Learning Mutational Patterns at Scale to Analyze Sequence Divergence in Novel Pathogens

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

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

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

Abstract

Influenza viruses constantly evolve[1], and mismatches between predicted and circulating strains impact vaccine effectiveness[2].A barrier to predicting the season-specific dominant strains is the limited ability to predict future mutations, or estimate the numerical likelihood of specific future strains. In this study, we introduce a biology-aware sequence similarity metric based on deep pattern recognition of emergent evolutionary constraints. We use our model in two applications. One, we calculate the odds of future mutations, outperforming WHO recommended flu vaccine compositions almost consistently over the past two decades. Two, we compute emergence risk of strains previously analyzed by the CDC's Influenza Risk Assessment Tool, showing a moderately strong linear correlation between our predictions and the CDC's, though our predictions require much less time and resources.

THE COVID-19 pandemic is one of the most devastating disasters of the past century. As researchers strive to develop effective therapeutics and vaccines to combat the SARS-CoV-2 virus, a looming question is whether we can prepare better for the next pandemic. 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[3–5] unquestionably drive up the odds of zoonotic spill-over. Nevertheless, efforts at tracking these effects have not improved our ability to quantify future risk of emergence of a specific strain from a specific host[6]. Tracking viral diversity in hosts such as bats or swines, while important, might not transparently map to emergence risk.

A key barrier to making progress in this direction has been the missing ability to estimate the likelihood of specific mutations in the future and thus to assess the risk of emergence from circulating strains. We urgently need tools to compute the likelihood of a wild strain spontaneously giving rise to another by random chance. Currently, this likelihood is qualitatively equated to sequence similarity, which is measured by the number of mutations it takes to change one strain to another. In reality, the odds of one sequence mutating to another is not just a function of how many mutations they differ by, 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*.

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 (selected for their known roles in cellular entry and exit[7]). 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.

A second application of the q-distance shows its utility in assessing risk effectively and quickly. We compare q-distance results to the CDC's Influenza Risk Assessment Tool (IRAT)[8], 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 takes months to analyze a single strain – hence the small number of analyzed strains – q-analysis can be done within seconds. Moreover, q-analysis only requires sequence data, while IRAT requires information for 10 risk elements, grouped into three categories: "properties of the virus," "attributes of the population," and "ecology & epidemiology of the virus"[8]. 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.

A successful completion of this study will have profound impact on bio-surveillance strategies. Empowered with the ability to rank newly collected strains by risk, we would be able to better judge pandemic risks, quantify the odds of a particular strain spilling to humans, and estimate its potential to lead to a global pandemic. And for strains already circulating in the human population, our tools will estimate the chances of new mutants, and their odds of escaping current vaccines. This study potentially represents an important step forward in modeling emerging pathogens, with uncharted impact on science and health, particularly as we prepare for the aftermath of COVID-19.

QNET ADVANTAGES AND RELATED LITERATURE

Numerous tools exist for ad hoc quantification of genomic similarity[9–14], 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[15]) and forecasting which amongst observed mutations will dominate the circulating population (Maher et al[16]). 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.

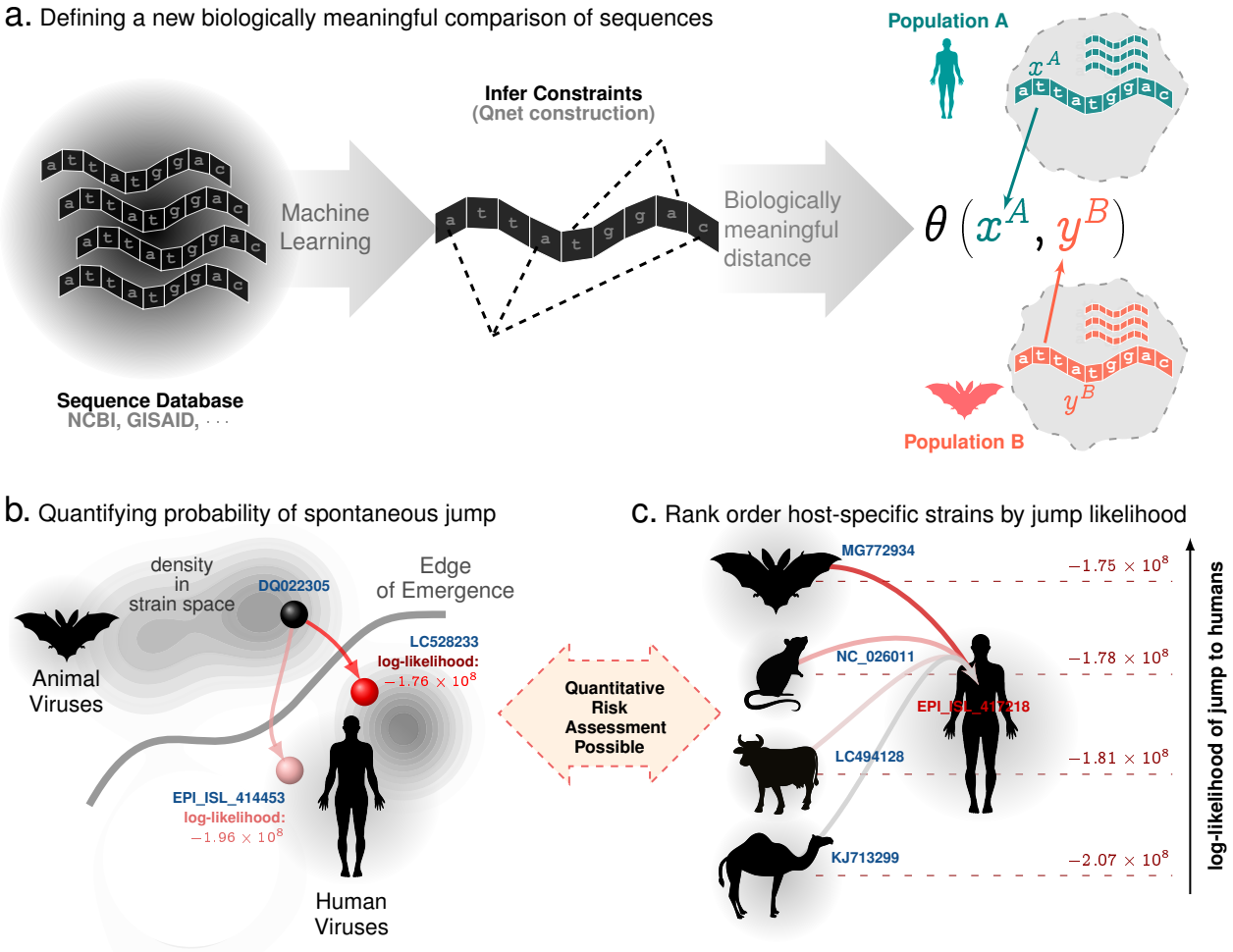

Fig. 1. **Key insights: ability to quantify risk and rank-order strains. Panel a.** Using sequence variations observed in large databases, we distill evolutionary constraints on a genomic sequence to induce a biology-aware metric for comparing subtle differences in mutating sequences. This metric (q-distance) adjusts to specific organisms, background populations and selection pressures, and reflects the true likelihood of a spontaneous jump from one sequence to the other. We can use this sequence level metric to compute distances between a sequence and a population, and two populations. **Panels b-c** illustrates that we can calculate bounds on the exact likelihood of a spontaneous jump between strains (panel b) and rank-order strains observed in a diverse set of hosts to accurately model future emergence risk (panel c).

METHODS

Aiming to validate our metric in the context of viral evolution, we begin by collecting over $98,000$ Influenza A HA/NA nucleotide sequences from two public databases (NCBI and GISAID; see SI-Table 3), uncovering a network of dependencies between individual mutations revealed through subtle variations of the aligned sequences. These dependencies define our organism-specific model referred to as the *quasi-species network* or the Qnet (see Fig. 1 and 2). 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.

Using aligned genomic sequences sampled from similar populations, *e.g.* HA from Human Influenza A in year 2008, we construct the Qnet via customized machine learning algorithms to learn models for predicting the mutational variations at each sequence index using other indices as features. For example, in Fig. 2a, 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. Finally, we define the q-distance (See Eq. (3) in Materials and Methods) as the square-root of the Jensen-Shannon (JS) divergence[17] of these conditional distributions from one sequence to another, averaged over the entire sequence. Invoking Sanov's theorem on large deviations[17], we show that the likelihood of spontaneous change is bounded above and below by a simple exponential function of the q-distance.

The mathematical intuition behind relating the new distance to change-probability is the same as in the prediction of a biased outcome when we sequentially toss a fair coin. With an overwhelming probability, such an experiment with a fair coin should result in roughly equal number of heads and tails. However, "large deviations" can happen, and the

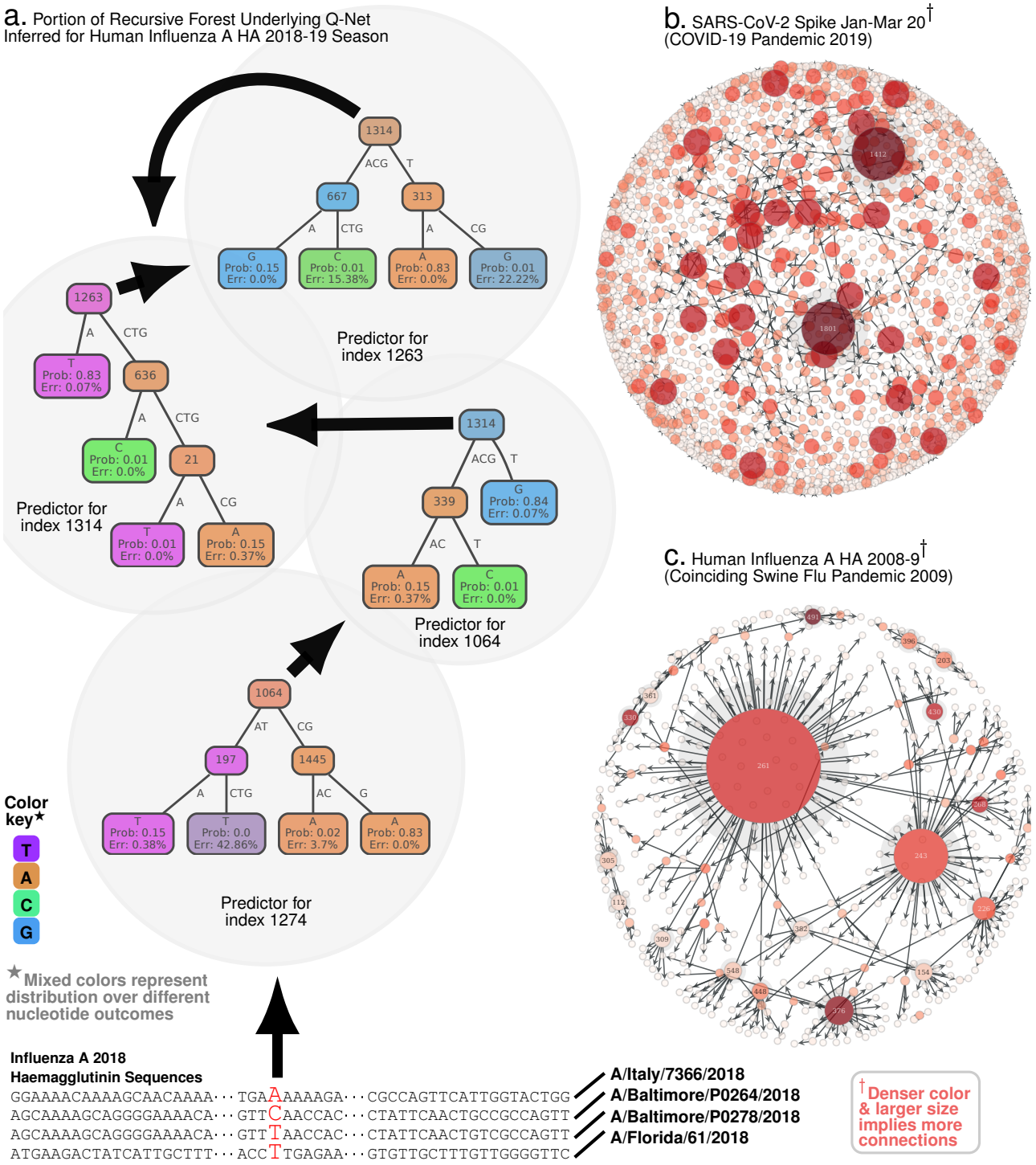

Fig. 2. **Qnet Computation scheme**. **Panel a**. As an example, beginning with aligned sequences, we calculate a conditional inference tree for index 1274, which involves indices 1064, 1445, 197 as predictive features. These features are automatically selected by the algorithm, as being maximally predictive of the base at 1274. Then, we compute predictors for each of these predictive indices, *e.g.* we show the inference tree computed for index 1064, which involves index 1314 and 339 as features. Continuing, we find that the predictor for 1314 involves indices 1263, 636 and 21, and that for 1263 involves 1314, 667 and 313. Note that recursive dependencies arise automatically: the predictor for 1263 depends on 1314, and that for 1314 depends on 1263. **Panels b-c** show Qnet dependency graphs for SARS-CoV-2 spike protein and Influenza A HA respectively, illustrating the distinct patterns of mutational constraints inferred. Both HA in Influenza A and the spike protein in SARS-CoV-2 are implicated in viral entry into host cells, and crucial for host specificity of infections. Additionally, the inferred structures underscore the significantly more complex dependencies in SARS-CoV-2 compared to Influenza A.

probability of such rare events is quantifiable[18] with existing theory. We show here that the likelihood of a spontaneous transition of a genomic sequence to a different variant by random chance may also be similarly bounded, given we have the Qnet as an estimated model of the evolutionary constraints.

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 [1,19]. 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 [20], 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 [2], have limited vaccine effectiveness in recent years [21]. For predicting future strains, we hypothesized that since the probability of a drift exponentially decreases with an increasing q-distance, the centroid of the strain distribution in our metric will change slowly. If true, the strain selected closest to the "q"-centroid will be a good approximation of next season's dominant strain. 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. 4. 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.

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. 4). 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).

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	35.02	67.44
H1N1	HA	Average	33.40	71.22
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	11.29	48.28
H1N1	NA	Average	15.46	52.14
H3N2	NA	North	13.92	8.57
H3N2	NA	South	14.77	22.73
H3N2	NA	Average	14.34	15.65

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. 4 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. 4 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 [19]). Additional improvement is possible if we recommend multiple strains every season for the vaccine cocktail (Fig. 4e,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. 3, 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.

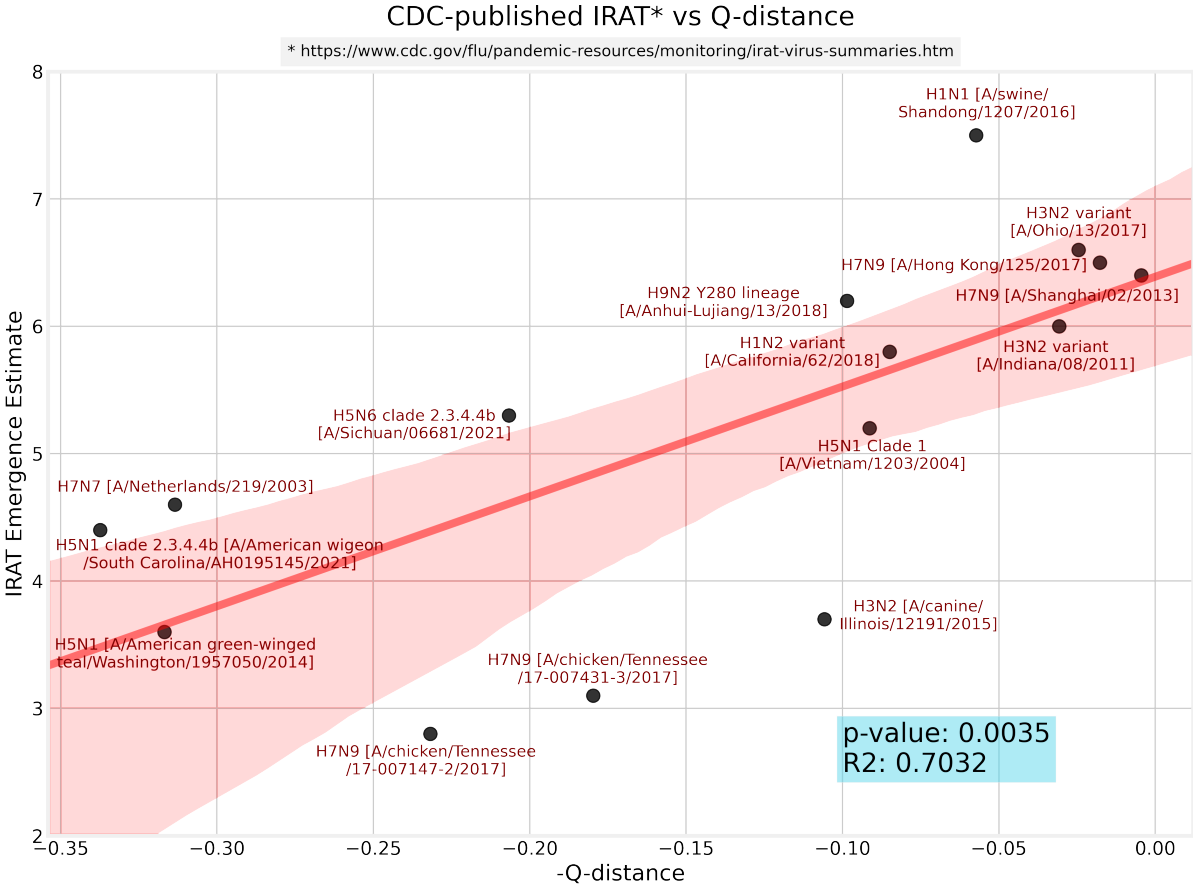

Fig. 3. **IRAT emergence risk vs. q-distance**. There is an approximate linear relationship between average q-distance from human circulating strains (averaged across both HA and NA) and IRAT emergence risk grade. Note that IRAT has released results for 23 strains to date, but only 15 are plotted on 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.

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. 5a 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. 5b-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[22,23] changes in local or regional properties such as charge, hydropathy, side chain solvent accessibility[24–27].

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. 5g-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[28–30]. 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.

Next, we briefly describe the details of the computational framework.

QNET 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[31], 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[17] between distributions, and among different possibilities, the q-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 q-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[32] 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 (Qnet)**.** *For a random field $X = \{X_i\}$ indexed by $i \in \{1, \cdots, N\}$, the Qnet is defined to be the set of predictors $\Phi = \{\Phi_i\}$, i.e., we have:*

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

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[33], although more general models are possible.

Biology-Aware Distance Between Sequences

Definition 2 (Q-distance – pseudo-metric 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 Qnet 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[34] 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[17], 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 Qnet 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 Qnet 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: Predicting Seasonal Strains

Analyzing the distribution of sequences using the q-distance allows us to estimate seasonal drift, which is particularly applicable to Influenza and Influenza-like viruses for which periodic adjustments of vaccine components are necessary to account for antigenic variations.

Our prediction is based on the following intuition: since the probability of spontaneous jump to a strain further away in the q-distance is exponentially lower, the q-centroid of the strain distribution (the centroid computed in the q-distance metric) observed over a season is expected to move slowly, and will be close to the dominant strain in the next season. Thus, we estimate the predicted dominant strain $\widehat{x}^{t+1}$ at time $t+1$, as a function of the observed population at time t as follows:

$$\widehat{x}^{t+1} = \underset{x \in P}{\arg\min} \sum_{y \in P^t} \theta(x, y) \tag{5}$$

where P^t is the sequence population at time t and $P = P^t \cup P^{t-1} \cup P^{t-2} \cup \cdots \cup P^1$. Here the unit of time is chosen to reflect the appropriate frequency over which vaccine components are re-assessed. In the case of Influenza, this is typically one year. Using this formulation, we test if the predicted strains are closer to the dominant strain in the classical edit distance, when compared against the WHO vaccine recommendations (See Fig. 4).

DATA SHARING

Working software is publicly available at https://pypi.org/project/quasinet/. 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[35]. We use two sequences databases: 1) National Center for Biotechnology Information (NCBI) virus[36] and 2) GISAID[37] 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] Dos Santos, G., Neumeier, E. & Bekkat-Berkani, R. Influenza: Can we cope better with the unpredictable? *Human vaccines & immunotherapeutics* **12**, 699–708 (2016).

[2] 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).

[3] 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).

[4] 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).

[5] Childs, J. Zoonotic viruses of wildlife: hither from yon. In *Emergence and Control of Zoonotic Viral Encephalitides*, 1–11 (Springer, 2004).

[6] 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).

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

[8] 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).

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

[10] 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).

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

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

[13] 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).

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

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

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

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

[18] 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).

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

[20] 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).

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

[22] Carugo, O. & Pongor, S. A normalized root-mean-spuare distance for comparing protein three-dimensional structures. *Protein science* **10**, 1470–1473 (2001).

[23] 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).

[24] Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology* **55**, 379–IN4 (1971).
[25] Shrake, A. & Rupley, J. A. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of molecular biology* **79**, 351–371 (1973).
[26] 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).
[27] 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).
[28] Tzarum, N. *et al.* Structure and receptor binding of the hemagglutinin from a human h6n1 influenza virus. *Cell host & microbe* **17**, 369–376 (2015).
[29] 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).
[30] 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).
[31] 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).
[32] Vanmarcke, E. *Random fields: analysis and synthesis* (World scientific, 2010).
[33] Hothorn, T., Hornik, K. & Zeileis, A. Unbiased recursive partitioning: A conditional inference framework. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS* **15**, 651–674 (2006).
[34] Manning, C. D., Manning, C. D. & Schütze, H. *Foundations of statistical natural language processing* (MIT press, 1999).
[35] McAuley, J., Gilbertson, B., Trifkovic, S., Brown, L. E. & McKimm-Breschkin, J. Influenza virus neuraminidase structure and functions. *Frontiers in microbiology* **10**, 39 (2019).
[36] Hatcher, E. L. *et al.* Virus variation resource–improved response to emergent viral outbreaks. *Nucleic acids research* **45**, D482–D490 (2017).
[37] Bogner, P., Capua, I., Lipman, D. J. & Cox, N. J. A global initiative on sharing avian flu data. *Nature* **442**, 981–981 (2006).

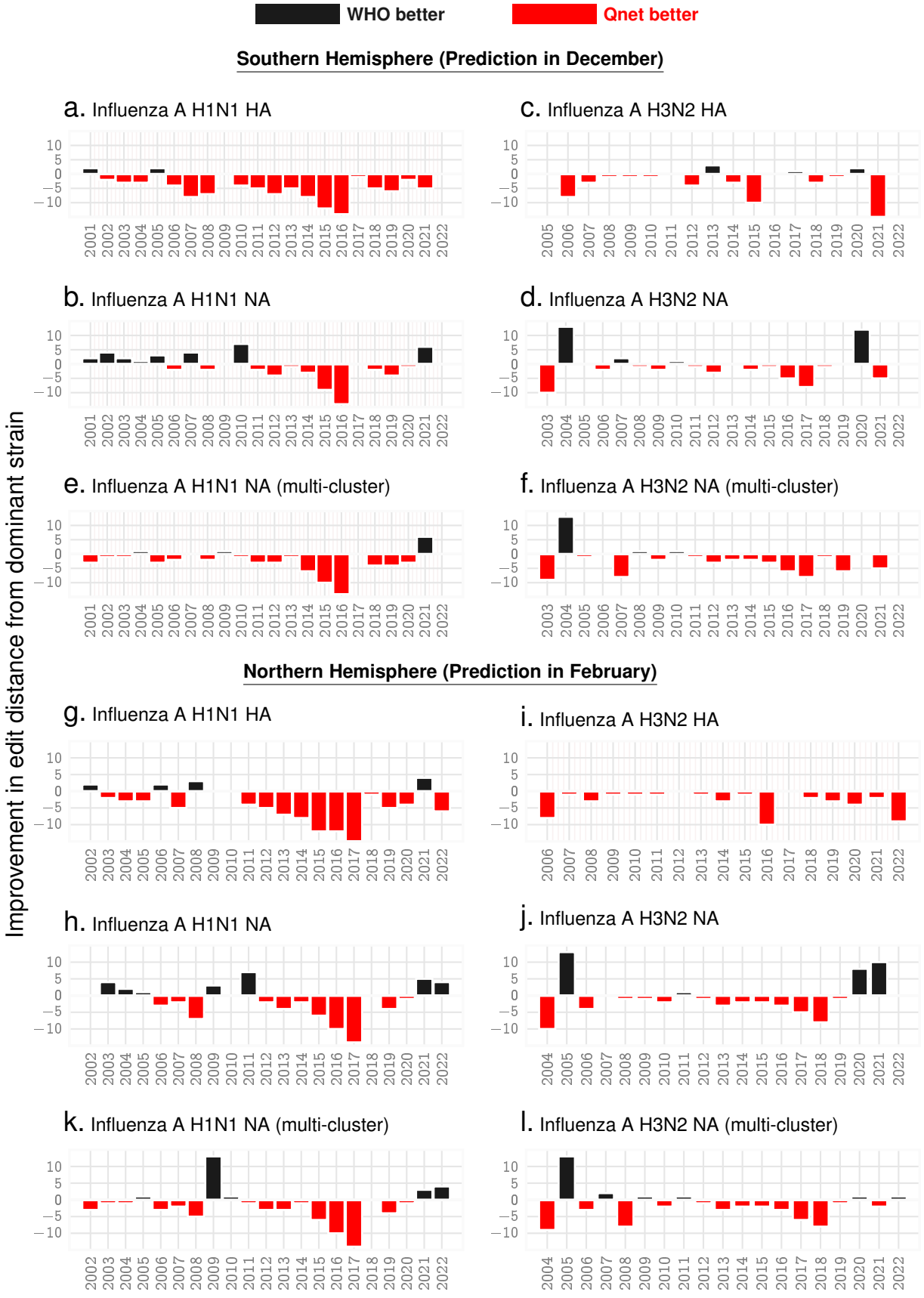

Fig. 4. **Seasonal predictions for Influenza A.** Relative out-performance of Qnet predictions against WHO recommendations for H1N1 and H3N2 sub-types for the HA and NA coding sequences over the both hemispheres. The negative bars (red) indicate the reduced edit distance between the predicted sequence 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.

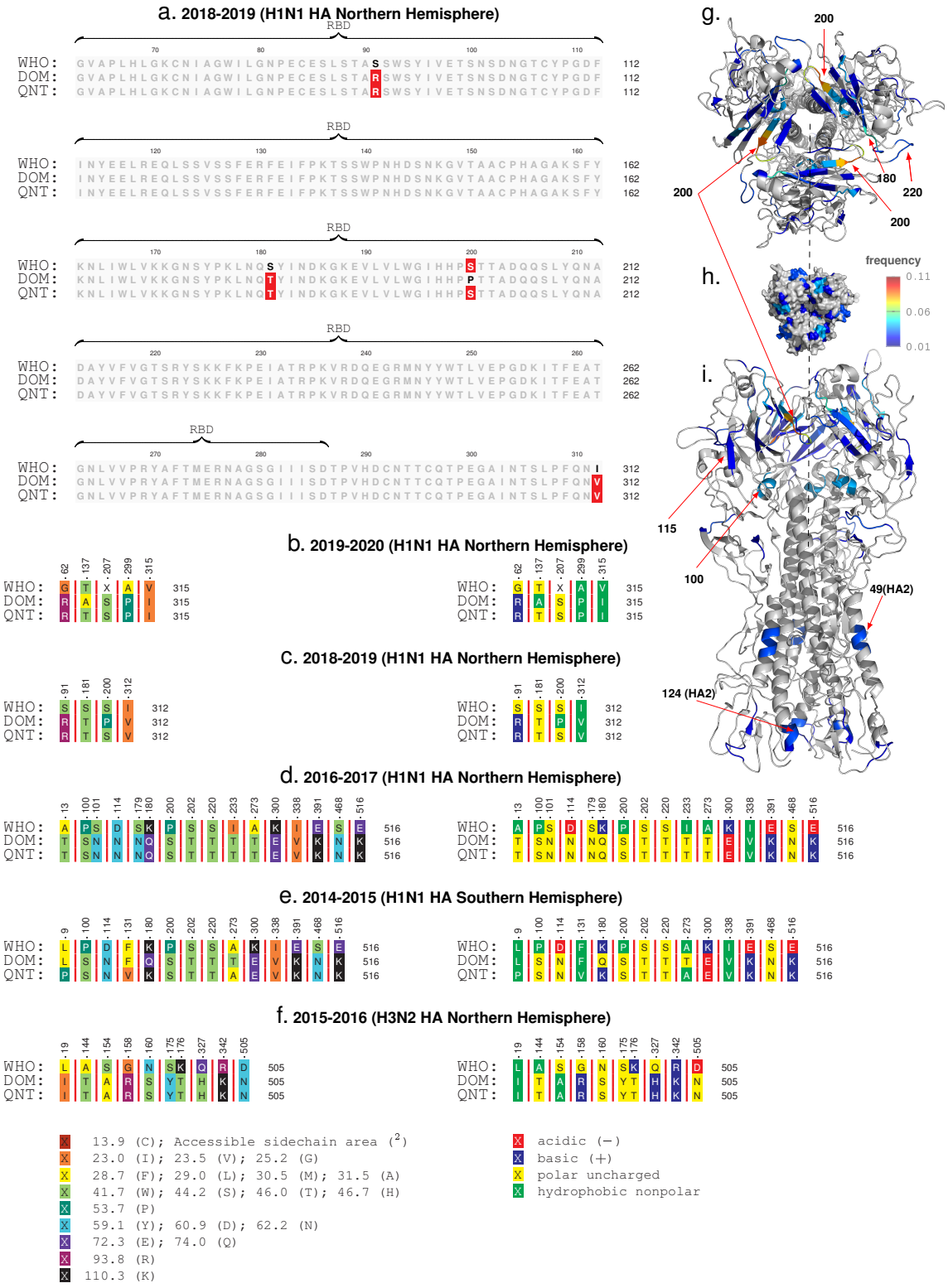

Fig. 5. **Sequence comparisons.** The observed dominant strain, we note that the correct Qnet deviations tend to be within the RBD, both for H1N1 and H3N2 for HA (panel a shows one example). Additionally, by comparing the type, side chain area, and the accessible side chain area, we note that the changes 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.