Second CMI-PB Prediction Challenge

Saonli Basu, Michael Anderson, Cheng-Chang Wu, Josey Sorenson, Katherine Li and Bhargob Kakoty

University of Minnesota

Second CMI-PB Challenge Dataset

- Training consists of 2020 and 2021 datasets.
- Contains cell frequency, gene expression, cytokine, and plasma antibody data.
- Predict ranks using baseline readouts of the 2022 dataset.
- Demographic data such as age, biological sex, vaccine priming status is provided.

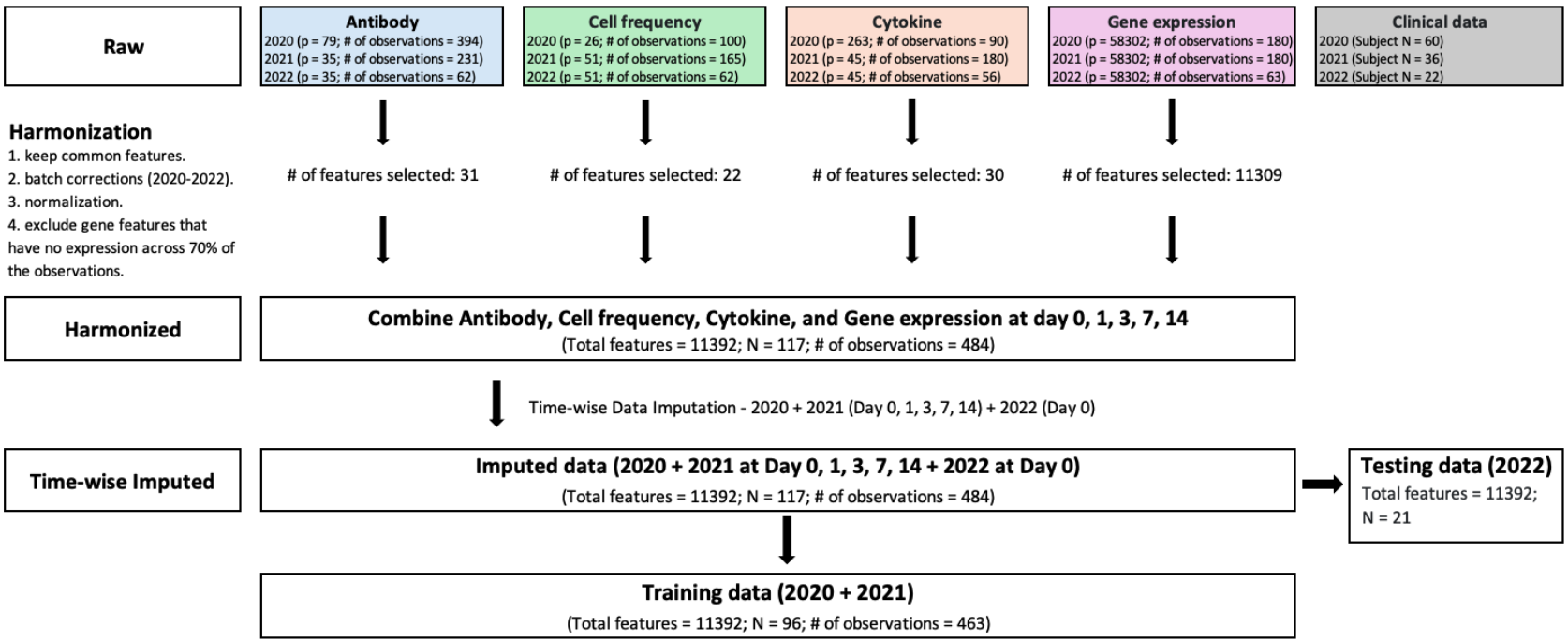
2020 and 2021 Data Demographics

	levels	aP	wP
N		47	47
Sex	Female	34	30
	Male	13	17
Ethnicity	Hispanic or Latino	12	11
	Not Hispanic or Latino	34	33
	Unknown	1	3
Race	Asian	15	12
	White	13	25
	American Indian	1	0
	Other Pacific Islander	2	0
	More than One Race	8	2
	Unknown or Not Reported	8	6
Age		20.2 (1.01)	30.29 (5.59)

2022 Data Demographics

	levels	aP	wP
N		12	9
Sex	Female	7	6
	Male	5	3
Ethnicity	Hispanic or Latino	3	0
	Not Hispanic or Latino	9	9
Race	Asian	3	2
	White	7	7
	More than One Race	1	0
	Unknown or Not Reported	1	0
Age		22.9 (2.93)	29.97 (3.42)

Processing Steps

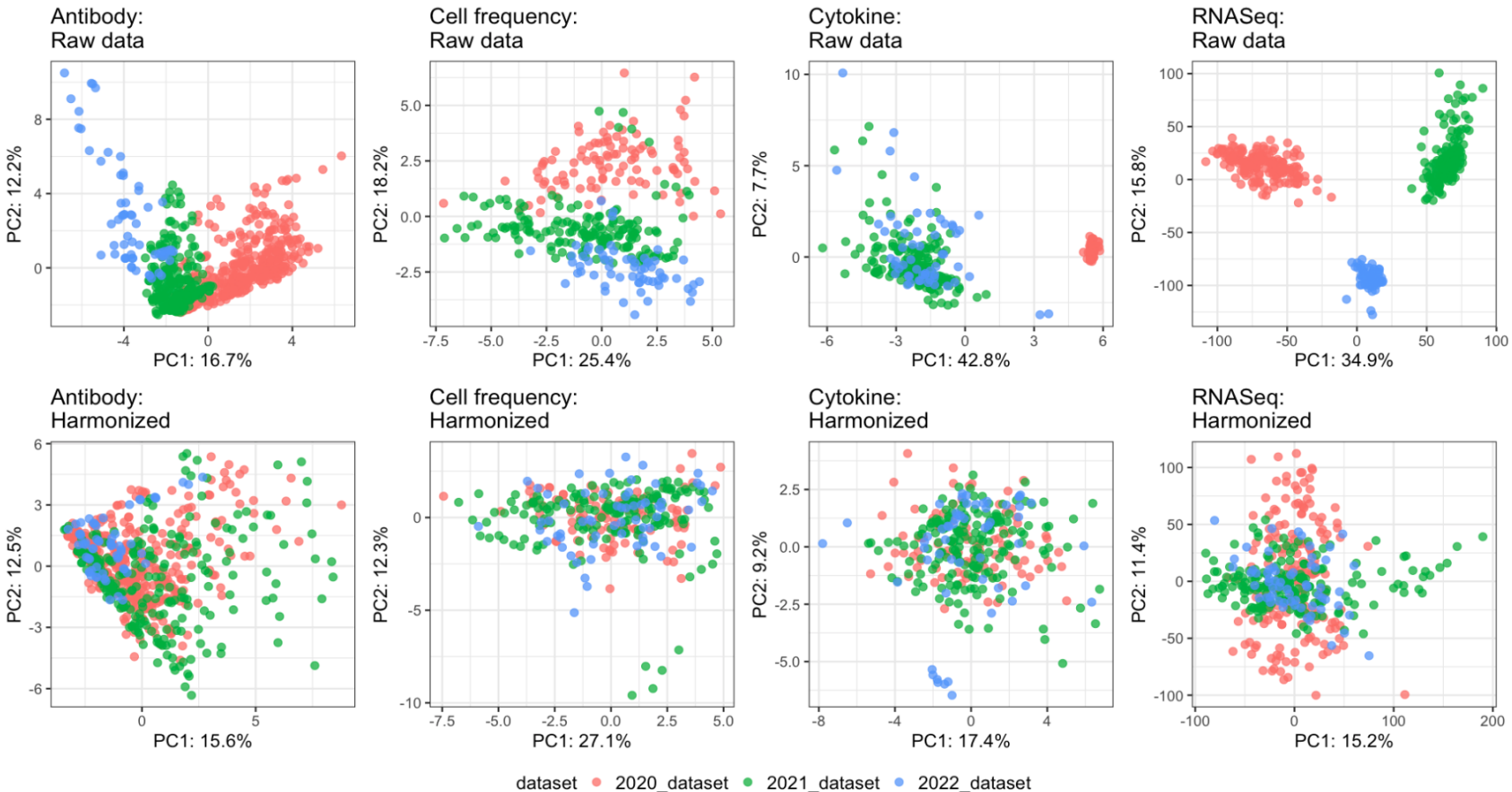


Processing Steps

- When examining the data, we noticed considerable batch effects for each year.
- We also noted several variables or observations have high missingness.
- Addressing these quality control steps was important.

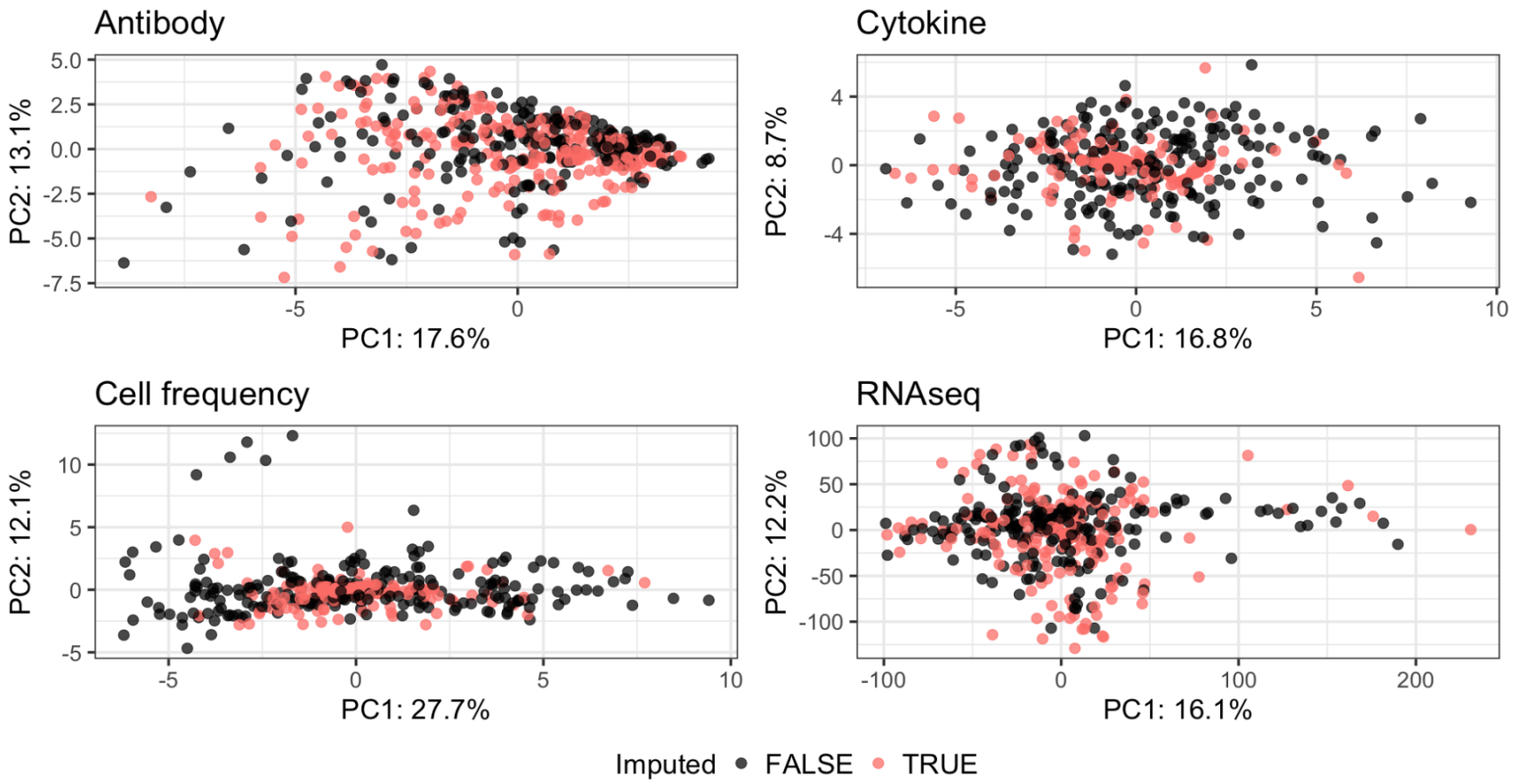
Harmonization

- Batch correction + Normalization



Imputation using Bayesian PCA Approach

- Combines 4 different modalities
- Imputation performed separately for planned days: 0, 1, 3, 7, and 14



Features

Features

	Training dataset		Test dataset	Harmonized dataset
	2020	2021	2022	
Antibody	79	35	35	31
Cell frequency	26	51	51	22
Cytokine	263	45	45	30
Gene expression	58302	58302	58302	11309

Task 1: Antibody titer tasks

- Rank the individuals by plasma antibody level 14 days post booster vaccinations.
- Rank the individuals by fold change of by plasma antibody level 14 days post booster vaccinations compared to the level at day 0.

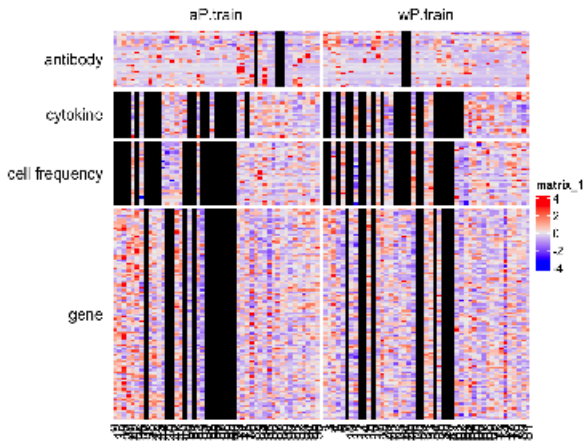
Task 2: Cell frequencies tasks

- Rank the individuals by predicted frequency of Monocytes on Day 1 after booster.
- Rank the individuals by fold change of predicted frequency of Monocytes on Day 1 after booster.

Task 3: Gene expression tasks

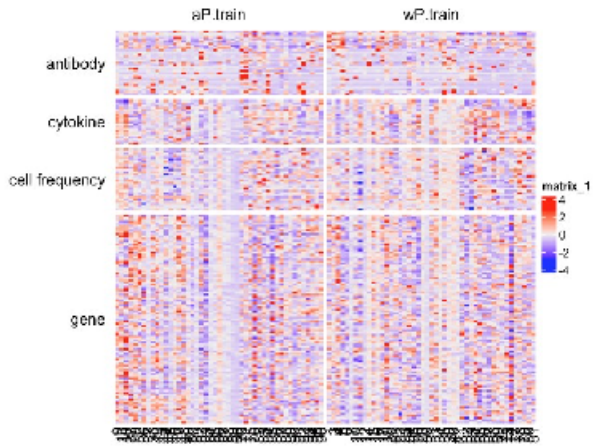
- Rank the individuals by predicted gene expression of CCL3 on Day 3 after booster.
- Rank the individuals by fold change of predicted gene expression of CCL3 on day 3 after booster.

Model Construction Using the Gene Expression Data



Harmonized Dataset (n=96)

Data imputation



Imputed Dataset (n=96)

$$X_1 \quad A_1$$
$$X_2 \quad A_2$$
$$X_3 \quad A_3 \quad \times W$$
$$X_4 \quad A_4$$

MCIA factors (k=10)

$$\approx$$

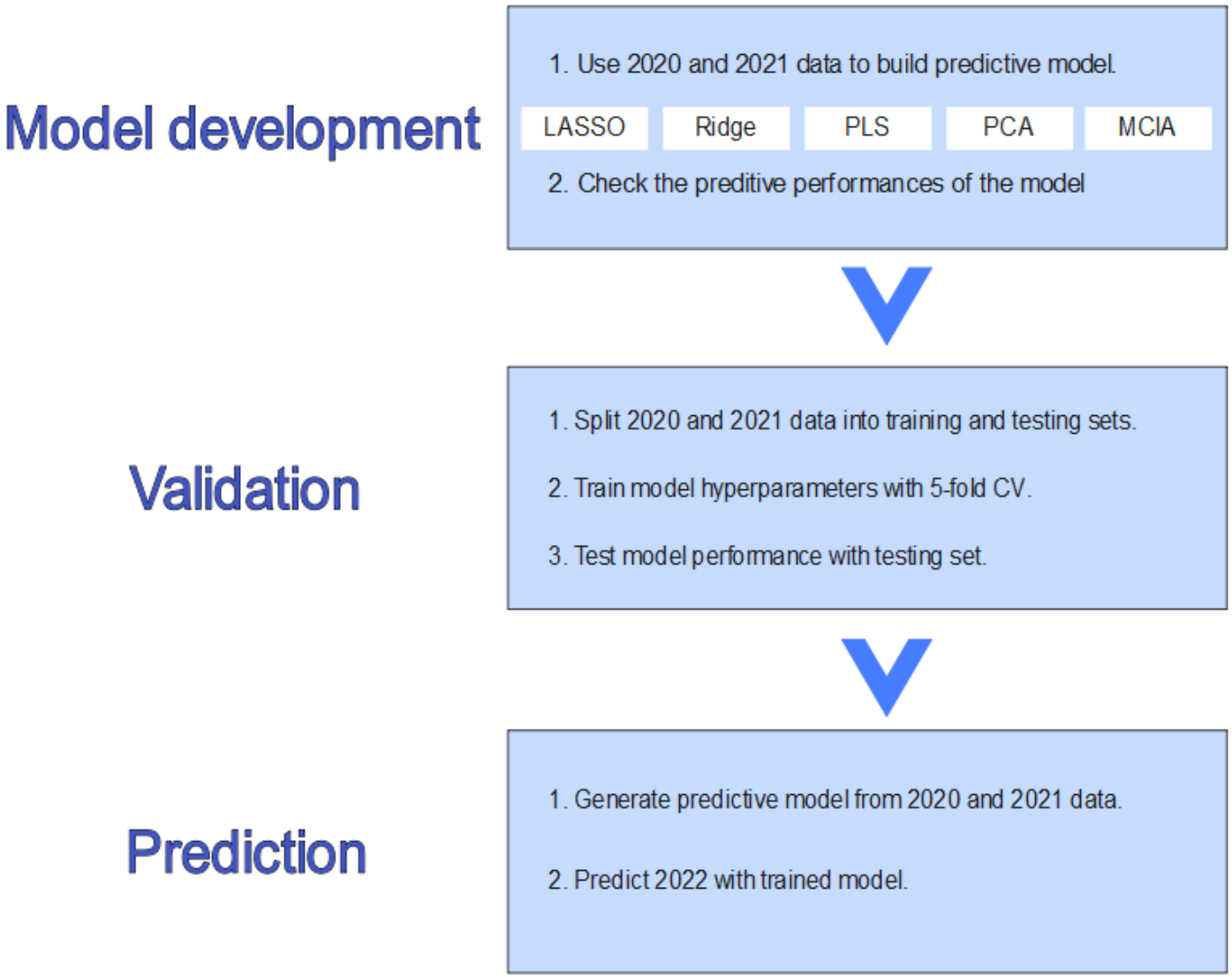
G

Calculate global scores (G) from loading scores (A)

$$y = X\beta + Zu + \epsilon$$

Linear Mixed Effects Model

Model Building and Prediction



LME using multi-omics factors

- Model multi-omics factor as the fixed covariates ($\tilde{X}_{ij}$ where i factor at time point j).
- Use a random intercept model to allow individual specific effect.
- Use cross validation and RMSE to assess model performance.

$$y = (\beta_0 + b_0) + t_j\beta_{time} + \tilde{X}_{1,j}\beta_1 + \cdots + \tilde{X}_{3,j}\beta_3$$
$$+\tilde{X}_{4,j}\beta_4 + \cdots + \tilde{X}_{10,j}\beta_{10} + \tilde{X}_{1,j}j\beta_{1,j} + \tilde{X}_{2,j}j\beta_{2,j} + \tilde{X}_{3,j}j\beta_{3,j} + \epsilon,$$
$$j = 1, ..., 5 \qquad (1)$$

$b_{01}, \ldots, b_{0N} \sim N(0, \sigma_b^2)$ independent of
$\epsilon_1, \ldots, \epsilon_N \sim N(0, \sigma^2)$.

Prediction Steps

- Estimate $\hat{b}_{01}, \ldots, \hat{b}_{0N}$ using BLUP.
- Global scores for test data subjects were computed for the test dataset by utilizing factor loadings from the training dataset and feature scores from the test dataset.
- Use Multi-omics factors for the test subjects and estimated regression coefficients in Equation 1 to predict Y for the test data subjects for each specific task.

Results

Number	Challenge	Spearman Corr
1.1)	IgG-PT-D14-titer-Rank	0.481818182
1.2)	IgG-PT-D14-FC-Rank	0.885714286
2.1)	Monocytes-D1-Rank	0.630074733
2.2)	Monocytes-D1-FC-Rank	0.297402597
3.1)	CCL3-D3-Rank	0.535064935
3.2)	CCL3-D3-FC-Rank	0.236363636