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 $Prx \to y$ 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):

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

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 $Pr x \rightarrow y$ 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

List of Abbreviations, Acronyms, and Symbols

Emergenet	Emergenet
IRAT	Influenza Risk Assessment Tool
CDC	Centre for Disease Control
WHO	World Health Organization
ML	Machine Learning
AI	Artificial Intelligence
HA	Hemaglutinnin
NA	Neuraminidase
RBD	Receptor Binding Site
CIT	Conditional Inference Trees
SME	Subject Matter Expert
NCBI	National Center for Biotechnology Information
GISAID	Global Initiative on Sharing All Influenza Data
NORAD	North American Air Defense
E-distance	Emergenet similarity between sequences
UQ	Uncertainty Quantification

Data Management Plan

Data sharing plan

Computing Environment: The UChicago computing environment provided by the Center for Research Informatics (CRI) will be sandboxed from the internet as well as other servers and data sources at UChicago. It will accessible only to the PI, research assistant(s), software developer(s), and/or system administrator(s) who require access during the course of the project. Box: Research data will be stored and preserved for the duration of the grant using Box, which uses AES 256-bit encryption and is also FedRamp authorized and HIPAA compliant (https://www.box.com/security). Box provides file versioning, which helps mitigate issues such as file corruption. UChicago is committed to using Box as the institutional cloud storage tool. If the university were to switch cloud storage solutions, we will meet the security needs of this and all other grant work supported by Box.

Data and research resources generated in this project research will be made available to the research community, which includes both scientific and consumer advocacy communities, and to the public. This includes all data and research resources generated during the project's period of performance, including:

Unique Data, defined as data that cannot be readily replicated. For this project, examples of unique data include curated models of genomic change for different sub-types of Influenza A, for different geographical locations. Final Research Data defined as recorded factual material commonly accepted in the scientific community as necessary to document and support research findings. In our context, examples are sequence ids of strains we use for our modeling, and the particulars of validations experiments, including the metadata needed to replicate those experiments in the laboratory. Research Resources include, but are not limited to, the full range of tools that we would develop and use in the laboratory. In this project, such resources include all developed software for modeling and prediction.

We will deposit software in Github repositories, allowing easy installation of such software in compatible systems. We will also deposit models, metadata and software copy at Zenodo for long-term citable access to the research resources and products.

No data sharing agreement is required for this project, since the underlying data on which we will learn our models are publicly accessible with minor restrictions.

Complete enumeration of sequence ids as obtained from NCBI and GISAID will be submitted, which is sufficient to replicate the results if using our developed software. Also descriptions of inferred and curated models will be made available. Example software programs based on our open-source library will be provided as well.

No specialized file format is necessary for this project. All files will be shared as text files, csv files or compressed versions of those.

No specialized transformation is necessary.

The effort of the postdoctoral associate funded on this project will carry out the requirements of this plan, and his salary will be partially covered under the proposed budget.

Technical Abstract

We plan to distill evolutionary constraints from rapidly expanding databases (GISAID & NCBI) of $> 10,000$ SARS-CoV-2 sequences, to predict epitopes and sequences of perturbed fusion proteins expected to emerge in future in the wild. Our central idea in this project is to model the constraints on the variations of the nucleotide sequences as a virus evolves by inferring a set of inter-dependent predictors known as the Quasinet or the Enet. The Enet framework is specifically designed for the analysis of biological sequences at scale, with the objective of modeling and prediction of dynamics unfolding in ultra-high dimensional sequence spaces. The key idea here is surprisingly simple: *we learn models for predicting the mutational variations at each index of the genomic sequence using other indices as features. Collectively, these predictors represent the emergent constraints that shape evolutionary changes from selection forces in the wild.*

Lay Abstract

We plan to distill evolutionary constraints from rapidly expanding databases (GISAID & NCBI) of $> 10,000$ SARS-CoV-2 sequences, to predict epitopes and sequences of perturbed fusion proteins expected to emerge in future in the wild. Our central idea in this project is to model the constraints on the variations of the nucleotide sequences as a virus evolves by inferring a set of inter-dependent predictors known as the Quasinet or the Enet. The Enet framework is specifically designed for the analysis of biological sequences at scale, with the objective of modeling and prediction of dynamics unfolding in ultra-high dimensional sequence spaces. The key idea here is surprisingly simple: *we learn models for predicting the mutational variations at each index of the genomic sequence using other indices as features. Collectively, these predictors represent the emergent constraints that shape evolutionary changes from selection forces in the wild.*

Statement of Work - 04/26/2023

PROPOSED START DATE 10/01/2023

Site 1:	University of Chicago 5801 S. Ellis Ave. Chicago, IL 60637 PI: Ishanu Chattopadhyay	Site 2:	University of Iowa 51 Newton Road Iowa City, IA 52242 Site PI: Balaji Manicassamy

Specific Aim 1: Formulate the E-distance	**Timeline (Months)**	**Site 1**	**Site 2**
Major Task 1			
Subtask T1.1: Formulate the Emergenet inference with UQ	1-3		
Subtask T1.2: Estimate sample complexity of the inference algorithm	2-4		
Subtask T1.3: Map mutations to physical time	3-6		
Milestones Achieved: Emergenet software beta release, uncertainty and sample complexity quantified,			
Major Task 2			
Subtask 1			
Subtask 2			
Subtask 3			
Milestones Achieved			

Specific Aim 2	**Timeline (Months)**	**Site 1**	**Site 2**
Major Task 1			
Subtask 1			
Subtask 2			
Subtask 3			
Milestones Achieved			
Major Task 2			
Subtask 1			
Subtask 2			
Subtask 3			
Milestones Achieved			

Specific Aim 3	**Timeline (Months)**	**Site 1**	**Site 2**
Major Task 1			
Subtask 1			
Subtask 2			
Subtask 3			
Milestones Achieved			
Major Task 2			
Subtask 1			
Subtask 2			
Subtask 3			
Milestones Achieved			

Impact Statement

The proposed research project is important and relevant to the FY23 PRMRP Topic Area of Infectious Diseases, as it aims to develop the BioNORAD platform to predict and identify the emergence of new strains of influenza viruses. This platform has the potential to significantly improve global surveillance and response capabilities for emerging pandemic threats, which is a growing concern in today's interconnected world.

The FY23 PRMRP Strategic Goal addressed in the proposed research is Epidemiology: Identify strategies for surveillance or develop modeling tools and/or biomarkers to predict outbreaks or epidemics. The BioNORAD platform aligns with this strategic goal by employing advanced machine learning algorithms and interdisciplinary insights to create an early warning system for pandemic threats. This system will enable proactive measures to protect both military and civilian populations, strengthening global health security.

The potential short-term impact of the proposed research includes the development of a robust, scalable, and actionable platform to predict and identify emerging influenza strains. This will enable more efficient allocation of resources for vaccine development, antiviral treatments, and other medical countermeasures, ultimately improving patient care and health outcomes.

In the long-term, the BioNORAD platform has the potential to revolutionize the field of infectious disease surveillance and response, leading to better preparedness for future pandemics. Moreover, the interdisciplinary nature of this project could foster innovations and breakthroughs in machine learning, information theory, evolutionary theory, epidemiology, and proteomics, with broad applications beyond influenza.

The proposed research has the potential to generate preliminary data that can be used as a foundation for future research projects. As the BioNORAD platform is developed and validated, the insights gained will inform the design of new surveillance strategies, modeling tools, and biomarkers for predicting outbreaks or epidemics. This will enable researchers to explore novel approaches to combating infectious diseases, ultimately contributing to improved global health security.

In summary, the proposed research project is highly relevant to the FY23 PRMRP Topic Area of Infectious Diseases and addresses the FY23 PRMRP Strategic Goal of Epidemiology. The development of the BioNORAD platform has the potential to make significant short-term and long-term impacts on the field of study and patient care, while also laying the foundation for future research projects in the areas of infectious disease surveillance, modeling, and prediction.

Relevance to Military Health Statement

The emergence of new strains of influenza viruses with the potential to cause pandemics is a global threat with significant implications for the health and safety of military personnel. The proposed BioNORAD platform is designed to predict and identify the emergence of new strains of influenza viruses, providing vital information for the Department of Defense (DoD) to take proactive measures to protect its personnel and assets. In this statement, we highlight the relevance of this grant to military health and why it is of interest to the DoD.

1) **Protecting Military Personnel and Assets:** Military personnel are often deployed in diverse geographical locations and close proximity to animal reservoirs, increasing their risk of exposure to novel influenza strains. The ability to preemptively identify and mitigate these threats is essential to safeguard the health of the deployed personnel and ensure the readiness and effectiveness of the military. The BioNORAD platform will enable the DoD to take proactive measures to protect its personnel and assets from emerging pandemic threats.
2) **Enhancing Military Medical Response Capabilities:** The development of the BioNORAD platform will provide the military with an advanced tool to better anticipate and respond to pandemic threats. Early identification of potential strains enables the development of targeted vaccines, antiviral treatments, and other medical countermeasures. This will significantly enhance the military's medical response capabilities, ensuring the health and well-being of its personnel.
3) **Strengthening Global Health Security:** The ability to predict and mitigate the spread of pandemic threats is a vital aspect of global health security. By developing the BioNORAD platform, the DoD will contribute to global efforts to prevent and respond to emerging infectious diseases. This will not only protect military personnel but also support civilian populations worldwide, strengthening international partnerships and cooperation.
4) **Reducing Economic and Operational Impacts:** Pandemics can have severe economic and operational consequences for the military. By enabling early detection and mitigation, the BioNORAD platform will help reduce the financial and operational burdens associated with major outbreaks. This will ensure that the DoD can continue to carry out its mission effectively during times of crisis.
5) **Promoting Interdisciplinary Collaboration and Innovation:** The development of the BioNORAD platform will bring together experts from various fields, including machine learning, information theory, evolutionary theory, epidemiology, and proteomics. This interdisciplinary collaboration will foster innovation and advance our understanding of the complex interactions between pathogens and their hosts. The knowledge and technologies generated by this project will have broad applications beyond influenza, with potential benefits for military health and biodefense efforts.

In summary, the development of the BioNORAD platform is highly relevant to military health and of significant interest to the Department of Defense. By enabling early identification and mitigation of emerging pandemic threats, the platform will protect military personnel and assets, enhance military medical response capabilities, strengthen global health security, reduce economic and operational impacts, and promote interdisciplinary collaboration and innovation. This project aligns with the FY23 PRMRP Portfolio Category: Infectious Diseases, FY23 PRMRP Topic: proteomics, and FY23 PRMRP Strategic Goal: Epidemiology: Identify strategies for surveillance or develop modeling tools and/or biomarkers to predict outbreaks or epidemics. The investment in the BioNORAD platform is a strategic step towards ensuring the health and safety of military personnel and the success of the DoD's mission in a world where pandemic threats are a growing concern.

Facilities, Existing Equipment, and Other Resources

The University of Chicago is a private non-profit institution located on the ethnically-diverse South Side of Chicago that has been a center of advanced learning and research since its inception in 1892. The University of Chicago is comprised of four graduate Divisions (Biological Sciences, Physical Sciences, Social Sciences, and Humanities), six professional schools (Chicago Booth School of Business, Divinity School, Harris School of Public Policy Studies, Law School, Pritzker School of Medicine, and School of Social Service Administration), the Graham School of General Studies, and the undergraduate College. The University has a unique history of organizing around research questions that cross disciplines rather than operating within disciplinary boundaries. The extent to which this strategy reflects University of Chicago is illustrated by its numerous interdisciplinary Committees, Centers and Institutes (described below). The University of Chicago maintains its commitment to scholarship, teaching, and research through its more than 2100 faculty members and a student population of approximately 15,600 with nearly 2/3 engaged in advanced research and professional study. Through the years, 86 Nobel Laureates (8 are current faculty), 44 members of the National Academy of Sciences, 169 members of the American Academy of Arts and Sciences, and 14 recipients of the National Medal of Science have been associated with the University as students, teachers or research investigators. The University of Chicago is ranked among the world's top universities by a number of criteria, including the amount of federal research funding received (despite a size much smaller than many of its academic peers). This spirit of discovery, innovation and public service provides a robust foundation for success.

South Shore Senior Health Center (SSSC): The SSSC is a 6800 sq. ft. university-owned geriatric facility 5 miles south of DCAM, with doorstep free parking in the heart of Chicago's South Side, specifically in the South Shore district. There are 13 patient exam rooms equipped with an examination bed, ophthalmoscope and otoscope, in addition to an on-site phlebotomist and capability for ECG. There is a large conference room for team meetings and support groups. The Memory Center team meets on Mondays at this site.

Center for Care and Discovery: Completed in 2013, the CCD is a ten-story adult hospital focused on cancer, advanced surgery, high tech imaging, the neurosciences and gastrointestinal procedures. The building is 1,200,000 GSF with floor plates of 102,000 GSF. The CCD includes 240 private patient rooms, 28 operating rooms (21 initially); an imaging department with 3 CT's, 2 MRI's, 1 fluoroscopy room, 2 general radiology rooms, 7 interventional radiology rooms; and a gastroenterology procedures suite with 11 GI procedure rooms, 2 fluoroscopy rooms and 2 bronchoscopy rooms. The CCD includes an inpatient kitchen, cafeteria, 7th floor sky lobby meeting rooms, ground floor retail space and clinical support services. Two floors are "shelled space" for future expansion of services. The building is centrally located between existing clinical facilities (DCAM, Comer) and new research facilities (KCBD, GCIS, Knapp). The facility is connected to both DCAM and Comer via above and below ground connections. There are procedure rooms, 2 fluoroscopy rooms and 2 bronchoscopy rooms. Neurology and neurosurgery inpatients as well as a Neuro- ICU are contiguously located on one floor of the CCD.

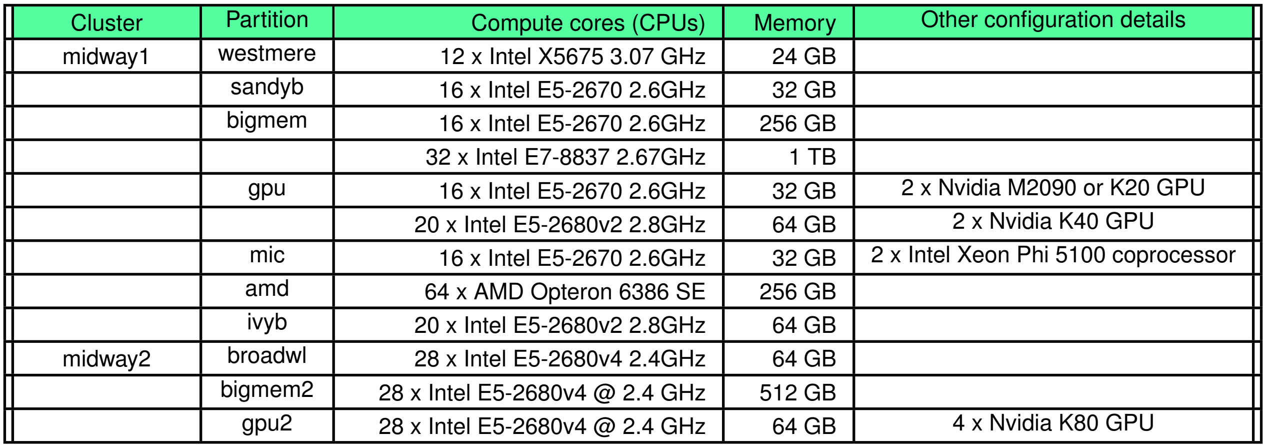
Computational Facilities

The principal investigators have access to extensive computational facilities available at the University of Chicago to carry out the tasks described.

Access to Clinical Data for AI-enabled Analytics: The ZeD lab (overseen by Professor Chattopadhyay) is housed within the Department of Medicine at the University of Chicago, and has access to the full range of high end computing resources offered by the University of Chicago. In addition, Prof. Chattppadyay's laboratory has access to the HIPAA compliant clinical data warehouse maintained by the Biological Sciences Division as detailed below:

The Clinical Research Data Warehouse: (CRDW) within the Biomedical Sciences Division of the University of Chicago is one of the deepest, richest, and most research-ready data repositories of its kind. Containing more than a decade of University of Chicago medical data, it seamlessly brings together multiple internal and external data sources to provide researchers with access to more than 12 million encounters for 2.3 million patients. The associated diagnoses, labs, medications, and procedures number in the tens of millions each. The CRDW is run on IBM Netezza Pure Data System for Analytics servers, a patented Asymmetric Massively Parallel Processing architecture designed to deliver exceptional query performance and modular scalability on highly complex mixed workloads.

TABLE 1
University of Chicago Research Computing Center Capabilities Summary



Cluster	Partition	Compute cores (CPUs)	Memory	Other configuration details
midway1	westmere	12 x Intel X5675 3.07 GHz	24 GB	
	sandyb	16 x Intel E5-2670 2.6GHz	32 GB	
	bigmem	16 x Intel E5-2670 2.6GHz	256 GB	
		32 x Intel E7-8837 2.67GHz	1 TB	
	gpu	16 x Intel E5-2670 2.6GHz	32 GB	2 x Nvidia M2090 or K20 GPU
		20 x Intel E5-2680v2 2.8GHz	64 GB	2 x Nvidia K40 GPU
	mic	16 x Intel E5-2670 2.6GHz	32 GB	2 x Intel Xeon Phi 5100 coprocessor
	amd	64 x AMD Opteron 6386 SE	256 GB	
	ivyb	20 x Intel E5-2680v2 2.8GHz	64 GB	
midway2	broadwl	28 x Intel E5-2680v4 2.4GHz	64 GB	
	bigmem2	28 x Intel E5-2680v4 @ 2.4 GHz	512 GB	
	gpu2	28 x Intel E5-2680v4 @ 2.4 GHz	64 GB	4 x Nvidia K80 GPU

In order to meet the acute need for data related to COVID-19, the CRDW team has constructed three data marts (de-identified, limited, and identified) to provide the most commonly requested data elements for this patient population. The initial instance of the COVID-19 data mart includes de-identified structured data on patient demographics, encounters, diagnoses, labs, medications, flow sheets, and procedures. Additional data will be added based on resource availability and urgency.

Cohort Discovery Tool: The purpose of this tool (SEE Cohorts) is to provide a secure web-based tool for the initial exploration of de-identified data. It allows researchers to search available data, build a cohort of patients, and view actual de-identified data within the interface. The data in SEE Cohorts is refreshed weekly.

Research Computing Center: The University of Chicago Research Computing Center (RCC) provides high-end research computing resources to researchers at the University of Chicago, which include high-performance computing and visualization resources; high-capacity storage and backup; software; high-speed networking; and hosted data sets. Resources are centrally managed by RCC staff who ensure the accessibility, reliability, and security of the compute and storage systems. A high-throughput network connects the Midway Compute Cluster to the UChicago campus network and the public internet through a number of high-bandwidth uplinks. To support data-driven research RCC hosts a number of large datasets to be accessed within the RCC compute environment.

RCC maintains three pools of servers for distributed high-performance computing. Ideal for tightly coupled parallel calculations, tightly-coupled nodes are linked by a fully non-blocking FDR-10 Infiniband interconnect. Loosely-coupled nodes are similar to the tightly-coupled nodes, but are connected with GigE rather than Infiniband and are best suited for high-throughput jobs. Finally, shared memory nodes contain much larger main memories (up to 1 TB) and are ideal for memory-bound computations. The types of CPU architectures RCC maintains are tabulated in Table 1.

RCC also maintains a number of specialty nodes:

- *Large shared memory nodes* - up to 1 TB of memory per node with either 16 or 32 Intel CPU cores. Midway is always expanding, but at time of writing RCC contains a total of 13,500 cores across 792 nodes, and 1.5 PB of storage.
- *Hadoop:* Originally developed at Google, Hadoop is a framework for large-scale data processing.
- *GPU Computing:* Scientific computing on graphics cards can unlock even greater amounts of parallelism from code. RCC GPU nodes each include two Nvidia Tesla-class accelerator cards and are integrated in the Infiniband network. RCC currently provides access to Fermi-generation M2090 GPU devices and Kepler-generation K20 and K40 devices.
- *Xeon Phi:* The Many Integrated-Core architecture (MIC) is Intel's newest approach to manycore computing. Researchers can experiment with these accelerators by using MIC nodes, each of which have two Xeon Phi cards, and are integrated into the Infiniband network.

Persistent and High-Capacity Storage. Storage is accessible from all compute nodes on Midway1 and Midway2 as well as outside of the RCC compute environment through various mechanisms, such as mounting directories as network drives on your personal computer or accessing data as a Globus Online endpoint (at the time of this writing, Globus Online is supported on Midway1). RCC takes snapshots of all home directories (users' private storage space) at regular intervals so that if any data is lost or corrupted, it can easily be recovered. RCC maintains GPFS Filesystem Snapshots for quick and easy data recovery. In the event of catastrophic storage failure, archival tape backups can be used to recover data from persistent storage locations on Midway. Automated snapshots of the home and project directories are available in case of accidental file deletion or other problems. Currently snapshots are available for these time periods: 1) 7 daily snapshots, 2) 4 weekly snapshots.

Tape Backups. Backups are performed on a nightly basis to a tape machine located in a different data center than the main storage system. These backups are meant to safeguard against events such as hardware failure or disasters that could result in the complete loss of RCC's primary data center.

Data Sharing. All data in RCC's storage environment is accessible through a wide range of tools and protocols. Because RCC provides centralized infrastructure, all resources are accessible by multiple users simultaneously, which makes RCC's storage system ideal for sharing data among your research group members. Additionally, data access and restriction levels can be put in place on an extremely granular level.

Data Security & Management. The HIPAA compliant security of the Research Computing Center's storage infrastructure, protected by two-factor authentication, gives users peace of mind that their data is stored, managed, and protected by HPC professionals. Midway's file management system allows researchers to control access to their data. RCC has the ability to develop data access portals for different labs and groups.

Publications and/or Patents

Patents:

□ Chattopadhyay, I. (2022). "Methods and systems for genomic based prediction of virus mutation" (Patent No. WO2022108965A1). World Intellectual Property Organization. URL: https://patents.google.com/patent/WO2022108965A1

TABLE 2
Pending Patent on Core Algorithm

Title	METHOD OF CREATING ZERO-BURDEN DIGITAL BIOMARKERS FOR DISORDERS AND EXPLOITING CO-MORBIDITY PATTERNS TO DRIVE EARLY INTERVENTION
Patent Application Type	International
International Filing Date	09/23/2020
International Application No.	PCT/US2020/052112
Publication Number	WO/2021/061702
Applicant	The University of Chicago
Priority Data	62/904,220, 09/23/2019 US 62/937,604, 11/19/2019 US
WIPO IP Portal Link	https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2021061702
IP Filing Plans	File non-provisional patent application in the United States and foreign jurisdictions by nationalization date 03/23/2022

Publications:

□ Huang, Yi, and Ishanu Chattopadhyay. "Universal risk phenotype of US counties for flu-like transmission to improve county-specific COVID-19 incidence forecasts." PLoS computational biology 17, no. 10 (2021): e1009363.

□ Dhanoa, J., Manicassamy, B. and Chattopadhyay, I., 2018. "Algorithmic Bio-surveillance For Precise Spatio-temporal Prediction of Zoonotic Emergence." arXiv preprint arXiv:1801.07807.

□ Chattopadhyay, Ishanu, Emre Kiciman, Joshua W. Elliott, Jeffrey L. Shaman, and Andrey Rzhetsky. "Conjunction of factors triggering waves of seasonal influenza." Elife 7 (2018): e30756.

□ Li, Jin, Timmy Li, and Ishanu Chattopadhyay. "Preparing For the Next Pandemic: Learning Wild Mutational Patterns At Scale For For Analyzing Sequence Divergence In Novel Pathogens." medRxiv (2020): 2020-07.

□ Chattopadhyay, Ishanu, Kevin Wu, Jin Li, and Aaron Esser-Kahn. "Emergenet: Fast Scalable Pandemic Risk Assessment of Influenza A Strains Circulating In Non-human Hosts." (2022).

Letters of Organizational Support

Letters of Collaboration

Intellectual and Material Property Plan

A. Intellectual Property (IP) Ownership and Management: The IP ownership and management for the BioNORAD project will be governed by a formal agreement signed by all participating organizations. The agreement will specify the following:

1) Ownership of any existing IP (background IP), such as Chattopadhyay, I. (2022). "Methods and systems for genomic based prediction of virus mutation" (Patent No. WO2022108965A1). World Intellectual Property Organization. URL: https://patents.google.com/patent/WO2022108965A1, will be retained by the originating organization.
2) New IP generated during the course of the project (foreground IP) will be jointly owned by the participating organizations, with the share of ownership determined by the contribution of each party to the development of the IP.
3) The participating organizations will identify a designated IP representative who will be responsible for managing IP issues and ensuring compliance with the agreement.
4) The IP agreement will include provisions for resolving disputes related to IP ownership and management.

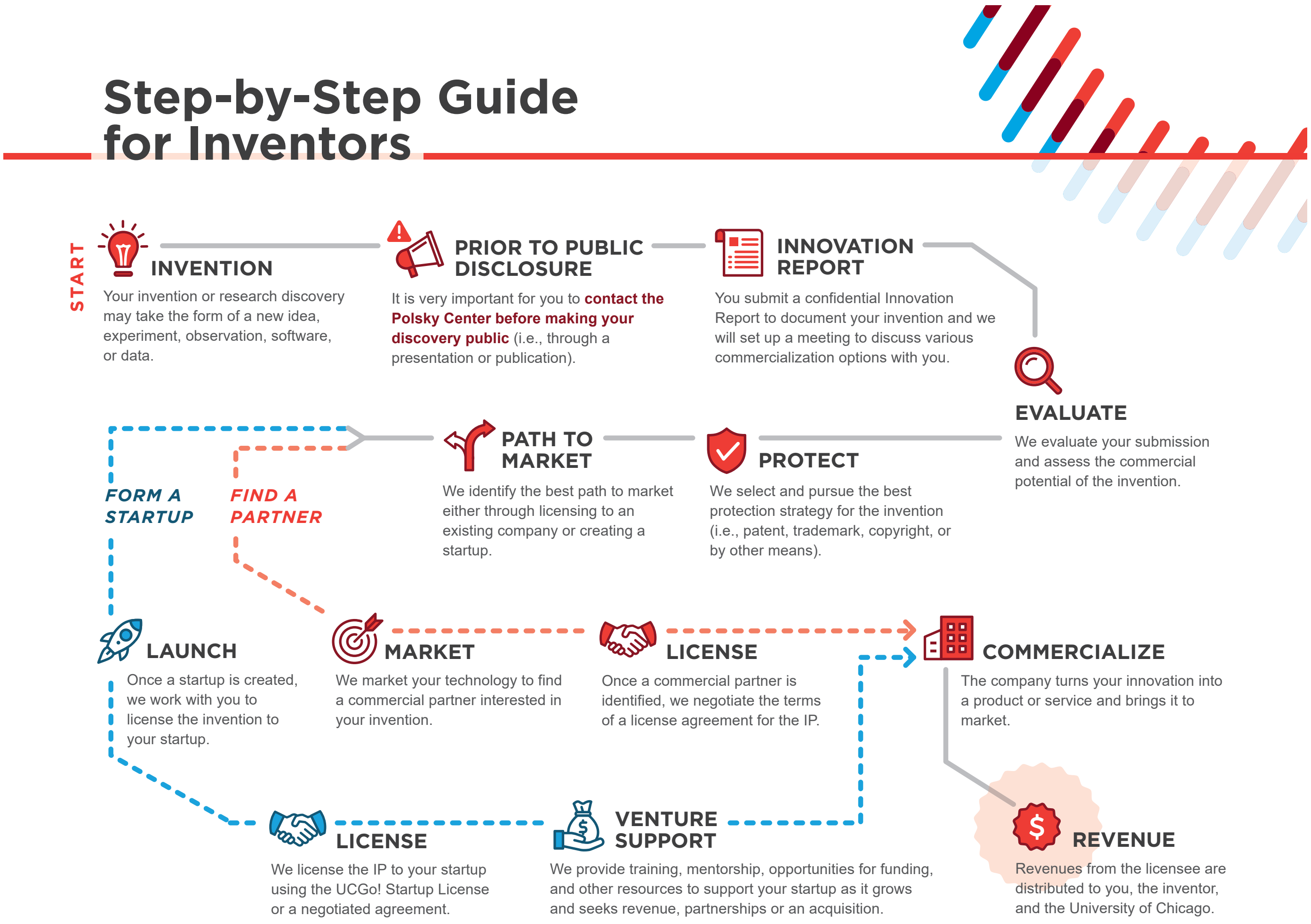
B. Licensing and Commercialization: The participating organizations will develop a strategy for licensing and commercializing the foreground IP for the BioNORAD platform, considering the following factors:

1) Evaluation of potential markets and applications for the platform, primarily focusing on global health organizations, governments, and pharmaceutical companies.
2) Identification of potential licensees and strategic partners.
3) Negotiation of licensing agreements, including royalties and other financial terms.
4) Development of a patent strategy, including filing and maintenance of patents in relevant jurisdictions.

C. Commercialization Strategy:

1) **Intellectual Property**: The participating organizations will develop and maintain a strong IP portfolio for the BioNORAD platform. This includes filing patent applications in key markets and ensuring that the IP is properly protected.
2) **Market Size**: The target market for the developed technology will be global health organizations, governments, and pharmaceutical companies involved in pandemic prevention and response. This market is expected to grow significantly due to increasing awareness of pandemic risks and the need for proactive measures.
3) **Financial Analysis**: The financial analysis will include a detailed assessment of the potential revenues, costs, and profitability of the BioNORAD platform. This will include projections for product pricing, market share, and revenue growth, as well as estimates of development costs, manufacturing expenses, and other operating costs.
4) **Strengths and Weaknesses**: The commercialization plan will identify the platform's strengths and weaknesses, as well as opportunities and threats in the market. This analysis will help the participating organizations to strategically position the platform in the market and address potential challenges.
5) **Barriers to the Market**: The commercialization plan will address potential barriers to market entry, such as competition, regulatory hurdles, and technology adoption challenges. Strategies will be developed to overcome these barriers and increase the chances of successful market penetration.
6) **Competitors**: The commercialization plan will include an analysis of the competitive landscape, identifying key competitors and their strengths and weaknesses. This will help the participating organizations to differentiate the BioNORAD platform and develop a competitive advantage.
7) **Management Team**: A strong management team will be assembled to lead the commercialization effort. This team will include individuals with experience in technology development, marketing, sales, and operations, as well as industry-specific expertise in pandemic prevention and response.
8) **Significance and Timeline**: The commercialization plan will outline the significance of the BioNORAD platform in addressing the challenges of emerging pandemic threats and the need for proactive measures. A timeline for the development and commercialization of the technology will be provided, along with milestones to track progress and measure success.

D. Inventions and IP Rights at The University of Chicago: The University of Chicago is committed to the open and timely dissemination of research outcomes. Investigators in the proposed activity recognize that promising new methods, technologies, strategies and software programming may arise during the course of



Step-by-Step Guide for Inventors

INVENTION
Your invention or research discovery may take the form of a new idea, experiment, observation, software, or data.

PRIOR TO PUBLIC DISCLOSURE
It is very important for you to **contact the Polsky Center before making your discovery public** (i.e., through a presentation or publication).

INNOVATION REPORT
You submit a confidential Innovation Report to document your invention and we will set up a meeting to discuss various commercialization options with you.

EVALUATE
We evaluate your submission and assess the commercial potential of the invention.

PROTECT
We select and pursue the best protection strategy for the invention (i.e., patent, trademark, copyright, or by other means).

PATH TO MARKET
We identify the best path to market either through licensing to an existing company or creating a startup.

MARKET
We market your technology to find a commercial partner interested in your invention.

LICENSE
Once a commercial partner is identified, we negotiate the terms of a license agreement for the IP.

COMMERCIALIZE
The company turns your innovation into a product or service and brings it to market.

REVENUE
Revenues from the licensee are distributed to you, the inventor, and the University of Chicago.

LAUNCH
Once a startup is created, we work with you to license the invention to your startup.

LICENSE
We license the IP to your startup using the UCGo! Startup License or a negotiated agreement.

VENTURE SUPPORT
We provide training, mentorship, opportunities for funding, and other resources to support your startup as it grows and seeks revenue, partnerships or an acquisition.

Fig. 1. Inventor pathway to commercialization at the University of Chicago

the research. The Investigators are aware of and agree to be guided by the principles for sharing research resources as described, for example, in the National Institutes of Health ”Principles and Guidelines for Recipients of NIH Research Grants and Contracts on Obtaining and Disseminating Biomedical Research Resources”.

While the investigators expect that research tools will be freely shared with the research community, opportunities for technology transfer through commercialization will be explored as appropriate. At the University of Chicago, its Polsky Center for Entrepreneurship and Innovation manages intellectual property (IP). The Polsky Center for Entrepreneurship and Innovation manages all technology transfer operations at the University of Chicago (See Figure 1).

Our Polsky Science and Technology group serves as the central resource for transforming groundbreaking ideas and faculty discoveries into new products, services, and ventures. We have a dedicated team of scientists with deep technical expertise who are exclusively focused on managing intellectual property and negotiating partnerships and licenses for technologies developed by faculty, researchers, and staff. The Polsky Center serves faculty, staff and students by commercializing inventions, ideas and software developed at the University to ensure that new knowledge benefits society.

Revenues from any commercial licenses will be shared with the inventor and reinvested in the research enterprise.

Data and Research Resources Sharing Plan

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] Huddleston, J. *et al.* Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza a/h3n2 evolution. *Elife* **9**, e60067 (2020).

[3] Reid, A. H. & Taubenberger, J. K. The origin of the 1918 pandemic influenza virus: a continuing enigma. *Journal of general virology* **84**, 2285–2292 (2003).

[4] Shao, W., Li, X., Goraya, M. U., Wang, S. & Chen, J.-L. Evolution of influenza a virus by mutation and re-assortment. *International journal of molecular sciences* **18**, 1650 (2017).

[5] Wille, M., Geoghegan, J. L. & Holmes, E. C. How accurately can we assess zoonotic risk? *PLoS biology* **19**, e3001135 (2021).

[6] 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).

[7] Pulliam, J. R. & Dushoff, J. Ability to replicate in the cytoplasm predicts zoonotic transmission of livestock viruses. *The Journal of infectious diseases* **199**, 565–568 (2009).

[8] Grewelle, R. E. Larger viral genome size facilitates emergence of zoonotic diseases. *bioRxiv* (2020).

[9] Grange, Z. L. *et al.* Ranking the risk of animal-to-human spillover for newly discovered viruses. *Proceedings of the National Academy of Sciences* **118**, e2002324118 (2021).

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

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

[12] Eng, C. L., Tong, J. C. & Tan, T. W. Predicting host tropism of influenza a virus proteins using random forest. *BMC medical genomics* **7**, 1–11 (2014).

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

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

[15] Woolthuis, R. G., van Dorp, C. H., Keşmir, C., de Boer, R. J. & van Boven, M. Long-term adaptation of the influenza a virus by escaping cytotoxic t-cell recognition. *Scientific reports* **6**, 1–8 (2016).

[16] Fan, K. *et al.* Role of itk signalling in the interaction between influenza a virus and t-cells. *Journal of general virology* **93**, 987–997 (2012).

[17] van de Sandt, C. E. *et al.* Differential recognition of influenza a viruses by m158–66 epitope-specific cd8+ t cells is determined by extraepitopic amino acid residues. *Journal of virology* **90**, 1009–1022 (2016).

[18] Berkhoff, E., Geelhoed-Mieras, M., Fouchier, R., Osterhaus, A. & Rimmelzwaan, G. Assessment of the extent of variation in influenza a virus cytotoxic t-lymphocyte epitopes by using virus-specific cd8+ t-cell clones. *Journal of General Virology* **88**, 530–535 (2007).

[19] Van de Sandt, C. E., Kreijtz, J. H. & Rimmelzwaan, G. F. Evasion of influenza a viruses from innate and adaptive immune responses. *Viruses* **4**, 1438–1476 (2012).

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

[21] Wood, J. M. *et al.* Reproducibility of serology assays for pandemic influenza h1n1: collaborative study to evaluate a candidate who international standard. *Vaccine* **30**, 210–217 (2012).

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

[23] 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).

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

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

[26] 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).

[27] 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).

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

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

[30] van der Meer, F. J. U. M., Orsel, K. & Barkema, H. W. The new influenza A H1N1 virus: balancing on the interface of humans and animals. *The Canadian veterinary journal = La revue veterinaire canadienne* **51**, 56–62 (2010).

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

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

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