Energy-based Modelling for Single-cell Data Annotation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Single-cell sequencing has provided profound insights into understanding heterogeneous cellular activities by measuring sequence information at the individual cell resolution. Accurately annotating a single-cell RNA sequencing (scRNA-seq) dataset is a crucial step for the single-cell analysis pipeline. In particular, previously unobserved cell types and cellular states frequently appear in scRNA-seq experiments and carry valuable information [1]. These highlights the need for reliable annotation tools with out-of-distribution (OOD) detection capability. In this work, we introduce energy-based models (EBMs), a family of probabilistic models, for scRNA-seq annotation and OOD detection, which results in more accurate, calibrated, and robust cell type predictions. Recent advancements in energy-based modelling have made it possible to design and deploy EBMs for joint discriminative and generative tasks [2]. Particularly, we present CLAMS, an EBM instance based on the joint energy-based model (JEM) [3], for single-cell data hybrid modelling. Our experiments reveal that hybrid modelling with EBMs maintains the strong discriminative power of baseline classifiers and outperforms the state-of-the-art by integrating generative capabilities in data annotation and OOD detection tasks. In addition, we provide a diagnosis of training JEM and propose effective regularization methods to boost JEM's performance. To the best of our knowledge, we are the first work that applies EBMs to single-cell data modelling.

1 Introduction

The advent of single-cell high-throughput sequencing has enabled researchers to examine cellular activities and functionalities at unprecedented resolution. Single-cell genomics makes it possible to study a group of cells within its micro-environment and resolve its heterogeneity. Many downstream analyses, such as trajectory inference, require a reliable annotation of cell types that appear in the dataset. Nevertheless, annotating single-cell data remains a laborious and time-consuming task, which usually requires both expert knowledge and well-established marker genes. A standard workflow consists of identifying a highly expressed marker gene, which is typically associated with a specific cell type, and labeling the corresponding cell or cell cluster [4].

Another challenge in single-cell type classification is that the current set of known cell types is not comprehensive. An effective classification approach should be able to warn the user if the observed cell is significantly different from the set of previously observed cell classes. For example, if a scientist is studying a cell population they should be warned if they are classifying a previously unseen rare cell type. A standard model with a softmax output would assign a probability distribution for labeling previously unseen cells even though it could be outside the learned distribution. Thus, an effective tool would have OOD detection capabilities, allowing it to alert the researcher when analyzing a previously unseen cell type.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

In our work, we use the recent advancements of EBMs and apply them to single-cell data modelling. EBMs can be easily integrated and reformulated to accomplish hybrid data modelling tasks, which simultaneously learn a discriminate classifier and a generative model. Specifically, we present a set of EBM data annotation tools with OOD capability and evaluate their performance through a diverse set of experiments on single-cell pancreas and peripheral blood mononuclear cell (PBMC) datasets. We find that EBMs produce well-calibrated cell type annotations and are able to detect both biological OOD samples. Our contributions in this work are four-fold:

- We introduce EBMs, a family of probabilistic models, to single-cell data modelling.
- We present CLAMS, a customized version of EBM, and accomplish robust and well-calibrated results for scRNA-seq annotation.
- We exploit generative data modelling with EBMs and demonstrate that our method outperforms state-of-the-art methods for OOD detection.

2 Methods

We implement EBMs including JEM, VERA, HDGE, and HDGE + JEM with author-suggested hyper-parameters. These models also serve as benchmark models in our work. In the rest of this section, we mainly discuss our model CLAMS as a carefully engineered EBM for scRNA-seq annotation and OOD detection. We demonstrate that our model is a suitable choice for such tasks.

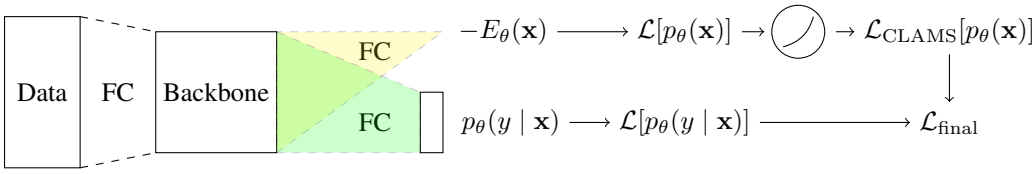


Figure 1: Block diagram of JEM training strategy used in this project: FC stands for fully-connected layer. We use a 1D residual network [5] as the backbone network architecture. JEM has a two-headed network output. When calculating the generative loss, the network parameters θ are frozen and SGLD follows the gradient of energy function $E_\theta(\mathbf{x})$ to iteratively update the generated samples in the replay buffer. When calculating the discriminative loss, the model behaves similarly to other neural network classifiers. We apply proposed regularization methods on the generative loss $\mathcal{L}[p_\theta(\mathbf{x})]$.

Here, we elaborate the design choices of our model CLAMS, discuss our diagnosis of JEM instability and present our resolutions. JEM jointly optimizes the cross-entropy discriminative loss $\mathcal{L}[p_\theta(y \mid \mathbf{x})]$ and a generative loss $\mathcal{L}[p_\theta(\mathbf{x})]$ as shown in Figure 1. With state-of-the-art training techniques, including dropout, weight decay, learning rate schedule, and early stopping, the training procedure of the classifier is fairly stable. Therefore, we hypothesized that the generative loss is the source of instability. In JEM, the generative loss is defined as:

$$\mathcal{L}[p_\theta(\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{fake}}}[E_\theta(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{real}}}[E_\theta(\mathbf{x})], \tag{1}$$

where $E_\theta(\cdot)$ is the energy value, or unnormalized density, assigned to a data point.

We identify that model divergence occurs when the energy values becoming negative resulting in gradient growth which in turn forms a positive feedback loop. Eventually, such a feedback loop causes gradient explosion resulting in a poorly performing model. We employ two regularization strategies to prevent the initiation of a positive feedback loop. First we clip the generative energy values using an activation function thus capping the magnitude of the negative values the energy function can take on. Second we introduce stochastic on-off regularization for the generative loss referencing the idea of a dropout unit. Our final model CLAMS is a joint energy based model and makes use of ELU clipping with stochastic regularization $p = 0.25$ for generative loss regularization. We expand on our JEM instability observations and interventions in Appendix C.

3 Experiments

We start by designing a scRNA-seq annotation experiment under a reference-query setup with different EBMs. In the experiment, we evaluate our models by annotation accuracy, calibration per-

formance, and OOD detection. We further develop a designated an one-class OOD detection experiment by leaving out classes in the dataset to inspect our model. Finally, we conduct an ablation study by model decomposition to assess the effectiveness of our regularization methods.

3.1 scRNA-seq Annotation

We include the primary results for this work in the main text, while ablation results can be found in the Appendix F. We train all models for a fixed 80 epochs and observe all training saturates before 80 epochs. We report results on the query dataset for the model with the best validation accuracy achieved during training. This can be viewed as early stopping and the divergence of model will not affect these numbers. We run repeated experiments over five random seeds. We summarize our results in Table 1. Our results demonstrate that EBM-based methods have consistently good performance on both datasets, and they tend to be robust, accurate, and well-calibrated. State-of-the-art biological tools does not have robust performance and only achieve competitive results on one dataset.

PBMC	n	Div.	In Acc	In ECE	$p(\mathbf{x})$	$p(y)$
CLAMS (Ours)	5	0	92.89 $\pm$ 0.72	5.39 $\pm$ 0.46	**0.92** $\pm$ 0.01	0.77 $\pm$ 0.02
JEM	5	100	90.32 $\pm$ 4.93	**2.61** $\pm$ 0.66	0.61 $\pm$ 0.28	0.86 $\pm$ 0.00
VERA	5	0	88.76 $\pm$ 2.04	6.56 $\pm$ 2.06	0.87 $\pm$ 0.05	0.80 $\pm$ 0.03
HDGE	5	0	**93.58** $\pm$ 0.30	3.45 $\pm$ 0.11	0.81 $\pm$ 0.01	0.78 $\pm$ 0.01
HDGE + JEM	5	100	92.29 $\pm$ 0.87	6.25 $\pm$ 2.34	0.42 $\pm$ 0.23	0.70 $\pm$ 0.09
ResNet	5	0	91.53 $\pm$ 1.41	4.73 $\pm$ 0.25	–	0.79 $\pm$ 0.02
$\text{SVM}_{\text{reject}}$	1	–	91.55	22.82	–	0.72
Ingest	1	–	81.55	–	–	–
scPred	1	–	93.01	5.42	–	**0.92**
Seurat v4	1	–	93.09	11.44	–	0.81
Pancreas	n	Div.	In Acc	In ECE	$p(\mathbf{x})$	$p(y)$
CLAMS (Ours)	5	0	88.38 $\pm$ 0.62	5.81 $\pm$ 0.75	0.89 $\pm$ 0.03	0.71 $\pm$ 0.07
JEM	5	100	86.61 $\pm$ 2.07	7.57 $\pm$ 0.86	**0.92** $\pm$ 0.06	0.54 $\pm$ 0.13
VERA	5	0	89.14 $\pm$ 0.86	9.59 $\pm$ 1.12	0.88 $\pm$ 0.13	0.57 $\pm$ 0.19
HDGE	5	0	89.05 $\pm$ 0.21	6.34 $\pm$ 0.45	0.79 $\pm$ 0.02	0.78 $\pm$ 0.02
HDGE + JEM	5	100	87.02 $\pm$ 1.63	6.42 $\pm$ 0.73	0.69 $\pm$ 0.13	0.57 $\pm$ 0.17
ResNet	5	0	88.13 $\pm$ 0.53	8.69 $\pm$ 0.42	–	0.67 $\pm$ 0.04
$\text{SVM}_{\text{reject}}$	1	–	**91.53**	13.28	–	0.59
Ingest	1	–	89.37	–	–	–
scPred	1	–	90.45	8.82	–	0.09
Seurat v4	1	–	89.81	**3.31**	–	0.87

Table 1: Annotation results: Div. is the chance of divergence (%); In Acc and In ECE are annotation accuracy and ECE score for in-distribution data; $p(\mathbf{x})$ and $p(y)$ are AUROCs of unnormalized density $p(\mathbf{x})$ and predictive class probability $p(y)$ score for OOD detection.

Furthermore, we observe a clear improvement of CLAMS over JEM in terms of almost all evaluation metrics. This evidences the effectiveness of our regularization methods and confirms our hypotheses regarding JEM training instability. Our model has slightly better performance against HDGE, which reformulates the hybrid data modelling as a contrastive learning problem [6]. Compared to HDGE, our clipping method and stochastic regularization methods can be simply implemented in one line on top of JEM. We argue that our model is as effective as HDGE but is much simpler. Lastly, we find that the un-normalized density $p(\mathbf{x})$ outperforms $p(y)$, which is the classical score for OOD detection in single-cell annotation problems. $p(\mathbf{x})$ also produces more reliable and more stable OOD detection, which are reflected as higher mean AUROC and lower standard deviation.

3.2 One-class OOD Detection

We further investigate the OOD detection capability of CLAMS by designing this follow-up OOD detection experiment: for a labeled dataset $\mathcal{D}^n$ with n classes, we first generate sub-datasets by leaving out each class. This yields n sub-dataset pairs $\mathcal{D}_i^{n-1}$ and $\tilde{D}_i^{n-1}, i \in [1, n]$. The leave-out class $\tilde{D}_i^{n-1}$ contains samples from purely unseen class in the training procedure and will serve as OOD samples. The models are trained on $\mathcal{D}_i^{n-1}$ and evaluated on the leave-out class $\tilde{D}_i^{n-1}$. We run five repetitions by setting matched seeds.

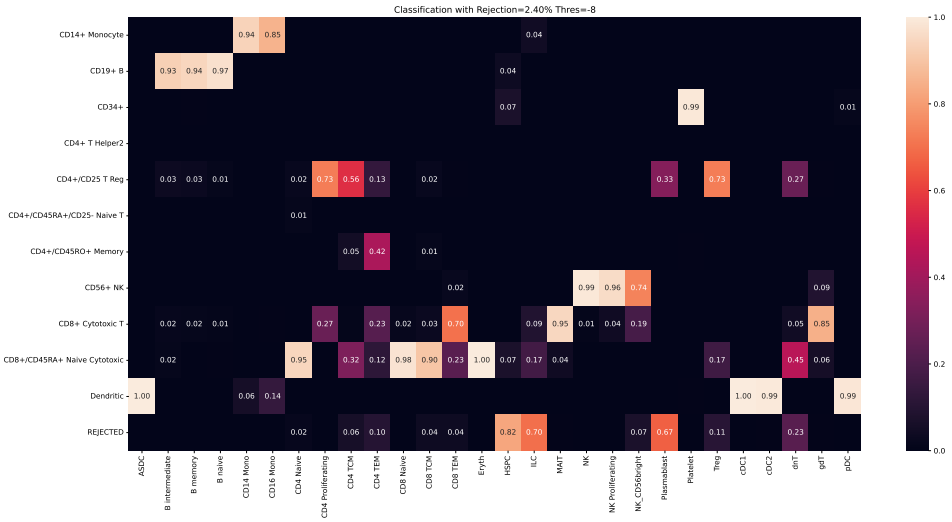


Figure 2: An annotation heatmap of PBMC dataset: rows are labels of the reference dataset and columns are labels of the query dataset. The heatmap is normalized column-wise.

PBMC	n	Div.	$p(\mathbf{x})$	$p(y)$	*Pancreas*	n	Div.	$p(\mathbf{x})$	$p(y)$
CLAMS (Ours)	55	0	**0.65** $\pm$ 0.29	0.53 $\pm$ 0.27	**CLAMS (Ours)**	95	0	0.83 $\pm$ 0.19	0.86 $\pm$ 0.09
JEM	55	100	0.63 $\pm$ 0.21	0.44 $\pm$ 0.26	JEM	95	100	0.73 $\pm$ 0.22	0.74 $\pm$ 0.16
VERA	55	0	0.51 $\pm$ 0.29	0.49 $\pm$ 0.26	VERA	95	0	0.78 $\pm$ 0.21	0.80 $\pm$ 0.13
HDGE	55	0	0.53 $\pm$ 0.26	0.54 $\pm$ 0.25	HDGE	95	0	0.85 $\pm$ 0.11	0.86 $\pm$ 0.10
HDGE + JEM	55	100	**0.65** $\pm$ 0.20	0.56 $\pm$ 0.25	HDGE + JEM	95	100	0.76 $\pm$ 0.19	0.78 $\pm$ 0.16
ResNet	55	0	–	0.49 $\pm$ 0.26	ResNet	95	0	–	0.82 $\pm$ 0.09
$\text{SVM}_{\text{reject}}$	11	–	–	0.64 $\pm$ 0.24	$\text{SVM}_{\text{reject}}$	19	–	–	**0.87** $\pm$ 0.06

Table 2: One-class OOD detection results: we do not include results for biological tools since by default they align the reference and query dataset, in which there is no overlapped cell type between them. Forcing such alignment will consequently cause overfitting and poor performance.

In Table 2, our results show that $p(\mathbf{x})$ possesses better OOD detection power comparing to classical $p(y)$ score on PBMC dataset and is as good as $p(y)$ on pancreas dataset. Our model CLAMS shows a clear distinction over JEM and has overall the best results. Our annotation OOD experiment in Section 3.1 focuses on an end-user scenario where the dataset to be annotated has a mix of in-distribution data and out-of-distribution data. Whereas, this one-class OOD experiment is a follow-up experiment under an artificial setting to analyze the model's capability to detect OOD data. Our results in Table 2 show that $p(\mathbf{x})$ has $\text{AUROC} > 0.5$ with $p - \text{value} < 0.0001$, which implies that our model can differentiate the leave-out class from the rest of dataset. However, a caveat of reading this result is that the similarity between underlying cell types may dilute AUROC values.

4 Conclusion

In this work, we utilize EBM and data hybrid modelling techniques to address scRNA-seq annotation and OOD detection problems, which are two essential steps in the single-cell data analysis pipeline. We exploit the generative power of EBMs for OOD detection while maintaining strong annotation accuracy and calibration performance. We also present CLAMS, a specific instance of EBM, which yields accurate and robust single-cell data analysis. Moreover, we pinpoint the source of JEM training instability being the implicitly problematic loss formulation and demonstrate that our regularization methods are simple yet effective in improving training and deployment performance. We hope that our work demonstrates the advantages of EBMs for single-cell data annotation under the presence of reference datasets and that our model, CLAMS, may become a competitive tool for this task. Furthermore, CLAMS also has the potential to be deployed for single-nuclei RNA sequencing (snRNA-seq) or single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq) data annotation since there exists an underlying similarity between the data formats.

References

[1] H. Pan, C. Xue, B. J. Auerbach, J. Fan, A. C. Bashore, J. Cui, D. Y. Yang, S. B. Trignano, W. Liu, J. Shi, *et al.*, "Single-cell genomics reveals a novel cell state during smooth muscle cell phenotypic switching and potential therapeutic targets for atherosclerosis in mouse and human," *Circulation*, vol. 142, no. 21, pp. 2060–2075, 2020.

[2] Y. Du and I. Mordatch, "Implicit generation and modeling with energy based models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[3] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," pp. 1–22, 12 2019.

[4] Z. A. Clarke, T. S. Andrews, J. Atif, D. Pouyabahar, B. T. Innes, S. A. MacParland, and G. D. Bader, "Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods," *Nature Protocols*, vol. 16, pp. 2749–2764, June 2021. Number: 6 Publisher: Nature Publishing Group.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[6] H. Liu and P. Abbeel, "Hybrid discriminative-generative training via contrastive learning," *arXiv preprint arXiv:2007.09070*, 2020.

[7] P. Panareda Busto and J. Gall, "Open set domain adaptation," in *Proceedings of the IEEE international conference on computer vision*, pp. 754–763, 2017.

[8] S. Garg, S. Balakrishnan, and Z. C. Lipton, "Domain adaptation under open set label shift," *arXiv preprint arXiv:2207.13048*, 2022.

[9] F. A. Vieira Braga, G. Kar, M. Berg, O. A. Carpaij, K. Polanski, L. M. Simon, S. Brouwer, T. Gomes, L. Hesse, J. Jiang, *et al.*, "A cellular census of human lungs identifies novel cell states in health and in asthma," *Nature medicine*, vol. 25, no. 7, pp. 1153–1163, 2019.

[10] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[11] H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen, "A benchmark of batch-effect correction methods for single-cell rna sequencing data," *Genome biology*, vol. 21, no. 1, pp. 1–32, 2020.

[12] L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni, "Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors," *Nature biotechnology*, vol. 36, no. 5, pp. 421–427, 2018.

[13] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, "Comprehensive integration of single-cell data," *Cell*, vol. 177, no. 7, pp. 1888–1902, 2019.

[14] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri, "Fast, sensitive and accurate integration of single-cell data with harmony," *Nature methods*, vol. 16, no. 12, pp. 1289–1296, 2019.

[15] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature methods*, vol. 15, no. 12, pp. 1053–1058, 2018.

[16] M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Müller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, *et al.*, "Benchmarking atlas-level data integration in single-cell genomics," *Nature methods*, vol. 19, no. 1, pp. 41–50, 2022.

[17] J. Alquicira-Hernandez, A. Sathe, H. P. Ji, Q. Nguyen, and J. E. Powell, "scpred: accurate supervised method for cell-type classification from single-cell rna-seq data," *Genome biology*, vol. 20, no. 1, pp. 1–17, 2019.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, and N. Yosef, "Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models," *Molecular systems biology*, vol. 17, no. 1, p. e9620, 2021.

[22] X. Shao, H. Yang, X. Zhuang, J. Liao, P. Yang, J. Cheng, X. Lu, H. Chen, and X. Fan, "scdeepsort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network," *Nucleic acids research*, vol. 49, no. 21, pp. e122–e122, 2021.

[23] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, Citeseer, 2011.

[24] W. Grathwohl, D. Duvenaud, and J. Kelly, "No mcmc for me: A mortized sampling for fast and stable training of energy - based models," pp. 1–35, 2021.

[25] X. Yang, H. Ye, Y. Ye, X. Li, and S. Ji, "Generative max-mahalanobis classifiers for image classification, generation and more," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 67–83, Springer, 2021.

[26] M. Baron, A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, A. M. Klein, D. A. Melton, and I. Yanai, "A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure," *Cell Systems*, vol. 3, pp. 346–360.e4, 2016.

[27] Åsa Segerstolpe, A. Palasantza, P. Eliasson, E. M. Andersson, A. C. Andréasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, D. M. Smith, M. Kasper, C. Ämmälä, and R. Sandberg, "Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes," *Cell Metabolism*, vol. 24, pp. 593–607, 2016.

[28] Y. J. Wang, J. Schug, K. J. Won, C. Liu, A. Naji, D. Avrahami, M. L. Golson, and K. H. Kaestner, "Single-cell transcriptomics of the human endocrine pancreas," *Diabetes*, vol. 65, pp. 3028–3038, 2016.

[29] M. J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M. A. Engelse, F. Carlotti, E. J. de Koning, and A. van Oudenaarden, "A single-cell transcriptome atlas of the human pancreas," *Cell Systems*, vol. 3, pp. 385–394.e3, 2016.

[30] ScanPy, "Integrating data using ingest and bbknn," 2022.

[31] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, 2017.

[32] 10x Genomics Inc, "Fresh 68k pbmcs (donor a)," 2016.

[33] 10x Genomics Inc, "10k pbmcs from a healthy donor (v3 chemistry)," 2018.

[34] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: large-scale single-cell gene expression data analysis," *Genome biology*, vol. 19, no. 1, pp. 1–5, 2018.

[35] G. Fumera and F. Roli, "Support vector machines with embedded reject option," in *International Workshop on Support Vector Machines*, pp. 68–82, Springer, 2002.

[36] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature biotechnology*, vol. 36, no. 5, pp. 411–420, 2018.

[37] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, pp. 1321–1330, PMLR, 2017.

[38] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, JMLR Workshop and Conference Proceedings, 2010.

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[40] X. Yang and S. Ji, "Jem++: Improved techniques for training jem," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6494–6503, 2021.

A Background and Related Works

Our work resides in the cross-discipline of single-cell genomics and machine learning. Therefore, we discuss prior works in both fields.

A.1 A Open Set Domain Adaption Problem

Single-cell data annotation can be framed as an open set domain adaption (OSDA) problem, where training data does not encode all information, and new classes may arrive in the deployment stage [7]. For example, a handwriting digit classifier was trained to recognize digits, but letters and punctuation marks may also be detected in the deployment. An OSDA problem has two distinct stages: a) align reference and query distributions, and b) learn a classifier [7, 8]. In the context of single-cell genomics, new cell types and cellular states are frequently detected in fresh single-cell datasets along their trajectories, which are often of interest to researchers given they sometimes carry information about previously unknown cellular functions [9, 1]. Consequently, the OOD detection capability of the classifier is essential to flag certain groups of cells for further investigation.

A.2 scRNA-seq Annotation

State-of-the-art tools often address the annotation problem as a two-stage problem. A new query dataset will be integrated with the reference dataset prior to the annotation. Various methods were proposed to align multiple single-cell transcriptomes. ComBat uses a Bayesian network to adjust and integrate data [10]. ComBat was originally proposed for microarray data, but has shown the capability to generalize to single-cell data [11]. Mutual nearest neighbors (MNN) extracts MNN pairs in both reference and query datasets and integrates them with linear correction vectors [12]. Seurat v4 also uses a MNN-based method to extract anchors for integration but filters MNNs with a shared neighbor graph [13]. Harmony iteratively learns correction vectors in a low-dimensional space with fuzzy clustering assignment [14]. scVI designs a multi-task variational auto-encoder to accomplish data integration and other fundamental tasks simultaneously [15]. It is reported that Harmony can be chosen as an initial attempt to integrate datasets due to its simplicity and efficiency, and scVI has particularly better performance on more complicated tasks [11, 16].

Once datasets are aligned and integrated, they can then be annotated with a classifier. cell type annotation is a well-studied problem in single-cell genomics. There are various methods based on discriminative models that have shown promise for this task. Ingest uses a kNN classifier to assign a known label based on kNN graph. scPred uses a calibrated support vector machine (SVM) classifier with a rejection option in a new feature space [17]. Seurat v4 creates a binary classification matrix and uses the similarity matrix of query cells and anchors [13]. With the advancement of DNN training, DNN-based methods have shown superior performance in many classical tasks such as image processing [18], data generation [19], natural language processing [20]. Such achievements have been gradually transferred to single-cell genomics data analysis. scANVI [21] develops a variational auto-encoder for single-cell probabilistic modelling and data annotation. scDeepSort is a graph neural network based single-cell annotation tool with a pre-trained model which does not require additional references [22]. DNN-based models usually tend to integrate multiple functionalities and features into a single model for end-to-end optimizations. In our work, we opt to design a tool specifically for scRNA-seq data annotation using an EBM.

Due to the complexity and heterogeneity of scRNA-seq data, annotating single-cell datasets remains a great challenge and an open problem. In our work, we introduce EBM-based methods, a family of probabilistic DNN models, for single-cell data annotation and OOD detection. We accomplish simultaneous discriminative and generative modelling for single-cell data, which yields an annotation tool and a data generator. In this work, we chose Harmony for data integration and a neural network with residual connection [5] as the classifier for data hybrid modelling.

A.3 Energy-Based Models and Hybrid Data modelling

EBMs are a very flexible class of models inspired by statistical mechanics and physics. For a data point $\mathbf{x}$, its energy value is determined through a function $E_\theta(\mathbf{x})$ parameterized by θ. With the use of the Boltzmann distribution, EBMs can be interpreted as parameterizing a probabilistic distribution

based on its energy value as:

$$p_\theta(\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{x}))}{Z(\theta)}, \tag{2}$$

where $Z(\theta)$ is the normalizing constant that is parameterized by θ, and $-E_\theta(\cdot)$ is defined as the negative energy function such that $E : \mathbb{R}^d \to \mathbb{R}$.

Since the energy function can be parameterized by any function, $E_\theta(\mathbf{x})$ can be a softmax classifier, and therefore EBMs can simultaneously produce a predictive model for the data and a generative model to capture the data distribution. One of the notable works in data hybrid modelling is JEM [3]. We propose CLAMS as an improved version of JEM. JEM defines a joint distribution of a data point $\mathbf{x}$ and the label y as:

$$p_\theta(\mathbf{x}, y) \triangleq \frac{\exp(f_\theta(\mathbf{x})[y])}{Z(\theta)}, \tag{3}$$

where $f_\theta(\mathbf{x})[y]$ is the yth logit of $\mathbf{x}$'s predictive class distribution. By marginalizing out y, $p_\theta(\mathbf{x})$ can be re-expressed in the standard EBM form:

$$p_\theta(\mathbf{x}) = \frac{\exp\left[\log\left(\sum_y \exp(f_\theta(\mathbf{x})[y])\right)\right]}{Z(\theta)}. \tag{4}$$

Therefore, JEM uncovers a hybrid model by optimizing the joint distribution $p_\theta(\mathbf{x}, y)$. To jointly train the classifier and generative components, we can formulate the maximum likelihood objective as the sum of discriminative and generative losses

$$\ell_{\text{ML}} = \log p_\theta(\mathbf{x}, y) = \log p_\theta(\mathbf{x}) + \log p_\theta(y \mid \mathbf{x}). \tag{5}$$

JEM adopts the standard EBM training strategy by retaining a replay buffer with stochastic gradient Langevin dynamics (SGLD) sampling, a gradient-guided Monte-Carlo Markov chain method [2, 3, 23].

Although effective for hybrid modelling, JEMs can be difficult to train. For instance, training instability is one major challenge that can make achieving high-quality generative and discriminative performance difficult [3, 24, 6, 25]. The model may diverge prior to saturation, which causes unsatisfactory performance in evaluation and deployment. We observed that this issue is especially common when working with single-cell datasets compared to images, the original JEM application area. The author of JEM used a heuristic approach to address this problem through check-pointing and re-initialization to a model prior to divergence [3]. Their succeeding work VERA resolved this issue by replacing SGLD sampling with variational inference, which consequently accelerates the training procedure and enables exact sample generation [24]. HDGE, another extension of JEM, re-interpreted the predictive distribution as contrasting learning and jointly optimizes it with the discriminative loss [6]. Both VERA and HDGE further improve JEM by replacing SGLD with other sampling methods. Nevertheless, our experiments demonstrate that VERA is poorly calibrated despite its accelerated training speed, and HDGE has to be integrated with JEM, as HDGE + JEM, to generate class-conditional samples. As a result, we focus on improving the original JEM formulation by identifying and resolving the root of the instability of the SGLD sampler.

B Experiments Appendix

B.1 Datasets

We run scRNA-seq annotation experiment with the following reference-query dataset pairs.

Tissue/Sample	Reference	Query	Source
pancreas	Baron [26], Segerstolpe [27], Wang [28]	Muraro [29]	[30]
PBMC	Zheng [31] (10x NextSeq)	10x NovaSeq	[32, 33]

Table 3: Datasets used in this work

We process datasets with standard ScanPy [34] workflow and select top highly variable genes for initial dimension reduction. To start, we run Harmony to integrate reference and query datasets prior

to the model training. The reference dataset is split into 8:1:1 for training, validation, and testing. Our models are trained on the training set, and the validation set is used for evaluation. Under our experimental setup, we have access to the labels of query datasets. We split cells in the query datasets into two groups: in-distribution data, where cell types appear in the reference dataset; and out-of-distribution data, where cell types do not appear in the reference dataset, i.e., new cell types.

B.2 Evaluation Criteria

We benchmark our models against three groups of models: EBMs (JEM [3], VERA [24], HDGE [6], HDGE + JEM [6]); typical classifers (ResNet [5], $\text{SVM}_{\text{reject}}$ [35]; and biological tools (Ingest, scPred [17], Seurat v4 [36]).

The core evaluation metrics are the annotation accuracy for in-distribution data and the score's area under the receiver operating characteristic (AUROC) for OOD data. Annotation accuracy reflects how well the model can annotate known cell types, and score AUROC measures how well a score differentiates OOD data from in-distribution data. We also analyze the calibration performance of a series of models. The calibration reflects how well the model's prediction confidence corresponds to its prediction accuracy. It is believed that, although modern DNN-based methods have accomplished exceptional performance, they tend to be poorly calibrated [3, 37].

ECE score is a useful metric to quantify the calibration performance of a classifier. It calculates the expected difference in prediction confidence and the corresponding prediction accuracy as follows:

$$\text{ECE} = \frac{1}{n}\sum_{i=1}^{m}|\mathcal{B}_i||\text{avg_acc}(\mathcal{B}_i) - \text{avg_conf}(\mathcal{B}_i)|, \qquad (6)$$

where n data points are split into m equal-sized bins $\mathcal{B}_i$ based on their prediction confidence. Figure 6 shows a histogram for ECE calculation, and the ECE score reflects the expected difference between the ideal values along the diagonal line and actual bin values. We expect a well-calibrated classifier to have a lower ECE score such that its prediction confidence is close to its prediction accuracy (along the red dashed diagonal line).

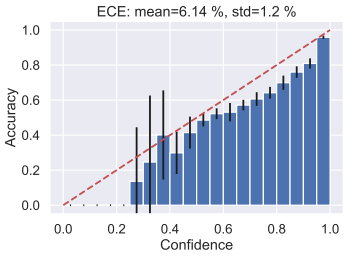


Figure 3: ECE histogram

B.3 Data Generation

In our work, we mainly focus on using the generative data modelling for the task of OOD detection rather than data generation. Nevertheless, we find that our model is able to produce meaningful single-cell data from the training set distribution. In Figure **??**, we include a t-SNE plot for the generated data from the PBMC dataset. Through visualizing samples from CLAMS replay buffer we are able to visually inspect the learned class distribution over the sample space. As opposed to class balanced datasets in the vision domain, single cell datasets can have severe differences in sample sizes between classes. We observe that classes with fewer number of cells are not learned as well as over-sampled classes.

Moreover, we propose a modified SGLD sampler with an imbalanced replay buffer to address class-imbalanced issue (see Appendix E). We do not use the modified version in this work and will further discuss this in our future works.

C JEM Training Instability

Theoretically, the generative loss $\mathcal{L}[p_\theta(\mathbf{x})]$ should oscillate around 0, which produces generated samples whose energy values are on a similar scale as the real samples. However, we noticed that this is not always true, and we presume that this could be the implicit problem of JEM formulation [25]. As shown in equation (1), the generative loss is defined as the difference between the expectation of fake energy values and real energy values. When $\mathcal{L}[p_\theta(\mathbf{x})]$ is negative, $\mathbb{E}_{\mathbf{x}\sim p_{\text{fake}}}[E_\theta(\mathbf{x})] < \mathbb{E}_{\mathbf{x}\sim p_{\text{real}}}[E_\theta(\mathbf{x})]$. The model parameters will be adjusted to either increase fake energy values or decrease the real energy values, *vice versa*. Empirically, we observe that the model

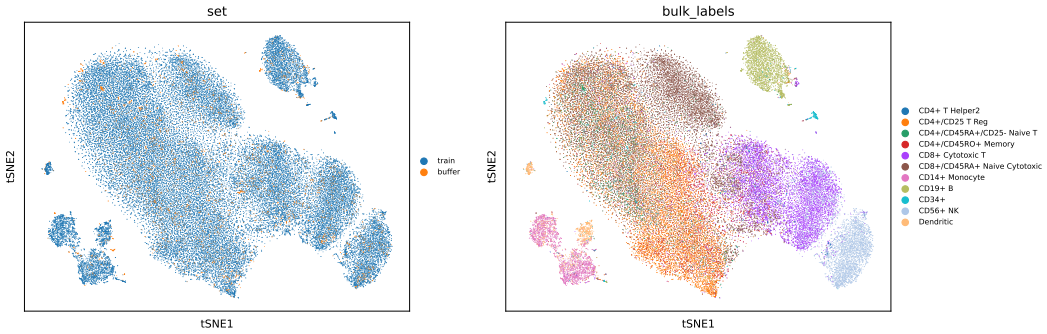


Figure 4: Generated data in t-SNE space: we generate 300 samples per class. On the left, data points are colored based on the source of data: training data are colored in blue, and generated data are colored in orange. On the right, data points are colored by cell types.

tends to behave as the former. As a result, such a training strategy forms a feedback loop, and the energy values will keep increasing in the training. Eventually, such a feedback loop will cause gradient explosion, and all samples in the replay buffer will become `NaN`. Under our assumption, to stabilize the training, we must introduce additional regularization methods to break the loop.

Clipping function	# of experiments	Chance of divergence (%)	Valid Acc
None (JEM)	189	70.90	82.79
ELU	415	**0**	**85.46**
ReLU	91	**0**	84.44
LeakyReLU	509	37.29	85.38

Table 4: Empirical evaluation of JEM training stability on PBMC dataset with a Bayesian sweep of hyper-parameters including model architecture, batch size, weight decay, dropout rate, learning rate, buffer size, SGLD step size, and SGLD standard deviation. We find that clipping functions can mostly stabilize the training. ELU and ReLU are robust to a large set of combinations of hyper-parameters.

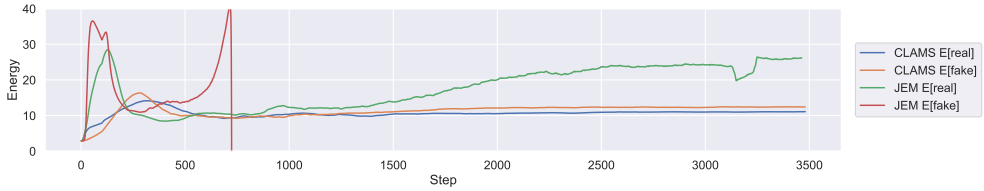


Figure 5: Energy values of CLAMS and JEM: with the presence of the feedback loop, the energy values tend to increase, which will eventually explode. With ELU clipping, the loss becomes smoother, and training is more stable.

Coincidentally, activation functions used in neural network training can be adopted as the clipping function to break the feedback loop. We tested three of them that have different tail behaviour in the negative half-space as shown in Figure 6: ELU: exponentially suppresses negatives values with an asymptote $y = -1$; LeakyReLU linearly suppresses negative values without any asymptote; ReLU eliminates all negative values. Empirically, we run a Bayesian hyper-parameter sweep to evaluate the model and clipping function under a pool of hyper-parameter combinations. We use the results to select the best combinations of hyper-parameters and evaluate the robustness of the clipping function. We do not observe any divergence in the training when regularized with ReLU and ELU and observed a lower chance of divergence when regularized with LeakyReLU. In consideration of other metrics, we selected ELU and conducted additional experiments based on it.

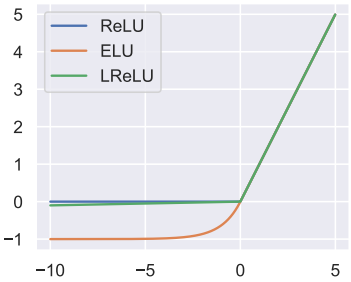


Figure 6: Clipping functions

Besides, when the generative loss is large and negative, the ELU clipping function suffers from the saturation problem [38], which results in almost no learning progress (see ELU $p = 1$ in Figure 7a, 7b). We, therefore, introduce additional regularizations to prevent the clipping unit from being saturated. Referencing the idea of a dropout unit [39], we test a stochastic on-off regularization unit and a hard-threshold on the clipped generative loss such that the generative loss is back-propagated only a fraction of the time. Stochastic regularization behaves exactly the same as a dropout unit acting on the generative loss $\mathcal{L}[p_\theta(\mathbf{x})]$; the hard-threshold acts as a binary switch and only allows the generative loss to pass through when $\mathcal{L}[p_\theta(\mathbf{x})] > t$. Our final model CLAMS is a JEM with ELU clipping and stochastic regularization $p = 0.25$.

D Ablation Study

Our model CLAMS consists of multiple regularization methods acting on the original JEM model, including clipping, stochastic regularization, and/or hard-threshold. To investigate the effect of these techniques, we break down our model and conduct an ablation study.

PBMC	n	Div.	In Acc	In ECE	$p(\mathbf{x})$	$p(y)$
ELU $\mathbf{p = 0.25}$ (Ours)	5	0	92.89 ± 0.72	5.39 ± 0.46	$\mathbf{0.92} \pm 0.01$	0.77 ± 0.02
JEM	5	100	90.32 ± 4.93	$\mathbf{2.61} \pm 0.66$	0.61 ± 0.28	0.86 ± 0.00
$p = 0.25$	5	100	$\mathbf{93.35} \pm 1.05$	6.59 ± 1.51	0.79 ± 0.14	0.78 ± 0.04
$p = 0.1$	5	0	91.81 ± 0.92	6.14 ± 0.50	0.73 ± 0.10	0.75 ± 0.03
$T = -1$	5	0	92.06 ± 0.93	5.06 ± 0.37	0.46 ± 0.15	0.77 ± 0.02
ELU	5	0	93.09 ± 0.33	4.87 ± 0.55	0.36 ± 0.11	0.80 ± 0.01
ELU $p = 0.1$	5	0	92.21 ± 0.87	4.83 ± 0.23	0.68 ± 0.05	0.78 ± 0.028
ELU $T = -1$	5	0	92.03 ± 1.18	5.04 ± 0.85	0.73 ± 0.15	0.77 ± 0.01

(a) Annotation

PBMC	n	Div.	$p(\mathbf{x})$	$p(y)$
ELU $\mathbf{p = 0.25}$ (Ours)	55	0	0.65 ± 0.29	0.53 ± 0.27
JEM	55	100	0.63 ± 0.21	0.44 ± 0.26
$p = 0.25$	55	94.55	0.61 ± 0.29	0.51 ± 0.27
$p = 0.1$	55	0	0.62 ± 0.30	0.51 ± 0.28
$T = -1$	55	0	0.75 ± 0.24	0.54 ± 0.28
ELU	55	0	$\mathbf{0.78} \pm 0.18$	0.55 ± 0.27
ELU $p = 0.1$	55	0	0.60 ± 0.29	0.51 ± 0.28
ELU $T = -1$	55	0	0.71 ± 0.24	0.54 ± 0.28

(b) One-class OOD detection

Table 5: Ablation study of PBMC dataset

In general, we find that any or any combinations of proposed regularization methods can evidently improve JEM's performance in terms of almost all evaluation metrics in our experiments. We observe a positive additive effect by combining loss clipping with other regularizations, which necessarily reduces the effect of saturation. Stochastic regularization with large p may not be sufficient

Pancreas	n	Div.	In Acc	In ECE	$p(\mathbf{x})$	$p(y)$
ELU $\mathbf{p = 0.25}$ (Ours)	5	0	88.38 ± 0.62	5.81 ± 0.75	0.89 ± 0.03	0.71 ± 0.07
JEM	5	100	86.61 ± 2.07	7.57 ± 0.86	$\mathbf{0.92} \pm 0.06$	0.54 ± 0.13
$p = 0.25$	5	20	87.31 ± 1.21	6.02 ± 1.46	$\mathbf{0.92} \pm 0.05$	0.78 ± 0.09
$p = 0.1$	5	0	87.88 ± 0.56	5.73 ± 1.16	0.90 ± 0.04	0.76 ± 0.04
$T = -1$	5	0	88.50 ± 0.39	$\mathbf{4.37} \pm 0.55$	0.86 ± 0.03	0.63 ± 0.10
ELU	5	0	86.89 ± 0.50	6.19 ± 0.52	0.91 ± 0.05	0.73 ± 0.03
ELU $p = 0.1$	5	0	$\mathbf{88.85} \pm 0.74$	4.50 ± 0.80	0.88 ± 0.05	0.71 ± 0.04
ELU $T = -1$	5	0	88.55 ± 0.37	4.96 ± 0.68	0.87 ± 0.04	0.71 ± 0.04

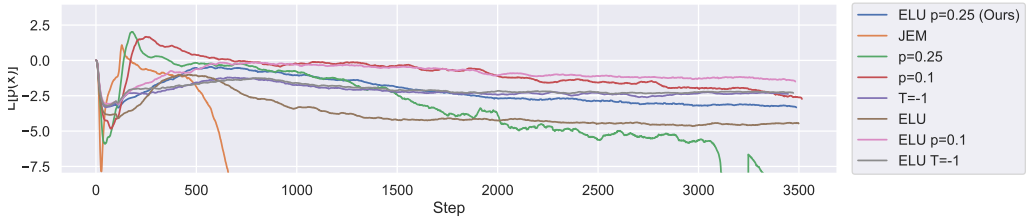
(a) Annotation

Pancreas	n	Div.	$p(\mathbf{x})$	$p(y)$
ELU $\mathbf{p = 0.25}$ (Ours)	95	0	0.83 ± 0.19	0.86 ± 0.09
JEM	95	100	0.73 ± 0.22	0.74 ± 0.16
$p = 0.25$	95	29.47	0.84 ± 0.15	0.84 ± 0.12
$p = 0.1$	95	0	0.84 ± 0.15	0.85 ± 0.10
$T = -1$	95	0	0.84 ± 0.19	0.86 ± 0.09
ELU	95	0	$\mathbf{0.85} \pm 0.15$	$\mathbf{0.87} \pm 0.09$
ELU $p = 0.1$	95	0	0.82 ± 0.20	0.86 ± 0.10
ELU $T = -1$	95	0	$\mathbf{0.85} \pm 0.17$	0.86 ± 0.09

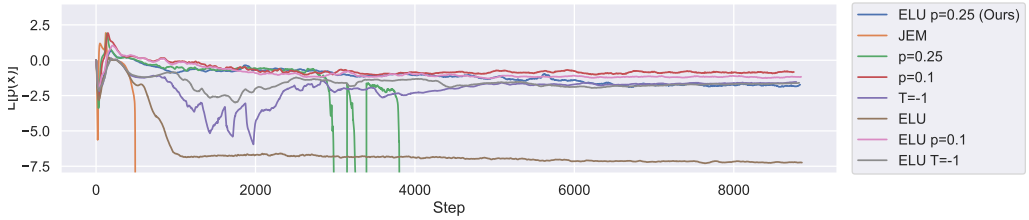
(b) One-class OOD detection

Table 6: Ablation study of pancreas dataset

to stabilize the training but we still observe a boost in results. An alternative interpretation to this is from the perspective of generative loss. In Figure 7, we observe that our regularization methods accomplish this by maintaining the loss in a reasonable range. Compared with the baseline JEM, the generative loss stays at higher values, and combinations of these regularization methods further push the generative loss upwards. Such observations confirm our hypothesis made in Section C that the generative loss is the source of instability, and breaking the implicit feedback loop will resolve the problem.



(a) Generative loss ($\mathcal{L}[p_\theta(\mathbf{x})]$) of pancreas dataset (MA100)

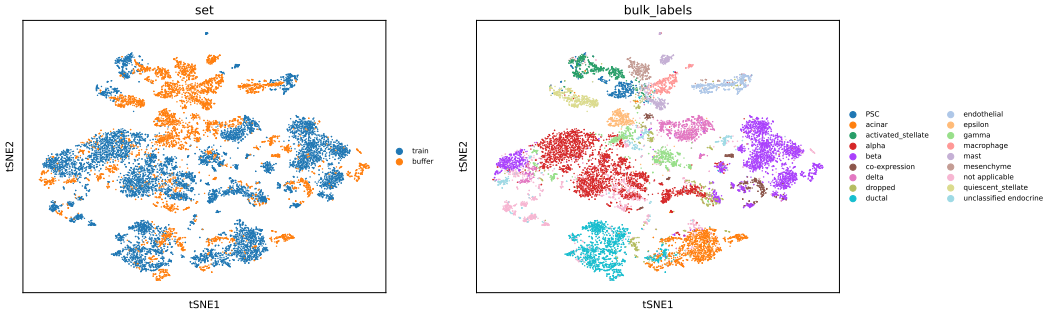


(b) Generative loss ($\mathcal{L}[p_\theta(\mathbf{x})]$) of PBMC dataset (MA100)

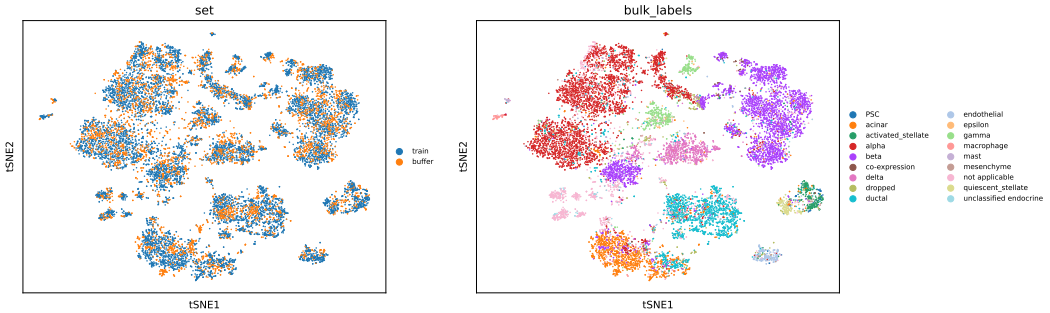
Figure 7: Generative loss ($\mathcal{L}[p_\theta(\mathbf{x})]$): we observe the generative loss have an impact on the training stability. The accumulation of negative loss throughout training tends to lead to divergence. Applying clipping with or without additional regularization can evidently stabilize the training. Also, the usage of additional regularization reduces the effect of the saturation problem, which is reflected in higher loss values.

E Modified SGLD Sampler

We notice that the SGLD sampler with a class-balanced replay buffer used in JEM may produce some artifacts when the training data is class-imbalanced (e.g., top orange clusters in Figure 8a). We look further into this part and find that those classes are small subgroups that only take up $\tilde{5}\%$ in the training set. In other words, we are trying to generate more data points than the dataset has. Works done by [3, 2, 40] mostly conduct experiments on CIFAR datasets where classes are balanced and no significant artifacts were observed. For class-imbalanced datasets, we test SGLD sampling with a modified class-imbalanced replay buffer, where classes that have more samples in the training set, have more samples in the replay buffer as well. This can also be considered as diluting minority classes in the replay buffer.



(a) Data generation with SGLD sampler and a balanced buffer



(b) Data generation with SGLD sampler and a imbalanced buffer (Ours)

Figure 8: Data generation: with the our imbalanced replay buffer, the mixing is improved (left) and, class distributions are also well-preserved (right).

F Additional Results

Clipping function	# of experiments	Chance of divergence (%)	Valid Acc
None (JEM)	393	15.77	90.91
ELU	808	**0**	**92.25**
ReLU	212	**0**	91.88
LeakyReLU	1164	1.34	91.99

Table 7: Empirical evaluation of JEM training stability on pancreas dataset with a Bayesian sweep of hyper-parameters including model architecture, batch size, weight decay, dropout rate, learning rate, buffer size, SGLD step size, SGLD standard deviation.

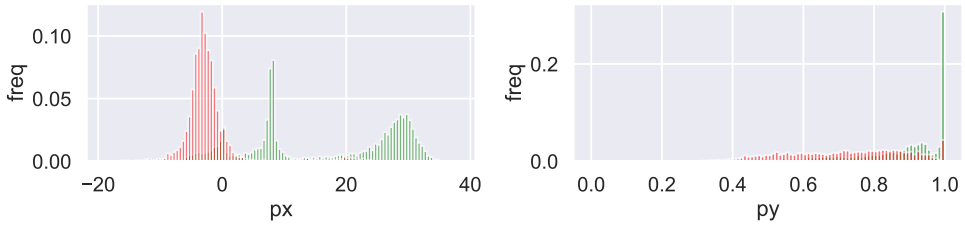


Figure 9: OOD scores on PBMC dataset: unnormalized density $p(\mathbf{x})$ (left), predictive class probability $p(y)$ (right). In-distribution data are colored in green; out-of-distribution data are colored in red.

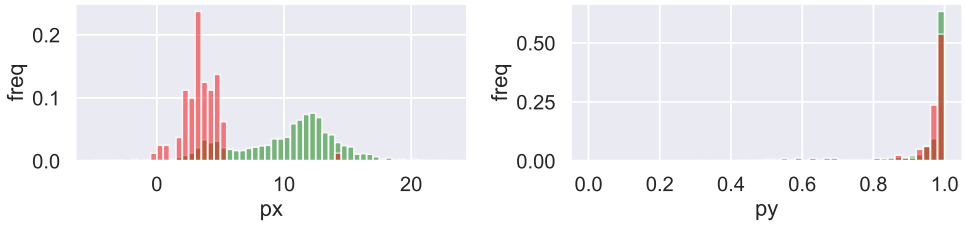


Figure 10: OOD scores on pancreas dataset