Rationale: Influenza A, partly on account of its segmented genome and its wide prevalence in animal hosts, has the ability to incorporate genes from multiple strains and (re)emerge as novel human pathogens[1,2], thus harboring a high pandemic potential. Strains spilling over into humans from animal reservoirs is thought to have triggered mild to devastating pandemics at least 4 times (1918 Spanish flu/H1N1, 1957 Asian flu/H2N2, 1968 Hong Kong flu/H3N2, 2009 swine flu/H1N1) in the past century[3]. One approach to mitigating such risk is to recognize animal strains that do not yet circulate in humans, but are likely to spill-over and quickly achieve human-to-human (HH) transmission capability. While global surveillance efforts collect wild specimens from diverse hosts/locations annually, our ability to reliably and scalably risk-rank individual strains remains limited[4]. CDC's current solution to this problem is the Influenza Risk Assessment Tool (IRAT)[5]. SMEs score strains based on the number of human infections, transmission in laboratory animals, receptor binding, population immunity, genomic analysis, antigenic relatedness, global prevalence, pathogenesis, and treatment options, which are averaged to obtain 2 scores (between 1 and 10) that estimate 1) the emergence risk and 2) the potential public health impact on sustained transmission. IRAT scores depend on multiple experimental assays, taking weeks/months to compile for a single strain. With tens of thousands of strains being collected annually, this results in a scalability bottleneck.

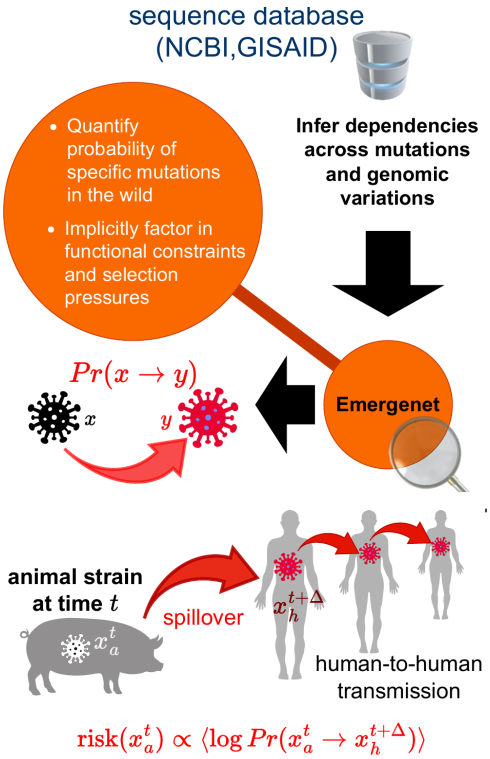

Fig. 1. Theoretical foundation of BioNORAD

Here we plan to develop a platform powered by novel pattern discovery algorithms to automatically parse out emergent evolutionary constraints operating on Influenza A viruses in the wild, to provide a less-heuristic theory-backed scalable solution to emergence prediction. We plan to show that this capability enables preempting strains which are *expected to be in future human circulation*, and approximate IRAT scores of non-human strains without experimental assays or SME scoring, in seconds as opposed to weeks or months. Our approach automatically takes into account the time-sensitive variations in selection pressures as the background strain circulation evolves, and will potentially be able to rank-order strains adaptively.

Additionally, we plan to validate our ability to predict future variations of viral proteins by showing that predicted variants of HA are functional, and maintain replicative fitness in cell cultures. Thus, bringing together rigorous data-driven modeling, and validation via tools from reverse genetics we plan to deliver an actionable and deployable platform (the BioNORAD) that optimally exploits the current biosurvellance capacity, *identifying when and where an imminent emergence event is likely, and if such novel strains are likely to achieve human-to-human transmission capability.*

Hypotheses: ***FY23 PRMRP Portfolio Category: Infectious Diseases | FY23 PRMRP Topic: proteomics | FY23 PRMRP Strategic Goal: Epidemiology: Identify strategies for surveillance or develop modeling tools and/or biomarkers to predict outbreaks or epidemics*** Our key hypotheses are as follows:

☐ 1) Learning cross-dependencies across point mutations in key proteins, *e.g.* those implicated in viral entry/exit (HA and NA) reveals the underlying rules of organization of their primary structures to forecast evolutionary trajectories, *i.e.*, predict future mutations and likelihood of jump events for animal strains. Importantly, while individual sequences do not encode enough information to predict its future mutations, discovering patterns from large sequence databases make such forecasts possible.

☐ 2) Current bio-surveillance produces sufficient data for meaningful pattern discovery, and inferred patterns of change can be assembled into an early warning system for pandemic threats, thus serving a similar function to the strategic goal of NORAD in the context of defending against geospatial pandemic threats, as opposed to protecting US airspace from adversarial intrusion.

Specific Aims: We have three key aims described below:

☐ **Aim 1: Formulate a novel metric of sequence similarity (E-distance) that reflects potential of spontaneous jump.** Devise a biologically meaningful metric for comparing two genomic sequences, that scales with the probability of one sequence spontaneously replicating to give rise to the other in the wild,

under a) realistic, b) time-dependent, and c) poorly understood selection pressures. Within this aim, we will deliver the implemented algorithm identifying the E-distance metric as function of sub-type, gene, region and time, which is demonstrably distinct from the classical edit distance. Since the E-distance reflects the odds of one sequence mutating to another, it is a function of not just how many mutations the two sequences are apart (the edit distance), but also how specific mutations incrementally affect fitness, and how possibly non-colocated mutations have emergent dependencies and compensatory epistatic effects. Without these explicit constraints, assessing a strain-specific jump-likelihood is open to subjective guesswork (current state-of-art). Our aim here is to show that a precise probabilistic calculation is theoretically possible and practically feasible, enabling an actionable framework for tracking evolutionary change. The major tasks within this aim are as follows: **T1.1 (Emergenet Development)** consisting of sub-tasks: (T1.1.1) Precisely formulate the Emergenet inference platform, that puts together ML algorithms capturing maximally predictive patterns of change and mutational dependencies, and (T1.1.2) provide uncertainty quantification for the inferred patterns. **T1.2 (Sample-complexity Estimation)** Investigate sample complexity of Emergenet, i.e., how much data is needed to reliably identify patterns. **T1.3 (Event Timeline Estimation)** comprising 2 subtasks: (T1.3.1) Map mutational change dynamics to "wall-time", to forecast *when* future variants will show up, and (T1.3.2) validate timeline predictions using records of past emergence events, by assessing the time-delay between Emergenet predictions observation of predicted mutations in historical strain populations.

□ **Aim 2: Validate E-distance as a similarity metric that can identify biologically meaningful sequence variations.** We aim to show that the E-distance may be used to differentiate between random perturbations in the genome (most of which would be deleterious, and not code for a viable protein), and perturbations that are biologically viable. This is a key distinguishing capability of the Emergenet platform, that can reliably identify possible future mutations, along with their precisely quantified likelihoods. We will demonstrate that perturbations predicted using this metric leads to viable and functional proteins. The major tasks within this aim are: **T2.1 (Quantify asymmetric transition probabilities between strains)** Key subtasks are: (T2.1.1) Infer probabilistic movement direction between strains, delineating the asymmetry of jump likelihood across strains, and (T2.1.2) chart multi-hop probabilistic trajectories from observed strains. **T2.2 (Laboratory Experiments for assessing fitness of predicted variants in cell culture)** consisting of the following subtasks: (T2.2.1) Shortlist HA variants with maximal emergence probability of H3N2 and H1N1 subtypes, (T2.2.2) Generate predicted HA variants using reverse genetics in human lung epithelial cell line (A549) and primary human lung cells, and (T2.2.3) evaluate generated variants for replicative fitness. These experimental assays will aim to demonstrate that strains with small E-distance from observed strains leads to viable variants, and that even small random perturbations causes a catastrophic fall in fitness.

□ **Aim 3: Develop a working implementation of BioNORAD.** We aim to demonstrate a prototype of the BioNORAD platform for analyzing Influenza A strains at scale for emergence and impact risk. Major tasks are : **T3.1 (IRAT score replication)** Replicate the published IRAT scores, along with uncertainty quantification, within seconds as a validation result. Key subtasks are: (T3.1.1) Investigate how each of the ten IRAT dimensions map to our Emergenet based risk, (T3.1.2) Evaluate if the IRAT scores would change if evaluated at different times, and (T3.1.3) incorporate timeline estimation in BioNORAD prototype to predict time to emergence. T3.2 (BioNORAD Results for Current/Recent Surveillance Data) which comprises the subtasks: (T3.2.1) analyze collected sequences at scale, by enumerating the risk profiles of all sequences collected recently within the last few years, and any new sequences that continue to be submitted to NCBI and GISAID. And (T3.2.2) set up an automated pipeline that pulls in sequence data of for new submissions, and publish a risk score automatically. We will collate this information in our pipeline to map the global risk, visualizing where and when an emergence event is likely, and for which strain/subtype/animal hosts.

Research Strategy and Feasibility: Our approach aims to reliably estimate the non-heuristic numerical probability $\Pr(x \to y)$ of a strain x spontaneously giving rise to y in the wild, thus preempting strains expected to be in future circulation, and approximating IRAT scores of non-human strains without detailed experimental assays or SME scoring. We plan to accomplish this by learning the complex cross-dependencies that constrain what a "valid alteration" of a AA sequence is, by first analyzing variations (point substitutions, indels) of residue sequences of key proteins implicated in cellular entry/exit[3,6], namely HA and NA, and then expanding the analysis to the complete viral genome. By representing these constraints within a predictive framework – the Emergenet (Enet) – we will estimate the odds of a specific mutation to arise in future, and consequently the probability of a specific strain spontaneously evolving into another (Fig. 1). For such explicit calculations we must first infer the variation of mutational probabilities and

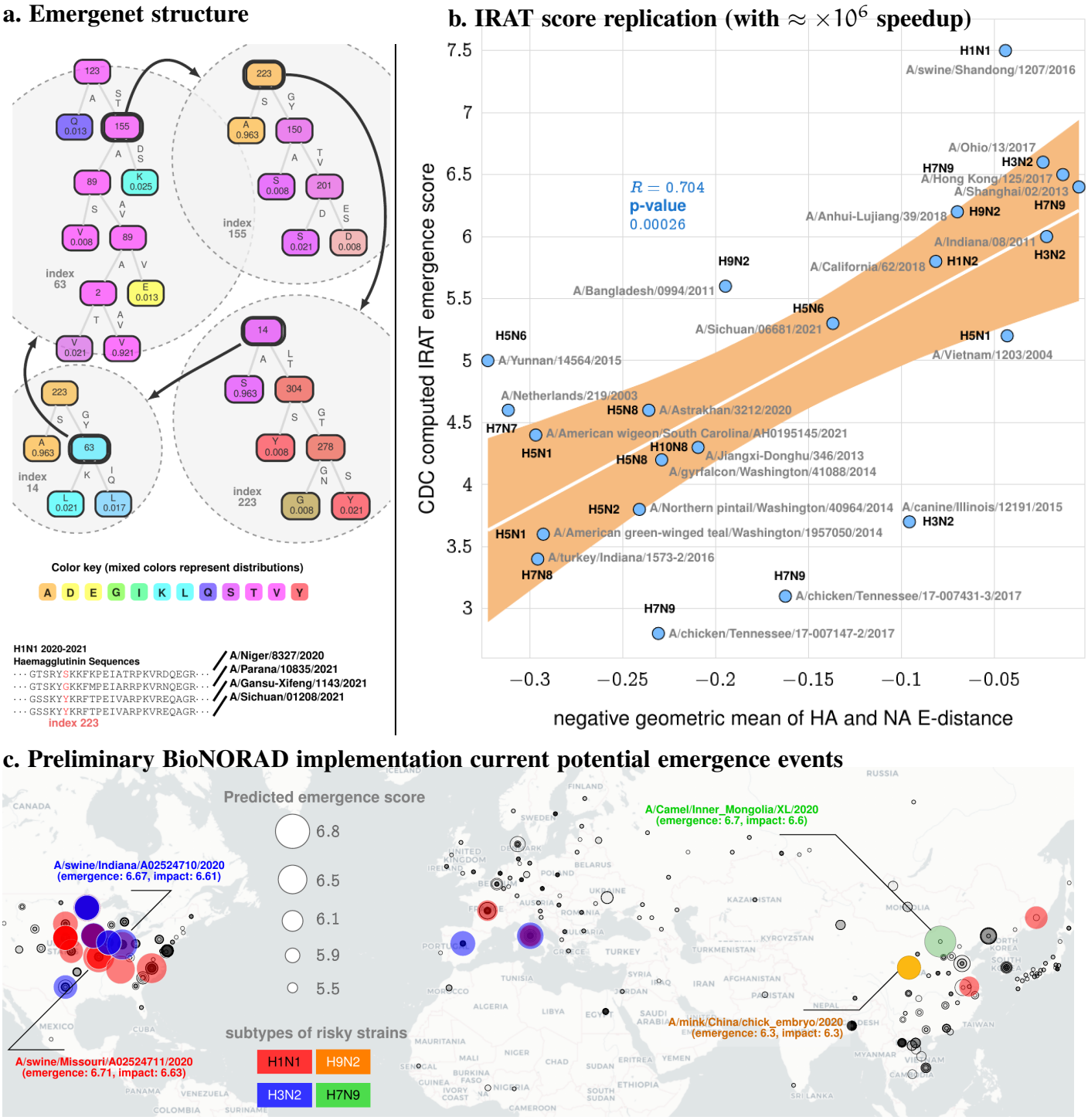

Fig. 2. a, a fragment of an example Emergenet showing the emergent dependencies captured. b, IRAT replication in preliminary analysis with only HA sequence data, c, a preliminary BioNORAD implementation with sequences collected over 2021-2022, showing the emerging threat-centers, subtypes.

potential residue replacements from one positional index to the next along the AA sequence. The many well-known classical DNA substitution models[7] or phylogeny inference tools assuming constant species-wise mutational characteristics, are not applicable here. Similarly, newer algorithms such as FluLeap[8] which identifies host tropism and species-level jump-risk[9] do not allow for strain-specific assessment.

The dependencies we expect to uncover are shaped by a functional necessity of conserving replicative fitness. Strains must be sufficiently common to be recorded in surveillance, implying that the sequences from public databases that we train with have high fitness. Lacking kinetic proofreading, Influenza A integrates

faulty nucleotides at a relatively high rate ($10^{-3} - 10^{-4}$) during replication[10,11]. However, few variations are actually viable, leading to emergent dependencies between such mutations. Furthermore, these fitness constraints are time-varying. The background strain distribution, and selection pressure from the evolution of cytotoxic T lymphocyte epitopes[12–16] in humans can change quickly. Our preliminary studies[17] suggest that we now have enough number of curated sequences in public databases to learn models that automatically factor in the evolving host immunity, and the current background environment.

Structurally, our model structure (an Emergenet) comprises an interdependent collection of local predictors, each aiming to predict the residue at a particular index using as features the residues at other indices (Fig. 2a). These individual predictors are implemented as conditional inference trees[18], in which nodal splits occur with a minimum pre-specified significance in differentiating the downstream child nodes. The set of residues acting as features in each such predictor is automatically identified, *e.g.*, in the fragment of the H1N1 HA Emergenet (2020-2021, Fig 2a), the predictor for residue 63 is dependent on residue 155, and the predictor for 155 is dependent on 223, the predictor for 223 is dependent on 14, and the residue at 14 is again dependent on 63, revealing a cyclic dependency. The complete Emergenet harbors a vast number of such relationships, and each internal node of a tree may be "expanded" to its own tree (since each index is "explained" by residues in other indices, which in turn have their own predictors). Owing to this recursive expansion, a complete Emergenet captures the intricate rules guiding evolutionary change as evidenced by our preliminary validations. Our research strategy comprises the following sequential steps:

(Step 1: Data and Emergenet Models) We will collect AA sequences for the key genes of all available strains from NCBI and GISAID databases, and construct Emergenets for each year, each gene and each chosen geographical region. Different spatial and temporal resolutions will be considered in the course of the project, and we will consider all 8 genes in the Influenza A genome: PB2, PB1, PA, HA, NP, NA, M and NS. With about two decades worth of data comprising nearly $> 380,000$ strains in NCBI and GISAID combined, when constructing models for each year, and the two hemispheres at a minimum, and the two subtypes H1N1 and H3N2, we end up with $8 \times 20 \times 2 \times 2 = 640$ Emergenet models. With new emerging subtypes, the number of inferred models will be higher, and these models together capture the totality of observable statistically significant patterns constraining Influenza A evolution.

(Step 2: E-distance metric calculation) Each Emergenet induces an intrinsic distance metric (E-distance) between strains, collected at the time and space associated with the model, defined as the square-root of JS divergence[19] of the conditional residue distributions, averaged over the sequence. Unlike the classical approach of measuring the number of edits between sequences, the E-distance is informed by the Emergenet-inferred dependencies, and adapts to the specific subtype, allele frequencies, and environmental variations. By our recent theoretical result[17], E-distance approximates the log-likelihood of spontaneous change $\mathtt{i.e.}$ $\log \Pr(\mathtt{x} \rightarrow \mathtt{y})$, and despite general correlation between E-distance and edit-distance, the E-distance between fixed strains can change if only the background environment changes.

(Step 3: Validation) We aim to validate the predictive ability of the Emergenet framework in two ways: a) predict dominant strains in seasonal epidemics and compare against historical WHO vaccine recommendations in how well we can preempt the actual circulation in future seasons in the two hemispheres, and b) carry out experimental assays in cell cultures to demonstrate that a perturbed strain is functional if the E-distance between the original and perturbed strain is small.

Seasonal strain-forecast validation. WHO recommendations for the flu shot is formulated about 6-7 months in advance based on global circulation[20]. In preliminary studies, our Emergenet-informed forecasts using only the HA gene outperform WHO/CDC recommended flu vaccine compositions consistently over the past two decades, for both H1N1 and H3N2 subtypes, individually in the two hemispheres (which have distinct recommendations[21]). For H1N1 HA, the Emergenet recommendation outperforms WHO by 52.07% on average over the last two decades, and 59.83% on average in the last decade, and by 65.79% in the period 2015-2019 (5 years pre-COVID-19). For H3N2 HA, the Emergenet recommendation outperforms WHO by 42.39% on average over the last two decades, and 35.00% on average in the last decade, and by 41.85% in the period 2015-2019. Despite limiting ourselves to only genotypic information, our approach distills emergent fitness-preserving constraints that outperform or match reported DMS-augmented strategies[17,22].

Assessment of fitness potential emerging zoonotic IAV variants in cell culture. Potential HA variants predicted by the Emergenet models will be generated using the reverse genetics system and evaluated for fitness against parental strains. Briefly, HA segments with potential mutations will be obtained through synthetic gene synthesis. We will assess the relative cell surface expression of parental HA and variants by

flow cytometry and western blotting. Next, we will generate recombinant viruses carrying mutant HA using an established reverse genetics system[23–31] by the Manicassamy lab at UIowa, and validate the recombinant viruses by performing NGS sequencing. To assess the replication fitness of recombinant viruses, we will perform single cycle and multicycle replication assays human lung epithelial cell line (A549) and primary human lung cells[32] (sourced from third-party vendors). In addition, we will assess the fitness of individual mutants by fitness competition assay with parental virus (1:1) and determine the relative ratio by high resolution melting (HRM) analysis[33–35] . These studies will help us determine the accuracy of Emergenet framework in predicting pandemic potential variants with enhanced fitness. *Choice of Cell-lines:* A549 cells and primary human lung cells are frequently used in IAV experiments due to their relevance to human infection, susceptibility to IAV, reproducibility, ease of cultivation, and compatibility with various molecular and cellular techniques[36]. These cells, derived from human lung carcinoma and lung tissue respectively, serve as appropriate models for studying IAV pathogenesis, and host interactions, as they express key host factors[37,38], and are compatible with molecular and cellular techniques[39]. Dr. Manicassamy has > 15 years of experience in working with human and zoonotic influenza viruses, and safely handling various human pathogens under enhanced BSL2 and BSL3 conditions.

(Step 4: BioNORAD Development) Determining the numerical odds of a spontaneous jump $\Pr(x \to y)$ allows us to frame the problem of forecasting dominant strain(s), and that of estimating the pandemic potential of an animal strain as mathematical propositions (albeit with some simplifying assumptions), with approximate solutions (Fig. 1). We will demonstrate that a pandemic risk score of a novel strain may be estimated by the probability of it giving rise to a well-adapted human strain. Our preliminary results[17] enable our ability to estimate the pandemic potential of novel animal strains, via a time-varying E-risk score $\rho_t(x)$ for a strain x not yet found to circulate in human hosts. We show that: $\rho_t(x) \triangleq -\frac{1}{|H^t|}\sum_{y \in H^t} \theta^{[t]}(x,y)$ scales as the average log-likelihood of $\Pr(x \to y)$ where y is any human strain of a similar subtype to x, and $\theta^{[t]}$ is the E-distance informed by the Emergenet computed from recent human strains H_t at time t of the same subtype as x, observed over the past year. The structure of dependencies revealed by the Emergenet inference makes it possible to estimate $\rho_t(x)$ explicitly. In the course of the project we will expand this risk score analytics to estimate the time to emergence in addition to spatial localization of such risk.

Preliminary Validation of BioNORAD. We constructed Emergenet models for HA and NA sequences using subtype-specific human strains, typically collected within the year prior to the assessment date. For rare human sub-types (H1N2, H7N7), we considered all subtype-specific human strains collected up to the assessment date to infer our Emergenet. For subtypes with little or no recorded human strains (H5N2, H5N6, H5N8, H7N8, H9N2, H10N8), we constructed the Emergenet using all human strains that match the HA subtype alone. This addresses the issue of "unknown unknowns": allowing Emergenet to assess threats posed by not-yet-human strains. We compute the E-risk for both HA and NA sequences (using the above relationship), finally reporting their geometric mean as our estimated risk. Considering IRAT emergence scores of 22 strains published by the CDC, we found strong out-of-sample support (correlation: 0.704, pvalue < 0.00026, Fig. 2b). Importantly, each E-risk score is computable in approximately 6 seconds as opposed to weeks/months, suggesting a *six order of magnitude speedup*. In the proposed study we will expand our analyis to incorporate all 8 genes. We show a preliminary implementation of the BioNORAD in Fig. 2c for all 6,066 strains retrieved in 2021/22, showing the localization of near-term threat events.

Innovation: Reported approaches to "predicting" mutations assume various models of DNA or AA substitution[7,40–44] ignoring the impact of a varying background and selection pressures. Importantly, a higher edit-similarity between strains do not imply a high likelihood of a jump. Current surveillance paradigms, and studies on habitat encroachment, climate change, and other ecological factors[45–47] have not improved our ability to actionably quantify future risk of emergence of a specific strain from a specific host at a specific place[48]. Recent advances in predicting seasonal strains[22] also do not generalize to predicting emergence events, especially for strains that do not yet circulate in humans. This project innovates and envisions a path to acquiring this transformative capability, which is currently well-beyond the state-of-art: the Emergenet algorithm is the first of its kind to learn an appropriate biologically meaningful comparison metric from data, without assuming any model of DNA or AA substitution, or a genealogical tree a priori, enabling an actionable pandemic early warning system. However, the potential impact of this work is not limited to pandemic preparedness, and can foster a more nuanced understanding of how viruses evolve and adapt over time. This could lead to the development of new therapeutic strategies that specifically target the evolutionary pathways of viruses, potentially revolutionizing the treatment of viral infections.