**Wellcome Data Science Ideathon – Proposal**

**BASELINE**

**Background**

In the complex and time-consuming realm of vaccine development, which typically spans 10-15 years, novel solutions are required, especially for elusive diseases like the Nipah virus (NiV). Conventional vaccination trials for NiV are impractical, requiring immense time and vaccine doses.

A potential accelerant to this process is utilizing 'correlates of protection'. These are measurable immune responses statistically linked with an organism's protection against a pathogen. Identifying these correlates and using them as surrogate endpoints can hasten vaccine development either by allowing an early conclusion to a trial or informing future vaccine success predictions.

Platforms like ImmPort and ImmuneSpace provide access to study data and tools to identify these correlates. However, improvements are needed. Firstly, making these tools more user-friendly for non-statisticians and non-coders can democratize access to data and analysis. Secondly, the integration of advanced machine learning methods can enhance the analysis of correlates across study datasets.

Ultimately, refining a platform for discovering and analysing correlates of protection could significantly expedite vaccine development, bringing us closer to combating infectious diseases with unprecedented speed and precision.

**Platform Overview**

The BASELINE platform, utilizing techniques such as Uniform Manifold Approximation and Projection (UMAP) assisted clustering in conjunction with the Gaussian Mixture Model (GMM), stands poised to extrapolate critical insights that can identify correlates of protection against extant and future infections as well as responses to vaccines.

One salient feature of our platform lies in its broad-spectrum data management capability, adept at accommodating diverse data streams. In order to elucidate the hierarchical interrelationships intrinsic to the provided data files, the platform employs graph-based data manipulation. This pivotal functionality enables users to discern common sample identifiers, thereby informing the selection of appropriate downstream data analysis techniques.

Users are afforded the option to engage in artificial intelligence-assisted analysis, leveraging an extensive baseline constructed from pre-existing data. Users can upload their anonymized datasets (RNA seq, NanoString, Fluorescence-Activated Cell Sorting (FACS), etc.). The data will subsequently be categorized (e.g., into responders and non-responders), facilitating the exploration of the 'new' study in the context of the extant baseline. The computational demands remain moderate due to the UMAP-GMM model, as the 'new' dataset will not undergo clustering but will be integrated into the existing model. The BASELINE model, further refining and enhancing its predictive capacity, will assimilate the new dataset into its base model. Notably, our platform can potentially accommodate an unlimited number of datasets, offering a comprehensive consolidation of datasets across diverse studies without constraints concerning the data type, be it NanoString, single-cell, or bulk RNA sequencing data.

The BASELINE model will be subsequently harnessed to predict antibody titers/responses, drawing upon metadata provided for patients. Following its development, the model will find application in clinical settings, categorizing individuals into clusters (response/no response), and thus providing clinicians with actionable insights. The model will be continually refined, incorporating additional studies into the BASELINE model to enhance prediction accuracy.

**Workplan**

Platform Refinement and Expansion

Over the course of the one-year grant period, we will focus on refining and expanding our platform, with a particular emphasis on user-friendliness, versatility, and advanced analytical capabilities. Here is a detailed timeline (Figure 1):

**Months 1-2: User-Experience (UX) Improvements**: We will initiate the project with a comprehensive UX review to identify areas where the platform's interface and usability can be enhanced. Feedback from existing users will be collated, and expert UX designers will be engaged to design improvements.

**Months 3-4: Implementation of UX Improvements**: Following the UX review, we will begin implementing the identified improvements. This will include reworking the user interface (UI), enhancing the navigation, and introducing more intuitive controls and help resources.

**Months 5-6: Incorporation of Advanced ML and Statistical Methods**: Concurrently, we will work on enhancing the platform's analytical capabilities. The incorporation of advanced machine learning methods like KNN, GMM, Louvain, and others will augment the platform's ability to intergrade and draw conclusions from vast and diverse datasets.

**Month 7: Testing and Iteration**: With the implementation of both UX improvements and advanced analytical capabilities, we will conduct thorough testing to ensure optimal performance. User testing will be carried out to ensure the changes have indeed improved the user experience. Any identified issues will be addressed, and the process will be iterated until a satisfactory level of performance and user experience is achieved.

**Months 8-9: Integration of Diverse Data Types and Structures**: Recognizing that different studies employ different kinds of data, we will focus on making our platform more versatile. Whether the data is from Nanostring, single cell/bulk RNA sequencing, or any other format, our platform will be able to handle it. This will involve developing algorithms capable of understanding and integrating these diverse data types.

**Month 10: Testing and Iteration**: Like step 4, testing will be conducted following the integration of diverse data types. This will ensure that the platform can effectively handle and analyse a wide variety of data formats.

**Month 11: Finalizing the Platform and Preparing for Deployment**: Upon successful testing, the platform will be finalized, and preparations will be made for deployment. This will include developing comprehensive user guides and training materials.

**Month 12: Deployment and User Training**: In the final month, the refined platform will be deployed. User training sessions will be organized to familiarize users with the new features and enhancements. Feedback will be gathered for future refinements.

**Throughout the project**: Data from various studies will be drawn together for the testing and optimising of the platform.

By following this rough timeline, we aim to create a user-friendly, versatile, and analytically robust platform that caters to a wide range of user needs and data types, thereby accelerating the discovery and analysis of correlates of protection in vaccine research.

**Collaboration with Biostatisticians**

Our team will work closely with biostatisticians, particularly those who advise regulatory authorities. Their expertise will guide the platform's development process and the interpretation of the findings. Regular workshops, training sessions, and continuous dialogue with these experts will enhance our understanding of regulatory requirements and statistical best practices.

**Adapting to New Diseases/Vaccines**

We will ensure our modelling pipelines' adaptability and flexibility to accommodate new diseases and vaccines. The continual learning algorithms employed by the platform will enable iterative learning from new data, thus facilitating the platform's adaptability to evolving challenges.

**Working with Government Stakeholders**

Transparent communication, active collaboration, and regulatory compliance will be central to our approach in working with government stakeholders. Independent audits, workshops, training sessions, and evidence-based recommendations will be employed to build trust and ensure the broad acceptance of our platform.

**Addressing Community Needs**

We recognize the importance of community engagement and trust. We will implement community outreach programs, town halls, and educational campaigns to address public concerns and gather input. A patient-centric approach will ensure the alignment of our decisions with community values and preferences.

**Collaboration with LMICs**

We aim to engage with researchers in Low- and Middle-Income Countries (LMICs) in a co-development approach to make the platform more responsive to their unique needs. Training workshops, dedicated support, and language localization features will be employed to democratize access to our platform in these regions.

**Data Governance Statement**

Our platform is committed to maintaining stringent data governance standards. We ensure strict adherence to privacy regulations like GDPR and HIPAA, utilize advanced encryption for data security, and enforce strict access controls. In addition, we focus on data quality and uphold transparency in our data handling practices, granting users full control over their data. Our goal is to provide a secure, compliant environment that fosters reliable vaccine development research.

**Expected Outcomes**

1. A refined, user-friendly, and versatile platform capable of analysing a wide range of data types and structures to predict correlates of protection.
2. Establishment of a robust, ongoing collaboration with biostatisticians and government stakeholders.
3. A platform capable of adapting to the evolving landscape of infectious diseases and vaccines.
4. Increased acceptance and trust of ML methods among key stakeholders and the public.
5. Wider accessibility and use of our platform in LMICs, contributing to global vaccine research.

In conclusion, our team is confident that the refinement and application of our platform will make a significant contribution to accelerating vaccine development, fostering global collaborations, and building trust in machine learning methods in vaccine research. We believe our work over the one-year grant period will yield vital advancements in the global fight against infectious diseases.

## Budget Template

**Salaries:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name Justification Period on project (months) % time Total** | | | |
| Gokhan Tut | Project lead | 12 | Full time - 54k |
| Duncan Murray | Data linkage | 12 | Part time -10% 10k |
| Wayne Croft | Clustering and AI | 12 | Part time -10% 5k |
| David Bone | UI design | 12 | Part time- 10% 3k |
|  |  |  |  |

**Materials and consumables:**

|  |  |  |
| --- | --- | --- |
| **Description Justification Total** | | |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Equipment:**

|  |  |  |
| --- | --- | --- |
| **Description Justification Total** | | |
| Laptops | Upgrading existing hardware | 8k |
| Monitors and other additional hardware | Upgrading existing hardware | 3k |
|  |  |  |
|  |  |  |
|  |  |  |

**Access charges:**

|  |  |  |
| --- | --- | --- |
| **Description Justification Total** | | |
| Open AI API and other services | Use of API and other services proved by Open AI to help accelerate our work | £1000 |
| Cloud services to host and run platform | We would need to execute final product on a cloud service to make it accessible by everyone | 8K |
|  |  |  |
|  |  |  |
|  |  |  |

**Travel and subsistence:**

|  |  |  |
| --- | --- | --- |
| **Description Justification Total** | | |
| Conferences | Would need to keep on top of the current AI trends and uses to make sure platform is modern and accessible | 3k |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Miscellaneous:**

|  |  |  |
| --- | --- | --- |
| **Description Justification Total** | | |
| External support | Incase when building and implementing platform there may be a requirement from an expert like cloud computing or website design | 5k |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Summary of costs requested:**

|  |  |
| --- | --- |
| **Description** | **Total** |
| **Salaries** | 72k |
| **Materials & Consumables** | 0 |
| **Equipment** | 11k |
| **Access charges** | 9k |
| **Travel and subsistence** | 3k |
| **Miscellaneous – other** | 5k |
| **Grand Total** | 100K |