nature cell biology

Perspective

https://doi.org/10.1038/s41556-025-01619-8

Biases in machine-learning models of human single-cell data

Received: 4 March 2024

Accepted: 9 January 2025

Published online: 19 February 2025

Check for updates

Theresa Willem[1,2] ✉, Vladimir A. Shitov[3,4], Malte D. Luecken[3,4], Niki Kilbertus[2,5,6], Stefan Bauer[2,5,6], Marie Piraud[2], Alena Buyx[1] & Fabian J. Theis[2,5,7] ✉

Recent machine-learning (ML)-based advances in single-cell data science have enabled the stratification of human tissue donors at single-cell resolution, promising to provide valuable diagnostic and prognostic insights. However, such insights are susceptible to biases. Here we discuss various biases that emerge along the pipeline of ML-based single-cell analysis, ranging from societal biases affecting whose samples are collected, to clinical and cohort biases that influence the generalizability of single-cell datasets, biases stemming from single-cell sequencing, ML biases specific to (weakly supervised or unsupervised) ML models trained on human single-cell samples and biases during the interpretation of results from ML models. We end by providing methods for single-cell data scientists to assess and mitigate biases, and call for efforts to address the root causes of biases.

Single-cell dataset sizes have recently begun to reach the population level, that is, encompassing hundreds or a thousand individuals[1–3]. At these scales, it is, for the first time, becoming possible to represent variation between samples from single-cell data using machine-learning (ML) techniques. Modelling sample variation has diagnostic and prognostic applications if the modelled variation is indicative of disease progression or remission. Because ML models are known for perpetuating biases from the datasets on which they are trained[1], it is important to pay attention to the multiple biases that can affect single-cell data-based ML tools and reflect on how these biases can be mitigated.

A plethora of biases have been described that can influence ML outputs in discriminative and hence problematic ways (for a more nuanced definition of bias and examples of harmful and less harmful biases, see Box 1), and there is no universally accepted definition of bias as a concept. Following previous definitions that have highlighted the unwanted social consequences of bias[4,5], we define bias as a systematic distortion of the ML model outputs that leads to ethically and/or socially unwanted effects, such as discrimination, misdiagnosis and under- or overtreatment.

Systematic distortions of ML model results can arise for multiple (unintentional) reasons. For example, owing to the underrepresentation of a certain genetic profile, the training data can contain too few examples from that group to produce outputs reliable for them, disadvantaging the underrepresented group[5–7]. Another phenomenon leading to bias, even when working with apparently perfectly balanced datasets, can be a historical inequity skewing the collected data[8] or too vague or broad definitions of the categorical data collected[9,10]. Such biases have been shown to confuse AI genomic studies, disadvantaging minority groups with less reliable predictions[11]. Owing to the pluralism of biases and sources of bias, a context-sensitive assessment of the relevant origins and types of bias in ML models is needed. The focus of this Perspective is biases related to ML models trained on human single-cell data. With this, we hope to contribute to the (computational) single-cell genomics community's understanding of possible biases in results obtained from ML models trained on human single-cell samples and how these biases arise, and to provide methods to mitigate these biases.

To identify single-cell ML-model-relevant biases, we first summarize recent advancements in the field of ML-based single-cell genomics

[1]TUM School for Medicine and Health, Institute of History and Ethics in Medicine, Technical University of Munich, Munich, Germany. [2]Helmholtz Munich, Munich, Germany. [3]Department of Computational Health, Institute of Computational Biology, Helmholtz Munich, Munich, Germany. [4]Comprehensive Pneumology Center (CPC) with the CPC-M bioArchive and Institute of Lung Health and Immunity (LHI), Helmholtz Munich; Member of the German Center for Lung Research (DZL), Munich, Germany. [5]School for Computation, Information and Technology, Technical University of Munich, Munich, Germany. [6]Munich Center for Machine Learning (MCML), Munich, Germany. [7]School of Life Sciences, Technical University of Munich, Munich, Germany. ✉e-mail: theresa.willem@helmholtz-munich.de; fabian.theis@helmholtz-munich.de

BOX 1

Bias definition and examples of harmful and less harmful biases

Etymologically descending from the French word 'biais', which translates to 'a slant, a slope, an oblique', biases refer to systematic distortions of datasets and statistical models (https://www.etymonline.com/word/bias). Such distortions, or tilts, reflect the data distribution in the dataset but contradict reality. Statistical predictions generated with biased data will reflect the biased data distribution and produce biased outputs. If unmitigated, biased outputs can have detrimental consequences for individuals who become, for instance, misclassified or otherwise affected by biased outputs[8], making bias an important ethical issue. It is, however, crucial to realize that different types of bias affect the ethicality of statistical models in different ways. Certain biases that align with the intended use of the predictive model and the context in which the training data were collected may be unproblematic or even essential for enhancing the predictive accuracy of the model. A desired bias would be created, for example, by deliberately oversampling patients of African descent in a study investigating the genetic basis of sickle cell anaemia. Because the sickle-cell trait is more common in this population, their overrepresentation provides richer data for identifying genetic variations associated with the condition. This bias is unproblematic because the sickle-cell trait is not causally affected by the region from which people originate. Therefore, research results obtained from such biased data can be applied to all people who share the same sickle-cell trait.

Certain other biases, however, are problematic. Mostly grounded in historically grown societal biases prevalent before data collection, problematic biases result from using biased proxy variables to predict the target variable[83], for example, predicting individuals' health risks based on proxies such as who was treated for which illness in the past. Access to care is socially biased owing to systematic medical disadvantages. For instance, Black patients have historically only received treatment when their conditions were worse on average than those of white patients. If data are collected on when and for what individuals received medical treatment, the racially biased cause–effect relationship is reflected in the data. If used to inform future medical decision-making, the predictive model trained on the biased data perpetuates the bias. One such model has been shown to assign lower risks to Black patients who were equally sick as white patients[8]. The application of the model in clinical practice thus resulted in undertreatment and denied access to care for Black patients. It has resulted in errors and, consequently, caused harm to individuals, thus violating the ethical principles of non-maleficence and fairness[84,85].

and briefly illustrate the development pipeline of ML models based on human single-cell samples. We then identify the biases that can occur by following the steps performed to train an ML model with human single-cell data. Such a pipeline-informed bias analysis can examine how various biases relate to and interfere with each other, potentially magnifying their effects and complicating their mitigation. All these biases should be considered when assessing the ethicality of ML models trained on human single-cell data.

Recent advances in ML for human single-cell data

Electronic health record data, such as health state, disease severity, organ function and laboratory test results, are a gateway for gaining valuable insights into variations within clinical conditions. The concurrent evolution of single-cell sequencing and its analysis methods leverage such newly available data and combine them with their donor's transcriptomics data to measure how differential expression of genes[12–14], or the differential abundance of cell types[15,16], relate to their donor's health status. Recently, ML methods have emerged to aid in modelling clinical covariates, such as disease status or severity, from single-cell transcriptomic data[17–20]. The first generation of such approaches required an analyst to specify a covariate of interest and did not allow unsupervised data exploration[15,17]. Thus, if structures in the data could be explained by other measured or unmeasured features or technical factors that the analyst does not take into account, they may go unnoticed. To address this challenge, recent single-cell analysis methods have been proposed to learn representations of human donor variation by generating an unsupervised embedding of single-cell samples. These so-called patient representation methods use large single-cell datasets to stratify healthy individuals and patients into multiple groups that capture the underlying molecular differences indicative of health status. They aim to use the generated insights to determine the diagnostic or prognostic potential in downstream analysis. Available patient representation methods include optimal transport-based methods, such as PhEMD[21], PILOT[22] and Diffusion Earth Mover's Distance[23], which model differences between donors as the amount of effort it would take to transform one transcriptomics signature to another. Variational inference-based methods, such as MrVi[24] and scPoli[25], train neural networks to embed cell-level information, correct batch effects and simultaneously learn patient variability. Other methods, such as GloScope[26] and IDEAS[27], calculate divergences between estimated probability densities of human transcriptomic profiles, while MultiscalePHATE[28] is a diffusion-based method that provides a representation of data on several levels of granularity.

The standard pipeline for ML models of human single-cell data

Data collection for a typical single-cell RNA-sequencing study with human samples starts in a clinic where tissue is taken from patients or healthy volunteers. These samples typically have to be stored before being processed in a laboratory. In the currently most widely used laboratory protocol developed by 10X Genomics[29], single cells are isolated and lysed, and mRNA is captured. The mRNA is then reverse-transcribed into DNA, amplified, and barcoded to track the cell and molecule of origin through further experimental steps. After sequencing the DNA, the sequencing reads are aligned to a reference transcriptome, remapped to the cell of origin using the cellular barcode, and counted. Computational analysis usually begins at this step, with count tables containing information about the number of gene copies for each cellular barcode. An analyst performs quality control and preprocessing of the data, including normalization, highly variable gene selection, and dimensionality reduction, followed by clustering for cell type identification and annotation of the obtained cell clusters. The preprocessed data can then be used for downstream analysis to derive information about gene, cellular or patient variation. Biases can be associated with any of the stages necessary to derive single-cell representations. To mitigate the biases that are specific to single-cell representations, it is crucial to disentangle those biases.

Biases along the pipeline of ML models of human single-cell data

Various biases affect single-cell model training. These biases start at sample collection and end with the interpretation of model results. We categorize these biases as societal, clinical, cohort, single-cell sequencing, ML and result-interpretation biases. We discuss these according to their emergence throughout the pipeline of ML models of human single-cell samples (Figs. 1 and 2). For an example of how multiple described biases can be present in a single, otherwise invaluable dataset, see Box 2.

https://doi.org/10.1038/s41556-025-01619-8

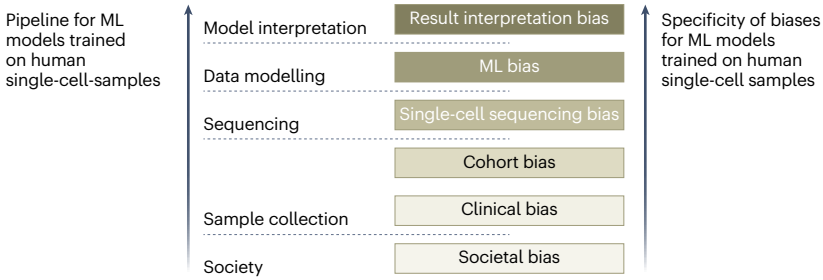

Fig. 1 | Biases occurring along the pipeline. Biases occurring along the pipeline increase in specificity for ML models of human single-cell samples.

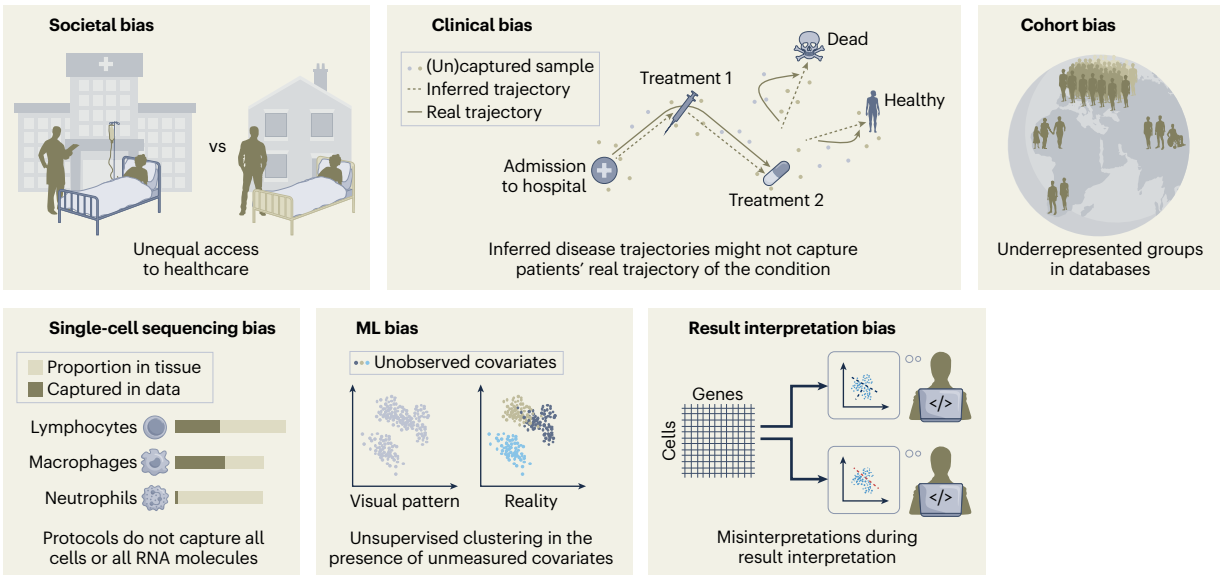

Fig. 2 | Illustration of societal, clinical, cohort, single-cell sequencing, ML and result-interpretation biases relevant in ML models of human single-cell data. Societal bias stems from structural inequalities within healthcare systems. Clinical bias arises from inconsistencies in clinical data collection and processing that introduce variability. Cohort bias stems from skewed cohort compositions, including demographic and clinical diversity. Single-cell sequencing bias stems from technical limitations and variability in sequencing methodologies that affect data quality. ML bias originates from the selection of algorithms and model parameters that can generate or amplify biases. Finally, result-interpretation bias stems from flexible visualization and interpretation tools that perpetuate or reinforce biased conclusions, influencing downstream scientific and clinical insights.

Societal bias

Societal biases are biases that arise with systematic social behaviour, such as discrimination based on sex, gender or race, and are deeply ingrained in human populations. These biases influence who is granted access to care and thus, as human single-cell samples are taken in clinics, also orchestrate who donates single-cell samples. One prominent example of a societal bias was an ML model that was used to manage the health of the US population but was biased against people of colour because of historical inequalities entrenched in the healthcare sector[8]. Similarly, other societal biases, including socioeconomic factors such as income, country of origin, area of living and education, to name but a few, are shaping current healthcare systems and determine which populations need and are granted access to care[30,31] and thus can or cannot become single-cell sample donors.

In addition to current societal biases, past societal biases must be considered. For an individual's genetics, it matters, for example, where their ancestors lived and which cultural traditions they adhered to in this area. For example, people who lived for several generations in places that provided them with starch-heavy produce have been shown to have different gene copy number variations than populations with other diets[32,33]. This shows that biological diversity in individuals forms through evolutionary pressure created by various factors, including societal processes. Although research connecting specific intergenerational ways of living and their modern kin's genetic expressions is still in its infancy, there is evidence that most genetic variants that influence disease risk have ancestry-specific evolutionary origins[34], making it relevant to consider societal biases—past and present—that might affect model performance in single-cell studies.

Clinical bias

In the clinic, at the point of RNA sample collection, clinical biases, introduced through varying clinical definitions, limit the reliability of the eventual dataset on which the single-cell analysis will be based. For example, an inclusion criterion for sample extraction might be that a participant is 'healthy'. Although 'healthy' is a blurry concept initially, tissue sample donors might suffer severe clinical conditions, such as lung cancer, but donate 'healthy' tissues that are unaffected by the disease. In other cases, donors might suffer genetically relevant comorbidities or are clinically considered 'healed'. In these cases, the samples are subsequently considered 'healthy' (or 'normal' or 'control') in the database, despite the specific health-related circumstances of the individual potentially having considerable effects on their transcriptomic profile. Another example of clinical bias is when sample collection is triggered by the change of a clinical parameter (Fig. 2 and Box 2). For example, when the condition of patients with COVID-19 deteriorates, blood might need to be drawn for testing. If the same blood is used as a sample for a single-cell database, it can lead to an oversampling of samples taken at a specific time point instead of the

BOX 2

Biases in peripheral blood mononuclear cell data from UK COVID-19 patients—an example of how biases can accumulate

To illustrate how biases can propagate to clinical conclusions, we refer to a COVID-19 single-cell dataset published by the Cambridge Institute of Therapeutic Immunology and Infectious Disease–National Institute of Health Research (CITIID–NIHR) COVID-19 BioResource Collaboration[86] and its subsequent analysis. The study explores single-cell RNA sequencing of peripheral blood mononuclear cells (PBMCs) in patients with COVID-19 from the UK. Although the dataset is highly valuable, several biases must be acknowledged.

The study focused on the progression of the disease rather than on recovery. Individuals with positive tests for SARS-CoV-2 were included, and patients with symptoms but with a negative test for the virus were intentionally excluded. Although this is a reasonable approach for a specific goal and in the condition of limited resources, it is important to recognize that the dataset misses a fraction of the clinical trajectory of some patients, where they are infected but do not yet produce a positive test. This could introduce a clinical bias. In terms of cohort, the dataset derives from samples collected within England, specifically from Cambridge, Newcastle and a smaller cohort in London. Donors' ethnicities are labelled as 'unknown', but the geographic concentration suggests limited ethnic diversity and thus probably induces a cohort bias, which restricts the generalizability of findings to more diverse global populations. For processing, each study site applied different laboratory and computational protocols. For example, every institute used a different PBMC isolation technique and different doublet-removal strategies performed by different individuals, thus introducing sequencing biases that contribute to batch effects in the data. One such batch effect distinguished the Cambridge cohort from the two other sites, which translated to the downstream ML models used: the PILOT authors observed an unsupervised ML bias in their model, in which "a variable describing the city of origin of the sample had a stronger association with the clustering than the status variable"[22]. The authors of this dataset had already anticipated this issue in the original paper[22], but it could not be mitigated. This shows that although biases associated with single-cell datasets can be detected and meaningfully discussed when anticipated, they bear the risk of going unnoticed and/or being unmanageable, thus contributing to interpretation biases.

The dataset from this study is one of the largest COVID-19 single-cell PBMC datasets published and is extensively used. While being an invaluable source of information, it also teaches us that reflecting on the entrenched biases is of utmost importance to hinder them from perpetuating along the analysis pipeline.

desired representation of the full clinical trajectory of the illness. In other words, cell samples can be biased by reflecting only certain disease states rather than providing insight into the full condition. Even if samples are taken in stable conditions of illness, varying sampling locations—for example, which anatomical location organ tissue is taken from across donors—can introduce clinical biases.

Other clinical biases to consider are, for example, how other covariates influence clinical samples in unintended ways. For example, processes that are closely linked to metabolism, such as circadian rhythm, currently remain largely unaccounted for in single-cell studies[35,36]. Lifestyle choices and socioeconomically influenced characteristics, such as smoking, drinking, cardiovascular health, biological age, stress and sun exposure, could equally introduce clinical biases in yet understudied ways.

Cohort bias

After RNA sample collection, one primary challenge in ML-based single-cell analysis is that sample sizes are often small. Some of the largest single-cell studies comprise hundreds of samples (for example, 982 individuals[1], 428 individuals[2] or 284 individuals[3]; for more information on single-cell studies and their characteristics, see the dynamic Single Cell Studies Database[37]). Although technological improvement in the single-cell field has led to an exponential increase in the number of cells, the median number of samples in an experiment was 11 in 2022[37]. This limits the ability of many studies to generalize their conclusions, because, if specific genetic profiles are underrepresented, the conclusions can be impaired for this specific group, giving rise to a potential cohort bias (Fig. 2). When ML models are trained on such imbalanced datasets, they prioritize learning the nuances of the overrepresented group while overlooking the diversity within others[8]. Consequently, when such models are used to generate patient clusters, they will be more proficient at distinguishing subgroups with similar genetic profiles, purely owing to the number of data points available, while probably failing to do the same for individuals of underrepresented genetic profiles. This bias extends to the interpretation of clusters, affecting their usefulness in stratifying individuals not belonging to the overrepresented group for personalized healthcare recommendations.

In addition, variances in data-collection practices of categorial data across study sites make it difficult to test for potential performance fluctuations among demographic groups that could perpetuate racism. Genomic sequencing was expected to end the long-running debates on the biological basis of racial categories[38,39], by showing that genomic profiles do not correlate with social constructs such as racial categories[9,10,40]. While this debate is still ongoing, there is increasing consensus that such categorical data should not be used for training genetic ML models[41]. Nevertheless, it is crucial to collect such data to, after models are trained, make sure certain demographic groups are not accidentally systematically disadvantaged, particularly if the same groups already suffer marginalization[42–45]. However, categorical data-collection practices seem to need updating. For example, in the CellxGene Discover database (Fig. 3 in ref. 29), we note a somewhat unclear breakdown of area of origin and an underrepresentation of certain groups; for example, there is one category for 'African American' and another for 'African American or Afro-Caribbean'. Such overlaps between data categories complicate the assessment of fairness in ML model outputs, especially for models trained on supplementary data from electronic health records, which may introduce additional biases.

Single-cell sequencing bias

Single-cell sequencing, that is, the determination of each sequence of nucleotides in the data, is a complex workflow that is, despite being very automized, prone to errors, and is limited by the technical hardware and software used. After single-cell samples are obtained, the first step of analysis is to prepare a sequencing library from the obtained sample. During this process, the sample is dissociated into its cellular components, cells are isolated into droplets or wells, the cellular RNA is tagged and captured, before being converted to double-stranded DNA and enriched with sequencing adapters that allow further processing of specific fragments[46].

This library preparation does not work equally well for all cell types (Fig. 2). For example, when the samples are highly heterogeneous in

the number of captured molecules or the dissociation process affects some cell types more than others, the cellular composition of the sample is altered. For example, stromal cells are more likely to break under the stress placed on them during dissociation, biasing the possibility of studying single-cell processes against diseases that are driven by stromal cells. Moreover, technical processes can infer differences in molecular quantity across samples, for example, when a particular read to a cell is misassigned (so-called index hopping)[47]. Such errors naturally overly affect cell states with smaller transcriptomes (for example, T lymphocytes), biasing opportunities to identify biological processes and diseases that are related to these cell types. Normalization methods are established as a crucial step in single-cell pipelines to counter differences in the number of reads per cell due to artefacts in sequencing or in the number of reads captured. Recently, however, scholars have (again) pointed to the limitations of these methods[48,49].

Moreover, during any sequencing process, so-called 'batch' effects can arise from handling cells and samples in distinct groups[50]. These batch effects can generate biases when they are collinear with the biological covariates of interest, such as the anatomical location from which the sample was taken, sex or age. In disease studies in particular, it is challenging to avoid collinearity of batch effects and disease conditions. Removing this technical effect during preprocessing may, however, also remove the biological signal, thus affecting the conclusions that can be drawn from such studies.

ML bias

Algorithms are not sources of biases per se, but a crucial issue when applying ML methods, such as those proposed for single-cell datasets, is the potentially uncontrollable amplification of the mentioned biases. Biases stemming from hidden covariates, batch effects[51,52] and unbalanced population compositions in the dataset can result in a model that focuses on overrepresented characteristics and/or populations[8], scaling the prevalent biases to ML biases. For example, clusters generated by an unsupervised clustering algorithm could be influenced by observed patient covariates such as batch effects, as well as unmeasured patient covariates, rather than the biological effect of interest (Fig. 2). During supervised data analysis, other distortions might occur that can be specific to this step and a particular model type. For example, simple linear regression is susceptible to outliers[53], tree-based models generalize poorly to unseen data ranges, and more complex models can overfit the training data and thus similarly generalize poorly to samples out of the distribution of the training data[54], making the results unreliable.

Although balanced accuracy[55] is applied to minimize such performance shortcomings in supervised models, reliable mitigation methods for unsupervised models are lacking, and resampling is currently the best available option to balance datasets. Resampling, however, is limited in utility, as the resampled groups can be unrepresentative and hence impair the validity of the conclusions.

Result-interpretation bias

Finally, analysis biases can distort result interpretation (Fig. 2). In recent debates, for example, tunable visualization methods have met push-back[56]. Specifically, relative distances between cells in plots created with dimensionality reduction tools, such as *t*-distributed stochastic neighbour embedding (*t*-SNE) and uniform manifold approximation and projection (UMAP) for dimension reduction have been discussed as overinterpretative[57]. By design, methods that visualize high-dimensional data in two dimensions cannot accurately reflect all distances inherent in the high-dimensional data in a two-dimensional visualization.

Indeed, these visualization methods can produce different images with variable distances between cell types depending on the stochasticity in the algorithm rather than the properties of the data, and, in both methods, the amount of space in the plot taken up by cells of a particular type is a function of their abundance and not their relative similarity. These effects distort visualization and can bias the interpretation of results when it is based on these plots alone. More research on analysis biases is needed, and it seems clear that the available, non-trivial tools can only be navigated with knowledge of both single-cell biology and the ML domain. The potential lack of experience in both these fields makes inadvertent mistakes and misinterpretations of variance in results likely. Especially when interpreting cellular clusters and trajectories based on potentially biased visualizations, invalid inferences are difficult to detect owing to a lack of ground-truth annotations in exploratory studies. In addition, other types of cognitive bias, such as confirmation bias, exist, which can further impair result interpretation. Together, this can lead to the publication and dissemination of analytically biased results.

Towards mitigating biases in single-cell model results

Genomics has a long history of dealing with various biases, such as population stratification and technical artefacts. In genome-wide association studies in particular, as well as pharmacogenomics and clinical research generally, much work has been conducted to reflect upon and mitigate biases[58–63]. Moreover, biases, especially those associated with insufficient diversity in genomic datasets, are increasingly discussed at contemporary ML conferences[64] and considered by initiatives such the Human Cell Atlas White Paper, which states "[d]iversity, inclusion and equity" [65] as values of the consortium and asserts that the first draft atlas includes "a minimum of 20 ethnically diverse samples" for each tissue. Contributing to these efforts, we advocate for several practical steps to counteract the biases in single-cell models trained on human samples.

Scrutinizing data categories to ensure comprehensive sampling

Data categories collected as metadata for single-cell databases exhibit considerable variability owing to differences in data-collection practices and categorical definitions across countries. Additionally, local legal restrictions can substantially influence the nature and extent of metadata collected. To alleviate cohort and clinical biases, it is crucial to undertake interdisciplinary dialogues in data category differences, advocate for collecting these metadata categories, and advance methods to correct the effects different metadata categories produce in different settings. Scholars involved in these efforts must carefully consider criteria relevant to single-cell models, informed by the potential use cases of the datasets, and should 'think outside the data', that is, actively consider omitted variables[66]. By communicating reflections on data categories, testing for biases across datasets will become more accessible.

Collecting large(r) and more diverse datasets

To alleviate cohort biases, data collectors should carefully consider relevant demographic criteria informed by potential use cases of the dataset. Moreover, single-cell database publishers should include a report on dataset diversity and balance metrics to encourage contributors to generate data with a better representation of the general population[67,68]. Established frameworks assisting database creators in documenting and communicating relevant information about their dataset are 'data sheets for datasets'[69] or 'data cards'[70].

Moreover, sufficiently large datasets need to be collected to allow such breakdowns. Researchers in the field should be able to test whether the performance of an ML model differs significantly between unitary groups (for example, by ancestry) and intersectional groups[71] (for example, younger and Afro-American ancestry or older and Afro-American ancestry) in the test dataset. Alternatively, data collectors and single-cell analysts should work together to create datasets and models that are particularly tailored for a specific context and hence can then be tested for reliability for the intended use. Emphasizing collaborative efforts, federated mechanisms for 'data collection

requests' would enable researchers to collectively identify underrepresented subgroups, communicate insights to data collectors, and encourage targeted data collection among the identified subgroups.

Advancing covariate testing and data correction

Clinical biases can best be post hoc mitigated by randomizing for the respective covariates. When this is not possible, or when datasets are too small to contain effectively randomized samples for multiple covariates of interest, one can instead model the effect and correct it in the data. This strategy is commonly used to correct for cell-cycle effects in single-cell data[72]. Similar approaches have been proposed to model (and correct for) the circadian rhythm[31] and could be extended to correct for diet using metabolic flux models that use cell data[73]. However, more research is needed to determine the scale of the impact on each covariate. This research should consider that biases based on covariates might be much harder to detect than in other disciplines. For example, sun damage will probably have a role in single-cell studies of human skin samples. However, compared to the processing of dermatological images, where sun damage is visually classifiable, dedicated studies are needed to detect such 'hidden factors' in single-cell data. Methods that can detect these 'hidden factors' and study how biases encrypt in latent spaces, particularly while their effects are collinear with the studied effect, are urgently warranted.

Promoting further efforts in single-cell processing

The biases imposed by the technical steps of library preparation and sequencing are a known and discussed issue. Every single-cell analysis workflow already includes steps to mitigate technical artefacts in the data, such as normalization, quality control and data correction[72,74]. Indeed, over 200 batch-effect correction methods are readily available[75]. Promising methods to mitigate batch effects during data generation already include multiplexing strategies that avoid library generation artefacts affecting samples differently by tagging and pooling all samples for joint library preparation[76–78]. Yet, correcting for batch effects collinearly with the biological effects of interest remains an open challenge, and choosing the right method with the optimal parameter settings requires being able to evaluate the performance of these methods. Although the latter is possible through evaluation frameworks, such as scIB[52] and Open Problems in Single-Cell Analysis[79], these frameworks rely on the availability of metadata that describe the possible origins of the batch effect (for example, laboratory protocols or equipment used). Thus, large-scale metadata collection, beyond addressing cohort biases, is also instrumental in reducing single-cell processing biases. Another particularly relevant metadata covariate that is beneficial to collect is the location where the samples were taken. Given inter-individual differences in anatomy and the often challenging contexts in which samples are taken, the anatomical location of a sample is often not recorded accurately. Finding standardized ways of recording such anatomical locations at high resolution, or developing methods to predict anatomical location from sampled data, would facilitate correcting for differences in sampling locations that affect the cellular compositions of samples. Furthermore, benchmarks themselves may introduce biases by favouring suboptimal methods when the ranking metrics are flawed. As an example, in one study[80], a simple method outperformed other popular data-integration tools based on conventional metrics, yet it distorted the biological structures in the data, such as developmental stage, structure and cell relationships. Living benchmarks, such as Open Problems in Single-Cell Analysis[79], have the potential to mitigate this bias by enabling community contributions that can update metrics and add new datasets. In summary, addressing the aforementioned problems will improve existing benchmarks and allow researchers to select the best-performing tools for more efficient single-cell data processing.

Reporting of limitations of intended-use-specific models

A recent review suggests that biases can be alleviated if benchmarking studies include the biases present and missing information in the datasets they assess[68]. Established frameworks to consider model biases in a context-dependent manner from a ready-trained model include the domain-agnostic 'model cards for model reporting'[81] framework and the healthcare-specific TRIPOD Statement[82] framework. By creating such cards or statements, developers collect and communicate detailed information about the model itself, its intended use, relevant demographic factors, training and evaluation data used, as well as metrics used and other considerations, in a written document and provide this information along with the model upon publication. This is a valuable starting point to identify biases entrenched in the data used for training and to generate new insights about biases; importantly, it also offers the prospect of educating those working with or utilizing the research created through the models about the entrenched biases.

Implementing fairness testing

Parity metrics are an easily applicable and established diagnostic tool to evaluate potential differences in model performance across different population groups, thus assessing the fairness of a model[5]. Such metrics can, for example, surface whether a predictive model has different error rates for different demographic groups or assigns one subgroup to a certain treatment much more frequently than another. Parity metrics analyse the overall bias of a model in terms of predictive performance; that is, they can typically not identify the source of existing biases (for example, distinguish between ML bias and cohort bias), but they can serve as useful pointers of where further scrutiny is needed. Similarly, the disparity of certain predictive metrics across subgroups does not need to constitute harmful bias per se (different treatments may be more effective for different populations). Standardizing this reporting enables the rapid uptake of subgroup-specific information, improves the likelihood of compliance reporting this information, and enables meta-analyses.

One caveat of parity testing is that the relevant subgroups must be known and manually selected beforehand. Similarly, when many metrics and many group combinations are tested against each other, the analyst has to consider multiple testing corrections. A related issue arises in intersectional parity testing, where multiple demographic factors, such as genetic ancestry, gender and age, lead to a combinatorially increasing number of intersectional subgroups, posing substantial statistical challenges to parity testing. Nevertheless, we encourage analysts to integrate (intersectional) parity testing directly into their model-building pipeline by—following best practices from unit testing—defining, continuously updating, and monitoring an extensive set of parity test scenarios tested throughout the model design and training phase. When computationally accounting for detected biases, it is, however, crucial to ensure that improving bias for one population group leads to overall better model performance and does not come at the detriment to another. Additionally, non-statistically significant results of the parity tests should not be interpreted as the absence of biases but rather as an indication that the data do not provide sufficient evidence to detect any potential biases.

Conclusions

ML methods are fast approaching the single-cell domain. In this Perspective we have discussed how societal biases, clinical biases, cohort biases, single-cell sequencing biases, (weakly or unsupervised) ML model biases, and result-interpretation biases emerge and influence study results achieved with ML models that are trained with human samples. The example of biases in PBMC data from UK patients with COVID-19 showed how multiple biases can be entrenched in a single dataset that otherwise is of immense value to the field (Box 2). We contribute to a growing body of literature that aims to mitigate single-cell associated biases by advocating for several ideas: (1) scrutinizing data categories to ensure comprehensive sampling, (2) collecting large(r) and more diverse datasets, (3) advancing covariate testing and data

correction, (4) promoting further efforts in single-cell processing, (5) reporting limitations of ML intended-use-specific models, and (6) implementing fairness testing. These recommendations are intended to serve as a starting point for a framework that promotes inclusivity, transparency and collaboration in single-cell analysis, fostering the reliability of findings in this rapidly evolving field.

References

1. Yazar, S. et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
2. Mathys, H. et al. Single-cell multiregion dissection of Alzheimer's disease. *Nature* **632**, 858–868 (2024).
3. Van Der Wijst, M. G. P. et al. Type I interferon autoantibodies are associated with systemic immune alterations in patients with COVID-19. *Sci. Transl. Med.* **13**, eabh2624 (2021).
4. Webster, C. S., Taylor, S., Thomas, C. & Weller, J. M. Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. *BJA Educ.* **22**, 131–137 (2022).
5. Barocas, S., Hardt, M. & Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities* (MIT Press, 2023).
6. Sun, W., Nasraoui, O. & Shafto, P. Evolution and impact of bias in human and machine learning algorithm interaction. *PLoS ONE* **15**, e0235502 (2020).
7. Miller, T. Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2017).
8. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
9. Yudell, M., Roberts, D., DeSalle, R. & Tishkoff, S. Taking race out of human genetics. *Science* **351**, 564–565 (2016).
10. Smedley, A. & Smedley, B. D. Race as biology is fiction, racism as a social problem is real: anthropological and historical perspectives on the social construction of race. *Am. Psychol.* **60**, 16–26 (2005).
11. Dai, B. et al. Racial bias can confuse AI for genomic studies. *Oncologie* **24**, 113–130 (2022).
12. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
13. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
14. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
15. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun.* **12**, 6876 (2021).
16. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).
17. Ravindra, N., Sehanobish, A., Pappalardo, J. L., Hafler, D. A. & Van Dijk, D. Disease state prediction from single-cell data using graph attention networks. In *Proc. ACM Conference on Health, Inference and Learning* 121–130 (ACM, 2020).
18. He, B. et al. CloudPred: predicting patient phenotypes from single-cell RNA-seq. In *Pacific Symposium on Biocomputing 2022* 337–348 (PSB, 2022).
19. Liu, C., Zhang, Y., Gao, X. & Wang, G. Identification of cell subpopulations associated with disease phenotypes from scRNA-seq data using PACSI. *BMC Biol.* **21**, 159 (2023).
20. Lotfollahi, M. et al. Biologically informed deep learning to query gene programs in single-cell atlases. *Nat. Cell Biol.* **25**, 337–350 (2023).
21. Chen, W. S. et al. Uncovering axes of variation among single-cell cancer specimens. *Nat. Methods* **17**, 302–310 (2020).
22. Joodaki, M. et al. Detection of Patlent-Level distances from single cell genomics and pathomics data with Optimal Transport (PILOT). *Mol. Syst. Biol.* **20**, 57–74 (2024).
23. Tong, A. et al. Diffusion Earth mover's distance and distribution embeddings. Preprint at https://doi.org/10.48550/arXiv.2102.12833 (2021).
24. Boyeau, P. et al. Deep generative modeling for quantifying sample-level heterogeneity in single-cell omics. Preprint at *bioRxiv* https://doi.org/10.1101/2022.10.04.510898 (2022).
25. De Donno, C. et al. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat. Methods* **20**, 1683–1692 (2023).
26. Wang, H., Torous, W., Gong, B. & Purdom, E. Visualizing scRNA-seq data at population scale with GloScop. *Genome Biol.* **25**, 259 (2024).
27. Zhang, M. et al. IDEAS: individual level differential expression analysis for single-cell RNA-seq data. *Genome Biol.* **23**, 33 (2022).
28. Kuchroo, M. et al. Multiscale PHATE identifies multimodal signatures of COVID-19. *Nat. Biotechnol.* **40**, 681–691 (2022).
29. CZI Single-Cell Biology Program et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.30.563174 (2023).
30. Gomez, D. et al. Gender-associated differences in access to trauma center care: a population-based analysis. *Surgery* **152**, 179–185 (2012).
31. Blair, I. V., Steiner, J. F. & Havranek, E. P. Unconscious (implicit) bias and health disparities: where do we go from here? *Perm. J.* **15**, 71–78 (2011).
32. Kowal, E. & Llamas, B. Race in a genome: long read sequencing, ethnicity-specific reference genomes and the shifting horizon of race. *J. Anthropol. Sci.* **96**, 91–106 (2019).
33. Perry, G. H. et al. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).
34. Benton, M. L. et al. The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.* **22**, 269–283 (2021).
35. Auerbach, B. J., FitzGerald, G. A. & Li, M. Tempo: an unsupervised Bayesian algorithm for circadian phase inference in single-cell transcriptomics. *Nat. Commun.* **13**, 6580 (2022).
36. Patel, V. R., Eckel-Mahan, K., Sassone-Corsi, P. & Baldi, P. CircadiOmics: integrating circadian genomics, transcriptomics, proteomics and metabolomics. *Nat. Methods* **9**, 772–773 (2012).
37. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**, baaa073 (2020).
38. Angier, N. Do races differ? Not really, genes show. *The New York Times* (22 August 2000).
39. Office of the Press Secretary. June 2000 White House Event. *The White House* (June 2000); https://www.genome.gov/10001356/june-2000-white-house-event
40. Saylor, K. W. & Martschenko, D. O. Promoting diagnostic equity: specifying genetic similarity rather than race or ethnicity. *J. Med. Ethics* **49**, 820–821 (2023).
41. McFarling, U. L. & Palmer, K. How race became ubiquitous in medical decision-making tools. *STAT* (4 September 2024); https://www.statnews.com/2024/09/04/embedded-bias-part-2-health-equity-racial-data-unintended-consequences/
42. Brandt, A. M. Racism and research: the case of the Tuskegee Syphilis Study. *Hastings Cent. Rep.* **8**, 21–29 (1978).
43. Leslie, C. Scientific racism: reflections on peer review, science and ideology. *Soc. Sci. Med.* **31**, 891–905 (1990).
44. Mohr, J. M. Oppression by scientific method: the use of science to 'other' sexual minorities. *J. Hate Stud.* **7**, 21 (2009).

45. Bashford, A. & Levine, P. *The Oxford Handbook of the History of Eugenics* (Oxford Univ. Press, 2010).
46. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
47. Ros-Freixedes, R. et al. Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. *Genet. Sel. Evol.* **50**, 64 (2018).
48. Booeshaghi, A. S., Hallgrímsdóttir, I. B., Gálvez-Merchán, Á. & Pachter, L. Depth normalization for single-cell genomics count data. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.06.490859 (2022).
49. Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nat. Methods* **20**, 665–672 (2023).
50. Davies, P., Jones, M., Liu, J. & Hebenstreit, D. Anti-bias training for (sc)RNA-seq: experimental and computational approaches to improve precision. *Briefings Bioinform.* **22**, bbab148 (2021).
51. Goh, W. W. B., Wang, W. & Wong, L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* **35**, 498–507 (2017).
52. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
53. Sun, H., Cui, Y., Wang, H., Liu, H. & Wang, T. Comparison of methods for the detection of outliers and associated biomarkers in mislabeled omics data. *BMC Bioinformatics* **21**, 357 (2020).
54. Loh, W. Fifty years of classification and regression trees. *Int. Stat. Rev.* **82**, 329–348 (2014).
55. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The balanced accuracy and its posterior distribution. In *Proc. 2010 20th International Conference on Pattern Recognition* 3121–3124 (IEEE, 2010).
56. Lähnemann, D. et al. Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
57. Chari, T. & Pachter, L. The specious art of single-cell genomics. *PLoS Comput. Biol.* **19**, e1011288 (2023).
58. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
59. Uffelmann, E. et al. Genome-wide association studies. *Nat. Rev. Methods Primers* **1**, 59 (2021).
60. Sul, J. H., Martin, L. S. & Eskin, E. Population structure in genetic studies: confounding factors and mixed models. *PLoS Genet.* **14**, e1007309 (2018).
61. Corpas, M. et al. Addressing ancestry and sex bias in pharmacogenomics. *Annu. Rev. Pharmacol. Toxicol.* **64**, 53–64 (2024).
62. Yakerson, A. Women in clinical trials: a review of policy development and health equity in the Canadian context. *Int. J. Equity Health* **18**, 56 (2019).
63. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475**, 163–165 (2011).
64. NeurIPS. NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines (2024).
65. Regev, A. et al. The Human Cell Atlas White Paper. Preprint at https://doi.org/10.48550/arXiv.1810.05192 (2018).
66. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**, 1–35 (2022).
67. Boakye Serebour, T. et al. Overcoming barriers to single-cell RNA sequencing adoption in low- and middle-income countries. Eur. *J. Hum. Genet.* **32**, 1206–1213 (2024).
68. Brooks, T. G., Lahens, N. F., Mrčela, A. & Grant, G. R. Challenges and best practices in omics benchmarking. *Nat. Rev. Genet.* **25**, 326–339 (2024).
69. Gebru, T. et al. Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).
70. Pushkarna, M., Zaldivar, A. & Kjartansson, O. Data cards: purposeful and transparent dataset documentation for responsible AI. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 1776–1826 (ACM, 2022).
71. Crenshaw, K. Mapping the margins: intersectionality, identity politics and violence against women of color. *Stanford Law Rev.* **43**, 1241–1299 (1991).
72. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
73. Wagner, A. et al. Metabolic modeling of single Th17 cells reveals regulators of autoimmunity. *Cell* **184**, 4168–4185.e21 (2021).
74. Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).
75. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14**, e1006245 (2018).
76. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
77. McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
78. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
79. Luecken, M. D. et al. Defining and benchmarking open problems in single-cell analysis. Preprint at *Research Square* https://doi.org/10.21203/rs.3.rs-4181617/v1 (2024).
80. Wang, H., Leskovec, J. & Regev, A. Metric mirages in cell embeddings. Preprint at *bioRxiv* https://doi.org/10.1101/2024.04.02.587824 (2024).
81. Mitchell, M. et al. Model cards for model reporting. In *Proc. Conference on Fairness, Accountability and Transparency—FAT* '19* 220–229 (ACM Press, 2019).
82. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* **162**, 55–63 (2015).
83. Obermeyer, Z. et al. *Algorithmic Bias Playbook* (Chicago Booth, 2021).
84. Floridi, L. *The Ethics of Artificial Intelligence: Principles, Challenges and Opportunities* (Oxford Univ. Press, 2023).
85. Beauchamp, T. L. & Childress, J. F. *Principles of Biomedical Ethics* (Oxford Univ. Press, 2019).
86. Cambridge Institute of Therapeutic Immunology and Infectious Disease–National Institute of Health Research (CITIID-NIHR) COVID-19 BioResource Collaboration et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021).

Acknowledgements

V.A.S. is supported by the Helmholtz Association under the joint research school 'Munich School for Data Science' (MUDS).

Author contributions

All authors contributed to the work presented in this manuscript. T.W. designed the original outline and wrote the first draft. V.A.S. contributed to the writing of the first draft and edited subsequent versions. M.D.L., N.K., S.B. and M.P., provided critical revisions and edits to multiple versions of the draft. A.B. and F.T. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Theresa Willem or Fabian J. Theis.

Peer review information *Nature Cell Biology* thanks Manuel Corpas and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.