## INNOVATING CROSS-DOMAIN SOLUTIONS TO DETECT EMERGING BIOLOGICAL THREATS

# AI-enabled Property Prediction For Characterization Of Novel Proteins

**Phillip Tomezsko** MIT Lincoln Laboratory **Ta-Hsuan Ong** MIT Lincoln Laboratory **Rafael Jaimes** MIT Lincoln Laboratory
**Josh Dettman** MIT Lincoln Laboratory **Avery Meyer** MIT Lincoln Laboratory

Rapid characterization of novel biological agents is critical to the nation's security posture. Recent advances in prediction of protein structure and associated function provide an opportunity to computationally quickly and securely characterize novel proteins. A technological revolution was kicked off in the field of structural biology by the performance of AlphaFold2 on protein folding in 2020. Further advances have come in prediction of protein function, prediction of protein complex formation and generation of novel protein sequences. This generation of AI tools for structural biology provide an opportunity to characterize a protein by sequence along, accelerating capability to rapidly understand novel proteins and reducing need for experiments.

The MITLL team has developed a flexible computational pipeline to characterize a variety of protein functions, as well as model protein complexes for custom analyses. In addition to flexibility, the pipeline also contains modules to analyze specific protein-classes in depth. The pipeline takes advantage of best in class, open-source AI tools such as AlphaFold2 for protein folding and Evolutionary Scale Modeling (ESM) as a protein language model. The input are protein sequences that could come from mass spectrometry or from nucleic acid sequencing.

To test the capabilities of the computational pipeline, we developed a module to predict enzymatic activity given an amino acid sequence and a module to classify conotoxins by pharmacological class. Using proteins detected from an internally derived mass spectrometry dataset, we identified 88% enzymes correctly by fusing the outputs of open-source algorithms ProteInfer and CLEAN. For conotoxins, we developed a classifier based on ESM and structural predictions from AlphaFold2. We tested this pipeline on an open-source transcriptomics dataset from cone snails. We achieved >80% accuracy on the annotated sequences. Additionally, we predicted the pharmacological class of other conotoxins from the dataset that were previously unannotated and created docking models using AlphaFold2 to confirm a realistic mode of binding.

The protein property prediction pipeline has been containerized and transitioned to a DoD sponsor to bolster their analytic capabilities for characterization of novel proteins. Future work will include small molecule and other biomolecules into the computational pipeline to include for analysis. Additional AI developments will be evaluated for inclusion in future iterations of the protein property prediction pipeline.