Predicting High versus Low anti–Pertussis-Toxin IgG Booster Responses from Baseline Non-Transcriptomic Immune Features in the CMI-PB 3rd Public Challenge

A Leakage-Controlled Systems-Immunology Machine-Learning Resource

K-Dense Web
contact@k-dense.ai

July 10, 2026

Abstract

Background. Infant priming with acellular (aP) versus whole-cell (wP) pertussis vaccine durably imprints the adult immune system, and the magnitude of the anti–pertussis-toxin (anti-PT) IgG recall after a Tdap booster varies widely between individuals. The Computational Models of Immunity–Pertussis Boosting (CMI-PB) resource provides dense, multimodal pre- and post-boost measurements that make it possible to ask whether the baseline (day-0) immune state alone predicts who will mount a strong antibody response. **Objective.** Using the CMI-PB 3rd public prediction-challenge dataset (2024 edition; 117 subjects sampled at days $-30/-14/0$ pre-boost and 1/3/7/14 post-boost), we built and rigorously validated a classifier that predicts a high- versus low anti-PT IgG fold-rise endpoint from baseline, non-transcriptomic features only (Luminex antibody titers, Olink cytokines, blood cell-population frequencies, and demographics; no RNA-seq). **Methods.** The endpoint was the day-14/day-0 ratio of normalized anti-PT IgG, dichotomized at the cohort median (ratio = 4.239); subjects lacking either time point were excluded ($n = 7$ of 118, leaving 111 labeled training subjects; 55 high vs. 56 low). Eighty-four cross-year–harmonized baseline features were modeled with an ElasticNet logistic regression and a Random Forest. Generalization was estimated with *nested* 10× repeated stratified 5-fold cross-validation (50 outer folds), where imputation, scaling, and hyperparameter selection (inner 5-fold grid search) were confined to each outer training split, eliminating hyperparameter-selection and preprocessing leakage. **Results.** Both models discriminated responders well above the 49.5% positive prevalence. The Random Forest achieved AUROC $= 0.841 \pm 0.081$ and AUPRC $= 0.838 \pm 0.087$; the ElasticNet logistic regression achieved AUROC $= 0.771 \pm 0.074$ and AUPRC $= 0.741 \pm 0.091$ (mean±SD over 50 folds). Pre-existing baseline anti-PT IgG was the single dominant predictor and carried a *negative* coefficient (higher baseline titer → lower fold-rise), a textbook ceiling effect, complemented by baseline B-cell and memory T-cell frequencies and selected Olink cytokines (IL-4, CCL3, IL-10). **Conclusions.** A parsimonious baseline immune signature—dominated by pre-existing anti-PT antibody and B-cell frequency—predicts Tdap booster responsiveness with good, leakage-free accuracy. Per-subject predictions for the 54 held-out challenge subjects and fully reproducible code are provided.

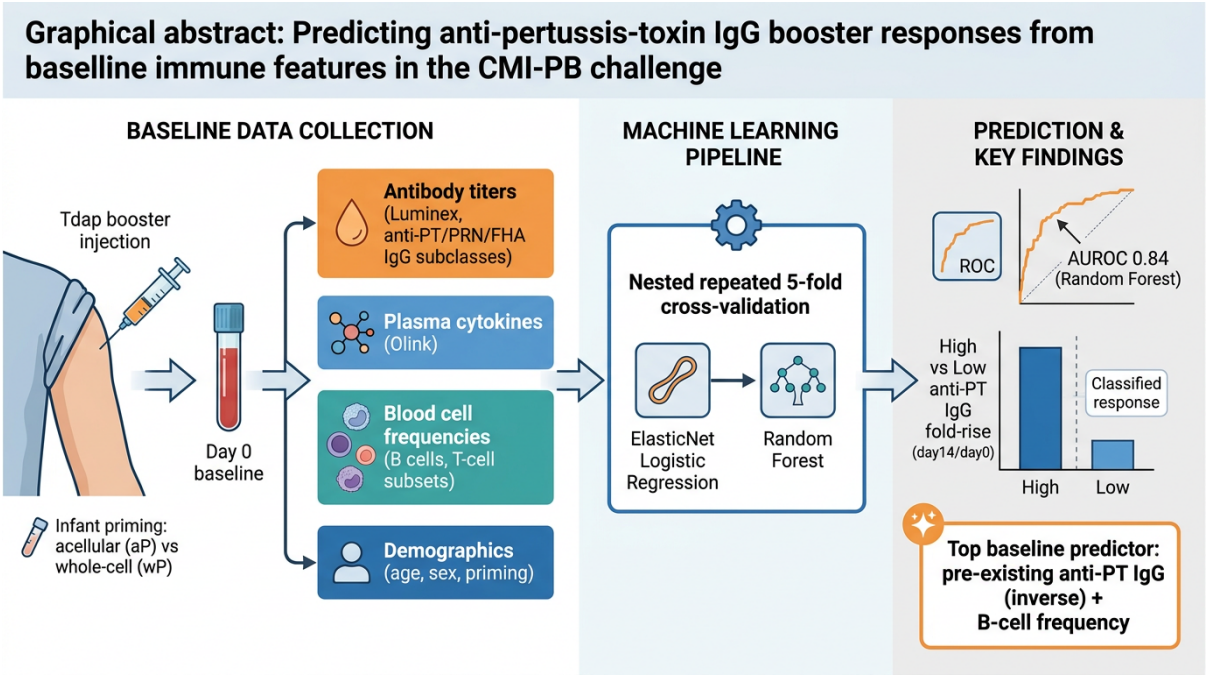


Figure 1. Graphical abstract. Baseline (day-0) antibody titers, plasma cytokines, blood cell frequencies, and demographics from adults receiving a Tdap booster (with acellular vs. whole-cell infant-priming background recorded) are passed through a leakage-controlled nested cross-validation pipeline comparing an ElasticNet logistic regression and a Random Forest. The models predict a high- versus low anti-PT IgG fold-rise (day-14/day-0) endpoint; the strongest baseline predictor is pre-existing anti-PT IgG (inverse association) together with B-cell frequency.

1 Introduction

1.1 Pertussis resurgence and the acellular-versus-whole-cell priming legacy

Despite sustained high vaccine coverage, whooping cough (*Bordetella pertussis*) has resurged across many high-income countries over the past two decades. Two intertwined explanations dominate the literature: comparatively rapid *waning* of vaccine-induced protection and durable *immunological imprinting* imposed by the type of vaccine received in infancy [1]. The switch, in the 1990s, from reactogenic whole-cell (wP) vaccines to better-tolerated acellular (aP) formulations coincided with this resurgence, and controlled follow-up showed that aP-conferred protection wanes faster than wP-conferred protection [1]. Crucially, the effect is not merely quantitative: infant priming leaves a qualitative signature on all later boosters. Children and adults primed with aP display persistently elevated IgG4 subclass responses and a more Th2-skewed recall, whereas wP priming preserves a broader Th1/Th17-biased phenotype and, often, stronger PT-specific responses [2, 3]. Because pertussis toxin (PT) is the antigen most tightly linked to protection—higher post-vaccination anti-PT IgG correlates with lower subsequent disease risk [4]—understanding what governs the *anti-PT IgG recall* after a Tdap booster is of direct translational interest.

1.2 Systems vaccinology and baseline prediction of vaccine response

A central theme of modern systems vaccinology is that the pre-vaccination immune state is not a blank slate: it measurably constrains how strongly an individual will respond. Landmark studies established that early or baseline transcriptional and cellular signatures forecast antibody

magnitude for yellow fever [5] and influenza [6], and a pan-vaccine meta-analysis of thirteen cohorts defined reproducible pre-vaccination "innate immune endotypes"—in particular an inflammatory, NF-κB-driven endotype—that predict antibody responses across platforms [7]. A recurrent and quantitatively important phenomenon in this literature is the *ceiling effect*: when response is expressed as a fold-rise, higher baseline antibody levels mechanically depress the achievable fold-change, so pre-existing titer is frequently the strongest single (inverse) predictor of fold-rise [8]. Any honest model of booster "responsiveness" must therefore both exploit and be interpreted in light of this baseline dependency.

1.3 The CMI-PB resource and the prediction task

The Computational Models of Immunity–Pertussis Boosting (CMI-PB) resource was created to turn these questions into a reproducible, community benchmark. It assembles dense, longitudinal, multi-modal measurements from adults receiving a Tdap booster—bulk PBMC RNA-seq, plasma cytokines (Olink), PBMC cell-subset frequencies, and multiplex Luminex antibody titers against PT, PRN, FHA, FIM2/3, and the diphtheria/tetanus control antigens—sampled at multiple pre-vaccination days (-30, -14, 0) and post-booster days (1, 3, 7, 14), with the aP/wP infant-priming background recorded for every subject [9]. The 3rd public prediction challenge (2024 edition) comprises $\sim$117 subjects across multiple study years and was released within the CMI-PB `cmi-pb-3rd-public-challenge-data-prep` resource; one of its core tasks is to predict the day-14 anti-PT IgG response, and the fold change relative to baseline, from the pre-vaccination state [9]. The resource is aligned with the broader HIPC/ImmPort Immune Signatures ecosystem [10], and the challenge organizers explicitly note that models must handle cross-year assay differences via harmonization and batch-aware preprocessing [9].

1.4 Objectives and scope

This report documents a self-contained analysis of the CMI-PB 3rd-challenge data with four deliberate constraints motivated by interpretability and reproducibility:

1. **Baseline only.** Predictors are restricted to day-0 measurements.
2. **Non-transcriptomic only.** We deliberately exclude the RNA-seq block, retaining antibody titers, Olink cytokines, cell frequencies, and demographics. This isolates the signal carried by clinically tractable, assay-portable readouts.
3. **A clearly defined, balanced endpoint.** A high- versus low anti-PT IgG fold-rise label dichotomized at the cohort median (Section 2.3).
4. **Leakage-free evaluation.** A *nested* repeated cross-validation design so that no aspect of preprocessing or hyperparameter selection can inflate the reported generalization estimates [11, 12].

We report AUROC and AUPRC against the observed prevalence, identify the baseline features that carry the most predictive signal, interpret them biologically (including the impact of infant priming), and release per-subject predictions and code.

2 Methods

2.1 Dataset, cohort, and provenance

Raw long-format tables were obtained from the official CMI-PB downloads server (3rd challenge, 2024 build) comprising training years `2020LD`, `2021LD`, `2022LD` (which contain day-14 outcomes)

and the challenge year `2023BD` (day-0 features only; the day-14 anti-PT IgG target is withheld). All tables are keyed by `specimen_id` and linked to subjects and time points through the `subject` and `specimen` tables. The analysis was restricted to adults (age at boost $\geq$ 18 years), a criterion satisfied by all subjects. In total 172 adults were available: 118 training (2020LD = 60, 2021LD = 36, 2022LD = 22) and 54 challenge (2023BD). Table 1 summarizes the cohort. Assay panels differed across years—2020LD used a broader Luminex/Olink panel and lacked the LEGENDplex cytokine assay entirely—which directly shaped the harmonization strategy (Section 2.4).

Table 1. Cohort composition of the CMI-PB 3rd public challenge (adults only). Training years carry the day-14 outcome; the 2023BD challenge year has baseline features only.

Year	Role	n subjects	aP / wP	F / M	Mean age (range)
2020LD	Training	60	28 / 32	42 / 18	25.1 (18–51)
2021LD	Training	36	19 / 17	24 / 12	25.2 (19–47)
2022LD	Training	22	13 / 9	13 / 9	25.2 (18–35)
2023BD	Challenge	54	27 / 27	33 / 21	26.6 (19–36)
Total	—	**172**	87 / 85	112 / 60	—

2.2 Baseline and post-boost specimen alignment

Baseline was defined as planned day 0 and the post-boost outcome as planned day 14. For each subject and target day, the specimen with the smallest absolute deviation between the *actual* and *planned* day relative to boost was selected (ties broken by the smallest `specimen_id`). Anti-PT IgG was extracted at both time points from the Luminex table (`isotype` = IgG, `antigen` = PT) as both raw MFI and per-antigen normalized MFI, and per-subject missingness at day 0 and day 14 was recorded.

2.3 Response endpoint definition

The primary quantity is the normalized anti-PT IgG fold-rise,

$$\text{fold\_rise\_ratio} = \frac{\text{anti-PT IgG}_{\text{norm}}(\text{day 14})}{\text{anti-PT IgG}_{\text{norm}}(\text{day 0})}. \tag{1}$$

The binary endpoint is

$$\text{responder} = \begin{cases} 1 \text{ (high)} & \text{if fold\_rise\_ratio} > m^*, \\ 0 \text{ (low)} & \text{if fold\_rise\_ratio} \leq m^*, \end{cases} \tag{2}$$

where m^* is the *cohort median* of the fold-rise ratio computed over all labeled training subjects. Subjects exactly at the median are assigned to the low (0) class by the $\leq$ rule.

Handling of subjects lacking either time point. Starting from the 118 training subjects, we excluded any subject lacking a valid normalized anti-PT IgG measurement at either day 0 or day 14 (6 missing day 0, 7 missing day 14; 7 unique subjects overall: IDs 2, 8, 37, 82, 87, 88, 113). These 7 subjects cannot receive a fold-rise label and are removed from the labeled modeling set; they are neither imputed nor used as training examples. This leaves **111 complete-case training subjects**, all with valid non-zero day-0 values (no infinite or undefined ratios). The resulting median threshold is $m^* = 4.2387$ (equivalently $\log_2 m^* = 2.0836$), yielding a near-balanced split of **55 high** and **56 low** responders (one subject lay exactly at the median and was assigned to low).

Sensitivity to the normalization choice. We repeated the labeling using the *raw* MFI ratio. Because per-antigen normalization is a strictly monotonic rescaling that cancels in a within-subject ratio, the raw-MFI median (4.2387) and the induced labels were identical to the normalized version (label agreement = 1.0; Spearman ρ between the two ratios = 1.0). The endpoint is therefore robust to this preprocessing decision.

2.4 Baseline feature engineering and cross-year harmonization

Because assay panels diverged across study years, we harmonized to the *intersection* feature set—per assay, only features measured in *every* cohort (2020LD/2021LD/2022LD/2023BD) were retained—so that identical columns are populated for training and challenge subjects. The LEGENDplex cytokine block was excluded entirely because it is absent from 2020LD (> 50% of training subjects), and retaining it would have required imputing an entire assay block for half the training set. The final feature space contains $K = 84$ day-0 features in four blocks (Table 2).

Table 2. Final harmonized baseline (day-0) feature set ($K = 84$). All antibody titers use per-antigen normalized MFI; Olink and cell frequencies use the cross-year intersection panels.

Block	Prefix	#	Content / encoding
Demographics	`DEMO_`	3	`age_at_boost` (continuous); `biological_sex_female` (F= 1/M= 0); `infancy_vac_wP` (wP= 1/aP= 0)
Antibody titers	`AB_`	31	Luminex `isotype_antigen` normalized MFI (IgG, IgG1–IgG4 × PT, PRN, FHA, FIM2/3, DT, TT, OVA)
Olink cytokines	`OLINK_`	30	Plasma protein concentrations (UniProt-keyed)
Cell frequencies	`CELL_`	20	PBMC population frequencies (% live cells)

2.5 Missing-data handling and leakage control

Antibody titers and demographics had no missing values in either cohort. The Olink and cell-frequency blocks exhibited *block-level* missingness (∼37% of training rows): a subject was either assayed for a whole panel or not at all, rather than having scattered random gaps. All imputation was performed with a median imputer and all standardization with a z-score scaler; crucially, in every cross-validation fold these transformers were fit *only* on that fold's training split and then applied to the held-out split (Section 2.7). The exported full-training and challenge matrices used train-only medians, so no information from the challenge cohort ever influenced fill values.

2.6 Predictive models

Two complementary classifiers were evaluated, each wrapped in a scikit-learn `Pipeline` of `SimpleImputer(median)` → `StandardScaler` → `classifier`:

- **ElasticNet logistic regression** (`penalty=elasticnet`, `solver=saga`), which yields a sparse, linear, sign-interpretable model and handles correlated antibody features gracefully [13]. Tuning grid: inverse-regularization $C \in \{0.01, 0.03, 0.1, 0.3, 1, 3\}$ and ℓ_1-ratio $\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.
- **Random Forest** [14], a nonlinear ensemble able to capture interactions and threshold effects. Tuning grid: `n_estimators`$\in \{300, 600\}$, `max_depth`$\in \{3, 5, \text{None}\}$, `min_samples_split`$\in \{2, 5, 10\}$.

The Random Forest also stands in for the gradient-boosted-tree family of models requested in the analysis brief; it was chosen as the representative tree ensemble because, on this small ($n = 111$), block-missing dataset, bagged forests are markedly less prone to over-fitting than boosting under an identical tuning budget.

2.7 Nested repeated cross-validation

A naive protocol that selects hyperparameters on all data and then reports cross-validated performance under those fixed settings leaks the model-selection choice into the evaluation and yields optimistically biased estimates [11, 12]. To avoid this we used a fully *nested* design:

- **Outer loop:** `RepeatedStratifiedKFold` with 5 splits × 10 repeats = **50 outer folds** (`random_state`= 42), preserving the high/low ratio in each fold and averaging over partition randomness.
- **Inner loop:** within each outer training split, a separate `StratifiedKFold` 5-fold `GridSearchCV` (scored by AUROC) selected the hyperparameters. Imputation and scaling were re-fit inside every inner and outer training split.
- **Evaluation:** the inner-selected model was refit on the outer training split and scored once on the untouched outer test fold. Hyperparameter selection is thus fully isolated from every held-out evaluation fold, so the reported AUROC/AUPRC are unbiased generalization estimates.

Discrimination was summarized by the area under the receiver-operating-characteristic curve (AUROC) and, because AUPRC is more informative than AUROC when the positive class is a meaningful minority and is directly interpretable against prevalence, the area under the precision–recall curve (AUPRC) [15]. Both are reported as mean ± SD across the 50 outer folds and compared to the trivial baselines (AUROC = 0.5; AUPRC = positive prevalence).

2.8 Feature importance and final (deployed) models

For interpretation, each model was additionally refit on all 111 training subjects. We report Random Forest Gini (mean decrease in impurity) importances and the signed ElasticNet coefficients (on standardized features). The distribution of inner-loop–selected hyperparameters across the 50 outer folds was tabulated as a stability check. These full-data refits are also the models deployed to score the 54 held-out challenge subjects; their full-data inner-CV scores are recorded for reference only and are *not* the headline performance estimates.

2.9 Software and reproducibility

Analyses used Python 3.12 with scikit-learn (≥1.5), pandas, and numpy [16], with a fixed random seed (42). The end-to-end workflow (`01_download_data.py` → `07_model_training.py`), the processed feature matrices, the per-fold metrics (`model_evaluation.json`), and the per-subject challenge predictions are released with this report.

3 Results

3.1 Cohort and endpoint

Of the 118 training subjects, 111 had complete day-0 and day-14 anti-PT IgG and were labeled (Section 2.3). The fold-rise distribution is broad and right-shifted (Figure 2): the median subject

boosted ∼4.2-fold, and the $\log_2$ fold-rise spans roughly −1.5 to +6, i.e. from a slight decline to a > 60-fold rise. The great majority of subjects sit above the "no change" line, confirming that a Tdap booster reliably recalls anti-PT IgG, but the *magnitude* is highly variable—exactly the variation the model must explain. The median split produced a near-perfectly balanced endpoint (55 high vs. 56 low; positive prevalence = 49.5%).

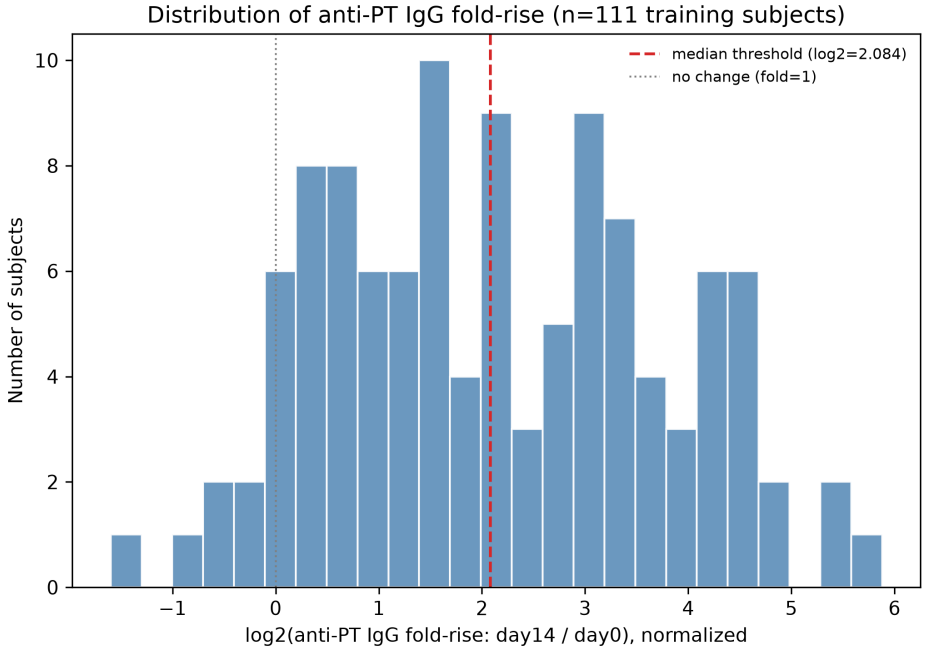
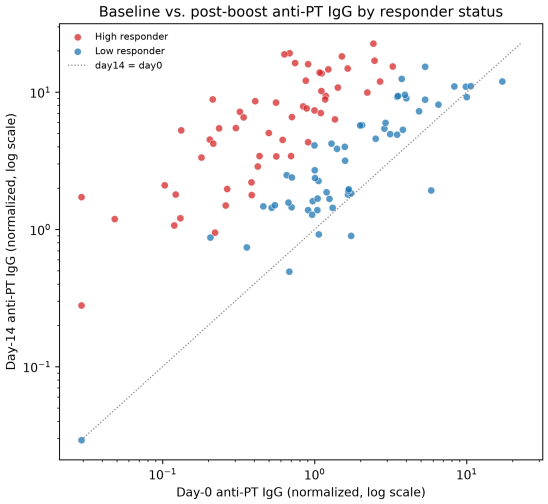
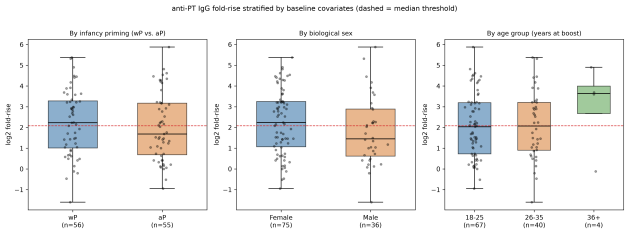


Figure 2. Distribution of the anti-PT IgG fold-rise ($\log_2$ day-14/day-0) across the 111 labeled training subjects. The red dashed line marks the cohort-median dichotomization threshold ($\log_2 m^* = 2.084$; $m^* = 4.239$); the dotted grey line marks no change (fold = 1).

The baseline-versus-post-boost scatter (Figure 3a) makes the ceiling effect visually explicit: high responders (red) cluster at *low* day-0 titers and rise steeply above the identity line, whereas low responders (blue) tend to start with higher baseline titers and change little. Endpoint associations with recorded covariates were modest but in the expected directions (Figure 3b): wP-primed subjects had a higher median $\log_2$ fold-rise (2.23 vs. 1.69 for aP) and a larger high-responder fraction (53.6% vs. 45.5%); females responded somewhat more strongly than males (53.3% vs. 41.7% high responders); and the small oldest age stratum (36+, $n = 4$) responded most strongly.



(a) Baseline vs. post-boost anti-PT IgG.

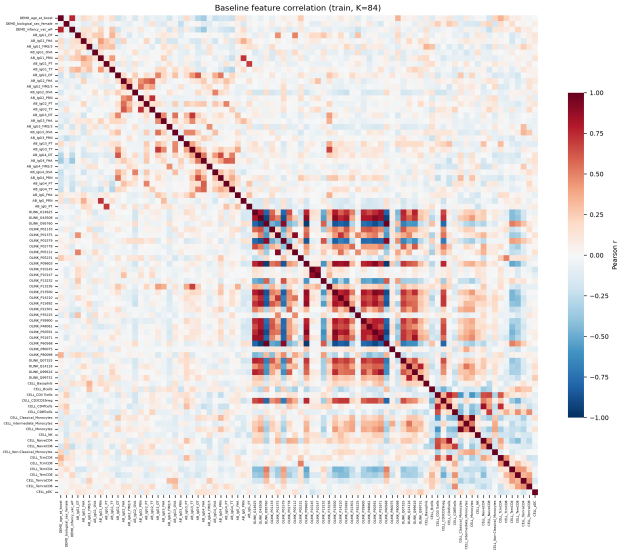


(b) Fold-rise stratified by baseline covariates.

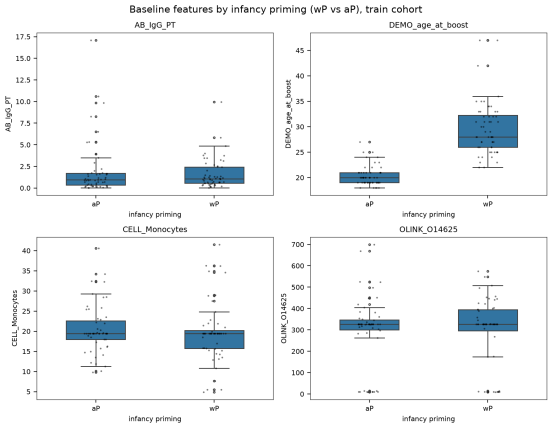
Figure 3. Endpoint behavior. (a) Day-14 vs. day-0 normalized anti-PT IgG on log–log axes, colored by responder status; high responders start low and rise steeply (ceiling effect). (b) $\log_2$ fold-rise by infant priming (wP vs. aP), biological sex, and age group; the red dashed line is the median threshold.

3.2 Feature space

The 84 harmonized baseline features form four biologically coherent blocks (Table 2). The correlation structure (Figure 4a) shows the expected within-isotype and within–cell-lineage correlation blocks (e.g. IgG subclasses against the same antigen; related T-cell subsets), motivating a regularized linear model alongside the tree ensemble. Distributions of representative features by infant priming (Figure 4b) reproduce known biology: wP-primed subjects were, as expected from the cohort design, older at boost, and baseline anti-PT IgG, monocyte frequency, and the Olink analyte CXCL11 (O14625) showed only modest priming-related shifts at baseline.



(a) Baseline feature correlation (84×84).



(b) Key features by infant priming.

Figure 4. Baseline feature space. (a) Pearson correlation heatmap of the 84 day-0 features (training cohort), revealing correlated antibody-subclass and cell-lineage blocks. (b) Distributions of anti-PT IgG, age at boost, monocyte frequency, and Olink CXCL11 stratified by wP vs. aP priming.

3.3 Cross-validated discrimination

Under the nested repeated cross-validation, both models discriminated high from low responders substantially above chance and above the 49.5% prevalence baseline (Table 3; Figures 5, 6). The Random Forest was the stronger baseline-feature model, reaching **AUROC** $= 0.841 \pm 0.081$ and **AUPRC** $= 0.838 \pm 0.087$. The ElasticNet logistic regression reached **AUROC** $= 0.771 \pm 0.074$ and **AUPRC** $= 0.741 \pm 0.091$. The AUPRC values are the most policy-relevant summary because they are directly comparable to the 0.495 no-skill line: the Random Forest improves precision-at-recall by roughly +0.34 AUPRC over that baseline, and the logistic model by roughly +0.25. The precision–recall curves (Figure 6) confirm that both models maintain high precision ($\gtrsim$0.8) across a wide recall range before degrading at very high recall, as expected for a balanced but modestly sized cohort.

Table 3. Leakage-free predictive performance (nested 10× repeated stratified 5-fold CV; 50 outer folds; mean ± SD). Positive (high-responder) prevalence = 49.5%, so the no-skill AUROC = 0.500 and no-skill AUPRC = 0.495.

Model	AUROC	AUPRC	ΔAUPRC vs. prevalence	Selected hyperparamete
No-skill baseline	0.500	0.495	—	—
ElasticNet logistic reg.	0.771 ± 0.074	0.741 ± 0.091	+0.246	$C = 3.0$, ℓ_1-ratio = 0.9
Random Forest	$\mathbf{0.841 \pm 0.081}$	$\mathbf{0.838 \pm 0.087}$	**+0.343**	n=600, depth=∞, split=

[†]Hyperparameters of the final full-data refit; the reported AUROC/AUPRC are from nested CV and do not depend on any single hyperparameter choice.

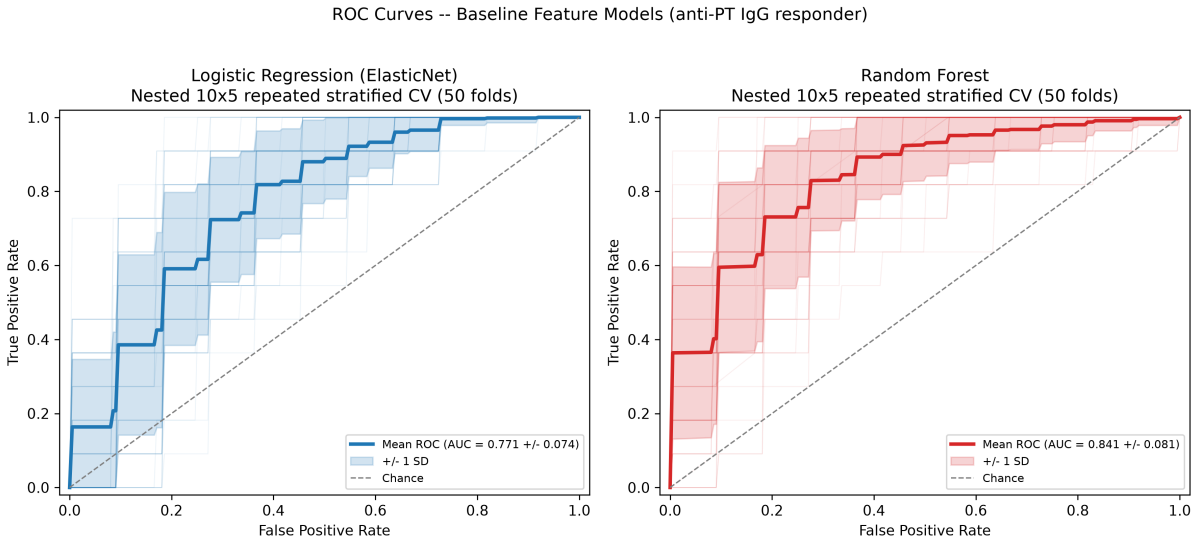


Figure 5. Receiver-operating-characteristic curves for the ElasticNet logistic regression (left) and Random Forest (right) under nested 10×5 repeated stratified CV. Thin lines are the 50 individual outer folds; the bold line is the mean ROC with a ±1 SD band; the dashed diagonal is chance.

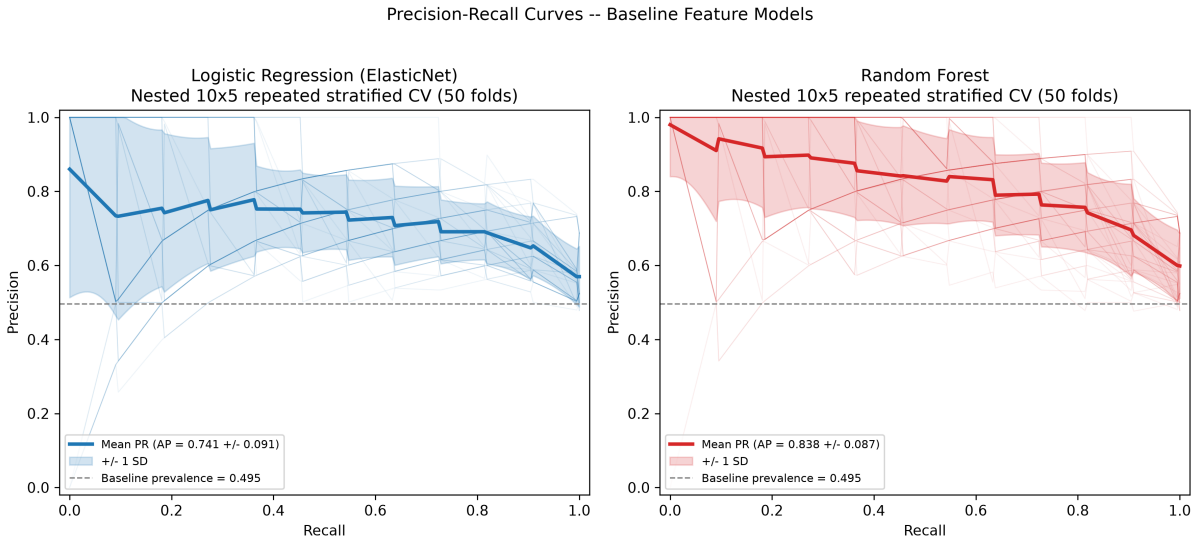


Figure 6. Precision–recall curves for both models under the same nested CV. The dashed horizontal line marks the 0.495 positive-class prevalence (no-skill AUPRC). Both models operate well above this line across most of the recall range.

3.4 Baseline features carrying the most signal

The two models agreed strikingly on which baseline features matter (Figure 7; Table 4). **Pre-existing baseline anti-PT IgG (`AB_IgG_PT`) was the single dominant predictor in both models**—the top Random Forest Gini feature (importance 0.105, more than double the next feature) and the largest-magnitude ElasticNet coefficient (−3.13). Its *negative* sign is the decisive interpretive result: subjects with higher baseline anti-PT IgG are predicted to be *low* fold-rise responders, the ceiling effect quantified. Anti-PT IgG1 (`AB_IgG1_PT`) ranked second in the forest, reinforcing the PT-titer signal.

Beyond PT titers, a consistent cellular signal emerged: baseline **B-cell frequency (`CELL_Bcells`)** was a top-4 feature in both models and carried a *positive* logistic coefficient (+1.66), i.e. more circulating B cells at baseline predicts a stronger recall. The logistic model further up-weighted central-memory CD4 (`CELL_TcmCD4`, +1.39) and effector-memory CD8 (`CELL_TemCD8`, +1.29) T-cell frequencies. Additional antibody features against other pertussis antigens (IgG3/IgG1 against PRN and FIM2/3, IgG2 against FHA/FIM2/3) contributed, as did a small set of Olink cytokines—most notably IL-4 (P05112), IL-10 (P22301), and CCL3/MIP-1α (P10147). Demographic variables (age, sex, priming) carried little independent weight once the molecular and cellular features were present (the ElasticNet shrank the priming and age coefficients toward zero), indicating that priming acts largely *through* the measured immune features rather than as an independent axis.

Table 4. Top baseline predictors of the high anti-PT IgG fold-rise endpoint. Left: Random Forest Gini importance. Right: signed ElasticNet logistic coefficients (standardized features); negative = predicts *low* responder. Olink UniProt IDs annotated with common protein names.

	Random Forest (Gini)			**ElasticNet logistic (coefficient)**	
#	Feature	Import.	#	Feature	Coef.
1	`AB_IgG_PT`	0.105	1	`AB_IgG_PT`	−3.13
2	`AB_IgG1_PT`	0.044	2	`AB_IgG3_FIM2/3`	−2.17
3	`AB_IgG2_FHA`	0.037	3	`CELL_Bcells`	+1.66
4	`CELL_Bcells`	0.035	4	`AB_IgG3_PRN`	−1.43
5	`AB_IgG3_PRN`	0.033	5	`CELL_TcmCD4`	+1.39
6	`AB_IgG3_FIM2/3`	0.029	6	`CELL_TemCD8`	+1.29
7	`AB_IgG4_PRN`	0.025	7	`AB_IgG1_PRN`	−1.22
8	`AB_IgG1_PRN`	0.023	8	`AB_IgG2_FIM2/3`	+1.21
9	`AB_IgG_FHA`	0.023	9	`AB_IgG_PRN`	+1.20
10	`AB_IgG4_PT`	0.022	10	`OLINK_P05112` (IL-4)	−1.20

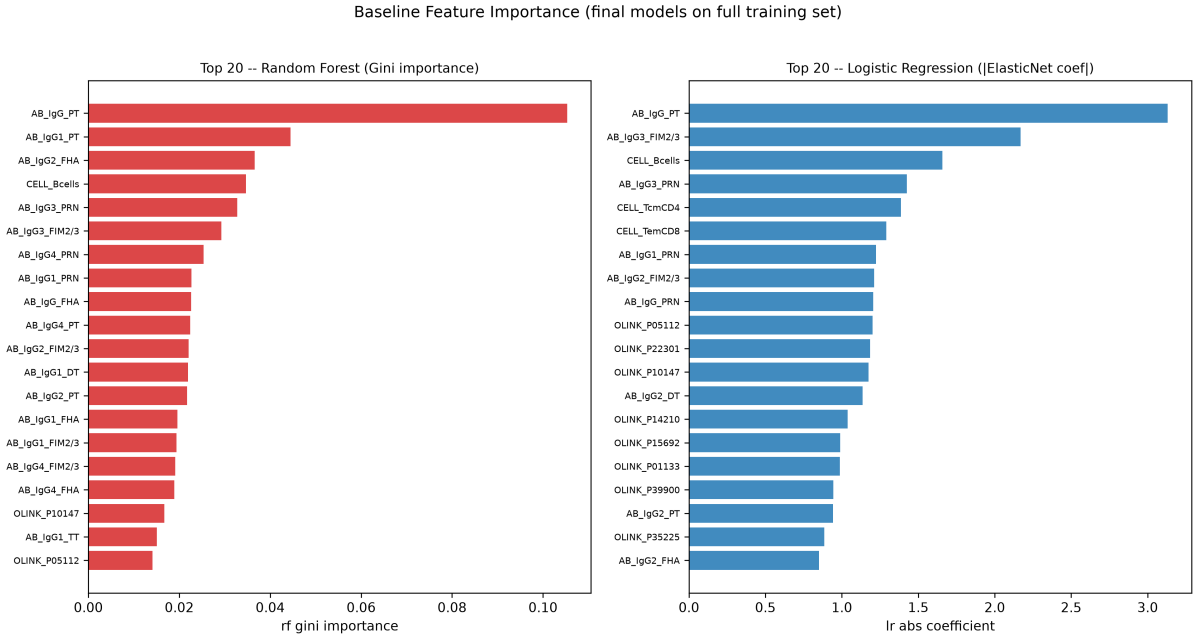


Figure 7. Baseline feature importance from the final models refit on all 111 training subjects. Left: top-20 Random Forest Gini importances. Right: top-20 ElasticNet logistic coefficients by absolute value. Anti-PT IgG dominates both rankings, followed by other pertussis-antigen titers, B-cell and memory T-cell frequencies, and selected Olink cytokines.

3.5 Hyperparameter-selection stability

Because the nested design re-selects hyperparameters in every outer fold, we inspected the stability of those choices. For the Random Forest, deep forests with 300–600 trees dominated (the most frequent setting, `max_depth=None, min_samples_split=2, n_estimators=300`, was chosen in 16% of folds, and unbounded-depth configurations collectively in the large majority), indicating a robust preference for low-bias trees. For the logistic model, moderate-to-strong regularization with a high ℓ_1-ratio recurred most often ($C = 0.1$, ℓ_1-ratio $= 0.9$ in 16% of folds), consistent with a sparse solution dominated by a few antibody and cell features. The spread of selected settings across folds underscores precisely why nested CV is necessary: no single configuration is universally optimal, and reporting performance under one fixed choice

would misstate generalization.

3.6 Per-subject challenge predictions

Both final models were refit on all 111 labeled training subjects and applied to the 54 held-out challenge (2023BD) subjects, for whom the day-14 outcome is withheld. Per-subject high-responder probabilities are provided for every challenge subject (Appendix A; machine-readable `predictions_challenge.csv`). At a 0.5 decision threshold the ElasticNet model flags 7 of 54 subjects and the Random Forest 3 of 54 as high responders; the Random Forest probabilities are more concentrated around the middle of the range (mean 0.37) while the sparse logistic model is more polarized (mean 0.15), reflecting the different decision geometries of the two model families. Because the challenge labels are withheld by design, these predictions are released as the deliverable rather than scored here.

4 Discussion

4.1 Baseline anti-PT IgG is the dominant, inverse predictor

The clearest biological message is that the pre-vaccination anti-PT IgG level itself is the strongest determinant of the fold-rise endpoint, and it acts *inversely*: subjects who already carry high anti-PT IgG at day 0 have the least room to boost. This is the ceiling effect formalized by Wang et al. [8], in which fold-change is negatively influenced by baseline titer while absolute change can still be positive, and it recurs across influenza, SARS-CoV-2, and other vaccines. Two practical consequences follow. First, the endpoint we (and the CMI-PB challenge) predict is genuinely a measure of *boostability*, not of protective end-titer; a low fold-rise in a subject with high baseline titer is not immunologically "bad." Second, any model of this endpoint that omitted baseline titer would be both biologically incomplete and statistically weaker—which is exactly why a leakage-controlled design that keeps baseline titer in the feature set, rather than partialling it out, is appropriate for the stated task.

4.2 Cellular predictors: B-cell and memory T-cell frequencies

After PT titer, the most reproducible signal was cellular. Higher baseline *B-cell frequency* predicted stronger recall in both models—mechanistically intuitive, since the anti-PT recall response is executed by memory B cells differentiating into antibody-secreting plasmablasts, and a larger circulating B-cell compartment plausibly reflects greater recall capacity. The logistic model's positive weighting of central-memory CD4 and effector-memory CD8 frequencies echoes the systems-vaccinology finding that baseline cell-population structure predicts post-vaccination antibody responses independently of age and pre-existing titer [6]. That a purely cellular, non-transcriptomic baseline readout carries independent predictive signal is encouraging for translation, because flow-cytometric frequencies are far more portable across laboratories than RNA-seq signatures.

4.3 Cytokine predictors and the inflammatory-endotype connection

Several Olink analytes entered the top features, most notably IL-4 (P05112, negative logistic coefficient), IL-10 (P22301), and CCL3/MIP-1α (P10147). The negative association of the Th2 cytokine IL-4 with a strong PT recall is consistent with the broader observation that a Th2/IgG4-skewed baseline milieu—the phenotype associated with acellular priming—accompanies a qualitatively different, and often less robust, PT-specific response [2, 3]. The

appearance of CCL3 is notable because the same chemokine was highlighted by da Silva Antunes et al. [3] as elevated in aP-primed adults, and it is one of the CMI-PB challenge's own prediction targets. More broadly, the presence of inflammatory-axis cytokines among the predictors is congruent with the pan-vaccine "inflammatory endotype" framework, in which a baseline innate-inflammatory state modulates antibody magnitude across vaccines [7]. Our cytokine effects are, however, individually small relative to the antibody-titer signal and should be read as supporting rather than primary.

4.4 Impact of infant priming

Infant priming exerted a real but modest and largely *indirect* effect on the endpoint. Unadjusted, wP-primed adults had a higher median fold-rise and high-responder fraction than aP-primed adults (Figure 3b), consistent with the durable imprinting literature [1–3]. Yet in the multivariable models the explicit priming indicator carried almost no independent weight—the ElasticNet shrank it toward zero—while priming-associated molecular features (IL-4, IgG subclass patterns against PRN/FIM2/3) were retained. The most parsimonious interpretation is that the priming background is encoded in, and acts through, the measured baseline antibody and cytokine profile, so that once those readouts are available the categorical aP/wP label adds little. This is a useful message for biomarker development: the informative quantity is the imprinted immune *state*, not the historical vaccine label per se. A caveat is that priming and age are confounded in this cohort (wP-primed subjects are older by design), so the two covariate effects cannot be fully disentangled here.

4.5 Methodological rigor: why the nested design matters

An earlier iteration of this analysis selected hyperparameters on the full dataset and then reported repeated-CV performance under those fixed settings—a protocol that leaks the model-selection decision into the evaluation and is known to inflate performance estimates [11, 12]. The present report corrects this with a fully nested design in which imputation, scaling, and grid-search hyperparameter selection are confined to each outer fold's training split. The hyperparameter-stability analysis makes the payoff concrete: because the optimal configuration varies fold to fold, any single fixed choice would misrepresent generalization. The reported AUROC/AUPRC are therefore conservative, unbiased estimates, and the modest gap between the nested-CV scores and the full-data inner-CV reference values (e.g. RF full-data reference AUROC 0.855 vs. nested 0.841) is exactly the small optimism that the nested design removes. Reporting AUPRC alongside AUROC follows best practice for interpretability against a defined prevalence [15].

4.6 Limitations

Several limitations bound the interpretation. (i) The labeled training set is small ($n = 111$) and drawn from a few academic cohorts of predominantly young adults, limiting power to detect small-effect predictors and generalizability to other ages and settings. (ii) Cross-year assay harmonization to the intersection panel discards features available in only some years (notably the broader 2020LD panels and the entire LEGENDplex block), trading breadth for a consistent, leakage-safe feature space; residual batch effects across years may persist despite normalization, as the challenge organizers themselves emphasize [9]. (iii) The Olink and cell-frequency blocks are block-missing for $\sim$37% of subjects; median imputation is principled but attenuates those blocks' apparent importance, so their true contribution may be underestimated. (iv) The endpoint is a median-dichotomized fold-rise, a pragmatic but arbitrary cut that discards magnitude information and is, by construction, dominated by the baseline-titer ceiling effect.

(v) Feature importances are associational, not causal, and the correlated antibody-subclass structure means individual rankings within the antibody block should be read as a group signal. (vi) We evaluated a linear model and one tree ensemble; the requested gradient-boosting family is represented only by the Random Forest.

4.7 Conclusions

Restricting predictors to the baseline, non-transcriptomic immune state, a leakage-controlled machine-learning pipeline predicts high- versus low anti-PT IgG booster responsiveness in the CMI-PB 3rd public challenge with good accuracy (Random Forest AUROC 0.841, AUPRC 0.838 against a 0.495 prevalence). The predictive signal is concentrated in a compact, interpretable, and clinically portable signature: pre-existing anti-PT IgG (strongly inverse, the ceiling effect), baseline B-cell and memory T-cell frequencies (positive), and a supporting set of pertussis-antigen antibodies and inflammatory/Th2 cytokines. Infant priming background influences the endpoint chiefly through these measured features rather than as an independent axis. Per-subject predictions for the 54 held-out challenge subjects and fully reproducible code accompany this report, providing a transparent baseline-feature benchmark for the CMI-PB community.

References

[1] Nicola P. Klein, Joan Bartlett, Ali Rowhani-Rahbar, Bruce Fireman, and Roger Baxter. Waning protection after fifth dose of acellular pertussis vaccine in children. *New England Journal of Medicine*, 367(11):1012–1019, 2012. doi: 10.1056/NEJMoa1200850.

[2] Saskia van der Lee, Lotte H. Hendrikx, Elisabeth A. M. Sanders, Guy A. M. Berbers, and Anne-Marie Buisman. Whole-cell or acellular pertussis vaccination in infancy determines IgG subclass profiles to DTaP booster vaccination. *Vaccine*, 36(2):220–226, 2018. doi: 10.1016/j.vaccine.2017.11.066.

[3] Ricardo da Silva Antunes, Ferran Soldevila, Mikhail Pomaznoy, Mariana Babor, Jason Bennett, Yuan Tian, et al. A system-view of Bordetella pertussis booster vaccine responses in adults primed with whole-cell versus acellular vaccine in infancy. *JCI Insight*, 6(7): e141023, 2021. doi: 10.1172/jci.insight.141023.

[4] Jann Taranger, Birger Trollfors, Teresa Lagergård, Valter Sundh, Danuta A. Bryla, Rachel Schneerson, and John B. Robbins. Correlation between pertussis toxin IgG antibodies in postvaccination sera and subsequent protection against pertussis. *Journal of Infectious Diseases*, 181(3):1010–1013, 2000. doi: 10.1086/315292.

[5] Troy D. Querec, Rama S. Akondy, Eva K. Lee, Weiping Cao, Helder I. Nakaya, Dirk Teuwen, et al. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature Immunology*, 10(1):116–125, 2009. doi: 10.1038/ni.1704.

[6] John S. Tsang, Pamela L. Schwartzberg, Yuri Kotliarov, Angelique Biancotto, Zhi Xie, Ronald N. Germain, et al. Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell*, 157(2):499–513, 2014. doi: 10.1016/j.cell.2014.03.031.

[7] Slim Fourati, Lewis E. Tomalin, Matthew P. Mulè, Daniel G. Chawla, Bram Gerritsen, Dmitry Rychkov, et al. Pan-vaccine analysis reveals innate immune endotypes predictive of antibody responses to vaccination. *Nature Immunology*, 23(12):1777–1787, 2022. doi: 10.1038/s41590-022-01329-5.

[8] Steven S. Wang, Paul M. Fernhoff, W. Harry Hannon, and Muin J. Khoury. Evaluation of vaccine-induced antibody responses: impact of baseline antibody levels. *Clinical and Vaccine Immunology*, 20(4):407–415, 2013. doi: 10.1128/CVI.00733-12.

[9] Pramod Shinde, Lisa Willemsen, Michael Anderson, Minori Aoki, Saonli Basu, Joan Barton, et al. Putting computational models of immunity to the test—an invited challenge to predict B. pertussis vaccination responses. *PLOS Computational Biology*, 21(1):e1012927, 2025. doi: 10.1371/journal.pcbi.1012927.

[10] Joann Diray-Arce, Helen E. R. Miller, Evan Henrich, Bram Gerritsen, Matthew P. Mule, Slim Fourati, et al. The immune signatures data resource, a compendium of systems vaccinology datasets. *Scientific Data*, 9:662, 2022. doi: 10.1038/s41597-022-01714-7.

[11] Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

[12] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. doi: 10.1186/1471-2105-7-91.

[13] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

[14] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[15] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

[16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A Per-subject challenge predictions

Predicted probabilities that each of the 54 held-out challenge (2023BD) subjects is a *high* anti-PT IgG fold-rise responder, from the ElasticNet logistic regression (LR) and Random Forest (RF) models refit on all 111 labeled training subjects. Sex: F=female, M=male; priming: aP/wP infant vaccine. The full machine-readable file is `predictions_challenge_annotated.csv`.

Subject	Priming	Sex	Age	P_{high} (LR)	P_{high} (RF)
119	aP	F	23	0.001	0.418
120	wP	F	27	0.003	0.376
121	aP	F	22	0.569	0.424
122	aP	F	23	0.000	0.281
123	wP	F	26	0.386	0.336
124	aP	M	22	0.979	0.430
125	wP	M	29	0.000	0.362
126	wP	M	29	0.000	0.297

continued on next page

Subject	Priming	Sex	Age	P_{high} (LR)	P_{high} (RF)
127	aP	F	26	0.279	0.377
128	wP	F	28	0.026	0.358
129	wP	M	31	0.000	0.321
130	wP	M	26	0.000	0.311
131	aP	F	24	0.424	0.360
132	wP	M	27	0.000	0.368
133	aP	F	25	0.177	0.326
134	wP	M	32	0.028	0.412
135	wP	M	27	0.072	0.310
136	wP	F	27	0.001	0.410
137	aP	F	24	0.000	0.367
138	aP	M	22	0.000	0.306
139	wP	F	29	0.006	0.330
140	aP	F	21	0.000	0.329
141	wP	F	26	0.510	0.486
142	aP	F	31	0.075	0.568
143	aP	F	19	0.029	0.369
144	aP	F	23	0.000	0.329
145	aP	M	20	0.985	0.357
146	wP	M	31	0.000	0.334
147	aP	F	23	1.000	0.335
148	wP	M	35	0.000	0.348
149	wP	F	32	0.034	0.424
150	wP	M	32	0.007	0.402
151	wP	F	31	0.397	0.640
152	wP	F	28	0.000	0.281
153	aP	F	25	0.245	0.541
154	aP	F	26	0.003	0.408
155	aP	F	26	0.000	0.312
156	aP	F	22	0.106	0.372
157	aP	F	26	0.000	0.284
158	aP	M	23	0.006	0.253
159	wP	F	29	0.003	0.415
160	aP	M	27	0.000	0.228
161	aP	F	30	0.684	0.355
162	aP	F	24	0.001	0.325
163	wP	F	30	0.036	0.470
164	wP	M	32	0.017	0.367
165	wP	F	30	0.019	0.428
166	aP	F	22	1.000	0.397
167	aP	M	26	0.016	0.446
168	wP	M	32	0.000	0.329
169	aP	M	20	0.005	0.410
170	wP	M	31	0.046	0.364
171	wP	F	20	0.040	0.377
172	wP	M	36	0.001	0.427