma Neutralization Breadth through Sequential Recognition of Multiple Epitopes and Immunotypes. *PLoS Pathogens* **9** (ed Trkola, A.) e1003738 (2013). doi:10.1371/journal.ppat.1003738.

154. Balla-Jhagjhoorsingh, S. S., Corti, D., Heyndrickx, L., *et al.* The N276 Glycosylation Site Is Required for HIV-1 Neutralization by the CD4 Binding Site Specific HJ16 Monoclonal Antibody. *PLoS ONE* **8** (ed Gray, C. M.) e68863 (2013). doi:10.1371/journal.pone.0068863.

155. Xie, H., Nie, J., Chen, Q., *et al.* Comparison of the genotypic and phenotypic properties of HIV-1 standard subtype B and subtype B/B' env molecular clones derived from infections in China. *Emerging Microbes & Infections* **7,** 1–13 (2018). doi:10.1038/s41426-018-0087-0.

156. Van Schooten, J., Schorcht, A., Farokhi, E., *et al.* Complementary antibody lineages achieve neutralization breadth in an HIV-1 infected elite neutralizer. *PLOS Pathogens* **18** (ed Doores, K. J.) e1010945 (2022). doi:10.1371/journal.ppat.1010945.

157. Hatada, M., Yoshimura, K., Harada, S., *et al.* Human immunodeficiency virus type 1 evasion of a neutralizing anti-V3 antibody involves acquisition of a potential glycosylation site in V2. *Journal of General Virology* **91,** 1335–1345 (2010). doi:10.1099/vir.0.017426-0.

158. Haddox, H. K., Dingens, A. S., Hilton, S. K., *et al.* Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife* **7** (2018). doi:10.7554/elife.34420.

159. Dingens, A. S., Arenz, D., Weight, H., *et al.* An Antigenic Atlas of HIV-1 Escape from Broadly Neutralizing Antibodies Distinguishes Functional and Structural Epitopes. *Immunity* **50,** 520–532.e3 (2019). doi:10.1016/j.immuni.2018.12.017.

160. Hodcroft, E. B. *CoVariants: SARS-CoV-2 Mutations and Variants of Interest.* https://covariants.org/ visited on 2024-09-10.

161. Miyakawa, K., Ota, M., Sano, K., *et al.* Emergence of Antigenic Variants in Bovine H5N1 Influenza Viruses. *Journal of Medical Virology* **97** (2025). doi:10.1002/jmv.70394.

162. Abramson, J., Adler, J., Dunger, J., *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630,** 493–500 (2024). doi:10.1038/s41586-024-07487-w.

163. DeLano, W. L. *The PyMOL Molecular Graphics System* version 3.1.0. 2002. http://www.pymol.org visited on 2026-01-18.

## Supplementary Material

**a**

| number of sequences | EVOtRec | UShER | VeryFastTree | CMAPLE | IQ-TREE |
|---|---|---|---|---|---|
| $1 \times 10^3$ | 1.4 ± 0.0 | 1.1 ± 0.1 | 305.7 ± 1.9 | 51.5 ± 1.3 | 11.4 ± 0.7 |
| $2 \times 10^3$ | 2.5 ± 0.1 | 4.1 ± 0.2 | 608.9 ± 12.8 | 182.0 ± 3.9 | 57.9 ± 2.5 |
| $5 \times 10^3$ | 6.0 ± 0.2 | 20.7 ± 2.6 | 1,539.5 ± 96.4 | 1,065.0 ± 27.0 | 480.5 ± 23.0 |
| $1 \times 10^4$ | 11.8 ± 0.4 | 66.4 ± 7.3 | 2,983.1 ± 134.3 | 3,594.2 ± 62.9 | 2,743.3 ± 119.9 |
| $2 \times 10^4$ | 24.2 ± 0.8 | 246.1 ± 26.7 | 6,166.0 ± 208.0 | 10,683.3 ± 167.0 | 13,731.7 ± 96.5 |
| $5 \times 10^4$ | 75.0 ± 4.8 | 2,066.6 ± 44.7 | 16,311.3 ± 480.5 | 50,126.0 ± 446.0 | - |
| $1 \times 10^5$ | 220.1 ± 8.9 | 8,038.7 ± 296.5 | 33,748.0 ± 1,342.5 | - | - |
| $2 \times 10^5$ | 715.4 ± 12.0 | 42,240.0 ± 812.0 | 85,018.7 ± 3,064.5 | - | - |

**b**

| number of sequences | EVOtRec | UShER | VeryFastTree | CMAPLE | IQ-TREE |
|---|---|---|---|---|---|
| $1 \times 10^3$ | 509.2 ± 0.3 | 67.3 ± 2.6 | 1,651.1 ± 6.5 | 1,380.3 ± 0.0 | 125.8 ± 0.0 |
| $2 \times 10^3$ | 946.1 ± 0.5 | 104.7 ± 0.0 | 2,973.7 ± 2.8 | 2,750.8 ± 0.0 | 485.4 ± 0.0 |
| $5 \times 10^3$ | 2,263.5 ± 0.2 | 255.7 ± 0.0 | 5,672.0 ± 4.0 | 7,324.5 ± 0.1 | 2,601.5 ± 0.7 |
| $1 \times 10^4$ | 4,430.1 ± 0.3 | 507.4 ± 0.0 | 9,082.2 ± 23.8 | 14,536.6 ± 13.2 | 8,781.0 ± 0.1 |
| $2 \times 10^4$ | 8,882.2 ± 0.6 | 1,010.8 ± 0.0 | 14,994.3 ± 553.7 | 29,466.9 ± 28.2 | 27,433.8 ± 0.9 |
| $5 \times 10^4$ | 22,779.1 ± 1.3 | 2,521.1 ± 0.0 | 31,274.9 ± 192.7 | 80,543.3 ± 72.0 | - |
| $1 \times 10^5$ | 47,889.8 ± 0.5 | 5,038.2 ± 0.0 | 61,485.9 ± 523.5 | - | - |
| $2 \times 10^5$ | 105,512.6 ± 2.7 | 10,072.5 ± 0.0 | 125,129.3 ± 692.4 | - | - |

**Supplementary Table 1. Benchmarking of EVOtRec vs. phylogenetic inference tools.** We analyzed the runtime and memory usage of EVOtRec in comparison with the following tools for the reconstruction of phylogenetic trees: UShER [21], VeryFastTree [49], CMAPLE [50] and IQ-TREE [48]. We applied these tools to eight SARS-CoV-2 genome alignments with 1k up to 200k distinct sequences, obtained by subsampling the SARS-CoV-2 whole genome alignment (Methods). Tables show the mean ± 95% CI (over 3 independent runs) of runtime in seconds (a) and peak memory usage (RAM) in megabytes (b). A dash (-) indicates that the runtime exceeded 24 hours, in which case no measurement was taken.

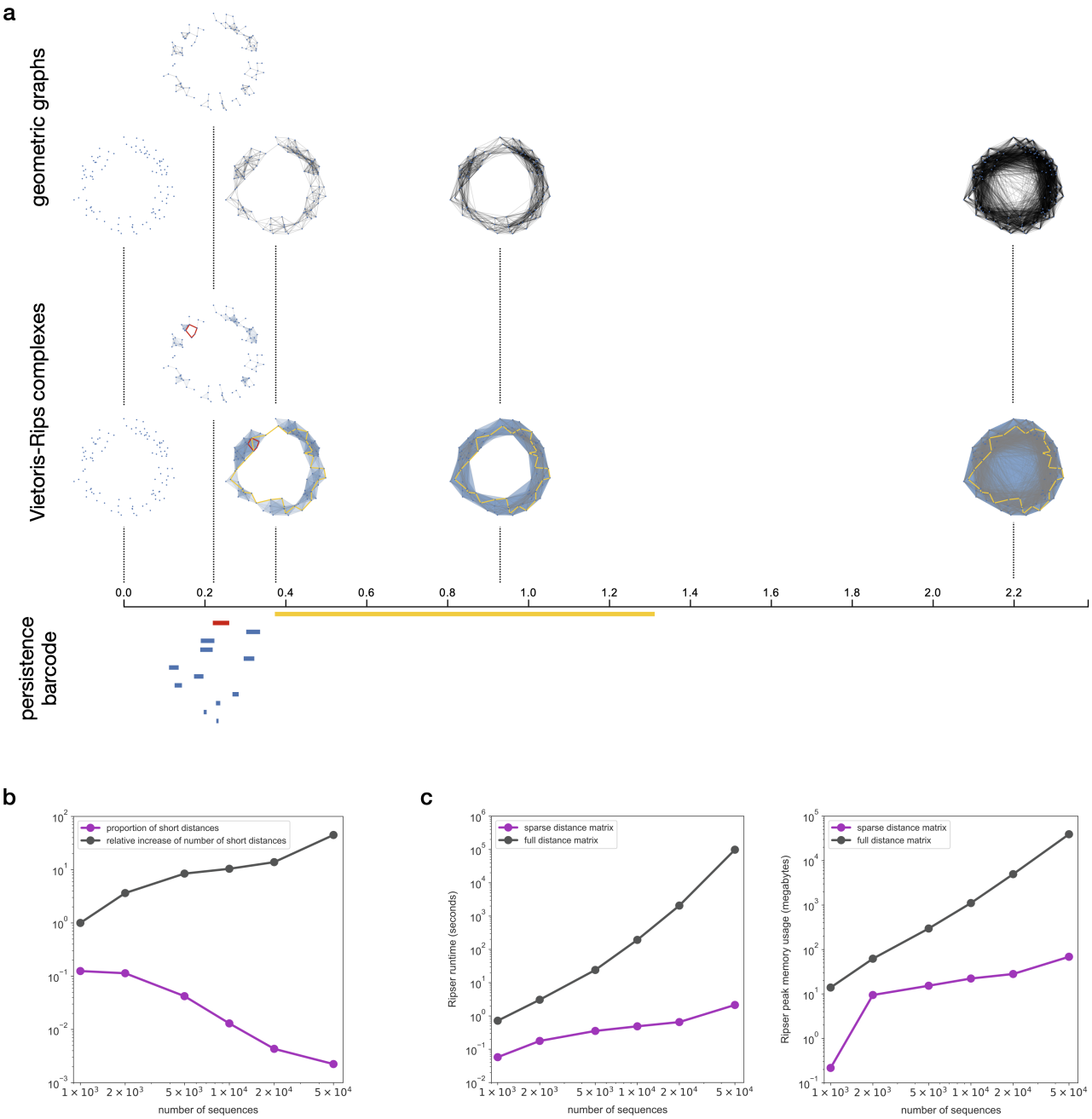
**Supplementary Table 2. Topological recurrence analysis of SARS-CoV-2.** Time series data (raw tRI, tRI, tRI significance level, prevalence, GISAID accession numbers) of the topological recurrence analysis of the following SARS-CoV-2 alignments: (i) whole genome alignment (starting date 30 December 2019); (ii) spike gene alignment (starting date 30 December 2019); (iii) Omicron spike gene alignment (starting date 01 November 2021); (iv) United Kingdom spike gene alignment (starting date 30 December 2019). CSV data files are available via https://doi.org/10.5281/zenodo.18332026.

**Supplementary Table 3. Topological recurrence analysis of H5N1.** Time series data (raw tRI, tRI, tRI significance level, prevalence, GISAID accession numbers) of the topological recurrence analysis of the following H5N1 alignments: (i) HA gene alignment (starting date 30 December 2021); (ii) PB2 gene alignment (starting date 30 December 2021). CSV data files are available via https://doi.org/10.5281/zenodo.18332026.

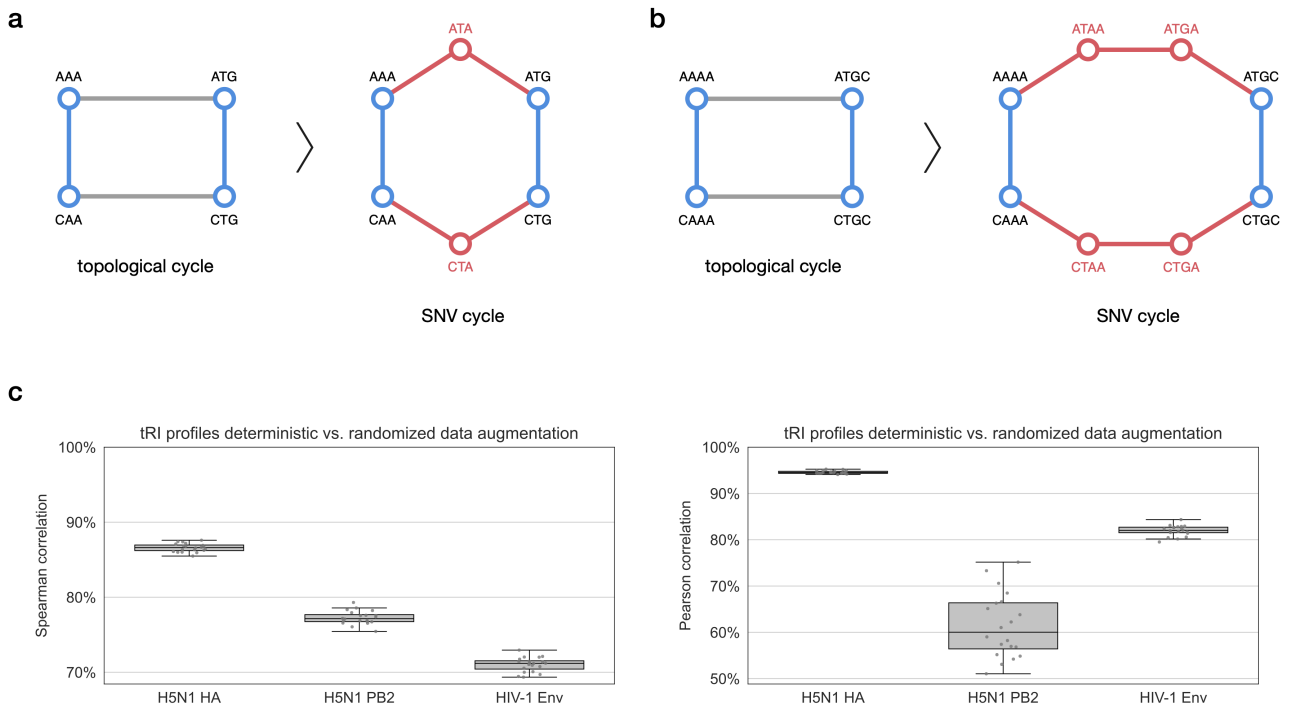
**Supplementary Table 4. Topological recurrence analysis of HIV-1.** Data (raw tRI, tRI, tRI significance level, prevalence, ENA accession numbers) of the topological recurrence analysis of the HIV-1 Env gene alignment. CSV data files are available via https://doi.org/10.5281/zenodo.18332026.

| mutation | tRI | prevalence | indication of adaptation | references |
|---|---|---|---|---|
| F317I | 244.0 | 0.035 | phylogeny-based analysis using codon models | Yoshida et al., 2011 |
| N302K | 244.0 | 0.037 | associated with antibody resistance | Witvrouw et al., 2005 |
| A281D | 214.5 | 0.013 | associated with antibody resistance | Watkins et al., 1993; Yoshimura et al., 2010; Galvez et al., 2024 |
| W17R | 188.1 | 0.0057 | potentially associated with antibody resistance | Upadhyay et al., 2018 |
| F353I | 183.0 | 0.013 | deep mutational scanning | Haddox et al., 2016 |
| W15R | 183.0 | 0.00047 | potentially associated with antibody resistance | Upadhyay et al., 2018 |
| K4M | 171.1 | 0.26 | potentially associated with antibody resistance | Upadhyay et al., 2018 |
| L765I | 148.5 | 0.018 | - | - |
| N234K | 122.0 | 0.012 | associated with antibody resistance | Férir et al., 2014 |
| V812G | 122.0 | 0.0045 | - | - |
| N356H | 104.3 | 0.049 | deep mutational scanning | Haddox et al., 2018 |
| C196G | 91.5 | 0.00027 | - | - |
| T49N | 82.5 | 0.59 | associated with antibody resistance | Mahalanabis et al., 2009; Crooks et al., 2021 |
| N230D | 67.6 | 0.47 | potentially associated with immune escape | Li et al., 2009 |
| L453I | 66.0 | 0.11 | - | - |
| I746S | 61.0 | 0.25 | - | - |
| F519I | 61.0 | 0.081 | associated with drug resistance | Anastassopoulou et al., 2009; Anastassopoulou et al., 2012 |
| N463K | 61.0 | 0.061 | - | - |
| D62E | 61.0 | 0.41 | deep mutational scanning | Haddox et al., 2016 |
| D57E | 61.0 | 0.0082 | - | - |
| V539E | 61.0 | 0.00050 | associated with antibody resistance | Dingens et al., 2019 |
| F522V | 61.0 | 0.0024 | - | - |
| Y795N | 61.0 | 0.0014 | - | - |
| K97E | 60.1 | 0.056 | associated with antibody resistance | Mielke et al., 2021 |
| V513A | 45.8 | 0.10 | deep mutational scanning | Haddox et al., 2016 |
| N230S | 41.3 | 0.011 | potentially associated with immune escape | Li et al., 2009 |
| V3A | 40.7 | 0.088 | - | - |
| N339Y | 32.1 | 0.016 | associated with antibody resistance | Dingens et al., 2019 |
| S334R | 30.5 | 0.011 | associated with antibody resistance | Caskey et al., 2017; Dingens et al., 2019 |
| M104R | 30.5 | 0.00023 | - | - |
| N241K | 30.5 | 0.0058 | deep mutational scanning | Haddox et al., 2016 |
| S158A | 30.5 | 0.00047 | associated with antibody resistance | MacLeod et al., 2016; Cale et al., 2017 |
| C28G | 30.5 | 0.0080 | - | - |
| S447A | 30.5 | 0.00040 | deep mutational scanning | Haddox et al., 2016 |
| N448K | 30.5 | 0.018 | associated with antibody escape | Schoofs et al., 2019 |
| N160D | 30.1 | 0.013 | associated with antibody resistance | Wibmer et al., 2013; Dingens et al., 2019 |
| T319A | 26.3 | 0.73 | potentially associated with antibody resistance | Hatada et al., 2010 |
| K282E | 26.3 | 0.0059 | deep mutational scanning | Haddox et al., 2016 |
| N276D | 26.3 | 0.029 | associated with antibody resistance | Balla-Jhagjhoorsingh et al., 2013 |
| N295T | 26.1 | 0.053 | associated with antibody resistance | Xie et al., 2018; Schooten et al., 2022 |

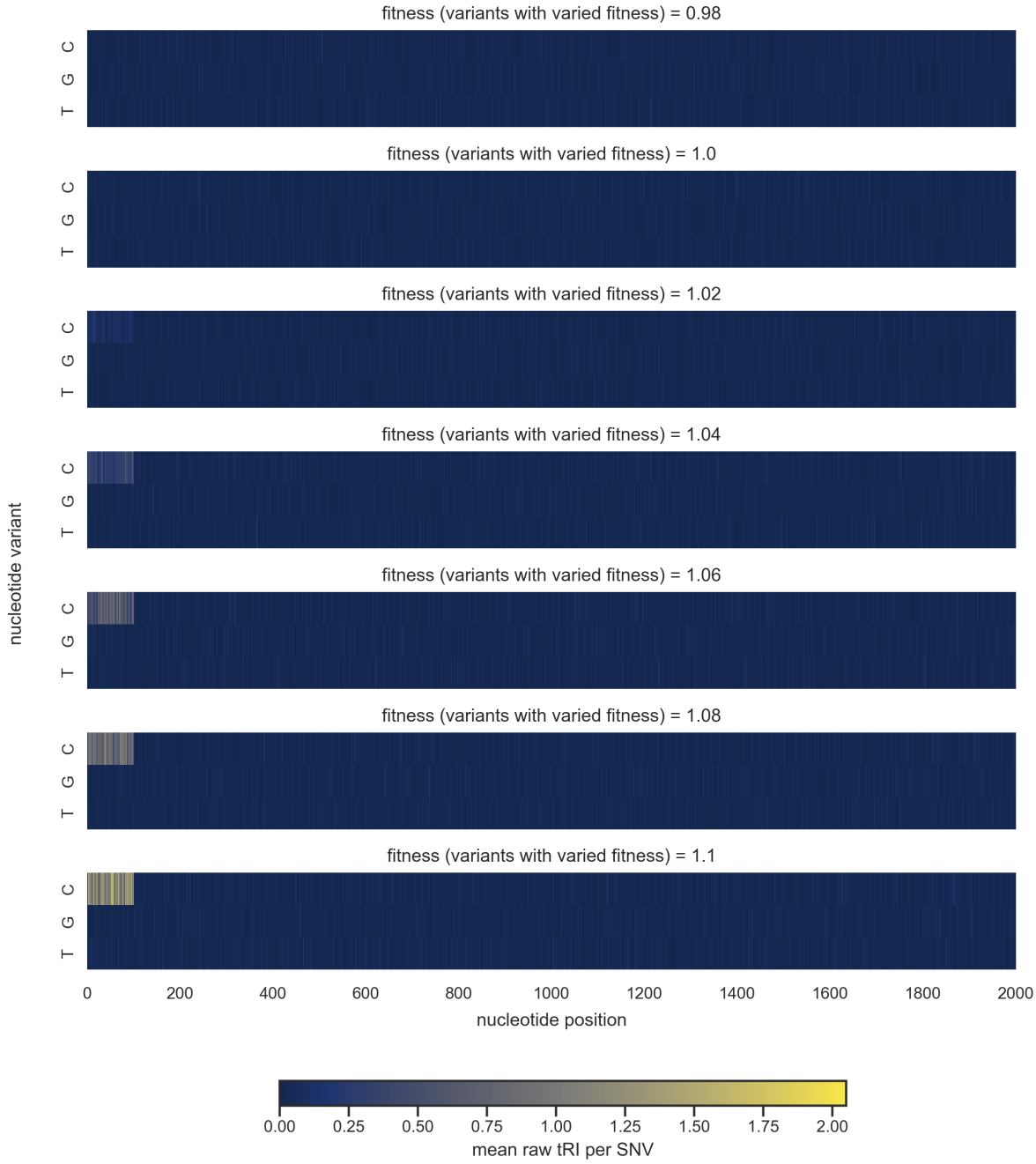
**Supplementary Table 5. Assessment of adaptation for HIV-1 amino acid changes with strong topological signal.** The table displays the top 40 amino acid changes on the HIV-1 envelope gene (Env) with strongest tRI signals in the EVOtRec analysis of the ENA Env gene alignment [126], together with their prevalence in the sequence alignment (Supplementary Table 4 and Methods). Among these, at least 70% (28 out of 40) have been associated with (potential) adaptation in the literature [106–112, 142–159].



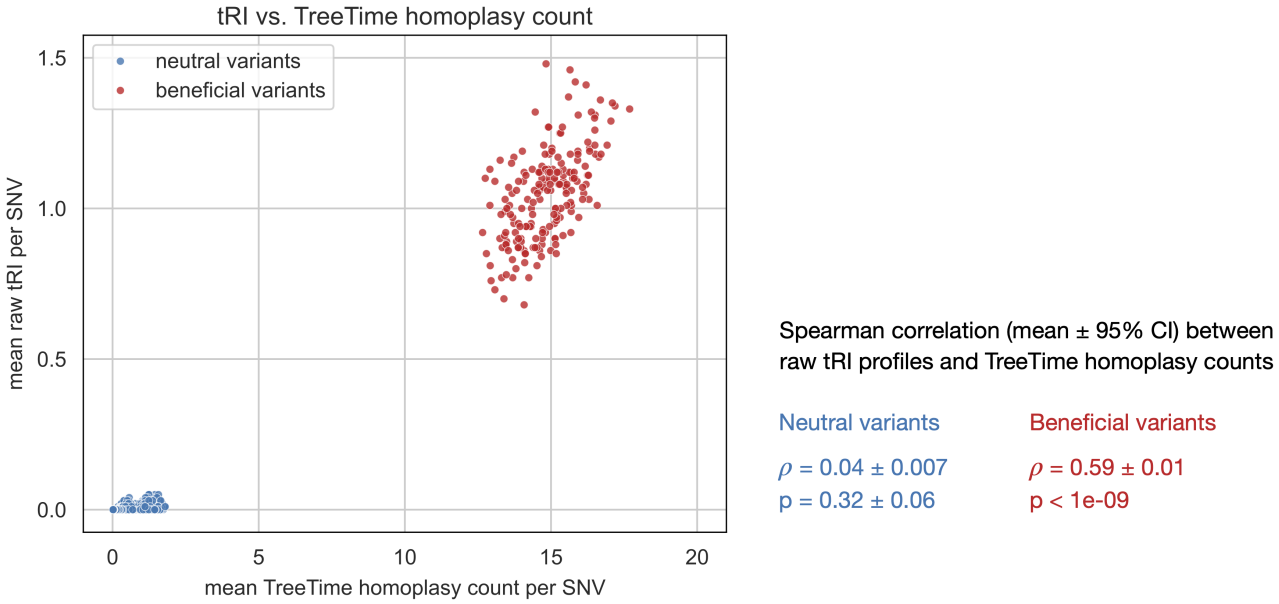

**Supplementary Figure 1. Computing persistent homology of genomic datasets.** EVOtRec uses Ripser [40] to compute the persistent homology of genomic datasets from genetic distance matrices (Methods). **(a)** Schematic example of how Ripser can be used to study the topology of genomic datasets. Each point represents a sample, and we display the geometric graphs (upper row), the resulting Vietoris–Rips complexes at different scales (middle row) and the persistence barcode in dimension one (lower row). If one only chooses one scale, one might either see nothing, or detect the small red cycle but miss the large yellow one, or vice versa. The basic idea of persistent homology to handle this issue is to characterize each cycle with its scale of appearance and disappearance: the red cycle induces a red bar in the barcode, and similarly for the yellow cycle. **(b, c)** EVOtRec's sparse distance format reduces the complexity of persistent homology computations. The format is based exclusively on short genetic distances, whose proportion among all observed distances rapidly decreases to less than 1% for alignments with more than 10k sequences (b). The usage of EVOtRec sparse distance matrices significantly reduces runtime and peak memory usage (RAM) for the computation of persistence barcodes with Ripser by several orders of magnitude as compared with standard full distance matrices (c). Runtimes in (c) are means over 5 independent runs.




**Supplementary Figure 2. Data augmentation via ancestral sequence reconstruction.** We applied data augmentation via ancestral sequence reconstruction in the H5N1 and HIV-1 genome alignments, as these alignments have lower sampling rate than the SARS-CoV-2 alignments (Methods). **(a-b)** For short edges (Hamming length = 2 or 3), synthetic sequences were added to the alignment to replace these edges with chains composed exclusively of unit-length edges. The procedure turns topological cycles containing edges of length 2 or 3 into SNV cycles, thereby amplifying the corresponding tRI signals. In the displayed schematic example with trinucleotide sequences (a), by adding two synthetic sequences (red) each of the horizontal edges of Hamming length 2 in the topological cycle is replaced by two unit-length edges (red), turning the topological cycle into an SNV cycle. In the displayed schematic example with tetranucleotide sequences (b), by adding four synthetic sequences (red) each of the horizontal edges of Hamming length 3 in the topological cycle is replaced by a chain of three unit-length edges (red), turning the topological cycle into an SNV cycle. **(c)** Box plot with overlaid strip plot showing the Spearman and Pearson correlation between tRI profiles generated with deterministic (as described in a, b) versus randomized data augmentation (Methods), for the H5N1 and HIV-1 alignments. Correlations were computed between a single alignment that was generated using deterministic data augmentation, and 20 alignments that were independently generated using randomized data augmentation. We found strong correlations across all three alignments, justifying the use of data augmentation in the preprocessing of the H5N1 and HIV-1 alignments.




**Supplementary Figure 3. Effects of variations in fitness on tRI signals.** We analyzed the dependence of the raw tRI on the fitness parameter in simulated synthetic sequence alignments generated with SANTA-SIM v1.0 [42]. Simulations were initialized with polyadenine sequences AA. . . A of length 2000nt (Methods). The fitness parameter assigned to C alleles at the first 100 nucleotide positions (variants with varied fitness) was uniformly set to one of the values 0.98, 1.00, 1.02, 1.04, 1.06, 1.08 and 1.10, making the mutations A1C, A2C, . . . , A100C deleterious, neutral or beneficial, respectively. The fitness parameter for all other mutations was kept at 1.00 (neutral variants). For each of these scenarios we ran 20 independent simulations. We performed a topological recurrence analysis with EVOtRec for each simulated alignment and subsequently computed the mean raw tRI per SNV (mean calculated by averaging over all 20 simulations for each SNV). The displayed heatmaps show the mean raw tRI per SNV, with the horizontal axis representing the 2000 nucleotide positions and the vertical axis representing the three mutation types A>C, A>G and A>T.

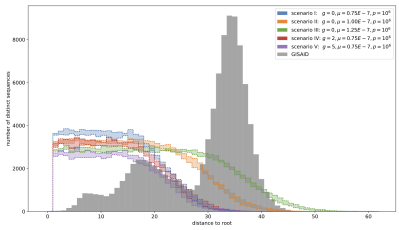



**Supplementary Figure 4. Comparison of tRI and phylogenetic homoplasy counts in simulated evolutionary scenarios.** We simulated 100 nucleotide sequence alignments with SANTA-SIM v1.0 [42], where 7.1% of single nucleotide mutations were assigned a fitness parameter of 1.3 (beneficial variants), while all other mutations were assigned a fitness parameter of 1.0 (neutral variants) (Methods). For each alignment, we (i) removed duplicate sequences and computed raw tRI signals with EVOtRec, and (ii) reconstructed phylogenetic trees with IQ-TREE v3.0.1 [48] and subsequently computed phylogenetic homoplasy counts with TreeTime v0.11.4 [51]. The scatter plot shows mean raw tRI and mean TreeTime homoplasy counts per SNV (means per SNV calculated by averaging over all 100 simulations). We found that (i) there were two clusters corresponding to whether the SNV was neutral or beneficial, (ii) mean raw tRI signals were smaller than mean TreeTime homoplasy counts, and (iii) several neutral SNVs exhibited no mean raw tRI signal but had positive TreeTime homoplasy count. Moreover, raw tRI profiles and TreeTime homoplasy counts showed a moderate-to-strong mean Spearman correlation of $\rho = 0.59 \pm 0.01$ across beneficial variants (mean $\pm$ 95% CI, calculated by averaging over all 100 simulations), but they were not correlated across neutral variants ($\rho = 0.04 \pm 0.007$).

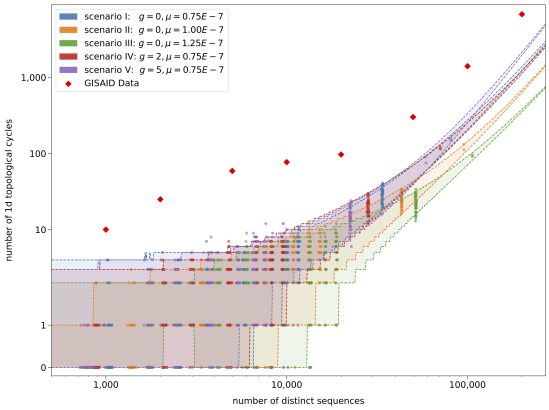
**a** Parameters for simulations of neutral evolutionary scenarios with SANTA-SIM

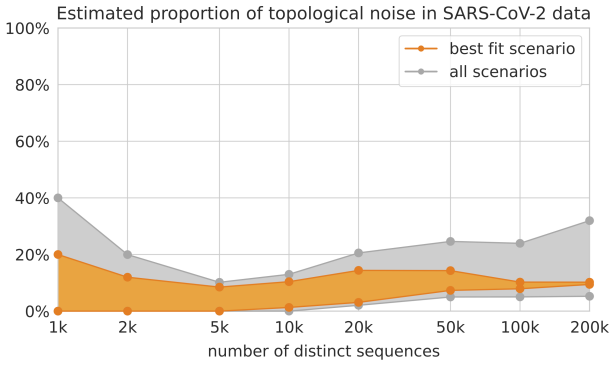
| scenario | mutation rate $\mu$ [$10^{-7}$ subst./(site×gen.)] | growth rate $g$ [1/gen.] | initial population $p$ / carrying capacity $c$ |
|---|---|---|---|
| I | 0.75 | 0 | $p = c =$ 1e2, 5e2, 1e3, 2.5e3, 5e3, 7.5e3, 1e4, 1e5, 1e6 |
| II | 1 | 0 | |
| III | 1.25 | 0 | |
| IV | 0.75 | 2 | $p = 100$ / $c =$ 1e2, 5e2, 1e3, 2.5e3, 5e3, 7.5e3, 1e4, 1e5, 1e6 |
| V | 0.75 | 5 | |

**b** Genetic distance from reference sequence in simulated neutral evolutionary scenarios vs. GISAID SARS-CoV-2 whole genome alignment

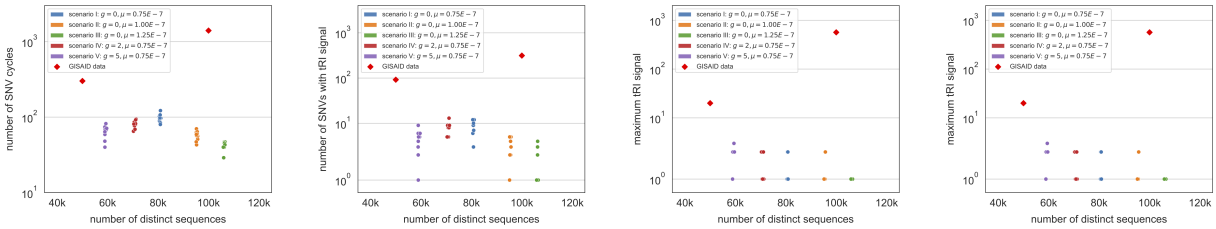



**c** Count of topological cycles in simulated neutral evolutionary scenarios vs. GISAID SARS-CoV-2 whole genome alignments

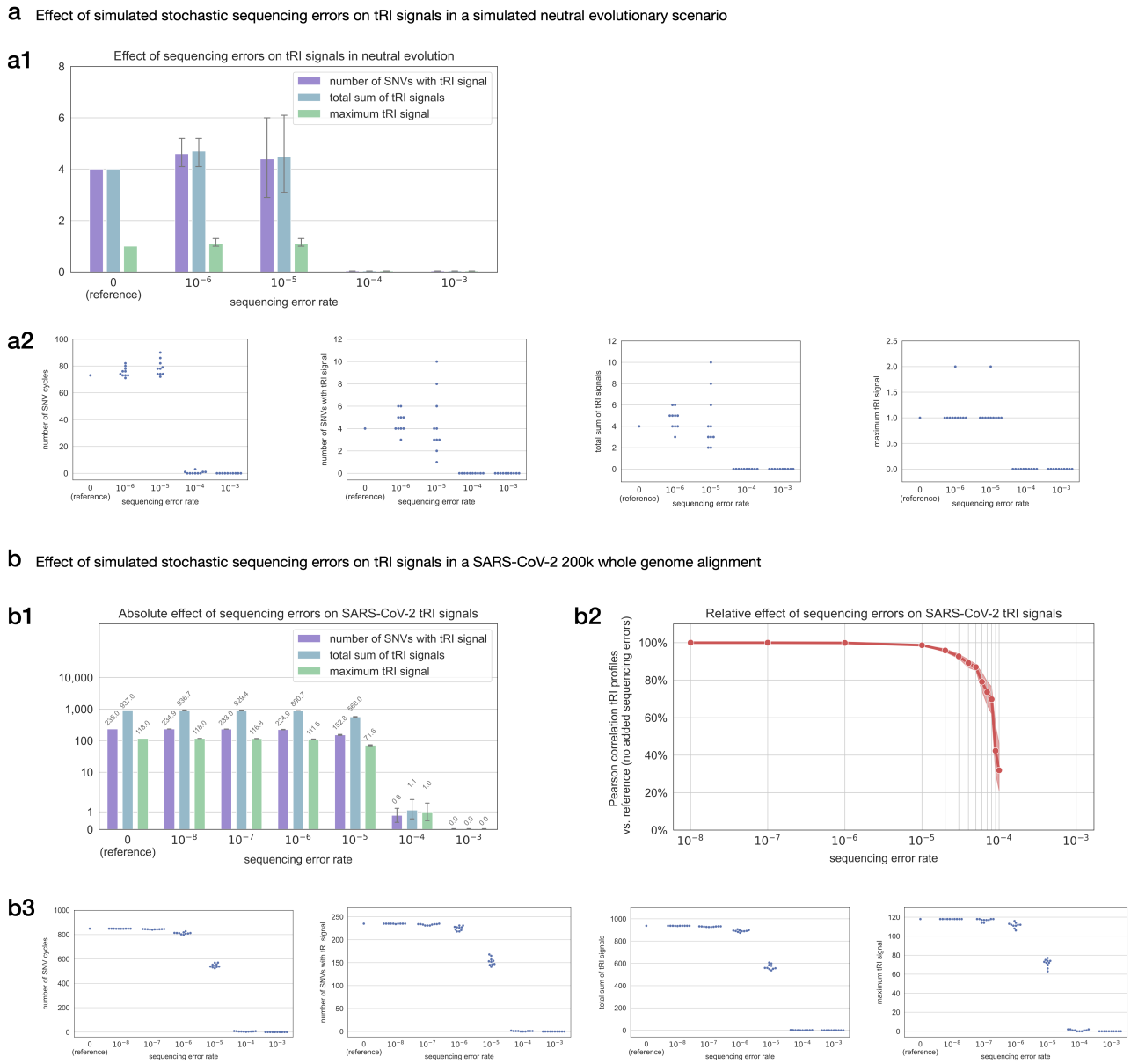


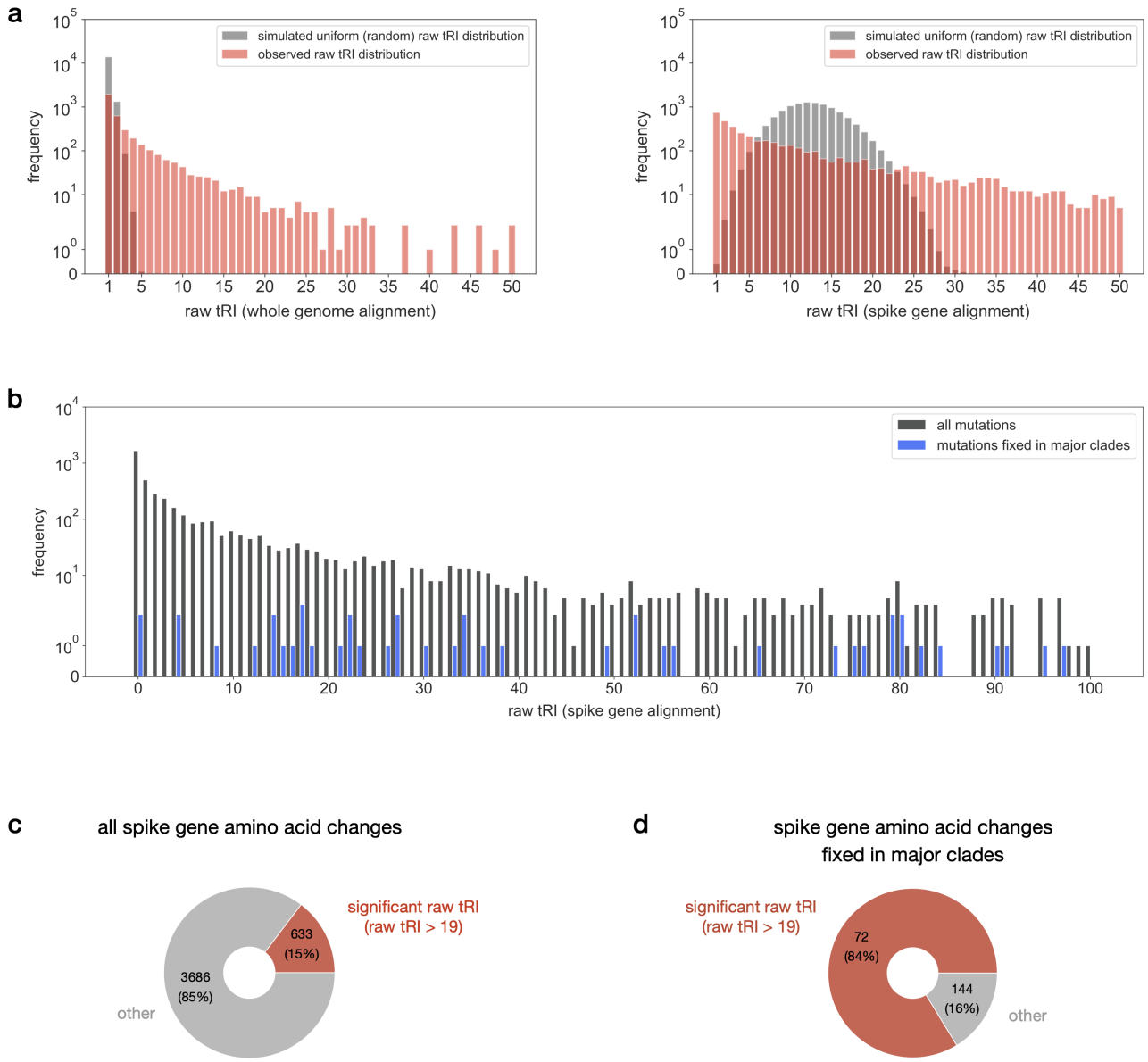

**d** Topological signals in simulated neutral evolutionary scenarios vs. GISAID SARS-CoV-2 whole genome alignments



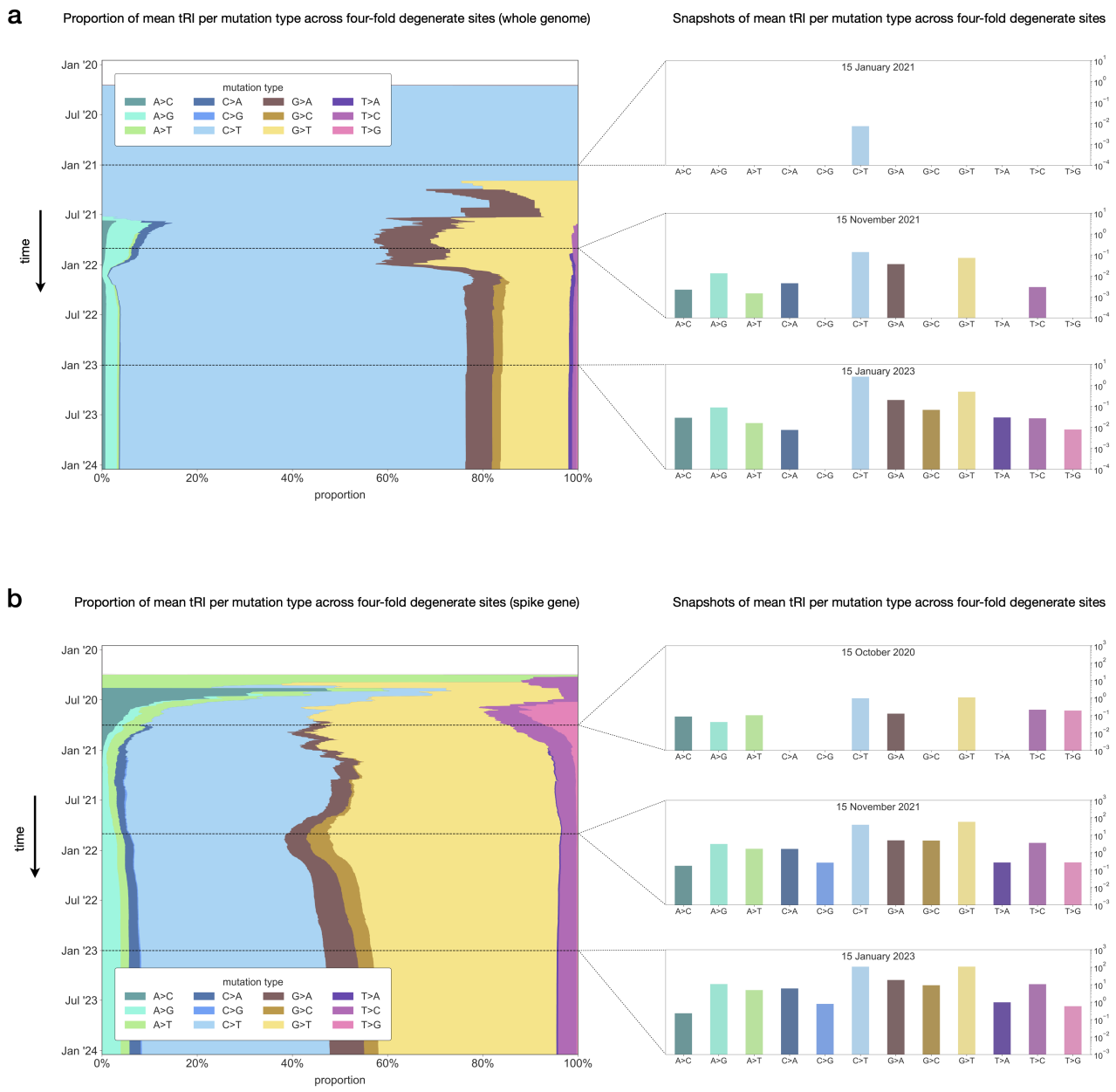

**Supplementary Figure 5. Robustness of tRI signals to topological noise.** We estimated topological background noise in SARS-CoV-2 whole genome alignments from topological signals in neutral evolutionary scenarios, based on synthetic alignments generated with SANTA-SIM v1.0 [42] for five distinct scenarios with varying growth rate $g$ and mutation rate $\mu$ (Methods). **(a)** SANTA-SIM parameter choices for the five simulated scenarios. Scenarios I-III vary over a range of mutation rates, while scenarios IV and V probe the effect of logistic population growth. **(b)** Comparison of genetic distances to the root in simulated alignments vs. genetic distances to the reference sequence hCoV-19/Wuhan/WIV04/2019 in the GISAID [58] SARS-CoV-2 200k whole genomes alignment. The alignment of scenario II best fits the SARS-CoV-2 data. **(c)** We computed topological cycles with Ripser [40] in GISAID SARS-CoV-2 whole genome alignments of size ranging from 1k up to 200k sequences, and in simulated alignments of varying size for the five distinct scenarios. The 95% prediction intervals for the number of cycles in each scenario are based on the extrapolation of a Panjer distribution for an increasing number of distinct sequences in simulated phylogenies. For each scenario, the validation dataset shown in the plot is well-described by the corresponding prediction intervals. We used the results in (c) to estimate the proportion of topological noise (95% confidence bands) in SARS-CoV-2 whole genome alignments. The best fit scenario is scenario II. **(d)** Summary statistics of topological signals in simulated alignments for scenarios I-IV and in GISAID SARS-CoV-2 whole genome alignments with 50k and 100k sequences. Scatter plots showing the number of SNV cycles, the number of SNVs with (positive) tRI signal, the total sum of tRI signals and the maximum tRI signal in simulated vs. SARS-CoV-2 alignments (Methods). The number of distinct sequences may vary across alignments because we removed duplicate sequences from the original simulated alignments, each comprising 200k sequences.



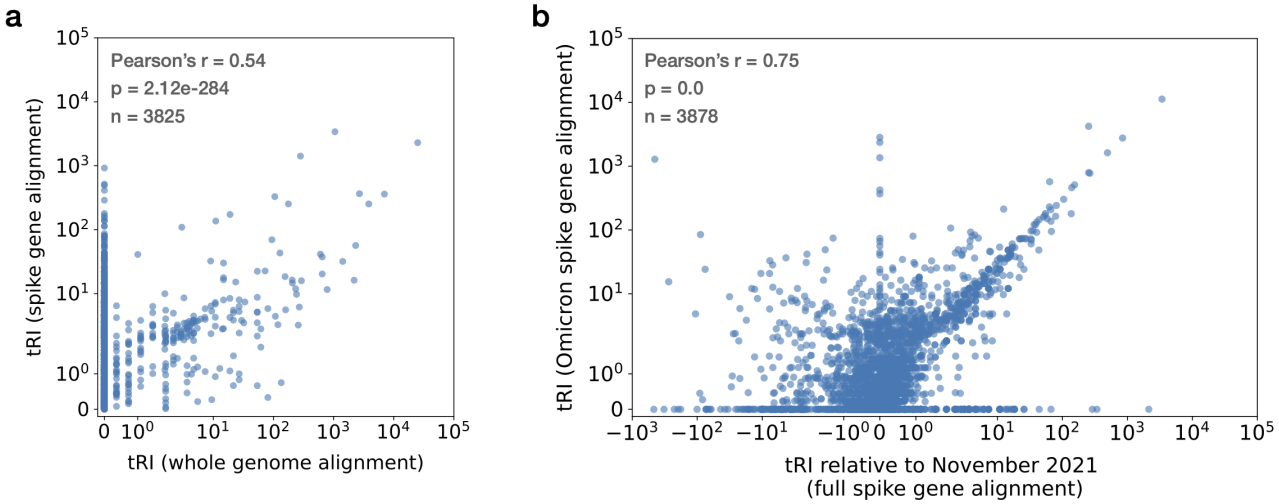

**Supplementary Figure 6. Robustness of tRI signals to stochastic sequencing errors.** We estimated the effect of stochastic sequencing errors on tRI by adding simulated stochastic sequencing errors to two reference alignments: a simulated neutral evolutionary scenario (a) generated with SANTA-SIM v1.0 [42], and a real SARS-CoV-2 200k whole genomes alignment (b) (Methods). We analyzed a wide range of error rates from $10^{-8}$ to $10^{-3}$ on a logarithmic scale. Rates of stochastic sequencing errors in consensus genomes typically lie within this range [43, 44]. We generated 10 alignments with added simulated sequencing errors for each scenario. **(a1)** Bar plot showing the absolute effect of sequencing errors on tRI (mean ± 95% CI) in neutral evolution. **(a2)** Swarm plots showing the number of SNV cycles and topological signals (as in a1) in the reference alignment and in alignments with added sequencing errors. **(b1)** Bar plot showing the absolute effect of added sequencing errors on tRI (mean ± 95% CI; bar value labels indicate means) for the original alignment without added sequencing errors (reference) and alignments with added sequencing errors. The presence of sequencing errors does not increase tRI signals, but decreases tRI signals for larger error rates. **(b2)** Line plot showing the Pearson correlation (mean ± 95% confidence band) between tRI profiles across the whole genome for alignments with added sequencing errors vs. those for the reference alignment. **(b3)** Swarm plots showing the number of SNV cycles and topological signals (as in b1 and b2) in the reference alignment and in alignments with added sequencing errors.



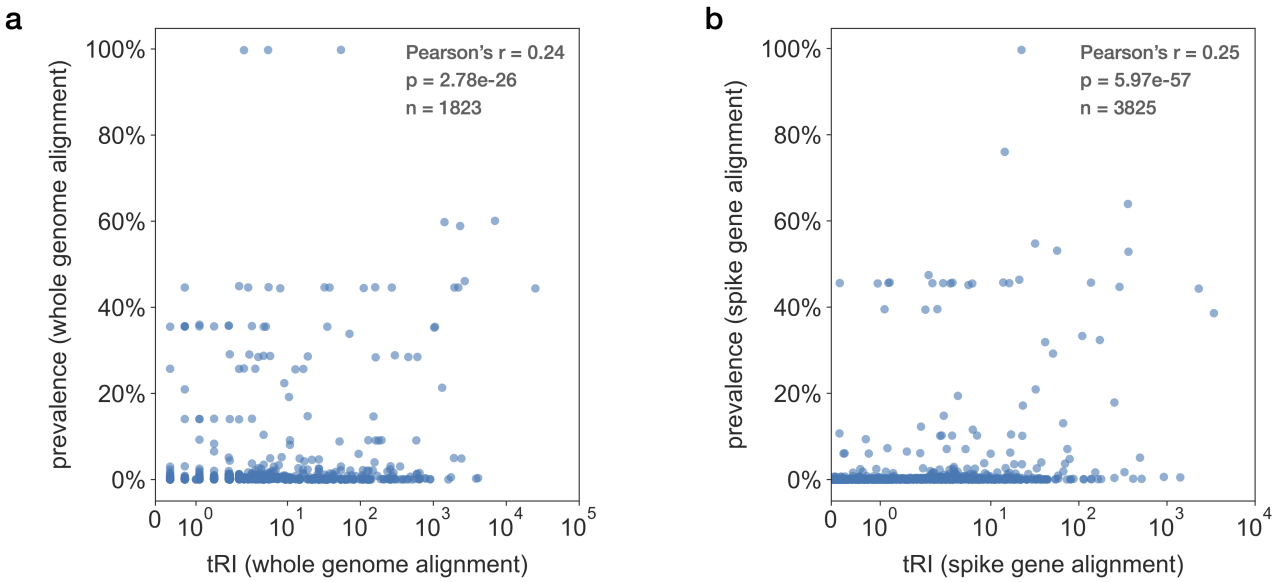

**Supplementary Figure 7. Statistical significance of SARS-CoV-2 tRI signals.** Raw tRI signals of the SARS-CoV-2 whole genome and spike gene alignments. **(a)** The frequency distributions of raw tRI signals in the whole genome alignment and spike gene alignment differ significantly from the frequency distributions obtained from a Monte Carlo permutation test under the null hypothesis that raw tRI signals are distributed uniformly across the genome (Kolmogorov–Smirnov test, $p < 0.05$; Methods). In the diagrams only raw tRI signals $\leq 50$ are shown. **(b)** Frequency distributions of raw tRI signals for all 4319 amino acid changes in the spike gene alignment, and for the subset of those 86 amino acid changes that became fixed in one of the Nextstrain clades Alpha (20I), Beta (20H), Gamma (20J), Delta (21A, 21I, 21J) or Omicron (21K, 21L, 22A, 22B, 22C, 22D, 22E, 22F, 23A, 23B, 23C, 23D, 23E, 23F, 23G, 23H) [20, 160]. Only raw tRI signals $\leq 100$ are shown. **(c-d)** Proportion of spike amino acid changes with significant raw tRI signal (raw tRI $> 19$) in all amino acid changes (c) and in amino acid changes fixed in one of the Nextstrain clades Alpha, Beta, Gamma, Delta or Omicron (d).




**Supplementary Figure 8. Proportions of tRI signals per nucleotide mutation type across four-fold degenerate sites in SARS-CoV-2.** Stacked area plots (left) showing the time-dependent relative proportion of mean tRI signals across all four-fold degenerate sites for each nucleotide mutation type on the whole genome (a) and on the Spike gene (b). Snapshot bar plots (right) taken at specific time points in the early, middle and late phase of the pandemic showing the absolute mean tRI across all four-fold degenerate sites for each nucleotide mutation type. Proportions from mid-2022 onwards in (a) are congruent with proportions obtained from mutation counts in SARS-CoV-2 phylogenetic trees [6].

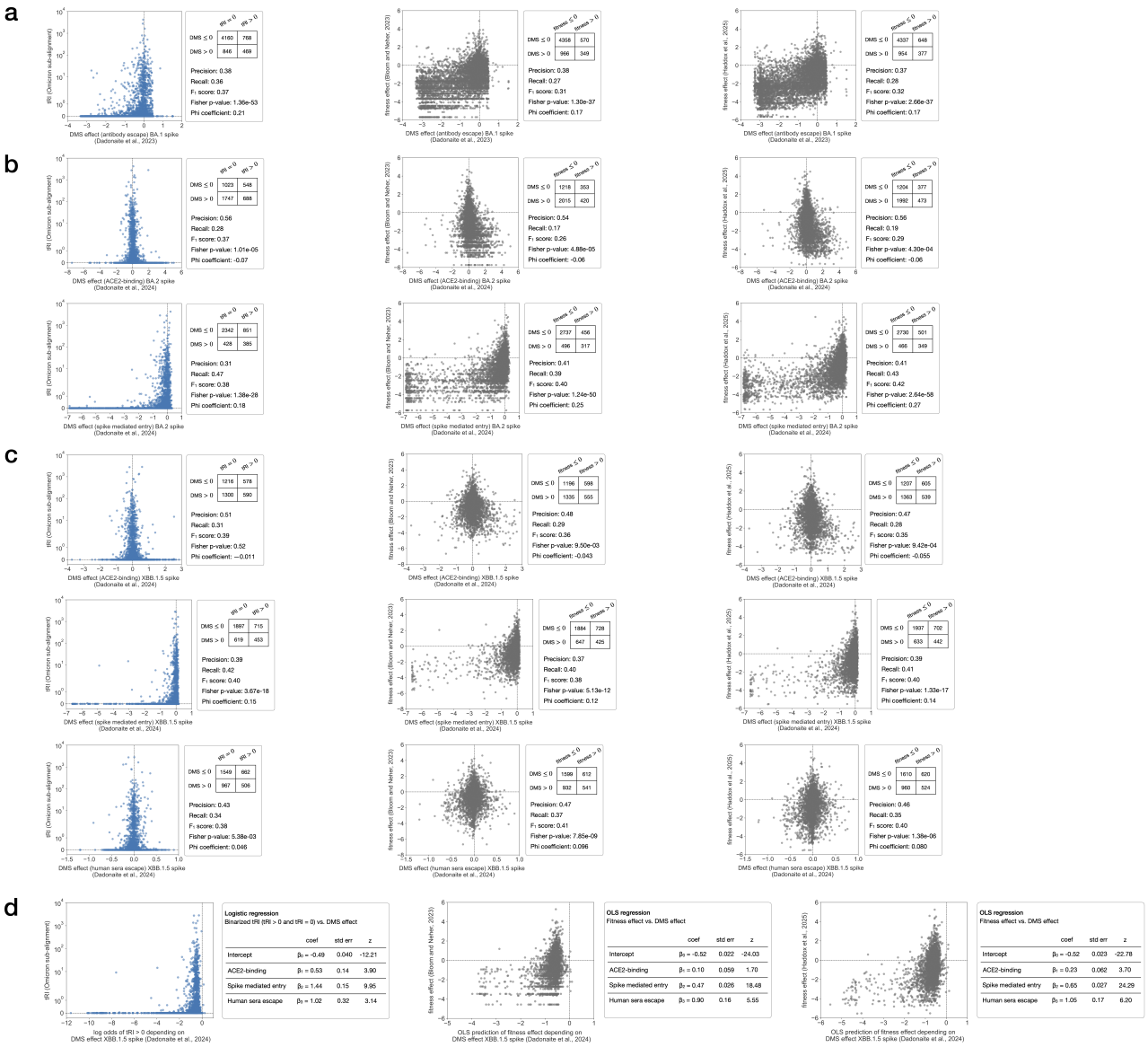



**Supplementary Figure 9. Restricting the topological recurrence analysis to sliding windows or sub-alignments.** **(a)** The topological recurrence analysis can be carried out for sliding windows on the genome, such as single genes or specific genomic regions (Methods). A comparison of tRI signals for SARS-CoV-2 spike gene SNVs obtained from the analysis of the whole genome alignment vs. the spike gene alignment illustrates the trade-off between tRI signal strength and genomic scope. Restricting the analysis to the spike gene alignment means that SNVs outside the spike gene do not contribute to genetic distance, which typically leads to the creation of more SNV cycles and hence stronger tRI signals. We found that tRI signals of the whole genome alignment and spike gene alignment are correlated with Pearsons's $r = 0.54$. **(b)** Since isolates in SNV cycles share the same genomic background, contributions to tRI signals can be attributed to unique genomic lineages (Figure 1). For example, we were able to extract tRI signals generated by SARS-CoV-2 Omicron sequences directly from tRI signals for the full spike gene alignment without the need to run the tRI analysis separately for a sub-alignment corresponding to the Omicron variant. This was achieved by counting only those tRI signals in the full spike gene alignment that arose after the onset of Omicron in November 2021. We found that tRI signals occurring after October 2021 in the full spike gene alignment (covering the period from January 2020 until February 2024) and tRI signals in the Omicron spike gene alignment (covering the period from November 2021 until February 2024) are strongly correlated with Pearson's $r = 0.75$.




**Supplementary Figure 10. Correlation between tRI and prevalence.** Scatter plots showing tRI vs. prevalence of SNVs on the last day of the whole genome alignment (a) and spike gene alignment (b) tRI/prevalence time series (29/30 January 2024). We found that tRI and prevalence were correlated with Pearson's $r = 0.24$ for the whole genome alignment and $r = 0.25$ for the spike gene alignment.

**Supplementary Figure 11. Time series tRI signals of selected SARS-CoV-2 and H5N1 amino acid changes. (a-g)** Panels displaying tRI/prevalence time series charts of selected SNVs giving rise to SARS-CoV-2 spike gene amino acid changes in Omicron sub-lineages at sites R346 (a), K444 (b), L452 (c), F456 (d), A475 (e), N460 (f) and F486 (g), showing tRI (red; area shaded in grey marks tRI significance level) and prevalence (blue) at daily resolution. Charts were generated with EVOtRec from the spike gene alignment covering the entire COVID-19 pandemic. We observed significant tRI for at least one amino acid change at each of these sites, indicating convergent evolution. This is in line with published studies [67–69] which demonstrated that amino acid changes at the sites reported in (a-g) are convergent and were acquired recurrently across several Omicron sub-lineages. **(h-i)** Panels displaying tRI/prevalence time series charts of selected SNVs giving rise to H5N1 HA and PB2 gene amino acid changes, showing tRI (red; area shaded in grey marks tRI significance level) and prevalence (blue) at daily resolution. (h) The HA mutations E2K, D104G, V147M and S336N have been associated with antibody neutralization in bovine H5N1 [161]; we found significant tRI signals for E2K, V147M and S336N. (i) Significant tRI signals of the mammalian-adaptive PB2 mutations E249G [117], A588T [118], D701N [77, 120] and D740N [78].

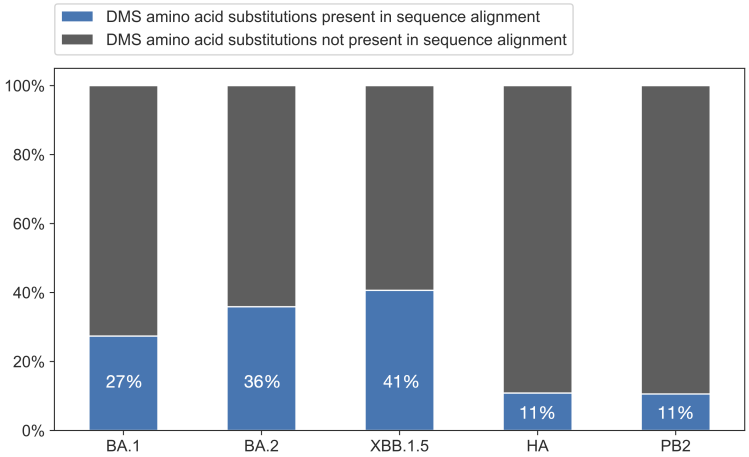



**e**

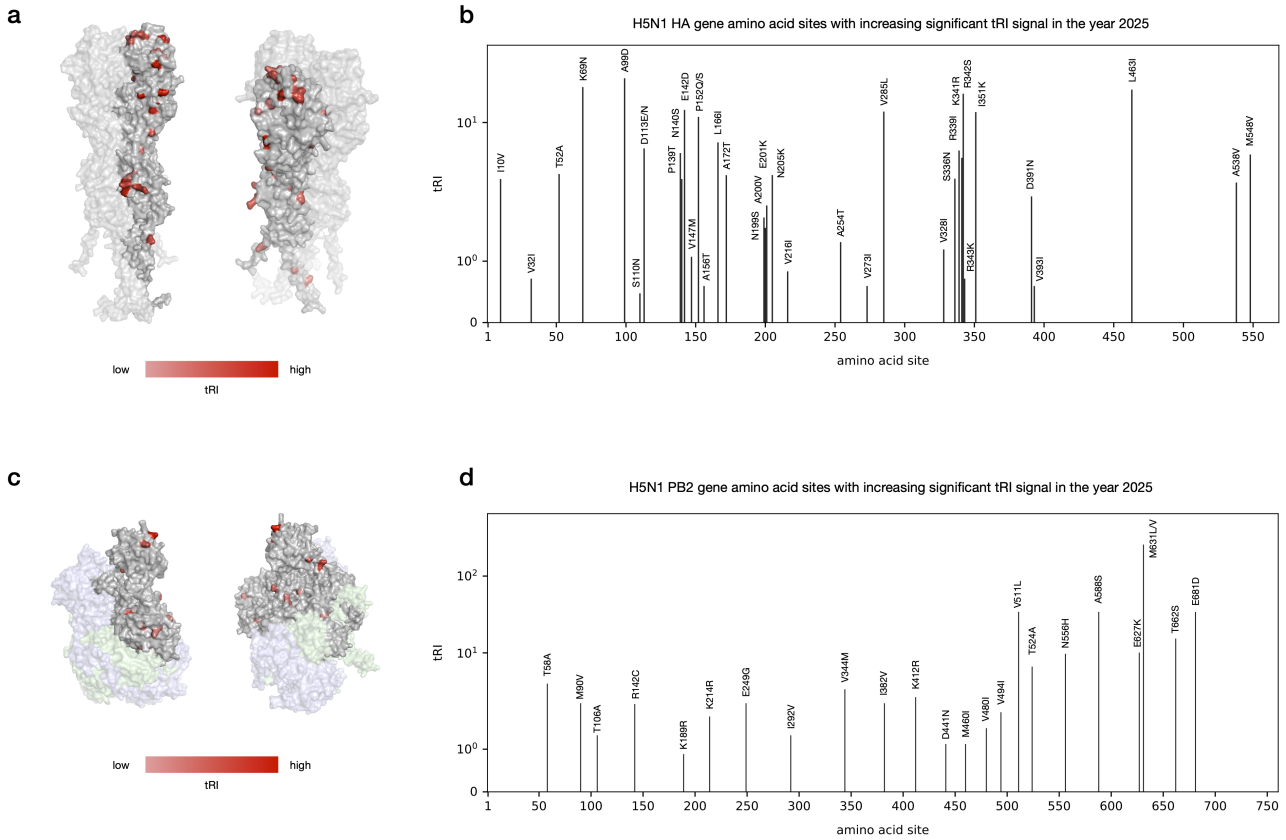
Performance of tRI vs. phylogeny-based fitness effect in predicting fitness increasing SARS-CoV-2 spike amino acid changes (positive DMS effect)

| | tRI | | | fitness effect | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Bloom and Neher, 2023 | | | Haddox et al., 2025 | | |
| **lineage (phenotype)** | precision | recall | $F_1$ score | precision | recall | $F_1$ score | precision | recall | $F_1$ score |
| BA.1 (antibody escape) | 0.38 | 0.36 | **0.37** | 0.38 | 0.27 | 0.31 | 0.37 | 0.28 | 0.32 |
| BA.2 (ACE2-binding) | 0.56 | 0.28 | **0.37** | 0.54 | 0.17 | 0.26 | 0.56 | 0.19 | 0.29 |
| BA.2 (spike mediated entry) | 0.31 | 0.47 | 0.38 | 0.41 | 0.39 | 0.40 | 0.41 | 0.43 | **0.42** |
| XBB.1.5 (ACE2-binding) | 0.51 | 0.31 | **0.39** | 0.48 | 0.29 | 0.36 | 0.47 | 0.28 | 0.35 |
| XBB.1.5 (spike mediated entry) | 0.39 | 0.42 | **0.40** | 0.37 | 0.40 | 0.38 | 0.39 | 0.41 | **0.40** |
| XBB.1.5 (human sera escape) | 0.43 | 0.34 | 0.38 | 0.47 | 0.37 | **0.41** | 0.46 | 0.35 | 0.40 |

**Supplementary Figure 12. SARS-CoV-2 topological signals correlate with positive selection in experimental data from deep mutational scanning. (a-d)** Correlation between EVOtRec tRI signals of SARS-CoV-2 spike amino acid changes in the Omicron spike gene alignment (Methods) and published pseudovirus deep mutational scanning (DMS) measurements [59, 72, 133–135] for various phenotypes of the BA.1 (a), BA.2 (b) and XBB.1.5 (c) sub-lineages. In each case in (a-c), we tested the performance of the tRI in identifying fitness-increasing amino acid changes (i.e. positive DMS effect). For this, we analyzed correlations statistically using contingency tables for binarized tRI (tRI $> 0$ vs. tRI $= 0$) and binarized DMS effect (DMS effect $> 0$ vs. DMS effect $\leq 0$) (a-c) and on a logistic regression analysis for the three phenotypes of XBB.1.5 (d) (Methods). For comparison with phylogeny-based methods, all panels include similar statistical analyses for binarized fitness effect estimates (fitness effect $> 0$ vs. fitness effect $\leq 0$) from [6, 17] using the datasets [136, 137] for BA.1 (a), [136, 138] for BA.2 (b) and [136, 139] for an ordinary least squares regression analysis for XBB.1.5 (c) (Methods). **(e)** Comparison of the performance of tRI vs. phylogeny-based fitness effect estimates [6, 17] in identifying fitness-increasing amino acid changes (i.e. positive DMS effect), based on the results from (a-c).




**Supplementary Figure 13.** Percentages of amino acid changes that have been tested in the DMS experiments in [59, 72–74] for the SARS-CoV-2 Omicron spike gene alignment (covering the sublineages BA.1, BA.2 and XBB.1.5) and the two avian influenza H5N1 alignments (covering the HA and PB2 genes).




**Supplementary Figure 14. Convergent H5N1 amino acid changes.** Amino acid changes on the H5N1 HA and PB2 genes that exhibited increasing significant tRI signal during the year 2025. Displayed tRI scores are taken from the last day of the HA and PB2 tRI time series (7/8 November 2025) (Supplementary Table 3 and Methods). **(a, c)** Two views of the H5N1 HA (a) and PB2 (c) proteins, with residues colored by sum of tRI scores across observed changes at the site. The HA protein is a homo-trimer; one sub-unit is shown with higher opacity and has the recurrently mutated residues highlighted. The polymerase complex is a hetero-trimer consisting of PB1 (pale green), PA (pale blue), and PB2 (light grey). Protein structures were predicted with AlphaFold3 [162] as homo-trimer (HA) and heterotrimer (PB2) based on the reference sequence A/American_Wigeon/South_Carolina/22-000345-001/2021, and visualized using PyMOL [163]. **(b, d)** Bar plots of amino acid changes and their tRI signals across the HA and PB2 gene.