Broad Fellows Application

Research Statement

Long-term research goals: My long-term research interests are to build the necessary experimental and computational components required to make real-time pathogen surveillance, prevention and treatment a *rationally predictive endeavour*. The core problems I have identified are: (1) the paucity of biochemical phenotype data that inform mechanistic knowledge of risk and pathogenesis, and (2) interpretable models to map from sequence data to quantitative measures of pathogen risk.

Research background: My research background, which has included both experimental and computational components, has provided me with an excellent set of tools with which to tackle this problem. My experimental training was in synthetic biology, where I was a member of the 2009 UBC iGEM team, and an advisor to the 2011 UCSF iGEM teams. I later switched into computational science under Prof. Jonathan A. Runstadler. My first area of focus has been on influenza disease ecology. Together with my colleagues in the Runstadler lab, we have investigated the role of reticulate evolution in influenza virus host switching (#cite: PNAS, EcoLetters), where I have won poster presentation awards (Broad Retreat, CEHS). The second (and more recent) focus is on the prediction of viral phenotype from genotype. The latter is where my current efforts are focused, in collaboration with the Harvard Intelligent & Probabilistic Systems group. Apart from these two main areas of focus, I have also collaborated with colleagues in the use of Bayesian phylogenetic methods to study influenza movement and reassortment in wild animals (#cite: PLoS Pathogens, Virus Genes), and developing analysis methods for viral phenotype data (#cite: SciRep, IGE papers). In collaboration with experimentally-oriented colleagues, we are currently building the experimental infrastructure to systematically characterize the biochemical activities of viral protein variants.

Broad Fellows vision: As a Broad Fellow, I envision assembling a team to build a real-time risk profiling dashboard for influenza. Using real-time sequencing data as an input, I envision this dashboard as being powered by machine learning, backed by experimental data, and delivering rapid and interpretable insights into the biology, pathology and epidemiology of emerging infections. With influenza as the proof of concept, the longer-term vision is for this to be modularity extensible to other pathogens as well. In pursuit of this goal, I plan to develop 2 main project thrusts, which I will elaborate on below.

Project Thrust 1: Develop epidemiologically relevant, scalable and rapid methods for experimentally phenotyping viruses.

Core problems: The goals of real-time pathogen surveillance are to map evolutionary trajectory and deduce its risk profile of a virus from its sequence. Doing so can help rationally guide medical and epidemiological decision-making in the event of new outbreaks. The sequencing technologies necessary to power this are being developed (#cite: MinION papers). The crucial missing link here is phenotype data that are relevant to epidemiology, and rapidly and scalably measurable. In the first project thrust, I aim to work with my team to develop technologies to rapidly phenotype influenza viruses polymerase (replication rate), hemagglutinin (viral entry and immune evasion), and neuraminidase (viral release and drug resistance). The core data problems I aim to solve with my team are:

1. The lack of systematically measured phenotype data paired with protein variant sequences.
2. The lack of a catalogue of safe, scalable, and standardized phenotyping assays to rapidly phenotype these proteins.
3. The lack of the concept of "signatures" of riskiness.

Proteins are the arbiters of viral function. Hence, I believe that a pathogen's measure of "riskiness" is composed of its proteins' biochemical activities. In this paradigm, the rational approach requires making systematic and properly standardized measurements, on epidemiologically relevant biochemcial phenotypes, for libraries of protein variants.

Proposed work: We will begin with phenotyping assays that are already amenable to systematic testing but have not yet been executed at scale. One low-hanging fruit is the polymerase minigenome assay, in which the influenza polymerase activity is read out using a luciferase reporter, and is amenable to high throughput liquid handling. In order to generate protein variant libraries, we will use a two-pronged approach. To learn from historical data, we will create a rational library of existing protein variants in the Influenza Research Database. To pre-emptively explore genotypic space, we will generate random mutants from contemporary protein variants that have been sampled in the past year.

As a medium term bet, and concurrent with ongoing systematic testing of the influenza polymerase, we will also explore the extension of this phenotyping system to other viral polymerases. In doing so, we aim to develop a modular, plug-and-play phenotyping system for rapidly phenotyping emerging viral outbreaks as they occur. In the long run, the goal is to develop and scale assays for other phenotypes, such as drug resistance, binding to cellular receptors, and *in vitro* antigenicity measurements using virus-like particles. In other words, multiple phenotypes across multiple viruses.

The data that we will generate as a team will have advantages that stand in contrast to the current available data. Firstly, it will be data relevant to understanding the mechanistic underpinnings of influenza risk and pathogenesis. This stands in contrast to more easily collectable proxies, such as the number of influenza-like illnesses (ILI) per year and viral load in patient cohorts, both of which are far removed from pathogenesis mechanisms. Understanding the biochemical underpinnings of pathogenesis opens opportunities for the development of drug treatments. Secondly, the data will be standardized and done systematically and at scale. This standardization has advantages over existing biochemical phenotype measurements, such as antigenic distance to vaccine strains, hemagglutinin receptor specificity and neuraminidase drug resistance, which have only been measured ad-hoc in response to new infections with non-standardized, merely conveniently available controls, and hence have no global standard of comparison for risk.

Potential Collaborations: I anticipate collaborating with Pardis Sabeti's group to explore the development of other molecular assays for emerging viruses. I also anticipate collaborating with Paul Blainey's group to leverage their microfluidics expertise in miniaturizing and automating the biochemical assays.

Short-term milestones: In the spirit and interest of open science, the protocols and data generated will be released freely through the Broad Institute, available for the research community through an in-house web-based interface, as well as through open access publications.

Project Thrust 2: Develop and deploy machine learning models that predict quantitative biochemical phenotype from sequence.

Machine learning offers us the capacity to learn the complex relationship between genotype and phenotype. Our second project thrust will take advantage of recent developments in deep learning to bring interpretability and learning capacity to machine learning on protein phenotypes.

Core problem: Over the past decade, machine learning has been applied to protein sequences (#cite) and structures (#cite) to predict properties such as drug resistance, but the state-of-the-art models suffer from the trade-off between learning capacity (model complexity) and interpretability, the latter being the biggest limiting factor in deployability in clinical settings. An additional limitation is that they are unable to accept variable-length sequences as input, which poses a problem for fast-evolving viral proteins that can undergo insertions and deletions, apart from substitutions, as part of their evolutionary trajectory.

Recent progress in deep learning has led to the development of convolutional deep networks that operate on chemical graphs to predict chemical properties of small molecules (#cite: David Duvenaud's work). In such graphs, nodes are atoms and edges are bonds. By converting the each variable length graph into a fixed length fingerprint, input sequences of variable length (e.g. chemical structures of varying numbers of atoms) can be used as inputs to supervised learning algorithms. By inspecting the maximally activated nodes and edges, the structural features most predictive of chemical properties (e.g. hydroxyl groups for solubility, sulfonyl groups for toxicity) can also interpretably visualized.

Proposed work: Protein structures are a natural extension of chemical graphs, where nodes are amino acids and edges are biochemical interactions between them. In collaboration with David Duvenaud of the Harvard Intelligent and Probabilistic Systems group, I am currently developing software that converts protein structures into a graph representation and a software package that enables general purpose deep learning on graphs. While both are currently unpublished work, I am actively working on these two pieces of software, with code publicly available on GitHub (#link), with considerable performance progress being made.

We will begin with maturing the graph deep learning software for use with our influenza replication phenotyping data; preliminary work on predicting HIV-1 protease drug resistance has shown great promise. The goal here is to identify the deep convolutional network architectures that can most accurately predict quantitative replication phenotype from modelled structure. We will leverage the computational resources that we have access to, including the Broad and BioMicroCenter (MIT) compute clusters, as well as on-demand commercially-available cloud compute capacity (where necessary).

A medium-term goal will be to pair the computational efforts with the diverse experimental data generated to train models for each protein's set of measures phenotypes. For single proteins/complexes with multiple measured phenotypes, we will experiment with multi-task learning, in which we simultaneously learn multiple measured phenotypes for the same sequence.

By training regression models on multiple viral phenotypes, we will gain the capacity to quantitatively map the risk profile of newly emerged viruses. As a long-term goal, I envision that these models, which are trained on mechanistically relevant data, can form the foundation of hierarchical models of pathogen biology and risk.

Beyond the application to viral phenotype prediction, machine learning on graph-structured data is a very new field of research, with only a handful of pre-print manuscripts available this year (#cite arXiv papers). Because of the novelty of this

field, we expect to make substantial contributions in the application of deep learning to graph-structured data in general.

Potential collaborations: I anticipate continuing to collaborate with David Duvenaud as he begins his tenure as a new faculty member at the University of Toronto. Additionally, this project thrust will open up new computational avenues for Broadies to leverage, for example, by training supervised machine learning tasks on metabolic and genomic interaction networks.

Short-term milestones: We will release the protein interaction network and graph fingerprinting software alongside manuscripts in open access publication avenues.

Forecasted Impact

The data and models that we develop will be a rich resource for the influenza research community, and as a Broad Fellow, I would welcome data reuse and sharing. Beyond their usefulness for epidemiological purposes (i.e. identifying signatures of a "risky" virus), I envision the data becoming useful in a variety of other settings.

One example is in drug repurposing. With the proposed scalable assays developed, I foresee collaborating with other groups at the Broad who are interested in drug repurposing efforts. For example, with our proposed phenotype collection, we may be able to identify existing non-toxic molecules that target multiple components of flu simultaneously, reducing the likelihood of drug resistance by opening avenues for combination therapy.

Another example is in drug development. In order to pre-emptively identify viral proteins that exhibit resistance to newly developed drugs, we can create new synthetic protein variants using contemporary strains as the starting sequence. Medicinal chemists may be able to use this data to pre-emptively test new versions of their drugs and validate their effects against the phenotype catalogue.

The final example is in furthering our basic understanding of pathogen evolution. This is a low-hanging fruit which I hope to pursue as soon as we have the data available. By using our trained deep learning models, we may (re-)examine historical trends of neuraminidase drug resistance or polymerase activity over time. We may also combine our predictions with Bayesian phylogenetic modelling to better understand how public health interventions affect the evolutionary trajectory of viral pathogens w.r.t. their epidemiologically relevant phenotypes.

Our ultimate goal is to make surveillance a holistic and rationally predictive endeavour. I believe that this work will positively impact pathogen genomic surveillance efforts by developing the necessary workflow, data and models for rational prediction of risk. I also foresee downstream research in drug development and pathogen surveillance through the public release of systematically measured pathogen phenotype data (i.e. "The Broad Phenotype Collection").

Planned Funding Avenues

In order to sustain this work beyond the Broad Fellows period, I will solicit funding from a variety of government and philantrophic sources. Apart from the NIH R21 proposal that I am co-writing with my advisor Jon and collaborator David, I foresee this being of interest to the DARPA Prophecy program, NIAID, and companies interested in drug development. Finally, to acquire a continued revenue stream for the research and development work, our team will explore the use of funding models for application programming interfaces (APIs) that allow access to value-added data and models, which may be of interest to other academic and commercial entities (#cite: database funding models).