nature machine intelligence

ARTICLES

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

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

Anthony Culos[1,2,10], Amy S. Tsai[1,10], Natalie Stanley[1,2], Martin Becker[1,2], Mohammad S. Ghaemi[1,2,3], David R. McIlwain[4], Ramin Fallahzadeh[1,2], Athena Tanada[1,2], Huda Nassar[1,2], Camilo Espinosa[1,2] Maria Xenochristou[1,2], Edward Ganio[1], Laura Peterson[1,5], Xiaoyuan Han[1], Ina A. Stelzer[1], Kazuo Ando[1], Dyani Gaudilliere[1], Thanaphong Phongpreecha[1,2,6], Ivana Marić[1,5], Alan L. Chang[1,2], Gary M. Shaw[5], David K. Stevenson[5], Sean Bendall[6], Kara L. Davis[5], Wendy Fantl[4,7,8], Garry P. Nolan[6], Trevor Hastie[2,9], Robert Tibshirani[2,9], Martin S. Angst[1,11], Brice Gaudilliere[1,5,11] and Nima Aghaeepour[1,2,5,11] ✉

The dense network of interconnected cellular signalling responses that are quantifiable in peripheral immune cells provides a wealth of actionable immunological insights. Although high-throughput single-cell profiling techniques, including polychromatic flow and mass cytometry, have matured to a point that enables detailed immune profiling of patients in numerous clinical settings, the limited cohort size and high dimensionality of data increase the possibility of false-positive discoveries and model overfitting. We introduce a generalizable machine learning platform, the immunological Elastic-Net (iEN), which incorporates immunological knowledge directly into the predictive models. Importantly, the algorithm maintains the exploratory nature of the high-dimensional dataset, allowing for the inclusion of immune features with strong predictive capabilities even if not consistent with prior knowledge. In three independent studies our method demonstrates improved predictions for clinically relevant outcomes from mass cytometry data generated from whole blood, as well as a large simulated dataset. The iEN is available under an open-source licence.

In response to an immunological challenge, immune cells act in concert to form a complex and dense cell-signalling network[1,2]. The single-cell evaluation of intracellular signalling responses is particularly valuable in characterizing this cellular network as it provides a functional assessment of an individual's immune system. In clinical settings, a deep understanding of functional immune responses not only provides diagnostic opportunities, but is also often the first step in developing immune therapies (recent examples include stratification of COVID-19 disease severity[3,4] and successful immune modulation in chronic lymphocytic leukaemia[5], neurodegeneration[6] and Ebola virus disease[7]).

Advanced flow cytometry technologies can characterize millions of single cells from a given patient, which enables the identification of signalling pathways, even in rare cell populations[8]. The recent advent of high-dimensional polychromatic flow cytometry[9,10] and mass cytometry[11,12] technologies have vastly increased our ability to study the human immune system with unprecedented functional depth by increasing the number of features measured per cell. However, the increased dimensionality relative to the cohort sizes in clinical studies, as well as the inherently complex networks of internal correlations between the measured cell types and pathways present unique computational challenges[13]. Translating complex and high-dimensional observations, like immunological data, into relevant computational models requires statistically rigorous analysis techniques. Multivariable modelling, in contrast to univariable analysis, can simultaneously consider all measured aspects of the immune system to improve predictions. However, multivariable modelling requires exponentially larger cohort sizes as the number of measurements grows (also know as the 'curse of dimensionality'[14–16]). This is especially true for more modern deep-learning-based models that provide general function approximation but require substantially larger sample sizes. In practice, increasing the cohort size by several orders of magnitude to power such analyses is often a significant challenge in resource-constrained settings. Moreover, multivariable analyses performed on all available measurements produce large complex models that are difficult to interpret and implement in resource-constrained settings[17] and often lack robustness[18].

Integration of prior knowledge has been broadly recognized as an effective approach for reducing model complexity and increasing

[1]Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Stanford, CA, USA. [2]Department of Biomedical Data Sciences, Stanford University, Stanford, CA, USA. [3]Digital Technologies Research Centre, National Research Council Canada, Toronto, Ontario, Canada. [4]Department of Microbiology and Immunology, Baxter Laboratory in Stem Cell Biology, Stanford University School of Medicine, Stanford, CA, USA. [5]Department of Pediatrics, Division of Neonatal and Developmental Medicine, Stanford University School of Medicine, Stanford, CA, USA. [6]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [7]Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford, CA, USA. [8]Department of Urology, Stanford University School of Medicine, Stanford, CA, USA. [9]Department of Statistics, Stanford University, Stanford, CA, USA. [10]These authors contributed equally: Anthony Culos, Amy S. Tsai. [11]These authors jointly supervised this work: Martin S. Angst, Brice Gaudilliere, Nima Aghaeepour. ✉e-mail: naghaeep@stanford.edu

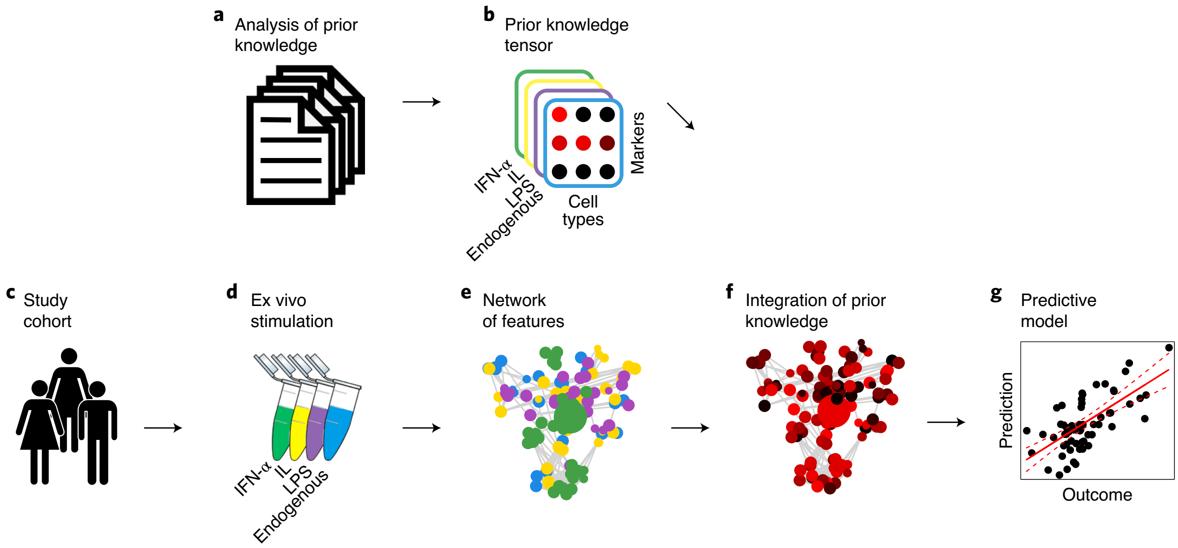

Fig. 1 | The immunological Elastic-Net analysis pipeline. a,**b**, Immunological prior knowledge for each feature, in response to each ex vivo stimulation condition, is extracted by a panel of experts (**a**) and encoded into a prior knowledge tensor to guide the model optimization process (**b**). **c**,**d**, Individuals within the cohort of study (**c**) provide blood samples, which are subsequently stimulated with ligands ex vivo to activate various signalling pathways of the immune system (**d**). **e**, This produces single-cell measurements of the immune system, resulting in a complex network of cell types and signalling pathways representing both innate and adaptive immunity. **f**,**g**, This dataset is then fed into the iEN algorithm (**f**) for predictive modelling of the outcome of interest (**g**).

robustness[19–31]. In biological sciences, examples of such knowledge integration include inference of biological networks[24,32] and causal pathway modelling[33,34]. In modern immunological datasets, however, integration of prior knowledge has been impractical due to the unstructured format of prior immunological datasets and the complex nature of the measured features. In this work, we propose a framework for integration of prior immunological knowledge into the model optimization process of the Elastic Net[35] (EN) algorithm (Fig. 1). Our decision to select the EN as a candidate method was primarily motivated by two factors. First, as a sparse model, the EN is broadly used when the number of features exceeds the number of observations, as is the case with single-cell profiling of the immune system. This sparsification not only improves performance, but also makes models more interpretable, which would prove indispensable for tasks that require accuracy from an interpretable model[17] (such as conflict forecasting[36], drug discovery[37], generalizing findings from animal models to human studies[38,39] and translational clinical studies). Second, while other algorithms can potentially be modified to account for prior knowledge, this modification for linear models such as the EN is a natural fit, as demonstrated by the Bayesian interpretation of the EN (Methods). Although other multivariable models have previously been extended to incorporate feature group information[21,40,41], none has integrated per-feature prior knowledge as needed in this work. In our immunological Elastic-Net (iEN) framework, the prior knowledge developed by a panel of expert immunologists on a per-feature basis is integrated into the EN algorithm during coefficient optimization (similar to adjusting Bayesian priors; Methods). The addition of knowledge-based immunological priors guides the sparsification process towards solutions more consistent with biological knowledge while still allowing all measured immune features to be included in the exploratory analysis.

In our experiments, the iEN outperformed the standard EN, as well as a broad range of standard machine learning algorithms, including the K-nearest-neighbours[42] (KNN), support vector machine[43] (SVM), random forest[44] (RF) and least absolute shrinkage and selection operator[45] (LASSO). Additionally, 'Super Learner'[46] (an ensemble of the aforementioned methods) and two alternative knowledge-integrated algorithms, Know-GRRF[25] and graper[41], were included in our evaluation. A two-layer repeated 10-fold cross-validation (CV) was used to determine model consistency and establish a clear comparison of model performance among the algorithms. In the first CV layer, the free parameters of the models were optimized (including a factoring controlling the impact of domain knowledge in the case of the iEN) and the second CV layer predicted previously unseen observations. These predictions were then aggregated and evaluated by an appropriate hypothesis test (Pearson or Wilcoxon rank-sum), root-mean-squared error (r.m.s.e.) and area under the receiver operator curve (AUROC) to measure model performance. This process was repeated with randomly generated CV folds to visualize the distribution of results subject to variations in the cohort.

In this Article, we have included two real-world clinical examples as well as a large simulation study. The first analysis, as an example of a continuous clinical outcome, identified components of maternal immune adaptations in a longitudinal term pregnancy (LTP) study, which included an independent validation cohort. The second example was a classification analysis of a categorical outcome, modelling patient and control populations for chronic periodontitis (ChP). The third example used synthetic data generated to replicate mass cytometry measurements to enable in-depth understanding of the iEN behaviour across varying cohort sizes. Additional analyses were performed to determine the effect of prior knowledge on general model behaviour and the stability of results subject to errors in the expert-defined prior knowledge. Each of the three examples was chosen to demonstrate the generalizability of the iEN algorithm, as well as its efficacy, in a range of real-world scenarios.

Knowledge integration

Prior knowledge tensors constructed before the analysis emphasized receptor-specific signalling responses describing canonical pathways activated downstream of the ex vivo stimulation conditions used in the mass cytometry assays. The biological priors used in the iEN model were created by an independent panel of immunologists, such that features more consistent with known biology have higher values (Supplementary Tables 1 and 2). For example, panel members broadly agreed on the prioritization of the phosphorylation of signal transducer and activator of transcription 1 (STAT1), STAT3 and STAT5 in all adaptive and innate immune cells in response to interferon-α (IFN-α) stimulation[47,48]; the phosphorylation of STAT1, STAT3, STAT5 and extracellular signal-regulated kinase 1/2 (ERK1/2) mitogen-activated protein kinase (MAPK) in all adaptive and innate immune cells in response to the interleukin (IL) cocktail containing IL-2 and IL-6[49,50]; the phosphorylation of P38 MAPK, MAPKAPK2, ERK1/2, ribosomal protein S6 (rpS6), cAMP-response element binding protein (CREB) and nuclear factor κ light-chain-enhancer of activated B cells (NF-κB) and total IκB signal in all innate immune cells (except plasmacytoid dendritic cells) and in regulatory T cells in response to the lipopolysaccharide (LPS) stimulation condition[51–55]. The resulting five tensors from each expert were then aggregated into a single median tensor used for iEN analyses. The use of median aggregation was to avoid bias by any one expert. An example of all measured immune features and those selected by this prior knowledge tensor is visualized in Fig. 2a.

These tensor values vary from zero to one, with one representing the immune features that are most consistent with prior knowledge according to the panel of experts. iEN regularized regression models are constructed by optimization of the objective function $L(\beta) = ||Y - X\phi\beta||^2 + \lambda[(1-\alpha)||\beta||^2/2 + \alpha||\beta||_1]$, where X is a matrix of p measured immune features (columns) for n patients (rows) and Y is a vector of the clinical outcomes of interest. The algorithm calculates the coefficients β that minimize the objective function subject to the $L_1 = ||\beta||_1$ and $L_2 = ||\beta||^2$ penalties. The combination of these penalty terms allow for the selection of the features correlated with the outcome of interest and the exclusion of redundant measurements, while also accounting for internally correlated measurements. Model optimization for iEN is controlled by three parameters: λ, α and φ. These parameters can be interpreted as the amount of sparsity in a model (λ), how sparsity is balanced between the L_1 and L_2 penalty terms (α), and the amount of prior knowledge prioritization (φ). Prioritization of biologically consistent features is accomplished through ϕ, which is a $p \times p$ diagonal matrix of the form $\mathrm{diag}(\phi) = \{\varphi_{1,1}, \varphi_{2,2}, \ldots, \varphi_{p,p}\}$ such that $\varphi_{i,i} = \{\mathrm{e}^{-\varphi(1-z_i)}$ where z_i is the score of the ith immune feature, and φ is the amount of prioritization attributed to the model}. This definition allows for a limited effect of φ on model coefficients while also increasing the impact of the features consistent with the prior knowledge tensor (Fig. 3). In the most extreme case, this results in only features with a tensor value of 1 being selected (Supplementary Fig. 1b). This behaviour allows the iEN to generate sparse models while still maintaining the exploratory nature of the EN, as opposed to a priori limiting the model to the features consistent with prior knowledge. Similar to the α and λ free parameters, prioritization of prior knowledge affects the sparsification (Fig. 3 and Supplementary Fig. 1) and optimization of the model (Supplementary Fig. 2). The two-layer 10-fold CV used for iEN optimization and estimation was implemented and parallelized over parameters α, λ and φ. Runtime analysis for this procedure is presented in Supplementary Fig. 3b.

Experiments

To evaluate the effectiveness of integrating prior immunological knowledge using the iEN algorithm, we have provided two real-world clinical examples as well as a simulated study.

Analysis of longitudinal term pregnancy. The first example investigated the adaptations of the maternal immune system during pregnancy that can be incorporated into predictive models of gestational age[56]. During a healthy pregnancy, the immune system strikes a delicate balance to enable tolerance towards the fetus and simultaneously mount a response to defend against pathogens. Abnormal immune system adaptations during pregnancy have been linked to adverse maternal and neonatal outcomes, such as pregnancy loss, preterm birth and preeclampsia[57–59]. This study aims to understand the immunological mechanisms behind term birth as a pivotal first step in understanding abnormal pregnancies and their impact on long-term outcomes[60]. In this example, a total of 54 blood samples from 18 women were studied during and six weeks after pregnancy. Three antepartum blood draws were collected at different gestational ages, with gestation being measured via ultrasound at the time of sample collection. The resulting 54 whole-blood profiles of the immune system were manually gated (Supplementary Fig. 4) into 24 cell types to measure the endogenous activity of 10 signalling markers, as well as the activity of these 10 markers in response to ex vivo stimulations with three different ligands, providing 960 immune features for analysis (Fig. 2a). Unsupervised analysis of each antepartum sample using the t-distributed stochastic neighbour embedding (t-SNE)[61] algorithm showed no readily identifiable patterns associated with gestational age (Supplementary Fig. 5a), motivating supervised iEN analysis. Predictive models were built using iEN, and model parameters were optimized to minimize the residual sum of squares of the predicted versus actual gestational age. iEN produced models of immune features that more accurately predicted gestational age than the similar EN analysis and other contemporary machine learning methods, which were agnostic to the immunologic priors (Fig. 4a,e and Supplementary Figs. 5b and 6a). iEN analyses of postpartum samples demonstrated that the immune system returns to a state that is similar to early pregnancy by six weeks postpartum (Supplementary Fig. 5c). An additional cohort of 10 women were prospectively studied and analysed as an independent validation set. Importantly, the validation cohort also demonstrated that iEN models produced substantially more accurate results than the EN algorithm (Fig. 4b,f and Supplementary Fig. 5e,f). Stepwise reduction of iEN and EN model coefficients revealed superior predictions in the validation cohort by the iEN compared to the EN algorithm for models of equal size (Supplementary Fig. 5d).

Analysis of chronic periodontitis. The second example investigated the classification of patients with ChP, a chronic inflammatory disease of the oral cavity. ChP is associated with severe systemic illnesses (such as heart disease, various malignancies and preterm labour)[62,63] and, in its most severe form, was estimated to affect ~11% of the global population in 2010[64]. A better understanding of the immunological manifestations of ChP is a critical first step for the development of immune therapies that may alter the course of systemic diseases associated with ChP. This dataset was generated from 28 participants, 14 diagnosed with ChP and 14 healthy controls. Blood samples from the study participants were analysed by mass cytometry and manually gated (Supplementary Fig. 4) for 18 cell types to measure 11 signalling markers in response to ex vivo stimulation with four different ligands; this provided a total of 792 immune features for analysis (Supplementary Fig. 7a). Unsupervised t-SNE visualization of the entire dataset showed no clear separation between case and control populations across all immune features, motivating supervised classification analysis using iEN (Supplementary Fig. 7b). For this example, iEN and EN algorithms were used to fit binomial models for the classification of patients and controls. All free parameters were optimized for AUROC[65]. The analysis results indicated that iEN outperformed the EN (Fig. 4c,g and Supplementary Fig. 7c) as well as other machine learning methods (Supplementary Figs. 6c and 7d).

Simulation study. The third example, a simulation study, demonstrated the advantages of prior knowledge integration in high-dimensional studies in relation to cohort size. Datasets were generated with 700 features per simulated patient while the number of patients varied from 100 to 1,000 in increments of 100. Data were generated with features that were random and uniform. A limited

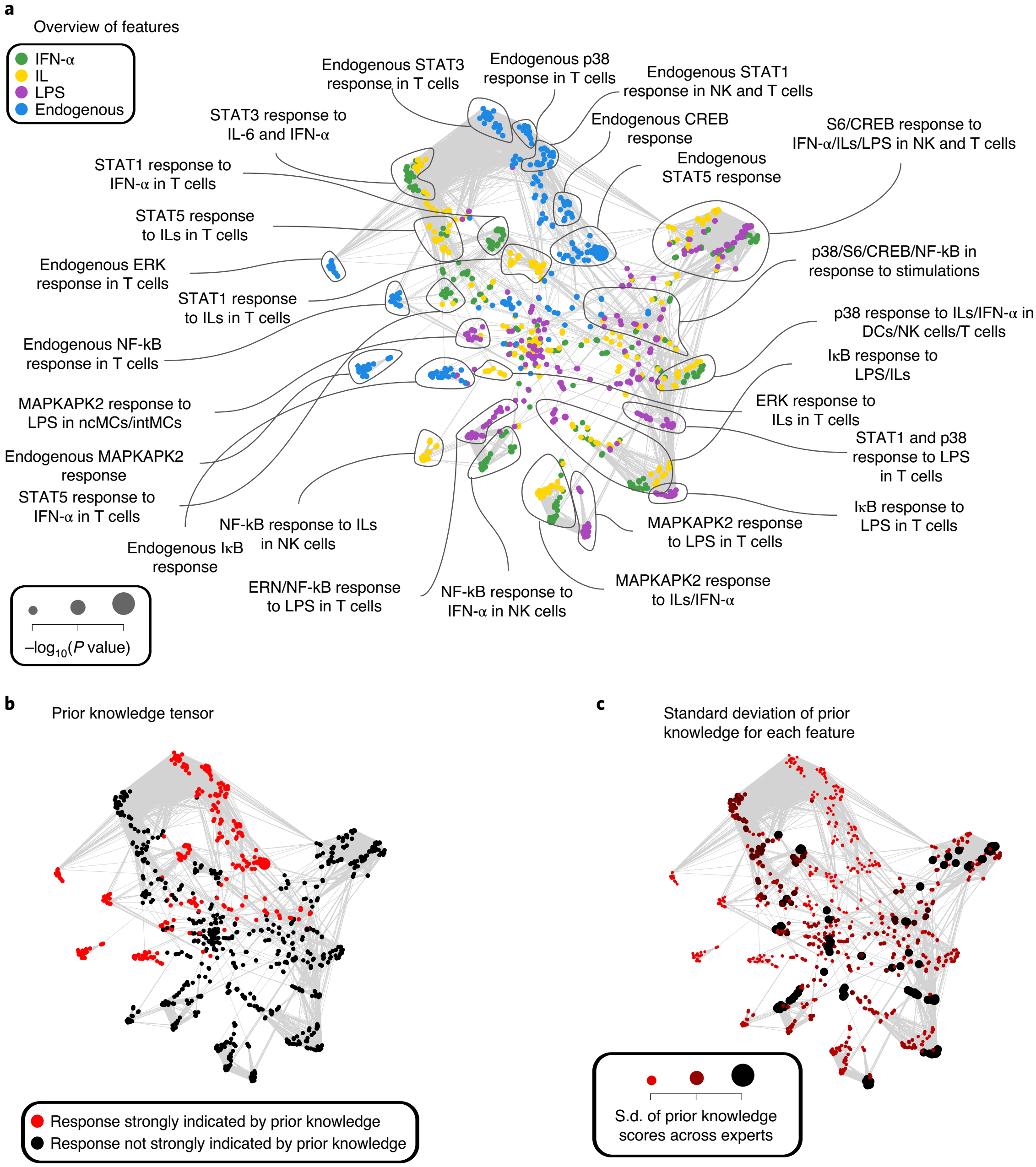
a
Overview of features
IFN-α
IL
LPS
Endogenous
Endogenous STAT3 response in T cells
Endogenous p38 response in T cells
Endogenous STAT1 response in NK and T cells
STAT3 response to IL-6 and IFN-α
Endogenous CREB response
S6/CREB response to IFN-α/ILs/LPS in NK and T cells
STAT1 response to IFN-α in T cells
Endogenous STAT5 response
STAT5 response to ILs in T cells
p38/S6/CREB/NF-kB in response to stimulations
Endogenous ERK response in T cells
STAT1 response to ILs in T cells
p38 response to ILs/IFN-α in DCs/NK cells/T cells
Endogenous NF-kB response in T cells
IκB response to LPS/ILs
MAPKAPK2 response to LPS in ncMCs/intMCs
ERK response to ILs in T cells
STAT1 and p38 response to LPS in T cells
Endogenous MAPKAPK2 response
STAT5 response to IFN-α in T cells
NF-kB response to ILs in NK cells
IκB response to LPS in T cells
Endogenous IκB response
MAPKAPK2 response to LPS in T cells
ERN/NF-kB response to LPS in T cells
NF-kB response to IFN-α in NK cells
MAPKAPK2 response to ILs/IFN-α
–log10(P value)
b
Prior knowledge tensor
Response strongly indicated by prior knowledge
Response not strongly indicated by prior knowledge
c
Standard deviation of prior knowledge for each feature
S.d. of prior knowledge scores across experts

Fig. 2 | Integration of immunological priors. a, Overview of the LTP study. A correlation network of intracellular signalling responses, measured in peripheral immune cells and coloured by ex vivo stimulation status, is visualized. Edges represent significant ($P < 0.05$) pairwise correlation after Bonferroni adjustment for multiple hypothesis correction. Node sizes represent the significance of correlations with the response variable (gestational age during term pregnancy). **b**, Immune features that were congruent with domain-specific knowledge as determined by a panel of five immunologists were refined into a tensor and used to determine the node colour of the correlation network. Here, immune features that have a value of 1 (full agreement among the panel) are coloured red and all other immune features are coloured black. **c**, The network is coloured by the standard deviation of scores assigned to each feature by the panel of immunologists. Overall, the consistency of the prior knowledge among panel members is higher in the features with a higher score, indicating a stronger agreement regarding the top features that should be prioritized by the algorithm and disagreements regarding features inconsistent with prior knowledge.

number of features were generated to have various degrees of correlation with the response variable. Specifically, 50 highly predictive features, 200 moderately predictive features and 450 randomly distributed features were assigned a corresponding biological prior. For a more detailed description of the data-generating process see the simulated data section in the Methods. Repeated 10-fold analysis of the simulated data with increasing population sizes displayed a convergent trend between the iEN and EN models as n increased (Fig. 4d,h, Supplementary Fig. 3a and Supplementary Table 6). As cohort size increased, iEN captured the highly predictive features faster than EN (Supplementary Fig. 3c). These results suggest that if the cohort size is too small to the point that model outputs are affected by noise, the advantage offered by prior knowledge also diminishes (left side, Fig. 4d,h). The results also indicate that integration of prior knowledge is of particular importance in settings where the number of instances (for example, cohort size) is limited yet a relatively large number of features are measured, as is common in translational systems immunology studies.

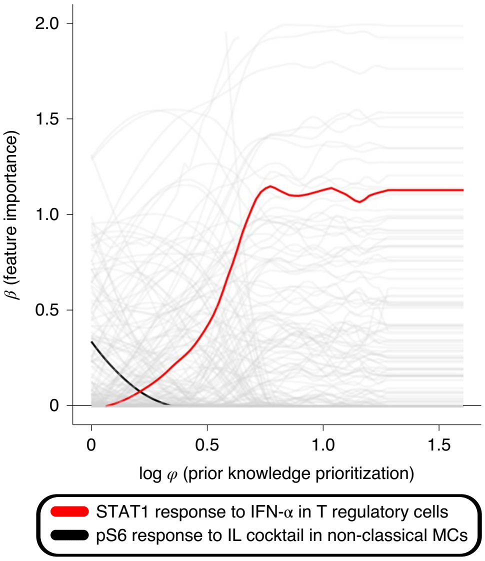

Fig. 3 | Prior knowledge and sparsification. Visualization of an example of the impact of prior immunological knowledge on various features in an iEN model. As φ is increased across the x axis (increased impact of prior knowledge), the contributions of each feature to the final model (y axis) change to select models consistent with immunological priors. We have highlighted two examples where a feature is emphasized or de-emphasized (in red and black, respectively) by prior knowledge. In this example, the STAT1 response to IFN-α stimulation in regulatory T cells is prioritized as STAT1 is downstream of the IFN-α/β receptor and is integral to their homeostasis and function[86]. Conversely, the prpS6 response to stimulation by IL-2 and IL-6 in non-classical monocytes is increasingly deprioritized as this signalling response is inconsistent with prior understanding of these signal transduction pathways in this cell type; IL-2 primarily drives T-cell differentiation through the Janus kinase (JAK)/STAT pathway[87]. Similarly, IL-6 primarily activates the JAK/STAT pathway and IL-6 receptors are expressed only in a subset of immune cells[88,89]. This confirms that integration of the priors can not only modify the algorithm's behaviour, but also that the intensity of this impact can be controlled through the φ free parameter.

Sensitivity analysis of the prior knowledge tensor. The iEN pipeline depends on the prior knowledge tensor. We therefore investigated the robustness of iEN by introducing errors into the prior knowledge tensor in each of the three examples. Introduction of moderate to substantial error into the prior tensor resulted in a consistent reduction in the predictive benefit of the iEN over the traditional EN model. Error was introduced stochastically and progressed towards uniform random noise, with 11 total incremental steps and 100 tensors generated per increment. As the prior tensor approaches a uniform random distribution (the highest amount of error in the prior knowledge matrix) the iEN and EN performances converge. These results remain consistent across the LTP, LTP validation, ChP and simulation studies (Fig. 5). Additional analyses performed on simulated data investigated the effect of errors and missing information in the prior knowledge tensor of individual experts (Supplementary Fig. 8). This analysis demonstrated that the algorithm is more sensitive to erroneous information introduced into the prior knowledge tensor than it is to missing information.

Empirical evaluation. In all analyses, the integration of expert knowledge improved the prediction of clinical outcomes in comparison to EN[35] with no prior knowledge (Fig. 4). iEN also outperformed standard machine learning algorithms including LASSO[45], RF[44], SVM[43,66], KNN[42,46], Super Learner[46] (an ensemble of the aforementioned methods) and the knowledge integrated Know-GRRF[25] and graper[41] methods (Supplementary Figs. 6 and 7d). Further comparison of iEN and EN models by selected features demonstrated a substantial overlap between the models; however, model comparison by coefficient weights displayed a substantial difference in the predictive importance of the features selected (Supplementary Fig. 9). The iEN results demonstrated a higher variability in predictions (Supplementary Table 5). This is probably caused by the additional φ parameter applied in iEN, resulting in a larger search space for potential model parameters. Hyperparameter selection frequency across all generated models displayed a consistent prioritization of prior knowledge via φ (Supplementary Fig. 10). The integration of prior knowledge allows the iEN to determine the predictive benefit of prioritizing canonical signalling pathways in a data-driven manner (Supplementary Fig. 2). Importantly, this enables iEN to function without excluding any of the features from consideration, even when scored as inconsistent with prior knowledge by the human experts. This behaviour allows for the iEN to contain the EN as an edge case when the prior knowledge is not beneficial. Similarly, when prioritization is most beneficial, the resulting model will contain only features scored as 1 in the prior tensor (those with complete consensus among the panel of experts).

Discussion

A structured collaboration between clinicians, biologists and computer scientists can lead to machine learning algorithms in the life sciences that achieve stronger results[67]. In this Article, we proposed a collaborative framework that enables integration of prior knowledge of cell signalling pathways in a machine learning algorithm to improve the predictions and robustness of the resulting models in clinical datasets. This knowledge-integrated approach to data analysis can be generalized under the 'learning using privileged information' paradigm described by Vapnik and Vashist[68] where external information is used at the time of training to improve the incurring decision rule. In our experiments, this strategy improved the accuracy of predictions in both translational clinical studies and simulated experiments, even when moderate amounts of noise were artificially added to the extracted prior knowledge.

Importantly, the data-driven approach implemented here allows for prior knowledge only to be incorporated when an improvement in the model is observed. Functionally, this reduces the regularization of the features consistent with prior knowledge, resulting in the development of sparse models that prioritize features in line with previous biological studies. This not only increases the predictive performance, but also facilitates biological interpretation and translation of the results. From a biological perspective, the iEN enhances the interpretability of the results as model features are enriched for cell-type and receptor-specific signalling responses (as opposed to over-reliance on visualization using dimension reduction tools[14]). These resulting multivariable models with increased biological consistency and improved predictive capabilities could also contribute to the development of robust and simplified assays for resource-limited settings. From a Bayesian perspective, this prior knowledge integration could be viewed as a shift in the prior distributions over β towards estimates that are more congruent with the true underlying distributions. A more explicit connection to the Bayesian setting is presented in the Methods.

This study has several limitations, which will guide our future research directions. First, definition of the prior knowledge tensor by individual human experts has the potential to induce a source of bias into the analysis. Although our analysis suggested that the method is robust to potential errors in the prior knowledge tensor, a more accurate and consistent definition of prior knowledge would improve this pipeline. In addition, the development of the prior

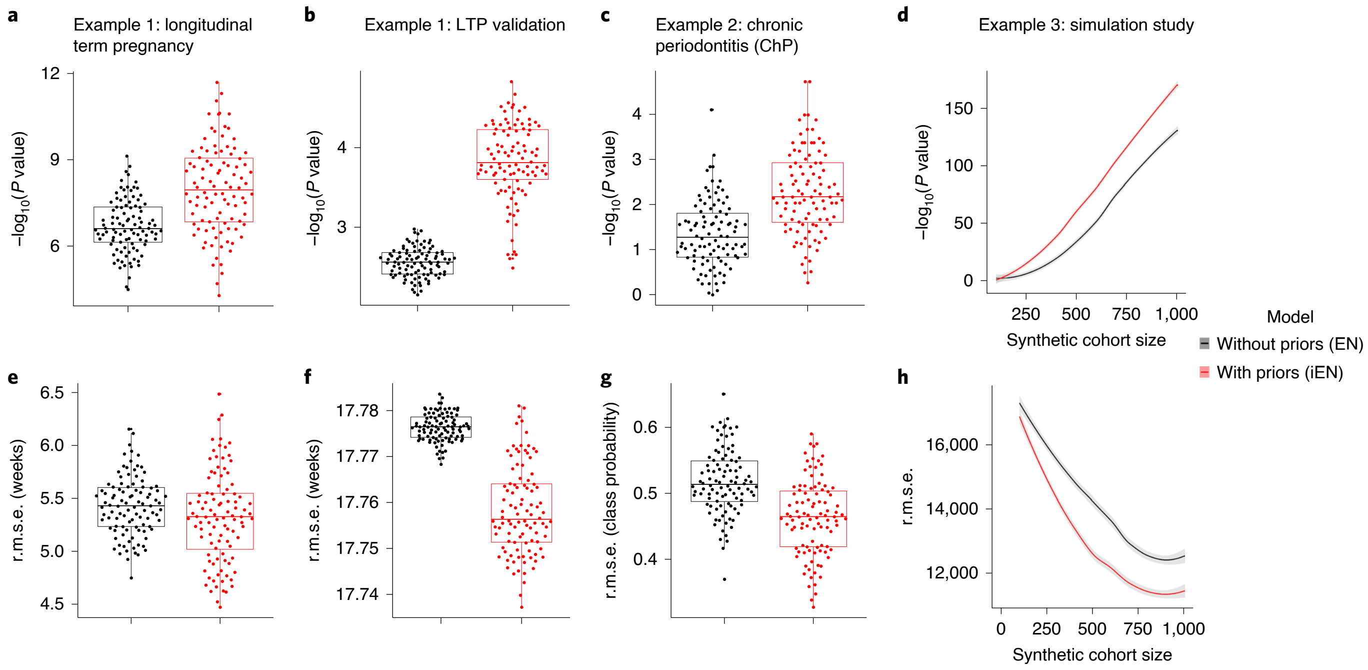

Fig. 4 | Incorporation of prior knowledge improves predictions in two clinical studies and a simulated experiment. a, Boxplot of Pearson correlation *P* values calculated on out-of-sample predictions from repeated 10-fold cross-validation of EN (black) and iEN (red) models for the LTP dataset. **b**, Validation of the LTP model on an independent validation cohort. These predictions were compared against the true response variable using $-\log_{10}$ Pearson correlation *P* values. **c**, Boxplots of Wilcoxon rank-sum test *P* values similarly calculated on out-of-sample predictions for the ChP dataset (null hypothesis: the sample-to-class assignment probabilities produced by the model are equal between the two outcome classes). Comparison of model performance for the respective datasets demonstrated improved predictions for the iEN, as shown by $-\log_{10}$ *P* values. **d**, A simulated study with varying cohort sizes of simulated 'patients' with 700 features demonstrated a larger gain (measured by $-\log_{10}$ Pearson's test *P* values) for the integration of prior immunological knowledge in datasets with a relatively small cohort size and a large number of features. Locally fitted polynomial curves of prediction performance over multiple cohort sizes are displayed with 95% confidence intervals (CIs). **e**–**h**, R.m.s.e. values to demonstrate the effect sizes of the models in **a**–**d**.

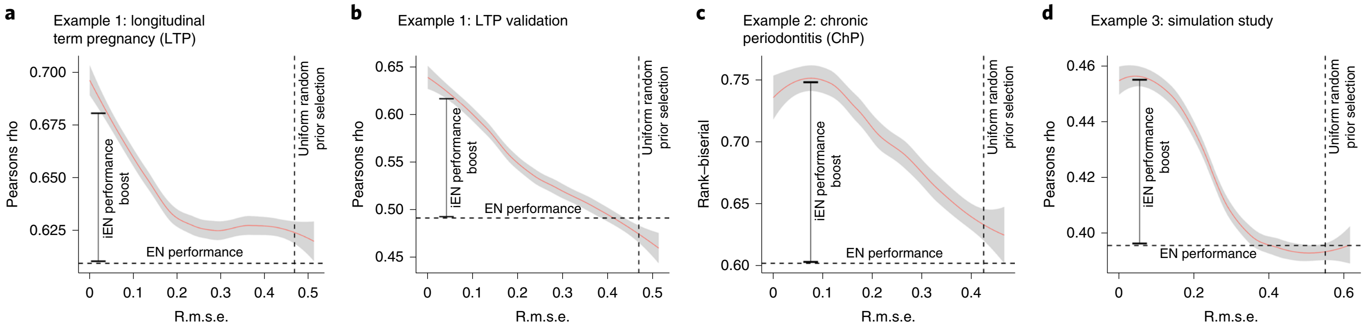

Fig. 5 | iEN is robust to errors in the prior knowledge tensor. a–**d**, Various levels of noise were artificially added to the prior knowledge values, as indicated by the r.m.s.e. values of the true prior values versus the simulated ones (*x* axis). As the value on the *x* axis increases, the amount of noise in the simulated prior increases until all priors are sampled from a uniform and random distribution (vertical dashed line). Reassuringly, at this point, the performance of iEN is close to that of the EN (with no priors), as indicated by a horizontal dashed line. Importantly, iEN continues to outperform the EN (horizontal dashed line) for even high amounts of error in the priors. All curves are locally fitted polynomials of predictive performance with the shaded region representing 95% CI.

knowledge is labour-intensive and requires careful and objective analysis of a broad range of studies. We believe stronger results can be achieved using text-mining strategies for direct extraction of prior knowledge from the literature (for example, see immuneXpresso[69]). Second, although our simulations point to trends in the advantages of prior knowledge integration in relation to cohort size, they cannot be used for specific guidance regarding the cohort size needed in future studies. Preliminary studies for determining the effect size and proper power analysis should be performed to guide this decision for a new study. Third, this work relies on manual analysis for the identification of all cell types (Supplementary Fig. 2) and mapping them to the prior knowledge. This process is labour-intensive, error-prone and may not identify all cell populations of interest[70]. In our future studies, we will combine state-of-the-art cell population identification algorithms[71–76] with our prior knowledge integrated to dynamically match clusters to the prior knowledge tensors for a more unbiased analysis. Fourth, this work only investigated the incorporation of prior knowledge into the EN algorithm. However, other methods can similarly be extended to incorporate expert knowledge (for example, see ref. [19] for a relevant extension of SVMs). In our future work, we will focus particularly on the incorporation of prior knowledge into deep learning methods[77,78]. Although these algorithms can model complex relationships that are valuable in high-throughput characterization of the immune system, the number of patients that are required for training a large neural network is often beyond the reach of typical immunological studies. We believe incorporation of prior immunological knowledge can reduce the number of patients required for implementation of deep learning approaches in clinical studies[79].

Additional research directions include exploring ensembles of prior knowledge integrated models, application of the iEN to domains outside of clinical immunology such as proteomics, metabolomics and transcriptomics, and application of domain-knowledge integrated models to multi-omic studies, which would provide a systems-level perspective on human biology[80].

Methods

Integration of immunological priors. The iEN framework extends the EN regularized regression method by integrating prior biological knowledge of cellular signal transduction into the coefficient optimization process. Consider an analysis with mass cytometry that generated features X, composed of observations (rows) $X^i = (x_{i1},...,x_{ip})^{\mathrm{T}}$ for $i = 1, 2, ..., n$, with each observation consisting of p measurements, where $p \in N$ and p is much greater than n. Corresponding to each observation is a value of interest y_i. Values of interest then constitute the response vector $Y = (y_1,...,y_n)^{\mathrm{T}}$. Response vectors are dataset-specific (for example, a vector of gestational age during pregnancy in the LTP example). A multivariable regression model can be constructed by computing the coefficients $\beta = (\beta_1,\beta_2,...,\beta_p)^{\mathrm{T}}$ that optimize the objective function, $L(\beta) = ||Y - X\beta||^2$. The EN method expands this definition with a linear combination of two regularization terms, the $L_1 = ||\beta||_1$ and $L_2 = ||\beta||^2$ penalties from the least absolute shrinkage and selection operator (LASSO) and ridge regression, respectively[81]. L_1 penalization reduces the model complexity and increases sparsity, while simultaneously selecting more descriptive features. However, it can select, at most, the number of observations when working in a high-dimensional space (specifically high-dimensional small observation size) and cannot select multiple, highly correlated features. L_2 penalization reduces the coefficient size and encourages the inclusion of highly correlated features, but cannot remove features completely. Incorporating both L_1 and L_2 regularization terms compensates for these issues. Penalization is applied to coefficients during model fitting and is determined by a penalization factor λ, as well as the ratio of penalization applied to each penalty term, α. The optimal ratio (α) and degree (λ) of penalization can be determined through optimization of the EN objective function:

$$L(\beta) = ||Y - X\beta||^2 + \lambda[(1-\alpha)||\beta||^2/2 + \alpha||\beta||_1]$$

EN models are agnostic to any information not included in X. However, the iEN incorporates a third parameter that encodes prior immunological knowledge: ϕ, a $p \times p$ diagonal matrix of the form $\mathrm{diag}(\phi) = \{\varphi_{1,1}, \varphi_{2,2}, \cdots, \varphi_{p,p}\}$ where $\varphi_{i,j} = 0\,\forall_{i \neq j}$. The ϕ factor guides models to be more consistent with the current understanding of signal transduction response. Biological priors compiled a priori by an independent panel of immunologists are used to prioritize certain signal transduction responses by scaling the features of X. The adapted model takes the form

$$L(\beta) = ||Y - X\phi\beta||^2 + \lambda[(1-\alpha)||\beta||^2/2 + \alpha||\beta||_1]$$

The biological priors, represented as a tensor of domain-specific knowledge, are manually constructed by a panel of experts. These biological priors are represented as a tensor of scores where features more consistent with known biology have higher values. These priors are a conservative indication of response from canonical signalling pathways that the field has a high level of confidence in observing. They are constructed as an $m \times l \times o$ tensor, $Z \in [0,1]^{m \times l \times o}$, where the associated mass cytometry assay consists of m cell types, l stimulations and o measured responses. An element in this tensor would correspond to a particular cell type and whether it will elicit a specific signalling response in response to each ex vivo stimulation. To make the connection between the prior tensor Z and the iEN parameter ϕ clear, consider the function $F(Z) \to \mathrm{diag}(\phi) \in \mathbb{R}^P_{[0,1]}$; that is to say, $\mathrm{diag}(\phi)$ is a vector of dimension $m \times l \times o = p$ with values in the range [0,1]. In other words, Z is transformed to a p-dimensional vector, where $\mathrm{diag}(\phi) = \{\varphi_{1,1}, \varphi_{2,2}, \cdots, \varphi_{p,p}\}$ such that $\varphi_{i,i} = \{\mathrm{e}^{-\varphi(1-z_i)}$, where z_i is the score of the ith immune feature, and φ is the amount of prioritization applied}. This formulation of $\varphi_{i,i} \in \mathrm{diag}(\phi)$ as $\mathrm{e}^{-\varphi(1-z_i)}$ affects features with lower prior value more than features with larger values. This definition allows for increased model stability than a formulation with $\varphi_{i,i} \in \mathrm{diag}(\phi)$ as $\mathrm{e}^{\varphi z_i}$ for large values of φ (Supplementary Fig. 1a,b).

Bayesian interpretation. The EN has a Bayesian representation[82] with priors over the estimates of β. This can help define the role of immunological priors in improving model predictions. The unnormalized version of this prior is reported as follows:

$$p(\beta|\lambda,\alpha) \propto \exp[-\lambda\{(1-\alpha)||\beta||^2 + \alpha||\beta||_1\}]$$

In the following, we show that the iEN has a similar interpretation, in which the prior distributions over β are altered according to the prioritization of biological knowledge, that is, the value of φ and the shape of Z. That is to say, our definition of the iEN can be represented as an alteration of the prior distribution over β given ϕ. The objective function of the iEN is $\hat{\beta} = \mathrm{argmin}(||Y - X\phi\beta||^2 + \lambda[(1-\alpha)||\beta||^2/2 + \alpha||\beta||_1])$. Now let $\bar{\beta} = \phi\beta$, then substitution for β results in the following optimization problem:

$$\hat{\beta} = \mathrm{argmin}(||Y - X\bar{\beta}||^2 + \lambda[(1-\alpha)||\phi^{-1}\bar{\beta}||^2/2 + \alpha||\phi^{-1}\bar{\beta}||_1])$$

From this formulation, the adjusted Bayesian prior for the iEN can be directly derived as follows:

$$p(\beta|\lambda,\alpha,\phi) \propto \exp[-\lambda\{(1-\alpha)||\phi^{-1}\bar{\beta}||^2 + \alpha||\phi^{-1}\bar{\beta}||_1\}]$$

To further illustrate the connection between iEN and the regular EN and their Bayesian interpretations, we show that the EN is a special case of iEN. For this, let us define the following two sets:

$$S_1 = \{P|\text{where}\, z_p = 1\}$$

$$S_2 = \{P|\text{where}\, z_p < 1\}$$

These sets indicate which estimates of $\hat{\beta}$ are affected by φ and which remain unaffected as previously defined. We can then subset the Z vector accordingly, with Z_{S1} being all biological priors of value one and Z_{S2} being all biological priors of value less than one. Therefore, we can separate the L_1 and L_2 norms according to these sets, reformulating the optimization problem as follows:

$$\begin{aligned}\hat{\beta} = \mathrm{argmin}\big(&\|Y - X\bar{\beta}\|^2 + \lambda\big[(1-\alpha)\|\bar{\beta}_{S_1}\|^2/2 + \alpha\|\bar{\beta}_{S_1}\|_1 \\ &+ (1-\alpha)(\mathrm{e}^{-\varphi(1-z_{s_2})})^{-2}\|\bar{\beta}_{S_2}\|^2/2 + \alpha(\mathrm{e}^{-\varphi(1-z_{s_2})})^{-1}\|\bar{\beta}_{S_2}\|_1\big]\big)\end{aligned}$$

Here, $\bar{\beta}_{S_1}$ and $\bar{\beta}_{S_2}$ represent the betas that are affected by the φ value. This allows for us to replace ϕ with $\mathrm{e}^{-\varphi(1-z)}$ respective of S_1 and S_2, which demonstrates how prioritization affects the estimation of β. Similar separation in the prior distribution also illustrates how priors over β are affected in the same manner:

$$\begin{aligned}p(\beta|\lambda,\alpha,\phi) \propto \exp\big[-\lambda\big\{&(1-\alpha)\|\bar{\beta}_{S_1}\|^2/2 + \alpha\|\bar{\beta}_{S_1}\|_1 \\ &+ (1-\alpha)(\mathrm{e}^{-\varphi(1-z_{s_2})})^{-2}\|\bar{\beta}_{S_2}\|^2/2 + \alpha(\mathrm{e}^{-\varphi(1-z_{s_2})})^{-1}\|\bar{\beta}_{S_2}\|_1\big\}\big]\end{aligned}$$

Because φ alters the prior distribution over β, this can be used to improve estimates of the true β in the Bayesian setting of iEN, as it does in the classic formulation. This also allows for the EN estimates to be included as a special case of the iEN when $\varphi = 0$.

Parameter optimization. iEN parameters were optimized over a grid of possible parameter values for each parameter (φ,α,λ). The φ search grid was generated from a logarithmic sequence with values between 0 and 100, which allows for the EN as a special case ($\varphi = 0$). Similarly, α was uniformly generated from values between 0 and 1. Generation of λ was done so that a large range of model sizes were tested during the 10-fold CV. Specifically, λ values were generated during the inner CV loop to avoid any possible information leak. The metrics used to justify parameter selection were the residual sum of squares for continuous response and the AUROC curve, as appropriate for each example presented.

Simulated data. All simulated data were generated using the '*simglm*' R package[83]. A total of 450 features were generated with a standard deviation of 15. However, 200 of these features had a mean value sampled from $N(N(0,10^2),15^2)$ to simulate features moderately associated with prior knowledge and 50 were sampled from $N(N(0,5^2),15^2)$, representing features highly associated with prior knowledge. The response variable was then generated as a linear combination of these 250 features. The additional 450 features represented features not associated with prior knowledge and were generated randomly using a normal distribution.

LTP cohort. Twenty-one pregnant women were included in the LTP study, all of whom received routine antepartum care at Lucile Packard Children's Hospital. Three patients were excluded from the study due to premature delivery (<37 weeks of gestation). Analysis was performed on the remaining 18 women who delivered at term (≥37 weeks of gestation). These 18 participants were aged 31.9 years (±3.4

(s.d.)). An independent cohort of 10 pregnant women who delivered at term were later enrolled as a validation cohort.

ChP cohort. A total of 30 patients were enrolled in the study of ChP: 15 healthy controls and 15 patients with ChP receiving treatment at Bell Dental Centre (San Leandro) and Stanford University School of Medicine. Two participants were excluded from the analysis, one patient due to autoimmune disease and one control due to onset of hand infection during the study. The final cohort consisted of 14 patient (aged 42.2 ± 10.5 years) and 14 control (36.5 ± 8.07 years) samples, each of which were split by gender: eight female and six male.

Whole blood sample processing. Whole blood samples were collected in 10-ml heparin-containing tubes and processed within 1 h of collection. Samples for the LTP cohort were stimulated with either 1 μg ml^{-1} of LPS, 100 ng ml^{-1} of IFN-α or a cocktail of 100 ng ml^{-1} IL (IL-2, IL-6), or they were left unstimulated to measure endogenous cellular activity. Samples for the ChP cohort were stimulated with LPS, IFN-α, tumour necrosis factor-α or a cocktail of IL-2, IL-4, IL-6 and granulocyte-macrophage colony-stimulating factor, or left unstimulated. Samples were fixed using a stabilization buffer (SmartTube) according to the manufacturer's instructions and stored at −80 °C until further processing.

Mass cytometry analysis. Post-thaw, fixed samples were added to an erythrocyte lysis buffer (SmartTube) and underwent two rounds of erythrocyte lysis. Cells were then barcoded as previously described[84]. In summary, 20-well barcode plates were prepared with a combination of two Pd isotopes out of a pool of six (^{102}Pd, ^{104}Pd, ^{105}Pd, ^{106}Pd, ^{108}Pd, ^{110}Pd) and added to the cells in 0.02% saponin/phosphate buffered saline. Samples were pooled and stained with metal-conjugated antibodies collectively to minimize experimental variation. The panels for the different cohorts are listed in Supplementary Tables 3 and 4. Intracellular staining was performed in methanol-permeabilized cells. Cells were incubated overnight at 4 °C with an iridium-containing intercalator (Fluidigm). Before mass cytometry analysis, cells were filtered through a 35-μm membrane and resuspended in a solution of normalization beads (Fluidigm).

Barcoded and stained cells were analysed on a Helios mass cytometer (Fluidigm) at an event rate of 500 to 1,000 cells per second. The data were normalized using Normalizer v0.1 MATLAB Compiler Runtime[85] and debarcoded with a single-cell MATLAB debarcoding tool[84]. Gating was performed using Cytobank (cytobank.org). Gating strategies for the different cohorts are shown in Supplementary Fig. 4.

Correlation network visualization. All datasets were visualized using correlation network structures. Each immune feature was denoted by a node and the network layout was calculated using the t-SNE algorithm applied to the complete adjacency matrix. For visualization purposes, only the edges with Bonferroni-corrected Pearson correlation P values of less than 0.05 were visualized.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this Article.

Data availability

Longitudinal term pregnancy raw data, processed data and source code for reproduction of the results are publicly available at http://flowrepository.org/id/FR-FCM-ZY3Q and http://flowrepository.org/id/FR-FCM-ZY3R for the original and validation studies, respectively. Similarly, chronic periodontitis raw data, processed data and source code for reproduction of the results are publicly available at https://flowrepository.org/id/FR-FCM-ZYT6.

Code availability

The iEN source code as well as scripts for reproduction of the results are available through: https://nalab.stanford.edu/immunological-elastic-net/ and https://github.com/Teculos/immunological-EN under an MIT licence with https://doi.org/10.5281/zenodo.3885868.

Received: 13 January 2020; Accepted: 26 August 2020;
Published online: 12 October 2020

References

1. Davis, M. M., Tato, C. M. & Furman, D. Systems immunology: just getting started. *Nat. Immunol.* **18**, 725–732 (2017).
2. Rieckmann, J. C. et al. Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat. Immunol.* **18**, 583–593 (2017).
3. Mathew, D. et al. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science* (2020); https://doi.org/10.1126/science.abc8511.
4. Wilk, A. J. et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070–1076 (2020).
5. Porter, D. L., Levine, B. L., Kalos, M., Bagg, A. & June, C. H. Chimeric antigen receptor-modified T cells in chronic lymphoid leukemia. *New Engl. J. Med.* **365**, 725–733 (2011).
6. Ryu, J. K. et al. Fibrin-targeting immunotherapy protects against neuroinflammation and neurodegeneration. *Nat. Immunol.* **19**, 1212–1223 (2018).
7. Saphire, E. O., Schendel, S. L., Gunn, B. M., Milligan, J. C. & Alter, G. Antibody-mediated protection against Ebola virus. *Nat. Immunol.* **19**, 1169–1178 (2018).
8. Krutzik, P. O. & Nolan, G. P. Intracellular phospho-protein staining techniques for flow cytometry: monitoring single cell signaling events. *Cytometry A* **55**, 61–70 (2003).
9. Nettey, L., Giles, A. J. & Chattopadhyay, P. K. OMIP-050: a 28-color/30-parameter fluorescence flow cytometry panel to enumerate and characterize cells expressing a wide array of immune checkpoint molecules. *Cytometry A* **93**, 1094–1096 (2018).
10. Chattopadhyay, P. K., Winters, A. F., Lomas, W. E., Laino, A. S. & Woods, D. M. High-parameter single-cell analysis. *Annu. Rev. Anal. Chem.* **12**, 411–430 (2019).
11. Bandura, D. R. et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81**, 6813–6822 (2009).
12. Bendall, S. C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
13. Finak, G. et al. Standardizing flow cytometry immunophenotyping analysis from the human immunophenotyping consortium. *Sci. Rep.* **6**, 20686 (2016).
14. Newell, E. W. & Cheng, Y. Mass cytometry: blessed with the curse of dimensionality. *Nat. Immunol.* **17**, 890–895 (2016).
15. Jain, A. K., Duin, P. W. & Mao, Jianchang Statistical pattern recognition: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 4–37 (2000).
16. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* 2nd edn (Springer, 2016).
17. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
18. Li, J., Liu, L., Le, T. D. & Liu, J. Accurate data-driven prediction does not mean high reproducibility. *Nat. Mach. Intell.* **2**, 13–15 (2020).
19. Krupka, E. & Tishby, N. Incorporating prior knowledge on features into learning. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics* (eds Meila, M. & Shen, X.) Vol. 2, 227–234 (PMLR, 2007).
20. Mollaysa, A., Strasser, P. & Kalousis, A. Regularising non-linear models using feature side-information. In *Proceedings of the 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) Vol. 70, 2508–2517 (PMLR, 2017).
21. Tai, F. & Pan, W. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics* **23**, 3170–3177 (2007).
22. Bergersen, L. C., Glad, I. K. & Lyng, H. Weighted LASSO with data integration. *Stat. Appl. Genet. Mol. Biol.* **10** (2011); https://doi.org/10.2202/1544-6115.1703
23. Handl, L., Jalali, A., Scherer, M., Eggeling, R. & Pfeifer, N. Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data. *Bioinformatics* **35**, i154–i163 (2019).
24. Zuo, Y., Yu, G. & Ressom, H. W. Integrating prior biological knowledge and graphical LASSO for network inference. In *2015 IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*) 1543–1547 (IEEE, 2015); https://doi.org/10.1109/BIBM.2015.7359905
25. Guan, X. & Liu, L. Know-GRRF: domain-knowledge informed biomarker discovery with random forests. In *Bioinformatics and Biomedical Engineering: 6th International Work-Conference, IWBBIO 2018, Granada, Spain, 2018, Proceedings, Part II* (eds Rojas, I. & Ortuño, F.) Vol. 10814, 3–14 (2018).
26. Shi, J., Zhang, S. & Qiu, L. Credit scoring by feature-weighted support vector machines. *J. Zhejiang Univ. Sci. C* **14**, 197–204 (2013).
27. Sarafianos, N., Vrigkas, M. & Kakadiaris, I. A. Adaptive SVM+: learning with privileged information for domain adaptation. In *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops* (*ICCVW*) 2637–2644 (IEEE, 2017); https://doi.org/10.1109/ICCVW.2017.313
28. Xing, H., Ha, M., Hu, B. & Tian, D. Linear feature-weighted support vector machine. *Fuzzy Inf. Eng.* **1**, 289–305 (2009).
29. Bhattacharya, G., Ghosh, K. & Chowdhury, A. S. Granger causality driven AHP for feature weighted knn. *Pattern Recogn.* **66**, 425–436 (2017).
30. Mollaysa, A., Kalousis, A., Bruno, E. & Diephuis, M. Learning to augment with feature side-information. In *Proceedings of the 11th Asian Conference on Machine Learning* (*PMLR*) Vol. 101, 173–187 (PMLR, 2019).
31. Ye, Y., Li, H., Deng, X. & Huang, J. Z. *Feature Weighting Random Forest for Detection of Hidden Web Search Interfaces* (ACL, 2008); https://www.aclweb.org/anthology/O08-6001.pdf

32. Zhang, W., Chien, J., Yong, J. & Kuang, R. Network-based machine learning and graph theory algorithms for precision oncology. *NPJ Precis. Oncol.* **1**, 25 (2017).
33. Sinha, S. Integration of prior biological knowledge and epigenetic information enhances the prediction accuracy of the Bayesian Wnt pathway. *Integr. Biol. (Camb.)* **6**, 1034–1048 (2014).
34. Fabris, F. & Freitas, A. A. New KEGG pathway-based interpretable features for classifying ageing-related mouse proteins. *Bioinformatics* **32**, 2988–2995 (2016).
35. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* (2005); https://doi.org/10.1111/j.1467-9868.2005.00527.x
36. Hegre, H., Metternich, N. W., Nygård, H. M. & Wucherpfennig, J. Introduction. *J. Peace Res.* **54**, 113–124 (2017).
37. Madhukar, N. S. et al. A Bayesian machine learning approach for drug target identification using diverse data types. *Nat. Commun.* **10**, 5221 (2019).
38. Sharpless, N. E. & Depinho, R. A. The mighty mouse: genetically engineered mouse models in cancer drug development. *Nat. Rev. Drug Discov.* **5**, 741–754 (2006).
39. Zhu, F., Nair, R. R., Fisher, E. M. C. & Cunningham, T. J. Humanising the mouse genome piece by piece. *Nat. Commun.* **10**, 1845 (2019).
40. Meier, L., Van De Geer, S. & Bühlmann, P. The group lasso for logistic regression. *J. R. Stat. Soc. B* **70**, 53–71 (2008).
41. Velten, B. & Huber, W. Adaptive penalization in high-dimensional regression and classification with external covariates using variational Bayes. *Biostatistics* (2019); https://doi.org/10.1093/biostatistics/kxz034.
42. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* (Springer, 2002); https://doi.org/10.1007/978-0-387-21706-2
43. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
44. Breiman, L. Random forests. *Mach. Learning* **45**, 5–32 (2001).
45. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
46. van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Stat. Appl. Genet. Mol. Biol.* **6**, 25 (2007).
47. Silvennoinen, O., Ihle, J. N., Schlessinger, J. & Levy, D. E. Interferon-induced nuclear signalling by Jak protein tyrosine kinases. *Nature* **366**, 583–585 (1993).
48. Ivashkiv, L. B. & Donlin, L. T. Regulation of type I interferon responses. *Nat. Rev. Immunol.* **14**, 36–49 (2014).
49. Boyman, O. & Sprent, J. The role of interleukin-2 during homeostasis and activation of the immune system. *Nat. Rev. Immunol.* **12**, 180–190 (2012).
50. Hunter, C. A. & Jones, S. A. IL-6 as a keystone cytokine in health and disease. *Nat. Immunol.* **16**, 448–457 (2015).
51. Beutler, B. A. TLRs and innate immunity. *Blood* **113**, 1399–1407 (2009).
52. Park, J. M. et al. Signaling pathways and genes that inhibit pathogen-induced macrophage apoptosis–CREB and NF-κB as key regulators. *Immunity* **23**, 319–329 (2005).
53. Kadowaki, N. et al. Subsets of human dendritic cell precursors express different toll-like receptors and respond to different microbial antigens. *J. Exp. Med.* **194**, 863–869 (2001).
54. Adib-Conquy, M., Scott-Algara, D., Cavaillon, J.-M. & Souza-Fonseca-Guimaraes, F. TLR-mediated activation of NK cells and their role in bacterial/viral immune responses in mammals. *Immunol. Cell Biol.* **92**, 256–262 (2014).
55. Caramalho, I. et al. Regulatory T cells selectively express Toll-like receptors and are activated by lipopolysaccharide. *J. Exp. Med.* **197**, 403–411 (2003).
56. Aghaeepour, N. et al. An immune clock of human pregnancy.*Sci. Immunol.* **2**, eaan2946 (2017).
57. Deshmukh, H. & Way, S. S. Immunological basis for recurrent fetal loss and pregnancy complications. *Annu. Rev. Pathol.* **14**, 185–210 (2018).
58. Arck, P. C. & Hecher, K. Fetomaternal immune cross-talk and its consequences for maternal and offspring's health. *Nat. Med.* **19**, 548–556 (2013).
59. Romero, R., Dey, S. K. & Fisher, S. J. Preterm labor: one syndrome, many causes. *Science* **345**, 760–765 (2014).
60. Paquette, A. G., Hood, L., Price, N. D. & Sadovsky, Y. Deep phenotyping during pregnancy for predictive and preventive medicine. *Sci. Transl. Med.* **12**, eaay1059 (2020).
61. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
62. Pihlstrom, B. L., Michalowicz, B. S. & Johnson, N. W. Periodontal diseases. *Lancet* **366**, 1809–1820 (2005).
63. Eke, P. I. et al. Update on prevalence of periodontitis in adults in the United States: NHANES 2009 to 2012. *J. Periodontol.* **86**, 611–622 (2015).
64. Kassebaum, N. J. et al. Global burden of severe periodontitis in 1990–2010: a systematic review and meta-regression. *J. Dent. Res.* **93**, 1045–1053 (2014).
65. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
66. Meyer, D., Dimitriadou, E., Hornik, K. & Leisch, F. *Package e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)* (TU Wien, 2019).
67. Littmann, M. et al. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nat. Mach. Intell.* (2020); https://doi.org/10.1038/s42256-019-0139-8.
68. Vapnik, V. & Vashist, A. A new learning paradigm: learning using privileged information. *Neural Netw.* **22**, 544–557 (2009).
69. Kveler, K. et al. Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed. *Nat. Biotechnol.* **36**, 651–659 (2018).
70. Aghaeepour, N. et al. Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10**, 228–238 (2013).
71. Lux, M. et al. flowLearn: fast and precise identification and quality checking of cell populations in flow cytometry. *Bioinformatics* **34**, 2245–2253 (2018).
72. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
73. Van Gassen, S. et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**, 636–645 (2015).
74. Qiu, P. et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29**, 886–891 (2011).
75. Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L. & Nolan, G. P. Automated mapping of phenotype space with single-cell data. *Nat. Methods* **13**, 493–496 (2016).
76. Stanley, N. et al. VoPo leverages cellular heterogeneity for predictive modeling of single-cell data. *Nat. Commun.* **11**, 3738 (2020).
77. Ding, X. et al. Prior knowledge-based deep learning method for indoor object recognition and application. *Syst. Sci. Control Eng.* **6**, 249–257 (2018).
78. Xu, Z., Liu, B., Wang, B., Sun, C. & Wang, X. Incorporating loose-structured knowledge into conversation modeling via recall-gate LSTM. In *Proceedings of the 2017 International Joint Conference on Neural Networks* (*IJCNN*) 3506–3513 (IEEE, 2017); https://doi.org/10.1109/IJCNN.2017.7966297
79. Diligenti, M., Roychowdhury, S. & Gori, M. Integrating prior knowledge into deep learning. In *Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications* (*ICMLA*) 920–923 (IEEE, 2017); https://doi.org/10.1109/ICMLA.2017.00-37
80. Ghaemi, M. S. et al. Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy. *Bioinformatics* **35**, 95–103 (2019).
81. Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
82. Hans, C. Elastic net regression modeling with the orthant normal prior. *J. Am. Stat. Assoc.* **106**, 1383–1393 (2011).
83. LeBeau, B. *simglm*: *Simulate Models Based on the Generalized Linear Model* (CRAN, 2019).
84. Zunder, E. R. et al. Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nat. Protoc.* **10**, 316–333 (2015).
85. Finck, R. et al. Normalization of mass cytometry data with bead standards. *Cytometry A* **83**, 483–494 (2013).
86. Pacella, I. et al. IFN-α promotes rapid human Treg contraction and late Th1-like Treg decrease. *J. Leukoc. Biol.* **100**, 613–623 (2016).
87. Metidji, A. et al. IFN-α/β receptor signaling promotes regulatory T cell development and function under stress conditions. *J. Immunol.* **194**, 4265–4276 (2015).
88. Scheller, J., Chalaris, A., Schmidt-Arras, D. & Rose-John, S. The pro- and anti-inflammatory properties of the cytokine interleukin-6. *Biochim. Biophys. Acta* **1813**, 878–888 (2011).
89. Heinrich, P. C. et al. Principles of interleukin (IL)-6-type cytokine signalling and its regulation. *Biochem. J.* **374**, 1–20 (2003).

Acknowledgements

This study was supported by the March of Dimes Prematurity Research Center at Stanford (22-FY18-808), the Bill and Melinda Gates Foundation (OPP1112382, OPP1189911 and OPP1113682), the Department of Anesthesiology, Perioperative and Pain Medicine at Stanford University, the Robertson Foundation, NIH (1R01HL13984401A1, R35GM138353), the American Heart Association (18IPA34170507 and 19PABHI34580007), the Food and Drug Administration (FDA; HHSF223201610018C), the Burroughs Wellcome Fund (1019816) and the National Institutes of Health (NIH; R01AG058417, R01HL13984401, R21DE02772801 and R61NS114926). N.A., D.R.M. and G.P.N. were supported by US FDA contract no. HHSF223201610018C. N.S. was supported by a Stanford Immunology Training Grant (5 T32 AI07290-33). B.G. was supported by the NIH (K23GM111657) and the Doris Duke Charitable Foundation (2018100). D.G. and B.G. were supported by NIH R21DE02772801. L.P. and X.H. were supported by the Stanford Maternal and Child Health Research Institute. The authors are solely responsible for the content of this Article, which does not necessarily represent the official views of the US Department of Health and Human Services (HHS).

Author contributions

A.C. conducted the data analysis and software development, generated figures and contributed to writing the manuscript. A.S.T. performed experiments, processed data, wrote the manuscript and produced figures. N.S., M.B., M.S.G., R.F., H.N., T.P., I.M., A.L.C., C.E. and M.X. contributed to the analysis plan, figure design and manuscript revisions. E.G., L.P., X.H., I.A.S., K.A. and D.G. designed and performed experiments. D.R.M., A.T., G.M.S., D.K.S., S.B., K.L.D., W.F., G.P.N., T.H. and R.T. contributed to the design and evaluation of the algorithm, and edited the manuscript. M.S.A. and B.G. coordinated the effort to collect and analyse biological data and interpreted the results, and contributed to writing the manuscript. N.A. conceived, designed and coordinated the study, interpreted data, and contributed to writing the manuscript. All authors read and approved the Article.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-020-00232-8.

Correspondence and requests for materials should be addressed to N.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Nima Aghaeepour

Last updated by author(s): Sept 3, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Manual gating was performed using http://cytobank.org. All other data processing was performed in R (v. 3.6.0) as described in the methods section.
Data analysis	All data processing was performed in R (v. 3.6.0) as described in the methods section. Please see the Software Policy Checklist for details.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All datasets are publicly available with no restrictions as described under the data availability section of the main text.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   □ Behavioural & social sciences   □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Not applicable
Data exclusions	No data was excluded
Replication	As described in the main text, all models were cross-validated and sensitivity analysis was performed using repeated randomization. In addition, the pregnancy study was validated using an independent set of patients.
Randomization	Not applicable
Blinding	Not applicable

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Human research participants
□	☒	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
□	☒	Flow cytometry
☒	□	MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Not applicable
Study protocol	For term pregnancy all information is associated with doi: 10.1126/sciimmunol.aan2946. Similarly, Chonic periodontitis study protocol is located at doi: 10.1177/0022034519857714
Data collection	Patient blood samples for the term pregnancy and chronic periodontitis study were collected at two Stanford facilities: the Lucile Packard Children's hospital, and the Bell Dental Center. Please see the articles linked above for details.
Outcomes	Please see the articles linked above.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Post-thaw, fixed samples were added to an erythrocyte lysis buffer (SmartTube Inc) and underwent two rounds of erythrocyte lysis. Cells were then barcoded. In summary, twenty-well barcode plates were prepared with a combination of 2 Pd isotopes out of a pool of six (102Pd, 104Pd, 105Pd, 106Pd, 108Pd, 110 563 Pd) and added to the cells in 0.02% saponin/ PBS. Samples were pooled and stained with metal565 conjugated antibodies collectively to minimize experimental variation. The panel for the different cohorts are listed in Supplementary Tables 3, and 4. Intracellular staining was performed in methanol-permeabilized cells. Cells were incubated overnight at 4°C with an iridium-containing intercalator (Fluidigm). Prior to mass cytometry analysis, cells were filtered through a 35µm membrane and resuspended in a solution of normalization beads (Fluidigm). Barcoded and stained cells were analysed on a Helios Mass Cytometer (Fluidigm) at an event rate of 500 to 1000 cells per second. The data was normalized using Normalizer v0.1 MATLAB Compiler Runtime89 and debarcoded with a single-cell MATLAB debarcoding tool8. Gating was performed using Cytobank (cytobank.org).
Instrument	Helios Mass Cytometer
Software	CyTOF Software v.7.0 and CellEngine. Longitudinal term pregnancy raw data, processed data, and source code for reproduction of the results are publicly available at http://flowrepository.org/id/FR-FCM-ZY3Q and http:// flowrepository.org/id/FR-FCM-ZY3R for the original and validation studies, respectively. Similarly, chronic periodontitis raw data, processed data, and source code for reproduction of the results are publicly available at https://flowrepository.org/id/ FR-FCM-ZYT6.
Cell population abundance	Not applicable
Gating strategy	Two-dimensional dot plots depicted for a representative sample. Gating was performed using ImmuneAtlas (https:// cellengine.com). Live singlet leukocytes were derived from DNA+CD235ab–CD61–cPARP– events, which are split into mononuclear cells (CD45+CD66–) and neutrophils (CD45+CD66+). Adaptive cell populations are gated from the mononuclear cells (CD19+CD3– B cells and CD19–CD3+ T cells). T cells are divided into TCRgd+, CD4+, and CD8+ subpopulations. Among the CD4+ T cells, T helper (Tbet+) regulatory T (FoxP3+CD25+), naïve (CD45RA+), and memory (CD45RA–) cells are identified. Among the CD8+ T cells, naïve (CD45RA+), memory (CD45RA–), Tbet+CD45RA+ and Tbet+CD45RA- cells are identified. The CD19–CD3– population is further split into CD7+ NK cells (CD56dimCD16+ and CD56+CD16–, respectively) and CD14+CD16– classical monocytes, CD14+CD16+ intermediate monocytes, and CD14–CD16+ non-classical monocytes. A subpopulation of myeloid-derived suppressor cells (HLA-DRlo) is derived from classical monocytes. Dendritic cell populations are defined as CD14–CD16–HLA-DR+, and subdivided into CD11c+ mDC, and CD123+ pDC. Please see supplemental materials for details.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.