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

Kevin Wu[1], Jin Li[1], Timmy 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
[4]Committee on Genetics, Genomics & Systems Biology, University of 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 A strains emerging into humans are suspected to 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 viable genotypic variations are shaped by evolutionary constraints using only genomic sequences of key viral proteins. Analyzing Hemaglutinnin (HA) and Neuraminidase (NA) sequences from nearly 100,000 unique strains, we estimate the likelihood of specific future mutations, yielding the numerical odds of one strain giving rise to a specific descendant 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 (HA: 34.8% improvement, NA: 12.5% improvement), we assess the pandemic potential of animal strains that do not yet circulate in humans. While the state-of-the-art Influenza Risk Assessment Tool (IRAT) from the CDC comprises multiple time-consuming experimental assays, our calculations take ≈ 6 seconds per strain, while strongly correlating with published IRAT scores (correlation=0.703, p-value $= 0.00026$). This six orders of magnitude speedup (weeks vs seconds) in identifying risky strains is a necessary step to exploit current surveillance capacity via scalably analyzing thousands of strains collected annually. Considering 6,066 wild Influenza A viruses sequenced post 2020, we identify individual strains of diverse subtypes, hosts and geo-locations posing maximal risk, with 6 having estimated emergence scores > 6.5. Such scalable risk-ranking can enable preemptive pandemic mitigation, including 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 that impact optimal pandemic response.**

INTRODUCTION

Influenza viruses constantly evolve[5], altering surface protein structures over a time scale of months 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 inoculating the population with a reformulated vaccine annually[5,7]. Among numerous factors that hinder optimal design of the seasonal 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 A strains spilling over into humans from animal reservoirs have triggered pandemics at least four times (1918 Spanish flu/H1N1, 1957 Asian flu/H2N2, 1968 hongkong flu/H3N2, 2009 swine flu/H1N1) in the past 100 years[1]. With the memory of the sudden emergence of COVID-19 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.

A 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. Despite global surveillance efforts to collect wild specimens from diverse hosts and geo-locations, our current ability to objectively, reliably and scalably evaluate such risk posed to humans by individual strains is limited[11]. Despite recent progress towards understanding emergence risk[12–14], scalable ranking of individual strains remains out of reach.

The Center for Disease Control's (CDC) current solution to preempt strain-specific emergence is the Influenza Risk Assessment Tool (IRAT)[15], which reflects evaluationsby subject matter experts (SME) from the CDC, the Food and Drug Administration (FDA), the Animal and Plant Health Inspection Service (APHIS), and the Agricultural Research Service (ARS). Each SME scores 1-3 elements from a set of ten factors comprising the number of recorded human infections, transmission in laboratory animals, receptor binding characteristics, population immunity, animal infections, genomic analysis, antigenic relatedness, global prevalence, pathogenesis, and treatment options. The point estimates are averaged across SMEs, scaled by predetermined weights, and summed to give an aggregate score for each of the two questions: 1) the emergence risk and 2) the potential public health impact on sustained transmission. The scores are potentially subjective, and involve multiple experimental assays, possibly taking weeks to compile for a single strain or lineage. This results in a scalability bottleneck: with global efforts collecting thousands of sequences annually, IRAT assessment is not fast enough to fully leverage current surveillance output.

Here we introduce a pattern recognition algorithm to automatically parse out emergent evolutionary constraints operating on Influenza A viruses in the wild, to numerically estimate 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.

To uncover relevant evolutionary constraints, we analyzed variations (point substitutions and indels) of the residue sequences of the two key proteins implicated in cellular entry and exit[1,16], namely HA and NA respectively. By representing these constraints within a predictive framework – the Emergenet– we estimated the odds of a specific mutation to arise in a given strain, and consequently the probability of a specific strain spontaneously evolving into another. 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[17] or standard phylogeny inference tools which assume a constant species-wise mutational characteristics, are not applicable to the problem at hand. Similarly, recently reported algorithms such as FluLeap[18] which identfies host tropism from sequence data, or estimating risk posed by different viral species[14] do not allow strain-specific risk assessment.

The genomic dependencies we uncover are shaped by a functional necessity of conserving/augmenting fitness in the wild. A strain 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 in RNA-polymerase, Influenza A integrates faulty nucleotides at a relatively high rate ($10^{-3} - 10^{-4}$) during replication[19,20]. However, few of these 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[21–25] 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 reflect the effect of 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 appearing at other indices (Fig. 1b). Thus, an Emergenet comprises atmost as many such position-specific predictors as the length of the sequence. These individual predictors are implemented as conditional inference trees[26], 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 indices acting as features in each predictor varies, $e.g.$, in the fragment of the H1N1 HA Emergenet (2020-2021, Fig 1b), the predictor for residue at index 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 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 this study we used HA and NA sequences from unique Influenza A strains in the public NCBI and GISAID databases recorded between 2000-2022 (98,299 in total). We construct Emergenets separately for H1N1 and H3N2 subtypes, and for each flu season, yielding 85 models in total for predicting seasonal dominance. This is advantageous since deep antigenic characterization tend to be substantially low-throughput compared to genome sequencing[27], and incorporation of deep mutational scanning (DMS) assays can 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 (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 $\log Pr(x \rightarrow y)$ may be approximated by the E-distance $\theta(x, y)$. The mathematical intuition relating E-distance to the loglikelihood of spontaneous change is similar to quantifying the odds of a rare biased outcome when we toss a fair coin[29]. Importantly, unlike the edit distance, the E-distance between two fixed sequences may change if only the background strain population changes (SI-Table **??** illustrates that the distance between fixed sequences vary with collection years.

Determining the numerical odds of spontaneous jump $Pr(x \rightarrow y)$ (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 emerging as the dominant one. 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 (See 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 (See Online Methods). The first term gets the solution close to the centroid of the current strain distribution (in the E-distance metric, which is different from the centroid if the standard edit distance is used), and the second term relates to how common the genomic patterns are amongst recent human 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[30]. For each year of the past two decades, we calculated strain forecasts using Eq. (12) with data available six months before the target season. We measured forecast performance by the number of mutations by which the predicted HA/NA sequences deviated from the dominant strain. Our Emergenet-informed forecasts outperform WHO/CDC recommended flu vaccine compositions almost consistently over the past two decades, for both H1N1 and H3N2 subtypes, individually in the northern and the southern hemispheres (which have distinct recommendations). The results broken down by hemisphere/protein/subtypes is given in Table 1. Fig. 2 illustrates the relative gains computed for both subtypes and the two hemispheres. Additional improvement is possible if we recommend multiple strains every season for the vaccine cocktail (Fig. 2e,f,k,l). The details of the specific strain recommendations made by the Emergenet approach for two subtypes (H1N1, H3N2), for two genes (HA, NA) and for the northern and the southern hemispheres over the previous two decades are enumerated in the Supplementary Tables SI-Table **??** through SI-Table 8. While it is recognized that even well-matched strains can fail to induce a strong immune response due to previous infection history of vaccine recipients[31], strain-matching is a crucial component to realizing high vaccine effectiveness[32], and we outperform the current practice (WHO/CDC) as well as recently reported prediction strategies using more standard computational and/or experimental frameworks[6,9].

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 (See 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 $Pr(x \rightarrow y)$ 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 past year of the assessment date, $e.g.$, the IRAT assessment date for A/swine/Shandong/1207/2016 is July 2020, and we use human H1N1 strains collected between 1/7/2019-6/30/2020 for the Emergenet construction. For sub-types with very few recorded human strains (H1N2, H5N1, H5N6, H7N7, H9N2), we consider all subtype-specific human strains collected upto the IRAT assessment date to construct our Emergenet. We then compute the average E-distance between the animal strain of interest and the recent human strains for both HA and NA sequences (using Eq. (2)), finally reporting their geometric mean as our estimated risk. Considering IRAT scores of 22 strains published by the CDC, we find strong out-of-sample support (correlation of -0.704, pvalue < 0.00026, Fig. 3, see Online Methods) for this claim. Importantly, each E-risk score is computable in approximately 6 seconds as opposed to potentially weeks taken by IRAT experimental assays.

The time-dependence of the E-risk reflects the impact of the background (See Table **??** and SI-Fig. 3), and we show that recomputing the risks 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 (-0.59, p-value 0.003).

To map the Emergenet distances to the more recognizable IRAT scores, we train a general linear model (GLM) from the the HA/NA-based E-risk values (See Online Methods). 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, despite our theoretical intuition primarily supporting the emergence phenomenon. With these maps we estimate the IRAT scores of all distinct Influenza A strains collected between 2020 through 2022 April (6066 strains in total), and identify the ones posing maximal risk (Fig. 3c). $1,773$ strains 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. 1), and collapsed all leaves within 15 edits, showing only the most risky strain within a collapsed group. This leaves 75 strains as shown in Fig. 1 (68 with emergence risk > 6.25, and 6 with emergence risk above 6.5, see Table 3). A substaintial number of risky strains are H1N1, followed by H3N2, with a small number of H7N9 and H9N2. Five 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 (See Fig. 4), which shows substatial residue replacements, in and out of the receptor binding domain (RBD).

Swines are known to be efficient mixing vessels[3,33**?** ,34], and hence unsuprisingly 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 again unsurprising, in the light of these hosts been recently suggested as efficient mixing vessels to breed human-compatible strains[36]. Thus, qualitatively 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.

In conclusion, while numerous tools exist for ad hoc quantification of genomic similarity[9,17,37–40], a smaller distance $i.e.$ a higher similarity of two 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]. Our approach is currently limited by the fact that the E-distance calculation is currently applicable 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 multi-variate regression analysis indicates that the most important factor for our approach to succeed is the diversity of the sequence dataset (see Supplementary Table 12), 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.

Online Methods

Next, we briefly describe the details of 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:*

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

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[26], 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.

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, $Pr(x \rightarrow y)$, is bounded by:

$$\omega^Q_y e^{\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \geqq Pr(x \rightarrow y) \geqq \omega^Q_y e^{-\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \tag{8}$$

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

Problem 1: 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 $x^{t+\delta}_\star$ 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:

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

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

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

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 $x^{t+\delta}_\star$ may be estimated as a solution to the optimization problem:

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

Problem 2: Measure of Pandemic Potential

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

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

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 teh E-distance induced by the Emergenet computed from the sequences in H^t.

The intuition here is that a lower bound of $\rho(x^t_a)$ scales as average log-likelihood of the x^t_a 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:

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

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:

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

$$\Rightarrow \rho(x^t_a) \geqq A \ln \left\langle Pr(x^t_a \rightarrow x^{t+\delta}_h) \right\rangle + C \tag{16}$$

Noting that A, C are not functions of x^t_a, 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.*

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

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:

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

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[46], 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:

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

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:

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

□

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:*

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

where the constants are given by:

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

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

Data Sharing

Working software is publicly available at https://pypi.org/project/emergenet/. Accession numbers of all sequences used, and acknowledgement documentation for GISAID sequences is available as supplementary information.

Data Source

In this study, we use sequences for the Hemaglutinnin (HA) and Neuraminidase (NA) for Influenza A (for subtypes H1N1 and H3N2), which are key enablers of cellular entry and exit mechanisms respectively[47]. We use two sequences 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 somewhat 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 98,299 sequences in our analysis, although not all were used due to some being duplicates (see SI-Table **??**).

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] Wille, M., Geoghegan, J. L. & Holmes, E. C. How accurately can we assess zoonotic risk? *PLoS biology* **19**, e3001135 (2021).

[12] 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).

[13] Grewelle, R. E. Larger viral genome size facilitates emergence of zoonotic diseases. *bioRxiv* (2020).

[14] 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).

[15] 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).

[16] Gamblin, S. J. & Skehel, J. J. Influenza hemagglutinin and neuraminidase membrane glycoproteins. *Journal of Biological Chemistry* **285**, 28403–28409 (2010).

[17] Posada, D. & Crandall, K. A. Modeltest: testing the model of dna substitution. *Bioinformatics (Oxford, England)* **14**, 817–818 (1998).

[18] 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).

[19] Ahlquist, P. Rna-dependent rna polymerases, viruses, and rna silencing. *Science* **296**, 1270–1273 (2002).

[20] Chen, R. & Holmes, E. C. Avian influenza virus exhibits rapid evolutionary dynamics. *Molecular biology and evolution* **23**, 2336–2341 (2006).

[21] 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).

[22] 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).

[23] 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).

[24] 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).

[25] 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).

[26] Hothorn, T., Hornik, K. & Zeileis, A. Unbiased recursive partitioning: A conditional inference framework. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS* **15**, 651–674 (2006).

[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] 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).

[30] 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).

[31] Cobey, S. *et al.* Poor immunogenicity, not vaccine strain egg adaptation, may explain the low h3n2 influenza vaccine effectiveness in 2012–2013. *Clinical Infectious Diseases* **67**, 327–333 (2018).

[32] Gouma, S., Weirick, M. & Hensley, S. E. Antigenic assessment of the h3n2 component of the 2019-2020 northern hemisphere influenza vaccine. *Nature communications* **11**, 1–5 (2020).

[33] 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).

[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] Fedotov, A. A., Harremoës, P. & Topsoe, F. Refinements of pinsker's inequality. *IEEE Transactions on Information Theory* **49**, 1491–1498 (2003).

[47] McAuley, J., Gilbertson, B., Trifkovic, S., Brown, L. E. & McKimm-Breschkin, J. Influenza virus neuraminidase structure and functions. *Frontiers in microbiology* **10**, 39 (2019).

[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).

Extended Data Table 2

Examples: Qnet induced distance varying for fixed sequence pair when background population changes (rows 1 -5), sequences with small edit distance and large q-distance, and the converse (rows 6-9)

	Edit dist.	Sequence A	Sequence B	Q-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 q-distance from sequence B, arising only from a change in the background population.

Extended Data Table 3

Correlation between q-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

Extended Data Table 4

Number of sequences collected from public databases

Database	Strain	No. of Sequences
NCBI	Influenza H1N1 HA	17,894
NCBI	Influenza H1N1 NA	16,637
NCBI	Influenza H3N2 HA	18,265
NCBI	Influenza H3N2 NA	14,699
GISAID	Influenza H1N1 HA	1,528
GISAID	Influenza H1N1 NA	1,490
GISAID	Influenza H3N2 HA	13,975
GISAID	Influenza H3N2 NA	13,811
Total		98,299

Extended Data Table 5
H1N1 HA Northern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet 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 6
H1N1 HA Southern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet 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 7
H3N2 HA Northern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet 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 8
H3N2 HA Southern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet 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 Table 9
Riskiest Strains Currently Circulating in Swine

H1N1 Strain	HA Risk	NA Risk	Overall Risk	Converted IRAT Score
A/swine/Tennessee/A02524414/2022	0.0201	0.0030	0.0077	6.2
A/swine/Missouri/A02750646/2022	0.0201	0.0070	0.0118	6.2
A/swine/Kansas/A02711847/2022	0.0201	0.0098	0.0141	6.2
A/swine/Iowa/A02636572/2022	0.0166	0.0225	0.0193	6.1
A/swine/Iowa/A02636308/2021	0.0143	0.0266	0.0195	6.1
A/swine/Illinois/A02750711/2022	0.0166	0.0233	0.0197	6.1
A/swine/Iowa/A02636616/2022	0.0166	0.0233	0.0197	6.1
A/swine/Oklahoma/A02246915/2022	0.0166	0.0233	0.0197	6.1
A/swine/Colorado/A02636469/2022	0.0166	0.0233	0.0197	6.1
A/swine/Iowa/A02636297/2021	0.0149	0.0267	0.0200	6.1
H3N2 Strain	HA Risk	NA Risk	Overall Risk	Converted IRAT Score
A/swine/Indiana/A02636492/2022	0.0104	0.0113	0.0108	6.2
A/swine/Indiana/A02636512/2022	0.0104	0.0113	0.0108	6.2
A/swine/Iowa/A02750695/2022	0.0110	0.0120	0.0115	6.2
A/swine/Oklahoma/A02711859/2022	0.0122	0.0114	0.0118	6.2
A/swine/Iowa/A02636351/2022	0.0121	0.0119	0.0120	6.2
A/swine/Iowa/A02636476/2022	0.0121	0.0120	0.0121	6.2
A/swine/Texas/A02636569/2022	0.0122	0.0120	0.0121	6.2
A/swine/Iowa/A02750726/2022	0.0123	0.0120	0.0121	6.2
A/swine/Iowa/A02750740/2022	0.0104	0.0156	0.0127	6.2
A/swine/Indiana/A02636521/2022	0.0104	0.0156	0.0127	6.2

* Converted IRAT Score computed using regression generated from the IRAT vs. Qnet comparison

Supplementary Methods: Notes on q-Distance & Supporting Results

The *q-distance* is a pseudo-metric since distinct sequences can induce the same distributions over each index, and thus evaluate to have a zero distance. This is actually desirable; we do not want our distance to be sensitive to changes that are not biologically relevant. The intuition is that not all sequence variations brought about by substitutions are equally important or likely. Even with no selection pressure, we might still see random variations at an index if such variations do not affect the replicative fitness. Under that scenario, the corresponding Φ_i will predict a flat distribution no matter what the input sequence is, thus contributing nothing to the overall distance. And even if two strains x, y have the same entry at some index i, the remaining residues might induce different distributions Φ_i based on the remote dependencies, $i.e.$, the entries in x_{-i}, y_{-i}. Also, it matters if the sequences come from two different background populations P, Q, $i.e.$, if the induced Qnets Φ^P, Φ^Q are different. Thus, if we construct Qnets for H1N1 Influenza A separately for the collection years 2008 and 2009, then the same exact sequence collected in the respective years might have a non-zero distance between them, reflecting the fact that the background population the sequences arose from are different, inducing possibly different expected mutational tendencies (See SI-Table **??**).

Next, we induce q-distance between a sequence and a population and between two populations.

Definition 5 (Pseudo-metric between populations)**.** *Using the notion of Hausdorff metric between sets:*

$$\forall x \in P, y \in Q,$$

$$\theta(x, Q) = \min_{y \in Q} \theta(x, y) \tag{30}$$

$$\theta(P, Q) = \max \left\{ \max_{x \in P} \theta(x, Q), \max_{y \in Q} \theta(y, P) \right\} \tag{31}$$

In-silico Corroboration of Qnet Constraints

We carry out in-silico experiments to corroborate that the constraints represented within an inferred Qnet are indeed reflective of the biology in play. We compare the results of simulated mutational perturbations to sequences from our databases (for which we have already constructed Qnets), 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 (See SI-Fig. 1). We find that in contrast to random variations, which rapidly diverge the trajectories, the Qnet 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 Qnet does indeed capture realistic constraints.

Multivariate Regression to Identify Factors in Strain Prediction

We investigate the key factors that contribute to our successful prediction of the dominant strain in the next season. We carry out a multivariate regression with data diversity, the complexity of inferred Qnet and the edit distance of the WHO recommendation from the dominant strain as independent variables. 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. Qnet 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 (See SI-Tab. 12).

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] Wille, M., Geoghegan, J. L. & Holmes, E. C. How accurately can we assess zoonotic risk? *PLoS biology* **19**, e3001135 (2021).

[12] 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).

[13] Grewelle, R. E. Larger viral genome size facilitates emergence of zoonotic diseases. *bioRxiv* (2020).

[14] 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).

[15] 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).

[16] Gamblin, S. J. & Skehel, J. J. Influenza hemagglutinin and neuraminidase membrane glycoproteins. *Journal of Biological Chemistry* **285**, 28403–28409 (2010).

[17] Posada, D. & Crandall, K. A. Modeltest: testing the model of dna substitution. *Bioinformatics (Oxford, England)* **14**, 817–818 (1998).

[18] 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).

[19] Ahlquist, P. Rna-dependent rna polymerases, viruses, and rna silencing. *Science* **296**, 1270–1273 (2002).

[20] Chen, R. & Holmes, E. C. Avian influenza virus exhibits rapid evolutionary dynamics. *Molecular biology and evolution* **23**, 2336–2341 (2006).

[21] 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).

[22] 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).

[23] 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).

[24] 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).

[25] 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).

[26] Hothorn, T., Hornik, K. & Zeileis, A. Unbiased recursive partitioning: A conditional inference framework. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS* **15**, 651–674 (2006).

[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] 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).

[30] 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).

[31] Cobey, S. *et al.* Poor immunogenicity, not vaccine strain egg adaptation, may explain the low h3n2 influenza vaccine effectiveness in 2012–2013. *Clinical Infectious Diseases* **67**, 327–333 (2018).

[32] Gouma, S., Weirick, M. & Hensley, S. E. Antigenic assessment of the h3n2 component of the 2019-2020 northern hemisphere influenza vaccine. *Nature communications* **11**, 1–5 (2020).

[33] 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).

[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] Fedotov, A. A., Harremoës, P. & Topsoe, F. Refinements of pinsker's inequality. *IEEE Transactions on Information Theory* **49**, 1491–1498 (2003).

[47] McAuley, J., Gilbertson, B., Trifkovic, S., Brown, L. E. & McKimm-Breschkin, J. Influenza virus neuraminidase structure and functions. *Frontiers in microbiology* **10**, 39 (2019).

[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).

SUPPLEMENTARY FIGURES & TABLES

SI Tab. 1
H1N1 NA Northern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2001-02	A/New Caledonia/20/99	A/New York/447/2001	A/Memphis/15/2000	4	4
2002-03	A/New Caledonia/20/99	A/Paris/0833/2002	A/New York/341/2001	1	5
2003-04	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	3	5
2004-05	A/New Caledonia/20/99	A/Singapore/14/2004	A/New York/223/2003	2	3
2005-06	A/New Caledonia/20/99	A/Taiwan/5524/2005	A/Florida/3e/2004	3	0
2006-07	A/New Caledonia/20/99	A/Massachusetts/08/2006	A/Sofia/361/2005	4	2
2007-08	A/Solomon Islands/3/2006	A/Tennessee/UR06-0106/2007	A/Sofia/490/2006	9	2
2008-09	A/Brisbane/59/2007	A/Sendai/TU66/2008	A/Maryland/04/2007	0	3
2009-10	A/Brisbane/59/2007	A/Thailand/SR08021/2009	A/Paris/910/2008	87	87
2010-11	A/California/7/2009	A/Finland/2460N/2010	A/Rome/709/2009	2	9
2011-12	A/California/7/2009	A/Tula/CRIE-GSYu/2011	A/Oman/SQUH-40/2010	4	2
2012-13	A/California/7/2009	A/Bangalore/697-32/2012	A/Nizhnii Novgorod/CRIE-ZCA/2011	4	0
2013-14	A/California/7/2009	A/Jiangsugusu/SWL1824/2013	A/LongYan/SWL33/2013	5	3
2014-15	A/California/7/2009	A/LongYan/SWL2457/2014	A/Utah/06/2013	9	3
2015-16	A/California/7/2009	A/Michigan/45/2015	A/Maryland/02/2014	14	4
2016-17	A/California/7/2009	A/Mexico/4436/2016	A/India/Pun151245/2015	14	0
2017-18	A/Michigan/45/2015	A/Illinois/37/2017	A/Utah/02/2016	3	3
2018-19	A/Michigan/45/2015	A/Kenya/47/2018	A/Maine/24/2017	4	0
2019-20	A/Brisbane/02/2018	A/Texas/7939/2019	A/Missouri/03/2018	1	0
2020-21	A/Hawaii/70/2019	A/Togo/897/2020	A/Texas/112/2019	0	5
2021-22	A/Victoria/2570/2019	A/Cote_d'Ivoire/3729/2021	A/Togo/0071/2021	1	5
2022-23	-1	-1	A/Lyon/820/2021	-1	-1

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

SI Tab. 2
H1N1 NA Southern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2001-02	A/New Caledonia/20/99	A/New York/447/2001	A/Canterbury/37/2000	4	6
2002-03	A/New Caledonia/20/99	A/Paris/0833/2002	A/New York/447/2001	1	5
2003-04	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	3	5
2004-05	A/New Caledonia/20/99	A/Singapore/14/2004	A/Memphis/5/2003	2	3
2005-06	A/New Caledonia/20/99	A/Taiwan/5524/2005	A/Canterbury/106/2004	3	6
2006-07	A/New Caledonia/20/99	A/Massachusetts/08/2006	A/Sofia/361/2005	4	2
2007-08	A/New Caledonia/20/99	A/Tennessee/UR06-0106/2007	A/Thailand/RMSC-UDN-20/2006	4	8
2008-09	A/Solomon Islands/3/2006	A/Sendai/TU66/2008	A/Tennessee/UR06-0151/2007	15	13
2009-10	A/Brisbane/59/2007	A/Thailand/SR08021/2009	A/Nebraska/07/2008	87	87
2010-11	A/California/7/2009	A/Finland/2460N/2010	A/Rome/709/2009	2	9
2011-12	A/California/7/2009	A/Tula/CRIE-GSYu/2011	A/Finland/2460N/2010	4	2
2012-13	A/California/7/2009	A/Bangalore/697-32/2012	A/Tula/CRIE-GSYu/2011	4	0
2013-14	A/California/7/2009	A/Jiangsugusu/SWL1824/2013	A/Oman/SQUH-63/2012	5	4
2014-15	A/California/7/2009	A/LongYan/SWL2457/2014	A/NanPing/SWL1640/2013	9	6
2015-16	A/California/7/2009	A/Michigan/45/2015	A/LongYan/SWL2457/2014	14	5
2016-17	A/California/7/2009	A/Mexico/4436/2016	A/Michigan/45/2015	14	0
2017-18	A/Michigan/45/2015	A/Illinois/37/2017	A/Mexico/4436/2016	3	3
2018-19	A/Michigan/45/2015	A/Kenya/47/2018	A/Kentucky/26/2017	4	2
2019-20	A/Michigan/45/2015	A/Texas/7939/2019	A/Kenya/47/2018	4	0
2020-21	A/Brisbane/02/2018	A/Togo/897/2020	A/Texas/7939/2019	6	5
2021-22	A/Victoria/2570/2019	A/Cote_D'Ivoire/1496/2021	A/NAGASAKI/8/2020	1	6
2022-23	-1	-1	A/Dakar/35/2021	-1	-1

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

SI Tab. 3
H3N2 NA Northern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2003-04	A/Moscow/10/99	A/Denmark/107/2003	A/New York/100/2002	13	3
2004-05	A/Fujian/411/2002	A/Hyogo/36/2004	A/New York/20/2003	3	16
2005-06	A/California/7/2004	A/Denmark/203/2005	A/Hong Kong/HKU20/2004	4	0
2006-07	A/Wisconsin/67/2005	A/Berlin/32/2006	A/Mexico/InDRE2227/2005	1	1
2007-08	A/Wisconsin/67/2005	A/Brazil/80/2007	A/Baden-Wuerttemberg/17/2006	8	7
2008-09	A/Brisbane/10/2007	A/Missouri/05/2008	A/Washington/01/2007	3	2
2009-10	A/Brisbane/10/2007	A/Oklahoma/09/2009	A/Wisconsin/24/2008	3	1
2010-11	A/Perth/16/2009	A/California/17/2010	A/New York/70/2009	2	3
2011-12	A/Perth/16/2009	A/Texas/14/2011	A/California/14/2010	3	2
2012-13	A/Victoria/361/2011	A/New York/02/2012	A/Singapore/C2011.493/2011	4	1
2013-14	A/Victoria/361/2011	A/Michigan/02/2013	A/New York/01/2012	3	1
2014-15	A/Texas/50/2012	A/Tehran/69634/2014	A/Boston/DOA2-176/2013	3	1
2015-16	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Thailand/CU-B10520/2014	3	0
2016-17	A/Hong Kong/4801/2014	A/North Carolina/62/2016	A/Delaware/02/2015	7	2
2017-18	A/Hong Kong/4801/2014	A/Texas/277/2017	A/New York/03/2016	8	0
2018-19	A/Singapore/INFIMH-16-0019/2016	A/Japan/NHRC_FDX70352/2018	A/Colorado/11/2017	4	3
2019-20	A/Kansas/14/2017	A/Washington/9757/2019	A/Guangxi-Fangcheng/54/2019	3	11
2020-21	A/Hong Kong/2671/2019	A/Bangladesh/1004005/2020	A/Maryland/02/2019	3	13
2021-22	A/Cambodia/e0826360/2020	A/Stockholm/10/2022	A/Bangladesh/1916/2020	2	2
2022-23	-1	-1	A/Iowa/20/2022	-1	-1

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

SI Tab. 4
H3N2 NA Southern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2003-04	A/Moscow/10/99	A/Denmark/107/2003	A/New York/101/2002	13	3
2004-05	A/Fujian/411/2002	A/Hyogo/36/2004	A/New York/20/2003	3	16
2005-06	A/Wellington/1/2004	A/Denmark/203/2005	A/Wellington/1/2004	2	2
2006-07	A/California/7/2004	A/Berlin/32/2006	A/Mexico/InDRE2227/2005	3	1
2007-08	A/Wisconsin/67/2005	A/Brazil/80/2007	A/Ohio/06/2006	8	10
2008-09	A/Brisbane/10/2007	A/Missouri/05/2008	A/Brazil/80/2007	3	2
2009-10	A/Brisbane/10/2007	A/Oklahoma/09/2009	A/Wisconsin/24/2008	3	1
2010-11	A/Perth/16/2009	A/California/17/2010	A/New York/70/2009	2	3
2011-12	A/Perth/16/2009	A/Texas/14/2011	A/Virginia/05/2010	3	2
2012-13	A/Perth/16/2009	A/New York/02/2012	A/Texas/14/2011	4	1
2013-14	A/Victoria/361/2011	A/Michigan/02/2013	A/New York/02/2012	3	3
2014-15	A/Texas/50/2012	A/Tehran/69634/2014	A/Michigan/02/2013	3	1
2015-16	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Tehran/69634/2014	3	2
2016-17	A/Hong Kong/4801/2014	A/North Carolina/62/2016	A/Parma/471/2015	7	2
2017-18	A/Hong Kong/4801/2014	A/Texas/277/2017	A/Guangdong/264/2016	8	0
2018-19	A/Singapore/INFIMH-16-0019/2016	A/Japan/NHRC_FDX70352/2018	A/Texas/277/2017	4	3
2019-20	A/Switzerland/8060/2017	A/Washington/9757/2019	A/Pennsylvania/317/2018	10	10
2020-21	A/South Australia/34/2019	A/Bangladesh/1004005/2020	A/Washington/9757/2019	1	13
2021-22	A/Hong Kong/2671/2019	A/India/PUN-NIV301718/2021	A/India/PUN-NIV301132/2021	6	4
2022-23	-1	-1	A/Michigan/UOM10045036720/2022	-1	-1

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

SI Tab. 5
H1N1 NA Northern Hemisphere (Multi-cluster)

Year	WHO Recommendation	WHO Error	Qnet Error 1	Qnet Error 2	Qnet Recommendation 1	Qnet Recommendation 2
2001-02	A/New Caledonia/20/99	4	1	6	A/New South Wales/26/2000	A/Canterbury/37/2000
2002-03	A/New Caledonia/20/99	1	0	5	A/Wellington/1/2001	A/New York/447/2001
2003-04	A/New Caledonia/20/99	3	2	8	A/Paris/0833/2002	A/Taiwan/141/2002
2004-05	A/New Caledonia/20/99	2	3	4	A/Memphis/5/2003	A/Hanoi/1004/2003
2005-06	A/New Caledonia/20/99	3	0	1	A/Denmark/130/2004	A/Paris/650/2004
2006-07	A/New Caledonia/20/99	4	2	8	A/Sofia/361/2005	A/Wellington/11/2005
2007-08	A/Solomon Islands/3/2006	9	4	8	A/Sofia/246/2006	A/New York/8/2006
2008-09	A/Brisbane/59/2007	0	13	19	A/Tennessee/UR06-0151/2007	A/Ohio/UR06-0178/2007
2009-10	A/Brisbane/59/2007	87	88	90	A/Sendai/TU66/2008	A/Japan/618/2008
2010-11	A/California/7/2009	2	1	6	A/South Carolina/WRAIR1645P/2009	A/Wisconsin/629-D00809/2009
2011-12	A/California/7/2009	4	1	3	A/England/21680633/2010	A/Hangzhou/178/2010
2012-13	A/California/7/2009	4	1	22	A/Joshkar-Ola/CRIE-BLP/2011	A/Rio Grande do Sul/578/2011
2013-14	A/California/7/2009	5	4	13	A/Thailand/MR10580/2012	A/Mexico/INMEGEN-INER 15/2012
2014-15	A/California/7/2009	9	3	7	A/Minnesota/02/2013	A/Helsinki/430/2013
2015-16	A/California/7/2009	14	4	7	A/Helsinki/808M/2014	A/Virginia/NHRC430739/2014
2016-17	A/California/7/2009	14	0	3	A/Michigan/45/2015	A/Colorado/30/2015
2017-18	A/Michigan/45/2015	3	3	8	A/Mexico/4436/2016	A/Arizona/03/2016
2018-19	A/Michigan/45/2015	4	0	4	A/California/NHRC_QV11073/2017	A/Minnesota/35/2017
2019-20	A/Brisbane/02/2018	1	0	2	A/Kenya/47/2018	A/Colorado/7682/2018
2020-21	A/Hawaii/70/2019	0	3	8	A/California/NHRC-OID_BOX-ILI-0012/2019	A/Indiana/30/2019
2021-22	A/Victoria/2570/2019	1	5	51	A/Togo/0071/2021	A/Yunnan-Mengzi/1462/2020
2022-23	-1	-1	-1	-1	A/Netherlands/10646/2022	A/Sydney/234/2022

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

SI Tab. 6
H1N1 NA Southern Hemisphere (Multi-cluster)

Year	WHO Recommendation	WHO Error	Qnet Error 1	Qnet Error 2	Qnet Recommendation 1	Qnet Recommendation 2
2001-02	A/New Caledonia/20/99	4	1	6	A/New South Wales/26/2000	A/Canterbury/37/2000
2002-03	A/New Caledonia/20/99	1	0	5	A/Wellington/1/2001	A/New York/447/2001
2003-04	A/New Caledonia/20/99	3	2	8	A/Paris/0833/2002	A/Taiwan/141/2002
2004-05	A/New Caledonia/20/99	2	3	4	A/Memphis/5/2003	A/Hanoi/1004/2003
2005-06	A/New Caledonia/20/99	3	0	1	A/Denmark/130/2004	A/Paris/650/2004
2006-07	A/New Caledonia/20/99	4	2	8	A/Sofia/361/2005	A/Wellington/11/2005
2007-08	A/New Caledonia/20/99	4	4	8	A/Sofia/246/2006	A/New York/8/2006
2008-09	A/Solomon Islands/3/2006	15	13	19	A/Tennessee/UR06-0151/2007	A/Ohio/UR06-0178/2007
2009-10	A/Brisbane/59/2007	87	88	90	A/Sendai/TU66/2008	A/Japan/618/2008
2010-11	A/California/7/2009	2	1	6	A/South Carolina/WRAIR1645P/2009	A/Wisconsin/629-D00809/2009
2011-12	A/California/7/2009	4	1	3	A/England/21680633/2010	A/Hangzhou/178/2010
2012-13	A/California/7/2009	4	1	22	A/Joshkar-Ola/CRIE-BLP/2011	A/Rio Grande do Sul/578/2011
2013-14	A/California/7/2009	5	4	13	A/Thailand/MR10580/2012	A/Mexico/INMEGEN-INER 15/2012
2014-15	A/California/7/2009	9	3	7	A/Minnesota/02/2013	A/Helsinki/430/2013
2015-16	A/California/7/2009	14	4	7	A/Helsinki/808M/2014	A/Virginia/NHRC430739/2014
2016-17	A/California/7/2009	14	0	3	A/Michigan/45/2015	A/Colorado/30/2015
2017-18	A/Michigan/45/2015	3	3	8	A/Mexico/4436/2016	A/Arizona/03/2016
2018-19	A/Michigan/45/2015	4	0	4	A/California/NHRC_QV11073/2017	A/Minnesota/35/2017
2019-20	A/Michigan/45/2015	4	0	2	A/Kenya/47/2018	A/Colorado/7682/2018
2020-21	A/Brisbane/02/2018	5	2	7	A/California/NHRC-OID_BOX-ILI-0012/2019	A/Indiana/30/2019
2021-22	A/Victoria/2570/2019	1	7	58	A/Togo/0155/2021	A/Shandong/00204/2021
2022-23	-1	-1	-1	-1	A/Switzerland/86136/2022	A/Wisconsin/04/2021

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

SI Tab. 7
H3N2 NA Northern Hemisphere (Multi-cluster)

Year	WHO Recommendation	WHO Error	Qnet Error 1	Qnet Error 2	Qnet Recommendation 1	Qnet Recommendation 2
2003-04	A/Moscow/10/99	13	4	5	A/Auckland/612/2002	A/New York/87/2002
2004-05	A/Fujian/411/2002	3	16	18	A/New York/20/2003	A/New York/12/2003
2005-06	A/California/7/2004	4	1	7	A/New York/358/2004	A/Singapore/36/2004
2006-07	A/Wisconsin/67/2005	1	3	8	A/Macau/557/2005	A/Hong Kong/HKU53/2005
2007-08	A/Wisconsin/67/2005	8	0	10	A/Wisconsin/42/2006	A/Wisconsin/44/2006
2008-09	A/Brisbane/10/2007	3	4	10	A/Missouri/06/2007	A/Japan/72/2007
2009-10	A/Brisbane/10/2007	3	1	7	A/Wisconsin/24/2008	A/Mississippi/UR07-0042/2008
2010-11	A/Perth/16/2009	2	3	8	A/New York/70/2009	A/Japan/883/2009
2011-12	A/Perth/16/2009	3	2	2	A/California/19/2010	A/Virginia/05/2010
2012-13	A/Victoria/361/2011	4	1	12	A/Texas/14/2011	A/Singapore/GP1684/2011
2013-14	A/Victoria/361/2011	3	1	5	A/Idaho/38/2012	A/Pavia/135/2012
2014-15	A/Texas/50/2012	3	1	1	A/Nevada/05/2013	A/Michigan/02/2013
2015-16	A/Switzerland/9715293/2013	3	0	4	A/Nicaragua/6866_14/2014	A/Iran/91244/2014
2016-17	A/Hong Kong/4801/2014	7	1	25	A/New Jersey/13/2015	A/California/NHRC_BRD41056N/2015
2017-18	A/Hong Kong/4801/2014	9	1	4	A/Guangdong/264/2016	A/Victoria/668/2016
2018-19	A/Singapore/INFIMH-16-0019/2016	3	2	4	A/Netherlands/3530/2017	A/Washington/17/2017
2019-20	A/Kansas/14/2017	3	4	10	A/England/538/2018	A/California/BRD12490N/2018
2020-21	A/Hong Kong/2671/2019	3	1	13	A/England/9738/2019	A/Washington/9757/2019
2021-22	A/Cambodia/e0826360/2020	2	3	7	A/Laos/527/2021	A/Michigan/UOM10045655748/2020
2022-23	-1	-1	-1	-1	A/Maine/02/2022	A/Michigan/UOM10042819294/2021

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

SI Tab. 8
H3N2 NA Southern Hemisphere (Multi-cluster)

Year	WHO Recommendation	WHO Error	Qnet Error 1	Qnet Error 2	Qnet Recommendation 1	Qnet Recommendation 2
2003-04	A/Moscow/10/99	13	4	5	A/Auckland/612/2002	A/New York/87/2002
2004-05	A/Fujian/411/2002	3	16	18	A/New York/20/2003	A/New York/12/2003
2005-06	A/Wellington/1/2004	2	1	7	A/New York/358/2004	A/Singapore/36/2004
2006-07	A/California/7/2004	3	3	8	A/Macau/557/2005	A/Hong Kong/HKU53/2005
2007-08	A/Wisconsin/67/2005	8	0	10	A/Wisconsin/42/2006	A/Wisconsin/44/2006
2008-09	A/Brisbane/10/2007	3	4	10	A/Missouri/06/2007	A/Japan/72/2007
2009-10	A/Brisbane/10/2007	3	1	7	A/Wisconsin/24/2008	A/Mississippi/UR07-0042/2008
2010-11	A/Perth/16/2009	2	3	8	A/New York/70/2009	A/Japan/883/2009
2011-12	A/Perth/16/2009	3	2	2	A/California/19/2010	A/Virginia/05/2010
2012-13	A/Perth/16/2009	4	1	12	A/Texas/14/2011	A/Singapore/GP1684/2011
2013-14	A/Victoria/361/2011	3	1	5	A/Idaho/38/2012	A/Pavia/135/2012
2014-15	A/Texas/50/2012	3	1	1	A/Nevada/05/2013	A/Michigan/02/2013
2015-16	A/Switzerland/9715293/2013	3	0	4	A/Nicaragua/6866_14/2014	A/Iran/91244/2014
2016-17	A/Hong Kong/4801/2014	7	1	25	A/New Jersey/13/2015	A/California/NHRC_BRD41056N/2015
2017-18	A/Hong Kong/4801/2014	9	1	4	A/Guangdong/264/2016	A/Victoria/668/2016
2018-19	A/Singapore/INFIMH-16-0019/2016	3	2	4	A/Netherlands/3530/2017	A/Washington/17/2017
2019-20	A/Switzerland/8060/2017	10	4	10	A/England/538/2018	A/California/BRD12490N/2018
2020-21	A/South Australia/34/2019	1	1	13	A/England/9738/2019	A/Washington/9757/2019
2021-22	A/Hong Kong/2671/2019	6	1	49	A/Darwin/11/2021	A/Hawaii/28/2020
2022-23	-1	-1	-1	-1	A/Congo/313/2021	A/Texas/12723/2022

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

SI Tab. 9
Influenza A Strains Evaluated by IRAT and Corresponding Qnet Computed Risk Scores

Influenza Virus	Subype	IRAT Date	IRAT Emergence Score	IRAT Impact Score	HA Sample	NA Sample	HA Avg. E-distance	NA Avg. E-distance	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

* HA strain is not available for A/duck/New York/1996, so this strain is omitted.

** Could not construct a Qnet of human sequence data available for that virus sub-type (less than 30 strains), so we constructed a Qnet using all human strains that match the HA sub-type, i.e. H5NX for H5N6.

*** These strains did not have enough human sequence data to generate a Qnet, even when only considering the HA sub-type. Thus, we estimated the risk score using every Qnet from the other IRAT strains, and took the average among NA and HA. Finally, we took the geometric mean of the resulting NA and HA averages.

SI Tab. 11
General linear model for evaluating effect of data diversity on Qnet performance

Variable Name	Description
qnet_complexity	Cumulative number of nodes in all predictors in the corresponding Qnet
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 ˜ qnet_complexity + data_diversity + qnet_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
qnet_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
qnet_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 ˜ qnet_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
qnet_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
===================================================================================
```

SI Tab. 12

General linear model evaluating Qnet emergence risk predictions against IRAT estimates

```
Model: IRAT_Emergence_Score ˜ Geometric_Mean
================================================================================
Dep. Variable:     IRAT_Emergence_Score   No. Observations:                   22
Model:                              GLM   Df Residuals:                       20
Model Family:                  Gaussian   Df Model:                            1
Link Function:                 identity   Scale:                         0.86853
Method:                            IRLS   Log-Likelihood:                -28.618
Date:                  Tue, 25 Oct 2022   Deviance:                       17.371
Time:                          00:58:27   Pearson chi2:                     17.4
No. Iterations:                       3   Pseudo R-squ. (CS):             0.5919
Covariance Type:              nonrobust
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          6.2467      0.356     17.529      0.000       5.548       6.945
Geometric_Mean    -8.1063      1.830     -4.429      0.000     -11.693      -4.519
==================================================================================
```

```
Model: IRAT_Emergence_Score ˜ Geometric_Mean + HA_Avg_Qdist*NA_Avg_Qdist
================================================================================
Dep. Variable:     IRAT_Emergence_Score   No. Observations:                   22
Model:                              GLM   Df Residuals:                       17
Model Family:                  Gaussian   Df Model:                            4
Link Function:                 identity   Scale:                         0.69369
Method:                            IRLS   Log-Likelihood:                -24.357
Date:                  Tue, 25 Oct 2022   Deviance:                       11.793
Time:                          00:58:59   Pearson chi2:                     11.8
No. Iterations:                       3   Pseudo R-squ. (CS):             0.7797
Covariance Type:              nonrobust
=============================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                     6.8403      0.442     15.459      0.000       5.973       7.708
Geometric_Mean              -23.7466      9.674     -2.455      0.014     -42.707      -4.786
HA_Avg_Qdist                  1.9097      3.979      0.480      0.631      -5.889       9.708
NA_Avg_Qdist                 -1.8133      2.826     -0.642      0.521      -7.353       3.726
HA_Avg_Qdist:NA_Avg_Qdist    54.2280     21.474      2.525      0.012      12.139      96.317
=============================================================================================
```

SI Tab. 13
Numbering Conversion to pdm09 and H3 Schemes

Query	H1N1pdm	H3
1	-	-
2	-	-
3	-	-
4	-	-
5	-	-
6	-	-
7	-	-
8	-	-
9	-	-
10	-	-
11	-	-
12	-	-
13	-	-
14	-	-
15	-	-
16	-	-
17	-	-
-	-	1
-	-	2
-	-	3
-	-	4
-	-	5
-	-	6
-	-	7
-	-	8
-	-	9
-	-	10
18	1	11
19	2	12
20	3	13
21	4	14
22	5	15
23	6	16
24	7	17
25	8	18
26	9	19
27	10	20
28	11	21
29	12	22
30	13	23
31	14	24
32	15	25
33	16	26
34	17	27
35	18	28
36	19	29
37	20	30
38	21	31
39	22	32
40	23	33
41	24	34
42	25	35
43	26	36
44	27	37
45	28	38
46	29	39
47	30	40
48	31	41
49	32	42
50	33	43
51	34	44
52	35	45
53	36	46
54	37	47
55	38	48
56	39	49
57	40	50
58	41	51
59	42	52
60	43	53
61	44	54
62	45	-
63	46	55
64	47	56
65	48	57
66	49	58
67	50	59
68	51	60
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
69	52	61
70	53	62
71	54	63
72	55	64
73	56	65
74	57	66
75	58	67
77	60	69
78	61	70
79	62	71
80	63	72
81	64	73
82	65	74
83	66	75
84	67	76
85	68	77
86	69	78
87	70	79
88	71	80
89	72	81
90	73	82
91	74	-
92	75	83
93	76	84
94	77	85
95	78	86
96	79	87
97	80	88
98	81	89
99	82	90
100	83	91
101	84	92
102	85	-
103	86	93
104	87	94
105	88	95
106	89	96
107	90	97
108	91	98
109	92	99
110	93	100
111	94	101
112	95	102
-	-	-
-	-	-
113	96	103
114	97	104
115	98	105
116	99	106
117	100	107
118	101	108
119	102	109
120	103	110
121	104	111
122	105	112
123	106	113
124	107	114
125	108	115
126	109	116
127	110	117
128	111	118
129	112	119
130	113	120
131	114	121
132	115	122
133	116	123
-	-	-
-	-	-
134	117	124
135	118	125
136	119	-
137	120	-
138	121	-
139	122	126
140	123	127
141	124	128
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
142	125	129
143	126	130
144	127	131
145	128	132
146	129	133
147	130	-
148	131	134
149	132	135
150	133	136
151	134	137
152	135	138
153	136	139
154	137	140
155	138	141
-	-	-
156	139	142
157	140	143
158	141	144
159	142	145
160	143	146
161	144	147
162	145	148
163	146	149
164	147	150
165	148	151
166	149	152
167	150	153
168	151	154
169	152	155
170	153	156
171	154	157
172	155	158
-	-	-
-	-	-
-	-	-
-	-	-
173	156	159
174	157	160
175	158	161
176	159	162
177	160	163
178	161	164
179	162	165
180	163	166
181	164	167
182	165	168
183	166	169
184	167	170
-	-	-
185	168	171
186	169	172
187	170	173
-	-	-
188	171	174
189	172	175
190	173	176
191	174	177
192	175	178
193	176	179
194	177	180
195	178	181
196	179	182
197	180	183
198	181	184
199	182	185
200	183	186
201	184	187
202	185	188
203	186	189
204	187	190
205	188	191
206	189	192
207	190	193
208	191	194
209	192	195
210	193	196
211	194	197
212	195	198
213	196	199
-	-	-
214	197	200
215	198	201
216	199	202
217	200	203
218	201	204
219	202	205
220	203	206
221	204	207
222	205	208
223	206	209
224	207	210
225	208	211
226	209	212
227	210	213
228	211	214
229	212	215
230	213	216
231	214	217
232	215	218
233	216	219
234	217	220
235	218	221
236	219	222
237	220	223
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
238	221	224
239	222	225
240	223	226
241	224	227
242	225	228
243	226	229
244	227	230
245	228	231
246	229	232
247	230	233
248	231	234
249	232	235
250	233	236
251	234	237
252	235	238
253	236	239
254	237	240
255	238	241
256	239	242
257	240	243
258	241	244
259	242	245
260	243	246
261	244	247
262	245	248
263	246	249
264	247	250
265	248	251
266	249	252
267	250	253
268	251	254
269	252	255
270	253	256
271	254	257
272	255	258
273	256	259
274	257	260
275	258	261
276	259	262
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
277	260	-
278	261	263
279	262	264
280	263	265
281	264	266
282	265	267
283	266	268
284	267	269
285	268	270
286	269	271
287	270	272
288	271	273
289	272	274
290	273	275
291	274	276
292	275	277
293	276	278
294	277	279
295	278	280
296	279	281
297	280	282
298	281	283
299	282	284
300	283	285
-	-	-
301	284	286
302	285	287
303	286	288
304	287	289
305	288	290
306	289	291
307	290	292
308	291	293
309	292	294
310	293	295
311	294	296
-	-	-
312	295	297
313	296	298