Effective Comparison of Single-Cell Embedding Visualizations

<u>Trevor Manz¹²</u>, Fritz Lekschas², Greg Finak², Evan Greene², Nils Gehlenborg¹

¹ Harvard Medical School | ² Ozette Technologies

Confusion Between Embedding Methods

Abundance Differences Between Samples



TRY IT YOURSELF: pip install cev

AND CHECK OUT THE EXAMPLES AT:







CD19 +	CD103 - CD28	- CD69) - PD1	-
HLADR -	GranzymeB -	CD25 -	ICOS -	TCRgd
CD38 -	CD127 - Tim3	-		

CD19 +	CD103 -	CD28 -	CD69	-	PD1	-	
HLADR -	Granzym	ieB - C	D25 -	ICO	S -	TCRgd	-
CD38 -	CD127 -	Tim3 -					

CD19 -	CD103 -	CD28 +	CD69 -	PD1 -		
HLADR - GranzymeB - CD25 - ICOS - TCF						
CD38 -	CD127 +	Tim3 -				

CD19 -	CD103 -	CD28 +	CD69	- PD1	-	
HLADR	Granzym	neB - CD	925 -	ICOS -	TCRgd	-
CD38 -	CD127 +	Tim3 -				

Data from Mair, F. et al., 2022. Nature.

WHAT? Single-cell embeddings transform cellular data of genetic, proteomic, and epigenomic features into lower-dimensional representations that aim to capture similarities among individual cells. Comparing these embeddings can expose patterns among cell clusters, providing insights into cell behavior, functions, and their role within broader biological contexts.

WHY? Comparing multiple embeddings is prevalent, but existing comparison methods are constrained to embeddings with point correspondences, which do not exist between experiments. Additionally, existing metrics struggle to reveal meaningful insights when comparing two embeddings of the same dataset.



CHALLENGES? Comparing multiple embeddings is a prevalent analysis task; however, existing pairwise comparison methods are constrained to embeddings with shared point correspondences, which do not exist between experiments. Existing metrics also struggle to reveal meaningful similarities and differences when comparing two embeddings of the same dataset.

CONCLUSIONS! We employ set-based similarity metrics on neighborhood graphs from single-cell embeddings, determining the similarity of shared labels at multiple levels between datasets. These metrics guide the visual exploration of such embeddings, enabling quick discovery of interesting phenotypes.

Label Confusion Set

METHODOLOGY

— Label Neighborhood Set



Create Delaunay graph of the embedded points
For each label: conduct *breadth-first search* for every point with that label

Construct Confusion Set

Points within one hop account to label confusion set
Remove points that are k standard deviations away

Construct Neighborhood Set

- Points with 1+ hop and not in the confusion set account for neighborhood set
- Scale neighborhood strength of each label by:

Average number of connections between all labels

Average distances of connections between all labels

