**Computational detection and characterization of epistatic interactions in influenza A haemagglutinin protein through surveillance sequencing and deep mutational scanning data**

Magdalene Walters1, David McCandlish2

1Undergraduate Research Program

2Cold Spring Harbor Laboratory

Human influenza virus is a widely circulating pathogen, with pandemic strains potentially infecting 20-40% of the global population in a single year. The surface proteins haemagglutinin (HA) and neuraminidase (NA) determine antigenic properties of the virus and are under acute positive selection for antigenically novel immune escape variants, causing rapid evolution. HA in particular typically accumulates 3-4 amino acid substitutions per year. Importantly, these substitutions have shown patterns of genetic interaction, requiring at least one permissive mutation for certain substitutions to exhibit adaptive effects. This project aims to detect and characterize these genetic interactions through a statistical framework informed by high-throughput mutagenesis data and HA surveillance sequence data. Surveillance sequence data is available through the Global Initiative for Sharing All Influenza Data (GISAID) EpiFlu database and effectively highlights temporal patterns in the genetic variation. Temporal heterogeneity in EpiFlu’s sampling density will be assessed by changes in synonymous variant diversity while the surveillance database’s error rate will be informed by high-throughput mutagenesis data. Most importantly, the large volume of sequence data yields a dense record of tolerable amino acid changes— this allows permissive substitutions to be identified by concomitant changes in the spectrum of observed variants.