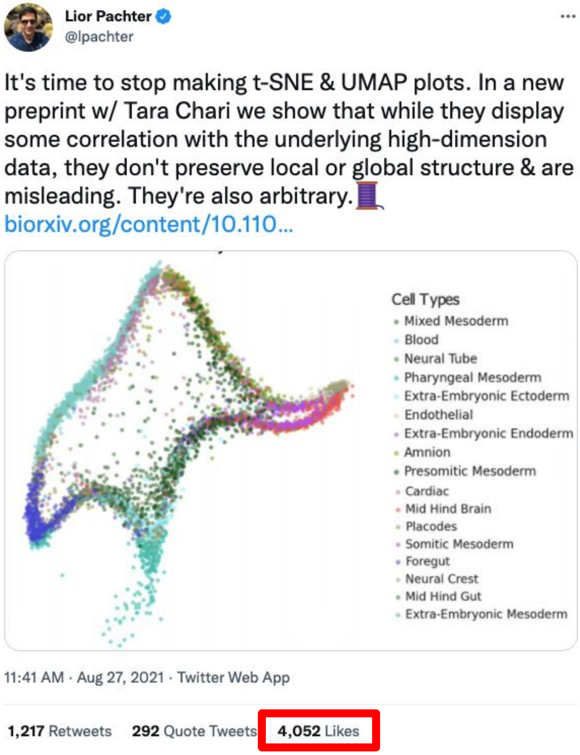
The Specious Art of Single-Cell Genomics

Tara Chari, Joeyta Banerjee, Lior Pachter

Presented by Anna Spiro & Pascal Sturmfels

Why this Paper? The Twitter arguments have been 🔥



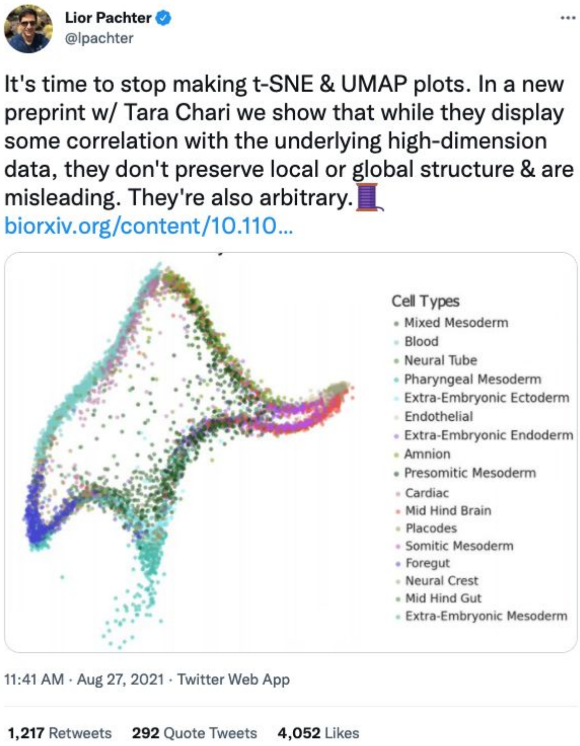
Lior Pachter
@lpachter

It's time to stop making t-SNE & UMAP plots. In a new preprint w/ Tara Chari we show that while they display some correlation with the underlying high-dimension data, they don't preserve local or global structure & are misleading. They're also arbitrary.🧵
biorxiv.org/content/10.110...

11:41 AM · Aug 27, 2021 · Twitter Web App

1,217 Retweets **292** Quote Tweets **4,052** Likes

Why this Paper? The Twitter arguments have been 🔥

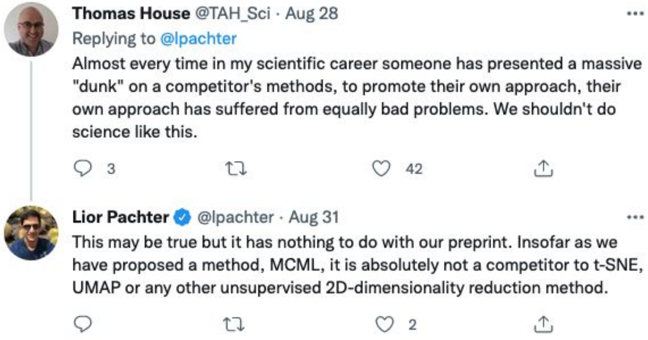


Lior Pachter @lpachter

It's time to stop making t-SNE & UMAP plots. In a new preprint w/ Tara Chari we show that while they display some correlation with the underlying high-dimension data, they don't preserve local or global structure & are misleading. They're also arbitrary.🧵
biorxiv.org/content/10.110...

11:41 AM · Aug 27, 2021 · Twitter Web App

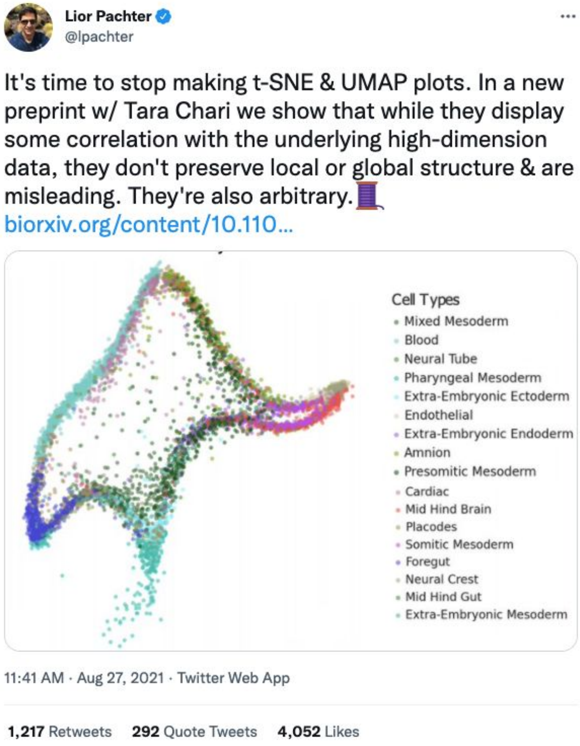
1,217 Retweets **292** Quote Tweets **4,052** Likes



Thomas House @TAH_Sci · Aug 28
Replying to @lpachter
Almost every time in my scientific career someone has presented a massive "dunk" on a competitor's methods, to promote their own approach, their own approach has suffered from equally bad problems. We shouldn't do science like this.

Lior Pachter @lpachter · Aug 31
This may be true but it has nothing to do with our preprint. Insofar as we have proposed a method, MCML, it is absolutely not a competitor to t-SNE, UMAP or any other unsupervised 2D-dimensionality reduction method.

Why this Paper? The Twitter arguments have been 🔥

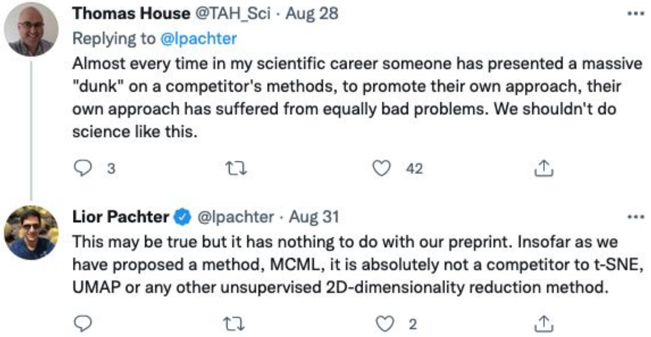


Lior Pachter
@lpachter

It's time to stop making t-SNE & UMAP plots. In a new preprint w/ Tara Chari we show that while they display some correlation with the underlying high-dimension data, they don't preserve local or global structure & are misleading. They're also arbitrary.🧵
biorxiv.org/content/10.110...

11:41 AM · Aug 27, 2021 · Twitter Web App

1,217 Retweets **292** Quote Tweets **4,052** Likes

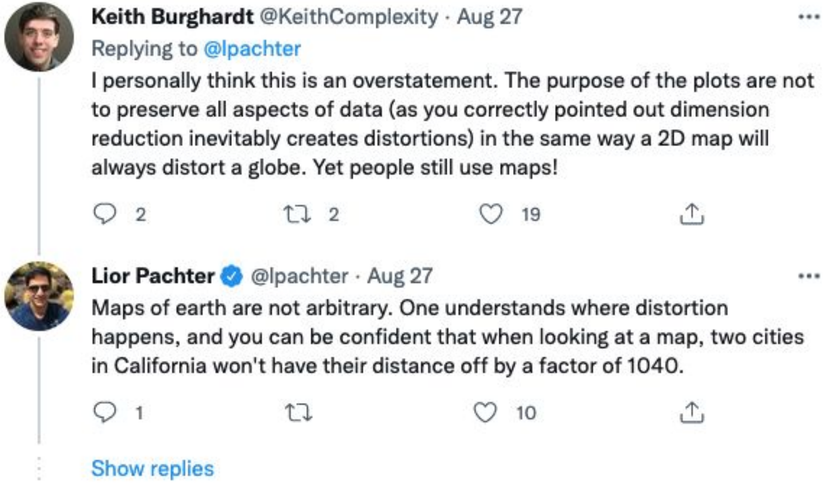


Thomas House @TAH_Sci · Aug 28
Replying to @lpachter
Almost every time in my scientific career someone has presented a massive "dunk" on a competitor's methods, to promote their own approach, their own approach has suffered from equally bad problems. We shouldn't do science like this.

3 42

Lior Pachter @lpachter · Aug 31
This may be true but it has nothing to do with our preprint. Insofar as we have proposed a method, MCML, it is absolutely not a competitor to t-SNE, UMAP or any other unsupervised 2D-dimensionality reduction method.

2



Keith Burghardt @KeithComplexity · Aug 27
Replying to @lpachter
I personally think this is an overstatement. The purpose of the plots are not to preserve all aspects of data (as you correctly pointed out dimension reduction inevitably creates distortions) in the same way a 2D map will always distort a globe. Yet people still use maps!

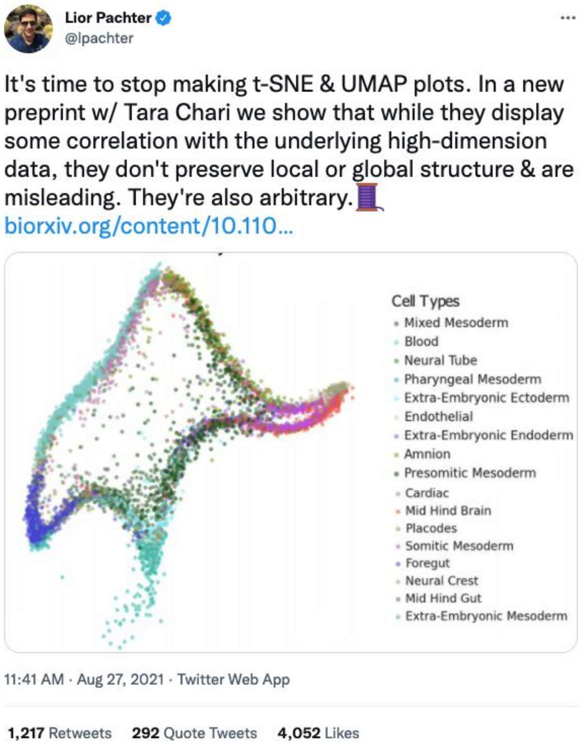
2 2 19

Lior Pachter @lpachter · Aug 27
Maps of earth are not arbitrary. One understands where distortion happens, and you can be confident that when looking at a map, two cities in California won't have their distance off by a factor of 1040.

1 10

Show replies

Why this Paper? The Twitter ... have been 🔥

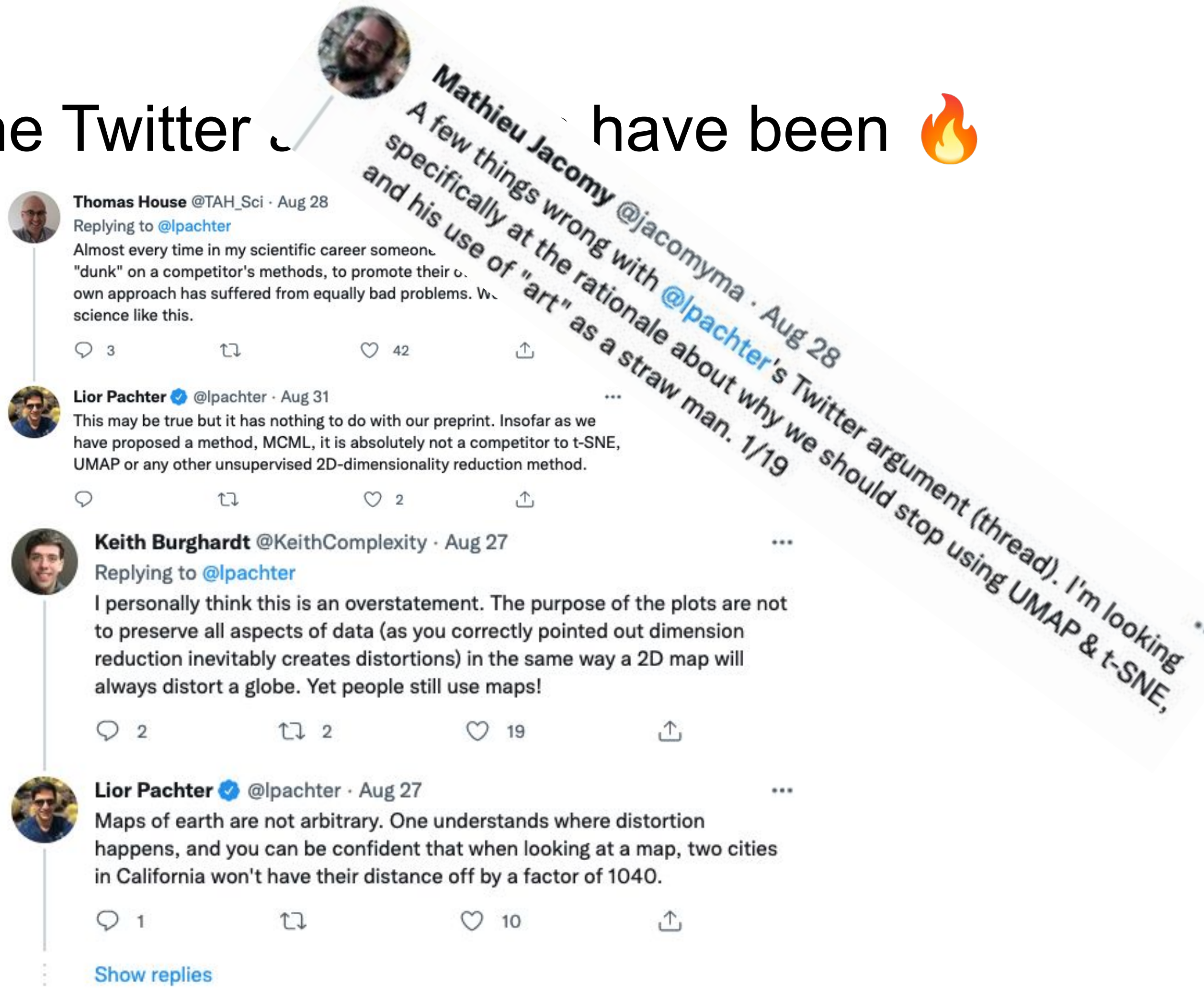


Lior Pachter @lpachter

It's time to stop making t-SNE & UMAP plots. In a new preprint w/ Tara Chari we show that while they display some correlation with the underlying high-dimension data, they don't preserve local or global structure & are misleading. They're also arbitrary.🧵 biorxiv.org/content/10.110...

11:41 AM · Aug 27, 2021 · Twitter Web App

1,217 Retweets **292** Quote Tweets **4,052** Likes



Thomas House @TAH_Sci · Aug 28
Replying to @lpachter
Almost every time in my scientific career someone "dunk" on a competitor's methods, to promote their o own approach has suffered from equally bad problems. W science like this.

3 42

Lior Pachter @lpachter · Aug 31
This may be true but it has nothing to do with our preprint. Insofar as we have proposed a method, MCML, it is absolutely not a competitor to t-SNE, UMAP or any other unsupervised 2D-dimensionality reduction method.

2

Keith Burghardt @KeithComplexity · Aug 27
Replying to @lpachter
I personally think this is an overstatement. The purpose of the plots are not to preserve all aspects of data (as you correctly pointed out dimension reduction inevitably creates distortions) in the same way a 2D map will always distort a globe. Yet people still use maps!

2 2 19

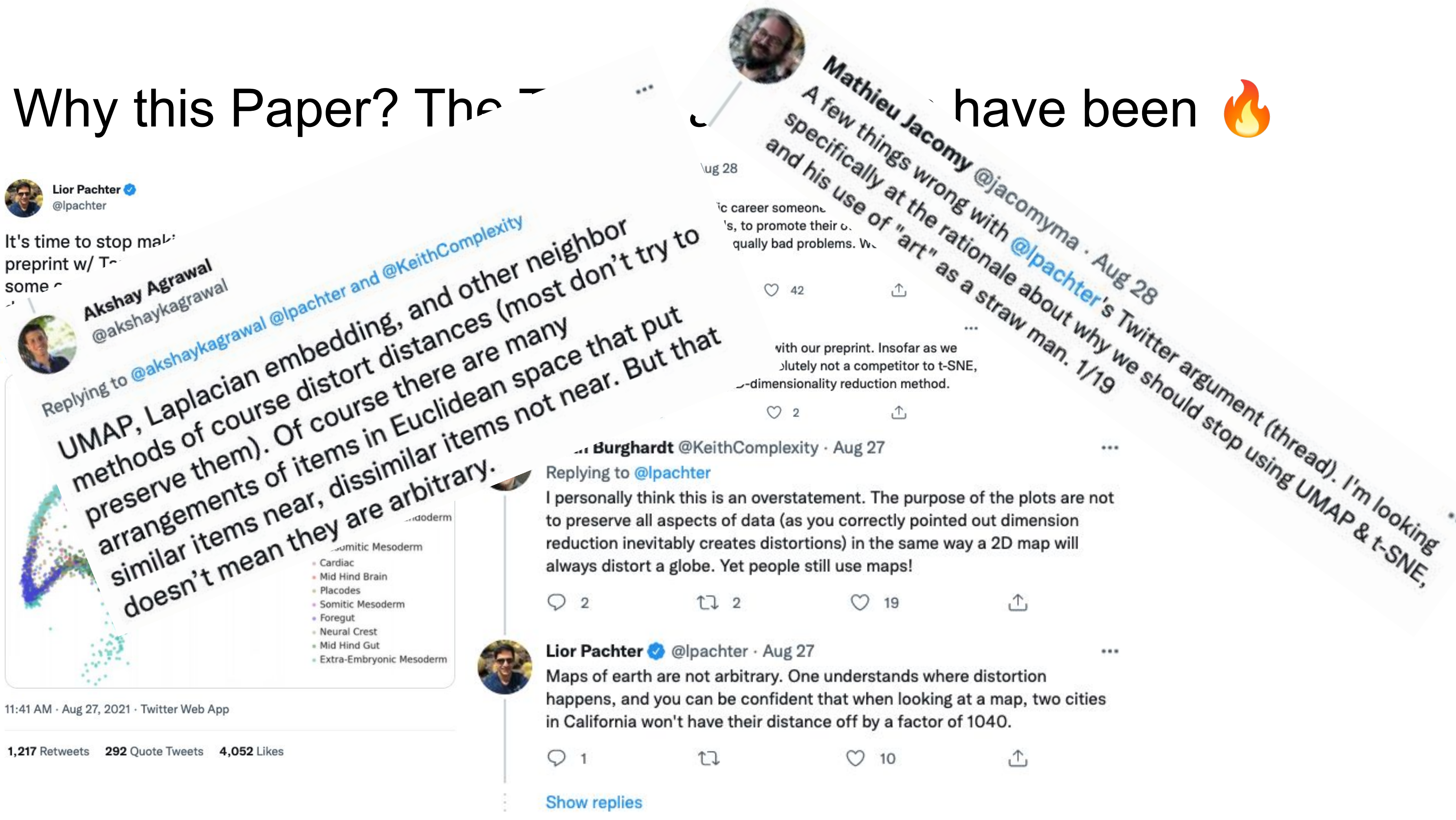
Lior Pachter @lpachter · Aug 27
Maps of earth are not arbitrary. One understands where distortion happens, and you can be confident that when looking at a map, two cities in California won't have their distance off by a factor of 1040.

1 10

Show replies

Mathieu Jacomy @jacomyma · Aug 28
A few things wrong with @lpachter's Twitter argument (thread). I'm looking specifically at the rationale about why we should stop using UMAP & t-SNE, and his use of "art" as a straw man. 1/19

Why this Paper? The have been 🔥



Lior Pachter @lpachter

It's time to stop mak
preprint w/ T
some

Akshay Agrawal @akshaykagrawal

Replying to @akshaykagrawal @lpachter and @KeithComplexity

UMAP, Laplacian embedding, and other neighbor methods of course distort distances (most don't try to preserve them). Of course there are many arrangements of items in Euclidean space that put similar items near, dissimilar items not near. But that doesn't mean they are arbitrary.

11:41 AM · Aug 27, 2021 · Twitter Web App

1,217 Retweets 292 Quote Tweets 4,052 Likes

ug 28

ic career someone
's, to promote their o
qually bad problems. W

♡ 42

with our preprint. Insofar as we
olutely not a competitor to t-SNE,
-dimensionality reduction method.

♡ 2

Burghardt @KeithComplexity · Aug 27

Replying to @lpachter

I personally think this is an overstatement. The purpose of the plots are not to preserve all aspects of data (as you correctly pointed out dimension reduction inevitably creates distortions) in the same way a 2D map will always distort a globe. Yet people still use maps!

2 2 ♡ 19

Lior Pachter @lpachter · Aug 27

Maps of earth are not arbitrary. One understands where distortion happens, and you can be confident that when looking at a map, two cities in California won't have their distance off by a factor of 1040.

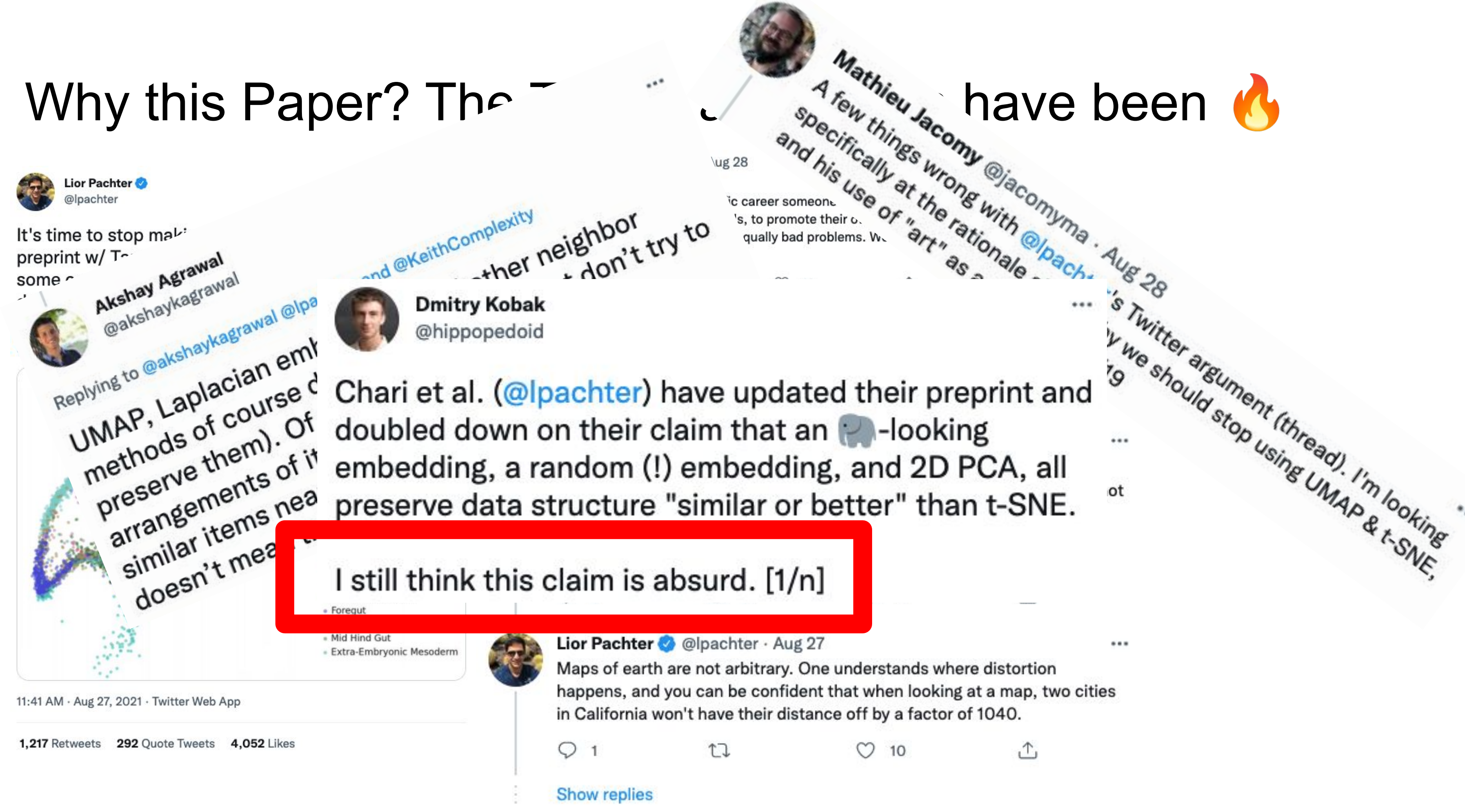
1 ♡ 10

Show replies

Mathieu Jacomy @jacomyma · Aug 28

A few things wrong with @lpachter's Twitter argument (thread). I'm looking specifically at the rationale about why we should stop using UMAP & t-SNE and his use of "art" as a straw man. 1/19

Why this Paper? The … have been 🔥



Lior Pachter @lpachter

It's time to stop mak…
preprint w/ T…
some …

Foregut
Mid Hind Gut
Extra-Embryonic Mesoderm

11:41 AM · Aug 27, 2021 · Twitter Web App

1,217 Retweets 292 Quote Tweets 4,052 Likes

Akshay Agrawal @akshaykagrawal

Replying to @akshaykagrawal @lpa… …nd @KeithComplexity

UMAP, Laplacian emb… …ther neighbor
methods of course d… …don't try to
preserve them). Of…
arrangements of it…
similar items nea…
doesn't mea…

…ug 28

…ic career someone…
…'s, to promote their o…
…qually bad problems. W…

Mathieu Jacomy @jacomyma · Aug 28

A few things wrong with @lpach… …'s Twitter argument (thread). I'm looking
specifically at the rationale … …y we should stop using UMAP & t-SNE,
and his use of "art" as a… …19

…ot

Dmitry Kobak @hippopedoid

Chari et al. (@lpachter) have updated their preprint and doubled down on their claim that an 🐘-looking embedding, a random (!) embedding, and 2D PCA, all preserve data structure "similar or better" than t-SNE.

I still think this claim is absurd. [1/n]

Lior Pachter @lpachter · Aug 27

Maps of earth are not arbitrary. One understands where distortion happens, and you can be confident that when looking at a map, two cities in California won't have their distance off by a factor of 1040.

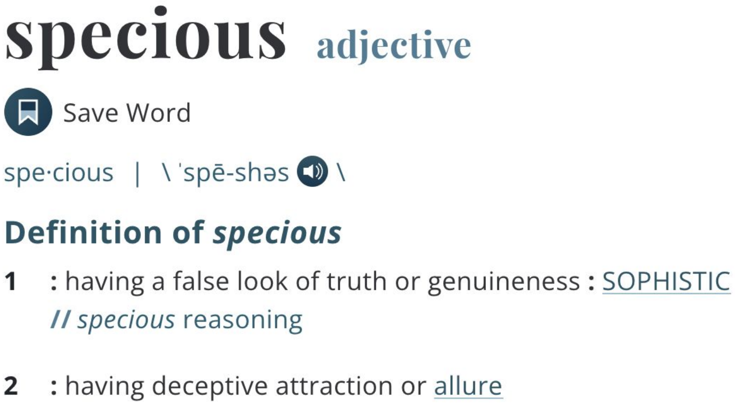
1 10

Show replies

Overview

In this paper, the authors:

- Discuss and quantify the distortions introduced in common dimensionality reduction practices
- Propose their own semi-supervised dimensionality reduction technique



specious adjective

Save Word

spe·cious | \ ˈspē-shəs \

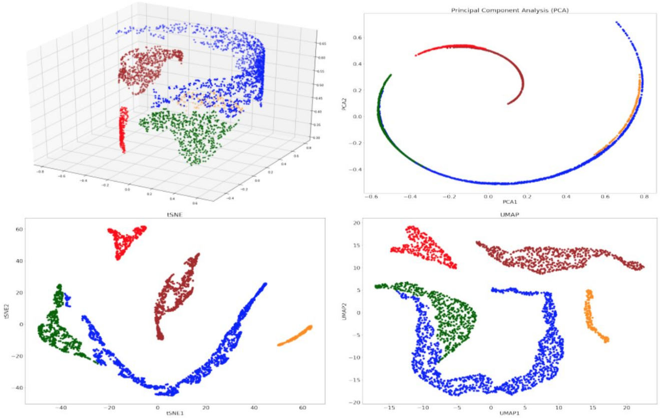
Definition of *specious*

1 : having a false look of truth or genuineness : SOPHISTIC
// *specious* reasoning

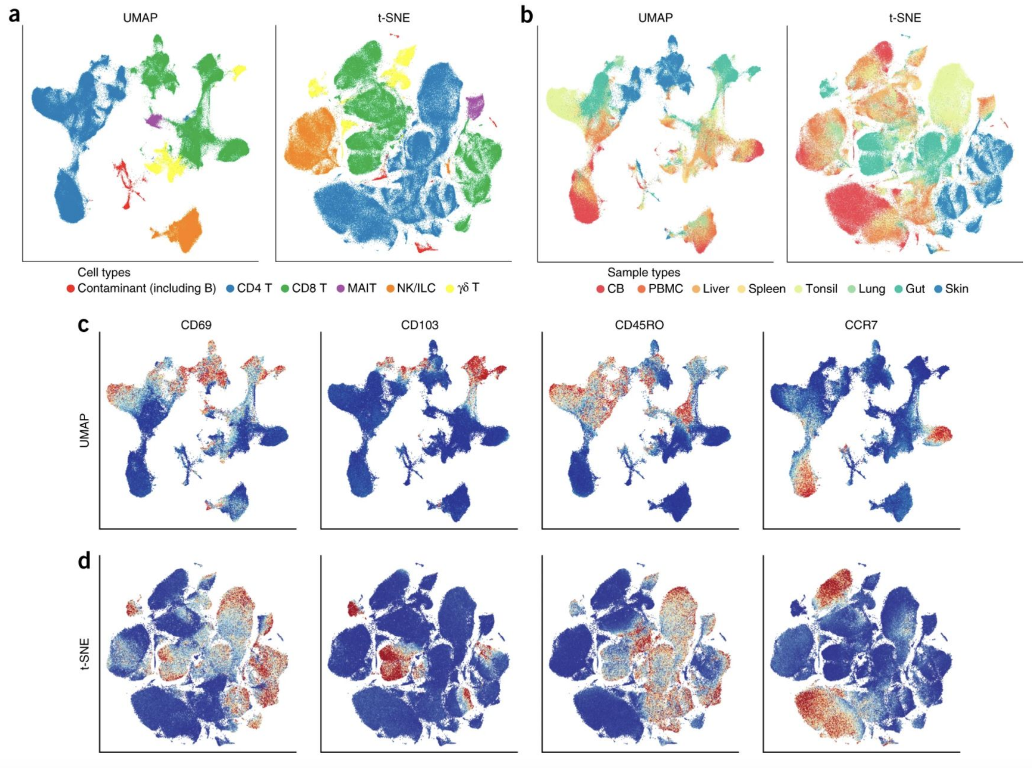
2 : having deceptive attraction or allure

Common dimensionality reduction practices

PCA, UMAP, t-SNE



t-SNE vs. UMAP: Global Structure, Oskolkov

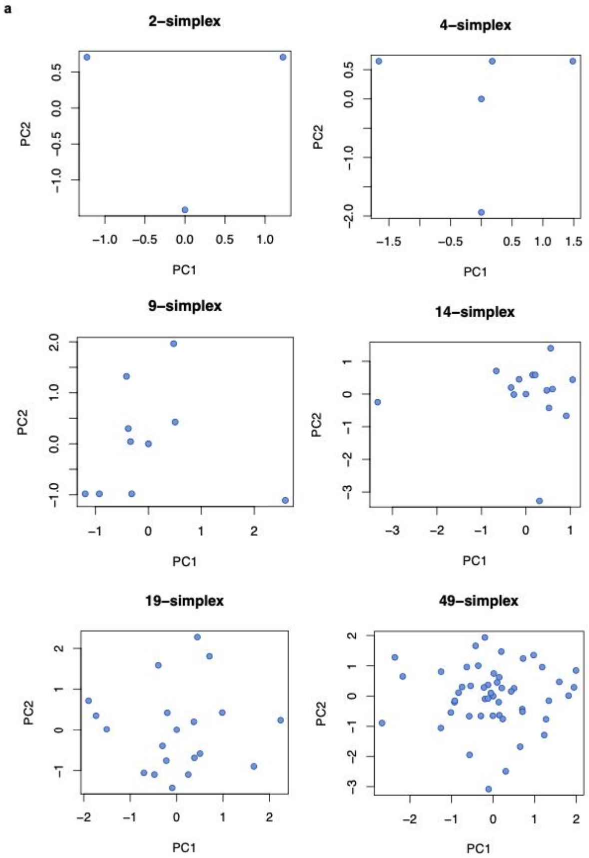


Becht et al. *Nature Biotechnology*

A difficult case of dimensionality reduction: the embedding of equidistant points

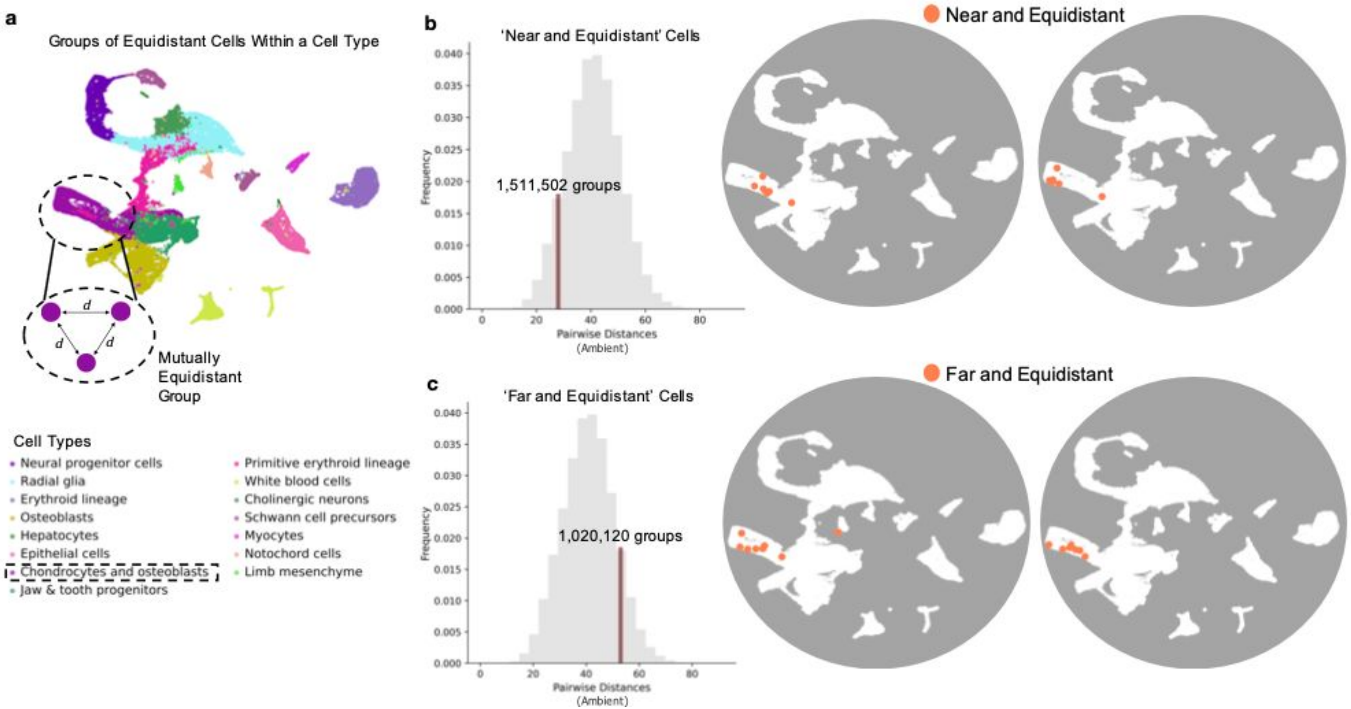
Math background:

- It is impossible to embed greater than
- n + 1 equidistant points in R^k for k ≤ n
- The ratio of the max distance D to the min distance d among n points in 2D grows as $O(\sqrt{n})$
- PCA projection of equidistant points is essentially a random projection (*Supplementary Figure 1*)

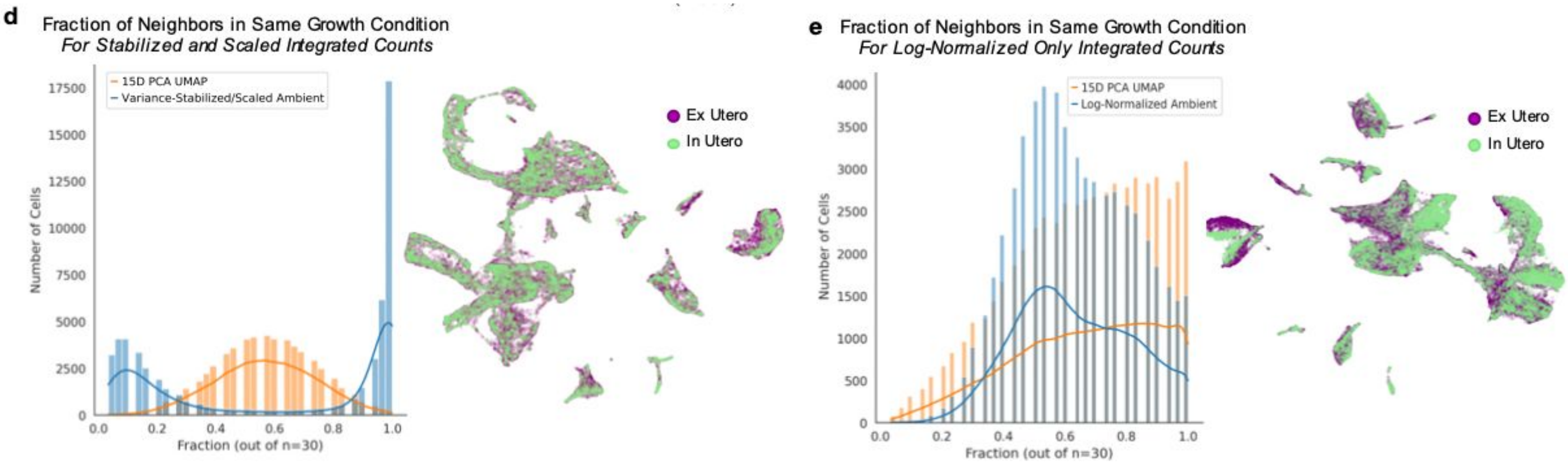


Near-equidistant points in biological data

Ex and in-utero mice embryo dataset (*Figure 1*)



Distortion of nearest neighbors



d Fraction of Neighbors in Same Growth Condition
For Stabilized and Scaled Integrated Counts

e Fraction of Neighbors in Same Growth Condition
For Log-Normalized Only Integrated Counts

Figure 1

Picasso algorithm

- To make a visual point about the distortion present in 2D embeddings, the authors present **Picasso**: an autoencoder whose loss function penalizes *distance between a user defined shape* as well as reconstruction error

Picasso algorithm

S represents the coordinates defining the desired shape, d = 2

D is an n x p pairwise distance matrix representing Euclidean distances between cell coordinates in latent space **Z** and shape coordinates **S** s.t.

$$d_{ij} = \|z_i - s_j\|_2$$

A is an n x p Boolean adjacency matrix that specifies an adjacent coordinate point for every cell (mapping n cells to p coordinates), A is determined by solving

$$min \sum_i \sum_j d_{ij} a_{ij}$$

a_{ij} = 1 IFF row i is assigned to column j

Picasso algorithm

$$L_{ShapeAware} = \sum A \odot D,$$

$$L = f * L_{ShapeAware} + (1 - f) * L_{Reconstruction}.$$

Picasso embeddings of biological data into arbitrary shapes

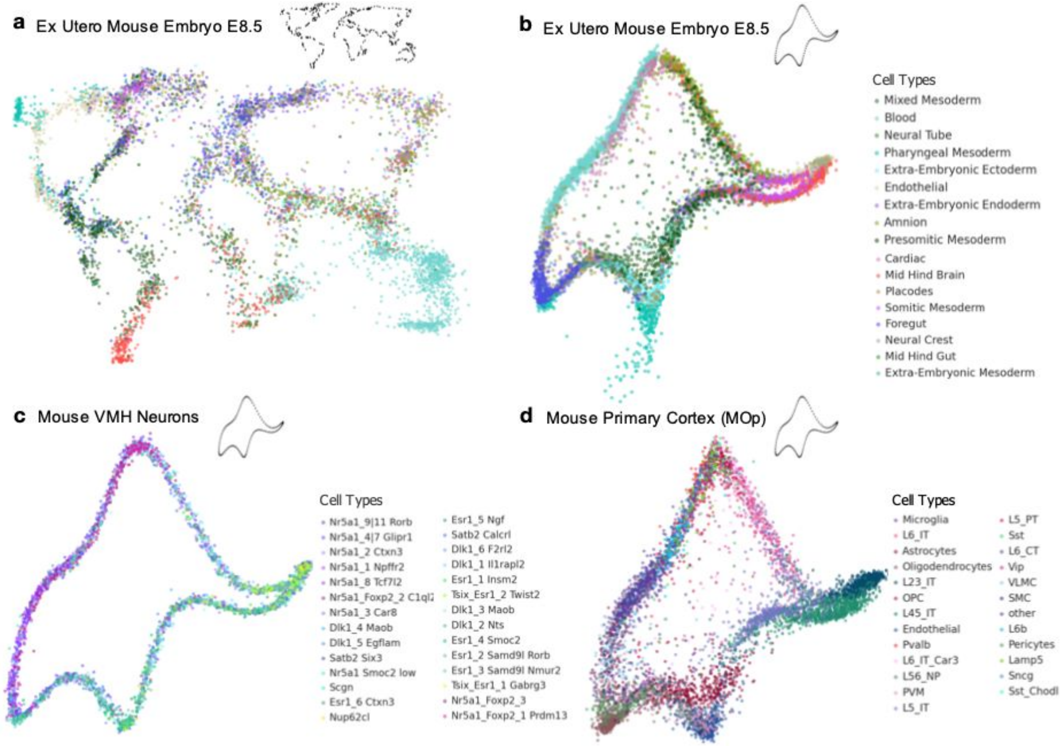
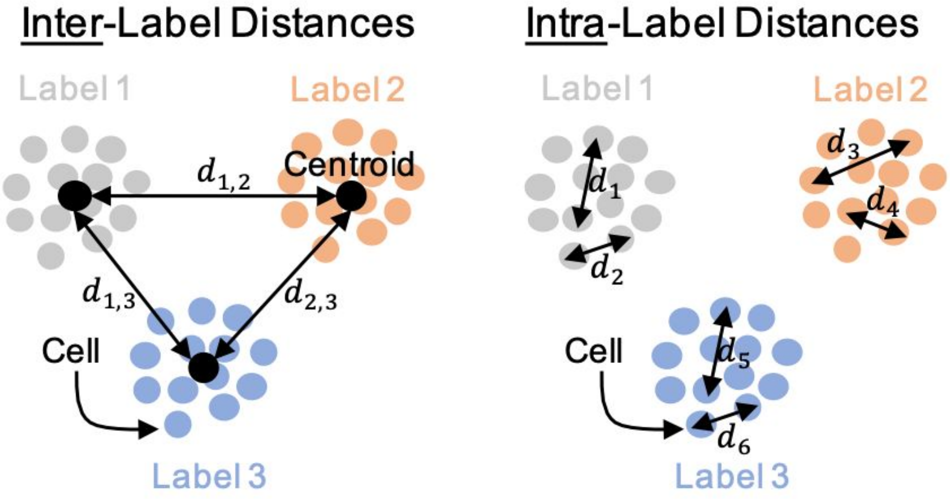


Figure 3: Picasso Embedding

Metric to assess preservation of biological structure



Supplementary *Figure 5*

Looking at iter- and intra-distances with respect to biological labels

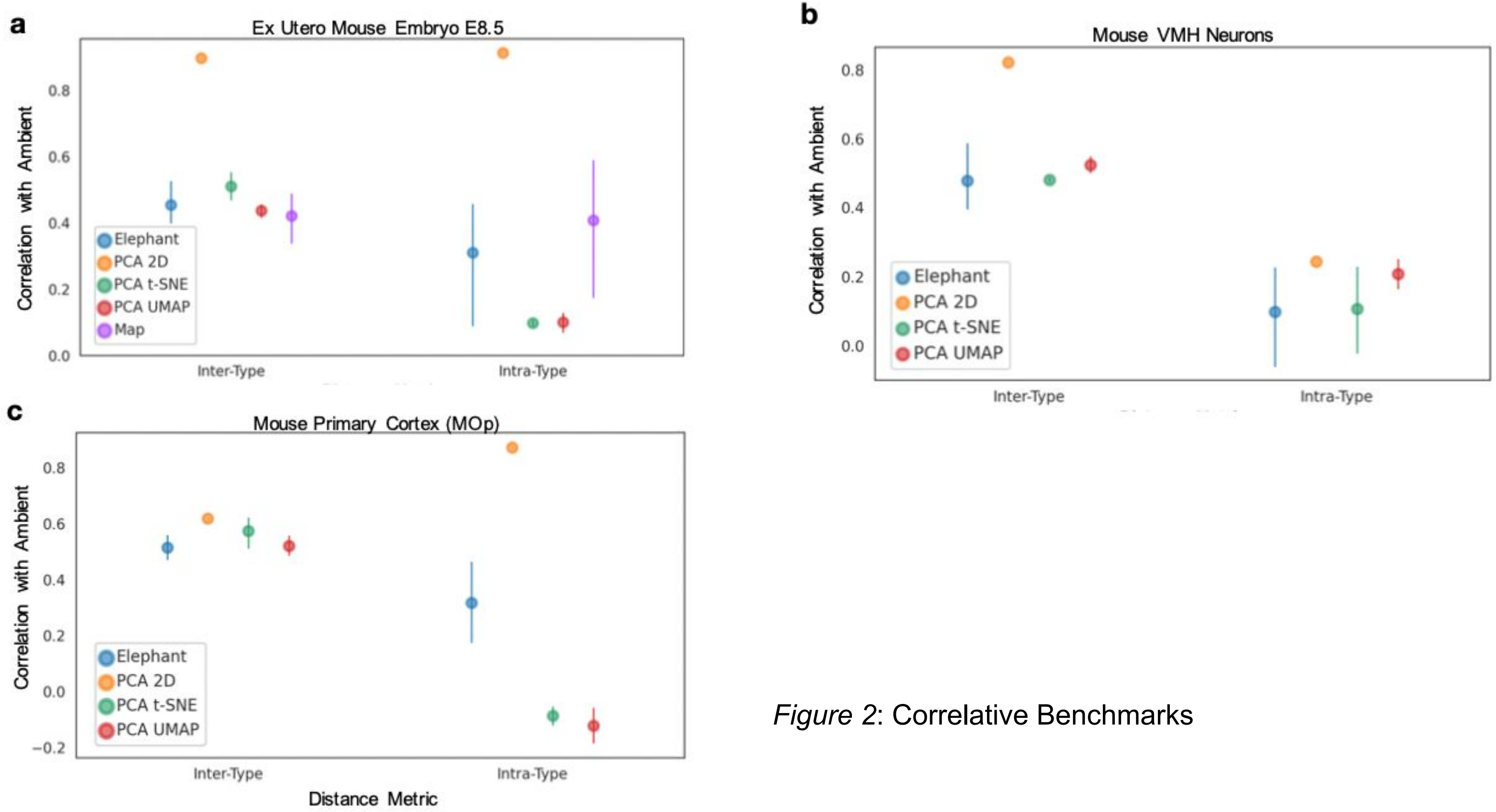


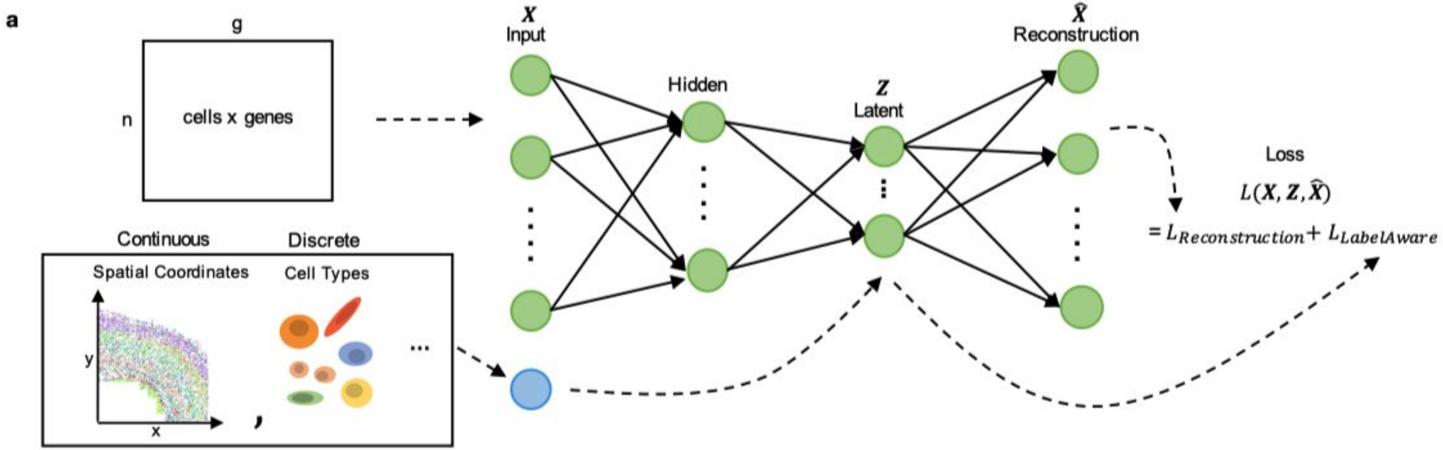
Figure 2: Correlative Benchmarks

So… what now?

- They argue that 2D embeddings, no matter how they are done, may induce unwanted distortions
- What should we do?
  - Stop putting 2D plots of genomics data in papers
  - Use their proposed method, MCML: “Multi-Class Multi-Label”
- The paper takes a bit of an odd turn here...

Motivation: Supervised Dimensionality Reduction

- "unsupervised dimensionality reduction, that does not account for the increasingly complex nature of multi-labeled genomics data including competing features in varying abundance, is likely to be suboptimal"
- Their method "MCML" can essentially be summed up as: use an autoencoder with an extra label-based clustering penalty: $L_{\text{LabelAware}}$



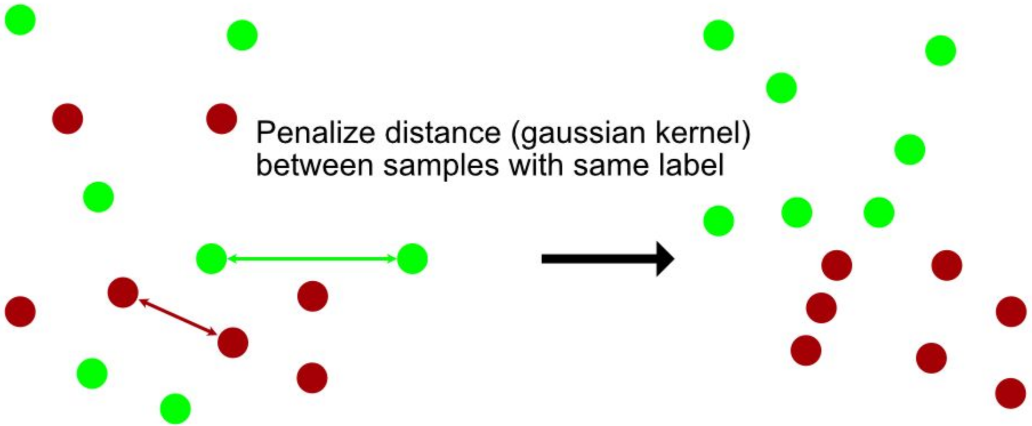
MCML: Multi-Class Multi-Label

- In more detail: let z_i be the latent embedding for the *i*th sample in the autoencoder. Then:

$$p_{ij} = \frac{exp(-\|z_i - z_j\|^2)}{\sum_j exp(-\|z_i - z_j\|^2)} \ , \ \sum p_i = 1.$$

$$L_{Discrete} = \sum_k \frac{\sum_{ij} p_{ij} \mathbb{1}_{ij}}{\sum_{ij} \mathbb{1}_{ij}} \text{ where } \mathbb{1}_{ij}(c_k) := \begin{cases} 1 & \text{if } \ c_{k,i} = c_{k,j} \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

MCML: Multi-Class Multi-Label



Penalize distance (gaussian kernel) between samples with same label

$$L_{Discrete} = \sum_{k} \frac{\sum_{ij} p_{ij} \mathbb{1}_{ij}}{\sum_{ij} \mathbb{1}_{ij}} \text{ where } \mathbb{1}_{ij}(c_k) := \begin{cases} 1 & \text{if } \; c_{k,i} = c_{k,j} \; , \\ 0 & \text{otherwise} \; . \end{cases}$$

MCML: Multi-Class Multi-Label

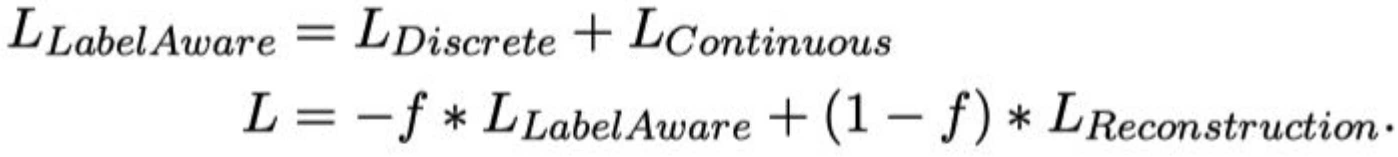
- For continuous labels, simply weight the distance using the label similarity

$$w_{ij} = \frac{exp(-\|c_{k,i} - c_{k,j}\|^2)}{\sum_j exp(-\|c_{k,i} - c_{k,j}\|^2)} \ , \ \sum w_i = 1.$$

$$L_{Cont} = \sum_k \frac{\sum_{ij} w_{ij} p_{ij}}{\sum_i max(w_{ij})}.$$

MCML: Multi-Class Multi-Label

- So their method is essentially… a very small modification to an autoencoder
- It seems unlikely this is a “novel” method
- Nonetheless, how do they evaluate their embeddings?



$$L_{LabelAware} = L_{Discrete} + L_{Continuous}$$

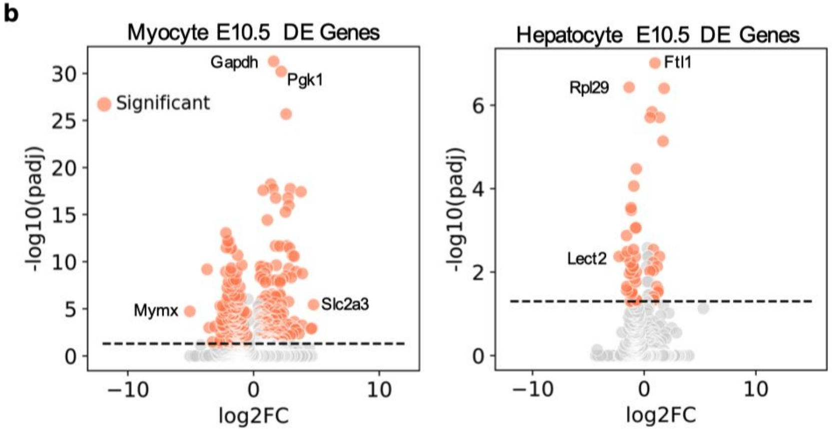
$$L = -f * L_{LabelAware} + (1-f) * L_{Reconstruction}.$$

Datasets

- Mouse embryogenesis (in-utero vs. ex-utero) + expression data
  - “Aguilera-Castrejon, A. et al. Ex utero mouse embryogenesis from pre-gastrulation to late organogenesis. en. Nature 593, 119–124 (May 2021).”
- C. Elegans embryogenesis (pseudo-time development) + expression data
  - Packer, J. S. et al. A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. en. Science 365 (Sept. 2019).
- Mouse primary motor cortex (spatial coordinates) + expression data
  - Zhang, M. et al. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics en. June 2020.

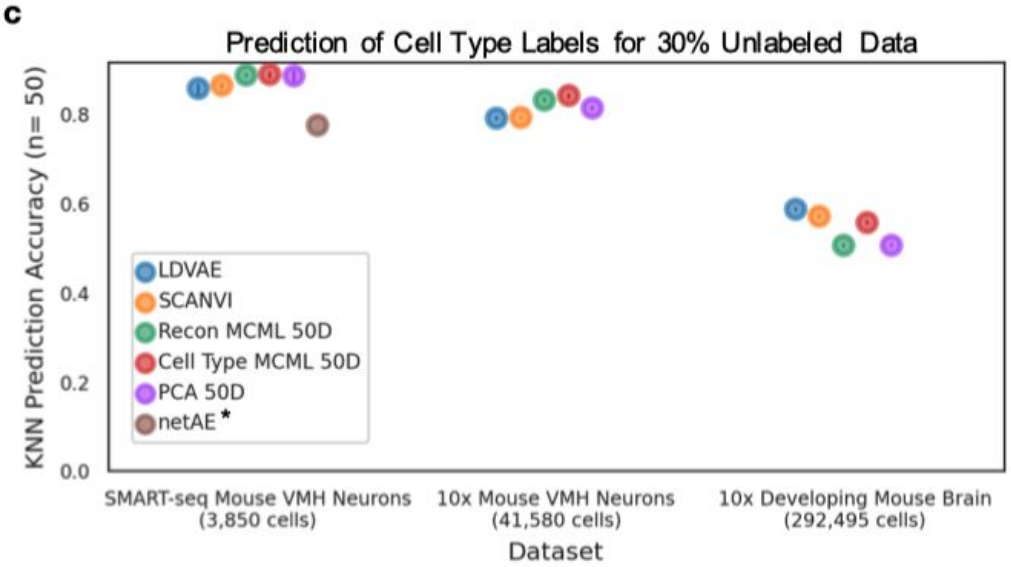
Identifying Cell Types

- They use MCML to identify cell-types that have the largest distances between in-utero and ex-utero clusters in the latent space
- They then use a standard DE pipeline to identify genes. They confirm findings from the original paper (myocytes) and highlight a cell type not discussed in the original paper (hepatocytes)



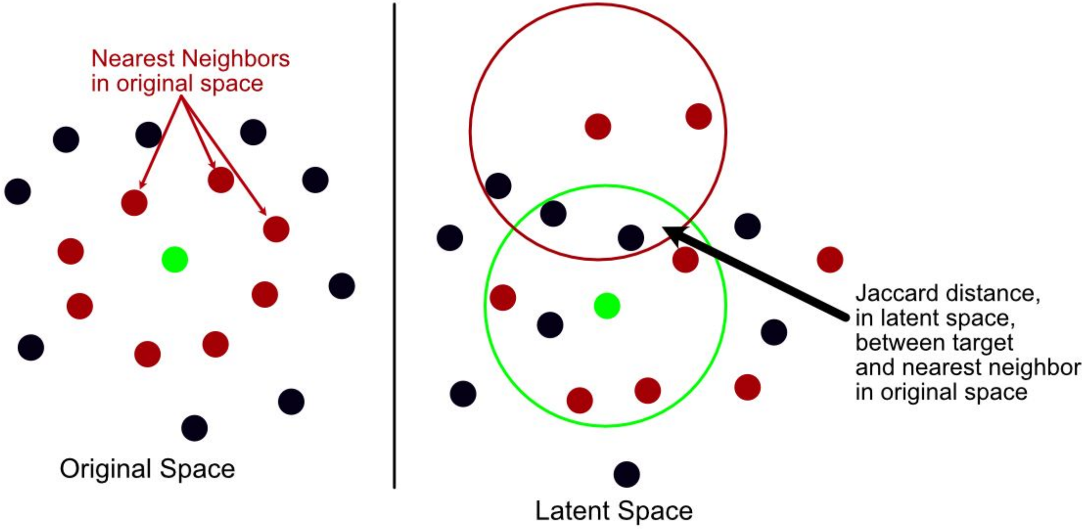
Prediction Accuracy

- They argue that cell type labels can be better predicted from their latent space as compared to other dimensionality reduction methods, although from the plot it is not entirely clear



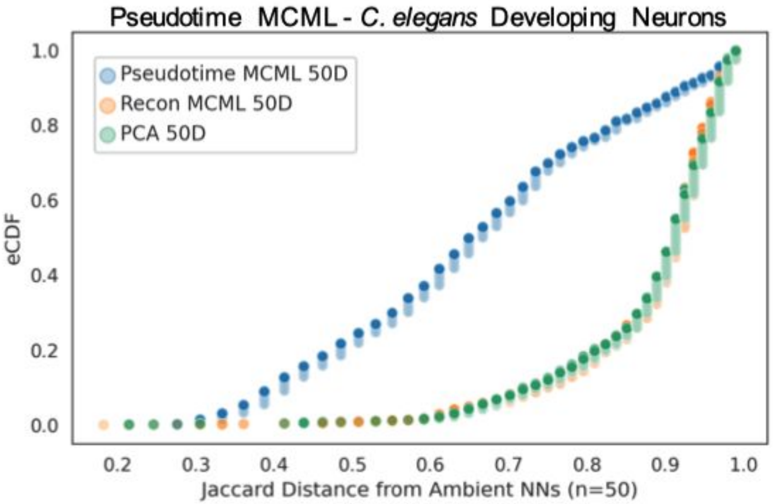
Another Metric: Jaccard Distance

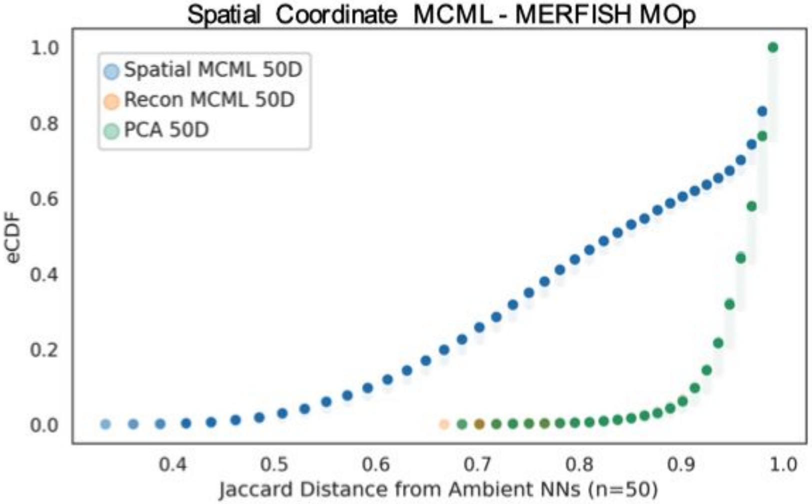
- A “good” embedding should preserve nearest neighbor structure
- Structure here is measured by Jaccard Distance between a sample and its original space nearest neighbors in the latent space


Nearest Neighbors in original space
Original Space
Latent Space
Jaccard distance, in latent space, between target and nearest neighbor in original space

Jaccard Distance eCDFs

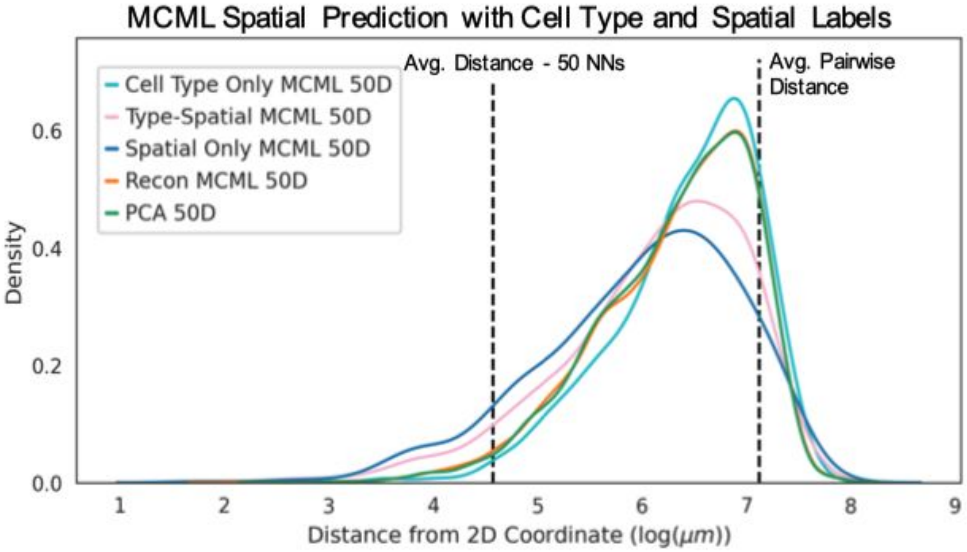
- Plots of the empirical CDF of (Jaccard Distances in the latent space) of a point with respect to its (nearest neighbors in the original space)
- In theory, more area under the curve = better





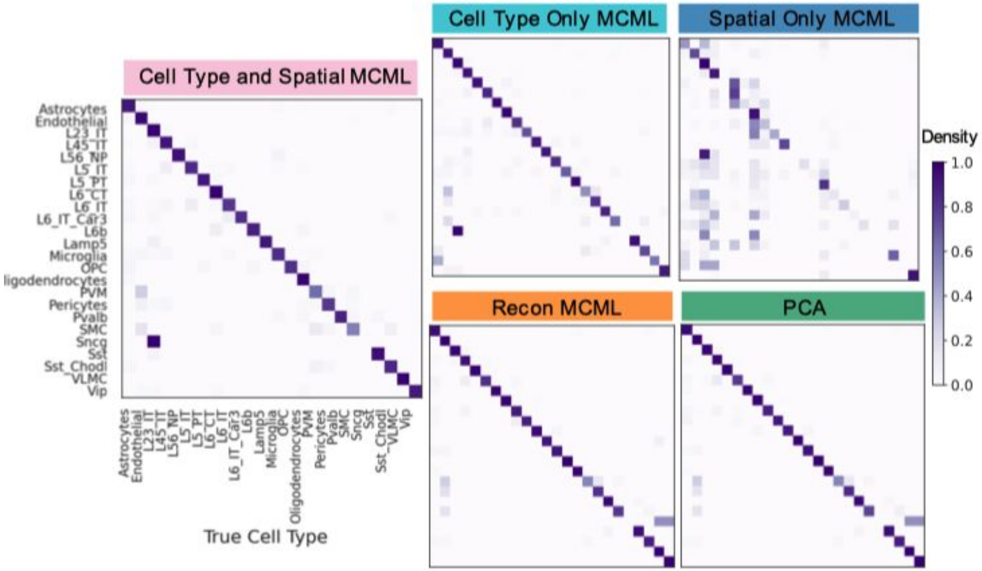
“Better” Prediction of Downstream Labels

- Predicting spatial coordinates of neurons in a 2D grid (not sure exactly how the labels were developed)
- Plot is a distribution of squared distances between predicted and true labels. Color scheme isn’t great but they do slightly better than just naive PCA



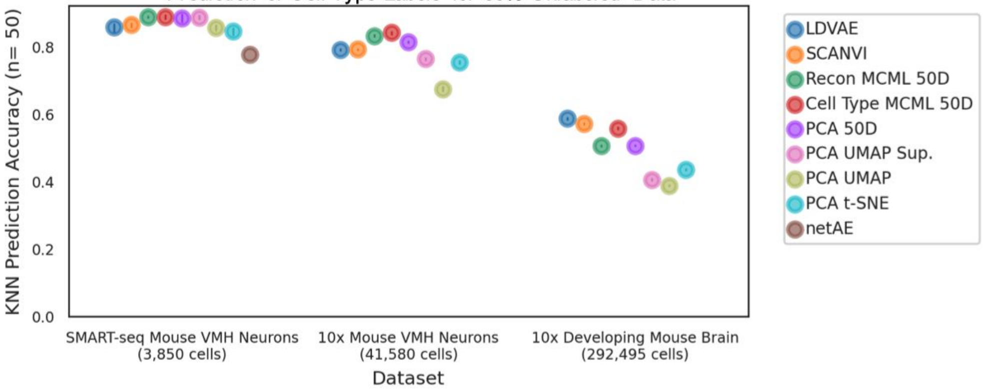
Predicting Cell Types

- Confusion matrix on predicting cell types
- This plot is… not super convincing

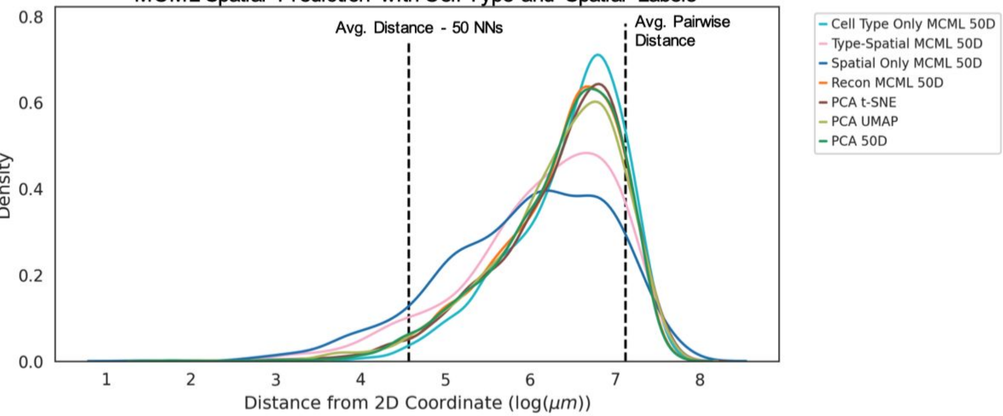


Some other random supplementary plots...

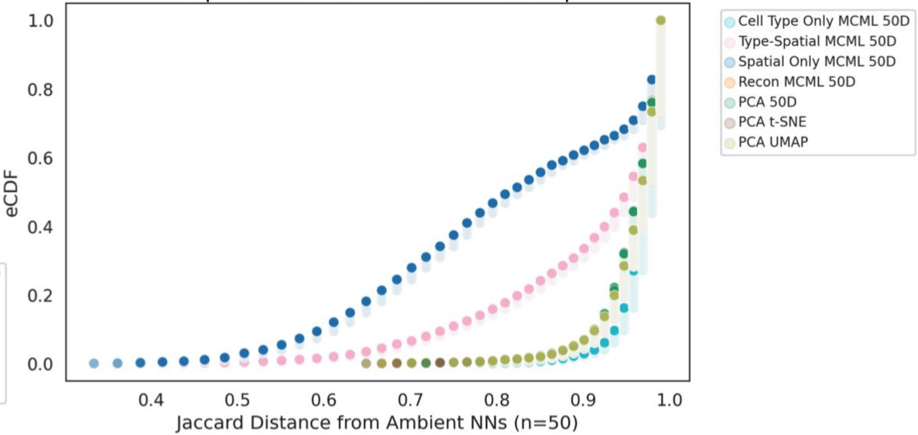
Prediction of Cell Type Labels for 30% Unlabeled Data



MCML Spatial Prediction with Cell Type and Spatial Labels



Spatial Coordinate MCML - MERFISH MOp



Conclusions - Part 1

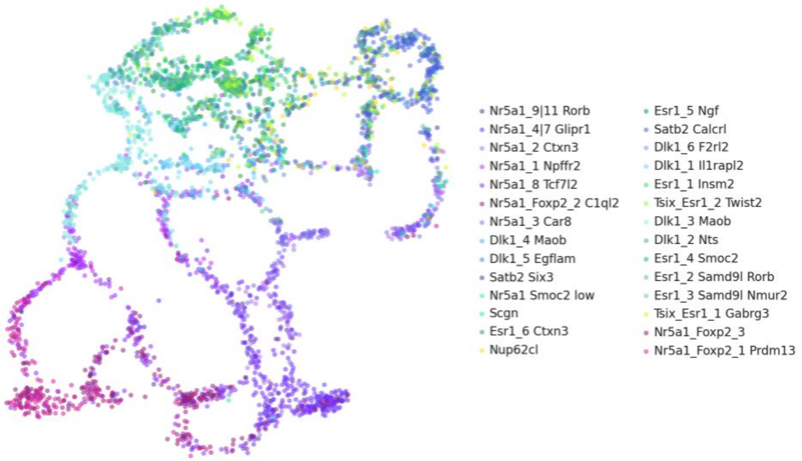
- Distortions are inevitable when projecting data (especially near-equidistant data) into lower dimensional space
- They use an autoencoder framework (Picasso) to fit cells into an arbitrary shape to show that (by some metrics), these arbitrary projections perform comparably to UMAP/t-SNE
- From these comparisons, they conclude a general inadequacy of UMAP/t-SNE projections for meaningful biological inference (especially for understanding patterns of variation within cell types)

Conclusions - Part 2

- They introduce an extension of an autoencoder they call “MCML”
- The loss function is based on intentionally grouping points that have the same labels in a lower dimensional space
- Experimental evaluation of the method is inconclusive, and baselines compared to in this paper were relatively weak
- In particular, clear metrics/standardized evaluations were lacking

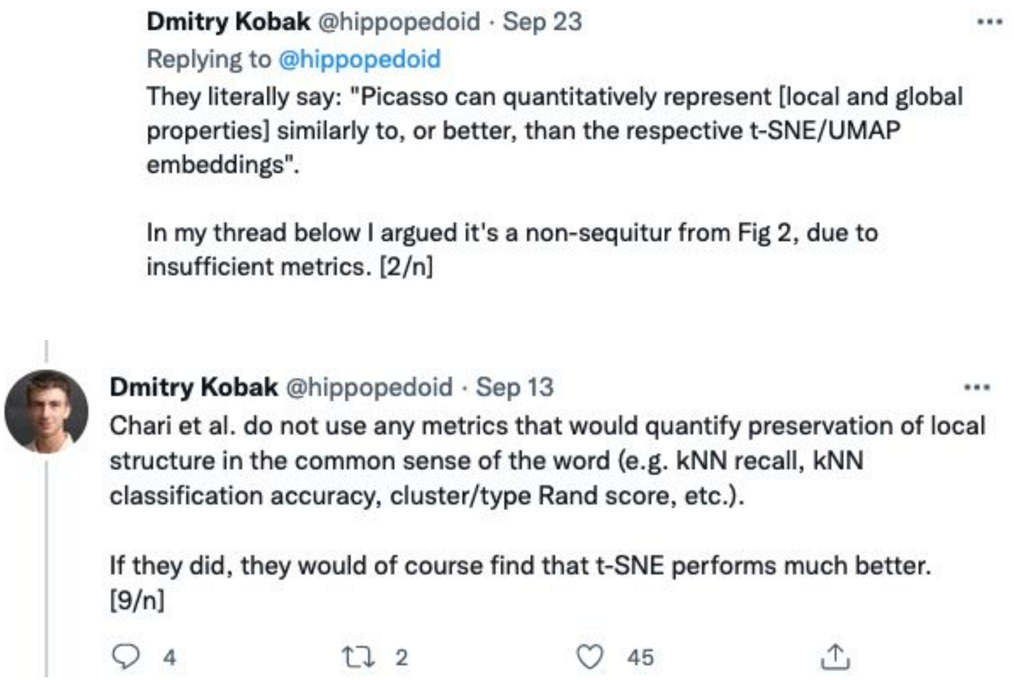
Discussion Questions

- What role does a 2D plot of high-dimensional data have in a scientific paper, if any?
- When you read a paper with such a plot, what do you take away from it, if anything? Is it useful, pretty but “specious”, or actively misleading?



Discussion Questions

- How do we “quantify” the performance or distortion of a 2D embedding plot?
- Does looking at distortion of equidistant cells or correlation of inter/intra class distances convince you? Are there any problems with doing this?



Dmitry Kobak @hippopedoid · Sep 23

Replying to @hippopedoid

They literally say: "Picasso can quantitatively represent [local and global properties] similarly to, or better, than the respective t-SNE/UMAP embeddings".

In my thread below I argued it's a non-sequitur from Fig 2, due to insufficient metrics. [2/n]

Dmitry Kobak @hippopedoid · Sep 13

Chari et al. do not use any metrics that would quantify preservation of local structure in the common sense of the word (e.g. kNN recall, kNN classification accuracy, cluster/type Rand score, etc.).

If they did, they would of course find that t-SNE performs much better. [9/n]

💬 4 🔁 2 ♡ 45

Discussion Questions

- What are the potential pitfalls of “label-aware” dimensionality reduction? Would you ever use such a method on your own data?
- Are there any kinds of biological signals that would be obscured by clustering points based on their given label (e.g. cell type)?

Discussion Questions

- If you were a reviewer, would you accept this paper? What feedback would you give?
- What is the role of social media in driving academic views, citations, acceptances? E.g., did this paper get more hype than it deserved because (1) it was tweeted from a popular account and (2) it has controversial opinions?

HEY, I SEE YOU JOINED INSTAGRAM.
YUP, I'M OFFICIALLY AN INSTAGRAD.
I'M POSTING SNAPSHOTS OF MY DAILY LIFE.
phd_comics
PHD Comics
Comics by Jorge Cham. Read more at:
phdcomics.com
RAMEN
RAMEN
JORGE CHAM © 2019
IT'S ALL PICTURES OF RAMEN.
I'M A LIFESTYLE UNFLUENCER.
WWW.PHDCOMICS.COM