**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]. Strain scoress based multiple experimentakl assays, the number of human infections if any, transmission in laboratory animals, receptor binding, population immunity, genomic analysis, antigenic relatedness, global prevalence, pathogenesis, and treatment options, are averaged to estimate emergence IRAT score, a number between 1 and 10. IRAT scores take weeks/months to compile for a single strain. With tens of thousands of strains being collected annually, this results in a scalability bottleneck.

Here we plan to develop a platform powered by pattern discovery and a novel generative AI framework to analyze all Influenza A strains currently in public repositories, to parse emergent evolutionary constraints operating in the wild. 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 any specific animal strain is close to achieving human adaptability and human-to-human transmission capability.*

Our broader goal to develop a general framework for foundational discovery in pathogen emergence beyond the case of Influenza A.

**Specific Aims:** **Aim 1: Develop a predictive framework for viral evolution and emergence risk assessment. (Lead: I. Chattopadhyay (UK))** We aim to quantify the probability of spontaneous viral mutation and cross-species transmission. By analyzing large-scale influenza genome datasets, the proposed Emergenet will infer cross-mutation dependencies and evolutionary constraints, enabling real-time forecasting of high-risk strains. The goal is to replace subjective expert-driven assessments with a probabilistic model capable of ranking strains by emergence potential, and identify a small actionable set of strains at teh edge of emergence.

**Aim 2: Experimentally validate the predictive accuracy of Emergenet. (Lead: S. Chattopadhyay (UK) and Manicassamy (UIowa))** This aim involves generating viral variants predicted by Emergenet through reverse genetics and assessing their fitness in human lung epithelial cells. The replication efficiency, antigenic properties, and transmission potential of these strains will be tested to determine if Emergenet's forecasts align with biological viability. Competitive fitness assays will evaluate whether mutations predicted by Emergenet enhance viral survival, distinguishing them from randomly introduced mutations that result in loss of function.

**Aim 3: Deploy BioNORAD for real-time pandemic risk surveillance. (Lead: I. Chattopadhyay (UK))** This aim integrates Emergenet's predictive capabilities into an automated global biosurveillance platform. BioNORAD will continuously score and rank emerging influenza strains in real-time using surveillance data from NCBI and GISAID. The system will dynamically update risk profiles, offering public health agencies an early warning system for emerging threats. By comparing BioNORAD's predictive rankings with actual outbreak data, the project will assess its ability to preemptively identify pandemic-potential strains.

**Aim 2 objectives do not constitute gain-of-function research.** The introduced mutations are limited to those observed in naturally circulating strains, ensuring no artificial enhancement of transmissibility or pathogenicity. All experimental work will comply with stringent biosafety regulations under BSL-2/BSL-3 conditions, with oversight from institutional biosafety committees. The primary objective is to validate Emergenet's ability to predict naturally occurring mutations, not to engineer novel viral properties. The focus is on understanding evolutionary trajectories rather than enhancing viral capabilities, aligning fully with ethical and biosafety standards.

**Foundational Questions Investigated:** What drives a viral strain to become pandemic after spill-over events, which tend to be common? Can these events be characterized and preempted? Is sequence-level

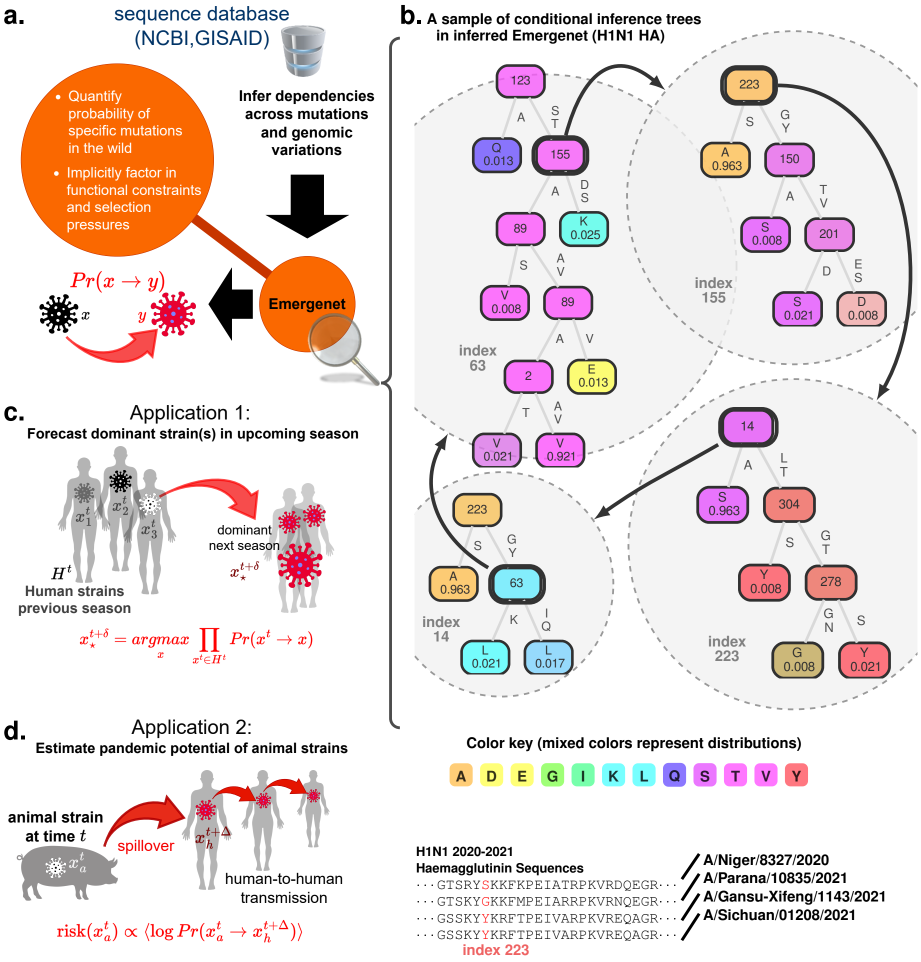
a.
sequence database
(NCBI,GISAID)
Quantify probability of specific mutations in the wild
Implicitly factor in functional constraints and selection pressures
Infer dependencies across mutations and genomic variations
Emergenet
$Pr(x \to y)$
$x$
$y$
b. A sample of conditional inference trees in inferred Emergenet (H1N1 HA)
123
A
S T
Q 0.013
155
A
D S
89
K 0.025
S
A V
V 0.008
89
A
V
2
E 0.013
T
A V
V 0.021
V 0.921
index 63
223
S
G Y
A 0.963
150
A
T V
S 0.008
201
D
E S
S 0.021
D 0.008
index 155
14
A
L T
S 0.963
304
S
G T
Y 0.008
278
G N
S
G 0.008
Y 0.021
index 223
223
S
G Y
A 0.963
63
K
I Q
L 0.021
L 0.017
index 14
Color key (mixed colors represent distributions)
A D E G I K L Q S T V Y
H1N1 2020-2021
Haemagglutinin Sequences
···GTSRYSKKFKPEIATRPKVRDQEGR···
···GTSKYGKKFMPEIARRPKVRNQEGR···
···GSSKYYKRFTPEIVARPKVREQAGR···
···GSSKYYKRFTPEIVARPKVREQAGR···
index 223
A/Niger/8327/2020
A/Parana/10835/2021
A/Gansu-Xifeng/1143/2021
A/Sichuan/01208/2021
c. Application 1:
Forecast dominant strain(s) in upcoming season
$x_1^t$
$x_2^t$
$x_3^t$
$H^t$
Human strains previous season
dominant next season
$x_\star^{t+\delta}$
$x_\star^{t+\delta} = argmax_x \prod_{x^t \in H^t} Pr(x^t \to x)$
d. Application 2:
Estimate pandemic potential of animal strains
animal strain at time $t$
$x_a^t$
spillover
$x_h^{t+\Delta}$
human-to-human transmission
$\text{risk}(x_a^t) \propto \langle \log Pr(x_a^t \to x_h^{t+\Delta}) \rangle$

Fig. 1: **Conceptual Scheme: Emergenet inference**. **Panel a** Variations of genomes for identical subtypes of Influenza A are analyzed to infer a recursive forest of conditional inference trees[6] – the Emergenet– which maximally captures the emergent dependencies between an a priori unspecified number of mutations. With these inferred dependencies we can estimate the numerical odds of specific mutations, and by extension, the numerical value of the probability of one strain giving rise to another in the wild, under complex selection pressures from the background. **Panel b** Snapshot of decision trees from the Emergenet inferred for H1N1 HA sequences collected in 2020-2021, which reveals a cyclic dependency. In general, every internal node of a component tree can be "expanded" into its own tree, underscoring the recursive structure of the Emergenet. **Panel c** First application: forecast dominant strain(s) for the next flu season, using only sequences collected up to six months prior and the inferred Emergenet, using data from the past year. **Panel d** Second application: estimation of the pandemic risk posed by individual animal strains that are still not known to circulate in humans.

information sufficient for such predictions? Do historical sequence variations reveal discernible patterns that enable forecasting of future mutations within current circulating populations?

**Feasibility & Preliminary Results:** The Emergenet framework has already demonstrated its predictive capability in forecasting the evolutionary trajectories of influenza A strains. By constructing a digital twin of sequence evolution, Emergenet quantitatively assesses the emergence potential of animal strains, providing a scalable alternative to labor-intensive expert assessments. Using a dataset of 220,151 hemagglutinin (HA) sequences, Emergenet outperforms WHO's seasonal vaccine recommendations for H1N1 and H3N2 subtypes over a 20-year period, improving antigenic match predictions by an average of 3.73 amino acids (28.40%). Furthermore, Emergenet's generative models correlate strongly with CDC's expert-assessed Influenza Risk Assessment Tool (IRAT) scores (Pearson's r = 0.721, p = $10^{-4}$) while achieving a computational speedup of at least five orders of magnitude, reducing assessment time from months to seconds[7].

To experimentally validate Emergenet, high-risk HA variants will be generated using the reverse genetics system developed by the BM at U of Iowa. HA segments with predicted mutations will be synthesized and assessed for cell surface expression via flow cytometry and western blotting. Recombinant viruses carrying mutant HA will be generated and validated using next-generation sequencing (NGS). Replication fitness of these recombinant viruses will be evaluated through single-cycle and multicycle replication assays in human lung epithelial cells (A549) and primary human lung cells. Additionally, competition assays between predicted high-fitness mutants and parental strains will be performed in a 1:1 ratio, with fitness outcomes determined via high-resolution melting (HRM) analysis. These experiments will assess the biological relevance of Emergenet's forecasts, determining whether the predicted mutations enhance viral fitness in human-relevant models.

To further extend the validation framework, SC at UK will conduct in vivo assessments using a well-characterized animal models to determine the pathogenicity and transmissibility of high-risk recombinant influenza variants. Disease severity will be evaluated through body weight loss, lung viral titers, histopathological analysis, and cytokine profiling, and transmission studies will assess whether predicted high-risk strains demonstrate increased transmissibility compared to parental strains in cohoused naive animals. These studies will provide physiological and immunological validation of Emergenet's predictions, ensuring that computationally inferred high-risk variants correspond to enhanced virulence and transmissibility in a mammalian host.

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

[7] Wu, K. Y., Li, J., Esser-Kahn, A. & Chattopadhyay, I. Emergenet: A digital twin of sequence evolution for scalable emergence risk assessment of animal influenza a strains. *arXiv preprint arXiv:2411.17154* (2024).