Project Narrative

Rationale:: Animal influenza viruses emerging into humans have triggered devastating pandemics in the past. Yet, our ability to evaluate the pandemic potential of individual strains that do not yet circulate in humans, remains limited. Here we propose to develop an experimentally validated platform called the Emergenet (Enet), to predict in near-real-time where and when new variants of concern would emerge, using only observed sequences of key viral proteins, procured in ongoing global surveillance of Influenza A viruses. We bring together new machine learning algorithms customized to the problem at hand, key insights from information theory, evolutionary theory, epidemiology and precise statistical uncertainty quantification to develop a rigorous framework, to track evolutionary trajectories of pathogens through a complex, poorly characterized, and dynamically changing fitness landscape. Our deliverable is best described as the foundations for a platform akin to bio-NORAD, *identifying when and where an imminent emergence event is likely, and if such novel strains are likely to achieve human-to-human transmission capability.*

Influenza viruses constantly evolve[1], sufficiently altering surface protein structures to evade the prevailing host immunity, and cause the recurring seasonal epidemic. These periodic infection peaks claim a quarter to half a million lives[2] globally. Additionally, Influenza A, partly on account of its segmented genome and its wide prevalence in animal hosts, can easily incorporate genes from multiple strains and (re)emerge as novel human pathogens[3], thus harboring a high pandemic potential. Strains spilling over into humans from animal reservoirs is thought to have triggered pandemics at least four times (1918 Spanish flu/H1N1, 1957 Asian flu/H2N2, 1968 Hong Kong flu/H3N2, 2009 swine flu/H1N1) in the past 100 years[4]. One approach to mitigating such risk is to identify animal strains that do not yet circulate in humans, but is likely to spill-over and quickly achieve human-to-human (HH) transmission capability. While global surveillance efforts collect wild specimens from diverse hosts and geo-locations annually, our ability to objectively, reliably and scalably risk-rank individual strains remains limited[5]. The Center for Disease Control's (CDC) current solution to this problem is the Influenza Risk Assessment Tool (IRAT)[6], which relies on time-consuming proteomics and transmission assays and potentially subjective evaluations by subject matter experts, taking weeks to months to compile for each strain of concern. With tens of thousands of strains being sequenced annually, this results in a scalability bottleneck.

Here we plan to develop a platform powered by novel pattern discovery and recognition 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 second as opposed to weeks or months. Our approach automatically takes into account the time-sensitive variations in selection pressures as the background strain circulation changes over time, and will potentially be able to rank-order strains adapatively. Additionally, we plan to validate our ability to predict future variations of viral proteins by showing that predicted variants of HA and NA fold correctly, and are functional, binding to the relevant human receptors in in-vivo laboratory experiments. Thus, bringing together rigorous data-driven modeling, and validation via tools from reverse genetics we plan to deliver an actionable and deployable platform that optimally exploits the current biosurvellance capacity.

The BioNORAD platform will enable proactive and actionable global surveillance for emerging pandemic threats from Influenza A. This importance of the ability to preempt pandemic risk to the national interest of the United States cannot be overstated, especially in the context of protecting DoD assest and personnel deployed in potentially high risk centers of emergence. Additionally, the BioNORAD will enable preemptive action including the inoculation of animal reservoirs before the first human infection, potentially eliminating the pandemic before it has a chance to trigger.

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 may be enumerated as follows:

□ 1) Learning patterns of cross-dependency between mutations and genomic change reveals enough of the underlying rules of organization of the primary structure of key viral proteins to meaningfully and actionably constrain the evolutionary trajectories of emerging pathogens. These inferred patterns can then be used to predict future mutations and likelihood of jump events for Influenza A viruses circulating in the wild in

animal reservoirs.

□ 2) The current global biosurveillance efforts produces sufficient data for sophisticated machine learning to carry out meaningful pattern discovery, to enable the development of a next-generation pro-active surveillance platform. Thus, observed patterns of change can be assembled into an early warning system for pandemic threats, and serves a similar function to the strategic goal of NORAD in the context of defending our airspace from adversarial intrusion.

Specific Aims:: Our specific aims are:

□ **Aim 1: Formulate the E-distance:** Devise a biologically meaningful metric of comparison between two genomic sequences, that scales with the probability of one sequence spontaneously replicating to give rise to the other in the wild under realistic, dynamic, and poorly understood selection pressures. Within this aim, our deliverables include the implemented algorithm that analyzes sequence databases, and identifies the E-distance metric of comparison. Since the E-distance reflects the odds of one sequence mutating to another in the wild, it is a function of not just how many mutations the two sequences are apart to begin with, but also how specific mutations incrementally affect fitness, and how possibly non-colocated mutations have emergent dependecies from how distant changes can compensate to maintain fitness. Without taking into account the constraints arising from the need to conserve function, assessing the jump-likelihood is open to subjective guesswork. Our aim here is to show that a precise calculation is possible, that then leads to a actionable framework for tracking evolutionary change. The major tasks within this aim are as follows: T1.1 Precisely formulate the Emergenet inference platform, and provide uncertainty qunatification for the inferred patterns. T1.2 Investigate the sample complexity of the model, $i.e.$, how much data is needed to acceptably identify meaningful patterns that constrain future change. T1.3 Map mutational change dynamics to "wall-time", to ultimately forecast *when* future variants will show up, or when an emergence event is likely. We plan to computationally validate these results using records of past emergence events.

□ **Aim 2:** Validate the E-distance as a simialrity metric on the strain space that differentiates between random perturbations in genomic organization (most of which would be deleterious, and not code for a viable protein), and perturbations that are biologically viable. This is a crucial capability of the Emergenet platform, that woudl make it possible to reliably identify possible future mutations, along with their precisely quantified likelihoods. We will show via in-vitro experiments, that perturbations predicted using this metric leads to viable and functional proteins. The major tasks within this aim are: T2.1 Refine our prelimnary result connecting the E-distance to the probability of spontaneous jump from one strain to another, connecting the inference unceratainty arising possibly from sample size limiations to the uncertainty in the jump probability estimates. T2.2 Laboratory experiments to show that small E-distance leads to vaiable proteins, and that random perturbations, even with a few edits, causes a dramatic fall in fitness.

□ **Aim 3** Develop and demonsrtae a working implementation of the BioNORAD platform for analyzing Influenza A strains at scale for emergence and impact risk. Major tasks are : T3.1 Replicate the published IRAT scores, along with uncertainty quantification, within seconds as a validation result. Investigate how each of the ten dimensions of IRAT comparison map to our Emergenet based risk. T3.2 Demonstrate that we can analyze collected sequences at scale, by enumerating the risk profiles of all sequences collected recently within teh last few years, and any new sequences that continue to be submitted to NCBI and GISAID. This will include setting up an automated pipeline that pulls out sequence data of new sequences, and published a risk score automatically. We will collate this information in our pipleine to map the global risk, visualizing where and when am emergenc event is likely, for what strain and subtype, and from which animal hosts.

Research Strategy and Feasibility:: One possible approach to mitigating pandemic risk is to identify animal strains that do not yet circulate in humans, but is likely to spill-over and quickly achieve human-to-human (HH) transmission capability. While global surveillance efforts collect wild specimens from diverse hosts and geo-locations annually, our ability to objectively, reliably and scalably risk-rank individual strains remains limited[5], despite some recent progress[7–9].

The Center for Disease Control's (CDC) current solution to this problem is the Influenza Risk Assessment Tool (IRAT)[6]. Subject matter experts (SME) score strains based on the number of human infections, infection and transmission in laboratory animals, receptor binding characteristics, population immunity, genomic analysis, antigenic relatedness, global prevalence, pathogenesis, and treatment options, which are averaged to obtain two scores (between 1 and 10) that estimate 1) the emergence risk and 2) the

potential public health impact on sustained transmission. IRAT scores are potentially subjective, and depend on multiple experimental assays, possibly taking weeks to compile for a single strain. This results in a scalability bottleneck, particularly with thousands of strains being sequenced annually.

Here we introduce a pattern recognition algorithm 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. Our approach is centred around numerically estimating the probability $Prx \rightarrow y$ of a strain x spontaneously giving rise to y. We show that this capability is key to preempting strains which are expected to be in future circulation, and 1) reliably forecast dominant strains of seasonal epidemics, and 2) approximate IRAT scores of non-human strains without experimental assays or SME scoring.

Emergenet Inference: To uncover relevant evolutionary constraints, we analyzed variations (point substitutions and indels) of the residue sequences of key proteins implicated in cellular entry and exit[4,10], namely HA and NA respectively. By representing these constraints within a predictive framework – the Emergenet (Enet) – we estimated the odds of a specific mutation to arise in future, and consequently the probability of a specific strain spontaneously evolving into another (Fig. **??**a). Such explicit calculations are difficult without first inferring the variation of mutational probabilities and the potential residue replacements from one positional index to the next along the protein sequence. The many well-known classical DNA substitution models[11] or standard phylogeny inference tools which assume a constant species-wise mutational characteristics, are not applicable here. Similarly, newer algorithms such as FluLeap[12] which identifies host tropism from sequence data, or estimation of species-level risk[9] do not allow for strain-specific assessment.

The dependencies we uncover are shaped by a functional necessity of conserving/augmenting fitness. Strains must be sufficiently common to be recorded, implying that the sequences from public databases that we train with have high replicative fitness. Lacking kinetic proofreading, Influenza A integrates faulty nucleotides at a relatively high rate ($10^{-3} - 10^{-4}$) during replication[13,14]. However, few variations are actually viable, leading to emergent dependencies between such mutations. Furthermore, these fitness constraints are not time-invariant. The background strain distribution, and selection pressure from the evolution of cytotoxic T lymphocyte epitopes[15–19] in humans can change quickly. With a sufficient number of unique samples to train on for each flu season, the Emergenet (recomputed for each time-period) is expected to automatically factor in the evolving host immunity, and the current background environment.

Structurally, 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. **??**b). Thus, an Emergenet comprises almost as many such position-specific predictors as the length of the sequence. These individual predictors are implemented as conditional inference trees[20], in which nodal splits have a minimum prespecified significance in differentiating the child nodes. Thus, each predictor yields an estimated conditional residue distribution at each index. The set of residues acting as features in each predictor are automatically identified, *e.g.*, in the fragment of the H1N1 HA Emergenet (2020-2021, Fig **??**b), 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, wherein each internal node of a tree may be "expanded" to its own tree. Owing to this recursive expansion, a complete Emergenet substantially captures the complexity of the rules guiding evolutionary change as evidenced by our out-of-sample validation.

In our first application (predicting future dominant strains) we used H1N1 and H3N2 HA and NA sequences from Influenza A strains in the public NCBI and GISAID databases recorded between 2000-2022 ($387,067$ in total, Supplementary Table S-**??**). We construct Emergenets separately for H1N1 and H3N2 subtypes, and for each flu season using HA sequences, yielding 84 models in total for predicting seasonal dominance. Using only sequence data is advantageous since deeper antigenic characterization tend to be substantially low-throughput compared to genome sequencing[21]. However, deep mutational scanning (DMS) assays have been shown to improve seasonal prediction[2]. Despite limiting ourselves to only genotypic information (and subtypes), our approach distills emergent fitness-preserving constraints that outperform reported DMS-augmented strategies.

Inference of the Emergenet predictors is our first step, which then induces an intrinsic distance metric between strains. The E-distance (i.e. Emergenet distance) (Eq. (**??**) in Online Methods) is defined as the square-root of the Jensen-Shannon (JS) divergence[22] 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. Central to our approach is the theoretical result (Theorem **??** in Online Methods) that the E-distance approximates the log-likelihood of spontaneous change $i.e.$ $\log Prx \rightarrow y$. Note that despite general correlation between E-distance and edit-distance, the E-distance between fixed strains can change if only the background environment changes (Supplementary Table S-**??**,S-**??**). In in-silico experiments, We find that while random mutations to genomic sequences produce rapidly diverging sets, Emergenet-constrained replacements produce sequences that are verifiably meaningful (In-silico Corroboration of Emergenet's Capability To Capture Biologically Meaningful Structure, Online Methods and Supplementary Fig. S-**??**).

Determining the numerical odds of a spontaneous jump $Prx \rightarrow y$ (Fig. **??**) 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. **??**c-d). Thus, a dominant strain for an upcoming season may be identified as one which maximizes the joint probability of simultaneously arising from each (or most) of the currently circulating strains (Fig. **??**c). This does not deterministically specify the dominant strain, but a strain satisfying this criterion has high odds of acquiring dominance. And, a pandemic risk score of a novel strain may be estimated by the probability of it giving rise to a well-adapted human strain. In the context of forecasting future dominant strain(s), we derive a search criteria (Predicting Dominant Seasonal Strains, Online Methods) from the above proposition, to identify historical strain(s) that are expected to be close to the next dominant strain(s):

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

where $x_\star^{t\delta}$ is a predicted dominant strain at time $t\ \delta$, H^t is the set of currently circulating human strains at time t observed over the past year, θ^t is the E-distance informed by the inferred Emergenet using sequences in H^t, ω_y is the estimated probability of strain y being generated by the Emergenet, and A is a constant dependent on the sequence length and significance threshold used. The first term gets the solution close to the centroid of the current strain distribution (in the E-distance metric, and not the standard edit distance), and the second term relates to how common the genomic patterns are amongst recent human strains.

Predicting Future Dominant Strains: Prediction of the future dominant strain as a close match to a historical strain allows out-of-sample validation against past World Health Organization (WHO) recommendations for the flu shot, which is reformulated about six months in advance based on a cocktail of historical strains determined via global surveillance[23]. For each year of the past two decades, we first computed three clusters of strains in the E-distance metric on their HA sequences. In each cluster, we calculated strain forecasts using Eq. (1) with data available six months before the target season, taking our first and second recommendations from the two largest clusters. We also calculated the top ten dominant strains for both HA and NA from the target season, ranked by closeness to the centroid in the strain space that season in the edit distance metric. We measured forecast performance by the average number of mutations by which the predicted HA/NA sequences deviated from the top ten dominant strains. Our Emergenet-informed forecasts outperform WHO/CDC recommended flu vaccine compositions consistently over the past two decades, for both H1N1 and H3N2 subtypes, individually in the northern and the southern hemispheres (which have distinct recommendations[24]). 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). The gains for H1N1 NA over the same time periods are 46.41%, 40.31%, and 54.85% respectively. 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. The gains for H3N2 NA over the same time periods are 46.90%, 42.31%, and 47.65% respectively (Extended Data Table **??**). Detailed predictions, along with historical strains closes to the observed dominant one are tabulated in Extended Data Tables **??** through **??**. Visually, Fig. **??** illustrates the relative gains computed for different subtypes and hemispheres.

Comparing the Emergenet inferred strain (ENT) against the one recommended by the WHO, we find that the residues that only the Emergenet recommendation matches correctly with dominant strain (DOM), while the WHO recommendation fails, are largely localized within the RBD, with $> 57\%$ occurring within the RBD on average (Extended Data Fig. **??**a), and 3) when the WHO strain deviates from the ENT/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 (Extended Data Fig.-**??**b-f), suggesting deviations in recognition characteristics[25,26]. Combined with the fact that we find circulating strains are almost always within a few edits of the DOM (Supplementary Fig. S-**??**), these observations suggest that hosts vaccinated with the ENT recommendation is can have season-specific antibodies that recognize a larger cross-section of the circulating strains.

Estimating Pandemic Risk of Non-human Strains: Our primary claim, however, is the 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 (Measure of Pandemic Potential, Online Methods):

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

scales as the average log-likelihood of $Prx \rightarrow y$ where y is any human strain of a similar subtype to x, and θ^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. As before, the Emergenet inference makes it possible to estimate $\rho_t x$ explicitly.

To map the Emergenet distances to more recognizable IRAT scores, we train a general linear model (GLM) from the the HA/NA-based E-risk values (Multivariate Regression to Identify Map from E-distance to Estimated IRAT scores, Online Methods and Supplementary Table S-**??**). Since the CDC-estimated IRAT impact scores are strongly correlated with their IRAT emergence scores (correlation of 0.8015), we also trained a separate GLM to estimate the impact score from the E-risk values (Supplementary Table S-**??**). Finally, we estimate the IRAT scores of all $6,066$ Influenza A strains sequenced globally between 2020 through 04/2022, and identify the ones posing maximal risk (Fig. **??**c). $1,773$ strains turn out to have a predicted emergence score > 6.0. However, many of these strains are highly similar, differing by only a few edits. To identify the sufficiently distinct risky strains, we constructed the standard phylogeny from HA sequences with score > 6 (Fig. **??**), and collapsed all leaves within 15 edits, showing only the most risky strain within a collapsed group. This leaves 75 strains (Fig. **??**), with 68 having emergence risk > 6.25, and 6 with risk above 6.5 (Extended Data Table **??**). Subtypes of the risky strains are overwhelmingly H1N1, followed by H3N2, with a small number of H7N9 and H9N2. Five maximally risky strains with emergence score > 6.58 are identified to be: A/swine/Missouri/A02524711/2020 (H1N1), A/Camel/Inner_Mongolia/XL/2020 (H7N9), A/swine/Indiana/A02524710/2020 (H3N2), A/swine/North Carolina/ A02479173/2020 (H1N1), and A/swine/Tennessee/ A02524414/2022 (H1N1). Additionally, A/mink/China/chick embryo/2020 (H9N2), with a lower estimated emergence score (6.26) is also important, as the most risky H9N2 strain in our analysis. We compare the HA sequences along with two dominant human strains in 2021-2022 season (Extended Data Fig. **??**), which shows substantial residue replacements, in and out of the receptor binding domain (RBD).

Innovation: While numerous tools exist for ad hoc quantification of genomic similarity[11,27–31], higher similarity between strains in these frameworks is not sufficient to imply a high likelihood of a jump. To the best of our knowledge, 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 amino acid substitution, or a genealogical tree a priori. While the effect of the environment and selection cannot be inferred from a single sequence, an entire database of observed strains, processed through the right lens, can parse out useful predictive models of these complex interactions. Our results are aligned with recent studies demonstrating effective predictability of future mutations for different organisms[32,33].

No animal work. Primary cells will be purchased from Lonza. So no IRB.

Assessment of fitness potential emerging zoonotic IAV variants in vitro cell culture. Potential HA variants identified in the Q-net algorithm 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 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. In addition, we 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 as