Proceedings of Machine Learning Research 179:1–19, 2022 Conformal and Probabilistic Prediction and Applications

# Uncertainty Estimation for Single-cell Label Transfer

**Robin Khatri** ROBIN.KHATRI@ZMNH.UNI-HAMBURG.DE
**Stefan Bonn** SBONN@UKE.DE
*Institute of Medical Systems Biology, Center for Biomedical AI (bAIome) – University Medical Center Hamburg-Eppendorf, Hamburg, Germany*

**Editor:** Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo and Lars Carlsson

## Abstract

Single-cell gene expression matrices require a cell type label for each cell for downstream analysis. A cell type label refers to a heterogeneous group to which a cell belongs. Machine learning algorithms that aim to automate the assignment of cell type labels train on reference datasets for which cell type labels are already defined. However, these methods are prone to error due to possible preprocessing errors and the dynamic nature of cellular states. Therefore, it is essential to measure the uncertainty associated with classifications. Here, we hypothesize that conformal prediction may provide a principled approach for this. We examine inductive conformal classifiers (ICPs) on the task of single-cell label transfer. ICPs lead to well-calibrated models that quantify uncertainties well. Results are motivating, and the uncertainties are intuitive and easy to interpret. We also consider a confidence-credibility conformal predictions setup that accurately predicts single labels with the desired error level. Such a model can also reject the classification of cell types unobserved in the reference dataset. However, the presence of unknown cell types violates the underlying assumption of a conformal predictor and is highly dependent on the quality of batch correction. We envision more work in detecting unknown cell types and using conformal predictions to evaluate batch correction methods.

**Keywords:** single-cell RNA-seq, single cell classification, conformal prediction

## 1. Introduction

Single-cell RNA sequencing (scRNA-seq) techniques measure mRNA expression from individual cells. The ability to analyze mRNAs at a single cell level has allowed biologists to identify new cell states and understand their dynamics and fate (Lähnemann et al., 2020). scRNA-seq results in a count matrix with genes and numeric expression of those genes per cell. In this count matrix, there are thousands of cells and genes (>10,000 variable genes). In order to analyze this resulting data, clustering tools such as Uniform Manifold Approximation and Projection (UMAP) are used to identify the cell types based on gene expression patterns of marker genes - the genes that are abundantly expressed only on cells belonging to specific cell types (Zhang et al., 2019). However, this process is time-consuming and requires an expert. As a result, it is imperative to simplify the process and utilize existing knowledge from already labeled datasets. We refer to this task as single-cell label transfer.

When dealing with single-cell datasets from different sources, we can term the already labeled dataset(s) as the reference and the dataset we are interested in labeling as the query. While both datasets may come from the same tissue and, as a result, may share a large number of cell types, the cells themselves may exhibit differences. These differences can either be biological or technical (Tran et al., 2020). A biological difference is between tissue

states (*e.g.* disease vs. healthy) and must be observed. Therefore, a method to identify cells based on a reference is inherently subjected to this, and will likely be uncertain for differentially distributed data points. The technical differences arise due to the use of different technologies and sequencing libraries, different ways to prepare tissue samples, etc. and can cause technical bias. Since our interest is in identifying cells in reference-based query, it is vital to remove this technical bias. Several methods termed batch correction or single-cell data integration methods have been proposed to perform this task (Korsunsky et al., 2019; Lotfollahi et al., 2019; Hie et al., 2019; Hausmann et al., 2022). Batch correction is a preprocessing step for many cell type classification methods. The objective is to obtain a representation of original gene expressions of reference and query in a joint space where the datasets are well integrated. Some batch correction methods such as Harmony (Korsunsky et al., 2019) and Scanorama (Hie et al., 2019) align clusters computed on principal components; the result is a co-embedding which can be used to identify cell types. Batch correction is not perfect and can lead to errors, especially in cases of partial overlap between cell types present between reference and query (Hie et al., 2019). Here, however, we assume that both datasets share similar distribution after correction with some noise depending on the batch correction method used.

The many sources of differences between a reference and a query single-cell dataset make it necessary to be aware of the uncertainty associated with the method that transfers knowledge from the reference to the query. In the case of single-cell label transfer, we are interested in the transfer of discrete cell type labels, and as such, the use of machine learning algorithms is natural. Here, we wanted to ask about the current state of identifying uncertainty associated with methods in use. In our experiments, we refer to single-cell label transfer as a two step process where we sequentially use batch correction and classification algorithms. State-of-the-art single-cell classification methods such as scPred (Alquicira-Hernandez et al., 2019) involve training an SVM and setting a heuristic threshold on output probabilities to filter out potentially wrong classifications. We argue that a heuristic-based threshold method depends entirely on the output probabilities and does not consider the likelihood of a data point to be from the dataset it was trained on. If we were to find cells where our model is confused between two or more cell types, we again have to resort to setting a threshold on the cell types other than the majority cell type label. It raises another question, if we were to change our algorithm, will the same heuristic be valid? Hence, this problem of choosing a probability threshold is largely dataset- and classification algorithm- dependent. A better approach is to identify uncertainties that consider how well the test samples conform to what model has observed and which can work agnostically to any classification algorithm.

Conformal prediction provides a natural approach to solve the aforementioned problem. Conformal classification is model agnostic and provides certain theoretical guarantees. Moreover, it is simple and general. Conformal classification has emerged as a method to measure distribution-free uncertainty in a range of applications. In terms of bioinformatics applications, it has been widely used to avoid the use of toxic drugs (Eklund et al., 2015; Alvarsson et al., 2021). We describe conformal classification in Section 2.1. In this work, we evaluate conformal classification on the task of single-cell label transfer using six human scRNA-seq datasets. We present the results in Section 3. We discuss those results and future directions in Section 4.

## 2. Methods

### 2.1. Inductive Conformal prediction (ICP)

The assumption of conformal prediction is the probabilistic exchangeability of data points. Exchangeability is a weaker assumption than IID. (Probabilistic) Exchangeability refers to likeliness of permutations of data order, *i.e.* for n data points $(x_1, y_1), (x_2, y_2)..., (x_n, y_n)$, all $n!$ permutations are equally likely (Shafer and Vovk, 2008). Here we consider a variant of conformal predictors, Inductive Conformal Predictors (ICP). We briefly describe the procedure of an ICP setup.

Given a classifier $h : x \rightarrow y$, where $(x, y)$ is a training example, we need a nonconformity measure $f_{nc}$ which quantifies how likely an instance is to be from training examples. Various non-conformity measures exist such as inverse probability function which relies on the output probability of the $h$, and is given as, $f_{nc}(x_i, j, h) = 1 - (h(x_i))_j$, where $(h(x_i))_j$ is the predicted probability of the data point $x_i$ to belong to class $j$.

To make predictions with ICP, we perform data split to measure non-conformities on a held-out dataset. In particular, the training set is divided into two distinct datasets, commonly referred to as *proper training set* and *calibration set*. Mondrian approach to calibration can be adopted to ensure similar guarantees for all classes by having one calibration set per class (Vovk et al., 2003).

After creating data splits, we train a machine learning model, which in this case is a classifier. Then, during calibration, the scores of non-conformity measures are evaluated using predictions for all data points within the calibration set from the trained classifier. Since we have access to the actual labels for each of these data points, non-conformity scores for the true class can be calculated. During prediction time, for each data point, non-conformity scores are computed for each class and assigned either one label, multiple labels, or no label, depending on the non-conformity score for a class and the non-conformity measures for all calibration data points within that class. Thereafter, $p$-value, $p_{(x,y)}$, of an instance $x$ for label $y$, can be given by the proportion of instances observed in calibration set that are either equally or more non-conforming when compared to this instance. At significance level $\epsilon$, all labels $y_i$ satisfying $p(x, y_i) > \epsilon$ are assigned to the test instance. In case where no calibration instance is less likely to belong to any label than the test instance, the prediction results in an empty set. These empty sets indicate an erroneous classification or test instance being from a different distribution than all instances in the calibration set. Proportion of prediction sets of different sizes differ with the confidence levels. Further, since $p$-values cannot be considered probabilites over a finite test set, we considered confidence-credibility framework to classify at desired error rates. The confidence-credibility framework is described in Section 2.1.1.

#### 2.1.1. Confidence-credibility framework

$p$-values obtained from a conformal classifier can not be used to guarantee an observed error rate over a finite test dataset. In order to relate $p$-values to probabilities over this dataset, we refer to what is known as confidence-credibility predictions (Papadopoulos, 2008). Here the aim is to assign a single label to each sample in the test set with some confidence and credibility. Confidence is given as 1 - second largest $p$-value. This is the highest confidence

at which the output is a single label. Credibility measures how likely a sample is to come from training set. It is defined as the largest $p$-value.

In order to use this setup to predict a single label while expecting at most $K$ errors on the test set, we can use the following procedure as described in Linusson et al. (2018):

1. For each data point in the test set of size $n$, make predictions on a test sample $i$ and obtain a triplet $(\hat{y}_i, \gamma_i, \mu_j)$, where $\hat{y}_i$ is the most likely single label, $\hat{\gamma}_i$ is the confidence for the label and $\mu_i$ is the credibility.
2. Set total number of errors tolerable on the test set, $K$, and obtain $\hat{k}_i = n(1 - \gamma_i)$.
3. Assign label $\hat{y}_i$ to the test sample $i$, if $\hat{k}_i \leq K$.

This procedure can be label-conditional to provide an expected maximum error for each label.

## 3. Experiments and results

In this section, we detail datasets used, define our experimental setup and present our results. In brief, for each experiment, we investigate calibration errors for ICPs to identify if a model is well-calibrated. Secondly, we evaluate the distribution and patterns of single, multi, and empty prediction sets. Finally, in order to make point predictions, we make use of the confidence-credibility setup and, as a comparison, provide its accuracy with base algorithm (SVM) with threshold = 0.7 as used in Lotfollahi et al. (2022). In all experiments, data split for proper training and calibration sets for ICP set up is 80%-20%.

### 3.1. Datasets

We used six scRNA-seq datasets from human peripheral blood mononuclear cells (PBMCs) and pancreas tissues from different sequencing technologies. The datasets are listed in Table 1.

| **Tissue** | **Dataset name** | **Original Source** | **Technology** |
|---|---|---|---|
| PBMC | PBMC 8k | 10x Genomics | 10x |
| | PBMC 6k | 10x Genomics | 10x |
| Pancreas | inDrop | Baron et al. (2016) | inDrop |
| | SS2 | Segerstolpe et al. (2016) | Smart-seq 2 |
| | CEL-Seq 2 | Muraro et al. (2016) | CEL-Seq 2 |
| | Fluidigm C1 | Lawlor et al. (2017) | Fluidigm C1 |
| | CEL-seq | Grün et al. (2016) | CEL-Seq |

Table 1: Details of datasets. The second column, *Dataset name* refers to the name of dataset as used in this work.

We downloaded PBMC datasets from the website of 10x Genomics[1] and we manually labeled cell types for each dataset separately. To label PBMC datasets, we used marker

1. https://www.10xgenomics.com

genes listed in Table 2 and tools from Scanpy (Wolf et al., 2018). In addition, we obtained the count matrices of pancreas datasets from Hie et al. (2019) along with cell type labels. Here, we restricted cell types to Bcells, Monocytes, CD4Tcells, CD8Tcells, and NK cells for PBMCs and to Alpha, beta, gamma, delta, acinar, ductal, endothelial, stellate cells for the pancreas dataset.

| **Cell type** | **Genes** |
|---|---|
| Bcells | MS4A1, CD19 |
| CD4Tcells | IL7R, CD4 |
| CD8Tcells | CD8A |
| Monocytes | LYZ, FCGR3A |
| NK | GNLY, NKG7 |

Table 2: Marker genes used to define cell types in PBMCs.

### 3.2. Preprocessing, feature selection and classification

First, we subset the reference and query datasets with the genes present in both. Then, to remove the low-quality cells, we filtered out genes expressed in less than 3 cells and cells that express less than 200 genes. We also removed cells expressing more than 4% mitochondrial genes. After that, we transformed the datasets to $\log_2(1+\text{CPM})$, where CPM refers to counts per million. Then we centered the dataset using Mean-variance scaling to have zero mean and unit variance. Finally, we subsetted the datasets to have 1,000 most highly variable genes (HVGs) and combined them. On the resulting data matrix, we computed principal components (PCs). After initial preprocessing, the number of cell types per dataset are provided in Tables 3 and 4.

| **Dataset** | **Bcells** | **CD4Tcells** | **CD8Tcells** | **Monocytes** | **NK** | **Total** |
|---|---|---|---|---|---|---|
| PBMC 6k | 704 | 2,240 | 714 | 1,397 | 301 | 5,536 |
| PBMC 8k | 992 | 1,975 | 890 | 1,870 | 320 | 6,047 |

Table 3: Number of cell types in PBMC datasets.

| **Dataset** | **Alpha** | **Beta** | **Gamma** | **Delta** | **Acinar** | **Ductal** | **Endo.** | **Stellate** | **Total** |
|---|---|---|---|---|---|---|---|---|---|
| inDrop | 2,249 | 3,048 | 260 | 613 | 272 | 898 | 689 | 362 | 8,391 |
| SS2 | 1,109 | 796 | 219 | 142 | 103 | 462 | 67 | 63 | 2,961 |
| CEL-Seq 2 | 885 | 600 | 125 | 199 | 170 | 315 | 31 | 101 | 2,426 |
| Fluidigm C1 | 241 | 300 | 12 | 21 | 6 | 27 | 12 | 13 | 632 |
| CEL-Seq | 220 | 341 | 21 | 66 | 162 | 402 | 37 | 22 | 1,271 |

Table 4: Number of cell types in pancreas datasets.

As described in Section 1, we can use batch correction methods before training a machine learning classifier to remove technical bias. Here, we use Harmony (Korsunsky et al., 2019) on 50 PCs for this purpose. Harmony first computes soft clusters for each cell on a PCA

embedding and then performs iterative correction of clusters to ensure that similar cells are closely clustered. As it is beyond the scope here to describe Harmony in detail, we refer the reader to the original publication for a detailed overview of Harmony. The final dataset, thus, contains 50 PCs that are corrected using Harmony. In our experiments, we report visualizations of UMAP embeddings to ensure sufficient data integration.

As an underlying algorithm, we use support vector machines (SVMs). SVMs have been used in single-cell classification previously and have achieved state-of-the-art performance (Alquicira-Hernandez et al., 2019). Further, a threshold can be set on output probabilities to only make confident classifications. Previous methods have used thresholded SVM to filter out potentially wrong predictions and have achieved good results (Alquicira-Hernandez et al., 2019; Lotfollahi et al., 2022). Therefore, SVM is a natural choice for us to use with conformal prediction for our task.

### 3.3. PBMCs

To begin with, we experimented on the two PBMC datasets. Distributions of the counts per cell type are largely similar between the two datasets, with CD4Tcells being the most abundant cell type (Table 3). We trained on one of the two datasets and tested on the other. Since the sources of training and test sets are different in these experiments, this could lead to validation issues of CP. To mitigate this, we first aligned training and test distributions using Harmony. Then, we visualize PC- and UMAP- embeddings before and after batch correction with Harmony in Figure 1. Visually, the datasets are well-integrated. Here, compared to Monocytes and B cells, several overlapping CD8Tcells, CD4Tcells, and NK cells exist. This difference is likely due to the biological similarities of CD4Tcells, CD8Tcells, and NK cells.

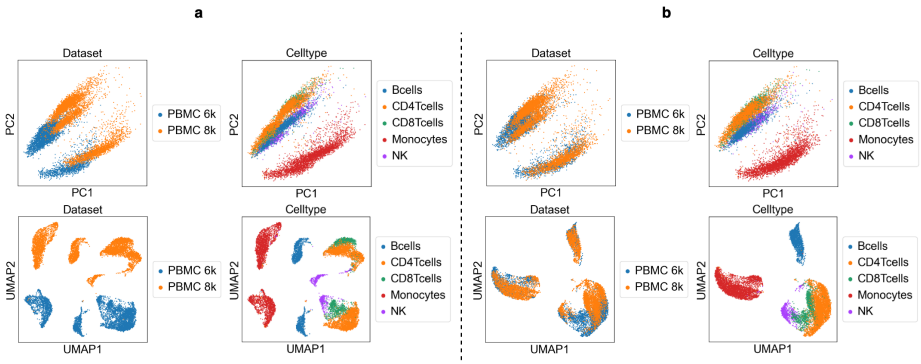


Figure 1: PC- and UMAP-embeddings for **a**: Original and **b**: Harmony corrected dataset.

We evaluated the base algorithm, SVM, on both original and batch corrected datasets. We considered two settings: experiment 1 - Training set: PBMC 6k, test set: PBMC 8k, and experiment 2 - Training set: PBMC 8k, test set: PBMC 6k. The confusion matrices of resulting models are given in Table 5. Without batch correction, relatively distinct cells, Bcells, and Monocytes are well-classified. However, in both experiments, the accuracies for CD8Tcells are almost 0 (experiment 1 - CD8Tcells: 0, and experiment 2 - CD8Tcells: 0.008). Almost all CD8Tcells are identified as CD4Tcells. After batch correction, this is slightly

better (experiment 1 - CD8Tcells: 0.12, and experiment 2 - CD8Tcells: 0.17). There can be two reasons for this poor performance on CD8Tcells. The first is the inherent difference in cell types (this is unlikely as both datasets are from healthy individuals and cell types are annotated using similar procedures), and the second is the improper batch correction or simply an inaccurate classifier. It should also be noted that a hierarchical classification approach can also be considered in this case to employ a two-step classification process where first, Monocytes and lymphocytes are classified. Then lymphocytes are classified into Tcells, NK, and Bcells (Alquicira-Hernandez et al., 2019). However, since we are interested in evaluating uncertainties, our objective lies in understanding whether ICP leads to similar performance as the underlying algorithm, and whether the resulting uncertainties make sense.

**(a) PBMC 6k to PBMC 8k**

| | | Bcells | CD4Tcells | CD8Tcells | Monocytes | NK |
|---|---|---|---|---|---|---|
| Before batch Correction | Bcells | 0.88 | 0.11 | 0 | 0.01 | 0 |
| | CD4Tcells | 0 | 0.99 | 0 | 0.01 | 0 |
| | CD8Tcells | 0 | 0.99 | 0 | 0.01 | 0 |
| | Monocytes | 0 | 0.04 | 0 | 0.96 | 0 |
| | NK | 0 | 0.52 | 0 | 0.21 | 0.27 |
| | | Bcells | CD4Tcells | CD8Tcells | Monocytes | NK |
| After batch Correction | Bcells | 0.99 | 0.01 | 0 | 0 | 0 |
| | CD4Tcells | 0 | 0.97 | 0.02 | 0.01 | 0 |
| | CD8Tcells | 0 | 0.88 | 0.12 | 0 | 0 |
| | Monocytes | 0 | 0.02 | 0 | 0.98 | 0 |
| | NK | 0 | 0.04 | 0.15 | 0.02 | 0.79 |

**(b) PBMC 8k to PBMC 6k**

| | | Bcells | CD4Tcells | CD8Tcells | Monocytes | NK |
|---|---|---|---|---|---|---|
| Before batch Correction | Bcells | 0.88 | 0.08 | 0 | 0.04 | 0 |
| | CD4Tcells | 0 | 0.97 | 0.02 | 0.01 | 0 |
| | CD8Tcells | 0 | 0.91 | 0.01 | 0.01 | 0.07 |
| | Monocytes | 0 | 0 | 0 | 1 | 0 |
| | NK | 0 | 0.11 | 0 | 0.04 | 0.85 |
| | | Bcells | CD4Tcells | CD8Tcells | Monocytes | NK |
| After batch Correction | Bcells | 0.98 | 0.01 | 0 | 0.01 | 0 |
| | CD4Tcells | 0 | 0.88 | 0.1 | 0.02 | 0 |
| | CD8Tcells | 0 | 0.44 | 0.17 | 0 | 0.39 |
| | Monocytes | 0 | 0 | 0 | 1 | 0 |
| | NK | 0 | 0.04 | 0.01 | 0 | 0.95 |

Table 5: Confusion matrices for experiments - 1 (Training set: PBMC 6k, Test set: BMC 8k) and 2 (Training set: PBMC 8k, Test set: BMC 6k) **a**: before and **b**: after batch correction.

Next, we classified the same datasets with ICP and evaluated the quality of the calibration. Figure 2 shows the error rates on calibration set across significance levels. For the non-Mondrian approach, the error rates differ widely between cell types. This is understandable as the number of cells differs per cell type (Tables 3). In contrast, for the Mondrian approach, errors on the calibration set are uniform across cell types. Next, we evaluated the performance of ICP in comparison with SVM and SVM with a threshold. Results are given in Table 6 for ICP at significance = 0.025. The choice of 0.025 is arbitrary

and is chosen to provide sufficient confidence. Here, for ICP, we include all predictions, regardless of the size of the prediction set. We compare both averages as well as overall accuracies. Average accuracy favors all cell types equally, regardless of size of the cell type cluster. Accuracies of ICPs are slightly better than the others.

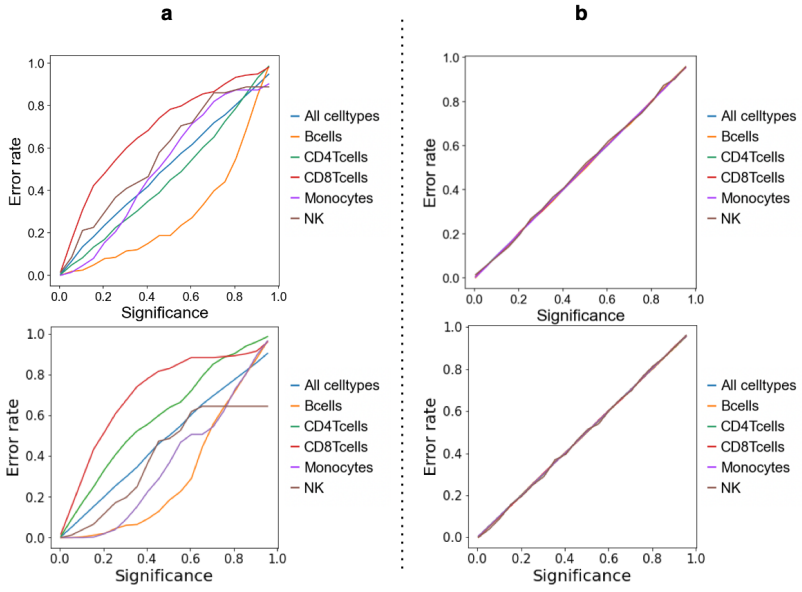


Figure 2: Error rates on calibration set for **a**: non-Mondrian and **b**: Mondrian approaches. The first row shows the results for experiment 1 and the second row shows the results for experiment 2.

| **Method** | **Test set** | **Average accuracy** | **Overall accuracy** |
|---|---|---|---|
| SVM | PBMC 8k | 0.771 | 0.844 |
| SVM | PBMC 6k | 0.616 | 0.834 |
| SVM (thr 0.7) | PBMC 8k | 0.842 | 0.868 |
| SVM (thr 0.7) | PBMC 6k | 0.742 | 0.863 |
| ICP (With SVM) | PBMC 8k | 0.935 | 0.854 |
| ICP (With SVM) | PBMC 6k | 0.790 | 0.868 |

Table 6: Comparison of SVM, SVM with threshold = 0.7 and ICP. For ICP, significance, $\epsilon$ = 0.025 and all prediction sets are considered. Average accuracy refers to average of per cell type accuracy and overall accuracy refers to accuracy across all cell types.

While ICP gives better average accuracy, it should be noted that to evaluate the results properly, we must consider both prediction set sizes and errors over the finite test set. To accomplish this, we first looked at the error rate per cell type over the test set (Figure 3). For experiment 1 (Test set: PBMC 8k), the error rates for CD8Tcells and NK are higher, while for experiment 2 (Test set: PBMC 6k), the error rates for CD8Tcells are higher than

error rates for other cell types. This is in line with the errors observed from the underlying algorithm. Since errors can arise from misclassifications, *i.e.* prediction set doesn't include the ground truth cell type, or no classification, *i.e.* empty prediction set. To assess this, we computed the ZeroC (fraction of prediction sets of size 0), OneC (fraction of prediction sets of size 1), and MultiC (fraction of prediction sets of size greater than 1), and we show them at different significance levels in Figure 4. For both experiments, across all cell types, the errors can be attributed to high ZeroC. In experiment 1, for both CD8Tcells and NK cells, ZeroC is higher even at low significance levels, indicating poor credibility and errors due to empty prediction sets. For experiment 2, this is only observed for CD8Tcells.

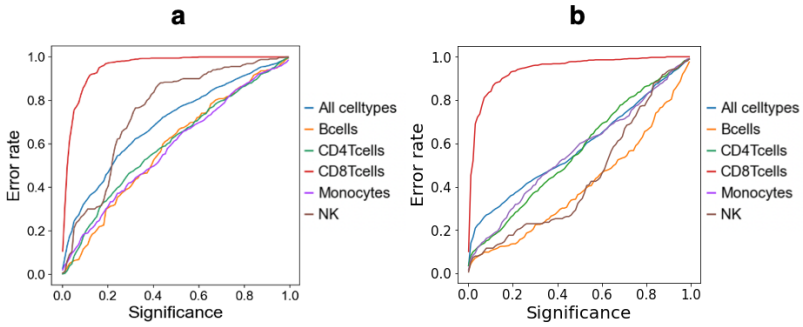


Figure 3: Error rates on test set for **a**: experiment 1 and **b**: experiment 2.

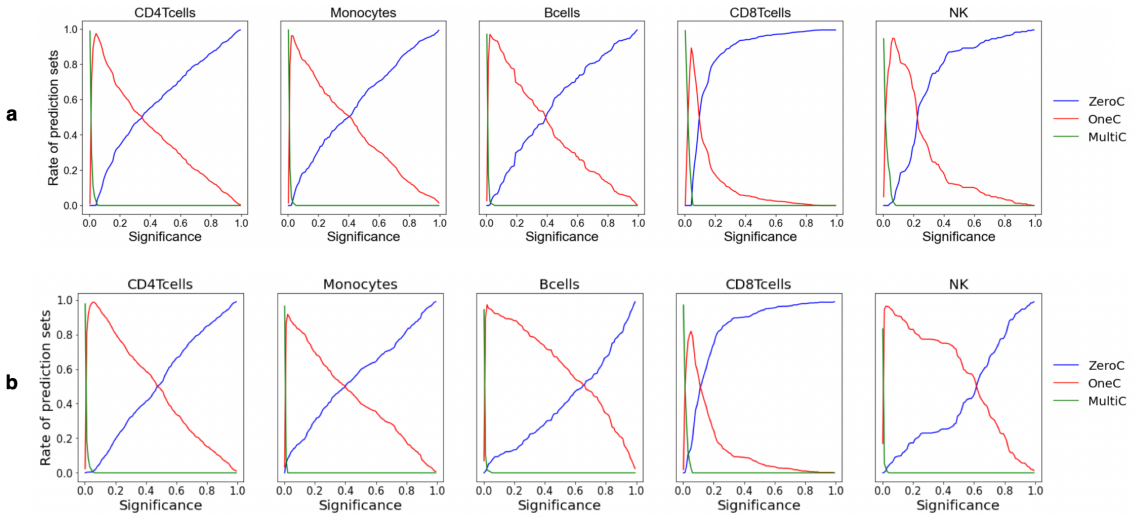


Figure 4: ZeroC, OneC and MultiC for **a**: experiment 1 and **b**: experiment 2.

Since errors are mainly attributed to empty prediction sets, we looked at the confidence-credibility predictions. Figure 5 shows the confidence and credibility scores on the UMAP embeddings for both experiments. CD8Tcells largely seem to be less credible. Therefore, we provide the average credibility per cell type in Table 7. We observed much lower credibility

for CD8Tcells (0.143) and NK cells (0.28) in experiment 1 and for CD8Tcells (0.164) in experiment 2. This is in line with the error rates on the test set (Figure 3).

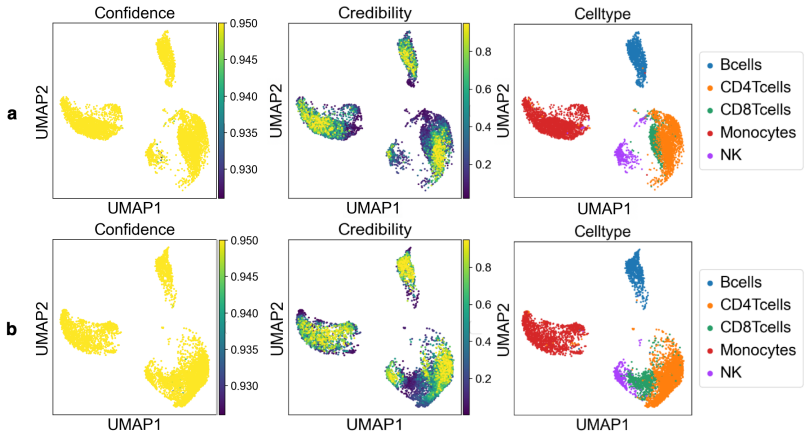


Figure 5: Confidence and credibility visualized on UMAP embeddings of test sets for **a**: experiment 1 and **b**: experiment 2.

| **Test set** | **Cell type** | **Average credibility** |
|---|---|---|
| PBMC 8k | Bcells | 0.417 |
| PBMC 8k | CD4Tcells | 0.415 |
| PBMC 8k | CD8Tcells | 0.143 |
| PBMC 8k | NK | 0.280 |
| PBMC 8k | Monocytes | 0.419 |
| PBMC 6k | Bcells | 0.60 |
| PBMC 6k | CD4Tcells | 0.482 |
| PBMC 6k | CD8Tcells | 0.164 |
| PBMC 6k | NK | 0.56 |
| PBMC 6k | Monocytes | 0.45 |

Table 7: Average credibility per cell type for experiments 1 and 2.

Using these confidence and credibility scores, we fixed $K$ at 0.025 (as a fraction of the size of the test set) and considered a label-conditional variant of the confidence-credibility setup defined in Section 2.1.1. $K = 0.025$ was chosen since at this level, model classified over 50% of all cells. We compared classification rates (*i.e.* Proportion of cells that were assigned a cell type label) and accuracy per cell type between CC-ICP (confidence-credibility setup of ICP) and SVM with a output probability threshold of 0.7. The results for experiments 1 and 2 are given respectively in Tables 8 and 9. While SVM with threshold provides more classifications and the classification rates are largely similar across cell types, for CC-ICP, only a limited number of CD8Tcells are classified (experiment 1: 0.11 and experiment 2: 0.466). However, the observed error rate is higher than the expected maximum error. To further evaluate this, we computed fractions of cells that were classified per cell type with increasing expected error rates. We show them over UMAP embeddings in Figure 6. As the expected error rate increases, more cells are classified. Experiments 1 and 2 are classified

at substantially different rates. However, the accuracy of the classified cells shows no such behavior. This is expected as $K$ balances the trade-off between the number of classifications and the expected error.

| **Method** | **Cell type** | **Classification rate** | **Accuracy** |
|---|---|---|---|
| SVM (thr 0.7) | Bcells | 0.982 | 1 |
| SVM (thr 0.7) | CD4Tcells | 0.956 | 0.988 |
| SVM (thr 0.7) | CD8Tcells | 0.808 | 0.068 |
| SVM (thr 0.7) | NK | 0.94 | 0.987 |
| SVM (thr 0.7) | Monocytes | 0.968 | 0.975 |
| CC-ICP | Bcells | 0.968 | 0.99 |
| CC-ICP | CD4Tcells | 0.564 | 0.98 |
| CC-ICP | CD8Tcells | 0.11 | 0.946 |
| CC-ICP | NK | 0.638 | 0.931 |
| CC-ICP | Monocytes | 0.649 | 0.99 |

Table 8: Comparison of classification rates and accuracies for SVM with threshold 0.7 and confidence-credibility inductive conformal predictions (CC-ICP) for experiment 1.

| **Method** | **Cell type** | **Classification rate** | **Accuracy** |
|---|---|---|---|
| SVM (thr 0.7) | Bcells | 0.974 | 0.99 |
| SVM (thr 0.7) | CD4Tcells | 0.859 | 0.897 |
| SVM (thr 0.7) | CD8Tcells | 0.757 | 0.146 |
| SVM (thr 0.7) | NK | 0.94 | 0.987 |
| SVM (thr 0.7) | Monocytes | 0.82 | 0.98 |
| CC-ICP | Bcells | 0.717 | 0.99 |
| CC-ICP | CD4Tcells | 0.656 | 0.911 |
| CC-ICP | CD8Tcells | 0.466 | 0.589 |
| CC-ICP | NK | 0.94 | 0.985 |
| CC-ICP | Monocytes | 0.768 | 0.948 |

Table 9: Comparison of classification rates and accuracies for SVM with threshold 0.7 and confidence-credibility inductive conformal predictions (CC-ICP) for experiment 2.

### 3.4. Pancreas

In this section, we focus on a single-cell label transfer setting in which information from multiple sources are utilized to increase available information. Here, we are interested in evaluating conformal prediction when reference and query consist of various datasets. On the one hand, this setting is more informative, but it can also be challenging due to increased sources of errors. Here, pancreas SS2 and inDrop form query, and the remaining pancreas datasets form reference (Table 4).

Similar to experiments on PBMCs, first, we looked at the quality of batch correction (Figure 7). Qualitatively, datasets once again seem well-integrated. Similar to our experiments on PBMCs, we next looked at the calibration error rates, test error rates, the sizes of prediction sets, and confidence and credibilities. For brevity, we present these results together in Figure 8. Here, there is no clear outlier looking directly at the test error rates

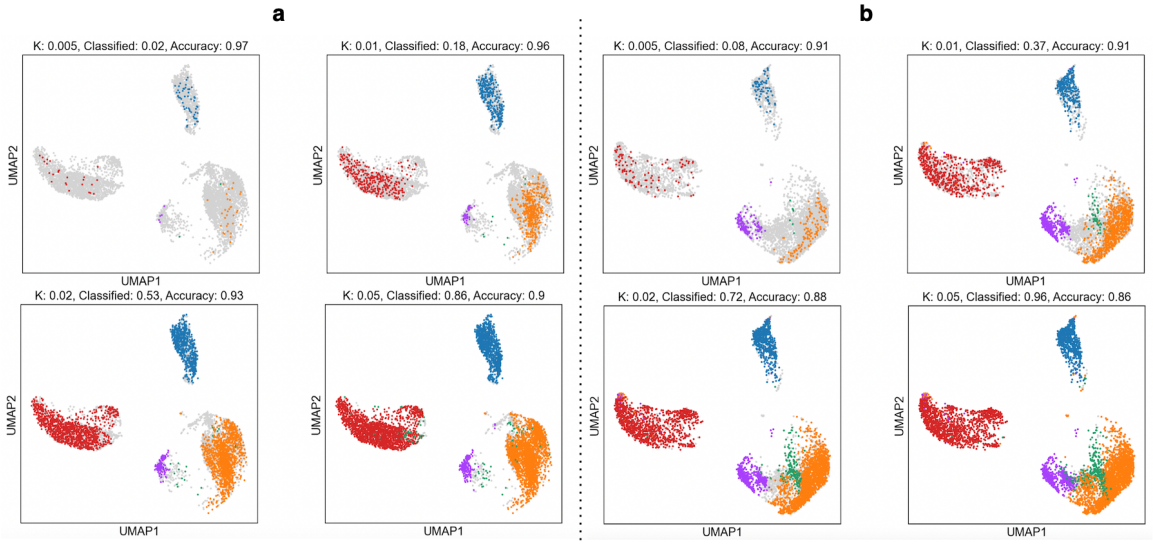


Figure 6: Classification and error rates for **a**: experiment 1 and **b**: experiment 2 for different values of $K$. $K$ is given as a fraction of the size of test set.

and the sizes of prediction sets. The comparison with other algorithms is given in Table 10 and average credibility per cell type is given in Table 11. As before, ICP, with all prediction sets considered, performs better than SVM and SVM with a threshold. For Stellate cells, the average credibility is the lowest (0.274), followed by alpha cells (0.312). We next looked at classification rates and accuracies at different values of $K$ (Figure 9). Here, the average accuracy per cell type is shown, and is much closer to the expected error rate.

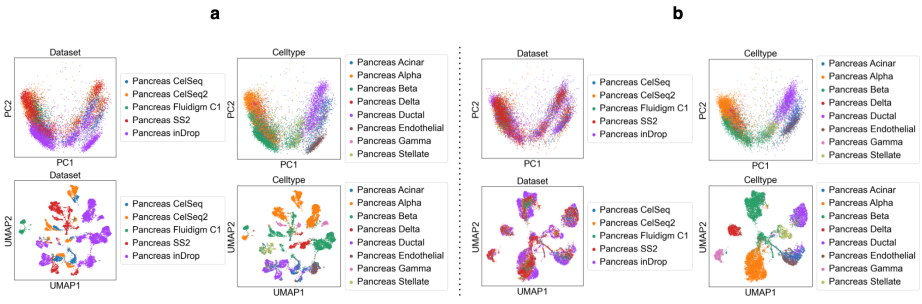


Figure 7: PC- and UMAP-embeddings for **a**: Original and **b**: Harmony corrected pancreas dataset.

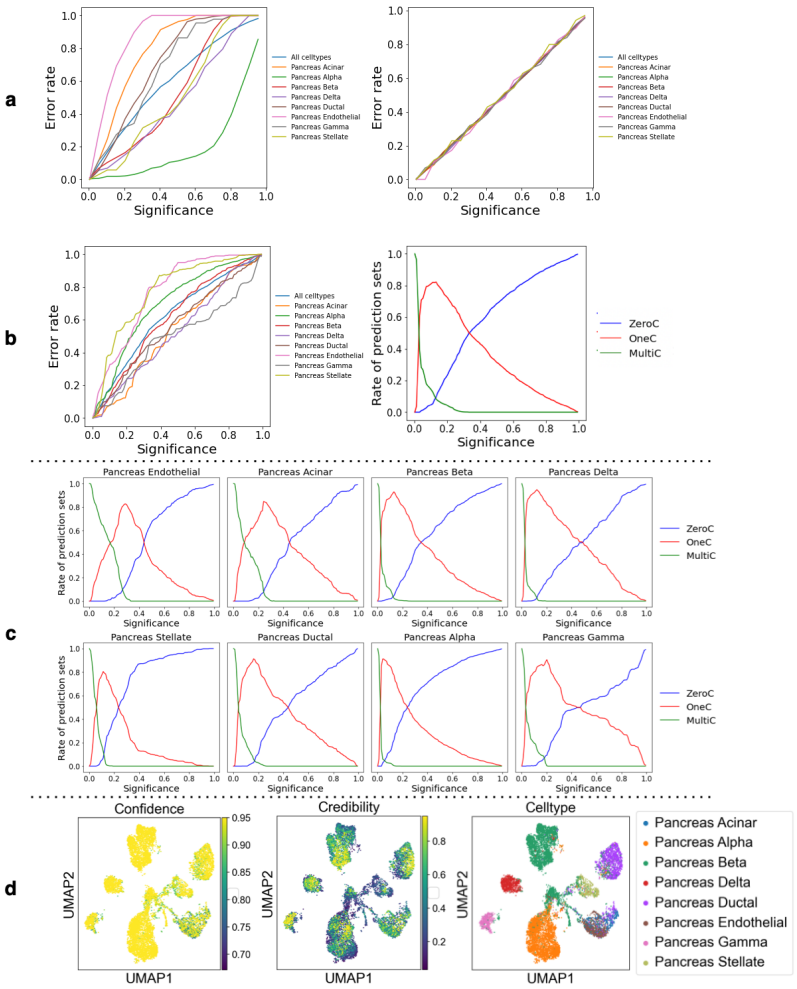


Figure 8: For pancreas dataset, **a**: Error rates on calibration set for non-Mondrian and Mondrian approaches. **b**: Error rates on test set. **c**: Sizes of prediction sets per cell type. **d**: Confidence and credibility visualized on UMAP embeddings of test set.

| **Method** | **Average accuracy** | **Overall accuracy** |
|---|---|---|
| SVM | 0.771 | 0.857 |
| SVM (thr 0.7) | 0.862 | 0.935 |
| ICP (With SVM) | 0.928 | 0.934 |

Table 10: Comparison of SVM, SVM with threshold = 0.7 and ICP. For ICP, significance $\epsilon = 0.025$, and all prediction sets are considered. Average accuracy refers to average of per cell type accuracy and overall accuracy refers to accuracy across all cell types.

| Cell type | Average credibility |
|---|---|
| Acinar | 0.508 |
| Alpha | 0.312 |
| Beta | 0.421 |
| Delta | 0.488 |
| Ductal | 0.488 |
| Endothelial | 0.473 |
| Gamma | 0.531 |
| Stellate | 0.274 |

Table 11: Average credibility per cell type.

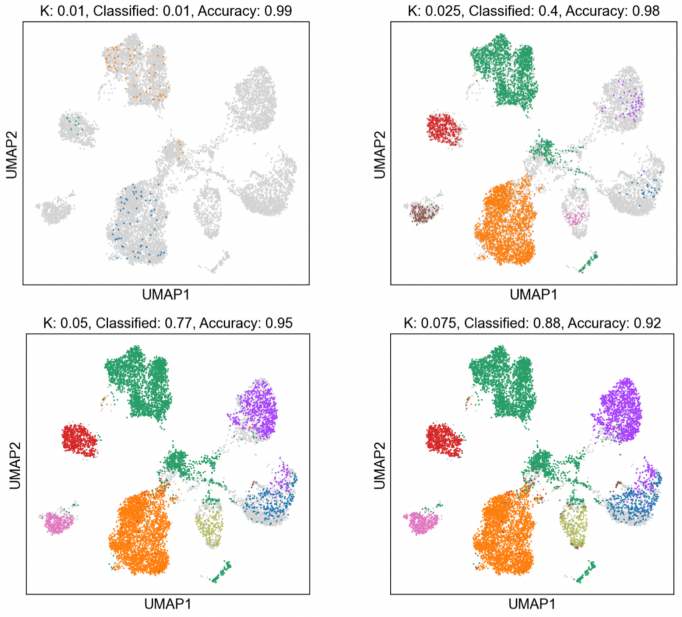


Figure 9: Classification and error rates for experiment on pancreas datasets for different $K$. $K$ is given as a fraction of the size of test set.

### 3.5. Classification of unknown cell types

Next, we wanted to investigate ICP predictions for detecting unknown cell types, *i.e.* cell types that are not observed in the training and calibration sets. While this setting violates the underlying assumption of conformal prediction, we still wanted to evaluate whether the resulting assignments for these unknown cells make sense, at least concerning the clustering, and whether the classification rates for these unknown cell types are lower compared to cell types observed in the reference. We removed each cell type from the training set while making no changes to the test set. For fair evaluation, we performed batch correction separately for each removal. This is essential as the presence of unknown cell types can affect the performance of batch correction algorithms (Hie et al., 2019). We compared assignments for the unknown cell type with average assignments on known cell types (*i.e.* cell types observed in training and test sets). The results are given in Table 12. The

results are different across cell types; however, the assignments on unknown cell types are considerably lower than those on the known cell types.

| **Unknow cell type** | **Assignment on Unknown** | **Average assignments on known** |
|---|---|---|
| Alpha | 0.08 | 0.42 |
| Delta | 0.59 | 0.64 |
| Beta | 0.68 | 0.82 |
| Ductal | 0.514 | 0.762 |
| Acinar | 0.501 | 0.761 |
| Gamma | 0.455 | 0.636 |
| Endothelial | 0.445 | 0.748 |
| Stellate | 0.379 | 0.705 |

Table 12: Classification of unknown cell types. Each row indicate an experiment where the cell type listed in the column "Unknown cell type" were removed from the training set.

### 3.6. Conformal prediction and Quality of batch correction

Lastly, we wanted to see if the higher error rate for particular cell types is due to the quality of batch correction. It should be noted that there is no "best" batch correction algorithm, and the quality of correction may differ from one dataset to other. Further, to evaluate batch correction, access to accurate cell type annotations in both the reference and query is needed. We argue that there is a relation between batch correction quality and uncertainties of conformal predictions. To confirm this, we considered two other batch correction algorithms, namely Scanorama (Hie et al., 2019) and scGen (Lotfollahi et al., 2019), and performed batch correction on PBMC datasets. We computed homogeneity scores (Rosenberg and Hirschberg, 2007) which measures the purity of clusters. High overlap between two cell type clusters would thus result in lower homogeneity. Scanorama gave the highest homogeneity score (0.812), and we considered corrected PCs from Scanorama for our experiment. We considered the setting with training dataset: PBMC 8k and test dataset: PBMC 6k. We observed a more consistent error rate over test set at different significance levels in Figure 10 and a more uniform average credibility per cell type in Table 14 compared to what we observed in Table 7 using Harmony. We envision further evaluations across other datasets and using more batch correction algorithms. However, this provides an insight into why the observed error rates could be higher for conformal prediction. It also evaluates data integration algorithms in relation to the label transfer task.

| Method | Homogeneity score |
|---|---|
| Harmony | 0.613 |
| Scanorama | 0.812 |
| scGen | 0.795 |

Table 13: Homogeneity scores for Harmony, Scanorama and scGen for correction of PBMC datasets.

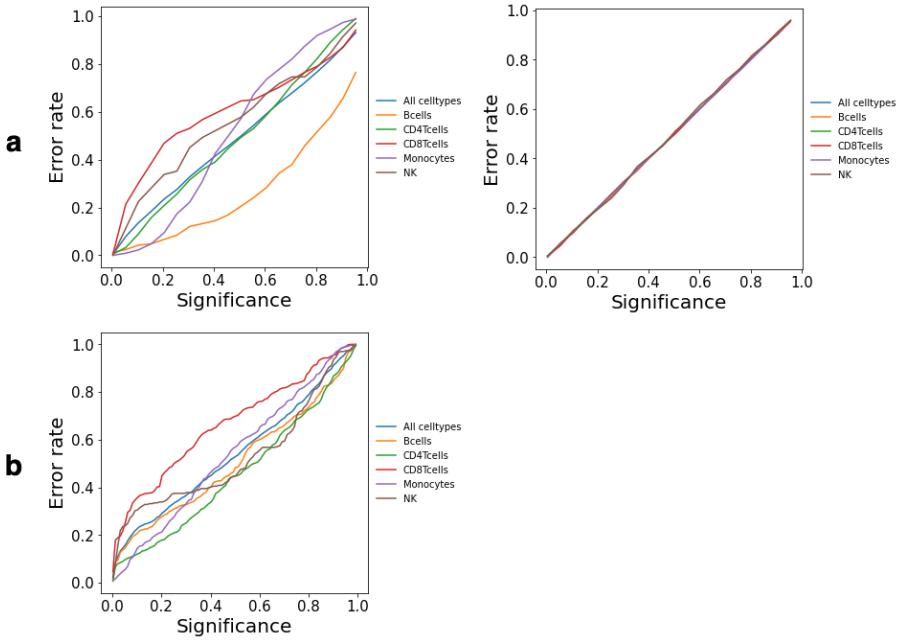


Figure 10: Results with using ICP on Scanorama corrected dataset: **a**: Error rates on the calibration set using non-mondrian and mondrian approaches. **b**: Error rates on test set.

| Cell type | Average credibility |
|---|---|
| Bcells | 0.587 |
| CD4Tcells | 0.460 |
| CD8Tcells | 0.440 |
| Monocytes | 0.498 |
| NK | 0.487 |

Table 14: Average credibility per cell type using ICP on Scanorama corrected datasets.

## 4. Discussion

We have presented a use case of conformal classification in single-cell label transfer. In both PBMC and pancreas datasets, conformal classifiers were well-calibrated. However, due to poor batch correction, the data may not be well integrated, and therefore may not wholly follow the exchangeability criterion. Consequently, we utilized the confidence-credibility framework and identified regions of uncertainty accurately. This setup allowed to classify cell types with desired error rates. Interestingly, these regions correspond to regions with cell type overlap and poor integration, such as CD8Tcells in the PBMC dataset and stellate cells in the pancreas datasets. These uncertainty patterns are informative as they reveal the uncertain nature of classification for those clusters and characterize cell clusters that an expert must look into. Further, we could repurpose ICP to predict unknown cell types. However, further evaluations using various cell types and tissues are needed to make conclusions. Nevertheless, ICP may help minimize errors on unknown cells, and these clusters may be identified by investigating uncertainty patterns.

We observed a relationship between the observed error of conformal classification and the quality of batch correction. We expect an ideal batch correction method to remove the technical bias and cluster similar cell types together. Hence, the difference between expected and observed error rates under the confidence-credibility setup may quantify batch correction quality specifically for the task of label transfer.

Lastly, we would like to mention possible future directions motivated by the experiments described here. First, technical improvements may be necessary to improve and to further assess the applicability of conformal prediction in label transfer. In this work, we did not consider any normalization method for restricting prediction sets, which is a natural next step. Moreover, we would be interested in utilizing conformal anomaly detection to detect unknown and rare cell types.

## Acknowledgments

R.K. was supported by the EU eRare project Maxomod and S.B. was supported by SFB 1192 C3.

## References

Jose Alquicira-Hernandez, Anuja Sathe, Hanlee P Ji, Quan Nguyen, and Joseph E Powell. scpred: accurate supervised method for cell-type classification from single-cell rna-seq data. *Genome biology*, 20(1):1–17, 2019.

Jonathan Alvarsson, Staffan Arvidsson McShane, Ulf Norinder, and Ola Spjuth. Predicting with confidence: using conformal prediction in drug discovery. *Journal of Pharmaceutical Sciences*, 110(1):42–49, 2021.

Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.

Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. The application of conformal prediction to the drug discovery process. *Annals of Mathematics and Artificial Intelligence*, 74(1):117–132, 2015.

Dominic Grün, Mauro J Muraro, Jean-Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike van den Born, Johan Van Es, Erik Jansen, Hans Clevers, et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell stem cell*, 19(2):266–277, 2016.

Fabian Hausmann, Can Ergen, Robin Khatri, Mohamed Marouf, Sonja Hänzelmann, Nicola Gagliani, Samuel Huber, Pierre Machart, and Stefan Bonn. Discern-deep single cell expression reconstruction for improved cell clustering and cell subtype and state detection. *bioRxiv*, 2022.

Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature biotechnology*, 37(6):685–691, 2019.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21 (1):1–35, 2020.

Nathan Lawlor, Joshy George, Mohan Bolisetty, Romy Kursawe, Lili Sun, V Sivakamasundari, Ina Kycia, Paul Robson, and Michael L Stitzel. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome research*, 27(2):208–222, 2017.

Henrik Linusson, Ulf Johansson, Henrik Boström, and Tuve Löfström. Classification with reject option using conformal prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 94–105. Springer, 2018.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1):121–130, 2022.

Mauro J Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon Van Gurp, Marten A Engelse, Francoise Carlotti, Eelco Jp De Koning, et al. A single-cell transcriptome atlas of the human pancreas. *Cell systems*, 3(4):385–394, 2016.

Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka, 2008.

Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.

Åsa Segerstolpe, Athanasia Palasantza, Pernilla Eliasson, Eva-Marie Andersson, Anne-Christine Andréasson, Xiaoyan Sun, Simone Picelli, Alan Sabirsh, Maryam Clausen, Magnus K Bjursell, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell metabolism*, 24(4):593–607, 2016.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21(1):1–32, 2020.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, et al. Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research*, 47(D1):D721–D728, 2019.