# Energy-based Modelling for Single-cell Data Annotation

Tianyi Liu [1,2] Philip Fradkin [1, 2] Lazar Atanackovic [1,2] Leo J. Lee [1,2]

[1] University of Toronto [2] Vector Institute

https://psi.toronto.edu

VECTOR INSTITUTE

## Contribution Summary

1. We introduced **Energy Based Models**, a family of probabilistic models, to **single-cell data** modelling.
2. We developed CLAMS, a customized version of EBM, and accomplished **robust and well-calibrated** results for scRNA-seq annotation.
3. We explored **generative modelling** with EBMs, and demonstrated that our model outperforms state-of-the-art methods for OOD detection.

## Energy-based Model Background

EBMs can be interpreted as parameterizing a probabilistic distribution based on its energy value as:

$$p_\theta(\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{x}))}{Z(\theta)}, \tag{1}$$

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

Class predictive distribution $p(y \mid \mathbf{x})$ becomes the main-stream method for OOD detection in single-cell annotation. Our work investigates generative OOD modelling while annotating cell types. One of the notable works of EBM is JEM (Grathwohl et al., 2019), which accomplishes simultaneous generative and discriminative data modelling. JEM defines a joint distribution of a data point $\mathbf{x}$ and its label $y$ as:

$$p_\theta(\mathbf{x}) = \frac{\exp\left[\log\left(\sum_y \exp(f_\theta(\mathbf{x})[y])\right)\right]}{Z(\theta)}. \tag{2}$$

where $f_\theta(\mathbf{x})[y]$ is the $y^{\text{th}}$ logit of $p(y \mid \mathbf{x})$.

The maximum likelihood objective can be formulated as the sum of generative and discriminative losses:

$$\ell_{\text{ML}} = \log p_\theta(\mathbf{x}, y) = \log p_\theta(\mathbf{x}) + \log p_\theta(y \mid \mathbf{x}). \tag{3}$$

JEM adopts the standard EBM training strategy by retaining a replay buffer with stochastic gradient Langevin dynamics (SGLD) sampling.

## CLAMS Formulation

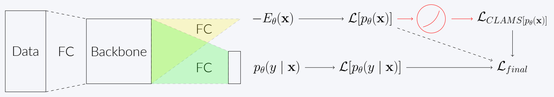


Figure 1. CLAMS block diagram

CLAMS boosts JEM's performance by applying regularizations on the generative loss $\mathcal{L}[p_\theta[(x)]$:

- Loss clipping with exponential linear unit.
- Stochastic regularization with $p = 0.25$.

## CLAMS Accuracy Calibration and Robustness

We ran scRNA-seq annotation experiment with pancreas and PBMC dataset pairs. To assess the effectiveness of our model, we evaluated three properties:

1. Effectiveness of the predictor (In-distribution accuracy (In Acc); Chance of Divergence (Div. %))
2. Predictor calibration (In ECE)
3. Model robustness (AUROCs of $p(\mathbf{x})$ and $p(y \mid \mathbf{x})$ score for OOD detection.)

| *PBMC* | $n$ | Div. | In Acc | In ECE | $p(\mathbf{x})$ | $p(y \mid \mathbf{x})$ |
|---|---|---|---|---|---|---|
| **CLAMS (Ours)** | 5 | 0 | 92.89 | 5.39 | **0.92** | 0.77 |
| JEM | 5 | 100 | 90.32 | **2.61** | 0.61 | 0.86 |
| VERA | 5 | 0 | 88.76 | 6.56 | 0.87 | 0.80 |
| HDGE | 5 | 0 | **93.58** | 3.45 | 0.81 | 0.78 |
| HDGE + JEM | 5 | 100 | 92.29 | 6.25 | 0.42 | 0.70 |
| ResNet | 5 | 0 | 91.53 | 4.73 | – | 0.79 |
| $\text{SVM}_{reject}$ | 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 |

## Label Transfer Experiment

We assessed a model trained on a reference dataset by evaluating its ability to generalize to a query dataset with a partially overlapping set of cell types and different class balances. We observed that CLAMS was able to almost perfectly identify B cells, T cells, dendritic cells, and NK cells. For cell types such as HSPC that are not present in the reference dataset, the model labelled them as OOD data points and rejected them.

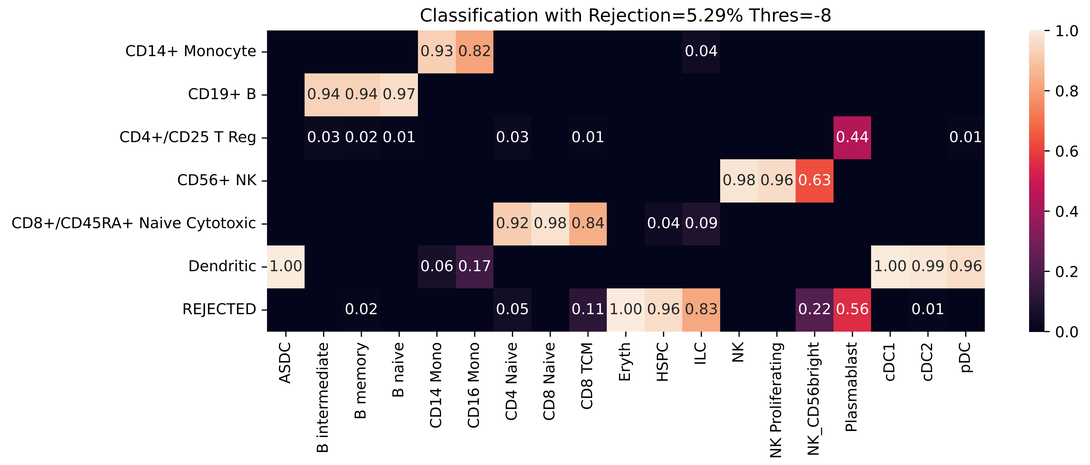


Figure 2. A partial annotation heatmap of PBMC dataset: rows are reference labels and columns are query labels.

## Learned Sample Visualization

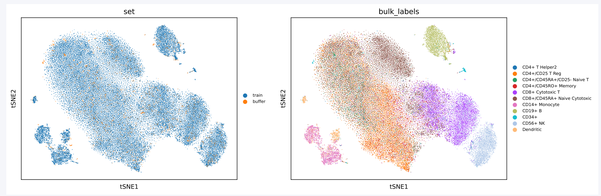


Figure 3. Generated data in $t$-SNE space: we generated 300 samples per class. On the left, data points are colored based on the source of data: training data are colored in blue, and generated data are colored in orange. On the right, data points are colored by cell types.

## Further Out of Distribution Experiments

We further investigated the OOD detection capability by designing an OOD detection experiment: we generated sub-datasets by leaving one out and evaluated the model's ability to identify leave-out classes. To reduce the effect of AUROC dilution due to underlying cell type correlation, we reported results on selected cell types.

| *PBMC* | $n$ | Div. | $p(\mathbf{x})$ | $p(y \mid \mathbf{x})$ |
|---|---|---|---|---|
| **CLAMS (Ours)** | 40 | 0 | **0.78** | 0.68 |
| JEM | 40 | 100 | 0.64 | 0.58 |
| VERA | 40 | 0 | 0.65 | 0.63 |
| HDGE | 40 | 0 | 0.66 | 0.68 |
| HDGE + JEM | 40 | 100 | 0.70 | 0.69 |
| ResNet | 40 | 0 | – | 0.63 |
| $\text{SVM}_{reject}$ | 8 | – | – | 0.75 |

## Key References

[1] 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.

Please find more details in our paper!