nature research

https://doi.org/10.1038/s42256-020-00232-8

Supplementary information

Integration of mechanistic immunological knowledge into a machine learning pipeline improves predictions

In the format provided by the authors and unedited

nature research

https://doi.org/10.1038/s42256-020-00232-8

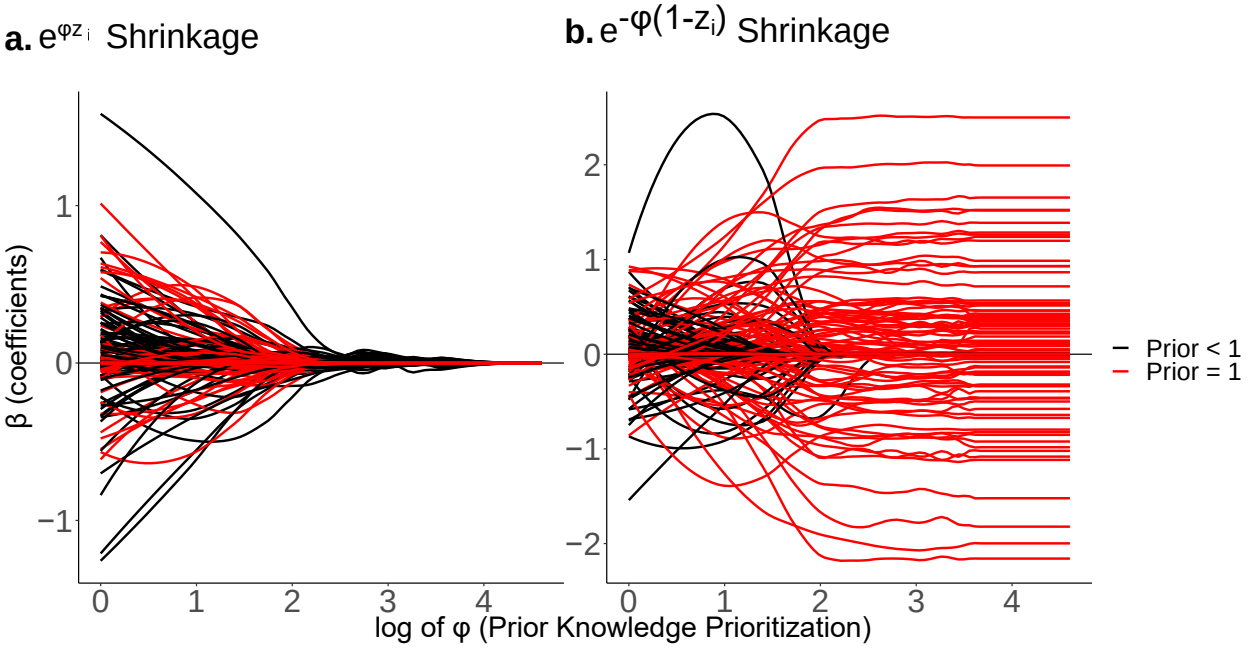
Supplementary information

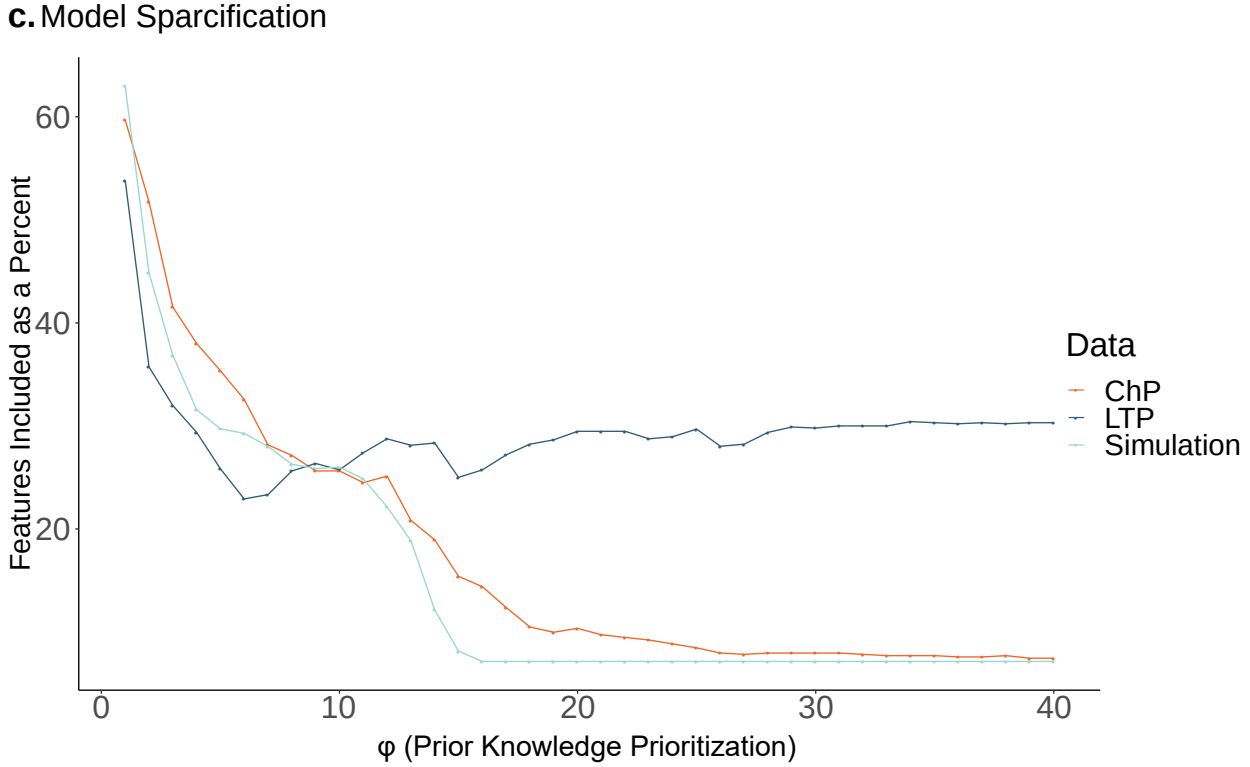
Integration of mechanistic immunological knowledge into a machine learning pipeline improves predictions

In the format provided by the authors and unedited

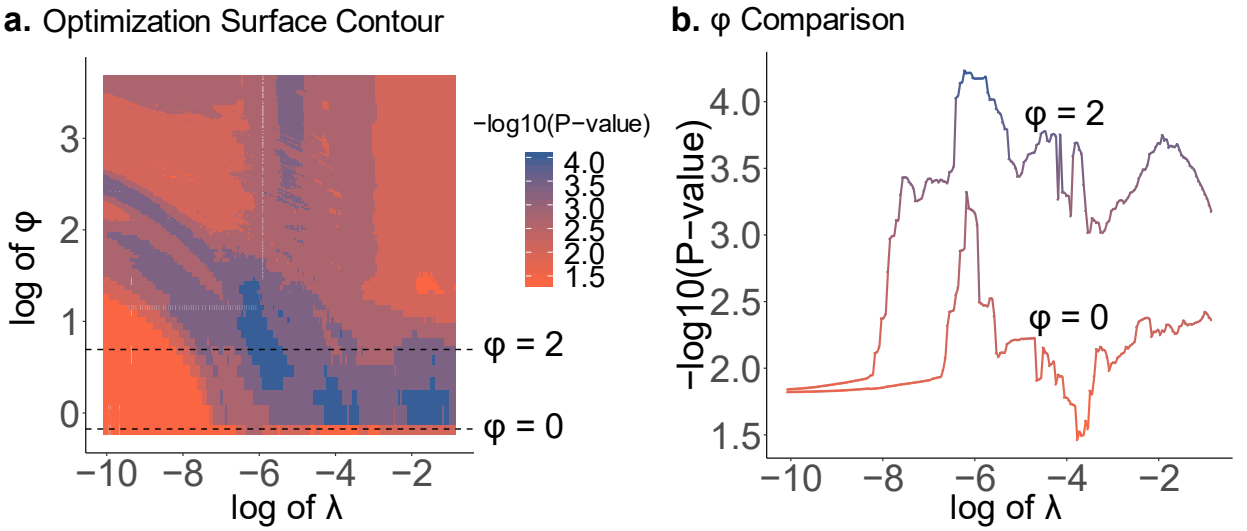
Supplemental Information

Supplemental Figures:

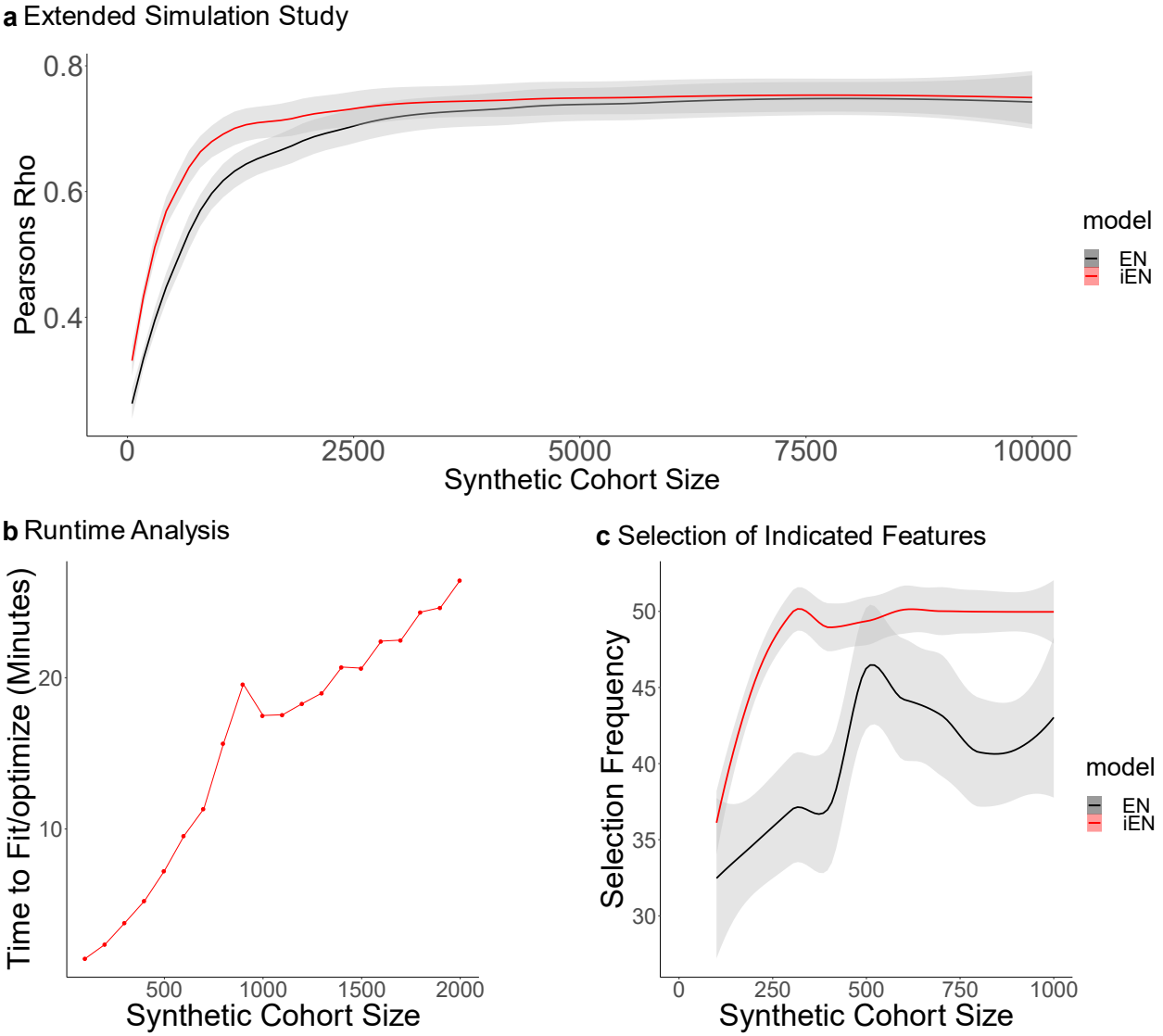




[**Supplemental Fig. 1 - Effect of feature prioritization on coefficients and model size: (a)** prior knowledge prioritization with $\varphi_{i,i} \in diag(\phi)$ as $e^{\varphi z_i}$ results in instability. This holds true for features with both high and low prior knowledge values. This instability is rooted in values truncating to infinity as ϕ increases, effectively removing all features in the order of highest prior value to lowest prior value. **(b)** Prioritization with $\varphi_{i,i} \in diag(\phi)$ as $e^{-\varphi(1-z_i)}$ results in stable progression of feature shrinkage as prioritization increases. This allows for the edge case where the model is constructed solely from features with prior value = 1. **(c)** Variation of φ to a model with fixed α and λ (α = 0.5, with λ selected such that each initial dataset had comparable model size) affects the number of features included, presented here as percent of features for each dataset. An increase of φ in all datasets follow similar trends, note that the ChP and Simulated studies truncate at approximately 7% of features included. This suggests that the simulated study reacts similarly to real mass cytometry data for evaluation of the model optimization procedure.]

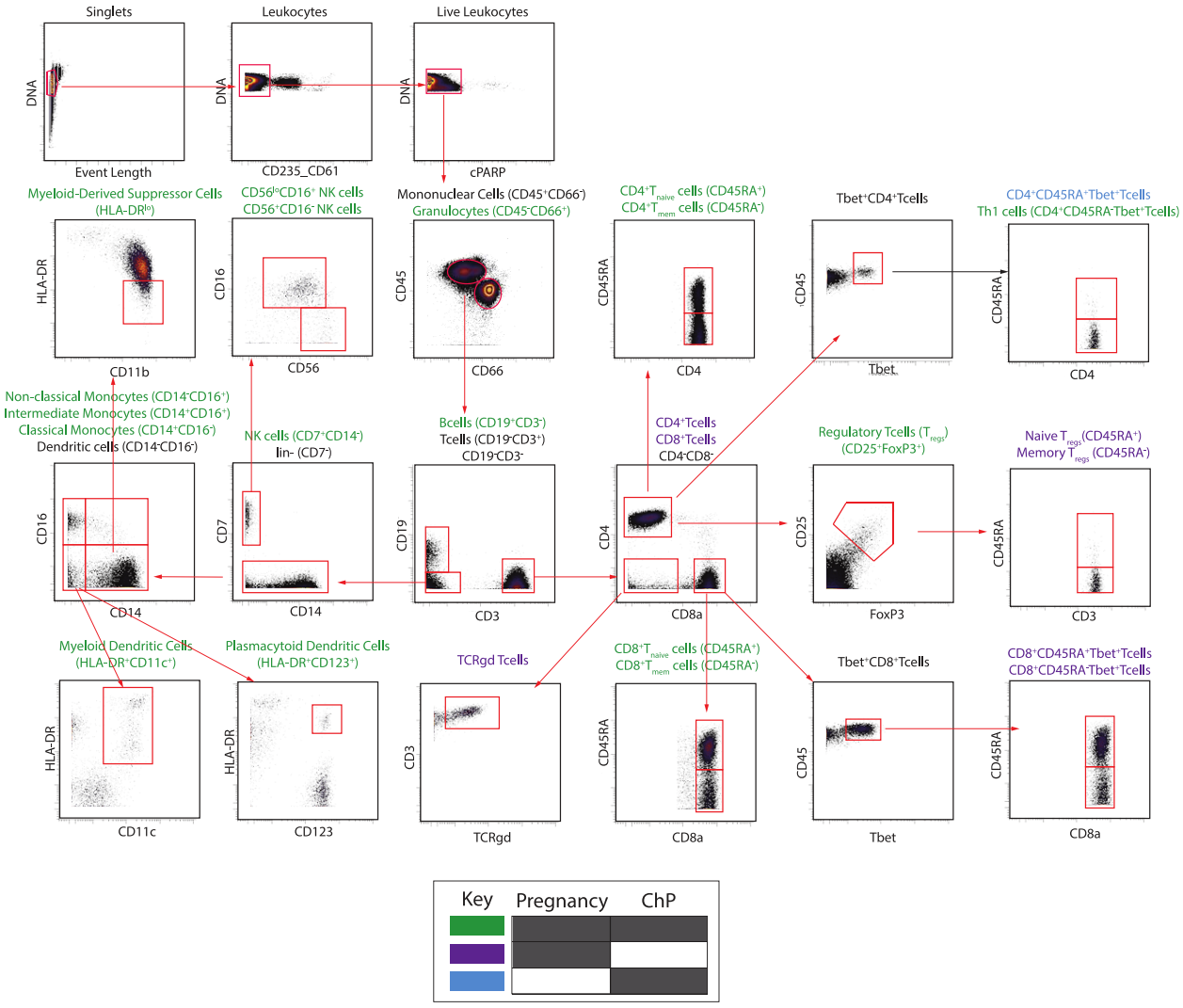


[**Supplemental Fig. 2 - Optimization landscape**: **(a)** Optimization of iEN models for the LTP study across a grid of φ and λ values (with α = 0.5) demonstrates the improvement in performance after incorporation of prior knowledge (φ = 0 is equivalent to the EN). The color gradient of this surface represents the -log10(P-value) of models trained on the original LTP data which predict the Validation dataset, for combinations of $\lambda \in$ [4.25e-5,0.425] and $\varphi \in$ [0,40]. This indicates that models fit with prioritization of expert knowledge (φ = 2) outperform the traditional EN (φ = 0). **(b)** A direct comparison of models, EN (φ = 0) vs. iEN (φ = 2) as indicated by the dashed lines in panel a, visualizing the performance difference over the λ sequence. This clearly demonstrates a robust improvement in the results when prior knowledge is integrated into the model.]

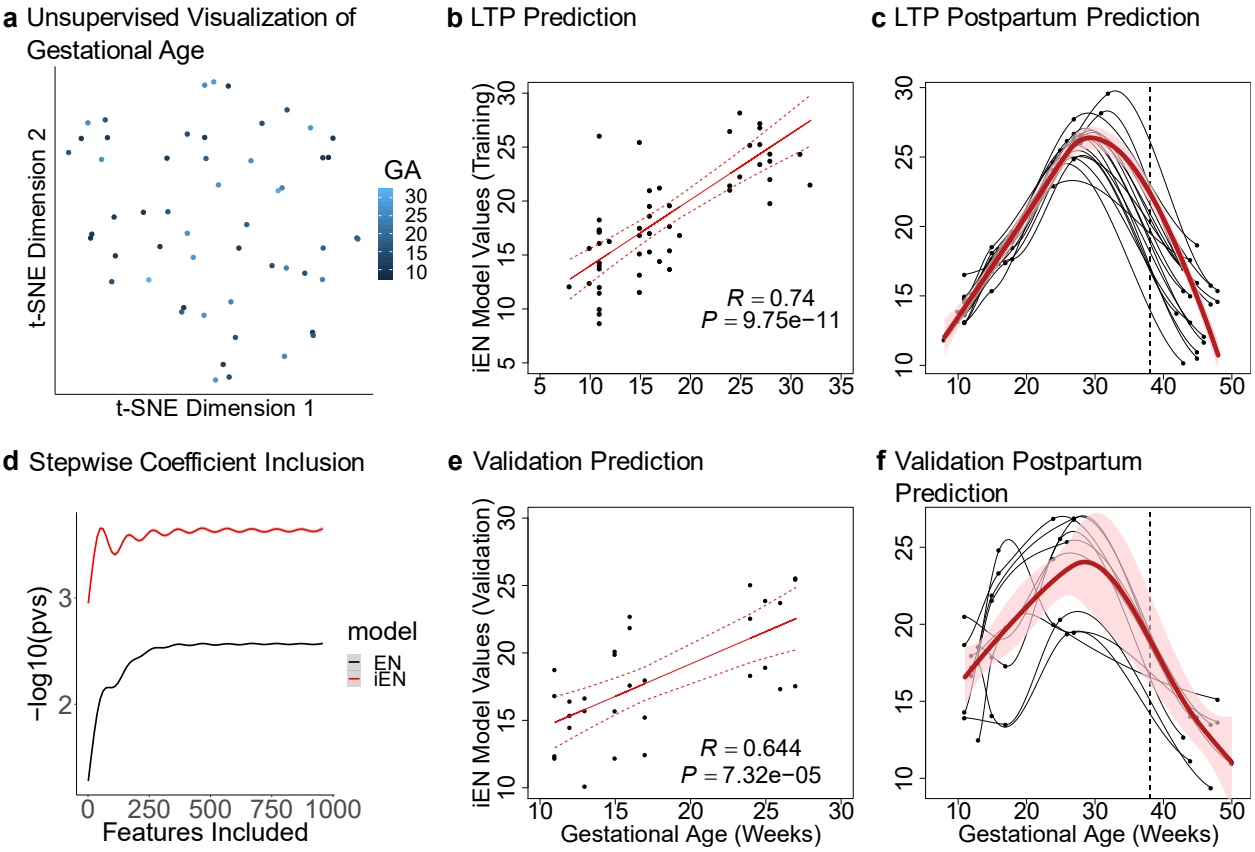


[**Supplemental Fig. 3 - Extended simulation study: (a)** Simulated data was used to compare iEN and EN across cohort sizes. Twenty simulated cohorts ranging from 50 to 10000 subjects (selected from a logarithmic sequence) were generated. Each cohort of simulated "patients" were fit 10 times using a randomized 10-fold cross-validation stratification. iEN outperformed EN in cohorts as large as 2000 patients. **(b)** Simulated data was used to determine the time requirement of the iEN algorithm with varying population sizes. This demonstrated a linear relationship between the cohort size and time to completion. Each model was optimized using the same parameter search space and 10-fold CV strategy used in the clinical studies. Here computation time for only one iteration is reported because the iEN software is parallelized over each φ and α values for high-performance computing. **(c)** Each simulation consisted of 50 highly predictive, 250 moderately predictive, and 450 vectors of uniform noise features. The y-axis visualizes the number of highly predictive features captured compared to synthetic cohort size (x-axis). These smoothed curves demonstrate that as cohort size increased, iEN captured the highly predictive features faster than EN. All smoothed curves are locally fitted polynomials with 95% CI.]

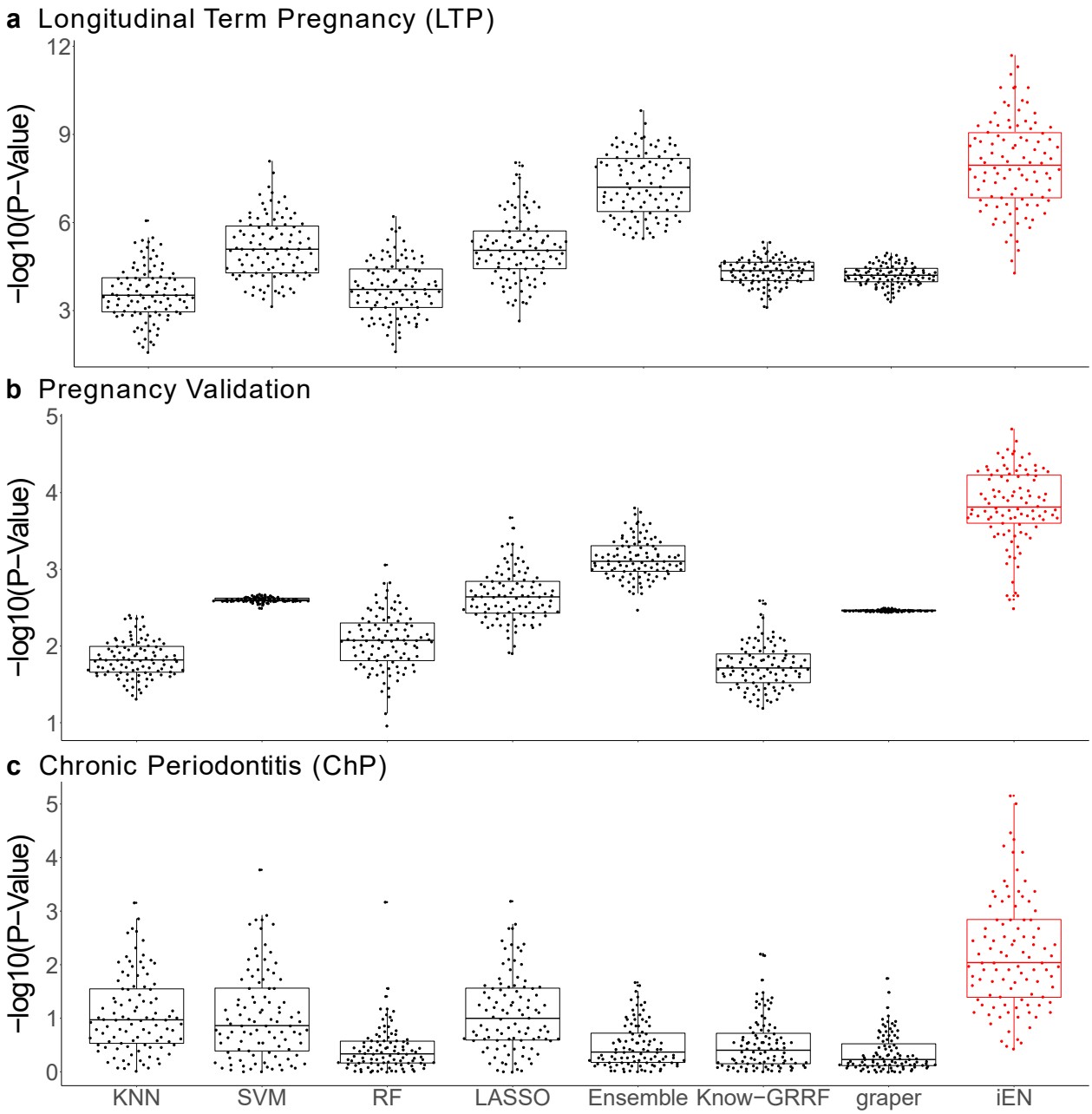
Gating Strategy



[**Supplemental Fig. 4 - Gating strategy for feature extraction from the measured single cells (combined for both clinical datasets):** Two-dimensional scatter plots shown for a representative patient sample. Gating was performed using Cytobank (www.cytobank.org) for the ChP and LTP cohorts. Populations noted in green were included in both analyses, purple populations were included in LTP analysis and blue populations were supplemented in only the ChP analysis.]

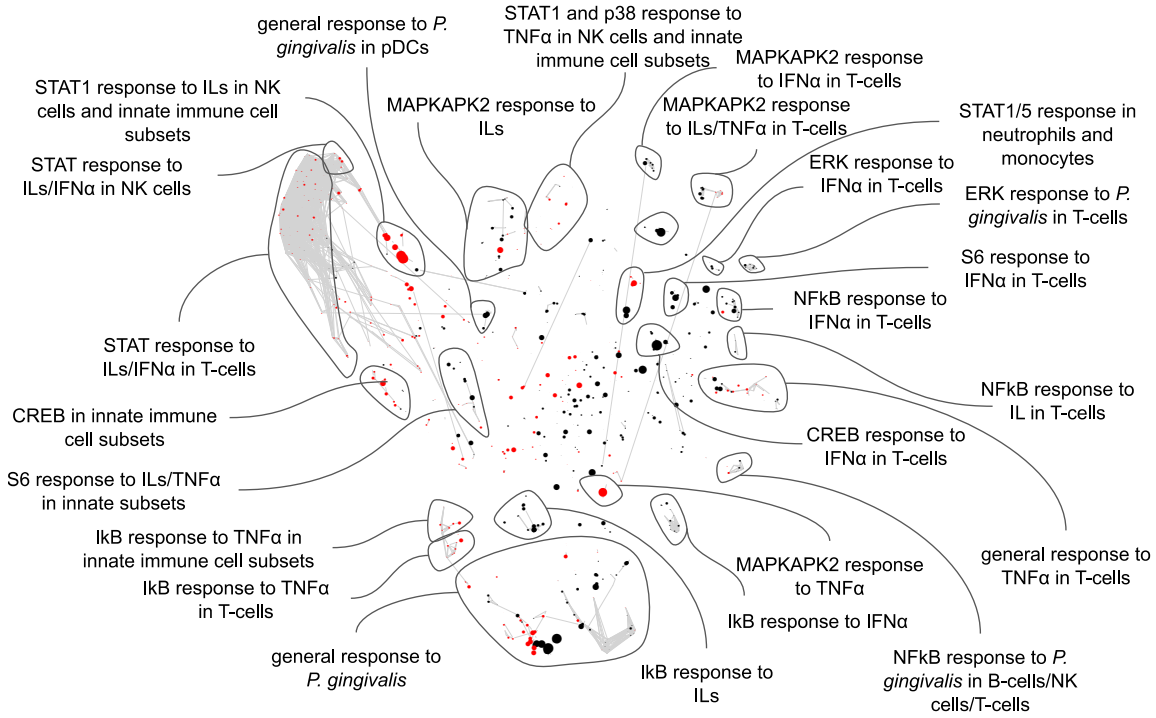


[**Supplemental Fig. 5 - Analysis of longitudinal term pregnancy**: **(a)** Unsupervised t-SNE visualization of each antepartum sample colored by gestational age showed no major connection between the measured features and gestational age, motivating supervised analysis for identification of immune features associated with the outcome. **(b)** Mean predictions from 10-fold cross-validated iEN models, as measured by adjusted R and associated P-value. These predictions also have a Spearman Rho of 0.724 with P-value of 5.814e-10. **(c)** Projecting mean coefficient models onto training data which contains postpartum samples demonstrates a return to baseline after delivery, with overlaid spline and 95% CI. **(d)** Smoothing splines of Pearson correlation p-value generated from each individual iEN and EN model, applied to the independent validation cohort with descending stepwise inclusion of features by coefficient size. **(e)** Predicted vs. actual gestational age values and adjusted R with P-value: Spearman rho for aggregate validation predictions are 0.621 with a P-value of 2.522e-4. **(f)** Similar to the training cohort, models applied to the validation cohort demonstrate a postpartum return to baseline, with overlaid locally fitted polynomial curve and 95% CI.]

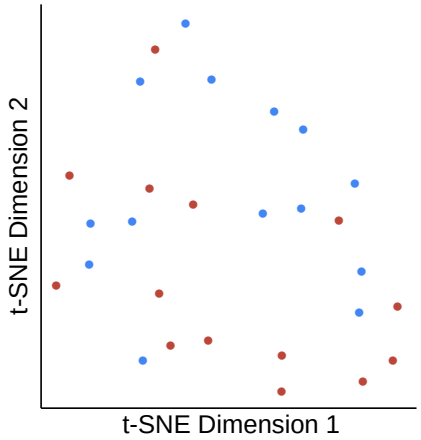


[**Supplemental Fig. 6 - Comparison against standard machine learning algorithms:** Expanded comparison of the IEN method against standard machine learning algorithms for **(a)** LTP**, (b)** LTP validation, and **(c)** ChP analysis. All algorithms were trained, tested, and optimized using a similar nested 10-fold CV approach.]

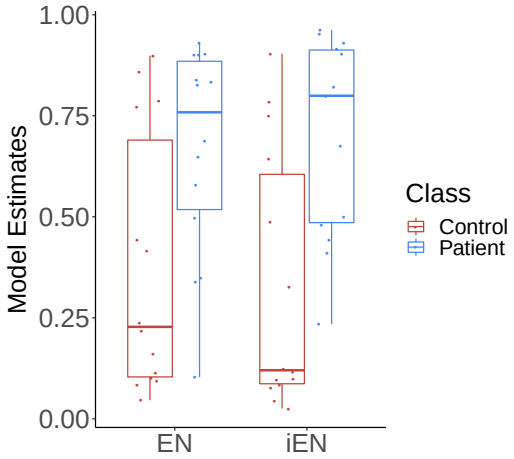
a Prior Tensor of ChP Features



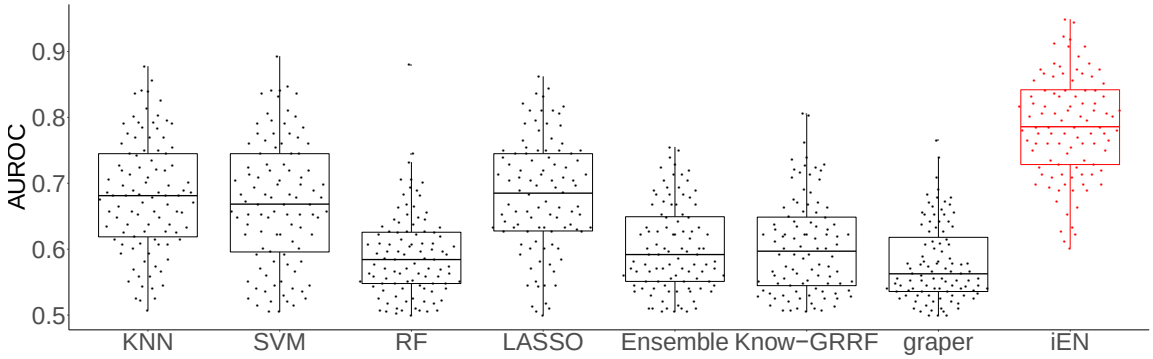
b Unsupervised Visualization of Case/Control Distributions



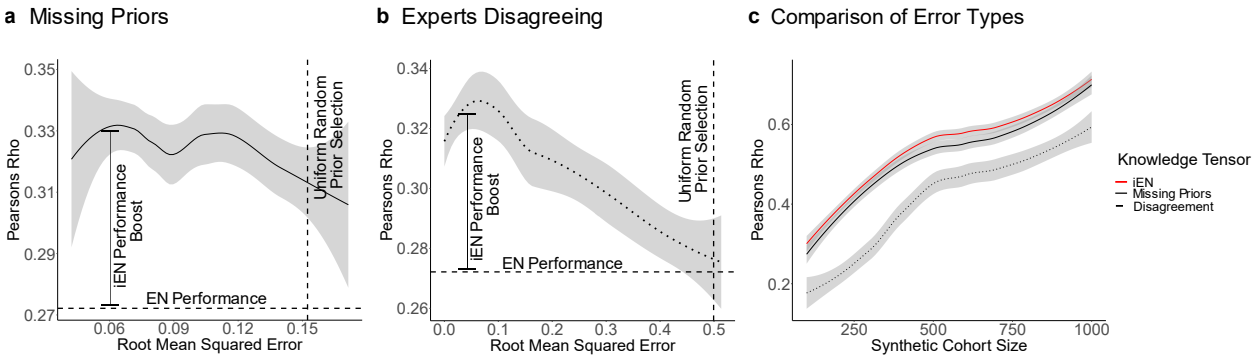
c Aggregate Class Predictions



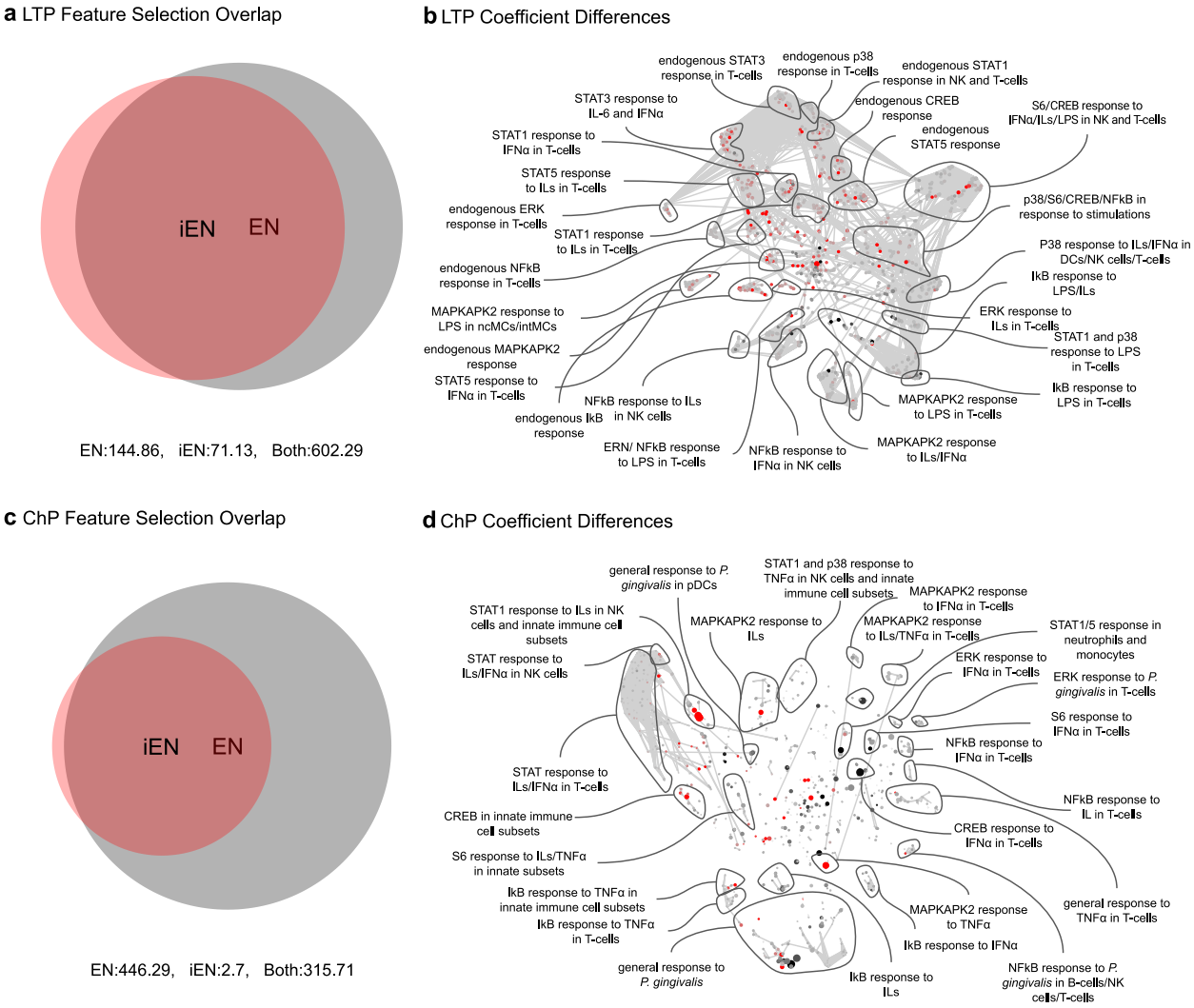
d Model Results



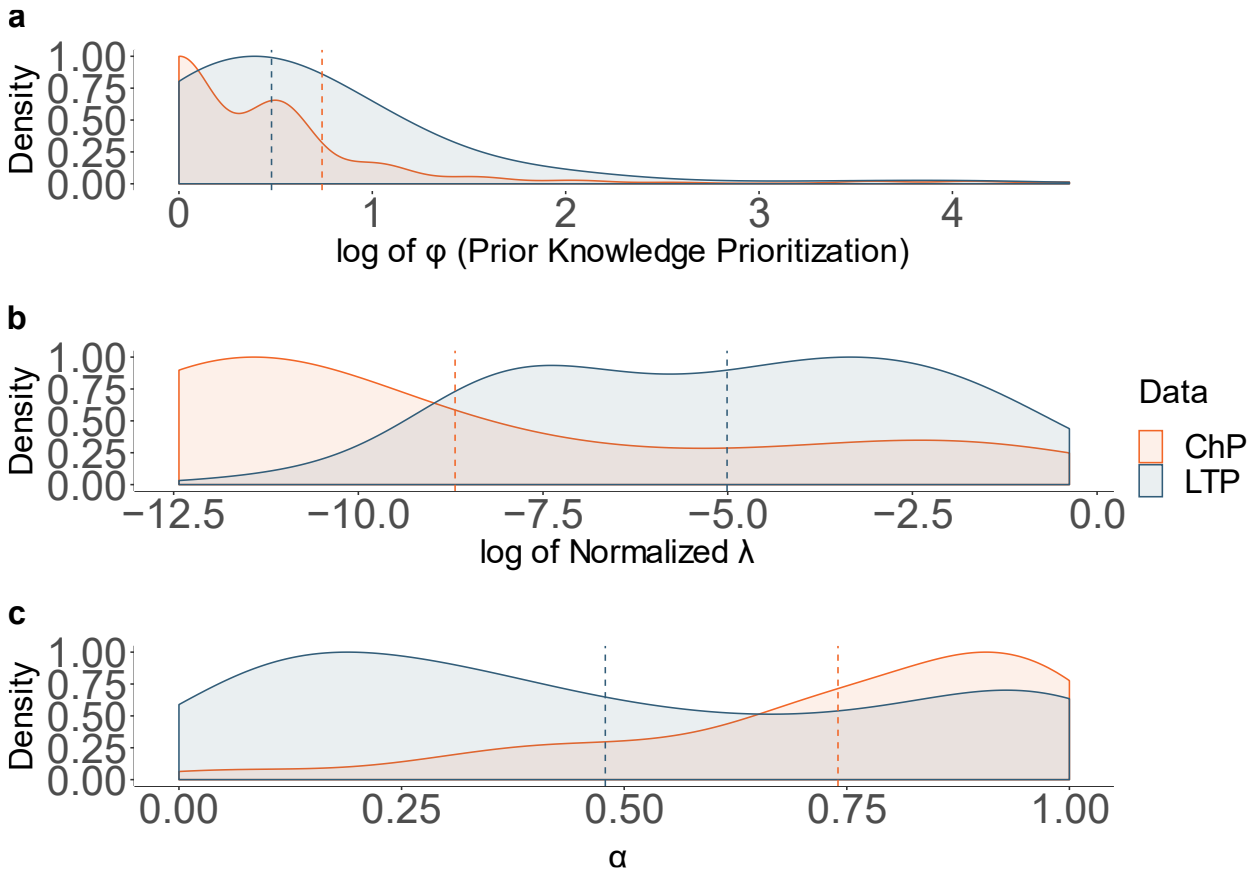
[**Supplemental Fig. 7 - Analysis of chronic periodontitis**: **(a)** A correlation network of intracellular signaling responses, measured in peripheral immune cells with nodes representing canonical signaling pathways (i.e. indicated by biological priors with a score > 0.5) displayed in red, otherwise black. Edges represent significant (P-value < 0.05) pairwise correlation after Bonferroni correction. Node size represents the significance of correlation to the response variable. **(b)** Unsupervised t-SNE visualization of patients showed no clear separation between control and patient populations across all immune features, motivating supervised classification analysis for identification of immune features associated with ChP. **(c)** Boxplot of mean out-of-sample prediction from 10-fold CV. Access to prior knowledge improved iEN's results compared to that of EN (with no access to prior knowledge).**(d)** For each ChP model, the Area Under the Receiver Operator Curve (AUROC) is visualized.]



[**Supplemental Fig. 8 - Simulation of Expert Error in Prior Knowledge:** To understand the robustness of the iEN we investigated disagreement among experts and missing information with simulated data. All simulations regarding missing information start from a baseline of 20% error. **(a)** Cellular signaling is a complex topic and it is likely that prior knowledge could be missing. We modeled this by randomly removing prior knowledge values across 9 simulated experts. The chance of random removal of prior values (representing lack of prior knowledge for that feature for an expert) increased until a random and uniform distribution was reached (vertical dashed line). **(b)** Similarly, experts may not agree on the prior knowledge values. We model this with increasing levels of error introduced to the 9 simulated experts until a uniform and random distribution is reached (vertical dashed line). **(c)** Both analyses, with uniform and random error, and the baseline iEN were repeated for various cohort sizes. These simulations demonstrate that iEN is robust to noise in the prior knowledge and that missing values in the prior knowledge have a smaller adverse impact than disagreement among experts. All curves are locally fitted polynomial curve with 95% CI.]



[**Supplemental Fig. 9 - iEN and EN feature selection and coefficient comparison: (a)** Comparison of features selected between each cross-validation iteration for LTP analysis showed significant overlap between the two models with slightly larger EN models. **(b)** Further comparison of iEN and EN LTP model coefficients averaged across all cross-validation iterations displays areas of the feature space more represented in their respective model. Immune features which are red are more represented in iEN models, while black nodes are more represented in EN models. **(c)** Comparison of features selected during ChP analysis showed similar overlap of features selected. However, EN models consistently used a larger number of features for the ChP than the LTP analysis. **(d)** ChP coefficients between iEN and EN models also displays a difference in feature representation.]



[**Supplemental Fig. 10 - Parameter selection during cross-validation:** Scaled distribution of optimized parameters, **(a)** φ, **(b)** λ, and **(c)** α, colored by the respective dataset. The mean of each distribution is overlaid as a dashed line. The consistency of parameter optimization across datasets and parameters show the stability of the model during the optimization process. Both φ and λ were generated on a log scale and are visualized accordingly.]

Supplemental Tables:

Table 1. Prior knowledge matrix for estimation of gestational age: All feature weights were limited to the range [0,1]. Responses most likely elicited by cells under certain stimulants (that were likely to be driven by the true underlying biology), were assigned a score of 1 whereas a 0 would indicate a feature representing cell types that should not elicit that particular response to that stimulant. The median value across all experts is reported here. See Methods section 5.1 for details.

Endogenous	pCREB	pERK1/2	IkB	pMK2	pNFkB	pP38	pS6	pSTAT1	pSTAT3	pSTAT5
Bcells	1	1	1	1	1	1	1	1	1	1
$CD16^{+}CD56^{-}$NKcells	1	1	1	1	1	1	1	1	1	1
$CD4^{+}Tcells_{mem}$	1	1	1	1	1	1	1	1	1	1
$CD4^{+}Tcells_{naive}$	1	1	1	1	1	1	1	1	1	1
$CD4^{+}$Tcells	1	1	1	1	1	1	1	1	1	1

CD45RA^{+}Tregs	1	1	1	1	1	1	1	1	1	1
CD45RA^{-}Tregs	1	1	1	1	1	1	1	1	1	1
CD56^{+}CD16^{-}NKcells	1	1	1	1	1	1	1	1	1	1
CD7^{+}NKcells	1	1	1	1	1	1	1	1	1	1
CD8^{+}Tcells$_{mem}$	1	1	1	1	1	1	1	1	1	1
CD8^{+}Tcells$_{naive}$	1	1	1	1	1	1	1	1	1	1
CD8^{+}Tcells	1	1	1	1	1	1	1	1	1	1
cMCs	1	1	1	1	1	1	1	1	1	1
Granulocytes	1	1	1	1	1	1	1	1	1	1
intMCs	1	1	1	1	1	1	1	1	1	1
M-MDSC	1	1	1	1	1	1	1	1	1	1
mDCs	1	1	1	1	1	1	1	1	1	1
ncMCs	1	1	1	1	1	1	1	1	1	1
pDCs	1	1	1	1	1	1	1	1	1	1
Tbet^{+}CD4^{+}Tcells$_{mem}$	1	1	1	1	1	1	1	1	1	1
Tbet^{+}CD8^{+}Tcells$_{mem}$	1	1	1	1	1	1	1	1	1	1
Tbet^{+}CD8^{+}Tcells$_{naive}$	1	1	1	1	1	1	1	1	1	1
TCR$\gamma\delta^{+}$Tcells	1	1	1	1	1	1	1	1	1	1
Tregs	1	1	1	1	1	1	1	1	1	1

IFN☒	pCREB	pERK1/2	IkB	pMK2	pNFkB	pP38	pS6	pSTAT1	pSTAT3	pSTAT5
Bcells	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
CD16^{+}CD56^{-}NKcells	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
CD4^{+}Tcells$_{mem}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
CD4^{+}Tcells$_{naive}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
CD4^{+}Tcells	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
CD45RA^{+}Tregs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
CD45RA^{-}Tregs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
CD56^{+}CD16^{-}NKcells	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
CD7^{+}NKcells	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
CD8^{+}Tcells$_{mem}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
CD8^{+}Tcells$_{naive}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
CD8^{+}Tcells	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
cMCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
Granulocytes	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
intMCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
M-MDSC	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
mDCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7

ncMCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
pDCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
$Tbet^{+}CD4^{+}Tcells_{mem}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
$Tbet^{+}CD8^{+}Tcells_{mem}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
$Tbet^{+}CD8^{+}Tcells_{naive}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
$TCR\gamma\delta^{+}Tcells$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7
Tregs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7

IL (IL-2/IL-6)	pCREB	pERK1/2	IkB	pMK2	pNFkB	pP38	pS6	pSTAT1	pSTAT3	pSTAT5
Bcells	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.8
$CD16^{+}CD56^{-}NKcells$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$CD4^{+}Tcells_{mem}$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$CD4^{+}Tcells_{naive}$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$CD4^{+}Tcells$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$CD45RA^{+}Tregs$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$CD45RA^{-}Tregs$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$CD56^{+}CD16^{-}NKcells$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$CD7^{+}NKcells$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$CD8^{+}Tcells_{mem}$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$CD8^{+}Tcells_{naive}$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$CD8^{+}Tcells$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
cMCs	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.6	0.8	0.7
Granulocytes	0.2	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.6	0.6
intMCs	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.6	0.8	0.7
M-MDSC	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.6	0.8	0.7
mDCs	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.6	0.8	0.7
ncMCs	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.6	0.7	0.7
pDCs	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.6	0.8	0.7
$Tbet^{+}CD4^{+}Tcells_{mem}$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$Tbet^{+}CD8^{+}Tcells_{mem}$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$Tbet^{+}CD8^{+}Tcells_{naive}$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
$TCR\gamma\delta^{+}Tcells$	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9
Tregs	0.3	0.6	0.3	0.2	0.3	0.2	0.2	0.7	0.8	0.9

LPS	pCREB	pERK1/2	IkB	pMK2	pNFkB	pP38	pS6	pSTAT1	pSTAT3	pSTAT5
Bcells	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	0	0
$CD16^{+}CD56^{-}NKcells$	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0	0	0
$CD4^{+}Tcells_{mem}$	0	0	0	0	0	0	0	0	0	0

$CD4^+Tcells_{naive}$	0	0	0	0	0	0	0	0	0	0
$CD4^+Tcells$	0	0	0	0	0	0	0	0	0	0
$CD45RA^+Tregs$	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	0	0
$CD45RA^-Tregs$	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	0	0
$CD56^+CD16^-NKcells$	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0	0	0
$CD7^+NKcells$	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0	0	0
$CD8^+Tcells_{mem}$	0	0	0	0	0	0	0	0	0	0
$CD8^+Tcells_{naive}$	0	0	0	0	0	0	0	0	0	0
$CD8^+Tcells$	0	0	0	0	0	0	0	0	0	0
cMCs	1	1	1	1	1	1	1	0.2	0.2	0.2
Granulocytes	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.2	0.1	0.1
intMCs	1	1	1	1	1	1	1	0.2	0.2	0.2
M-MDSC	1	1	1	1	1	1	1	0.2	0.2	0.2
mDCs	0.9	0.8	0.8	1	0.8	1	0.8	0.2	0.1	0.1
ncMCs	1	1	1	1	1	1	1	0.2	0.2	0.2
pDCs	0	0	0	0	0	0	0	0	0	0
$Tbet^+CD4^+Tcells_{mem}$	0	0	0	0	0	0	0	0	0	0
$Tbet^+CD8^+Tcells_{mem}$	0	0	0	0	0	0	0	0	0	0
$Tbet^+CD8^+Tcells_{naive}$	0	0	0	0	0	0	0	0	0	0
$TCR\gamma\delta^+Tcells$	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	0	0
Tregs	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	0	0

Table 2. Prior knowledge matrix for prediction of Chronic Periodontitis: All feature weights were limited to the range [0,1]. Responses most likely elicited by cells under certain stimulants (that were likely to be driven by the true underlying biology), were assigned a score of 1 whereas a 0 would indicate a feature representing cell types that should not elicit that particular response to that stimulant. The median value across all experts is reported here. See Methods section 5.1 for details.

IFNa	pCREB	pERK1/2	IkB	pMK2	pNFkB	pP38	pS6	pSTAT1	pSTAT3	pSTAT5	pSTAT6
Granulocytes	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.5
cMCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
ncMCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
intMCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
M-MDSC	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
mDCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
pDCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
$CD7^+NKcells$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6
$CD16^+CD56^-NKcells$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6
$CD56^+CD16^-NKcells$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6

$CD4^+Tcells_{mem}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6
$CD4^+Tcells_{naive}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
$Tbet^+CD4^+Tcells_{mem}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
$Tbet^+CD4^+CD45RA^+Tcells$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
Tregs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
$CD8^+Tcells_{mem}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6
$CD8^+Tcells_{naive}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6
Bcells	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7

IL Cocktail	pCREB	pERK1/2	IkB	pMK2	pNFkB	pP38	pS6	pSTAT1	pSTAT3	pSTAT5	pSTAT6
Granulocytes	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.5
cMCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
ncMCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
intMCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
M-MDSC	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
mDCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
pDCs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
$CD7^+NKcells$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6
$CD16^+CD56^-NKcells$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6
$CD56^+CD16^-NKcells$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6
$CD4^+Tcells_{mem}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6
$CD4^+Tcells_{naive}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
$Tbet^+CD4^+Tcells_{mem}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
$Tbet^+CD4^+CD45RA^+Tcells$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
Tregs	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7
$CD8^+Tcells_{mem}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6
$CD8^+Tcells_{naive}$	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.6
Bcells	0.4	0.4	0.4	0.4	0.4	0.4	0.2	1	0.8	0.7	0.7

P. gingivalis	pCREB	pERK1/2	IkB	pMK2	pNFkB	pP38	pS6	pSTAT1	pSTAT3	pSTAT5	pSTAT6
Granulocytes	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.2	0.1	0.1	0.1
cMCs	1	1	1	1	1	1	1	0.2	0.2	0.2	0.2
ncMCs	1	1	1	1	1	1	1	0.2	0.2	0.2	0.2
intMCs	1	1	1	1	1	1	1	0.2	0.2	0.2	0.2
M-MDSC	0.9	0.8	0.8	0.9	0.8	1	0.8	0.2	0.2	0.2	0.2
mDCs	0.9	0.8	0.8	1	0.8	1	0.8	0.2	0.1	0.1	0.1
pDCs	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0	0	0	0
$CD7^+NKcells$	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0	0	0	0
$CD16^+CD56^-NKcells$	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0	0	0	0

$CD56^{+}CD16^{-}$NKcells	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0	0	0	0
$CD4^{+}Tcells_{mem}$	0	0	0	0	0	0	0	0	0	0	0
$CD4^{+}Tcells_{naive}$	0	0	0	0	0	0	0	0	0	0	0
$Tbet^{+}CD4^{+}Tcells_{mem}$	0	0	0	0	0	0	0	0	0	0	0
$Tbet^{+}CD4^{+}CD45RA^{+}$Tcells	0	0	0	0	0	0	0	0	0	0	0
Tregs	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	0	0	0
$CD8^{+}Tcells_{mem}$	0	0	0	0	0	0	0	0	0	0	0
$CD8^{+}Tcells_{naive}$	0	0	0	0	0	0	0	0	0	0	0
Bcells	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	0	0	0

TNFa	pCREB	pERK1/2	IkB	pMK2	pNFkB	pP38	pS6	pSTAT1	pSTAT3	pSTAT5	pSTAT6
Granulocytes	0.8	0.7	0.8	0.8	0.8	0.8	0.7	0.1	0.2	0.2	0.2
cMCs	0.9	0.8	1	0.8	1	0.8	0.8	0.1	0.2	0.2	0.2
ncMCs	0.9	0.8	1	0.8	1	0.8	0.8	0.1	0.2	0.2	0.2
intMCs	0.9	0.8	1	0.8	1	0.9	0.8	0.1	0.2	0.2	0.2
M-MDSC	0.8	0.8	1	0.8	1	0.8	0.8	0.1	0.2	0.2	0.2
mDCs	0.9	0.8	1	0.8	1	0.8	0.8	0.1	0.2	0.2	0.2
pDCs	0.8	0.8	1	0.8	1	0.8	0.8	0.1	0.2	0.2	0.2
$CD7^{+}$NKcells	0.8	0.8	1	0.8	1	0.8	0.8	0.1	0.2	0.2	0.2
$CD16^{+}CD56^{-}$NKcells	0.8	0.8	1	0.8	1	0.8	0.8	0.1	0.2	0.2	0.2
$CD56^{+}CD16^{-}$NKcells	0.8	0.8	1	0.8	1	0.8	0.8	0.1	0.2	0.2	0.2
$CD4^{+}Tcells_{mem}$	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.1	0.2	0.2	0.2
$CD4^{+}Tcells_{naive}$	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.1	0.2	0.2	0.2
$Tbet^{+}CD4^{+}Tcells_{mem}$	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.1	0.2	0.2	0.2
$Tbet^{+}CD4^{+}CD45RA^{+}$Tcells	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.1	0.2	0.2	0.2
Tregs	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.1	0.2	0.2	0.2
$CD8^{+}Tcells_{mem}$	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.1	0.2	0.2	0.2
$CD8^{+}Tcells_{naive}$	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.1	0.2	0.2	0.2
Bcells	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.1	0.2	0.2	0.2

Table 3. Panel used for the mass cytometry assay of chronic periodontitis patient whole blood samples:

Antibody	Manufacturer	Symbol	Atomic Mass	Clone	Comment
Barcode 1	Trace Sciences	Pd	102		Barcode
Barcode 2	Trace Sciences	Pd	104		Barcode

Barcode 3	Trace Sciences	Pd	105		Barcode
Barcode 4	Trace Sciences	Pd	106		Barcode
Barcode 5	Trace Sciences	Pd	108		Barcode
Barcode 6	Trace Sciences	Pd	110		Barcode
CD235ab	Biolegend	In	113	HIR2	Phenotype
CD61	BD	In	113	VI-PL2	Phenotype
CD45	Biolegend	In	115	HI30	Phenotype
CD66	BD	La	139	CD66a-B1.1	Phenotype
CD7	BD	Pr	141	M-T701	Phenotype
CD19	Biolegend	Nd	142	HIB19	Phenotype
CD45RA	Biolegend	Nd	143	HI100	Phenotype
CD11b	Fluidigm	Nd	144	ICRF44	Phenotype
CD4	Fluidigm	Nd	145	RPA-T4	Phenotype
CD8a	Fluidigm	Nd	146	RPA-T8	Phenotype
CD11c	Fluidigm	Sm	147	Bu15	Phenotype
CD123	Biolegend	Nd	148	6H6	Phenotype
pCREB	Cell Signaling Technology	Sm	149	87G3	Phenotype
pSTAT5	Fluidigm	Nd	150	47	Function
pp38	BD	Eu	151	36/p38	Function
TCRγδ	Fluidigm	Sm	152	11F2	Phenotype
pSTAT1	Fluidigm	Eu	153	58D6	Function
pSTAT3	Cell Signaling Technology	Sm	154	M9C6	Function
pS6	Cell Signaling Technology	Gd	155	D57.2.2E	Function
CD24	Biolegend	Gd	156	ML5	Phenotype
CD38	Biolegend	Gd	157	HIT2	Phenotype
CD33	Fluidigm	Gd	158	WM53	Phenotype
pMAPKAPK2	Fluidigm	Tb	159	27B7	Function

Tbet	Fluidigm	Gd	160	4B10	Function
cPARP	BD	Dy	161	F21-852	Function
FoxP3	Fluidigm	Dy	162	PCH101	Phenotype
IκB	Fluidigm	Dy	164	L35A5	Function
CD16	Fluidigm	Ho	165	3G8	Phenotype
pNF-κB	Fluidigm	Er	166	K10-895.12.50	Function
pERK1/2	Fluidigm	Er	167	D13.14.4E	Function
pSTAT6	Fluidigm	Er	168	18	Function
CD25	Biolegend	Tm	169	M-A251	Phenotype
CD3	Fluidigm	Er	170	UCHT1	Phenotype
CD27	BD	Yb	171	M-T271	Phenotype
CCR2	Biolegend	Yb	173	K036C2	Phenotype
HLA-DR	Fluidigm	Yb	174	L243	Phenotype
CD56	BD	Yb	176	NCAM16.2	Phenotype
DNA1	Fluidigm	Ir	191		DNA
DNA2	Fluidigm	Ir	192		DNA

Table 4. Panel used for the mass cytometry assay of pregnant women whole blood samples:

Antibody	Manufacturer	Symbol	Atomic Mass	Clone	Comment
Barcode 1	Trace Sciences	Pd	102		Barcode
Barcode 2	Trace Sciences	Pd	104		Barcode
Barcode 3	Trace Sciences	Pd	105		Barcode
Barcode 4	Trace Sciences	Pd	106		Barcode
Barcode 5	Trace Sciences	Pd	108		Barcode
Barcode 6	Trace Sciences	Pd	110		Barcode
CD235ab	Biolegend	In	113	HIR2	Phenotype
CD61	BD	In	113	VI-PL2	Phenotype

CD45	Biolegend	In	115	HI30	Phenotype
CD66	BD	La	139	CD66a-B1.1	Phenotype
CD7	BD	Pr	141	M-T701	Phenotype
CD19	Fluidigm	Nd	142	HIB19	Phenotype
CD45RA	Fluidigm	Nd	143	HI100	Phenotype
CD11b	Fluidigm	Nd	144	ICRF44	Phenotype
CD4	Fluidigm	Nd	145	RPA-T4	Phenotype
CD8a	Fluidigm	Nd	146	RPA-T8	Phenotype
CD11c	Fluidigm	Sm	147	Bu15	Phenotype
CD123	Biolegend	Nd	148	6H6	Phenotype
pCREB	Cell Signaling Technology	Sm	149	87G3	Phenotype
pSTAT5	Fluidigm	Nd	150	47	Function
pp38	CST	Eu	151	36/p38/pT18	Function
TCRγδ	Fluidigm	Sm	152	11F2	Phenotype
pSTAT1	Fluidigm	Eu	153	58D6	Function
pSTAT3	BD	Sm	154	4/P pY705	Function
pS6	Cell Signaling Technology	Gd	155	D57.2.2E	Function
CD33	Fluidigm	Gd	158	WM53	Phenotype
pMAPKAPK2	Fluidigm	Tb	159	27B7	Function
Tbet	Fluidigm	Gd	160	4B10	Phenotype
FoxP3	Fluidigm	Dy	162	PCH101	Phenotype
IκB	Fluidigm	Dy	164	L35A5	Function
CD16	Fluidigm	Ho	165	3G8	Phenotype
pNF-κB	Fluidigm	Er	166	K10-895.12.50	Function
pERK1/2	CST	Er	167	D13.14.4E	Function
CD25	Biolegend	Tm	169	M-A251	Phenotype
CD3	Fluidigm	Er	170	UCHT1	Phenotype
CD15	Fluidigm	Yb	172	W6D3	Phenotype

HLA-DR	Fluidigm	Yb	174	L243	Phenotype
CD14	Fluidigm	Yb	175	M52E	Phenotype
CD56	Fluidigm	Yb	176	NCAM16.2	Phenotype
DNA1	Fluidigm	Ir	191		DNA
DNA2	Fluidigm	Ir	192		DNA

Table 5. Summary statistics for the EN and iEN models:

	EN		iEN	
	Median of -log10 p-values	SD of -log10 p-values	Median of -log10 p-values	SD of -log10 p-values
LTP	**6.61**	**0.934**	**7.95**	**1.54**
VAL	**2.56**	**0.18**	**3.81**	**0.48**
ChP	**1.25**	**0.86**	**2.04**	**1.03**

Table 6. Quantification of the cohort size required to achieve various Pearson Rho intervals in the simulation study:

	(0, 0.175]	(0.175, 0.29]	(0.29, 0.405]	(0.405, 0.52]	(0.52, 0.635]	(0.635, 0.751)
Average iEN Cohort Size	88.32	122.51	58.04	224.73	292.93	3620.87
Average EN Cohort Size	109.14	220.69	101.28	276.86	989.92	5129.79
Difference in Average Required Cohort Size	-20.82	-98.19	-43.24	-53.13	-696.99	-1508.92
% Reduction in Size	19.08	44.49	42.69	19.19	70.41	29. 41