Emergenet: Fast Scalable Pandemic Risk Assessment of Influenza A Strains Circulating In Non-human Hosts

Kevin Wu[1], Jin Li[1], Aaron Esser-Kahn[2,3], and Ishanu Chattopadhyay[1,4,5]★

[1]Department of Medicine, University of Chicago, IL, USA
[2]Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL, USA
[3]Committee on Immunology, University of Chicago, Chicago, IL, USA
[5]Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, IL, USA

★To whom correspondence should be addressed: e-mail: ishanu@uchicago.edu.

Abstract: **Animal influenza viruses emerging into humans have triggered devastating pandemics in the past[1–4]. Yet, our ability to evaluate the pandemic potential of individual strains that do not yet circulate in humans, remains limited. In this study we introduce the Emergenet, to computationally learn how new variants emerge, shaped by evolutionary constraints using only observed genomic sequences of key viral proteins. Analyzing $399,476$ Haemagglutinnin (HA) and Neuraminidase (NA) sequences from public databases, we estimate the likelihood of specific future mutations, and the numerical odds of specific descendants arising via natural processes. After validating our model to forecast the dominant strain(s) for seasonal flu, with Emergenet-based forecasts significantly outperforming WHO recommendations almost consistently over the past two decades for H1N1/H3N2 subtypes, individually in the Northern/Southern hemispheres (average match-improvement 52.07% over two decades, 59.83% over the last decade, and 65.79% over the pre-COVID-19 five year period for H1N1 HA), we assess the pandemic potential of animal strains not yet known to transmit in humans. While the state-of-the-art Influenza Risk Assessment Tool (IRAT) from the CDC to assess such risk includes time-consuming experimental assays, our calculations take $\approx 6sec/$strain, yet strongly correlating with published IRAT scores (correlation: 0.703, p-value: 0.00026). This six orders of magnitude speedup is necessary to exploit current surveillance capacity, and analyze thousands of strains collected annually. Considering $6,066$ wild Influenza A animal viruses sequenced post-2020, we identify risky strains of diverse subtypes, hosts and geo-locations, with six having estimated emergence scores > 6.5. Such scalable risk-ranking can enable preemptive pandemic mitigation, including the targeted inoculation of animal hosts before the first human infection, and outline new public health measures that are potentially effective notwithstanding possible vaccine hesitancy in humans.**

Introduction

Influenza viruses constantly evolve[5], sufficiently altering surface protein structures to evade the prevailing host immunity, and cause the recurring seasonal epidemic. These periodic infection peaks claim a quarter to half a million lives[6] globally, and currently our response hinges on annually inoculating the human population with a reformulated vaccine[5,7]. Among numerous factors that hinder optimal design of the flu shot, failing to correctly predict the future dominant strain dramatically reduces vaccine effectiveness[8]. Despite recent advances[6,9] such predictions remain imperfect. In addition to the seasonal epidemic, influenza strains spilling over into humans from animal reservoirs have triggered pandemics at least four times (1918 Spanish flu/H1N1, 1957 Asian flu/H2N2, 1968 Hong Kong flu/H3N2, 2009 swine flu/H1N1) in the past 100 years[1]. With the memory of the sudden SARS-CoV-2 emergence fresh in our minds, a looming question is whether we can preempt and mitigate such events in the future. Influenza A, partly on account of its segmented genome and its wide prevalence in common animal hosts, can easily incorporate genes from multiple strains and (re)emerge as novel human pathogens[3,10], thus harboring a high pandemic potential.

One possible approach to mitigating such risk is to identify animal strains that do not yet circulate in humans, but is likely to spill-over and quickly achieve human-to-human (HH) transmission capability. While global surveillance efforts collect wild specimens from diverse hosts and geo-locations annually, our ability to objectively, reliably and scalably risk-rank individual strains remains limited[12], despite some recent progress[13–15].

The Center for Disease Control's (CDC) current solution to this problem is the Influenza Risk Assessment Tool (IRAT)[16]. Subject matter experts (SME) score strains based on the number of human infections, infection and transmission in

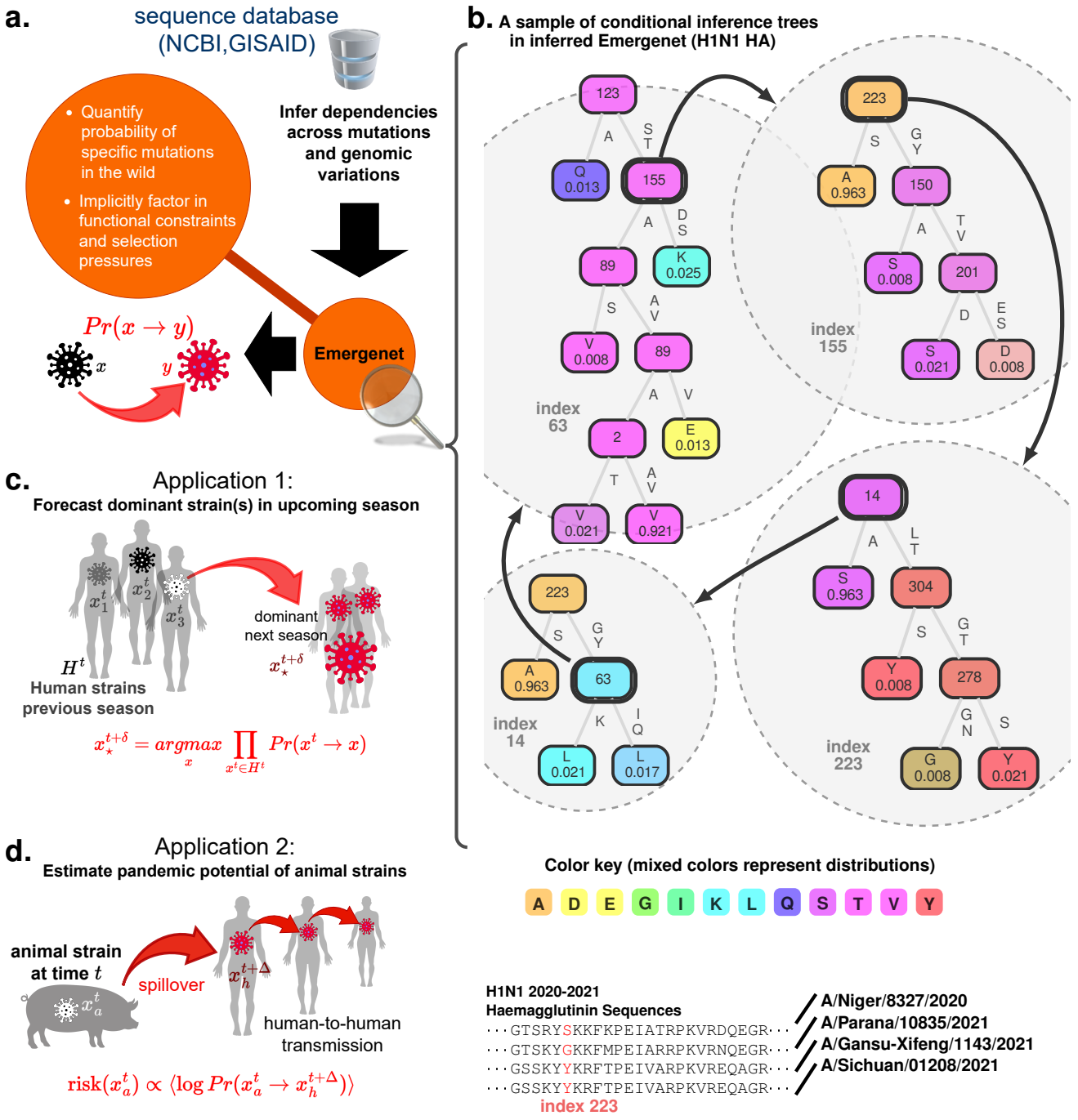


Fig. 1. **Emergenet inference and applications**. **a**, Variations of genomes for identical subtypes of Influenza A are analyzed to infer a recursive forest of conditional inference trees[11] – the Emergenet– which maximally captures the emergent dependencies between an a priori unspecified number of mutations. With these inferred dependencies we can estimate the numerical odds of specific mutations, and by extension, the numerical value of the probability of one strain giving rise to another in the wild, under complex selection pressures from the background. b, Snapshot of decision trees from the Emergenet inferred for H1N1 HA sequences collected in 2020-2021, which reveals a cyclic dependency. In general, every internal node of a component tree can be "expanded" into its own tree, underscoring the recursive structure of the Emergenet. c, First application: forecast dominant strain(s) for the next flu season, using only sequences collected up to six months prior and the inferred Emergenet, using data from the past year. d, Second application: estimation of the pandemic risk posed by individual animal strains that are still not known to circulate in humans (see Eq. (13) in Online Methods).

laboratory animals, receptor binding characteristics, population immunity, genomic analysis, antigenic relatedness, global prevalence, pathogenesis, and treatment options, which are averaged to obtain two scores (between 1 and 10) that estimate 1) the emergence risk and 2) the potential public health impact on sustained transmission. IRAT scores are potentially subjective, and depend on multiple experimental assays, possibly taking weeks to compile for a single strain. This results in a scalability bottleneck, particularly with thousands of strains being sequenced annually.

Here we introduce a pattern recognition algorithm to automatically parse out emergent evolutionary constraints operating on Influenza A viruses in the wild, to provide a less-heuristic theory-backed scalable solution to emergence

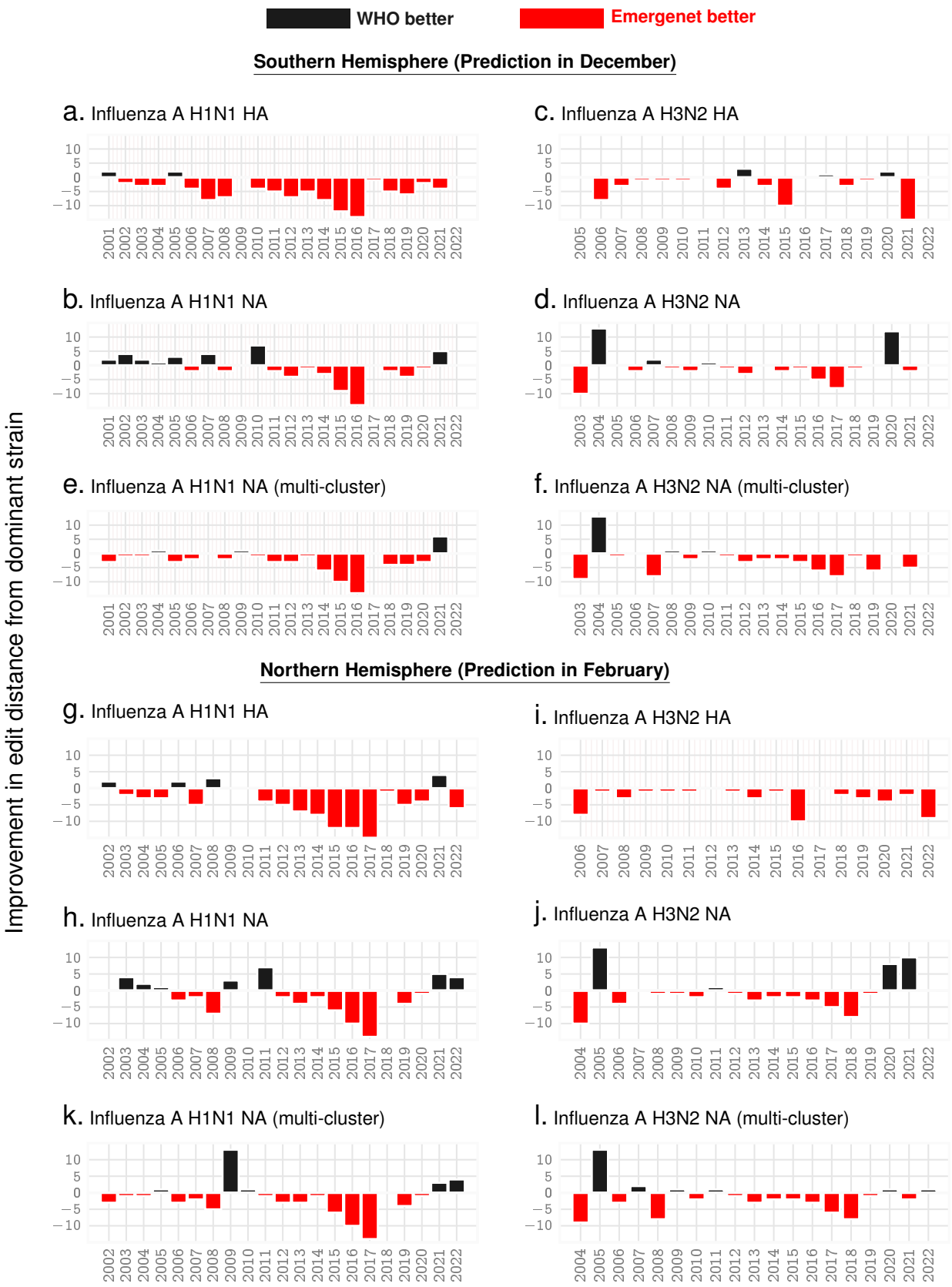


Fig. 2. **Seasonal predictions for Influenza A.** Relative out-performance of Emergenet predictions against WHO recommendations for H1N1 and H3N2 sub-types for the HA and NA coding sequences over the both hemispheres. The negative bars (red) indicate the reduced edit distance between the predicted sequence and the actual dominant strain that emerged that year. Note that the recommendations for the north are given in February, while that for the south are given at the previous December, keeping in mind that the flu season in the south begins a few months early (e.g. for the 2021-2022 flu season, southern data in the table is labelled '2021' and northern is labelled '2022').

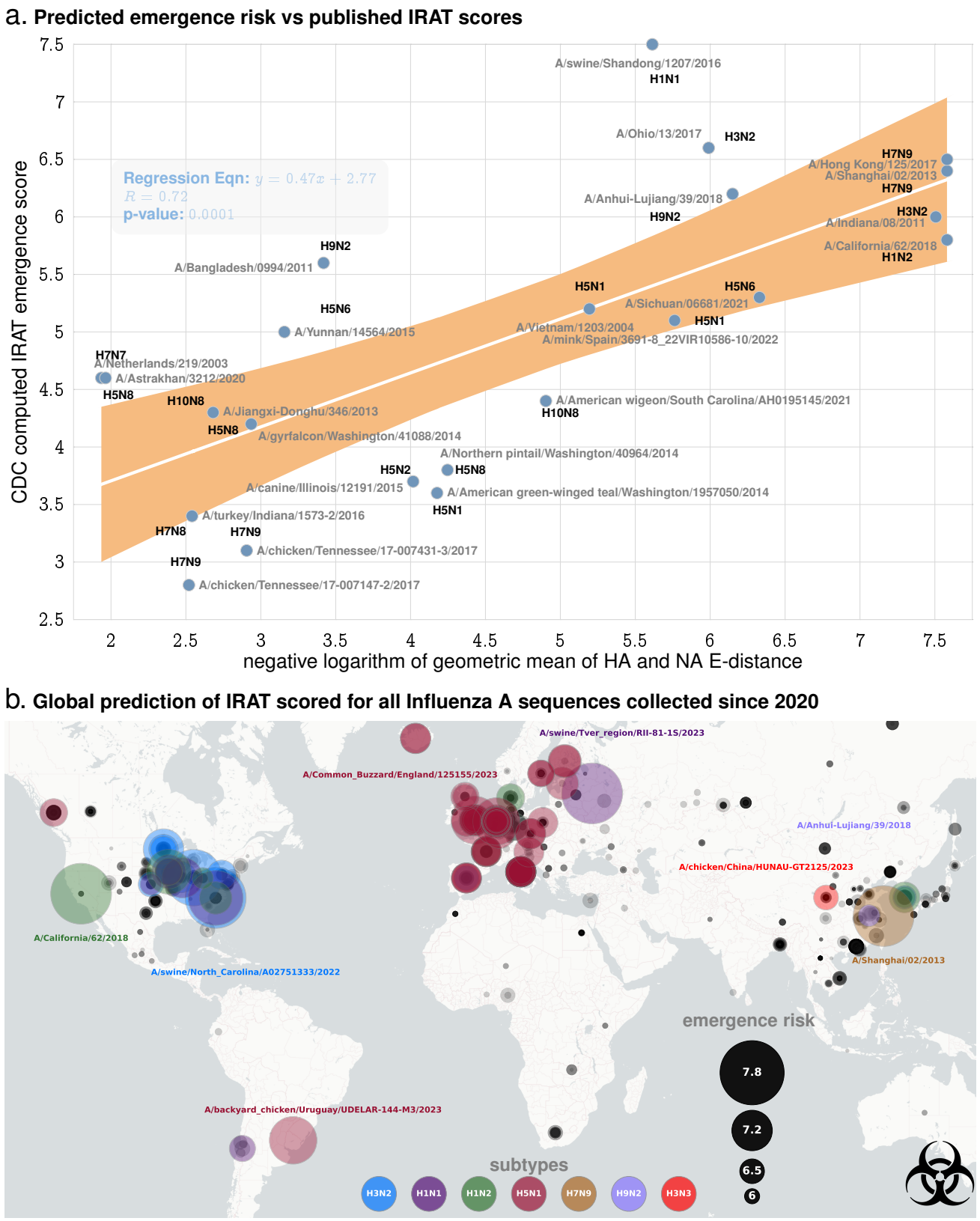


Fig. 3. **Emergenet based estimation of IRAT score**. a, We find an approximate linear relationship between average E-distance from human circulating strains (negative geometric mean of the E-distance for HA and NA sequences) and the published IRAT emergence scores calculated by the CDC. b, Estimation of the IRAT emergence score via fitting a GLM model to the E-distancees estimated from the Emergenet. c, Estimation of IRAT impact scores via fitting a separate GLM model to the the E-distancees. d, Identifying risky Influenza A strains amongst those collected between 01/2020-09/2022 using our approach.

prediction. Our approach is centred around numerically estimating the probability $Pr(x \rightarrow y)$ of a strain x spontaneously giving rise to y. We show that this capability is key to preempting strains which are expected to be in future circulation, and 1) reliably forecast dominant strains of seasonal epidemics, and 2) approximate IRAT scores of non-human strains without experimental assays or SME scoring.

Fig. 4. Phylogeny constructed with edit distances, with Influenza A strains collected between 01/2020-09/2023, with estimated IRAT emergence risk > 6.0, and collapsing leaves which differ by less than 20 edits in the HA, displaying the most risky strains in the collapsed group. The top risky strains are marked with a red arrowhead, which comes from diverse animal hosts, and geographic regions.

Emergenet Inference

To uncover relevant evolutionary constraints, we analyzed variations (point substitutions and indels) of the residue sequences of key proteins implicated in cellular entry and exit[1,17], namely HA and NA respectively. By representing these constraints within a predictive framework – the Emergenet (Enet) – we estimated the odds of a specific mutation to arise in future, and consequently the probability of a specific strain spontaneously evolving into another (Fig. 1a). Such explicit calculations are difficult without first inferring the variation of mutational probabilities and the potential residue replacements from one positional index to the next along the protein sequence. The many well-known classical DNA substitution models[18] or standard phylogeny inference tools which assume a constant species-wise mutational characteristics, are not applicable here. Similarly, newer algorithms such as FluLeap[19] which identifies host tropism from sequence data, or estimation of species-level risk[15] do not allow for strain-specific assessment.

The dependencies we uncover are shaped by a functional necessity of conserving/augmenting fitness. Strains must be sufficiently common to be recorded, implying that the sequences from public databases that we train with have high replicative fitness. Lacking kinetic proofreading, Influenza A integrates faulty nucleotides at a relatively high rate ($10^{-3} - 10^{-4}$) during replication[20,21]. However, few variations are actually viable, leading to emergent dependencies between such mutations. Furthermore, these fitness constraints are not time-invariant. The background strain distribution, and selection pressure from the evolution of cytotoxic T lymphocyte epitopes[22–26] in humans can change quickly. With a sufficient number of unique samples to train on for each flu season, the Emergenet (recomputed for each time-period) is expected to automatically factor in the evolving host immunity, and the current background environment.

Structurally, an Emergenet comprises an interdependent collection of local predictors, each aiming to predict the residue at a particular index using as features the residues at other indices (Fig. 1b). Thus, an Emergenet comprises almost as many such position-specific predictors as the length of the sequence. These individual predictors are implemented as conditional inference trees[11], in which nodal splits have a minimum pre-specified significance in differentiating the child nodes. Thus, each predictor yields an estimated conditional residue distribution at each index. The set of residues acting as features in each predictor are automatically identified, $e.g.$, in the fragment of the H1N1 HA Emergenet (2020-2021, Fig 1b), the predictor for residue 63 is dependent on residue 155, and the predictor for 155 is dependent on 223, the predictor for 223 is dependent on 14, and the residue at 14 is again dependent on 63, revealing a cyclic dependency. The complete Emergenet harbors a vast number of such relationships, wherein each internal node of a tree may be "expanded" to its own tree. Owing to this recursive expansion, a complete Emergenet substantially captures the complexity of the rules guiding evolutionary change as evidenced by our out-of-sample validation.

In our first application (predicting future dominant strains) we used H1N1 and H3N2 HA and NA sequences from Influenza A strains in the public NCBI and GISAID databases recorded between 2000-2022 (387,067 in total, Supplementary Table S-1). We construct Emergenets separately for H1N1 and H3N2 subtypes, and for each flu season using HA sequences, yielding 84 models in total for predicting seasonal dominance. Using only sequence data is advantageous since deeper antigenic characterization tend to be substantially low-throughput compared to genome sequencing[27]. However, deep mutational scanning (DMS) assays have been shown to improve seasonal prediction[6]. Despite limiting ourselves to only genotypic information (and subtypes), our approach distills emergent fitness-preserving constraints that outperform reported DMS-augmented strategies.

Inference of the Emergenet predictors is our first step, which then induces an intrinsic distance metric between strains. The E-distance (i.e. Emergenet distance) (Eq. (5) in Online Methods) is defined as the square-root of the Jensen-Shannon (JS) divergence[28] of the conditional residue distributions, averaged over the sequence. Unlike the classical approach of measuring the number of edits between sequences, the E-distance is informed by the Emergenet-inferred dependencies, and adapts to the specific subtype, allele frequencies, and environmental variations. Central to our approach is the theoretical result (Theorem 1 in Online Methods) that the E-distance approximates the log-likelihood of spontaneous change $i.e.\ \log Pr(x \to y)$. Note that despite general correlation between E-distance and edit-distance, the E-distance between fixed strains can change if only the background environment changes (Supplementary Table S-2,S-3). In in-silico experiments, We find that while random mutations to genomic sequences produce rapidly diverging sets, Emergenet-constrained replacements produce sequences that are verifiably meaningful (In-silico Corroboration of Emergenet's Capability To Capture Biologically Meaningful Structure, Online Methods and Supplementary Fig. S-1).

Determining the numerical odds of a spontaneous jump $Pr(x \to y)$ (Fig. 1) allows us to frame the problem of forecasting dominant strain(s), and that of estimating the pandemic potential of an animal strain as mathematical propositions (albeit with some simplifying assumptions), with approximate solutions (Fig. 1c-d). Thus, a dominant strain for an upcoming season may be identified as one which maximizes the joint probability of simultaneously arising from each (or most) of the currently circulating strains (Fig. 1c). This does not deterministically specify the dominant strain, but a strain satisfying this criterion has high odds of acquiring dominance. And, a pandemic risk score of a novel strain may be estimated by the probability of it giving rise to a well-adapted human strain. In the context of forecasting future dominant strain(s), we derive a search criteria (Predicting Dominant Seasonal Strains, Online Methods) from the above proposition, to identify historical strain(s) that are expected to be close to the next dominant strain(s):

$$x_\star^{t+\delta} = \underset{y \in \cup_{\tau \leq t} H^\tau}{\arg\min} \left(\sum_{x \in H^t} \theta^{[t]}(x, y) - \left| H^t \right| A \ln \omega_y \right) \tag{1}$$

where $x_\star^{t+\delta}$ is a predicted dominant strain at time $t+\delta$, H^t is the set of currently circulating human strains at time t observed over the past year, $\theta^{[t]}$ is the E-distance informed by the inferred Emergenet using sequences in H^t, ω_y is the estimated probability of strain y being generated by the Emergenet, and A is a constant dependent on the sequence length and significance threshold used. The first term gets the solution close to the centroid of the current strain distribution (in the E-distance metric, and not the standard edit distance), and the second term relates to how common the genomic patterns are amongst recent human strains.

Predicting Future Dominant Strains

Prediction of the future dominant strain as a close match to a historical strain allows out-of-sample validation against past World Health Organization (WHO) recommendations for the flu shot, which is reformulated about six months in advance based on a cocktail of historical strains determined via global surveillance[29]. For each year of the past two decades, we first computed three clusters of strains in the E-distance metric on their HA sequences. In each cluster, we calculated strain forecasts using Eq. (1) with data available six months before the target season, taking our first and second recommendations from the two largest clusters. We also calculated the top ten dominant strains for both HA and NA from the target season, ranked by closeness to the centroid in the strain space that season in the edit distance metric. We measured forecast performance by the average number of mutations by which the predicted HA/NA sequences deviated from the top ten dominant strains. Our Emergenet-informed forecasts outperform WHO/CDC recommended flu vaccine compositions consistently over the past two decades, for both H1N1 and H3N2 subtypes, individually in the northern and the southern hemispheres (which have distinct recommendations[7]). For H1N1 HA, the Emergenet recommendation outperforms WHO by 52.07% on average over the last two decades, and 59.83% on average in the last decade, and by 65.79% in the period 2015-2019 (5 years pre-COVID-19). The gains for H1N1 NA over the same time periods are 46.41%, 40.31%, and 54.85% respectively. For H3N2 HA, the Emergenet recommendation outperforms WHO by 42.39% on average over the last two decades, and 35.00% on average in the last decade, and by 41.85% in the period 2015-2019. The gains for H3N2 NA over the same time periods are 46.90%, 42.31%, and 47.65% respectively (Extended Data Table 1). Detailed predictions, along with historical strains closes to the observed dominant one are tabulated in Extended Data Tables 2 through 5. Visually, Fig. 2 illustrates the relative gains computed for different subtypes and hemispheres.

Comparing the Emergenet inferred strain (ENT) against the one recommended by the WHO, we find that the residues that only the Emergenet recommendation matches correctly with dominant strain (DOM), while the WHO recommendation fails, are largely localized within the RBD, with $> 57\%$ occurring within the RBD on average (Extended Data Fig. 1a), and 3) when the WHO strain deviates from the ENT/DOM matched residue, the "correct" residue is often replaced in the WHO recommendation with one that has very different side chain, hydropathy and/or chemical properties (Extended Data Fig.-1b-f), suggesting deviations in recognition characteristics[30,31]. Combined with the fact that we find circulating strains are almost always within a few edits of the DOM (Supplementary Fig. S-2), these observations suggest that hosts vaccinated with the ENT recommendation is can have season-specific antibodies that recognize a larger cross-section of the circulating strains.

Estimating Pandemic Risk of Non-human Strains

Our primary claim, however, is the ability to estimate the pandemic potential of novel animal strains, via a time-varying E-risk score $\rho_t(x)$ for a strain x not yet found to circulate in human hosts. We show that (Measure of Pandemic Potential, Online Methods):

$$\rho_t(x) \triangleq -\frac{1}{|H^t|}\sum_{y\in H^t}\theta^{[t]}(x,y) \tag{2}$$

scales as the average log-likelihood of $Pr(x \to y)$ where y is any human strain of a similar subtype to x, and $\theta^{[t]}$ is the E-distance informed by the Emergenet computed from recent human strains H_t at time t of the same subtype as x, observed over the past year. As before, the Emergenet inference makes it possible to estimate $\rho_t(x)$ explicitly.

To validate our score against CDC-estimated IRAT emergence scores, we construct Emergenet models for HA and NA sequences using subtype-specific human strains, typically collected within the year prior to the assessment date, *e.g.*, the assessment date for A/swine/Shandong/1207/2016 is 06/2020, and we use human H1N1 strains collected between 01/07/2019 and 06/30/2020 for the Emergenet inference. For sub-types with less recorded human strains (H1N2, H7N7), we consider all subtype-specific human strains collected up to the assessment date to infer our Emergenet. For subtypes with very few or no recorded human strains even without a lower date bound (H5N2, H5N6, H5N8, H7N8, H9N2, H10N8), we construct the Emergenet using all human strains that match the HA subtype, *e.g.* H5Nx for H5N2, H5N6, and H5N8. This addresses the general concern that Emergenet may not be able assess the threat posed by the viruses that we have yet to detect in sufficient numbers; the strains for which we used this method (marked by ** in Supplementary Table S-6) fit along the fit line in Fig. 3. We then compute the E-risk for both HA and NA sequences (using Eq. (2)), finally reporting their geometric mean as our estimated risk for the strain. Considering IRAT emergence scores of 22 strains published by the CDC, we find strong out-of-sample support (correlation of 0.704, pvalue < 0.00026, Fig. 3) for this claim. Importantly, each E-risk score is computable in approximately 6 seconds as opposed to potentially weeks taken by IRAT experimental assays and SME evaluation. Additionally, using a subtype-

specific Emergenet modulates the metric of comparison of genomic sequences, adapting it to the specific subtype of the virus.

The time-dependence of the E-risk reflects the impact of the changing background, and recomputing the risk estimates using Emergenets constructed from the recent circulating strains instead of using those from when the IRAT assessments took place at the CDC, worsens the correlation (0.597, p-value 0.003, see Supplementary Table S-**??**).

To map the Emergenet distances to more recognizable IRAT scores, we train a general linear model (GLM) from the the HA/NA-based E-risk values (Multivariate Regression to Identify Map from E-distance to Estimated IRAT scores, Online Methods and Supplementary Table S-4). Since the CDC-estimated IRAT impact scores are strongly correlated with their IRAT emergence scores (correlation of 0.8015), we also trained a separate GLM to estimate the impact score from the E-risk values (Supplementary Table S-5). Finally, we estimate the IRAT scores of all $6,066$ Influenza A strains sequenced globally between 2020 through 04/2022, and identify the ones posing maximal risk (Fig. 3c). $1,773$ strains turn out to have a predicted emergence score > 6.0. However, many of these strains are highly similar, differing by only a few edits. To identify the sufficiently distinct risky strains, we constructed the standard phylogeny from HA sequences with score > 6 (Fig. 4), and collapsed all leaves within 15 edits, showing only the most risky strain within a collapsed group. This leaves 75 strains (Fig. 4), with 68 having emergence risk > 6.25, and 6 with risk above 6.5 (Extended Data Table 8). Subtypes of the risky strains are overwhelmingly H1N1, followed by H3N2, with a small number of H7N9 and H9N2. Five maximally risky strains with emergence score > 6.58 are identified to be: A/swine/Missouri/A02524711/2020 (H1N1), A/Camel/Inner_Mongolia/XL/2020 (H7N9), A/swine/Indiana/A02524710/2020 (H3N2), A/swine/North Carolina/ A02479173/2020 (H1N1), and A/swine/Tennessee/ A02524414/2022 (H1N1). Additionally, A/mink/China/chick embryo/2020 (H9N2), with a lower estimated emergence score (6.26) is also important, as the most risky H9N2 strain in our analysis. We compare the HA sequences along with two dominant human strains in 2021-2022 season (Extended Data Fig. 2), which shows substantial residue replacements, in and out of the receptor binding domain (RBD).

Swines are known to be efficient mixing vessels[3,32–34], and hence unsurprisingly host a large fraction of the risky strains ($> 80\%$ over 6.0, to over 50% over 6.5). Also, as expected, most of these swine strains are of H1N1 subtype, with the other subtypes having emerged into humans more recently. Our finding that a H7N9 poses substantial risk is likewise not surprising: HH transmission has been suspected in Asian-lineage H7N9 strains, and are rated by IRAT as having the greatest potential to cause a pandemic[35]. The finding of the most risky H9N2 strain in a mink is also unsurprising, in the light of these hosts been recently suggested as efficient mixing vessels to breed human-compatible strains[36]. Thus, qualitatively our results are well aligned with the current expectations; nevertheless the ability to quantitatively rank specific strains which pose maximal risk is a crucial new capability enabling proactive pandemic mitigation efforts.

Conclusion

While numerous tools exist for ad hoc quantification of genomic similarity[9,18,37–40], higher similarity between strains in these frameworks is not sufficient to imply a high likelihood of a jump. To the best of our knowledge, the Emergenet algorithm is the first of its kind to learn an appropriate biologically meaningful comparison metric from data, without assuming any model of DNA or amino acid substitution, or a genealogical tree a priori. While the effect of the environment and selection cannot be inferred from a single sequence, an entire database of observed strains, processed through the right lens, can parse out useful predictive models of these complex interactions. Our results are aligned with recent studies demonstrating effective predictability of future mutations for different organisms[41,42].

The E-distance calculation is currently limited to analogous sequences (such as point variations of the same protein from different viral subtypes), and the Emergenet inference requires a sufficient diversity of observed strains. A multivariate regression analysis indicates that the most important factor for our approach to succeed is the diversity of the sequence dataset (Multivariate Regression to Understand Data Characteristics Necessary For Emergenet Modeling, Online Methods and Supplementary Table S-6), which would exclude applicability to completely novel pathogens with no related human variants, and ones that evolve very slowly. Nevertheless, the tools reported here can improve effectiveness of the annual flu shot, and perhaps allow for the development of preemptive vaccines to target risky animal strains before the first human infection in the next pandemic. Apart from outlining new precision public health measures to avert pandemics, such strategies might also help to non-controversially counter the impact of vaccine hesitancy which has interfered with optimal pandemic response in recent times.

ONLINE METHODS

We briefly describe the proposed computational framework.

Emergenet Framework

We do not assume that the mutational variations at the individual indices of a genomic sequence are independent (See Fig 1a). Irrespective of whether mutations are truly random[43], since only certain combinations of individual mutations are viable, individual mutations across a genomic sequence replicating in the wild appear constrained, which is what is

explicitly modeled in our approach.

Consider a set of random variables $X = \{X_i\}$, with $i \in \{1, \cdots, N\}$, each taking value from the respective sets Σ_i. Here each X_i is the random variable modeling the "outcome" *i.e.* the AA residue at the i^{th} index of the protein sequence. A sample $x \in \prod_1^N \Sigma_i$ is an ordered N-tuple, which is a specific strain in this context, consisting of a realization of each of the variables X_i with the i^{th} entry x_i being the realization of random variable X_i.

We use the notation x_{-i} and $x^{i,\sigma}$ to denote:

$$x_{-i} \triangleq x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_N \tag{3a}$$

$$x^{i,\sigma} \triangleq x_1, \cdots, x_{i-1}, \sigma, x_{i+1}, \cdots, x_N, \sigma \in \Sigma_i \tag{3b}$$

Also, $\mathscr{D}(S)$ denotes the set of probability measures on a set S, *e.g.*, $\mathscr{D}(\Sigma_i)$ is the set of distributions on Σ_i.

We note that X defines a random field[44] over the index set $\{1, \cdots, N\}$.

Definition 1 (Emergenet)**.** *For a random field $X = \{X_i\}$ indexed by $i \in \{1, \cdots, N\}$, the Emergenet is defined to be the set of predictors $\Phi = \{\Phi_i\}$, i.e., we have:*

$$\Phi_i : \prod_{j \neq i} \Sigma_j \to \mathscr{D}(\Sigma_i), \tag{4}$$

where for a sequence x, $\Phi_i(x_{-i})$ estimates the distribution of X_i on the set Σ_i.

We use conditional inference trees as models for predictors[11], although more general models are possible.

Biology-Aware Distance Between Sequences

The mathematical form of our metric is not arbitrary; JS divergence is a symmetricised version of the more common KL divergence[28] between distributions, and among different possibilities, the E-distance is the simplest metric such that the likelihood of a spontaneous jump (See Eq. (8) in Methods) is provably bounded above and below by simple exponential functions of the E-distance.

Definition 2 (E-distance: adaptive biologically meaningful dissimilarity between sequences)**.** *Given two sequences $x, y \in \prod_1^N \Sigma_i$, such that x, y are drawn from the populations P, Q inducing the Emergenet Φ^P, Φ^Q, respectively, we define a pseudo-metric $\theta(x, y)$, as follows:*

$$\theta(x, y) \triangleq \mathbf{E}_i \left(\mathbb{J}^{\frac{1}{2}} \left(\Phi_i^P(x_{-i}), \Phi_i^Q(y_{-i}) \right) \right) \tag{5}$$

where $\mathbb{J}(\cdot, \cdot)$ is the Jensen-Shannon divergence[45] and $\mathbf{E}_i$ indicates expectation over the indices.

The square-root in the definition arises naturally from the bounds we are able to prove, and is dictated by the form of Pinsker's inequality[28], ensuring that the sum of the length of successive path fragments equates the length of the path.

Membership Degree

For our modeling to be reliable, we need a quantitative test of how well the Emergenet represents the data. Here, we formulate an explicit membership test to ascertain if individual samples may indeed be generated by the Emergenet with sufficiently high probability.

Definition 3 (Membership probability of a sequence)**.** *Given a population P inducing the Emergenet Φ^P and a sequence x, we can compute the membership probability of x:*

$$\omega_x^P \triangleq Pr(x \in P) = \prod_{j=1}^{N} \left(\Phi_j^P(x_{-j})|_{x_j} \right) \tag{6}$$

x_j is the j^{th} entry in x, and is thus an element in the set Σ_j. Since we are mostly concerned with the case where Σ_j is a finite set, $\Phi_j^P(x_{-j})|_{x_j}$ is the entry in the probability mass function corresponding to the element of Σ_j which appears at the j^{th} index in sequence x.

We can carry out this calculation for a sequence x known to be in the population P as well, which allows us to define the membership degree ω_x^P.

Definition 4 (Membership degree)**.** *Let X be a random field representing a population P, ie.. $X = x$ is a randomly drawn sequence from P. Then the membership degree ω^P is a function of the random variable X:*

$$\omega^P(X) \triangleq \prod_{j=1}^{N} \left(\Phi_j^P(X_{-j})|_{X_j} \right) \tag{7}$$

Note that ω^P takes values in the unit interval $[0, 1]$, and the probability x is a member of the population P is $\omega^P(X = x)$, denoted briefly as ω_x^P or ω_x if P is clear from context.

Since $\omega^P(X)$ is a random variable, we can now compute sets of sequences that better represent the population P, and ones that are on the fringe. We can also evaluate using a pre-specified significance-level if a particular sequence is not from the population P.

Theoretical Probability Bounds

The Emergenet framework allows us to rigorously compute bounds on the probability of a spontaneous change of one strain to another, brought about by chance mutations. While any sequence of mutations is equally likely, the "fitness" of the resultant strain, or the probability that it will even result in a viable strain, or not. Thus the necessity of preserving function dictates that not all random changes are viable, and the probability of observing some trajectories through the sequence space are far greater than others. The Emergenet framework allows us to explore this constrained dynamics, as revealed by a sufficiently large set of genomic sequences.

The mathematical intuition relating E-distance to the log-likelihood of spontaneous change is similar to quantifying the odds of a rare biased outcome when we toss a fair coin. While for an unbiased coin, the odds of roughly 50% heads is overwhelmingly likely, large deviations do happen rarely, and it turns out that the probability of such rare deviations can be explicitly quantified with existing statistical theory[46]. Generalizing to non-uniform conditional distributions inferred by the Emergenet, the likelihood of a spontaneous transition by random chance may also be similarly bounded.

We show in Theorem 1 in the supplementary text that at a significance level α, with a sequence length N, the probability of spontaneous jump of sequence x from population P to sequence y in population Q, $Pr(x \to y)$, is bounded by:

$$\omega_y^Q e^{\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \geqq Pr(x \to y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \tag{8}$$

where ω_y^Q is the membership probability of strain y in the target population, N is the sequence length, and α is the statistical signifacnce level.

Predicting Dominant Seasonal Strains

Analyzing the distribution of sequences observed to circulate in the human population at the present time allows us to forecast dominant strain(s) in the next flu season as follows:

Let $x_\star^{t+\delta}$ be a dominant strain in the upcoming flu season at time $t+\delta$, where H^t is the set of observed strains presently in circulation in the human population (at time t). We will assume that the Emergenet is constructed using the sequences in teh set H^t, and remains unchanged upto $t+\delta$. Since this set is a function of time, the inferred Emergenet also changes with time, and the induced E-distance is denoted as $\theta^{[t]}(\cdot,\cdot)$.

From the RHS bound established in Theorem 1 (See Eq. (8) above) in the supplementary text, we have:

$$\ln \frac{Pr(x \to x^{t+\delta})}{\omega_{x^{t+\delta}}} \geqq -\frac{\sqrt{8}N^2}{1-\alpha}\theta^{[t]}(x, x^{t+\delta}) \tag{9}$$

$$\Rightarrow \sum_{x\in H^t} \ln \frac{Pr(x \to x^{t+\delta})}{\omega_{x^{t+\delta}}} \geqq \sum_{x\in H^t} -\frac{\sqrt{8}N^2}{1-\alpha}\theta^{[t]}(x, x^{t+\delta}) \tag{10}$$

$$\Rightarrow \sum_{x\in H^t} \theta^{[t]}(x, x^{t+\delta}) - \left|H^t\right| A \ln \omega_{x^{t+\delta}} \geqq A \ln \frac{1}{\prod_{x\in H^t} Pr(x \to x^{t+\delta})} \tag{11}$$

where $A = \frac{1-\alpha}{\sqrt{8}N^2}$, where N is the sequence length considered, and α is a fixed significance level. Since minimizing the LHS maximizes the lower bound on the probability of the observed strains simultaneously giving rise to $x^{t+\delta}$, a dominant strain $x_\star^{t+\delta}$ may be estimated as a solution to the optimization problem:

$$x_\star^{t+\delta} = \underset{y\in\cup_{\tau\leq t}H^\tau}{\arg\min} \sum_{x\in H^t} \theta^{[t]}(x,y) - \left|H^t\right| A \ln \omega_y \tag{12}$$

Measure of Pandemic Potential

We measure the potential of an animal strain x_a^t to spillover and become HH capable as a human strain $x_h^{t+\delta}$, via the proposed E-risk defined as follows:

$$\rho(x_a^t) \triangleq -\frac{1}{|H^t|}\sum_{x\in H^t} \theta^{[t]}(x_a^t, x) \tag{13}$$

where as before H^t is the set of human strains observed recently (we take this as strains collected within the past year), and $\theta^{[t]}$ is the E-distance induced by the Emergenet computed from the sequences in H^t.

The intuition here is that a lower bound of $\rho(x_a^t)$ scales as average log-likelihood of the x_a^t giving rise to a human strains in circulation at time t. Since the strains in H^t are already HH capable, a high average likelihood of producing a similar strain has a high potential of being a HH cabale novel variant, which is a necessary condition of a pandemic strain. To

establish the lower bound, we note that from Theorem 1 (See Eq. (8) above) in the supplementary text, we have:

$$\sum_{y\in H^t} \ln \left| \frac{Pr(x_a^t \to y)}{\omega_y} \right| \leqq -\frac{\sqrt{8}N^2}{1-\alpha} \left|H^t\right| \rho(x_a^t) \tag{14}$$

Denoting, $A = \frac{1-\alpha}{\sqrt{8}N^2}$, $A \ln(\prod_{y\in H^t} \omega_y) = C$, and $\langle\cdot\rangle$ as the geometric mean function, we have:

$$\Rightarrow \rho(x_a^t) \geqq A \ln \left(\prod_{y\in H^t} Pr(x_a^t \to y) \right)^{1/\left|H^t\right|} + C \tag{15}$$

$$\Rightarrow \rho(x_a^t) \geqq A \ln \left\langle Pr(x_a^t \to x_h^{t+\delta}) \right\rangle + C \tag{16}$$

Noting that A, C are not functions of x_a^t, we conclude that a lower bound of the proposed risk measure $\rho(\cdot)$ scales with the average loglikelihood of producing strains close to a circulating human strain at the current time.

Proof of Probability Bounds

Theorem 1 (Probability bound)**.** *Given a sequence x of length N that transitions to a strain $y \in Q$, we have the following bounds at significance level α.*

$$\omega_y^Q e^{\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \geqq Pr(x \to y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \tag{17}$$

where ω_y^Q is the membership probability of strain y in the target population Q (See Def. 3), and $\theta(x, y)$ is the q-distance between x, y (See Def. 2).

Proof. Using Sanov's theorem[28] on large deviations, we conclude that the probability of spontaneous jump from strain $x \in P$ to strain $y \in Q$, with the possibility $P \neq Q$, is given by:

$$Pr(x \to y) = \prod_{i=1}^{N} \left(\Phi_i^P(x_{-i})|_{y_i}\right) \tag{18}$$

Writing the factors on the right hand side as:

$$\Phi_i^P(x_{-i})|_{y_i} = \Phi_i^Q(y_{-i})|_{y_i} \left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}} \right) \tag{19}$$

we note that $\Phi_i^P(x_{-i})$, $\Phi_i^Q(y_{-i})$ are distributions on the same index i, and hence:

$$|\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}| \leqq \sum_{y_i\in\Sigma_i} |\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}| \tag{20}$$

Using a standard refinement of Pinsker's inequality[47], and the relationship of Jensen-Shannon divergence with total variation, we get:

$$\theta_i \geqq \frac{1}{8}|\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}|^2 \Rightarrow \left| 1 - \frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}} \right| \leqq \frac{1}{a_0}\sqrt{8\theta_i} \tag{21}$$

where a_0 is the smallest non-zero probability value of generating the entry at any index. We will see that this parameter is related to statistical significance of our bounds. First, we can formulate a lower bound as follows:

$$\log\left(\prod_{i=1}^{N} \frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) = \sum_i \log\left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) \geqq \sum_i \left(1 - \frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}}\right) \geqq \frac{\sqrt{8}}{a_0}\sum_i \theta_i^{1/2} = -\frac{\sqrt{8}N}{a_0}\theta \tag{22}$$

Similarly, the upper bound may be derived as:

$$\log\left(\prod_{i=1}^{N} \frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) = \sum_i \log\left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) \leqq \sum_i \left(\frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}} - 1\right) \leqq \frac{\sqrt{8}N}{a_0}\theta \tag{23}$$

Combining Eqs. 22 and 23, we conclude:

$$\omega_y^Q e^{\frac{\sqrt{8}N}{a_0}\theta} \geqq Pr(x \to y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N}{a_0}\theta} \tag{24}$$

Now, interpreting a_0 as the probability of generating an unlikely event below our desired threshold ($i.e.$ a "failure"), we note that the probability of generating at least one such event is given by $1 - (1 - a_0)^N$. Hence if α is the pre-specified significance level, we have for $N >> 1$:

$$a_0 \approx (1 - \alpha)/N \tag{25}$$

Hence, we conclude, that at significance level $\geqq \alpha$, we have the bounds:

$$\omega_y^Q e^{\frac{\sqrt{8}N^2}{1-\alpha}\theta} \geqq Pr(x \to y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N^2}{1-\alpha}\theta} \tag{26}$$

□

Remark 1. *This bound can be rewritten in terms of the log-likelihood of the spontaneous jump and constants*

independent of the initial sequence x as:

$$|\log Pr(x \to y) - C_0| \leqq C_1\theta \tag{27}$$

where the constants are given by:

$$C_0 = \log \omega_y^Q \tag{28}$$

$$C_1 = \frac{\sqrt{8}N^2}{1-\alpha} \tag{29}$$

In-silico Corroboration of Emergenet's Capability To Capture Biologically Meaningful Structure

We compare the results of simulated mutational perturbations to sequences from our databases (for which we have already constructed Emergenets), and then use NCBI BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to identify if our perturbed sequences match with existing sequences in the databases (Supplementary Fig. S-1). We find that in contrast to random variations, which rapidly diverge the trajectories, the Emergenet constraints tend to produce smaller variance in the trajectories, maintain a high degree of match as we extend our trajectories, and produces matches closer in time to the collection time of the initial sequence, suggesting that the Emergenet does indeed capture realistic constraints.

Multivariate Regression to Understand Data Characteristics Necessary For Emergenet Modeling

We investigate the key factors that contribute to modeling a set of strains well within the Emergenet framework. We carry out a multivariate regression with data diversity, the complexity of inferred Emergenet and the edit distance of the WHO recommendation from the dominant strain as independent variables (See Supplementary Table S-6 for definitions). Here we define data diversity as the number of clusters we have in the input set of sequences, such that any two sequences five or less mutations apart are in the same cluster. Emergenet complexity is measured by the number of decision nodes in the component decision trees of the recursive forest.

We select several plausible structures of the regression equation, and in each case conclude that data diversity has the most important and statistically significant contribution (Supplementary Table S-6).

Multivariate Regression to Identify Map from E-distance to Estimated IRAT scores

We train separate General Linear Models (GLM) to estimate IRAT scores (emergence and impact) with average E-distance of a sequence of interest from a set of human strains, considering HA and NA sequences separately, using the CDC computed IRAT scores as the dependent variables. We also include the geometric mean of the HA and NA based E-distances as a potential explanatory variables. We use a standard Gaussian model family with identity link function to keep our model that maps E-distances to the IRAT scores as simple as possible (see Supplementary Table S-4).

Data and Software Sharing

Working open-source software (requiring Python 3.x) is publicly available at https://pypi.org/project/emergenet/. All inferred Emergenet models inferred is available at https://doi.org/10.5281/zenodo.7387861. Accession numbers of all sequences used in this study is provided as supplementary information (seq_metadata.xlsx).

Data Source

We use two public sequence databases: 1) National Center for Biotechnology Information (NCBI) virus[48] and 2) GISAID[49] databases. The former is a community portal for viral sequence data, aiming to increase the usability of data archived in various NCBI repositories. GISAID has a more restricted user agreement, and use of GISAID data in an analysis requires acknowledgment of the contributions of both the submitting and the originating laboratories (Corresponding acknowledgment tables are included as supplementary information). We collected a total of 399, 476 sequences in our analysis (see Supplementary Table S-1).

References

[1] Shao, W., Li, X., Goraya, M. U., Wang, S. & Chen, J.-L. Evolution of influenza a virus by mutation and re-assortment. *International journal of molecular sciences* **18**, 1650 (2017).

[2] Mills, C. E., Robins, J. M. & Lipsitch, M. Transmissibility of 1918 pandemic influenza. *Nature* **432**, 904–906 (2004).

[3] Reid, A. H. & Taubenberger, J. K. The origin of the 1918 pandemic influenza virus: a continuing enigma. *Journal of general virology* **84**, 2285–2292 (2003).

[4] Landolt, G. A. & Olsen, C. W. Up to new tricks–a review of cross-species transmission of influenza a viruses. *Animal Health Research Reviews* **8**, 1–21 (2007).

[5] Dos Santos, G., Neumeier, E. & Bekkat-Berkani, R. Influenza: Can we cope better with the unpredictable? *Human vaccines & immunotherapeutics* **12**, 699–708 (2016).
[6] Huddleston, J. *et al.* Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza a/h3n2 evolution. *Elife* **9**, e60067 (2020).
[7] Boni, M. F. Vaccination and antigenic drift in influenza. *Vaccine* **26**, C8–C14 (2008).
[8] Tricco, A. C. *et al.* Comparing influenza vaccine efficacy against mismatched and matched strains: a systematic review and meta-analysis. *BMC medicine* **11**, 153 (2013).
[9] Neher, R. A., Russell, C. A. & Shraiman, B. I. Predicting evolution from the shape of genealogical trees. *Elife* **3**, e03568 (2014).
[10] Vergara-Alert, J. *et al.* The ns segment of h5n1 avian influenza viruses (aiv) enhances the virulence of an h7n1 aiv in chickens. *Veterinary research* **45**, 1–11 (2014).
[11] Hothorn, T., Hornik, K. & Zeileis, A. Unbiased recursive partitioning: A conditional inference framework. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS* **15**, 651–674 (2006).
[12] Wille, M., Geoghegan, J. L. & Holmes, E. C. How accurately can we assess zoonotic risk? *PLoS biology* **19**, e3001135 (2021).
[13] Pulliam, J. R. & Dushoff, J. Ability to replicate in the cytoplasm predicts zoonotic transmission of livestock viruses. *The Journal of infectious diseases* **199**, 565–568 (2009).
[14] Grewelle, R. E. Larger viral genome size facilitates emergence of zoonotic diseases. *bioRxiv* (2020).
[15] Grange, Z. L. *et al.* Ranking the risk of animal-to-human spillover for newly discovered viruses. *Proceedings of the National Academy of Sciences* **118**, e2002324118 (2021).
[16] CDC. Influenza risk assessment tool (irat) — pandemic influenza (flu) — cdc. https://www.cdc.gov/flu/pandemic-resources/national-strategy/risk-assessment.htm. (Accessed on 07/02/2021).
[17] Gamblin, S. J. & Skehel, J. J. Influenza hemagglutinin and neuraminidase membrane glycoproteins. *Journal of Biological Chemistry* **285**, 28403–28409 (2010).
[18] Posada, D. & Crandall, K. A. Modeltest: testing the model of dna substitution. *Bioinformatics (Oxford, England)* **14**, 817–818 (1998).
[19] Eng, C. L., Tong, J. C. & Tan, T. W. Predicting host tropism of influenza a virus proteins using random forest. *BMC medical genomics* **7**, 1–11 (2014).
[20] Ahlquist, P. Rna-dependent rna polymerases, viruses, and rna silencing. *Science* **296**, 1270–1273 (2002).
[21] Chen, R. & Holmes, E. C. Avian influenza virus exhibits rapid evolutionary dynamics. *Molecular biology and evolution* **23**, 2336–2341 (2006).
[22] Woolthuis, R. G., van Dorp, C. H., Keşmir, C., de Boer, R. J. & van Boven, M. Long-term adaptation of the influenza a virus by escaping cytotoxic t-cell recognition. *Scientific reports* **6**, 1–8 (2016).
[23] Fan, K. *et al.* Role of itk signalling in the interaction between influenza a virus and t-cells. *Journal of general virology* **93**, 987–997 (2012).
[24] van de Sandt, C. E. *et al.* Differential recognition of influenza a viruses by m158–66 epitope-specific cd8+ t cells is determined by extraepitopic amino acid residues. *Journal of virology* **90**, 1009–1022 (2016).
[25] Berkhoff, E., Geelhoed-Mieras, M., Fouchier, R., Osterhaus, A. & Rimmelzwaan, G. Assessment of the extent of variation in influenza a virus cytotoxic t-lymphocyte epitopes by using virus-specific cd8+ t-cell clones. *Journal of General Virology* **88**, 530–535 (2007).
[26] Van de Sandt, C. E., Kreijtz, J. H. & Rimmelzwaan, G. F. Evasion of influenza a viruses from innate and adaptive immune responses. *Viruses* **4**, 1438–1476 (2012).
[27] Wood, J. M. *et al.* Reproducibility of serology assays for pandemic influenza h1n1: collaborative study to evaluate a candidate who international standard. *Vaccine* **30**, 210–217 (2012).
[28] Cover, T. M. & Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, New York, NY, USA, 2006).
[29] Agor, J. K. & Özaltın, O. Y. Models for predicting the evolution of influenza to inform vaccine strain selection. *Human vaccines & immunotherapeutics* **14**, 678–683 (2018).
[30] Carugo, O. & Pongor, S. A normalized root-mean-spuare distance for comparing protein three-dimensional structures. *Protein science* **10**, 1470–1473 (2001).
[31] Righetto, I., Milani, A., Cattoli, G. & Filippini, F. Comparative structural analysis of haemagglutinin proteins from type a influenza viruses: conserved and variable features. *BMC bioinformatics* **15**, 363 (2014).
[32] Ma, W., Kahn, R. E. & Richt, J. A. The pig as a mixing vessel for influenza viruses: human and veterinary implications. *Journal of molecular and genetic medicine: an international journal of biomedical research* **3**, 158 (2009).
[33] Nelson, M. I. & Worobey, M. Origins of the 1918 pandemic: revisiting the swine “mixing vessel” hypothesis. *American journal of epidemiology* **187**, 2498–2502 (2018).
[34] Baumann, J., Kouassi, N. M., Foni, E., Klenk, H.-D. & Matrosovich, M. H1N1 Swine Influenza Viruses Differ from Avian Precursors by a Higher pH Optimum of Membrane Fusion .
[35] Qi, X. *et al.* Probable person to person transmission of novel avian influenza a (h7n9) virus in eastern china, 2013: epidemiological investigation. *Bmj* **347** (2013).
[36] Sun, H. *et al.* Mink is a highly susceptible host species to circulating human and avian influenza viruses. *Emerging microbes & infections* **10**, 472–480 (2021).
[37] Goldberger, A. L. & Peng, C.-K. Genomic classification using an information-based similarity index: application to

the sars coronavirus. *Journal of Computational Biology* **12**, 1103–1116 (2005).
[38] Huelsenbeck, J. P. & Crandall, K. A. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and systematics* **28**, 437–466 (1997).
[39] van der Meer, F. J. U. M., Orsel, K. & Barkema, H. W. The new influenza A H1N1 virus: balancing on the interface of humans and animals. *The Canadian veterinary journal = La revue veterinaire canadienne* **51**, 56–62 (2010).
[40] Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125 (2009).
[41] Mollentze, N., Babayan, S. A. & Streicker, D. G. Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS biology* **19**, e3001390 (2021).
[42] Maher, M. C. *et al.* Predicting the mutational drivers of future sars-cov-2 variants of concern. *Science Translational Medicine* **14**, eabk3445 (2022).
[43] Hernández-Orozco, S., Kiani, N. A. & Zenil, H. Algorithmically probable mutations reproduce aspects of evolution, such as convergence rate, genetic memory and modularity. *Royal Society open science* **5**, 180399 (2018).
[44] Vanmarcke, E. *Random fields: analysis and synthesis* (World scientific, 2010).
[45] Manning, C. D., Manning, C. D. & Schütze, H. *Foundations of statistical natural language processing* (MIT press, 1999).
[46] Varadhan, S. S. Large deviations. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, 622–639 (World Scientific, 2010).
[47] Fedotov, A. A., Harremoës, P. & Topsoe, F. Refinements of pinsker's inequality. *IEEE Transactions on Information Theory* **49**, 1491–1498 (2003).
[48] Hatcher, E. L. *et al.* Virus variation resource–improved response to emergent viral outbreaks. *Nucleic acids research* **45**, D482–D490 (2017).
[49] Bogner, P., Capua, I., Lipman, D. J. & Cox, N. J. A global initiative on sharing avian flu data. *Nature* **442**, 981–981 (2006).
[50] Tzarum, N. *et al.* Structure and receptor binding of the hemagglutinin from a human h6n1 influenza virus. *Cell host & microbe* **17**, 369–376 (2015).
[51] Lazniewski, M., Dawson, W. K., Szczepińska, T. & Plewczynski, D. The structural variability of the influenza a hemagglutinin receptor-binding site. *Briefings in functional genomics* **17**, 415–427 (2018).
[52] Garcia, N. K., Guttman, M., Ebner, J. L. & Lee, K. K. Dynamic changes during acid-induced activation of influenza hemagglutinin. *Structure* **23**, 665–676 (2015).

Extended Data Table 1
Out-performance of Emergenet recommendations over WHO for Influenza A vaccine composition

			Two decades			One decade			2015-2019		
Sub-type	Gene	Hemi-sphere	WHO Error	Emergenet Error	Improve-ment (%)	WHO Error	Emergenet Error	Improve-ment (%)	WHO Error	Emergenet Error	Improve-ment (%)
H1N1	HA	North	12.67	8.76	30.83	4.38	1.19	72.83	2.52	0.33	86.79
H1N1	HA	South	13.57	9.00	33.68	4.67	1.62	65.31	2.52	0.62	75.47
H1N1	HA	Average	13.12	8.88	32.25	4.53	1.40	69.07	2.52	0.48	81.13
H3N2	HA	North	7.65	4.71	38.46	5.00	2.94	41.18	1.82	0.88	51.61
H3N2	HA	South	7.59	4.82	36.43	4.94	3.00	39.29	1.82	0.94	48.39
H3N2	HA	Average	7.62	4.77	37.44	4.97	2.97	40.24	1.82	0.91	50.00
H1N1	NA	North	8.29	6.90	16.67	2.62	1.10	58.18	2.10	0.48	77.27
H1N1	NA	South	9.14	8.38	8.33	3.00	1.43	52.38	2.10	0.76	63.64
H1N1	NA	Average	8.72	7.64	12.50	2.81	1.27	55.28	2.10	0.62	70.46
H3N2	NA	North	4.21	3.63	13.75	2.11	1.79	15.00	1.32	0.32	76.00
H3N2	NA	South	4.68	4.16	11.24	2.58	2.05	20.41	1.32	0.42	68.00
H3N2	NA	Average	4.44	3.90	12.50	2.34	1.92	17.70	1.32	0.37	72.00

Extended Data Table 2
H1N1 HA Northern Hemisphere

Year	WHO Recommendation	Dominant Strain	Emergenet Recommendation	WHO Error	Emergenet Error
2001-02	A/New Caledonia/20/99	A/Canterbury/41/2001	A/Dunedin/2/2000	4	6
2002-03	A/New Caledonia/20/99	A/Taiwan/567/2002	A/New York/241/2001	3	1
2003-04	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	5	2
2004-05	A/New Caledonia/20/99	A/Thailand/Siriraj-Rama-TT/2004	A/New York/222/2003	7	4
2005-06	A/New Caledonia/20/99	A/Niedersachsen/217/2005	A/Canterbury/106/2004	8	10
2006-07	A/New Caledonia/20/99	A/India/34980/2006	A/Auckland/619/2005	6	1
2007-08	A/Solomon Islands/3/2006	A/Norway/1701/2007	A/New York/8/2006	8	11
2008-09	A/Brisbane/59/2007	A/Pennsylvania/02/2008	A/Kentucky/UR06-0476/2007	2	2
2009-10	A/Brisbane/59/2007	A/Singapore/ON1060/2009	A/Hong Kong/549/2008	119	119
2010-11	A/California/7/2009	A/England/01220740/2010	A/New York/14/2009	5	1
2011-12	A/California/7/2009	A/Punjab/041/2011	A/Kansas/01/2010	7	2
2012-13	A/California/7/2009	A/British Columbia/001/2012	A/Moscow/WRAIR4308T/2011	11	4
2013-14	A/California/7/2009	A/Moscow/CRIE-32/2013	A/Helsinki/1199/2012	10	2
2014-15	A/California/7/2009	A/Thailand/CU-C5169/2014	A/Maryland/02/2013	12	0
2015-16	A/California/7/2009	A/Georgia/15/2015	A/Utah/3691/2014	14	2
2016-17	A/California/7/2009	A/Hawaii/21/2016	A/Adana/08/2015	16	0
2017-18	A/Michigan/45/2015	A/Michigan/291/2017	A/Beijing-Huairou/SWL1335/2016	5	4
2018-19	A/Michigan/45/2015	A/Washington/55/2018	A/India/C1721549/2017	6	1
2019-20	A/Brisbane/02/2018	A/Kentucky/06/2019	A/New Jersey/01/2018	5	1
2020-21	A/Hawaii/70/2019	A/Togo/905/2020	A/Italy/8949/2019	4	8
2021-22	A/Victoria/2570/2019	A/Ireland/20935/2022	A/Togo/45/2021	9	3
2022-23	-1	-1	A/Netherlands/00068/2022	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

Extended Data Table 3
H1N1 HA Southern Hemisphere

Year	WHO Recommendation	Dominant Strain	Emergenet Recommendation	WHO Error	Emergenet Error
2001-02	A/New Caledonia/20/99	A/Canterbury/41/2001	A/South Canterbury/50/2000	4	6
2002-03	A/New Caledonia/20/99	A/Taiwan/567/2002	A/Canterbury/41/2001	3	1
2003-04	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	5	2
2004-05	A/New Caledonia/20/99	A/Thailand/Siriraj-Rama-TT/2004	A/Memphis/5/2003	7	4
2005-06	A/New Caledonia/20/99	A/Niedersachsen/217/2005	A/Canterbury/106/2004	8	10
2006-07	A/New Caledonia/20/99	A/India/34980/2006	A/Niedersachsen/217/2005	6	2
2007-08	A/New Caledonia/20/99	A/Norway/1701/2007	A/Thailand/CU68/2006	14	6
2008-09	A/Solomon Islands/3/2006	A/Pennsylvania/02/2008	A/Kentucky/UR06-0476/2007	9	2
2009-10	A/Brisbane/59/2007	A/Singapore/ON1060/2009	A/Belem/241/2008	119	119
2010-11	A/California/7/2009	A/England/01220740/2010	A/Singapore/ON1060/2009	5	1
2011-12	A/California/7/2009	A/Punjab/041/2011	A/England/01220740/2010	7	2
2012-13	A/California/7/2009	A/British Columbia/001/2012	A/Punjab/041/2011	11	4
2013-14	A/California/7/2009	A/Moscow/CRIE-32/2013	A/India/P122045/2012	10	5
2014-15	A/California/7/2009	A/Thailand/CU-C5169/2014	A/Jiangsuhailing/SWL1382/2013	12	4
2015-16	A/California/7/2009	A/Georgia/15/2015	A/Thailand/CU-C5169/2014	14	2
2016-17	A/California/7/2009	A/Hawaii/21/2016	A/Georgia/15/2015	16	2
2017-18	A/Michigan/45/2015	A/Michigan/291/2017	A/Beijing-Huairou/SWL1335/2016	5	4
2018-19	A/Michigan/45/2015	A/Washington/55/2018	A/Michigan/291/2017	6	1
2019-20	A/Michigan/45/2015	A/Kentucky/06/2019	A/Washington/55/2018	7	1
2020-21	A/Brisbane/02/2018	A/Togo/905/2020	A/Italy/8451/2019	10	8
2021-22	A/Victoria/2570/2019	A/Abidjan/457/2021	A/Togo/0298/2021	9	5
2022-23	-1	-1	A/Cote_D'Ivoire/1270/2021	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

Extended Data Table 4
H3N2 HA Northern Hemisphere

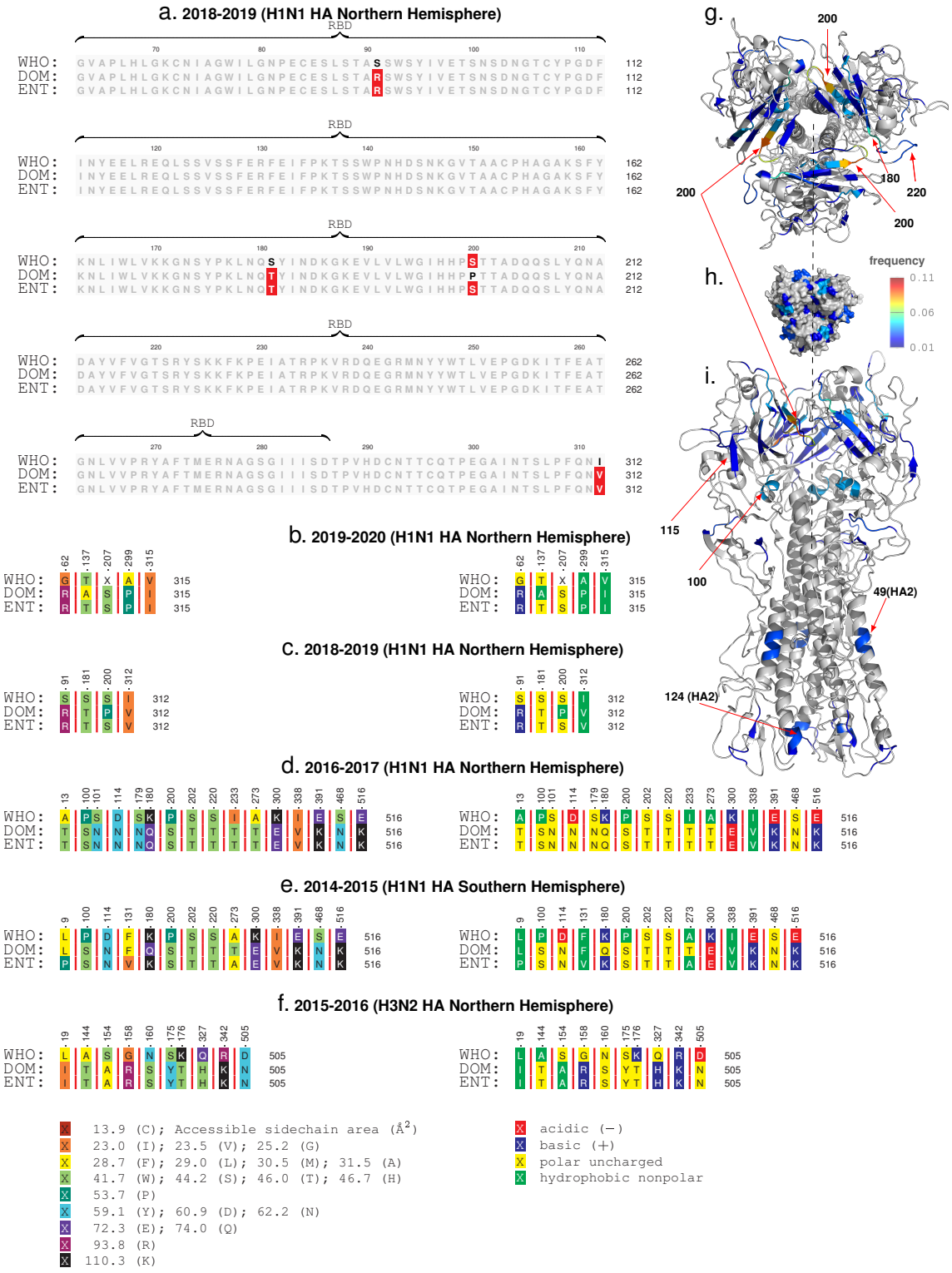
Year	WHO Recommendation	Dominant Strain	Emergenet Recommendation	WHO Error	Emergenet Error
2005-06	A/California/7/2004	A/Denmark/195/2005	A/Tairawhiti/369/2004	10	2
2006-07	A/Wisconsin/67/2005	A/New York/5/2006	A/South Australia/22/2005	5	4
2007-08	A/Wisconsin/67/2005	A/Tennessee/11/2007	A/Colorado/05/2006	8	5
2008-09	A/Brisbane/10/2007	A/Massachusetts/13/2008	A/Virginia/UR06-0021/2007	3	2
2009-10	A/Brisbane/10/2007	A/Hawaii/14/2009	A/Manhean/03/2008	7	6
2010-11	A/Perth/16/2009	A/Utah/12/2010	A/Philippines/5/2009	8	7
2011-12	A/Perth/16/2009	A/Piaui/14202/2011	A/Singapore/C2010.310/2010	4	4
2012-13	A/Victoria/361/2011	A/Alborz/927/2012	A/Tehran/895/2012	4	3
2013-14	A/Victoria/361/2011	A/Delaware/01/2013	A/Singapore/H2012.934/2012	4	1
2014-15	A/Texas/50/2012	A/Alborz/72205/2014	A/Nebraska/03/2013	10	9
2015-16	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Ontario/01/2014	10	0
2016-17	A/Hong Kong/4801/2014	A/Guangdong/12/2016	A/Oregon/02/2015	0	0
2017-18	A/Hong Kong/4801/2014	A/Maryland/25/2017	A/New York/03/2016	3	1
2018-19	A/Singapore/INFIMH-16-0019/2016	A/Vermont/04/2018	A/Ontario/038/2017	8	5
2019-20	A/Kansas/14/2017	A/Kentucky/27/2019	A/California/7330/2018	16	12
2020-21	A/Hong Kong/2671/2019	A/India/Pun-NIV289524/2021_Jan	A/California/NHRC-OID_FDX100215/2019	16	14
2021-22	A/Cambodia/e0826360/2020	A/Human/New_York/PV60641/2022	A/India/Pun-NIV291000/2021_Jan	14	5
2022-23	-1	-1	A/Ireland/14993/2022	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

Extended Data Table 5
H3N2 HA Southern Hemisphere

Year	WHO Recommendation	Dominant Strain	Emergenet Recommendation	WHO Error	Emergenet Error
2005-06	A/Wellington/1/2004	A/Denmark/195/2005	A/Waikato/21/2004	3	3
2006-07	A/California/7/2004	A/New York/5/2006	A/South Australia/22/2005	12	4
2007-08	A/Wisconsin/67/2005	A/Tennessee/11/2007	A/New York/923/2006	8	5
2008-09	A/Brisbane/10/2007	A/Massachusetts/13/2008	A/Tennessee/11/2007	3	2
2009-10	A/Brisbane/10/2007	A/Hawaii/14/2009	A/Manhean/03/2008	7	6
2010-11	A/Perth/16/2009	A/Utah/12/2010	A/Hawaii/14/2009	8	7
2011-12	A/Perth/16/2009	A/Piaui/14202/2011	A/Utah/12/2010	4	4
2012-13	A/Perth/16/2009	A/Alborz/927/2012	A/Piaui/14202/2011	8	4
2013-14	A/Victoria/361/2011	A/Delaware/01/2013	A/Callao/IPE00830/2012	4	7
2014-15	A/Texas/50/2012	A/Alborz/72205/2014	A/Delaware/01/2013	10	7
2015-16	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Alborz/72205/2014	10	0
2016-17	A/Hong Kong/4801/2014	A/Guangdong/12/2016	A/Parma/471/2015	0	0
2017-18	A/Hong Kong/4801/2014	A/Maryland/25/2017	A/Ontario/196/2016	3	4
2018-19	A/Singapore/INFIMH-16-0019/2016	A/Vermont/04/2018	A/Texas/279/2017	8	5
2019-20	A/Switzerland/8060/2017	A/Kentucky/27/2019	A/Santa Catarina/1200/2018	13	12
2020-21	A/South Australia/34/2019	A/India/Pun-NIV289524/2021_Jan	A/Kentucky/27/2019	12	14
2021-22	A/Hong Kong/2671/2019	A/Darwin/9a/2021	A/India/PUN-NIV301718/2021	19	1
2022-23	-1	-1	A/Latvia/04-86261/2022	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric



Extended Data Figure 1. **Sequence comparisons.** Comparing the Emergenet (ENT) and the WHO recommendation (WHO), and the observed dominant strain (DOM), we note that the correct Emergenet predictions tend to be within the RBD, both for H1N1 and H3N2 for HA (panel a shows one example). Additionally, by comparing the type, side chain area, and the accessible side chain area, we note that DOM and ENT are often close in important chemical properties, while WHO deviations are not (panel b-f). Panels g-i show the localization of the deviations in the molecular structure of HA, where we note that the changes are most frequent in the HA1 sub-unit (the globular head), and around residues and structures that have been commonly implicated in receptor binding interactions *e.g* the $\approx$ 200 loop, the $\approx$ 220 loop and the $\approx$ 180-helix[50–52].

Extended Data Table 6
Influenza A Strains Evaluated by IRAT and Corresponding Emergenet Computed Risk Scores

Influenza Virus	Subtype	IRAT Date	IRAT Emergence Score	IRAT Impact Score	HA Sample	NA Sample	HA E-risk	NA E-risk	Geom. Mean	Emergenet Emergence Score	Emergenet Impact Score
A/swine/Shandong/1207/2016	H1N1	Jul 2020	7.5	6.9	1000	1000	-0.0941	-0.0205	0.0440	6.0	6.2
A/Ohio/13/2017	H3N2	Jul 2019	6.6	5.8	1000	1000	-0.0184	-0.0306	0.0238	6.3	6.2
A/Hong Kong/125/2017	H7N9	May 2017	6.5	7.5	437	437	-0.0296	-0.0058	0.0131	6.6	6.5
A/Shanghai/02/2013	H7N9	Apr 2016	6.4	7.2	178	178	-0.0055	-0.0036	0.0044	6.7	6.6
A/Anhui-Lujiang/39/2018	H9N2	Jul 2019	6.2	5.9	31	30	-0.0290	-0.1681	0.0698	5.2	5.0
A/Indiana/08/2011	H3N2	Dec 2012	6.0	4.5	1000	1000	-0.0523	-0.0091	0.0218	6.4	6.5
A/California/62/2018	H1N2	Jul 2019	5.8	5.7	55	55	-0.1089	-0.0610	0.0815	5.4	5.5
A/Bangladesh/0994/2011***	H9N2	Feb 2014	5.6	5.4			-0.2078	-0.1823	0.1947	4.3	4.9
A/Sichuan/06681/2021	H5N6	Oct 2021	5.3	6.3	45	45	-0.3616	-0.0518	0.1369	5.2	6.4
A/Vietnam/1203/2004	H5N1	Nov 2011	5.2	6.6	258	246	-0.1673	-0.0111	0.0430	6.2	6.7
A/Yunnan/14564/2015**	H5N6	Apr 2016	5.0	6.6	344	331	-0.3482	-0.2987	0.3225	4.9	6.5
A/Astrakhan/3212/2020**	H5N8	Mar 2021	4.6	5.2	381	365	-0.1603	-0.3472	0.2359	3.9	4.4
A/Netherlands/219/2003	H7N7	Jun 2012	4.6	5.8	46	46	-0.2757	-0.3521	0.3115	4.6	5.8
A/American wigeon/South Carolina/AH0195145/2021	H5N1	Mar 2022	4.4	5.1	335	323	-0.1722	-0.5114	0.2967	4.0	4.7
A/Jiangxi-Donghu/346/2013***	H10N8	Feb 2014	4.3	6.0			-0.2088	-0.2101	0.2094	4.3	4.8
A/gyrfalcon/Washington/41088/2014**	H5N8	Mar 2015	4.2	4.6	341	328	-0.1532	-0.3424	0.2290	3.9	4.3
A/Northern pintail/Washington/40964/2014**	H5N2	Mar 2015	3.8	4.1	341	328	-0.1529	-0.3799	0.2410	3.9	4.3
A/canine/Illinois/12191/2015	H3N2	Jun 2016	3.7	3.7	1000	1000	-0.0607	-0.1509	0.0957	4.9	4.8
A/American green-winged teal /Washington/1957050/2014	H5N1	Mar 2015	3.6	4.1	326	314	-0.1911	-0.4482	0.2927	4.1	4.9
A/turkey/Indiana/1573-2/2016**	H7N8	Jul 2017	3.4	3.9	495	494	-0.1130	-0.7738	0.2957	3.4	3.9
A/chicken/Tennessee/17-007431-3/2017	H7N9	Oct 2017	3.1	3.5	496	495	-0.1027	-0.2569	0.1624	4.1	4.2
A/chicken/Tennessee/17-007147-2/2017	H7N9	Oct 2017	2.8	3.5	496	495	-0.2095	-0.2541	0.2307	4.2	4.8

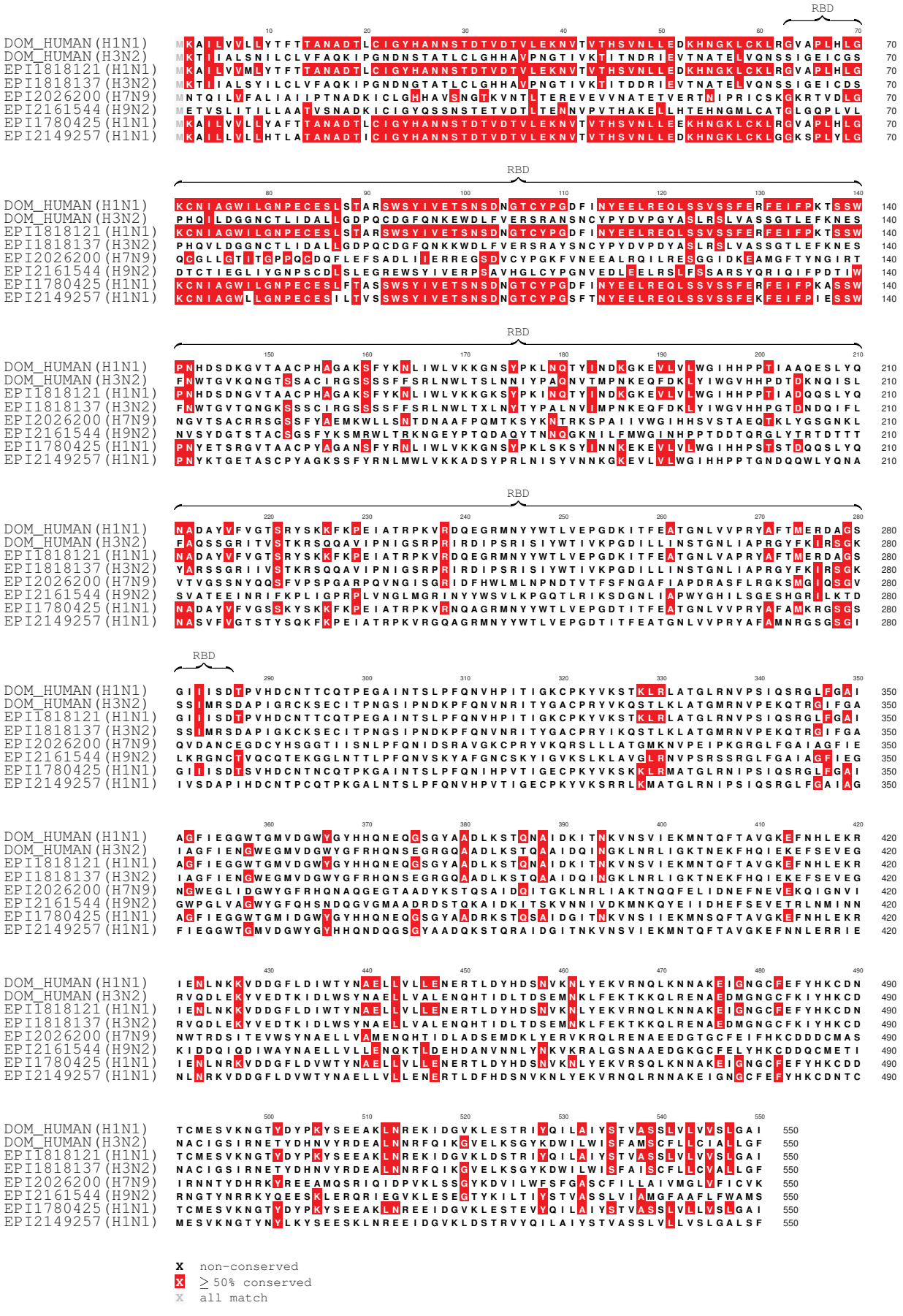
** Emergenet constructed using all human strains that match the HA sub-type, $e.g.$, H5Nx for H5N6.
*** distance estaimated averaging over those obtained by considering all Emergenets from other subtypes.

Extended Data Table 7
Count of identified strains above estimated emergence risk threshold

subtype	No. with minimum estimated IRAT emergence score				
	6.0	6.2	6.3	6.4	6.5
H1N1	62 (83%)	57 (81%)	53 (82%)	5 (50%)	4 (67%)
H3N2	11 (15%)	11 (16%)	11 (17%)	4 (40%)	1 (17%)
H7N9	1 (1%)	1 (1%)	1 (1%)	1 (10%)	1 (17%)
H9N2	1 (1%)	1 (1%)	0	0	0

Extended Data Table 8
Influenza A Strains Evaluated by IRAT and Corresponding Emergenet Computed Risk Scores

strain	sub-type	HA accession	NA accession	predicted IRAT impact	predicted IRAT emergence
A/swine/Missouri/A02524711/2020	H1N1	EPI1818121	EPI1818122	6.6318	6.7067
A/Camel/Inner_Mongolia/XL/2020	H7N9	EPI2026200	EPI2026202	6.6286	6.7026
A/swine/Indiana/A02524710/2020	H3N2	EPI1818137	EPI1818138	6.6070	6.6757
A/swine/North_Carolina/A02479173/2020	H1N1	EPI1780425	EPI1780426	6.5878	6.6517
A/swine/Tennessee/A02524414/2022	H1N1	EPI2149257	EPI2149258	6.5379	6.5893
A/swine/Minnesota/A02635976/2021	H1N1	EPI1912208	EPI1912209	6.4841	6.5220
A/swine/Chile/VN1401-5054/2020	H3N2	EPI1974975	EPI1974978	6.4578	6.4891
A/swine/Italy/56910/2020	H3N2	EPI2142217	EPI2142173	6.4537	6.4840
A/swine/Minnesota/A02245643/2020	H3N2	EPI1769178	EPI1769179	6.4370	6.4631
A/swine/Iowa/A02479005/2020	H1N1	EPI1777621	EPI1777622	6.4103	6.4297
A/swine/Iowa/A02524874/2020	H3N2	EPI1907838	EPI1907839	6.3836	6.3962
A/swine/Indiana/A02636638/2022	H1N1	EPI2153370	EPI2153371	6.3830	6.3954
A/swine/Illinois/A02479007/2020	H3N2	EPI1777629	EPI1777630	6.3769	6.3877
A/swine/Spain/44579-1/2020	H3N2	EPI1930744	EPI1930748	6.3737	6.3838
A/swine/Minnesota/A02248037/2021	H1N1	EPI1912188	EPI1912189	6.3694	6.3784
A/swine/Minnesota/A02245699/2020	H3N2	EPI1833007	EPI1833008	6.3692	6.3781
A/swine/Iowa/A02635917/2021	H1N1	EPI1911753	EPI1911754	6.3687	6.3774
A/swine/Minnesota/A02711801/2022	H1N1	EPI2153420	EPI2153421	6.3674	6.3759
A/swine/Illinois/A02635936/2021	H1N1	EPI1911791	EPI1911792	6.3671	6.3754
A/swine/South_Dakota/A02524453/2020	H1N1	EPI1765555	EPI1765556	6.3658	6.3738
A/swine/Minnesota/A02248061/2021	H1N1	EPI1912494	EPI1912495	6.3624	6.3696
A/swine/Iowa/A02636439/2022	H1N1	EPI2147475	EPI2147476	6.3616	6.3686
A/swine/Iowa/A02524875/2020	H1N1	EPI1907858	EPI1907859	6.3610	6.3678
A/swine/Minnesota/A02248060/2021	H1N1	EPI1912500	EPI1912501	6.3592	6.3656
A/swine/Nebraska/A02636117/2021	H1N1	EPI1932937	EPI1932938	6.3582	6.3644
A/swine/Iowa/A02524513/2020	H1N1	EPI1832647	EPI1832648	6.3580	6.3641
A/swine/Iowa/A02524724/2020	H1N1	EPI1818387	EPI1818388	6.3579	6.3640
A/swine/Iowa/A02635719/2021	H1N1	EPI1910907	EPI1910908	6.3579	6.3640
A/swine/Nebraska/A02479337/2020	H1N1	EPI1769116	EPI1769117	6.3579	6.3640
A/swine/Iowa/A02479383/2020	H1N1	EPI1771027	EPI1771028	6.3579	6.3640
A/swine/Nebraska/A02479212/2020	H1N1	EPI1775884	EPI1775885	6.3579	6.3640
A/swine/Minnesota/A02245424/2020	H1N1	EPI1780207	EPI1780208	6.3579	6.3640
A/swine/Minnesota/A02479051/2020	H1N1	EPI1778572	EPI1778573	6.3579	6.3640
A/swine/Missouri/A02525065/2021	H1N1	EPI1908581	EPI1908582	6.3579	6.3640
A/swine/Missouri/A02524951/2020	H1N1	EPI1908429	EPI1908430	6.3579	6.3640
A/swine/Iowa/A02524892/2020	H1N1	EPI1907881	EPI1907882	6.3579	6.3640
A/swine/Nebraska/A02524954/2020	H1N1	EPI1908393	EPI1908394	6.3579	6.3640
A/swine/Iowa/A02524994/2020	H1N1	EPI1908427	EPI1908428	6.3579	6.3640
A/swine/Iowa/A02525313/2021	H1N1	EPI1910761	EPI1910762	6.3579	6.3640
A/swine/Iowa/A02524646/2020	H1N1	EPI1817164	EPI1817165	6.3579	6.3640
A/swine/Nebraska/A02479186/2020	H1N1	EPI1774141	EPI1774142	6.3578	6.3638
A/swine/Iowa/A02479156/2020	H1N1	EPI1780249	EPI1780250	6.3574	6.3634
A/swine/Iowa/A02479229/2020	H1N1	EPI1775914	EPI1775915	6.3574	6.3633
A/swine/Iowa/A02479303/2020	H1N1	EPI1768639	EPI1768640	6.3567	6.3625
A/swine/Minnesota/A02710691/2021	H1N1	EPI2146090	EPI2146091	6.3561	6.3618
A/swine/Iowa/A02635881/2021	H1N1	EPI1911668	EPI1911669	6.3560	6.3616
A/swine/Iowa/A02525354/2021	H1N1	EPI1910789	EPI1910790	6.3560	6.3616
A/swine/Iowa/A02524739/2020	H1N1	EPI1818383	EPI1818384	6.3543	6.3595
A/swine/Iowa/A02635823/2021	H1N1	EPI1911263	EPI1911264	6.3543	6.3594
A/swine/Minnesota/A02711797/2022	H3N2	EPI2153382	EPI2153383	6.3537	6.3587
A/swine/Iowa/A02479141/2020	H1N1	EPI1780241	EPI1780242	6.3531	6.3579
A/swine/Iowa/A02635955/2021	H1N1	EPI1912240	EPI1912241	6.3516	6.3561
A/swine/Iowa/A02750621/2022	H1N1	EPI2161576	EPI2161577	6.3509	6.3552
A/swine/Illinois/A02525253/2021	H3N2	EPI1910375	EPI1910376	6.3506	6.3548
A/swine/Iowa/A02245587/2020	H1N1	EPI1775817	EPI1775818	6.3497	6.3537
A/swine/Iowa/A02636145/2021	H1N1	EPI1932055	EPI1932930	6.3487	6.3525
A/swine/Iowa/A02636114/2021	H1N1	EPI1931853	EPI1931854	6.3487	6.3525
A/swine/Iowa/A02525217/2021	H1N1	EPI1909087	EPI1909088	6.3487	6.3525
A/swine/Iowa/A02636496/2022	H1N1	EPI2148086	EPI2148087	6.3485	6.3522
A/swine/Iowa/A02635871/2021	H1N1	EPI1911656	EPI1911657	6.3473	6.3507
A/swine/Iowa/A02479067/2020	H1N1	EPI1778734	EPI1778735	6.3469	6.3501
A/swine/Minnesota/A02246459/2021	H1N1	EPI1912518	EPI1912519	6.3465	6.3496
A/swine/Minnesota/A02525348/2021	H1N1	EPI1910795	EPI1910796	6.3400	6.3415
A/swine/Iowa/A02479343/2020	H1N1	EPI1769114	EPI1769115	6.3387	6.3399
A/canine/Texas/21-011409-001/2021	H3N2	EPI1896555	EPI1896557	6.3093	6.3030
A/swine/Kansas/A02245381/2020(H1N1)	H1N1	EPI1777723	EPI1777724	6.2958	6.2860
A/swine/Iowa/A02246996/2021	H1N1	EPI2146133	EPI2146134	6.2817	6.2684
A/mink/China/chick_embryo/2020	H9N2	EPI2161544	EPI2161548	6.2787	6.2646



Extended Data Figure 2. HA sequence comparison with dominant human strains (DOM_HUMAN H1N1, H3N2) with Emergenet estimated top 5 risky strains (2020-2022 April) along with the teh most risky H9N2 strain (A/mink/China/chick embryo/2020), showing substantial differences from the circulating strains both in and out of the RBD.

SUPPLEMENTARY FIGURES & TABLES

S-Tab. 1
Number of Influenza sequences collected from public databases

Database	Subtype	No. HA Sequences	No. NA Sequences	Total
GISAID	H1N1	1,3536	13,501	27,037
GISAID	H1N2	857	857	1,714
GISAID	H3N2	40,257	40,096	8,0353
GISAID	H5N1	1,970	1,943	3,913
GISAID	H5N2	22	24	46
GISAID	H5N6	186	186	372
GISAID	H5N8	1,449	1401	2,850
GISAID	H7N1	3	3	6
GISAID	H7N2	2	2	4
GISAID	H7N3	101	99	200
GISAID	H7N5	1	1	2
GISAID	H7N6	8	8	16
GISAID	H7N7	57	56	113
GISAID	H7N8	4	4	8
GISAID	H7N9	1265	1264	2,529
GISAID	H9N2	312	312	624
GISAID	H10N8	1	1	2
NCBI	H1N1	17,902	16,645	34,547
NCBI	H3N2	18,257	14,691	32,948
NCBI	H5N1	1	1	2
NCBI	H7N7	1	1	2
NCBI	H7N9	2	2	4
	Total	96,195	91,099	187,294

S-Tab. 2
Examples: Emergenet induced distance varying for fixed sequence pair when background population changes (rows 1 -5), sequences with small edit distance and large E-distance, and the converse (rows 6-9)

	Edit dist.	Sequence A	Sequence B	Emer-genet E-dist.	Year A*	Year B*
1	18	A/Singapore/23J/2007	A/Tennessee/UR06-0294/2007	0.0111	2007	2007
2	18	A/Singapore/23J/2007	A/Tennessee/UR06-0294/2007	0.0094	2008	2008
3	18	A/Singapore/23J/2007	A/Tennessee/UR06-0294/2007	0.0027	2009	2009
4	18	A/Singapore/23J/2007	A/Tennessee/UR06-0294/2007	0.0025	2010	2010
5	18	A/Singapore/23J/2007	A/Tennessee/UR06-0294/2007	0.6163	2007	2010
6	11	A/Naypyitaw/M783/2008	A/Singapore/201/2008	0.8852	2008	2008
7	15	A/Cambodia/W0908339/2012	A/Singapore/DMS1233/2012	0.2737	2012	2012
8	126	A/South Dakota/03/2008	A/Singapore/10/2008	0.3034	2008	2008
9	141	A/Jodhpur/3248/2012	A/Cambodia/W0908339/2012	0.2405	2012	2012

*Year A and year B correspond to the assumed collection years for sequences A and B respectively for the purpose of this example. Sequence A in row 1 is collected in 2007, but is assumed to be from different years in rows 2-4 to demonstrate the change in E-distance from sequence B, arising only from a change in the background population.

S-Tab. 3
Correlation between E-distance and edit distance between sequence pairs

Phenotypes	Correlation
Influenza H1N1 HA	0.76
Influenza H1N1 NA	0.74
Influenza H3N2 HA	0.85
Influenza H3N2 NA	0.79

S-Tab. 4
General linear model evaluating Emergenet emergence risk predictions against IRAT estimates

Model: IRAT_Emergence_Score ~ Geometric_Mean

Dep. Variable:	IRAT_Emergence_Score	No. Observations:	20
Model:	GLM	Df Residuals:	18
Model Family:	Gaussian	Df Model:	1
Link Function:	identity	Scale:	0.76392
Method:	IRLS	Log-Likelihood:	-24.632
Date:	Fri, 18 Nov 2022	Deviance:	13.751
Time:	20:17:29	Pearson chi2:	13.8
No. Iterations:	3	Pseudo R-squ. (CS):	0.7407
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	6.4084	0.342	18.711	0.000	5.737	7.080
Geometric_Mean	-9.9819	1.925	-5.185	0.000	-13.755	-6.209

Model: IRAT_Emergence_Score ~ Geometric_Mean + HA_E-risk:NA_E-risk

Dep. Variable:	IRAT_Emergence_Score	No. Observations:	20
Model:	GLM	Df Residuals:	17
Model Family:	Gaussian	Df Model:	2
Link Function:	identity	Scale:	0.74617
Method:	IRLS	Log-Likelihood:	-23.826
Date:	Fri, 18 Nov 2022	Deviance:	12.685
Time:	20:18:29	Pearson chi2:	12.7
No. Iterations:	3	Pseudo R-squ. (CS):	0.7678
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	6.7809	0.460	14.736	0.000	5.879	7.683
Geometric_Mean	-19.3647	8.078	-2.397	0.017	-35.198	-3.531
HA_E-risk:NA_E-risk	31.5402	26.392	1.195	0.232	-20.187	83.267

S-Tab. 5
General linear model evaluating Emergenet impact risk predictions against IRAT estimates

Model: IRAT_Impact_Score ~ Geometric_Mean

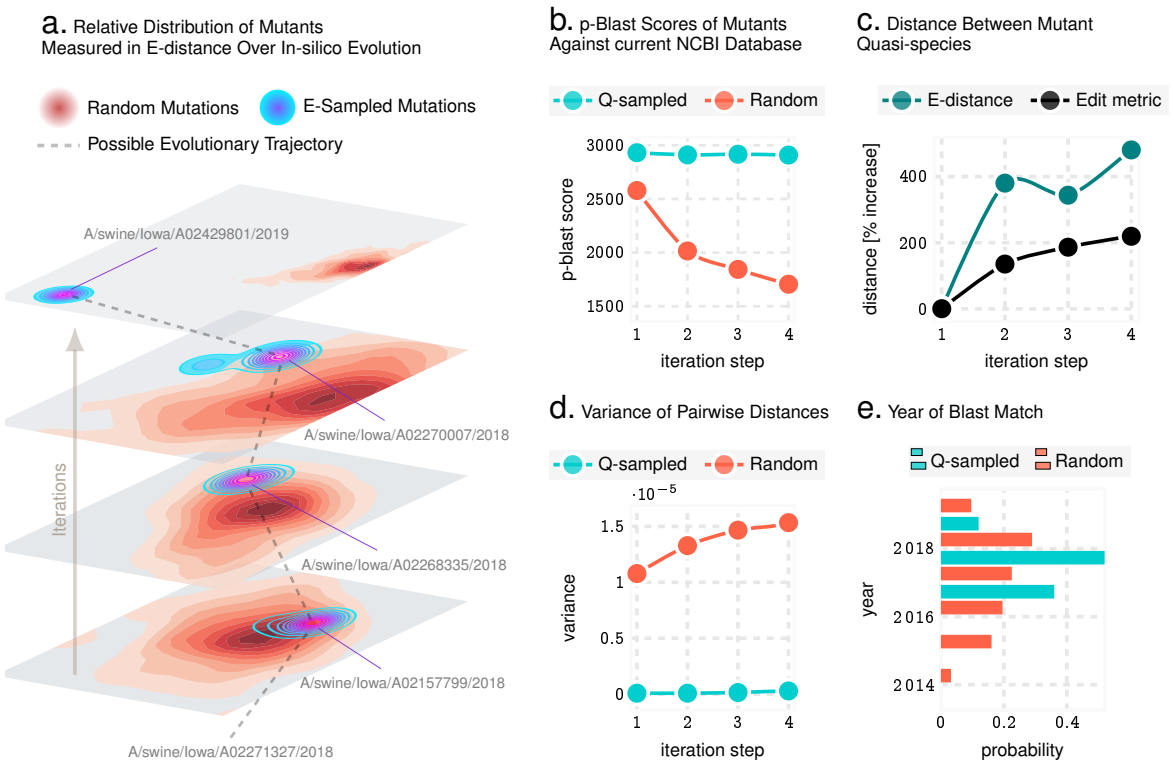
Dep. Variable:	IRAT_Impact_Score	No. Observations:	20
Model:	GLM	Df Residuals:	18
Model Family:	Gaussian	Df Model:	1
Link Function:	identity	Scale:	1.0201
Method:	IRLS	Log-Likelihood:	-27.525
Date:	Fri, 18 Nov 2022	Deviance:	18.362
Time:	20:20:03	Pearson chi2:	18.4
No. Iterations:	3	Pseudo R-squ. (CS):	0.4382
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	6.3736	0.396	16.104	0.000	5.598	7.149
Geometric_Mean	-7.5193	2.225	-3.380	0.001	-11.879	-3.159

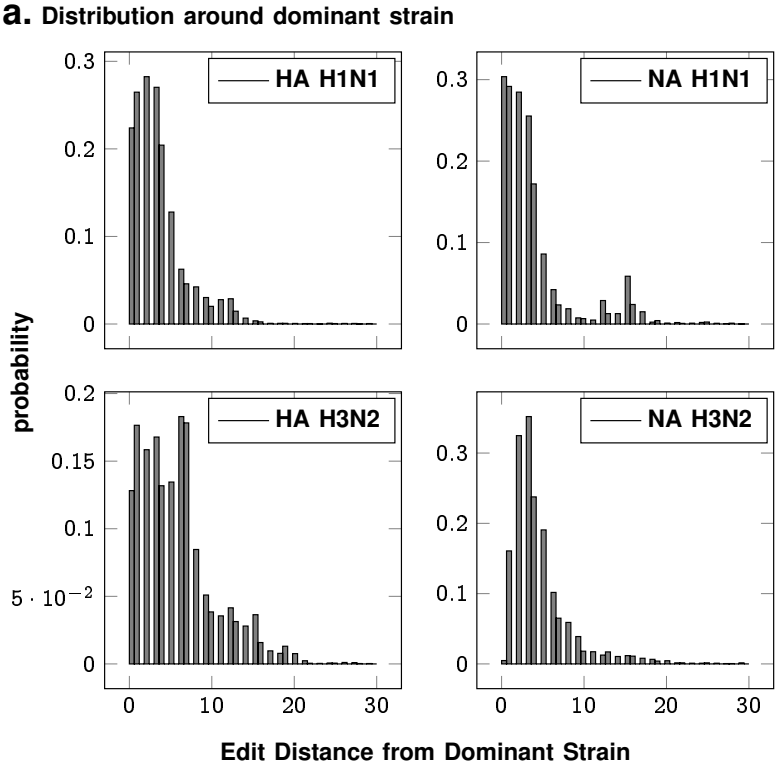
Model: IRAT_Impact_Score ~ Geometric_Mean + HA_E-risk:NA_E-risk

Dep. Variable:	IRAT_Impact_Score	No. Observations:	20
Model:	GLM	Df Residuals:	17
Model Family:	Gaussian	Df Model:	2
Link Function:	identity	Scale:	1.0345
Method:	IRLS	Log-Likelihood:	-27.093
Date:	Fri, 18 Nov 2022	Deviance:	17.587
Time:	20:20:50	Pearson chi2:	17.6
No. Iterations:	3	Pseudo R-squ. (CS):	0.4584
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	6.6913	0.542	12.350	0.000	5.629	7.753
Geometric_Mean	-15.5210	9.512	-1.632	0.103	-34.165	3.123
HA_E-risk:NA_E-risk	26.8979	31.076	0.866	0.387	-34.010	87.806



S-Fig. 1. **E-distance validation in-silico using Influenza A sequences from NCBI database. Panel a** illustrates that the Emergenet induced modeling of evolutionary trajectories initiated from known haemagglutinnin (HA) sequences are distinct from random paths in the strain space. In particular, random trajectories have more variance, and more importantly, diverge to different regions of the landscape compared to Emergenet predictions. **Panels b-e** show that unconstrained Q-sampling produces sequences maintain a higher degree of similarity to known sequences, as verified by blasting against known HA sequences, have a smaller rate of growth of variance, and produce matches in closer time frames to the initial sequence. **Panel c** shows that this is not due to simply restricting the mutational variations, which increases rapidly in both the Emergenet and the classical metric.



S-Fig. 2. **No. of mutations from the seasonal dominant strain over the years** The quasispecies that circulates each season for each sub-type is tightly distributed around the dominant strain on average.

S-Tab. 6
General linear model for evaluating effect of data diversity on Emergenet performance

Variable Name	Description
enet_complexity	Cumulative number of nodes in all predictors in the corresponding Emergenet
data_diversity	Number of clusters in set of input sequence where each sequence in a specific cluster is separated by at least 5 mutations from sequences not in the cluster
ldistance_WHO	Deviation of WHO predicted strain from the dominant strain

```
model:dev ~ enet_complexity + data_diversity + enet_complexity * data_diversity + ldistance_WHO
             Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                    dev   No. Observations:                  235
Model:                            GLM   Df Residuals:                      230
Model Family:                Gaussian   Df Model:                            4
Link Function:               identity   Scale:                          23.214
Method:                          IRLS   Log-Likelihood:                -700.43
Date:                Thu, 11 Jun 2020   Deviance:                       5339.2
Time:                        16:45:46   Pearson chi2:                 5.34e+03
No. Iterations:                     3   Covariance Type:             nonrobust
=================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Intercept                        -0.1116      1.090     -0.102      0.918      -2.248       2.025
enet_complexity                   0.0005      0.000      1.075      0.282      -0.000       0.001
data_diversity                    0.3197      0.126      2.531      0.011       0.072       0.567
enet_complexity:data_diversity -6.932e-05   5.01e-05     -1.383      0.167      -0.000    2.89e-05
ldistance_WHO                    -0.0348      0.035     -1.007      0.314      -0.102       0.033
=================================================================================================


model:dev ~ enet_complexity + data_diversity + ldistance_WHO
             Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                    dev   No. Observations:                  235
Model:                            GLM   Df Residuals:                      231
Model Family:                Gaussian   Df Model:                            3
Link Function:               identity   Scale:                          23.306
Method:                          IRLS   Log-Likelihood:                -701.41
Date:                Thu, 11 Jun 2020   Deviance:                       5383.6
Time:                        16:45:47   Pearson chi2:                 5.38e+03
No. Iterations:                     3   Covariance Type:             nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.0841      0.665      1.630      0.103      -0.219       2.387
enet_complexity -4.12e-05      0.000     -0.156      0.876      -0.001       0.000
data_diversity      0.1788      0.075      2.392      0.017       0.032       0.325
ldistance_WHO     -0.0695      0.024     -2.930      0.003      -0.116      -0.023
==============================================================================
```