**DPA26BZ03-DV014**
**Real-Time Pathogen-Host Interactome Prediction**
**Frequently Asked Questions (FAQs)**

1. Training data source: Is the use of C. elegans host-pathogen genetic screen data as the primary training dataset for a GNN predictor of pathogen-host protein-protein interactions consistent with the scope of Topic DPA26BZ03-DV014? The topic specifies prediction from pathogen protein sequence alone — our approach uses pathogen protein sequence as the sole input at inference time, but uses C. elegans experimentally characterized interactions as training supervision. Is this distinction acceptable?
   **A: The inference approach is compliant, but the model must be able to successfully predict interactions for human hosts, and not just for specific pathways.**

2. Wet-lab validation requirement: Does the Phase I scope require experimental wet-lab validation (e.g., infection assays, co-immunoprecipitation, yeast two-hybrid) of the computational predictions, or is computational validation (cross-pathogen leave-one-out accuracy benchmarking; comparison to published interaction databases such as STRING, BioGRID, and IntAct) sufficient for the Phase I deliverable? We can include C. elegans infection lifespan assays as a validation arm if experimental validation is required, but wish to confirm whether this is necessary for Phase I.
   **A: This solicitation is a Direct to Phase 2 topic, while computational validation data is sufficient to show Phase 1 requirements, please see above for applicability to human hosts.**

3. Scope of "host-pathogen interactions": Does the topic scope encompass prediction of host immune pathway activation/suppression profiles (e.g., p38 MAPK activation, NF-κB suppression) in addition to direct protein-protein binding interactions? Our platform predicts both: (a) direct pathogen protein – human host protein binding interactions, and (b) the downstream immune pathway consequences of those interactions. Is the second layer of prediction (pathway-level) within scope, or should the Phase I deliverable focus exclusively on direct protein-protein interaction prediction?
   **A: This solicitation is a Direct to Phase 2 topic, there are no Phase 1 deliverables, However, these predictions are in scope, but immune pathway interactions only are insufficient as an outcome of this model.**

4. Does the term "protein sequence data" mean the primary sequence structure, i.e. the order of amino acids? Is this definition more broad? If so, please provide clarification**.**
   **A: This may refer to any type of protein data, but would need to be data that can be related to the structure and function of the protein and will provide the order of amino acids. Protein sequences may be inferred from nucleic acid sequences.**

5. The "protein sequence data" can come in different forms based on acquisition technology and can be partial or complete. Does the solicitation assume that all "protein sequence data" will be partial? Complete sequence data would imply known, well characterized proteins.
   **A: It may refer to either type of data**

6. Implicit to this proposal is acquisition of the "protein sequence data." Does this proposal seek methods and validation to acquire this "protein sequence data" such as the use of Mass Spectrometry and Edman degradation?
   **A: Methods for protein sequence acquisition are outside the scope of this solicitation.**

7. If the proposal does not require "protein sequence data" acquisition, is a data set being provided for the purpose of this solicitation?
   **A: Data should be acquired by the Proposer from publicly available data sources.**

8. Regarding the acquisition of protein sequence data, if required:

   Is there a medium of acquisition that is preferred? For example, samples from a deployed or austere environment could include air, water, waste material, and biological samples from both human/animal and non-animal sources. The complexity of the peptide sequences could differ significantly between different sources. For example, skin or blood samples, would be more complex than a sputum or urine sample. Similarly, samples of air or human/animal would also contain different complexities of protein composition.
   **A: The aim of the solicitation is to predict pathogen:host interactions. The sample source is not relevant to the pathogen's proteome. This is not focused on pathogen surveillance.**

9. Is the intent of this proposal to investigate a platform for sample acquisition that allows the generation of "protein sequence data" from diverse samples?
   **A: Please see answer to Q8.**