Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space

Lei Xiong[1,2,4,5], Kang Tian[1,2,5], Yuzhe Li[1,3], Qiangfeng Cliff Zhang[1,2,*]

[1] MOE Key Laboratory of Bioinformatics, Beijing Advanced Innovation Center for Structural Biology & Frontier Research Center for Biological Structure, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing, China 100084

[2] Tsinghua-Peking Center for Life Sciences, Beijing, China 100084

[3] Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China 100871

[4] Shanghai Qi Zhi Institute, Shanghai 200030, China

[5] Co-first authorship

[*] Correspondence: qczhang@tsinghua.edu.cn (Q.C.Z.)

ABSTRACT

Computational tools for integrative analyses of diverse single-cell experiments are facing formidable new challenges including dramatic increases in data scale, sample heterogeneity, and the need to informatively cross-reference new data with foundational datasets. Here, we present SCALEX, a deep-learning method that integrates single-cell data by projecting cells into a batch-invariant, common cell-embedding space in a truly online manner (*i.e.*, without retraining the model). SCALEX substantially outperforms online iNMF and other state-of-the-art non-online integration methods on benchmark single-cell datasets of diverse modalities, (e.g., scRNA-seq, scATAC-seq), especially for datasets with partial overlaps, accurately aligning similar cell populations while retaining true biological differences. We showcase SCALEX's advantages by constructing continuously expandable single-cell atlases for human, mouse, and COVID-19 patients, each assembled from diverse data sources and

growing with every new data. The online data integration capacity and superior performance makes SCALEX particularly appropriate for large-scale single-cell applications to build-upon previously hard-won scientific insights.

INTRODUCTION

Single-cell experiments enable the decomposition of samples into their constituent, diverse cell-types and cell states[1-4]. Many computational tools have been developed for integrative analysis of single-cell datasets, all seeking to separate biological variations from non-biological noise, such as batch effects of different donors, conditions, and/or analytical platforms[5,6]. The scope of the integration task is expanding rapidly with technical advances for single-cell studies, which continue to grow larger and larger in scale, now exceeding 1 million cells in some cases[7,8]. Moreover, the range of examined sample types is also increasing, and datasets now often include highly heterogenous cell subsets[9,10]. Most importantly, as single-cell studies become more routine, new studies should be informatively cross-referenced to foundational research stuides[7,8,11-15]. Thus, there is a growing need for integration tools that can manage single-cell data of large-scale and complex cell-type compositions while also supporting accurate alignment to and exploration within existing datasets.

Most current single-cell data integration methods (e.g., Seurat[16-18], MNN[19], Harmony[20], Conos[21], Scanorama[22], BBKNN[23], *etc.*) are based on the searching across batches for cell-correspondence, for instance similar individual cells or cell anchors/clusters. These methods suffer from three limitations. First, they are prone to mixing cell populations that only exist in some batches, which becomes a severe problem for the integration of complex datasets that contain non-overlapping cell populations in each batch (*i.e.*, partially overlapping data)[16,17]. Second, they require computational resources that increase dramatically as the number of cells and of batches increase, making these methods increasingly unsuitable for today's large-scale

53 single-cell datasets[7,8,11-15]. Finally, these methods can only remove batch effects from
54 the current dataset being assessed. Each time a new dataset is added, it requires an
55 entirely new integration process that changes the existing integration results of
56 previous studies. This requirement severely limits a tool's ability to continuously
57 integrate arriving new single-cell data without recalculating existing integrations from
58 scratch, a capacity referred to as "online" data integration[24].

59 Online data integration ability is becoming increasingly crucial with today's
60 single-cell experiments. The recently developed tool, online iNMF[24], an online version
61 of LIGER[25], iteratively applies integrative non-negative matrix factorization (iNMF) to
62 decouple the shared and dataset-specific factors related to cell identities, and thus is
63 able to incorporate new data with existing datasets on-the-fly. Another recently
64 developed package, scvi-tools[26], combining scVI[27] with scArches[28], applies a
65 conditional variational autoencoder (VAE)[29] framework to model the inherent
66 distribution of the input single-cell data for data integration. However, the conditional
67 VAE design of scVI requires model augmentation and retraining when integrating new
68 data, meaning that scVI is not an online method.

69 Here, we developed SCALEX as a method for online integration of heterogeneous
70 single-cell data based on a VAE framework. The encoder of SCALEX is designed to be
71 a data projection function that only preserves batch-invariant biological data
72 components when projecting single-cells. Importantly, the projection function is a
73 generalized one that requires no retraining on new data, thus allowing SCALEX to
74 integrate single-cell data in an online manner. Working with an extensive collection of
75 benchmark datasets, we demonstrate that SCALEX substantially outperforms online
76 iNMF as well as non-online single-cell data integration tools, in terms of integration
77 accuracy, scalability, and computationally efficiency. The advantages make SCALEX
78 particularly appropriate for the integration and research utilization of today's

single-cell datasets, which continue to grow along with the ongoing explosion of single-cell studies in biology and medicine.

RESULTS

SCALEX implements a generalized encoder that enables online integration of single-cell data

To enable online integration, the fundamental design concept underlying SCALEX is to implement a *generalized* projection function that disentangles the batch-related components away from the batch-invariant components of single-cell data and projects the batch-invariant components into a common cell-embedding space. We previously applied VAE and designed SCALE (Single-Cell ATAC-seq Analysis via Latent feature Extraction) to model and analyze single-cell ATAC-seq data[30]. We found that the encoder of SCALE has the potential to disentangle cell-type-related and batch-related features in a low-dimensional embedding space.

Here, to obtain a generalized encoder for data projection without retraining, SCALEX includes three specific design elements (Fig. 1a, Fig. S1, Methods). First, SCALEX implements a batch-free encoder that extracts only biological-related latent features (z) from input single-cell data (x) and a batch-specific decoder[29] that reconstructs the original data from z by incorporating batch information back during data reconstruction. Supplying batch information only to the decoder focuses the encoder exclusively on learning the batch-invariant biological components, which is crucial for the encoder generalizability. In contrast, scVI includes a set of batch-conditioned parameters into its encoder, which restrains the encoder from the generalizability with new batches and thus precludes online data integration. Second, SCALEX includes a Domain-Specific Batch Normalization (DSBN)[31] layer using multi-branch Batch Normalization[32] in its decoder to support incorporation of batch-specific variations during single-cell data reconstruction. Third, the SCALEX

105 encoder employs a mini-batch strategy that samples data from all batches (instead of
106 single batches), which more tightly follows the overall distribution of the input data.
107 Note that each mini-batch is subjected to a Batch Normalization layer in the encoder to
108 adjust the deviation of each mini-batch and to align it to the overall input distribution.

109 We conducted extensive analyses of SCALEX hyperparameters and also tested the
110 specific contributions of each design element by implementing a set of SCALEX
111 test-variants, each lacking an individual or a combination of the design elements, and
112 evaluating their performance for single-cell data integration (Supplementary Note). We
113 found that each design element is crucial for the integration performance of SCALEX.
114 More importantly, the combination of these design elements renders the encoder of
115 SCALEX a generalized function capable of accurate projection of single cell data from
116 different batches into a batch-invariant cell-embedding space, making SCALEX a truly
117 online data integration method.

118 SCALEX integration is substantially more accurate than state-of-the-art 119 single-cell data integration methods

120 We extensively assessed the basic data integration performance of SCALEX (Methods),
121 following the evaluative framework proposed in a recent comparative study[33]. We
122 examined multiple well-curated scRNA-seq datasets, including human *pancreas* (eight
123 batches of five studies)[34-38], *heart* (two batches of one study)[39] and *liver* (two
124 studies)[40,41]; as well as human non-small-cell lung cancer (*NSCLC*, four studies)[42-45]
125 and peripheral blood mononuclear cells (*PBMC*; two batches assayed by two different
126 protocols)[16]. Our comparison included online iNMF and other state-of-the-art
127 non-online single-cell data integration methods, including Seurat v3, Harmony, MNN,
128 Conos, BBKNN, Scanorama, LIGER (*i.e.*, batch iNMF), and scVI. We evaluated the
129 integration performance of these tools based on the benchmark datasets by Uniform

130 Manifold Approximation and Projection (UMAP)[46] embedding visualization as well as
131 a series of scoring metrics [19,20,47-49].

132 With UMAP embedding, we note that all of the raw datasets displayed strong batch
133 effects, with cell-types that were common in different batches separately distributed.
134 Overall, SCALEX, Seurat v3, and Harmony achieved the best integration performance
135 for most of the datasets by merging common cell-types across batches while keeping
136 disparate cell-types apart (Fig. S2). MNN, scVI, and Conos integrated many datasets
137 but left some common cell-types not well-aligned. Online iNMF, LIGER, BBKNN,
138 and Scanorama often had unmerged common cell-types, and sometimes incorrectly
139 mixed distinct cell-types together. For example, considering the T cell populations
140 between the two batches in the *PMBC* dataset (Fig. 1b), while SCALEX, Seurat v3,
141 Harmony, MNN, scVI integrations were effective, online iNMF misaligned some of
142 the CD4 naïve T cells with CD8 naïve T cells, and misaligned some NK cells with CD8
143 T cells.

144 SCALEX substantially outperformed all of the other methods for cell-type
145 clustering, as assessed by the adjusted Rand Index (ARI)[47] and the Normalized Mutual
146 Information (NMI)[48] (Fig. 1c, Fig. S3-S4). To quantify cell-type separation and batch
147 mixing we used two paired metrics (Methods): a pair comprising the Silhouette score[49]
148 and the batch entropy mixing score[19], as well as a pair comprising the cell-type and
149 integration local inverse Simpson's Indexes (cLISI and iLISI)[20]. Overall, SCALEX
150 achieved the highest scores for cell-type separation, and tied with Seurat v3 and
151 Harmony as the best-performing methods on the batch mixing metrics (Fig. S4).
152 Interestingly, we observed that both LIGER and online iNMF often scored the lowest
153 for cell-type separation yet the highest for batch mixing. However, after careful
154 investigation, we concluded that a higher batch mixing score does not necessarily
155 indicate better data integration, but instead often indicates an issue of over-correction,
156 which we consider in-depth in a dedicated subsection below.

157 **SCALEX is scalable to Atlas-level datasets and accommodates diverse data**
158 **modalities**

159 Single-cell datasets that contain a large number of cells and consist of heterogenous and
160 complex samples from multiple tissues have been termed "Atlas-level" datasets in a
161 recent comparative study[33]. These Atlas-level datasets are posing new challenges to
162 data integration tools. We tested the scalability and computation efficiency of
163 SCALEX by applying it to a typical Atlas-level dataset, the Human Fetal Atlas dataset,
164 which contains 4,317,246 cells from two data batches, GSE156793 and GSE134355
165 (Methods)[8,15] (Fig. S5a,b). SCALEX accurately integrated these two batches, showing
166 good alignment of the same cell-types (Fig. 1d). In addition to SCALEX, only BBKNN,
167 Scanorama, and scVI can be used to integrate this Atlas-level dataset, however, their
168 integrations does not separate and align the cell-types well, as indicated by the UMAP
169 embeddings (Fig. S5c) and the low cell-type separation and batch mixing scores (Fig.
170 S5d).

171 We compared the computational efficiency of different methods using
172 down-sampled datasets (of 10 K, 50 K, 250 K, 1 M, and 4M cells) from this Human
173 Fetal Atlas dataset. Both SCALEX and online iNMF consumed very efficient runtime
174 and memory that increased only linearly with data size. scVI also is scalable to 4M cells
175 with acceptable memory usage, whereas Seurat v3, Harmony, Conos, and LIGER
176 consumed runtime and/or memory that increased exponentially, thus did not scale
177 beyond 1M cells on a workstation of 64 CPU cores and 256 GB memory (Fig. 1e).
178 Notably, the deep learning framework of SCALEX enables it to run very efficiently on
179 GPU devices, requiring much reduced runtime (using about 20 minutes and 90 GB of
180 memory on the 4M dataset).

181 SCALEX can be used to integrate other modalities of single-cell data (*e.g.*,
182 scATAC-seq[50,51], CITE-Seq[52], *etc.*) and cross-modality data (*e.g.*, simultaneous

183 analysis of scRNA-seq and scATAC-seq) (Methods). SCALEX substantially
184 outperformed all other methods for integration of mouse brain scATAC-seq datasets
185 (two batches assayed by snATAC and 10X)[53] (Fig. 1f, Fig. S6a-c), and performed well
186 for integration of additional single-cell data modalities including CITE-seq[52] and
187 spatial transcriptome MERFISH data[54] (Fig. S6d,e). We also used SCALEX to
188 integrate a cross-modality dataset (scRNA-seq and scATAC-seq)[55,56] and found that
189 SCALEX correctly integrated the two modalities of data and distinguished rare cells
190 that are specific to the scRNA-seq data, including pDC and platelet cells (Fig. 1g),
191 doing so better than other methods according to both UMAP embeddings and multiple
192 analytical metrics (Fig. S7).

193 **SCALEX integrates partially overlapping datasets without over-correction**

194 Many recent single-cell datasets, especially Atlas-level datasets, feature high sample
195 heterogeneity and complex cell-type compositions[9,10]. These datasets often contain
196 partially overlapping batches where each batch contains some non-overlapping cell
197 populations. For example, the *liver* dataset is a partially overlapping dataset where the
198 hepatocyte population contains multiple subtypes specific to different batches: three
199 subtypes are specific to LIVER_GSE124395, and two other subtypes only appear in
200 LIVER_GSE115469 (Fig. S8).

201 This partial overlap problem presents a major challenge for single-cell data
202 integration and often leads to an issue of over-correction (*i.e.*, mixing of distinct
203 cell-types), especially for those local cell similarity-based methods[16,17]. For example,
204 Seurat v3 mixed the hepatocyte-CXCL1, hepatocyte-CYP2A13, and
205 hepatocyte-TAT-AS1 cells and Harmony mixed the hepatocyte-CYP2A13 and
206 hepatocyte-TAT-AS1 cells (Fig. 2a). As a global integration method that projects cells
207 into a common cell-embedding space, SCALEX is expected to be less sensitive to this
208 problem. Indeed, we noticed that SCALEX correctly maintained the five hepatocyte

209 subtypes apart (as did scVI. Fig. 2a). Unexpectedly, despite being a global method,
210 online iNMF severely suffered from over-correction, mixing all five hepatocyte
211 subtypes, and even mixing B cells and NK cells (Fig. 2a), presumably because its
212 matrix factoring algorithm forced the alignment of distinct cell-types.

213 We defined an over-correction score, a metric to measure this over-correction
214 problem based on the percent of cells with inconsistent cell-types in the neighborhood
215 for each cell (Methods). Formally, the over-correction score is a negative index, *i.e.*, the
216 higher the over-correction score, the more severe the extent of inaccurate mixing of
217 cell-types. For the benchmark datasets, SCALEX had the lowest over-correction scores
218 (Fig. 2b), whereas online iNMF yielded extremely high over-correction scores.

219 To systematically characterize the performance of different methods on partially
220 overlapping datasets, we constructed test datasets with a range of common cell-types,
221 that we generated based on down-sampling of the six major cell-types in the *pancreas*
222 dataset (Methods). SCALEX integration was accurate for all cases, aligning the same
223 cell-types without over-correction, whereas Seurat v3, Harmony, and online iNMF
224 frequently mixed distinct cell-types (Fig. 2c-d). Although scVI showed one of the
225 lowest levels of over-correction when integrating partially overlapping datasets, it is
226 prone to mistakenly splitting one cell-type into many small groups. We noted that the
227 severity of over-correction and error-splitting is amplified as the overlapping number
228 decreases (Fig. S9). When there were no common cell-types, both Seurat v3 and
229 Harmony collapsed the six cell-types into three, mixing alpha with gamma cells, beta
230 with delta cells, and acinar with ductal cells to varying extents, whereas scVI split alpha
231 cells into 6 groups. We repeated this down-sampling analysis from the 12 cell-types in
232 the *PBMC* dataset and observed similar results of over-correction and error-splitting
233 (Fig. S10).

234 **SCALEX increases the scope and resolution of an existing cell space by adding**
235 **new data through online projection**

236 The generalizability of SCALEX's encoder to project cells from various sources into a
237 common cell-embedding space without model retraining allows SCALEX to integrate
238 new single-cell data with existing data in an online manner. We tested the online data
239 integration performance of SCALEX for newly arriving data based on the *pancreas*
240 dataset. Prior to projection, we first used SCALEX to integrate the *pancreas* dataset
241 and this accurately removed the strong batch effect that was evident in the raw data (Fig.
242 3a, Fig. S11a-b).

243 We subsequently projected three new batches of scRNA-seq data[57-59] for pancreas
244 tissues (Fig. 3b) into this "pancreas cell space" using the same SCALEX encoder
245 trained on the original *pancreas* dataset. After projection, most of the cells in the new
246 batches were accurately aligned to the correct cell-types in the pancreas cell space,
247 enabling their accurate annotation by cell-type label transfer (Fig. 3c, Method). We
248 benchmarked projection accuracy by calculating the ARI, the NMI, and the F1 score to
249 evaluate cell-type annotation by label transfer with cell-type information in the original
250 studies (Methods). We compared the results with online iNMF and scVI, the only two
251 tools that are able to project cells into an *existing* cell-space (note that data projection of
252 scVI needs model retaining through scArches). SCALEX achieved the highest
253 projection accuracy in comparisons with online iNMF and scVI (Fig. 3c). scVI also
254 achieved high accuracy, projecting most cells onto right locations, with only a few
255 exceptions of alpha and ductal cells (Fig. S11c). Online iNMF mixed distinct cell-types
256 when incorporating new batches, *e.g*. projecting some alpha cells onto the locations of
257 gamma and delta cells, and delta cells onto beta cells (Fig. S11c), which in turn led to
258 wrong annotations during label transfer (Fig. 3c).

259 The ability to project new single-cell data into an existing cell-embedding space
260 allows SCALEX to readily enrich (*i.e.*, to add biological resolution) this cell space with
261 additional informative details. To verify this, we projected two additional melanoma
262 data batches (SKCM_GSE72056, SKCM_GSE123139)[10,60] onto the previously
263 constructed PBMC space. Again, SCALEX correctly projected all common cell-types
264 were onto the same locations in the PBMC cell space (Fig. 3d), but online iNMF mixed
265 tumor cells with plasma, monocyte and CD8 T cells, and scVI split the CD8T cells into
266 several distinct groups (Fig. S12). Importantly, we noticed that for the tumor and
267 plasma cells only present in the melanoma data batches, SCALEX did not project these
268 cells onto any existing cell populations in the PBMC space; rather, it projected them
269 onto new locations close to similar cells, with the plasma cells projected to a location
270 near B cells, and the tumor cells projected to a location near HSC cells (Fig. 3e). This
271 indicates that SCALEX can enrich an existing cell space with new cell-types through
272 data projection.

273 SCALEX projection also enables *post hoc* annotation of unknown cell-types in an
274 existing cell space using new data. For instance, we noted a group of previously
275 uncharacterized cells in the *pancreas* dataset (Fig. 3a). We found that these cells
276 displayed high expression levels of known epithelial gene markers (Methods). We
277 therefore assembled a collection of epithelial cells from the *bronchial epithelium*
278 dataset[61], and then projected these epithelial cells onto the pancreas cell space. We
279 found that a group of antigen-presenting airway epithelial (SLC16A7+ epithelial) cells
280 were projected onto the same location of the uncharacterized cells (Fig. 3f). These data,
281 together with the observation that both cell populations showed similar marker gene
282 expression (Fig. 3g), suggest that these uncharacterized cells are also SLC16A7+
283 epithelial cells. Note that online iNMF and scVI were not able to identify this small
284 group of epithelial cells, because they were split into several smaller groups and/or
285 were often mixed with other cell-types (Fig. S2). SCALEX thus enables discovery

286 science in cell biology by supporting exploratory analysis with large numbers of
287 diverse datasets.

288 **SCALEX integration constructs expandable single-cell atlases**

289 The ability to combine heterogenous data into a common cell-embedding space makes
290 SCALEX a powerful tool to construct a single-cell atlas from a collection of diverse
291 datasets. We applied SCALEX integration to three large and complex datasets: the
292 Mouse Atlas dataset (comprising multiple organs from two studies assayed by 10X,
293 Smart-seq2, and Microwell-seq[12,14]), the Human Atlas dataset (comprising multiple
294 organs from two studies assayed by 10X and Microwell-seq[15,62]), and the Human Fetal
295 Atlas dataset[8,15] (Fig. S13).

296 Despite the strong batch effects in the raw data, SCALEX accurately integrated the
297 three batches of the Mouse Atlas data into a common cell-embedding space (Fig. 4a-c,
298 Fig. S14a). Common cell-types were well-aligned at the same position in the cell space,
299 including B, T, and endothelial cells presented in all tissues, and proximal tubule,
300 urothelial, and hepatocytic cells from particular tissues. Distinct cell-types were
301 located separately, such as sperm, Leydig, and small intestine cells from the
302 Microwell-seq data, keratinocyte stem cells and large intestine cells from the
303 Smart-seq2 data, indicating that biological variations were well preserved (Fig S14b,c).
304 We compared SCALEX with all other methods and found that SCALEX performed the
305 best for cell-type clustering, especially for avoiding over-correction (Fig. 4d,e, Fig.
306 S13b).

307 Importantly, atlases generated with SCALEX can be further expanded by
308 projecting new single-cell data to support comparative studies of cells both in the
309 original atlas and in the new data. To illustrate this utility, we projected two additional
310 data batches of aged mouse tissues from *Tabula Muris Senis* (Smart-seq2 and 10X)[13]
311 and two single tissue datasets (lung and kidney)[63] onto the SCALEX Mouse Atlas cell

312 space. We found that cells in the new data batches were correctly projected onto the
313 locations of the same cell-types in the cell-embedding space of the initial atlas (Fig. 4f)
314 as confirmed by the accurate cell-type annotations for the new data by label transfer
315 (Fig. 4g. Methods).

316 Following the same strategy, we constructed a SCALEX Human Atlas by
317 integration of multiple tissues from two studies (GSE134355, GSE159929) (Fig.
318 S15a,b). SCALEX effectively eliminated the batch effects in the original data and
319 integrated the two datasets (Fig. S15c,d). Again, we were able to correctly project two
320 additional human skin datasets (GSE130973, GSE147424)[64,65] onto the Human Atlas
321 cell space (Fig. S15e), and accurately annotated these projected skin cells (Fig. S15f.
322 Methods). In sum, these results illustrate that SCALEX enables: i) researchers to
323 evaluate their project-specific single cell datasets by leveraging existing information in
324 large-scale (and ostensibly well annotated) cell atlases; and ii) atlas creators to
325 informatively integrate new datasets and derive new biological insights from new
326 research programs.

327 An integrative SCALEX COVID-19 PBMC Atlas revealed distinct immune 328 responses among COVID-19 patients

329 Many single-cell studies have been conducted to analyze COVID-19 patient immune
330 responses[66-73]. However, these studies often suffer from small sample size and/or
331 limited sampling of various disease states[67,73]. For a comprehensive study, we used
332 SCALEX to generate a COVID-19 PBMC Atlas, integrating data from nine COVID-19
333 studies, involving a total of 860,746 single cells in 10 batches[66-72] (Fig. 5a, Fig. S16a,b).
334 We identified 22 cell-types, each of which has support from gene expression data for
335 canonical markers (Fig. 5b,c, Fig. S16c,d. Methods). Cells across different studies were
336 integrated accurately with the same cell-types aligned together, confirming the

337 integration performance of SCALEX (Fig. 5c, Fig. S16d), which was much better than
338 the other methods in terms of cell-type clustering (Fig. S16a,b).

339 Interestingly, we found that some cell subpopulations were differentially
340 associated with patient status (Fig 5d). A subpopulation of CD14 monocytes
341 (CD14-ISG15-Mono) was characterized by its high expression of Type I
342 interferon-stimulated genes (ISGs) and genes enriched with immune-response-related
343 GO terms (Fig 5e,f). The frequency of CD14-ISG15-Mono cells increased significantly
344 from mild/moderate to severe patients (Fig. 5g, Fig. S17a. Methods). Within the
345 COVID-19 patients, we observed a significant decrease in ISG gene expression in
346 CD14-ISG15-Mono cells between the mild/moderate and severe cases, suggesting an
347 immune exhaustion-like response in severe COVID-19 patients (Fig. 5e).

348 Additionally, a neutrophil subpopulation (NCF1-Immature_Neutrophil),
349 characterized by decreased expression of the genes responsible for neutrophil
350 activation but elevated expression of genes enriched with viral-process-related GO
351 terms, was specifically enriched in severe verse mild/moderate patients (Fig. S17b,c).
352 A plasma cell subpopulation (MZB1-Plasma), characterized by decreased expression
353 of genes related to antibody production and enriched for immune and inflammatory
354 response-related GO terms, were also enriched in severe patients (Fig. S17d,e). Thus,
355 the SCALEX COVID-19 PBMC atlas, generated by integrating a highly diverse
356 collection of single-cell data from individual studies, identified multiple immune
357 cell-types that become progressively dysfunctional during COVID-19 disease
358 progression. Importantly, these cell trends were not and could not have been detected in
359 the small-scale, individual studies that served as the basis for our SCALEX COVID-19
360 PBMC atlas.

361 **Online integration of the SCALEX COVID-19 PBMC Atlas with the SC4**
362 **consortium study**

363 Our analysis based on the SCALEX COVID-19 PBMC Atlas yielded findings
364 consistent with two conclusions from the Single Cell Consortium for COVID-19 in
365 China (SC4) study, a recent large-scale effort that generated a single-cell atlas of over 1
366 million cells from 171 COVID-19 patients and 25 healthy controls[7] (Fig. S18a). First,
367 both studies observed the same set of immune cell subpopulations which displayed
368 differential associations with COVID-19 severity. The proportions of CD14 monocytes,
369 megakaryocytes, plasma cells, and pro T cells were elevated with increasing disease
370 severity, while the proportion of pDC and mDC cells decreased (Fig. 5g). Second,
371 based on calculating the same cytokine score and inflammatory score (defined in the
372 SC4 study) for the cells in our SCALEX COVID-19 PBMC Atlas, we confirmed that
373 the megakaryocytes and monocyte subpopulations are associated with cytokine storms
374 triggered by SARS-Cov2 infection and are further elevated in severe patients[74] (Fig. 5j.
375 Methods).

376 SCALEX's online integration capacity enables us to project the SC4 consortium
377 dataset into the cell space of the SCALEX COVID-19 PBMC Atlas. We found that the
378 cell-types of two atlases were well-aligned (Fig. 5h,i, Fig. S18b,c). Integration of the
379 SC4 data further substantially improved both the scope and resolution of the SCALEX
380 COVID-19 PBMC Atlas. First, this data added macrophages and epithelial cells to the
381 cell space, enabling investigation of their potential involvement in COVID-19. The
382 integration also supported more precise characterization of specific cell subpopulations.
383 For example, the megakaryocyte population, not distinguished in either the SCALEX
384 COVID-19 PBMC Atlas or the SC4 Atlas (Fig. S18f), were divided into two
385 subpopulations in the combined atlas after projection of SC4 (Fig. 5h). An exploratory
386 functional analysis of the differentially expressed genes in these two newly delineated
387 megakaryocyte subpopulations (TUBA8-Mega and IGKC-Mega, Fig. S18d,e) revealed
388 enrichment for the GO terms "humoral immune response" for IGKC-Mega cells, yet
389 enrichment for "negative regulation of platelet activation" for TUBA8-Mega cells (Fig.
390 5k). These results illustrate how the continuously expandable single-cell atlases

generated using SCALEX capitalize on existing large-scale data resources and also facilitate the discovery of new biological and biomedical insights.

DISCUSSION

Single cell studies are becoming more and more prevalent, growing larger and larger in scale, and expanding in the scope of sample types, often with quite heterogenous cell subsets. Thus, there is a great need for data integration tools to accurately and efficiently handle these Atlas-level datasets[33]. Further, there is also a need for online integration capacity to continuously incorporate incoming new data with existing integrations without having to recalculate from scratch[24]. By design, SCALEX learns a generalized projection function to project heterogeneous single-cell data into a common cell-embedding space, enabling it to achieve *bona fide* online data integration. SCALEX is also computationally efficient, and preserves biological variations and avoids over-correction when integrating partially overlapping datasets.

These features make SCALEX particularly useful for Atlas-level datasets, allowing the integration of many single-cell studies to support ongoing, very large-scale research programs throughout the life sciences and biomedicine. We speculate that use of SCALEX to project single-cell datasets from highly diverse cancer types to construct a pan-cancer single-cell atlas may lead to the discovery of previously unknown cell-types that are common to divergent carcinomas and that function in pathogenesis, malignant progression, and/or metastasis.

REFERENCES

415 1 Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann,
416 S. A. The technology and biology of single-cell RNA sequencing. *Mol Cell* **58**,
417 610-620, doi:10.1016/j.molcel.2015.04.005 (2015).
418 2 Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell
419 heterogeneity. *Nat Rev Immunol* **18**, 35-45, doi:10.1038/nri.2017.76 (2018).
420 3 Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to
421 mechanism. *Nature* **541**, 331-338, doi:10.1038/nature21350 (2017).
422 4 Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of
423 human immune cell development and intratumoral T cell exhaustion. *Nature*
424 *Biotechnology* **37**, 925-936, doi:10.1038/s41587-019-0206-z (2019).
425 5 Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects
426 in high-throughput data. *Nat Rev Genet* **11**, 733-739, doi:10.1038/nrg2825
427 (2010).
428 6 Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and
429 technical variability in single-cell RNA-sequencing experiments. *Biostatistics*
430 **19**, 562-578, doi:10.1093/biostatistics/kxx053 (2018).
431 7 Ren, X. *et al.* COVID-19 immune features revealed by a large-scale
432 single-cell transcriptome atlas. *Cell*, doi:10.1016/j.cell.2021.01.053 (2021).
433 8 Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**,
434 doi:10.1126/science.aba7721 (2020).
435 9 Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast
436 Tumor Microenvironment. *Cell* **174**, 1293-1308 e1236,
437 doi:10.1016/j.cell.2018.05.060 (2018).
438 10 Li, H. *et al.* Dysfunctional CD8 T Cells Form a Proliferative, Dynamically
439 Regulated Compartment within Human Melanoma. *Cell* **176**, 775-789.e718,
440 doi:https://doi.org/10.1016/j.cell.2018.11.043 (2019).
441 11 Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, doi:10.7554/eLife.27041
442 (2017).
443 12 Tabula Muris, C. *et al.* Single-cell transcriptomics of 20 mouse organs creates
444 a Tabula Muris. *Nature* **562**, 367-372, doi:10.1038/s41586-018-0590-4
445 (2018).
446 13 Tabula Muris, C. A single-cell transcriptomic atlas characterizes ageing
447 tissues in the mouse. *Nature* **583**, 590-595, doi:10.1038/s41586-020-2496-1
448 (2020).
449 14 Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**,
450 1091-1107 e1017, doi:10.1016/j.cell.2018.02.001 (2018).
451 15 Han, X. *et al.* Construction of a human cell landscape at single-cell level.
452 *Nature* **581**, 303-309, doi:10.1038/s41586-020-2157-4 (2020).
453 16 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating
454 single-cell transcriptomic data across different conditions, technologies, and
455 species. *Nat Biotechnol* **36**, 411-420, doi:10.1038/nbt.4096 (2018).

17 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).

18 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell*, doi:https://doi.org/10.1016/j.cell.2021.04.048 (2021).

19 Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421-427, doi:10.1038/nbt.4091 (2018).

20 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*, doi:10.1038/s41592-019-0619-0 (2019).

21 Barkas, N. *et al.* Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat Methods* **16**, 695-698, doi:10.1038/s41592-019-0466-z (2019).

22 Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol*, doi:10.1038/s41587-019-0113-3 (2019).

23 Polanski, K. *et al.* BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964-965, doi:10.1093/bioinformatics/btz625 (2020).

24 Gao, C. *et al.* Iterative Refinement of Cellular Identity from Single-Cell Data Using Online Learning. *bioRxiv*, 2020.2001.2016.909861, doi:10.1101/2020.01.16.909861 (2020).

25 Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887 e1817, doi:10.1016/j.cell.2019.05.006 (2019).

26 Gayoso, A. *et al.* scvi-tools: a library for deep probabilistic analysis of single-cell omics data. *bioRxiv*, 2021.2004.2028.441833, doi:10.1101/2021.04.28.441833 (2021).

27 Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053-1058, doi:10.1038/s41592-018-0229-2 (2018).

28 Lotfollahi, M. *et al.* Query to reference single-cell integration with transfer learning. *bioRxiv*, 2020.2007.2016.205997, doi:10.1101/2020.07.16.205997 (2020).

29 Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114* (2013).

30 Xiong, L. *et al.* SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat Commun* **10**, 4576, doi:10.1038/s41467-019-12630-7 (2019).

31 Chang, W.-G., You, T., Seo, S., Kwak, S. & Han, B. Domain-Specific Batch Normalization for Unsupervised Domain Adaptation. *arXiv:1906.03950* (2019).

32 Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167* (2015).

498 33 Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell
499 genomics. *bioRxiv*, 2020.2005.2022.111161, doi:10.1101/2020.05.22.111161
500 (2020).
501 34 Lawlor, N. *et al.* Single-cell transcriptomes identify human islet cell signatures
502 and reveal cell-type-specific expression changes in type 2 diabetes. *Genome*
503 *Res* **27**, 208-222, doi:10.1101/gr.212720.116 (2017).
504 35 Segerstolpe, A. *et al.* Single-Cell Transcriptome Profiling of Human
505 Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab* **24**, 593-607,
506 doi:10.1016/j.cmet.2016.08.020 (2016).
507 36 Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas.
508 *Cell Syst* **3**, 385-394 e383, doi:10.1016/j.cels.2016.09.002 (2016).
509 37 Grun, D. *et al.* De Novo Prediction of Stem Cell Identity using Single-Cell
510 Transcriptome Data. *Cell Stem Cell* **19**, 266-277,
511 doi:10.1016/j.stem.2016.05.010 (2016).
512 38 Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse
513 Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**,
514 346-360 e344, doi:10.1016/j.cels.2016.08.011 (2016).
515 39 Litvinukova, M. *et al.* Cells of the adult human heart. *Nature* **588**, 466-472,
516 doi:10.1038/s41586-020-2797-4 (2020).
517 40 Aizarani, N. *et al.* A human liver cell atlas reveals heterogeneity and epithelial
518 progenitors. *Nature* **572**, 199-204, doi:10.1038/s41586-019-1373-2 (2019).
519 41 MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals
520 distinct intrahepatic macrophage populations. *Nat Commun* **9**, 4383,
521 doi:10.1038/s41467-018-06318-7 (2018).
522 42 Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor
523 microenvironment. *Nat Med* **24**, 1277-1289, doi:10.1038/s41591-018-0096-5
524 (2018).
525 43 Song, Q. *et al.* Dissecting intratumoral myeloid cell plasticity by single cell
526 RNA-seq. *Cancer Med* **8**, 3072-3085, doi:10.1002/cam4.2113 (2019).
527 44 Zilionis, R. *et al.* Single-Cell Transcriptomics of Human and Mouse Lung
528 Cancers Reveals Conserved Myeloid Populations across Individuals and
529 Species. *Immunity* **50**, 1317-1334 e1310, doi:10.1016/j.immuni.2019.03.009
530 (2019).
531 45 Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and
532 cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* **11**,
533 2285, doi:10.1038/s41467-020-16164-1 (2020).
534 46 McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold
535 Approximation and Projection for Dimension Reduction. *arXiv:1802.03426*
536 (2018).
537 47 Hubert, L. & Arabie, P. Comparing partitions. *Journal of Classification* **2**,
538 193-218, doi:10.1007/BF01908075 (1985).

539 48 Amelio, A. & Pizzuti, C. in *Proceedings of the 2015 IEEE/ACM International*
540 *Conference on Advances in Social Networks Analysis and Mining 2015*
541 1584–1585 (Association for Computing Machinery, Paris, France, 2015).
542 49 Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and
543 validation of cluster analysis. *Journal of Computational and Applied*
544 *Mathematics* **20**, 53-65, doi:https://doi.org/10.1016/0377-0427(87)90125-7
545 (1987).
546 50 Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin
547 accessibility by combinatorial cellular indexing. *Science* **348**, 910-914,
548 doi:10.1126/science.aab1601 (2015).
549 51 Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of
550 regulatory variation. *Nature* **523**, 486-490, doi:10.1038/nature14590 (2015).
551 52 Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in
552 single cells. *Nat Methods* **14**, 865-868, doi:10.1038/nmeth.4380 (2017).
553 53 Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with
554 SnapATAC. *Nat Commun* **12**, 1337, doi:10.1038/s41467-021-21583-9 (2021).
555 54 Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA
556 imaging. Spatially resolved, highly multiplexed RNA profiling in single cells.
557 *Science* **348**, aaa6090, doi:10.1126/science.aaa6090 (2015).
558 55 Genomics, X. 10k Peripheral blood mononuclear cells (PBMCs) from a
559 healthy donor, Single Cell ATAC Dataset by Cell Ranger 1.0.1. (2018).
560 56 Genomics, X. 10k PBMCs from a Healthy Donor (v3 chemistry), Single Cell
561 Gene Expression Dataset by Cell Ranger 3.0.0. (2018).
562 57 Wang, Y. J. *et al.* Single-Cell Transcriptomics of the Human Endocrine
563 Pancreas. *Diabetes* **65**, 3028-3038, doi:10.2337/db16-0405 (2016).
564 58 Enge, M. *et al.* Single-Cell Analysis of Human Pancreas Reveals
565 Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* **171**,
566 321-330 e314, doi:10.1016/j.cell.2017.09.004 (2017).
567 59 Xin, Y. *et al.* Pseudotime Ordering of Single Human beta-Cells Reveals States
568 of Insulin Production and Unfolded Protein Response. *Diabetes* **67**, 1783-1794,
569 doi:10.2337/db18-0365 (2018).
570 60 Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma
571 by single-cell RNA-seq. *Science* **352**, 189-196, doi:10.1126/science.aad0501
572 (2016).
573 61 Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals
574 the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377-381,
575 doi:10.1038/s41586-018-0394-6 (2018).
576 62 He, S. *et al.* Single-cell transcriptome profiling of an adult human cell atlas of
577 15 major organs. *Genome Biol* **21**, 294, doi:10.1186/s13059-020-02210-0
578 (2020).

579 63 Kimmel, J. C. *et al.* Murine single-cell RNA-seq reveals cell-identity- and
580 tissue-specific trajectories of aging. *Genome Res* **29**, 2088-2103,
581 doi:10.1101/gr.253880.119 (2019).
582 64 Sole-Boldo, L. *et al.* Single-cell transcriptomes of the human skin reveal
583 age-related loss of fibroblast priming. *Commun Biol* **3**, 188,
584 doi:10.1038/s42003-020-0922-4 (2020).
585 65 He, H. *et al.* Single-cell transcriptome analysis of human skin identifies novel
586 fibroblast subpopulation and enrichment of immune subsets in atopic
587 dermatitis. *J Allergy Clin Immunol* **145**, 1615-1628,
588 doi:10.1016/j.jaci.2020.01.042 (2020).
589 66 Schulte-Schrepping, J. *et al.* Severe COVID-19 Is Marked by a Dysregulated
590 Myeloid Cell Compartment. *Cell* **182**, 1419-1440.e1423,
591 doi:10.1016/j.cell.2020.08.001 (2020).
592 67 Lee, J. S. *et al.* Immunophenotyping of COVID-19 and influenza highlights
593 the role of type I interferons in development of severe COVID-19. *Sci*
594 *Immunol* **5**, doi:10.1126/sciimmunol.abd1554 (2020).
595 68 Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in
596 patients with severe COVID-19. *Nat Med* **26**, 1070-1076,
597 doi:10.1038/s41591-020-0944-y (2020).
598 69 Guo, C. *et al.* Single-cell analysis of two severe COVID-19 patients reveals a
599 monocyte-associated and tocilizumab-responding cytokine storm. *Nat*
600 *Commun* **11**, 3924, doi:10.1038/s41467-020-17834-w (2020).
601 70 Yao, C. *et al.* Cell-Type-Specific Immune Dysregulation in Severely Ill
602 COVID-19 Patients. *Cell Rep* **34**, 108590, doi:10.1016/j.celrep.2020.108590
603 (2021).
604 71 Zhang, J. Y. *et al.* Single-cell landscape of immunological responses in
605 patients with COVID-19. *Nat Immunol* **21**, 1107-1118,
606 doi:10.1038/s41590-020-0762-x (2020).
607 72 Ballestar, E. *et al.* Single cell profiling of COVID-19 patients: an international
608 data resource from multiple tissues. *medRxiv*, 2020.2011.2020.20227355,
609 doi:10.1101/2020.11.20.20227355 (2020).
610 73 Bernardes, J. P. *et al.* Longitudinal Multi-omics Analyses Identify Responses
611 of Megakaryocytes, Erythroid Cells, and Plasmablasts as Hallmarks of Severe
612 COVID-19. *Immunity* **53**, 1296-1314 e1299,
613 doi:10.1016/j.immuni.2020.11.017 (2020).
614 74 Chen, G. *et al.* Clinical and immunological features of severe and moderate
615 coronavirus disease 2019. *J Clin Invest* **130**, 2620-2629,
616 doi:10.1172/JCI137244 (2020).
617 75 Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization.
618 *arXiv:1412.6980* (2014).

76 Danese, A., Richter, M. L., Fischer, D. S., Theis, F. J. & Colomé-Tatché, M. EpiScanpy: integrated single-cell epigenomic analysis. *bioRxiv*, doi:10.1101/648097 (2019).

77 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15, doi:10.1186/s13059-017-1382-0 (2018).

78 Stuart, T., Srivastava, A., Lareau, C. & Satija, R. Multimodal single-cell chromatin analysis with Signac. *bioRxiv*, doi:10.1101/2020.11.09.373613 (2020).

79 Abdi, H. & Williams, L. J. Principal component analysis. *WIREs Comput. Stat.* **2**, 433-459, doi:10.1002/wics.101 (2010).

80 Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 5233, doi:10.1038/s41598-019-41695-z (2019).

Methods

Overview of the SCALEX model. SCALEX applies a variational autoencoder (VAE) to project the different batches of datasets into the same batch-invariant low-dimensional embeddings by learning a batch-free encoder and a batch-specific decoder simultaneously. Since the encoder and decoder are coupled to learn a batch-free encoder, a batch label (b) is only exposed to the decoder within the domain-specific batch normalization (DSBN), thus the decoder captures the batch information while the encoder learns the domain-invariant features. In the encoder, SCALEX takes the input expression profile (x) across all the batches as a whole mixture distribution without distinguishing their batch sources and extracts their mean (μ) and variance (σ^2) of the latent representations (z) in a 10-dimension embedding space to learn their global data structure. A standard multivariate Gaussian prior is used for z, while the approximated distribution of z is re-parameterized by $z = \mu + \sigma * \varepsilon$, where ε is sampled from $\mathbb{N}(0, I)$. In the decoder, SCALEX maps the latent representations with batch label (b) back to their original profile. To enable the decoder to capture the batch-specific variations, a DSBN layer is applied to learn a batch-specific normalization for each batch label (b), before transforming them back

to their original profile with the new batch variations. To learn the global distribution to avoid overcorrection on partially overlapping datasets, within each mini-batch in the training process, SCALEX randomly samples data from all batches and trains on them together with Batch Normalization to smooth the batch-specific shifts and align to the global distribution. Once trained, the encoder of SCALEX is generalized to any batches and serves as a universal function for globally mapping different batches of datasets into the same batch-invariant space.

Training SCALEX is to maximize the log-likelihood of the observed single-cell sequencing data x:

$$\begin{aligned} logp(x) &= log \int_z p(x,z)dz \\ &\geq E_{q(z|x)}\left[log\frac{p(x,z)}{q(z|x)}\right] \\ &= \mathcal{L}_{ELBO}(x) \end{aligned}$$

Then the loss function is transformed into the evidence lower bound (ELBO). While the ELBO can be further decomposed into two terms:

$$\mathcal{L}_{ELBO}(x) = E_{q(z|x)}[logp(x|z)] - D_{KL}(q(z|x)||p(z))]$$

The first term is the reconstruction term, which minimizes the distance between the generated output data x' and the original input data x, calculated as the binary cross entropy between x' and x. The second term is the regularization term, which minimizes the Kullback-Leibeler divergence between posterior distribution $\mathbb{N}(\mu,\sigma)$ and prior distribution $\mathbb{N}(0,I)$ of latent representations z. To enable a more flexible alignment under the latent space, we adjusted the coefficient of the second term to 0.5 after hyper-parameter optimization via a grid search; thus, the final loss function is:

$$\mathcal{L}_{ELBO}(x) = E_{q(z|x)}[logp(x|z)] - 0.5 * D_{KL}(q(z|x)||p(z))]$$

The overall network architecture of SCALEX consists of an encoder and a decoder. The encoder is a two-layer neural network (fully connected [1024]-BN-ReLU-fully connected[10]) for mean (μ) and variance (σ^2) of the 10-dimension latent representations z using a reparameterization to obtain latent representations z, and the decoder has only one layer (no hidden layer), directly connecting latent representations z to the output x (fully connected-DSBN-Sigmoid) with domain-specific batch normalization, where the latent representations z and batch ID are provided as input, and a Sigmoid activation function. We used the Adam[75] optimizer with a 5e-4 weight decay and betas (0.9, 0.999, the exponential decay rate for the first and second moment parameters) to optimize the model under the learning rate 2e-4. We adopted mini-batch strategy to iteratively optimize the model, in each mini-batch, we randomly sampled data from all batches instead of from the same batch, and the mini-batch size for training input is 64. The maximum number of training iterations is 30,000 and an early stopping is triggered when there has been no improvement for 10 epochs. The hyper-parameters are chosen after a grid search. SCALEX is very robust with all of these hyper-parameters, all of the results in this manuscript are produced under the same parameters.

Domain-specific batch normalization (DSBN). Batch normalization (BN)[32] is a widely used training technique in deep neural networks to reduce internal covariate shifting. A BN layer whitens activations within a mini-batch of samples followed by scaling and shifting with learned affine parameters γ and β. For a mini-batch of samples: $\mathcal{B} = \{x_{1\ldots m}\}$;

$$\mu_{\mathcal{B}} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_{\mathcal{B}}^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2$$

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$$

$$y_i = \gamma \hat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i)$$

Where μ_B is the mini-batch mean, σ_B^2 is the mini-batch variance, $\hat{x}_i$ is the normalized output by μ_B and σ_B^2, y_i is the BN output by scaling and shifting $\hat{x}_i$ with parameters γ and β, and ϵ is a constant added to the mini-batch variance for numerical stability.

Domain specific batch normalization (DSBN)[31] is a combination of multiple sets of BN specific to each domain. DSBN learns domain-specific affine parameters γ_d and β_d for each domain, *d* is the domain label; here, domain represents different batches. In the neural network, DSBN serves like multi-channel BN and switches to the corresponding BN given the domain label *d*. The DSBN layer can be written as:

$$y_d = \gamma_d \hat{x}_d + \beta_d \equiv \mathrm{DSBN}_{\gamma_d,\beta_d}(x_d, d)$$

where *d* is the batch label, and γ_d and β_d are domain-specific affine parameters for domain *d*.

DSBN could capture the domain-specific information by estimating mini-batch statistics by learning affine parameters for each domain separately, thus enabling the network to learn the domain-invariant features.

Preprocessing for scRNA-seq. We downloaded gene expression matrices and preprocessed them using the following procedure: i). Cells with fewer than 600 genes and genes present in fewer than 3 cells were filtered out. ii). Total counts of each cell were normalized to 10,000. iii). Values of each gene were subjected to log transformation with an offset of 1. iv). The top 2,000 highly variable genes were identified. v). Values of each gene were normalized to the range of 0-1 within each

715 batch by the *MaxAbsScaler* function in the *scikit-learn* package in Python. The
716 processed matrix was used as input for the SCALEX model for downstream
717 differential gene expression analysis.

718 For the *human fetal atlas* dataset, we collected two batches (batch GSE156793,
719 which contains 4,062,980 cells by sciRNA-seq3, and batch GSE134355, which
720 contains 254,266 cells by Microwell-seq). We then selected the cells from the
721 common tissues (1,369,619 cells) for integration and computational efficiency
722 benchmarking (down-sampled from different data sizes including 10K, 50K, 250K,
723 1M, and 4M).

724 **Preprocessing for scATAC-seq.** We downloaded open chromatin profile matrices
725 (peaks or bins), merged them by peaks (or bins), and processed them using the
726 following procedure: i). The combined matrix was binarized and filter bins with fewer
727 than 3 cells. ii). The top 30,000 most variable peaks (or bins) were selected using the
728 *select_var_feature* function in the *EpiScanpy*[76] package. iii). Total counts of each cell
729 were normalized to the median of the total counts of all cells by using the
730 *normalize_total* function, with parameters *target_sum*="None" in the *Scanpy*[77]
731 package. iv). Values of each peak (or bin) were normalized to the range of 0-1 within
732 each batch by the *MaxAbsScaler* function in the *scikit-learn* package in Python. The
733 processed matrix was used as input for the SCALEX model.

734 **Preprocessing for cross-modality data (scRNA-seq and scATAC-seq).** We first
735 created a gene activity matrix by the *GeneActivity* function in the *Signac*[78] R package
736 to quantify the activity of each gene from scATAC-seq data. We then combined gene
737 activity score matrix with scRNA-seq data matrix as two individual "batches" for
738 integration. The subsequent preprocessing followed the same preprocessing used for
739 the scRNA-seq data (above).

Visualization. UMAP algorithm[46] was used for visualization. We applied the *neighbors* function from the Python package *Scanpy* with the parameters *n_neighbors*=30 and *metric*="Euclidean" for computing the neighbor graph, followed by *umap* function with *min_dist*=0.1 to visualize cells in a two-dimensional space.

Adjusted Rand Index. The Rand Index (RI) computes a similarity score between two clustering assignments by considering matched and unmatched assignment pairs, independent of the number of clusters. The Adjusted Rand Index (ARI) score is calculated by "adjust for chance" with RI as follows:

$$ARI = \frac{RI - Expected_RI}{\max(RI) - Expected_RI}$$

If given the contingency table, then ARI can also be represented by:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

The ARI score is 0 for random prediction and 1 for perfectly matching.

Normalized mutual information.

$$NMI = \frac{I(P;T)}{\sqrt{\mathrm{H(P)H(T)}}}$$

Where P and T are categorical distributions for the predicted and real clustering, I is the mutual entropy, and H is the Shannon entropy.

Silhouette score. We used the silhouette score to assess the separation of biological populations with the function *silhouette_score* in the *scikit-learn* package in Python. The silhouette score was computed by combining the average intra-cluster distance (a) and the average nearest-cluster (b) for each cell.

$$silhouette\ score = (b - a)/\max(a, b)$$

Here, we took UMAP embeddings as input to calculate silhouette score.

758 **Batch entropy mixing score.** Batch entropy mixing score (adapted from "entropy of
759 batch mixing"[19]) was used to access the regional mixing of cells from different
760 batches, with a high score suggesting that cells from different batches are well mixed
761 together.

762 The batch entropy mixing score was computed as follows:

763 (1) Calculated the proportion *Pi* of cell numbers in each batch to the total cell
764 numbers.
765 (2) Randomly chose 30 cells from all batches.
766 (3) Calculated the 30 nearest neighbors for each randomly chosen cell.
767 (4) The regional mixing entropies for each cell were defined as:

$$pi' = \frac{pi/Pi}{\sum_{i=1}^{n} pi/Pi}$$

$$\mathrm{E} = \sum_{i=0}^{n} pi' log(pi')$$

768 where *pi* is the proportion of cells from batch *i* in a given region, such that $\sum_{i=0}^{n} pi =$
769 1, *pi*' is a correction item to eliminate the deviation caused by the different cell
770 numbers in different batches. The total mixing entropy was then calculated as the sum
771 of the regional mixing entropies.

772 (5) Repeated (2)-(4) for 10 iterations with different randomly chosen cells and
773 calculated the average, E, as the final batch entropy mixing score.

774 Note that to mitigate the effect of misalignment of batch-specific cell-types, we
775 calculated the batch entropy mixing score only based on cells from cell-types that are
776 common in different batches.

Local inverse Simpsons Index (LISI). The LISI metric was proposed by Korsunsky *et al. 2019* [20] to assess batch and cell-type mixing. We calculated integration LISI (iLISI) and cell-type LISI (cLISI) values using the *compute_lisi* function in the *lisi* R package. UMAP embeddings, batch labels, and cell-type labels were used as input in calculation. Briefly,

$$LISI(x_i) = \frac{1}{\sum_{y \in Y} P(y|x_i)^2}$$

where $x_i \in \{x_1, x_2, \dots, x_N\}$ is the *i*-th cell's UMAP embeddings in the dataset of size *N*, and *Y* is the set of unique values with respect to the type of LISI we are computing (*i.e.*, *Y* is the values of "batch label" for calculating iLISI and the value of "cell-type label" for calculating cLISI). The probability $P(y|x_i)$ refers to the "relative abundance" of the covariate *y* within KNN (k-nearest neighborhood) of x_i. A Gaussian kernel-based distribution of neighborhoods was used and the perplexity was fixed to 30.

Over-correction score. We defined an over-correction score to assess the level of over-correction problem, based on calculating the percentage of cells with inconsistent cell-types in each cell's neighborhood. We calculated the over-correction score over all cells, and for each cell i we averaged the frequency of the k-nearest neighboring cells with distinct cell-types to the cell i (see the following equation).

$$over_correction\ score = 1 - \frac{1}{n * k} \sum_{i=1}^{n} \sum_{j=1}^{k} \boldsymbol{I}(cell_type_i, cell_type_j)$$

where n is the total cell number, k represents the k-nearest neighbors of each cell, the cell-type of the cell i is $cell_type_i$, the cell-type of the neighboring cell j is $cell_type_j$, and $\boldsymbol{I}$ is an indicative function defined as:

$$\boldsymbol{I}(cell_type_i, cell_type_j) = \begin{cases} 1 & if\ cell_type_i = cell_type_j \\ 0 & if\ cell_type_i \neq cell_type_j \end{cases}$$

797 Formally, the over-correction score is a negative index, i.e., the higher the
798 over-correction score, the more severe the extent of inaccurate mixing of cell-types.

799 **Comparison with other integration methods.** We compared SCALEX to nine other
800 batch effect removal methods (see below for specific details of each method). For
801 each dataset as input for all methods, we performed the same filtration, followed by
802 method-specific normalization, batch correction and visualization. Note that for visual
803 comparison, we also included the embeddings of the raw input data, wherein we
804 performed dimensionality reduction by Principal Component Analysis (PCA)[79]
805 followed by UMAP visualization to see the batch effects. No correction function was
806 used. All parameters were kept as default values.

807 Scanorama (v1.6). We performed the preprocessing pipelines as stated above (as the
808 same below), and used the *Scanpy* and *scanorama* Python packages for integration.
809 For the *highly_variable_genes* function, we set *flavor*="seurat", *batch_key*="batch",
810 and *n_top_genes*=2,000. After extracting highly variable genes, we divided the
811 datasets according to the batch labels and formed a new list of datasets as the input for
812 the *correct_scanpy* function. The integration matrix was kept for downstream analysis.
813 All other parameters were kept their default values.

814 BBKNN (v1.3.12). We used *Scanpy* and *bbknn* Python packages and followed the
815 suggested pipelines for integration. For the *highly_variable_genes* function, we set
816 *flavor*="seurat", *batch_key*="batch", and *n_top_genes*=2,000. After selecting cell
817 neighbors at the low-dimensional space from the PCA analysis, we performed the
818 *bbknn* function with *neighbors_within_batch*=5, *n_pcs*=20, and *trim*=0. All other
819 parameters were default.

820 scVI (scvi-tools, v0.11): We used the *scvi* Python package and followed the suggested
821 pipelines. Batch information was added to the VAE model by setting n_batch.

822 Seurat v3 (v3.2.3): We used the *Seurat* R package and followed the standard
823 integration workflow. We normalized different batches of a dataset separately. For the
824 *FindVariableFeatures* function, we set *selection.method*="vst" and *nfeatures*=2,000
825 to select 2,000 highly variable genes for each batch of a dataset. For the
826 *FindIntegrationAnchors* function, we set *k.filter*=100. All other parameters were kept
827 at default values. If the number of input cells in a dataset exceeded 50,000, we
828 employed the reciprocal PCA and reference-based integration to improve
829 computational efficiency.

830 Harmony (v1.0): We used the *harmony* R package. We created a *Seurat* object with
831 all cells and performed the standard workflow. After PCA, we used the *RunHarmony*
832 function for integration. All parameters were default.

833 Conos (v1.3.1): We used the *Conos* R package. For each batch of dataset, we used the
834 *basicSeuratProc* and *RunTSNE* functions for preprocessing. After that, we built a
835 joint graph using the *buildGraph* function with *k*=30 and *k.self*=5. All other
836 parameters were default.

837 MNN (FastMNN, v0.3.0): We used the *SeuratWrappers* R package. We created a
838 Seurat object with all cells and performed the standard workflow. Then we used the
839 *RunFastMNN* function with default parameters for integration.

840 Online iNMF and LIGER (LIGER, v1.0.0): We used the *rliger* R package. For the
841 online iNMF method, we used the online_iNMF function with *k*=20,
842 *miniBatch_size*=5,000 and *max.epochs*=5. For the LIGER method, we used the
843 *optimizeALS* function with *k*=20. All other parameters were the default values.
844 Different from other methods, online iNMF only loads one mini-batch from the whole

845 data in the HDF5 file format (converted from the orginal data format by the *rhdf5* R
846 package) for a memory-efficient implementation; accordingly, a file conversion issue
847 with the down-sampled *human fetal atlas* dataset of 4M data size prevented online
848 iNMF from calculating computational efficiency with the 4M.

849 **Cell-type annotation by clustering.** This type of annotation was used for *de novo*
850 annotation of a single-cell dataset. We used a Leiden clustering[80] method for cell
851 clustering (specifically employing the *leiden* function from the Python package
852 *Scanpy* with default parameters). Then for each cluster, we annotate its cell-type
853 based on: i) cell-type annotations of each cell in the original study, if available, or ii)
854 expression levels of canonical marker genes in each cell. A majority vote strategy was
855 used when needed. Similar to Ren et al. 2021, we also employed a hierarchical
856 annotation strategy, *i.e.*, we first clustered all cells in a dataset into several major
857 clusters, then for some big clusters, we further clustered them into minor clusters
858 respectively.

859 **Single cell projection.** We defined single cell projection as the operation to convert
860 high-dimensional single-cell data (*e.g.*, gene expression profiles in scRNA-seq or open
861 chromatin profiles in scATAC-seq) to low-dimensional representations in the common
862 SCALEX cell-embedding space using the trained encoder.

863 **Cell-type annotation by label transfer.** This type of annotation was used for
864 annotation of a new single-cell data batch using the annotations in a large single-cell
865 dataset as a reference, or for *post hoc* annotations of unknown cell population(s) in a
866 large dataset using new batches of data of known cell-types. Both scenarios require
867 "single cell projection" (see details below).

868 The basic idea of cell-type annotation by label transfer is based-on that the same
869 cell-types will occupy the same locations in the low-dimensional SCALEX
870 cell-embedding space, thus cell-type annotation in one data batch can be transferred to

another data batch, for the cells positioned at the same locations. Technically, we used the *KNeighborsClassifier* function from the *scikit-learn* package to train a prediction model, using the representations (in the low-dimensional cell-embedding space) of the single-cell data with known cell-type labels as input. We then used this model to make cell-type predictions for cells without annotations using their representations (in the low-dimensional cell-embedding space) as input. For comparison, label transfer for online iNMF follows the same procedures as SCALEX by predicting the cell-type based-on the projected locations.

Similarity matrix and confusion matrix. We used similarity matrix to evaluate the congruence of two different batches for the same cell-types in the common cell-embedding space. Technically, we merged all cells with the same cell-type label and calculated an average representation (in the low-dimensional cell-embedding space) for the cell-type. This was repeated for all cell-types. We then calculated the similarity matrix $S=[S_{ij}]$ for the cell-type similarities of the two batches, where S_{ij} is the Pearson correlation coefficient between the average representation of cell-type i in data_batch_1 and the average representation of cell-type j in data_batch_2.

We used the confusion matrix to evaluate the accuracy of cell-type annotations (prediction) when a gold-standard annotation is available, which is typical for "cell-type annotation by label transfer" (see above). In cell-type annotation by label transfer, we predict the cell-types for a single-cell data_batch_1, using the annotations in another data_batch_2. When data_batch_1 was already annotated with cell-types, we can calculate the confusion matrix $C=[C_{ij}]$ to compare the cell-type predictions with the existing cell-type annotations, where C_{ij} equals the percentage of cells known to be in cell-type i and predicted to be in cell-type j.

F1 score. F1 score is a weighted average of the precision and recall as a classification metric, which is computed by:

$$F1\ score = 2 * (\text{precision} * \text{recall})/(\text{precision} + \text{recall})$$

Generation of partially overlapping datasets. To simulate partially overlapping datasets from the *pancreas* dataset, we used the pancreas_celseq2 and pancreas_smartseq2 data batches, and worked with only six cell-types (alpha, beta, ductal, acinar, delta, gamma). For each simulated partially overlapping dataset, we randomly selected three to six cell-types from each batch, and counted the number of the common cell-types, which was used as the indicator for the overlapping level (whole integers, 0 to 6). We required the union of cell-types in the newly simulated partially overlapping dataset to cover all six cell-types.

For the *PBMC* dataset, we used both of the two data batches and worked with twelve cell-types (B, CD4 T, CD4 naive T, CD8 T, CD8 naive T, DC, HSC, Megakaryocyte, NK, monocyte-CD14, monocyte-FCGR3A, pDC). We used the same down-sampling strategy as for the *pancreas* dataset (above).

Analysis of changes in cell-type frequency across multiple conditions. To identify differences in cell-type frequency among the scRNA-seq data from the mild/moderate, severe, convalescent COVID-19 patients, as well as the healthy and influenza patient controls, we applied a Dirichlet-multinomial regression model. This model accounts for the constraint that the cell frequencies in a scRNA-seq data are not independent of each other. In detail, we normalized the regression coefficients to a standard normal distribution and calculated a z score, and then conducted significance testing based on the regression model generated by the *DirichReg* function in the R package *DirichletReg* (v 0.7).

Differential gene expression analysis and Gene Ontology term enrichment analysis. Differential gene expression analysis was performed on all expressed genes using the *rank_genes_groups* function with *method*="t-test" in the *Scanpy* package, for two certain cell-types in a COVID-19 single-cell atlas. A gene was considered differentially expressed when a log2-fold change was >1 in the two conditions in

comparison, and the Benjamini-Hochberg adjusted P-value was < 0.01. The top 200 highly expressed genes sorted by scores (implemented in *Scanpy*) of each cell-type were used as the input for GO analysis, and enriched GO terms were acquired for each group of cells of the "GO_Biological_Process_2018" dataset using the Python package *GSEApy*.

Inflammatory and cytokine score analysis. We defined the inflammatory score and the cytokine score for each cell following Ren et al. 2021, based on the expression of a defined collection of cytokine genes and inflammatory-response-related genes, which were from Liberzon et al. 2015 (Supplementary Table 1). We then calculated the cytokine and inflammatory scores from the raw gene expression profile using the *score_genes* function implemented in the *Scanpy*.

Ablation studies using test-variants of SCLAEX

We evaluate the contributions of each design element by the performance of different ablation variant models on integration and projection tasks (Fig. S18, Fig. S19).

Baseline/full SCLAEX. The baseline or the full SCALEX model is described in the paper, including an encoder without batch label, mini-batch sampling from all batches with Batch Normalization, a decoder with DSBN, and beta=0.5. All other variant models are compared to this baseline model.

A variant of encoder with batch label. The baseline model with an encoder with batch label as input.

A variant of sampling by batch without Batch Normalization. The baseline model without the Batch Normalization layer in the encoder as well as each mini-batch is sampled by batch instead of from all batches.

A variant of decoder without DSBN. The baseline model without the DSBN layer in the decoder.

Software availability.

SCALEX is available at https://github.com/jsxlei/SCALEX.

Data availability.

All data analyzed in this study are publicly available; the data sources are detailed in Supplementary Table 2.

Acknowledgements

We thank Jianbin Wang, Jin Gu and Fuchou Tang for helpful comments and advice. This work is supported by the State Key Research Development Program of China (Grant No. 2018YFA0107603 and 2019YFA0110002); the National Natural Science Foundation of China (Grants No. 91740204, 91940306, and 31761163007); the Beijing Advanced Innovation Center for Structural Biology; and the Tsinghua-Peking Joint Center for Life Sciences.

Author contributions

Q.C.Z. conceived and supervised the project. L.X. designed and implemented the SCALEX model. L.X. and K.T. validated the SCALEX model. L.X., K.T., and Y.L. analyzed the results. L.X. and Q.C.Z. wrote the manuscript, with inputs from all the authors.

Competing interests

The authors declare no competing interests.

971 **Figures**

972 **Fig. 1 | The design and performance of SCALEX for single-cell data integration.**
973 **a**, SCALEX models the global structure of single-cell data using a variational
974 autoencoder (VAE) framework. **b**, UMAP embeddings of the *PBMC* dataset before
975 and after integration by indicated methods. Cells are colored by batch (left) and
976 cell-type (right). Misalignments are highlighted with red circles. **c**, Scatter plot
977 comparing SCALEX and the other state-of-the-art single-cell data integration tools in
978 terms of the ARI score (y-axis) and the NMI score (x-axis), based on the Leiden
979 clustering results in the latent space across the indicated benchmark datasets. **d**,
980 UMAP embeddings of the SCALEX integration of the Human Fetal Atlas dataset
981 after integration by SCALEX, colored by batch and cell-type. **e**, Comparison of
982 computation efficiency based on datasets of different sizes sampled from the whole
983 Human Fetal Atlas dataset) including runtime (left) and memory usage (right). Online
984 iNMF was not successfully tested on 4M data due to a HDF5 file conversion issue for
985 large data (Methods). **f**, UMAP embeddings of the mouse brain scATAC-seq dataset
986 before (left) and after integration (middle, right); colored by data batch or cell-types. **g**,
987 UMAP embeddings of the PBMC scRNA-seq and scATAC-seq cross-modality
988 dataset before (left) and after SCALEX integration (middle, right); colored by batch
989 or cell-type.

990 **Fig. 2 | Comparison of integration performance over partially overlapping**
991 **datasets by different methods. a**, UMAP embeddings of the integration results by
992 the indicated methods based on the *liver* dataset. **b**, Over-correction score of different
993 methods based on the indicated benchmark datasets. **c**, Over-correction score of the
994 indicated methods based on the simulated datasets, with decreased numbers of
995 common cell-types (obtained by down-sampling the *pancreas* and *PBMC* dataset). **d**,
996 UMAP embeddings of the integration results by the indicated methods based on the
997 simulated *panceas* datasets with different numbers of specified common cell-types.

998 Over-corrections are highlighted with red circles and error-splittings with purple
999 circles, respectively.

1000 **Fig. 3 | Projecting heterogenous data into a common cell-embedding space. a**,
1001 UMAP embeddings of the *pancreas* dataset after integration by SCALEX, colored by
1002 cell-type and by batch. **b**, UMAP embeddings of the common cell space obtained by
1003 using SCALEX to project three additional indicated *pancreas* data batches onto the
1004 *pancreas* dataset. Cells are colored by cell-type with light gray shadows representing
1005 the original *pancreas* dataset. **c**, Confusion matrix between ground truth cell-types
1006 and those annotated by different methods. ARI, NMI and F1 scores (top) measure the
1007 annotation accuracy. **d**, UMAP embeddings of the common space obtained by using
1008 SCALEX to project the two projected melanoma data batches onto the *PBMC* dataset,
1009 colored by cell-types with light gray shadows represent the original *PBMC* dataset. **e**,
1010 UMAP embeddings of the common cell space that includes the original *PBMC*
1011 dataset and the two projected melanoma data batches. **f**, Annotating an
1012 uncharacterized small cell population in the *pancreas* dataset by projection of the
1013 bronchial epithelium data batches into the pancreas cell space. Only the
1014 uncharacterized cells in the *pancreas* dataset (left) and the SLC16A7+ epithelial cells
1015 in the bronchial epithelium data batches (right) are colored. **g**, Heatmap showing the
1016 normalized expression of the top-10 ranking specific genes for the uncharacterized
1017 cell population in different cell-types.

1018 **Fig. 4 | Construction of an expandable mouse single-cell atlas. a,** Datasets acquired
1019 using different technologies (Smart-seq2, 10X, and Microwell-seq) covering various
1020 tissues used for construction of the mouse atlas. **b**, UMAP embeddings of the Mouse
1021 Atlas dataset colored by batch and tissue. **c**, UMAP embeddings of the Mouse Atlas
1022 after SCALEX integration, colored by cell-type. **d**, Scatter plot showing a quantitative
1023 comparison of the ARI score (y-axis) and the NMI score (x-axis) based-on the Leiden
1024 clustering results on the latent space based on the Human Atlas, Mouse Atlas, and

1025 Human Fetal Atlas datasets. **e**, Comparison of over-correction score by the indicated
1026 methods based on the Human Atlas, Mouse Atlas, and Human Fetal Atlas datasets. **f**,
1027 UMAP embeddings of the common cell space obtained by using SCALEX to project
1028 the two *Tabula Muris Senis* data batches and two mouse tissues (lung and kidney)
1029 data batches onto the Mouse Atlas dataset. Cells are colored by cell-type with light
1030 gray shadows representing the original Mouse Atlas dataset. **g**, Confusion matrix of
1031 the cell-type annotations by SCALEX and those in the original studies. Color bar
1032 represents the percentage of cells in confusion matrix C_{ij} known to be cell-type i and
1033 predicted to be cell-type j.

1034 **Fig. 5 | Online integration of COVID-19 PBMC Atlas. a**, The COVID-19 PBMC
1035 Atlas dataset composition, including healthy controls and influenza patients, as well
1036 as mild/moderate, severe, and convalescent COVID-19 patients. **b**,**c** UMAP
1037 embeddings of the COVID-19 PBMC Atlas after SCLAEX integration colored by
1038 cell-type (**b**), and by batch (**c**). **d**, UMAP embeddings of the COVID-19 PBMC Atlas
1039 separated by disease state. Cells are colored by cell-type with light gray shadows
1040 representing the other disease state cells. **e**, Stacked violin plot of
1041 differentially-expressed ISGs among CD14 monocytes across disease states. **f**, GO
1042 terms enriched in the differentially-expressed genes for CD14-IL1B-Mono and
1043 CD14-ISG15-Mono cells. **g**, Cell-type frequency across healthy and influenza
1044 controls, and among mild/moderate, severe, and convalescent COVID-19 patients.
1045 Dirichlet-multinomial regression was used for pairwise comparisons, ***$p<0.001$,
1046 **$p<0.01$, *$p<0.05$. **h**, UMAP embeddings of the common cell space obtained by
1047 using SCALEX to project the SC4 Atlas (Single Cell Consortium for COVID-19 in
1048 China). Cells are colored by cell-type from label transfer based-on the locations in the
1049 COVID-19 PBMC Atla*s* dataset (left) and cell clusters in original SC4 study (middle),
1050 and Unified UMAP embeddings combining the SCALEX COVID-19 PBMC Atlas
1051 and the SC4 Atlas (right). **i**, Similarity matrix of meta-cell representations for
1052 cell-types between the SCALEX COVID-19 PBMC Atlas and SC4 in the common

1053 cell-embedding space after SCALEX projection. Color bar represents the Pearson
1054 correlation coefficient between the average meta-cell representation of two cell-types
1055 from a respective data batch. **j**, UMAP embeddings of the SCALEX COVID-19
1056 PBMC Atlas colored by the cytokine score and the inflammatory score. **k**, GO terms
1057 enriched in the differentially-expressed genes for TUBA8-Mega and IGKC-Mega
1058 cells.

1059 **Supplementary figures**

1060 **Fig. S1 | Three design elements to learn a generalized encoder. a**, Assessing
1061 encoders with or without batch labels. **b**, Assessing decoders with or without DSBN.
1062 **c**, Assessing mini-batch sampling by batch without Batch Normalization or sampling
1063 from all batches with Batch Normalization. **d**, Detailed architecture of SCALEX and
1064 information flow.

1065 **Fig. S2 | Comparisons of integration performance of indicated methods across**
1066 **the indicated benchmark datasets.** UMAP embeddings of the indicated methods
1067 across the indicated benchmark datasets colored by batches and cell-type.
1068 Misalignments are highlighted with red circles.

1069 **Fig. S3 | Comparisons of clustering results of indicated methods across the**
1070 **indicated benchmark datasets.** UMAP embeddings of clustering results of the
1071 indicated benchmark datasets after integration by indicated methods; colored by
1072 Leiden (left) and Louvain (right).

1073 **Fig. S4 | Comparisons of integration performance by quantification metrics.**
1074 Scatter plots showing the ARI and NMI scores based-on the Leiden and Louvain
1075 clustering results, the Silhouette and batch entropy mixing scores, and the cLISI/
1076 iLISI scores.

1077 **Fig. S5 | Comparisons of integration performance of indicated methods based on**
1078 **Human Fetal Atlas dataset. a**, UMAP embeddings of the Human Fetal Atlas before
1079 integration. **b,** Dotplot of canonical marker genes for each cell-type. Dot color
1080 represents average expression level, and dot size represents the proportion of cells in
1081 the group expressing the marker. **c**, UMAP embeddings of the indicated methods
1082 (only SCALEX, BBKNN, Scanorama, and scVI are scalable to the Human Fetal Atlas
1083 dataset). Cells are colored by batch (left), cell-type (middle), and Leiden clustering
1084 (right). **d**, Scatter plot showing the ARI and NMI scores based-on the Leiden
1085 clustering results, the Silhouette and batch entropy mixing scores, and the iLISI and
1086 1/cLISI scores. Note that online iNMF was not successfully tested on the Human
1087 Fetal Atlas due to a HDF5 file conversion issue for large data.

1088 **Fig. S6 | Comparisons of integration performance of indicated methods based on**
1089 **scATAC-seq dataset and other modality datasets. a**, UMAP embeddings of the
1090 mouse brain scATAC-seq dataset before and after integration by indicated methods;
1091 colored by batch (left) and cell-type (right). **b**, Cell-type-specific peaks for the mouse
1092 brain scATAC-seq dataset. **c**, Scatter plot showing the ARI and NMI scores based-on
1093 the Leiden and Louvain clustering results, the Silhouette and batch entropy mixing
1094 scores, and the iLISI and 1/cLISI scores. **d**,**e**, UMAP embedding before and after
1095 SCALEX integration over CITE-seq (which measures abundance of both RNA and
1096 protein in single cells) spleen lymph dataset and cross-modality mouse brain dataset
1097 between spatial transcriptome (MERFISH) dataset and RNA-seq (10X) dataset.

1098 **Fig. S7 | Comparisons of integration performance of indicated methods based on**
1099 **cross-modality dataset. a**, UMAP embeddings of the PBMC cross-modality dataset
1100 by indicated methods. Cells are colored by batch or cell-type. **b**, Scatter plot showing
1101 a quantitative comparison of the silhouette score (y-axis) and the batch entropy
1102 mixing score (x-axis) (left), ARI (y-axis), and NMI (x-axis) based-on Leiden

1103 clustering (middle) and Louvain clustering (right), the Silhouette and batch entropy
1104 mixing scores, and the iLISI and 1/cLISI scores for the PBMC cross-modality dataset.

1105 **Fig. S8 | Canonical marker genes of different cell-types and UMAP embeddings**
1106 **of the *liver* dataset. a,** UMAP embeddings of the *liver* dataset, colored by batch (left)
1107 and cell-type (right) after SCALEX integration. **b,** Dotplot of canonical marker genes
1108 for each cell-type. Dot color represents average expression level, and dot size
1109 represents the proportion of cells in the group expressing the marker. **c**, Normalized
1110 marker gene expression on the UMAP embeddings of the five hepatocyte subtypes.
1111 Color bar represents the expression level.

1112 **Fig. S9 | Comparisons of integration performance based on partially overlapping**
1113 **simulated *pancreas* dataset. a,** Partially overlapping datasets were generated by
1114 down-sampling the *pancreas* dataset, consisted of common cell-types with a
1115 decreased overlapping number (ranging from 0 to 6). Integration results for SCALEX,
1116 Seurat, Harmony, and online iNMF are shown in the UMAP embeddings colored by
1117 batches (left) and cell-types (right) respectively (overlapping number decreases from
1118 6 to 0). Misalignments are highlighted with red circles. **b**, Scatter plot showing a
1119 quantitative comparison of the indicated methods in terms of the ARI score (y-axis)
1120 and the NMI score (x-axis), based on the Leiden (top) and Louvain (bottom)
1121 clustering results in the latent space based on simulated *pancreas* datasets.

1122 **Fig. S10 | Comparisons of integration performance based on partially**
1123 **overlapping simulated *PBMC* dataset. a**, Partially overlapping datasets were
1124 generated by down-sampling the *PBMC* dataset, consisted of common cell-types with
1125 a decreased overlapping number (ranging from 0 to 6). Integration results for
1126 SCALEX, Seurat, Harmony, and online iNMF are shown in the UMAP embeddings
1127 colored by batches (left) and cell-types (right) respectively (overlapping number
1128 decreases from 6 to 0). Misalignments are highlighted with red circles. **b**, Scatter plot

1129 showing a quantitative comparison of the indicated methods in terms of the ARI score
1130 (y-axis) and the NMI score (x-axis), based on the Leiden (top) and Louvain (bottom)
1131 clustering results in the latent space based on simulated *PBMC* datasets.

1132 **Fig. S11 | Projection of three additional pancreas data batches onto the *pancreas***
1133 **dataset. a**, UMAP embeddings of the *pancreas* dataset and the three additional
1134 pancreas data batches and the bronchial epithelium data batches (data from three
1135 donors) before integration. Cells are colored by batch. **b**, Dot plot of canonical
1136 markers of cell-types of reference *pancreas* dataset; dot color represents average
1137 expression level, and dot size represents the proportion of cells in the group
1138 expressing the marker. **c,** UMAP embeddings of the common cell space obtained by
1139 using online iNMF (top) and scVI (bottom) to project the three additional indicated
1140 pancreas data batches onto the *pancreas* dataset. Cells are colored by cell-type with
1141 light gray shadows representing the original *pancreas* dataset. Misalignments are
1142 highlighted with red circles. Note that here for convenient comparison of the
1143 projected data with the existing data, we showed the UMAP embeddings of the
1144 *pancreas* dataset combined with the projected data, which are visualized differently
1145 from the UMAP embeddings of the *pancreas* dataset alone in Fig. S2 as UMAP
1146 visualizations change with embedding data.

1147 **Fig. S12 | Projection of two melanoma datasets onto the *PBMC* dataset. a**, UMAP
1148 embeddings of the *PBMC* dataset and the two additional melanoma data batches
1149 before integration. Cells are colored by batch. **b**, Dot plot of canonical markers of
1150 cell-types of each batch; dot color represents average expression level, while dot size
1151 represents the proportion of cells in the group expressing the marker. **c**, UMAP
1152 embeddings of the common cell space obtained by using online iNMF (top) and scVI
1153 (bottom) to project the two additional indicated melanoma data batches onto the
1154 *PBMC* dataset. Cells are colored by cell-type with light gray shadows representing the
1155 original *PBMC* dataset. Misalignments are highlighted with red circles. Note that here

1156 for convenient comparison of the projected data with the existing data, we showed the
1157 UMAP embeddings of the *PBMC* dataset combined with the projected data, which are
1158 visualized differently from the UMAP embeddings of the *PBMC* dataset alone in Fig.
1159 S2 as UMAP visualizations change with embedding data.

1160 **Fig. S13 | Comparisons of integration across Atlas-level datasets. a**, UMAP
1161 embeddings of the indicated methods across the indicated datasets. Results are blank
1162 for those methods not scalable to this data size. **b**, Scatter plots showing the
1163 comparisons of indicated methods in terms of the ARI and NMI scores based-on the
1164 Louvain clustering results, the Silhouette and batch entropy mixing scores, and the
1165 cLISI/ iLISI scores based on the indicated Atlas-level datasets.

1166 **Fig. S14 | The SCALEX Mouse Atlas. a**, UMAP embeddings of the Mouse Atlas
1167 dataset before integration, colored by batch. **b**, UMAP embeddings of three mouse
1168 atlas data batches (Smart-seq2, 10X, and Microwell-seq) after integration, colored by
1169 cell-type; the light gray shadows represent the original Mouse Atlas dataset. **c**,
1170 Dotplot of the top 5 cell-type-specific genes for each cell-type in the Mouse Atlas
1171 dataset. Dot color represents average expression level, while dot size represents the
1172 proportion of cells in the group expressing the marker. Note that here for convenient
1173 comparison of the projected data with the existing data, we showed the UMAP
1174 embeddings of the Mouse Atlas combined with the projected data, which are
1175 visualized differently from the UMAP embeddings of the Mouse Atlas alone in Fig.
1176 S13 as UMAP visualizations change with embedding data.

1177 **Fig. S15 | The SCALEX Human Atlas. a,** The Human Atlas dataset acquired using
1178 different technologies (Smart-seq2, 10X, and Microwell-seq) covering various tissues
1179 used for construction of the human atlas. **b-c**, UMAP embeddings of the Human Atlas
1180 dataset colored by batch and cell-type, before (**b**) and after integration (**c**). **d**,
1181 Similarity matrix of meta-cell representations for cell-types in the two data batches in

1182 the common cell-embedding space after SCALEX integration between two batches.
1183 Color bar represents the Pearson correlation coefficient between the average meta-cell
1184 representation of two cell-types from a respective data batch. **e**, UMAP embeddings
1185 of the common cell space obtained by using SCALEX to project the two additional
1186 indicated human skin data batches onto the Human Atlas dataset. Cells are colored by
1187 cell-type with light gray shadows representing the original Human Atlas dataset. **f**,
1188 Confusion matrix of the cell-type annotations by SCALEX and those in the original
1189 study. Color bar represents the percentage of cells in confusion matrix C_{ij} known to be
1190 in cell-type i and predicted to be in cell-type j. Note that here for convenient
1191 comparison of the projected data with the existing data, we showed the UMAP
1192 embeddings of the Human Atlas combined with the projected data, which are
1193 visualized differently from the UMAP embeddings of Human Atlas alone in Fig. S13
1194 as UMAP visualizations change with embedding data.

1195 **Fig. S16 | COVID-19 immune landscape. a**, UMAP embeddings of the COVID-19
1196 PBMC Atlas dataset before and after integration by indicated methods. Cells are
1197 colored by batch (left), cell-type (middle) and Leiden clustering results (right). **b**,
1198 Scatter plots showing the comparisons of indicated methods in terms of the ARI and
1199 NMI scores based-on the Leiden and Louvain clustering results, the Silhouette and
1200 batch entropy mixing scores, and the cLISI/ iLISI scores. Bar plot showing the
1201 comparisons of indicated methods in terms of the over-correction scores based on the
1202 COVID-19 PBMC Atlas dataset. **c**, Dotplot of canonical marker genes for each
1203 cell-type. Dot color represents average expression level, while dot size represents the
1204 proportion of cells in the group expressing the marker. **d**, UMAP embeddings of the
1205 COVID-19 PBMC atlas in individual batches after SCALEX integration, colored by
1206 cell-type; the light gray shadows represent the other batches of COVID-19 PBMC
1207 atlas. Note that here for convenient comparison of the projected data with the existing
1208 data, we showed the UMAP embeddings of the SCALEX COVID-19 PBMC Atlas
1209 combined with the projected data, which are visualized differently from the UMAP

1210 embeddings of SCALEX COVID-19 PBMC Atlas alone as UMAP visualizations
1211 change with embedding data.

1212 **Fig. S17 | COVID-19 heterogeneous dysfunctional immune response. a**,
1213 Frequency of cell distributions across healthy people and influenza patient controls,
1214 and among mild/moderate, severe, and convalescent COVID-19 patients.
1215 Dirichlet-multinomial regression was used for pairwise comparisons, ***$p<0.001$,
1216 **$p<0.01$, *$p<0.05$. **b**, Stacked violin plot of differentially-expressed genes between
1217 PNPLA2-Immature_Neutrophil and NCF1-Immature_Neutrophil cells. **c**, GO terms
1218 enriched in the differentially-expressed genes for PNPLA2-Immature_Neutrophil and
1219 NCF1-Immature_Neutrophil cells. **d**, Stacked violin plot of differentially-expressed
1220 genes between PRDM1-Plasma and MZB1-Plasma. **e**, GO terms enriched in the
1221 differentially-expressed genes for PRDM1-Plasma and MZB1-Plasma cells.

1222 **Fig. S18 | Projection of the SC4 Atlas onto the SCLAEX COVID-19 PBMC Atlas.**
1223 **a-b**, UMAP embeddings of the SC4 Atlas before integration (**a**) and after projection
1224 onto the SCALEX COVID-19 PBMC Atlas (**b**). **c**, Separate UMAP embeddings of
1225 each SC4 data batch, after being projected onto the SCALEX COVID-19 PBMC
1226 space, colored by cell-type. light gray shadows represent the COVID-19 PBMC Atlas.
1227 **d**, UMAP embeddings of the TUBA8-Mega and IGKC-Mega cells. **e**, UMAP
1228 embeddings of the differentially-expressed genes of TUBA8-Mega and IGKC-Mega
1229 cells. **f**, UMAP embeddings of the common cell space obtained by using scVI to
1230 project the *SC4 Atlas* onto the COVID-19 PBMC Atlas dataset. Cells are colored by
1231 cell-type with light gray shadows representing the original COVID-19 PBMC Atlas
1232 dataset. Misalignments are highlighted with red circles.

1233 **Fig. S19 | Ablation studies of different SCALE architectures for single-cell data**
1234 **integration. a**, UMAP embeddings after integration by SCALEX and other SCALEX
1235 test-variants. **b**, Scatter plot showing a quantitative comparison of the performance of

1236 the full and different test-variants of SCALEX using the ARI score (y-axis) and the
1237 NMI score (x-axis), the Silhouette score (y-axis) and the batch entropy mixing score
1238 (x-axis), and the 1/cLISI score (y-axis) and the iLISI score (x-axis), across the
1239 indicated benchmark datasets.

1240 **Fig. S20 | Ablations studies of different SCALEX architectures for single-cell**
1241 **data projection.** Left: UMAP embeddings of three projected pancreas data batches
1242 projected onto the pancreas space using different SCALEX architectures, colored by
1243 cell-type; light gray shadows represent the original *pancreas* dataset. Right: UMAP
1244 embeddings of the two projected melanoma data batches projected onto the PBMC
1245 space using different SCALEX architectures, colored by cell-type; light gray shadows
1246 represent the original *PBMC* dataset.

1247 **Fig. S21 | Comparison of SCALEX with beta=0.5 and beta=1 across the**
1248 **indicated benchmark datasets. a**, UMAP embeddings of SCALEX (beta=0.5) and
1249 SCALEX (beta=1) across the indicated benchmark datasets; colored by batch (left)
1250 and cell-type (right). **b**, Scatter plot showing a quantitative comparison of the
1251 performance of SCALEX (beta=0.5) and SCALEX (beta=1) using the ARI score
1252 (y-axis) and the NMI score (x-axis), the Silhouette score (y-axis) and the batch
1253 entropy mixing score (x-axis), and the 1/cLISI score (y-axis) and the iLISI score
1254 (x-axis), across the indicated benchmark datasets.

Fig 1

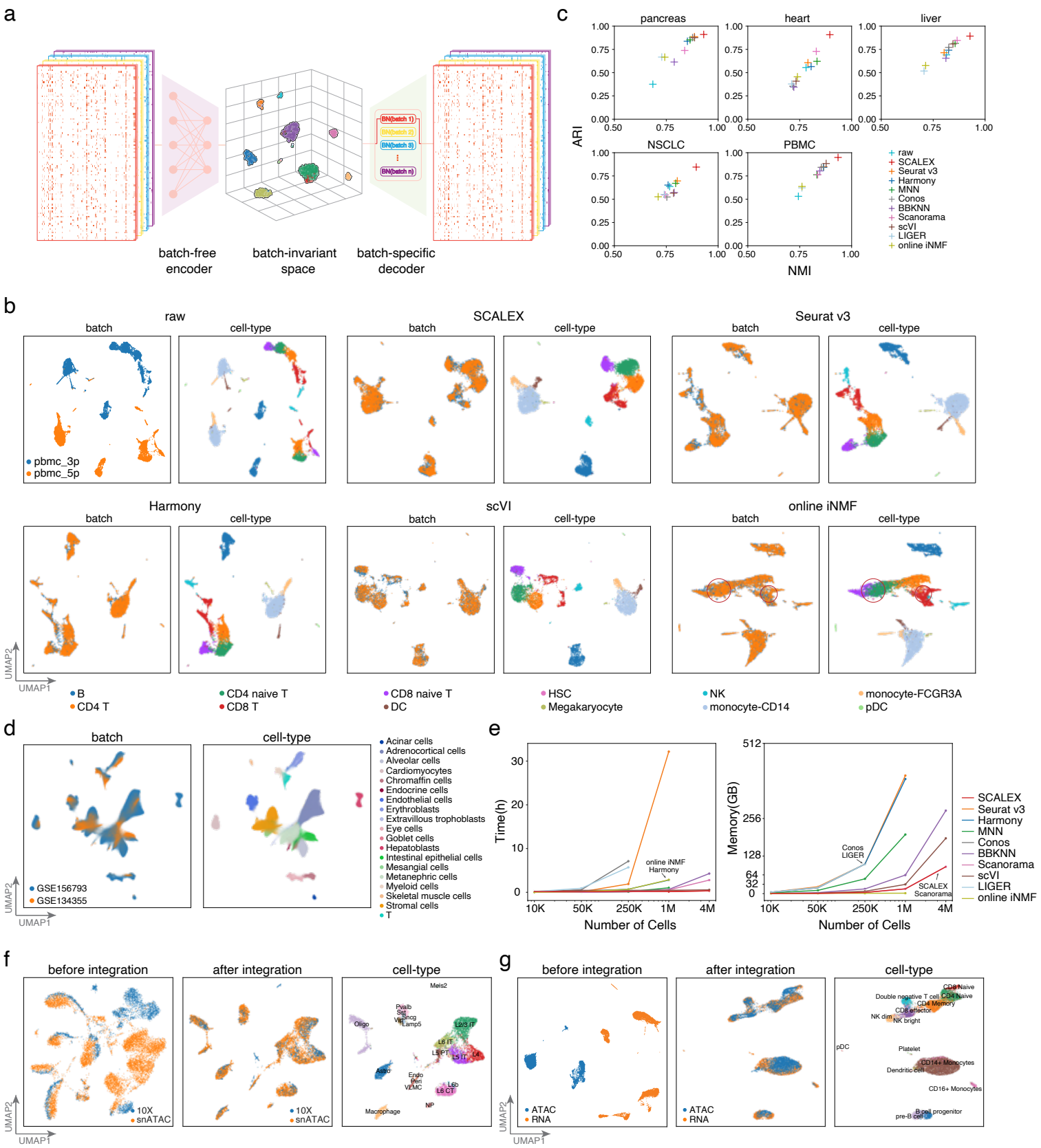

a
batch-free encoder
batch-invariant space
batch-specific decoder
BN(batch 1)
BN(batch 2)
BN(batch 3)
BN(batch n)
c
pancreas
heart
liver
NSCLC
PBMC
ARI
NMI
raw
SCALEX
Seurat v3
Harmony
MNN
Conos
BBKNN
Scanorama
scVI
LIGER
online iNMF
b
batch
cell-type
pbmc_3p
pbmc_5p
UMAP2
UMAP1
B
CD4 T
CD4 naive T
CD8 T
CD8 naive T
DC
HSC
Megakaryocyte
NK
monocyte-CD14
monocyte-FCGR3A
pDC
d
GSE156793
GSE134355
Acinar cells
Adrenocortical cells
Alveolar cells
Cardiomyocytes
Chromaffin cells
Endocrine cells
Endothelial cells
Erythroblasts
Extravillous trophoblasts
Eye cells
Goblet cells
Hepatoblasts
Intestinal epithelial cells
Mesangial cells
Metanephric cells
Myeloid cells
Skeletal muscle cells
Stromal cells
T
e
Time(h)
Memory(GB)
Number of Cells
online iNMF
Harmony
Conos
LIGER
SCALEX
Scanorama
f
before integration
after integration
10X
snATAC
Meis2
Macrophage
NP
g
ATAC
RNA
Platelet
CD14+ Monocytes
Dendritic cell
CD16+ Monocytes
pDC
B cell progenitor
pre-B cell

Fig 2

a
raw
SCALEX
Seurat v3
Harmony
scVI
online iNMF
batch
cell-type
GSE115469
GSE124395
UMAP2
UMAP1
B
NK-GNLY_low
cholangiocytes
erythroid
hepatocyte-CYP2A13
hepatocyte-SCD
kupffer
stellate
NK-GNLY_high
T
endothelial
hepatocyte-CXCL1
hepatocyte-GSTA1
hepatocyte-TAT-AS1
mature B
b
Over-correction Score
pancreas
PBMC
liver
NSCLC
heart
SCALEX
Seurat v3
Harmony
MNN
Conos
BBKNN
Scanorama
scVI
LIGER
online_iNMF
c
simulated from pancreas
simulated from PBMC
Over-correction Score
overlapped cell-type number / total cell-type number
d
overlap_0
overlap_1
overlap_2
overlap_4
pancreas_celseq
pancreas_smartseq2
acinar
alpha
beta
delta
ductal
gamma

Fig 3

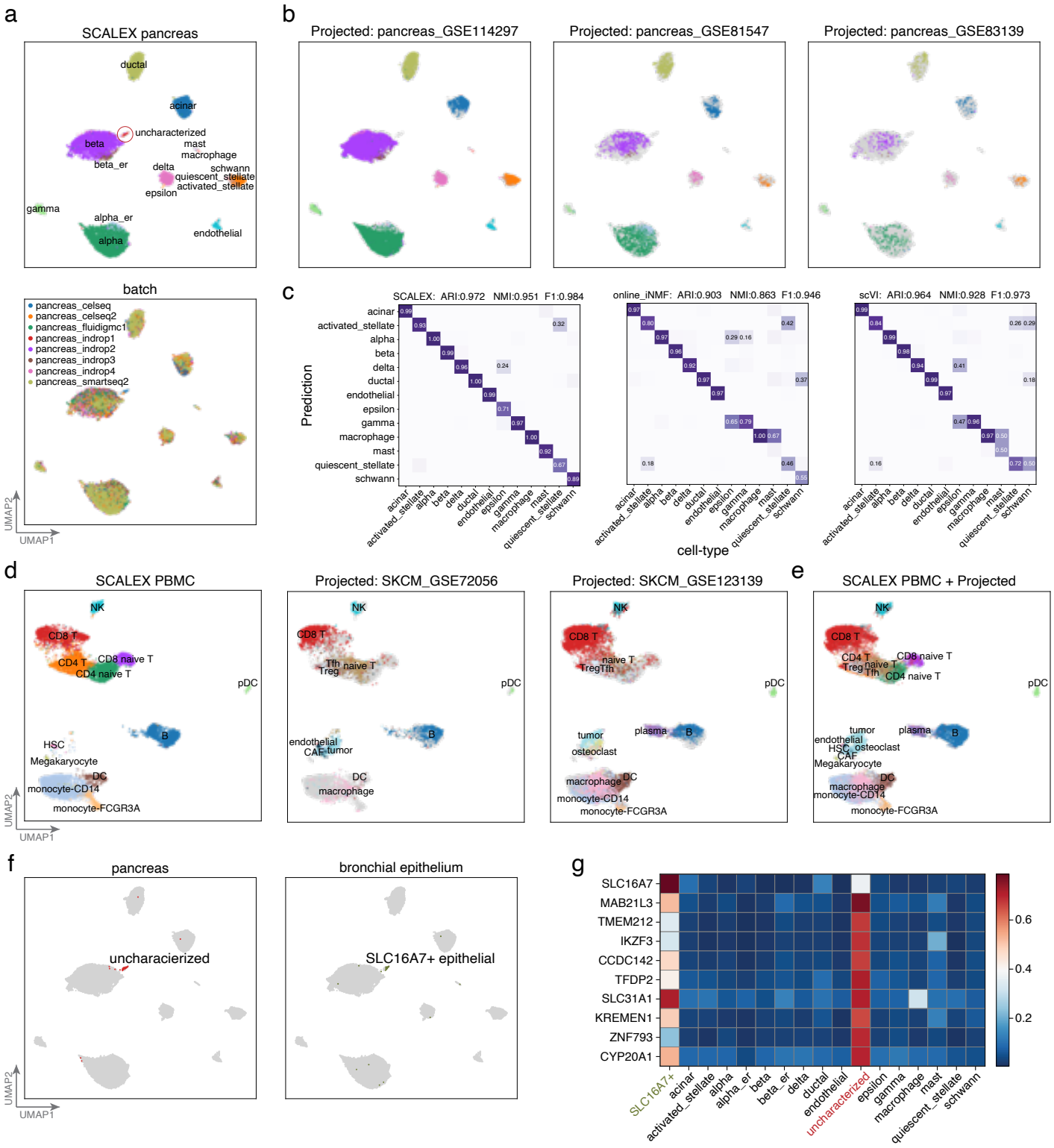

a
SCALEX pancreas
ductal
acinar
uncharacterized
beta
mast
macrophage
beta_er
delta
schwann
quiescent_stellate
activated_stellate
epsilon
gamma
alpha_er
endothelial
alpha
batch
pancreas_celseq
pancreas_celseq2
pancreas_fluidigmc1
pancreas_indrop1
pancreas_indrop2
pancreas_indrop3
pancreas_indrop4
pancreas_smartseq2
UMAP2
UMAP1
b
Projected: pancreas_GSE114297
Projected: pancreas_GSE81547
Projected: pancreas_GSE83139
c
SCALEX: ARI:0.972 NMI:0.951 F1:0.984
online_iNMF: ARI:0.903 NMI:0.863 F1:0.946
scVI: ARI:0.964 NMI:0.928 F1:0.973
Prediction
cell-type
d
SCALEX PBMC
NK
CD8 T
CD4 T
CD8 naive T
CD4 naive T
pDC
B
HSC
Megakaryocyte
DC
monocyte-CD14
monocyte-FCGR3A
Projected: SKCM_GSE72056
Tfh
Treg
naive T
endothelial
CAF
tumor
macrophage
Projected: SKCM_GSE123139
TregTfh
plasma
osteoclast
e
SCALEX PBMC + Projected
f
pancreas
uncharacierized
bronchial epithelium
SLC16A7+ epithelial
g
SLC16A7
MAB21L3
TMEM212
IKZF3
CCDC142
TFDP2
SLC31A1
KREMEN1
ZNF793
CYP20A1
SLC16A7+
acinar
activated_stellate
alpha
alpha_er
beta
beta_er
delta
ductal
endothelial
uncharacterized
epsilon
gamma
macrophage
mast
quiescent_stellate
schwann
0.6
0.4
0.2

Fig 4

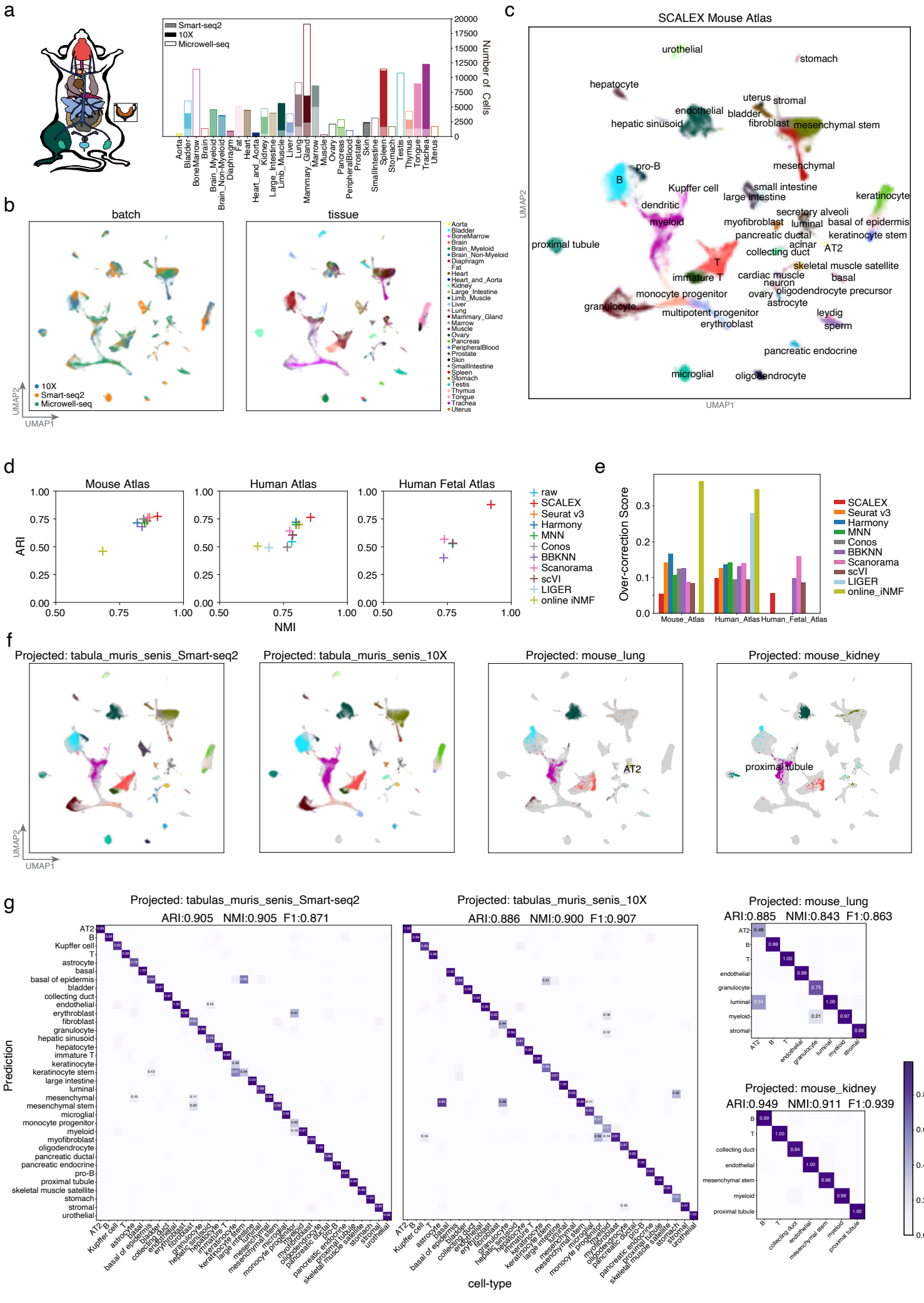
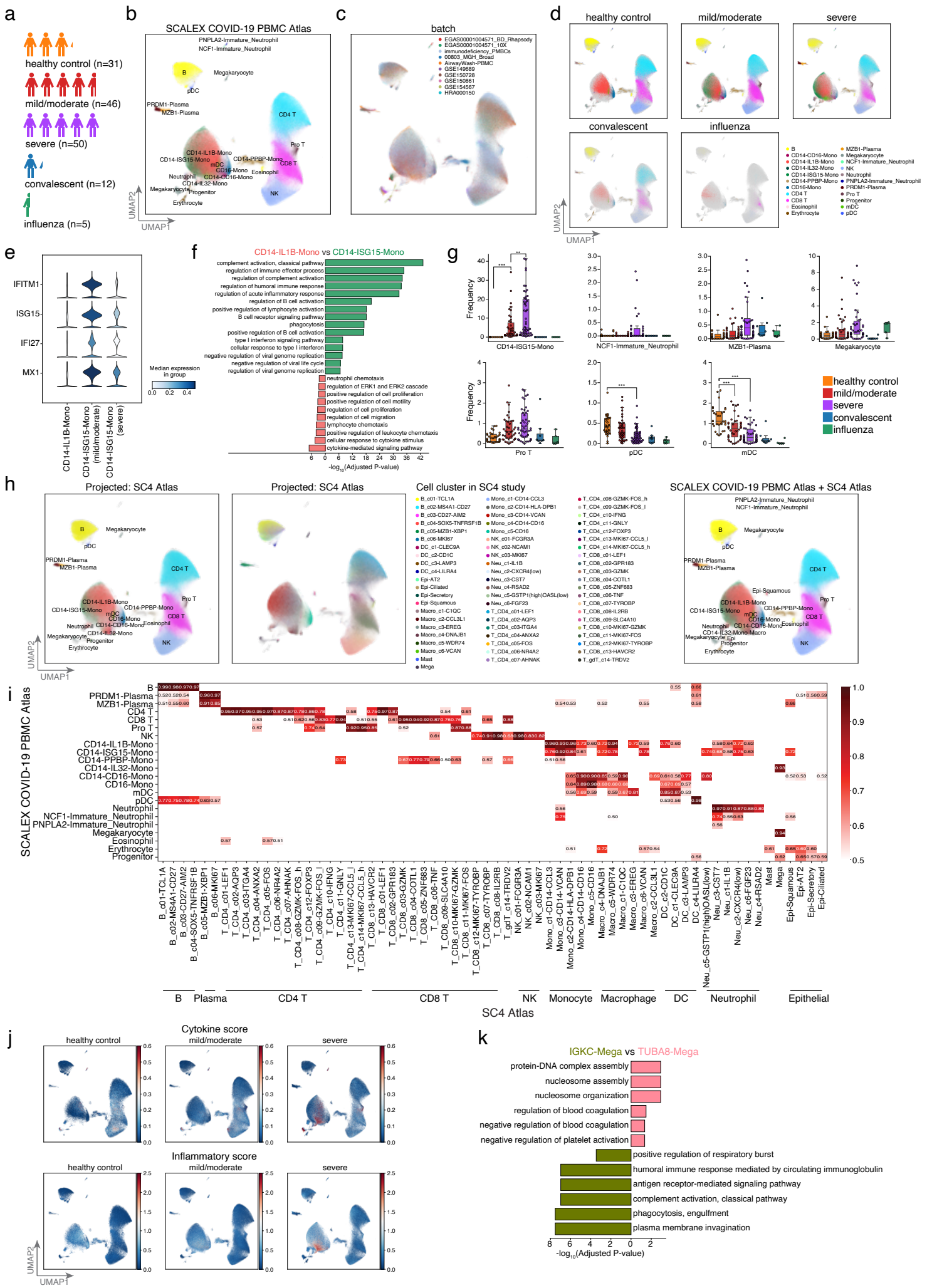
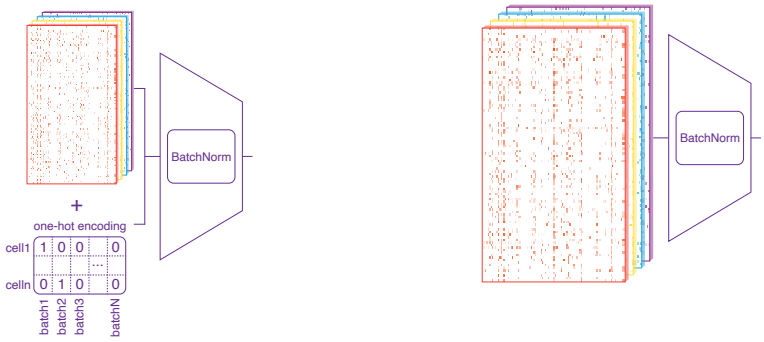

a
Smart-seq2
10X
Microwell-seq
Number of Cells
Aorta
Bladder
BoneMarrow
Brain
Brain_Myeloid
Brain_Non-Myeloid
Diaphragm
Fat
Heart
Heart_and_Aorta
Kidney
Large_Intestine
Limb_Muscle
Liver
Lung
Mammary_Gland
Marrow
Muscle
Ovary
Pancreas
PeripheralBlood
Prostate
Skin
SmallIntestine
Spleen
Stomach
Testis
Thymus
Tongue
Trachea
Uterus
b
batch
tissue
UMAP2
UMAP1
10X
Smart-seq2
Microwell-seq
c
SCALEX Mouse Atlas
urothelial
stomach
hepatocyte
uterus
stromal
endothelial
bladder
hepatic sinusoid
fibroblast
mesenchymal stem
pro-B
mesenchymal
B
Kupffer cell
small intestine
large intestine
keratinocyte
dendritic
secretory alveoli
myeloid
myofibroblast
luminal
basal of epidermis
pancreatic ductal
keratinocyte stem
proximal tubule
acinar
AT2
collecting duct
T
skeletal muscle satellite
immature T
cardiac muscle
basal
neuron
monocyte progenitor
ovary
oligodendrocyte precursor
astrocyte
granulocyte
multipotent progenitor
leydig
erythroblast
sperm
pancreatic endocrine
microglial
oligodendrocyte
UMAP2
UMAP1
d
Mouse Atlas
Human Atlas
Human Fetal Atlas
ARI
NMI
raw
SCALEX
Seurat v3
Harmony
MNN
Conos
BBKNN
Scanorama
scVI
LIGER
online iNMF
e
Over-correction Score
Mouse_Atlas
Human_Atlas
Human_Fetal_Atlas
SCALEX
Seurat v3
Harmony
MNN
Conos
BBKNN
Scanorama
scVI
LIGER
online_iNMF
f
Projected: tabula_muris_senis_Smart-seq2
Projected: tabula_muris_senis_10X
Projected: mouse_lung
Projected: mouse_kidney
AT2
proximal tubule
UMAP2
UMAP1
g
Projected: tabulas_muris_senis_Smart-seq2
ARI:0.905 NMI:0.905 F1:0.871
Projected: tabulas_muris_senis_10X
ARI:0.886 NMI:0.900 F1:0.907
Projected: mouse_lung
ARI:0.885 NMI:0.843 F1:0.863
Projected: mouse_kidney
ARI:0.949 NMI:0.911 F1:0.939
Prediction
cell-type

Fig 5


a
healthy control (n=31)
mild/moderate (n=46)
severe (n=50)
convalescent (n=12)
influenza (n=5)
b
SCALEX COVID-19 PBMC Atlas
c
batch
d
healthy control
mild/moderate
severe
convalescent
influenza
UMAP1
UMAP2
e
IFITM1
ISG15
IFI27
MX1
Median expression in group
CD14-IL1B-Mono
CD14-ISG15-Mono (mild/moderate)
CD14-ISG15-Mono (severe)
f
CD14-IL1B-Mono vs CD14-ISG15-Mono
complement activation, classical pathway
regulation of immune effector process
regulation of complement activation
regulation of humoral immune response
regulation of acute inflammatory response
regulation of B cell activation
positive regulation of lymphocyte activation
B cell receptor signaling pathway
phagocytosis
positive regulation of B cell activation
type I interferon signaling pathway
cellular response to type I interferon
negative regulation of viral genome replication
negative regulation of viral life cycle
regulation of viral genome replication
neutrophil chemotaxis
regulation of ERK1 and ERK2 cascade
positive regulation of cell proliferation
positive regulation of cell motility
regulation of cell proliferation
regulation of cell migration
lymphocyte chemotaxis
positive regulation of leukocyte chemotaxis
cellular response to cytokine stimulus
cytokine-mediated signaling pathway
-log10(Adjusted P-value)
g
Frequency
CD14-ISG15-Mono
NCF1-Immature_Neutrophil
MZB1-Plasma
Megakaryocyte
Pro T
pDC
mDC
healthy control
mild/moderate
severe
convalescent
influenza
h
Projected: SC4 Atlas
Projected: SC4 Atlas
Cell cluster in SC4 study
SCALEX COVID-19 PBMC Atlas + SC4 Atlas
i
SCALEX COVID-19 PBMC Atlas
B
Plasma
CD4 T
CD8 T
NK
Monocyte
Macrophage
DC
Neutrophil
Epithelial
SC4 Atlas
j
Cytokine score
healthy control
mild/moderate
severe
Inflammatory score
healthy control
mild/moderate
severe
UMAP1
UMAP2
k
IGKC-Mega vs TUBA8-Mega
protein-DNA complex assembly
nucleosome assembly
nucleosome organization
regulation of blood coagulation
negative regulation of blood coagulation
negative regulation of platelet activation
positive regulation of respiratory burst
humoral immune response mediated by circulating immunoglobulin
antigen receptor-mediated signaling pathway
complement activation, classical pathway
phagocytosis, engulfment
plasma membrane invagination
-log10(Adjusted P-value)

Supplementary Fig 1

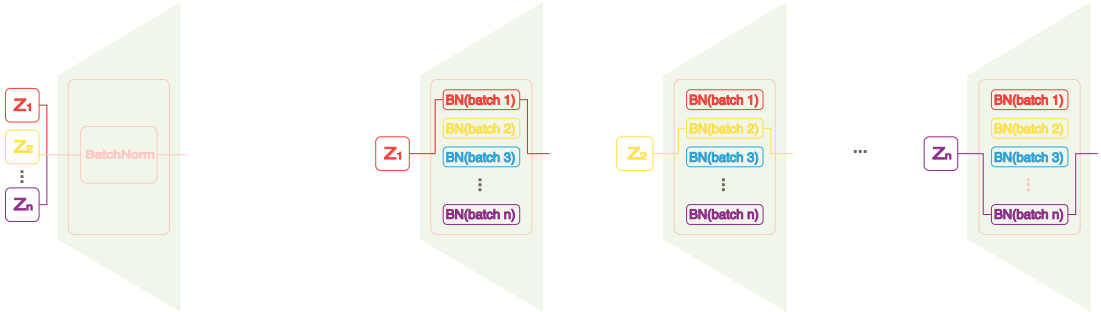
a



Encoder with batch label

Encoder without batch label

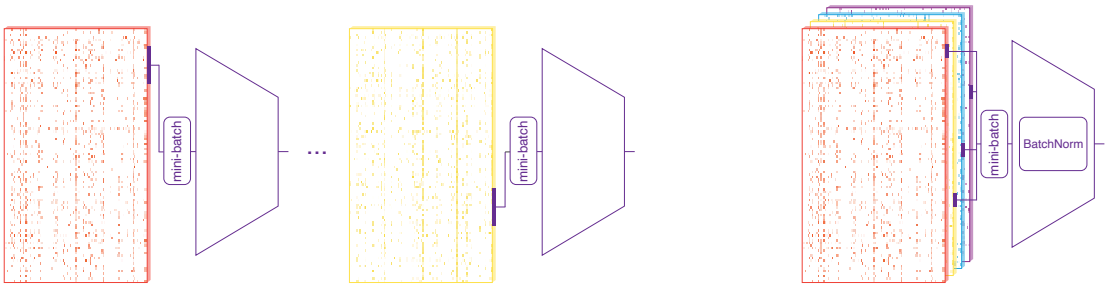
b



Batch Normalization

Domain-specific Batch Normalization (DSBN)

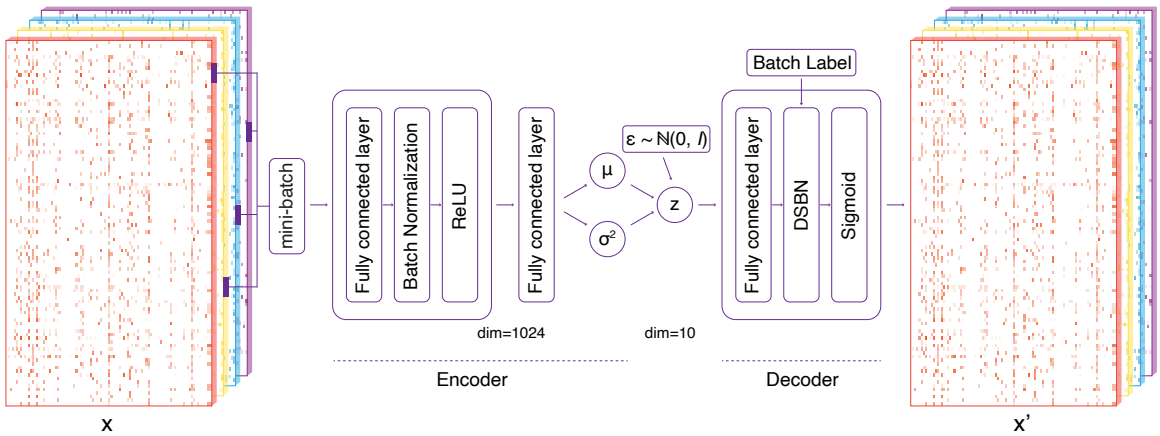
c



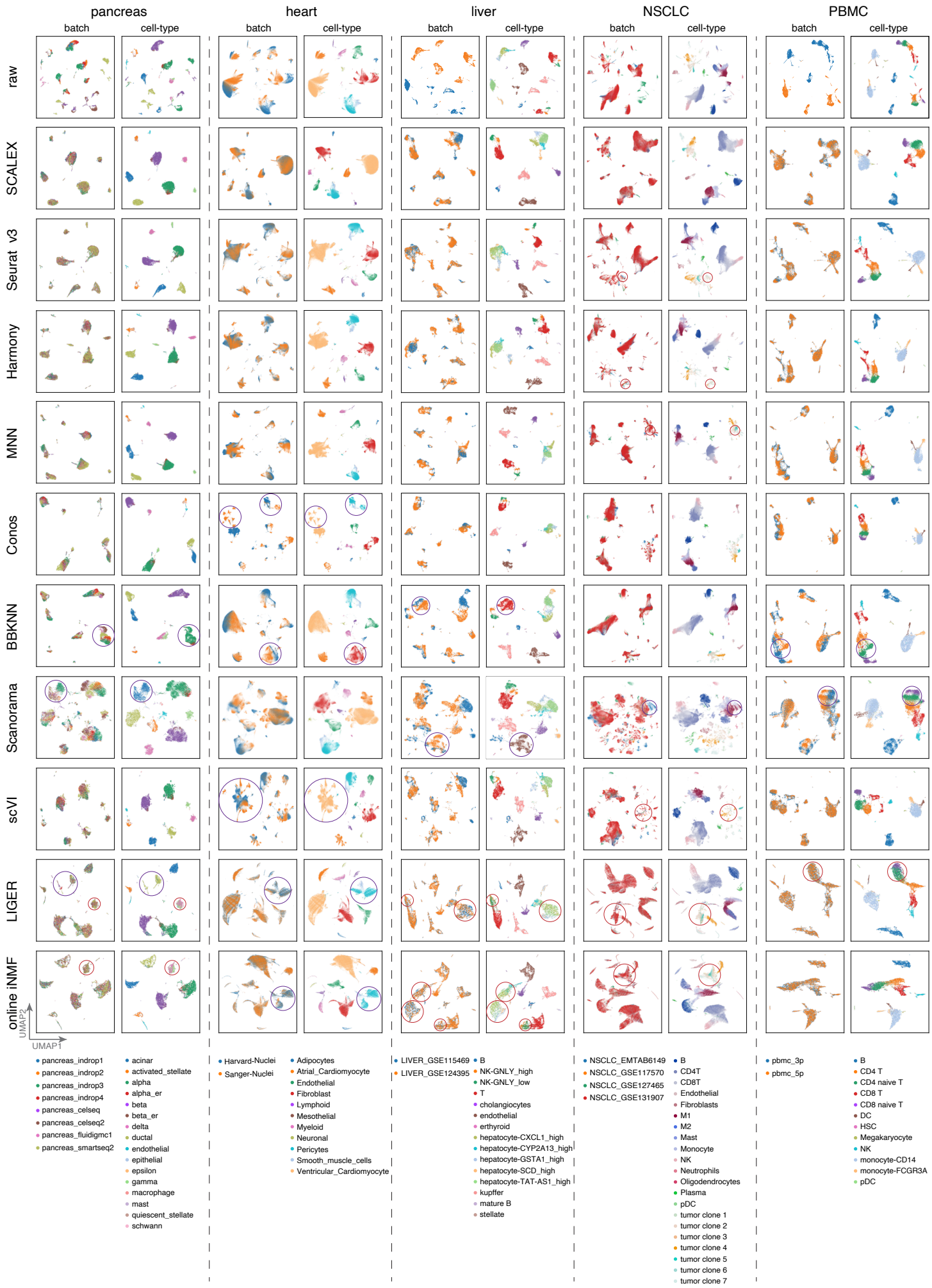
Mini-batch sampling by batch without Batch Normalization

Mini-batch sampling from all batches with Batch Normalization

d

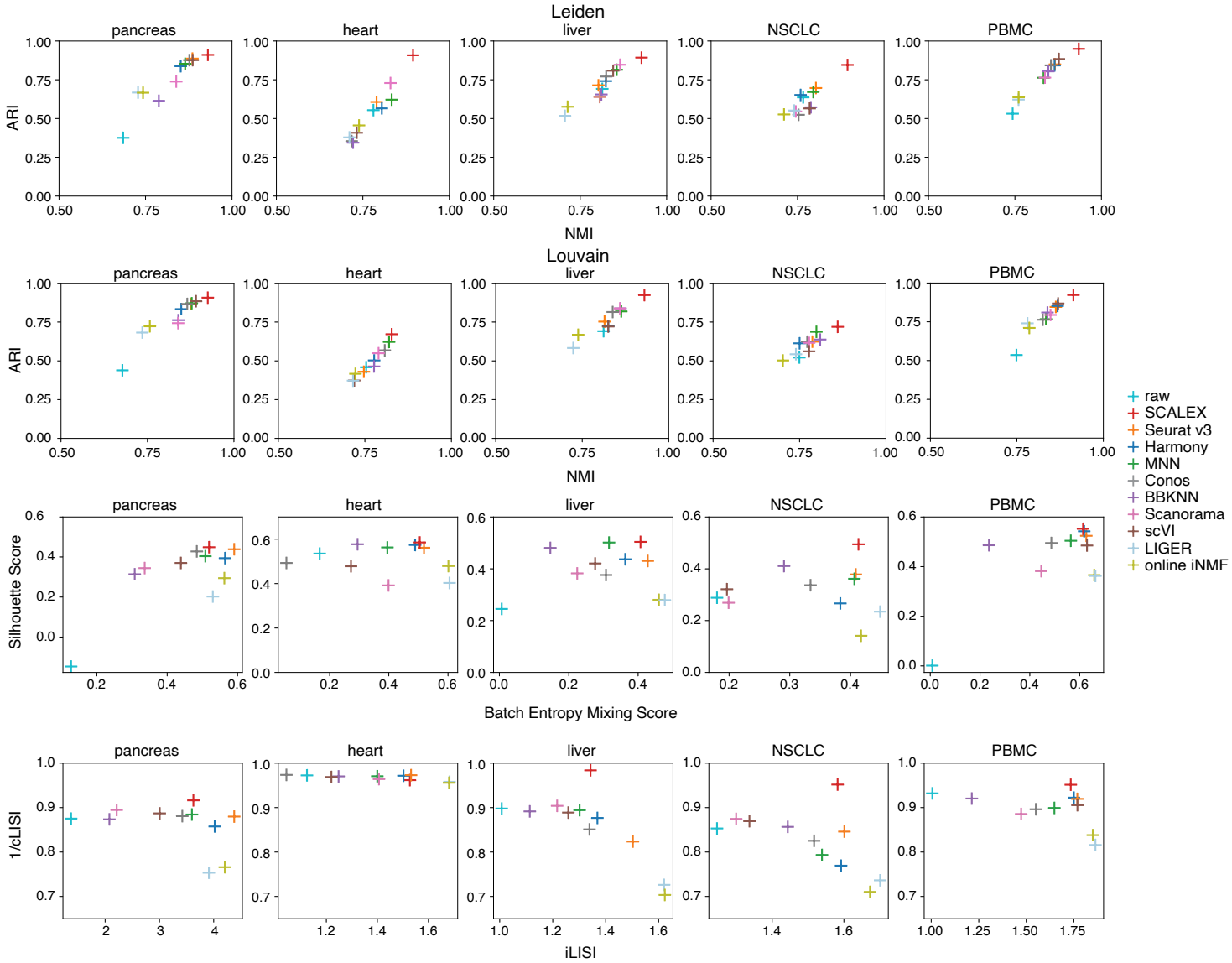


Supplementary Fig 2

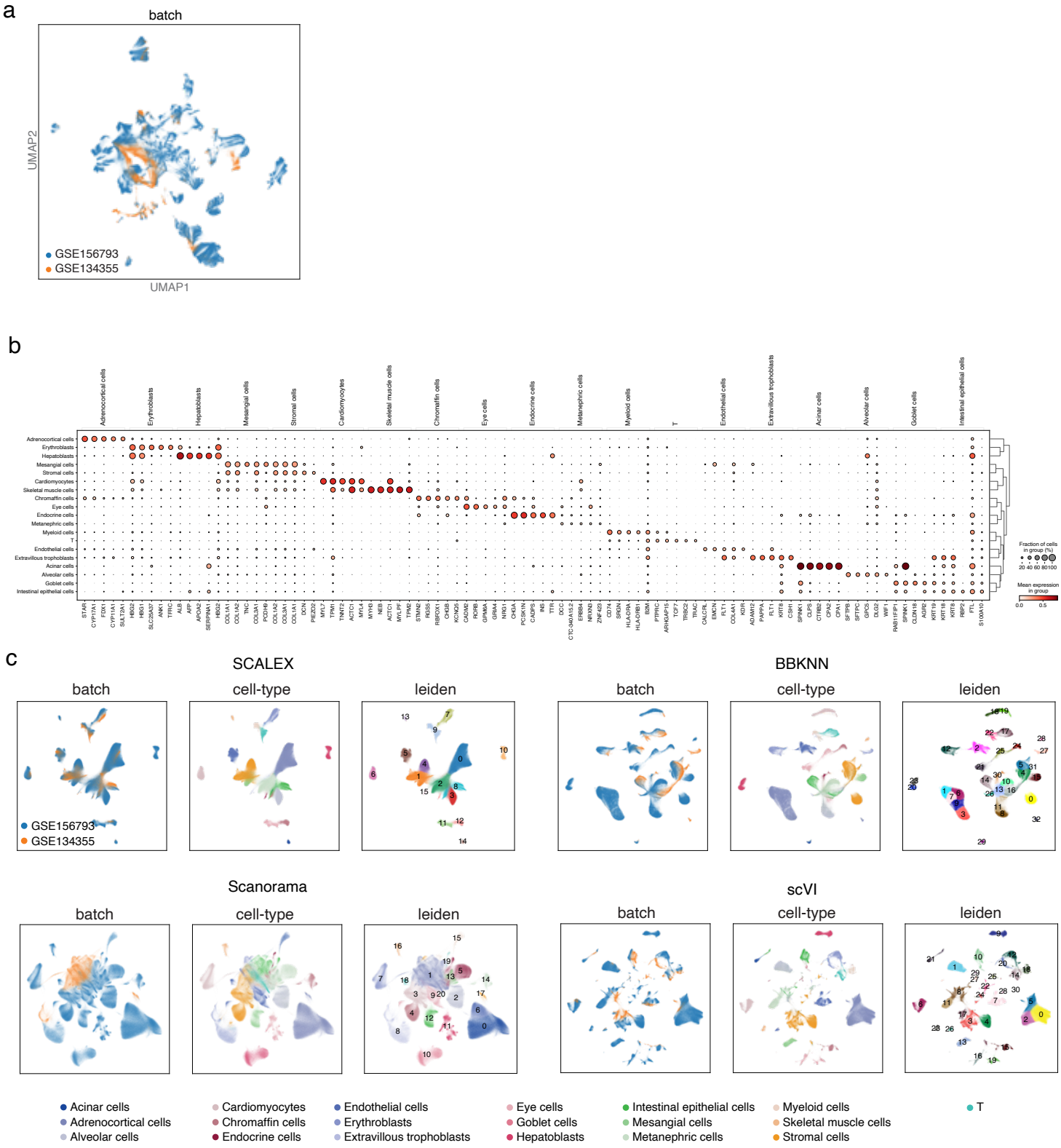


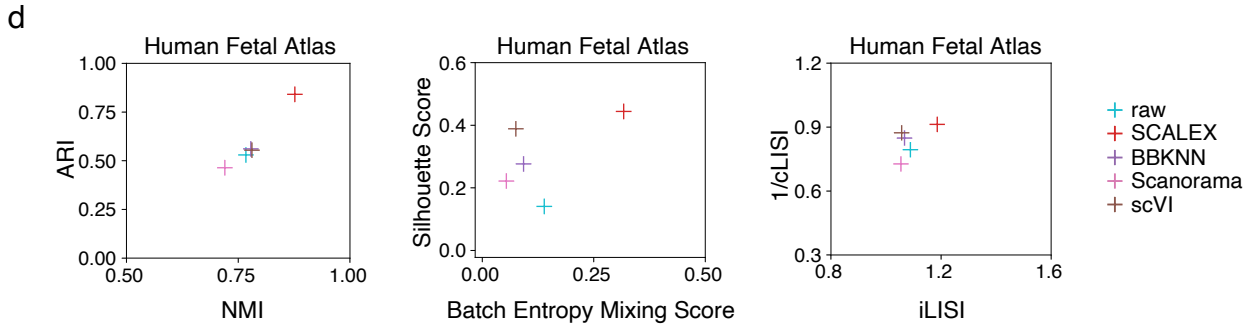
Supplementary Fig 3

Supplementary Fig 4

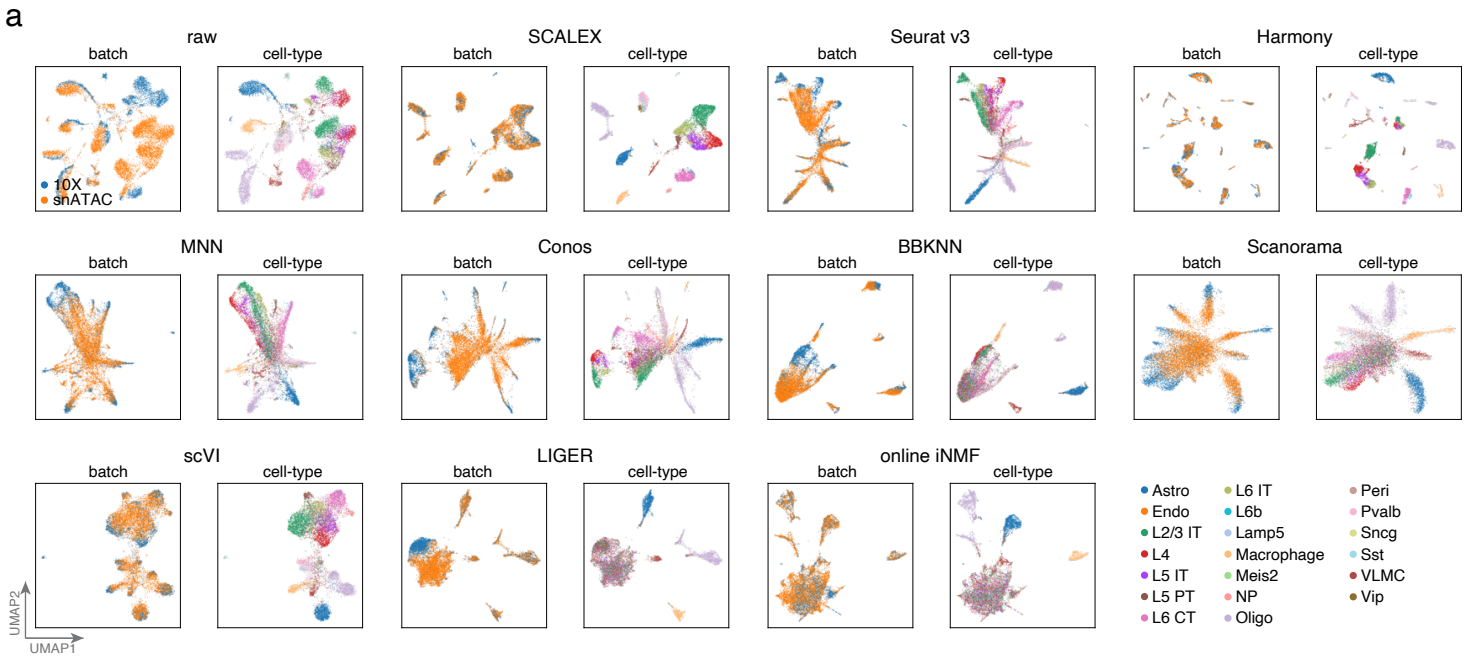


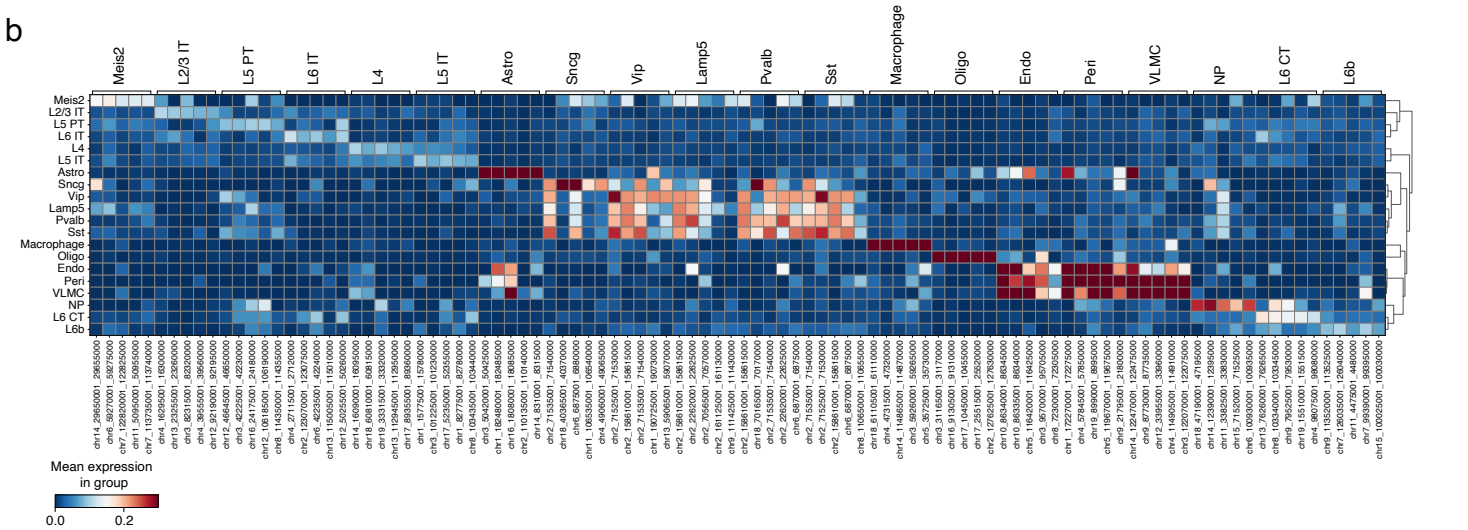
Supplementary Fig 5

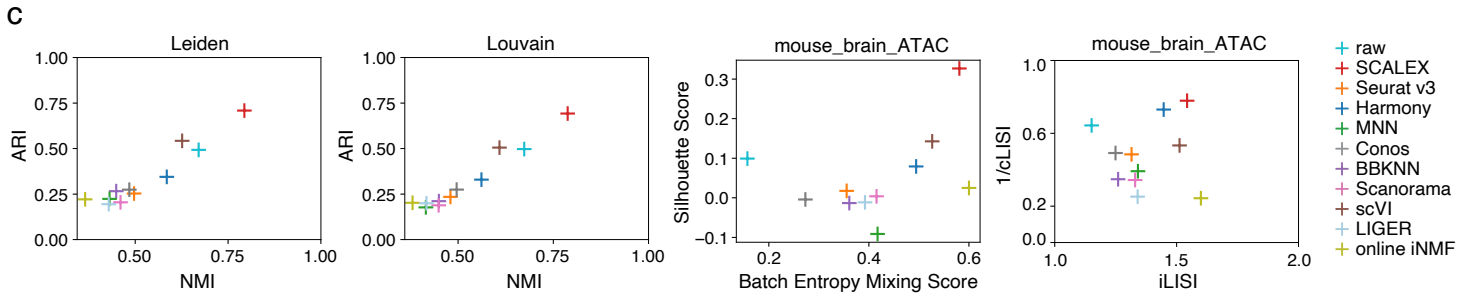


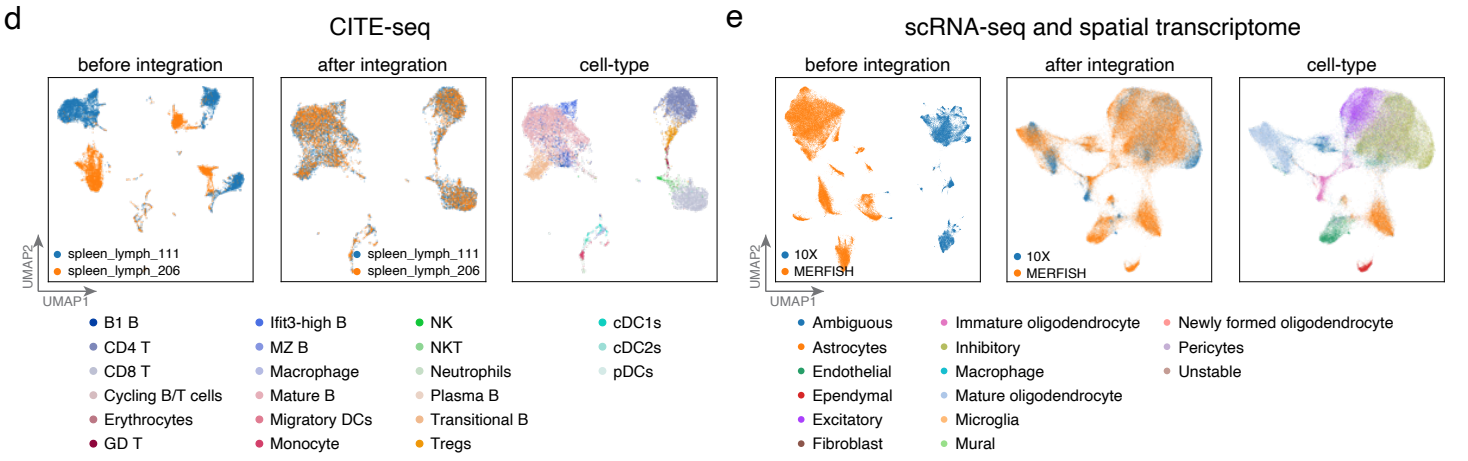


Supplementary Fig 6

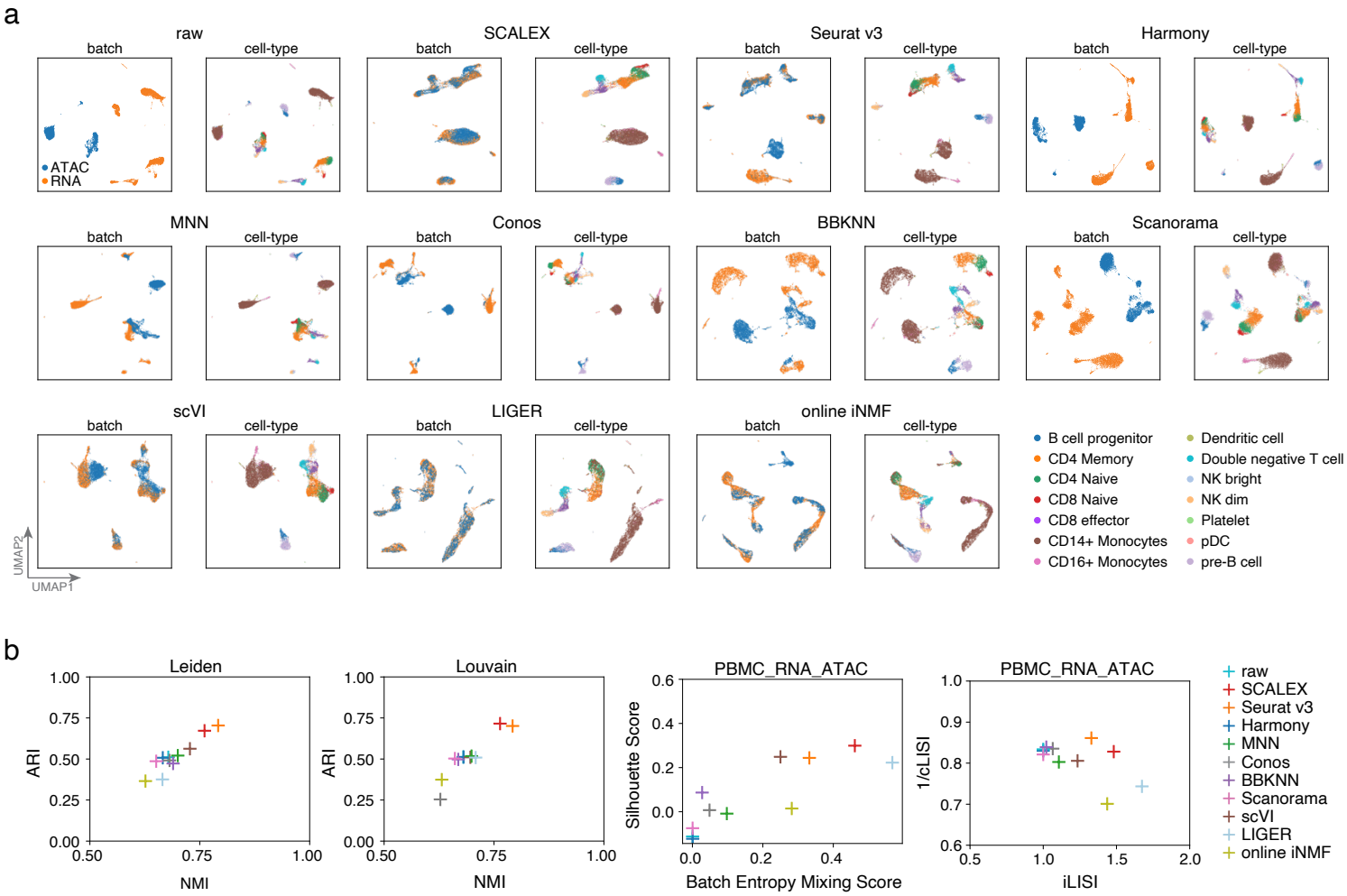




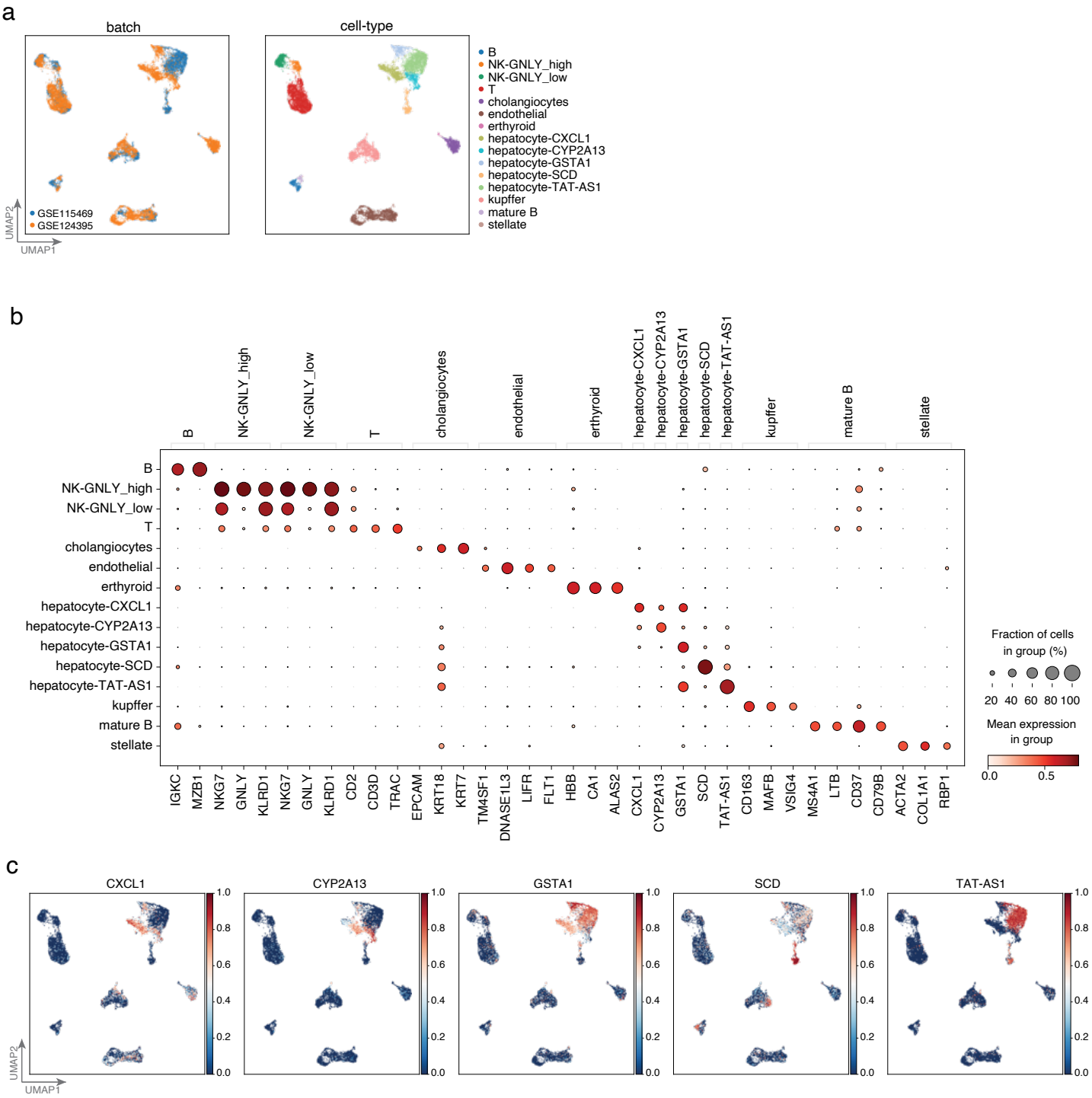




Supplementary Fig 7



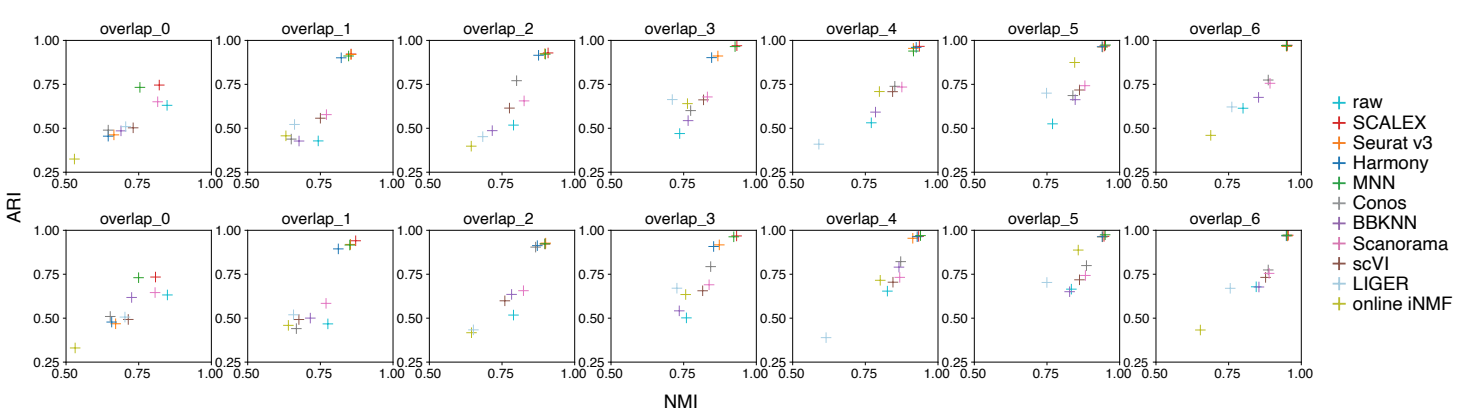
Supplementary Fig 8



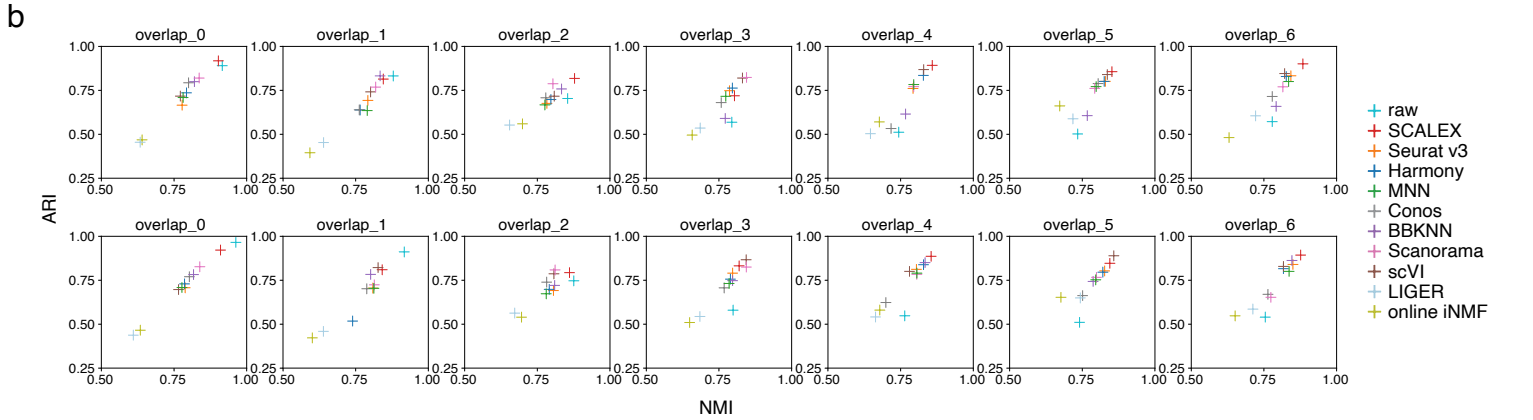
Supplementary Fig 9

a

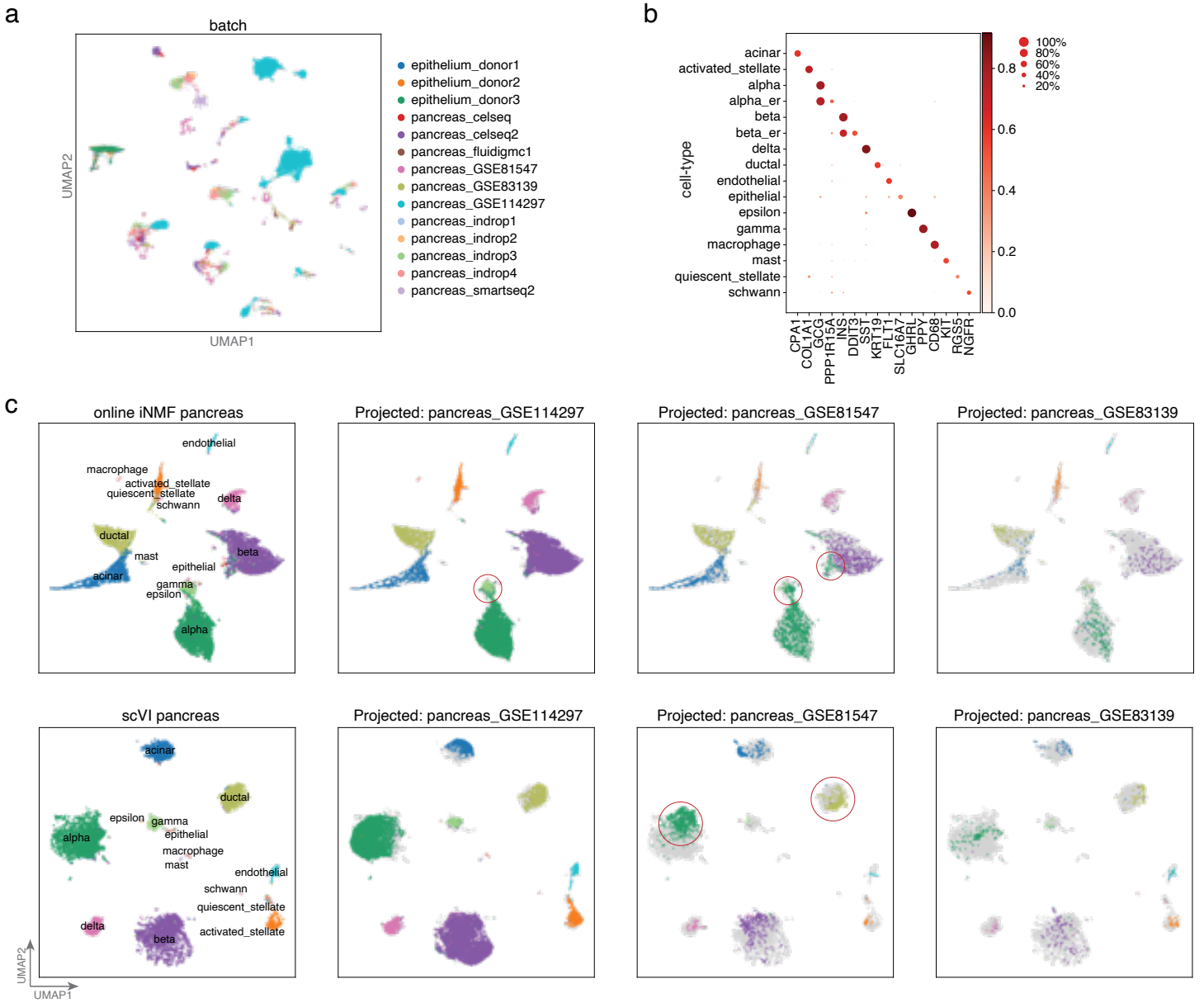
b



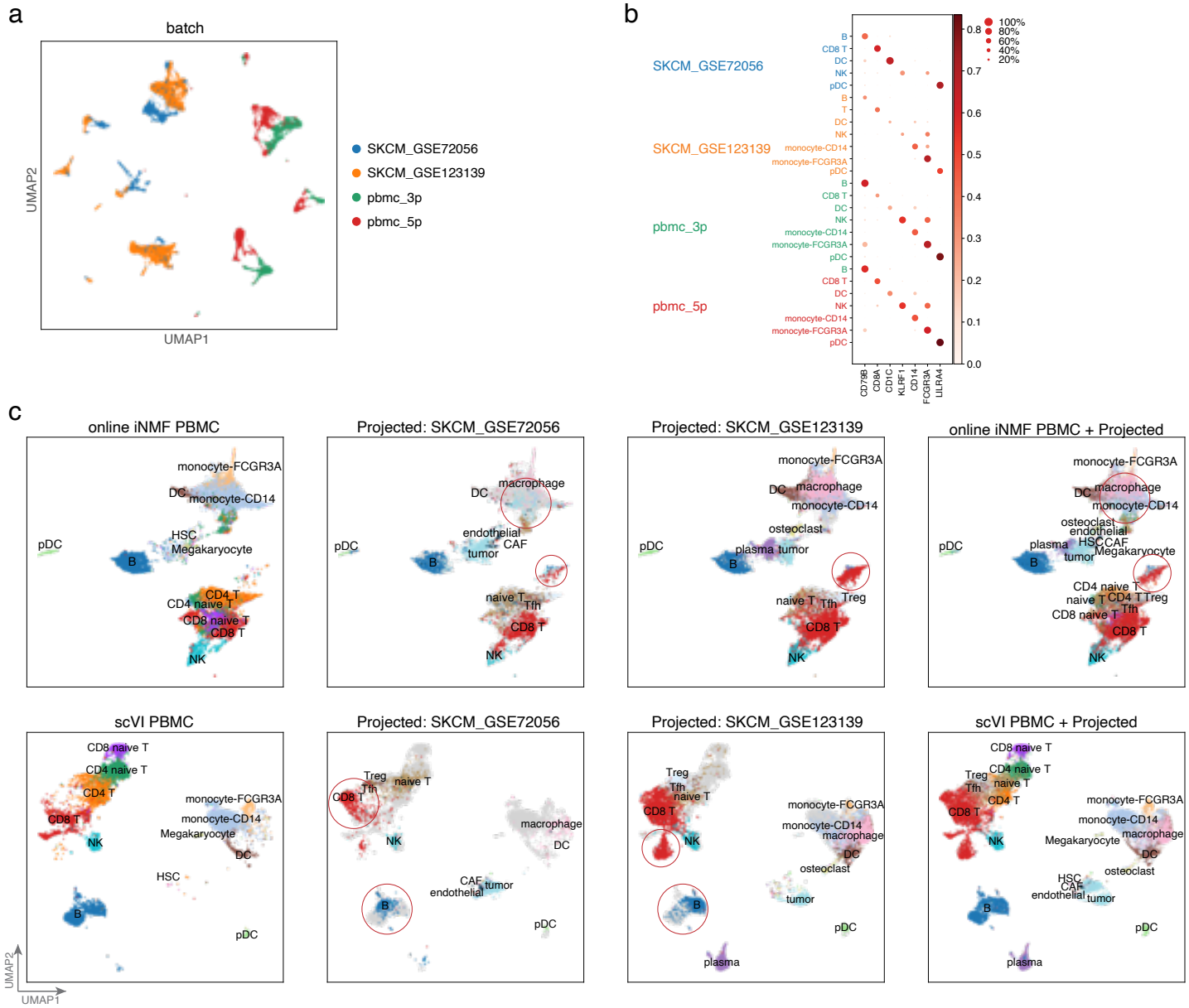
Supplementary Fig 10



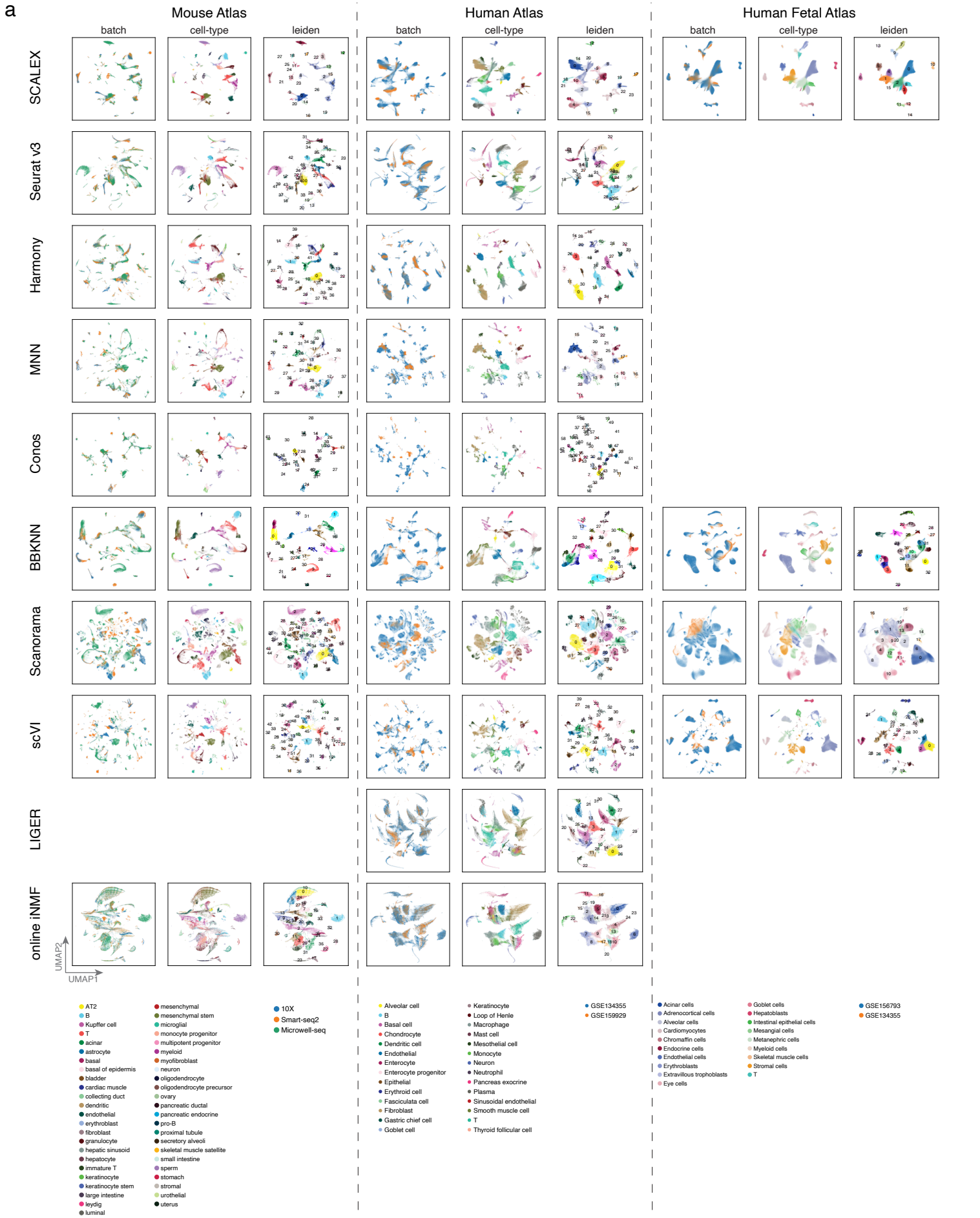
Supplementary Fig 11

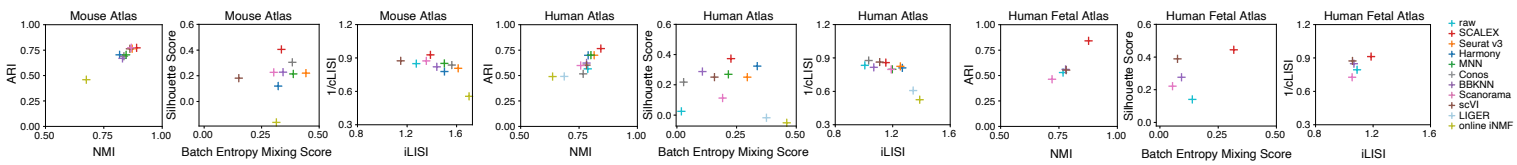


Supplementary Fig 12

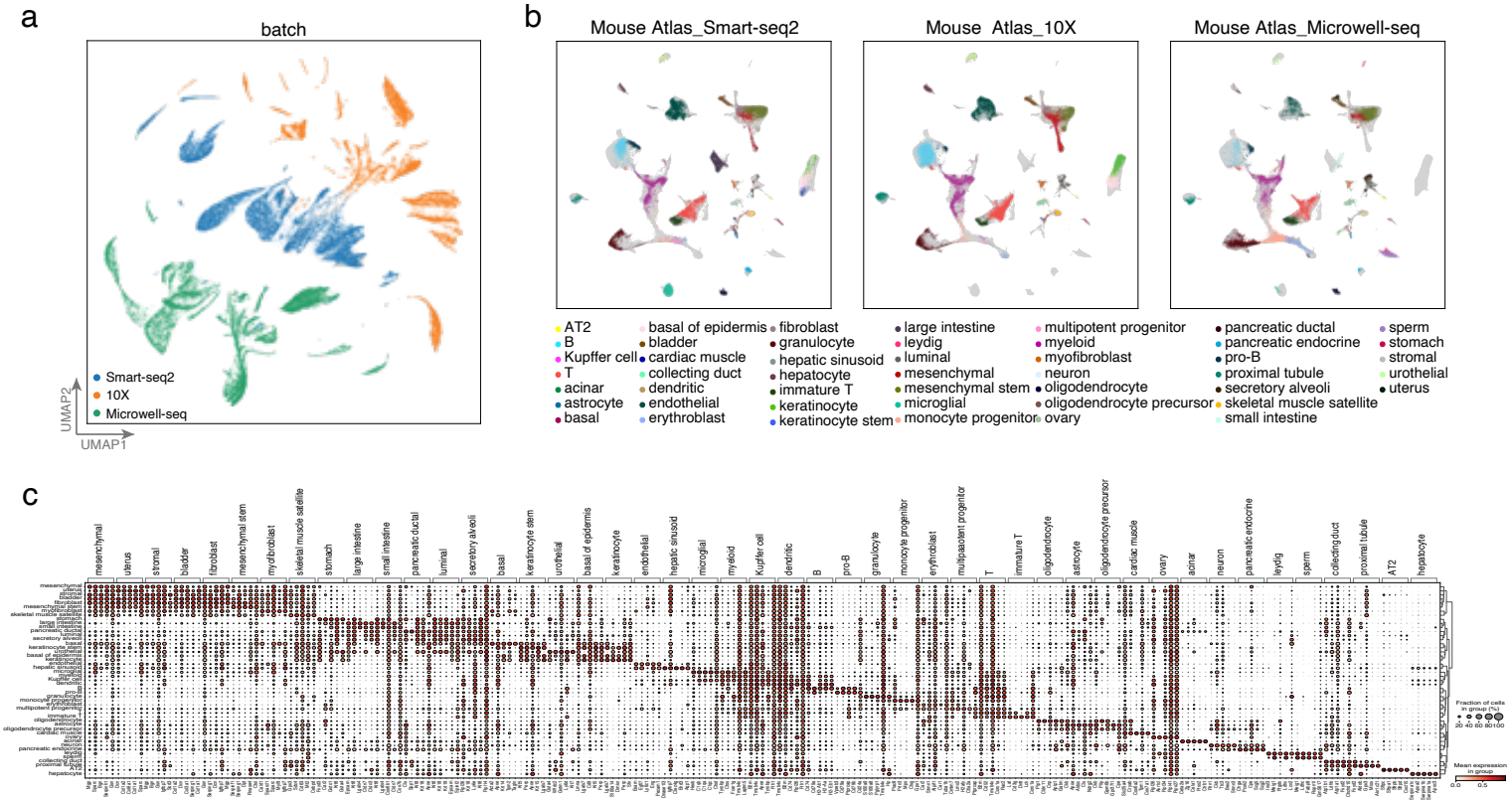


Supplementary Fig 13

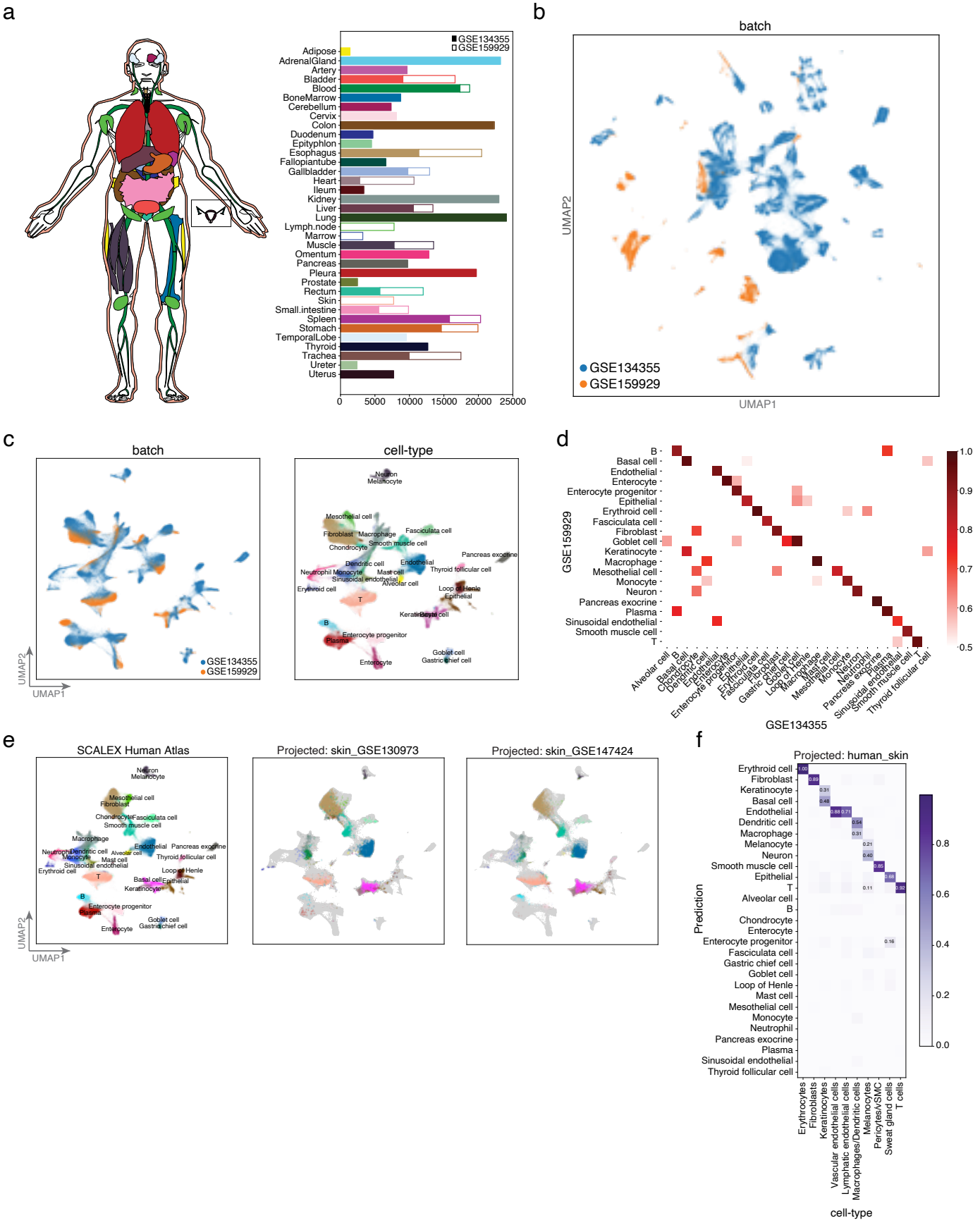




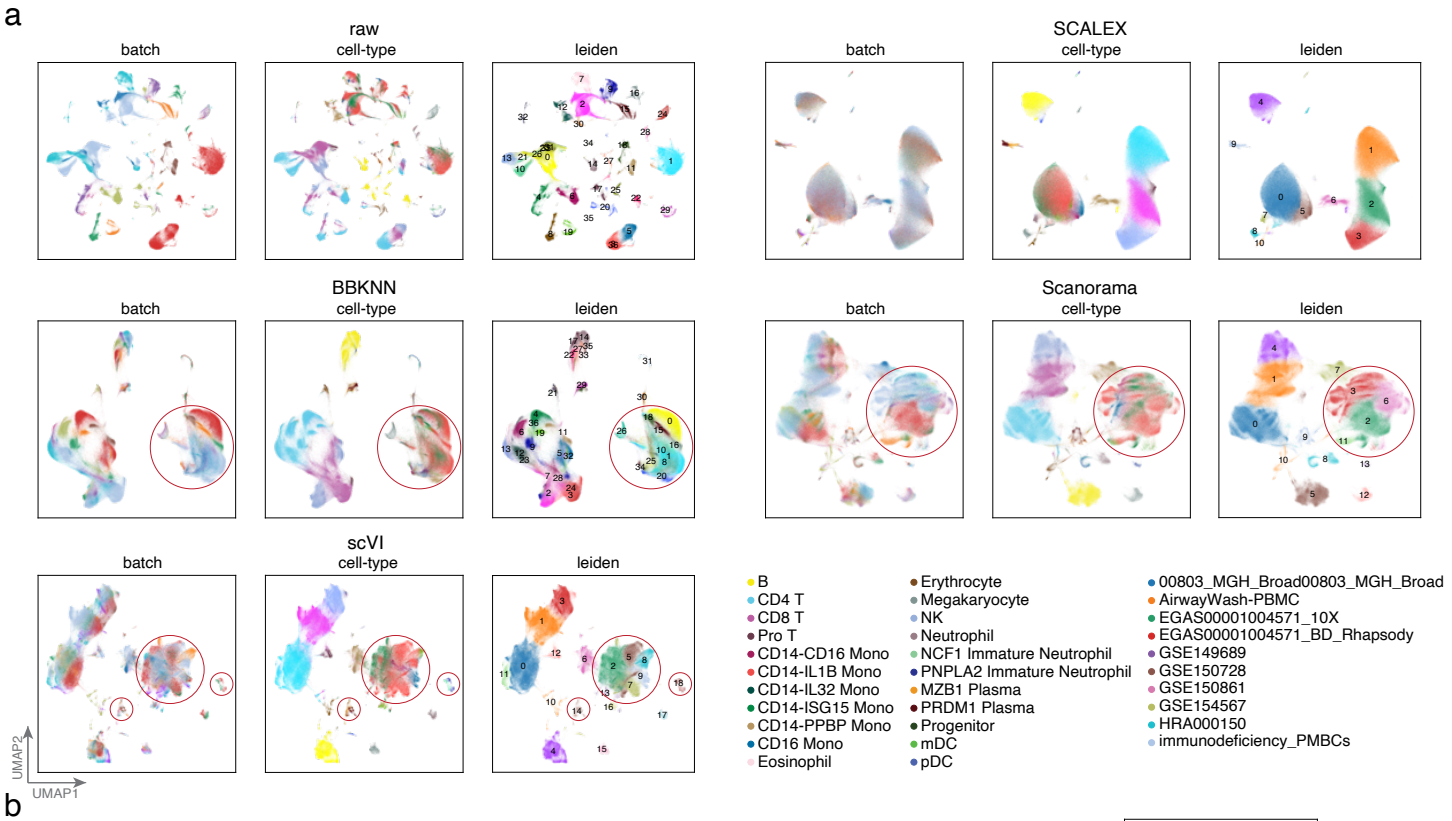
Supplementary Fig 14

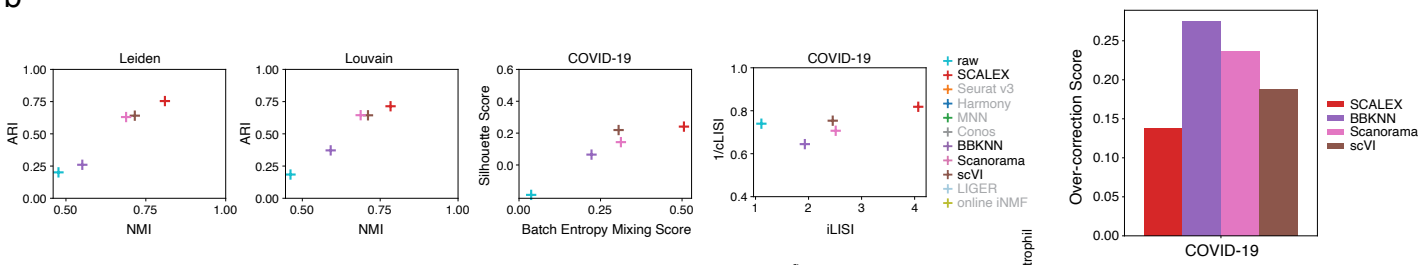


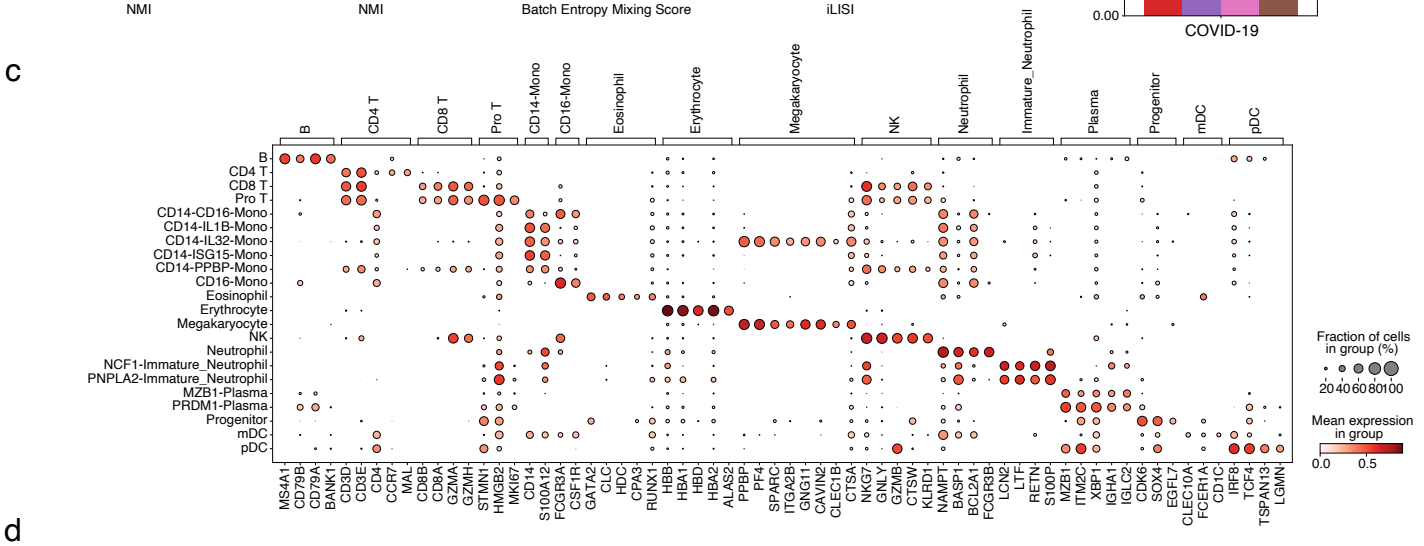
Supplementary Fig 15

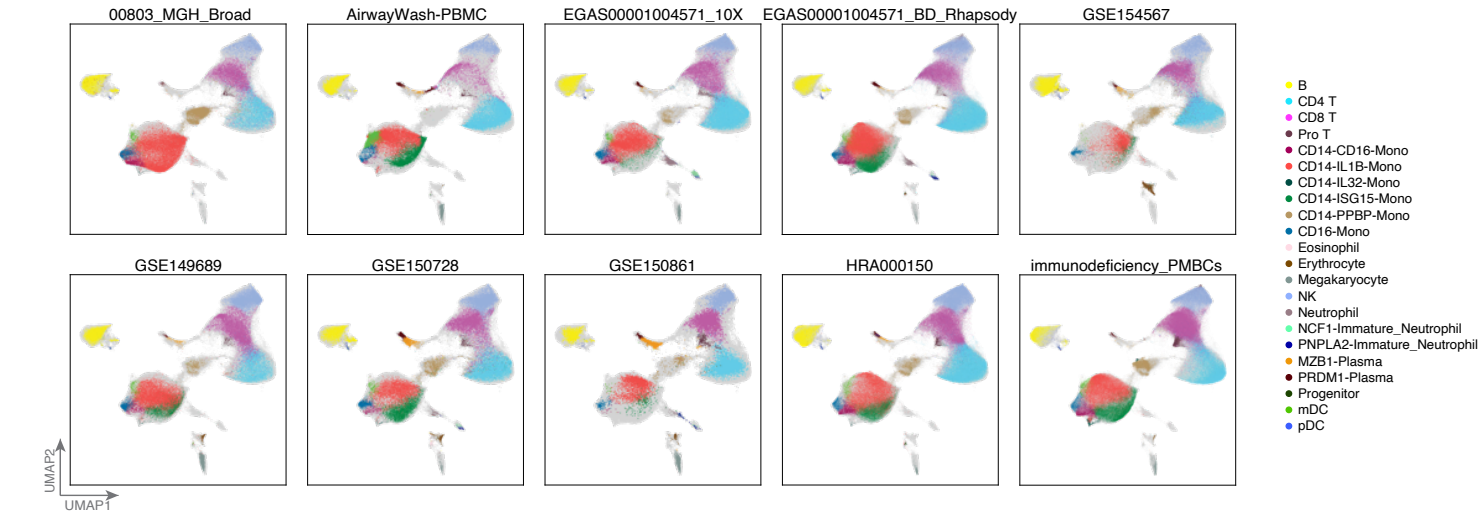


Supplementary Fig 16

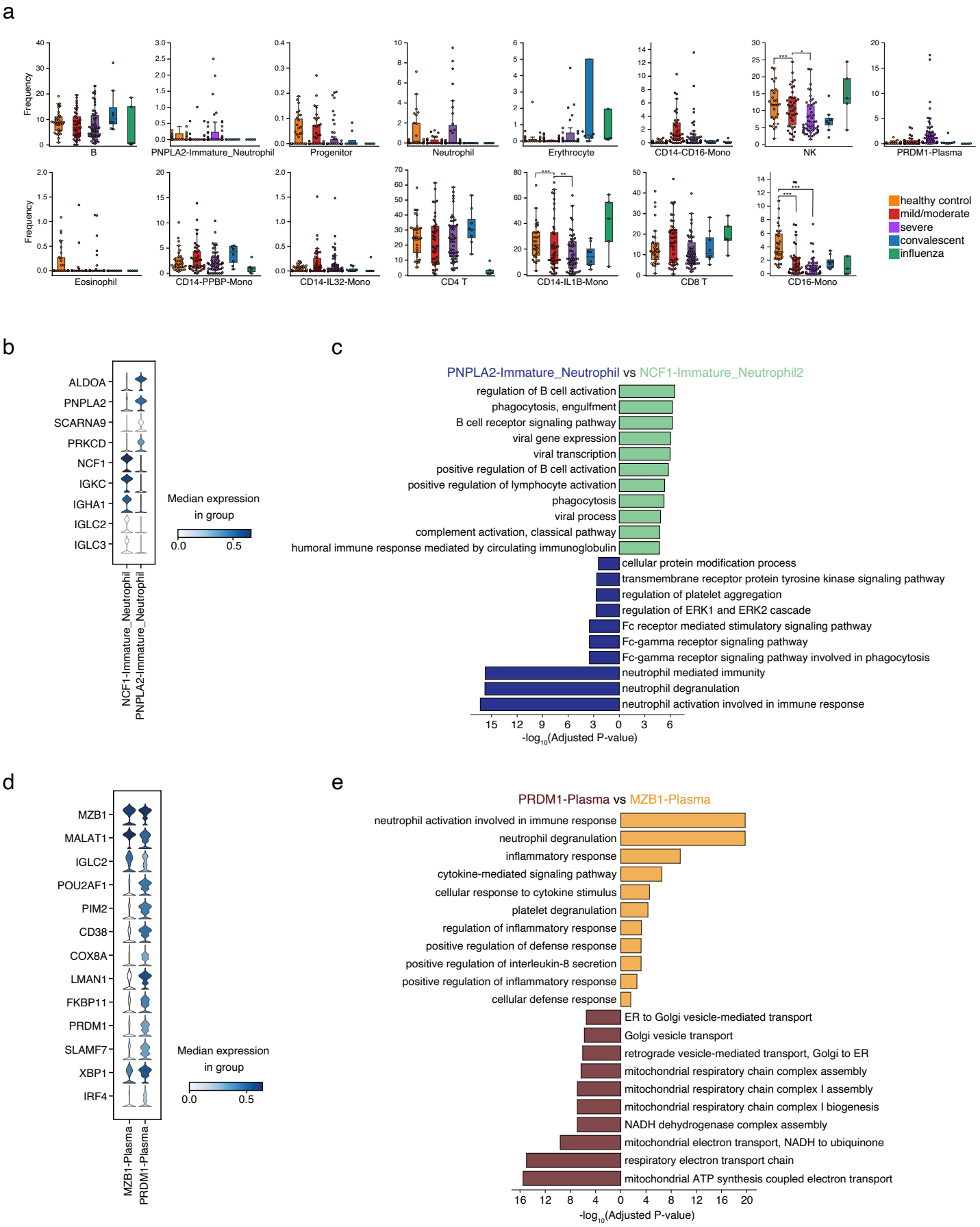




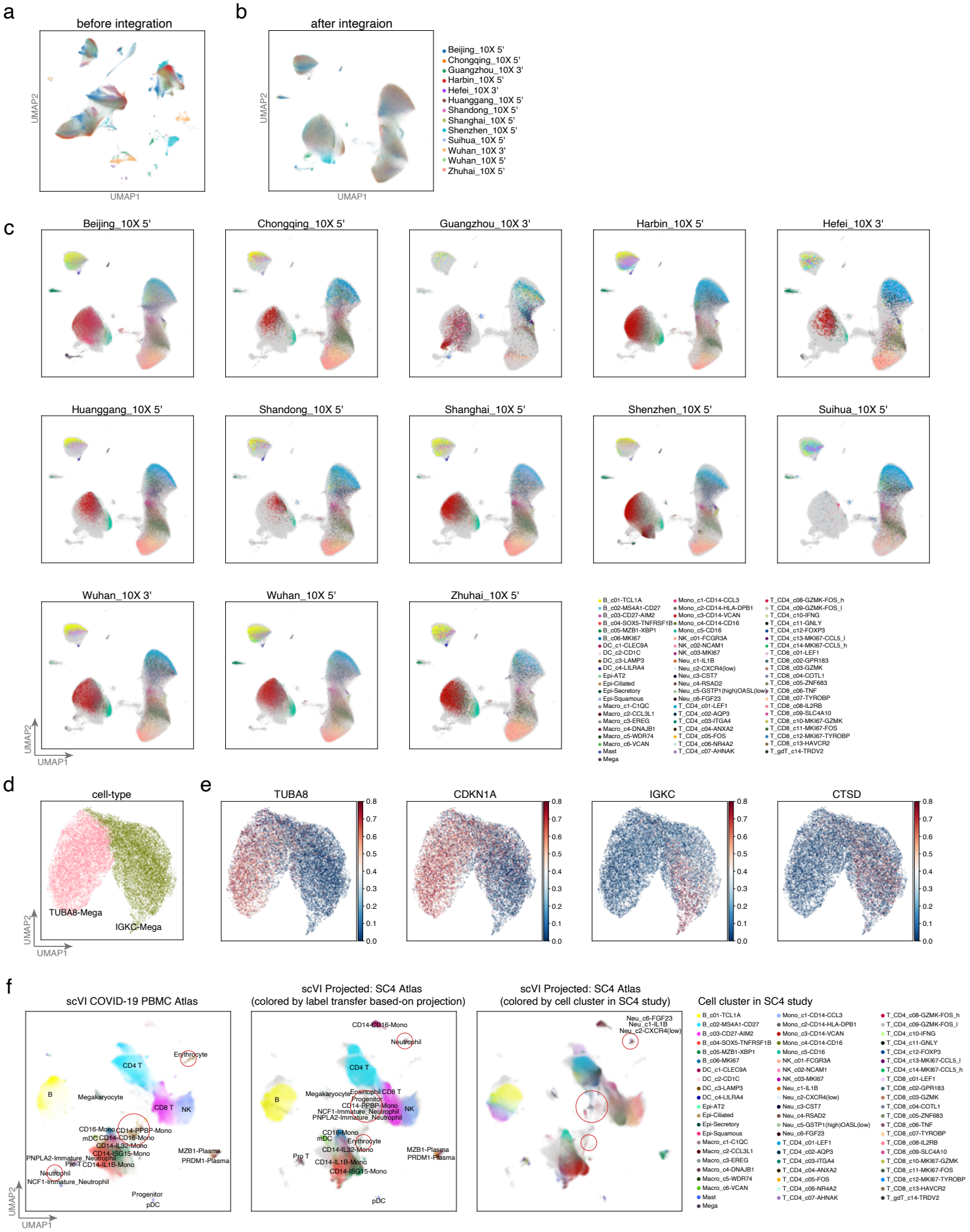




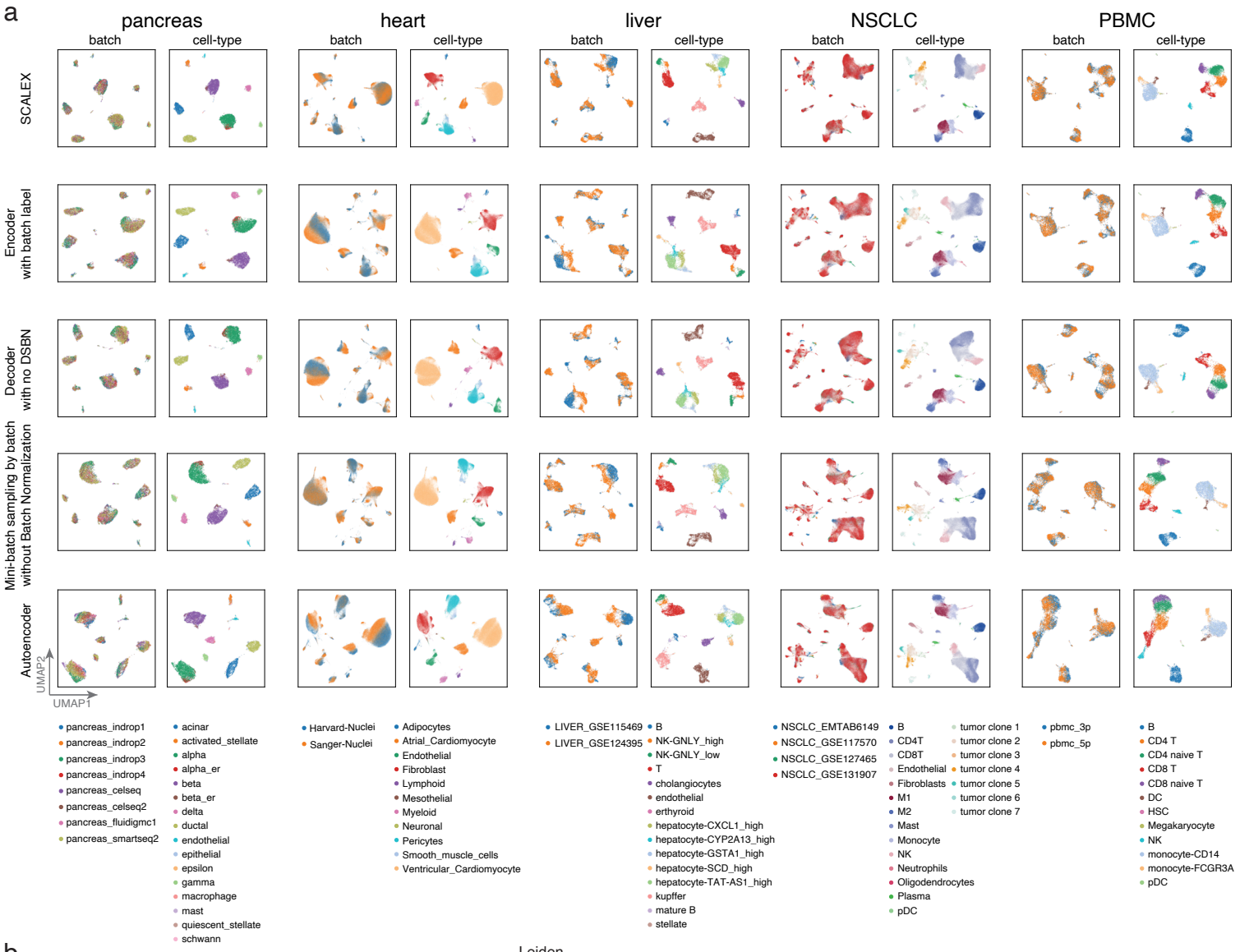
Supplementary Fig 17

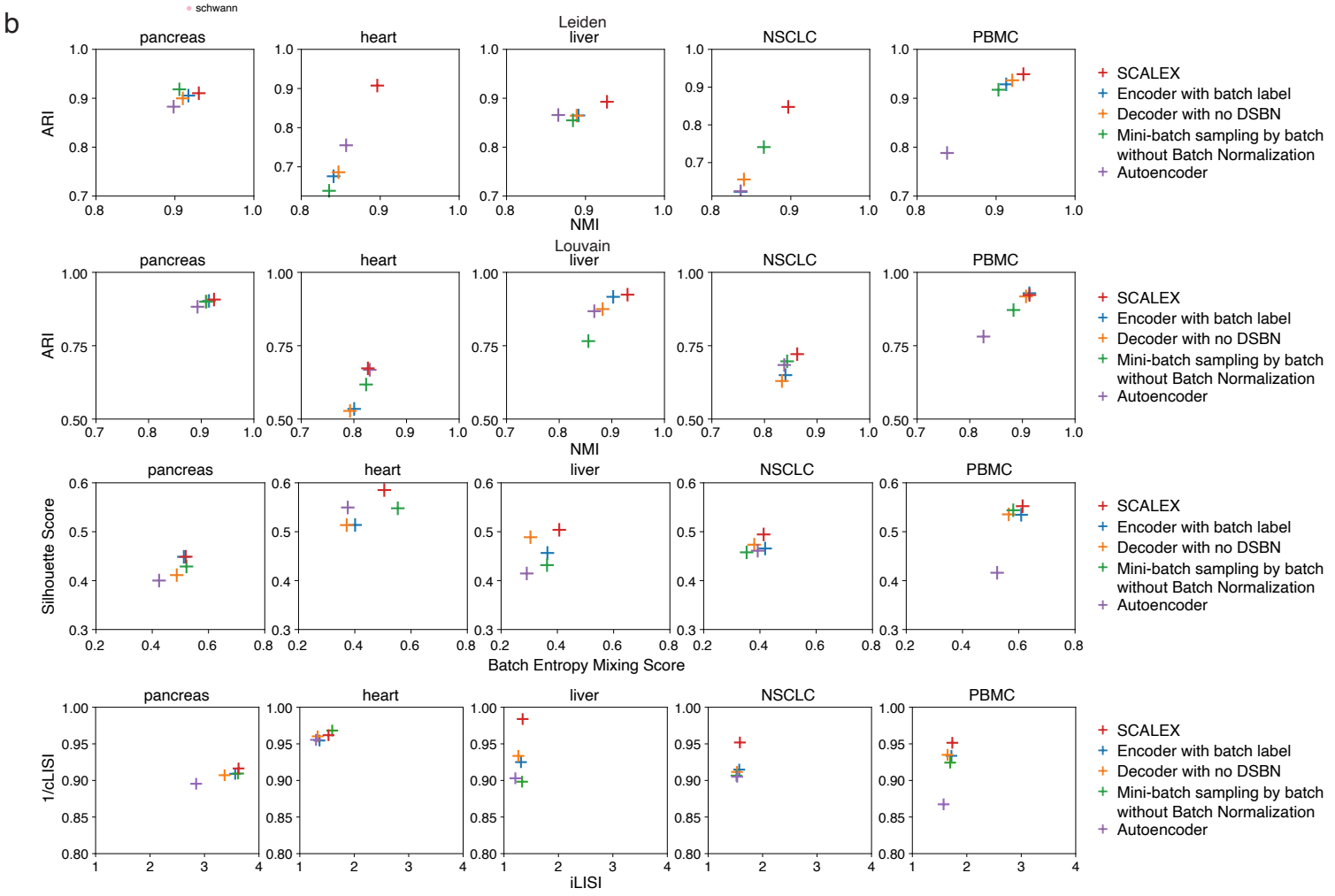


Supplementary Fig 18



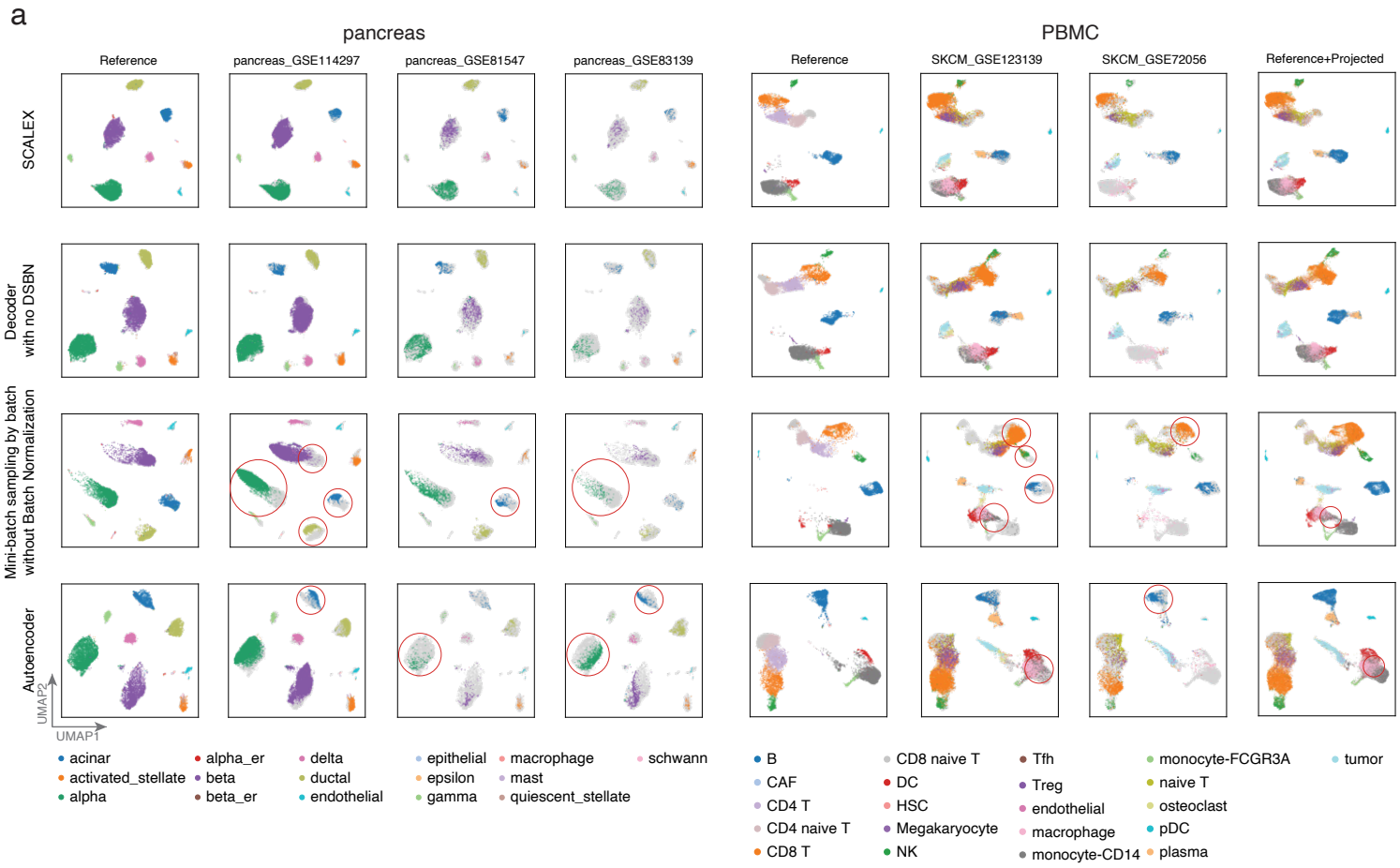
Supplementary Fig 19



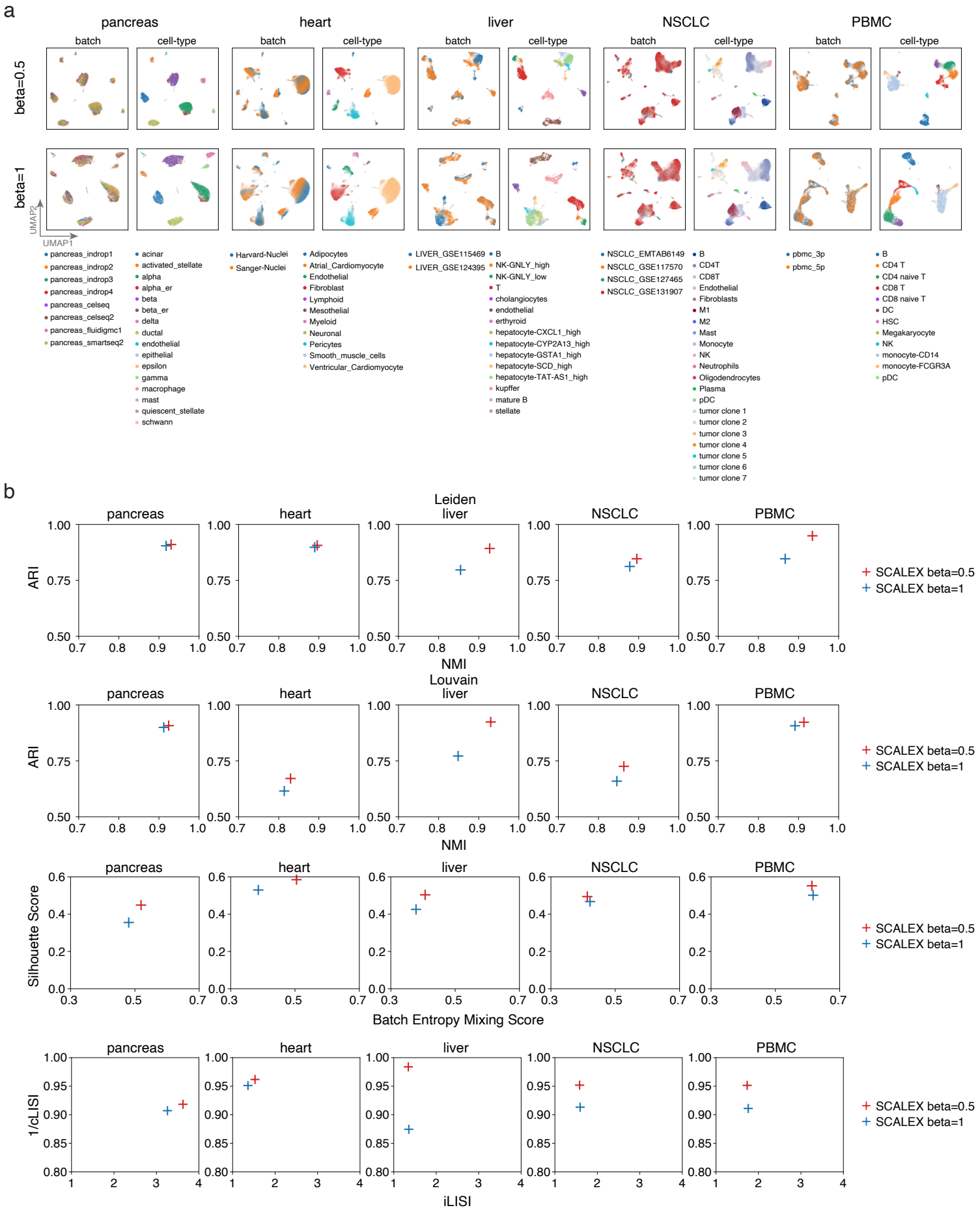


Supplementary Fig 20

a



Supplementary Fig 21



Supplementary Note: Ablation Study

To accomplish an accurate generalized encoder, the design of the full SCALEX included the following specific innovations:

1) an asymmetric autoencoder that inputs batch information only to the decoder (*i.e.*, never to the encoder) (See diagram in Fig. S1);

2) a DSBN layer in the decoder to release the encoder from the burden of capturing the batch-specific variations;

3) a mini-batching strategy that samples data from all batches simultaneously (rather than single batches iteratively) and thus more tightly follows the same overall distribution of the full input dataset; this strategy includes a Batch Normalization layer in the encoder that adjusts the deviation of each mini-batch and aligns them to the overall input distribution.

We conducted ablation studies to investigate the contributions of each design element of SCALEX. That is, we analyzed the performance (for integration and projection tasks) of the full SCALEX and four SCALEX "test-variants", each with a distinct network architecture (Fig. S19-21). These can be summarized as follows:

Full SCALEX (referred as "Baseline" in the following). This Baseline model includes an encoder without batch labeling, sampling from all batches with Batch Normalization, decoder with DSBN, and beta=0.5. All other ablations are compared relative to this.

Encoder with batch label. A test-variant with Baseline including an encoder with batch label as input. We found that the integration performance of this variant is similar with the full SCALEX, showing only a slight reduction in the evaluation scores (Fig. S19). However, the real issue is this: this addition of batch information at the beginning precludes online integration of newly arriving data. Put another way, this SCALEX test-variant is not capable of integrating single-cell data in a truly online manner.

Decoder without DSBN. A test-variant of Baseline removing the DSBN layer from decoder. The DSBN layer combines multiple batch-specific Batch Normalization layers to capture the batch-specific information; this approach has been demonstrated as effective for domain adaption, as it provides a weak alignment across different domains. We observed an obvious drop in the integration performance (in terms of all evaluation scores) and a slight drop in the

projection performance of this SCALEX test-variant, based-on the UMAP embeddings (Fig. S19,20).

Sampling by batch without the Batch Normalization (BN) layer in the encoder. Removing the Batch Normalization layer from the encoder, and each mini-batch is sampled by batch instead of from all the batches. The integration performance of this test-variant was obviously worse than full SCALEX (Fig. S19). The projection performance also dropped obviously, with clear deviations from the common cell-embedding space (Fig. S20).

Regular autoencoder. A test-variant that uses a regular autoencoder instead of a VAE framework. This variant performed the worst among all the variants we tested, for both the integration and projection tasks (Fig. S19,20).

Note that we also explored altering the Beta factor. Replacing the beta factor as 1 instead of 0.5. The integration performance of test variant of SCALEX for beta factor of 1 is worse than the SCALEX for beta factor of 1 (Fig. S21).