INNOVATING CROSS-DOMAIN SOLUTIONS TO DETECT EMERGING BIOLOGICAL THREATS

Learning The Language Of Proteins And Predicting The Impact Of Mutations

Bin Hu Los Alamos National Laboratory **Kaelyn Gibson** Los Alamos National Laboratory **Po-E Li** Los Alamos National Laboratory **Valerie Li** Los Alamos National Laboratory **Patrick Chain** Los Alamos National Laboratory

Mutations in proteins directly impact their structure and function. Understanding the “language” of proteins, or the sequence to function (genotype-phenotype) relationship has many real-world applications. One set of applications includes those in biodefense, such as biological threat detection and biosurveillance, antibody engineering, and medical countermeasure development. In this study, we present a novel language model-based approach that can rapidly analyze vast collections of sequences, and make near real-time functional predictions that compare favorably to those made using conventional bioinformatic and experimental methods. Our findings reveal that tailored protein language models can predict protein mutation phenotypes, such as binding affinity or level of expression, when they are trained with high-throughput functional data. Protein language models applied to viral genomes can also discern the lineage within a family (e.g., sarbecovirus sequences). Coupled with sequenced-based biosurveillance, this type of model may provide early warning signals of potential zoonotic spillovers (i.e. host jumping) or “escape” from existing medical countermeasures posed by novel mutations. This research not only underscores the potential of ML and language models in addressing pressing challenges in understanding the mapping of sequence to function, but further elucidates their potential application in accelerating the response to biological threats as they evolve.

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. The authors appreciate funding from CDC, Los Alamos National Laboratory (LDRD Director Initiated Research, 20240734DI, 20210767DI, 20200732ER). Part of this work was funded by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act. This project has been funded in part with Federal funds under Interagency Agreement No. 22FED2200087IPD (RRJJ) between the Centers for Disease Control and Prevention, Influenza Division and the US Department of Energy, Los Alamos National Laboratory. The GPU cluster was purchased by the bioassurance funding through DOD. This work is approved for public release under LA-UR-24-21857.

doi: https://doi.org/10.1101/2024.04.05.588133