ALLEN INSTITUTE for BRAIN SCIENCE

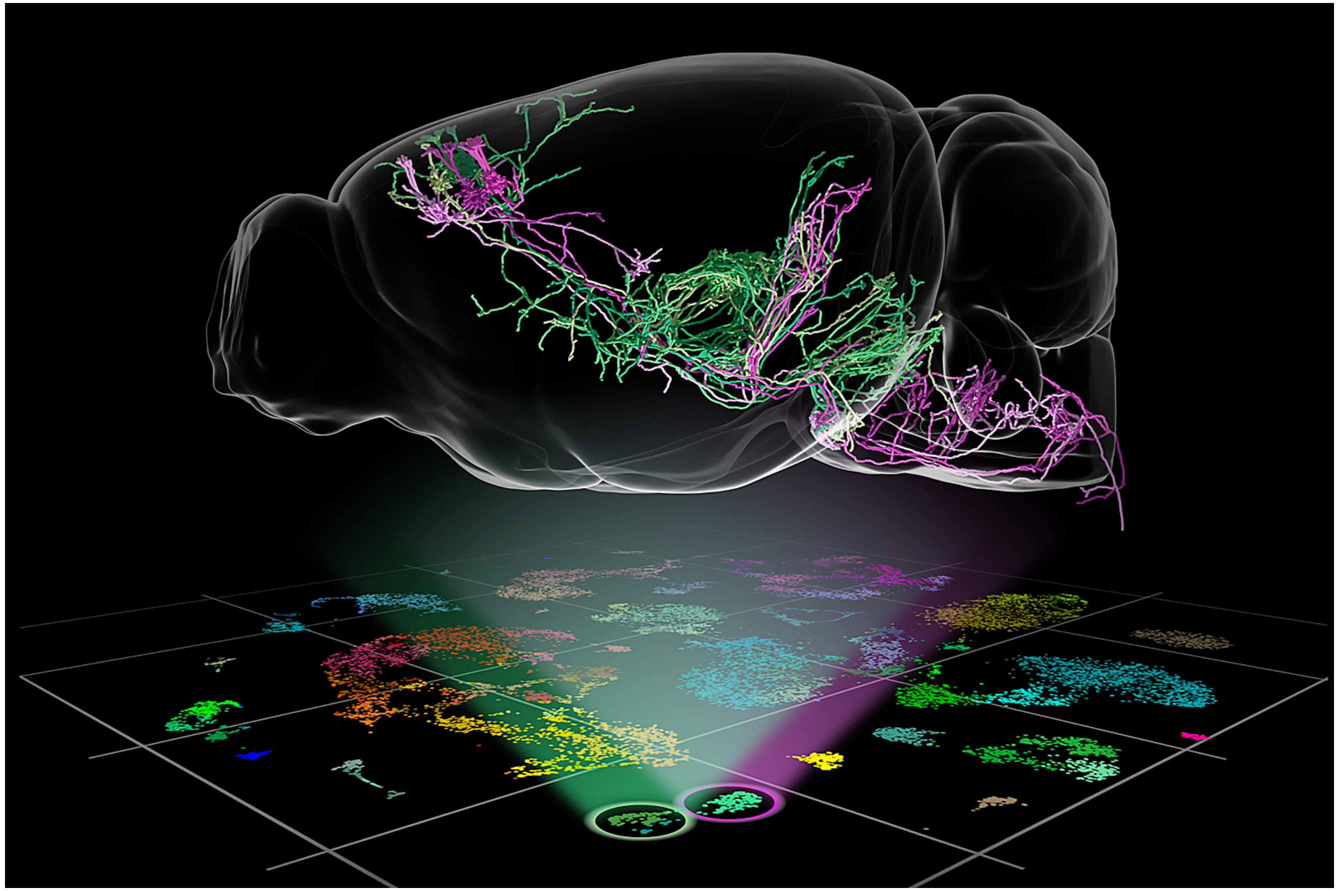
Joint identification of neuron types and type-specific activity-regulated genes with coupled autoencoders

04/12/2021

Yeganeh Marghi

Motivation

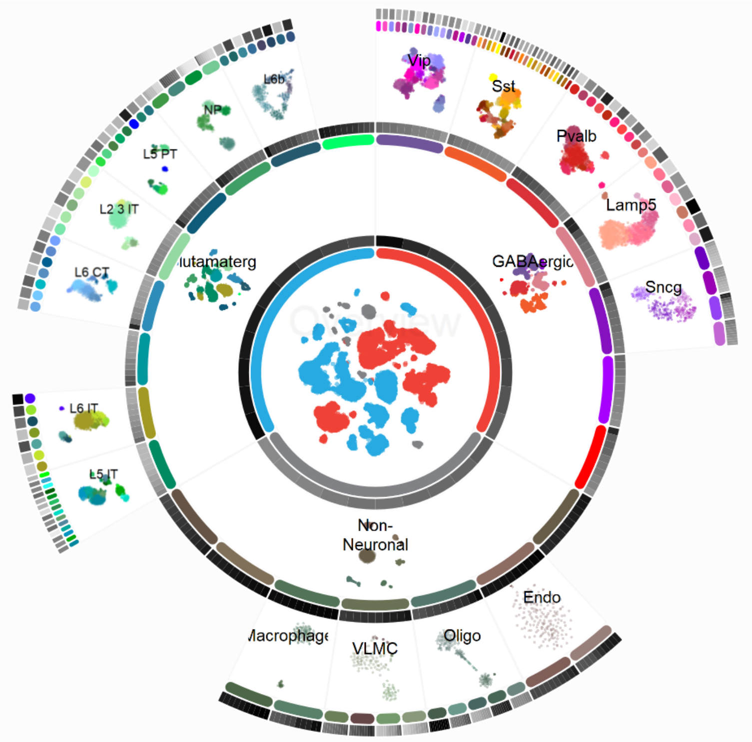
- From in-depth analysis of cells to understanding the brain
- Studying the whole brain at single-cell resolution by single-cell omics
- The potential to unravel the molecular programs underlying the cellular diversity



Allen Institute for Brain Science, Press releases, 2018.

Motivation

- From in-depth analysis of cells to understanding the brain
- Studying the whole brain at single-cell resolution by single-cell omics
- The potential to unravel the molecular programs underlying the cellular diversity
- Measurement noise and biological variation cause significant challenges

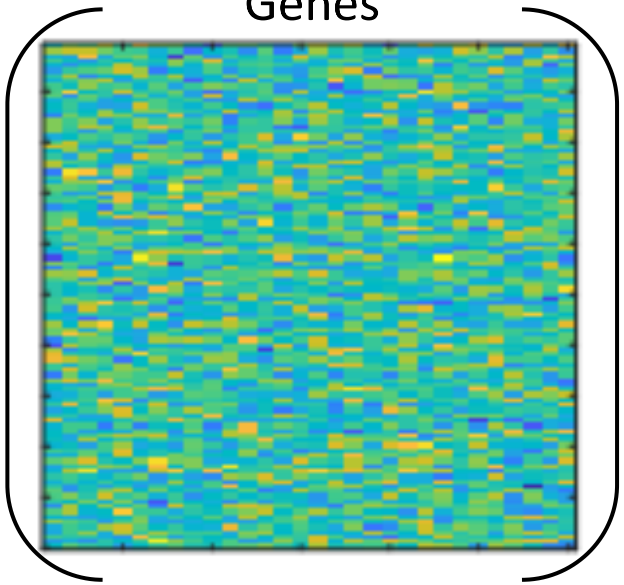


Allen Institute for Brain Science, Press releases, 2018.

X

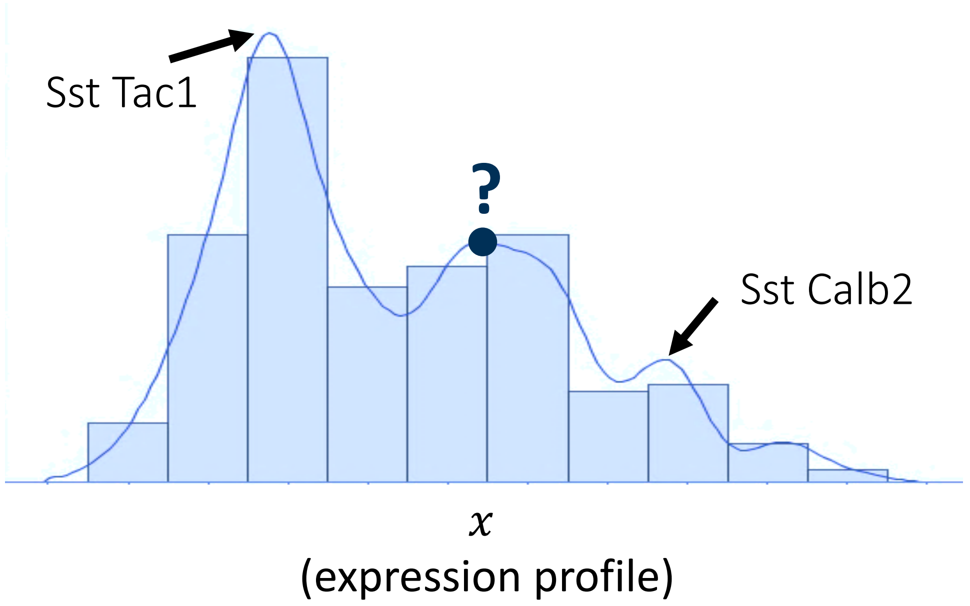
Genes

Cells



Single-cell data: a mixture landscape

Mixture models: measurement is a function of two (random) variables.

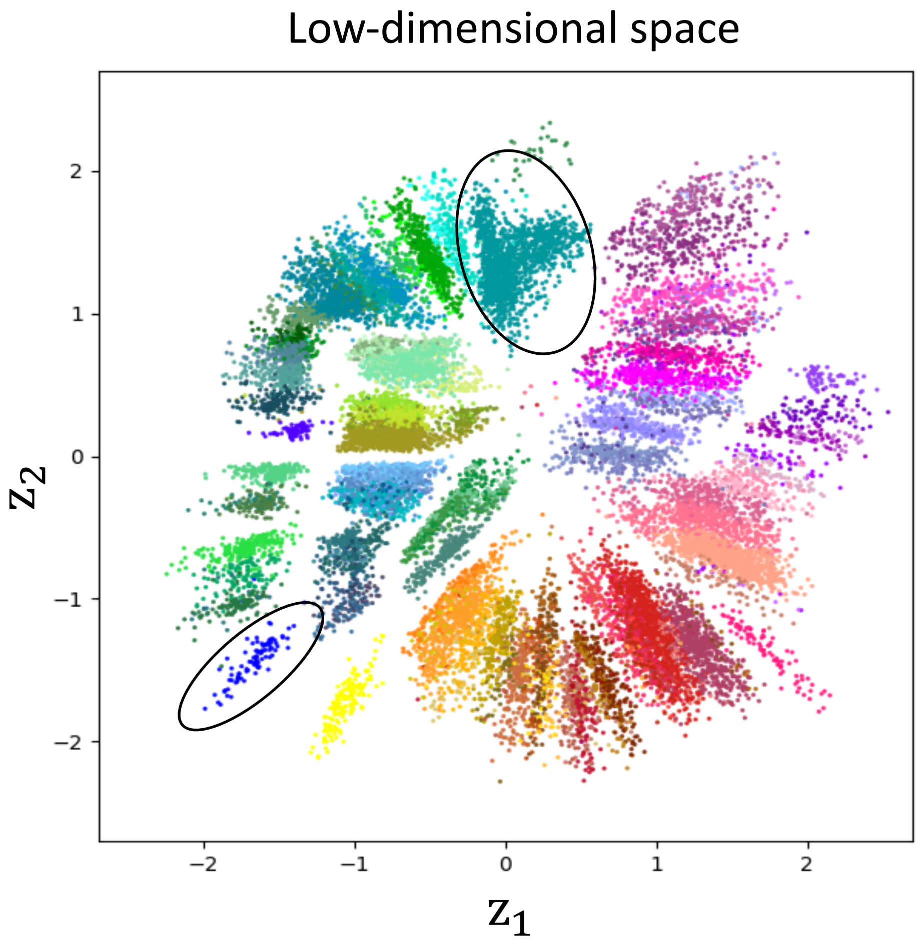


x: scRNA-seq data
c: cell type (discrete factor)
s: cell type-dependent variations (continuous factor)

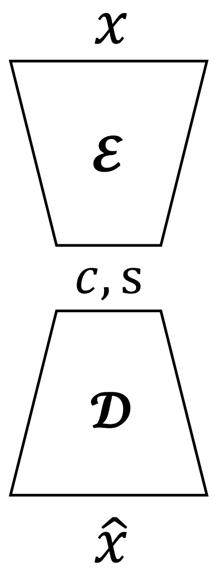
$$x = f(c, s)$$

Single-cell data: a mixture landscape

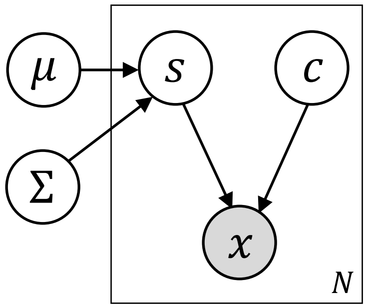
Mixture models: measurement is a function of two (random) variables.

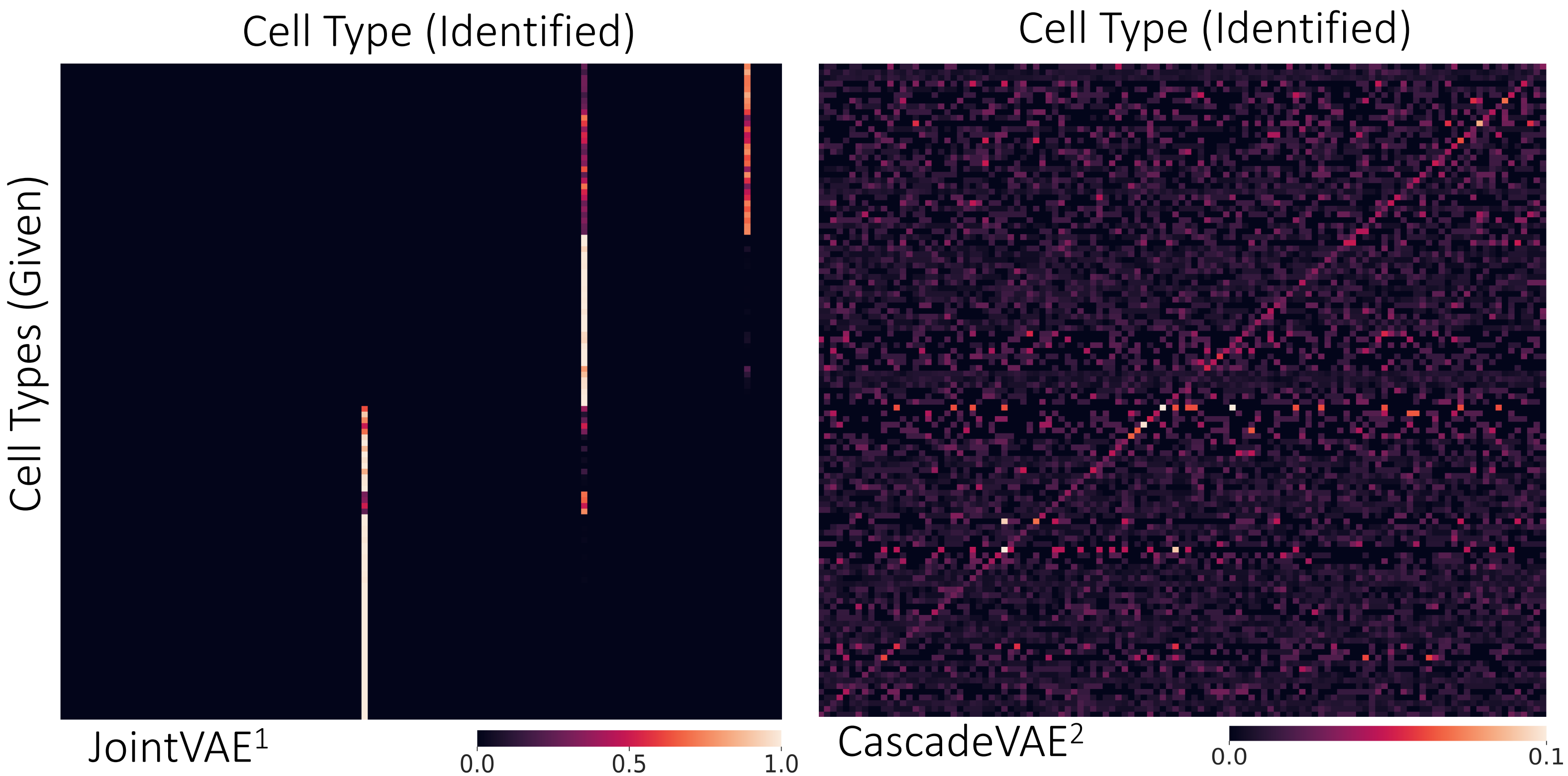


Mixture representation learning: variational approach



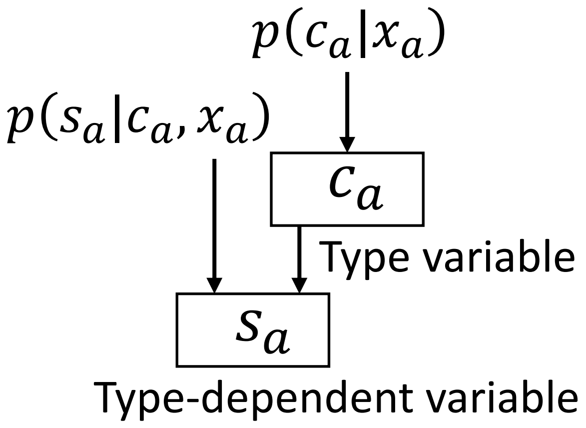
$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{s},\mathbf{c}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s},\mathbf{c})\right] - D_{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})\|p(\mathbf{s})\right) - D_{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{c}|\mathbf{x})\|p(\mathbf{c})\right)$$

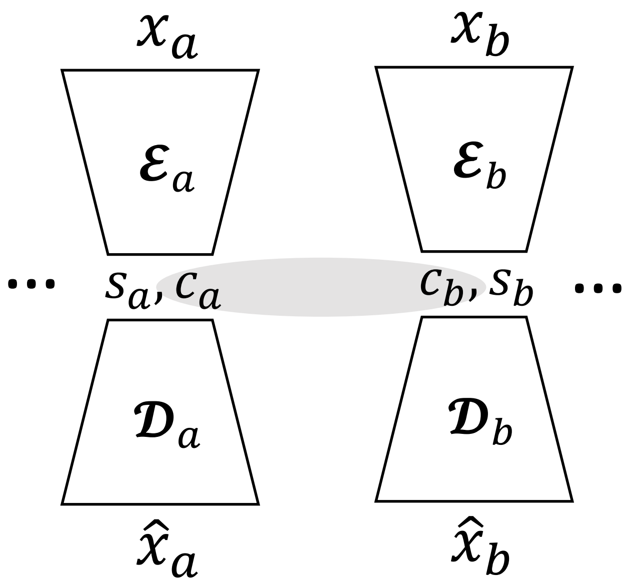


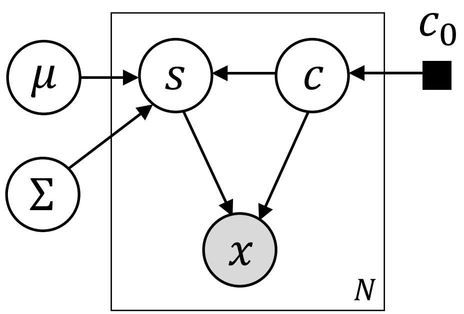


1. Dupont, Emilien. "Learning disentangled joint continuous and discrete representations." *NeurIPS*, 2018.
2. Jeong, Yeonwoo, and Hyun Oh Song. "Learning discrete and continuous factors of data via alternating disentanglement." ICML, 2019.

Mixture representation learning: variational approach





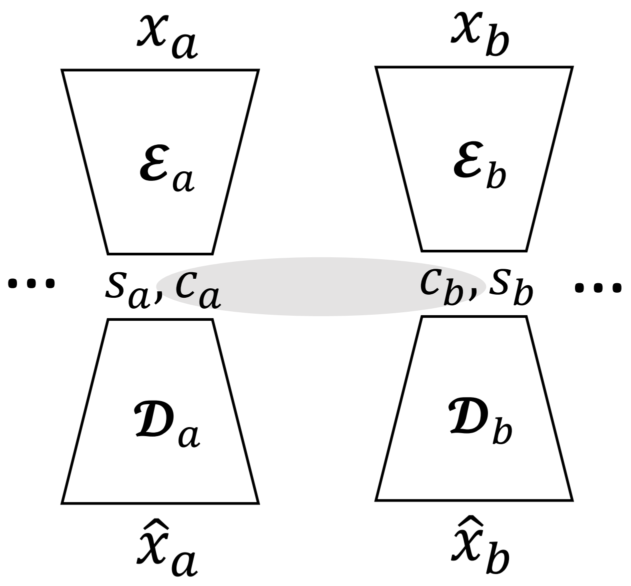


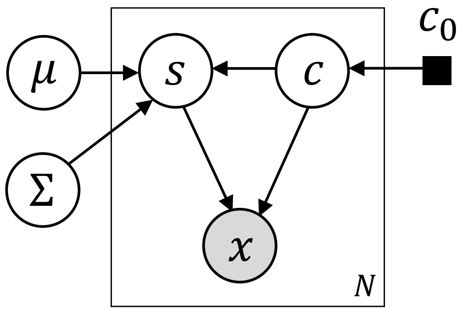
$$x_a = f(c_a, s_a)$$

$$f(c_a, s_a) = p(c_a)p(s_a|c_a)$$

c_a: cell type (discrete)
s_a: cell type-dependent variations (continuous)

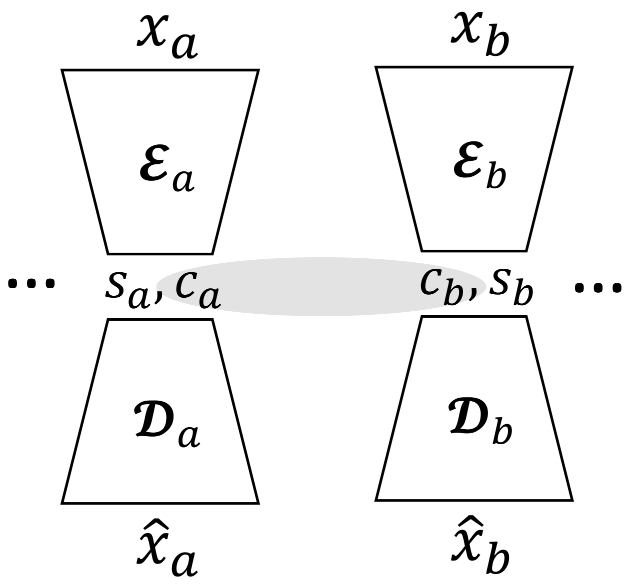
Coupled mixture VAE framework (cpl-mixVAE)

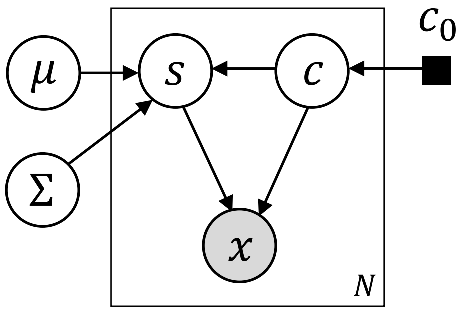




Objective function: $$\max \sum_{a=1}^{A} (A-1)\left(\mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}[\log p(\mathbf{x}_a|\mathbf{s}_a,\mathbf{c}_a)] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}[D_{KL}(q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a))]\right)$$
$$-\sum_{a<b} \mathbb{E}_{q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)}\mathbb{E}_{q(\mathbf{s}_b|\mathbf{c}_b,\mathbf{x}_b)}[D_{KL}(q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b)\|p(\mathbf{c}_a,\mathbf{c}_b))]$$
$$\text{s.t. } \mathbf{c}_a = \mathbf{c}_b \quad \forall a,b \in [1,A],\ a<b$$

Coupled mixture VAE framework (cpl-mixVAE)





Objective function: $\max \sum_{a=1}^{A} \mathbb{E}_{q(\mathbf{s}_a, \mathbf{c}_a | \mathbf{x}_a)} \left[\log p(\mathbf{x}_a | \mathbf{s}_a, \mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a | \mathbf{x}_a)} \left[D_{KL}\left(q(\mathbf{s}_a | \mathbf{c}_a, \mathbf{x}_a) \| p(\mathbf{s}_a | \mathbf{c}_a)\right)\right] + H(\mathbf{c}_a | \mathbf{x}_a)$

$$s.t.\ \mathbb{E}_{q(\mathbf{c}_a | \mathbf{x}_a)} \left[d^2(\mathbf{c}_a, \mathbf{c}_0)\right] < \epsilon$$

Coupled mixture VAE framework (cpl-mixVAE)

Proposition 1. *Consider the problem of mixture representation learning in a multi-arm VAE framework. For* $A > B \geq 1$ *and* $\forall m$,

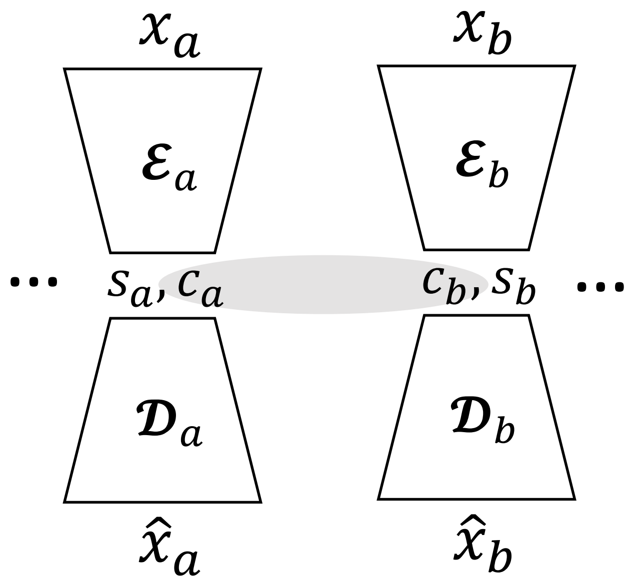
$$\mathcal{C}_m^A(m) > \mathcal{C}_m^B(m).$$

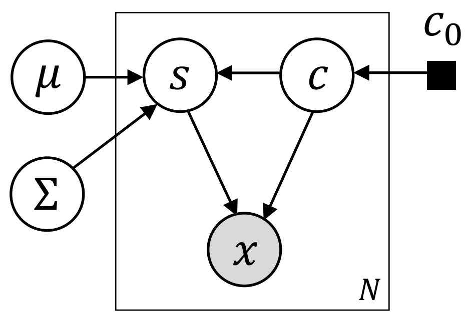
Proposition 2. *In the A-arm VAE framework, there exists an* A *such that* $\forall m, n,\ m \neq n$,

$$\mathcal{C}_m^A(m) > \mathcal{C}_m^A(n),$$

independent of the relative abundances of categories.

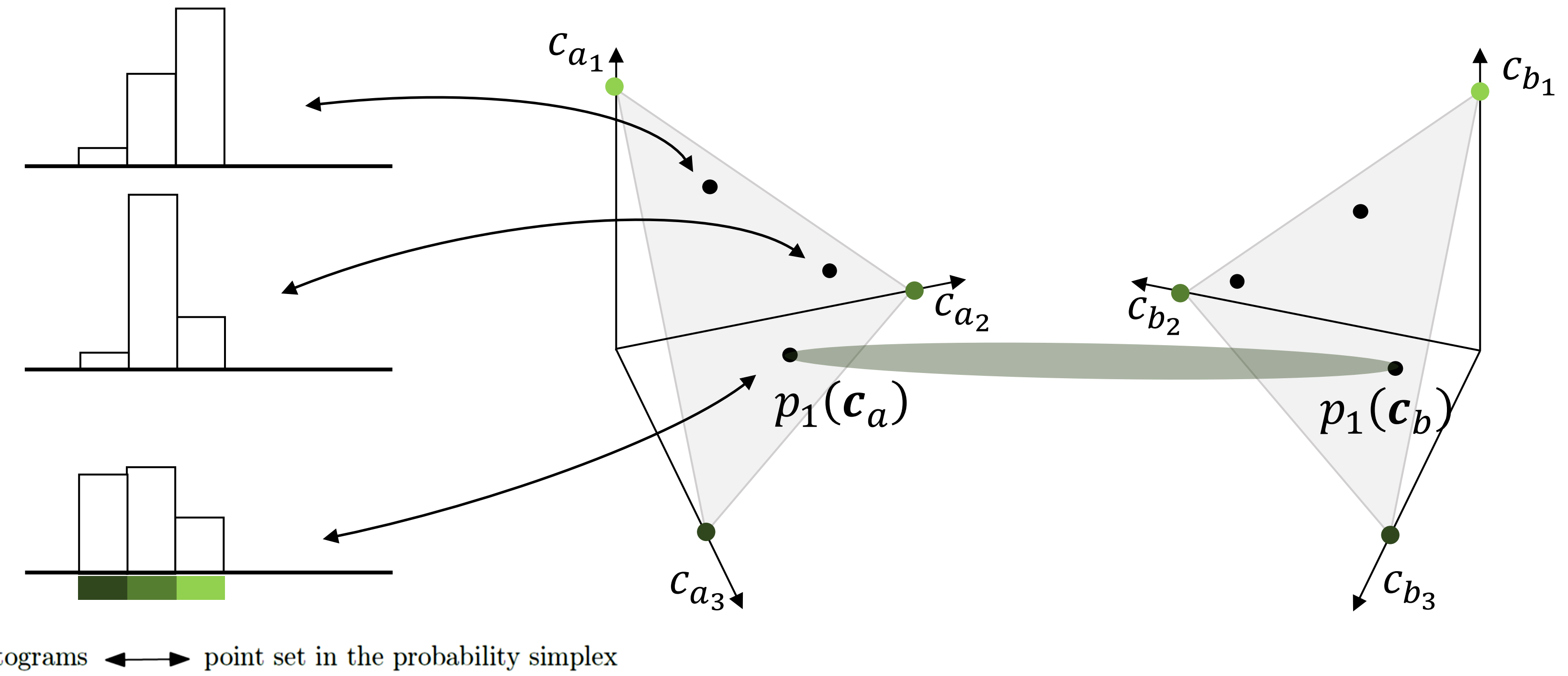
$$\mathcal{C}_m(k) = \mathbb{E}_{\mathbf{x}|m}\left[\log p(c = k|\mathbf{x})\right]$$





Marghi, Yeganeh M., Rohan Gala, and Uygar Sümbül. "Joint Learning of Discrete and Continuous Variability with Coupled Autoencoding Agents." *arXiv preprint arXiv:2007.09880* (2020).

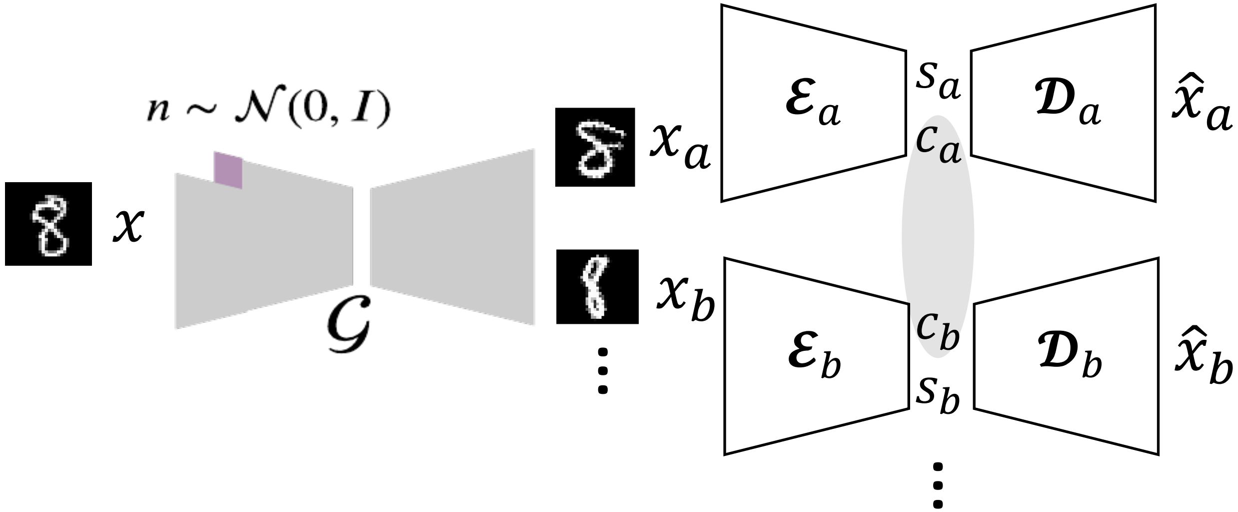
Consensus assignment



Using Aitchison geometry: $d(\boldsymbol{c}_a, \boldsymbol{c}_b) = D_A(\boldsymbol{c}_a, \boldsymbol{c}_b), \quad \boldsymbol{c}_a, \boldsymbol{c}_b \in \boldsymbol{S}^K$

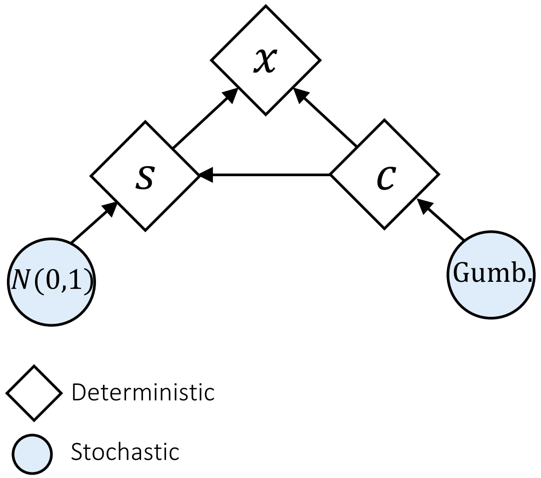
Analogy in machine learning

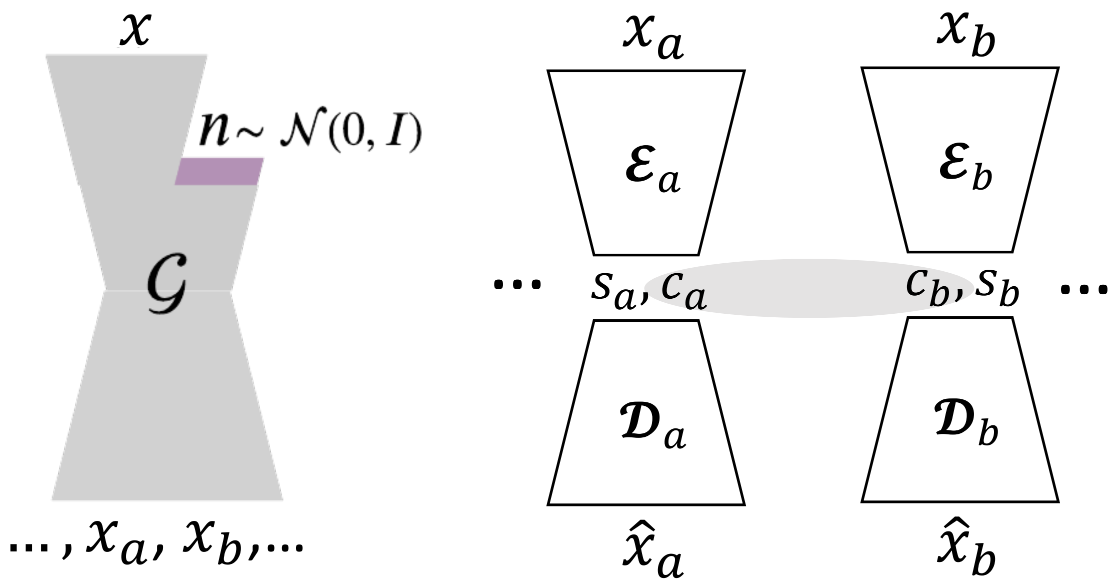
The MNIST dataset



A-arm VAE framework

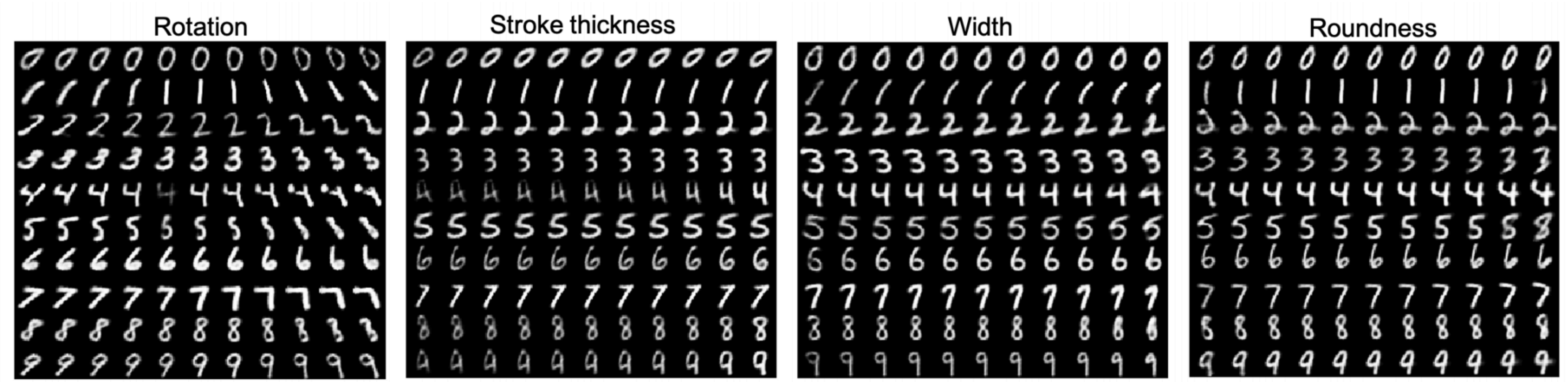
Network implementation



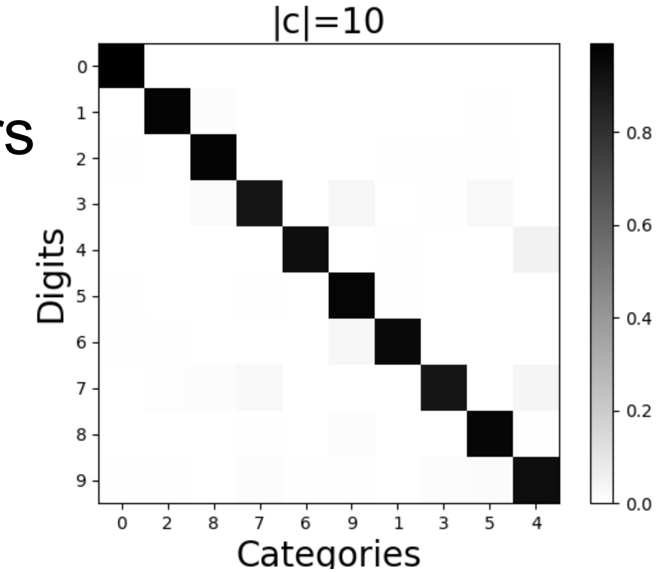


Benchmark dataset: interpretation of c & s

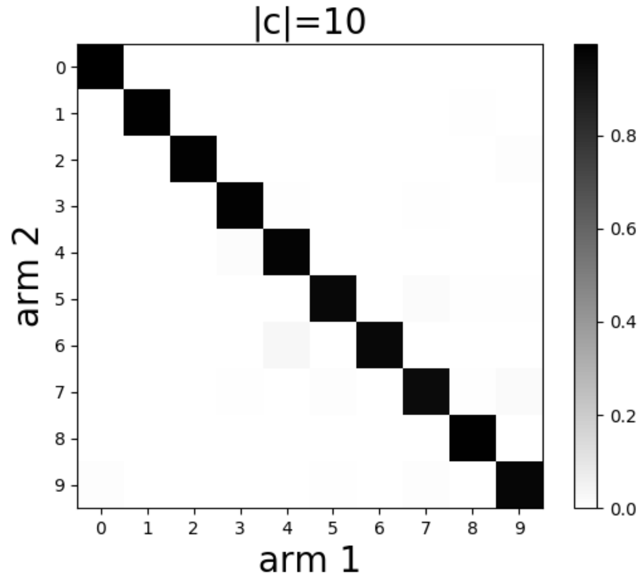
Continuous factors



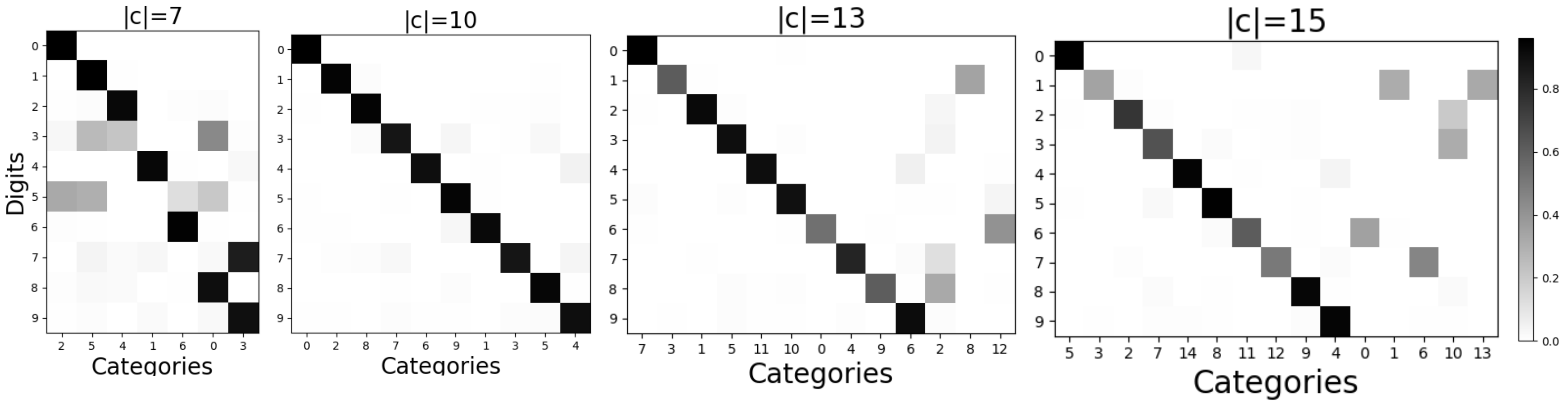
Discrete factors

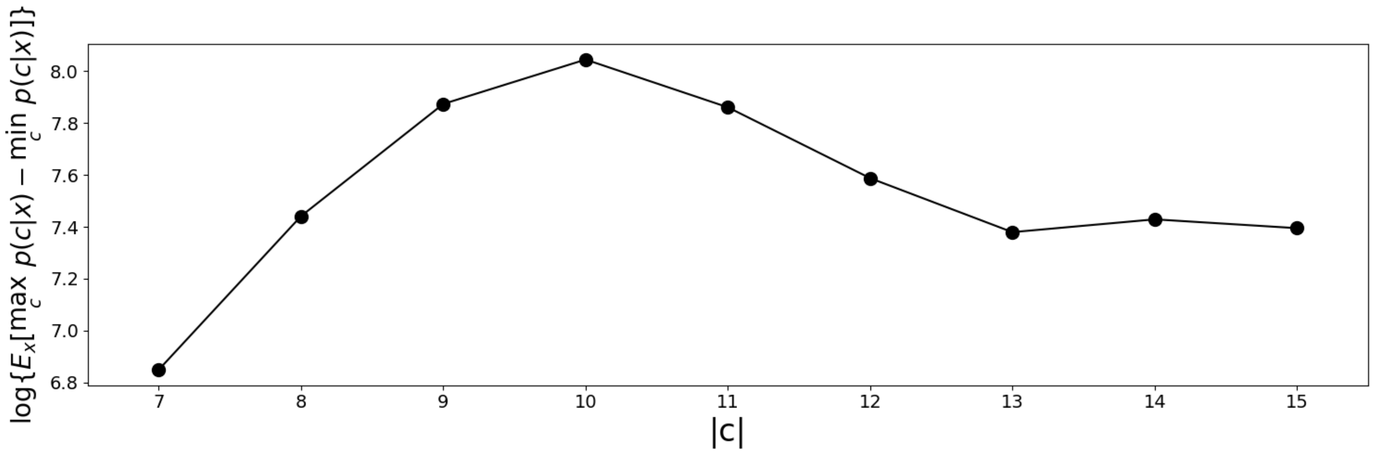


Consensus among arms

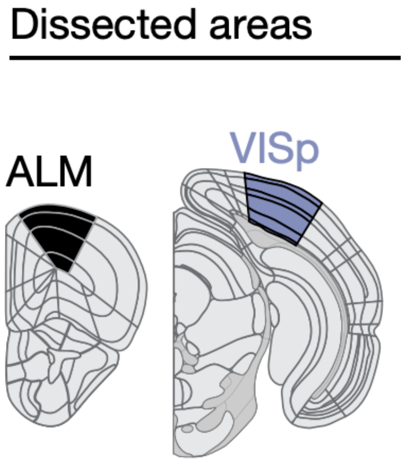


Benchmark dataset: unknown $|c|$

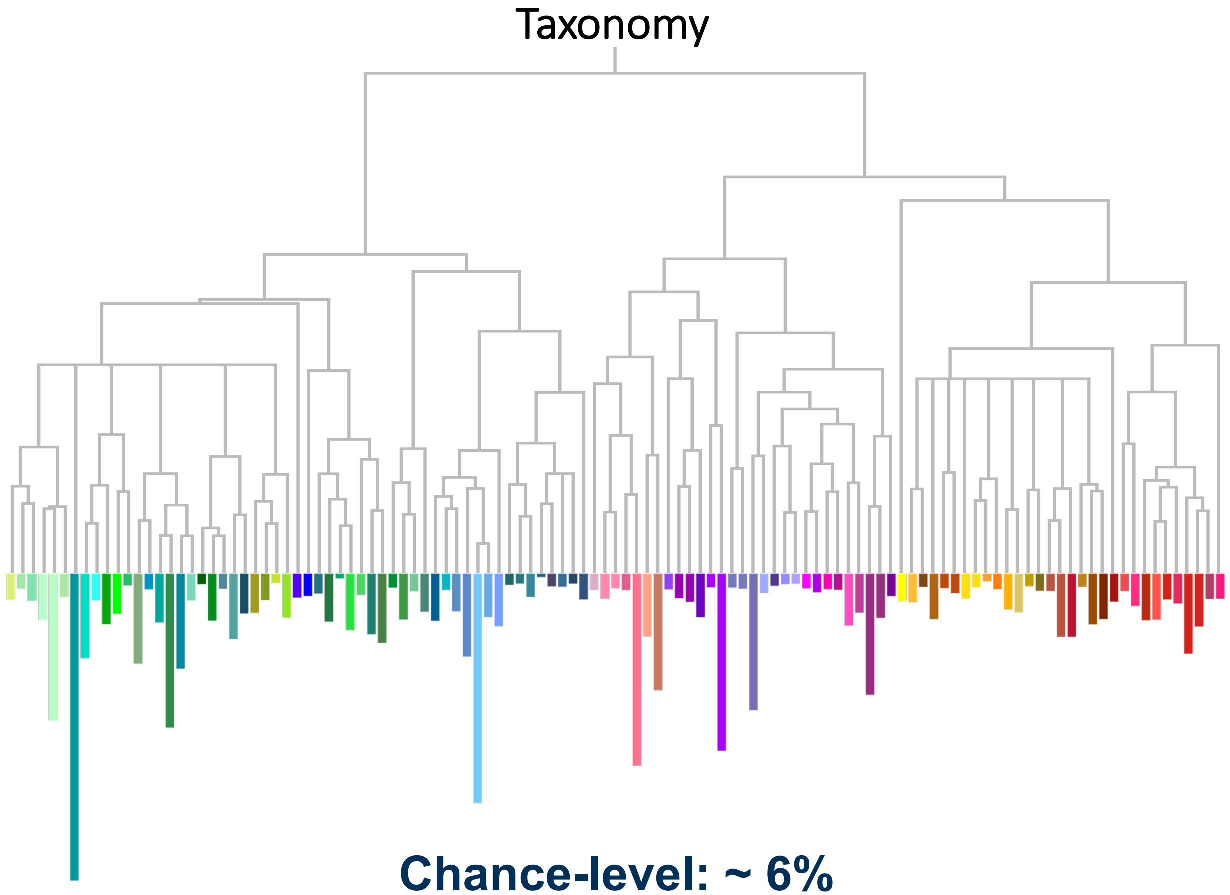




scRNA-seq dataset (Tasic et al., 2018)

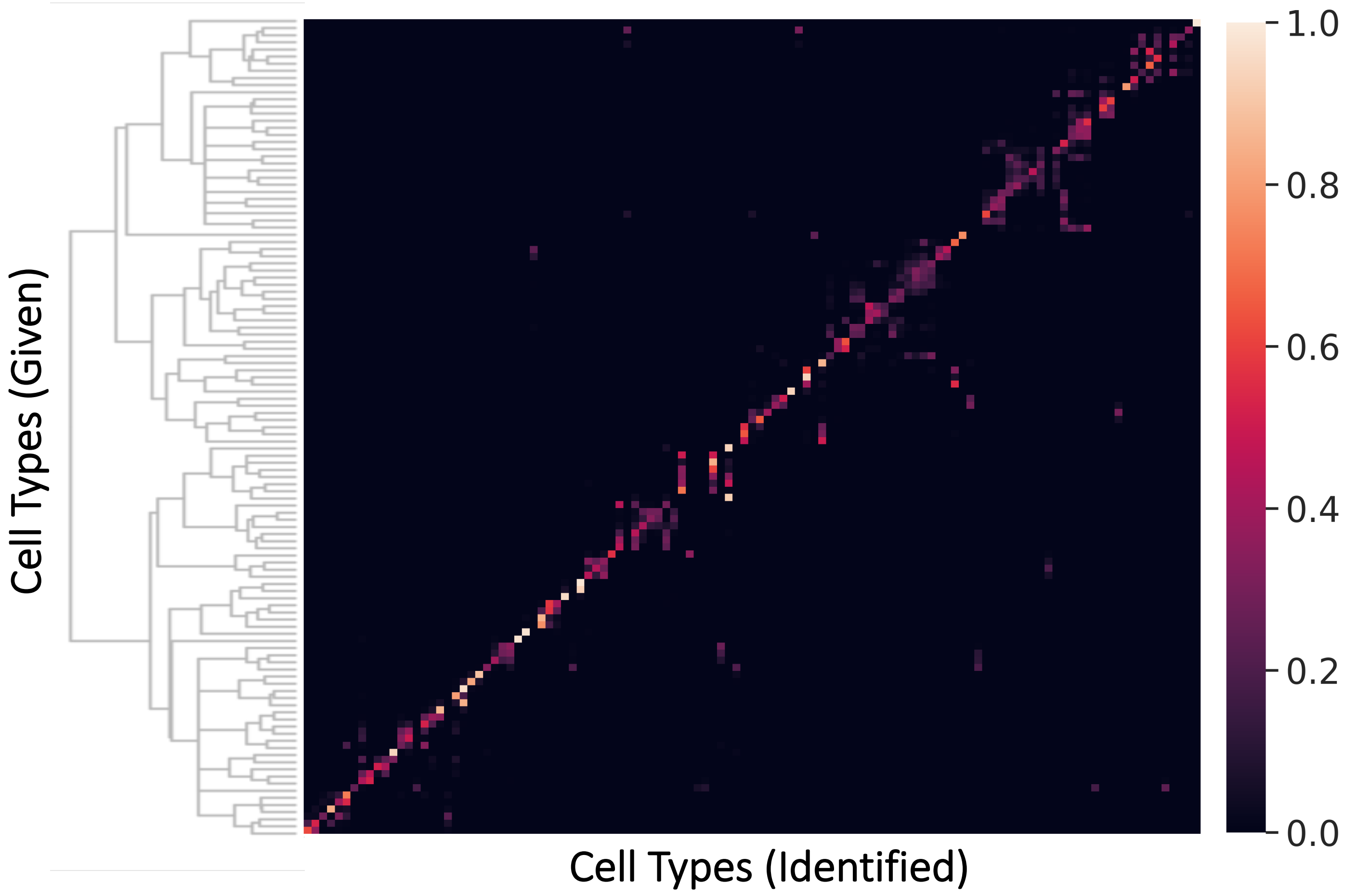


- transcriptomic profiles for **22,365** cells
- **115** excitatory and inhibitory neuron types
- 5000 DE genes

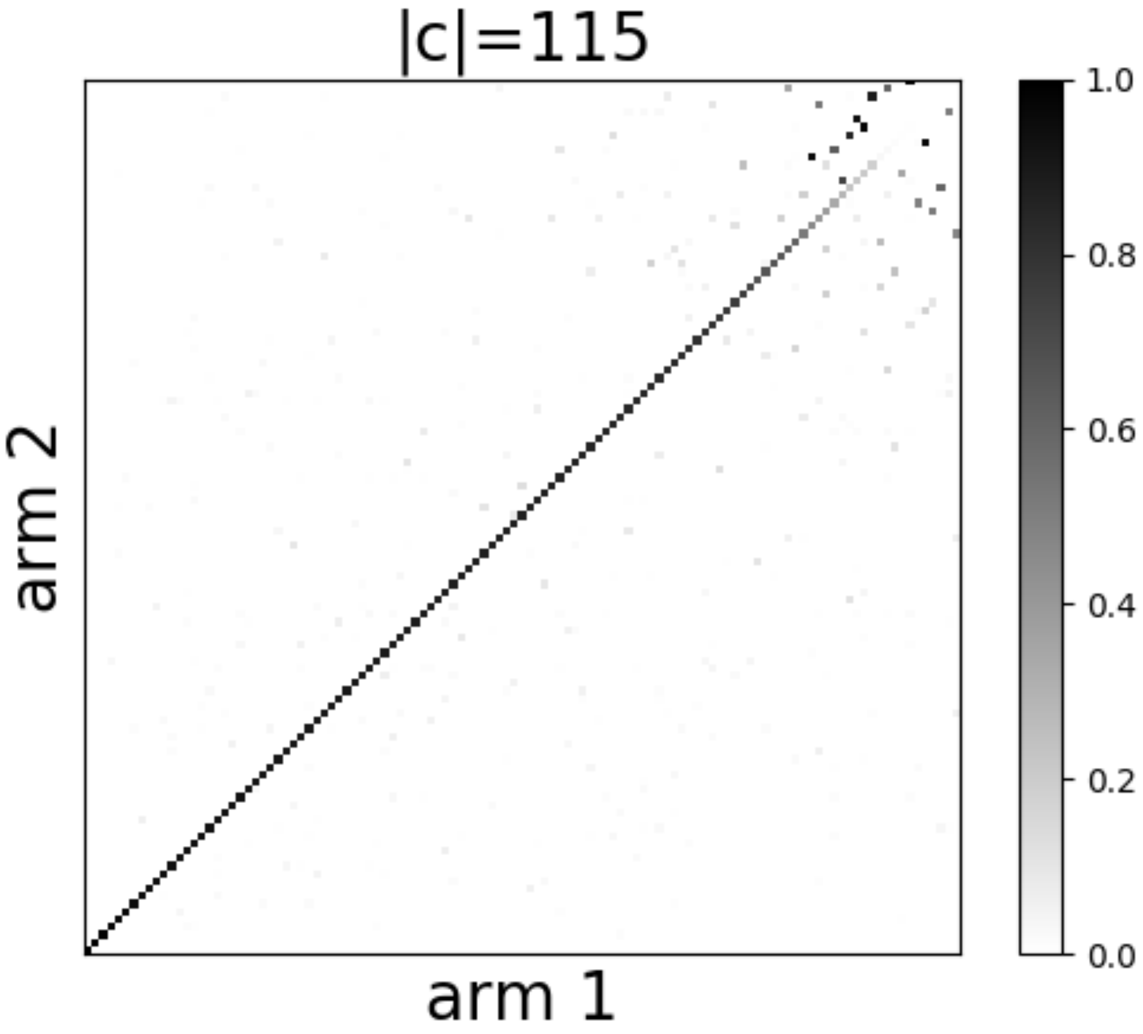


Chance-level: ~ 6%

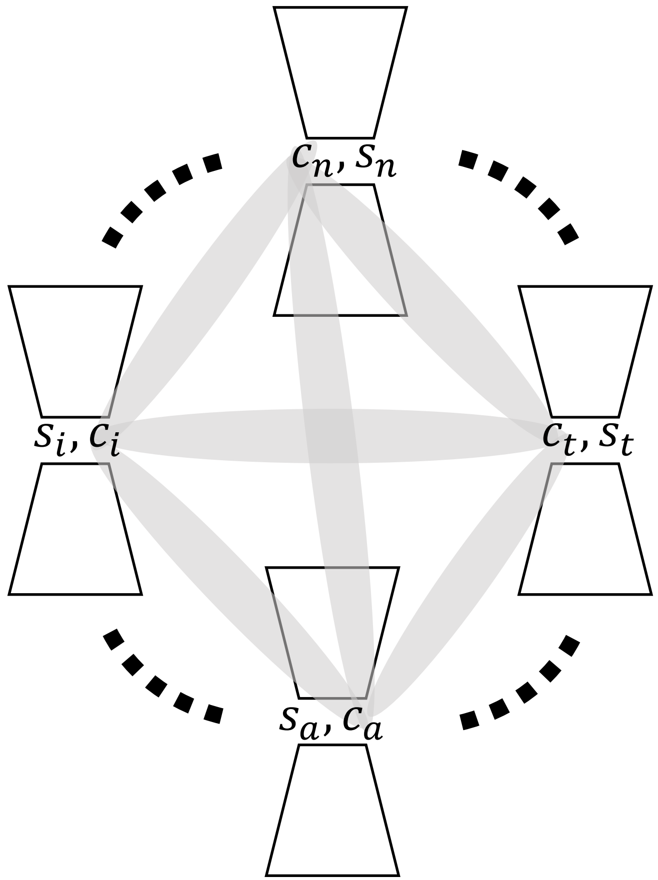
scRNA-seq dataset: transcriptomic identities

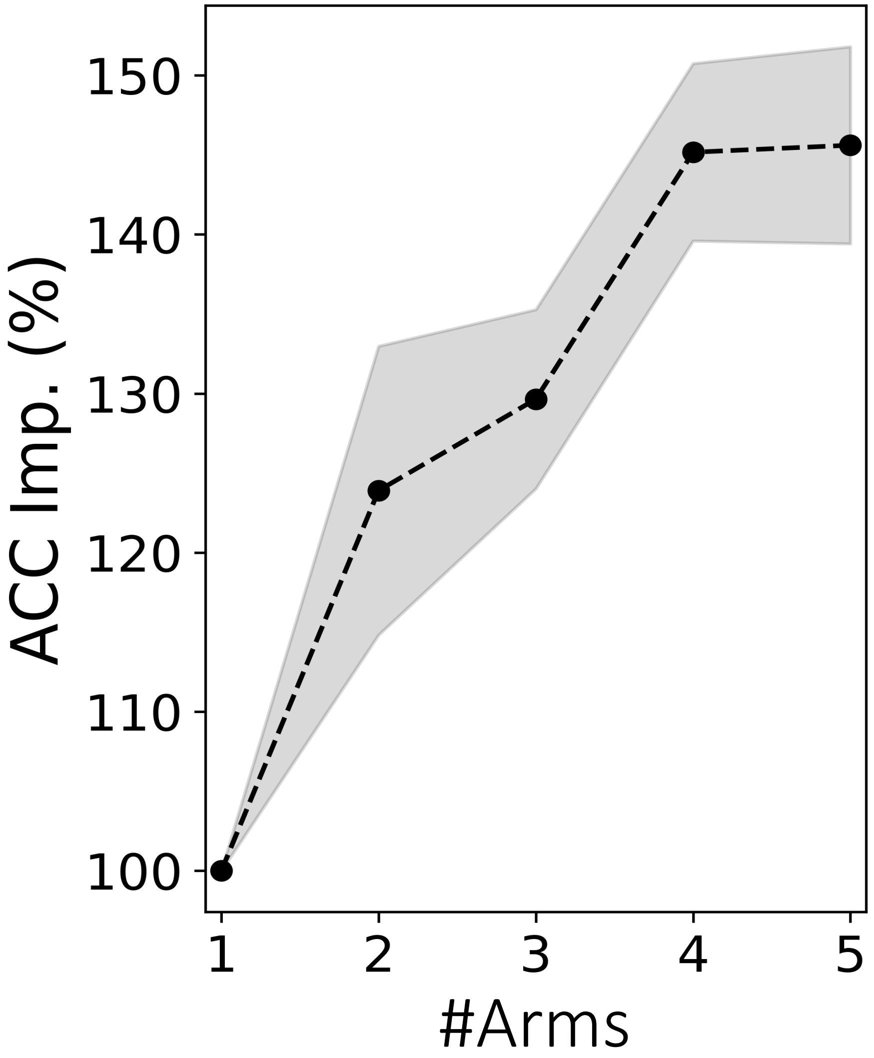


scRNA-seq dataset: transcriptomic identities

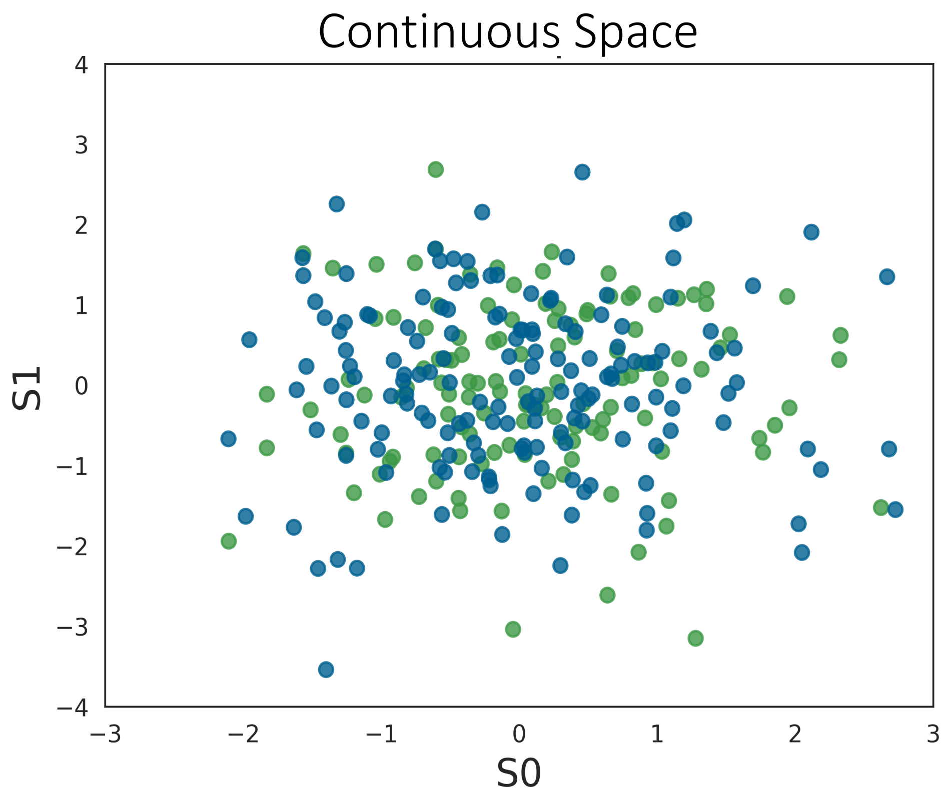


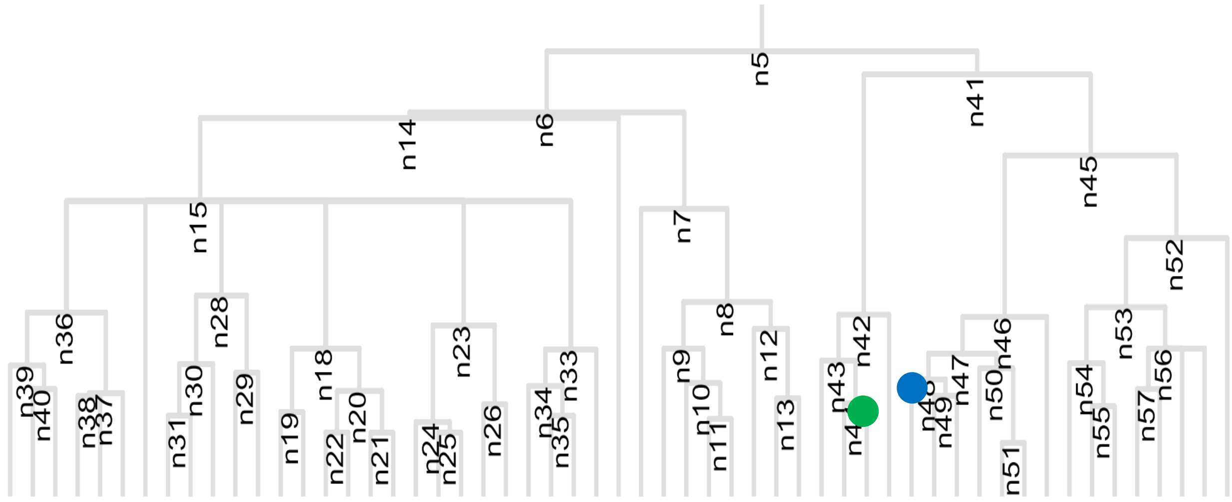
scRNA-seq dataset: more than 2 arms



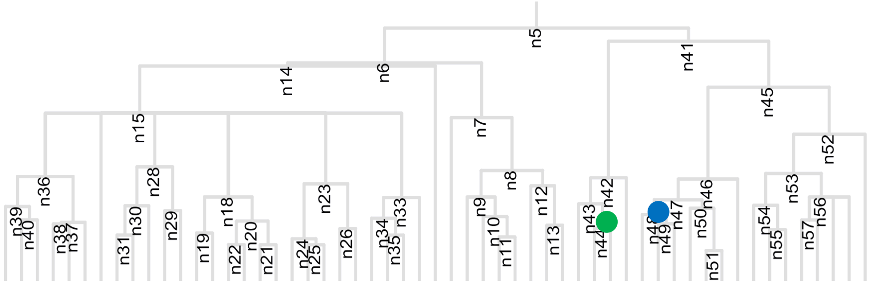


Identifying genes regulating continuous variability

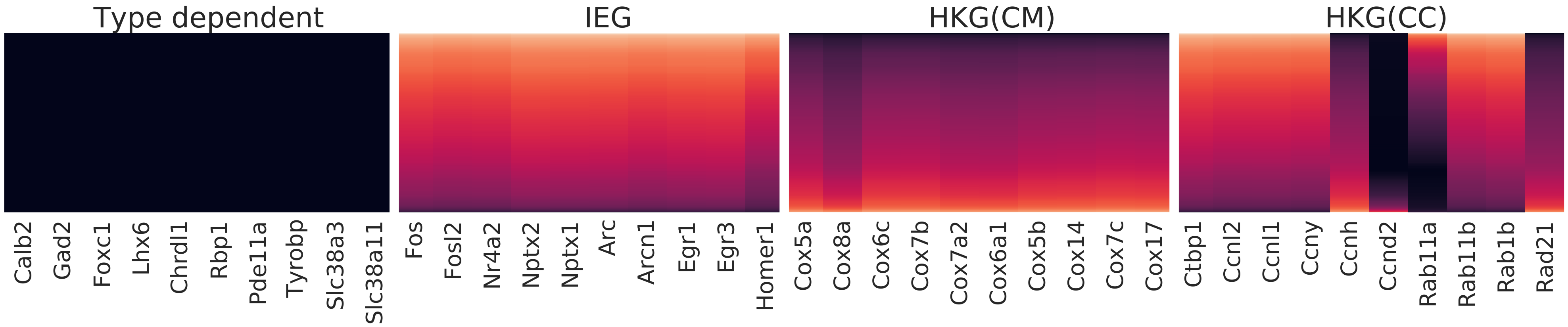




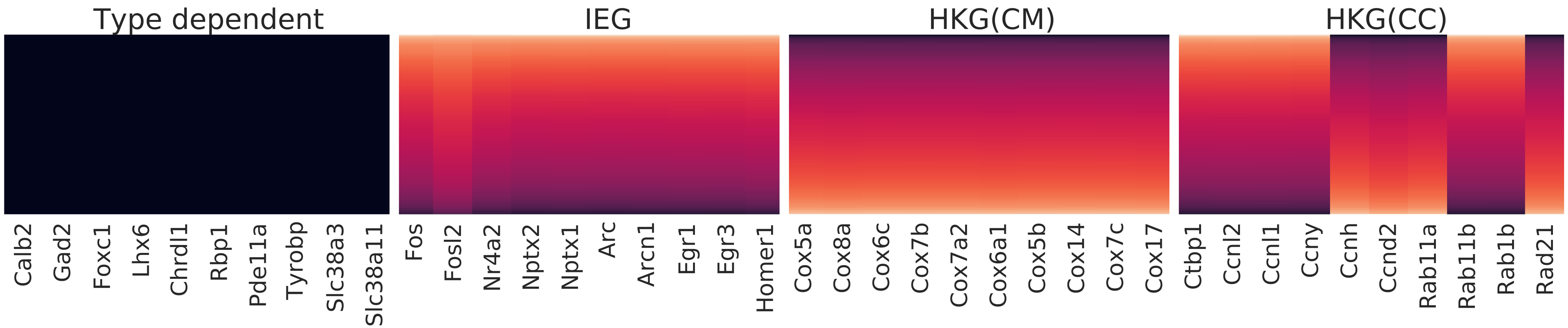
Identifying genes …



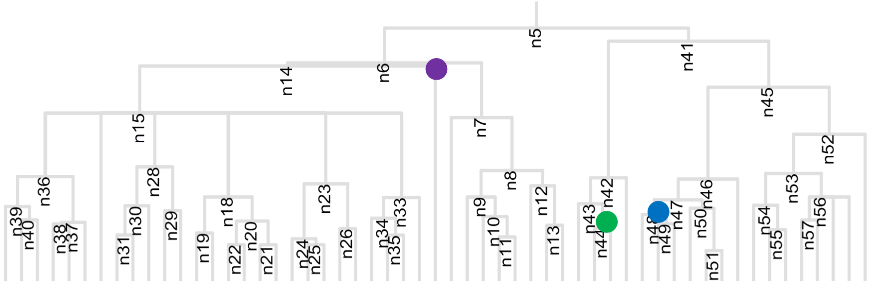
L5 NP ALM Trhr Nefl (n44)

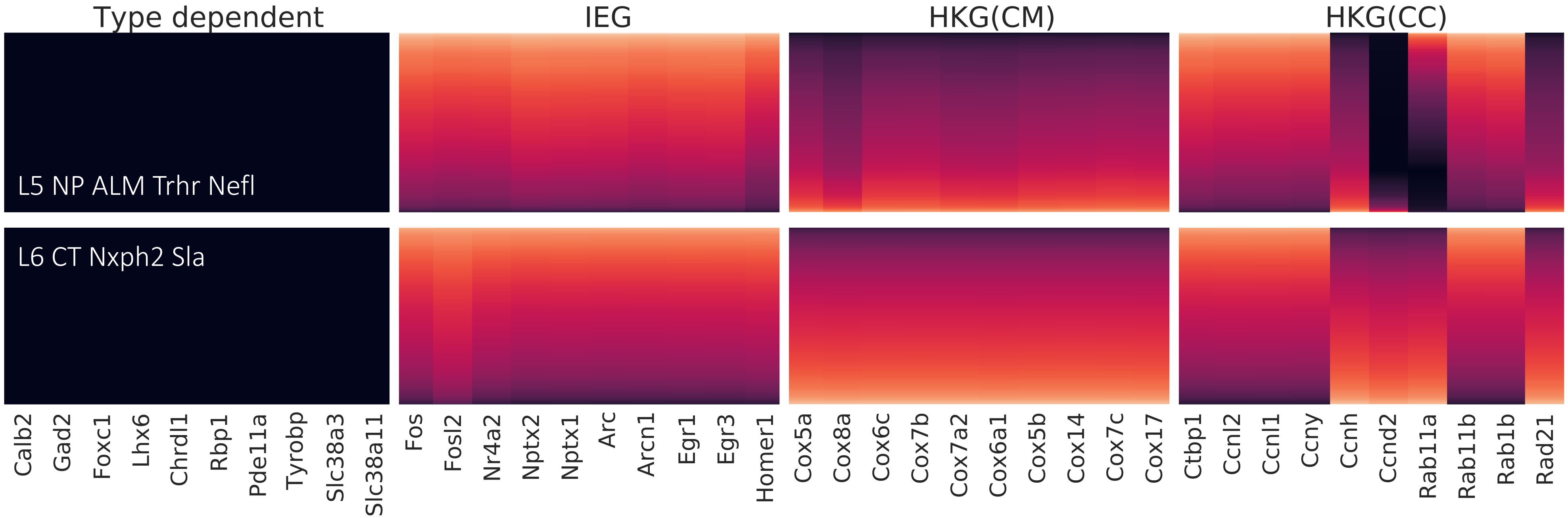


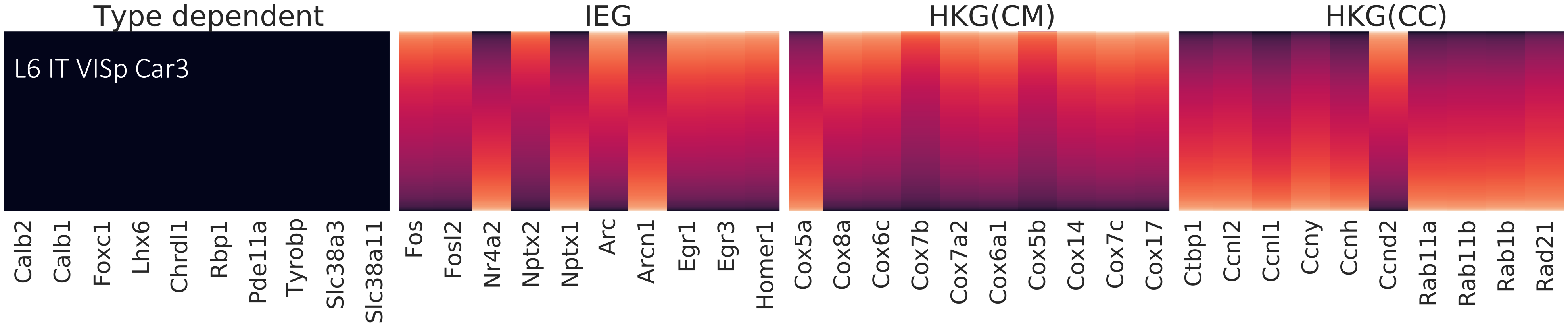
L6 CT Nxph2 Sla (n48)



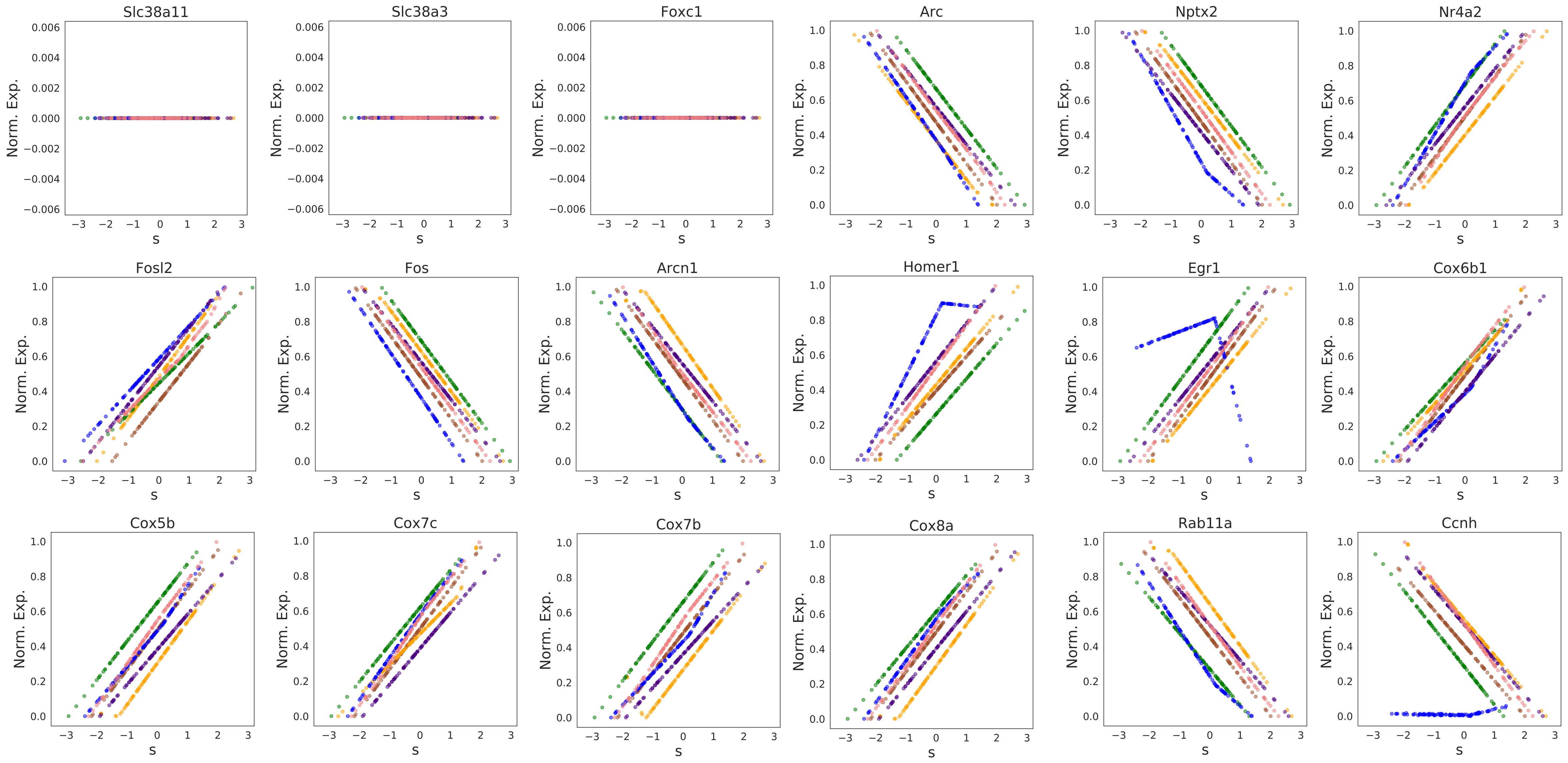
Identifying genes …





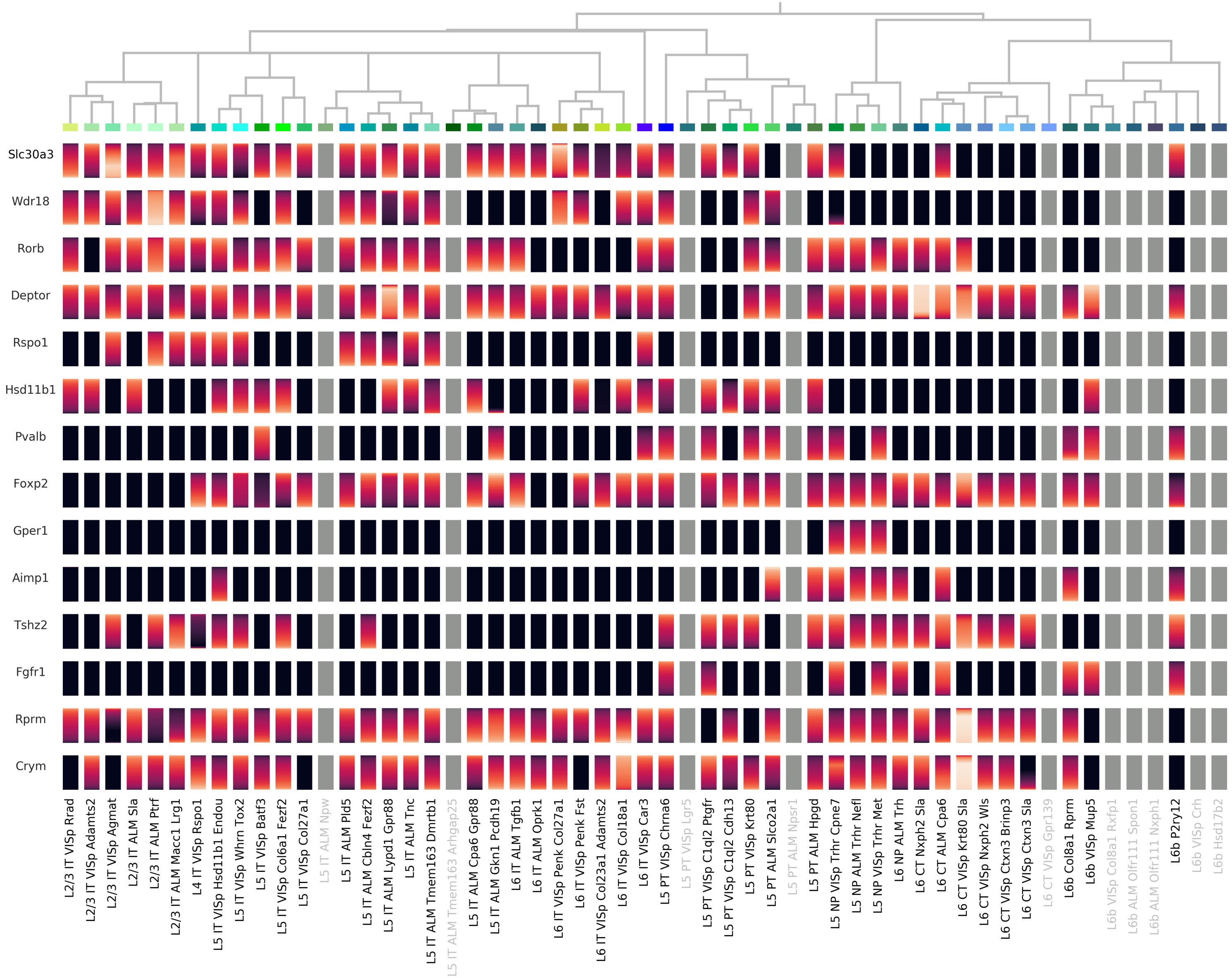


Robustness of type-dependent variabilities



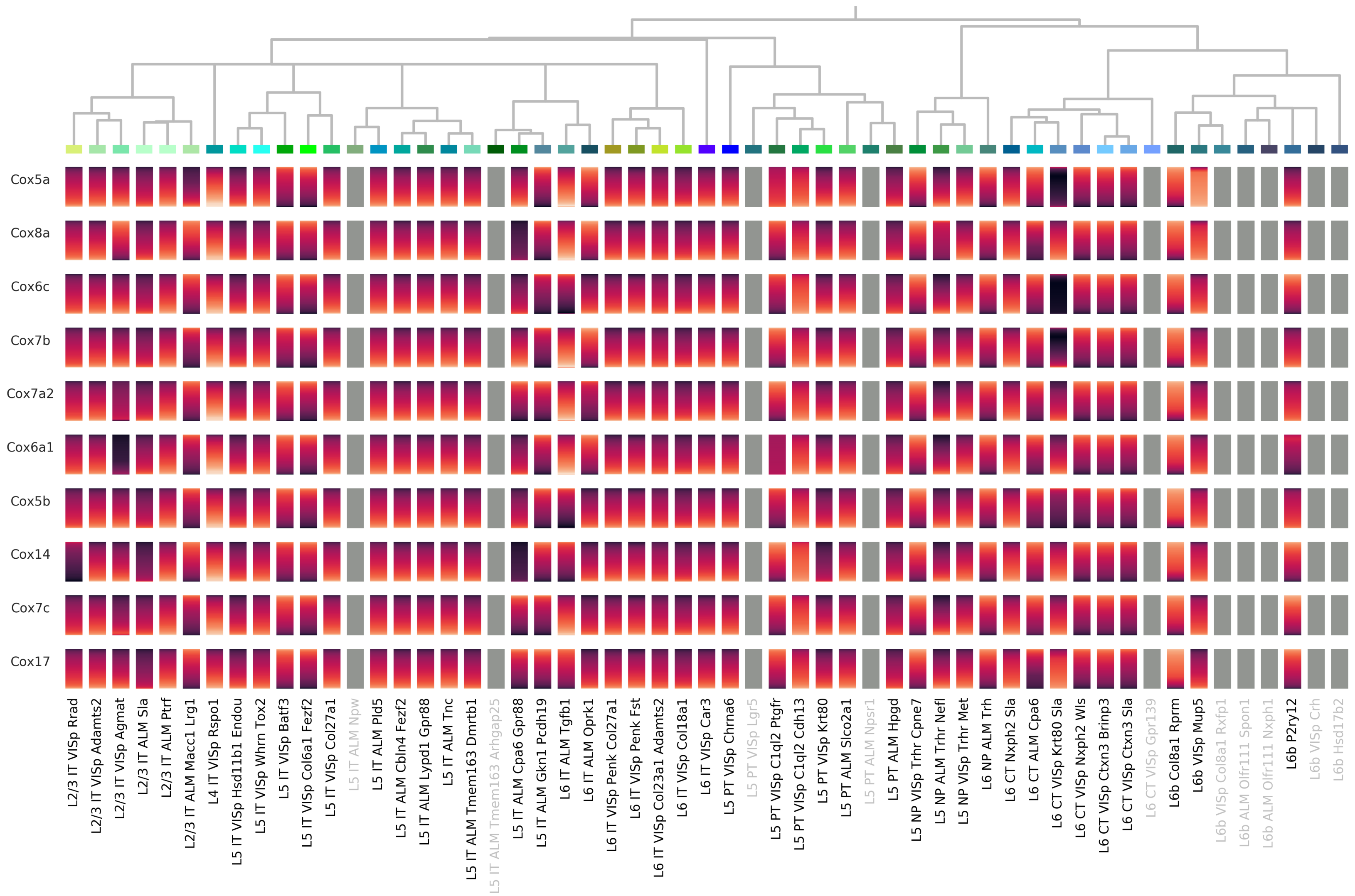
Glutamatergic cells

Marker genes



Glutamatergic cells

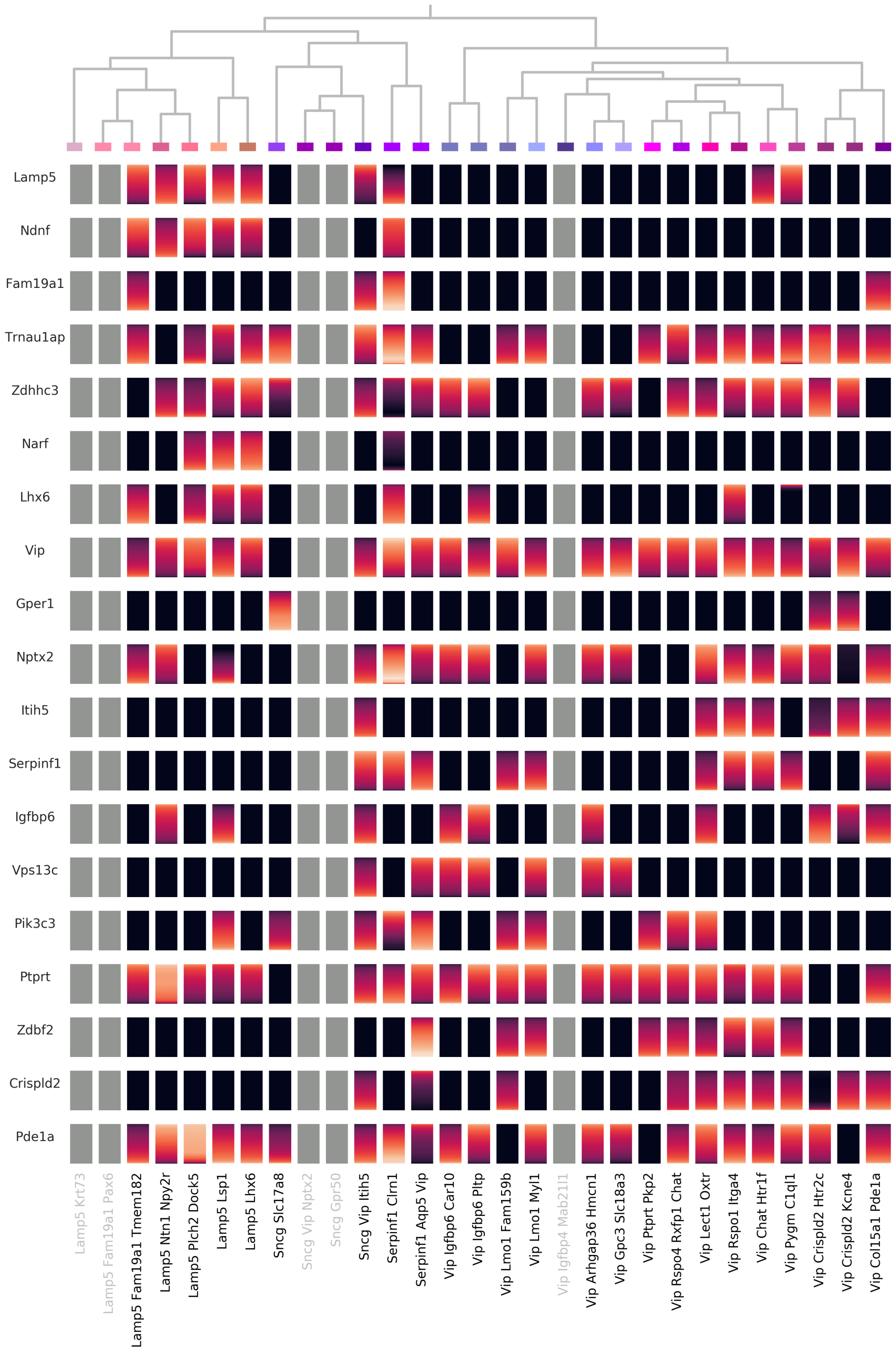
House-keeping genes



ALLEN INSTITUTE for BRAIN SCIENCE

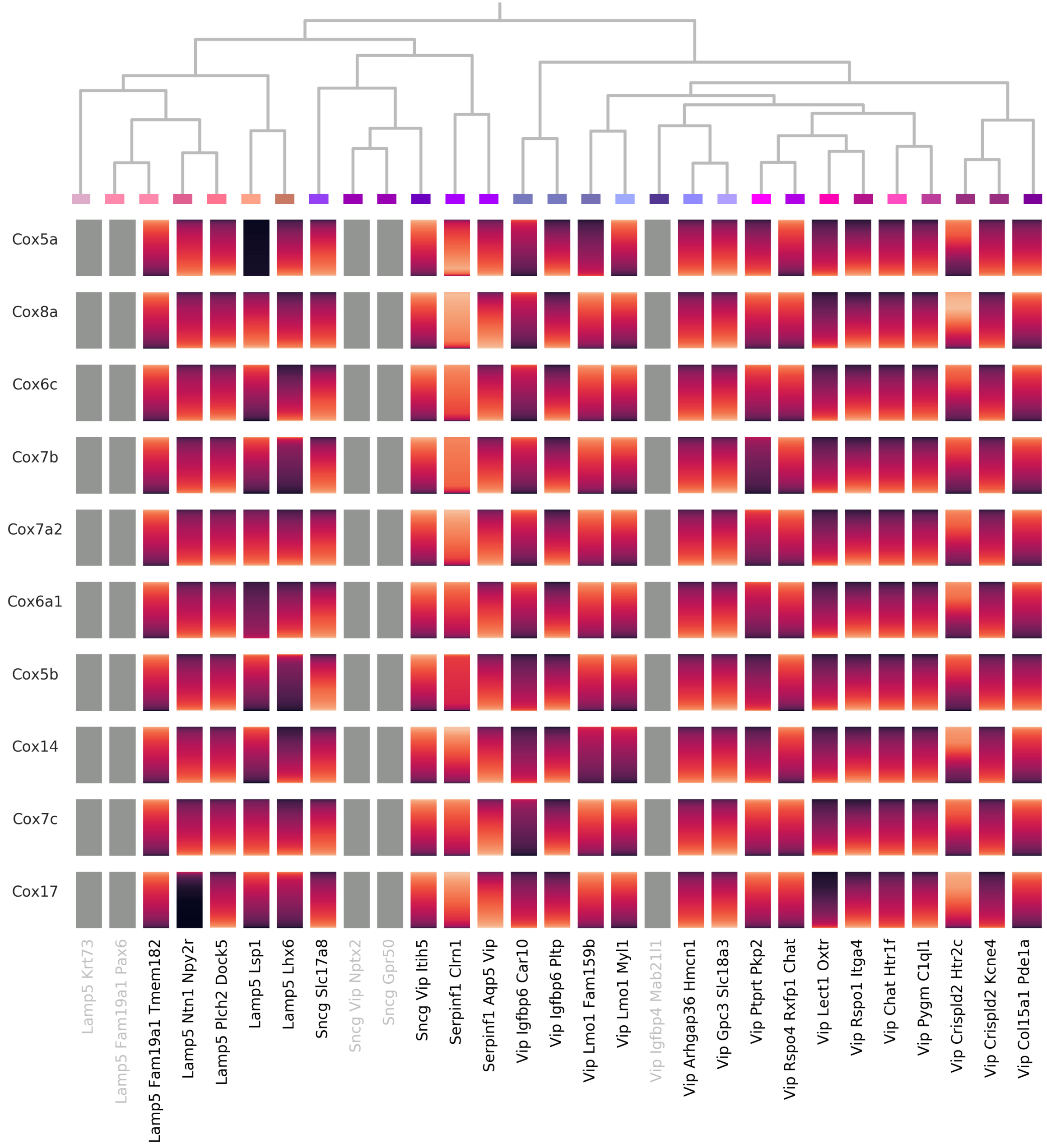
GABAergic cells

Marker genes



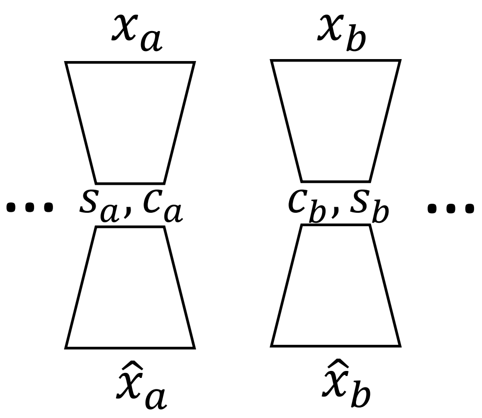
GABAergic cells

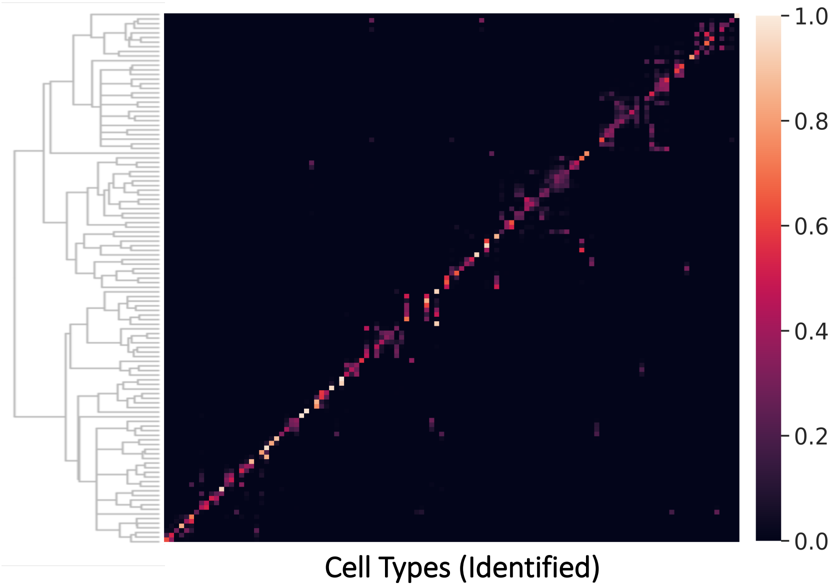
House-keeping genes

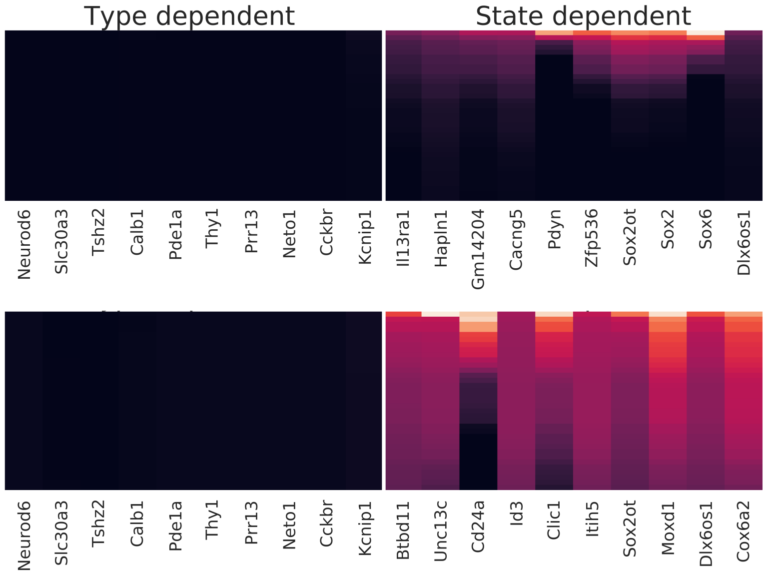


Summary

- Introducing cpl-mixVAE as a general framework to apply the power of collective decision making in unsupervised joint learning of discrete and continuous generative factors.
- Determining the neuronal cell types in an unsupervised setting, while identifying the genes implicated in regulating biologically relevant neuronal states.
- Studying (differential) gene expression variabilities using the type-dependent continuous factor.



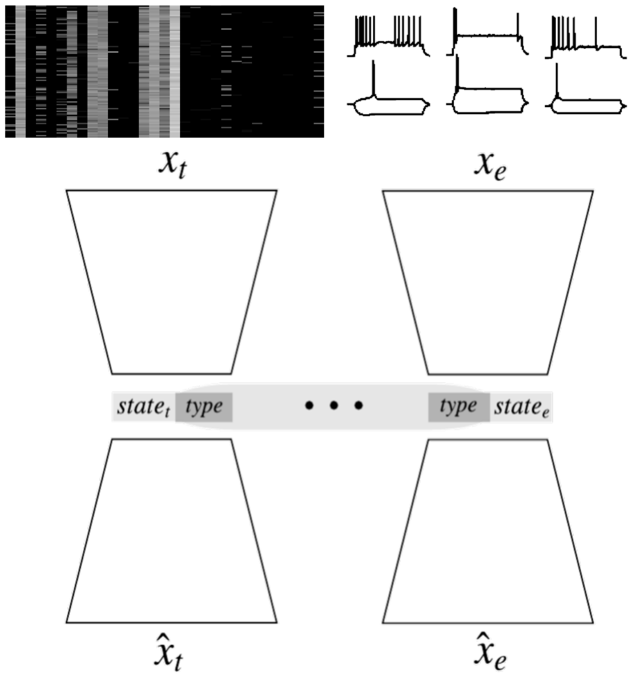




Future studies

- Multi-modal datasets (Joint identification of cell types and states in different modalities)

- Trajectory-based differential expression analysis for single-cell sequencing data



ALLEN INSTITUTE for BRAIN SCIENCE

THANK YOU

Team:

Uygar Sümbül

Rohan Gala

Olga Gliko

Fahimeh Baftizadeh

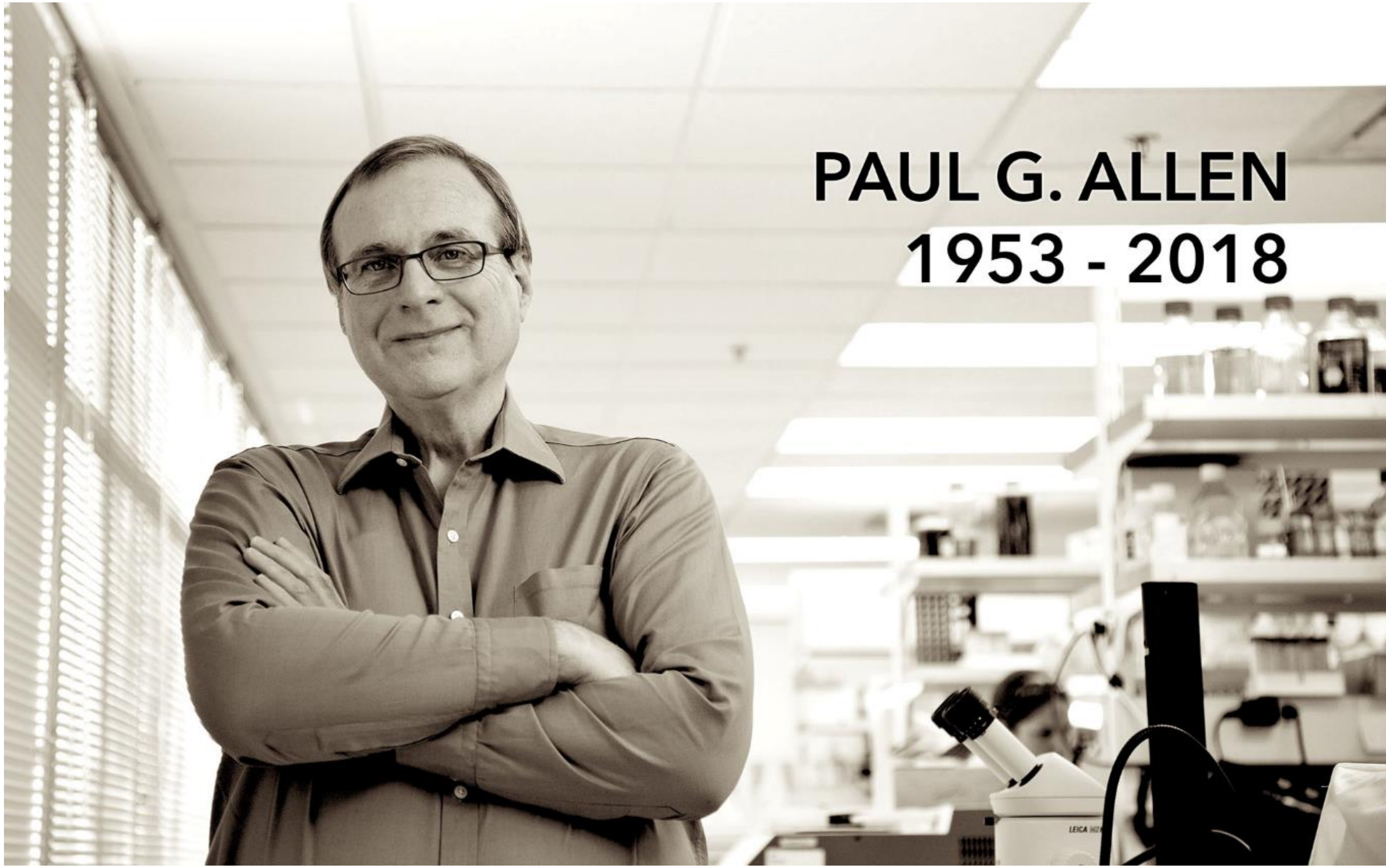
ALLEN INSTITUTE

ALLEN INSTITUTE for BRAIN SCIENCE

THANK YOU

We wish to thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement, and support.

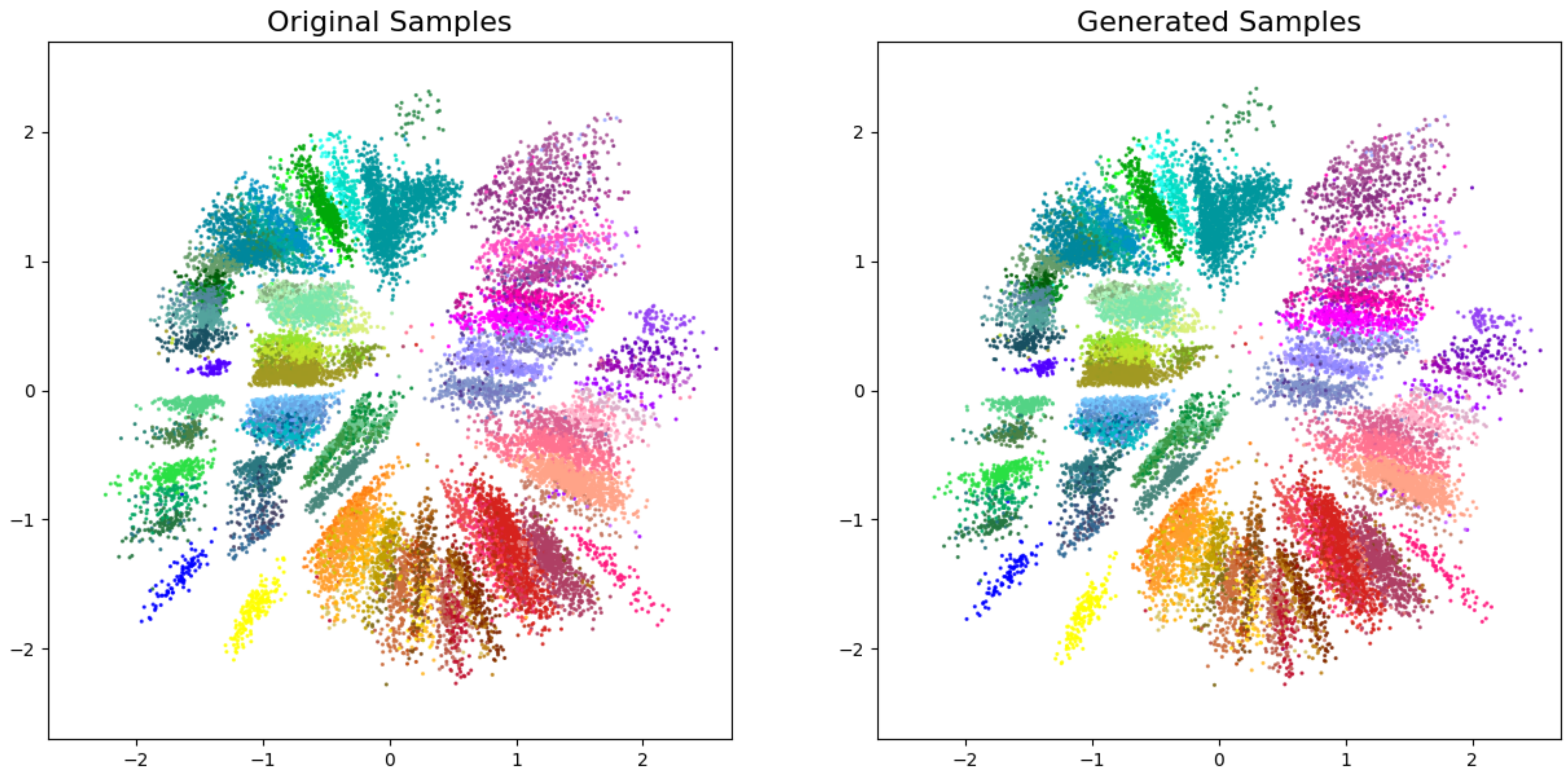
brain-map.org



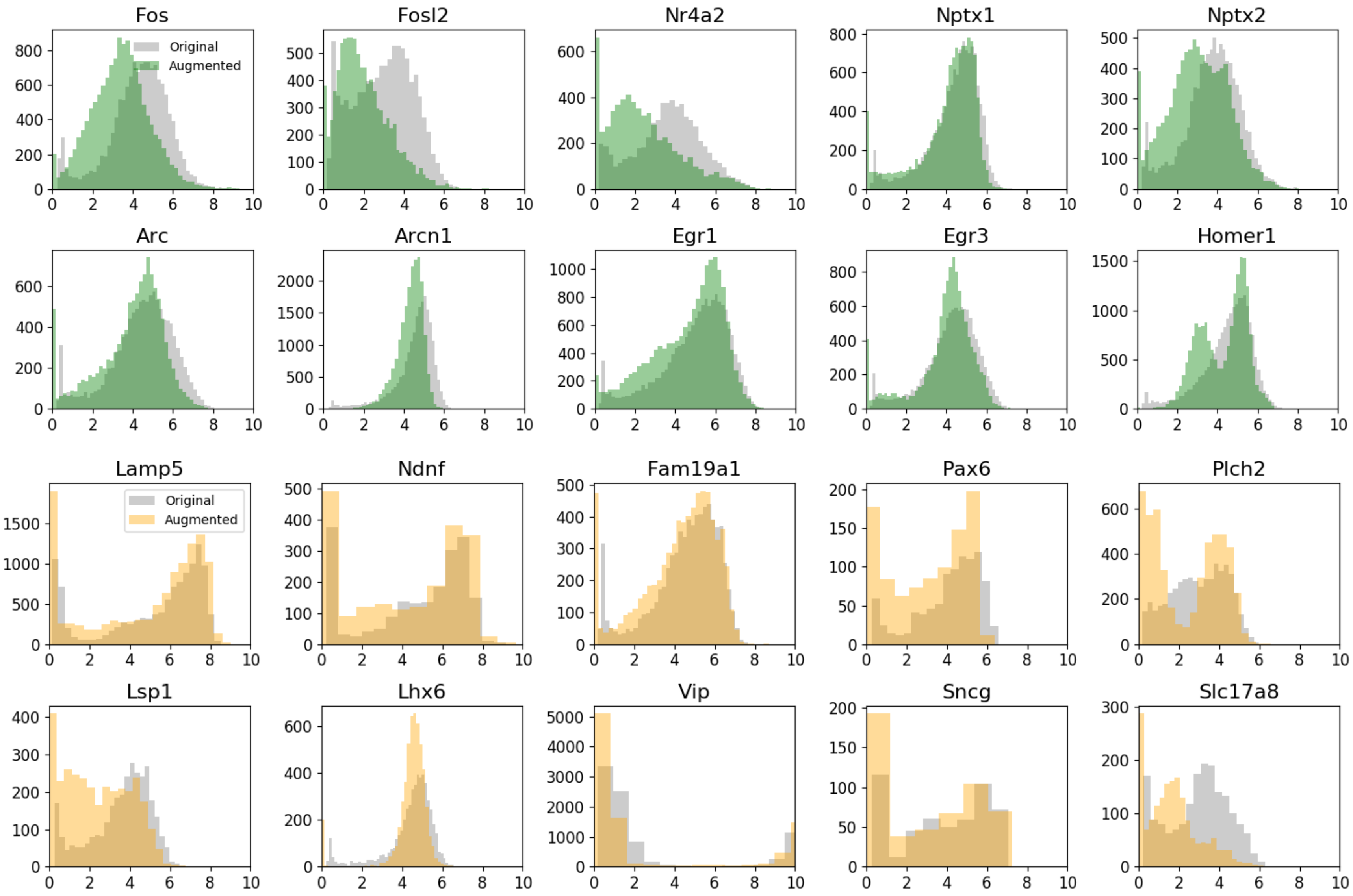
ALLEN INSTITUTE | ALLEN INSTITUTE for BRAIN SCIENCE

Supplement

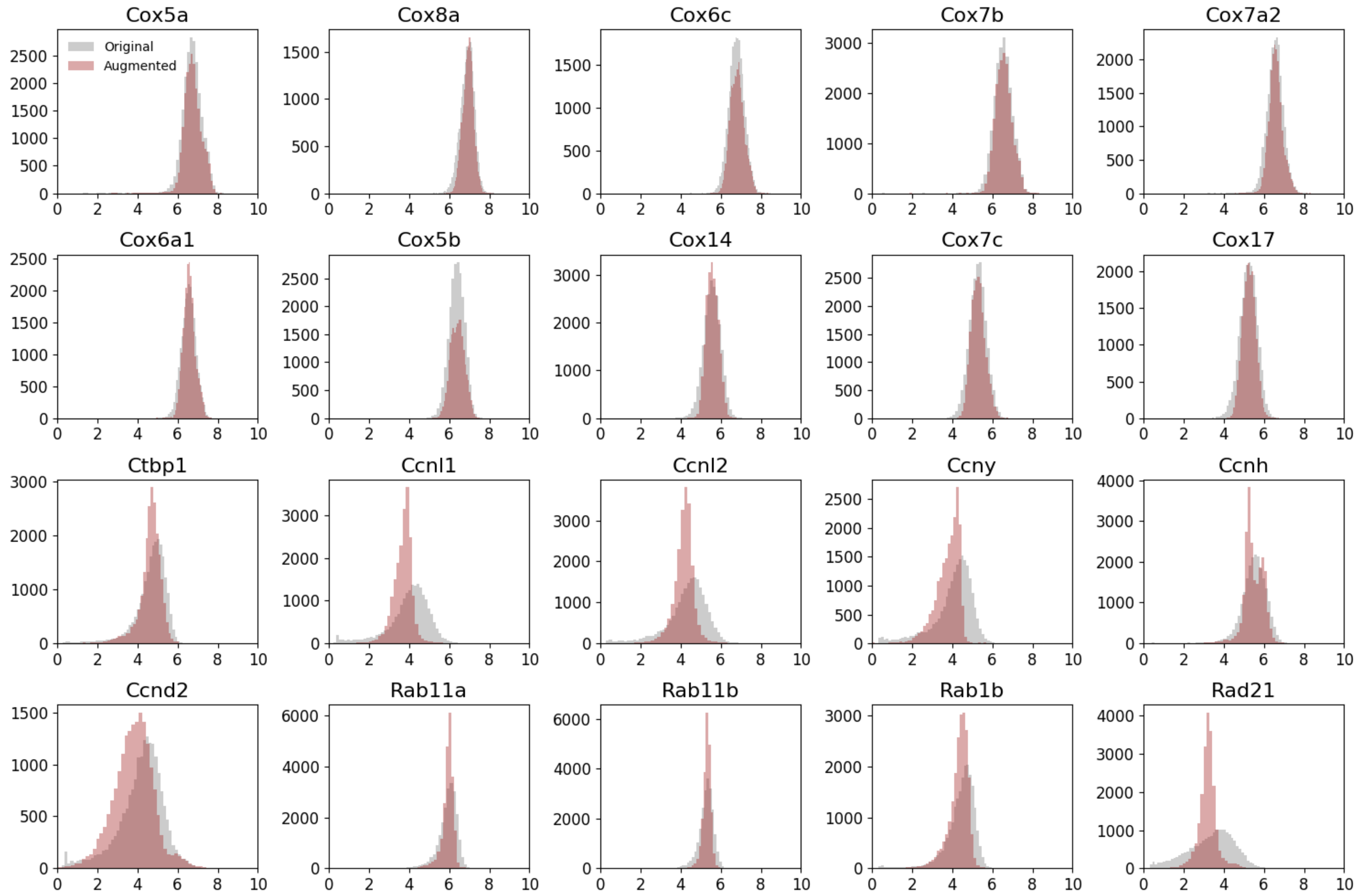
Single-cell generator



Single-cell generator



Single-cell generator



All datasets: overall performance

Dataset	Chance-level	$\|c\|$	$\|s\|$	Method	ACC (%) ↑ (mean ± s.d.)	Computation ↑ (iteration/sec)	Disentanglement score
MNIST	10.0%	10	2	InfoGAN	77.87 ± 21.68	12.2	–
			10	JointVAE	68.99 ± 11.76	74.1	
				CascadeVAE	81.41 ± 09.54	23.8	
				cpl-mixVAE	**84.56 ± 06.47**	17.5	
dSprite	33.3%	3	6	JointVAE	44.79 ± 03.88	52.6	74.51 ± 05.17
				CascadeVAE	78.84 ± 15.65	15.4	90.49 ± 05.28
				cpl-mixVAE	**96.30 ± 09.15**	20.6	89.98 ± 04.09
scRNA-seq	06.3%	115	2	JointVAE	12.53 ± 01.83	28.6	–
				CascadeVAE	02.69 ± 00.05	03.4	
				cpl-mixVAE	**38.78 ± 01.26**	10.1	

Consensus assignment