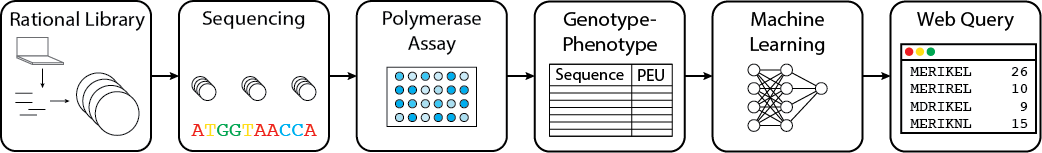
**Introduction:** One goal of viral genomic surveillance is to use a virus’ sequence to predict how it may cause . Unfortunately, the biggest gap hampering this vision is a lack of viral protein sequence variants paired with systematically measured phenotype data, on which computational models can be developed or trained to learn the sequence-function mapping. As a Rowland Junior Fellow, I aim to (1) develop the necessary experimental infrastructure to rapidly determine viral phenotype from its sequence, and (2) integrate the data generation pipeline with new advances in machine learning to enable accurate and rapid prediction of viral risk from sequence. My long-term goal is to predict *in silico* viral risk from its component proteins’ sequences, without requiring experimentation on whole live viruses (Fig. 1).



**Fig. 1:** Data generation and modelling pipeline from protein sequence to biochemical activity.

**(1) Viral phenotyping:** Health risk due to viral infection is partially composed of a virus’ component biochemical properties, which are determined by its proteins. In order to address the key missing link of systematically characterized viral phenotype data, I will lead my group to systematically and rapidly characterize influenza virus protein variants in a safe (i.e. not involving live viruses) and scalable manner. To do so, we will test individual component protein mutants for their *in vitro* biochemical properties. One example is the influenza polymerase assay, based on a well-established and modular genetic system, that uses luciferase expression as a readout. As a Rowland Jr. Fellow, I will work with my team to develop the experimental infrastructure to scale these assays in a high-throughput fashion. We will use synthetic genome assembly, high throughput sequencing methods, and robotics to scale our library generation and measurement steps for well-established phenotype assays. We will begin with the influenza polymerase for replication rate and neuraminidase for drug resistance. My medium-term goals are to develop new assays for systematically measuring other aspects of influenza’s risk, and leverage implemented genetic systems to rapidly test other emerging virals. My long-term goal is to have a plug-and-play, modular genetic system for testing any new viral proteins and its variants within days of sequencing.

**(2) Computational prediction:** Predicting viral protein activity from sequence can be cast as learning a non-linear mapping from genotype space to phenotype space. The viral phenotyping data generated in (1) constitute a gold standard, densely measured dataset for precisely this task. We will use supervised learning algorithms to learn the mapping from genotype space to phenotype space, while also partnering with current collaborators to develop new learning algorithms for this task. With our experimental pipeline, data, and models, I aim to build an integrated “viral forecasting” system that maps viral risk trajectories, predict future viral properties, and test/validate these predictions.

**Funding Avenues:** In pursuit of these goals, I have written two Broad*Next10* grant applications, one of which was co-written with colleagues at the Broad Institute, to develop such standardized, safe and scalable assays for the influenza polymerase and neuraminidase. Both our grants were awarded, totalling $80,000 in funds. Additionally, my current advisor and I are collaborating with the Harvard Intelligent and Probabilistic Systems group on an NIH R21 grant to fund these efforts further. Other planned funding sources include DARPA’s Prophecy program, philanthropic groups (Gates and Simons foundations), and the NIH/NIAID. We will also explore data access/licensing models with interested industry partners to enable our research and engineering efforts to be self-sustaining.