

# Defining and benchmarking open problems in single-cell analysis



**S**ingle-cell genomics has enabled the study of biological processes at an unprecedented scale and resolution. These studies were enabled by innovative data generation technologies coupled with emerging computational tools specialized for single-cell data. As single-cell technologies have become more prevalent, so has the development of new analysis tools, which has resulted in over 1,700 published algorithms<sup>1</sup> (as of February 2024). Thus, there is an increasing need to continually evaluate which algorithm performs best in which context to inform best practices<sup>2,3</sup> that evolve with the field.

In many fields of quantitative science, public competitions and benchmarks address this need by evaluating state-of-the-art methods against known criteria, following the concept of a common task framework<sup>4</sup>. Here, we present Open Problems, a living, extensive, community-guided platform including 12 current single-cell tasks that we envisage raising standards for the selection, evaluation and development of methods in single-cell analysis.

In single-cell genomics, as in many other domains, it is typical for analysis algorithms to be evaluated using benchmarks. However, such benchmarks are often of limited use as the field suffers from a lack of standardized procedures for benchmarking<sup>5</sup>, leading to different assessments of the same method and producing different outcomes. Bespoke benchmarks set up by method developers to evaluate newly developed algorithms often include datasets and metrics chosen to highlight the advantages of their tools, which has been shown to lead to less objective assessments<sup>6,7</sup>. Even if datasets and metrics are standardized, historical analysis shows that when benchmarks are implemented by the same groups introducing new methods, the evaluations tend to inflate performance of the newest models via custom hyperparameter selection and data processing<sup>8</sup>.

To provide more uniform and neutral assessment, groups can perform specialized benchmarking studies independently of method development. Tools such as

registered reports, which promote neutrality of benchmarking results by design, have recently gained in popularity to enable such studies. These efforts aim to systematically evaluate the current state of the art in a given area and may be less biased. However, their results are static and inevitably age. These frameworks are typically not designed for extensibility or interoperability, limiting the value of reusing a framework to perform additional systematic benchmarks<sup>5</sup>. This inability to reuse infrastructure leads to repeats of non-standardized benchmarks that cannot provide the guidance that users need. For example, at least four benchmarks of batch integration methods exist<sup>9–12</sup>, each of which uses different sets of datasets and metrics and suggests different optimal methods (Fig. 1a). Similar issues have been reported across other single-cell topics, where datasets and metrics typically have less than 10% overlap between benchmarks<sup>13</sup>.

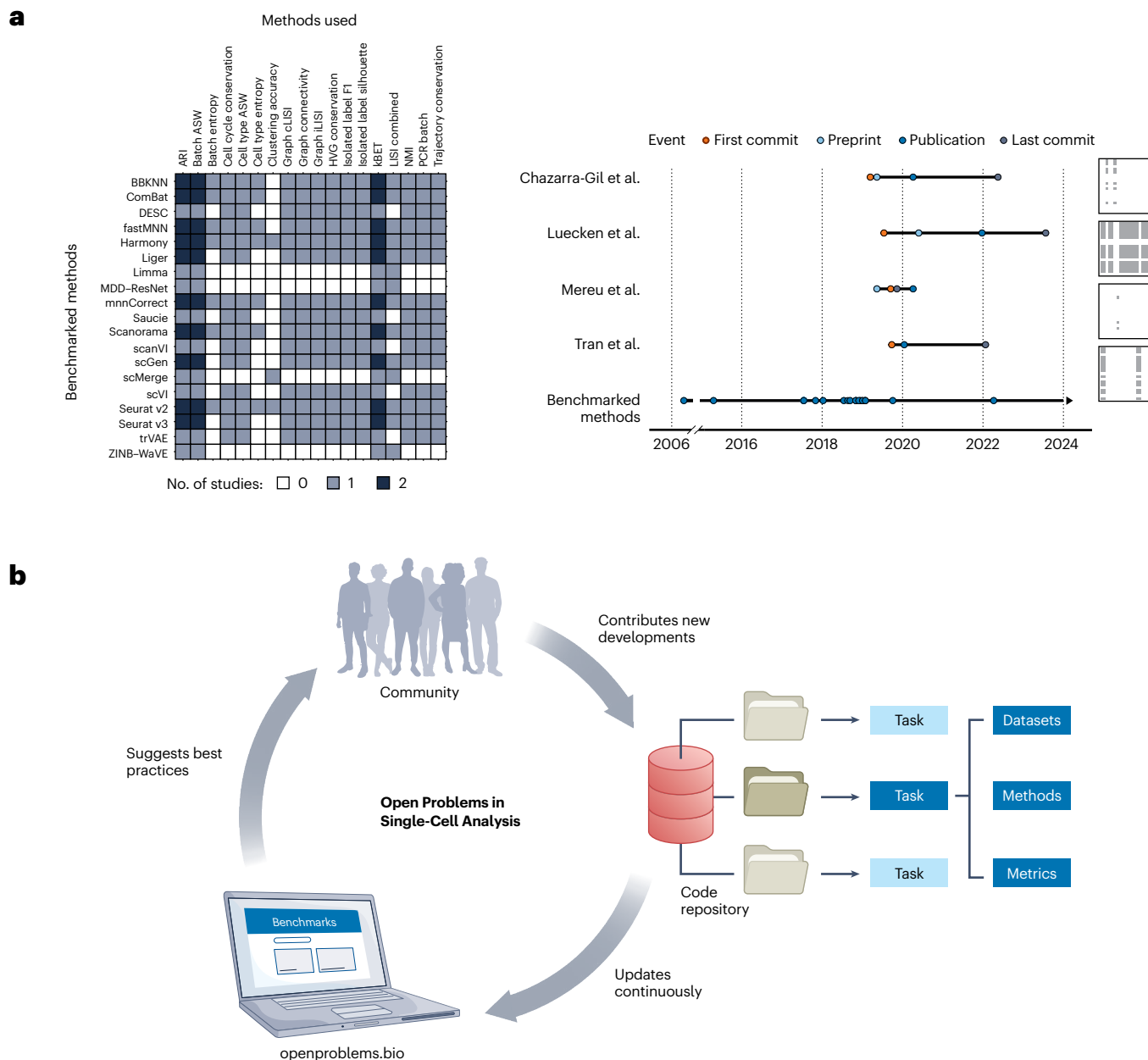
Ideally, benchmarks that guide users and promote method innovation use consistently applied datasets and metrics that are established independently of method development and with ongoing community participation<sup>5,6,13</sup>. Such community participation around quantified tasks requires continual updates, a process that is hard to realize in the typical result–paper framework that defines the modern scientific process.

To achieve this goal, we developed the Open Problems in Single-Cell Analysis (Open Problems) platform. The Open Problems platform is an open-source, extensible, living benchmarking framework that enables quantitative evaluation of best practices in single-cell analysis. It combines a permissively licensed GitHub repository (<https://github.com/openproblems-bio/openproblems>) with community-defined tasks, an automated benchmarking workflow, and a website to explore the results. Currently, Open Problems includes 12 defined tasks, in which 81 datasets are used to evaluate 171 methods using 37 metrics. These tasks were defined by community engagement, including on the public GitHub repository, in weekly community meetings, and at a hackathon in March 2021 with over

50 participants. This broad involvement has already led to new benchmarking insights and best practice recommendations while improving and standardizing previously published benchmarks. We envisage Open Problems' community-defined standards for progress in single-cell data science raising the bar for the selection and evaluation of methods, providing targets for method innovation, and enabling developers without single-cell expertise to contribute to the field.

To enable truly living benchmarks, we designed a standardized and automated infrastructure that allows members of the single-cell community to contribute to Open Problems in a seamless manner (Supplementary Methods). Each Open Problems task consists of datasets, methods and metrics (Fig. 1b). Datasets define both the input and the ground truth for a task, methods attempt to solve the task, and metrics evaluate the success of a method on a given dataset. We provide cloud infrastructure to enable centralized benchmarking when new methods, datasets or metrics are added to our platform. Within each task, every method is evaluated on every dataset using every metric, and each method is then ranked on a per-dataset basis by the average normalized metric score and presented in a summary table on the Open Problems website (<https://openproblems.bio>). Normalization is used to make metric ranges comparable for comparison and visualization of method results without affecting the metric's ability to highlight method outliers (Supplementary Methods).

Community engagement on the platform is centered around an open discussion forum, open code contribution opportunities, and task leadership. Task leaders are community members who have contributed substantially to a task, assume organizational responsibilities for the task, and are ultimately responsible for task definition, task maintenance and facilitation of community contributions. Task definitions, choices of metrics and implementations of methods are discussed on our GitHub repository and can be easily amended by pull requests, which are reviewed by task leaders and the core infrastructure team.



**Fig. 1 | The Open Problems in Single-cell Analysis living benchmarking platform. a**, Overview and timeline of published benchmarks of single-cell batch integration. Four publications have benchmarked 19 methods using 18 metrics. Light gray and black squares indicate whether one or two benchmarks include this method–metric combination (left). Arrows indicate the range of publication times of methods included in the benchmark. **b**, The Open Problems

platform. The Open Problems platform consists of tasks that are broken down into datasets, methods and metrics. The community contributes code to these tasks in the platform, which uses these contributions to extend the benchmarks that are run and pushed to the Open Problems website. The community can then consult the website for guidance on method selection.

While this community-centered approach may lead to suboptimal metrics being contributed, it also facilitates a self-cleansing process whereby metrics can be removed or amended if limitations or biases are uncovered. In this manner, Open Problems promotes the longevity of hosted benchmarks.

To enable seamless community involvement in Open Problems, we have designed our platform to leverage cloud infrastructure, which provides reproducibility, accessibility and automation (Supplementary Fig. 1). Each task is organized as a directory, with subdirectories for datasets, methods, metrics and

utilities. Each task must contain at least one dataset, one metric and two baseline methods, which provide upper and lower bounds for performance of the task. Components (that is, dataset loader, method or metric) are Viash components<sup>14</sup> that exist as a single script (implemented in Bash, Python or R) and

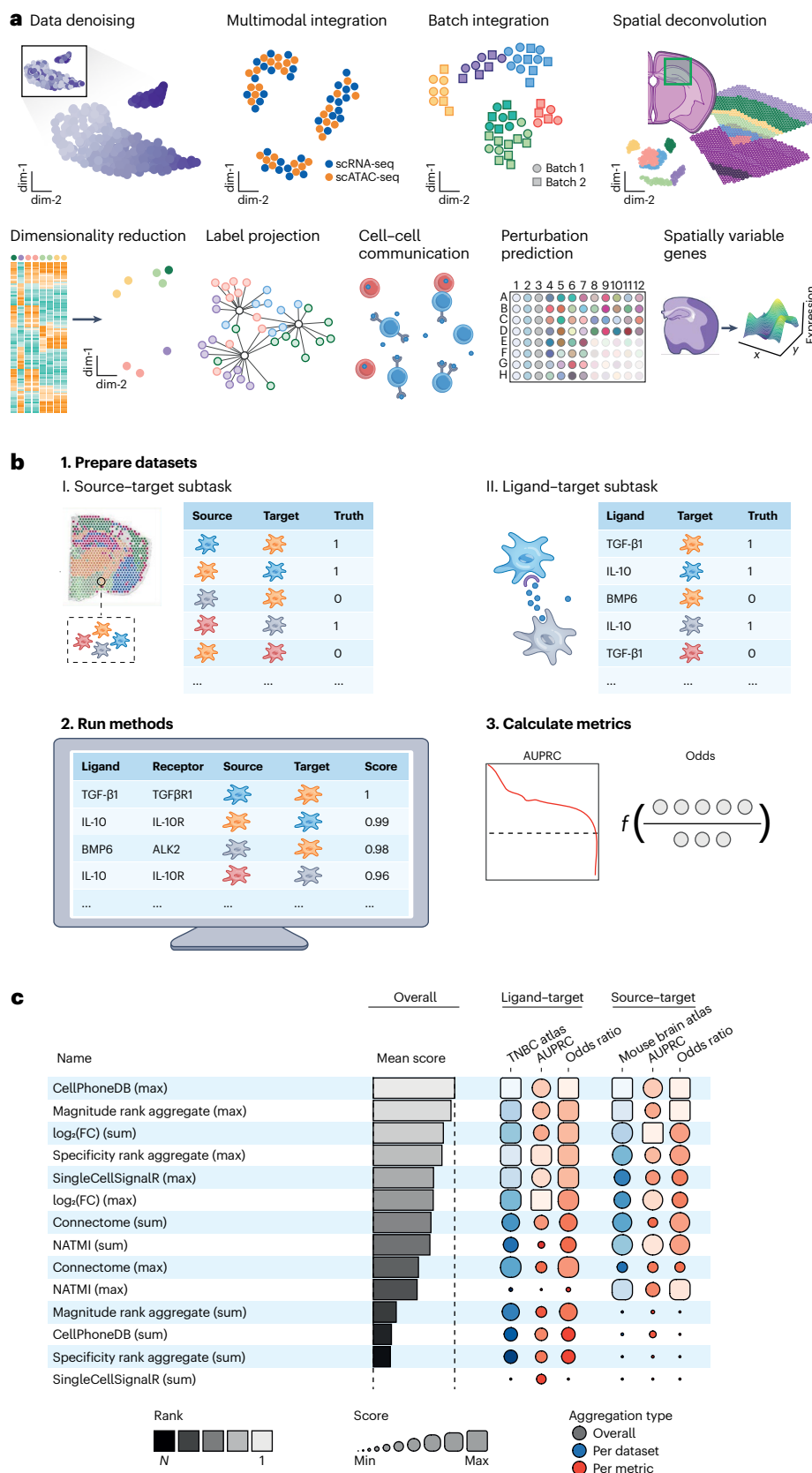
some metadata (a YAML file named config.vsh.yaml) in the relevant subdirectory. Adding a new method is as simple as opening a pull request to the repository and adding a new file that follows the API for that task. When a community member adds a component, the new contribution is automatically tested in the cloud. When all tests pass and the new contribution is accepted, the results from the new contribution are automatically submitted to the Open Problems website. To maximize reproducibility, each component is run in a versioned Docker container defined by the method contributor, and all data are downloaded from public repositories, including Figshare and CELLxGENE<sup>15</sup>.

We started by defining 9 Open Problems tasks (Fig. 2a), which extends to 12 with the inclusion of subtasks. While several tasks were directly informed by published benchmarking papers (for example, batch correction<sup>9</sup> or cell-cell communication<sup>16</sup>), others were defined by method developers in the single-cell community (for example, spatial decomposition). The Open Problems platform was seeded with these tasks as they represent a cross-section of important and (mostly) well-researched tasks in single-cell genomics. We envisage this set of tasks being a starting point for further community development to address and refine further open problems in single-cell analysis.

A typical task setup can be exemplified by the cell-cell communication (CCC) task

**Fig. 2 | Task overview, setup and results.**

**a**, Overview of the 9 tasks currently included in the Open Problems platform. Batch integration and cell-cell communication (CCC) consist of 3 and 2 subtasks, respectively, making up the current total of 12 tasks. **b**, The CCC task. This task includes two subtasks defined by different types of ground truth: spatial cell type co-localization in the source-target subtask and cytokine profiling in the ligand-target subtask. Methods are run on each subtask to score the likelihood of interaction between source and target cell types or ligand and target cell types. Finally, the area under the precision-recall curve (AUPRC) and the odds ratio of true to false positive interactions in the top 5% of predicted pairs are used to score method outputs (Supplementary Note 1.1). **c**, Collated results of the two CCC subtasks. Methods are ranked using the mean of the overall score for each subtask (shown as “TNBC atlas” and “Mouse brain atlas” blue boxes, respectively). These overall scores are computed as the mean of all scaled metric results (red boxes). Linear scaling is performed using random and perfect baseline methods, whose performances are set to 0 and 1, respectively (see Supplementary Methods). The results shown here are from Open Problems v1 (ref. 24).



(Fig. 2b and Supplementary Note 1.1). The goal of cell–cell communication inference methods is to infer which cell types are communicating within a tissue to mediate tissue function. Typical algorithms base predictions on the expression of ligand and receptor genes in dissociated single-cell data<sup>17</sup>. Ground-truth data for cellular communication are challenging to obtain. Thus, this task is divided into two subtasks that use different proxies for this ground truth: spatial colocalization (source–target subtask) and cytokine activity (ligand–target subtask). As the CCC methods included in this task (Supplementary Methods)<sup>18–21</sup> typically score ligand–receptor pairs using either their expression magnitude or cell-type specificity, *mean* and *max* aggregation functions are used to score interaction strengths between source and target cell types (source–target task) or ligands and target cell types (ligand–target task). The outputs of these methods are finally evaluated using the area under the precision–recall curve and odds ratios. These metrics measure how well ground truth source–target (co-localized cell types) or ligand–target (cytokine activity within a cell type) pairs are prioritized when ranking all interactions and how many true pairs are found in the top 5%, respectively.

While the CCC task was contributed to Open Problems on the basis of a published benchmark<sup>16</sup>, the task definition and metrics evolved with input from the community and the Open Problems team. This process has enabled the Open Problems results to generate insight beyond the initial publication (Fig. 2c), which focused predominantly on the comparison of CCC databases and showed variable method performance across tasks. In the CCC Open Problems task, we find that methods that rely on expression magnitude outperform approaches that rely on expression specificity. Indeed, the top performers across tasks are CellPhoneDB and LIANA’s ensemble model of expression magnitude scoring methods. Furthermore, *max* aggregation of ligand–receptor scores outperformed *mean* aggregation across tasks and methods. This improved inference of cellular communication using only the top-predicted interactions suggests that methods are better at prioritizing a small fraction of relevant interactions while being prone to noise when their full interaction rankings are considered. Thus, analysts interpreting CCC results may likewise want to focus only on the most high-scoring predictions when inferring which cell types interact (Supplementary Note 1.1).

Using this combination of expert knowledge and community input, in this manuscript we also provide best-practice recommendations for preprocessing and method selection for label projection, dimensionality reduction for 2D visualization, batch integration, spatial decomposition, denoising and matching of cellular profiles across modalities (Supplementary Note 1). For example, on all four reference datasets currently included in the Open Problems label projection task, a simple logistic regression model outperforms more complex methods that explicitly model batch effects, even when noise is added to the training data (Supplementary Note 1.2). Moreover, we also show that it is easier to correct for batch effects in single-cell graphs than in latent embeddings or expression matrices (Supplementary Note 1.4), that denoising methods perform best with non-standard preprocessing approaches that better stabilize variance (Supplementary Note 1.6) and that simple models tend to outperform more complex ones for perturbation prediction (Supplementary Note 1.8).

Overall, Open Problems tasks are continually updated benchmarks that increase in robustness as new methods are developed and more complex datasets become available. Our vision is that these benchmarks will form the basis for best-practice recommendations by groups such as Single-Cell Best Practices (<https://www.sc-best-practices.org/>).

Open Problems living benchmarking tasks also function as a quantifiable target for the development of new methods. This problem definition is particularly useful for the wider machine learning community that may lack domain knowledge (that is, single-cell expertise). Leveraging the batch integration and matching modality tasks as a basis, we previously set up popular competitions for multimodal data integration at NeurIPS 2021 (refs. 22,23) and 2022, with over 260 and 1,600 participants, respectively. In these competitions, the developers of multiple top performers had no previous experience with single-cell data, yet were able to submit solutions that substantially outperformed state-of-the-art methods<sup>22</sup>. We envisage the Open Problems platform driving method development by improving the accessibility of open challenges in single-cell analysis via defined tasks. To promote this, Open Problems enables method developers to submit both prototype and final solutions to the platform for automated evaluation against the current state of the art. Open Problems results, which are made available under a Creative Commons Attribution

(CC-BY) license, can then be included in the respective method papers. Similarly, entirely new benchmarks can be implemented as tasks, run via Open Problems, and published separately while remaining updatable.

Taken together, the Open Problems platform is a community resource that quantitatively defines open challenges in single-cell analysis, determines the current state-of-the-art solutions, promotes method development to improve on these solutions, and monitors progress toward these goals. Open Problems addresses issues observed in custom and decentralized benchmarking by providing standardized but flexible infrastructure and task definitions. Thereby, Open Problems enables broader accessibility for scientists to contribute to the advancement of the field of single-cell analysis. We envisage Open Problems bringing about a shift in perspective on method selection for data analysts and method evaluation for developers, supporting a transition toward higher standards for methods in single-cell data science.

## Code availability


All Open Problems code is publicly available at <https://www.github.com/openproblems-bio/openproblems>. This code includes data loaders for all datasets used, with associated meta-data on where this data came from. Code to reproduce the figures is publicly available at <https://github.com/openproblems-bio/nbt2025-manuscript>. Detailed information on all datasets is available at <https://openproblems.bio/datasets>. Documentation for the platform and contribution guides can be found at <https://openproblems.bio/documentation>.

Malte D. Luecken<sup>1,2,55</sup>, Scott Gigante<sup>3,55</sup>, Daniel B. Burkhardt<sup>4,55</sup>, Robrecht Cannoodt<sup>5,6,7,55</sup>, Daniel C. Strobl<sup>1,8,9</sup>, Nikolay S. Markov<sup>10</sup>, Luke Zappia<sup>1,5,11</sup>, Giovanni Palla<sup>1,9</sup>, Wesley Lewis<sup>12</sup>, Daniel Dimitrov<sup>13</sup>, Michael E. Vinyard<sup>14,15,16</sup>, D. S. Magruder<sup>17</sup>, Michaela F. Mueller<sup>1,2,9</sup>, Alma Andersson<sup>18,19,20</sup>, Emma Dann<sup>21</sup>, Qian Qin<sup>15</sup>, Dominik J. Otto<sup>22,23,24</sup>, Michal Klein<sup>25</sup>, Olga Borisovna Botvinnik<sup>26,27</sup>, Louise Deconinck<sup>10,6,7</sup>, Kai Waldrant<sup>10,5</sup>, Sai Nirmayi Yasa<sup>5</sup>, Artur Szalata<sup>10,11</sup>, Andrew Benz<sup>28</sup>, Zhijian Li<sup>15,16</sup>, Open Problems Jamboree Members\*, Jonathan M. Bloom<sup>29</sup>, Angela Oliveira Pisco<sup>26,30</sup>, Julio Saez-Rodriguez<sup>13</sup>, Drausin Wulsin<sup>3</sup>, Luca Pinello<sup>16</sup>, Yvan Saeys<sup>16,7,31</sup>,



**Fabian J. Theis** <sup>1,11,32,56</sup> &  
**Smita Krishnaswamy** <sup>12,17,33,56</sup> 

<sup>1</sup>Institute of Computational Biology, Helmholtz Munich, Neuherberg, Germany. <sup>2</sup>Institute of Lung Health & Immunity, Helmholtz Munich; Member of the German Center for Lung Research (DZL), Munich, Germany. <sup>3</sup>Immunai, New York, USA. <sup>4</sup>NVIDIA, Santa Clara, CA, USA. <sup>5</sup>Data Intuitive, Lebbeke, Belgium. <sup>6</sup>Data Mining and Modelling for Biomedicine group, VIB Center for Inflammation Research, Ghent, Belgium. <sup>7</sup>Department of Applied Mathematics, Computer Science, and Statistics, Ghent University, Ghent, Belgium. <sup>8</sup>Institute of Clinical Chemistry and Pathobiochemistry, School of Medicine, Technical University of Munich, Munich, Germany. <sup>9</sup>TUM School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany. <sup>10</sup>Division of Pulmonary and Critical Care Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA. <sup>11</sup>Department of Mathematics, School of Computing, Information and Technology, Technical University of Munich, Munich, Germany. <sup>12</sup>Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. <sup>13</sup>Faculty of Medicine and Heidelberg University Hospital, Institute for Computational Biomedicine, Heidelberg University, Heidelberg, Germany. <sup>14</sup>Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. <sup>15</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>16</sup>Molecular Pathology Unit, Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA. <sup>17</sup>Department of Computer Science, Yale University, New Haven, CT, USA. <sup>18</sup>Genentech Inc, South San Francisco, CA, USA. <sup>19</sup>Gene Technology, Royal Institute of Technology (KTH), Stockholm, Sweden. <sup>20</sup>Science for Life Laboratory (SciLifeLab), Solna, Sweden. <sup>21</sup>Wellcome Sanger Institute, Cambridge, UK. <sup>22</sup>Basic Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. <sup>23</sup>Computational Biology Program, Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. <sup>24</sup>Translational Data Science IRC, Fred Hutchinson Cancer Center, Seattle, WA, USA. <sup>25</sup>Apple, Paris, France. <sup>26</sup>Data Sciences Platform, Chan Zuckerberg Biohub, San Francisco, CA, USA. <sup>27</sup>Bridge Bio Pharma, Palo Alto, CA, USA. <sup>28</sup>Cellarity, Inc, Somerville, MA, USA. <sup>29</sup>Department of Mathematics, Massachusetts Institute of

Technology, Cambridge, MA, USA. <sup>30</sup>Insitro, South San Francisco, USA. <sup>31</sup>VIB Center for AI & Computational Biology (VIB.AI), Ghent, Belgium. <sup>32</sup>Cellular Genetics Programme, Wellcome Sanger Institute, Hinxton, UK. <sup>33</sup>Department of Genetics, Yale University, New Haven, CT, USA. <sup>55</sup>These authors contributed equally: Malte D. Luecken, Scott Gigante, Daniel B. Burkhardt, Robrecht Cannoodt. <sup>56</sup>These authors jointly supervised this work: Fabian J. Theis, Smita Krishnaswamy. \*A list of authors and their affiliations appears at the end of the paper.  e-mail: [fabian.theis@helmholtz-munich.de](mailto:fabian.theis@helmholtz-munich.de); [smita.krishnaswamy@yale.edu](mailto:smita.krishnaswamy@yale.edu)

Published online: 1 July 2025

## References

- Zappia, L., Phipson, B. & Oshlack, A. *PLOS Comput. Biol.* **14**, e1006245 (2018).
- Heumos, L. et al. *Nat. Rev. Genet.* **24**, 550–572 (2023).
- Luecken, M. D. & Theis, F. J. *Mol. Syst. Biol.* **15**, e8746 (2019).
- Donoho, D. J. *Comput. Graph. Stat.* **26**, 745–766 (2017).
- Sonrel, A. et al. *Genome Biol.* **24**, 119 (2023).
- Brooks, T. G., Lahens, N. F., Mrčela, A. & Grant, G. R. *Nat. Rev. Genet.* **25**, 326–339 (2024).
- Buchka, S., Hapfelmeier, A., Gardner, P. P., Wilson, R. & Boulesteix, A.-L. *Genome Biol.* **22**, 152 (2021).
- Musgrave, K., Belongie, S. & Lim, S.-N. In *Computer Vision – ECCV 2020* (eds Vedaldi, A. et al.) *Lecture Notes in Computer Science* Vol. 12370 (Springer, 2020); [https://doi.org/10.1007/978-3-030-58595-2\\_41](https://doi.org/10.1007/978-3-030-58595-2_41)
- Luecken, M. D. et al. *Nat. Methods* **19**, 41–50 (2022).
- Chazarra-Gil, R., van Dongen, S., Kiselev, V. Y. & Hemberg, M. *Nucleic Acids Res.* **49**, e42 (2021).
- Tran, H. T. N. et al. *Genome Biol.* **21**, 12 (2020).
- Mereu, E. et al. *Nat. Biotechnol.* **38**, 747–755 (2020).
- Cao, Y. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.12.19.572303> (2025).
- Cannoodt, R. et al. *J. Open Source Softw.* **9**, 6089 (2024).
- CZI Cell Science Program et al. *Nucleic Acids Res.* **53**, D886–D900 (2025).
- Dimitrov, D. et al. *Nat. Commun.* **13**, 3224 (2022).
- Armingol, E., Baghdassarian, H. M. & Lewis, N. E. *Nat. Rev. Genet.* **25**, 381–400 (2024).
- Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. *Nat. Protoc.* **15**, 1484–1506 (2020).
- Hou, R., Denisenko, E., Ong, H. T., Ramiłowski, J. A. & Forrest, A. R. R. *Nat. Commun.* **11**, 5011 (2020).
- Raredon, M. S. B. et al. *Sci. Rep.* **12**, 4187 (2022).
- Cabello-Aguilar, S. et al. *Nucleic Acids Res.* **48**, e55 (2020).
- Lance, C. et al. In *Proc. NeurIPS 2021 Competitions and Demonstrations Track* 162–176 (NeurIPS, 2022).
- Luecken, M. D. et al. In *Proc. Neural Information Processing Systems Track on Datasets and Benchmarks 1* (NeurIPS, 2021); <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/158f3069a435b314a80bdc024f8e422-Abstract-round2.html>
- Gigante, S. et al. Openproblems-Bio/Openproblems: V1.0.0. *Zenodo* <https://doi.org/10.5281/ZENODO.13769879> (2024).

## Acknowledgements

We received continual support in many ways from Jonah Cool, Ivana Williams and Fiona Griffin from the Chan Zuckerberg Initiative for this project, without whom we would not have come this far. We would also like to thank Mohammad Lotfollahi for early discussions

on Open Problems. E.V.B. would like to thank the Caltech Bioengineering Graduate program and Paul W. Sternberg for support. This work was supported by the Chan Zuckerberg Initiative Foundation (grant CZIF2022-007488, Human Cell Atlas Data Ecosystem) and the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation (grant number 2021-235155) awarded to M.D.L., D.B.B., S.G., F.J.T. and S.K. This work was co-funded by the European Union (ERC, DeepCell -101054957, to A.S. and F.J.T.). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. G.P. is supported by the Helmholtz Association under the joint research school Munich School for Data Science and by the Joachim Herz Foundation. Throughout this work, W.L. was supported by the US National Institutes of Health under Continuing Education Training Grants (T15). D.D. was supported by the European Union's Horizon 2020 Research and Innovation Program (860329 Marie-Curie ITN "STRATEGY-CKD"). M.E.V. is supported by the US National Institutes of Health under a Ruth L. Kirschstein National Research Service Award (1F31CA257625) from the National Cancer Institute. E.D. is supported by Wellcome Sanger core funding (WT206194). This work was supported by the Research Foundation Flanders (FWO) (1SF3822N to L.D.). B.R. is supported by the Bavarian state government with funds from the Hightech Agenda Bavaria. This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. C.B.G.-B. was supported by a PhD fellowship from Fonds Wetenschappelijk Onderzoek (FWO, 11F1519N). V.K. was supported by Wellcome Sanger core funding. G.L.M. received support from Swiss National Science Foundation grant PZ00P3\_193445 and Chan Zuckerberg Initiative grants number 2022-249212 and 2019-002427. D.R. was supported by the National Cancer Institute of the US National Institutes of Health (2U24CA180996).

## Author contributions

M.D.L., S.G., and D.B.B. conceived the idea. M.D.L., S.G., D.B.B., R.C., and O.B.B. developed the infrastructure. M.D.L., S.G., D.B.B., R.C., D.C.S., N.S.M., L.Z., G.P., W.L., D.D., M.E.V., M.F.M., A.A., E.D., Q.Q., A.S., A.B., and Z.L. formalized a benchmarking task. M.D.L., S.G., D.B.B., R.C., D.C.S., N.S.M., L.Z., G.P., W.L., D.D., M.E.V., D.S.M., M.F.M., A.A., E.D., Q.Q., D.J.O., M.K., O.B.B., K.W., S.N.Y., A.S., A.B., Z.L., C.A.-E., E.d.V.B., A.T.C., B.D., C.E., V.K., H.S., V.S. and A.T. contributed to the codebase. M.D.L., S.G., R.C., D.C.S., N.S.M., L.Z., G.P., W.L., D.D., L.D. and K.W. analyzed the results. M.D.L., S.G., D.B.B., J.M.B., A.O.P., J.S.-R., D.W., L.P., Y.S., F.J.T. and S.K. provided resources and supervised the work. M.D.L., S.G., D.B.B., R.C., D.C.S., N.S.M., L.Z., G.P., W.L. and D.D. coordinated the research. M.D.L., S.G., D.B.B., F.J.T. and S.K. acquired funding for the work. M.D.L., S.G., D.B.B., R.C., D.C.S., N.S.M., L.Z., G.P., W.L., D.D., M.E.V., M.F.M., A.A., E.D., Q.Q., D.J.O., M.K., O.B.B., A.S., A.B., Z.L., B.R., J.M.B., A.O.P., C.A.-E., E.d.V.B., A.B., C.B.G.-B., A.T.C., B.D., C.E., S.F., A.G., S.H., Y.J., V.K., G.L.M., M.G.L., R.L., D.R., H.S., V.S., A.T., G.X. and C.X. contributed to benchmarking task definition. M.D.L., S.G., D.B.B., R.C., D.C.S., N.S.M., L.Z., G.P., W.L., D.D., M.E.V. and D.S.M. prepared the manuscript. D.C.S., N.S.M., L.Z., G.P., W.L., D.D., M.E.V., D.S.M. and M.F.M. contributed equally as second authors. All authors reviewed the manuscript.

## Competing interests

M.D.L. consults for CatalYm GmbH, contracted for the Chan Zuckerberg Initiative and received speaker fees from Pfizer and Janssen Pharmaceuticals. S.G. has equity interest in Immunai Inc. D.B.B. is a paid employee of and has equity interest in NVIDIA. R.C. has equity interest in Data Intuitive BV. L.Z. has consulted for Lamin Labs GmbH. W.L. consulted for Protein Evolution Incorporated. From 2019 to 2022, A.A. was a consultant for 10x Genomics. From October 2023, E.D. has been a consultant for EnsoCell Therapeutics. O.B.B. is currently an employee of Bridge Bio Pharma. A.S. consults for Cellarity Inc. and Exvivo Labs Inc. A.B. is a paid employee of and has equity interest in Cellarity, Inc. J.B. has equity interest

in Cellarity, Inc. J.S.-R. reports funding from GSK, Pfizer and Sanofi and fees or honoraria from Trave Therapeutics, Stadapharm, Astex, Owkin, Pfizer and Grunenthal. D.W. has equity interest in Immunai Inc. F.J.T. consults for Immunai Inc., Singularity Bio B.V., CytoReason Ltd and Cellarity, and has ownership interest in Dermagnostix GmbH and Cellarity. S.K. is a visiting professor at Meta and scientific advisor at Ascent

Bio, Inc. E.d.V.B has ownership interest in Retro Biosciences and ImYoo Inc and is employed by ImYoo Inc. A.T.C. is an employee of Orion Medicines. B.D. is a paid employee of and has equity interest in Cellarity Inc. A.G. is currently an employee of Google DeepMind. Google DeepMind has not directed any aspect of this study nor exerts any commercial rights over the results. R.L. is an employee of Genentech.

V.S. has ownership interest in Altos Labs and Vesalius Therapeutics. A.T. has an ownership interest in Dreamfold.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41587-025-02694-w>.

## Open Problems Jamboree Members

**Bastian Rieck<sup>34,35</sup>, Constantin Ahlmann-Eltze<sup>36</sup>, Eduardo da Veiga Beltrame<sup>37</sup>, Carmen Bravo González-Blas<sup>38</sup>, Ann T. Chen<sup>26,39</sup>, Benjamin DeMeo<sup>28,40</sup>, Can Ergen<sup>41</sup>, Swann Floc'hlay<sup>38</sup>, Adam Gayoso<sup>41</sup>, Stephanie Hicks<sup>42,43,44</sup>, Yuge Ji<sup>1</sup>, Vitalii Kleshchevnikov<sup>21</sup>, Gioele La Manno<sup>45</sup>, Maximilian G. Lombardo<sup>46</sup>, Romain Lopez<sup>18,47</sup>, Dario Righelli<sup>48</sup>, Hirak Sarkar<sup>49,50</sup>, Valentine Svensson<sup>51</sup>, Alexander Tong<sup>17,52,53</sup>, Galen Xing<sup>41,54</sup> & Chenling Xu<sup>41</sup>**

<sup>34</sup>Institute of AI for Health, Helmholtz Munich, Neuherberg, Germany. <sup>35</sup>Department of Informatics, University of Fribourg, Fribourg, Switzerland.

<sup>36</sup>Genome Biology Unit, EMBL, Heidelberg, Germany. <sup>37</sup>Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates. <sup>38</sup>VIB Center for Brain & Disease Research, Leuven, Belgium. <sup>39</sup>Orion Medicines, Foster City, CA, USA. <sup>40</sup>Department of Biomedical Informatics, Harvard University, Cambridge, MA, USA. <sup>41</sup>Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA. <sup>42</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. <sup>43</sup>Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. <sup>44</sup>Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA. <sup>45</sup>Brain Mind Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. <sup>46</sup>Chan Zuckerberg Initiative, Redwood City, CA, USA. <sup>47</sup>Department of Genetics, Stanford University, Stanford, CA, USA. <sup>48</sup>Department of Statistical Sciences, University of Padua, Padua, Italy. <sup>49</sup>Department of Computer Science, Princeton University, Princeton, NJ, USA. <sup>50</sup>Princeton Ludwig Institute, Princeton University, Princeton, NJ, USA. <sup>51</sup>Altos Labs, San Diego, CA, USA. <sup>52</sup>Mila-Quebec AI Institute, Montreal, Quebec, Canada. <sup>53</sup>Université de Montréal, Montréal, Quebec, Canada. <sup>54</sup>Gladstone-UCSF Institute of Genomic Immunology, San Francisco, CA, USA.