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

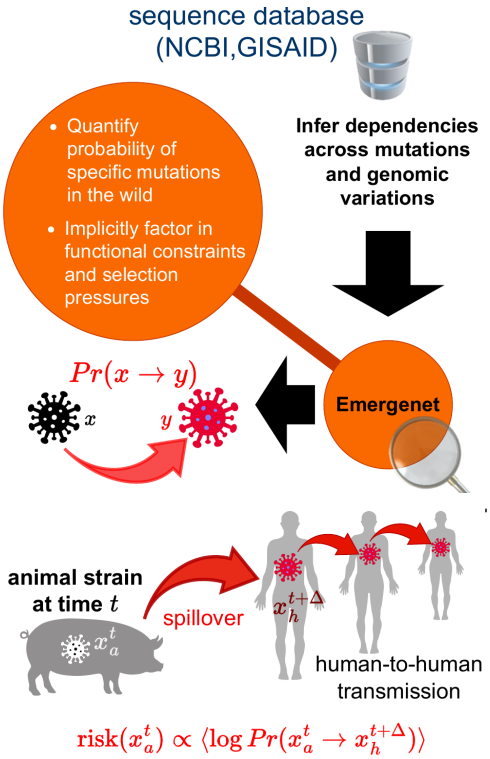

Fig. 1. Theoretical foundation of BioNORAD

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

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

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

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

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

Specific Aims: We have three key aims described below:

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

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

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

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

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

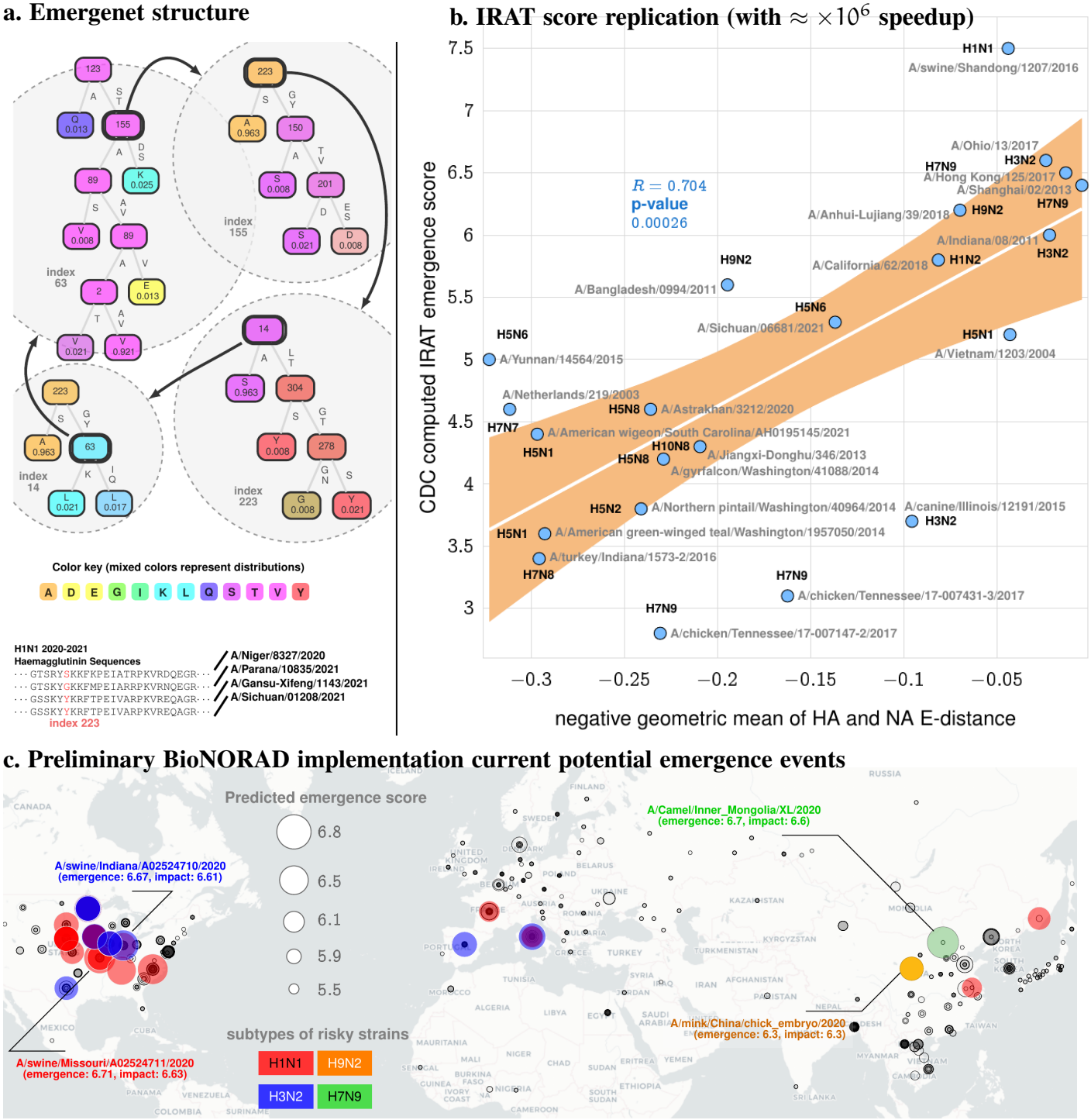

Fig. 2. a, a fragment of an example Emergenet showing the emergent dependencies captured. b, IRAT replication in preliminary analysis with only HA sequence data, c, a preliminary BioNORAD implementation with sequences collected over 2021-2022, showing the emerging threat-centers, subtypes.

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

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

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

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

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

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

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

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

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

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

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

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

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

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
H1N1, H3N2, H5N1, H5N2, H9N2, H5N8, H7N8, H5N6, H10N8	Influenza A strains

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. Handling of Restricted Data: In cases where genomic sequences cannot be shared publicly due to legal or ethical constraints, the project will ensure that the handling and sharing of such data is in compliance with relevant regulations and guidelines. Sequence IDs and metadata will be shared, allowing other researchers to request access to the restricted sequences through appropriate channels. The project will have provisions in place for using sequences that are not publicly posted if the researchers or laboratories who collected the sequences do not grant permission to share them publicly. In such cases, the project team will collaborate with the data owners to determine the most appropriate way to utilize and share the data, ensuring that all parties' interests are respected and protected.

G. Monitoring Adherence to the Data Management Plan: Adherence to the Data Management Plan will be monitored throughout the project by a dedicated postdoctoral researcher, who will dedicate 5% of their time to this task. The postdoctoral researcher will work under the supervision of the Principal Investigator, ensuring that the Data Management Plan is properly executed and that all project members are aware of their data management responsibilities.

Monitoring will include periodic reviews of data storage, sharing, and preservation practices, as well as ensuring that data quality assurance and quality control measures are effectively implemented. Any deviations from the plan will be addressed promptly, and the plan will be updated as needed to accommodate

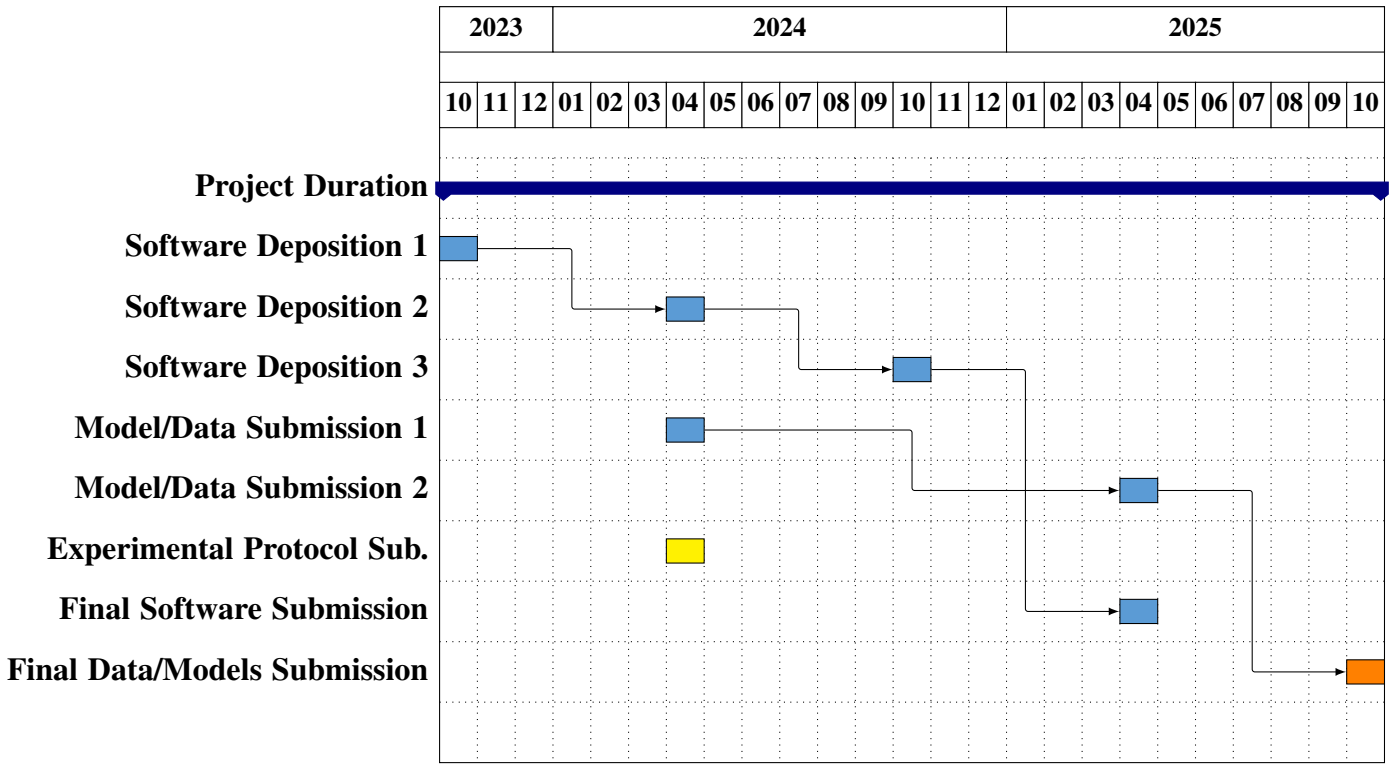

Fig. 3. DMP checkpoints. Software is deposited at Github repositories. Models and data are submitted at Zenodo.

new data types, formats, or sharing requirements.

H. 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.

I. 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. A total of 5% of the project's effort will be allocated to data management, including the postdoctoral researcher's time and any other administrative tasks related to data management.

The justification for any decisions regarding data preservation and sharing will be provided in the plan. This approach ensures that the research products are appropriately managed and shared, while also taking into account the costs and administrative burden associated with data management.

J. Periodic Review of the Data Management Plan: The Data Management Plan will be reviewed every six months by the Principal Investigator (PI) and co-Principal Investigator (co-PI) to ensure that it remains up-to-date and relevant to the project's needs. During these reviews, the PI and co-PI will assess the current data management practices and identify any necessary updates or changes to the plan.

These periodic reviews will help ensure that the Data Management Plan continues to be effective in guiding the project's data management activities and that it evolves as needed to accommodate new data types, formats, or sharing requirements. Any updates to the plan will be communicated to all project members to ensure that everyone remains informed about the project's data management expectations and responsibilities.

K. Compliance with DoD Instructions: This Data Management Plan adheres to the guidelines set forth in Section 3.c. Enclosure 3 of the Department of Defense (DoD) Instructions 3200.12. By following these guidelines, the project ensures that all data management practices are in compliance with the requirements of the DoD and that the data generated during the project will be appropriately managed, stored, and shared.

TECHNICAL ABSTRACT

Our 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.

Influenza A viruses, with their segmented genome and high prevalence in animal hosts, can recombine genes from animal strains and (re)emerge as novel human pathogens. Pandemics triggered by animal Influenza A strains spilling over into humans have occurred multiple times in the past century. Mitigating such risk involves identifying animal strains that have high odds of spilling over into humans and rapidly achieving human-to-human transmission capability. While global surveillance efforts collect wild specimens annually, our ability to reliably and scalably risk-rank individual strains according their pandemic potential remains limited. CDC's current solution to this problem is the Influenza Risk Assessment Tool (IRAT), which is based on subject matter experts scoring strains-of-concern on ten different phenotypic aspects related to transmission and pathogenicity on the basis of multiple experimental assays. Thus scoring each strain can take weeks to months. With tens of thousands of strains being collected annually, this presents a scalability bottleneck.

Here we propose to develop the BioNORAD platform to proactively identify the risks posed by emerging Influenza A strains with high pandemic potential, akin to the strategic function of NORAD in defending the North American airspace. Powered by novel pattern discovery algorithms, our proposed algorithms will automatically parse emergent evolutionary constraints on Influenza A viruses in the wild. This platform will provide a less-heuristic, theory-backed, experimentally validated, scalable solution to emergence prediction, preempting strains expected in future human circulation and approximating IRAT scores of non-human strains a million times faster (seconds as opposed to weeks/months) without experimental assays or SME scoring.

Our central insight here is that to preempt strains expected to be in future circulation, we need to reliably estimate the non-heuristic numerical probability of a strain x spontaneously giving rise to strain y in the wild. To accomplish this, we aim to learn the complex cross-dependencies that constrain what a "valid alteration" of an amino acid (AA) sequence looks like (random perturbations are expected to be large deleterious), by first analyzing variations of residue sequences of key proteins implicated in cellular entry/exit, namely HA and NA, and then expanding the analysis to the complete viral genome. Our core algorithm, the Emergenet, leverages the inferred cross-dependencies to estimate the odds of a specific mutation arising in the future, and consequently the probability of a specific strain spontaneously evolving into another.

We will validate our ability to predict future mutations by showing that Emergenet predicted HA variants express correctly, are functional, and maintain fitness unlike random modifications, via demonstrating proper folding, cell surface expression of variants by flow cytometry as well as evaluate the fitness of recombinant viruses in 1:1 competition experiments with the unperturbed parental strain.

Finally these models will be assembled into a integrated platform (the BioNORAD) that shows global near-time pandemic threat events as they emerge (from observation of new animal strains, as they are uploaded to the platform), along with estimates of the threat severity and estimated time to emergence.

Current surveillance paradigms, while crucial for mapping disease ecosystems, fail to address the challenge that a higher edit-similarity between strains does not not imply a high likelihood of jump. Also, recent advances in predicting seasonal strains do not generalize to predicting emergence events, especially strains that do not yet circulate in humans, underscoring the problem of unknown unknowns. This project innovates and envisions a path to acquiring this transformative capability well-beyond the state-of-art.

The BioNORAD platform is crucial to US national interest, particularly in the context of protecting DoD assets and personnel deployed in potentially high-risk centers of emergence. Additionally, the platform will enable preemptive action, such as the inoculation of animal reservoirs before the first human infection, potentially eliminating the pandemic before it has a chance to trigger. The investment in the BioNORAD platform is thus 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. The platform's ability to automatically factor in the evolving host immunity and the current background environment sets it apart from current state-of-the-art methods, making it a vital asset in the fight against pandemic threats and safeguarding military personnel, assets, and global health security.

LAY ABSTRACT

Our 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.

Influenza A viruses have the potential to cause devastating pandemics, and ongoing surveillance efforts are insufficient to reliably identify and rank strains with high pandemic potential. The current solution to this problem, the Influenza Risk Assessment Tool (IRAT) by the CDC, is subjective and slow requiring multiple experimental assays, and thus cannot keep up with the scale of current global biosurveillance efforts.

To address this challenge, we propose to develop the BioNORAD platform to proactively identify the risks posed by emerging Influenza A strains with high pandemic potential. The platform will use novel pattern discovery algorithms to automatically parse evolutionary constraints operating on Influenza A viruses in the wild, and provide an experimentally validated scalable solution to emergence prediction.

The Emergenet algorithm, at the core of the BioNORAD platform, is the first of its kind to learn an adaptive biologically meaningful comparison metric from data, enabling an actionable pandemic early warning system. By reliably estimating the numerical probability of a strain giving rise to another strain in the wild, the platform can preempt strains expected to be in future circulation, and approximate IRAT scores of non-human strains without experimental assays or subject matter expert scoring.

We will validate this core algorithm by demonstrating that Emergenet predicted variations of genes from strains observed in the wild express correctly on host cell surfaces, are functional, and do not lose replicative fitness, and hence present potential future variants.

Ultimately the inferred models will be used to power the BioNORAD platform that can track global threats from potential near-time emergence events as new strains are uploaded to the platform, along with estimates of the threat severity and estimated time to emergence.

This represents a transformative capability, addressing the problem of unknown unknowns in the field of emergence modeling: state of the art tools are severely limited in their ability to predict or track emergence of novel strains. This knowledge gap renders the current substantial bio-surveillance efforts largely ineffective for precise actionable prediction of emergence events. These surveillance paradigms, while crucial for mapping disease ecosystems, fail to address the challenge of higher edit-similarity between strains not implying a high likelihood of a jump. Ongoing efforts at tracking and modeling ecological factors driving up the odds of zoonotic spill-over, while expensive, have not improved our ability to quantify future risk of emergence of a specific strain from a specific host. Recent reported advances in predicting seasonal epidemic strain distributions do not generalize to predicting emergence events as well, especially for novel subtypes that do not yet circulate in humans.

By automating the analysis of influenza strains, the BioNORAD platform has the potential to revolutionize our ability to prevent pandemics before they emerge. It is a transformative technology that is urgently needed in a world where pandemics are a growing concern. By providing early warning of emerging strains, the platform can help prevent widespread outbreaks and save lives.

Furthermore, the BioNORAD platform has the potential to enable more proactive vaccination efforts. By identifying animal reservoirs of emerging strains before they jump to humans, public health officials could vaccinate these animals and prevent the emergence of pandemic strains, saving countless lives.

The BioNORAD platform can play a crucial role in protecting DoD assets and personnel deployed in potentially high-risk centers of emergence, and may make prophylactic inoculation possible against future strains. Thus, 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.

Our project is a collaborative effort between experts in computer science, biology, and epidemiology, and all tools, software and data generated will be shared in accordance with DoD mandates. This interdisciplinary approach is essential for developing a platform that can effectively address the complex problem of predicting pandemic risk. By bringing together experts from different fields, the platform can take advantage of cutting-edge advances in machine learning, bioinformatics, and epidemiology.

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 sequence similarity metric E-distance	**Timeline (Months)**	**Site 1**	**Site 2**
Major Task 1.1: Emergenet Development			
Subtask T1.1.1: Precisely formulate the Emergenet inference platform, that puts together ML algorithms capturing maximally predictive patterns of change and mutational dependencies	1-3	✓	
Subtask T1.1.2: Provide uncertainty quantification for the inferred patterns represented in the Emergenet models	2-6	✓	
Milestones Achieved: M1) Emergenet software beta release M2) Uncertainty quantification of inferred models			
Major Task 1.2: Sample-complexity for Emergenet inference			
Milestones Achieved: M3) Sample-complexity estimates complete			
Major Task 1.3: Event Timeline Estimation			
Subtask T1.3.1: Map mutational change dynamics to "wall-time" to forecast *when* future variants will show up	3-9	✓	
Subtask T1.3.2: Computationally validate timeline predictions using records of past emergence events	9-12	✓	
Milestones Achieved: M4) Software update for timeline estimation released			

Specific Aim 2: Validate E-distance as a similarity metric on strain space identifying biologically valid sequence variations	**Timeline (Months)**	**Site 1**	**Site 2**
Major Task 2.1: Quantify asymmetric transition probabilities between strains			
Subtask 2.1.1 Develop analytical framework to identify probabilistic movement direction between strains	9-12	✓	
Subtask 2.1.2 Develop analytics to chart multi-step probabilistic trajectories from observed strains	12-15	✓	
Milestones Achieved: M5) Software update for future trajectory calculation			
Major Task 2: Assessment of fitness of potential emerging zoonotic IAV variants in cell cultures			
Subtask 2.2.1 Shortlist HA variants with maximal emergence probability of H3N2 and H1N1 subtypes	6-9	✓	✓
Subtask 2.2.2 Generate predicted HA variants using reverse genetics in human lung epithelial cell line (A549) and primary human lung cells	9-22		✓
Subtask 2.2.3 Evaluate generated variants for replicative fitness	15-24		✓
Milestones Achieved: M6) Emergenet experimental validation complete			

Specific Aim 3: Develop a working implementation of BioNORAD	Timeline (Months)	Site 1	Site 2
Major Task 3.1 IRAT score replication			
Subtask T3.1.1 Investigate how each of the ten dimensions of IRAT comparison map to our Emergenet based risk	12-18	✓	
Subtask T3.1.2 Evaluate sensitivity of IRAT scores to risk-assessment timepoint	15-18	✓	
Subtask T3.1.3 Incorporate event timeline estimation in BioNORAD prototype to predict time to emergence	18-24	✓	
Milestones Achieved: M7) IRAT score-based validation of BioNORAD prototype			
Major Task 2: Results for Current/Recent Surveillance Data			
Subtask T3.2.1 Demonstrate we can analyze collected sequences at scale, by enumerating the risk profiles of all sequences collected recently within teh last few years	18-24	✓	
Subtask T3.2.2 Set up an automated pipeline that pulls in sequence data of for new submissions, and publish a risk score automatically	18-24	✓	
Milestones Achieved: M8) Working version of BioNORAD platform demonstrated M9) Final project report submitted			

DELIVERABLES

☐ Emergenet software

☐ BioNORAD implementation

☐ Experimental protocols for fitness assessment of Emergenet predicted strains

TIMELINE

The project timeline, indicating various milestones and tasks is illustrated in Fig. 4

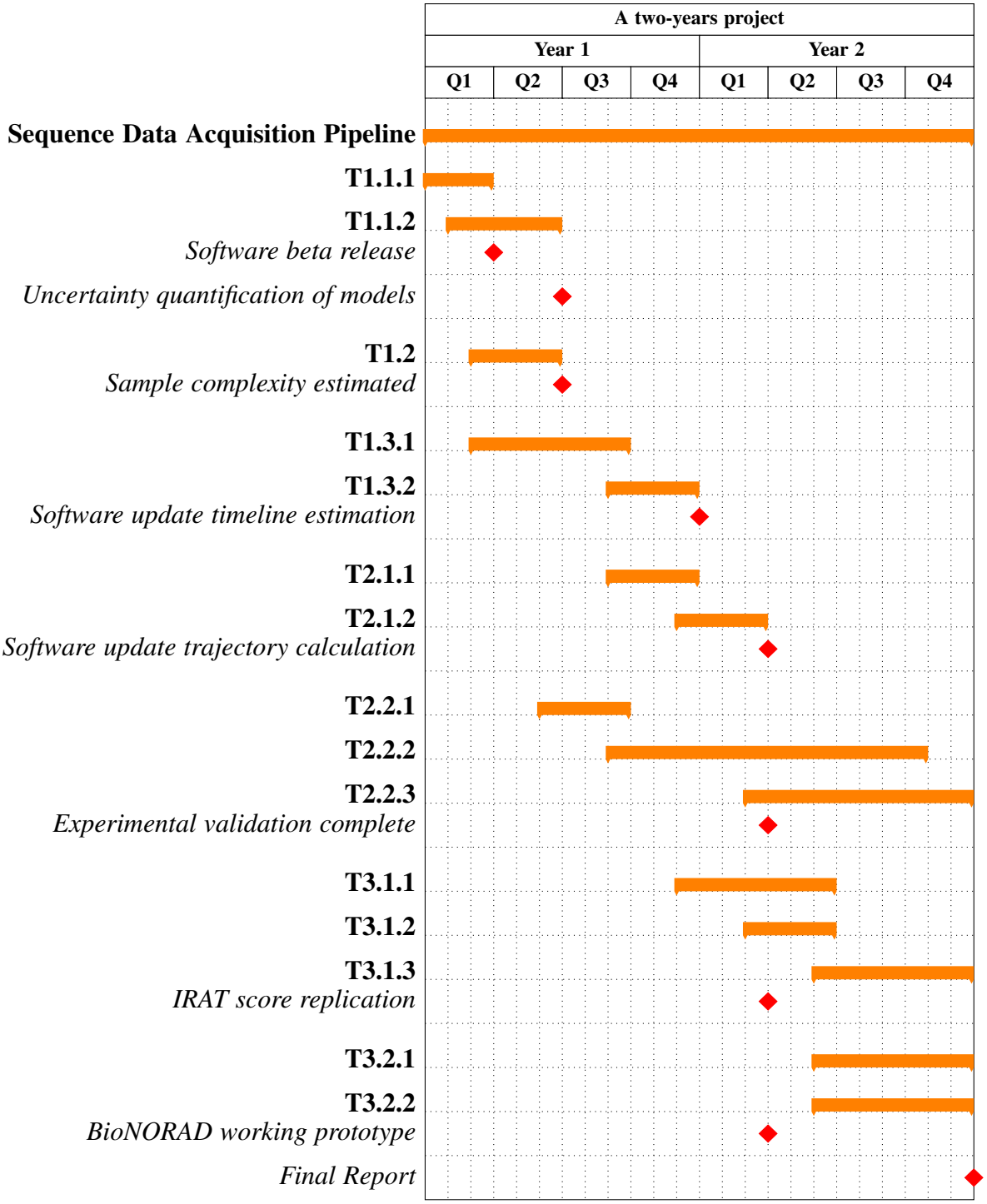

Fig. 4. Gantt chart of project timeline

Impact Statement

Relevance to FY23 PRMRP Topic Area: The proposed research project is highly relevant to the proteomics topic area, specifically as it applies to infectious diseases caused by zoonotic emergence. The Emergenet algorithm can distill the emergent rules of organization of key viral proteins, uncovering how proteins respond to dynamic fitness landscapes and adapt to new hosts while evading host immune defenses. These investigations can provide key insights into protein evolution and reveal novel patterns that can aid in the discovery of vaccines and prevent the emergence of new virus strains.

Relevance to FY23 PRMRP Strategic Goal: The proposed research project addresses the strategic goal of epidemiology, with a focus on identifying strategies for surveillance or developing modeling tools and/or biomarkers to predict outbreaks or epidemics. The BioNORAD platform provides a theory-backed, experimentally validated, and scalable solution to emergence prediction, preempting strains expected in future human circulation and approximating Influenza Risk Assessment Tool (IRAT) scores of non-human strains without experimental assays or subject matter expert scoringa million times faster. As a comparison, while 22 strains have been scored by the IRAT since 2013, we can analyze and score the entire GISAID and NCBI database in under a week with a moderately powerful computing cluster. This represents a significant improvement in our ability to predict pandemics caused by Influenza A viruses.

Potential impact, either short-term or long-term on the field of study and/or patient care: The potential impact of the proposed research on the field of study and patient care is immense. Currently, pandemic preparedness and response is reactive and slow, often leading to significant loss of life and economic damage. The ability to accurately predict the emergence of high-risk pandemic strains before they have the opportunity to trigger a pandemic would revolutionize our approach to pandemic preparedness and response, allowing us to rapidly develop and produce vaccines tailored to the specific strain before the pandemic even begins, potentially saving countless lives.

Furthermore, by understanding the evolutionary constraints and cross-dependencies that give rise to high-risk strains, we can begin to develop targeted interventions to prevent these strains from emerging in the first place. This could involve inoculating animal reservoirs with vaccines tailored to these high-risk strains or developing targeted surveillance efforts in animals before they have the opportunity to jump to humans.

The potential impact of this work is not limited to pandemic preparedness, and can foster a more nuanced understanding of how viruses evolve and adapt over time. This could lead to the development of new therapeutic strategies that specifically target the evolutionary pathways of viruses, potentially revolutionizing the treatment of viral infections. In addition, the development of the BioNORAD platform has implications beyond Influenza A viruses. The methods developed as part of this work can be applied to other viruses with pandemic potential, such as coronaviruses and other respiratory viruses.

Overall, the potential impact of this work is significant, both in the short-term and long-term. By developing a better understanding of the evolutionary forces that give rise to high-risk pandemic strains and by providing an early warning system to identify emerging strains before they have the opportunity to trigger a pandemic, we can revolutionize our approach to pandemic preparedness and response, making this research project a critical investment in the future of our society.

Potential to generate preliminary data that can be used as a foundation for future research projects: The BioNORAD platform's ability to predict the emergence of novel Influenza A strains with high pandemic potential, as well as estimate their evolutionary trajectories, will generate a wealth of data that can be used as a foundation for future research projects. The knowledge and insights gained from this research will facilitate the development of more effective and efficient biosurveillance methods and tools, as well as the design of new therapeutics and vaccines that target specific viral strains. Furthermore, this work will provide a framework for future studies aimed at predicting the emergence of other infectious diseases with pandemic potential.

In addition, the Emergenet algorithm's ability 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, represents a significant advance in the field of machine learning. This has the potential to impact a wide range of research areas beyond infectious disease epidemiology, such as protein structure prediction and drug design. The proposed research will not only advance our understanding of Influenza A evolution and pandemic risk assessment but also contribute to the broader scientific community's knowledge of machine learning and artificial intelligence.

Relevance to Military Health Statement

The proposed research project is highly responsive to the healthcare needs of military service members, veterans, and beneficiaries, specifically in regards to pandemic preparedness and response. 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. Here we briefly discuss the incidence and/or prevalence of influenza in the general population as well as in military service members, veterans, and beneficiaries, describe the population(s)/dataset(s) to be used in the proposed research project, and provide a description of how the knowledge, information, products, technologies, or applications gained from the research could be implemented in a dual-use capacity to benefit both military and civilian populations.

Incidence and Prevalence of Influenza: Influenza is a highly infectious respiratory illness caused by the influenza virus. It affects millions of people worldwide each year and is responsible for a significant number of hospitalizations and deaths. According to the Centers for Disease Control and Prevention (CDC), in the United States alone, the estimated number of influenza-associated illnesses ranges from 9 million to 45 million annually, resulting in between 140,000 and 810,000 hospitalizations and between 12,000 and 61,000 deaths in a typical flu season[49]. Military service members, veterans, and beneficiaries are not immune to the effects of influenza, and the Department of Defense (DoD) reports that influenza-like illness is one of the top five reasons for medical encounters among service members[50]. In addition, influenza can have a significant impact on military readiness, with outbreaks resulting in decreased personnel availability and decreased mission effectiveness[51].

Population and Dataset: The proposed research project will utilize a combination of publicly available genetic sequence data and epidemiological data from influenza surveillance programs. Specifically, the Global Initiative on Sharing All Influenza Data (GISAID) and the National Center for Biotechnology Information (NCBI) databases will be used to collect genetic sequence data on influenza viruses collected from human and animal hosts. Epidemiological data will be obtained from the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC), including information on the geographic distribution of influenza strains and the incidence and prevalence of disease.

The use of these datasets is highly appropriate for the proposed research project. The GISAID and NCBI databases contain a wealth of genetic sequence data on influenza viruses collected from around the world. These databases are regularly updated and publicly available, ensuring that the research can be replicated and extended by other researchers. Additionally, the epidemiological data obtained from WHO and CDC are widely recognized as authoritative sources of information on influenza surveillance and are used by public health agencies around the world to track the spread of influenza.

Accessing these datasets is feasible, as they are publicly available and regularly updated. The use of these datasets also enables the research to be conducted at scale, allowing for the identification of emerging strains and the analysis of their evolutionary trajectories in a timely and efficient manner.

Dual-Use Capacity: The knowledge and technologies gained from this research could be implemented in a dual-use capacity to benefit both military and civilian populations. The ability to predict and preempt the emergence of pandemic strains of influenza viruses has significant implications for global public health and biosurveillance efforts. By developing the BioNORAD platform, we can not only protect military personnel but also contribute to global efforts to prevent and respond to emerging infectious diseases.

The knowledge gained from this research project could also be applied to other infectious diseases with pandemic potential, such as coronaviruses and other respiratory viruses. The methods and algorithms developed as part of this work can be applied to other viruses with pandemic potential, potentially improving our ability to predict and respond to future pandemics, enabling us to act quickly and decisively to prevent the spread of the virus and save lives.

In conclusion, the proposed research project is highly responsive to the health care needs of military Service Members, Veterans, and/or beneficiaries. Influenza viruses are a major threat to military personnel deployed overseas, and the ability to predict and preempt pandemic strains of influenza viruses has significant implications for the health and safety of military personnel. The use of publicly available datasets and the interdisciplinary collaboration involved in this research project make it a highly feasible and appropriate approach to the problem. Finally, the dual-use potential of the research has significant implications for global public health and biosurveillance efforts, making it a critical investment in the future of both military and civilian populations.

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.

Computational Facilities: The principal investigator has 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:

Research Computing Center (RCC) at the University of Chicago: The Research Computing Center (RCC) at the University of Chicago is a hub for researchers across various fields, offering advanced computing resources, storage systems, and a wide range of software packages to facilitate high-performance computing. This document provides an extensive overview of the RCC's facilities and capabilities, focusing on the Midway2 supercomputing cluster, storage and backup systems, software, and networking infrastructure.

Midway2 Supercomputing Cluster: Midway2 is the second-generation high-performance computing cluster at RCC, replacing the first-generation cluster, Midway, which was decommissioned in 2019. Midway2 comprises a large pool of servers, software, and storage that researchers can utilize to enhance the efficiency and scale of their computational science. The RCC provides resources for distributed computing and shared memory computing, as well as emerging technologies like accelerators and big-data systems.

University of Chicago researchers can access RCC resources for free. To extend RCC resources with additional storage and computation, researchers can refer to the Cluster Partnership Program.

Cluster Computing Resources: The RCC maintains three pools of servers for parallel and distributed high-performance computing:

1) Tightly-coupled nodes: Ideal for tightly coupled parallel computing tasks, these nodes on Midway2 have a fully non-blocking FDR and EDR Infiniband, providing up to 100Gbps interconnect.
2) Loosely-coupled nodes: Similar to the tightly-coupled nodes but connected with 40Gbps GigE instead of Infiniband, these nodes are best suited for distributed tasks.
3) Shared memory nodes: These nodes contain much larger main memories (up to 1 TB) and are ideal for memory-bound tasks.

RCC maintains Intel Broadwell (28 cores @ 2.4 GHz with 64 GB memory per node) and Intel Skylake (40 cores @ 2.4 GHz with 96 GB memory per node) CPU architectures.

RCC also maintains specialty nodes, such as large shared memory nodes with up to 1TB of memory per node and 16, 28, or 32 Intel CPU cores. At the time of writing, Midway2 contains a total of 16,016 cores across 572 nodes and 2.2 PB of storage.

Emerging Technology: RCC's emerging technology resources allow researchers to be at the cutting edge of scientific computing. These include:

1) Hadoop: A framework for large-scale data processing based on Google's paper and initially developed at Yahoo. Researchers can experiment with RCC's Hadoop infrastructure to become familiar with big data techniques.
2) GPU Computing: Scientific computing on graphics cards can unlock even greater amounts of parallelism from code. GPU nodes on Midway each have four Nvidia K80 accelerator cards and are integrated into the Infiniband network.

Storage and Backup: RCC hosts and maintains various storage systems, including persistent high-capacity storage that can be shared among a research group or remain private to each individual user, and high-performance storage for temporarily staging and quickly accessing data.

Persistent and High-Capacity Storage:

1) Home Directory: Each RCC user has a home directory for storing small, frequently used items such as source code, binaries, and scripts. The home directory is only accessible by its owner and is suitable for storing files that do not need to be shared with others. Data in the home directory can be accessed from Midway and remotely via different protocols.
2) Project Space: Principal investigators can request a project space for their research group. These directories are used for longer-term storage of data/files shared by members of a research group/project and are accessible from all RCC compute systems and remotely.

High-Performance Scratch Space: ScratchSpace: Hosted on RCC's high-performance storage system, scratch space is intended for staging data required/generated by computational processes running on the cluster. Unlike home and project directories, scratch space is neither snapshotted nor backed up and may be periodically purged. Users are responsible for ensuring any important data in the scratch space is replicated in a location providing persistent storage, such as project or home directories.

Backup and Data Recovery:

1) Filesystem Snapshots: Automated snapshots of home, project, and cds directories are available for recovering data in case of accidental file deletion or other problems. Typically, 7 daily and 4 weekly snapshots are available. However, RCC may reduce the number of snapshots during periods of high space usage.
2) Tape Backup: Nightly backups to a tape machine located in a different data center safeguard against hardware failure or disasters. These backups are intended for disaster recovery only, and users should rely on filesystem snapshots for regular data recovery. Users should also avoid using special characters in filenames, as they are not supported by the backup system.

Software: RCC installs, configures, and maintains hundreds of software packages on the Midway cluster. Some of the key software packages include:

1) Compilers: GNU and Intel C/C++/Fortran compiler suites, Nvidia's CUDA compiler for GPU computing, and compilers for Java, Julia, Go, and Haskell.
2) Interactive programming environments: Open-source and commercial programming environments like Python, MATLAB, Mathematica, Stata, and R. These environments often include pre-installed libraries and packages.
3) Data processing tools: Programs for dealing with large-scale data formats (HDF5 and NetCDF), data-movement programs (Globus), and database software (PostgreSQL and Hadoop).
4) Numerical libraries: Intel's Math Kernel Library (MKL) and OpenBLAS, the GNU Scientific Library (GSL), and FFTW (a Fourier transform library).
5) Community codes: Commonly-used scientific software such as LAMMPS, Gromacs, YT, lfrit, QIIME, and genetic analysis programs like SAMtools, Bowtie, BLAST, GATK, PLINK, and TopHat.

RCC can build and install open-source software upon request and help with negotiating licensing agreements, purchasing commercial software, or migrating purchased commercial software to RCC systems.

Networking: RCC's Midway supercomputing cluster is connected to the University of Chicago network backbone through a 10 Gbps network uplink. The university network connects to Internet2 at 10 Gbps and has 10 Gbps connections to other commercial networks. All Midway file transfer nodes and login nodes uplink through the Midway switch at 20 Gbps to the University of Chicago campus network backbone.

The research networks at the University of Chicago are deployed at two campus core distribution points, connecting via two 10-Gigabit Ethernet circuits to MREN and the CIC OmniPop. The connectivity provides the university with flexibility and capacity to connect to other institutions and share research data and resources.

The University of Chicago has built a Network infrastructure to establish a Science DMZ, which is distinct from the general-purpose campus network and purpose-built for data-intensive science. The Science DMZ includes support for virtual circuits, software-defined networking, and 100 gigabit Ethernet. RCC compute resources are now connected with a 40 Gbps Ethernet connection to the UChicago Science DMZ, and tests are being performed to ensure proper network traffic segregation.

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.

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 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.

University of Iowa 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 expres-

sion.

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

Abstract: A method includes receiving a first plurality of aligned genomic sequences of a virus from a database. The aligned genomic sequences have a first common background. The method includes calculating a Qnet for each genomic sequence of the first plurality of aligned genomic sequences. The Qnet for each sequence is calculated by calculating a conditional inference tree for each index of the aligned genomic sequences using other indices in the aligned genomic sequences as predictive features, and calculating predictors for indices that were used as predictive features when calculating the conditional inference tree for each index.

Relevant 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. DOI: https://doi.org/10.1371/journal.pcbi.1009363

☐ Dhanoa, J., Manicassamy, B. and Chattopadhyay, I., 2018. "Algorithmic Bio-surveillance For Precise Spatio-temporal Prediction of Zoonotic Emergence." arXiv preprint arXiv:1801.07807. Preprint DOI: https://arxiv.org/abs/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. DOI: https://doi.org/10.7554/eLife.30756

☐ 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. Preprint DOI: https://doi.org/10.1101/2020.07.17.20156364

☐ 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). Preprint DOI: https://doi.org/10.21203/rs.3.rs-2336091/v1

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES

DEPARTMENT OF MEDICINE
OFFICE OF THE CHAIRMAN
5841 S. Maryland Avenue, MC 6092 Chicago, IL 60637-1470
Phone 773-702-1051; Fax 773-702-4427
Everett E. Vokes, M.D.
Chairman, Department of Medicine
John E. Ultmann Distinguished Service Professor
Physician in Chief, University of Chicago Medicine & Biological Sciences
evokes@medicine.bsd.uchicago.edu

March 31, 2023

Department of Defense
Peer Reviewed Medical Research Program
Discovery Award

Re: Ishanu Chattopadhyay, PhD

Dear Reviewers:

On behalf of the Department of Medicine and Section of Hospital Medicine at the University of Chicago, I am delighted to be able to provide this letter expressing our enthusiastic support for Dr. Ishanu Chattopadhyay and his application for a Department of Defense Discovery Award titled "BioNORAD: Fast Scalable Pandemic Risk Assessment of Influenza A Strains Circulating in Non-human Hosts." Our Department is wholly supportive of Dr. Chattopadhyay and his outstanding research program.

Dr. Chattopadhyay is an accomplished and productive member of our faculty who currently holds an appointment as an Assistant Professor of Medicine. His well established research program focuses on large-scale data analysis, machine learning, and automated model discovery with the goal of unraveling complex phenomena in biology, medicine, the epidemiology of complex diseases including screening and diagnosis, and clinical decision making. In the current application, Dr. Chattopadhyay seeks to leverage his expertise to pursue innovative work seeking to develop a platform powered by novel pattern discovery and recognition algorithms that would be able to automatically parse out and identify emergent evolutionary strains of Influenza A viruses in the wild to provide global surveillance for potential emerging pandemic threats. This work has clear significance given past examples of Influenza A pandemics and the recent global experience with COVID-19. I am confident that you will Dr. Chattopadhyay is ideally suited to lead the proposed work.

As Chairman of the Department of Medicine, you have my assurance that we will continue to provide Dr. Chattopadhyay with all of the necessary physical resources, including laboratory space and equipment as well as administrative support necessary to accomplish his research aims. Dr. Chattopadhyay directs the Zero Knowledge Discovery (ZeD) laboratory which is housed within the Department of Medicine and has access to a full range of high end computing resources necessary for the work proposed in this application. In addition, Dr. Chattopadhyay will have access to all of the outstanding shared scientific resources and core facilities available to investigators at The University of Chicago. This includes The University of Chicago Research Computing Center which provides additional high-end research computing resources including high-capacity storage and backup.

In summary, Dr. Chattopadhyay has been highly successful and productive throughout his career as a researcher. I am continually impressed by the quality and innovative nature of his work and am fully supportive of the current proposal. I hope you are as impressed as I am by the rigor and originality of his research, the significance of the questions he wants to ask, and the talent and ability he has already demonstrated in his career to date.

Sincerely,

Everett E. Vokes, M.D.
John E. Ultmann Distinguished Service Professor
Chairman, Department of Medicine
Physician in Chief, University of Chicago Medicine & Biological Sciences

UNIVERSITY of IOWA
CARVER COLLEGE OF MEDICINE
University of Iowa Health Care

Department of Microbiology and Immunology

Balaji Manicassamy, Ph.D.
Associate Professor
51 Newton Rd, 3-430 BSB
Iowa City, IA 52242
319-335-7590 Tel
319-335-9006 Fax
balaji-manicasssamy@uiowa.edu

April 21, 2023

Ishanu Chattopadhyay, PhD
Assistant Professor
Department of Medicine
Institute for Genomics and Systems Biology
University of Chicago,

RE: Collaboration on your CDMRP application - **BioNORAD: Fast Scalable Pandemic Risk Assessment of Influenza A Strains Circulating In Non-human Hosts**

Dear Ishanu,

I am excited to continue to collaborate with you on this CDMRP application for developing algorithms to detect emerging influenza A viruses. As you know well, my research is focused on studying the host pathways essential for influenza A viruses and developing novel host-directed therapeutics against influenza viruses. We have expertise in the influenza virus reverse genetics system and routinely use this system to generate recombinant viruses for our studies. As we have discussed, my lab will generate and evaluate the predicted emerging strains in tissue culture. My group has a dedicated USDA inspected enhanced BSL2 suites to perform studies with zoonotic influenza viruses. I look forward to our continuous and productive collaborations. I wish you the best of luck with your application.

Sincerely,

Balaji Manicassamy, Ph.D
Associate Professor
Department of Microbiology & Immunology

Intellectual and Material Property Plan

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

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 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. Intellectual Products Expected:

☐ Software implementations of Emergenet inference algorithm

☐ Generation of sequence data from the Emergenet models specying novel strains expected to arise in future

☐ Patents filed by UChicago

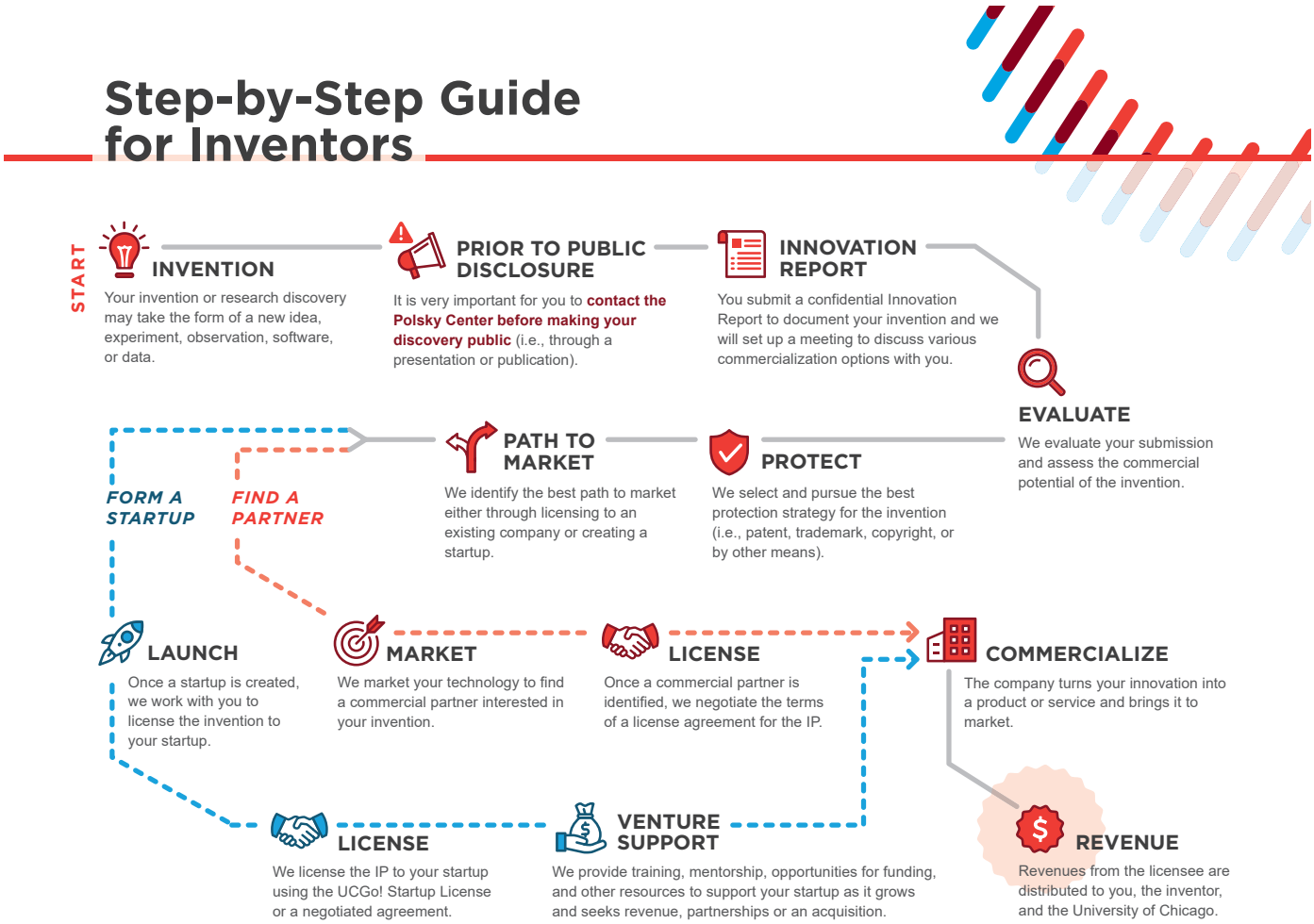

Fig. 5. Inventor pathway to commercialization at the University of Chicago

E. 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 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 5).

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

The BioNORAD project aims to develop a platform for early identification and mitigation of emerging pandemic threats, particularly those caused by novel strains of influenza viruses. The project will generate several types of data and research resources that can be valuable to the scientific community. In this section, we describe how we plan to share these data and research resources with the research community.

Types of Data and Research Resources: The BioNORAD project will generate several types of data and research resources, including:

- **Genomic data:** The project will generate predicted genomic sequences of key viral proteins of Influenza A.
- **Machine learning models:** The project will develop machine learning models for predicting the emergence and spread of novel strains of influenza viruses. These models will be based on the genomic data acquired during the project from public sequence databases.
- **Software tools:** The project will develop software tools for analyzing genomic and clinical data, as well as for visualizing and interpreting the results of machine learning models. These tools will be valuable for researchers working on influenza viruses and other emerging infectious diseases.

Sharing Plan: We plan to share the data and research resources generated by the BioNORAD project with the research community as widely as possible, while safeguarding the privacy of participants and protecting confidential and proprietary data and third-party intellectual property.

Data sharing: We plan to deposit the genomic data generated by the project in publicly accessible repositories with long-term DOI access. These repositories will provide a persistent and citable location for the data and ensure that they are discoverable and accessible to the research community. We will use the Zenodo platform to deposit the data, which will provide long-term archiving and DOI access. We will also ensure that the data are compliant with data sharing policies of funding agencies such as the National Institutes of Health and the Department of Defense. Genomic sequence data will be also deposited to NCBI.

Model sharing: We plan to deposit the machine learning models developed during the project in public repositories such as GitHub, with open-source licenses, making them freely available to the research community. We will provide documentation and instructions for running the models, as well as sample data for testing and validation. We will also encourage the research community to contribute to the development of the models by providing feedback and suggesting improvements.

Software sharing: We plan to deposit the software tools developed during the project in public repositories such as GitHub, with open-source licenses, making them freely available to the research community. We will provide documentation and instructions for using the tools, as well as sample data for testing and validation. We will also encourage the research community to contribute to the development of the tools by providing feedback and suggesting improvements.

Feasibility and Appropriateness: The data generated from this project will be shared through various channels to ensure maximum accessibility to the research community. In addition to depositing the software in a GitHub repository, models and data will be deposited in Zenodo, an open-access repository for scientific data, with a long-term DOI access. The datasets generated will be organized and labeled appropriately to enable easy retrieval, interpretation, and reuse by other researchers. The datasets will be accompanied by comprehensive documentation, including descriptions of the methods and procedures used to generate the data, as well as any relevant metadata.

To further facilitate the sharing of data and resources, the project team will follow best practices for data management, including adhering to FAIR (Findable, Accessible, Interoperable, Reusable) data principles. These practices include using standardized data formats and metadata, creating clear documentation for data and code, and ensuring that data and code are appropriately versioned and stored in secure and reliable repositories.

The project team recognizes that data sharing carries certain risks, particularly regarding the protection of sensitive information. However such risk is minimal in this project, which will not involve any clinical information or human participants.

In summary, the BIONORAD project is committed to promoting data and research resource sharing to enhance collaboration and accelerate the translation of research results into practical applications. The project team will ensure that data and resources generated during the project's period of performance are

made widely available while safeguarding the any proprietary information that we may use (including sequence data that do not carry permission to be posted publicly). By sharing and leveraging data and resources, this project will contribute to a more expeditious translation of research results into knowledge, products, and procedures to improve military health and global health security.

Implementation in a Dual-Use Capacity: The knowledge, information, products, technologies, or applications gained from the proposed research could be implemented in a dual-use capacity to benefit the civilian population that also addresses a need related to military health. The research findings will be published in peer-reviewed journals, presented at scientific conferences, and disseminated to the research community. The findings will also be communicated to the wider community through various media, such as press releases, newsletters, and social media platforms. The knowledge generated by this research will have broad applications beyond the military, with potential benefits for the civilian population, including the development of vaccines, diagnostic tools, and treatments for infectious diseases.

In summary, we will make every effort to ensure that all data and research resources generated during the project's period of performance will be made publicly available while safeguarding the privacy of participants and protecting confidential and proprietary data and third-party intellectual property. We recognize that the unique data generated by our project will be valuable to the research community, and we will make every effort to share this data with the research community. The research resources generated during the project will also be made publicly available. The knowledge, information, products, technologies, or applications gained from the research could be implemented in a dual-use capacity to benefit the civilian population that also addresses a need related to military health.

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] Chattopadhyay, I., Wu, K., Li, J. & et al. Emergenet: Fast scalable pandemic risk assessment of influenza a strains circulating in non-human hosts (2022). URL https://doi.org/10.21203/rs.3.rs-2336091/v1. 2336091.

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

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

[20] Agor, J. K. & Özaltın, O. Y. Models for predicting the evolution of influenza to inform vaccine strain selection. *Human vaccines & immunotherapeutics* **14**, 678–683 (2018).

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

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

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

[24] 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).

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

[26] 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).

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

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

[29] 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).

[30] 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).

[31] 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).

[32] 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).

[33] 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).

[34] 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).

[35] 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).

[36] 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).

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

[38] 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).

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

[40] 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).

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

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

[43] 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).

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

[45] 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).

[46] 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).

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

[48] 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).

[49] for Disease Control, C. & Prevention. Seasonal influenza (2021). URL https://www.cdc.gov/flu/about/index.html.

[50] of Defense, U. D. Report on effects of a changing climate to the department of defense (2020). URL https://media.defense.gov/2020/Jul/21/2002462222/-1/-1/1/Report_on_Effects_of_a_Changing_Climate_to_the_Department_of_Defense.PDF.

[51] Wells, K. B. & Lewis, M. J. Influenza: Evolution, detection, and response. *Clinical Microbiology Reviews* **18**, 80–103 (2005).