Swarm immunology: harnessing blockchain technology and artificial intelligence in human immunology

Joachim L. Schultze[1,2,3]✉, *Maren Büttner*[3,4] *and Matthias Becker*[1,2]

Human immunology may soon benefit from the use of artificial intelligence and blockchain technologies. Here, we discuss how Swarm Learning could foster collaborative worldwide immunology studies that fully respect local data privacy regulations by sharing insights, not data.

For decades, immunological research has benefited from highly standardized animal models. Yet, with increasing knowledge the translation from model systems to human diseases seems to be more and more problematic and often fails[1]. At the same time, technological advances in genomics down to the single-cell level, the introduction of artificial intelligence (AI) into biomedical research, and novel approaches to model human disease — including organoids or lab-on-a-chip approaches — are poised to revolutionize medicine, including human immunology[2]. Methods such as single-cell RNA-sequencing (RNA-seq) and mass cytometry provide important new insights, yet at the same time require novel analytical approaches, particularly when it comes to scaling to large clinical multi-centre studies. Here, machine learning (that is, the branch of AI that improves models automatically using data) is the prerequisite for automated scaling and uncovering the molecular patterns in single-cell data. Leveraging the full potential of machine learning algorithms — for example, for disease classification or stratification from high-throughput data — requires inclusion of hundreds of patients to accommodate the potential biases owing to factors such as local experimental batch, age, sex, genetic background or ethnicity[3]. Collecting the data is in itself a laborious task, and few centres in the world are able to conduct these kinds of studies on their own. Although millions of samples of blood and biological tissues are taken each year, sharing the data from these samples is greatly restricted owing to personal data protection laws. The legislation has rightfully put high bars here to protect the health data of the individual; however, these laws simultaneously discourage scientific progress.

To overcome such limitations, we recently developed Swarm Learning (SL) as a fully decentralized machine learning principle to facilitate the integration of data from several sites under full consideration of data privacy regulations[4]. Conceptually, SL is a decentralized approach to train a joint machine learning model through parameter sharing while keeping private patient data safe locally (FIG. 1a). Every participating site is a node in the Swarm network and participates in the model training with local data. Data security, confidentiality and sovereignty are ensured through private permissioned blockchain technology (see Related links for an explanation of blockchains). New nodes can enter the Swarm network via a blockchain smart contract, regulating the conditions for Swarm network membership in a fully automated electronic fashion. New Swarm members agree to the collaboration terms, obtain the model and perform local training until joint training goals have been reached. This approach offers new opportunities to overcome the limitations for collaborative sciences as several research sites may easily join forces to tackle the same research question but with much larger data available for analysis without sharing primary data between sites.

Learning a joint model on data at various sites requires an agreement on the dataset and its pre-processing as well as models jointly agreed upon. To achieve high quality input the datasets require a minimum level of standardization in sample handling, selection of measured features and data pre-processing. In genomics research, the human reference genome with accurate gene annotations is the common reference, which then allows for the alignment of RNA-seq data against the reference. For humans, all data span the same feature space with over 30,000 genes. By contrast, the number of measured features seen with antibodies in flow cytometry and mass cytometry, as well as in CITE-seq and Ab-seq, is in the order of magnitude of 10 to 100 (FIG. 1b), while the number of possible surface molecules is over 1,000 (REF.[5]). Notably, not all surface molecules have an available antibody counterpart. The experimental limitations for cell surface protein marker technologies thus demand thorough marker selection. The panel design is usually specific for the research question and the cell type of interest — that is, a T cell panel incorporates different markers than a B cell panel, with little to no overlap. When the data provided by different sites is very different in the selected markers, even when the

[1] Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany.

[2] PRECISE Platform for Single Cell Genomics and Epigenomics, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE) and the University of Bonn, Bonn, Germany.

[3] Genomics and Immunoregulation, Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany.

[4] Institute of Computational Biology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany.

✉e-mail: Joachim.schultze@dzne.de

https://doi.org/10.1038/s41577-022-00740-1

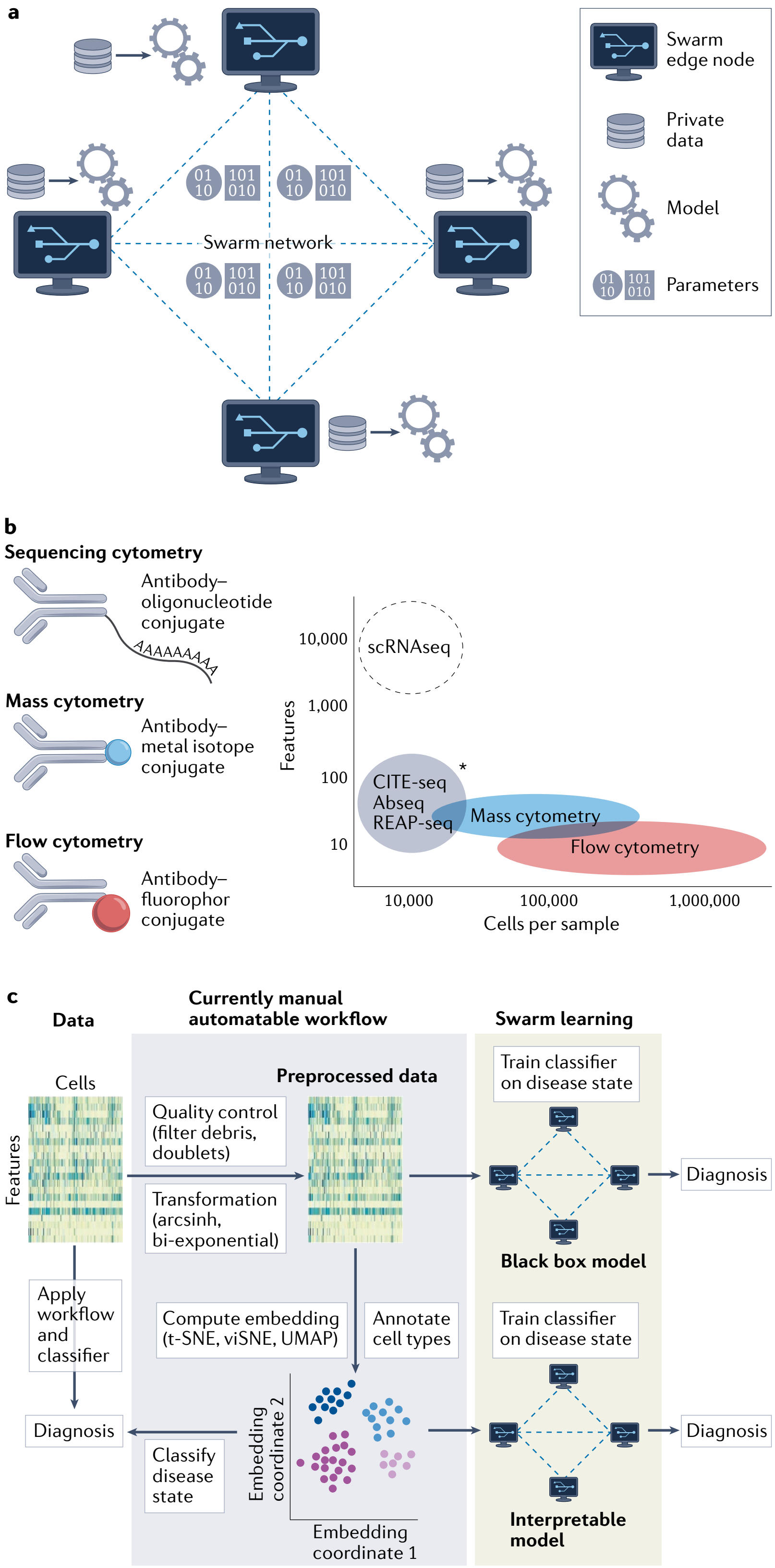

Fig. 1 | **Swarm Learning. a** | Swarm Learning principle. Data remain at the participating site while all sites jointly perform model parameter estimation in a Swarm network. **b** | Single-cell methods based on antibody tagging differ in the number of features measured and throughput. Methods marked with as asterisk (*) represent only the antibody features, but are usually coupled with single-cell RNA-sequencing, whose number of features is displayed in the grey dashed circle. **c** | Workflow for single-cell data analysis in immunology. Dashed arrows denote classification models, which can be set up as Swarm Learning models. Grey box highlights potentially automatable processing steps.

same disease is measured, joint modelling using these data becomes challenging. Here, the key for the broader application of SL is the standardization of panels and antibody concentrations. For instance, clinical diagnostics in leukaemia have been successfully standardized by the EuroFlow consortium[6] and subsequently commercialized. Thus, owing to the higher level of standardization, the diagnostic community could already benefit from SL, further optimizing test development by accessing and analysing large datasets with innovative AI applications. Furthermore, the use of ensemble models for classification on several panels from the same samples would allow for more flexibility in the marker choice[7]. Any future application of machine learning to flow cytometry will benefit from standardization in data pre-processing (FIG. 1c). For instance, flow cytometry data pre-processing involves a fine-tuned compensation owing to the spectral overlap in the fluorescent dyes followed by normalization, which is handled mostly manually. Especially when we want to combine data from different modalities from flow cytometry and mass cytometry, as well as from CITE-seq and Ab-seq studies, the input data need to adhere to a transferable standard. What is true for cell surface marker analysis would similarly apply to other typical data types in human immunology, for example, plasma-based protein markers or ex vivo immune activation panels.

SL supports different kinds of models and a broad range of applications. Deep learning models, especially variational autoencoders, have shown superior performance when handling high-throughput, high-dimensional single-cell data, for instance, in data integration tasks[8]. Moreover, they can be used for building reference atlases at one site, sharing the model of the data and integrating new data at a different site[9]. While this approach relies on a single entity that creates the reference, it indicates the potential of distributed deep learning models for SL in a fully decentralized setup. The advantage of these models is an intuitive interpretability of the learned latent space, which allows us to classify cells, not just entire samples. We are convinced that this level of granularity will be critical for the development of immune-based biomarkers and can only be reached by integrating large enough datasets from many different institutions and hospitals, but without sharing primary data in an SL setting.

Collectively, SL opens a new perspective for science in the clinical context. In a sufficiently large Swarm network, one would be able to use all types of observed perturbations in humans, such as response to vaccination or infectious diseases, to infer causal principles of the human immune system from the vast amount of data. A concerted systems immunology initiative may easily collect human samples in a global setup, and create large human cohorts providing enough data to study molecular mechanisms of human disease. Such enlarged cohorts are key for successful clinical applications, from disease classification using machine learning to unbiased biomarker discovery. For instance, the COVID-19 pandemic has accelerated such collaborative endeavours in the German COVID-19 Omics Initiative (DeCOI), and may serve as a blueprint for future pandemics[4,10].

As a next step, we will have to show that heterogeneous immune data are indeed applicable to SL principles at scale. Furthermore, such SL-enabled international activities will greatly benefit from improvements of data standardizations within human immunology. The development of platforms that allow easy access to SL projects will facilitate the field. Lastly, if successful, immune biomarker and AI-based disease classification and stratification needs approval by the authorities prior to becoming standard of care, which in itself will require further efforts and developments. Nevertheless, the start of a truly integrative era of human immunology research is now in sight.

1. Pulendran, B. & Davis, M. M. The science and medicine of human immunology. *Science* **369**, eaay4014 (2020).
2. Rajewsky, N. et al. LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* **587**, 377–386 (2020).
3. Hu, Z., Tang, A., Singh, J., Bhattacharya, S. & Butte, A. J. A robust and interpretable end-to-end deep learning model for cytometry data. *Proc. Natl Acad. Sci. USA* **117**, 21373–21380 (2020).
4. Warnat-Herresthal, S. et al. Swarm Learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).
5. Bausch-Fluck, D. et al. The in silico human surfaceome. *Proc. Natl Acad. Sci. USA* **115**, E10988–E10997 (2018).
6. van Dongen, J. J. M. et al. EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. *Leukemia* **26**, 1908–1975 (2012).
7. Aghaeepour, N. et al. Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10**, 228–238 (2013).
8. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
9. Lotfollahi, M. et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
10. Schulte-Schrepping, J. et al. Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell* **182**, 1419–144 (2020).

Acknowledgements
This work was partially funded by the BMBF (NaFoUniMedCovid19-COVIM: 01KX2021) and under Germany's Excellence Strategy (DFG – EXC2151 – 390873048).

Competing interests
The authors declare no competing interests.

RELATED LINKS
What is blockchain technology? https://blog.chain.link/what-is-blockchain/
DeCOI, German COVID-19 OMICS Initiative: https://decoi.eu/