scNym: Semi-supervised adversarial neural networks for single cell classification

Anonymous Authors[1]

Abstract

Single cell genomics experiments can reveal the keystone cellular actors in complex tissues. However, annotating cell type and state identities for each molecular profile in these experiments remains an analytical bottleneck. Here, we present scNym, a semi-supervised adversarial neural network that learns to transfer cell identity annotations from one experiment to another. scNym uses the semi-supervised MixMatch framework and domain adversarial training to take advantage of information in both the labeled and unlabeled datasets. scNym offers superior performance to baseline approaches in transferring cell identity annotations across experiments performed with different technologies or in distinct biological conditions. We demonstrate with ablation experiments that semi-supervision and adversarial training techniques improved both the performance and calibration of scNym models. We also show that scNym models are well-calibrated and interpretable with saliency methods, allowing for review of model decisions by domain experts.

1. Introduction

Single cell genomics allows for simultaneous molecular profiling of thousands of diverse cells (Trapnell, 2015). To derive biological insight from these data, each single cell molecular profile must be annotated with a cell identity, such as a cell type or state label. This task is traditionally performed manually, which can be time-consuming and error prone. Existing automated tools (Abdelaal et al., 2019) learn relationships between cell identity and molecular features from a training set with existing labels $(x, y) \sim \mathcal{D}$ without considering the unlabeled target dataset $u \sim \mathcal{U}$ in the learning process. Results from the semi-supervised learning literature suggest that incorporating information from $\mathcal{U}$ during training can improve the performance of prediction models (Oliver et al., 2018).

To make use of information in the unlabeled target dataset for cell type classification, we have developed a semi-supervised, adversarial neural network model scNym. In the typical supervised learning framework, the model touches the target unlabeled dataset to predict labels only after training has concluded. In contrast, our semi-supervised learning framework trains the model parameters on both the labeled and unlabeled data in order to leverage the structure in the target dataset, whose measurements may have been influenced by myriad sources of biological and technical bias and batch effects.

scNym uses the unlabeled target data through a combination of MixMatch semi-supervision (Berthelot et al., 2019) and by training a domain adversary (Ganin et al., 2016). The MixMatch semi-supervision combines MixUp data augmentations (Zhang et al., 2017), pseudolabeling of the target data (Lee, 2013; Verma et al., 2019), and an interpolation consistency penalty to improve generalization across the training and target domains.

By training a domain classification model as an adversary, scNym models learn a domain adapted embedding of the training and target datasets in addition to a performant identity classifier (Fig. 1A). Our model requires no prior manual specification of marker genes and yields a well-calibrated, continuous metric of classification confidence. We also provide model interpretation methods (Springenberg et al., 2014) to analyze which genes drive cell type classification decisions.

2. Approach

We train scNym models $f_\theta : x \to p(y|x)$ on normalized mRNA abundance profiles and labels (x, y) from the training dataset and unlabeled profiles u from the target dataset. We implement f_θ as a four-layer neural network with batch normalization, ReLU activations, dropout regularization, and a softmax activation on the final layer. Each hidden layer has 256 units. We sample minibatches $\mathbf{X}$ from the training set and $\mathbf{U}$ from the target set. At each training

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at the ICML 2020 Workshop on Computational Biology (WCB). Do not distribute.

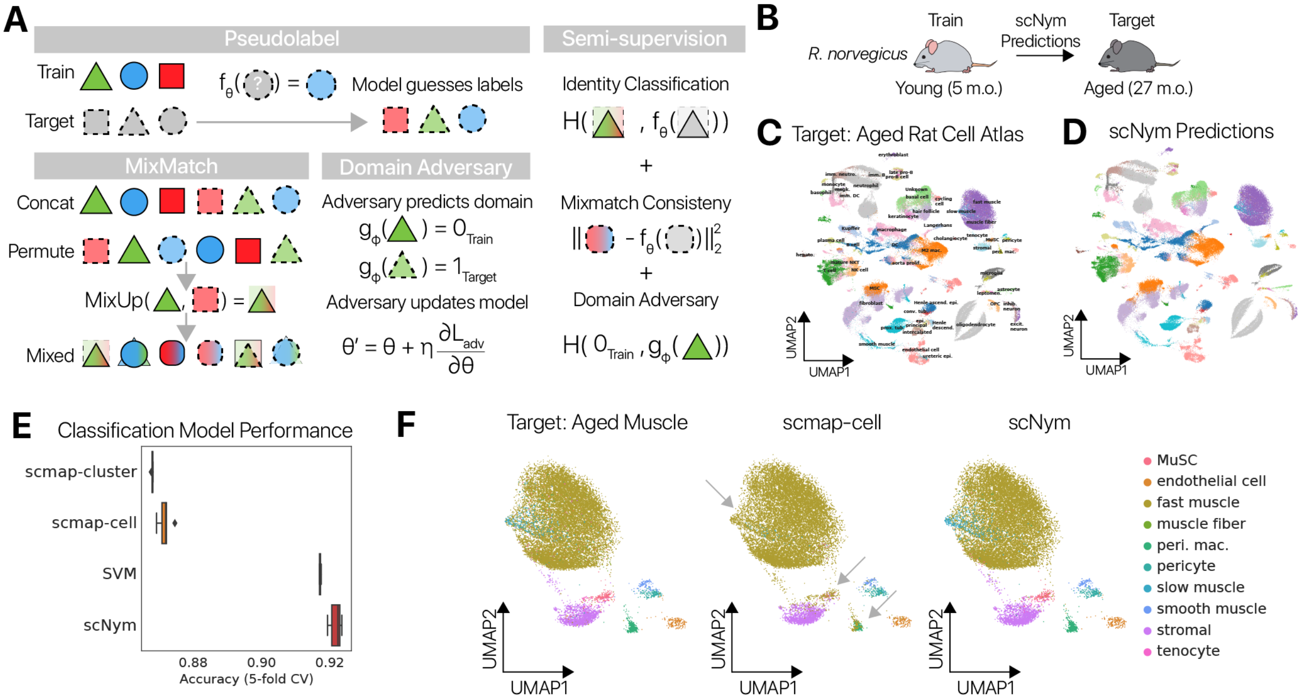

Figure 1. **scNym transfers cell identity annotations between sequencing technologies in the mouse lung. (A)** Graphical depiction of the scNym training procedure. Pseudolabels are generated for each observation in the unlabeled target data using model predictions and augmented with MixUp. An adversary is also trained to discriminate training and target observations. We train model parameters using a combination of supervised classification, interpolation consistency, and adversarial objectives. **(B)** scNym models were trained on young cells in the Rat Aging Cell Atlas and used to predict labels for aged cells. **(C)** Ground truth cell type annotations for the aged cells of the Rat Aging Cell Atlas shown in a UMAP projection. **(D)** scNym predicted cell types in the target dataset. scNym predictions match ground truth annotation in the majority (>90%) of cases. **(E)** Accuracy scores for scNym models and other state-of-the-art classification models. We find that scNym yields significantly higher accuracy scores than baseline methods ($p < 0.01$, Wilcoxon Rank Sums). Note: multiple existing methods could not complete this task. **(F)** Skeletal muscle cells labeled with ground truth annotations (left), scmap-cell predictions (center), and scNym predictions (right) are displayed in a UMAP projection. scNym accurately predicts multiple cell types that are confused by scmap-cell (arrows).

iteration, our model f_θ is used to generate pseudolabels $z_i = f_\theta(u_i)$ for unlabeled examples in a minibatch. We minimize pseudolabel entropy with a temperature sharpening operation $z_i = z_i^2 / \sum z_i^2$.

After pseudolabel generation, we further augment samples using MixUp (Zhang et al., 2017) weighted averages across the pseudolabeled minibatch and a labeled minibatch. We keep track of the dominant sample in each mixed pair and preserve labeled versus pseudolabeled identities on the mixed outputs. We sample the MixUp parameter from a symmetric Beta distribution $\lambda \sim \text{Beta}(\alpha, \alpha)$ where $\alpha = 0.3$.

We then apply a supervised cross-entropy loss $L_{\text{sup}} = \mathbb{E}[H(y_m, f_\theta(x_m)]$ to mixed, labeled examples (x_m, y_m) and a semi-supervised mean squared error penalty $L_{\text{SSL}} = \mathbb{E}[\|z_m - f_\theta(u_m)\|_2^2]$ on the difference between mixed pseudolabels z_m and model predictions on the mixed, pseudolabeled observations u_m, where $\|\cdot\|_2$ is the ℓ^2-norm. These losses are balanced by a weight function $\lambda_{\text{SSL}}(t) \rightarrow [0, 1]$ that we scale over 100 epochs of training with a sigmoid schedule.

We additionally train a domain adversary model (Ganin et al., 2016) g_ϕ at each iteration. We implement g_ϕ as a two-layer neural network with a softmax activation on the final layer. We train the adversary to predict the domain of origin $d_i \in \{0, 1\}$ for each point given the penultimate layer embedding of the classification model $v_i = f_\theta(x_i)^{(l-1)}$, such that $g_\phi : v_i \rightarrow \hat{d}_i$. We optimize the adversary g_ϕ with standard gradient descent and a cross-entropy loss $L_{\text{adv}} = \mathbb{E}[H(d_i, g_\phi(v_i))]$, but we use the "gradient reversal trick," to update the classifier parameters θ using the *inverse* of the adversary's gradients: $\theta_t \leftarrow \theta_{t-1} + \eta w_t \frac{\partial L_{\text{adv}}}{\partial \theta}$, where $w_t \rightarrow [0, 0.1]$ is a gradient weight we scale with a sigmoid schedule over 20 epochs. Our final loss is then:

$$L(\theta, \mathbf{X}, \mathbf{U}, t) = L_{\text{sup}} + \lambda_{\text{SSL}}(t) L_{\text{SSL}} + L_{\text{adv}}$$

Algorithm 1 scNym training for one epoch.

```
**Input:** train set $\mathcal{D} = \{x_i, y_i\}_i^N$, target set $\mathcal{U} = \{u_i\}_i^{N'}$
**Models:** Classifier $f_\theta$, Adversary $g_\phi$
**for** minibatches $\mathbf{X} \in \mathcal{D}$ **do**
    Draw unlabeled minibatch $\mathbf{U} \sim \mathcal{U}$
    Pseudolabel target examples $z_i = f_\theta(u_i)$
    Concat. batches $\mathbf{W} = \mathbf{X} :: \mathbf{U}; (w, q) \sim \mathbf{W}$
    MixUp $w_m = \mathrm{Mix}_\lambda(w_i, w_k); q_m = \mathrm{Mix}_\lambda(q_i, q_k)$
    Split batches $\mathbf{X}' = \mathbf{W}_{1:N}; \mathbf{U}' = \mathbf{W}_{N:N'}$
    Adversarial prediction $\hat{d}_i = g_\phi(f_\theta(x_i)^{(l-1)})$
    Compute $L_{\text{sup}} = \mathbb{E}_{(x_m, y_m) \sim \mathbf{X}'}[H(y_m, f_\theta(x_m))]$
    Compute $L_{\text{SSL}} = \mathbb{E}_{(u_m, z_m) \sim \mathbf{U}'}[\|z_m - f_\theta(u_m)\|_2^2]$
    Compute $L_{\text{adv}} = \mathbb{E}_{u \sim \mathbf{U}; x \sim \mathbf{X}}[H(d_i, \hat{d}_i)]$
    Backpropagate $\nabla L = L_{\text{sup}} + \lambda_{\text{SSL}}(t) L_{\text{SSL}} + L_{\text{adv}}$
    $\theta_t \leftarrow \theta_{t-1} - \eta[\frac{\partial L_{\text{sup}}}{\partial \theta} + \frac{\partial L_{\text{SSL}}}{\partial \theta} - w_t \frac{\partial L_{\text{adv}}}{\partial \theta}]$
    $\phi_t \leftarrow \phi_{t-1} - \eta[\frac{\partial L_{\text{adv}}}{\partial \theta}]$
**end for**
```

We minimize the loss over 400 training epochs with the Adadelta optimizer using a weight decay 10^{-5}, initial learning rate $\eta = 1.0$, and early stopping based on a validation set. We outline a single epoch of our training procedure in (Algorithm 1).

3. Results

3.1. Cell type classification benchmark tasks

We evaluated the performance of scNym to transfer cell identity annotations in five distinct tasks. These tasks were chosen to capture diverse kinds of technological and biological variation that complicate annotation transfer. All of our tasks represent a true cell type transfer across different experiments, in contrast to some efforts that report within-experiment hold-out accuracy.

We first evaluated cell type annotation transfer between animals of different ages. We trained scNym models on cells from young rats (5 months old) from the Rat Aging Cell Atlas (Ma et al., 2020) and predicted on cells from aged rats (27 months old, Fig. 1B). We found that predictions from our scNym model trained on young cells largely matched the ground truth annotations (92.2% accurate) on aged cells (Fig. 1C, D).

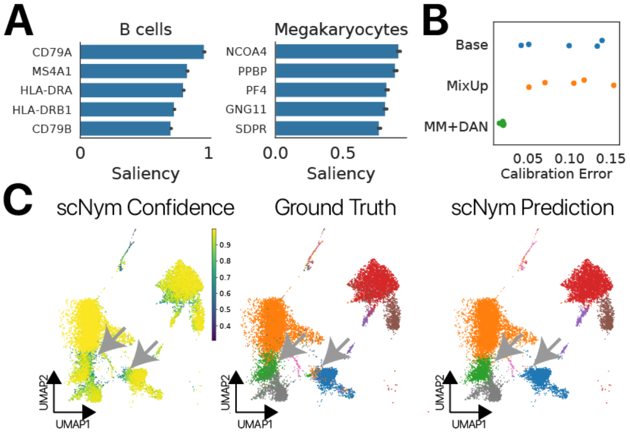

Figure 2. **scNym models are interpretable and well-calibrated. (A)** We used rectified backpropagation to derive saliency maps for each stimulated PBMCs classified with scNym models trained on unstimulated PBMCs. We recover known marker genes of many cell types (e.g. *CD79A* for B cells, *PPBP* for magekaryocytes). **(B)** MixMatch and adversarial training reduce the expected calibration error of scNym models for stimulated PBMC predictions. **(C)** Low confidence scores highlight incorrect scNym predictions (arrows) in stimulated PBMCs.

We compared scNym performance on this task to six state-of-the-art single cell RNA-seq cell identity annotation methods (Kiselev et al., 2018; Alquicira-Hernandez et al., 2019; Tan & Cahan, 2019; Abdelaal et al., 2019; de Kanter et al., 2019). scNym produced significantly improved labels over these methods, some of which could not complete this large task on our hardware (256GB RAM; Wilcoxon Rank Sums, $p < 0.01$, Fig. 1E, Table 1). We found that some of the largest differences in accuracy between scNym and the commonly used scmap-cell method were in the skeletal muscle. scNym models accurately classified multiple cell types in the muscle that are confused by scmap-cell, demonstrating that the increased accuracy of scNym is meaningful for downstream analyses (Fig. 1F).

We next tested the ability of scNym to classify cell identities after perturbation. We trained on human PBMCs in the absence of stimulation and predicted on PBMCs after stimulation with IFNβ (Kang et al., 2017). scNym achieved high accuracy ($> 91\%$), superior to baseline methods (Table 1).

To evaluate the ability of scNym models to transfer labels across different experimental technologies, we trained scNym models on single cell profiles from ten mouse tissues in the "Tabula Muris" captured using the 10x Chromium technology and predicted labels for cells from the same experiment captured using Smart-Seq2 (SS2) (Tabula Muris Consortium, 2018). We found that scNym predictions were highly accurate ($> 90\%$) and superior to baseline methods (Table 1).

In a second cross-technology task, we trained scNym on mouse lung data from the Tabula Muris and predicted on lung data from the "Mouse Cell Atlas," a separate mouse cell atlas effort that used a distinct single cell RNA-seq technology (Han et al., 2018). We found that scNym yielded superior classification accuracy ($> 90\%$) on mouse lung cell types in the Mouse Cell Atlas despite experimental batch effects and differences in the sequencing technologies (Table 1). We also trained scNym models to transfer

	scmap-cell	scmap-cluster	SVM	singleCellNet	scPred	CHETAH	scNym
Young to Old Rat	87.2	86.8	91.7	N/A	N/A	N/A	**92.2**
hPBMC Cross-Stim	38.3	78.4	81.9	90.3	60.5	49.9	**91.7**
TM 10x to MCA	89.7	83.3	88.4	80.9	62.4	85.5	**91.4**
TM 10x to SS2	92.3	88.3	93.1	85.9	70.1	86.9	**93.6**
Spatial Txn	81.8	76.7	92.1	87.7	**92.3**	56.6	91.6

Table 1. **Comparison of model performance across tasks.** Mean accuracy across 5-fold training split is reported. Bold text marks best models per task ($p < 0.05$, Wilcoxon Rank Sums). N/A indicates that a model could not complete the task on our hardware (256 GB of RAM).

regional identity annotations in spatial transcriptomics data from mouse brain sections (10x Genomics, 2020) and found performance competitive with baseline methods (Table 1). Together, these results demonstrate that scNym models can transfer cell type annotations across technologies and experimental environments with high performance.

3.2. Ablation Experiments

We ablated different components of our scNym model to determine which features were responsible for high performance. We found that semi-supervision with MixMatch and training with a domain adversary improved model performance (Wilcoxon Rank Sums, $p < 0.05$). This result was observed for multiple tasks. We hypothesized that scNym models might benefit from domain adaptation through the adversarial model. We found that training and target domains were significantly more mixed in scNym embeddings, supporting this hypothesis (Wilcoxon Rank Sums on entropy of batch mixing, $p < 0.05$). These results suggest that semi-supervision and adversarial training improve the accuracy of cell type classifications.

3.3. scNym models are interpretable using saliency methods

To interpret the classification decisions of our scNym models, we developed saliency analysis tools to identify genes that influence model decisions (Springenberg et al., 2014). We found that salient genes included known markers of specific cell types (e.g. *CD79A* for B cells, *GNLY* for NK cells), in addition to genes that may not have been chosen heuristically (Fig. 2A). This result provides confidence that our models are learning biologically meaningful representations that are likely to transfer across experiments.

3.4. scNym models are well-calibrated

We also investigated the calibration of our scNym models by comparing the prediction confidence scores to prediction accuracy (Thulasidasan et al., 2019). We found that MixMatch improved model calibration, such that high confidence predictions are more likely to be correct (Fig. 2B). scNym confidence scores can therefore be used to highlight cells that may benefit from manual review, further improving the annotation exercise when it contains a domain expert in the loop (Fig. 2C).

scNym confidence scores can also highlight unseen cell types in the target dataset using a modified training procedure that incorporates pseudolabel thresholding, inspired by FixMatch (Sohn et al., 2020). We simulated an experiment where we “discover” multiple cell types by predicting annotations on the Tabula Muris brain cell data using models trained on non-brain tissues. New cell types not present in the training data were given low confidence scores, highlighting these cells as potential cell type discoveries ($> 95\%$ recall).

4. Conclusion

Our benchmark results demonstrate that scNym models provide high performance across a range of cell identity classification tasks, including cross-age, cross-perturbation, and cross-technology. scNym performs better than six state-of-the-art baseline methods across varied tasks and is the only method here with high performance (>90% accuracy) on all tasks. Through ablation experiments, we show that MixMatch semi-supervision and domain adversarial training improve model performance. These methods also improve model calibration, such that scNym confidence scores can be used to identify cells for manual refinement. Our saliency analysis experiments show that scNym models learn intuitive, biologically relevant features for cell type classification and highlight features that drive decisions for low confidence cells during manual curation.

Our results collectively demonstrate that semi-supervised and adversarial learning methods provide promising performance benefits for cell identity classification models and motivate further adaptation of these techniques for single cell genomics applications. We aim to enable widespread use of these methods via the release of their open source implementations, tutorials, and pre-trained models for mouse, rat, and human cell types available from `anonymous.url`.

References

10x Genomics. 10x Genomics Single Cell Datasets, 2020. URL https://support.10xgenomics.com/single-cell-gene-expression/datasets.

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., and Mahfouz, A. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, 20(1): 194, September 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1795-z. URL https://doi.org/10.1186/s13059-019-1795-z.

Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q., and Powell, J. E. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology*, 20(1):264, December 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1862-5. URL https://doi.org/10.1186/s13059-019-1862-5.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. MixMatch: A Holistic Approach to Semi-Supervised Learning. *NeurIPS*, pp. 5049–5059, 2019.

de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T., and Holstege, F. C. P. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Research*, 47(16):e95, September 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz543. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6895264/.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-Adversarial Training of Neural Networks. *arXiv:1505.07818 [cs, stat]*, May 2016. URL http://arxiv.org/abs/1505.07818. arXiv: 1505.07818.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G.-C., Chen, M., and Guo, G. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, 173(5):1307, May 2018.

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C. M., Gate, R. E., Mostafavi, S., Marson, A., Zaitlen, N., Criswell, L. A., and Ye, C. J. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, December 2017.

Kiselev, V. Y., Yiu, A., and Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nature methods*, 15(5):359–362, April 2018.

Lee, D. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *ICML Workshop on Challenges in Representation Learning*, 2013.

Ma, S., Sun, S., Geng, L., Song, M., Wang, W., Ye, Y., Ji, Q., Zou, Z., Wang, S., He, X., Li, W., Esteban, C. R., Long, X., Guo, G., Chan, P., Zhou, Q., Belmonte, J. C. I., Zhang, W., Qu, J., and Liu, G.-H. Caloric Restriction Reprograms the Single-Cell Transcriptional Landscape of Rattus Norvegicus Aging. *Cell*, pp. 1–41, February 2020.

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. *NeurIPS*, pp. 3235–3246, 2018.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *arXiv:2001.07685 [cs, stat]*, January 2020. URL http://arxiv.org/abs/2001.07685. arXiv: 2001.07685.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for Simplicity: The All Convolutional Net. *arXiv*, December 2014.

Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727): 367–372, October 2018.

Tan, Y. and Cahan, P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Systems*, pp. 1–31, July 2019. doi: 10.1016/j.cels.2019.06.004. URL https://doi.org/10.1016/j.cels.2019.06.004. Publisher: The Authors.

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks. pp. 13888–13899, 2019.

Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498, October 2015.

Verma, V., Lamb, A., Kannala, J., Bengio, Y., and Lopez-Paz, D. Interpolation Consistency Training for Semi-Supervised Learning. *arXiv*, March 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In *ICLR*, October 2017.