EXPLORING QUERY-TO-REFERENCE MAPPING CHALLENGES FOR AUTOMATED SINGLE-CELL ATLAS-BASED DIAGNOSTICS

Francesco Craighero[1,*], **Davide Maspero**[2,*], **Laura Jiménez-Gracia**[2], **Sergio Aguilar-Fernández**[2], **Maria Boulougouri**[1], **Juan C. Nieto**[2,†] **& Holger Heyn**[2,†]
[1] Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland;
[2] Centro Nacional de Analisis Genomico (CNAG), Spain

francesco.craighero@epfl.ch, juan.nieto@cnag.eu, holger.heyn@cnag.eu

[*] Co-first authors; [†] Joint senior authors

ABSTRACT

Single-cell atlases are built by integrating multiple heterogeneous datasets into a common embedding space. The aim is reducing the dataset-specific biases or batch effects, while capturing the overall cellular composition and biological variability. One of the envisioned applications is automated diagnostics, where atlases are used as references to predict the phenotype of unseen patients. Here, we developed a diagnostic tool from a multi-disease atlas of inflammation. Moreover, we provided a benchmark of state-of-the-art integration methods for mapping and classifying unseen patients. In our tests, all the methods performed well when query batch effects are well represented in the reference, but mostly failed otherwise. Notably, linear integration approaches demonstrated superior robustness and reduced hyperparameter sensitivity compared to more powerful variational autoencoder-based methods. These findings highlight two fundamental challenges: the selection of the optimal integration method and the management of previously unobserved batch effects when classifying new query patients. As a viable solution, we designed and tested a Centralized experimental scenario where reference and query datasets are generated in the same center, demonstrating a potential pathway toward reliable atlas-based diagnostics.

1 INTRODUCTION

The rapidly growing availability of single-cell transcriptomic studies (scRNA-seq) Angerer et al. (2017), combined with the development of computational tools scaling to millions of cells Wolf et al. (2018), represents an unprecedented opportunity to uncover new cellular biology. Yet, substantial challenges remain in fully harnessing its potential. Indeed, the publicly available data is composed of heterogeneous datasets generated under different experimental conditions and from a limited number of individuals Hrovatin et al. (2025). As a consequence, significant efforts have been devoted into optimizing and standardizing the integration of multiple studies into a single resource, the so-called "atlas" Heumos et al. (2023); Hrovatin et al. (2025); Luecken et al. (2022; 2024). A large-scale, curated atlas is not only a powerful tool to generate more robust and generalizable analyses but also represents a reference to facilitate the analysis of new studies Lotfollahi et al. (2024).

Integrating multiple studies is a key challenge when building an atlas Lähnemann et al. (2020), as it requires eliminating technical biases, or "batch effects", such as those arising from differences in the chemistry assay or the sequencing platform, while preserving the underlying biological variability. Data integration methods tackle this issue by learning a harmonized representation space, such as with Variational Autoencoders Kingma & Welling (2014). Moreover, recently developed tools also enable mapping new query studies onto an existing reference atlas De Donno et al. (2023); Kang et al. (2021); Lotfollahi et al. (2022; 2024). Crucially, the data integration method must correct batch effects in the query to successfully map it onto common atlas representation.

Single-cell reference atlases serve as a valuable resource for various downstream tasks, including automating the otherwise time-consuming process of cell annotation. One novel promising application is patient classification Rood et al. (2022), which, when applied to unseen query studies, turns a reference atlas into a diagnostic tool by diagnosing new patients based on the condition of their corresponding cells. Consequently, effective data integration is fundamental to generalize to new, unseen patients. However, many challenges must be considered when envisioning and designing a reference atlas. Indeed, studies usually include a limited number of individuals Hrovatin et al. (2025), and might contain only one or few diseases, leading to an undesirable overlap between batch effects and the relevant biological information for the downstream tasks. Accordingly, research on designing multi-disease, multi-study atlas-based diagnostic tools is still ongoing.

Given the lack of an off-the-shelf solution for building single-cell atlases Hrovatin et al. (2025), the research community has devoted significant efforts into defining best practices Heumos et al. (2023); Hrovatin et al. (2025) and benchmarks Luecken et al. (2022; 2024). In particular, existing data integration methods have been evaluated in the context of reference atlas building Luecken et al. (2022), but research on effective evaluation metrics is still ongoing Wang et al. (2024). Since specific use cases for reference atlases require ad-hoc experimental settings and workflows, novel applications such as patient classification still lack the same level of guideline coverage as more standard tasks, like cell annotation via label transfer Kang et al. (2021); Lotfollahi et al. (2022).

In this work, we focus on a curated atlas of Peripheral Blood Mononuclear Cells (PBMCs) developed as part of a broader study on disease-specific inflammatory cell profiles Jiménez-Gracia et al. (2023). Specifically, we introduce a workflow to classify new patients given a reference atlas. Moreover, we present the first robust evaluation of state-of-the-art data integration and query-to-reference mapping methods from the lenses of patient classification, offering practical guidelines to both improve the experimental design for reference building and select the most appropriate integration method. Importantly, the evaluation is performed in two distinct settings with increasing levels of difficulty: (1) on the "unseen-patients" and (2) the "unseen-studies" datasets, with patients belonging to either the same or to different studies than the ones in the reference, respectively. In this work, we demonstrate that the second setting introduces crucial challenges that may affect the atlas-building design choices. Last, we discuss a centralized dataset scenario as a viable experimental setup for building complex, multi-disease, multi-study atlas-based diagnostic tools.

2 TOWARDS ATLAS-BASED DIAGNOSTICS

Given the potential of atlases as diagnostic tools Rood et al. (2022), several approaches have been proposed for patient classification (see appendix A.1). However, given the complexity of the experimental setup, there is still a lack of best-practices for developing multi-disease, multi-study atlas-based diagnostic tools. Specifically, we identify three critical challenges that need to be addressed:

Challenge 1 Single-cell studies usually focus on few diseases, and they do not always include healthy controls. What is the best method to integrate those studies by removing the dataset-level variability without eliminating the biological, disease-specific signal?

Challenge 2 A diagnostic tool is expected to correctly classify new patients, potentially with new sources of batch effects. How should we design reference atlases and patient classifiers to account for unseen sources of batch effects?

Challenge 3 The choice of the data integration and the query-to-reference mapping approach is crucial. How sensitive are the most widely used methods to the hyperparameters choice when applied to unseen query data? How should we select the best data integration method to effectively map unseen sources of batch effects?

3 REFERENCE AND QUERY DATASETS GENERATION

3.1 INFLAMMATION ATLAS AND QUERY DATASETS

The single-cell atlas of inflammation considered here was designed as a comprehensive resource to enhance the understanding of both physiological and pathological inflammation through the study

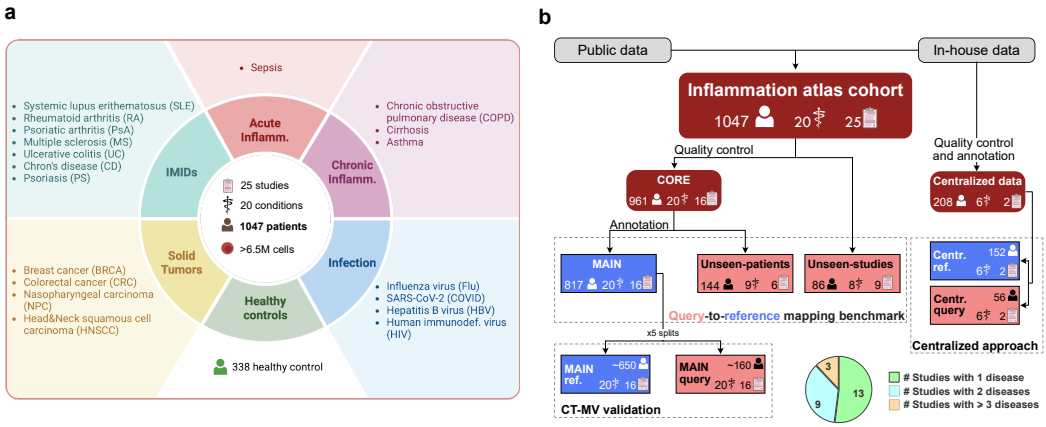

Figure 1: **Inflammation atlas and reference-query splits generation for each scenario.** (**a**) Overview of the atlas of inflammation. (**b**) Experimental design and data splitting schema for each of the scenarios: CT-MV validation, query-to-reference mapping benchmark, and centralized approach. For each split we report the number of patients, conditions, and studies.

of circulating immune cells using single-cell transcriptomic data Jiménez-Gracia et al. (2023). This project includes data from 1047 patients undergoing acute and chronic inflammatory processes, as well as healthy donors, covering 20 conditions with over 6.5 million cells. The availability of many patients with different conditions makes the atlas a perfect setting for designing and evaluating a diagnostic tool. The list of conditions and details about data sources are available in appendix A.3.

3.2 Data splitting and scenarios

Starting from the whole atlas of inflammation, we defined multiple independent splits that were used as reference and query datasets in different scenarios (fig. 1b). We first split the studies in the inflammation atlas cohort into two datasets: "CORE" and "unseen-studies", including 961 and 86 samples (patients), and 20 and 8 conditions, respectively. Next, data quality control was performed independently for each of the two datasets, which involved removing low-quality libraries, samples, cells, and underrepresented genes, as well as excluding doublets. Then, the CORE was further split into "MAIN" and "unseen-patients", with 817 and 144 samples, respectively, with 20 and 9 conditions each. First, samples were stratified based on their study, chemistry, and disease. From each of these groups, we then randomly selected 20% of samples to be part of the unseen-patients dataset, provided they amounted to at least 5 samples. As detailed in the following sections, we will define a scenario where a five-fold cross-validation on the MAIN dataset generates five reference-query datasets pairs, and one scenario where the MAIN dataset will be used as a reference and the unseen-patients and unseen-studies datasets as queries. Given the availability of two studies sequenced in the same center (with the same chemistry assay and sequencing platform) covering 6 diseases, we also defined a "Centralized" dataset. In this scenario, we defined the reference and query dataset by stratifying samples based on the pool used to generate the RNA libraries, which comprised 152 and 56 samples each. The distribution of studies, chemistries, and patients for each disease and dataset are available in fig. A.1.

3.3 Annotation

As reported in Jiménez-Gracia et al. (2023), the MAIN dataset and the Centralized reference were preprocessed and annotated independently.

MAIN The MAIN dataset (Inflammation Atlas) contains 4,435,922 cells, grouped into 15 main cell populations (*level-1*, appendix A.3.4) and 64 subpopulations (*level-2*). The final gene universe is composed of 8,253 genes, obtained by aggregating: 3,283 Highly Variable Genes (HVGs), 6,868 Differentially Expressed Genes (DEGs) between healthy and each inflammatory condition, and 1,364 manually curated immune-specific genes.

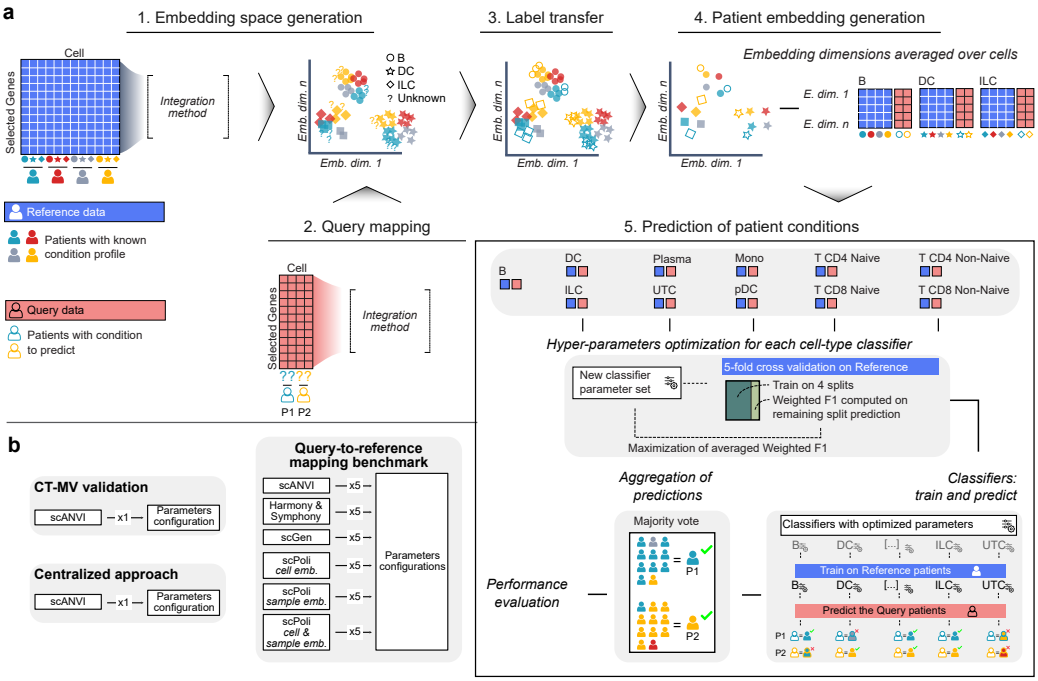

Figure 2: **CT-MV patient classifier pipeline and chosen data integration approaches.** (**a**) CT-MV patient classifier pipeline: (1) a chosen integration method is trained on the reference dataset to define the cell embeddings, (2) new patients are mapped into the reference, (3) cell-types are transferred from the annotated reference to the query dataset, (4) cell embeddings are aggregated into patient embeddings, (5) the CT-MV classifier is trained on reference patient embeddings are then applied to predict the condition of the query ones. (**b**) Data integration methods used in each of the three settings: CT-MV validation, query-to-reference mapping benchmark, and centralized approach.

Centralized reference The Centralized reference consists of 152 patients and 855,417 cells, with 15 main cell populations (*level-1*). The gene universe is the same as in the MAIN dataset.

4 METHODS

In this section, we outline the key steps of our patient classifier pipeline, which is designed to function in a multi-disease, multi-study setting (fig. 2a). First, a chosen data integration method is used to harmonize the annotated studies composing the reference, correcting their batch effects, and defining the cell embeddings in the learned representation space. The query dataset is then mapped into the reference to define the query cell embeddings. Since cells in the query dataset are not annotated, labels are transferred from the reference. After building the reference, we can define the patient classifier (referred to as CT-MV for Cell-Type Majority Vote), which combines classifiers applied independently to each cell type through a majority voting scheme. The CT-MV classifier is first trained on the reference embeddings and then tested on the left-out query ones.

4.1 DATA INTEGRATION, LABEL TRANSFER, AND QUERY-TO-REFERENCE MAPPING

Starting from a reference dataset, we applied four state-of-the-art integration methods, scGen Lotfollahi et al. (2019), scANVI Lotfollahi et al. (2022), scPoli De Donno et al. (2023), and Harmony Korsunsky et al. (2019)) (detailed in appendix A.2), to obtain the batch-corrected embedding and then map the query datasets. Note that each integration method was applied independently to each query dataset. Here, the choice of the batch-effect covariate is crucial, as the studies overlap with the target diseases (see fig. A.1). To preserve disease variability after integration, it may be beneficial to choose a broader batch covariate that is not confounded by the disease Hrovatin et al.

(2025). In our approach, unless specified otherwise, we used the broader chemistry assay covariate (appendix A.3.1) as the primary source of batch effect, instead of the finer study and sample-level covariates. Integration methods that use cell types to improve integration used *level-2*, while label transfer was performed on *level-2* and labels were then aggregated into *level-1*. When using the Centralized datasets, the only available *level-1* was used.

4.2 CT-MV Patient Classifier

4.2.1 Patient Embeddings

After obtaining the corrected embeddings from a data integration approach, cell-wise embeddings have to be merged into patient-wise embeddings p. The aggregation is performed at the cell type level (*level-1*) by computing the mean embedding across cells belonging to the same sample and cell type. In detail, let c_i^t be the embedding of the i-th cell of cell-type $t \in T$, then the mean cell type embedding is $\overline{c^t} = 1/n \sum_1^n c_i^t$. Accordingly, the cell type t embedding is defined as $c^t \coloneqq \overline{c^t}$ and the default patient embedding (referred to as ***cell***, as based on cell embeddings) is the set of all cell type embeddings $p \coloneqq \{c^t\}_{t \in T}$. In contrast to the other integration methods, scPoli also returns an embedding at the patient level p_{sc}. Consequently, we were able to test two additional alternatives to the default patient embedding (*cell*): directly using scPoli's learned patient embeddings p_{sc}, i.e., $p \coloneqq p_{sc}$ (***sample***), and re-defining cell-type embeddings as the concatenation ($\|$) of each cell-type mean embedding with the learned patient embedding, that is $c^t \coloneqq \overline{c^t} \parallel p_{sc}$ (***cell&sample***).

4.2.2 CT-MV and Hyperparameter Tuning

For each cell-type t, we train one classifier on the reference's cell type embeddings c^t. Only for scPoli in the *sample* embedding setting, we train just one classifier on the learned patient embeddings p_{sc}. We tested the following classifier types: both the linear and non-linear Support Vector Classifier (LinearSVC and SVC), and the nearest neighbors classifier (kNN). The final CT-MV classifier prediction is the majority vote among the cell-type classifiers. In the case of a tie between different conditions, we conservatively rejected the prediction of the classifiers. Moreover, if a given query patient does not have any cells annotated in a given cell type, the corresponding prediction is set as "Not Available". This label is not taken into account during majority voting, and is considered a wrong prediction when evaluating the performances of that cell type. For each classifier type, we trained different hyperparameter configurations (table A.1) and evaluated their performance using five-fold cross-validation on the reference patient embeddings. The best configurations were selected according to the weighted F1-score (WF1), independently for each cell type. Lastly, the best classifier is then retrained on the whole reference dataset. Notice that we consider the classifier type part of the CT-MV hyperparameter tuning.

4.3 Scenarios and Query-to-reference Mapping Benchmark

Having covered the data splits (section 3.2) and the CT-MV patient classifier, we can now introduce the three experimental scenarios presented in this work. In the first scenario (*CT-MV validation*), we validated our CT-MV classifier on five reference-query combinations obtained through a five-fold cross-validation on the MAIN dataset, stratifying by disease. Since selecting the best-performing data integration method and its hyperparameters is computationally expensive and, as we will later discuss, the best method might not generalize to all types of query dataset, in this scenario we employed only scANVI with pre-defined hyperparameters (3 layers, 256 hidden nodes, 30 embedding dimensions), as it is a top-performing method in existing benchmarks Luecken et al. (2022). After validating the patient classifier, in the second scenario (*Query-to-reference mapping benchmark*) we then benchmarked each of the four selected data integration approaches with multiple hyperparameter configurations (table A.2). In this second scenario, we first integrated the MAIN reference and mapped both the unseen-patients and unseen-studies query datasets, respectively. In our last scenario (*Centralied approach*), we integrated the Centralized reference and then mapped the Centralized query onto it. Here, compared to the unseen-patients and unseen-studies query datasets, the main source of batch effect is given by the library sequencing pool (appendix A.3.1), rather than the study. As this scenario serves as a proof-of-concept, we again used only our pre-defined scANVI for integration and mapping.

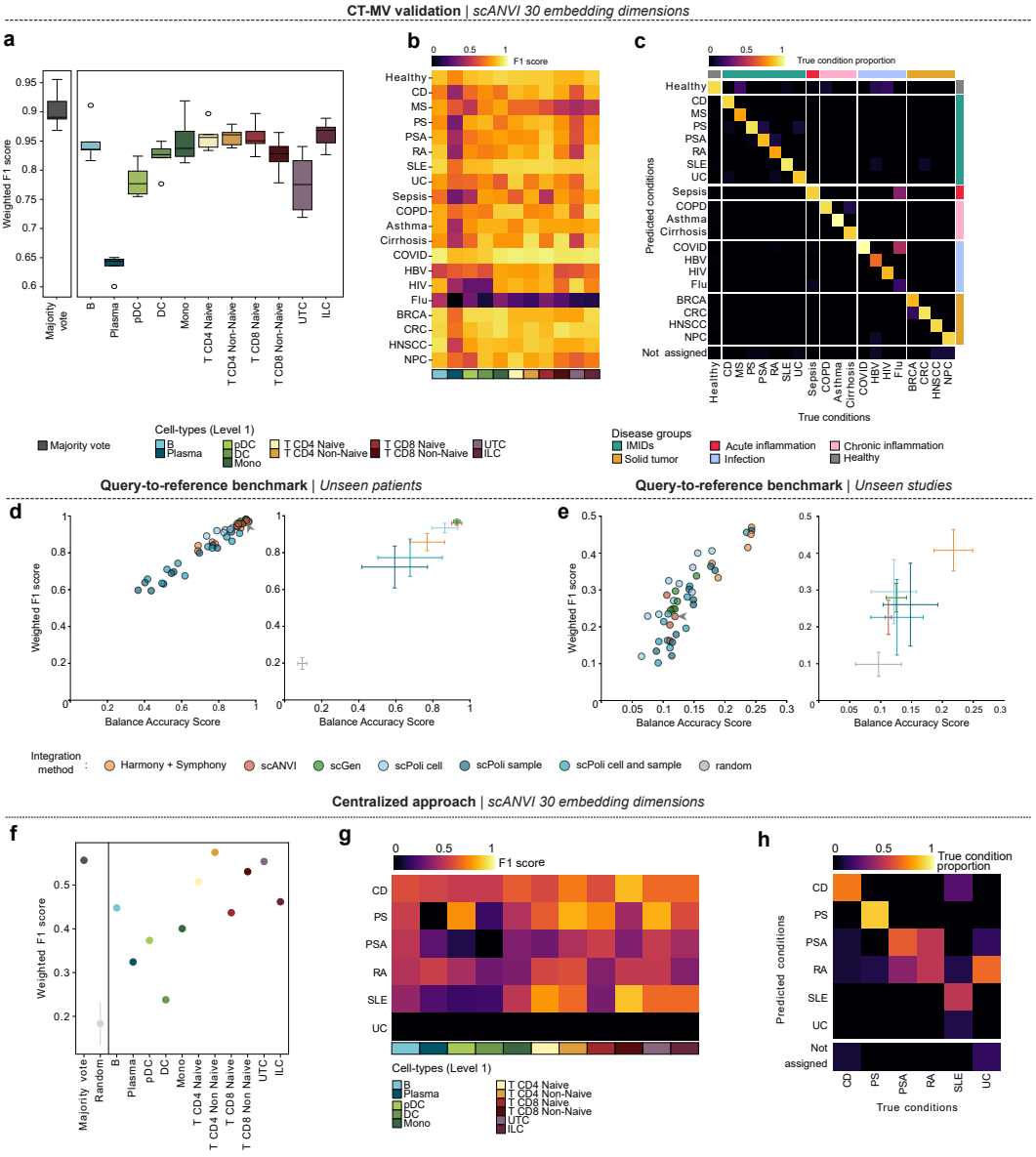

Figure 3: **CT-MV classifier performance across the three scenarios**. (**a**-**c**) CT-MV performance evaluation in the five-fold cross validation, showing (**a**) distribution of Weighted F1-scores for each left-out split; (**b**) F1-score for each combination of cell-type and disease, after aggregating all the predictions of the left-out folds; (**c**) Normalized confusion matrices displaying proportion of predictions belonging to each true condition after aggregating all the predictions of the left-out folds. (**d** and **e**) distribution of Weighted F1 and Balanced Accuracy Score for all the configurations of each data integration approach on the unseen patients and studies scenarios, respectively. (**f**-**h**) Performance evaluation in the Centralized dataset scenario, showing (**f**) Weighted F1-scores for left-out pool observation. Mean and standard deviation of weighted F1 score of 100 random condition assignments is reported; (**g**) F1-score for each combination of cell-type and disease; (**h**) Normalized confusion matrices displaying proportion of predictions belonging to each true condition.

5 RESULTS

5.1 VALIDATION OF THE CT-MV APPROACH

The CT-MV classification strategy obtained high performance in the cross-validation scenario (figs. 3a to 3c), achieving an average Weighted F1-score (WF1) of 0.90 ± 0.03 (minimum 0.87) and Balanced Accuracy Score (BAS) of 0.85 ± 0.07 (minimum 0.79) across five independent runs. Flu was the only disease that failed to be classified effectively (Recall: 0.18). Training a classifier for each cell type separately allowed us to evaluate their relevance in distinguishing inflammatory diseases, particularly for those with lower overall performance scores. Certain diseases, such as COVID, COPD, and Asthma, were particularly well classified by lymphoid and myeloid cells. HIV was best classified by naive lymphoid cells (i.e., Naive CD4 and CD8 T cells, and B cells) with an F1 of 0.83, consistent with the virus's tropism for infecting mainly CD4 T cells. In contrast, dendritic cell types (i.e., DC and pDC) did not allow correct disease assignment (F1 of 0.29). Overall, Plasma cells and UTC showed the lowest BAS (0.53 and 0.67) and WF1 (0.64 and 0.78), underscoring the strength of our majority voting approach encompassing multiple cell types embeddings.

5.2 CURRENT QUERY-TO-REFERENCE MAPPING APPROACHES FAIL ON UNSEEN STUDIES

In line with our previous results with scANVI, all integration methods demonstrated high performance in the unseen-patients dataset (fig. 3d). Specifically, scANVI, scPoli, and scGen achieved similarly high BAS (> 0.94) and WF1 (> 0.97), followed by Harmony (BAS: 0.92 and WF1: 0.94) in their best-performing configurations. However, all approaches experienced a decline in predictive power on the unseen-studies dataset (fig. 3e), with Harmony performing best (BAS: 0.24 and WF1: 0.47). Notably, scPoli achieved the best results on the unseen-studies dataset when the learned patient-wise embedding (*sample*) was used for classification (BAS: 0.24 and WF1: 0.46), but the same model underperformed on the unseen-patients dataset (BAS: 0.37 and WF1: 0.60). This suggests that not all the patient variability is encoded in scPoli-learned embeddings (p_{sc}). Additionally, *sample* embeddings may be less prone to overfitting due to their smaller dimensionality compared to the *cell* and *cell&sample* patient embeddings. By evaluating different hyperparameters for each approach, we estimated each method's sensitivity to the chosen configuration. Harmony emerged as the best-performing and least sensitive to hyperparameter choices on the unseen-studies dataset, with a BAS of 0.22 ± 0.03 and a WF1 of 0.41 ± 0.06 across the considered hyperparameters. In conclusion, all tools showed limited generalization in the most challenging task of classifying patients from the unseen-studies dataset, even after hyperparameter adjustments. In settings where hyperparameter tuning and validation are not feasible due to the lack of enough data spanning all target conditions and covariates (**challenges 2-3**), tools like Harmony and Symphony might be preferable to more complex VAEs. Indeed, while linear approaches have less representational power than VAEs, they are less prone to overfitting and more robust to hyperparameter choices.

5.3 A CENTRALIZED PATIENT CLASSIFIER

The notable performance decline between the unseen-patients and unseen-studies query datasets can be attributed to the more pronounced batch effects in the latter, such as differences in chemistry assays, sequencing platform, and research centers. Again, the fact that studies contain only few diseases prevents for directly correcting by the study covariate (**challenge 1**), leading to our design choice of correcting by the chemistry assay. These variations hindered the integration methods trained on the MAIN reference from fully generalizing to the query (**challenge 2**). Moreover, the significantly different performance rankings obtained from the two query datasets highlighted the challenge of selecting an integration method that generalizes across multiple unseen query datasets (**challenge 3**). In the final Centralized approach scenario, we investigated a setting where query data was generated at the same center, using a single chemistry assay and sequencing platform, but sequenced in a different library pool (figs. 3f to 3h). Although not directly comparable to the performance on the unseen-studies dataset, the WF1 and BAS increased to 0.56 and 0.53, respectively, demonstrating a notable enhancement in generalization performance. Our preliminary results indicate that diagnostic tools based on centralized, single-chemistry studies, including both the reference studies and future ones to be mapped onto it, could offer a superior experimental setup. In this scenario, the primary sources of batch effects would be shared between the reference and query datasets, promoting a better classifier generalization (overcoming **challenge 2**). Similarly, data in-

tegration could be selected through a cross-validation on the reference, given enough coverage of diseases across pools (overcoming **challenge 3**).

6 DISCUSSION

Patient classification is a promising application of large-scale reference atlases, which involves predicting a new patient's condition from its high-resolution single-cell transcriptomic profile signal. A powerful atlas-based diagnostic tool can only be introduced if diseases and their batch effects are well-represented in the reference atlas. However, single-cell studies usually contain few patients and diseases, requiring careful design of their integration, as study batch effects overlap with the disease's biological signal (**challenge 1**, also discussed in Hrovatin et al. (2025); Willem et al. (2025)). Moreover, new patients might belong to studies with unseen technological variations, such as a different chemistry assay or sequencing platform than those in the reference, complicating their mapping onto the reference (**challenge 2**). All of these aspects make the choice of integration and mapping methods crucial. However, we currently lack metrics and guidelines for selecting them in the context of query-to-reference mapping and diagnostics (**challenge 3**).

In this work, we conducted an analysis using data from a single-cell atlas of inflammation Jiménez-Gracia et al. (2023), which, due to its extensive range of inflammatory conditions, serves as an ideal setting for an atlas-based diagnostic tool. Accordingly, we defined the CT-MV patient classifier approach and confirmed its good performance on five reference-query combinations of the MAIN dataset. Considering the large availability of studies and diseases in the atlas, we evaluated two scenarios using the current state-of-the-art integration and reference mapping approaches with multiple hyperparameter configurations, named the unseen-patients and unseen-studies query datasets. While CT-MV performs well across different integration methods for left-out patients belonging to the same studies included in the MAIN reference, all methods poorly generalize to novel unobserved studies (**challenge 2**). We attribute the performance drop to the fact that the unseen-studies dataset contains batch effects that are under-represented in the reference, preventing the generalization of the CT-MV classifier. Interestingly, while VAE-based approaches perform well in the easier scenarios, the linear approach Harmony leads to better performance on the unseen-studies dataset, while also being less susceptible to the hyperparameter configuration. The performance difference between the two scenarios highlights the need for large-scale benchmarks and metrics to rank integration and mapping methods in the context of query-to-reference mapping (**challenge 3**).

Last, we defined a viable solution for the three challenges where reference and query are sequenced in the same center, with the main source of batch effect being the library sequencing pool (appendix A.3.1). While a large-scale atlas with diseases represented across multiple studies would undoubtedly improve diagnostic tool performance, we argue that a centralized approach could achieve similar positive outcomes with fewer patients. Indeed, although not directly comparable, CT-MV obtained better performance on the Centralized query dataset compared to unseen-studies one, suggesting a lower impact of unseen batch effects than in the first case (overcoming **challenge 2**). In a Centralized scenario, samples from different diseases could be sequenced in the same pool to improve their distribution and break the batch-disease relation (overcoming **challenge 1**). Moreover, the better representation of batch effects in the reference could ease the selection of the best integration approach. While cross-validation of the MAIN reference may not adequately represent integration performance on unseen studies due to insufficient disease-batch variability, this approach would be more viable in our Centralized scenario (overcoming **challenge 3**).

We believe that additional efforts should be devoted to standardizing and benchmarking the workflow for atlas-based diagnostics. While this work provides an initial benchmark and experimental design guidelines, additional research is needed to address various contexts, including different diseases, batch effect sources, experimental setups (such as batch covariate selection and baselines), and evaluation protocols (incorporating custom metrics for diagnostic tools). Interestingly, the atlas-based approach discussed here represents an alternative to the recent surge of single-cell foundation models Cui et al. (2024), which leverage the availability of large-scale data to learn both technical and biological signals. Future research should consider the advantages and disadvantages of both approaches within the context of automated diagnostics.

MEANINGFULNESS STATEMENT

Meaningful representations can be defined through the principle of compression. Specifically, a meaningful representation of an input can be viewed as its lossy compression. During the learning process, the target task is crucial in identifying which information is relevant and should be preserved, and which can be discarded. The integration of single-cell datasets aims to compress them into a representation that retains only the biological signal, eliminating the technical and stochastic noise. A key challenge addressed in this work is developing a generalizable compression method that retains the disease signal in single-cell transcriptomic data even on new unseen query datasets.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Human blood processed in-house for this project was pre-selected and included within other ongoing studies. All the studies included were conducted in accordance with ethical guidelines and all patients provided written informed consent. Ethical committees and research project approved each study included in this manuscript.

FUNDING

This project has received funding from the European Union's H2020 research and innovation program under grant agreement No. 848028 (DoCTIS; Decision On Optimal Combinatorial Therapies In Imids Using Systems Approaches). F.C. is funded by the Swiss National Science Foundation (SNSF) grant No CRSII5_205884/1. D.M. is supported by the Juan de la Cierva Fellowship (JDC2022-049637-I) from the Spanish Ministry of Science and Innovation and the European Union "NextGenerationEU"/PRTR. L.J.-G. has held an FPU PhD fellowship (FPU19/04886) from the Spanish Ministry of Universities. M.B is funded by the Graph Neural Networks for Explainable Artificial Intelligence ERA-NET + EJP (20CH21_195579) grant. Part of the computational analyses were supported by the Google Cloud Research Credits program with the award GCP19980904.

COMPETING INTERESTS

H.H. is co-founder and shareholder of Omniscope, scientific advisory board member of Nanostring and MiRXES, consultant to Moderna and Singularity and has received honorarium from Genentech. J.C.N. is scientific consultant to Omniscope.

REFERENCES

David J. Ahern, Zhichao Ai, Mark Ainsworth, Chris Allan, Alice Allcock, Brian Angus, M. Azim Ansari, Carolina V. Arancibia-Cárcamo, Dominik Aschenbrenner, Moustafa Attar, J. Kenneth Baillie, Eleanor Barnes, Rachael Bashford-Rogers, Archana Bashyal, Sally Beer, Georgina Berridge, Amy Beveridge, Sagida Bibi, Tihana Bicanic, Luke Blackwell, Paul Bowness, Andrew Brent, Andrew Brown, John Broxholme, David Buck, Katie L. Burnham, Helen Byrne, Susana Camara, Ivan Candido Ferreira, Philip Charles, Wentao Chen, Yi-Ling Chen, Amanda Chong, Elizabeth A. Clutterbuck, Mark Coles, Christopher P. Conlon, Richard Cornall, Adam P. Cribbs, Fabiola Curion, Emma E. Davenport, Neil Davidson, Simon Davis, Calliope A. Dendrou, Julie Dequaire, Lea Dib, James Docker, Christina Dold, Tao Dong, Damien Downes, Hal Drakesmith, Susanna J. Dunachie, David A. Duncan, Chris Eijsbouts, Robert Esnouf, Alexis Espinosa, Rachel Etherington, Benjamin Fairfax, Rory Fairhead, Hai Fang, Shayan Fassih, Sally Felle, Maria Fernandez Mendoza, Ricardo Ferreira, Roman Fischer, Thomas Foord, Aden Forrow, John Frater, Anastasia Fries, Veronica Gallardo Sanchez, Lucy C. Garner, Clementine Geeves, Dominique Georgiou, Leila Godfrey, Tanya Golubchik, Maria Gomez Vazquez, Angie Green, Hong Harper, Heather A. Harrington, Raphael Heilig, Svenja Hester, Jennifer Hill, Charles Hinds, Clare Hird, Ling-Pei Ho, Renee Hoekzema, Benjamin Hollis, Jim Hughes, Paula Hutton, Matthew A. Jackson-Wood, Ashwin Jainarayanan, Anna James-Bott, Kathrin Jansen, Katie Jeffery, Elizabeth Jones, Luke Jostins, Georgina Kerr, David Kim, Paul Klenerman, Julian C. Knight, Vinod Kumar, Piyush Kumar Sharma, Prathiba Kurupati, Andrew Kwok, Angela Lee, Aline Linder, Teresa Lockett, Lorne Lonie, Maria Lopopolo, Martyna Lukoseviciute, Jian Luo, Spyridoula Marinou, Brian Marsden, Jose Martinez, Philippa C. Matthews, Michalina Mazurczyk, Simon McGowan, Stuart McKechnie, Adam Mead, Alexander J. Mentzer, Yuxin Mi, Claudia Monaco, Ruddy Mon-

tadon, Giorgio Napolitani, Isar Nassiri, Alex Novak, Darragh P. O'Brien, Daniel O'Connor, Denise O'Donnell, Graham Ogg, Lauren Overend, Inhye Park, Ian Pavord, Yanchun Peng, Frank Penkava, Mariana Pereira Pinho, Elena Perez, Andrew J. Pollard, Fiona Powrie, Bethan Psaila, T. Phuong Quan, Emmanouela Repapi, Santiago Revale, Laura Silva-Reyes, Jean-Baptiste Richard, Charlotte Rich-Griffin, Thomas Ritter, Christine S. Rollier, Matthew Rowland, Fabian Ruehle, Mariolina Salio, Stephen Nicholas Sansom, Raphael Sanches Peres, Alberto Santos Delgado, Tatjana Sauka-Spengler, Ron Schwessinger, Giuseppe Scozzafava, Gavin Screaton, Anna Seigal, Malcolm G. Semple, Martin Sergeant, Christina Simoglou Karali, David Sims, Donal Skelly, Hubert Slawinski, Alberto Sobrinodiaz, Nikolaos Sousos, Lizzie Stafford, Lisa Stockdale, Marie Strickland, Otto Sumray, Bo Sun, Chelsea Taylor, Stephen Taylor, Adan Taylor, Supat Thongjuea, Hannah Thraves, John A. Todd, Adriana Tomic, Orion Tong, Amy Trebes, Dominik Trzupek, Felicia Anna Tucci, Lance Turtle, Irina Udalova, Holm Uhlig, Erinke van Grinsven, Iolanda Vendrell, Marije Verheul, Alexandru Voda, Guanlin Wang, Lihui Wang, Dapeng Wang, Peter Watkinson, Robert Watson, Michael Weinberger, Justin Whalley, Lorna Witty, Katherine Wray, Luzheng Xue, Hing Yuen Yeung, Zixi Yin, Rebecca K. Young, Jonathan Youngs, Ping Zhang, and Yasemin-Xiomara Zurke. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell*, 185(5):916–938.e58, March 2022. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2022.01.012.

Philipp Angerer, Lukas Simon, Sophie Tritschler, F. Alexander Wolf, David Fischer, and Fabian J. Theis. Single cells make big data: New challenges and opportunities in transcriptomics. *Current Opinion in Systems Biology*, 4:85–91, August 2017. ISSN 2452-3100. doi: 10.1016/j.coisb.2017.07.004.

Pierre Boyeau, Justin Hong, Adam Gayoso, Martin Kim, José L. McFaline-Figueroa, Michael I. Jordan, Elham Azizi, Can Ergen, and Nir Yosef. Deep generative modeling of sample-level heterogeneity in single-cell genomics, May 2024.

Anthony R. Cillo, Cornelius H. L. Kürten, Tracy Tabib, Zengbiao Qi, Sayali Onkar, Ting Wang, Angen Liu, Umamaheswar Duvvuri, Seungwon Kim, Ryan J. Soose, Steffi Oesterreich, Wei Chen, Robert Lafyatis, Tullia C. Bruno, Robert L. Ferris, and Dario A. A. Vignali. Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. *Immunity*, 52(1):183–199.e9, January 2020. ISSN 1074-7613. doi: 10.1016/j.immuni.2019.11.014.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0.

Carlo De Donno, Soroor Hediyeh-Zadeh, Amir Ali Moinfar, Marco Wagenstetter, Luke Zappia, Mohammad Lotfollahi, and Fabian J. Theis. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nature Methods*, pp. 1–10, October 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-02035-2.

Wei Dong, Charikar Moses, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pp. 577–586, New York, NY, USA, March 2011. Association for Computing Machinery. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963487.

Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, Daniel C. Strobl, Juan Henao, Fabiola Curion, Herbert B. Schiller, and Fabian J. Theis. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, August 2023. ISSN 1471-0064. doi: 10.1038/s41576-023-00586-w.

Karin Hrovatin, Lisa Sikkema, Vladimir A. Shitov, Graham Heimberg, Maiia Shulman, Amanda J. Oliver, Michaela F. Mueller, Ignacio L. Ibarra, Hanchen Wang, Ciro Ramírez-Suástegui, Peng He, Anna C. Schaar, Sarah A. Teichmann, Fabian J. Theis, and Malte D. Luecken. Considerations for building and using integrated single-cell atlases. *Nature Methods*, 22(1):41–57, January 2025. ISSN 1548-7105. doi: 10.1038/s41592-024-02532-y.

Yale Jiang, Brian R. Rosborough, Jie Chen, Sudipta Das, Georgios D. Kitsios, Bryan J. McVerry, Rama K. Mallampalli, Janet S. Lee, Anuradha Ray, Wei Chen, and Prabir Ray. Single cell RNA

sequencing identifies an early monocyte gene signature in acute respiratory distress syndrome. *JCI Insight*, 5(13), July 2020. ISSN 0021-9738. doi: 10.1172/jci.insight.135678.

Laura Jiménez-Gracia, Davide Maspero, Sergio Aguilar-Fernández, Francesco Craighero, Sara Ruiz, Domenica Marchese, Ginevra Caratù, Marc Elosua-Bayes, Mohamed Abdalfatah, Angela Sanzo-Machuca, Ana M. Corraliza, Ramon Massoni-Badosa, Hoang A. Tran, Rachelly Normand, Jacquelyn Nestor, Yourae Hong, Tessa Kole, Petra van der Velde, Frederique Alleblas, Flaminia Pedretti, Adrià Aterido, Martin Banchero, German Soriano, Eva Román, Maarten van den Berge, Azucena Salas, Jose Manuel Carrascosa, Antonio Fernández Nebro, Eugeni Domènech, Juan Cañete, Jesús Tornero, Javier Pérez-Gisbert, Ernest Choy, Giampiero Girolomoni, Britta Siegmund, Antonio Julià, Violeta Serra, Roberto Elosua, Sabine Tejpar, Silvia Vidal, Martijn C. Nawijn, Sara Marsal, Pierre Vandergheynst, Alexandra-Chloé Villani, Juan C. Nieto, and Holger Heyn. Interpretable Inflammation Landscape of Circulating Immune cells, November 2023.

Joyce B. Kang, Aparna Nathan, Kathryn Weinand, Fan Zhang, Nghia Millard, Laurie Rumker, D. Branch Moody, Ilya Korsunsky, and Soumya Raychaudhuri. Efficient and precise single-cell reference atlas mapping with Symphony. *Nature Communications*, 12(1):5890, October 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25957-x.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, December 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0619-0.

David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, Kieran R. Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccio, Giacomo Corleone, Bas E. Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharina Jahn, Thamar Jessurun Lobo, Emma M. Keizer, Indu Khatri, Szymon M. Kielbasa, Jan O. Korbel, Alexey M. Kozlov, Tzu-Hao Kuo, Boudewijn P.F. Lelieveldt, Ion I. Mandoiu, John C. Marioni, Tobias Marschall, Felix Mölder, Amir Niknejad, Alicja Rączkowska, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J. Theis, Huan Yang, Alex Zelikovsky, Alice C. McHardy, Benjamin J. Raphael, Sohrab P. Shah, and Alexander Schönhuth. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):31, February 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-1926-6.

Anastasia Litinetskaya, Maiia Shulman, Soroor Hediyeh-zadeh, Amir Ali Moinfar, Fabiola Curion, Artur Szałata, Alireza Omidi, Mohammad Lotfollahi, and Fabian J. Theis. Multimodal weakly supervised learning to identify disease-specific changes in single-cell atlases, July 2024.

Yang Liu, Shuai He, Xi-Liang Wang, Wan Peng, Qiu-Yan Chen, Dong-Mei Chi, Jie-Rong Chen, Bo-Wei Han, Guo-Wang Lin, Yi-Qi Li, Qian-Yu Wang, Rou-Jun Peng, Pan-Pan Wei, Xiang Guo, Bo Li, Xiaojun Xia, Hai-Qiang Mai, Xue-Da Hu, Zemin Zhang, Yi-Xin Zeng, and Jin-Xin Bei. Tumour heterogeneity and intercellular networks of nasopharyngeal carcinoma at single cell resolution. *Nature Communications*, 12(1):741, February 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21043-4.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, August 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0494-8.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, Sergei Rybakov,

Alexander V. Misharin, and Fabian J. Theis. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1):121–130, January 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01001-7.

Mohammad Lotfollahi, Yuhan Hao, Fabian J. Theis, and Rahul Satija. The future of rapid and automated single-cell data analysis using reference mapping. *Cell*, 187(10):2343–2358, May 2024. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2024.03.009.

Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, January 2022. ISSN 1548-7105. doi: 10.1038/s41592-021-01336-8.

Malte D Luecken, Scott Gigante, Daniel B Burkhardt, Robrecht Cannoodt, Daniel C Strobl, Nikolay S Markov, Luke Zappia, Giovanni Palla, Wesley Lewis, Daniel Dimitrov, et al. Defining and benchmarking open problems in single-cell analysis. *Research Square*, 2024.

Jerome C. Martin, Christie Chang, Gilles Boschetti, Ryan Ungaro, Mamta Giri, John A. Grout, Kyle Gettler, Ling-shiang Chuang, Shikha Nayar, Alexander J. Greenstein, Marla Dubinsky, Laura Walker, Andrew Leader, Jay S. Fine, Charles E. Whitehurst, M. Lamine Mbow, Subra Kugathasan, Lee A. Denson, Jeffrey S. Hyams, Joshua R. Friedman, Prerak T. Desai, Huaibin M. Ko, Ilaria Laface, Guray Akturk, Eric E. Schadt, Helene Salmon, Sacha Gnjatic, Adeeb H. Rahman, Miriam Merad, Judy H. Cho, and Ephraim Kenigsberg. Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell*, 178(6):1493–1508.e20, September 2019. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2019.08.008.

Jordi Martorell-Marugán, Raúl López-Domínguez, Juan Antonio Villatoro-García, Daniel Toro-Domínguez, Marco Chierici, Giuseppe Jurman, and Pedro Carmona-Sáez. Explainable deep neural networks for predicting sample phenotypes from single-cell transcriptomics. *Briefings in Bioinformatics*, 26(1):bbae673, January 2025. ISSN 1477-4054. doi: 10.1093/bib/bbae673.

Pragnesh Mistry, Shuichiro Nakabo, Liam O'Neil, Rishi R. Goel, Kan Jiang, Carmelo Carmona-Rivera, Sarthak Gupta, Diana W. Chan, Philip M. Carlucci, Xinghao Wang, Faiza Naz, Zerai Manna, Amit Dey, Nehal N. Mehta, Sarfaraz Hasni, Stefania Dell'Orso, Gustavo Gutierrez-Cruz, Hong-Wei Sun, and Mariana J. Kaplan. Transcriptomic, epigenetic, and functional analyses implicate neutrophil diversity in the pathogenesis of systemic lupus erythematosus. *Proceedings of the National Academy of Sciences*, 116(50):25222–25228, December 2019. doi: 10.1073/pnas.1908576116.

Mukta G. Palshikar, Rohith Palli, Alicia Tyrell, Sanjay Maggirwar, Giovanni Schifitto, Meera V. Singh, and Juilee Thakar. Executable models of immune signaling pathways in HIV-associated atherosclerosis. *npj Systems Biology and Applications*, 8(1):1–15, September 2022. ISSN 2056-7189. doi: 10.1038/s41540-022-00246-5.

Richard K. Perez, M. Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C. Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, Andrew Lu, Mike Thompson, Nadav Rappoport, Andrew Dahl, Cristina M. Lanata, Mehrdad Matloubian, Lenka Maliskova, Serena S. Kwek, Tony Li, Michal Slyper, Julia Waldman, Danielle Dionne, Orit Rozenblatt-Rosen, Lawrence Fong, Maria Dall'Era, Brunilda Balliu, Aviv Regev, Jinoos Yazdany, Lindsey A. Criswell, Noah Zaitlen, and Chun Jimmie Ye. Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589):eabf1970, April 2022. doi: 10.1126/science.abf1970.

P. Ramachandran, R. Dobie, J. R. Wilson-Kanamori, E. F. Dora, B. E. P. Henderson, N. T. Luu, J. R. Portman, K. P. Matchett, M. Brice, J. A. Marwick, R. S. Taylor, M. Efremova, R. Vento-Tormo, N. O. Carragher, T. J. Kendall, J. A. Fallowfield, E. M. Harrison, D. J. Mole, S. J. Wigmore, P. N. Newsome, C. J. Weston, J. P. Iredale, F. Tacke, J. W. Pollard, C. P. Ponting, J. C. Marioni, S. A. Teichmann, and N. C. Henderson. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature*, 575(7783):512–518, November 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1631-3.

Xianwen Ren, Wen Wen, Xiaoying Fan, Wenhong Hou, Bin Su, Pengfei Cai, Jiesheng Li, Yang Liu, Fei Tang, Fan Zhang, Yu Yang, Jiangping He, Wenji Ma, Jingjing He, Pingping Wang, Qiqi Cao, Fangjin Chen, Yuqing Chen, Xuelian Cheng, Guohong Deng, Xilong Deng, Wenyu Ding, Yingmei Feng, Rui Gan, Chuang Guo, Weiqiang Guo, Shuai He, Chen Jiang, Juanran Liang, Yi-min Li, Jun Lin, Yun Ling, Haofei Liu, Jianwei Liu, Nianping Liu, Shu-Qiang Liu, Meng Luo, Qiang Ma, Qibing Song, Wujianan Sun, GaoXiang Wang, Feng Wang, Ying Wang, Xiaofeng Wen, Qian Wu, Gang Xu, Xiaowei Xie, Xinxin Xiong, Xudong Xing, Hao Xu, Chonghai Yin, Dongdong Yu, Kezhuo Yu, Jin Yuan, Biao Zhang, Peipei Zhang, Tong Zhang, Jincun Zhao, Peidong Zhao, Jianfeng Zhou, Wei Zhou, Sujuan Zhong, Xiaosong Zhong, Shuye Zhang, Lin Zhu, Ping Zhu, Bin Zou, Jiahua Zou, Zengtao Zuo, Fan Bai, Xi Huang, Penghui Zhou, Qinghua Jiang, Zhiwei Huang, Jin-Xin Bei, Lai Wei, Xiu-Wu Bian, Xindong Liu, Tao Cheng, Xiangpan Li, Pingsen Zhao, Fu-Sheng Wang, Hongyang Wang, Bing Su, Zheng Zhang, Kun Qu, Xiaoqun Wang, Jiekai Chen, Ronghua Jin, and Zemin Zhang. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell*, 184(7):1895–1913.e19, April 2021. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2021.01.053.

Miguel Reyes, Michael R. Filbin, Roby P. Bhattacharyya, Kianna Billman, Thomas Eisenhaure, Deborah T. Hung, Bruce D. Levy, Rebecca M. Baron, Paul C. Blainey, Marcia B. Goldberg, and Nir Hacohen. An immune-cell signature of bacterial sepsis. *Nature Medicine*, 26(3):333–340, March 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0752-4.

Jennifer E. Rood, Aidan Maartens, Anna Hupalowska, Sarah A. Teichmann, and Aviv Regev. Impact of the Human Cell Atlas on medicine. *Nature Medicine*, 28(12):2486–2496, December 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-02104-7.

Adam K. Savage, Miriam V. Gutschow, Tony Chiang, Kathy Henderson, Richard Green, Monica Chaudhari, Elliott Swanson, Alexander T. Heubeck, Nina Kondza, Kelli C. Burley, Palak C. Genge, Cara Lord, Tanja Smith, Zachary Thomson, Aldan Beaubien, Ed Johnson, Jeff Goldy, Hamid Bolouri, Jane H. Buckner, Paul Meijer, Ernest M. Coffey, Peter J. Skene, Troy R. Torgerson, Xiao-jun Li, and Thomas F. Bumol. Multimodal analysis for human ex vivo studies shows extensive molecular changes from delays in blood processing. *iScience*, 24(5), May 2021. ISSN 2589-0042. doi: 10.1016/j.isci.2021.102404.

David Schafflick, Chenling A. Xu, Maike Hartlehnert, Michael Cole, Andreas Schulte-Mecklenbeck, Tobias Lautwein, Jolien Wolbert, Michael Heming, Sven G. Meuth, Tanja Kuhlmann, Catharina C. Gross, Heinz Wiendl, Nir Yosef, and Gerd Meyer zu Horste. Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nature Communications*, 11(1):247, January 2020. ISSN 2041-1723. doi: 10.1038/s41467-019-14118-w.

Lisa Sikkema, Ciro Ramírez-Suástegui, Daniel C. Strobl, Tessa E. Gillett, Luke Zappia, Elo Madissoon, Nikolay S. Markov, Laure-Emmanuelle Zaragosi, Yuge Ji, Meshal Ansari, Marie-Jeanne Arguel, Leonie Apperloo, Martin Banchero, Christophe Bécavin, Marijn Berg, Evgeny Chichelnitskiy, Mei-i Chung, Antoine Collin, Aurore C. A. Gay, Janine Gote-Schniering, Baharak Hooshiar Kashani, Kemal Inecik, Manu Jain, Theodore S. Kapellos, Tessa M. Kole, Sylvie Leroy, Christoph H. Mayr, Amanda J. Oliver, Michael von Papen, Lance Peter, Chase J. Taylor, Thomas Walzthoeni, Chuan Xu, Linh T. Bui, Carlo De Donno, Leander Dony, Alen Faiz, Minzhe Guo, Austin J. Gutierrez, Lukas Heumos, Ni Huang, Ignacio L. Ibarra, Nathan D. Jackson, Preetish Kadur Lakshminarasimha Murthy, Mohammad Lotfollahi, Tracy Tabib, Carlos Talavera-López, Kyle J. Travaglini, Anna Wilbrey-Clark, Kaylee B. Worlock, Masahiro Yoshida, Maarten van den Berge, Yohan Bossé, Tushar J. Desai, Oliver Eickelberg, Naftali Kaminski, Mark A. Krasnow, Robert Lafyatis, Marko Z. Nikolic, Joseph E. Powell, Jayaraj Rajagopal, Mauricio Rojas, Orit Rozenblatt-Rosen, Max A. Seibold, Dean Sheppard, Douglas P. Shepherd, Don D. Sin, Wim Timens, Alexander M. Tsankov, Jeffrey Whitsett, Yan Xu, Nicholas E. Banovich, Pascal Barbry, Thu Elizabeth Duong, Christine S. Falk, Kerstin B. Meyer, Jonathan A. Kropski, Dana Pe'er, Herbert B. Schiller, Purushothama Rao Tata, Joachim L. Schultze, Sara A. Teichmann, Alexander V. Misharin, Martijn C. Nawijn, Malte D. Luecken, and Fabian J. Theis. An integrated cell atlas of the lung in health and disease. *Nature Medicine*, 29(6):1563–1577, June 2023. ISSN 1546-170X. doi: 10.1038/s41591-023-02327-2.

Marina Terekhova, Amanda Swain, Pavla Bohacova, Ekaterina Aladyeva, Laura Arthur, Anwesha Laha, Denis A. Mogilenko, Samantha Burdess, Vladimir Sukhov, Denis Kleverov, Barbora

Echalar, Petr Tsurinov, Roman Chernyatchik, Kamila Husarcikova, and Maxim N. Artyomov. Single-cell atlas of healthy human blood unveils age-related loss of NKG2C+GZMB-CD8+ memory T cells and accumulation of type 2 memory T cells. *Immunity*, 56(12):2836–2854.e9, December 2023. ISSN 1074-7613. doi: 10.1016/j.immuni.2023.10.013.

Hanchen Wang, Jure Leskovec, and Aviv Regev. Metric Mirages in Cell Embeddings, April 2024.

Shaobo Wang, Qiong Zhang, Hui Hui, Kriti Agrawal, Maile Ann Young Karris, and Tariq M. Rana. An atlas of immune cell exhaustion in HIV-infected individuals revealed by single-cell transcriptomics. *Emerging Microbes & Infections*, 9(1):2333–2347, January 2020. ISSN null. doi: 10.1080/22221751.2020.1826361.

Theresa Willem, Vladimir A. Shitov, Malte D. Luecken, Niki Kilbertus, Stefan Bauer, Marie Piraud, Alena Buyx, and Fabian J. Theis. Biases in machine-learning models of human single-cell data. pp. 1–9, 2025. ISSN 1476-4679. doi: 10.1038/s41556-025-01619-8. URL `https://www.nature.com/articles/s41556-025-01619-8`.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, February 2018. ISSN 1474-760X. doi: 10.1186/s13059-017-1382-0.

Guangzhi Xiong, Stefan Bekiranov, and Aidong Zhang. ProtoCell4P: An explainable prototype-based neural network for patient classification using single-cell RNA-seq. *Bioinformatics*, 39(8): btad493, August 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad493.

Chao Zhang, Jiesheng Li, Yongqian Cheng, Fanping Meng, Jin-Wen Song, Xing Fan, Hongtao Fan, Jing Li, Yu-Long Fu, Ming-Ju Zhou, Wei Hu, Si-Yu Wang, Yuan-Jie Fu, Ji-Yuan Zhang, Ruo-Nan Xu, Ming Shi, Xueda Hu, Zemin Zhang, Xianwen Ren, and Fu-Sheng Wang. Single-cell RNA sequencing reveals intrahepatic and peripheral immune characteristics related to disease phases in HBV-infected patients. *Gut*, 72(1):153–167, January 2023. ISSN 0017-5749, 1468-3288. doi: 10.1136/gutjnl-2021-325915.

A Appendix

A.1 Patient Classifiers

To date, multiple approaches for patient classification have been proposed Boyeau et al. (2024); De Donno et al. (2023); Litinetskaya et al. (2024); Martorell-Marugán et al. (2025); Xiong et al. (2023). Notably, few of them are designed to handle both data integration and classification Boyeau et al. (2024), while also enabling query-to-reference mapping on either transcriptomic De Donno et al. (2023) or multi-omic data Litinetskaya et al. (2024).

A.2 Data Integration Methods

scGEN. scGen is defined by two main components: a Variational AutoEncoder (VAE) and a latent space arithmetic method. After training the VAE on the reference dataset, the latent space arithmetic corrects for the batch effect induced by the chemistry assay used. Within each cell type, scGen first selects the mean μ_{max} of the most populated batch, and then corrects each batch with mean μ_0 by adding $\delta = \mu_{max} - \mu_0$ to each cell's embedding. The final corrected count matrix corresponds to the generated count matrix from the arithmetic-corrected embeddings. Following scGen's tutorials, we will refer to corrected embeddings to those obtained by using the corrected expression matrix as input. To perform scGen's batch correction on the query dataset, we first need to transfer the annotation from the reference. Following a similar approach employed in the Human Lung Cell Atlas Sikkema et al. (2023) and introduced in Lotfollahi et al. (2022), we transfer the labels using (approximate) nearest neighbor Dong et al. (2011). Query cells are annotated with the most probable cell type on the 10 nearest neighbors in the annotated reference.

scANVI. scANVI is an extension of the VAE-based scVI Lopez et al. (2018) model that also exploits cell-type information to perform self-supervised learning. We first trained scVI and scANVI sequentially on the reference dataset. Then, we fine-tuned[1] scANVI on the query dataset to transfer the labels and extract the query cell embeddings.

scPoli. In contrast to other VAE-based integration methods such as scANVI, scPoli encodes both the chemistry covariate and the sample as a learnable conditional embedding, and characterizes each cell-type as a prototype in the latent embedding to facilitate label transfer. First, scPoli is pre-trained on the reference dataset and its conditions, and then fine-tuned to optimize the prototypes. In the query mapping phase, we freeze the model and learn the new conditional embeddings belonging to the query dataset. The label transfer is performed by simply assigning the cell type corresponding to the closest prototype in the latent embedding space.

Harmony and Symphony. Harmony and Symphony Kang et al. (2021) are two related methods for data integration and query-to-reference mapping. In contrast to the previous three methods, these are not VAE-based, but instead compute batch-corrected embeddings in the Principal Component Analysis (PCA) space. Each cell is first assigned to one or more soft-clusters, favoring clusters representing multiple datasets. Cluster-specific linear correction factors are then applied to the corresponding cells as a linear mixture model, leading to batch-corrected embeddings. Symphony computes soft-cluster assignments for the query cells based on their proximity to the reference cluster centroids, and then corrects their embeddings using Harmony's mixture model. Following Symphony's defaults, annotations are transferred to the query dataset using nearest neighbors in the embedding space.

A.3 Inflammation Atlas

A.3.1 Chemistry assays and sequencing pools

In this project, we are considering only data generated using 10x Chromium Next GEM single-cell kits, which include $3'$ or $5'$ mRNA amplification *chemistry assays*. Specifically, we are using version 2 and version 3 for the $3'$ assays, and version 1 and version 2 for the $5'$ assays. *Sequencing pools* refer to libraries where multiple samples have been pooled together and processed with a single 10x kit.

A.3.2 Inflammatory Conditions

Most of the samples were processed in-house, including patients with the following conditions: Rheumatoid Arthritis (RA), Psoriatic Arthritis (PSA), Crohn's Disease (CD), Ulcerative Colitis (UC), Psoriasis (PS), Systemic Lupus Erythematosus (SLE), Asthma, Chronic Obstructive Pulmonary Disease (COPD), Breast Cancer (BRCA), Cirrhosis, Colorectal Cancer (CRC), and COVID-19. Additionally, publicly available datasets were incorporated to complete the cohort, covering conditions the following conditions: Sepsis, Head and Neck Squamous Cell Carcinoma (HNSCC), Hepatitis B Virus (HBV), Multiple Sclerosis (MS), Nasopharyngeal Cancer (NPC), and Human Immunodeficiency Virus (HIV), COVID-19, and flu. Both public and in-house data included healthy controls.

A.3.3 Sources

In-house data sources Vall d'Hebron Research Institute within the DoCTIS consortia (`https://doctis.eu/`), Institut Hospital del Mar d'Investigacions Mèdiques, University Medical Center Groningen, Vall d'Hebron Institute of Oncology, Biomedical Research Institut Sant Pau, Katholieke Universiteit Leuven, Biomedical Research Institut Sant Pau.

Public datasets Reyes et al. (2020); Jiang et al. (2020); Cillo et al. (2020); Zhang et al. (2023); Schafflick et al. (2020); Liu et al. (2021); Palshikar et al. (2022); Wang et al. (2020); Perez et al.

[1]We followed the tutorial in `https://docs.scvi-tools.org/en/stable/tutorials/notebooks/scrna/query_hlca_knn.html`

(2022); Savage et al. (2021); Mistry et al. (2019); Ramachandran et al. (2019); Martin et al. (2019); Ahern et al. (2022); Ren et al. (2021); Terekhova et al. (2023).

A.3.4 CELL TYPE (*level-1*)

Peripheral blood mononuclear cell (PBMCs) includes myeloid and lymphoid compartments. The Lymphoid compartment includes B cells, Plasma cells, T CD4 Naive cells, T CD4 Non-Naive cells, T CD8 Naive cells, T CD8 Non-Naive cells, Unconventional T Cells (UTC), and Innate Lymphoid Cells (ILC). While the myeloid compartment includes plasmacytoid Dendritic Cells (pDC), Denditric Cells (DC), and Monocytes (Mono).

A.4 APPENDIX FIGURES

Method	Parameter	Search Space
`sklearn.svm.LinearSVC`	scaler	[True, False]
	fit_intercept	[True, False]
	class_weigth	[balanced, None]
	C	$[1e^{-3}, 1e^{-2}, \ldots, 1e^{5}]$
`sklearn.svm.SVC`	scaler	[True, False]
	kernel	[sigmoid, rbf, poly]
	class_weight	[balanced, None]
	C	$[1e^{-3}, 1e^{-2}, \ldots, 1e^{5}]$
`sklearn.neighbors.KNeighborsClassifier`	scaler	[True, False]
	metric	[cosine, euclidean]
	class_weight	[uniform, distance]
	n_neighbors	[1, 2, 3, . . . , 5]

Table A.1: **CT-MV classifier hyperparameter configurations.** Hyperparameter configurations for each classifier method employed for the CT-MV classifier hyperparameter tuning. Note that the classifier method itself is considered as a hyperparameter. By scaler, we are referring to `sklearn.preprocessing.StandardScaler`.

Method	Parameter set	Configurations
scANVI	(n_latent, n_hidden)	$(256, 20), (256, 30), (256, 50), (256, 100), (512, 200)$
scGen	n_latent	$20, 30, 50, 100, 200$
scPoli	(embedding_dims, latent_dim)	$([3, 20], 20),\ ([3, 50], 20),\ ([3, 100], 20),$ $([3, 20], 30),\ ([3, 100], 30),$ $([3, 20], 50),\ ([3, 100], 50),$ $([3, 20], 100),\ ([3, 100], 100)([3, 100], 200)$
Harmony	n_comps	$20, 30, 50, 100, 200$

Table A.2: **Data integration methods configurations**. Configurations for each data integration method employed in the query-to-reference mapping benchmark scenario.

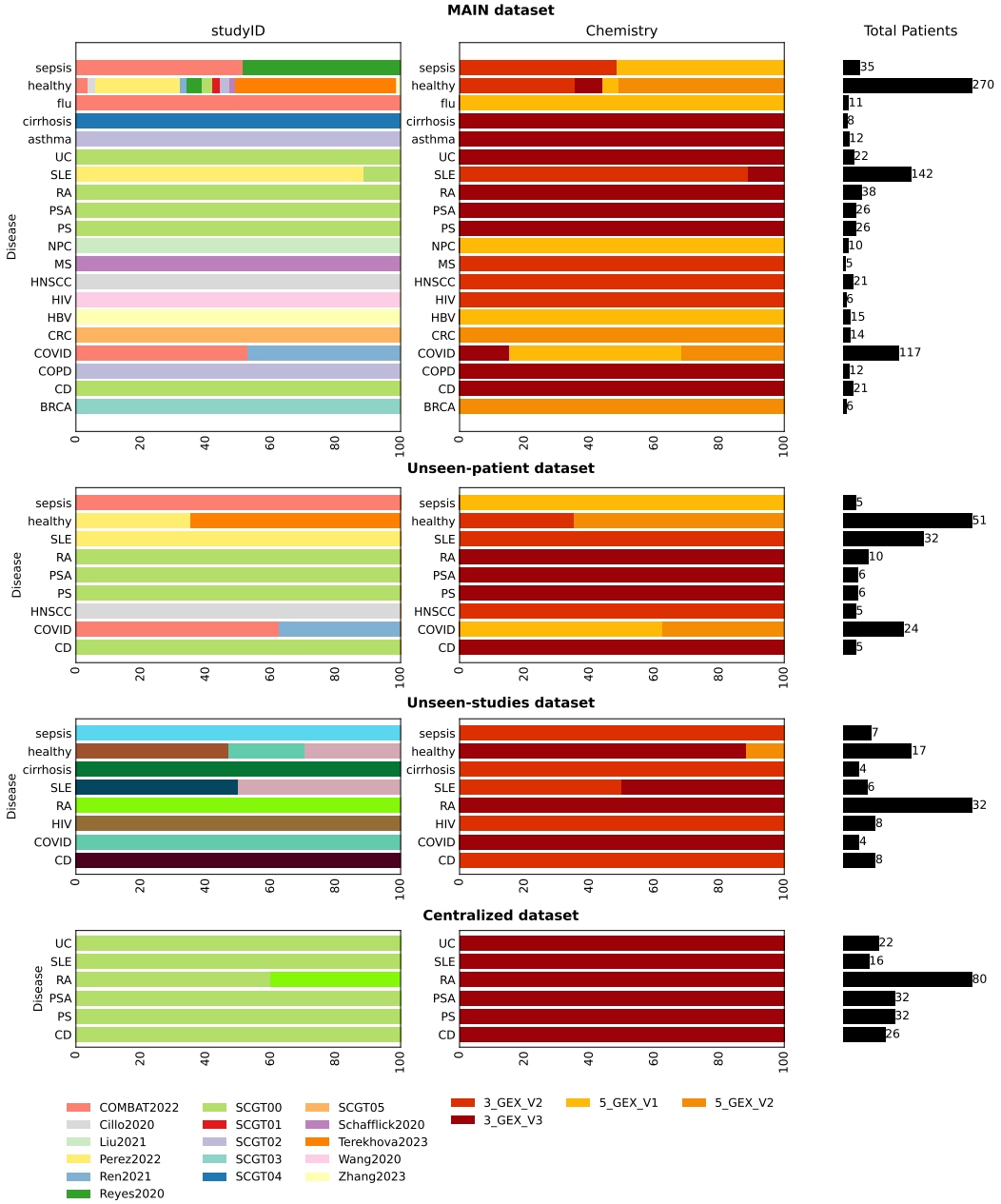

Figure A.1: **Patient distribution**. Distribution of patients across different datasets categorized by disease type and study ID. The main dataset, unseen-patient dataset, unseen-studies dataset, and centralized dataset are shown with their respective patient counts and chemistry data.