Predictive Modeling of Immune Responses to Pertussis Vaccination

Project by: Peng Cheng, Javier Garcia, Weikang Guan, Brian Qian
Advisors: Barry Grant (Professor), Jason Hsiao (PhD student)

UC San Diego
JACOBS SCHOOL OF ENGINEERING
MAS Data Science and Engineering

CMI-PB COMPUTATIONAL MODELS OF IMMUNITY — PERTUSSIS BOOST —

PROBLEM

Pertussis, commonly known as whooping cough, is a highly contagious lung infection caused by the bacterium Bordetella pertussis. Vaccination is the primary strategy for controlling the spread of pertussis. There are two main types of vaccines: whole-cellular (wP) and acellular (aP). The efficacy of vaccine-induced immunity can diminish over time. Several factors influence this, including the type of vaccine received and the individual's age at vaccination.

Vaccine Types

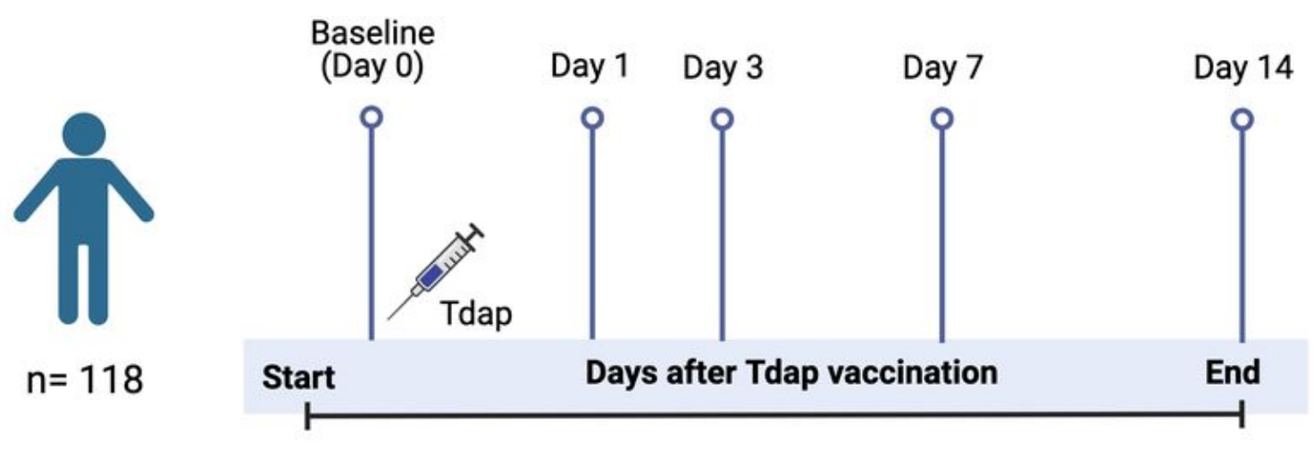
1. **Whole-cellular (wP) vaccine:** Contains inactivated whole bacteria. Known for higher efficacy but associated with more adverse reactions.
2. **Acellular (aP) vaccine:** Contains purified components of the bacteria. It has fewer side effects but may have lower long-term efficacy.

Challenges

Balancing the safety and efficacy of pertussis vaccines remains a critical challenge. wP vaccines offer robust initial protection but can cause severe side effects. aP vaccines are safer but may not provide as durable immunity.

Research Importance

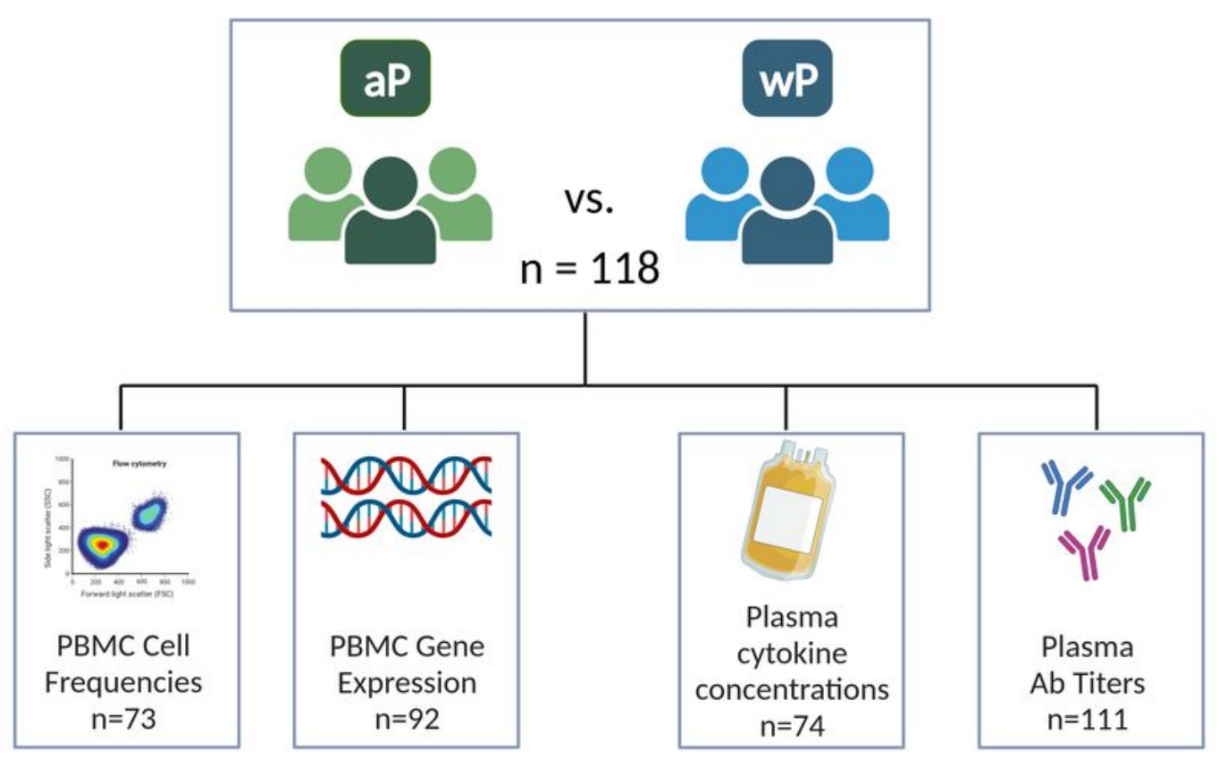
Ongoing research and monitoring of vaccine effectiveness are crucial. Surveillance data help in understanding the duration of immunity and the potential need for booster doses.



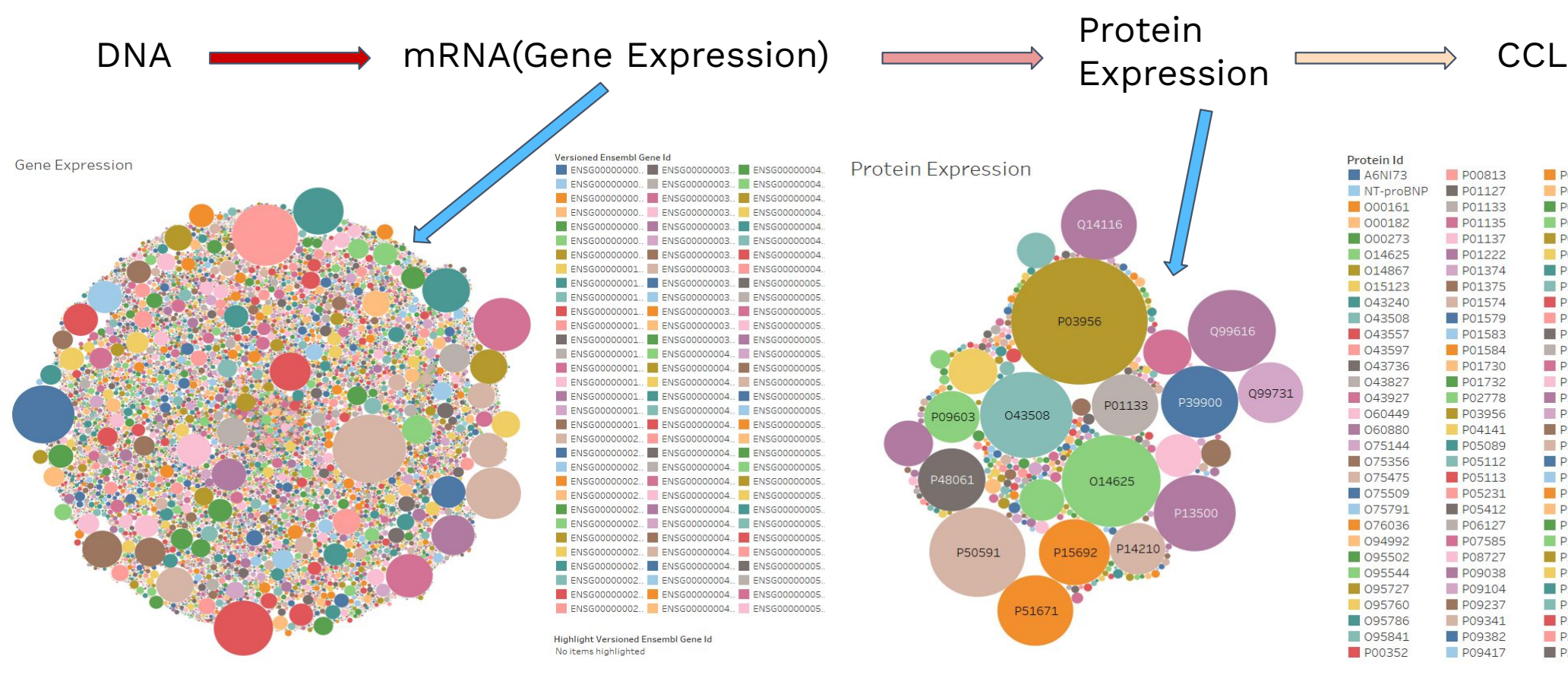
Objective: Analyze immune responses post-Tdap booster vaccination.

Subjects: 118 individuals contributing 500+ blood specimens.

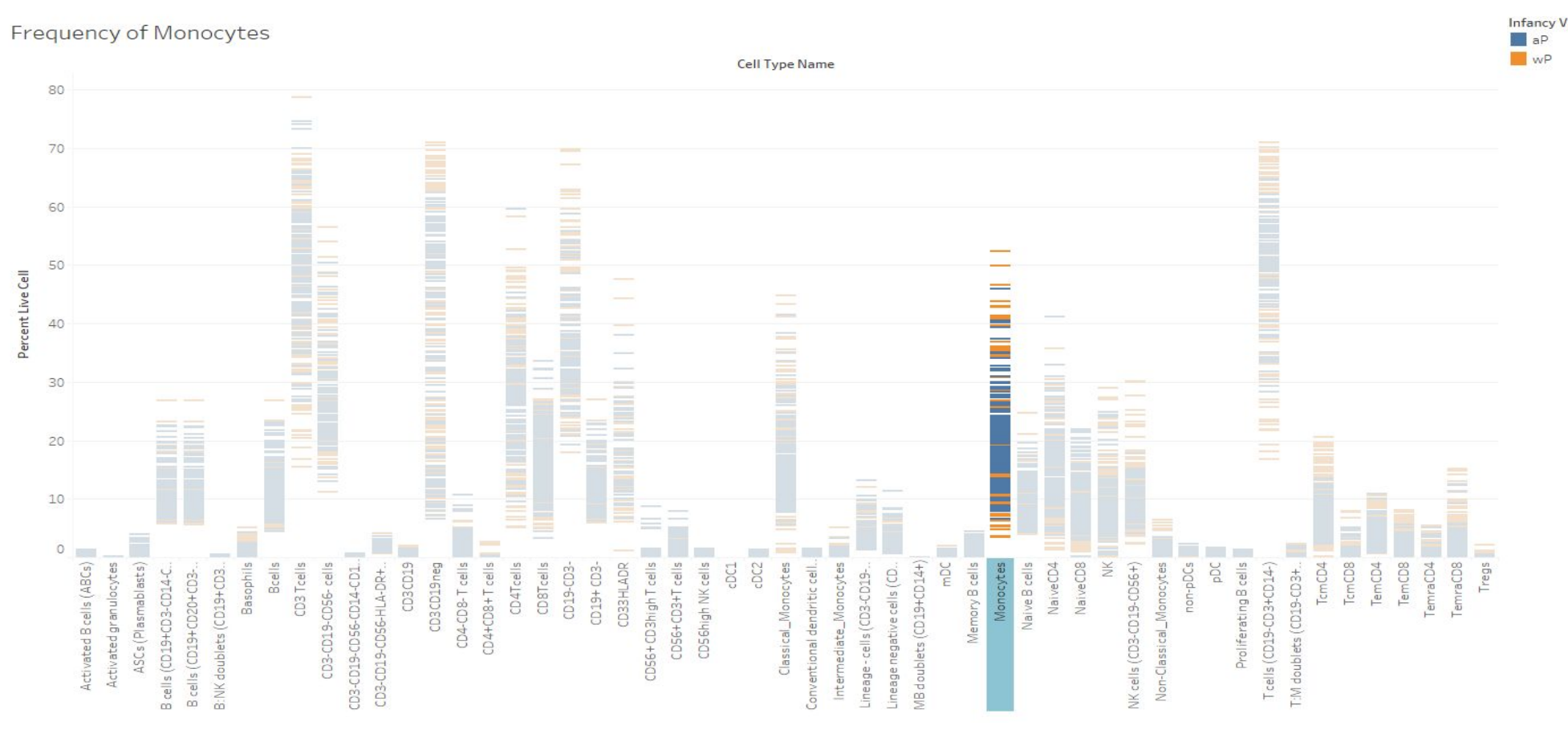
Timeframe: Pre- and post-vaccination (days 1, 3, 7, and 14).


aP
wP
vs.
n = 118
PBMC Cell Frequencies n=73
PBMC Gene Expression n=92
Plasma cytokine concentrations n=74
Plasma Ab Titers n=111

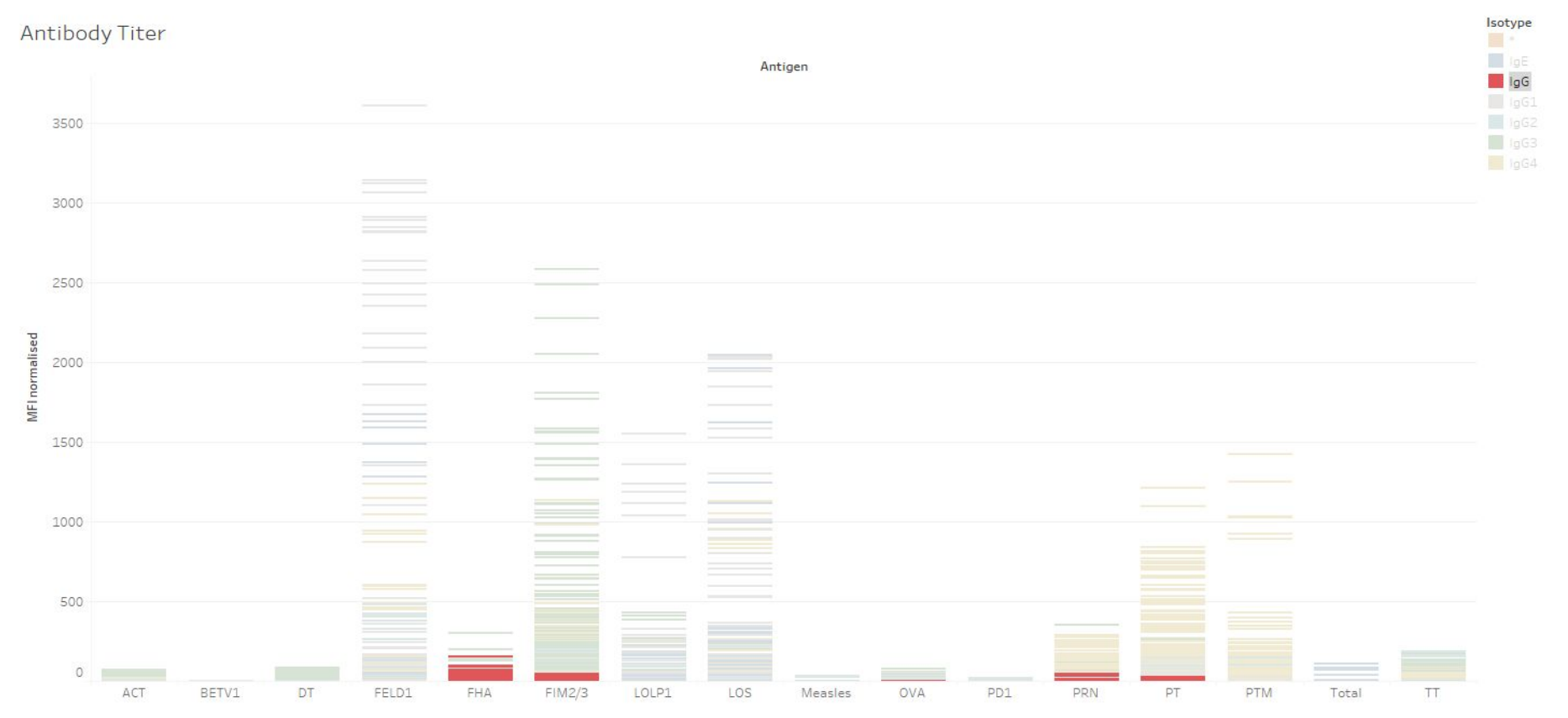
OVERVIEW OF DATA



DNA is transcribed into messenger RNA, which carries instructions from genes to produce proteins at the right times and in the right amounts. CCL3 is crucial in the inflammatory process and is significant for pertussis research. Numerous genes and proteins are expressed, with CCL3 being one of the key genes under study.



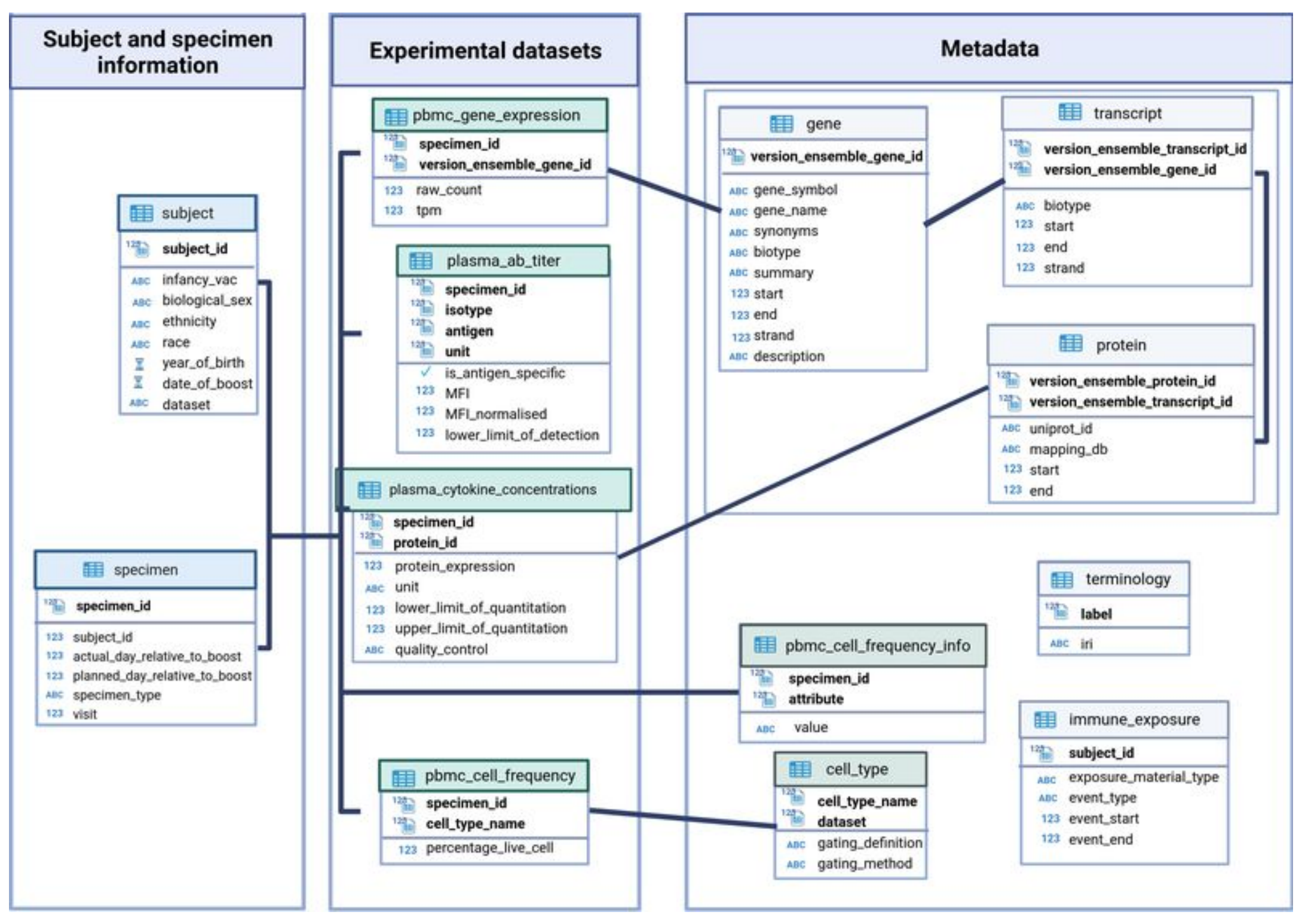
We collect many features for different cells but are only predicting monocytes. Monocytes, a type of white blood cell, defend the body against infections and are collected from blood samples. Other cell types can be used to predict monocyte values.



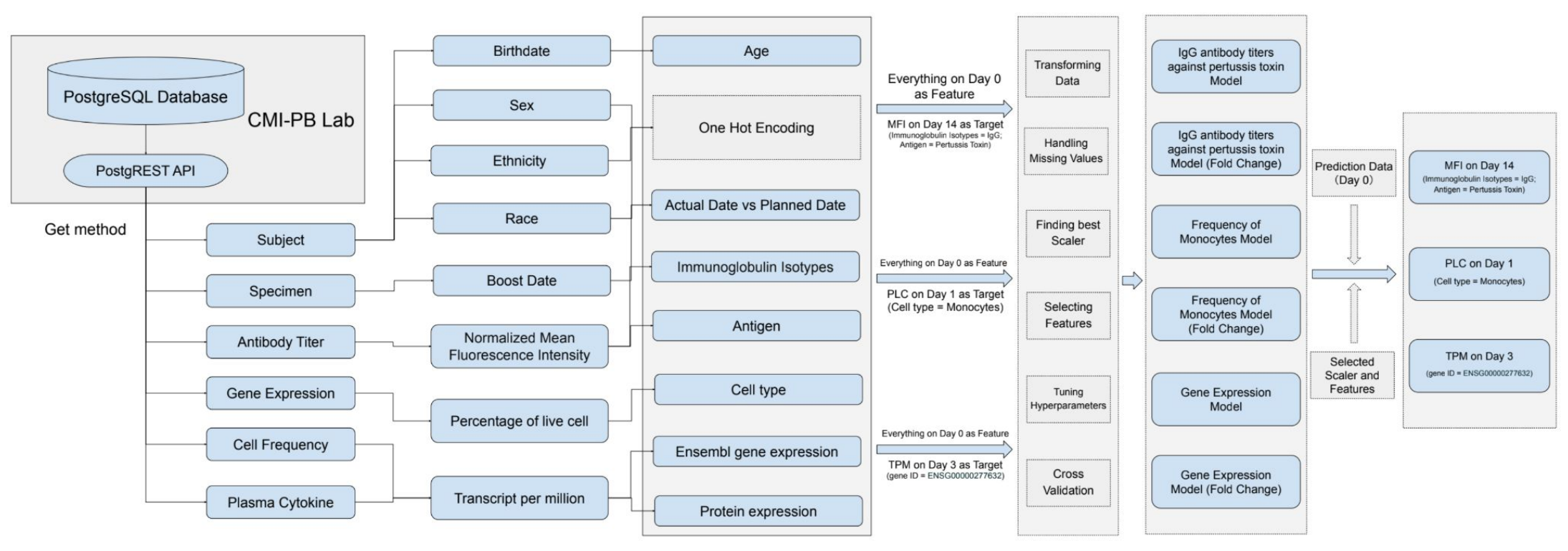
Antigens and IgG are a common subset of immunoglobulins (antibodies). Mean fluorescence intensity (MFI) quantitatively measures antibody titer (concentration of antibodies in the blood).

METHODS AND TECHNIQUES

PostgreSQL database provided by CMI-PB



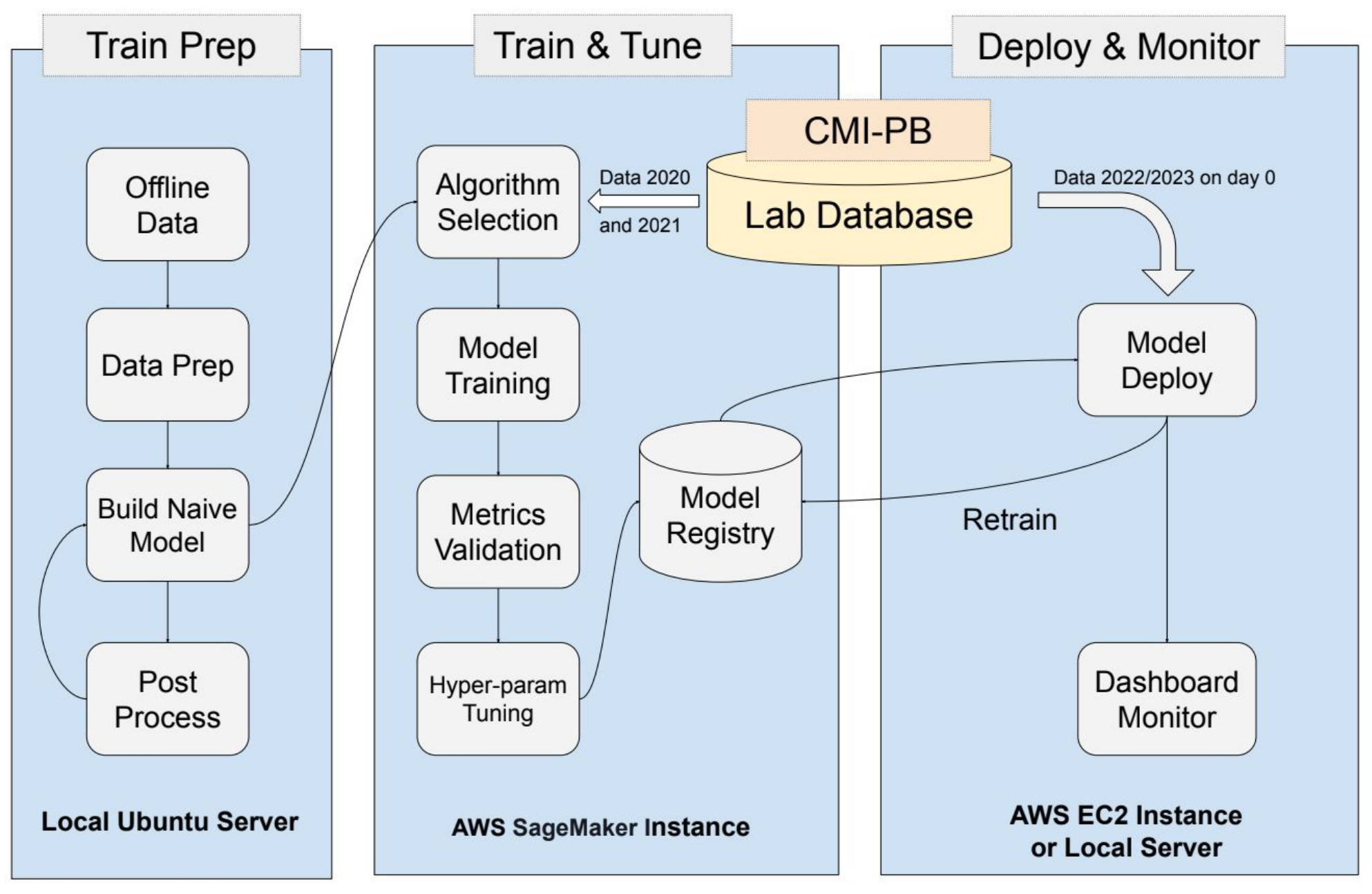
We extract the most critical data from these files. The subject data includes demographic information about the individual. The specimen data comprises the blood sample details, such as the collection date and the identity of the sample's owner. Experimental tables capture the measurements from the blood sample, indicating the body's status at the collection time.



Data preprocessing involves handling missing values, managing outliers, and scaling features to ensure the data is ready for model training. Each model requires specific feature selection, training algorithms, and evaluation metrics to ensure accurate and reliable predictions:

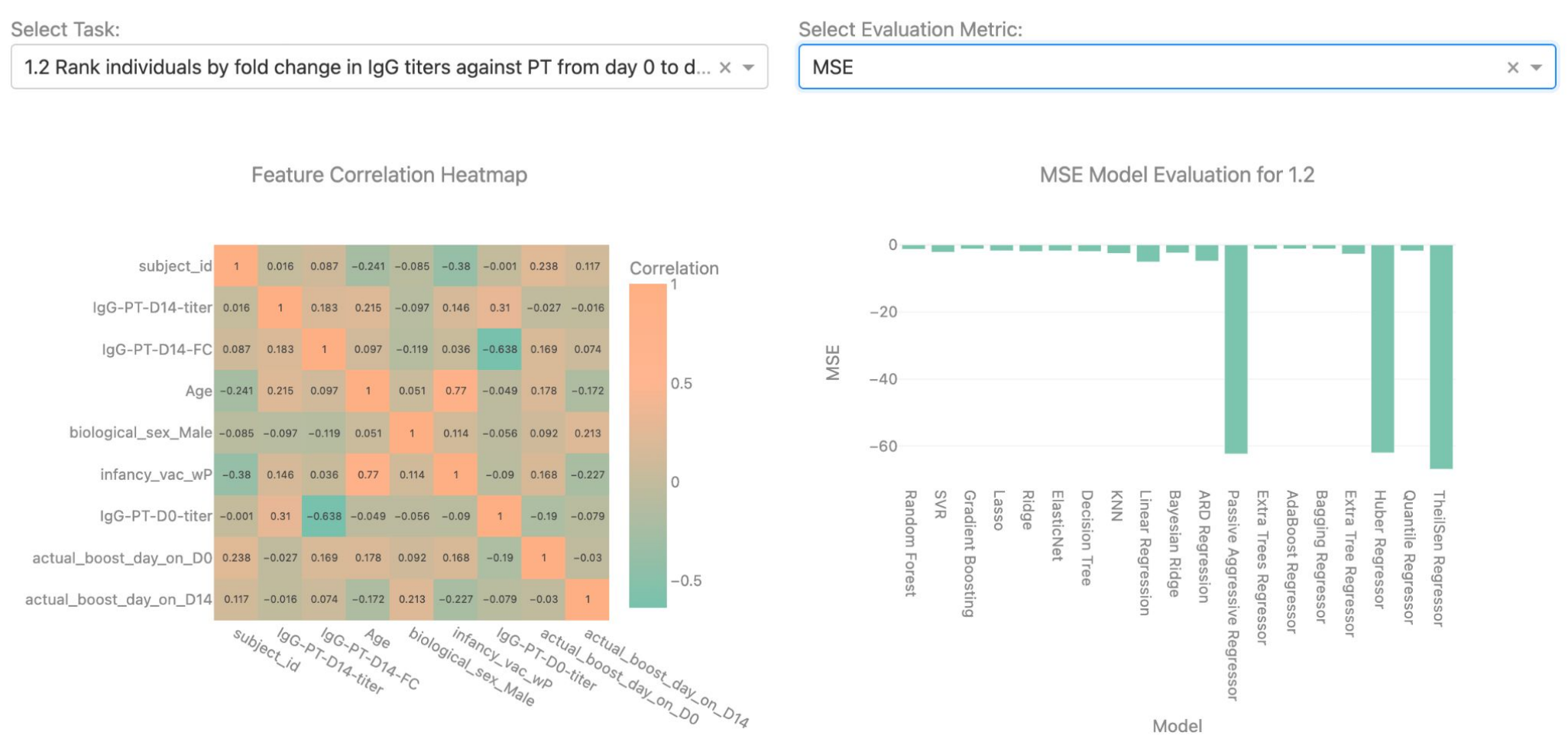
1. *Antibody model*: MFI on day 14 after vaccination
2. *Monocytes Frequency model*: Monocytes on Day 1 after vaccination
3. *Gene Expression model*: TPM on Day 3 after vaccination

SOLUTION ARCHITECTURE

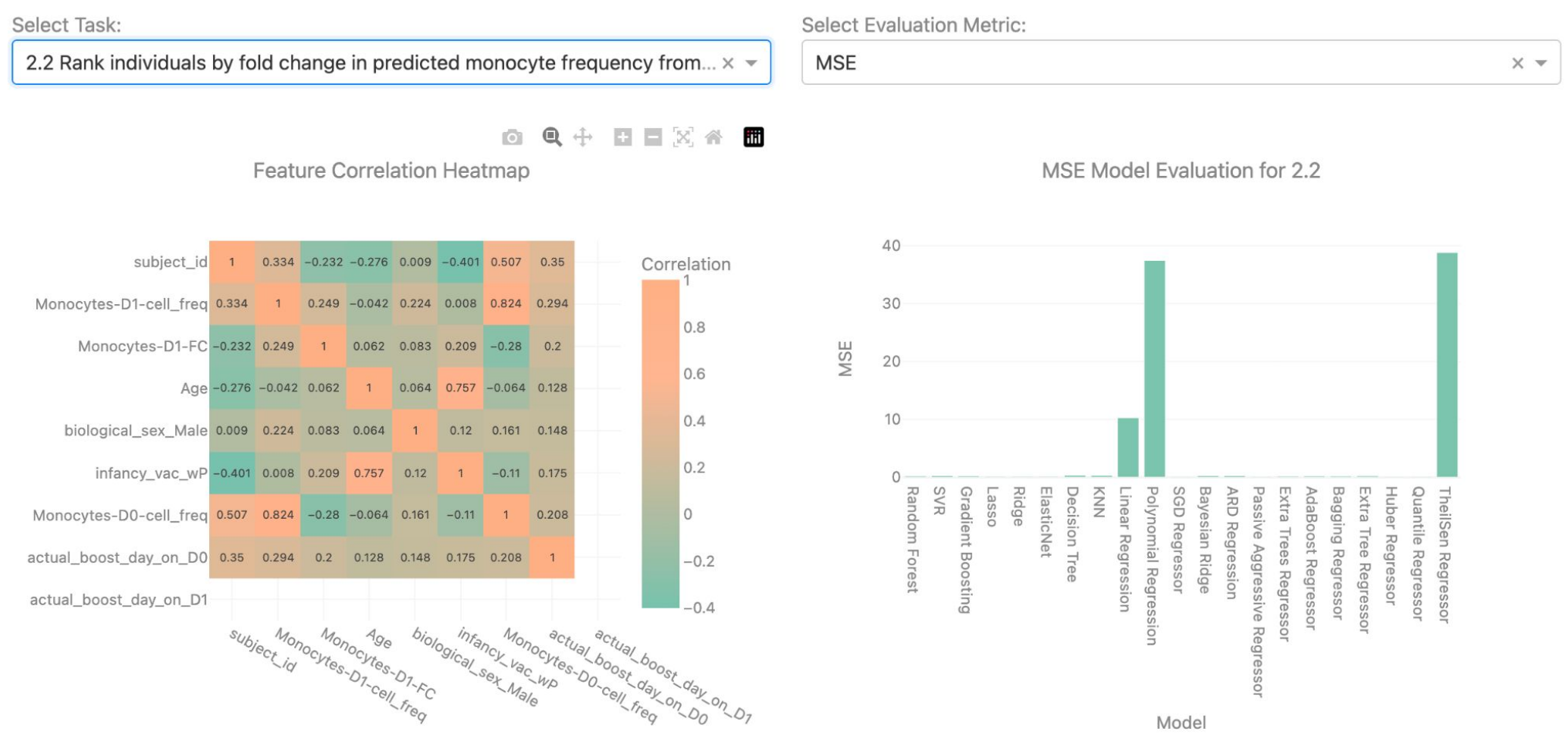


We initially utilized data from 2020-2021 to build naive models, leveraging static data to identify effective model types and key features. Through this process, we gained insights into which models performed best and which features were most significant for our predictions. We integrated a feedback cycle in AWS SageMaker, utilizing API data from 2020-2022. This allowed us to train our models on the cloud, benefiting from scalable computing resources and streamlined workflows. Once satisfied with our models' performance, we saved the best-performing models to a model registry in an S3 bucket for future use and deployment.

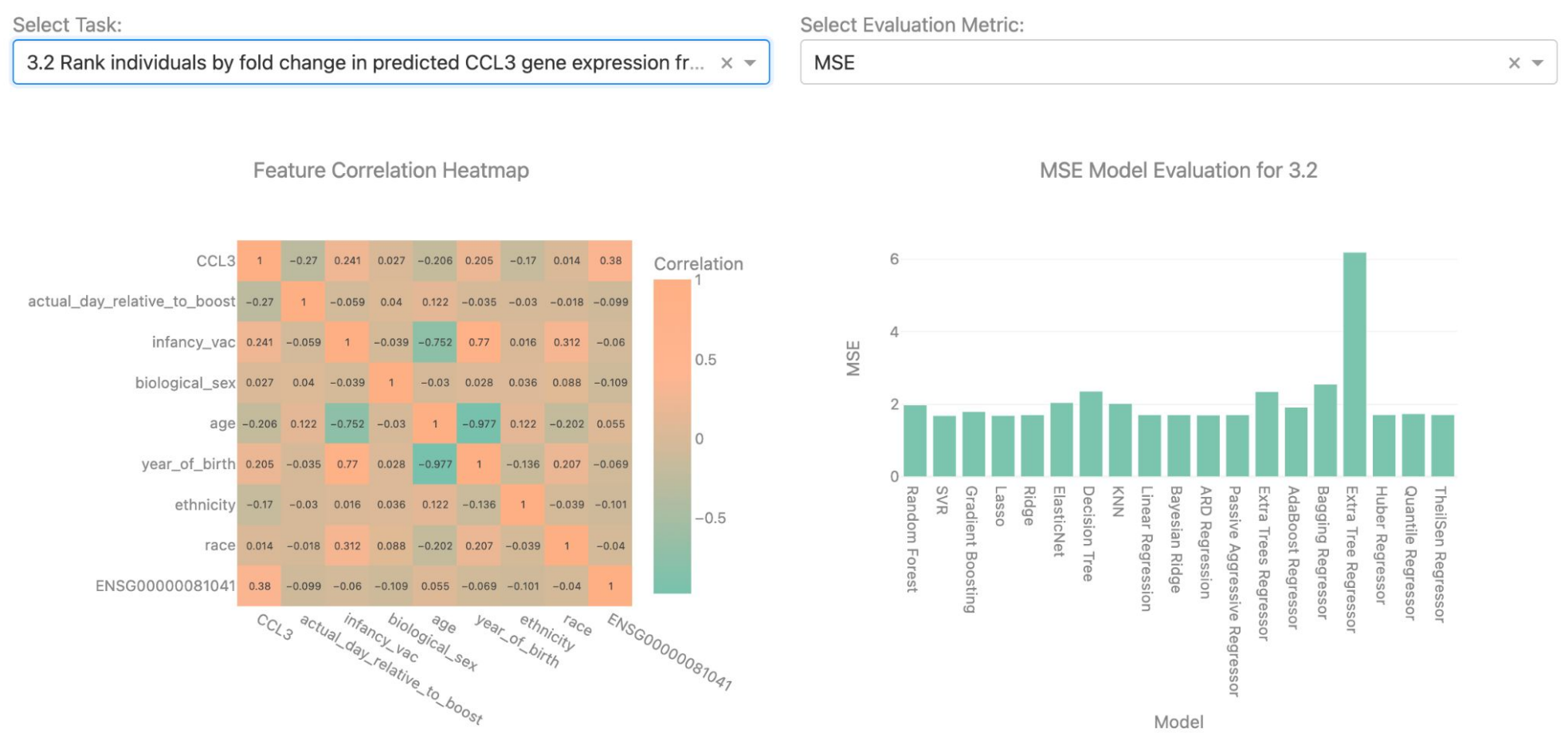
RESULTS


Select Task: 1.2 Rank individuals by fold change in IgG titers against PT from day 0 to d...
Select Evaluation Metric: MSE

Task 1.2


Select Task: 2.2 Rank individuals by fold change in predicted monocyte frequency from...
Select Evaluation Metric: MSE

Task 2.2


Select Task: 3.2 Rank individuals by fold change in predicted CCL3 gene expression fr...
Select Evaluation Metric: MSE

Task 3.2

KEY INSIGHTS

Support Vector Regression (SVR): The model performed well in predicting antibody titer levels. SVR is effective in high-dimensional spaces and can handle nonlinear relationships, making it suitable for handling our complex biological data.

Extra Tree Regressor and Gradient Boosting: These ensemble methods outperformed other methods in predicting the Log2 fold change of antibody levels at day 14 relative to day 0. They were able to reduce overfitting through bagging and boosting techniques, making them a good fit for our dataset.

Gradient Boosting for Monocyte Frequency: Gradient Boosting showed the best performance in predicting monocyte frequency at day 1. The approach of this model helps minimize the prediction error, making it applicable to a variety of tasks.

ElasticNet for Gene Expression Levels: Initially, ElasticNet performed best in predicting CCL3 gene expression levels at day 3. However, simpler models such as Stochastic Gradient Descent(SGD) Regressor and TheilSen Regressor outperformed ElasticNet when predicting the Log2 fold change of the target variable.