Data Management Plan

A. Introduction: This Data Management Plan outlines the procedures for managing, storing, and sharing data generated during the course of the project on the analysis of hundreds of thousands of genomic sequences of Influenza A of various subtypes from public repositories (NCBI and GISAID). The research products include models and software, as well as example programs demonstrating how to access and read these models and apply them to raw data. An online system will also be developed to pull in new sequences from public databases, compute emergent risk, and display the results publicly. Experimental validation will involve specific protocols, cell lines, and procedures, which can be shared in an appropriate manner.

B. Data Types and Formats: This project will involve the following types of data:

- Genomic sequences of Influenza A subtypes
- Metadata, including sequence IDs and other relevant information
- Models and software for generating, inferring, and reading the models
- Example programs demonstrating model access and application to raw data
- Experimental validation data, including protocols, cell lines, and procedures

All data will be stored in open and widely used formats, such as FASTA for genomic sequences, JSON or CSV for metadata, and standard programming languages like Python for software and example programs.

C. Data Acquisition and Processing: Data will be acquired from public repositories, such as NCBI and GISAID, and processed using custom-built software tools. These tools will parse the raw data and metadata, analyze the genomic sequences, and generate models to assess emergent risk. Quality assurance and quality control measures will be in place during data collection, analysis, and processing to ensure the reliability of the results.

D. Data Storage and Preservation: Data, models, and software will be stored on secure servers with appropriate backup and version control systems. Genomic sequences and metadata will be deposited in public repositories like NCBI and GISAID, while models and other data will be deposited at Zenodo for long-term access with DOI identifiers. Example programs and software tools will be hosted on GitHub repositories, ensuring easy access and collaboration.

E. Data Sharing: Data sharing will be achieved through a combination of methods:

- Metadata and sequence IDs will be shared publicly, while respecting any restrictions on the genomic sequences themselves
- Models, software, and example programs will be available on GitHub repositories, allowing for easy access, collaboration, and updates
- Tools developed during the project will be easily installed from code registries like PyPI
- Experimental validation data, including protocols, cell lines, and procedures, will be shared in an appropriate and secure manner, ensuring compliance with any legal or ethical requirements
- An online system will provide public access to emergent risk assessment based on new sequences from public databases

F. Handling of Restricted Data: In cases where genomic sequences cannot be shared publicly due to legal or ethical constraints, the project will ensure that the handling and sharing of such data is in compliance with relevant regulations and guidelines. Sequence IDs and metadata will be shared, allowing other researchers to request access to the restricted sequences through appropriate channels. The project will have provisions in place for using sequences that are not publicly posted if the researchers or laboratories who collected the sequences do not grant permission to share them publicly. In such cases, the project team will collaborate with the data owners to determine the most appropriate way to utilize and share the data, ensuring that all parties' interests are respected and protected.

G. Monitoring Adherence to the Data Management Plan: Adherence to the Data Management Plan will be monitored throughout the project by a dedicated postdoctoral researcher, who will dedicate 5% of their time to this task. The postdoctoral researcher will work under the supervision of the Principal Investigator, ensuring that the Data Management Plan is properly executed and that all project members are aware of their data management responsibilities.

Monitoring will include periodic reviews of data storage, sharing, and preservation practices, as well as ensuring that data quality assurance and quality control measures are effectively implemented. Any deviations from the plan will be addressed promptly, and the plan will be updated as needed to accommodate

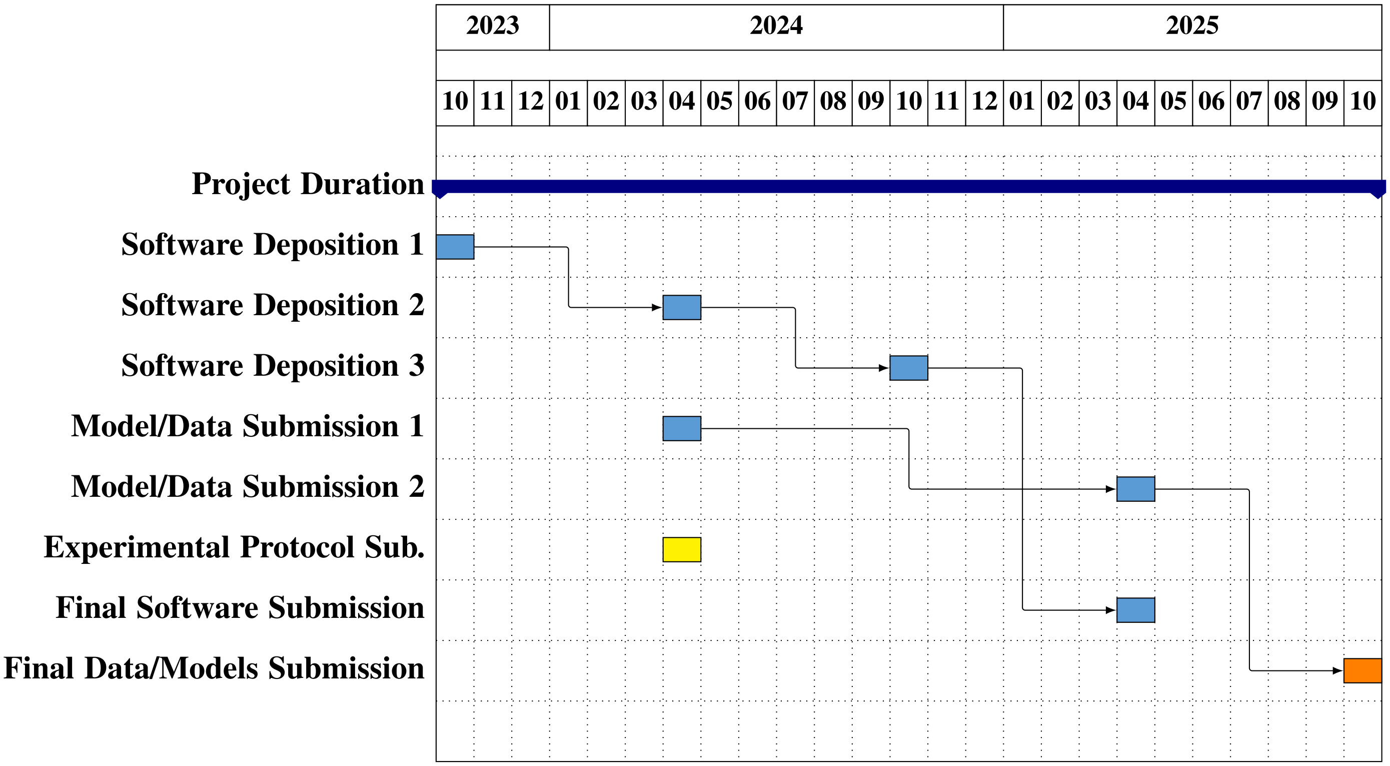

Fig. 3. DMP checkpoints. Software is deposited at Github repositories. Models and data are submitted at Zenodo.

new data types, formats, or sharing requirements.

H. Preservation Timeframe: Data preservation will be maintained for a minimum of 10 years following the completion of the project. This timeframe will ensure that the research products remain accessible and usable by the scientific community for future research and development.

I. Costs and Administrative Burden: The Data Management Plan takes into consideration the balance between the value of data preservation and other factors such as associated costs and administrative burden. Data storage, preservation, and sharing costs will be factored into the project budget. A total of 5% of the project's effort will be allocated to data management, including the postdoctoral researcher's time and any other administrative tasks related to data management.

The justification for any decisions regarding data preservation and sharing will be provided in the plan. This approach ensures that the research products are appropriately managed and shared, while also taking into account the costs and administrative burden associated with data management.

J. Periodic Review of the Data Management Plan: The Data Management Plan will be reviewed every six months by the Principal Investigator (PI) and co-Principal Investigator (co-PI) to ensure that it remains up-to-date and relevant to the project's needs. During these reviews, the PI and co-PI will assess the current data management practices and identify any necessary updates or changes to the plan.

These periodic reviews will help ensure that the Data Management Plan continues to be effective in guiding the project's data management activities and that it evolves as needed to accommodate new data types, formats, or sharing requirements. Any updates to the plan will be communicated to all project members to ensure that everyone remains informed about the project's data management expectations and responsibilities.

K. Compliance with DoD Instructions: This Data Management Plan adheres to the guidelines set forth in Section 3.c. Enclosure 3 of the Department of Defense (DoD) Instructions 3200.12. By following these guidelines, the project ensures that all data management practices are in compliance with the requirements of the DoD and that the data generated during the project will be appropriately managed, stored, and shared.