Emergenet: Fast Scalable Pandemic Risk Estimation of Influenza A Strains Collected 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 BioloScalegy, 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: **Novel Influenza A strains emerging into humans from animal reservoirs pose an elevated risk of triggering global pandemics, as evidenced by multiple similar occurrences in the past[1–4] due to the possibility of large antigenic shifts from the circulating human strains. Yet, our current ability to scalably evaluate the pandemic potential of individual strains that do not yet circulate in humans remains limited. In this study, we introduce a computational approach, known as the Emergenet, to learn how viable genotypic variations are shaped by emergent evolutionary constraints using only genomic sequence data. Analyzing Hemaglutinnin (HA) and Neuraminidase (NA) amino acid sequences from nearly 100,000 unique Influenza A strains from NCBI and GISAID public databases, our proposed algorithm merges machine learning and large deviation theory to estimate the likelihood of specific future mutations, ultimately yielding the numerical odds of one parent strain giving rise to a specific descendant via natural evolutionary processes. After validating our model on the problem of forecasting the dominant strain(s) of the upcoming flu season, with Emergenet-based forecasts significantly outperforming World Health Organization (WHO) recommended flu vaccine compositions almost consistently over the past two decades for H1N1 and H3N2 subtypes, individually in the Northern and the Southern hemispheres (% improvement), we assess the pandemic potential of novel animal strains that do not yet circulate in humans. While the state-of-the-art Influenza Risk Assessment Tool (IRAT) from the Center for Disease Control (CDC) comprises multiple time-consuming experimental assays, our proposed risk score can be evaluated in seconds for each new strain, while strongly correlating with the published IRAT scores (R^2 =). This substantial speedup (weeks vs seconds) in identifying risky strains is necessary to fully exploit the current surveillance capacity via scalably analyzing tens of thousands of strains collected every year. Thus, our results potentially enables meaningful preemptive pandemic mitigation strategies, the relevance of which cannot be overstated in the aftermath of SARS-CoV-2 emergence.**

INTRODUCTION

Influenza viruses constantly evolve[5], producing sequence alterations over a time scale of months that perturb surface protein structures sufficiently to evade the prevailing host immunity, and cause the recurring seasonal flu epidemic. These periodic infection peaks claim a quarter to half a million lives[6] globally, and currently our response hinges on yearly innoculation of the population with a reformulated vaccine[5,7]. Failing to correctly predict the dominant strain in the upcoming season reduces vaccine effectiveness[8] dramatically, and despite recent advances[6,9] such predictions remain imperfect. In addition to the seasonal epidemic, novel Influenza A strains spilling over into humans from animal reservoirs have caused global 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 sudden emergence of COVID-19 and the ensuing devastating pandemic 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 is common animal hosts, can easily incorporate genes from multiple strains and (re)emerge as novel human pathogens[3,10], thus harboring a high potential of triggering the next pandemic.

A possible approach to mitigating such risk is to identify specific strains in animal hosts that do not yet circulate in humans but have the potential to spill-over and quickly achieve human-to-human (HH) transmission capability.

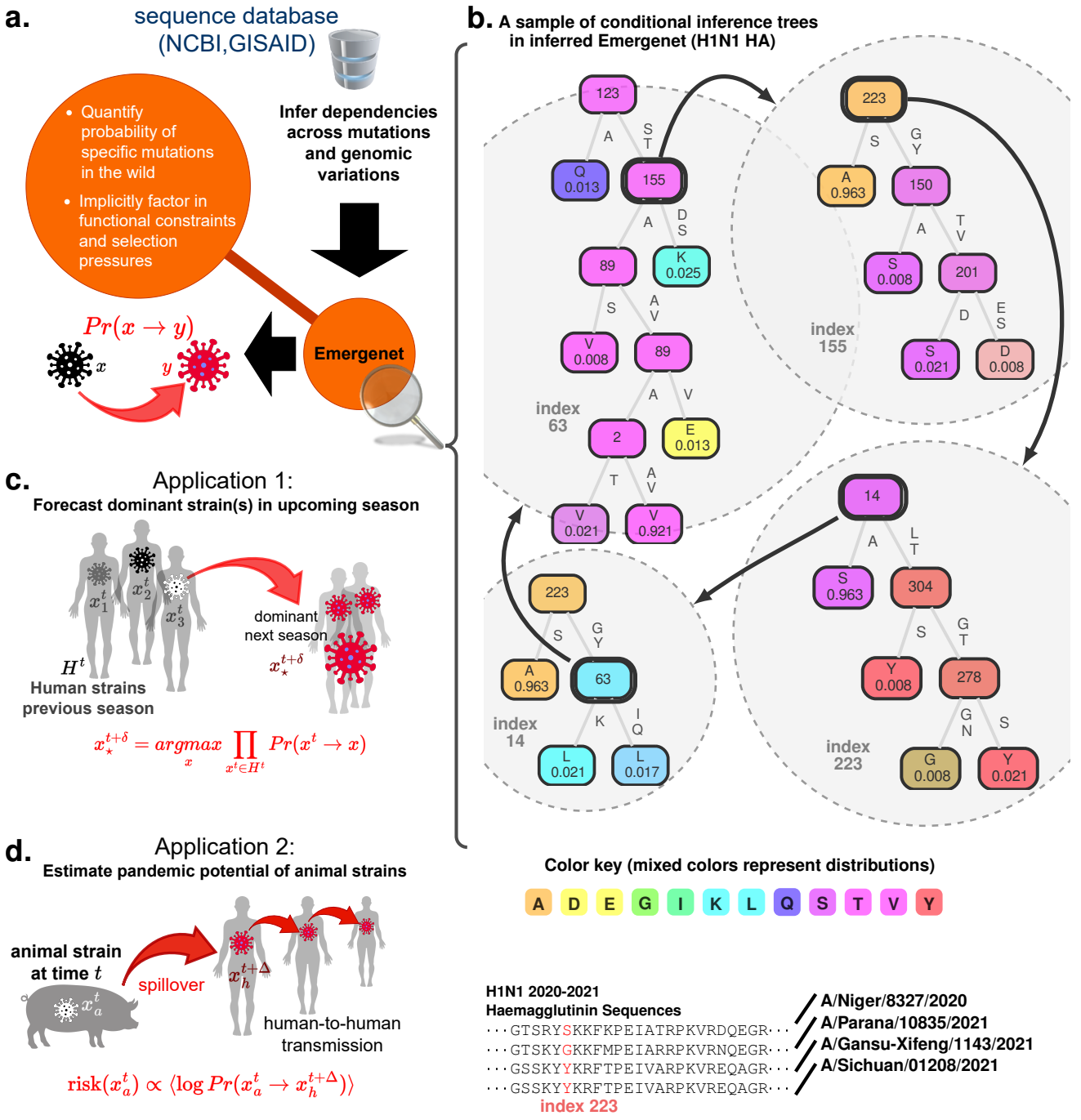


Fig. 1. **Emergenet inference and applications**. **a**, Variations of genomes for identical subgroups of Influenza A are analyzed to infer a recursive forest of conditional inference trees[11] – the Emergenet– which maximally captures the emergent dependencies between an a priori unspecified number of mutations, deletions and insertions. With these inferred dependencies we can estimate the numerical odds of specific mutations, and by extension, the numerical value of the probability of one strain giving rise to another in the wild, under complex selection pressures from the background. b, Snapshot of decision trees from the Emergenet constructed for H1N1 haemagglutinnin 2018 sequences. Note that the decision tree predicting the bases at index 1274 uses the bases at 1064, 1445, 197 as features. These features are automatically selected, as being maximally predictive of the bases be at 1274. Then, we compute predictors for each of these feature indices, *e.g.* trees for index 1064, which involves index 1314 and 339 as features. Continuing, we find that the trees for index 1314 involves indices 1263, 636 and 21, and that for 1263 involves 1314, 667 and 313. The predictor for 1263 depends on 1314, and that for 1314 depends on 1263, revealing the recursive structure of Emergenet. c, First application: With Emergenet induced ability to quantify mutation probabilities, we forecast dominant strain(s) for the next flu season, using only sequences collected in the previous season (and the inferred Emergenet, using data from the past year). d, Second application: estimation of the risk of a global pandemic posed by individual animal strains that are still not known to circulate in humans.

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 animal strains is limited.

#1. IRAT description

CDC's current solution to this problem is the Influenza Risk Assessment Tool (IRAT), which uses a combination of ten weighted risk elements, aiming to factor in 1) expert-selected properties of the virus, 2) attributes of the population, and

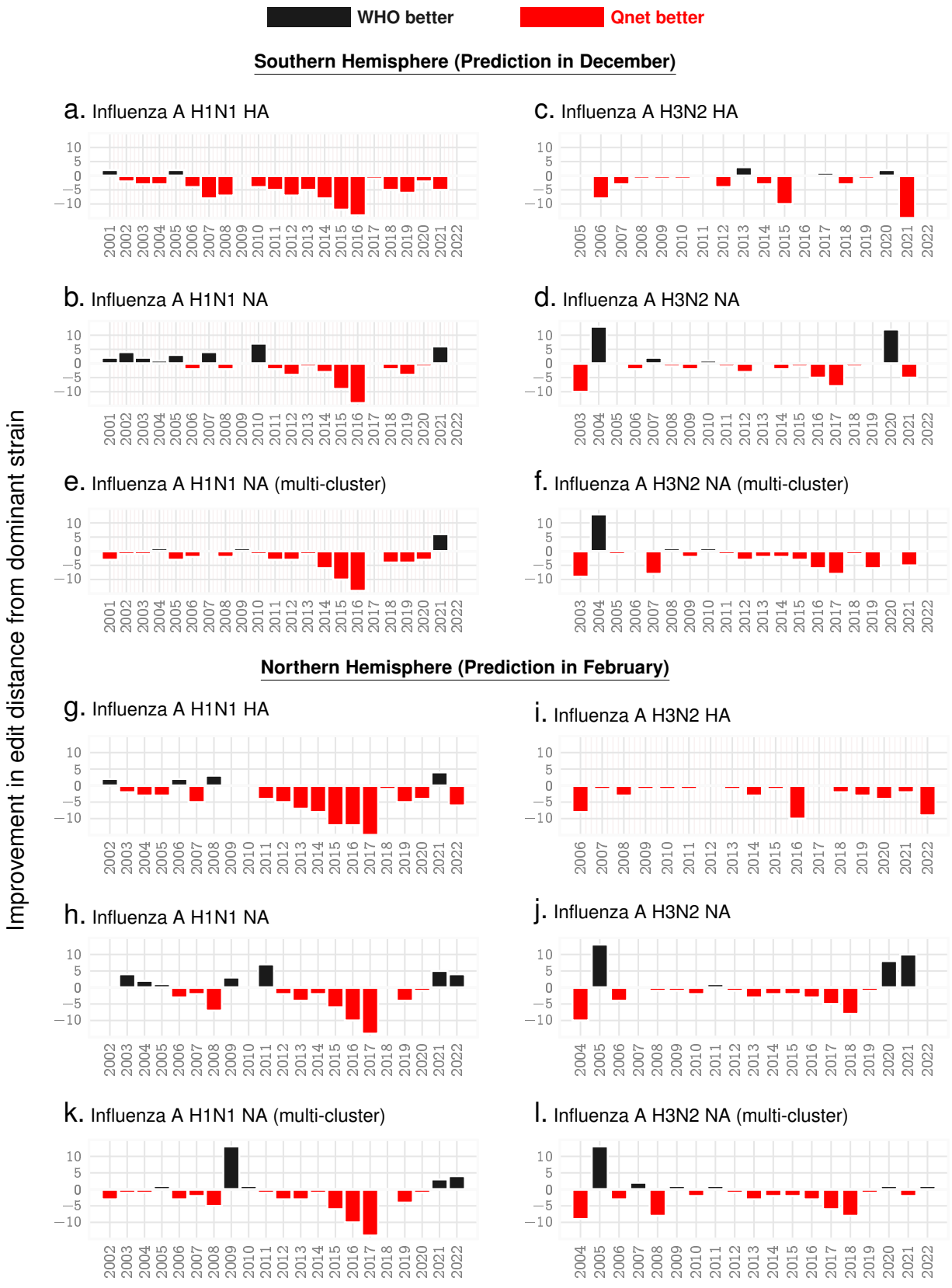


Fig. 2. **Seasonal predictions for Influenza A.** Relative out-performance of Qnet predictions against WHO recommendations for H1N1 and H3N2 sub-types for the HA and NA coding sequences over the both hemispheres. The negative bars (red) indicate the reduced edit distance between the predicted sequence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

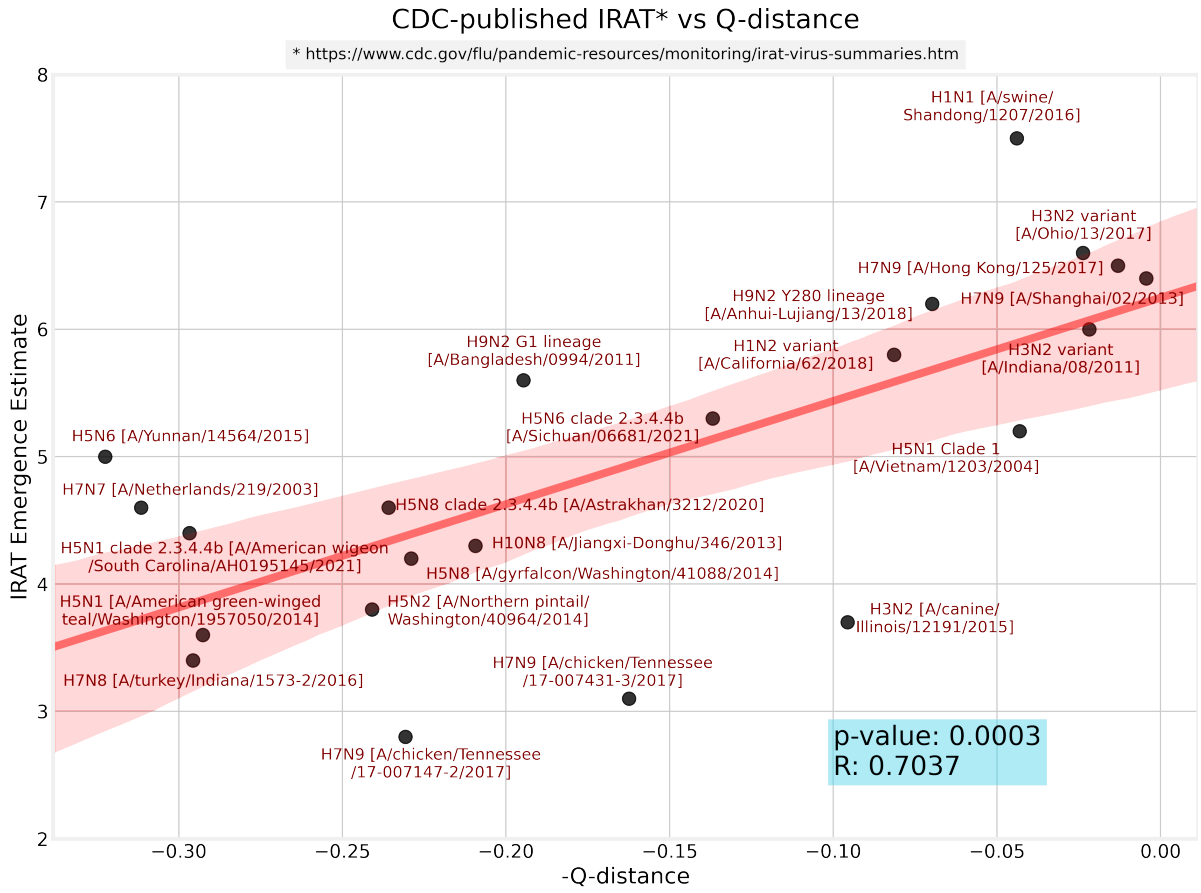


Fig. 3. **IRAT emergence risk vs. q-distance**. There is an approximate linear relationship between average q-distance from human circulating strains (averaged across both HA and NA) and IRAT emergence risk grade. Note that IRAT has released results for 23 strains to date, but only 15 are plotted on the graph. This is because the strains not pictured have less than 30 human strains of the same sub-type, so a sufficiently representative Qnet could not be trained.

3) ecology and epidemiological characteristics[12]. The IRAT score assigns a grade between 1 and 10 individually for emergence risk and public health impact to Influenza A viruses that not currently circulating among humans. Evaluating the IRAT score involves multiple experimental assays for each strain, possibly taking weeks to return the final evaluation for a single strain. This causes a scalability bottleneck: while the current global biosurveillance efforts annually collect thousands of sequences, most of these sequences will never be analyzed in time. IRAT assessment protocols are not fast enough to leverage the full capacity of current surveillance output, and thus have low odds of successfully preempting a pandemic.

#2. algo intro 1

In this study, we introduce a pattern recognition algorithm to automatically parse out emergent evolutionary constraints operating on Influenza A viruses in the wild. These constraints help us to numerically estimate the probability of a strain x spontaneously giving rise to strain y in the wild, *i.e.*, estimate $Pr(x \rightarrow y)$, which then allows us to approximate IRAT scores of non-human strains without direct experimental assays, and forecast dominant strains of seasonal epidemics. To uncover relevant evolutionary constraints, we analyze observed variations (point substitutions and indels) of the amino acid (AA) sequences of the two key proteins implicated in cellular entry and exit[13], namely HA and NA respectively. By representing these constraints within a predictive framework – the Emergenet– we estimate the odds of a specific mutation to arise in a given strain, and consequently the probability of a specific strain spontaneously giving rise to another. Such explicit calculations are difficult without first inferring the emergent constraints at play, as well as the variation of mutational probabilities and the potential residue replacements from one positional index to the next along the AA sequence of a protein. The many well-known classical DNA substitution models[14] or standard approaches to phylogenetic tree inference which assume a constant mutation rate or some variation thereof, do not address these issues, and are not applicable to the problem at hand.

The dependencies we uncover are shaped by a functional necessity of conserving/augmenting fitness. 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} to 10^{-4}) during replication[15,16]. However, few of these variations are actually viable, with only specific patterns maintaining/gaining fitness; leading to emergent dependencies between such changes. Furthermore, these fitness constraints are not time-invariant. The background distribution of strains, and selection pressure from the evolution of cytotoxic T lymphocyte epitopes[17–21] 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 track these

TABLE 1
Influenza A Strains Evaluated by IRAT and Corresponding Qnet Computed Risk Scores

Influenza Virus	Subype	IRAT Date	IRAT Emer-gence Score	IRAT Impact Score	HA Qnet Sample	NA Qnet Sample	HA Avg. Qdist	NA Avg. Q-dist.	Geom. Mean
A/swine/Shandong/1207/2016	H1N1	Jul 2020	7.5	6.9	1000	1000	0.0941	0.0205	0.0440
A/Ohio/13/2017	H3N2	Jul 2019	6.6	5.8	1000	1000	0.0184	0.0306	0.0238
A/Hong Kong/125/2017	H7N9	May 2017	6.5	7.5	437	437	0.0296	0.0058	0.0131
A/Shanghai/02/2013	H7N9	Apr 2016	6.4	7.2	178	178	0.0055	0.0036	0.0044
A/Anhui-Lujiang/39/2018	H9N2	Jul 2019	6.2	5.9	31	30	0.0290	0.1681	0.0698
A/Indiana/08/2011	H3N2	Dec 2012	6	4.5	1000	1000	0.0523	0.0091	0.0218
A/California/62/2018	H1N2	Jul 2019	5.8	5.7	55	55	0.1089	0.0610	0.0815
A/Bangladesh/0994/2011***	H9N2	Feb 2014	5.6	5.4	-1	-1	0.2078	0.1823	0.1947
A/Sichuan/06681/2021	H5N6	Oct 2021	5.3	6.3	45	45	0.3616	0.0518	0.1369
A/Vietnam/1203/2004	H5N1	Nov 2011	5.2	6.6	258	246	0.1673	0.0111	0.0430
A/Yunnan/14564/2015**	H5N6	Apr 2016	5	6.6	344	331	0.3482	0.2987	0.3225
A/Astrakhan/3212/2020**	H5N8	Mar 2021	4.6	5.2	381	365	0.1603	0.3472	0.2359
A/Netherlands/219/2003	H7N7	Jun 2012	4.6	5.8	46	46	0.2757	0.3521	0.3115
A/American wigeon/South Carolina/AH0195145/2021	H5N1	Mar 2022	4.4	5.1	335	323	0.1722	0.5114	0.2967
A/Jiangxi-Donghu/346/2013***	H10N8	Feb 2014	4.3	6	-1	-1	0.20878	0.2101	0.2094
A/gyrfalcon/Washington/41088/2014**	H5N8	Mar 2015	4.2	4.6	341	328	0.1532	0.3424	0.2290
A/Northern pintail/Washington/40964/2014**	H5N2	Mar 2015	3.8	4.1	341	328	0.1529	0.3799	0.2410
A/canine/Illinois/12191/2015	H3N2	Jun 2016	3.7	3.7	1000	1000	0.0607	0.1509	0.0957
A/American green-winged teal/Washington/1957050/2014	H5N1	Mar 2015	3.6	4.1	326	314	0.1911	0.4482	0.2927
A/turkey/Indiana/1573-2/2016**	H7N8	Jul 2017	3.4	3.9	495	494	0.1130	0.7738	0.2957
A/chicken/Tennessee/17-007431-3/2017	H7N9	Oct 2017	3.1	3.5	496	495	0.1027	0.2569	0.1624
A/chicken/Tennessee/17-007147-2/2017	H7N9	Oct 2017	2.8	3.5	496	495	0.2095	0.2541	0.2307
A/duck/New York/1996 *	H1N1	Nov 2011	2.3	2.4	1000	1000	-1	-1	-1

* -1 indicates missing data, either from lack of human sequence data available for that virus sub-type (less than 30 strains) or missing IRAT sequence data (in the case of A/duck/New York/1996)

constraints, automatically reflecting the effect of evolving host immunity, and the current background strain distribution.

#3. Add details replacing XXX

We infer the Emergenet using the AA sequences of HA and NA proteins from all unique Influenza A strains in the NCBI and GISAID databases between the years XXX to present time (2022 April), which leads to a set of 98XXX strains in total. We construct Emergenets separately for H1N1 and H3N2 subtypes, and for each flu season, constructing in total XXX models. Structurally, an Emergenet comprises an interdependent collection of local predictors: each aiming to model the observed amino acid "outcome" at a specific positional index of the proteins using as features (input variables) the residues appearing at other locations (Fig. 1b). The algorithm automatically identifies the set of features (AA positions) that influence the outcome at a particular index, implying that an Emergenet comprises atmost as many such position-specific predictors as the length of the AA sequence. These individual predictors are implemented as conditional inference trees[11], in which nodal splits have a minimum pre-specified significance in differentiating the child nodes. Inferring residue predictors at each index as functions of the rest of the AA sequence essentially estimates the conditional residue distribution at each index. The set of indices acting as features in each tree varies; for example, as shown in the fragment of the H1N1 HA Emergenet (2020-2021) in Fig 1b, the predictor for index 63 is dependent on the AA residue at index 155, and the predictor for index 155 is dependent on index 223, the predictor for index 223 is dependent on index 14, and the residue at index 14 is dependent on index 63, revealing a cyclic dependency. The complete Emergenet harbors a vast number of such relationships, wherein each internal node of a tree may be recursively "expanded" to its own tree. Due to this recursive expansion, a complete Emergenet is hard to visualize, nevertheless it captures the complexity of the rules guiding evolutionary change substantially better than

earlier attempts, as evidenced by our validation results.

#4. WHy dont we use phenotypic information

As described, for Emergenet inference we use AA sequences with no additional phenotypic annotation, other than identifying the host animal (and time and place of collection). This is advantageous, since antigenic characterization of Influenza A strains tend to be substantially laborious and low-throughput compared to genome sequencing[22]; however, incorporation of phenotypic information, $e.g.$ from deep mutational scanning assays, has been shown to improve prediction of seasonal strains[6]. Despite limiting ourselves to only genotypic information our approach is able to distill deep emergent interdependencies that uncovers a rich structure of fitness-preserving constraints that then outperform reported phenotypic information augmented strategies.

Inference of the Emergenet components is the first step in our approach, which induces the definition of a new distance metric between genomic sequences. The E-distance (Eq. (5) in Methods) is defined as the square-root of the Jensen-Shannon (JS) divergence[23] of the position-specific conditional distributions, averaged over the entire sequence length. Unlike the classical edit distance measuring the number of mutations that the sequences differ by, the E-distance is informed by the dependencies inferred by the Emergenets, and adapts to the specific organism, allele frequencies, and variations in the background population. Central to our approach is the theoretical result (Theorem 1 in Supplementary text) that $Pr(x \to y)$ is bounded above and below by simple exponential functions of the E-distance $\theta(x, y)$. The mathematical intuition behind relating E-distance to the change-probability $Pr(x \to y)$ is similar to the prediction of a biased outcome when we toss a fair coin. With an overwhelming probability, a sequence of such tosses should result in roughly equal number of heads and tails. However, large deviations, $i.e$, substantial deviations from the expected fraction do happen, and the probability of such rare events is explicitly quantifiable[24] with elementary results from large deviation theory. Generalizing to non-uniform conditional probabilities inferred by the Emergenet, we show that the likelihood of a spontaneous transition $Pr(x \to y)$ by random chance may also be similarly bounded. As an important distinction from the edit distance, the E-distance between two fixed sequences may change even if only the background population changes (SI-Table 1, example where the distance between two fixed sequences vary when we vary their collection years, and hence the environment or the background strain distribution). Thus, we can not only estimate the risk of emergence of a particular animal strain, but can track how that risk evolves over time [Figure ref XXX].

The ability of the Emergenet framework to determine the numerical odds of spontaneous jump $Pr(x \to y)$ (Fig. 1) suggests that we are able to frame the problem of forecasting dominant strain(s), and that of estimating the pandemic potential of an animal strain as precise mathematical propositions (albeit with some simplifying assumptions), with demonstrable approximate solutions (Fig. 1c-d). Thus, a dominant strain for an upcoming flu season may be identified as one which maximizes the joint probability of simultaneosly 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 strain. 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. We validate our proposed solutions for these problems in out-of-sample data. In the context of the first problem namely the forecast of dominant strain(s) for the next flu season, we obtain the following search problem (See Methods), to identify a historical strain that is expected to be close to the upcoming season's dominant strain:

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

Prediction of the future dominant strain as a close historical strain allows direct validation of the approach with past WHO recommendations. Notably, the flu shot is annually prepared at least six months in advance, and is based on a cocktail of historical strains determined by the WHO via global surveillance[25], hoping to match the circulating strain(s) in the upcoming season (recommendations for the northern hemisphere are given in February, while that for the southern hemisphere are given at the end of December the previous year). For each year of the past two decades, we calculated strain forecasts using Eq. (10) with data available 6 months before the target season. We measured forecast performance by the number of mutations by which the predicted HA/NA AA sequence sequence deviated from the realized dominant strain, which we approximated as the one closest to the centroid of the observed strains in the target season in the sense of the edit distance $i.e.$ number of mutations. 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. 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 Text in Tables SI-Table 4 through SI-Table 16. 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[26], strain-matching is a crucial component to realizing high vaccine effectiveness[27]. Thus, our results may improve the effectiveness of the flu shot via sophisticated pattern-recognition, outperforming current practice (WHO/CDC) as well as recently reported prediction strategies using more standard computational and/or experimental frameworks[6,9], without using detailed phenotypic information such as deep mutational scanning data[6,9].

#5. concrete numbers

Our primary claim, however, is the ability to estimate the pandemic potential of novel animal strains, via our proposed

risk score $\rho(x)$ for a strain x not yet found to circulate in human hosts. We show that the following measure:

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

scales as the average log-likelihood of $Pr(x \to y)$ where y is any human strain of a similar subtype to x. As before, the Emergenet inference makes it possible to estimate $\rho(x)$ explicitly. For each strain previously analyzed by IRAT, we construct Emergenet models for HA and NA segments using all human strains of the same subtypes circulating in the year prior to risk assessment. For example, the A/swine/Shandong/1207/2016 strain was assessed by IRAT in July 2020, so we use human H1N1 strains collected between July 1, 2019 - June 30, 2020. For sub-types with few recorded human strains (H1N2, H5N1, H5N6, H7N7, H9N2), we consider all human strains of the corresponding subtypes collected upto the IRAT evaluation date. We then compute the average E-distance between a given animal strain and the circulating human strains for both HA and HA segments (using Eq. (2)), with finally reporting the geometric mean of the estimates as the estimated risk. Our analysis shows that this risk scales as the average log-likelihood of jumping to a similar human strain, and thus is expected to correlate with the IRAT emergence score. Considering IRAT scores of 22 strains published by the CDC, we find strong support (correlation of -0.7032 ($p < 0.005$), Fig. 3) for this claim. Importantly, the risk score is computable in seconds as opposed to weeks taken by IRAT experimental assays, and this dramatic reduction in time and cost opens the door to fully exploiting the current surveillance capacity.

Salient points to be noted about the IRAT approximation approach are: 1) risk is a function of time; the background distribution of strains matters, and we show that recomputing the risks for a different time changes (worsens) the correlation with the IRAT estimate. We also show how the risks evolve over time according to our model (See Table **??** and SI-Fig. 3) 2) We train a linear model to estimate the IRAT emergence score from the risk values. Also, since the IRAT impact score is strongly correlated with the emergence score (correlation=XXXX), we trained a linear model to estimate the impact score as well from the risk. We show these results in Table **??**. 3) Available IRAT scores are well approximated within our framework, albeit with four outliers. For three of these, the approximation would be better if IRAT had underestimated their risk, which include one H3N2 CIV and two H7N9. The zoonotic potential of H3 CIV is unknown[28]. Since canines has been proven to be susceptible to infection with at least three different influenza A subtypes (H3N8, H3N2 and H1N1)[29–31] and can also be infected with diverse human influenza viruses, these animals have the potential to act as mixing vessels in which novel viruses are generated by reassortment. Compared to strains in which the full genome has evolved in a non-human host, such reassortant viruses may more easily overcome the species barrier to cause an outbreak in humans[32]. Thus, the risk for the 2015 H3N2 IL CIV might have been underestimated by IRAT. Risk from the H7N9 strains could also have been underestimated. To date, a total of 1,568 laboratory-confirmed human infections with avian influenza A(H7N9) virus including 616 fatal cases have been reported to WHO since early 2013. The last case of human infection with avian influenza A(H7N9) reported to WHO in the Western Pacific Region was in 2019. Of the 1,568 human infections with avian influenza A(H7N9), 33 have reported mutations in the hemagglutinin gene indicating a change to high pathogenicity in poultry, and while the current virus does not transmit easily among humans, the WHO recognizes the need for close monitoring to identify changes in the virus and transmission behaviourhttps://www.who.int/docs/default-source/wpro---documents/emergency/surveillance/avian-influenza/ai_20220930.pdf?sfvrsn=22ea0816_18. This might be corrobortated by the fact that the risk from our analysis increase when recomputed with current background, rather than when the IRAT scores were generated (E-distance has decreases). Similar arguments apply for the H1N1 swine strain (A/swine/Shandong/1207/2016) that was given the highest IRAT risk score. Our analysis identifies this as an outlier, and suggest the risk was overestimated, which might again be corroborated by the decrease in the risk when recomputed with current background.

#6. we can have a linera estimation of both emergenec score and impact score

With our liner models trained, we compute a simulated IRAT analysis of all Influenza A strains collected over past two years (XXXXX sequences in total). (Table **??** shows the top 5 strains of each available subtype). The geolocation of the strains, along with risk information is shown in Fig. **??**.

Summary & Conclusion: Numerous tools exist for ad hoc quantification of genomic similarity[9,14,33–36], for which a smaller distance (higher similarity) does not necessarily imply that a feasible trajectory exists from one strain to the other in the wild, or that the likelihood of a jump is high. These measures tend to be variations of edit distances between sequences, and are not aware of selection pressures and evolutionary dynamics. Despite the diverse techniques explored in these domains, the missing piece is effectively learning which changes are likely in the wild, conditioned on possibly the entire sequence of the current strain. Our algorithm is arguably the first of its kind to learn an appropriate 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 lense , can parse out predictive models of these complex interactions. Our results are well aligned With emerging studies making clear that predicting future mutations are indeed feasible for different organisms[37,38] with the correct approach.

Calculation of E-distance is currently limited to similar sequences (such as 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 Text, SI Table 20), which would exclude applicability to completely novel pathogens, 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 premeptive vaccines to target risky strains in the animal hosts before the first

human infection in the next pandemic.

METHODS

Next, we briefly describe the details of the 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[39], 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. The mathematical form of our metric is not arbitrary; JS divergence is a symmetricised version of the more common KL divergence[23] between distributions, and among different possibilities, the E-distance is the simplest metric such that the likelihood of a spontaneous jump (See Eq. (6) in Methods) is provably bounded above and below by simple exponential functions of the E-distance.

Consider a set of random variables $X = \{X_i\}$, with $i \in \{1, \cdots, N\}$, each taking value from the respective sets Σ_i. A sample $x \in \prod_1^N \Sigma_i$ is an ordered N-tuple, 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[40] over the index set $\{1, \cdots, N\}$. To clarify the biological picture, we refer to the sample x as an amino acid or nucleotide sequence, identifying the entry at each index with the corresponding protein residue or the nucleotide base pair.

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

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

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

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

Biology-Aware Distance Between Sequences

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[41] 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[23], ensuring that the sum of the length of successive path fragments equates the length of the path, making it possible to use standard algorithms for q-phylogeny construction.

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

$$\omega_y^Q e^{\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \geqq Pr(x \to y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \tag{6}$$

where ω_y^Q is the membership probability of strain y in the target population.

The ability to estimate the probability of spontaneous jump between sequences in terms of θ has crucial implications. It allows us to 1) construct a new phylogeny that directly relates the probability of jumps rather than the number of mutations between descendants. 2) simulate realistic trajectories in the sequence space from any given initial strain, and 3) estimate drift in the sequence space by analyzing the statistical characteristics of the diffusion occurring in the strain space.

Application 1: Predicting Seasonal Strains

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

Let $x_\star^{t+\delta}$ be a dominant strain in the upcoming flu season at time $t+\delta$, where H^t is the set of observed strains presently in circulation in the human population (at time t). We will assume that the Emergenet remains unchanged upto $t+\delta$. From the RHS bound established in Theorem 1 (See Eq. (6) above) in the supplementary text, we have:

$$\ln \frac{Pr(x \to x^{t+\delta})}{\omega_{x^{t+\delta}}} \geqq -\frac{\sqrt{8}N^2}{1-\alpha}\theta(x, x^{t+\delta}) \tag{7}$$

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

$$\Rightarrow \sum_{x\in H^t} \theta(x, x^{t+\delta}) - \left|H^t\right| A \ln \omega_{x^{t+\delta}} \geqq A \ln \frac{1}{\prod_{x\in H^t} Pr(x \to x^{t+\delta})} \tag{9}$$

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

$$x_\star^{t+\delta} = \underset{y\in\cup_{\tau\leq t}H^\tau}{\arg\min} \sum_{x\in H^t} \theta(x, y) - \left|H^t\right| A \ln \omega_y \tag{10}$$

Application 2: Measure of Pandemic Potential

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

$$\rho(x_a^t) \triangleq -\frac{1}{\left|H^t\right|} \sum_{x\in H^t} \theta(x_a^t, x) \tag{11}$$

The intuition here is that a lower bound of $\rho(x_a^t)$ scales as average log-likelihood of the x_a^t giving rise to a human strains in circulation at time t. Since the strains in H^t are already HH capable, a high average likelihood of producing a similar strain has a high potential of being a HH cabale novel variant, which is a necessary condition of a pandemic strain. To establish the lower bound, we note that from Theorem 1 (See Eq. (6) above) in the supplementary text, we have:

$$\sum_{y\in H^t} \ln \left|\frac{Pr(x_a^t \to y)}{\omega_y}\right| \leqq -\frac{\sqrt{8}N^2}{1-\alpha} \left|H^t\right| \rho(x_a^t) \tag{12}$$

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

$$\Rightarrow \rho(x_a^t) \geqq A \ln \left(\prod_{y\in H^t} Pr(x_a^t \to y) \right)^{1/\left|H^t\right|} + C \tag{13}$$

$$\Rightarrow \rho(x_a^t) \geqq A \ln \left\langle Pr(x_a^t \to x_h^{t+\delta}) \right\rangle + C \tag{14}$$

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

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[42]. We use two sequences

databases: 1) National Center for Biotechnology Information (NCBI) virus[43] and 2) GISAID[44] 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 3).

REFERENCES

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

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

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

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

[5] Dos Santos, G., Neumeier, E. & Bekkat-Berkani, R. Influenza: Can we cope better with the unpredictable? *Human vaccines & immunotherapeutics* **12**, 699–708 (2016).

[6] Huddleston, J. *et al.* Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza a/h3n2 evolution. *Elife* **9**, e60067 (2020).

[7] Boni, M. F. Vaccination and antigenic drift in influenza. *Vaccine* **26**, C8–C14 (2008).

[8] Tricco, A. C. *et al.* Comparing influenza vaccine efficacy against mismatched and matched strains: a systematic review and meta-analysis. *BMC medicine* **11**, 153 (2013).

[9] Neher, R. A., Russell, C. A. & Shraiman, B. I. Predicting evolution from the shape of genealogical trees. *Elife* **3**, e03568 (2014).

[10] Vergara-Alert, J. *et al.* The ns segment of h5n1 avian influenza viruses (aiv) enhances the virulence of an h7n1 aiv in chickens. *Veterinary research* **45**, 1–11 (2014).

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

[12] 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).

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

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

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

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

[17] 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).

[18] 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).

[19] 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).

[20] 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).

[21] 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).

[22] 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).

[23] Cover, T. M. & Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, New York, NY, USA, 2006).

[24] 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).

[25] 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).

[26] 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).

[27] 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).

[28] Martinez-Sobrido, L. *et al.* Characterizing emerging canine h3 influenza viruses. *PLoS pathogens* **16**, e1008409

(2020).
[29] Chang, C., New, A., Taylor, J. & Chiang, H. Influenza virus isolations from dogs during a human epidemic in taiwan. Tech. Rep., NAVAL MEDICAL RESEARCH UNIT NO 2 MANILA (PHILIPPINES) (1976).
[30] Houser, R. & Heuschele, W. Evidence of prior infection with influenza a/texas/77 (h3n2 (virus in dogs with clinical parainfluenza. *Canadian Journal of Comparative Medicine* **44**, 396 (1980).
[31] Chen, Y. *et al.* Emergence and evolution of novel reassortant influenza a viruses in canines in southern china. *MBio* **9**, e00909–18 (2018).
[32] Lowen, A. C. Constraints, drivers, and implications of influenza a virus reassortment. *Annual review of virology* **4**, 105–121 (2017).
[33] 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).
[34] Huelsenbeck, J. P. & Crandall, K. A. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and systematics* **28**, 437–466 (1997).
[35] 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).
[36] Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125 (2009).
[37] Mollentze, N., Babayan, S. A. & Streicker, D. G. Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS biology* **19**, e3001390 (2021).
[38] Maher, M. C. *et al.* Predicting the mutational drivers of future sars-cov-2 variants of concern. *Science Translational Medicine* **14**, eabk3445 (2022).
[39] 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).
[40] Vanmarcke, E. *Random fields: analysis and synthesis* (World scientific, 2010).
[41] Manning, C. D., Manning, C. D. & Schütze, H. *Foundations of statistical natural language processing* (MIT press, 1999).
[42] McAuley, J., Gilbertson, B., Trifkovic, S., Brown, L. E. & McKimm-Breschkin, J. Influenza virus neuraminidase structure and functions. *Frontiers in microbiology* **10**, 39 (2019).
[43] Hatcher, E. L. *et al.* Virus variation resource–improved response to emergent viral outbreaks. *Nucleic acids research* **45**, D482–D490 (2017).
[44] Bogner, P., Capua, I., Lipman, D. J. & Cox, N. J. A global initiative on sharing avian flu data. *Nature* **442**, 981–981 (2006).
[45] Rulli, M. C., Santini, M., Hayman, D. T. & D'Odorico, P. The nexus between forest fragmentation in africa and ebola virus disease outbreaks. *Scientific reports* **7**, 41613 (2017).
[46] Chua, K. B., Chua, B. H. & Wang, C. W. Anthropogenic deforestation, el niiio and the emergence of nipah virus in malaysia. *Malaysian Journal of Pathology* **24**, 15–21 (2002).
[47] Childs, J. Zoonotic viruses of wildlife: hither from yon. In *Emergence and Control of Zoonotic Viral Encephalitides*, 1–11 (Springer, 2004).
[48] Fair, J. & Fair, J. Viral forecasting, pathogen cataloging, and disease ecosystem mapping: Measuring returns on investments. *Current Topics in Microbiology and Immunology* 75–83 (2019).
[49] Cdc vaccine effectiveness studies (2020). URL https://www.cdc.gov/flu/vaccines-work/effectiveness-studies.htm.
[50] Carugo, O. & Pongor, S. A normalized root-mean-spuare distance for comparing protein three-dimensional structures. *Protein science* **10**, 1470–1473 (2001).
[51] Righetto, I., Milani, A., Cattoli, G. & Filippini, F. Comparative structural analysis of haemagglutinin proteins from type a influenza viruses: conserved and variable features. *BMC bioinformatics* **15**, 363 (2014).
[52] Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology* **55**, 379–IN4 (1971).
[53] Shrake, A. & Rupley, J. A. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of molecular biology* **79**, 351–371 (1973).
[54] Momen-Roknabadi, A., Sadeghi, M., Pezeshk, H. & Marashi, S.-A. Impact of residue accessible surface area on the prediction of protein secondary structures. *Bmc Bioinformatics* **9**, 357 (2008).
[55] Adamczak, R., Porollo, A. & Meller, J. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins: Structure, Function, and Bioinformatics* **59**, 467–475 (2005).
[56] Tzarum, N. *et al.* Structure and receptor binding of the hemagglutinin from a human h6n1 influenza virus. *Cell host & microbe* **17**, 369–376 (2015).
[57] Lazniewski, M., Dawson, W. K., Szczepińska, T. & Plewczynski, D. The structural variability of the influenza a hemagglutinin receptor-binding site. *Briefings in functional genomics* **17**, 415–427 (2018).
[58] Garcia, N. K., Guttman, M., Ebner, J. L. & Lee, K. K. Dynamic changes during acid-induced activation of influenza hemagglutinin. *Structure* **23**, 665–676 (2015).

DISCUSSION

In the aftermath of the COVID-19 pandemic that caused one of the most devastating disasters of the past century, a looming question is whether we can prepare for, preempt and mitigate such events in the future. Evolving viruses, whether currently circulating in the human population, or in animal reservoirs that might spillover and attain human-to-human transmission capability, pose an ever-present epidemic risk. Current surveillance paradigms, while crucial for mapping disease ecosystems, are limited in their ability to address this challenge. Habitat encroachment, climate change, and other ecological factors[45–47] unquestionably drive up the odds of zoonotic spill-overs. Nevertheless, current efforts at tracking these effects have not improved our ability to quantify future risk of emergence[48]. Tracking viral diversity in animal hosts, while important, often does not transparently map to emergence risk. This is particularly true for Influenza A, which partly on account of its segmented genome, can easily incorporate genes from multiple strains and emerge as novel human pathogens, and thus harbor a high pandemic potential. While large antigenic shifts in Influenza A are relatively rare, even the smaller seasonal sequence alterations in cause sufficient variation in the surface proteins to evade existing immunity, and require yearly reformulation of the flu vaccine.

However, for the vaccine to be effective, we need to predict the dominant circulating strain of the upcoming season with sufficient accuracy. Currently, the composition of the flu shot is decided at least six months in advance of the seasonal infection peak, and targets three to four historical strains as recommended by the CDC/WHO, who identify these specific strains by sampling the current circulation[25], hoping to match the dominant strain(s) in the upcoming season. A variety of hard-to-model effects hinder this prediction, which, despite observed cross-reactive effects[8], have had limited vaccine effectiveness in recent years[49]. Rank-ordering strains which do not yet circulate in humans according to either their spillover risk or their pandemic potential, has proven to be even more difficult [REF]. CDC's current, somewhat subjective, solution to this problem is the Influenza Risk Assessment Tool (IRAT), which uses a combination of ten weighted risk elements, including 1) properties of the virus, 2) attributes of the population, and 3) ecology and epidemiological characteristics of the virus[12] that are expert-selected. Evaluating these factors involve several experimental assay for each strain, taking possibly weeks to return the final IRAT score for a single strain. Thus, we have a scalability problem: with the current global biosurveillance efforts collecting tens of thousands of sequences every year, IRAT assessment is simply not fast enough to preempt a pandemic.

DISCUSSION & SEQUENCE COMPARISONS

For further discussion, we looked at our Qnet predictions more closely. Comparing the Qnet inferred strain (QNT) against the one recommended by the WHO, we find: 1) the residues that only the QNT matches correctly with DOM (while the WHO fails) are largely localized within the receptor binding domain (RBD), with $> 57\%$ occurring within the RBD on average (see Fig. 1a for a specific example), and 2) when the WHO strain deviates from the QNT/DOM matched residue, the "correct" residue is often replaced in the WHO recommendation with one that has very different side chain, hydropathy and/or chemical properties (see Fig. 1b-f), suggesting deviations in recognition characteristics. Combined with the fact that we find circulating strains are almost always within a few edits of the DOM (see SI-Fig. 2), these observations suggest that hosts vaccinated with the QNT recommendation is more likely to have season-specific antibodies that are more likely to recognize a larger cross-section of the circulating strains.

High season-to-season genomic variation in the key Influenza capsidic proteins is driven by two opposing influences: 1) the need to conserve function limiting random mutations, and 2) hyper-variability to escape recognition by neutralizing antibodies. Even a single residue change in the surface proteins might dramatically alter recognition characteristics, brought about by unpredictable[50,51] changes in local or regional properties such as charge, hydropathy, side chain solvent accessibility[52–55].

Focusing on the average localization of the QNT to WHO deviations in the HA molecular structure, the changes are observed to primarily occur in the HA1 sub-unit (see Fig. 1g-i, HA0 numbering used, other numbering conversions are given in SI-Table 21), with the most frequent deviations occurring around the ≈ 200 loop, the ≈ 220 loop, the ≈ 180 helix, and the ≈ 100 helix, in addition to some residues in the HA2 sub-unit (≈ 49 & ≈ 124). Unsurprisingly, the residues we find to be most impacted in the HA1 sub-unit (the globular top of the fusion protein) have been repeatedly implicated in receptor binding interactions[56–58]. Thus, we are able to fine tune the future recommendation over the state of the art, largely by modifying residue recommendations around the RBD and structures affecting recognition dynamics.

It is well known that the influenza viral RNA-polymerase represents the lack of proofreading function. Thus, the integration of faulty nucleotides often occurs during the viral replication process with a rate of 10^{-3} to 10^{-4}, which results in high mutation rates [39,40].

Due to its crucial role in receptor recognition and attachment, IAV HA is considered to be a principal determinant of the host-range. The specificity of the HA of avian influenza viruses is for $\alpha - 2,3$ SA receptors found in the intestinal tract of the bird, whereas $\alpha - 2,6$ SA receptors are predominantly found in the upper respiratory tract of humans. Recently, it has been shown that mutations in the HA protein alter its receptor-binding preference that allows the highly pathogenic avian H5N1 IAV to transmit between mammals [41]. Therefore, it is not surprising that multiple changes in gene segments of the avian influenza virus could result in its adaptation to humans [1]. On the other hand, owing

to having both $\alpha - 2,3$ and $\alpha - 2,6$ linkages, pigs and several avian species (pheasants, turkeys, quails) may act as mixing vessels and can generate re-assortment viruses [42,43].

Influenza proteins must evade immune recognition while maintaining their ability to function and interact with host cellular factors [44]. The three mechanisms by which influenza viruses undergo evolutionary change include mutation (antigenic drift), re-assortment (antigenic shift), and, in rare instances, recombination. The different virus lineages are predominantly host specific, but there are periodic exchanges of influenza virus gene segments between species, giving rise to pandemics of disease in humans, lower animals, and birds [45]. Influenza virus evolution proceeds via re-assortment and mutation, and such evolution can influence the host specificity and pathogenicity of these viruses [46]. Genetic variations of influenza A virus lead to possible changes in upcoming epidemiological behavior and may result in human pandemics.

Significant mutations in antigenic sites resulting from constant point mutations in the influenza virus contribute to the gradual evolution of the virus, leading to antigen migration to produce new influenza virus subtypes to escape the immune pressure of the population [47]. All subtypes of influenza A virus antigenic drift can occur, but such antigenic drift often occurs in the general human influenza. Immune escape can be achieved by mutation in IAV proteins such as HA and/or NA. The minimal structural changes can occur in these surface proteins and so the immune protection of the host (acquired through previous infections or immunization) will no longer be effective against the invading virus. As a consequence, the immune system is unable to identify the newly changed virus variants and the recognition pattern of the antigen-antibody-interaction is not fully functional anymore. In addition, amino acid substitutions in HA protein can change the receptor preference of influenza virus. Some studies have shown that the G186V mutation in HA protein was noted as a potential adaptation of avian H7 to human-type receptors [48,49]. In A/Vietnam/1203/2004 (H5N1) virus, K58I substitution in HA protein is associated with increased viral replication of upper respiratory tracts in mice and ferrets [50]. Remarkably, the K58I substitution combined with a G219S mutation in HA protein increased the overall affinities of binding to $\alpha - 2,3$ and $\alpha - 2,6$ SA of the A/Anhui/1/13 (H7N9) virus [51]. Furthermore, there is a R292K mutation in NA protein in H7N9 virus strains which had been isolated from a patient after drug treatment. This substitution was found to promote drug resistance; in particular, it gave a high resistance to oseltamivir which is the most commonly used anti-influenza drug [52]. Antigenic drifts are the main reason for new variants and cause annual influenza outbreaks. Although these changes may not lead to pandemics, antigenic drift over a period of time can make a strain considerably different from the original pandemic virus.

It has been confirmed that the long-term evolution of cytotoxic T lymphocyte (CTL) epitopes is associated with CTL-mediated clearance of infection and it is thought that the selection pressures imposed by CTL immunity shape the long-term evolution of IAV [53,54]. Viruses mutate amino acid residues within CTL epitopes to evade CTL recognition [55]. Under certain circumstances, amino acid substitutions occur at the anchoring residues, while in other cases they occur at the T cell receptor contact residues [56]. For instance, mutations at the anchored residues of the CTL epitope have been described in the human leukocyte antigen (HLA)-B* 2705 restricted NP383–391 epitope, which has the R-to-G substitution at position 384 (R384G) [57,58]. This replacement significantly reduced the in vitro virus-specific CTL response in HLA-B* 2705-positive individuals.

4.2. Re-Assortment It has been well recognized that the segmented genome of the influenza virus allows the exchange of RNA segments between genotypically different influenza viruses, resulting in the production of new strains and/or subtypes [67], which is referred to as re-assortment. A pandemic IAV can be produced by transmission from animals to humans or by reconfiguration between avian influenza viruses and human influenza viruses [68]. As the influenza virus has a segmented genome, re-assortment is an important mechanism for generation of the ”novel” virus [69]. Thus, re-assortment of the virus achieves a new antigenic pattern known as ”antigenic shift”. Pandemic influenza emerges as a result of such major genetic changes of IAV. These modifications occur due to mechanistic errors during the replication of viral RNA polymerase, evolutionary pressure, the novel environment of the host, immune pressure, or antiviral drug pressure [70]. Two of the three major human influenza pandemics in the twentieth century (1957 and 1968) and this century (2009) were due to the re-assortment between the human IAV and other host species.

There is evidence indicating that the HA, NA, and PB1 genes of the H2N2 1957 pandemic strain in addition to the HA and PB1 fragments of the H3N2 1968 pandemic strain are both avian, and the remaining fragments may come directly from 1918 [67]. The first influenza pandemic in this century, the influenza A H1N1 virus, is a re-assortant caused by a multiple mixed recombination between the European H1N1 swine influenza virus, North American H1N2 swine influenza virus, North American avian influenza virus, and H3N2 influenza virus [71].

In addition to mutation and re-assortment, IAVs still have another relatively rare means of evolution called recombination. Genetic recombination is one of the primary processes that produce the genetic diversity upon which natural selection acts. Recombination in IAVs can occur through two main mechanisms: one is the non-homologous recombination that occurs between two different RNA fragments [81,82]; the other is the controversial homologous recombination, often considered to be absent or very rare, which is thought to participate in template switching while the polymerase is copying the RNA.

Wild waterfowl and shorebirds belong to the main natural host species of IAV [88]. IAV has been able to establish the successful infection of a variety of animals, including avian and mammalian species, and its evolution has led to the emergence of IAV in human beings for a long time [89]. Since the pandemic outbreak of influenza virus in 1918, the

re-assortment of influenza virus has occurred among bird and human viruses. As described above, the re-assortment of influenza viruses has resulted in the pandemic of H2N2 in 1957 and of H3N2 in 1968 [90]. During the year 2009, there was an outbreak of H1N1 in humans that caused the first pandemic of influenza through human transmission in the 21st century [91].

Usually, an avian influenza subtype does not infect humans and a human influenza subtype is unable to infect the birds. However, swine acts as a virus mixer vessel, leading to the generation of new influenza viruses, which can infect both humans and poultry. The mutation and re-assortment of the IAV genome are susceptible to forming new subtypes of influenza virus that may result in widely propagated and destructive pandemics due to the lack of immunity to the emerging pathogen [67]. For example, the outbreak of H5N1 avian influenza in 1997 and the outbreak of H1N1 swine influenza in 2009 caused great panic and brought serious economic losses to the breeding industry.

#7. review more craefully pheno-typic info used in teh literature and why it is claimed to be necessary in those papers. Why dont we need it

BRIEF METHODS

A key barrier to making progress on both the problems cited above, namely predicting the dominant strain(s) in seasonal flu, as well as estimating the numerical odds of an animal strain to spillover and attain HH capability, is our limited understanding of the emergent dependencies across individual mutations that constrain evolutionary trajectories. Thus, to the best of our knowledge, the state of the art has no tools to estimate the numerical likelihood of specific mutations in the future, and in general the likelihood of a wild strain spontaneously giving rise to another by random chance. Currently, this likelihood is often qualitatively equated to sequence similarity, which is measured by the number of mutations it takes to change one strain to another. However, the odds of one sequence mutating to another is not just a function of how many mutations separate them, but also of how specific mutations incrementally affect fitness. Ignoring the constraints arising from the need to conserve function makes any assessment of the mutation likelihood open to subjective bias. Here, we show that a precise calculation is possible when sequence similarity is evaluated via a new biologically-aware metric, which we call the *q-distance*.

Some recent efforts have recognized this gap, and have attempted to predict future dominant strain by incorporating other phenotypic details.

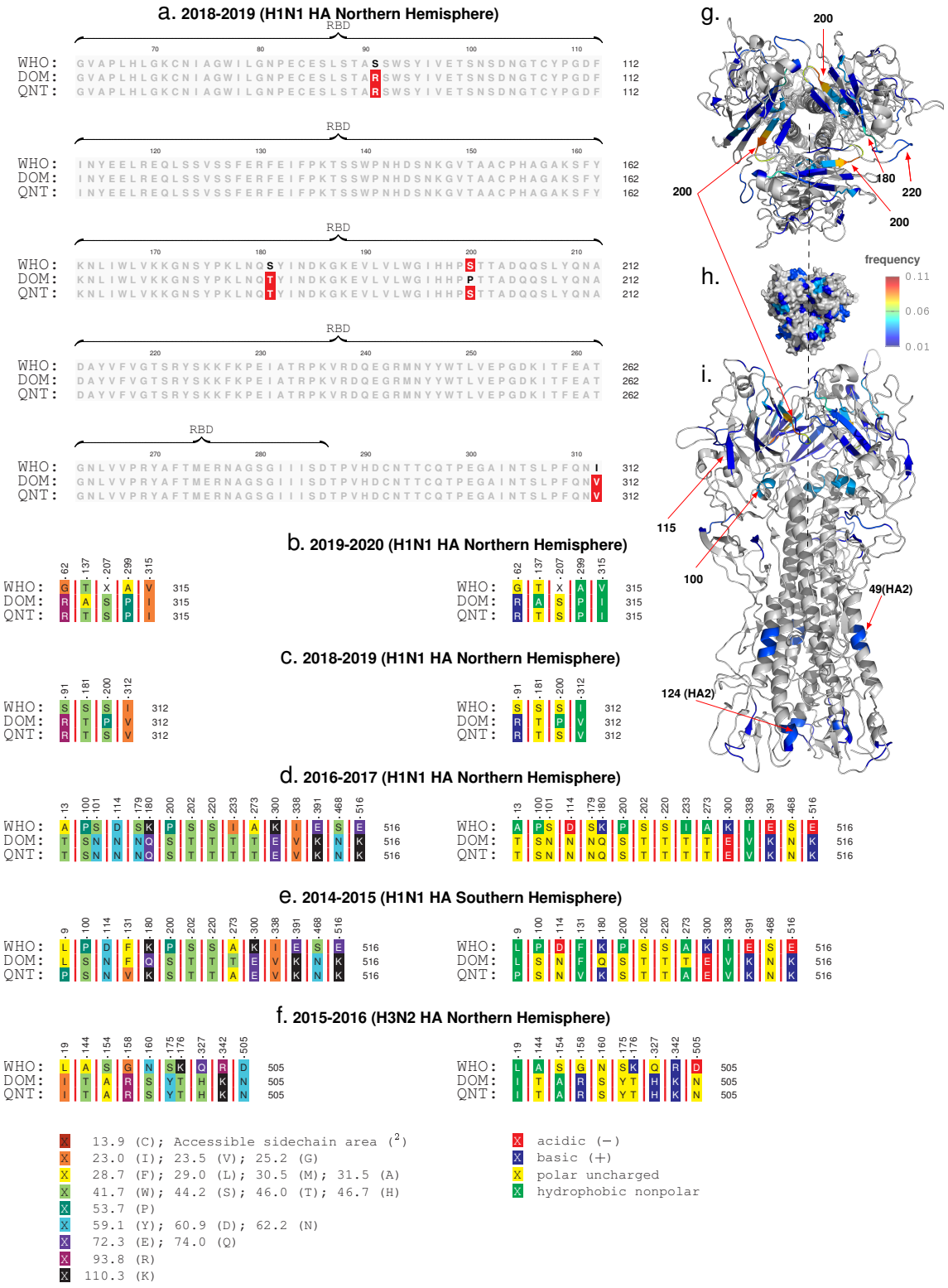
As an applications of the q-distance, we show that we can improve seasonal forecasts for the future dominant circulating strain by learning from the mutational patterns of key surface proteins: Hemaglutinnin (HA) and Neuraminidase (NA) for Influenza A. We outperform the WHO's recommendations for the flu-shot composition consistently over past two decades, measured as the number of mutations that separate the predicted from the dominant circulating strain in each season. Our recommendations repeatedly end up closer to the dominant circulating strain, illustrating the potential of our approach to correctly predict evolutionary trajectories.

We also show that this new metric allows us to assess the risk posed by novel strains effectively and quickly. We compare q-distance results to the CDC's Influenza Risk Assessment Tool (IRAT)[12], which gives a grade between 1-10 for emergence risk and public health impact to Influenza A viruses not currently circulating among humans. Our results show strong negative correlations between IRAT emergence risk grades and q-distances to the nearest human strains to the strains in question. However, while IRAT may take weeks to analyze a single strain – hence the small number of analyzed strains – q-analysis can be done within milliseconds for each new strain. Moreover, q-analysis only requires sequence data, while IRAT requires information for 10 risk elements, grouped into three categories: 1) properties of the virus, 2) attributes of the population, and 3) ecology and epidemiology of the virus[12]. Thus, our method could potentially be a low-cost, efficient substitute to IRAT, which could used at scale to rank the risk of emergence of non-circulating strains.

Discussion? Thus, the tool proposed in this study can profoundly impact bio-surveillance strategies. The ability to rank newly collected strains by risk at scale, allows actionable estimates of pandemic risks via quantifying the odds of a particular strain spilling into to the human population. Additionally, for strains already circulating in humans, our tools can estimate the odds of specific new mutan varaints emerging, and their ability to escape current vaccines.

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

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



Extended Data Figure 1. **Sequence comparisons.** The observed dominant strain, we note that the correct Qnet deviations tend to be within the RBD, both for H1N1 and H3N2 for HA (panel a shows one example). Additionally, by comparing the type, side chain area, and the accessible side chain area, we note that the changes often have very different properties (panel b-f). Panels g-i show the localization of the deviations in the molecular structure of HA, where we note that the changes are most frequent in the HA1 sub-unit (the globular head), and around residues and structures that have been commonly implicated in receptor binding interactions *e.g* the $\approx$ 200 loop, the $\approx$ 220 loop and the $\approx$ 180-helix.