SIMON: open-source knowledge discovery platform

Adriana Tomic[1,2*], Ivan Tomic[3*], Levi Waldron[4,5], Ludwig Geistlinger[4,5], Max Kuhn[6], Rachel L. Spreng[7], Lindsay C. Dahora[7], Kelly E. Seaton[7], Georgia Tomaras[7], Jennifer Hill[1], Niharika A. Duggal[8], Ross D. Pollock[9], Norman R. Lazarus[9], Stephen D.R. Harridge[9], Janet M. Lord[8,10], Purvesh Khatri[2,11], Andrew J. Pollard[1#] and Mark M. Davis[2,12,13#]

[1]Oxford Vaccine Group, Department of Pediatrics, University of Oxford, Oxford, UK. [2]Institute of Immunity, Transplantation and Infection, Stanford University School of Medicine, Stanford, CA, USA. [3]Deep Medicine, Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK. [4]Graduate School of Public Health and Health Policy and [5]Institute for Implementation Science and Population Health, City University of New York, NY, USA. [6]RStudio, PBC, Boston, MA, USA. [7]Duke Human Vaccine Institute, Duke University, Durham, NC, USA. [8]MRC-Versus Arthritis Centre for Musculoskeletal Ageing Research, Institute of Inflammation and Ageing, University of Birmingham Research Labs, Birmingham, UK. [9]Centre for Human and Applied Physiological Sciences, King's College London, UK. [10]NIHR Birmingham Biomedical Research Centre, University Hospital Birmingham NHS Foundation Trust and University of Birmingham, UK. [11]Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, CA, USA. [12]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA. [13]Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA.

*Co-first and corresponding authors: Adriana Tomic (info@adrianatomic.com) and Ivan Tomic (info@ivantomic.com)

#Co-senior authors: Andrew J Pollard (andrew.pollard@paediatrics.ox.ac.uk) and Mark M. Davis (mmdavis@stanford.edu)

Abstract:

Data analysis and knowledge discovery has become more and more important in biology and medicine with the increasing complexity of the biological datasets, but necessarily sophisticated programming skills and in-depth understanding of algorithms needed pose barriers to most biologists and clinicians to perform such research. We have developed a modular open-source software SIMON to facilitate the application of 180+ state-of-the-art machine learning algorithms to high-dimensional biomedical data. With an easy to use graphical user interface, standardized pipelines, automated approach for machine learning and other statistical analysis methods, SIMON helps to identify optimal algorithms and provides a resource that empowers non-technical and technical researchers to identify crucial patterns in biomedical data.

Main text:

Over the past several years, due to the technological breakthroughs in genome sequencing[1], high-dimensional flow cytometry[2-4], mass cytometry[5,6], and multi-parameter microscopy[7,8] the amount and complexity of biological data has become increasingly intractable and it is no longer feasible to extract knowledge without using sophisticated computer algorithms. Therefore, researchers are in need of novel computational approaches that can cope with complexity and heterogeneity of data in an objective and unbiased way. Machine learning (ML), a subset of artificial intelligence, is a computational approach developed to identify patterns from the data in order to make predictions on new data[9]. ML has a profound impact on biological research[10-12], including genomics[13], proteomics[14-16], cell image analysis[17], drug discovery and development[18], and cell phenotyping[6,19,20] which revolutionized our understanding of biological complexity. Recently, using systems-level analysis of genetic, transcriptional, and proteomic signatures to predict patients' response to vaccines[21,22], therapies and disease progression[23-27], ML has become primary computational approach used in the 'precision medicine'[28].

The biggest challenge is the proper application of ML methods and the translation of the results into meaningful insights. The analysis of massive datasets and extraction of knowledge using ML requires knowledge of many different computational libraries for data pre-processing and cleaning, data partitioning, model building and tuning, evaluation of the performance for the model and minimizing overfitting[11]. Tools to achieve these tasks have been mainly developed either in R (https://www.r-project.org/)[29,30] or Python (www.python.org/)[31], which have today become leading statistical programming languages in data science. Because R and Python are free and open-source, they have been quickly adopted by a large community of programmers who are building new libraries and improving existing ones. As of May 2020, there are 15,658 R packages available in the CRAN package repository (https://cran.r-project.org/). Many of the packages offer different modeling functions and have different syntax for model training, predictions and determination of variable importance. Due to the lack of a unified method for proper application of ML process, even experienced bioinformaticians struggle with these time-consuming ML tasks. To provide a uniform interface and standardize the process of building predictive models, ML libraries were developed, for example mlr3[32] (https://mlr3.mlr-org.com), the **c**lassification **a**nd **re**gression **t**raining (caret)[30,33] (https://rdrr.io/cran/caret), scikit-learn[34] (https://scikit-learn.org), mlPy[35] (https://mlpy.fbk.eu), SciPy (https://www.scipy.org/) including also ones for deep learning, such as TensorFlow (https://www.tensorflow.org/), PyTorch (https://pytorch.org/) and Keras (https://keras.io/). Since those libraries do not have a graphical user interface, usage requires extensive programming experience and general knowledge of R or Python making it inaccessible for many life science researchers. Therefore, there

is an increased effort to harmonize those libraries and develop a software that will facilitate application of ML in life sciences.

The software should provide a standardized ML method for data pre-processing, data partitioning, building predictive models, evaluation of model performance and selection of features. Moreover, such software should be adapted for biological datasets that have high percentage of missing values[36], have imbalanced participant distributions (i.e. having a high number of infected subjects, but only a relatively small number of healthy controls)[37] or suffer from a *"curse of dimensionality"*, i.e. poor predictive power, as can be observed when the number of features is much greater than the number of samples[38]. Additionally, beyond ML process, the software should support exploratory analysis and visualization of the results using user-friendly graphical interface. The fast-paced technological development dramatically increased size of biological datasets and computational power needed for analysis. Therefore, open-source web-based software supporting cloud processing architecture is essential. Additionally, software should support an automated ML[39] (autoML) process that rapidly builds high-performance predictive models by identifying optimal ML method, including selection of an appropriate algorithm, optimization of model hyperparameters and evaluation of the best-performing models. AutoML improves the efficiency of ML process and resulting models often outperform hand-designed models[39,40].

To address these challenges, we developed SIMON (Sequential Iterative Modeling "Over Night"), a free and open-source software for application of ML in life sciences that facilitates production of high-performing ML models and allows researchers to focus on knowledge discovery process. SIMON provides a user-friendly, uniform interface for building and evaluating predictive models using a variety of ML algorithms. Currently, there are 182 different ML algorithms available (**Supplementary table 1**). The entire ML process which is based on the caret[33] library, from model building and evaluation to feature selection in SIMON is fully automated. This allows advanced ML users to focus on other important aspects necessary to build highly accurate models, such as data preprocessing, feature engineering and model deployment. It also makes the entire ML process more accessible to domain-knowledge experts that formulated the research hypothesis and collected the data, but lack programming ML skills Additionally, to prevent optimistic accuracy estimates and to optimize the model for generalization to unseen data, SIMON introduces unified process for model training, hyperparameter tuning and model evaluation by generation of training, validation and test sets. Training set is used for building models, validation set is used for hyperparameter tuning and finally, models are evaluated in an unbiased way using the test, also known as holdout set that has never been used in training. Beside the standardized ML process, the initial install version offers a set of core components specifically suited for analysis of biomedical data, such as multi-set intersection function for integration of data with many missing values[41] (https://cran.r-

project.org/web/packages/mulset/index.html), method for identifying differentially expressed genes using significance analysis in microarrays (SAM)[42], graphical representation of the clustering analysis important for detection of batch effects, graphical display of the correlation analysis and graphical visualizations of the ML results that can be downloaded as publication-ready figures in scalable vector graphics (SVG) format. Finally, SIMON is available in two versions as a single mode and a server version. The single mode is developed as a SIMON Docker container (https://www.docker.com/), ensuring code reproducibility and solving installation compatibility issues across major operating systems (Windows, MacOS and Linux). In both versions parallel computing is supported which is essential for more efficient ML analysis by distributing the workload across several processors. To promote collaboration, data sharing and support distributed cloud processing, SIMON is also available as a server version. The server version can be installed on a private or a public Linux cloud service. Distributed cloud processing (multiNode) is implemented utilizing OpenStack, a free and open source cloud computing platform (https://www.openstack.org/). The advantage of the server versions is that it has multiNode capability which allows users to distribute workload on multiple computers simultaneously to optimize SIMON performance. The multiNode process can be used to horizontally scale analysis to large infrastructure, such as high performance computing clusters to meet the computational needs and accommodate parallel processing of large amounts of data. Additionally, in the server version, users can configure data storage either on a local server or in a cloud using service which is interoperable with Amazon Web Services S3 application programming interface (AWS S3 API)[43]. SIMON is also translated into multiple languages by collaborative open-source effort. SIMON source code is regularly updated, and both source code and compiled software are available from the project's website at http://www.genular.org/. Overall, SIMON is designed to provide a uniform knowledge discovery interface adaptable to the increasing size of biomedical datasets allowing data scientists, bioinformaticians and domain-knowledge experts to solve biological research questions.

We demonstrate the accuracy, ease of use and power of SIMON on five different biomedical datasets and build predictive models for arboviral infection severity (SISA)[44], the identification of the cellular immune signature associated with a high-level of physical activity (Cyclists)[45], the determination of the humoral responses that mediate protection against *Salmonella* Typhi infection (VAST)[46], early-stage detection of colorectal cancer from microbiome data (Zeller)[47,48], and for the detection of liver hepatocellular carcinoma cells (LIHC)[49] (**Fig. 1 b, c, d, e, Supplementary protocol**). To build models using the SISA dataset (described in the **Supplementary methods** and available as **Supplementary table 2**), 11 ML algorithms were used, five from the original publication[44] (treebag, k nearest neighbors, random forest, stochastic generalized boosting model and neural network) and additionally, '*sda',* shrinkage discriminant analysis;

'hdda', high dimensional discriminant analysis; *'svmLinear2'*, support vector machine with linear kernel; *'pcaNNet'*, neural networks with feature extraction; *'LogitBoost',* boosted logistic regression and naïve Bayes. Due to the unified ML process for training, tuning and evaluating predictive models, users can test a variety of ML algorithms in SIMON. Since the same training and test sets are used by different algorithms, resulting models can be compared and the best performing models can be selected. After manually setting initial parameters for data partitioning, predictor and outcome variables, exploratory classes, pre-processing and selecting ML algorithms (**Fig. 1a**), SIMON automatically performs all necessary ML analysis steps to build, tune and evaluate predictive models. The process of building all 11 models on the SISA dataset in SIMON finished in 59 sec on a standard laptop (Intel® Core™ i7 Processor 7700HQ and 16 GB of RAM). In SIMON, users can evaluate model performance using standard performance measurements such as accuracy, sensitivity, specificity, precision, recall, area under the receiver operating characteristic curve (AUROC), precision-recall area under curve (prAUC), and logarithmic loss (LogLoss) on training and holdout, test sets (**Fig. 1b**). The shrinkage discriminant analysis model ('*sda'*) had the highest AUROC of 0.97 on the training set and also performed well on the holdout, test set (test AUROC 0.96) (**Fig. 1c, Supplementary table 3,** the model is available as the **Supplementary data 1**).

To demonstrate SIMON's capabilities for analyzing biomedical datasets with missing data, we applied SIMON to (i) the Cyclists dataset studying the impact of physical activity on the immune system in adulthood[45] (**Supplementary table 4**) and (ii) the VAST dataset collected from a clinical trial which was undertaken to evaluate typhoid vaccine efficacy[50] (**Supplementary table 5**). Description of both datasets is available in **Supplementary methods.** The percentage of missing values was 8% in the Cyclists dataset and 21% in the VAST dataset either due to the exclusion of samples not passing quality control criteria or the lack of sample volume to repeat experiments and obtain reportable data. To build models using the datasets with missing values, we used the multi-set intersection (*mulset*) function[41] to identify shared features between donors and generate resamples (**Supplementary protocol**). Because *mulset* function generates multiple resamples from the initial dataset based on shared features, it is useful for removal of missing values and can be used for integration of data collected from different assays and across clinical studies[41]. For the Cyclists dataset, the *mulset* function generated 146 resamples. The models were built for each of the 146 resamples using five ML algorithms (naïve Bayes, svmLinear2, pcaNNet, logistic regression and hdda) to identify immune cell subsets enriched in the cohort of master cyclists. The analysis finished in 41 min and 24 sec. The model with the highest performance measures was built with naïve Bayes on the resample with 96 donors that shared 31 features (train AUROC 0.99 and test AUROC 1) (**Fig. 1d, Supplementary table 6 and Supplementary data 2**). The *mulset* function generated 206 resamples from the initial VAST dataset with varying number of donors and features. Resamples with less than 10

donors in the test set were removed prior ML process to prevent too optimistic predictive estimates using the holdout set. Therefore, the ML analysis was performed on 58 resamples using same five ML algorithms as for the Cyclists dataset. The entire analysis finished in 31 min and 1 sec. The top performing model was built on the resample with 47 donors that shared 13 features with the naïve Bayes algorithm (train AUROC 0.73 and test AUROC 0.71) (**Fig. 1d**, **Supplementary table 7 and Supplementary data 3**).

We also applied SIMON to (i) a dataset with a large number of features measured using whole-metagenome shotgun sequencing of fecal samples (Zeller dataset, **Supplementary table 8**), and (ii) the liver hepatocellular carcinoma dataset from TCGA with an imbalanced sample distribution of tumor and adjacent normal tissue samples (LIHC dataset, **Supplementary table 9**). Both datasets are described in **Supplementary methods**. For the Zeller dataset, models were built using ML algorithms known to perform well in the situations where more features were measured than individuals, such as shrinkage discriminant analysis[51], high dimensional discriminant analysis[52] and neural network with feature extraction[53]. Two additional algorithms were included, svmLinear2 and LogitBoost. The complete analysis was performed in less than 1 min (0:38 min). The sda algorithm built the model with the highest performance (train AUROC 0.86 and test AUROC 0.81) having a higher performance measure than the published LASSO linear regression model[47] (train AUROC 0.84 and test AUROC 0.85) (**Fig. 1e, Supplementary table 10 and Supplementary data 4**). For the LIHC dataset we used same five ML algorithms as for Zeller dataset and analysis finished in 11 min and 30 sec. For such highly imbalanced dataset the precision-recall AUC (prAUC)[54] is a much better performance measurement than AUROC that reported near-perfect performance (**Fig. 1e**). The prAUC provides information how well the model correctly detects cancer cells, while it is less stringent on the evaluation of healthy cells. To avoid obtaining overly optimistic prediction results (often observed on imbalanced datasets), we ranked models based on the prAUC of the training set (**Supplementary table 11)**. The model that had the best performance was built using the svmLinear2 algorithm (train prAUC 0.83) and it also performed well on the holdout, test set (prAUC 0.73) (**Supplementary data 5**).

The *drowsiness* contributed the most to the top-performing SISA model, confirming the findings from the original study[44] (**Supplementary table 12**). To standardize the process for evaluation of the features and their contribution to the models, we implemented the variable importance score evaluation functions from the caret library[33]. This allows users to compare features selected across models. In the case of SISA dataset, *drowsiness* contributed the most in all of the models built (**Supplementary table 13**), indicating the importance of this symptom and its correlation with hospitalization. The features that contributed the most to the Cyclists model were total memory, unswitched memory and naïve B cells, recent thymic emigrants, CD8+ T cells with TEMRA phenotype, and regulatory T cells (CD25+ Foxp3+ CD4+ T cells)

(**Supplementary table 14**). In comparison to age-matched physically inactive individuals (non-cyclists), the master cyclists had increased frequencies of recent thymic emigrants, naïve B cells and CD3 cells, and decreased frequencies of memory B cells and CD8 T cells with TEMRA phenotype, confirming that ageing of immune systems i.e. immunosenescence can be reduced by high levels of physical activity[45] (**Fig. 1f, Supplementary figure 1**). To further explore the relationship between selected features, users can perform correlation analysis to reveal highly correlated features (**Fig. 1g, Supplementary protocol**). Naïve and memory B cells were identified as being highly correlated (**Fig. 1g**), as expected since these subsets were determined from the same flow cytometry plots and their relationship is inversely correlated. Removal of those highly correlated features can help to build more accurate models. Removal of naïve B cells resulted in building predictive model with the same performance measurements as the model built on the entire dataset (train AUROC 0.99 and test AUROC 1) (**Supplementary table 15**), while removal of total memory B cells lowered the accuracy estimates (train AUROC 0.98 and test AUROC 1) (**Supplementary table 16**), indicating the importance of memory B cells to discriminate between master cyclists and non-cyclists. In the VAST dataset, individuals with higher IgA, IgA1, IgA2 and IgG2 titers against native Vi polysaccharide (nViPS) antigen and higher IgA and IgG3 titers against biotinylated Vi polysaccharide (ViBiot) on the day of the challenge were protected against the typhoid challenge supporting the data from univariate analysis[46] (**Supplementary table 17 and Supplementary figure 2**). Moreover, using the clustering function of SIMON's exploratory analysis module, we can quickly identify that the IgA2 signature dominates the responses after vaccination with a purified Vi polysaccharide (Vi-PS), while the IgG2 signature was dominant for the Vi tetanus toxoid conjugate (Vi-TT) vaccine[46] (**Fig. 1h, Supplementary protocol**). For the Zeller dataset, the same features as originally reported[40] contributed the most to the model, including *Fusobacterium nucleatum* and *Peptostreptococcus stomatis* (**Supplementary table 18**). The features that contributed the most to the LIHC model were well-known genes identified to be upregulated in LIHC such as *GABRD* and *PLVAP*[55] and genes enriched in adjacent normal tissue samples *ANGPTL6*[56], *VIPR1*[57] and *OIT3*[58] as a typical signature for healthy liver tissue (**Supplementary table 19**, **Supplementary figure 3**).

Overall, SIMON is a powerful software platform for data mining that facilitates pattern recognition and knowledge extraction from high-quality, heterogenous biological and clinical data, especially where there is missing data, an imbalanced distribution and/or high dimensionality. It can be used for identification of genetic, microbial and immunological correlates of protection and help guiding further analysis of the biomedical data.

Figure

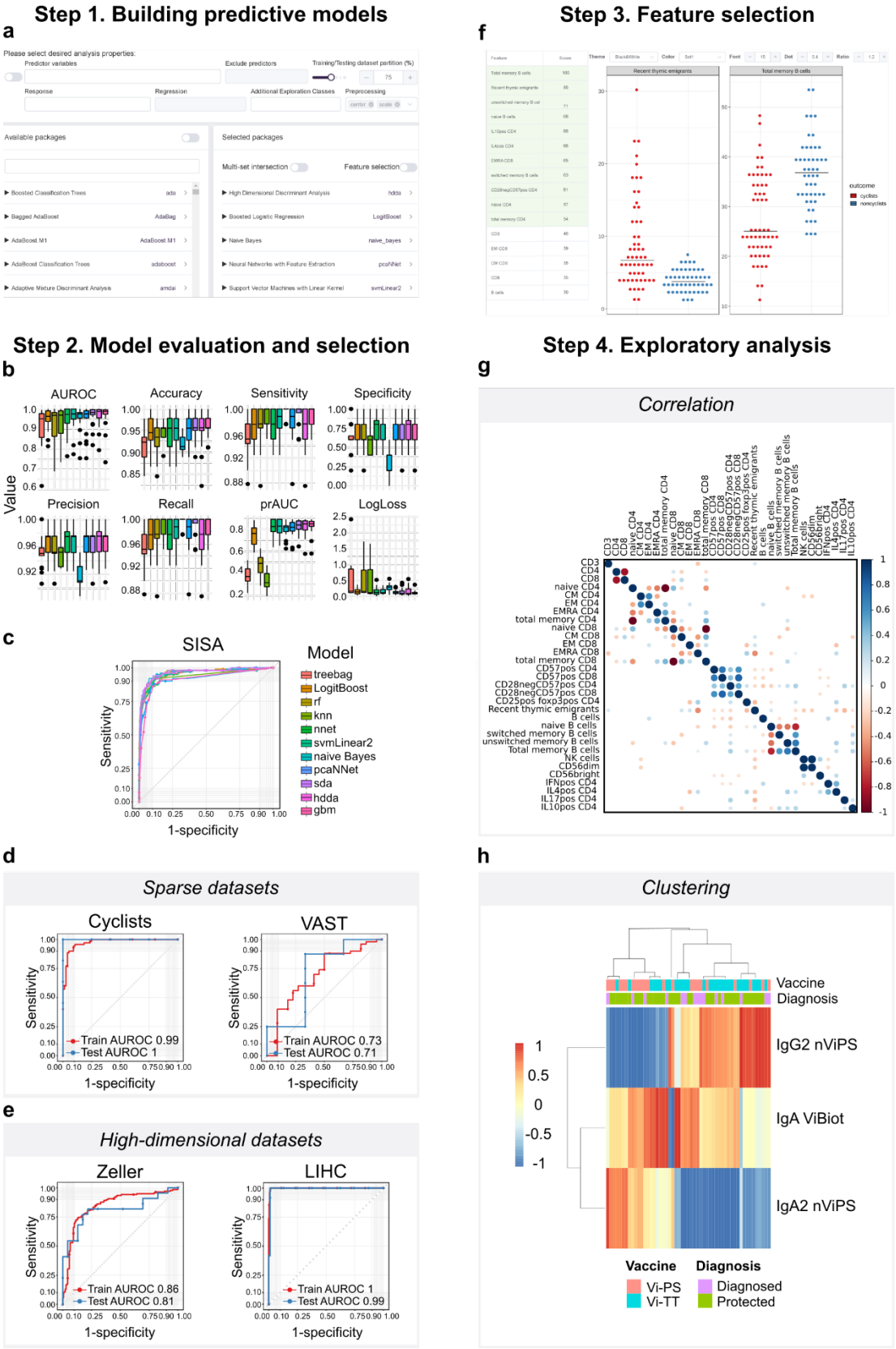

Figure 1. SIMON machine learning workflow. Step 1. Building predictive models. (a) Screenshot of the SIMON graphical user interface demonstrating input selection for machine learning analysis, such as predictors and response (outcome) variables, additional exploration classes, training/test split, preprocessing functions and desired machine learning algorithms. **Step 2. Model evaluation and selection.** Comparison of **(b)** box plots of performance measurements calculated for 11 predictive models and **(c)** ROC curves built on the SISA dataset. Comparison of ROC curves calculated from the training and test sets on **(d)** datasets with missing values (Cyclists and VAST) and **(e)** high-dimensional datasets (Zeller and LIHC). **Step 3. Feature selection. (f)** Screenshot of the SIMON interface showing variable importance scores calculated for each feature and graphical visualization of the selected features from the Cyclists dataset. **Step 4. Exploratory analysis. (g)** Correlation analysis on the Cyclists dataset. **(h)** Clustering analysis on the VAST dataset.

References

1 Stuart, T. & Satija, R. Integrative single-cell analysis. *Nature reviews. Genetics* **20**, 257-272 (2019).

2 Nolan, J. P. & Condello, D. Spectral flow cytometry. *Current protocols in cytometry* **Chapter 1**, Unit1 27 (2013).

3 Gregori, G. *et al.* Hyperspectral cytometry at the single-cell level using a 32-channel photodetector. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **81**, 35-44 (2012).

4 Futamura, K. *et al.* Novel full-spectral flow cytometry with multiple spectrally-adjacent fluorescent proteins and fluorochromes and visualization of in vivo cellular movement. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **87**, 830-842 (2015).

5 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. *Analytical chemistry* **81**, 6813-6822 (2009).

6 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).

7 Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nature medicine* **20**, 436-442 (2014).

8 Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods* **11**, 417-422 (2014).

9 Bishop, C. M. *Pattern Recognition and Machine Learning*. (Springer-Verlag New York, 2006).

10 Yip, K. Y., Cheng, C. & Gerstein, M. Machine learning and genome annotation: a match meant to be? *Genome biology* **14**, 205 (2013).

11 Chicco, D. Ten quick tips for machine learning in computational biology. *BioData mining* **10**, 35 (2017).

12 Deo, R. C. Machine Learning in Medicine. *Circulation* **132**, 1920-1930 (2015).

13 Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nature reviews. Genetics* **16**, 321-332 (2015).

14 Bonetta, R. & Valentino, G. Machine learning techniques for protein function prediction. *Proteins* **88**, 397-413 (2020).

15 Jurtz, V. *et al.* NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of immunology* **199**, 3360-3368 (2017).

16 Lin, H. H., Ray, S., Tongchusak, S., Reinherz, E. L. & Brusic, V. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC immunology* **9**, 8 (2008).

17 Kan, A. Machine learning applications in cell image analysis. *Immunology and cell biology* **95**, 525-530 (2017).

18 Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature reviews. Drug discovery* **18**, 463-477 (2019).

19 Newell, E. W., Sigal, N., Bendall, S. C., Nolan, G. P. & Davis, M. M. Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes. *Immunity* **36**, 142-152 (2012).

20 Horowitz, A. *et al.* Genetic and environmental determinants of human NK cell diversity revealed by mass cytometry. *Science translational medicine* **5**, 208ra145 (2013).

21 Chaudhury, S. *et al.* Identification of Immune Signatures of Novel Adjuvant Formulations Using Machine Learning. *Scientific reports* **8**, 17508 (2018).

22 Chaudhury, S. *et al.* Combining immunoprofiling with machine learning to assess the effects of adjuvant formulation on human vaccine-induced immunity. *Human vaccines & immunotherapeutics* **16**, 400-411 (2020).

23 Warsinske, H. C. *et al.* Assessment of Validity of a Blood-Based 3-Gene Signature Score for Progression and Diagnosis of Tuberculosis, Disease Severity, and Treatment Response. *JAMA network open* **1**, e183779 (2018).

24 Robinson, M. *et al.* A 20-Gene Set Predictive of Progression to Severe Dengue. *Cell reports* **26**, 1104-1111 e1104 (2019).

25 Sweeney, T. E. *et al.* A community approach to mortality prediction in sepsis via gene expression analysis. *Nature communications* **9**, 694 (2018).

26 Mayhew, M. B. *et al.* A generalizable 29-mRNA neural-network classifier for acute bacterial and viral infections. *Nature communications* **11**, 1177 (2020).

27 Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* **13**, 8-17 (2015).

28 Beckmann, J. S. & Lew, D. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. *Genome medicine* **8**, 134 (2016).

29 R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2013).

30 Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**, 1-26 (2008).

31 Guttag, J. V. *Introduction to Computation and Programming Using Python: With Application to Understanding Data*. Second edn, 472 / 466 (The MIT Press, 2016).

32 Lang, M. *et al.* mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software* **4**, 1903 (2019).

33 caret: Classification and Regression Training v. 6.0-80 (R package, 2018).

34 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).

35 Albanese, D. *et al.* mlpy: Machine Learning Python. *arXiv* **1202.6548** (2012). <https://arxiv.org/abs/1202.6548>.

36 Bell, M. L., Fiero, M., Horton, N. J. & Hsu, C. H. Handling missing data in RCTs; a review of the top medical journals. *BMC medical research methodology* **14**, 118 (2014).

37 Pes, B. Handling Class Imbalance in High-Dimensional Biomedical Datasets. *8th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE).* 150-155 (2019 IEEE).

38 Bellman, R. E. *Dynamic programming*. (Princeton University Press, 1957).

39 Automated Machine Learning: Methods, Systems, Challenges in *The Springer Series on Challenges in Machine Learning* (eds F. Hutter, L. Kotthoff, & J. Vanschoren) (Springer, 2018). <http://automl.org/book>.

40 Thornton, C., Hutter, F., Hoos, H. H. & Leyton-Brown, K. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. *KDD: Knowledge Discovery and Data Mining 2013.* 847–855.

41 Tomic, A. *et al.* SIMON, an Automated Machine Learning System, Reveals Immune Signatures of Influenza Vaccine Responses. *J Immunol* **203**, 749-759 (2019).

42 Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116-5121 (2001).

43 Murty, J. *Programming Amazon Web Services: S3, EC2, SQS, FPS, and SimpleDB*. 604 (O'Reilly Media, 2009).

44 Sippy, R. *et al.* Severity Index for Suspected Arbovirus (SISA): Machine learning for accurate prediction of hospitalization in subjects suspected of arboviral infection. *PLoS neglected tropical diseases* **14**, e0007969 (2020).

45 Duggal, N. A., Pollock, R. D., Lazarus, N. R., Harridge, S. & Lord, J. M. Major features of immunesenescence, including reduced thymic output, are ameliorated by high levels of physical activity in adulthood. *Aging cell* **17** (2018).

46 Dahora, L. C. *et al.* IgA and IgG1 Specific to Vi Polysaccharide of Salmonella Typhi Correlate With Protection Status in a Typhoid Fever Controlled Human Infection Model. *Frontiers in immunology* **10**, 2582 (2019).

47 Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology* **10**, 766 (2014).

48 Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nature methods* **14**, 1023-1024 (2017).

49 Geistlinger, L. *et al.* Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in bioinformatics* (2020).

50 Jin, C. *et al.* Efficacy and immunogenicity of a Vi-tetanus toxoid conjugate vaccine in the prevention of typhoid fever using a controlled human infection model of Salmonella Typhi: a randomised controlled, phase 2b trial. *The Lancet* **390**, 2472–2480 (2017).

51 Mkhadri, A. Shrinkage parameter for the modified linear discriminant analysis. *Pattern Recognition Letters* **16**, 267-275 (1995).

52 Bouveyron, C., Girard, S. & Schmid, C. High-Dimensional Discriminant Analysis. *Communications in Statistics - Theory and Methods* **36**, 2607–2623 (2007).

53 Ripley, B. D. *Pattern Recognition and Neural Networks*. (Cambridge University Press, 1996).

54 Davis, J. & Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23rd International Conference on Machine Learning.* (2006).

55 Sarathi, A. & Palaniappan, A. Novel significant stage-specific differentially expressed genes in hepatocellular carcinoma. *BMC cancer* **19**, 663 (2019).

56 Oike, Y. *et al.* Angiopoietin-related growth factor antagonizes obesity and insulin resistance. *Nat Med* **11**, 400-408 (2005).

57 Lu, S., Lu, H., Jin, R. & Mo, Z. Promoter methylation and H3K27 deacetylation regulate the transcription of VIPR1 in hepatocellular carcinoma. *Biochemical and biophysical research communications* **509**, 301-305 (2019).

58 Xu, Z. G. *et al.* A novel liver-specific zona pellucida domain containing protein that is expressed rarely in hepatocellular carcinoma. *Hepatology* **38**, 735-744 (2003).