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 can have large antigenic shifts, and is thought to have precipitated devastating pandemics in the past[1–4]. Yet, our current ability to scalably evaluate the pandemic potential of individual strains remains severely limited. In this study, we introduce a computational approach to parse out emergent evolutionary constraints using only genomic sequence data. A representation of these inferred dependencies along with the associated computational framework is referred to as the Emergenet. 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 computational tools estimate the likelihood of a specific future mutation, 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 outperforming World Health Organization (WHO) 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 (% 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 several 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 potentially key to fully exploiting the current biosurvellinace capacity via scalably analyzing tens of thousands of strains collected every year, thus enabling 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. Designing the optimal flu shot requires predicting the dominant strain(s) in the upcoming season, and deviations between the predicted and the circulating strain(s) reduce vaccine effectiveness[7] dramatically. Unfortunately, despite recent advances in forecasting tools[6,8], prediction of the dominant strain(s) remains imperfect.

In addition to the seasonal flu epidemic posing a serious health concern, novel Influenza A strains spilling over into humans from animal reservoirs can cause global pandemics, as demonstrated 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 prepare for, 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, has historically demonstrated its ability to easily incorporate genes from multiple strains and emerge as novel human pathogens[3,10], thus harboring a high potential of triggering the next pandemic.

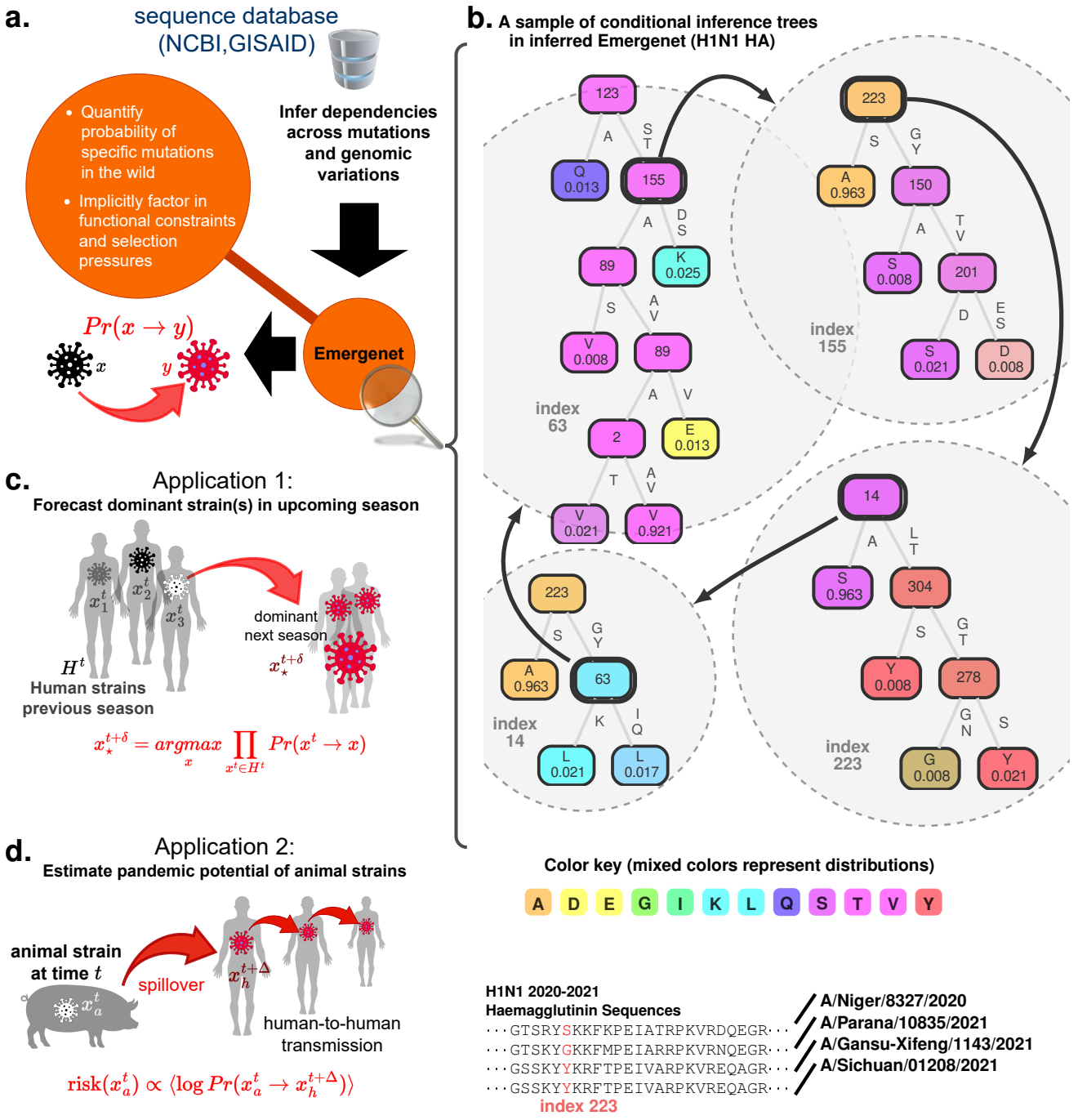

Fig. 1. **Emergenet inference and applications**. **a**, Variations of genomes for identical subgroups of Influenza A are analyzed to infer a recursive forest of decision trees (conditional inference trees[9]), the Emergenet, which maximally captures the emergent dependencies between an a priori unspecified 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.

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. Despite global surveillance efforts to collecting wild specimens from diverse hosts and geo-locations, our current ability to objectively, reliably and scalably evaluate the risk posed to humans by individual animal strains is limited.

TABLE 1
Out-performance of Qnet recommendations over WHO for Influenza A vaccine composition

Subtype	Gene	Hemisphere	Two decades (% Improvement)	One decade (% Improvement)
H1N1	HA	North	31.78	75.00
H1N1	HA	South	34.66	66.28
H1N1	HA	Average	33.22	70.64
H3N2	HA	North	38.76	42.50
H3N2	HA	South	36.72	38.67
H3N2	HA	Average	37.74	40.58
H1N1	NA	North	19.64	56.00
H1N1	NA	South	11.83	50.00
H1N1	NA	Average	15.74	53.00
H3N2	NA	North	13.92	8.57
H3N2	NA	South	11.36	15.91
H3N2	NA	Average	12.64	12.24

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 the 1) properties of the virus, 2) attributes of the population, and 3) ecology and epidemiological characteristics of the virus[11] that are largely expert-selected. The IRAT score assigns a grade between 1-10 for emergence risk and public health impact to Influenza A viruses not currently circulating among humans. However, evaluating the IRAT risk elements involve multiple experimental assays for each strain, possibly taking weeks to return the final score for a single strain. Thus, we have a scalability problem: while the current global biosurveillance efforts are collecting tens of thousands of sequences every year, 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.

In this study, we introduce a machine learning algorithm to automatically parse out emergent evolutionary constraints operating on Influenza A viruses in the wild, via analyzing observed variations (point substitutions and indels) of the amino acid (AA) sequences of the two key proteins implicated in cellular entry and exit[12], HA and NA respectively. Uncovering and representing these constraints within a predictive framework, which we refer to as the Emergenet, allows us to estimate the numerical odds of a specific mutation to arise from a given strain spontaneously in the wild. Explicit quantification of these likelihoods then yields a numerical estimate of the probability of a specific strain of interest giving rise to another. Such numerical estimates are impossible to obtain without first inferring the emergent constraints at play specific to a strain, as well as the variation of mutational probabilities from one positional index to the next along the AA sequence. Thus, the many well-known classical DNA substitution models[13] or standard approaches to phylogenetic tree inference, do not address these issues.

The dependencies between local variations that the Emergenet uncovers, arises from a functional necessity of conserving or augmenting fitness. This follows from the fact that a strain needs to be present in sufficiently high numbers in circulation to be observed and recorded, implying that the sequence data from public databases that we train with represents strains with high replicative fitness. A lack of proofreading function in influenza viral RNA-polymerase leads to the integration of faulty nucleotides during the viral replication process with a rate of 10^{-3} to 10^{-4}, which results in high mutation rates[14,15]. However, not all variations of the viral genome are equally viable. Only specific patterns of such variations can maintain or gain replicative fitness; thus leading to emergent dependencies between such mutations. Furthermore, these fitness constraints are not time-invariant. The background distribution of strains, and selection pressure from long-term evolution of cytotoxic T lymphocyte (CTL) epitopes[16–20] 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 changing constraints that shape the viral evolutionary trajectories, automatically reflecting the effect of functional constraints, the impact of evolving host immunity, and the current strain distribution.

The ability of the Emergenet framework to determine the numerical odds of specific sequence variations suggests that we might be able to frame the problem of forecasting future dominant strain(s), and that of estimating the pandemic potential of an observed animal strain as precise mathematical questions (albeit with some simplifying assumptions), with demonstrated approximate solutions (Fig. 1). 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). 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.

#1. Add details replacing XXX

We infer our models using the AA sequences of HA and NA proteins from all available 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 only consider strains for which both HA and NA sequences are available, and construct Emergenets separately for H1N1 and H3N2 subtypes, and for each flu season, constructing in total XXX Emergenets. Structurally, an Emergenet comprises an interdependent forest of local predictors: each such predictor aims to model

#2. add details and example of structure, fixing fig 1 to AA from ATGC

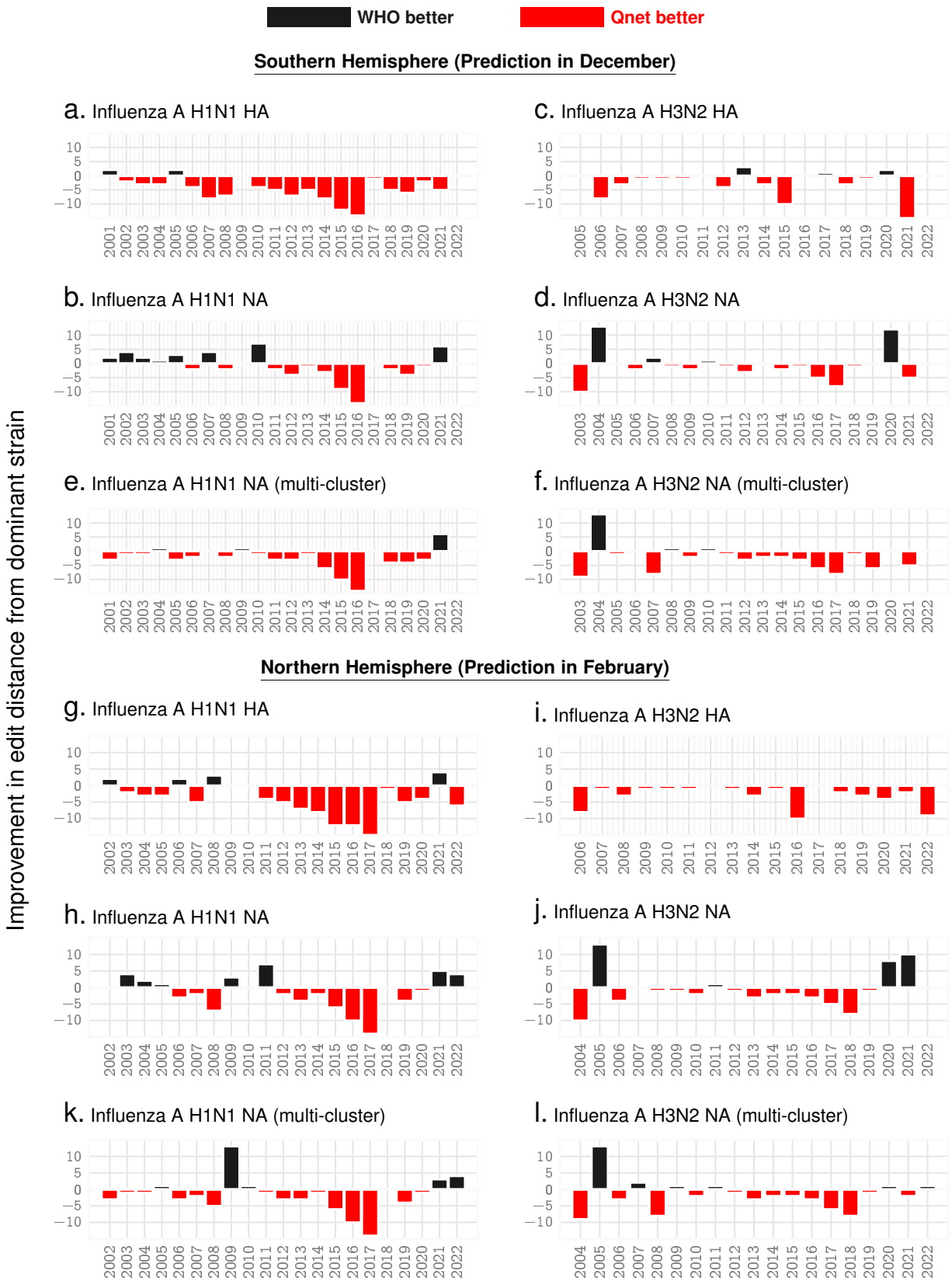

Fig. 2. **Seasonal predictions for Influenza A.** Relative out-performance of Qnet predictions against WHO recommendations for H1N1 and H3N2 sub-types for the HA and NA coding sequences over the both hemispheres. The negative bars (red) indicate the reduced edit distance between the predicted sequence 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.

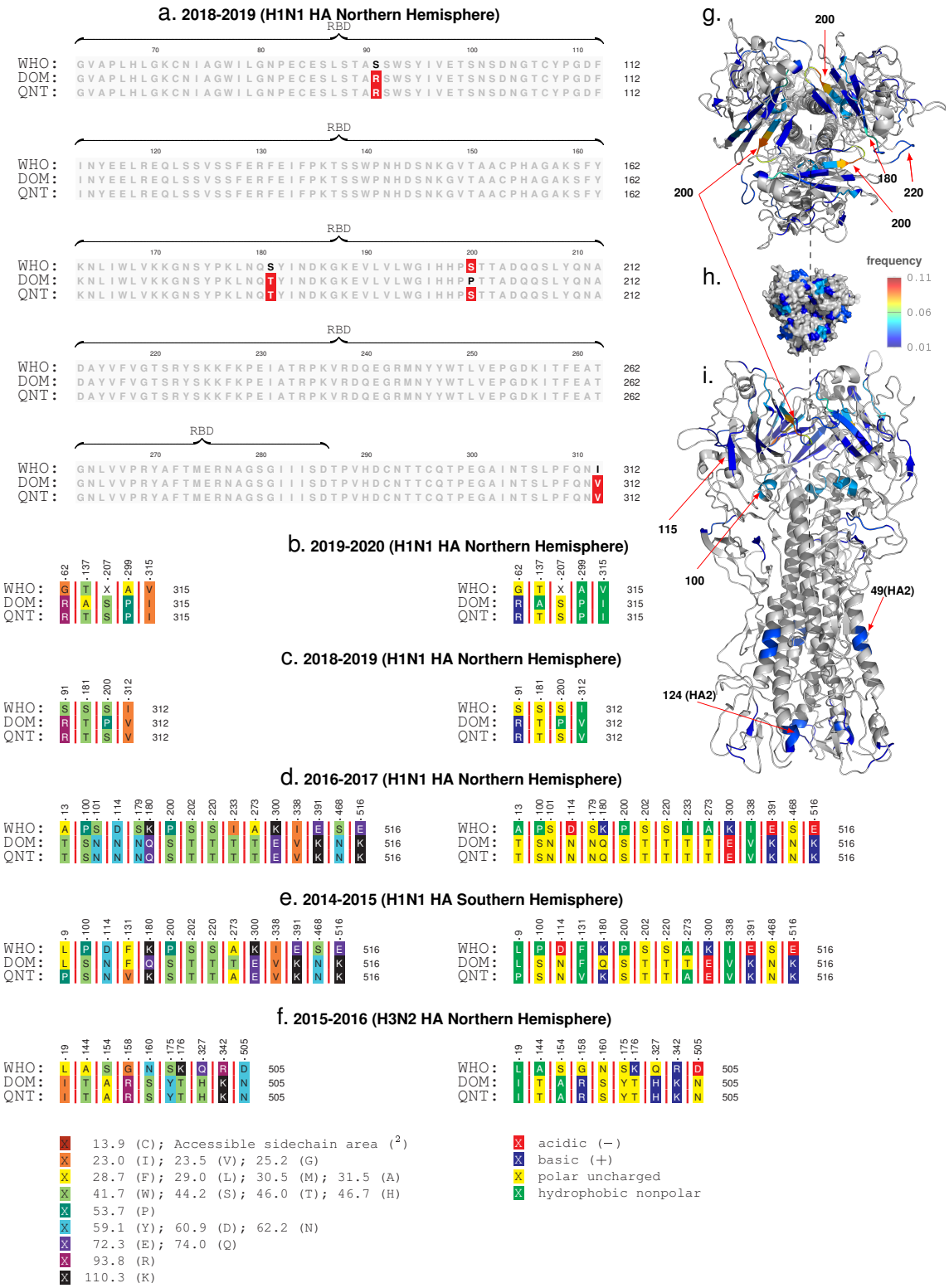

Fig. 3. **Sequence comparisons.** The observed dominant strain, we note that the correct Qnet deviations tend to be within the RBD, both for H1N1 and H3N2 for HA (panel a shows one example). Additionally, by comparing the type, side chain area, and the accessible side chain area, we note that the changes 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.

the observed amino acid "outcome" at a specific positional index of the protein of interest, using as features the bases appearing at other locations of the protein sequence (Fig. 1b). The algorithm automatically identifies the set of features (AA positions) that influence the outcome at a prticular position, implying that an Emergenet comprises atmost as many such position-specific predictors as the sequence length of the protein. Currently, the component predictors of an Emergenet are implemented as conditional inference trees[9],that make sure each node split during the tree construction has a minimum level of statistical significance in differentiating the resulting child nodes. Importantly, the

#3. WHy dont we use phenotypic information

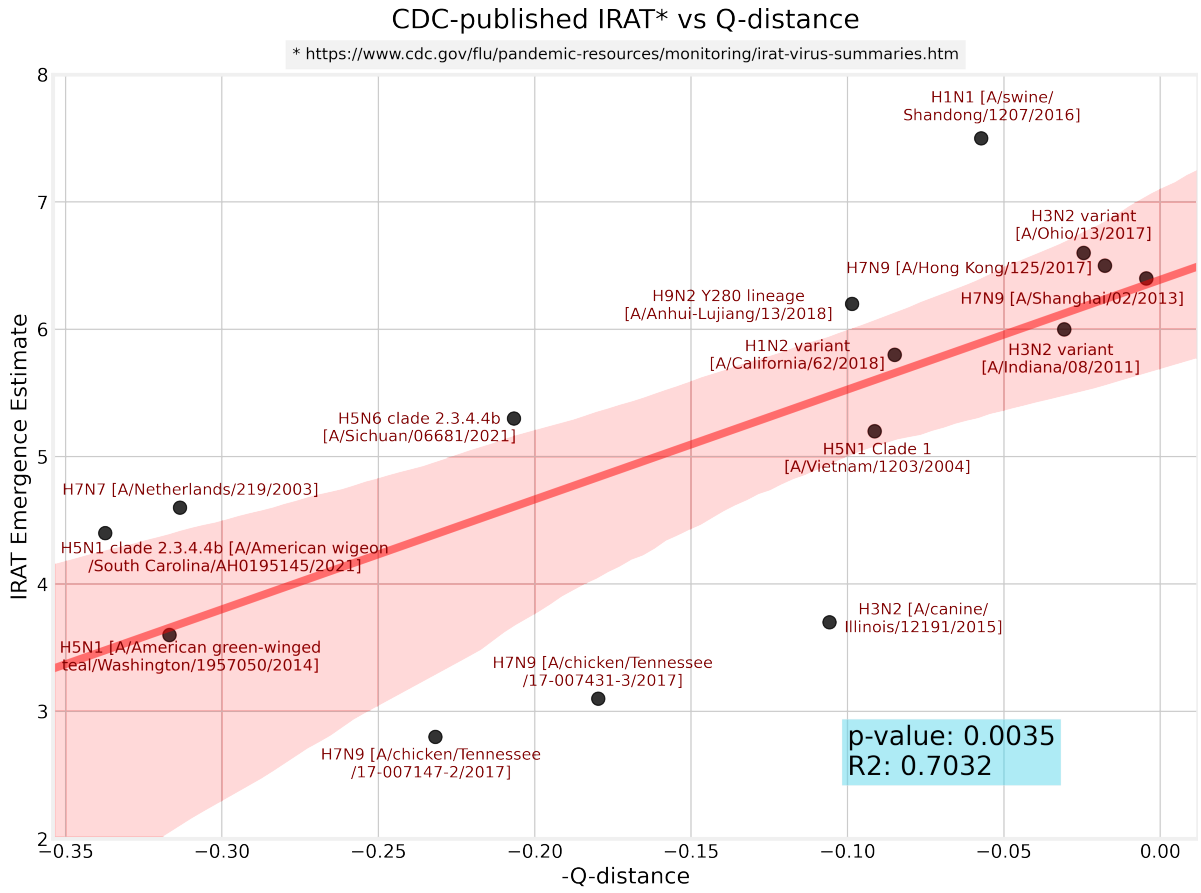

Fig. 4. **IRAT emergence risk vs. q-distance**. There is an approximate linear relationship between average q-distance from human circulating strains (averaged across both HA and NA) and IRAT emergence risk grade. Note that IRAT has released results for 23 strains to date, but only 15 are plotted on 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.

Emergenet component models are inferred purely from the amino acid sequences of the proteins of interest with no additional phenotypic annotation, other than identifying the host animal (and time and place of collection). Antigenic characterization of different Influenza A strains tend to be substantially laborious and low-throughput compared to genome sequencing[21], but incorporation of such 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 may be leveraged to outperform strategies explored in the literature. Thus, our modeling approach is a substantial departure from the state of the art; we do not assume models of mutation a priori, and we make use of little information that is specific to Influenza A genomic sequences, or its life cycle, beyond focusing on HA and NA which are known to be the main targets of neutralizing antibodies[1].

Our results demonstrate improved solutions to two important problems described above, namely: 1) forecast of dominant strain(s) for the next flu season, and 2) identification of high risk strains amongst those collected in non-human hosts in the wild. 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. 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[22], strain-matching is a crucial component to realizing high vaccine effectiveness[23]. Our results outline a new approach to 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[6,8]. Our primary claim, however, is the ability to estimate the pandemic potential of novel strains, via our proposed risk score. The risk score is demonstrated to closely tracks the IRAT estimate for Influenza A viruses, specifically ones that do not yet circulate in humans, and is computable in seconds as opposed to weeks taken by the detailed experimental assays. This dramatic reduction in time and cost potentially opens the door to fully exploiting the current biosurvellinace capacity, scalably analyzing risk from the tens of thousands of strains collected every year, and hence moving towards meaningful preemptive pandemic mitigation strategies.

Result requirements: 0) Check if changes happen on incorporating the additional term from eqn 9

1) percent improvement data table augnment with comparison against literature data (in number of average mutational

deviation)

2) table for fig 2

3) table/figure for risk scores of all h3n2/h1n1 strains in swine collected in 2021/2022

4) How many sequences were used from which databases. Where is this info described (need this to appear in teh abstract as well)

5) all qnets should be submitted to zenodo

It seems the old results were generated using a different version of quasinet package. Maybe we should indeed regenerate the results to make sure we can answer reviewer questions.

Target results:

1. HA NA recommendations, northern outhern hemispheres (earlier we used 2 years of data, and chose the time over which we consider last season sequences as different for north vs the south, we did not use separate qnets over north/south)

2. check if recommendations match or errors match at least 3. check similar approach for latest years 4. must complete IRAT prediction table (using geometric mean between ha and na estimates seems to do better. try)

RESULTS

We begin by collecting $> 98,000$ Influenza A HA/NA nucleotide sequences from two public databases (NCBI: https://www.ncbi.nlm.nih.gov/genome/viruses/ and GISAID: https://www.gisaid.org/ see SI-Table 3), uncovering a network of dependencies between individual mutations revealed through subtle variations of the aligned sequences. The representation of these dependencies as a recursive forest of conditional inference trees is referred to as the Emergenet (see Fig. 1). Using aligned genomic sequences sampled from similar populations, *e.g.* HA from Human Influenza A in year 2008, we learn models for predicting the mutational variations at each sequence index using other indices as features. For example, in Fig. **??**a, the predictor for index 1274 uses variation at index 1064 as a feature, and the predictor for index 1064 uses index 1314 as a feature, and so on – ultimately uncovering a recursive dependency structure. The Qnet predicts the nucleotide distribution over the base alphabet at any specific index, conditioned on the nucleotides making up the rest of the sequence of the gene or genome fragment under consideration. Aside from this example, amino acids sequences can also be used to train the Qnet.

#4. update the index refs

Finally, we define the q-distance (see Eq. (3) in Materials and Methods) as the square-root of the Jensen-Shannon (JS) divergence[24] of these conditional distributions from one sequence to another, averaged over the entire sequence. The q-distance, informed by the dependencies modeled by the inferred Qnets, adapts to the specific organism, allele frequencies, and variations in the background population. Invoking Sanov's theorem on large deviations[24], we show that the likelihood of spontaneous change is bounded above and below by a simple exponential function of the q-distance.

Importantly, the q-distance between two sequences may change even if only the background population changes (see SI-Table 1, where the distance between two fixed sequences vary when we vary their collection years). Sequences may have a large q-distance and a small edit distance, and vice versa (although on average the two distances tend to be positively correlated, see SI-Table 2). Hence for tracking drift in Influenza A, we construct a seasonal Qnet for each sub-type and protein that we consider.

Our first application aims to predict dominant strains for the seasonal flu epidemic. Periodic adjustment of the Influenza vaccine components is necessary to account for antigenic drift[5,25]. The flu shot in each hemisphere 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[26], hoping to match the circulating strain(s) in the upcoming flu season. A variety of hard-to-model effects hinder this prediction, which, despite observed cross-reactive effects[7], have limited vaccine effectiveness in recent years[27].

We then computed the dominant strain in each season as the centroid of the strain distribution observed in a given season in the classical sense (no. of mutations), since the edit distance metric is widely used and offers a point of comparison between WHO predictions and Qnet predictions. Note that the 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, keeping in mind that the flu season in the south begins a few months early, as described in Fig. 2. Finally, we computed the edit distance (no. of mutations) between the dominant strain and the WHO and Qnet predictions.

Our second application aims to compare emergence risk grades given by the CDC through its Influenza Risk Assessment Tool (IRAT) with results using our q-distance metric. While IRAT uses a combination of 10 weighted risk elements evaluated slowly over the course of several months per strain, we attempt to quantify emergence risk quickly with the q-distance metric. We looked at the same strains that were analyzed by IRAT. For each strain previously analyzed by IRAT, we construct Qnet models for HA and NA segments using all human strains of the same variety 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 will use human H1N1 strains circulating between July 1, 2019 through June 30, 2020. For sub-types with few human strains (H1N2, H5N1, H5N6, H7N7, H9N2), we only use the upper bound of the date. We then compute the average q-distance between the strain in question and the circulating human strains for both HA and HA segments. Seven of the 23 strains are not included in our comparison due to having zero or too few human strains in the sample space to construct a Qnet; see Supplementary Text, SI-Table 16. We hypothesize that a lower average q-distance between the strain in question and circulating human strains should correspond to a higher emergence risk. Hence, we expect to see a high negative correlation between q-distance and IRAT grade, which assigns 1 to be the lowest risk and 10 to be the highest risk.

Concrete Results

We tested the hypothesis of our first application, computing the strain closest to the "q"-centroid for each flu season and selecting that strain as the prediction for the next season's dominant strain. We performed this analysis on past two decades of sequence data for Influenza A (H1N1 and H3N2) with promising results: the q-distance based prediction demonstrably outperforms WHO recommendations by reducing the distance between the predicted and the dominant strain (Fig. 2). Recall that we identify the dominant strain to be the one that occurs most frequently, computed as the centroid of the strain distribution observed in a given season in the classical sense (no. of mutations).

The Qnet single-cluster predictions consistently outperform the WHO recommendations. For H1N1 HA, the Qnet induced recommendation outperforms the WHO suggestion by $> 33\%$ on average over the last two decades, and $> 71\%$ on average in the last decade. The gains for H1N1 NA over the same time periods are $> 15\%$ and $> 52\%$, respectively. For H3N2 HA, the Qnet induced recommendation outperforms the WHO suggestion by $> 37\%$ on average over the last two decades, and $> 40\%$ in the last decade. The gains for H3N2 NA over the same time periods are $> 14\%$ and $> 15\%$, respectively. Finding multi-cluster predictions has the potential to yield even more improved results, as seen in Fig. 2 and SI-Table 12 through SI-Table 15.

The full table of single-cluster results with improvement broken down by hemisphere is given in Table 1. Fig. 2 illustrates the relative gains computed for both subtypes and the two hemispheres (since the flu season occupy distinct time periods and may have different dominant strains in the northern and southern hemispheres[25]). 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 Qnet 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 15.

We hypothesized in our second application that there will be a high negative correlation between q-distance and IRAT emergence grade. Plotting our results in Fig. 4, we find a correlation of -0.7032 ($p < 0.005$), which is statistically and substantively significant. We can conclude, therefore, that a lower averae q-distance to currently circulating human strains corresponds to a higher risk of emergence with respect to the CDC's grades. Due to the small number of Influenza A strains that have been analyzed by IRAT, we should be wary of the realistic statistical significance of our results. Achieving a moderately high correlation coefficient and p-value is nevertheless a positive result, and further uncovers the potential of our model to quantify risk of emergence.

For further analysis, we also performed q-analysis on IRAT H1- and H3- sub-types by taking average q-distance between the target strain and all human-circulating strains available, with no upper or lower collection date bound. We expected the correlation to be worse than with bounded strains, since a strain being "close" to humans at some point in the past does not necessarily mean being close now. Indeed, our results showed almost no correlation to the IRAT emergence risk scores. Bounded results for H1- and H3- sub-types yielded a correlation of -0.6916, while unbounded results yielded a correlation of 0.0545; see SI-Fig. 3.

Given the efficiency of the q-distance computations, we can track how risk of emergence changes over time by continually updating the current human-circulating strains each year. For exact average q-distance and Qnet sample size statistics, please see SI-Table 16 in the Supplementary Text.

DISCUSSION

Emergenet Advantages and Related Literature

Numerous tools exist for ad hoc quantification of genomic similarity[8,13,28–31], which are not inherently biologically meaningful – a smaller edit distance between two strains does not necessarily imply that a feasible trajectory exists from one to the other in the wild. These measures tend to be variations of distances between symbolic sequences, and are not aware of selection pressures and evolutionary dynamics. Despite the diverse techniques and concepts explored in these domains, the key 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 the first of its kind to learn an appropriate metric of comparison from data, without assuming any model of DNA or amino acid substitution, or a genealogical tree a priori, and is designed to be aware of the impact of the host environment and background epidemiology.

This is a major improvement over the existing state of art in phylogeny construction from sequences, which generally

assume a model for character substitution (either for nucleotides or amino acid residues) ignoring the effect of selection and the existence of long-range complex dependencies in viable mutations along the genomic sequence. Notably, even relatively complex substitution models ($e.g.$ ones that allow site specific mutation rates) do not capture the effect of individual changes that may dramatically alter fitness in the environment. Our proposed approach, on the other hand, learns from and leverages these patterns, using sophisticated pattern discovery via novel machine learning algorithms. While the effect of the environment and selection cannot be inferred from a single sequence, an entire database of observed strains, processed through sophisticated learning, can parse out predictive models of these complex interactions. Our only intuitively well-justified assumption on the evolutionary dynamics is that more fit strains end up with more progenies (follows from definition of Malthusian fitness), and are thus more likely to be sampled in surveillance efforts (again, intuitively obvious). Thus, in a strict mathematical sense, the distance we propose is not a distance between two strains x, y, but a distance between strain x^A (x in a background environment A) and a strain y^B (y in a background environment B). Indeed we can show that the distance between the same pair of strains of Influenza A HA is different based on if they are collected in 2008 vs in 2009, reflecting that the background environment and circulating diversity changed over the two years. Thus, our distance metric is fundamentally different from measures that exist in the literature. In particular, our mathematical framework leads to the key result that the q-distance is a scaled representation of the log-likelihood of spontaneous jump between strains. This interpretation is missing in existing tools, and makes way for leveraging the q-distance to model emergence of new strains. Thus, we can predict entirely new sequences – which differ by a non-trivial number of edits from any observed strain – that still lead to functional proteins, as demonstrated in our preliminary studies.

Very recently, two articles have explored the possibility of predicting pathogenicity from genomic sequences (Mollentze[32]) and forecasting which amongst observed mutations will dominate the circulating population (Maher et al[33]). These studies provide strong pieces of evidence that challenge the idea that forecasting future variants of virus strains is impossible, while aligns with our goals. While their questions overlap with our framework, our approach is distinct and more ambitious. For example, Mollentze uses classical sequence similarity; extended to include similarity to human housekeeping genes hoping to identify viruses evading the human immune system more easily. The demonstrated performance is poor (incorrectly tagging all SARS-related coronaviruses as potentially pathogenic), implying unactionable specificity. On the other hand, Maher outright assumes mutations to be independent. Features are found manually, are specific to SARS-CoV-2, and the authors take a meta-analysis-esque route, compiling together a "kitchen-sink" of features via standard machine learning. Importantly, these approaches only aim to predict point mutations, with the gargantuan complexity of tracking a more complete strain through a high-dimensional sequence space well beyond their conceptual limits. Thus, even the question if whether a yet-to-be-seen strain is indeed a valid biological encoding of a virus (which is simpler to determining risk posed by such future variants) cannot answered by our peers, limiting such approaches to analyzing mutations already seen, or strains already collected. Additionally, generalizability and actionability is suspect, given that Maher's features are SARS-CoV-2 specific, and Mollentze's similarity to housekeeping genes might not be universal. Finally, both these approaches apply to a mutation or a combination of mutations that already exist, and cannot predict new mutations, or new strains.

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[34–36] 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[37]. 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[26], 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[7], have had limited vaccine effectiveness in recent years[27]. 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[11] 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. 3a 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. 3b-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[38,39] changes in local or regional properties such as charge, hydropathy, side chain solvent accessibility[40–43].

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. 3g-i, HA0 numbering used, other numbering conversions are given in SI-Table 18), 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[44–46]. 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.

Limitations & Conclusion

Calculation of q-distance is currently limited to similar and aligned sequences, $e.g.$ Influenza strains from different sub-types, hosts or seasons. Furthermore, we need a sufficient diversity of observed strains to successfully construct the Qnet. 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 17). Arguably, simply reducing the edit distance from the dominant strain is not guaranteed to translate to a better immunological protection. Nevertheless, consistent improvement in this metric achieved purely via computational means suggests the possibility of improvement over current practice.

In conclusion, we introduce a data-driven distance metric to track subtle deviations in sequences. We show that we can use the q-distance metric to make recommendations for the flu-shot composition, outperforming the WHO's recommendations in relation to the dominant strain. We also show that we can roughly replicate the CDC's IRAT grades for emergence risk of strains not currently circulating among humans in an efficient manner that can be scaled to rank many more strains than is currently done. The ability to predict future flu strains via subtle variations in a limited set of immunologically important residues suggest that the tools developed here could lead to more effective escape-resistant vaccines, which could be essential in preempting and mitigating the next pandemic.

Online 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[47], 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[24] between distributions, and among different possibilities, the E-distance is the simplest metric such that the likelihood of a spontaneous jump (See Eq. (4) 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:

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

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

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[48] 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:*

$$\Phi_i : \prod_{j \neq i} \Sigma_j \to \mathscr{D}(\Sigma_i), \tag{2}$$

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[9], 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:*

$$\theta(x, y) \triangleq \mathbf{E}_i \left(\mathbb{J}^{\frac{1}{2}} \left(\Phi_i^P(x_{-i}), \Phi_i^Q(y_{-i}) \right) \right) \tag{3}$$

where $\mathbb{J}(\cdot, \cdot)$ is the Jensen-Shannon divergence[49] 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[24], 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:

$$\omega_y^Q e^{\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \geqq Pr(x \to y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N^2}{1-\alpha}\theta(x,y)} \tag{4}$$

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. (4) above) in the supplementary text, we have:

$$\ln \frac{Pr(x \to x^{t+\delta})}{\omega_{x^{t+\delta}}} \geqq -\frac{\sqrt{8}N^2}{1-\alpha}\theta(x, x^{t+\delta}) \tag{5}$$

$$\Rightarrow \sum_{x \in H^t} \ln \frac{Pr(x \to x^{t+\delta})}{\omega_{x^{t+\delta}}} \sum_{x \in H^t} \geqq -\frac{\sqrt{8}N^2}{1-\alpha}\theta(x, x^{t+\delta}) \tag{6}$$

$$\Rightarrow \sum_{x \in H^t} \theta(x, x^{t+\delta}) - \left|H^t\right| A \ln \omega_{x^{t+\delta}} \geqq A \ln \frac{1}{\prod_{x \in H^t} Pr(x \to x^{t+\delta})} \tag{7}$$

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:

$$x_\star^{t+\delta} = \underset{y \in \cup_{\tau \leq t} H^\tau}{\arg\min} \sum_{x \in H^t} \theta(x, y) - \left|H^t\right| A \ln \omega_y \tag{8}$$

Further noting that the second term on the right is at least an order of magnitude smaller compared to the first term, a good approximation for the optimization may be stated as:

$$x_\star^{t+\delta} = \underset{y \in \cup_{\tau \leq t} H^\tau}{\arg\min} \sum_{x \in H^t} \theta(x, y) \tag{9}$$

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:

$$\rho(x_a^t) \triangleq -\frac{1}{\left|H^t\right|} \sum_{x \in H^t} \theta(x_a^t, x) \tag{10}$$

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. (4) above) in the supplementary text, we have:

$$\sum_{y \in H^t} \ln \left| \frac{Pr(x_a^t \to y)}{\omega_y} \right| \leqq -\frac{\sqrt{8}N^2}{1-\alpha} \left|H^t\right| \rho(x_a^t) \tag{11}$$

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:

$$\Rightarrow \rho(x_a^t) \geqq A \ln \left(\prod_{y \in H^t} Pr(x_a^t \to y) \right)^{1/\left|H^t\right|} + C \tag{12}$$

$$\Rightarrow \rho(x_a^t) \geqq A \ln \left\langle Pr(x_a^t \to x_h^{t+\delta}) \right\rangle + C \tag{13}$$

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[50]. We use two sequences databases: 1) National Center for Biotechnology Information (NCBI) virus[51] and 2) GISAID[52] 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] 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).

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

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

[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] 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).

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

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

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

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

[16] 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).

[17] 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).

[18] 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).

[19] 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).

[20] 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).

[21] 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).

[22] 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).

[23] 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).

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

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

[26] 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).

[27] Cdc vaccine effectiveness studies (2020). URL https://www.cdc.gov/flu/vaccines-work/effectiveness-studies.htm.

[28] 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).

[29] Huelsenbeck, J. P. & Crandall, K. A. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and systematics* **28**, 437–466 (1997).

[30] 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).

[31] Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125 (2009).

[32] Mollentze, N., Babayan, S. A. & Streicker, D. G. Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS biology* **19**, e3001390 (2021).

[33] Maher, M. C. *et al.* Predicting the mutational drivers of future sars-cov-2 variants of concern. *Science Translational Medicine* **14**, eabk3445 (2022).

[34] 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).

[35] 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).

[36] Childs, J. Zoonotic viruses of wildlife: hither from yon. In *Emergence and Control of Zoonotic Viral Encephalitides*, 1–11 (Springer, 2004).
[37] 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).
[38] Carugo, O. & Pongor, S. A normalized root-mean-spuare distance for comparing protein three-dimensional structures. *Protein science* **10**, 1470–1473 (2001).
[39] 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).
[40] Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology* **55**, 379–IN4 (1971).
[41] Shrake, A. & Rupley, J. A. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of molecular biology* **79**, 351–371 (1973).
[42] 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).
[43] 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).
[44] Tzarum, N. *et al.* Structure and receptor binding of the hemagglutinin from a human h6n1 influenza virus. *Cell host & microbe* **17**, 369–376 (2015).
[45] 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).
[46] 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).
[47] 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).
[48] Vanmarcke, E. *Random fields: analysis and synthesis* (World scientific, 2010).
[49] Manning, C. D., Manning, C. D. & Schütze, H. *Foundations of statistical natural language processing* (MIT press, 1999).
[50] McAuley, J., Gilbertson, B., Trifkovic, S., Brown, L. E. & McKimm-Breschkin, J. Influenza virus neuraminidase structure and functions. *Frontiers in microbiology* **10**, 39 (2019).
[51] Hatcher, E. L. *et al.* Virus variation resource–improved response to emergent viral outbreaks. *Nucleic acids research* **45**, D482–D490 (2017).
[52] Bogner, P., Capua, I., Lipman, D. J. & Cox, N. J. A global initiative on sharing avian flu data. *Nature* **442**, 981–981 (2006).

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.

#5. review more craefully phenotypic 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)[11], 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[11]. 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.