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.

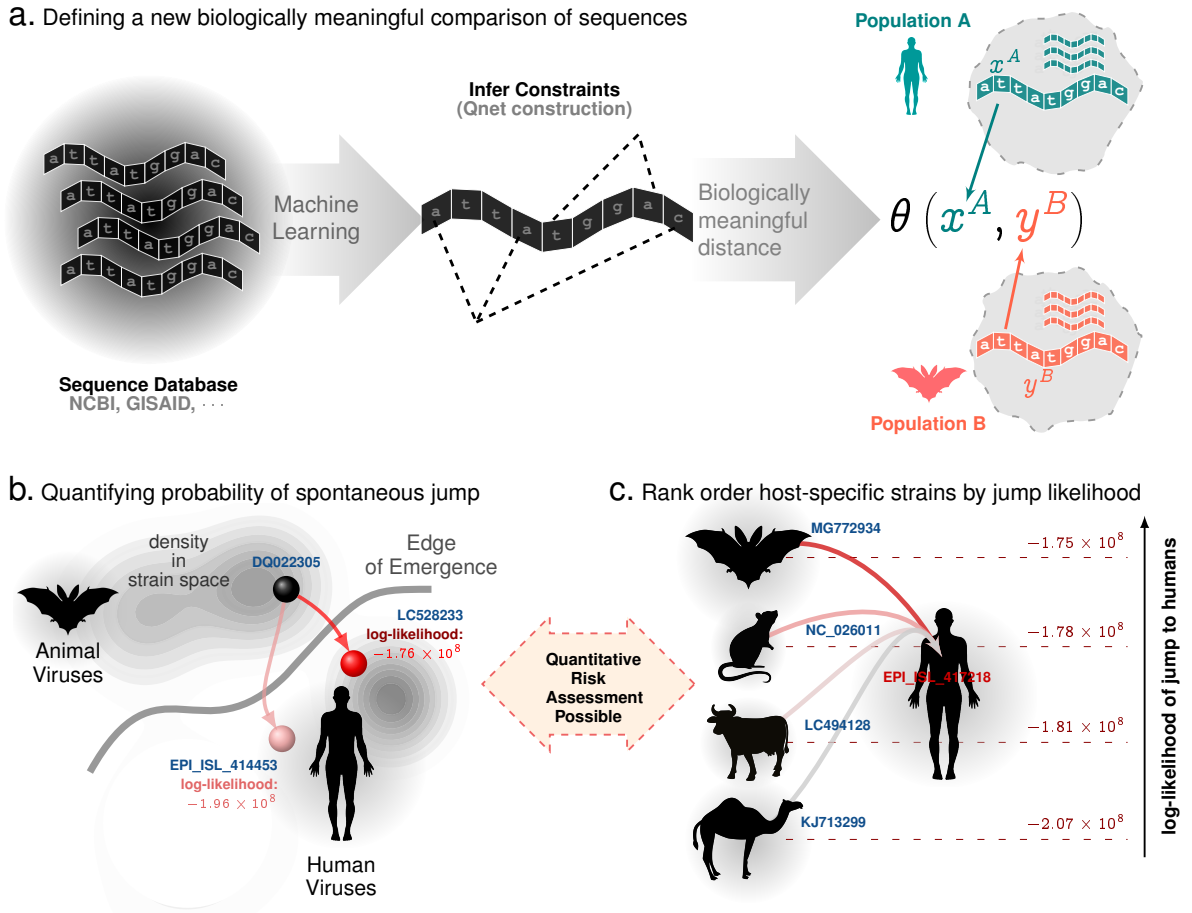


Fig. 1. **Key insights: ability to quantify risk and rank-order strains. Panel a.** Using sequence variations observed in large databases, we distill evolutionary constraints on a genomic sequence to induce a biology-aware metric for comparing subtle differences in mutating sequences. This metric (q-distance) 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

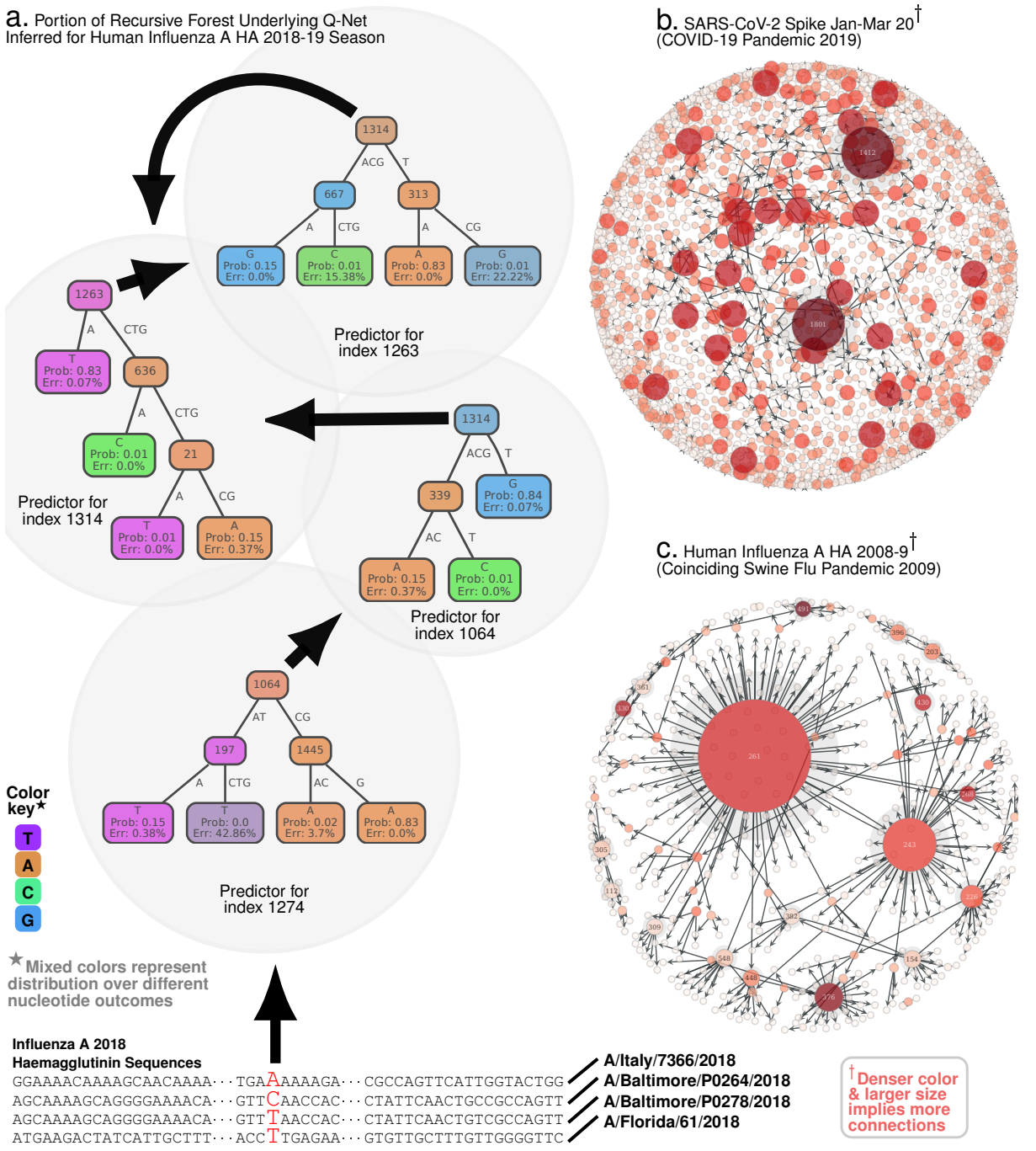


Fig. 2. **Qnet Computation scheme**. **Panel a**. As an example, beginning with aligned sequences, we calculate a conditional inference tree for index 1274, which involves indices 1064, 1445, 197 as predictive features. These features are automatically selected by the algorithm, as being maximally predictive 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.

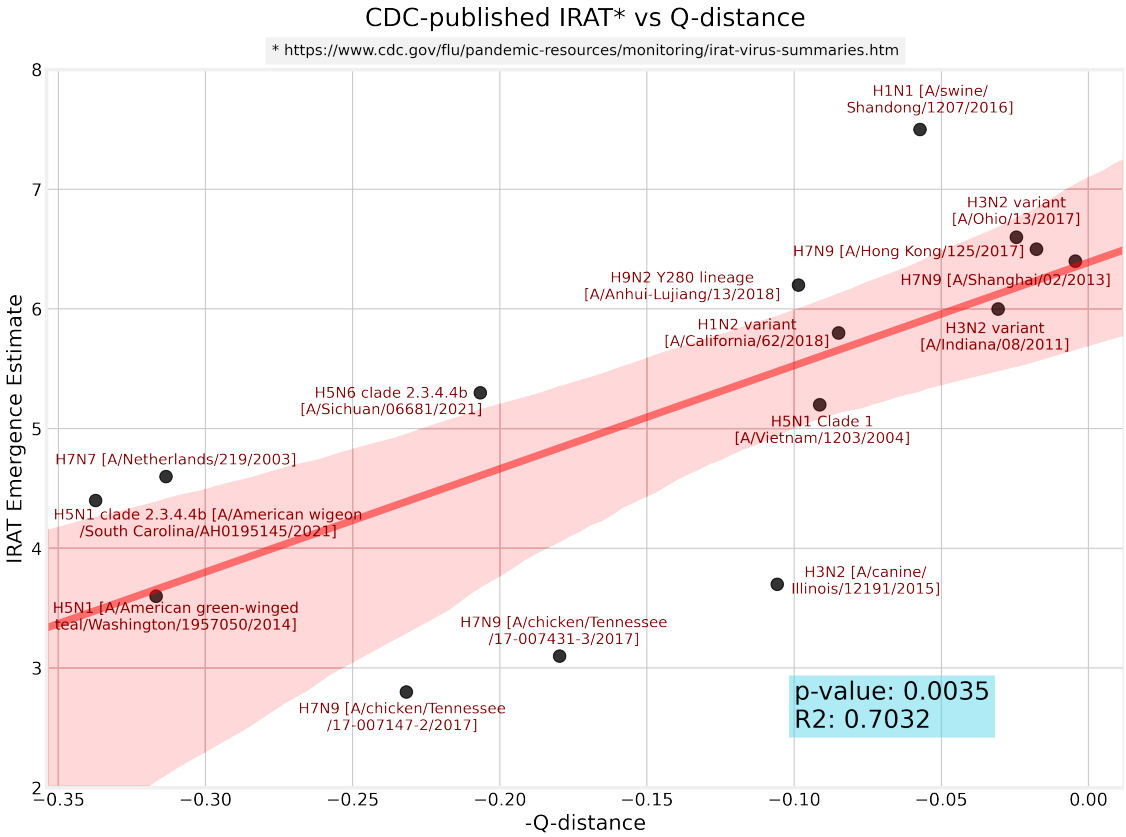


Fig. 3. **IRAT emergence risk vs. q-distance**. There is an approximate linear relationship between average q-distance from human circulating strains (averaged across both HA and NA) and IRAT emergence risk grade. Note that IRAT has released results for 23 strains to date, but only 15 are plotted on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

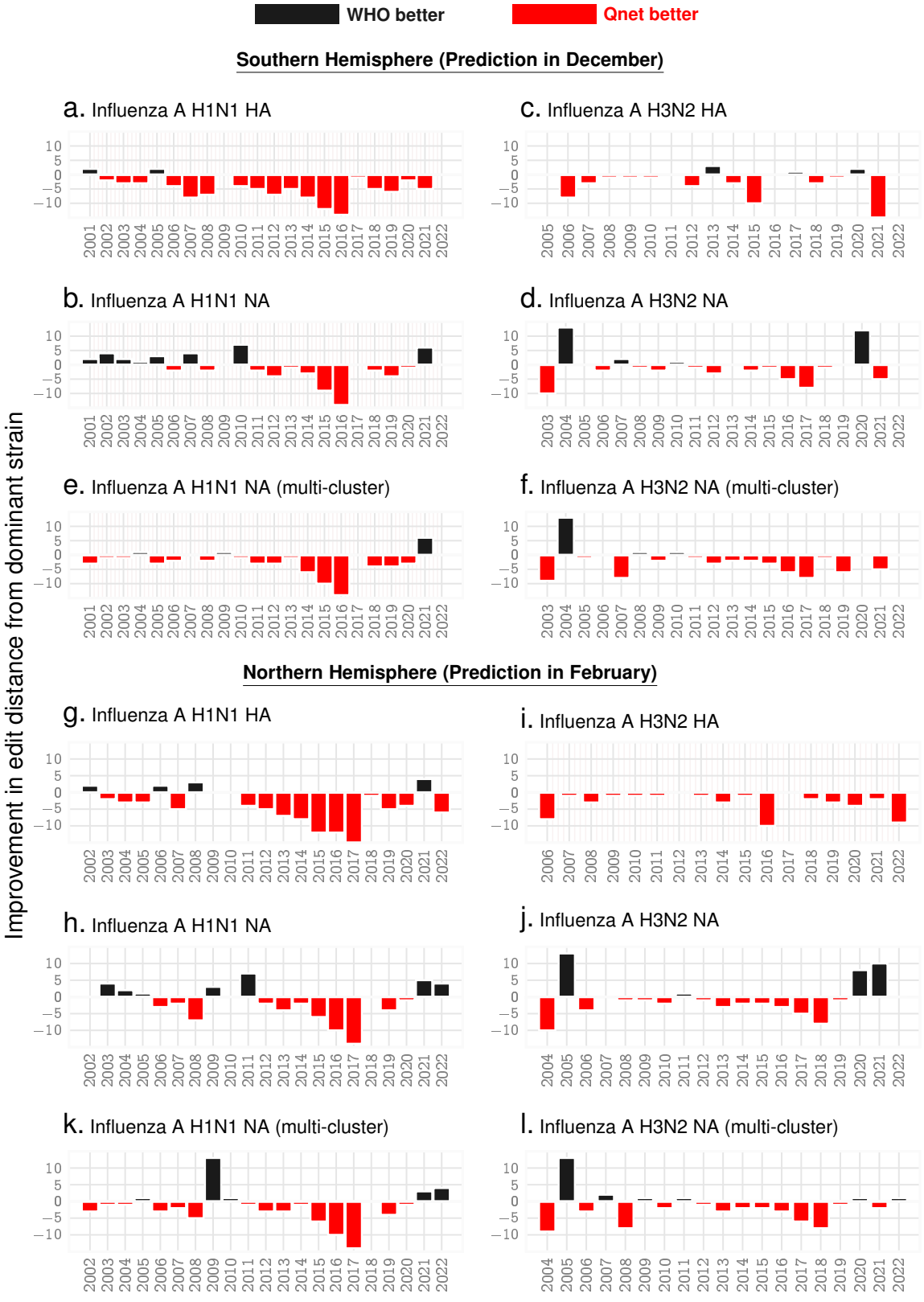


Fig. 4. **Seasonal predictions for Influenza A.** Relative out-performance of Qnet predictions against WHO recommendations for H1N1 and H3N2 sub-types for the HA and NA coding sequences over the both hemispheres. The negative bars (red) indicate the reduced edit distance between the predicted sequence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

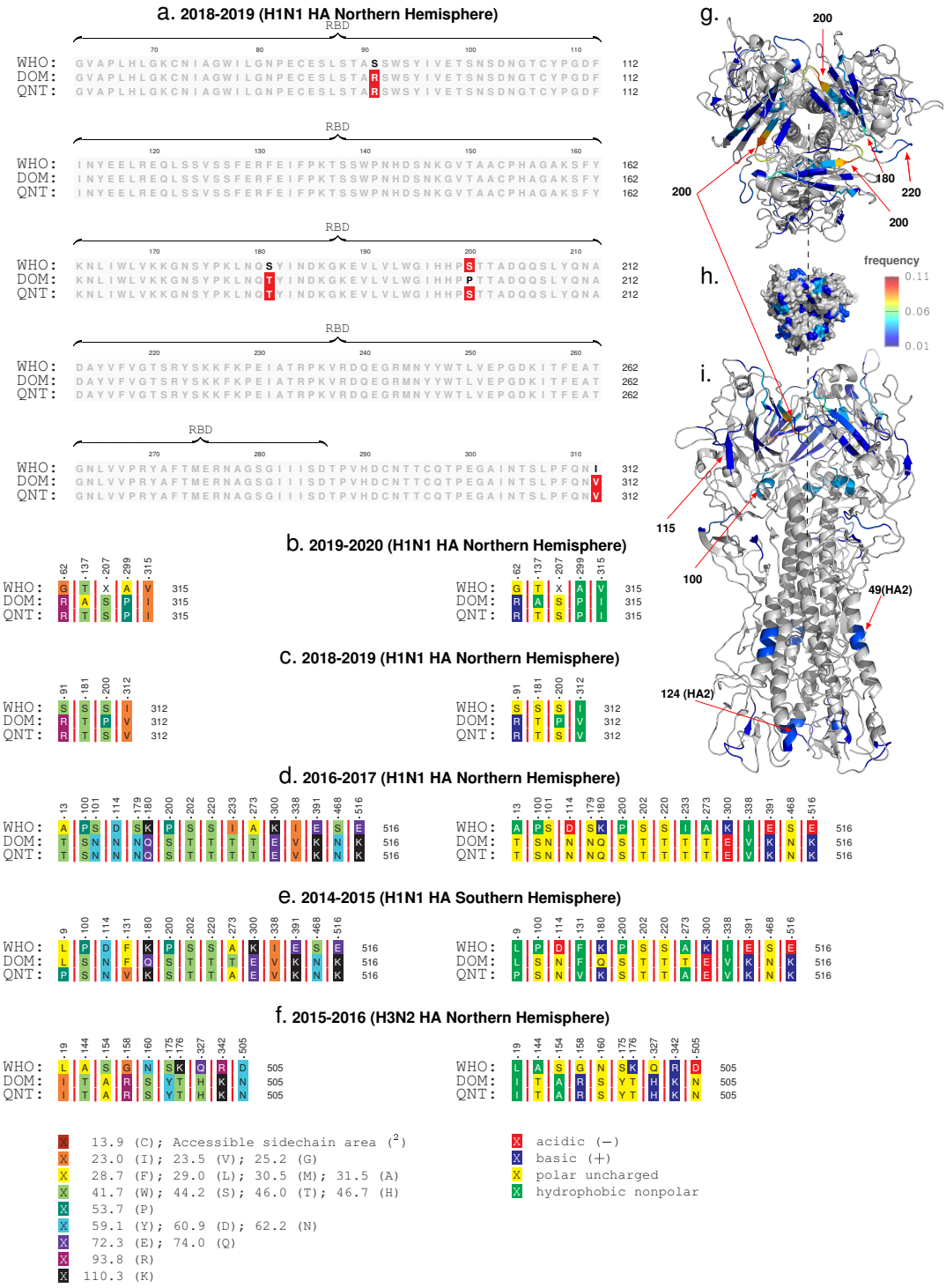


Fig. 5. **Sequence comparisons.** The observed dominant strain, we note that the correct Qnet deviations tend to be within the RBD, both for H1N1 and H3N2 for HA (panel a shows one example). Additionally, by comparing the type, side chain area, and the accessible side chain area, we note that the changes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

Supplementary Text: 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.

SUPPLEMENTARY METHODS: NOTES ON Q-DISTANCE & SUPPORTING RESULTS

The *q-distance* is a pseudo-metric since distinct sequences can induce the same distributions over each index, and thus evaluate to have a zero distance. This is actually desirable; we do not want our distance to be sensitive to changes that are not biologically relevant. The intuition is that not all sequence variations brought about by substitutions are equally important or likely. Even with no selection pressure, we might still see random variations at an index if such variations do not affect the replicative fitness. Under that scenario, the corresponding Φ_i will predict a flat distribution no matter what the input sequence is, thus contributing nothing to the overall distance. And even if two strains x, y have the same entry at some index i, the remaining residues might induce different distributions Φ_i based on the remote dependencies, $i.e.$, the entries in x_{-i}, y_{-i}. Also, it matters if the sequences come from two different background populations P, Q, $i.e.$, if the induced Qnets Φ^P, Φ^Q are different. Thus, if we construct Qnets for H1N1 Influenza A separately for the collection years 2008 and 2009, then the same exact sequence collected in the respective years might have a non-zero distance between them, reflecting the fact that the background population the sequences arose from are different, inducing possibly different expected mutational tendencies (See SI-Table 1).

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

Definition 1 (Pseudo-metric between populations)**.** *Using the notion of Hausdorff metric between sets:*

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

$$\theta(x, Q) = \min_{y \in Q} \theta(x, y) \tag{1}$$

$$\theta(P, Q) = \max \left\{ \max_{x \in P} \theta(x, Q), \max_{y \in Q} \theta(y, P) \right\} \tag{2}$$

In-silico Corroboration of Qnet Constraints

We carry out in-silico experiments to corroborate that the constraints represented within an inferred Qnet are indeed reflective of the biology in play. We compare the results of simulated mutational perturbations to sequences from our databases (for which we have already constructed Qnets), and then use NCBI BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to identify if our perturbed sequences match with existing sequences in the databases (See SI-Fig. 1). We find that in contrast to random variations, which rapidly diverge the trajectories, the Qnet constraints tend to produce smaller variance in the trajectories, maintain a high degree of match as we extend our trajectories, and produces matches closer in time to the collection time of the initial sequence — suggesting that the Qnet does indeed capture realistic constraints.

Significance Test for Population Membership

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

Definition 2 (Membership probability of a sequence)**.** *Given a population P inducing the Qnet Φ^P and a sequence x, we can compute the membership probability of x:*

$$\omega_x^P \triangleq Pr(x \in P) = \prod_{j=1}^{N} \left(\Phi_j^P(x_{-j})|_{x_j}\right) \tag{3}$$

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

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

Definition 3 (Membership degree)**.** *Let X be a random field representing a population P, ie.. $X = x$ is a randomly drawn sequence from P. Then the membership degree ω^P is a function of the random variable X:*

$$\omega^P(X) \triangleq \prod_{j=1}^{N} \left(\Phi_j^P(X_{-j})|_{X_j}\right) \tag{4}$$

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

Since $\omega^P(X)$ is a random variable, we can now compute sets of sequences that better represent the population P, and ones that are on the fringe. We can also evaluate using a pre-specified significance-level if a particular sequence is not from the population P, thus identifying if we need to recompute the predictors Φ, or split the base population. We can set up a hypothesis testing scenario to determine if sequences are indeed from a test population, as follows:

Given a population P, inducing a Qnet Φ^P, and a sequence x, we assume the null hypothesis is $x \notin P$. We reject the null hypothesis at a pre-specified significance α, if

$$Pr(\omega^P(X) \geqq \omega^P(X = x)) \leqq \alpha \tag{5}$$

The fraction of newly observed sequences that do not reject the null hypothesis can then be used as an estimate of the species-specific divergence in population characteristics.

Proof of Probability Bounds

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

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

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

Proof. Using Sanov's theorem[1] on large deviations, we conclude that the probability of spontaneous jump from strain $x \in P$ to strain $y \in Q$, with the possibility $P \neq Q$, is given by:

$$Pr(x \to y) = \prod_{i=1}^{N} \left(\Phi_i^P(x_{-i})|_{y_i}\right) \tag{7}$$

Writing the factors on the right hand side as:

$$\Phi_i^P(x_{-i})|_{y_i} = \Phi_i^Q(y_{-i})|_{y_i} \left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) \tag{8}$$

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

$$|\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}| \leqq \sum_{y_i \in \Sigma_i} |\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}| \tag{9}$$

Using a standard refinement of Pinsker's inequality[2], and the relationship of Jensen-Shannon divergence with

total variation, we get:

$$\theta_i \geqq \frac{1}{8}|\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}|^2 \Rightarrow \left|1 - \frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}}\right| \leqq \frac{1}{a_0}\sqrt{8\theta_i} \tag{10}$$

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

$$\log\left(\prod_{i=1}^{N}\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) = \sum_i \log\left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) \geqq \sum_i\left(1 - \frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}}\right) \geqq \frac{\sqrt{8}}{a_0}\sum_i \theta_i^{1/2} = -\frac{\sqrt{8}N}{a_0}\theta \tag{11}$$

Similarly, the upper bound may be derived as:

$$\log\left(\prod_{i=1}^{N}\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) = \sum_i \log\left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) \leqq \sum_i\left(\frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}} - 1\right) \leqq \frac{\sqrt{8}N}{a_0}\theta \tag{12}$$

Combining Eqs. 11 and 12, we conclude:

$$\omega_y^Q e^{\frac{\sqrt{8}N}{a_0}\theta} \geqq Pr(x \to y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N}{a_0}\theta} \tag{13}$$

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

$$a_0 \approx (1-\alpha)/N \tag{14}$$

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

$$\omega_y^Q e^{\frac{\sqrt{8}N^2}{1-\alpha}\theta} \geqq Pr(x \to y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N^2}{1-\alpha}\theta} \tag{15}$$

□

Remark 1. *This bound can be rewritten in terms of the log-likelihood of the spontaneous jump and constants independent of the initial sequence x as:*

$$|\log Pr(x \to y) - C_0| \leqq C_1\theta \tag{16}$$

where the constants are given by:

$$C_0 = \log\omega_y^Q \tag{17}$$

$$C_1 = \frac{\sqrt{8}N^2}{1-\alpha} \tag{18}$$

Multivariate Regression to Identify Factors in Strain Prediction

We investigate the key factors that contribute to our successful prediction of the dominant strain in the next season. We carry out a multivariate regression with data diversity, the complexity of inferred Qnet and the edit distance of the WHO recommendation from the dominant strain as independent variables. Here we define data diversity as the number of clusters we have in the input set of sequences, such that any two sequences five or less mutations apart are in the same cluster. Qnet complexity is measured by the number of decision nodes in the component decision trees of the recursive forest.

We select several plausible structures of the regression equation, and in each case conclude that data diversity has the most important and statistically significant contribution (See SI-Tab. 17).

REFERENCES

[1] Cover TM, Thomas JA. Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). New York, NY, USA: Wiley-Interscience; 2006.

[2] Fedotov AA, Harremoës P, Topsoe F. Refinements of Pinsker's inequality. IEEE Transactions on Information Theory. 2003;49(6):1491–1498.

SUPPLEMENTARY FIGURES & TABLES

SI Tab. 1
EXAMPLES: QNET INDUCED DISTANCE VARYING FOR FIXED SEQUENCE PAIR WHEN BACKGROUND POPULATION CHANGES (ROWS 1 -5), SEQUENCES WITH SMALL EDIT DISTANCE AND LARGE Q-DISTANCE, AND THE CONVERSE (ROWS 6-9)

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

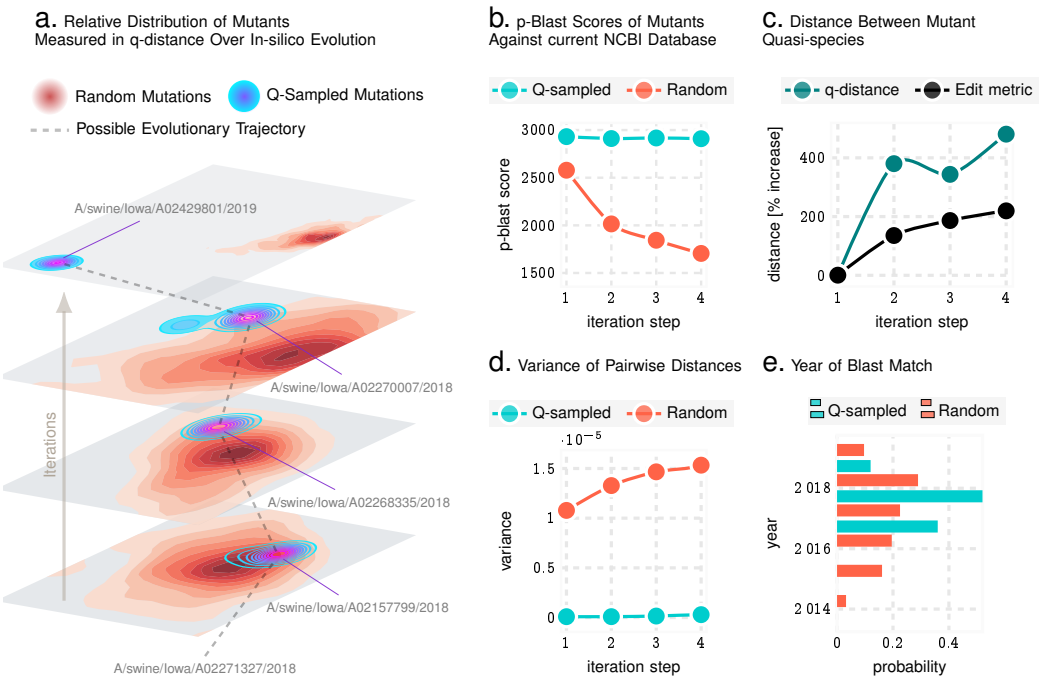
*Year A and year B correspond to the assumed collection years for sequences A and B respectively for the purpose of this example. Sequence A in row 1 is collected in 2007, but is assumed to be from different years in rows 2-4 to demonstrate the change in q-distance from sequence B, arising only from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

SI Tab. 2
CORRELATION BETWEEN Q-DISTANCE AND EDIT DISTANCE BETWEEN SEQUENCE PAIRS

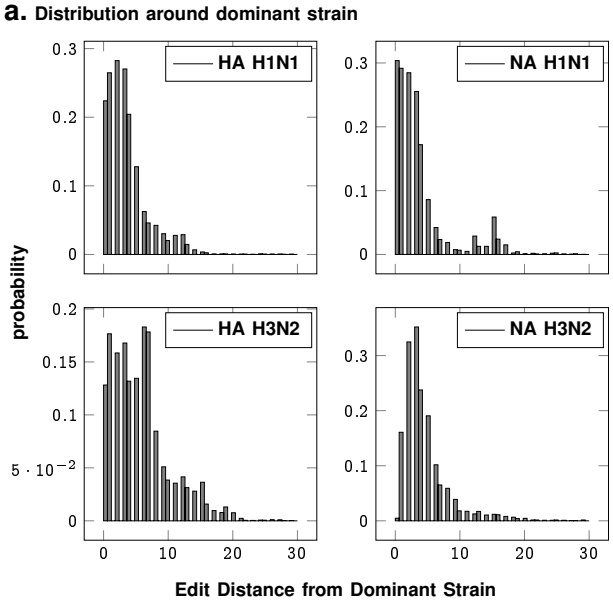
Phenotypes	Correlation
Influenza H1N1 HA	0.76
Influenza H1N1 NA	0.74
Influenza H3N2 HA	0.85
Influenza H3N2 NA	0.79

SI Tab. 3
NUMBER OF SEQUENCES COLLECTED FROM PUBLIC DATABASES

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



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



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

SI Tab. 4
H1N1 HA NORTHERN HEMISPHERE

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

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

SI Tab. 5
H1N1 HA SOUTHERN HEMISPHERE

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2001-02	A/New Caledonia/20/99	A/Canterbury/41/2001	A/South Canterbury/50/2000	4	6
2002-03	A/New Caledonia/20/99	A/Taiwan/567/2002	A/Canterbury/41/2001	3	1
2003-04	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	5	2
2004-05	A/New Caledonia/20/99	A/Thailand/Siriraj-Rama-TT/2004	A/Memphis/5/2003	7	4
2005-06	A/New Caledonia/20/99	A/Niedersachsen/217/2005	A/Canterbury/106/2004	8	10
2006-07	A/New Caledonia/20/99	A/India/34980/2006	A/Niedersachsen/217/2005	6	2
2007-08	A/New Caledonia/20/99	A/Norway/1701/2007	A/Thailand/CU68/2006	14	6
2008-09	A/Solomon Islands/3/2006	A/Pennsylvania/02/2008	A/Kentucky/UR06-0476/2007	9	2
2009-10	A/Brisbane/59/2007	A/Singapore/ON1060/2009	A/Belem/241/2008	119	119
2010-11	A/California/7/2009	A/England/01220740/2010	A/Singapore/ON1060/2009	5	1
2011-12	A/California/7/2009	A/Punjab/041/2011	A/England/01220740/2010	7	2
2012-13	A/California/7/2009	A/British Columbia/001/2012	A/Punjab/041/2011	11	4
2013-14	A/California/7/2009	A/Moscow/CRIE-32/2013	A/India/P122045/2012	10	5
2014-15	A/California/7/2009	A/Thailand/CU-C5169/2014	A/Jiangsuhailing/SWL1382/2013	12	4
2015-16	A/California/7/2009	A/Georgia/15/2015	A/Thailand/CU-C5169/2014	14	2
2016-17	A/California/7/2009	A/Hawaii/21/2016	A/Georgia/15/2015	16	2
2017-18	A/Michigan/45/2015	A/Michigan/291/2017	A/Beijing-Huairou/SWL1335/2016	5	4
2018-19	A/Michigan/45/2015	A/Washington/55/2018	A/Michigan/291/2017	6	1
2019-20	A/Michigan/45/2015	A/Kentucky/06/2019	A/Washington/55/2018	7	1
2020-21	A/Brisbane/02/2018	A/Togo/905/2020	A/Italy/8451/2019	10	8
2021-22	A/Victoria/2570/2019	A/Abidjan/457/2021	A/Togo/35/2021	9	4
2022-23	-1	-1	A/Cote_D'Ivoire/1270/2021	-1	-1

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

SI Tab. 6
H1N1 NA NORTHERN HEMISPHERE

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

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

SI Tab. 7
H1N1 NA SOUTHERN HEMISPHERE

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2001-02	A/New Caledonia/20/99	A/New York/447/2001	A/Canterbury/37/2000	4	6
2002-03	A/New Caledonia/20/99	A/Paris/0833/2002	A/New York/447/2001	1	5
2003-04	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	3	5
2004-05	A/New Caledonia/20/99	A/Singapore/14/2004	A/Memphis/5/2003	2	3
2005-06	A/New Caledonia/20/99	A/Taiwan/5524/2005	A/Canterbury/106/2004	3	6
2006-07	A/New Caledonia/20/99	A/Massachusetts/08/2006	A/Sofia/361/2005	4	2
2007-08	A/New Caledonia/20/99	A/Tennessee/UR06-0106/2007	A/Thailand/RMSC-UDN-20/2006	4	8
2008-09	A/Solomon Islands/3/2006	A/Sendai/TU66/2008	A/Tennessee/UR06-0151/2007	15	13
2009-10	A/Brisbane/59/2007	A/Thailand/SR08021/2009	A/Nebraska/07/2008	87	87
2010-11	A/California/7/2009	A/Finland/2460N/2010	A/Rome/709/2009	2	9
2011-12	A/California/7/2009	A/Tula/CRIE-GSYu/2011	A/Finland/2460N/2010	4	2
2012-13	A/California/7/2009	A/Bangalore/697-32/2012	A/Tula/CRIE-GSYu/2011	4	0
2013-14	A/California/7/2009	A/Jiangsugusu/SWL1824/2013	A/Oman/SQUH-63/2012	5	4
2014-15	A/California/7/2009	A/LongYan/SWL2457/2014	A/NanPing/SWL1640/2013	9	6
2015-16	A/California/7/2009	A/Michigan/45/2015	A/LongYan/SWL2457/2014	14	5
2016-17	A/California/7/2009	A/Mexico/4436/2016	A/Michigan/45/2015	14	0
2017-18	A/Michigan/45/2015	A/Illinois/37/2017	A/Mexico/4436/2016	3	3
2018-19	A/Michigan/45/2015	A/Kenya/47/2018	A/Kentucky/26/2017	4	2
2019-20	A/Michigan/45/2015	A/Texas/7939/2019	A/Kenya/47/2018	4	0
2020-21	A/Brisbane/02/2018	A/Togo/897/2020	A/Texas/7939/2019	6	5
2021-22	A/Victoria/2570/2019	A/Cote_D'Ivoire/1496/2021	A/Togo/0155/2021	1	7
2022-23	-1	-1	A/Dakar/35/2021	-1	-1

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

SI Tab. 8
H3N2 HA NORTHERN HEMISPHERE

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2005-06	A/California/7/2004	A/Denmark/195/2005	A/Tairawhiti/369/2004	10	2
2006-07	A/Wisconsin/67/2005	A/New York/5/2006	A/South Australia/22/2005	5	4
2007-08	A/Wisconsin/67/2005	A/Tennessee/11/2007	A/Colorado/05/2006	8	5
2008-09	A/Brisbane/10/2007	A/Massachusetts/13/2008	A/Virginia/UR06-0021/2007	3	2
2009-10	A/Brisbane/10/2007	A/Hawaii/14/2009	A/Manhean/03/2008	7	6
2010-11	A/Perth/16/2009	A/Utah/12/2010	A/Philippines/5/2009	8	7
2011-12	A/Perth/16/2009	A/Piaui/14202/2011	A/Singapore/C2010.310/2010	4	4
2012-13	A/Victoria/361/2011	A/Alborz/927/2012	A/Tehran/895/2012	4	3
2013-14	A/Victoria/361/2011	A/Delaware/01/2013	A/Singapore/H2012.934/2012	4	1
2014-15	A/Texas/50/2012	A/Alborz/72205/2014	A/Nebraska/03/2013	10	9
2015-16	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Ontario/01/2014	10	0
2016-17	A/Hong Kong/4801/2014	A/Guangdong/12/2016	A/Oregon/02/2015	0	0
2017-18	A/Hong Kong/4801/2014	A/Maryland/25/2017	A/New York/03/2016	3	1
2018-19	A/Singapore/INFIMH-16-0019/2016	A/Vermont/04/2018	A/Ontario/038/2017	8	5
2019-20	A/Kansas/14/2017	A/Kentucky/27/2019	A/California/7330/2018	16	12
2020-21	A/Hong Kong/2671/2019	A/India/Pun-NIV289524/2021_Jan	A/California/NHRC-OID_FDX100215/2019	16	14
2021-22	A/Cambodia/e0826360/2020	A/Human/New_York/PV60641/2022	A/India/Pun-NIV291000/2021_Jan	14	5
2022-23	-1	-1	A/Denmark/370/2022	-1	-1

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

SI Tab. 9
H3N2 HA SOUTHERN HEMISPHERE

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2005-06	A/Wellington/1/2004	A/Denmark/195/2005	A/Waikato/21/2004	3	3
2006-07	A/California/7/2004	A/New York/5/2006	A/South Australia/22/2005	12	4
2007-08	A/Wisconsin/67/2005	A/Tennessee/11/2007	A/New York/923/2006	8	5
2008-09	A/Brisbane/10/2007	A/Massachusetts/13/2008	A/Tennessee/11/2007	3	2
2009-10	A/Brisbane/10/2007	A/Hawaii/14/2009	A/Manhean/03/2008	7	6
2010-11	A/Perth/16/2009	A/Utah/12/2010	A/Hawaii/14/2009	8	7
2011-12	A/Perth/16/2009	A/Piaui/14202/2011	A/Utah/12/2010	4	4
2012-13	A/Perth/16/2009	A/Alborz/927/2012	A/Piaui/14202/2011	8	4
2013-14	A/Victoria/361/2011	A/Delaware/01/2013	A/Callao/IPE00830/2012	4	7
2014-15	A/Texas/50/2012	A/Alborz/72205/2014	A/Delaware/01/2013	10	7
2015-16	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Alborz/72205/2014	10	0
2016-17	A/Hong Kong/4801/2014	A/Guangdong/12/2016	A/Parma/471/2015	0	0
2017-18	A/Hong Kong/4801/2014	A/Maryland/25/2017	A/Ontario/196/2016	3	4
2018-19	A/Singapore/INFIMH-16-0019/2016	A/Vermont/04/2018	A/Texas/279/2017	8	5
2019-20	A/Switzerland/8060/2017	A/Kentucky/27/2019	A/Santa Catarina/1200/2018	13	12
2020-21	A/South Australia/34/2019	A/India/Pun-NIV289524/2021_Jan	A/Kentucky/27/2019	12	14
2021-22	A/Hong Kong/2671/2019	A/Darwin/9a/2021	A/India/PUN-NIV301718/2021	19	1
2022-23	-1	-1	A/Saint-Martin/00754/2022	-1	-1

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

SI Tab. 10
H3N2 NA NORTHERN HEMISPHERE

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2003-04	A/Moscow/10/99	A/Denmark/107/2003	A/New York/100/2002	13	3
2004-05	A/Fujian/411/2002	A/Hyogo/36/2004	A/New York/20/2003	3	16
2005-06	A/California/7/2004	A/Denmark/203/2005	A/Hong Kong/HKU20/2004	4	0
2006-07	A/Wisconsin/67/2005	A/Berlin/32/2006	A/Mexico/InDRE2227/2005	1	1
2007-08	A/Wisconsin/67/2005	A/Brazil/80/2007	A/Baden-Wuerttemberg/17/2006	8	7
2008-09	A/Brisbane/10/2007	A/Missouri/05/2008	A/Washington/01/2007	3	2
2009-10	A/Brisbane/10/2007	A/Oklahoma/09/2009	A/Wisconsin/24/2008	3	1
2010-11	A/Perth/16/2009	A/California/17/2010	A/New York/70/2009	2	3
2011-12	A/Perth/16/2009	A/Texas/14/2011	A/California/14/2010	3	2
2012-13	A/Victoria/361/2011	A/New York/02/2012	A/Singapore/C2011.493/2011	4	1
2013-14	A/Victoria/361/2011	A/Michigan/02/2013	A/New York/01/2012	3	1
2014-15	A/Texas/50/2012	A/Tehran/69634/2014	A/Boston/DOA2-176/2013	3	1
2015-16	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Thailand/CU-B10520/2014	3	0
2016-17	A/Hong Kong/4801/2014	A/North Carolina/62/2016	A/Delaware/02/2015	7	2
2017-18	A/Hong Kong/4801/2014	A/Texas/277/2017	A/New York/03/2016	8	0
2018-19	A/Singapore/INFIMH-16-0019/2016	A/Japan/NHRC_FDX70352/2018	A/Colorado/11/2017	4	3
2019-20	A/Kansas/14/2017	A/Washington/9757/2019	A/Guangxi-Fangcheng/54/2019	3	11
2020-21	A/Hong Kong/2671/2019	A/Bangladesh/1004005/2020	A/Maryland/02/2019	3	13
2021-22	A/Cambodia/e0826360/2020	A/Stockholm/10/2022	A/Darwin/9/2021	2	2
2022-23	-1	-1	A/Michigan/UOM10042819294/2021	-1	-1

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

SI Tab. 11
H3N2 NA SOUTHERN HEMISPHERE

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2003-04	A/Moscow/10/99	A/Denmark/107/2003	A/New York/101/2002	13	3
2004-05	A/Fujian/411/2002	A/Hyogo/36/2004	A/New York/20/2003	3	16
2005-06	A/Wellington/1/2004	A/Denmark/203/2005	A/Wellington/1/2004	2	2
2006-07	A/California/7/2004	A/Berlin/32/2006	A/Mexico/InDRE2227/2005	3	1
2007-08	A/Wisconsin/67/2005	A/Brazil/80/2007	A/Ohio/06/2006	8	10
2008-09	A/Brisbane/10/2007	A/Missouri/05/2008	A/Brazil/80/2007	3	2
2009-10	A/Brisbane/10/2007	A/Oklahoma/09/2009	A/Wisconsin/24/2008	3	1
2010-11	A/Perth/16/2009	A/California/17/2010	A/New York/70/2009	2	3
2011-12	A/Perth/16/2009	A/Texas/14/2011	A/Virginia/05/2010	3	2
2012-13	A/Perth/16/2009	A/New York/02/2012	A/Texas/14/2011	4	1
2013-14	A/Victoria/361/2011	A/Michigan/02/2013	A/New York/02/2012	3	3
2014-15	A/Texas/50/2012	A/Tehran/69634/2014	A/Michigan/02/2013	3	1
2015-16	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Tehran/69634/2014	3	2
2016-17	A/Hong Kong/4801/2014	A/North Carolina/62/2016	A/Parma/471/2015	7	2
2017-18	A/Hong Kong/4801/2014	A/Texas/277/2017	A/Guangdong/264/2016	8	0
2018-19	A/Singapore/INFIMH-16-0019/2016	A/Japan/NHRC_FDX70352/2018	A/Texas/277/2017	4	3
2019-20	A/Switzerland/8060/2017	A/Washington/9757/2019	A/Pennsylvania/317/2018	10	10
2020-21	A/South Australia/34/2019	A/Bangladesh/1004005/2020	A/Washington/9757/2019	1	13
2021-22	A/Hong Kong/2671/2019	A/India/PUN-NIV301718/2021	A/Darwin/11/2021	6	1
2022-23	-1	-1	A/Texas/12723/2022	-1	-1

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

SI Tab. 12
H1N1 NA NORTHERN HEMISPHERE (MULTI-CLUSTER)

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

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

SI Tab. 13
H1N1 NA SOUTHERN HEMISPHERE (MULTI-CLUSTER)

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

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

SI Tab. 14
H3N2 NA NORTHERN HEMISPHERE (MULTI-CLUSTER)

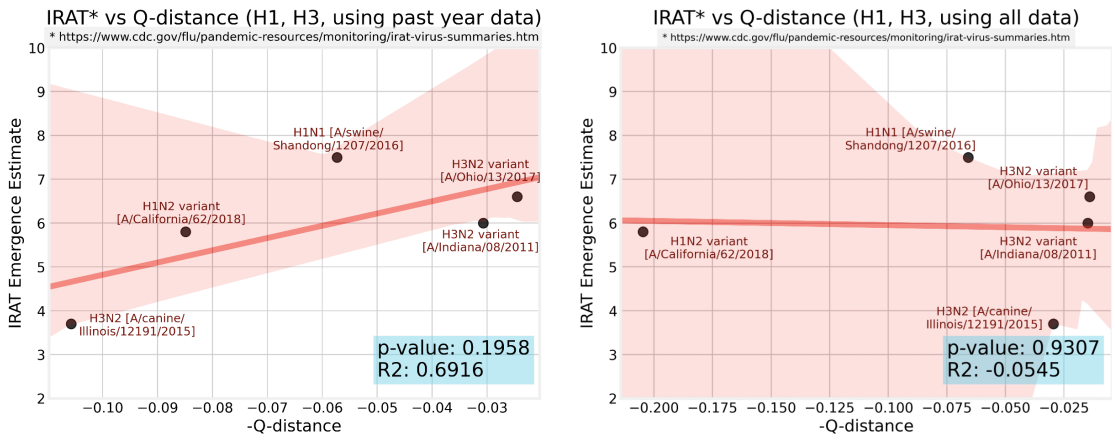
Year	WHO Recommendation	WHO Error	Qnet Error 1	Qnet Error 2	Qnet Recommendation 1	Qnet Recommendation 2
2003-04	A/Moscow/10/99	13	4	5	A/Auckland/612/2002	A/New York/87/2002
2004-05	A/Fujian/411/2002	3	16	18	A/New York/20/2003	A/New York/12/2003
2005-06	A/California/7/2004	4	1	7	A/New York/358/2004	A/Singapore/36/2004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

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

SI Tab. 15
H3N2 NA SOUTHERN HEMISPHERE (MULTI-CLUSTER)

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

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



SI Fig. 3. **IRAT vs. Q-distance relationship for H1- and H3- sub-types, using past year data vs. using all data.** On the result when computing average q-distance between the target strain and the circulating human strains from the past year, and on the right is the result when using all available human strains of that sub-type. Evidently, the former has a much higher correlation, since a strain being "close" to humans at some point does not necessarily mean being close now.

SI Tab. 16
INFLUENZA A STRAINS EVALUATED BY IRAT AND CORRESPONDING QNET COMPUTED RISK SCORES

Influenza Virus	Subype	IRAT Date	IRAT Emergence Score	IRAT Impact Score	HA Qnet Sample	NA Qnet Sample	HA Avg. Q-dist	NA Avg. Q-dist.	Both Avg. Q-dist.
A/swine/Shandong/1207/2016	H1N1	Jul 2020	7.5	6.9	1000	1000	0.094135	0.020530	0.057332
A/duck/New York/1996	H1N1	Nov 2011	2.3	2.4	1000	1000	-1	-1	-1
A/California/62/2018	H1N2	Jul 2019	5.8	5.7	55	55	0.108902	0.060951	0.084926
A/Ohio/13/2017	H3N2	Jul 2019	6.6	5.8	1000	1000	0.018431	0.030623	0.024527
A/Indiana/08/2011	H3N2	Dec 2012	6.0	4.5	1000	1000	0.052311	0.009103	0.030707
A/canine/Illinois/12191/2015	H3N2	Jun 2016	3.7	3.7	1000	1000	0.060665	0.150856	0.105761
A/American wigeon/South Carolina/AH0195145/2021	H5N1	Mar 2022	4.4	5.1	335	323	0.172180	0.511387	0.337368
A/American green-winged teal/Washington/1957050/2014	H5N1	Mar 2015	3.6	4.1	326	314	0.191127	0.448228	0.316856
A/Vietnam/1203/2004	H5N1	Nov 2011	5.2	6.6	258	246	0.167254	0.011074	0.091342
A/Northern pintail/Washington/40964/2014	H5N2	Mar 2015	3.8	4.1	-1	-1	-1	-1	-1
A/Sichuan/06681/2021	H5N6	Oct 2021	5.3	6.3	45	45	0.361591	0.051794	0.206692
A/Yunnan/14564/2015	H5N6	Apr 2016	5.0	6.6	16	16	-1	-1	-1
A/Astrakhan/3212/2020	H5N8	Mar 2021	4.6	5.2	-1	-1	-1	-1	-1
A/gyrfalcon/Washington/41088/2014	H5N8	Mar 2015	4.2	4.6	-1	-1	-1	-1	-1
A/Netherlands/219/2003	H7N7	Jun 2012	4.6	5.8	46	46	0.275671	0.352060	0.313455
A/turkey/Indiana/1573-2/2016	H7N8	Jul 2017	3.4	3.9	-1	-1	-1	-1	-1
A/chicken/Tennessee/17-007431-3/2017	H7N9	Oct 2017	3.1	3.5	496	495	0.102686	0.256855	0.179693
A/chicken/Tennessee/17-007147-2/2017	H7N9	Oct 2017	2.8	3.5	496	495	0.209532	0.254090	0.231788
A/Hong Kong/125/2017	H7N9	May 2017	6.5	7.5	437	437	0.029599	0.005775	0.017687
A/Shanghai/02/2013	H7N9	Apr 2016	6.4	7.2	178	178	0.005495	0.003556	0.004525
A/Bangladesh/0994/2011	H9N2	Feb 2014	5.6	5.4	13	12	-1	-1	-1
A/Anhui-Lujiang/39/2018	H9N2	Jul 2019	6.2	5.9	31	30	0.029024	0.168090	0.098557
A/Jiangxi-Donghu/346/2013	H10N8	Feb 2014	4.3	6.0	-1	-1	-1	-1	-1

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

SI Tab. 17
GENERAL LINEAR MODEL FOR EVALUATING EFFECT OF DATA DIVERSITY ON QNET PERFORMANCE

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

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


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

SI Tab. 18
NUMBERING CONVERSION TO PDM09 AND H3 SCHEMES

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