Rationale: Influenza A, partly on account of its segmented genome and its wide prevalence in animal hosts, can incorporate genes from multiple strains and (re)emerge as novel human pathogens[1,2], thus harboring a high pandemic potential. Strains spilling over into humans from animal reservoirs is thought to have triggered pandemics at least 4 times (1918 Spanish flu/H1N1, 1957 Asian flu/H2N2, 1968 Hong Kong flu/H3N2, 2009 swine flu/H1N1) in the past century[3]. One approach to mitigating such risk is to 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 reliably and scalably risk-rank individual strains remains limited[4]. CDC's current solution to this problem is the Influenza Risk Assessment Tool (IRAT)[5]. SMEs score strains based on the number of human infections, infection and transmission in laboratory animals, receptor binding, 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 depend on multiple experimental assays, taking weeks/months to compile for a single strain. With tens of thousands of strains being collected annually, this results in a scalability bottleneck.

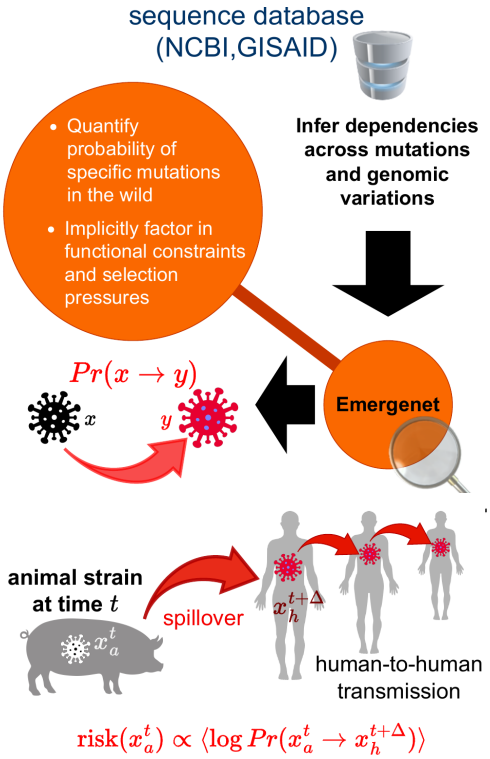

Fig. 1. Theoretical foundation of BioNORAD

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

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

☐ 1) Learning patterns of cross-dependencies across point mutations in key proteins implicated in viral entry and exit (HA and NA) reveals enough of the underlying rules of organization of their primary structures to actionably forecast evolutionary trajectories, i.e., predict future mutations and likelihood of jump events for wild Influenza A viruses in animal reservoirs. Importantly, while a single sequence does not encode enough information to predict its future mutations, discovering patterns from large sequence databases is a viable approach to making such forecasts, as observed past variations are indeed functions of fitness and selection.

☐ 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, thus serving a similar function to the strategic goal of NORAD in the context of defending against geospatial pandemic threats, as opposed to protecting US airspace from adversarial intrusion.

Specific Aims: We have three key aims described below: .

☐ **Aim 1: Formulate a novel metric of sequence similarity (E-distance) that reflects potential of**

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

☐ **Aim 2: Validate E-distance as a similarity metric on strain space that can identify biologically valid sequence variations.** We aim to show that the E-distance may be used to differentiate 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 (Quantify uncertainties in jump-probabilities)** Refine our preliminary result connecting the E-distance to the probability of spontaneous jump from one strain to another, connecting the inference uncertainty arising possibly from sample size limitations to the uncertainty in the jump probability estimates. **T2.2 (Laboratory Experiments)** 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 a working implementation of BioNORAD.** We aim to demonstrate a prototype of the BioNORAD platform for analyzing Influenza A strains at scale for emergence and impact risk. Major tasks are : **T3.1 (IRAT score replication)** Replicate the published IRAT scores, along with uncertainty quantification, within seconds as a validation result. Key subtasks are: (T3.1.1) Investigate how each of the ten dimensions of IRAT comparison map to our Emergenet based risk, and (T3.1.2) Evaluate if the IART scores would change if evaluated at different times e.g. now as opposed to when they were published. T3.2 (BioNORAD Results for Current/Recent Surveillance Data) which comprises the subtasks: (T3.2.1) demonstrating 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. And (T3.2..2) set up an automated pipeline that pulls in sequence data of for new submissions, and publish a risk score automatically. We will collate this information in our pipleine to map the global risk, visualizing where and when am emergenc event is likely, and for which strain/subtype/animal hosts.

Research Strategy and Feasibility: Our approach aims to reliably estimate the non-heuristic numerical probability $\Pr(x \to y)$ of a strain x spontaneously giving rise to y in the wild, thus preempting strains expected to be in future circulation, and approximating IRAT scores of non-human strains without detailed experimental assays or SME scoring. We plan to accomplish this by learning the complex cross-dependencies that constrain what a "valid alteration" of a AA sequence is, by first analyzing variations (point substitutions, indels) of residue sequences of key proteins implicated in cellular entry/exit[3,6], namely HA and NA, and then expanding the analysis to the complete viral genome. By representing these constraints within a predictive framework – the Emergenet (Enet) – we will estimate the odds of a specific mutation to arise in future, and consequently the probability of a specific strain spontaneously evolving into another (Fig. 1). For such explicit calculations we must first infer the variation of mutational probabilities and potential residue replacements from one positional index to the next along the AA sequence. The many well-known classical DNA substitution models[7] or phylogeny inference tools assuming constant species-

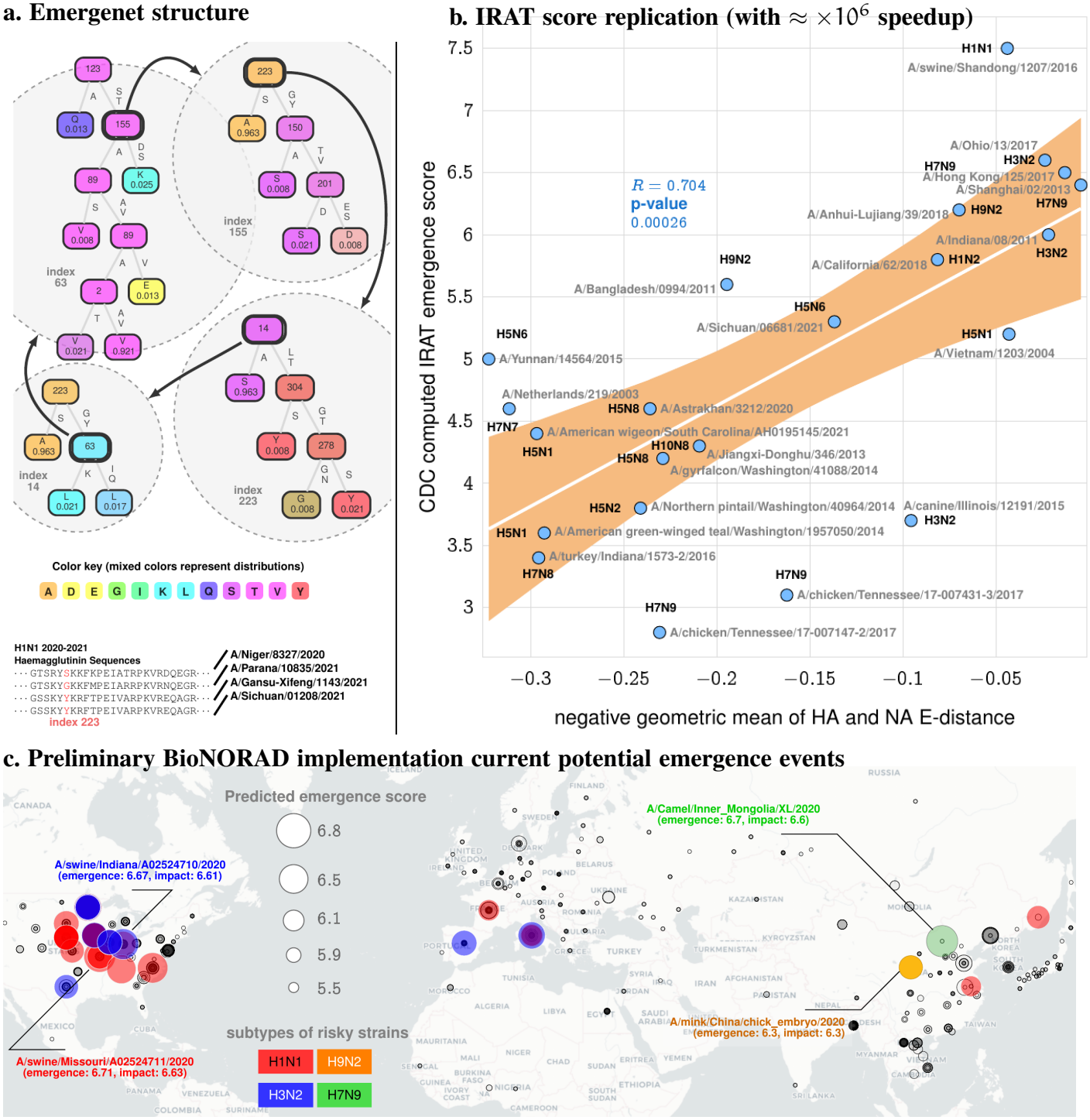

Fig. 2. a, a fragment of an example Emergenet showing the emergent dependencies captured. b, IRAT replication in preliminary analysis with only HA sequence data, c, a preliminary BioNORAD implementation with sequencesd collected over 2021-2022, showing the emerging threat-centers, subtypes.

wise mutational characteristics, are not applicable here. Similarly, newer algorithms such as FluLeap[8] which identifies host tropism and species-level jump-risk[9] do not allow for strain-specific assessment.

The dependencies we expect to uncover are shaped by a functional necessity of conserving replicative fitness. Strains must be sufficiently common to be recorded in surveillance, implying that the sequences from public databases that we train with have high fitness. Lacking kinetic proofreading, Influenza A integrates faulty nucleotides at a relatively high rate ($10^{-3} - 10^{-4}$) during replication[10,11]. However, few variations are actually viable, leading to emergent dependencies between such mutations. Furthermore, these fitness

constraints are time-varying. The background strain distribution, and selection pressure from the evolution of cytotoxic T lymphocyte epitopes[12–16] in humans can change quickly. Our preliminary studies [REF] suggest that we now have enough number of curated sequences in public databases to learn models that automatically factor in the evolving host immunity, and the current background environment.

Structurally, our model structure (an Emergenet) comprises an interdependent collection of local predictors, each aiming to predict the residue at a particular index using as features the residues at other indices (Fig. 2a). These individual predictors are implemented as conditional inference trees[17], in which nodal splits occur with a minimum pre-specified significance in differentiating the downstream child nodes. The set of residues acting as features in each such predictor is automatically identified, *e.g.*, in the fragment of the H1N1 HA Emergenet (2020-2021, Fig 2a), the predictor for residue 63 is dependent on residue 155, and the predictor for 155 is dependent on 223, the predictor for 223 is dependent on 14, and the residue at 14 is again dependent on 63, revealing a cyclic dependency. The complete Emergenet harbors a vast number of such relationships, and each internal node of a tree may be "expanded" to its own tree (since each index is "explained" by residues in other indices, which in turn have their own predictors). Owing to this recursive expansion, a complete Emergenet captures the intricate rules guiding evolutionary change as evidenced by our preliminary validations. Our research strategy comprises the following sequential steps:

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

(Step 2: E-distance metric calculation) Each Emergenet induces an intrinsic distance metric (E-distance) between strains, collected at the time and space associated with the model, defined as the square-root of JS divergence[18] of the conditional residue distributions, averaged over the sequence. Unlike the classical approach of measuring the number of edits between sequences, the E-distance is informed by the Emergenet-inferred dependencies, and adapts to the specific subtype, allele frequencies, and environmental variations. By our recent theoretical result [REF], E-distance approximates the log-likelihood of spontaneous change i.e. $\log \Pr(\mathtt{x} \to \mathtt{y})$, and despite general correlation between E-distance and edit-distance, the E-distance between fixed strains can change if only the background environment changes.

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

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

Assessment of fitness potential emerging zoonotic IAV variants in vitro cell culture. Potential HA variants predicted by the Emergenet models will be generated using the reverse genetics system and evaluated for fitness against parental strains. Briefly, HA segments with potential mutations will be obtained through synthetic gene synthesis. We will assess the relative cell surface expression of parental HA and variants by flow cytometry and western blotting. Next, we will generate recombinant viruses carrying mutant HA using an established reverse genetics system[22–30] 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[31] (sourced from third-party vendors). In addition, we will assess the fitness of individual mutants by fitness competition assay with parental virus (1:1) and determine the relative ratio by high resolution melting (HRM) analysis[32–34] . These studies will help us determine the accuracy of Emergenet framework in predicting pandemic potential variants with enhanced fitness. *Choice of Cell-lines:* A549 cells and primary human lung cells are frequently used in in-vitro IAV experiments due to their relevance to human infection, susceptibility to IAV, reproducibility, ease of cultivation, and compatibility with various molecular and cellular techniques[35]. These cells, derived from human lung carcinoma and lung tissue respectively, serve as appropriate models for studying IAV pathogenesis, and host interactions, as they express key host factors[36,37], and are compatible with molecular and cellular techniques[38]. Dr. Manicassamy has $>$ 15 years of experience in working with human and zoonotic influenza viruses, and safely handling various human pathogens under enhanced BSL2 and BSL3 conditions.

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

Preliminary Validation of BioNORAD. We construct Emergenet models for HA and NA sequences using subtype-specific human strains, typically collected within the year prior to the assessment date, *e.g.*, for A/swine/Shandong/1207/2016 assessed on 06/2020 we use human H1N1 strains collected between 01/07/2019-06/30/2020 for Emergenet inference. For sub-types with few recorded human strains (H1N2, H7N7), we consider all subtype-specific human strains collected up to the assessment date to infer our Emergenet. For subtypes with very few or no recorded human strains even without a lower date bound (H5N2, H5N6, H5N8, H7N8, H9N2, H10N8), we construct the Emergenet using all human strains that match the HA subtype, *e.g.* H5Nx for H5N2, H5N6, and H5N8. This addresses the general concern of "unknown unknowns": allowing Emergenet to assess threats posed by rare or not-yet-human strains. We compute the E-risk for both HA and NA sequences (using the above relationship), finally reporting their geometric mean as our estimated risk for the strain. Considering IRAT emergence scores of 22 strains published by the CDC, we find strong out-of-sample support (correlation of 0.704, pvalue $<$ 0.00026, Fig. 2b). Importantly, each E-risk score is computable in approximately 6 seconds as opposed to potentially weeks taken by IRAT experimental assays and SME evaluation, suggesting a *six order of magnitude speedup*. In the proposed study we will expand our anayis to the complete IAV genome, incorporating all 8 genes, and quantify the modeling uncertainties, and the statistical association of our score with the ten individual scoring elements of IRAT. We show a preliminary implementation of the BioNORAD in Fig. 2c for all 6,066 strains retrieved in 2021/22, showing the localization of potential near-term threat events.

Innovation: While tools exist for ad-hoc quantification of genomic similarity[7,39–43], higher edit-similarity between strains do not imply a high likelihood of a jump. Current surveillance paradigms, while crucial for mapping disease ecosystems, fail to address this challenge. Habitat encroachment, climate change, and other ecological factors[44–46] unquestionably drive up the odds of zoonotic spill-over. Nevertheless, efforts at tracking and modeling these effects till date have not improved our ability to quantify future risk of emergence of a specific strain from a specific host[47]. Recent advances in predicting seasonal epidemic strains[21] do not generalize to predicting emergence events, especially strains that do not yet circulate in humans. This project innovates and envisions a path to acquiring this transformative capability, which is currently well-beyond the state-of-art: the Emergenet algorithm is the first of its kind to learn an appropriate biologically meaningful comparison metric from data, without assuming any model of DNA or AA substitution, or a genealogical tree a priori,enabling an actionable pandemic early warning system.

List of Abbreviations, Acronyms, and Symbols

BioNORAD	Proposed early warning system for pandemic risk (this proposal)
AA residue	Amino Acid residue
AI	Artificial Intelligence
CDC	Centre for Disease Control
CIT	Conditional Inference Trees
DMS	Deep mutational scanning
GISAID	Global Initiative on Sharing All Influenza Data
HA	Hemaglutinnin
IAV	Influenza A virus
IRAT	Influenza Risk Assessment Tool
JS Divergence	Jensen-Shannon Divergence
ML	Machine Learning
NA	Neuraminidase
NCBI	National Center for Biotechnology Information
NORAD	North American Air Defense
RBD	Receptor Binding Site
SME	Subject Matter Expert
UChicago	University of Chicago
UIowa	University of Iowa
UQ	Uncertainty Quantification
WHO	World Health Organization
E-distance	Emergenet similarity between sequences
FluLeap	ML algorithm that uses sequence data to classify influenza viruses as either avian or human
PRMRP	Peer Reviewed Medical Research Program
BSL2	Bio-safety Level 2
BSL3	Bio-safety Level 3
$\theta(x, y)$	E-distance between strains x, y
$\theta^{[t]}(x, y)$	E-distance between strains x, y calculated at time t
H_t	recent human strains H_t at time t of similar subtypes if available. If the specific sub-type is rare, we progressively widen the definition to all strains with similar target gene.
$\rho_t(x)$	time-varying E-risk score for emergence risk of strain x
$\Pr(x \to y)$	Probability of strain x spontaneously mutating to produce strain y in the wild
x_a^t	animal strain observed at time t
x_h^t	human strain observed at time t

DATA MANAGEMENT PLAN

A. Introduction: This Data Management Plan outlines the procedures for managing, storing, and sharing data generated during the course of the project on the analysis of hundreds of thousands of genomic sequences of Influenza A of various subtypes from public repositories (NCBI and GISAID). The research products include models and software, as well as example programs demonstrating how to access and read these models and apply them to raw data. An online system will also be developed to pull in new sequences from public databases, compute emergent risk, and display the results publicly. Experimental validation will involve specific protocols, cell lines, and procedures, which can be shared in an appropriate manner.

B. Data Types and Formats: This project will involve the following types of data:

- Genomic sequences of Influenza A subtypes
- Metadata, including sequence IDs and other relevant information
- Models and software for generating, inferring, and reading the models
- Example programs demonstrating model access and application to raw data
- Experimental validation data, including protocols, cell lines, and procedures

All data will be stored in open and widely used formats, such as FASTA for genomic sequences, JSON or CSV for metadata, and standard programming languages like Python for software and example programs.

C. Data Acquisition and Processing: Data will be acquired from public repositories, such as NCBI and GISAID, and processed using custom-built software tools. These tools will parse the raw data and metadata, analyze the genomic sequences, and generate models to assess emergent risk. Quality assurance and quality control measures will be in place during data collection, analysis, and processing to ensure the reliability of the results.

D. Data Storage and Preservation: Data, models, and software will be stored on secure servers with appropriate backup and version control systems. Genomic sequences and metadata will be deposited in public repositories like NCBI and GISAID, while models and other data will be deposited at Zenodo for long-term access with DOI identifiers. Example programs and software tools will be hosted on GitHub repositories, ensuring easy access and collaboration.

E. Data Sharing: Data sharing will be achieved through a combination of methods:

- Metadata and sequence IDs will be shared publicly, while respecting any restrictions on the genomic sequences themselves
- Models, software, and example programs will be available on GitHub repositories, allowing for easy access, collaboration, and updates
- Tools developed during the project will be easily installed from code registries like PyPI
- Experimental validation data, including protocols, cell lines, and procedures, will be shared in an appropriate and secure manner, ensuring compliance with any legal or ethical requirements
- An online system will provide public access to emergent risk assessment based on new sequences from public databases

F. Preservation Timeframe: Data preservation will be maintained for a minimum of 10 years following the completion of the project. This timeframe will ensure that the research products remain accessible and usable by the scientific community for future research and development.

G. Costs and Administrative Burden: The Data Management Plan takes into consideration the balance between the value of data preservation and other factors such as associated costs and administrative burden. Data storage, preservation, and sharing costs will be factored into the project budget. The justification for any decisions regarding data preservation and sharing will be provided in the plan.

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 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.

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 principal investigators have access to extensive computational and experimental facilities available at the University of Chicago and the University of Iowa to carry out the projects outlined.

University of Chicago facilities, equipment, and other resources: Argonne Leadership Computing Facility (ALCF).

In partnership with Rick Stevens and Argonne National Lab, we will also have access to Argonne's suite of high-performance machines, including ThetaGPU with 28 NVIDIA DGX-3 state-of-the-art AI systems, Polaris, which will be installed in spring, 2021 with 1024 NVIDIA A100 GPUs able to support 7,000 simultaneous training sessions at once for problems that need hyper parameter optimization or for exploring properties of models. Aurora will be installed Summer 2022 and will host 60,000 Intel PVC-XT, performing AI inference at something like 64x faster at over 100 EFs per second, able to train up to 480K models in parallel and some very large models, supporting training models as high as 100 trillion parameters. Human brains have on order 100 trillion synapses (1011 neurons and 1.8-3.2 x 1014 synapses) and so this machine is, in some sense, a brain scale computer, likely the fastest on the planet when installed, and by far the fastest devoted to AI. Finally, we will have access to the National AI Accelerator Assessment Testbed (N3AT) Facility, designed to assess the usefulness of novel hardware designs from dozens of startups for advancing key research problems in AI. These include computers from Cerebras, SambaNova, Graphcore, Groq and many others. These models have the demonstrative ability to training transformer models, famously large in parameters, much more efficiently than within standard computing arrays (see Data and Computation). We will use Globus (www.globus.org) as a platform for data management, which provides managed data transfer solutions for high-speed, reliable, secure file transfer and replication among any variety of storage systems, including on-premise file systems and object stores, desktops and laptops, and cloud storage.

ALCF also houses a range of other high performance computing capabilities spearheaded by its 10-petaflop IBM Blue Gene/Q system, Mira. Mira is comprised of 48 racks, with 1,024 nodes per rack; 768 terabytes of RAM; 35 PB of storage; and 384 I/O nodes.

ALCF also hosts large storage systems. Mira data systems consist of 384 I/O nodes that connect to 16 storage area networks (SANs) that control 8,960 disk drives with a total capacity of 28.8 PB of raw storage and a maximum aggregate transfer speed of 240 GB/s. ALCF computing resources share two 10,000-slot libraries using LTO4 tape technology. The LTO tape drives have built-in hardware compression with compression ratios typically between 1.25:1 and 2:1, depending on the data, giving an effective capacity of 16-24 PB. Two parallel file systems—PVFS and GPFS—are used to access the storage.

ALCF also operates a Globus-based data management and sharing pilot service called Petrel. Petrel provides 1PB of storage that can be used by Argonne researchers and users of Argonne facilities to store and share data with collaborators without the need for local account management.

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.

Compute Infrastructure. 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.

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

- *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 security of the Research Computing Center's storage infrastructure 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.

The Institute for Molecular Engineering at the University of Chicago house a vibrant research community of multidisciplinary scientists that regularly collaborates to make significant scientific contributions. UChicago also features several translational resources, such as the Human Tissue Research Center, and the Transgenic Animal Center.

The University of Chicago Comprehensive Cancer Center (UCCCC): One of only two NCI-designated

Comprehensive Cancer Centers in Illinois, the UCCCC has a reputation for excellence and innovation and a commitment to address cancer through clinical and basic science cancer research and training, clinical cancer care, and expertise in population research. UCCCC researchers have access to a comprehensive set of shared technologies with the University of Chicago Biological Sciences Division (BSD), including 13 Core facilities. The UCCCC offers a wealth of intellectual, technological, and financial resources to pursue a comprehensive, collaborative research program involving more than 215 renowned scientists and clinicians.

Translational and collaborative research: The University of Chicago's strong physical sciences division, including my home department of Chemistry, is located in direct proximity to the medical school and hospital system. Indeed, I chose to start my independent career here at UChicago specifically so that I could develop a group whose work could impact human health. I have now witnessed firsthand the benefits of this proximity and have developed several strategic collaborations with clinicians and clinical researchers to develop new technologies. The University devotes substantial resources to translational research, which provides a clear path to move from the bench to the clinic. This includes the University of Chicago Innovation Exchange (https://innovation.uchicago.edu), which provides seed money and expertise to translate basic science discoveries into commercial ventures and to foster collaborations between the basic science divisions, medical school, and national labs. Therefore, the University of Chicago is an exceptional location to pioneer paradigm-shifting biomedical technologies.

University of Chicago Core Facilities The University of Chicago offers extensive access to cutting-edge research technologies and expertise needed for my work via a well-funded and expertly staffed set of core facilities:

- The Microscopy Core – Maintains an extensive array of microscopes that our group can access, including a variety of confocal and 2-photon fluorescence scopes. The primary microscope we currently use is a Nikon Confocal featuring multiple laser excitation sources, a 2-photon light source, an automated stage, and a heated chamber. This will be very useful for the time course, live cell imaging experiments associated with the proposed research when our group's microscope is insufficient.
- The Biophysics Core - Provides access and training to state-of-the-art biophysics equipment, such as a Biacore, Plate Readers, and a ProteOn XPR36 protein interaction array system.
- Transgenic/ES Cell Technology Mouse Core Facility - Provides genetically manipulated mice through transgenic technology or embryonic stem (ES) cell manipulation. The facility provides a comprehensive set of technical services and a fully operational construction and gene targeting service.

University of Iowa, Facilities and Other Resources: **Laboratory Space.** Dr. Manicassamy's Laboratory is housed in the Department of Microbiology & Immunology on the second floor of the Bowen Science Building (BSB; Core 400) at The University of Iowa. This state of the art facility is comprised of 1,700 square feet of newly renovated laboratory space. In addition to the infrastructure available in the Department of Microbiology & Immunology, Dr. Manicassamy's research is supported by excellent core facilities at the University of Iowa, including Next Generation Sequencing, Bioinformatics Core, Flow Cytometry, Central Microscopy Research Facility, Small Animal Imaging, DNA sequencing. Dr. Manicassamy has full-time administrative support through the Department of Microbiology & Immunology.

Office Space. Dr. Manicassamy has 200 square feet of separate office space adjunct to the laboratory in BSB. PC and Mac computers, computational network, laser printers, color printers, and scanners are available. Manuscripts and desktop publishing of papers can be prepared in several offices available to scientists working in BSB.

Scientific Environment. The University of Iowa has a highly collaborative research community with several leading Virologists and Immunologists, including Drs. Stanley Perlman (Coronaviruses pathogenesis/host responses), Mark Stinski (Herpes Virus), John Harty (T cell responses to Infection), Kevin Legge (dendritic cell-T cell responses to Influenza virus), Steven Varga (Host responses to RSV), Gail Bishop (Viral Immunology), Wendy Maury (Filovirus entry), Richard Roller (Molecular Herpes virology), Jack Stapleton (HIV/HCV pathogenesis), and Hillel Haim (HIV evolution/pathogenesis). Weekly Microbiology seminar series, and the Virology and Immunology journal clubs provide excellent opportunities for students and postdocs for scientific interactions. In addition, research program in pulmonary biology provide an excellent scientific forum for discussion and collaboration. The Levitt Center for Viral Pathogenesis provides funding for seminar speakers and student travel and is another venue for interactions with research interests related to the project. Moreover, the students and postdocs are supported by several NIH T32 training grants.

Animal Care Unit (ACU) at The University of Iowa is in full compliance with all NIH guidelines and regulations pertaining to the care and use of experimental animals (PHS Assurance no. A3021-01). The ACU has enjoyed accreditation from the American Association of Accreditation of Laboratory Animal Care since 1994. The ACU maintains centralized animal housing facilities, which are staffed by highly trained individuals to provide husbandry and research support services. In addition to providing daily animal care, all ordering and receipt of animals, quarantine and health monitoring is performed by the ACU. The ACU also provides research support services, such as anesthesia and surgical support, rodent breeding assistance, diagnostic laboratory services, and investigator training. The ACU veterinarians are faculty members of The University of Iowa. They are available to assist investigators and their staff with all aspects of their animal research activities.

Biosafety Facility. The select agent containment laboratories are housed on the 5 rd floor of the Carver Biomedical Research Building (CBRB) and on the 4 th floor of the Medical Laboratories (ML) Building, University of Iowa, Iowa City, IA. BSL3 facilities are CDC certified and is managed by Ms. Dana Reis (Director). Dr. Manicassamy has one BSL3 suite with one class IIb biosafety cabinet and one Animal BSL3 suite available for work on highly pathogenic influenza viruses and coronaviruses.

Major Equipment in Dr. Manicassamy's laboratory::

1) **Biosafety cabinets:** Sterile environment for handling infectious materials, safeguarding researcher and samples.
2) **Tissue culture incubators:** Optimal temperature, humidity, and gas conditions for cell and tissue growth.
3) **PCR machines and real-time PCR machines:** Amplify specific DNA sequences, monitor amplification process in real-time.
4) **Table top centrifuges:** Separate liquid sample components based on density.
5) **Microscopes:** Visualize microscopic organisms, cells, and minute structures.
6) **-80 freezers and -20 freezers:** Long-term storage and preservation of biological samples.
7) **Bacterial incubators:** Optimal conditions for bacterial growth and development.
8) **Bacterial shakers:** Facilitate aeration and mixing of bacterial cultures.
9) **Thermal cyclers:** Precise temperature control for DNA amplification and molecular biology applications.
10) **UV-Spectrophotometers:** Measure light absorbance, determine concentration and purity of biomolecules.
11) **Fluorescence microscope:** Visualize and analyze samples using fluorescence, study cellular structures and biomolecular interactions.
12) **Gel doc unit:** Capture and document images of DNA, RNA, and protein samples in gel electrophoresis experiments.

UI Department of Microbiology and Immunology Resources:

1) **3 flow cytometers:** Measure and analyze physical and chemical properties of cells or particles in fluid.
2) **Fluorescence plate readers:** Measure fluorescence intensity in microplate format, high-throughput analysis of cellular and biochemical events.
3) **Zeiss inverted fluorescent and confocal microscopes:** High-resolution imaging of biological samples, visualize living cells and tissues.
4) **qRT-PCR thermocyclers:** Quantify RNA transcripts in real-time, provide insights into gene expression.
5) **Fuji CCD imaging system:** Capture high-resolution images of fluorescent and chemiluminescent samples, sensitive and accurate detection of biomolecules.
6) **High-speed and ultracentrifuges with varied rotors:** Rapid separation of samples based on size, shape, and density, accommodate various rotor types.
7) **Typhoon imaging system:** Detect, quantify, and analyze proteins, nucleic acids, and biomolecules in gels, membranes, and microplates.
8) **ELISA plate readers:** Measure absorbance of enzyme-linked immunosorbent assays (ELISA), quantify proteins, peptides, and hormones.
9) **TopCount:** High-throughput quantification of radioactivity in samples, study biochemical and cellular processes.
10) **LiCor imaging system:** Sensitive and accurate detection of fluorescent and chemiluminescent samples in various formats.

11) **Darkroom facilities:** Controlled processing of light-sensitive materials, such as photographic films and imaging plates.
12) **Cold and warm rooms:** Temperature-controlled spaces for storage and experiments requiring specific temperature conditions.

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

H. 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.

I. 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.

J. 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.

K. 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

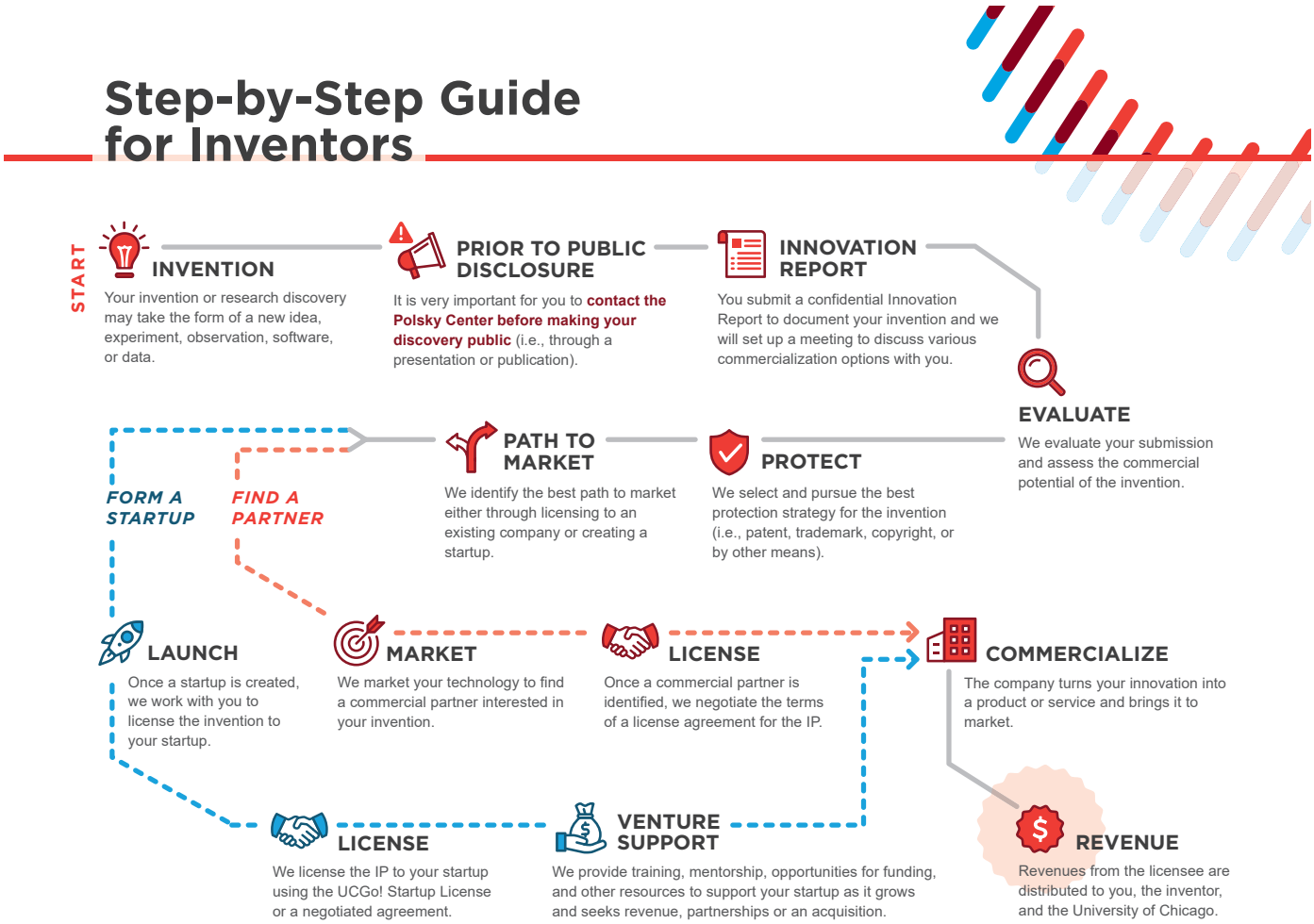

Fig. 3. Inventor pathway to commercialization at the University of Chicago

of 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 3).

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

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.

REFERENCES

[1] 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).

[2] Dos Santos, G., Neumeier, E. & Bekkat-Berkani, R. Influenza: Can we cope better with the unpredictable? *Human vaccines & immunotherapeutics* **12**, 699–708 (2016).

[3] 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).

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

[5] 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).

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

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

[8] 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).

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

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

[12] 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).

[13] 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).

[14] 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).

[15] 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).

[16] 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).

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

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

[19] 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).

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

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

[22] Zhao, X. *et al.* Expanding the tolerance of segmented Influenza A Virus genome using a balance compensation strategy. *PLoS Pathog* **18**, e1010756 (2022).

[23] Ganti, K., Han, J., Manicassamy, B. & Lowen, A. C. Rab11a mediates cell-cell spread and reassortment of influenza A virus genomes via tunneling nanotubes. *PLoS Pathog* **17**, e1009321 (2021).

[24] Han, J. *et al.* Host factor Rab11a is critical for efficient assembly of influenza A virus genomic segments. *PLoS Pathog* **17**, e1009517 (2021).

[25] Kandasamy, M., Furlong, K., Perez, J. T., Manicassamy, S. & Manicassamy, B. Suppression of Cytotoxic T Cell Functions and Decreased Levels of Tissue-Resident Memory T Cells during H5N1 Infection. *J Virol* **94** (2020).

[26] Li, P. *et al.* Luciferase. *Viruses* **10** (2018).

[27] Tundup, S. *et al.* Endothelial cell tropism is a determinant of H5N1 pathogenesis in mammalian species. *PLoS Pathog* **13**, e1006270 (2017).

[28] Perez, J. T., a Sastre, A. & Manicassamy, B. Insertion of a GFP reporter gene in influenza virus. *Curr Protoc Microbiol* **Chapter 15**, 1–15 (2013).

[29] Manicassamy, B. *et al.* Analysis of in vivo dynamics of influenza virus infection in mice using a GFP reporter virus. *Proc Natl Acad Sci U S A* **107**, 11531–11536 (2010).

[30] Manicassamy, B. *et al.* Protection of mice against lethal challenge with 2009 H1N1 influenza A virus by 1918-like and classical swine H1N1 based vaccines. *PLoS Pathog* **6**, e1000745 (2010).

[31] Medina, R. A. *et al.* Glycosylations in the globular head of the hemagglutinin protein modulate the virulence and antigenic properties of the H1N1 influenza viruses. *Sci Transl Med* **5**, 187ra70 (2013).

[32] Ganti, K., Han, J., Manicassamy, B. & Lowen, A. C. Rab11a mediates cell-cell spread and reassortment of influenza a virus genomes via tunneling nanotubes. *PLoS Pathogens* **17**, e1009321 (2021).

[33] Wittwer, C. T., Reed, G. H., Gundry, C. N., Vandersteen, J. G. & Pryor, R. J. High-resolution genotyping by amplicon melting analysis using lcgreen. *Clinical chemistry* **49**, 853–860 (2003).

[34] Marshall, N., Priyamvada, L., Ende, Z., Steel, J. & Lowen, A. C. Influenza virus reassortment occurs with high frequency in the absence of segment mismatch. *PLoS pathogens* **9**, e1003421 (2013).

[35] Matrosovich, M. *et al.* Avian influenza a viruses differ from human viruses by recognition of sialyloligosaccharides and gangliosides and by a higher conservation of the ha receptor-binding site. *Virology* **233**, 224–234 (1997).

[36] Shinya, K. *et al.* Influenza virus receptors in the human airway. *Nature* **440**, 435–436 (2006).

[37] Chan, M. C. *et al.* Tropism and innate host responses of the 2009 pandemic h1n1 influenza virus in ex vivo and in vitro cultures of human conjunctiva and respiratory tract. *The American journal of pathology* **176**, 1828–1840 (2010).

[38] Neumann, G. *et al.* Generation of influenza a viruses entirely from cloned cdnas. *Proceedings of the National Academy of Sciences* **96**, 9345–9350 (1999).

[39] 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).

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

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

[42] 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).

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

[44] Rulli, M. C., Santini, M., Hayman, D. T. & D'Odorico, P. The nexus between forest fragmentation in africa and ebola virus disease outbreaks. *Scientific reports* **7**, 41613 (2017).

[45] Chua, K. B., Chua, B. H. & Wang, C. W. Anthropogenic deforestation, el niiio and the emergence of nipah virus in malaysia. *Malaysian Journal of Pathology* **24**, 15–21 (2002).

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

[47] Fair, J. & Fair, J. Viral forecasting, pathogen cataloging, and disease ecosystem mapping: Measuring returns on investments. *Current Topics in Microbiology and Immunology* 75–83 (2019).