BASILISC: BAYESIAN ALIGNMENT OF SIAMESE-INTEGRATED LATENTS FOR INFERRING SINGLE-CELL CONCORDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell assays are the established method to study the genome one cell at a time. Unimodal assays measure one molecular layer, and the newer multimodal assays measure several. Most data are single-modality, so integrating them relies on models trained on the multimodal assays, where the cell correspondence across modalities is known. These models return a matching, but rarely a confidence measure. We present BASILISC (Bayesian Alignment of Siamese-Integrated Latents for Inferring Single-cell Concordance), an uncertainty-aware siamese network for matching unimodal single-cell embeddings. It places the uncertainty on the correspondence itself: for a cross-modal pair of cells, it predicts a distribution over whether the two share a cell type, and at the cell level over which cell is the partner. To our knowledge, this is the first uncertainty score on a learned cross-modal correspondence evaluated against a known ground-truth pairing. BASILISC learns a biologically meaningful similarity across modalities that reflects cell lineage: an unseen cell type is matched to its lineage relative, matching still works on such held-out cell types (balanced accuracy 0.85 on unseen vs. 0.93 on seen types), and the predictive uncertainty rises on these out-of-distribution cells. The uncertainty score also selects high-quality matches for downstream multimodal analysis: discarding the least certain fifth of the pairs raises accuracy from 0.93 to 0.98. At the cell level, a deep ensemble of the architecture reaches 0.062 matching accuracy on the benchmark test set, above the competition winner, and the posterior read on that assignment flags the wrong matches that cross a cell-type boundary better than a deterministic read of the same scores (AUROC 0.68 against 0.64).

1 INTRODUCTION

In complex systems, a single measurement modality captures only part of an entity; to solve prediction problems that depend on the whole, the representations learned from each modality have to be brought into correspondence. The cell is such an entity, and to understand the biological processes that underlie cellular identity and disease, we need a meaningful multimodal representation of the cell state. Cross-modal matching models usually learn two encoders, one representing each data modality, and predict a point-estimate matching score for the correspondence of a given pair. Such a deterministic model can tell us which pairs are the best matches, but not how close the runner-up was, or whether the assignment would survive a different draw of the training data, or whether a data point lies where the encoders were never trained. In this work, we focus on the problem of multimodal representation matching. We place a posterior on the encoders and the uncertainty score on the correspondence itself, read the uncertainty of the assignment from the disagreement of the posterior draws, and evaluate that uncertainty where the true pairing is known: paired single-cell multi-omics data.

Single-cell assays are important means of studying the individual cell, in contrast to bulk assays, which average over all cells in a sample and thereby can report only general trends (Tanay & Regev, 2017; Luecken & Theis, 2019). Most single-cell datasets are unimodal, whereas multimodal datasets are more costly and have become available only recently (Stoeckius et al., 2017; Chen et al., 2019; Ma et al., 2020). A unimodal assay measures a single aspect of the cell, such as how much each gene

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

is expressed (scRNA-seq) (Zheng et al., 2017) or which regions of the chromatin are accessible, that is, depleted of nucleosomes and therefore available for regulatory proteins to bind (scATAC-seq) (Buenrostro et al., 2013; 2015). With unimodal assays, each assay is run on a separate but comparable cell population and the two are matched computationally afterward; no true cell-level correspondence exists, so the matching cannot be checked against a ground truth. Multimodal assays instead measure two or more modalities simultaneously in the same cell, yielding a paired dataset in which the modalities are linked through their cell of origin, providing us with the ground truth for the cross-modal matching task. Therefore, paired datasets generated by multimodal single-cell assays are used to train models that can, in turn, be used to integrate unpaired unimodal datasets (Luecken et al., 2021; Lance et al., 2022).

The matching of cell populations across modalities that have been assayed in different cells is termed diagonal integration (Argelaguet et al., 2021). From a machine learning perspective the methods for diagonal integration fall into two categories. In the first, a representation is learned for each modality separately and the correspondence is established afterward as a separate step, by optimal transport couplings (Demetci et al., 2022; Cao et al., 2022; 2020), mutual nearest neighbors (Haghverdi et al., 2018; Stuart et al., 2019), or a mapping fitted between the two spaces (Wu et al., 2021). In the second, a single embedding space is learned for both modalities at once, either by a shared latent-variable generative model such as a multimodal variational autoencoder (Rautenstrauch & Ohler, 2024; Gayoso et al., 2021; Ashuach et al., 2023; Gong et al., 2021; Cao & Gao, 2022), by matrix factorization with shared factors (Welch et al., 2019; Argelaguet et al., 2020), or by manifold alignment (Liu et al., 2019; Cao et al., 2020), which matches the geometry of the two manifolds under objectives that preserve local neighborhood structure and require no cell-level correspondence. In both cases a correspondence score for a given cross-modal input pair is obtained from the learned space (Tu et al., 2022; Gossi et al., 2023; Liu et al., 2023), and in most existing methods it is obtained without measuring the confidence of the model in the predicted match score. On one benchmark dataset (Luecken et al., 2021) which we also use here, the strongest methods for the matching task are the models CLUE (Tu et al., 2022), MatchCLOT (Gossi et al., 2023), and the siamese matcher MinNet (Liu et al., 2023), all of which train two encoders with a contrastive objective and read the match off the similarity matrix, by argmax or by entropic optimal transport (Cuturi, 2013). A soft transport coupling looks like a distribution over partners, but it is a normalized similarity: it carries no posterior and no calibration, and it says nothing about how far the model would have to be wrong for the match to change.

Genomic measurements are high-dimensional, sparse, and noisy. Uncertainty quantification provides a measure of confidence in the model's prediction that can reflect the shortcomings of the data. The uncertainty score indicates which predictions are reliable enough to carry into downstream analysis (Abdar et al., 2021; Hüllermeier & Waegeman, 2021). Uncertainty quantification has been studied extensively for classification and regression in data-rich domains such as computer vision (Kendall & Gal, 2017; Filos et al., 2019; Nado et al., 2022; Franzen et al., 2024), but less so in computational genomics (Saran et al., 2026; Engelmann et al., 2022; Bajwa et al., 2024; Zhou et al., 2026). The standard methods place a distribution over the weights or over stochastic forward passes: deep ensembles (Lakshminarayanan et al., 2017), Monte Carlo dropout (Gal & Ghahramani, 2016), and mean-field variational inference (Blundell et al., 2015; Wen et al., 2018), and the resulting score is used to abstain on unreliable predictions (Geifman & El-Yaniv, 2017) and to flag inputs from outside the training distribution (Hendrycks & Gimpel, 2017; Ovadia et al., 2019). In single-cell analysis, where uncertainty is reported at all, it is placed on the cell type label (De Donno et al., 2023; Engelmann et al., 2022; López-De-Castro et al., 2025) or on the latent representation of one cell (Gayoso et al., 2021; Ashuach et al., 2023); none of these places it on the correspondence between two cells. Moreover, representation matching differs in that the prediction is a relation between two inputs, so the uncertainty is a property of the pair rather than of either element, and it may arise from either encoder or from the geometry of the space in which the two are compared. Related questions have been addressed in retrieval and metric learning through probabilistic embeddings and calibrated similarity scores (Oh et al., 2019; Chun et al., 2021), but uncertainty on a cross-modal correspondence has received limited attention, and is rarely evaluated against a known ground-truth pairing.

We present BASILISC, Bayesian Alignment of Siamese-Integrated Latents for Inferring Single-cell Concordance. BASILISC learns biologically meaningful relations between single-cell modalities and places the uncertainty score on the correspondence itself, and we ask whether that score iden-

tifies the matches a model should decline to make. We build it at two levels of the same problem: matching at the cell-type level and the cell level. At both levels we report predictive performance for the matching tasks, compare it to the published benchmark methods, and report what the uncertainty score adds: whether it separates wrong matches from correct ones, whether highly uncertain matching pairs point to underlying biology, and whether it flags a cell type the model never saw during training. Our contributions are three. First, we place an uncertainty score on the correspondence between the learned representations of two modalities, rather than on a label, a latent, or a translated profile, and evaluate it against a known ground-truth pairing; to our knowledge this has not been done before. Second, we show what the score delivers at two levels of granularity, the cell type and the cell, on three data types: it is higher on wrongly matched pairs and on cells of a type the model never saw, so a matcher can decline a match and flag an out-of-distribution input; every comparison is against a deterministic read of the same assignment and pre-registered against a stated bar, and we state where the score is limited. Third, we show that the uncertainty reflects a source: the pairs the model is unsure of are cells of one lineage family, the learned similarity follows lineage, and the correct matches with the highest uncertainty are candidates for cells in transition, where chromatin opens ahead of transcription.

The rest of the paper is organized as follows. Section 2 defines the matching task, Section 3 the two BASILISC models, and Section 4 the experimental setup. Section 5 reports our results on evaluating matching predictive performance and the uncertainty quantification at the cell-type level and at the cell level, on the multiome benchmark, with the replication on CITE-seq in the supplement. Section 6 concludes the paper, states what the uncertainty scores on correspondences do and do not deliver, the limitations of the study, and an outlook to future research.

2 Multimodal single-cell data integration

We treat diagonal integration as a supervised correspondence problem. Given a cell observed in one modality and a cell observed in another, the model predicts a score between 0 and 1 to indicate whether the two are the same cell, or, at a coarser level, cells of the same type. Training uses paired data, where every cell is measured in both modalities and the pairing is known; at test time the model is asked about cells it has not seen. The match score has to carry across the technical variation of a new dataset, generalize to rare and unseen cell types, and support the two uses a matching model has in practice: transferring labels from a reference atlas to a new dataset, and finding, for a query cell, its counterpart in the other modality.

The data we use in our experiments are the multimodal single-cell benchmark of the NeurIPS 2021 integration competition (Luecken et al., 2021; Lance et al., 2022), a curated collection of human bone marrow mononuclear cells from ten donors profiled at four sites with two assay types that together cover three types of single cell modalities. The multiome assay (10x Genomics Chromium Single Cell Multiome) measures gene expression (GEX) and chromatin accessibility (ATAC) in the same nucleus, and the CITE-seq assay (Stoeckius et al., 2017) measures gene expression (GEX) and surface protein abundance (ADT) in the same cell. The three modalities carry different kinds of signals (supplement, Section A), so an integration method needs to represent each modality on its own terms and learn a mapping between the modality-specific representations. Moreover, each site–donor combination is a technical batch, and the systematic difference between batches, from sample handling, sequencing depth, and the donor, is the batch effect that every integration method has to survive: remove the batch effect while preserving biologically meaningful signal in data. A summary of both datasets is given in Table S1 in the supplement, together with the number of cells per type and per batch (Tables S3, S4, and S5).

The competition's match-modality task asks, for each cell in one modality of a held-out test set, which cell in the other modality is its partner. The test set is the $20{,}009$ nuclei of one site, in three batches, and the score is the average probability the method places on the true partner (the cell level in our experiments); a one-to-one assignment scores exactly its accuracy. The winning entry, CLUE (Tu et al., 2022), reached 0.058, and MatchCLOT (Gossi et al., 2023), published after the competition, 0.0755; both are deterministic. We use the same test set and score at the cell level, so that our accuracy is read beside theirs.

3 METHOD

BASILISC is a siamese network with a posterior over its weights, trained on paired single-cell data to score the correspondence between a cell observed in one modality and a cell observed in the other. We build two versions of BASILISC's architecture, one to predict the matching at cell-type level and one for cell level. We compare different methods for making BASILISC uncertainty-aware: the posterior is approximated by mean-field variational inference (MFVI, the default), by Monte Carlo dropout, or by a deep ensemble of five networks; the architecture is the same in each case, and the deterministic network is the control. The cell-type level BASILISC is a classifier over frozen single-modality embeddings that answers whether two cells are of the same type, and the cell level BASILISC matcher trained end-to-end from raw counts that answers which cell in the other modality is the same cell. Both produce, for every prediction, an uncertainty score read from the disagreement of the posterior draws, and both are compared with a deterministic version of the same network. Figures 1 and 2 show the two architectures for the multiome data; the same architectures are used for CITE-seq, with the ATAC leg replaced by an ADT leg.

3.1 CELL-TYPE LEVEL: A BAYESIAN SIAMESE CLASSIFIER OVER FROZEN EMBEDDINGS

Inputs. Each cell enters as a 20-dimensional embedding computed by Liam (Rautenstrauch & Ohler, 2024), a variational autoencoder for single-cell data, trained here on one modality alone with no cell-type labels and no access to the pairing; one model for RNA and one for ATAC (RNA and ADT for CITE-seq). The embeddings are frozen and never fine-tuned, and the two latent spaces are unaligned (supplement), so the correspondence between them is entirely learned by the network.

Architecture (Figure 1). The network has one leg per modality, a shared block, and a merged head. Each leg is a linear layer $20 \to 64$ with GELU activation(Hendrycks & Gimpel, 2016) and layer normalization(Ba et al., 2016), with its own weights, since the two inputs live in different spaces and the three modalities have very different data characteristics. The shared block, a linear layer $64 \to 64$ with GELU and layer normalization, is applied to the output of both legs with the same weights, which is what makes the network siamese. The merged head concatenates the two outputs and maps them through a linear layer $128 \to 64$ with GELU and a another $64 \to 1$ to a logit, whose sigmoid is the predicted probability p that the two cells are of the same type.

Posterior. In the Bayesian version, the weights and biases of the two merged layers are replaced by a mean-field Gaussian posterior with a $\mathcal{N}(0, 1)$ prior, trained by variational inference (Blundell et al., 2015) with the local reparameterization(Kingma et al., 2015) trick. We call this the *base* placement. In a second placement, *legs*, every linear layer of the network is variational, legs and shared block included. We also train the same architecture with Monte Carlo dropout (dropout 0.2 on the input and the hidden layer of the merged head, kept active at test time) (Gal & Ghahramani, 2016) and as a deep ensemble of five deterministic networks (Lakshminarayanan et al., 2017), so that the posterior is compared with the deterministic network as well as two other common approximation methods.

3.2 CELL LEVEL: A BAYESIAN CONTRASTIVE MATCHER FROM RAW COUNTS

Inputs and architecture (Figure 2). Each cell enters as its raw count vector, 13,431 genes for GEX and 116,490 peaks for ATAC (134 proteins for ADT). The two legs are encoders shaped as Liam's single-modality encoders, $n_{\text{in}} \to 128 \to 128 \to 20$ with layer normalization and ReLU, and they are initialized from the weights of the trained single-modality Liam models, so that training starts from an encoder that already represents each modality on its own. Above the legs sits one shared projection, $20 \to 64 \to 32$, through which both modalities pass with the same weights, as the shared block of the cell-type-level model does. Its 32-dimensional output is ℓ_2-normalized.

Posterior. In the Bayesian version, every linear layer of the encoders and of the shared projection is variational, with the same mean-field Gaussian, prior and training as above, except the first layer of each encoder, which stays a point estimate (supplement). At evaluation, one weight sample is drawn for the whole network and applied to every cell, so that a draw is one function and the draws can be compared as assignments. Liam's dropout of 0.1 is kept in both arms as a training regularizer and is off at evaluation.

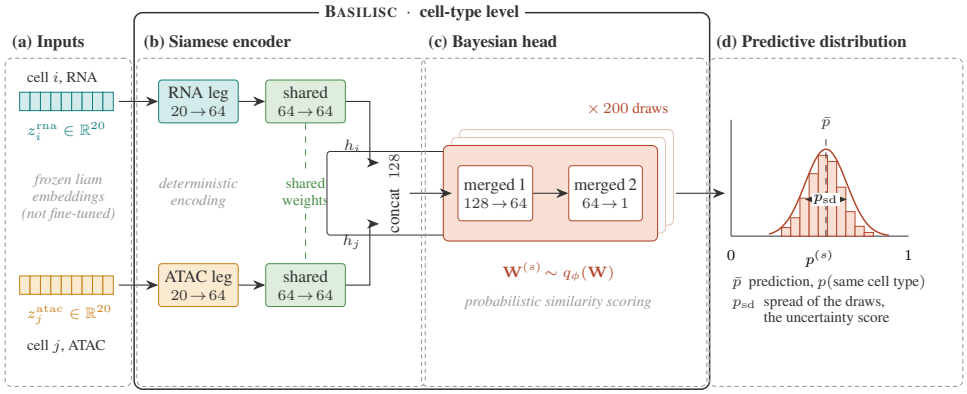


Figure 1: **BASILISC at the cell-type level: from frozen embeddings to a predictive distribution.** The boxed region, panels (b) and (c), is the siamese model with deterministic legs and shared layers and a probabilistic head; (a) is the frozen embedding as input and (d) is a readout of the predictive distribution over 200 posterior draws.

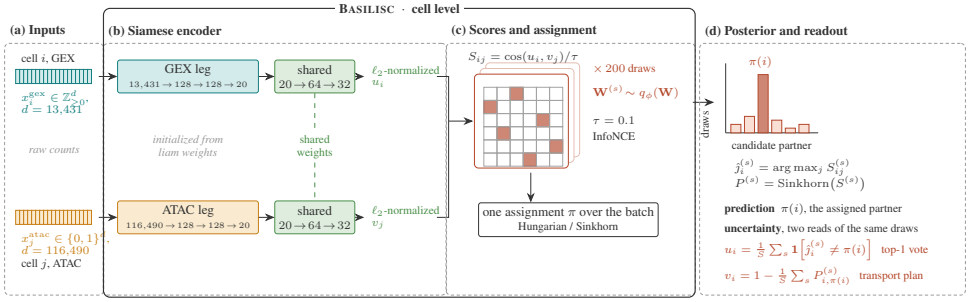


Figure 2: **BASILISC at the cell level: from raw counts to an assignment and its uncertainty.** The boxed region, panels (b) and (c), is the trained model, with legs initialized from single-modality Liam weights and a shared projection; (a) is the raw-count input, and (d) is the readout: one assignment π over the batch, and the two uncertainty scores u_i and v_i read from 200 posterior draws per seed, pooled over the five seeds.

Cell partner assignment. For a test batch, each of five seeds gives a score matrix over all (GEX, ATAC) pairs of the batch; we average the five and solve one linear assignment problem (Kuhn, 1955) on the mean matrix, which returns a one-to-one assignment π. The prediction for cell i is its assigned partner $\pi(i)$, and the matching accuracy is the share of cells whose assigned partner is their own; under the benchmark's match-modality score, a one-to-one assignment scores exactly its accuracy, so our number is read beside the published ones. See the supplement for details of training, calculating uncertainty scores, and deterministic controls.

4 EXPERIMENTAL FRAMEWORK

The experiments ask, three things of the uncertainty score: whether it separates wrong predictions from correct ones, whether abstaining on it removes the errors that matter, and whether it is high on cells of a type the model never trained on. Each question is asked under a controlled construction of the data, and each answer is read beside the deterministic control, so that what the posterior adds is separated from what the network already knew.

Arms. At the cell-type level we compare the Bayesian network with the posterior on the merged head (BNN), the same network with the posterior on every layer, Monte Carlo dropout (MCD), a five-member deep ensemble (ENS), and the deterministic network (DET). Three baselines use the same frozen embeddings and the same pair labels with no network: k-nearest-neighbor label

transfer, orthogonal Procrustes with Platt scaling, and a logistic pair classifier (supplement). At the cell level we compare the MFVI matcher (BNN) with the deterministic matcher (DET), each as a single seed and as a five-seed deep ensemble, against the published scores of CLUE (Tu et al., 2022) and MatchCLOT (Gossi et al., 2023) on the same test set.

Data constructions. We train and evaluate on five constructions of the same data, which differ in what is withheld at training. In *mix*, every cell type and batch is seen in training and the test pairs are new cells of seen types. In leave-one-cell-type-out (*LOCTO*), one cell type is removed from training and appears only at test time, while its close relatives stay in training; this is the hardest novelty case, because the held-out type has a near neighbor the model knows. In leave-one-family-out (*LOFO*), a whole lineage family (Table S2) is removed, so the held-out types have no close relative in training, and leave-two-families-out (*LTFO*) removes two families. In the few-shot variant of LOFO, $k \in \{0, 5, 10, 20, 30, 50\}$ cells of each type of the held-out family are returned to training. At the cell level we use the competition split (*challenge*) and its LOCTO variant. Every construction is run for five seeds. The full definitions, the held-out lists, and the pair populations are given in the supplement.

5 Results

At the cell-type level every number is a mean over five seeds on the multiome data unless stated, and the held-out designs pool 22 types $\times$ five seeds $\times$ two pair populations, 220 runs. At the cell level every number is on the 20,009 test cells of the competition split, five seeds, 200 posterior draws per seed. Two measures recur. Error detection is the AUROC of an uncertainty score for separating wrong predictions from correct ones. Novelty detection is the AUROC of the score for separating cells of the held-out type from seen cells, read on positive pairs at the cell-type level. Both are reported beside the score's deterministic control: the confidence margin at the cell-type level, the Sinkhorn transport weight at the cell level (supplement, Section B.5). Every analysis was specified before it was run. We recorded the expected outcome and the bar it had to clear in a dated log released with the code.

5.1 Cell-type level

Classification is solved by the representation; the probability is what the network adds. On pairs of seen types the model reaches AUROC 0.976 $\pm$ 0.002 on cross-batch pairs and 0.977 $\pm$ 0.008 within batch, accuracy 0.932, from about 70 training cells per type; holding one type out leaves the seen types at 0.975 $\pm$ 0.006 (Table S8). The frozen embeddings already carry most of this: a linear pair classifier on the same inputs reaches 0.917, and on the held-out type's pairs it matches the network (balanced accuracy 0.882 against 0.861). What the network adds in distribution is a better probability, Brier 0.048 to 0.053 for the network arms against 0.063 for the best baseline and NLL 0.168 to 0.191 against 0.218, in 5 of 5 seeds; what it adds out of distribution is the rest of this section.

The uncertainty finds the model's errors, and abstaining on it removes them. Per run, p_{sd} separates misclassified pairs from correct ones at AUROC 0.856 $\pm$ 0.016; median p_{sd} is 0.010 on correct pairs and 0.079 on errors. Discarding the least certain fifth removes 63% of the errors and raises accuracy on the rest from 0.932 to 0.975; the gain at half coverage is positive in all ten runs (Figure S2). On this job the margin ties the posterior, 0.858 against 0.856, with the same risk-coverage curve to the third decimal. This is what the delta method predicts: to first order $p_{\mathrm{sd}} = \bar{p}(1-\bar{p})\,\sigma_z$, and per pair p_{sd} correlates with the margin at Spearman 0.95 to 0.99 at every posterior width we trained (Table S7).

A type never seen in training. On the held-out type's own pairs (47% positive) AUROC is 0.920 $\pm$ 0.057 against 0.975 on the seen types in the same runs, and recall 0.816 against 0.947. Per run, p_{sd} separates unseen positives from seen ones at 0.627 $\pm$ 0.138, above 0.5 in 19 of 22 types; the deterministic network's margin gives 0.564 and 0.582 on the two populations, and every sampled arm beats it (t from 3.8 to 11.5). The uncertainty is specific (Figure 3C–E). Among an unseen type's negatives, median p_{sd} is 0.057 when the partner is the nearest seen type and 0.0015 for any other type, a 38-fold ratio, separated per run at AUROC 0.81 $\pm$ 0.08. The model is certain about what the unseen cell is not and unsure only about its relative. For label transfer from an atlas this is the useful

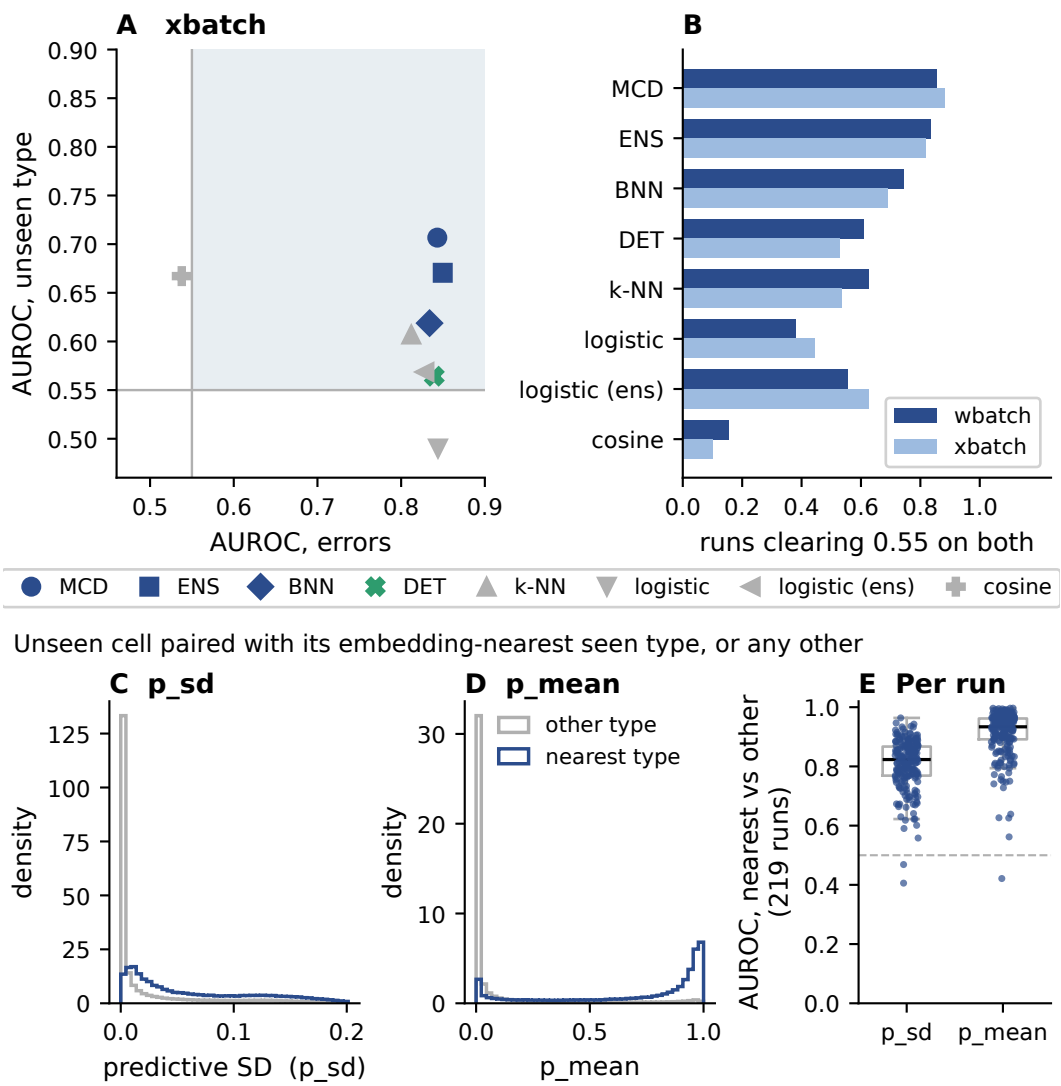


Figure 3: Uncertainty at the cell-type level. (A) Each method's own uncertainty score on the 110 cross-batch LOCTO runs: AUROC for detecting its own errors on seen-type pairs against AUROC for detecting the held-out type's positives among seen positives; the shaded region is above 0.55 on both. (B) The share of runs in which one score clears 0.55 on both jobs, for both pair populations. (C–E) An unseen type's negatives, split by whether the partner is the seen type nearest to it in the input embedding or any other seen type: (C) predictive SD and (D) predictive mean of the two groups, pooled over 219 runs; (E) per-run AUROC of each score for separating the two groups, one point per run, dashed line at chance.

case: a cell of a type the atlas lacks is not assigned a random label; it is assigned its nearest relative's label with a raised uncertainty, so the transfer can be flagged for review rather than accepted or refused outright.

One score does both jobs, and no baseline does. Figure 3A–B shows each method's own score on error detection and on novelty detection over 110 LOCTO runs per population (Table S9). On novelty every posterior beats the margin: MC dropout 0.707, the five-member ensemble 0.671, MFVI 0.619, against 0.564 for the deterministic network. The cosine similarity of the frozen embeddings flags unseen types at 0.667, so we do not claim that the posterior flags novelty better than every alternative. The share of runs in which one score clears 0.55 on both jobs is 88% and 85% for MC dropout, 82% and 84% for the ensemble, 69% and 75% for MFVI, 53% and 61% for the

deterministic network, 45% and 38% for the linear pair classifier, and 10% and 15% for the cosine. Each baseline has one half: the cosine finds novelty and is at chance on its own errors (0.54), the linear map finds its errors and is at chance on novelty (0.49). The deterministic arm is the identical network with sampling switched off, so the difference between 53% and 69% is the posterior alone. Where the posterior sits decides what it carries. Over a 200-fold range of posterior width classification does not move while novelty detection rises from 0.584 with a collapsed posterior to 0.627 at the standard ELBO weight and saturates above half that width (Table S7); placing the same posterior on every layer instead of the merged block alone leaves classification and error detection unchanged and raises novelty detection from 0.627 to 0.683, $+0.056$ [$+0.036$, $+0.076$] over 220 paired runs.

The uncertainty reflects lineage. The correspondence the model learns carries the structure of the biology. The 22×22 matrix of mean predictions from the LOCTO runs reads as a lineage tree (Figure S3): every held-out type's nearest neighbor among the seen types is a lineage relative, so a type the model never saw is matched to its family. Against the embedding's centroid cosine, Spearman is $+0.68$ over 462 type pairs and the nearest neighbor is identical for 16 of 22 types; at matched row entropy the model puts 0.573 of each row on the held-out type's own family and the embedding 0.568 ($p = 0.48$), so the model preserves the embedding's structure and extends it to new types rather than sharpening it. The uncertainty follows the same structure. An unseen cell is uncertain only against its relative, a 38-fold difference in median p_{sd} (Figure 3C–E), and a wrong match at this level is a match to a relative. Cells of one lineage family resemble each other across modalities, and cells with the same label are not all in one state: chromatin opens ahead of transcription along a differentiation path (Ma et al., 2020; Martin et al., 2021), so two cells of one type can differ in how far accessibility has run ahead of expression; the correct matches the model is least sure of are those of the types it half-takes for a neighbor, and one reading, offered as an interpretation, is that the uncertainty picks up cells in transition.

Without a relative in training the model abstains. Holding out a whole family instead of one type, recall on the family's positives falls from 0.82 to 0.49 and median $\bar{p}$ from 0.86 to 0.48, while the novelty separation of p_{sd} rises from 0.63 to 0.73. Returning k cells of the held-out family to training brings recall to 0.93 at $k = 20$ on both populations, against 0.94 on seen types, and adding cells stops helping between 20 and 30 (Figure S4; 1,260 runs). Which cells are returned does not matter: the cells most confusable with a seen type are worse than random by 0.015 recall, the cells that spread the family out no better ($p = 0.40$). Every claim of this level replicates on the CITE-seq data, 38 types, 390 runs, at AUROC 0.970 and 0.963 on the two pair populations (Table S10).

5.2 Cell level

Accuracy, and where the errors go. The deep ensemble matches 0.0624 of the test cells to their own partner, 95% CI [0.0591, 0.0656], above CLUE's 0.058 (Tu et al., 2022) with the interval clear of it; MatchCLOT's 0.0755 (Gossi et al., 2023) is not reached. Five MFVI seeds averaged reach 0.0478 [0.0447, 0.0507]; the mean-field posterior costs accuracy at cell level. The errors are structured, and equally so in both arms: 88.8% and 88.4% of wrong assignments land on another cell of the query's own type, against 14.3% with the partners permuted within batch, and of the 11% that leave the type, 76.2% and 74.6% go to a lineage relative against 15.3% permuted (Figure S5). An accuracy of 0.06 at the level of the cell sits on near-perfect matching at the level of the type.

Which errors the posterior sees. Pooled over the three test batches, the transport-plan posterior v_i separates wrong assignments from correct ones at AUROC 0.804 against 0.792 for the Sinkhorn weight of the same mean matrix at its best temperature, $+0.013$ [$+0.009$, $+0.017$]: above zero and under the $+0.02$ we set in advance as the bar for helping (Table 1). Splitting the errors by kind changes the picture (Figure S6). On the 11% that cross types, which we call harmful because a partner of another type carries the wrong label into anything downstream, the posterior leads in every batch, u_i against the control $+0.060$, $+0.031$, $+0.027$, all above the bar, and still leads with the control's temperature chosen on that target. On the 89% that stay in the type it does not (-0.007 to -0.043): a wrong partner of the same type earns nearly as much transport mass as the right one, so the deterministic plan is confident either way, and that is where its pooled advantage comes from. The posterior trades detection of the harmless errors for detection of the harmful ones. Keeping the most certain half of the assignments by u_i removes 32.8% of the cross-type errors; by the Sinkhorn

Table 1: Detecting a wrong assignment, by the kind of error. AUROC of each score for separating wrong assignments of the given kind from correct matches, Bayesian arm, pooled over the three test batches of the competition split (20,009 cells; 19,053 wrong, of which 16,852 same-type and 2,201 cross-type; of the cross-type errors 1,642 go to a lineage relative and 559 to another family).

score	any wrong	same-type	cross-type	cross-type, partner is	
				a relative	another family
transport-plan posterior v_i	0.804	0.821	0.679	0.676	0.686
top-1 vote u_i	0.777	0.790	0.678	0.672	0.697
Sinkhorn weight, best T ($T = 0.5$)	0.792	0.812	0.635	0.646	0.603
Sinkhorn weight, $T = 1$	0.785	0.804	0.639	0.641	0.631
seed stability	0.701	0.707	0.656	0.654	0.662
row margin	0.604	0.602	0.618	0.606	0.652

weight, 25.1%. Every score rates a cross-type error as more certain than a same-type one (posterior 0.35 on that contrast): the harmful errors are the confident errors. Its edge grows with the distance of the wrong partner. On errors to a lineage relative (1,642) u_i reaches 0.672 against 0.646 for the control; on errors to another family (559) 0.697 against 0.603, per batch $+0.038$, $+0.017$, $+0.013$ on relatives and $+0.117$, $+0.073$, $+0.071$ on other families (Figure S7). The control is more confident about a wrong match to another family than to a relative; the posterior is less.

A type never seen in training, at the cell level. Holding one of eight types out of training, held-out cells still get a same-type partner in 0.60 to 0.96 of cases, within 0.05 of the model trained on all types in the deterministic arm and within 0.11 in the Bayesian arm (Figure S8; 50 draws per seed). As a novelty score the Sinkhorn weight separates held-out from seen cells at 0.67 to 0.85 for the five types with at least 751 test cells and is below chance for pDC, CD4$^+$ T naive, and HSC: a small held-out type has few same-type competitors in its batch, so its assignment is more confident than a seen cell's. The posterior adds $+0.016$ on average, above the bar in 3 of 8 types. Assignment confidence carries a type-size dependence that no score we tested removes; the novelty claim of this paper is made at the cell-type level.

6 Discussion

We put uncertainty on the correspondence in a multimodal matching model, at two levels of granularity, the cell type and the cell, on three single-cell data types, and asked where it works and where it is limited. At the cell-type level one score is higher on the pairs the model gets wrong and on the cells of a type it never trained on, and no deterministic baseline does both; an unseen type is matched to its lineage relative and is uncertain only against that relative, so the uncertainty reflects the cell lineage structure the model has learned without ever seeing it. At the cell level, the posterior flags the wrong matches that cross a type boundary, the harmful ones, better than a deterministic read of the same scores, and its lead grows with the distance of the wrong partner; both matchers reach the published methods' accuracy. Where it is limited: on seen types the predictive SD ranks pairs as the confidence margin does, so abstaining on it works but is not a gain over a deterministic network. The separation of an unseen type is above chance in 19 of 22 types and rises when its whole lineage is absent, but it is modest, and MC dropout and a plain cosine similarity flag unseen types as well or better. At the cell level the score does not detect a never-seen type, because a small held-out type has few competitors in its batch, and on the errors that stay inside a type the deterministic transport plan is the better detector. Our empirical framework comes with limitations, which we set out here. The cell-type level works on frozen embeddings, so what the model learns about lineage is the embedding's structure carried into a correspondence, and the classification itself is within reach of a linear map on the same inputs. The Sinkhorn control's temperature is chosen on the test set, which favors the control. Our study covers three single-cell modalities, but it's limited to one tissue type, which we leave for future studies. Three things follow: the posterior should sit on the encoders as well as on the correspondence head, since on every layer it raised novelty detection at no cost to classification; a novelty score for one-to-one assignment has to correct for the number of competitors a cell has; and the correct matches with a high uncertainty are where to look for where chromatin and transcriptome part ways.

AI USE STATEMENT

The research question, the model design and the first implementation predate this work and were done completely without AI assistance. That earlier version was written manually in TensorFlow Probability and run on a subset of the Multiome data. From that code, we used AI assistants to port the implementation to PyTorch, refactor it, and scale it to the full datasets. The cell-level model and most of the experiments reported here were implemented the same way. We also used AI assistants for feedback on experimental design, for computing the summary statistics and tables in the supplement, for searching for related work, for producing and editing the figures. The main manuscript text is written manually and polished, and complemented by AI. Parts of the supplement text are generated by AI and verified manually. An author was in the loop at every step of any task mentioned here where AI was used, and we reviewed all AI-assisted work. We did not use generative AI to generate synthetic research data, scope the project, develop theoretical models, design the research question, design the paired dataset generation settings, design experimental scenarios, design Siamese model architecture.

ETHICS STATEMENT

The aim of this work is to learn a richer multimodal representation of the cell, which can in turn help in understanding biological processes and disease mechanisms. We foresee no ethical issue with the present work. The data are the publicly released benchmark datasets described in Section 2; no personal data were processed.

REPRODUCIBILITY STATEMENT

Every reported number in the paper is produced by code that will be made available in the accompanying code base at the time of publication. Random seeds are fixed and reported for all experiments for reproducibility. Every analysis carried unit tests, and each test was verified to break on a deliberately broken input. Experiments were run on a shared Linux HPC cluster under Slurm, on NVIDIA GPUs ranging from Tesla V100 to H100 (16 to 80 GB); one cell-type-level model trains in about one GPU-minute and one cell-level model in 10 to 45 minutes, depending on the card, so the full set of runs is reproducible on a single GPU within days.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2021.05.008. URL https://www.sciencedirect.com/science/article/pii/S1566253521001081.

Xiuli An, Vincent P. Schulz, Jie Li, Kunlu Wu, Jing Liu, Fumin Xue, Jingping Hu, Narla Mohandas, and Patrick G. Gallagher. Global transcriptome analyses of human and murine terminal erythroid differentiation. *Blood*, 123(22):3466–3477, May 2014. ISSN 1528-0020. doi: 10.1182/blood-2014-01-548305. URL http://dx.doi.org/10.1182/blood-2014-01-548305.

Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multimodal single-cell data. *Genome Biology*, 21(1):111, May 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02015-1. URL http://dx.doi.org/10.1186/s13059-020-02015-1.

Ricard Argelaguet, Anna S. E. Cuomo, Oliver Stegle, and John C. Marioni. Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, 39(10):1202–1215, May 2021. ISSN 1546-1696. doi: 10.1038/s41587-021-00895-7. URL http://dx.doi.org/10.1038/s41587-021-00895-7.

Tal Ashuach, Mariano I. Gabitto, Rohan V. Koodli, Giuseppe-Antonio Saldi, Michael I. Jordan, and Nir Yosef. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231, 2023. doi: 10.1038/s41592-023-01909-9.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Ayesha Bajwa, Ruchir Rastogi, Pooja Kathail, Richard W. Shuai, and Nilah M. Ioannidis. Characterizing uncertainty in predictions of genomic sequence-to-activity models. In *Proceedings of the 19th Machine Learning in Computational Biology (MLCB) meeting*, volume 261 of *Proceedings of Machine Learning Research*, pp. 279–297. PMLR, 2024.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks, 2015.

Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, October 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2688. URL http://dx.doi.org/10.1038/nmeth.2688.

Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015. ISSN 1476-4687. doi: 10.1038/nature14590. URL http://dx.doi.org/10.1038/nature14590.

Kai Cao, Xiangqi Bai, Yiguang Hong, and Lin Wan. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, 36(Supplement 1):i48–i56, 2020. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaa443. URL http://dx.doi.org/10.1093/bioinformatics/btaa443.

Kai Cao, Yiguang Hong, and Lin Wan. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics*, 38(1):211–219, 2022. doi: 10.1093/bioinformatics/btab594.

Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, May 2022. ISSN 1546-1696. doi: 10.1038/s41587-022-01284-4. URL http://dx.doi.org/10.1038/s41587-022-01284-4.

Song Chen, Blue B. Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12):1452–1457, 2019. doi: 10.1038/s41587-019-0290-0.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8411–8420. IEEE, 2021. doi: 10.1109/cvpr46437.2021.00831. URL http://dx.doi.org/10.1109/cvpr46437.2021.00831.

Frank Cichocki, Bartosz Grzywacz, and Jeffrey S. Miller. Human nk cell development: One road or many? *Frontiers in Immunology*, 10:2078, August 2019. ISSN 1664-3224. doi: 10.3389/fimmu.2019.02078. URL http://dx.doi.org/10.3389/fimmu.2019.02078.

Darren A. Cusanovich, Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015. doi: 10.1126/science.aab1601.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Carlo De Donno, Soroor Hediyeh-Zadeh, Amir Ali Moinfar, Marco Wagenstetter, Luke Zappia, Mohammad Lotfollahi, and Fabian J. Theis. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nature Methods*, 20(11):1683–1692, 2023. doi: 10.1038/s41592-023-02035-2.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Scot: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, January 2022. ISSN 1557-8666. doi: 10.1089/cmb.2021.0446. URL http://dx.doi.org/10.1089/cmb.2021.0446.

Yi Ding, Marieke Lavaert, Simon Grassmann, Victor I. Band, Liang Chi, Arundhoti Das, Sumit Das, Christelle Harly, Susannah C. Shissler, Justin Malin, Dingkang Peng, Yongge Zhao, Jinfang Zhu, Yasmine Belkaid, Joseph C. Sun, and Avinash Bhandoola. Distinct developmental pathways generate functionally distinct populations of natural killer cells. *Nature Immunology*, 25(7):1183–1192, 2024. ISSN 1529-2916. doi: 10.1038/s41590-024-01865-2. URL http://dx.doi.org/10.1038/s41590-024-01865-2.

Jan Engelmann, Leon Hetzel, Giovanni Palla, Lisa Sikkema, Malte D. Luecken, and Fabian J. Theis. Uncertainty quantification for atlas-level cell type transfer. *arXiv preprint arXiv:2211.03793*, 2022.

Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks, 2019.

Jannik Franzen, Claudia Winklmayr, Vanessa E. Guarino, Christoph Karg, Xiaoyan Yu, Nora Koreuber, Jan P. Albrecht, Philip Bischoff, and Dagmar Kainmueller. Arctique: An artificial histopathological dataset unifying realism and controllability for uncertainty quantification. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2411.07097.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, 2016.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L. Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, 2021. doi: 10.1038/s41592-020-01050-x.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Boying Gong, Yun Zhou, and Elizabeth Purdom. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biology*, 22:351, 2021. doi: 10.1186/s13059-021-02556-z.

Federico Gossi, Pushpak Pati, Panagiotis Chouvardas, Adriano Luca Martinelli, Marianna Kruithof-de Julio, and Maria Anna Rapsomaniki. Matching single cells across modalities with contrastive learning and optimal transport. *Briefings in Bioinformatics*, 24(3):bbad130, 2023. doi: 10.1093/bib/bbad130.

Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, April 2018. ISSN 1546-1696. doi: 10.1038/nbt.4091. URL http://dx.doi.org/10.1038/nbt.4091.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Peng Huang, Yongzhong Zhao, Jianmei Zhong, Xinhua Zhang, Qifa Liu, Xiaoxia Qiu, Shaoke Chen, Hongxia Yan, Christopher Hillyer, Narla Mohandas, Xinghua Pan, and Xiangmin Xu. Putative regulators for the continuum of erythroid differentiation revealed by single-cell transcriptome of human BM and UCB cells. *Proceedings of the National Academy of Sciences*, 117(23):12868–12876, May 2020. ISSN 1091-6490. doi: 10.1073/pnas.1915085117. URL http://dx.doi.org/10.1073/pnas.1915085117.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021. doi: 10.1007/s10994-021-05946-3.

Isabel E. Ishizuka, Michael G. Constantinides, Herman Gudjonson, and Albert Bendelac. The innate lymphoid cell precursor. *Annual Review of Immunology*, 34(1):299–316, May 2016. ISSN 1545-3278. doi: 10.1146/annurev-immunol-041015-055549. URL http://dx.doi.org/10.1146/annurev-immunol-041015-055549.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Christopher Lance, Malte D. Luecken, Daniel B. Burkhardt, Robrecht Cannoodt, Pia Rautenstrauch, Anna Laddach, Aidyn Ubingazhibov, Zhi-Jie Cao, Kaiwen Deng, Sumeer Khan, Qiao Liu, Nikolay Russkikh, Gleb Ryazantsev, Uwe Ohler, Angela Oliveira Pisco, Jonathan Bloom, Smita Krishnaswamy, and Fabian J. Theis. Multimodal single cell data integration challenge: results and lessons learned. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pp. 162–176. PMLR, 2022. URL https://proceedings.mlr.press/v176/lance22a.html.

Chaozhong Liu, Linhua Wang, and Zhandong Liu. Single-cell multi-omics integration for unpaired data by a siamese network with graph-based contrastive loss. *BMC Bioinformatics*, 24:5, 2023. doi: 10.1186/s12859-022-05126-7.

Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. Jointly embedding multiple single-cell omics measurements. In *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, volume 143 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 10:1–10:13, 2019. doi: 10.4230/LIPIcs.WABI.2019.10.

Marcos López-De-Castro, Alberto García-Galindo, José González Gomariz, and Rubén Armañanzas. Conformal inference for reliable single cell RNA-seq annotation. *Bioinformatics*, 41(10):btaf521, 2025. doi: 10.1093/bioinformatics/btaf521.

Leif S. Ludwig, Caleb A. Lareau, Erik L. Bao, Satish K. Nandakumar, Christoph Muus, Jacob C. Ulirsch, Kaitavjeet Chowdhary, Jason D. Buenrostro, Narla Mohandas, Xiuli An, Martin J. Aryee, Aviv Regev, and Vijay G. Sankaran. Transcriptional states and chromatin accessibility underlying human erythropoiesis. *Cell Reports*, 27(11):3228–3240.e7, 2019. ISSN 2211-1247. doi: 10.1016/j.celrep.2019.05.046. URL http://dx.doi.org/10.1016/j.celrep.2019.05.046.

Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019. doi: 10.15252/msb.20188746.

Malte D. Luecken, Daniel Bernard Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann T. Chen, Louise Deconinck, Angela M. Detweiler, Alejandro A. Granados, Angela Oliveira Pisco, Smita Krishnaswamy, Fabian J. Theis, and Jonathan M. Bloom. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=gN35BGa1Rt.

Sai Ma, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K. Kartha, Tristan Tay, Travis Law, Caleb Lareau, Ya-Chieh Hsu, Aviv Regev, and Jason D. Buenrostro. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183(4):1103–1116.e20, November 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.09.056. URL http://dx.doi.org/10.1016/j.cell.2020.09.056.

Eric W. Martin, Jana Krietsch, Roman E. Reggiardo, Rebekah Sousae, Daniel H. Kim, and E. Camilla Forsberg. Chromatin accessibility maps provide evidence of multilineage gene priming in hematopoietic stem cells. *Epigenetics & Chromatin*, 14(1):2, January 2021. ISSN 1756-8935. doi: 10.1186/s13072-020-00377-1. URL http://dx.doi.org/10.1186/s13072-020-00377-1.

Mónica Itzel Martínez-Gutiérrez, Irma Cañedo Solares, Fernando Gómez-Chávez, Liliana Monserrat Molina-López, Regina Flores-Rabasa, Juan Alejandro Magdaleno-Villanueva, José Israel León-Pedroza, and Dolores Correa. Do human b-1 lymphocytes truly exist? *Frontiers in Immunology*, 17:1810144, April 2026. ISSN 1664-3224. doi: 10.3389/fimmu.2026.1810144. URL http://dx.doi.org/10.3389/fimmu.2026.1810144.

Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael W. Dusenberry, Sebastian Farquhar, Qixuan Feng, Angelos Filos, Marton Havasi, Rodolphe Jenatton, Ghassen Jerfel, Jeremiah Liu, Zelda Mariet, Jeremy Nixon, Shreyas Padhy, Jie Ren, Tim G. J. Rudner, Faris Sbahi, Yeming Wen, Florian Wenzel, Kevin Murphy, D. Sculley, Balaji Lakshminarayanan, Jasper Snoek, Yarin Gal, and Dustin Tran. Uncertainty baselines: Benchmarks for uncertainty robustness in deep learning, 2022.

Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embeddings. In *International Conference on Learning Representations*, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Amit A. Patel, Yan Zhang, James N. Fullerton, Lies Boelen, Anthony Rongvaux, Alexander A. Maini, Venetia Bigley, Richard A. Flavell, Derek W. Gilroy, Becca Asquith, Derek Macallan, and Simon Yona. The fate and lifespan of human monocyte subsets in steady state and systemic inflammation. *Journal of Experimental Medicine*, 214(7):1913–1923, 2017. ISSN 1540-9538. doi: 10.1084/jem.20170355. URL http://dx.doi.org/10.1084/jem.20170355.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Anna Maria Ranzoni, Andrea Tangherloni, Ivan Berest, Simone Giovanni Riva, Brynelle Myers, Paulina M. Strzelecka, Jiarui Xu, Elisa Panada, Irina Mohorianu, Judith B. Zaugg, and Ana Cvejic. Integrative single-cell rna-seq and atac-seq analysis of human developmental hematopoiesis. *Cell Stem Cell*, 28(3):472–487.e7, March 2021. ISSN 1934-5909. doi: 10.1016/j.stem.2020.11.015. URL http://dx.doi.org/10.1016/j.stem.2020.11.015.

Pia Rautenstrauch. liam_manuscript_reproducibility, 2024. Version v0.0.0.

Pia Rautenstrauch and Uwe Ohler. Liam tackles complex multimodal single-cell data integration challenges. *Nucleic Acids Research*, 52(12):e52–e52, 06 2024. ISSN 0305-1048. doi: 10.1093/nar/gkae409. URL https://doi.org/10.1093/nar/gkae409.

Patrick Fernandes Rodrigues and Roxane Tussiwand. Novel concepts in plasmacytoid dendritic cell (pdc) development and differentiation. *Molecular Immunology*, 126:25–30, October 2020. ISSN 0161-5890. doi: 10.1016/j.molimm.2020.07.006. URL http://dx.doi.org/10.1016/j.molimm.2020.07.006.

Patrick Fernandes Rodrigues, Llucia Alberti-Servera, Anna Eremin, Gary E. Grajales-Reyes, Robert Ivanek, and Roxane Tussiwand. Distinct progenitor lineages contribute to the heterogeneity of plasmacytoid dendritic cells. *Nature Immunology*, 19(7):711–722, 2018. ISSN 1529-2916. doi: 10.1038/s41590-018-0136-9. URL http://dx.doi.org/10.1038/s41590-018-0136-9.

Sepideh Saran, Mahsa Ghanbari, and Uwe Ohler. Uncertainty-aware deep learning for genomics applications: Insights from an empirical study. *arXiv preprint arXiv:2608.11054*, 2026.

Wen Shao, Yifeng Wang, Qian Fang, Wenjuan Shi, and Hai Qi. Epigenetic recording of stimulation history reveals blimp1–bach2 balance in determining memory b cell fate upon recall challenge. *Nature Immunology*, 25(8):1432–1444, 2024. ISSN 1529-2916. doi: 10.1038/s41590-024-01900-2. URL http://dx.doi.org/10.1038/s41590-024-01900-2.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017. doi: 10.1038/nmeth.4380.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.05.031. URL http://dx.doi.org/10.1016/j.cell.2019.05.031.

Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020. doi: 10.1038/s41587-019-0379-5.

Amos Tanay and Aviv Regev. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637):331–338, 2017. doi: 10.1038/nature21350.

Xinming Tu, Zhi-Jie Cao, Chen-rui Xia, Sara Mostafavi, and Ge Gao. Cross-linked unified embedding for cross-modality representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://openreview.net/forum?id=Tfb73TeKnJ-.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Joshua D. Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z. Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887.e17, 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.05.006. URL http://dx.doi.org/10.1016/j.cell.2019.05.006.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches, 2018.

Kevin E. Wu, Kathryn E. Yost, Howard Y. Chang, and James Zou. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences*, 118(15):e2023070118, 2021. doi: 10.1073/pnas.2023070118.

Masashi Yukawa, Sajjeev Jagannathan, Sushmitha Vallabh, Andrey V. Kartashov, Xiaoting Chen, Matthew T. Weirauch, and Artem Barski. Ap-1 activity induced by co-stimulation is required for chromatin opening during t cell activation. *Journal of Experimental Medicine*, 217(1):e20182009, October 2020. ISSN 1540-9538. doi: 10.1084/jem.20182009. URL http://dx.doi.org/10.1084/jem.20182009.

Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017. doi: 10.1038/ncomms14049.

Jessica Zhou, Kaeli Rizzo, Trevor Christensen, Ziqi Tang, and Peter K. Koo. Uncertainty-aware genomic deep learning with knowledge distillation. *npj Artificial Intelligence*, 2026. doi: 10.1038/s44387-025-00053-3. Preprint: bioRxiv 2024.11.13.623485.