0:00:03.58 –> 0:00:06.09 - Right, so I think while we’re waiting,
0:00:06.09 –> 0:00:09.473 I’ll just give a very brief introduction about Jingshu.
0:00:10.52 –> 0:00:14.2 Jingshu is an Assistant Professor from the Stats Department
0:00:14.2 –> 0:00:16.85 at University of Chicago.
0:00:16.85 –> 0:00:21.65 And today she’s going to present some very exciting work
0:00:21.65 –> 0:00:25.233 on trajectory inference for the single cell data.
0:00:26.7 –> 0:00:28.893 I’m very excited to hear about her work.
0:00:30.99 –> 0:00:31.823 I think…
0:00:33.5 –> 0:00:35.23 Lets wait for two minutes
0:00:35.23 –> 0:00:39.72 and then we’ll start with Jingshu’s work.
0:00:39.72 –> 0:00:43.747 So if you have any other related questions about Jingshu
0:00:44.59 –> 0:00:48.32 before the talk start, you’re free to ask as well.
0:00:48.32 –> 0:00:50.73 (chuckles)
0:00:50.73 –> 0:00:53.05 - Hi Lynn. Can you make me a co-host.
0:00:53.05 –> 0:00:56.423 - Oh right. Thanks for reminding me.
0:00:58.26 –> 0:00:59.093 Let me see.
0:01:13.858 –> 0:01:16.191 So one more minute to start.
0:01:41.502 –> 0:01:43.15 - So you can see my screen, right?
0:01:43.15 –> 0:01:46.593 - Yes. I can see your screen. Looks good to me.
0:01:47.68 –> 0:01:50.333 Maybe I’ll hand it over to you now Jingshu I think,
0:01:52.227 –> 0:01:54.193 if (indistinct) late I think,
0:01:55.81 –> 0:01:58.16 they can ask questions if you miss any details.
0:01:59.34 –> 0:02:01.38 - Yes. Okay.
0:02:01.38 –> 0:02:03.9 So thanks everyone for coming
0:02:03.9 –> 0:02:07.523 and settling for the introduction invitation.
0:02:08.41 –> 0:02:12.69 Today I will talk about the Single-Cell RNA Sequencing Data
0:02:12.69 –> 0:02:15.61 and how we can learn the cell dynamics
0:02:15.61 –> 0:02:18.433 from the single-cell RNA sequencing.
So the single-cell RNA sequencing is a relatively newly-developed technology for measuring the RNA expression levels in individual cells. The traditional microarrays or bulk RNA sequencing matches the gene expressions as the average across all cells in a tissue. However, a cell is made of many cells and a tissue is made of many cells and the cell population is typically not homogenous and the cells can have different functions. So in contrast, in single-cell RNA sequencing, we have measured the transcriptional profile in each individual cell. We can expand this vector of gene expressions for a tissue to a matrix, and each entry is a detailed RNA count for a particular gene or a particular cell. The benefit is that we have a more detailed understanding of what is going on in the tissue. The benefit of single-cell RNA sequencing is that it can give you a relatively unbiased and complete picture of the cell population. This is particularly useful when the cell population is complicated.
And as an application of the method that I will introduce today in this lecture, in this talk, I will focus on the study of the mouse neocortex. This is a cartoon showing the migration and generation of the projection neurons in the mouse neocortex. Yeah, you guys see that this is quite a complicated process and there are still a lot of things that are unknown about the neuronal diversity and the mechanism of how the projection neurons are generated. And the goal is that we want to use the single-cell RNA sequencing so that we can have a more complete understanding of this, the neuronal diversity and the neuron development. So you can see that here in this cartoon, this shapes, there are different shapes and colors, to represent different cell types in the neocortex, as the cells are experiencing the continuous dynamic changes actually in the real cell population, it is much complicated than that. There is not clear boundaries between different cell types and there may be... There even, it’s not a clear definition of cell type. So, what we hope is that we want to use single-cell RNA sequencing to first recover the trajectory of the dynamic changes or the developmental process that the cells are experiencing. Specifically we focus on two datasets. One data set, we name it as data set A.
So this is a data set that is recently collected by my collaborator.

And so we have samples... The cells from the mouse neocortex at six different embryonic days.

And before our data, there is another dataset we call it, we name it data set B.

And this dataset is a smaller dataset than ours but they are... They have also sequenced a very similar brain region of the mouses and they have a sequence of cells from four different embryonic days.

So you can see that our, most of the days that are sequenced in our dataset and with the other dataset B, do not overlap.

And so it would be beneficial if we can have with... If we can combine the two there datasets and so that we can make use of the cells from both studies.

For instance, for our dataset, we don’t have these cells from the day 11, which is quite important day.

For example here, day 11 are the day that, there are projection neurons that are, beginning time, well, there are projection neurons that are generated.

And so this E11 cells are sequenced from the other dataset.

So it would be beneficial if we can perform a choice analysis of the two datasets and learn a shared developmental trajectory as these two datasets,
are actually sequencing the same mouse brain region. So as you may have imagined, if we don’t do anything, if we just concatenate the cells from two datasets and treat them as datasets from the same lab, then these two datasets actually will not merge because of the bash effects between the two datasets. Because these are from two labs and they have different sequencing machines so the cells become different, though they are coming from the same brain region. And this is a figure called the UMAP which is a two-dimensional projection of the high dimensional, observed single-cell RNA sequencing data so that we can have a visualization of the cell population. And using our marker which is called vitae that I will introduce later or we can merge the cells from two different sources. And as I will show later, we can also keep the uniqueness, the unique characteristics that only exist in one of the datasets. So we can keep the biological meaningful differences between the two datasets. And our method is actually not just, data integration approach. So what we can do, is that we can also simultaneously, learn a shared trajectory structure
and we can at the same time do the disintegration or more generally correct for confounding effects such as the data source and other various like cell cycles.

And in the...

In this figure, the arrows show the direction of the developmental process and the line width represents the score for an edge. So it shows how confident we are in whether there’s a transition between the two states that the line connects.

So our method actually belongs to a larger group of computational tools for single-cell RNA sequencing which is called the trajectory inference. So it is called trajectory inference that is different from statistical inference.

In general, the trajectory inference approaches, they will infer or they start with a type of the underlying trajectory structure and the cell fate decisions in biological process, such as cell differentiation as what we have already seen in the mouse neocortex, and some other biological process, such as immune response, cancer expansion and many more are using single-cell RNA sequencing data.
they will assume a specific type of the trajectory structure for the underlying developmental process, such as a linear structure, a linear topology or a bifurcating, a bifurcation or tree-like trajectory. And as the cell populations that we are trying to understand, become more and more complicated, recent methods also try to infer, the type of the trajectory structure from the observed single-cell RNA sequencing data. And whilst we have learned the trajectory structure, then this trajectory inference approaches, will computationally project and order the cells along the trajectory. And the right order of the cells along the trajectory, are called the pseudotime of the cells. So the trajectory inference, is also called the pseudotime analysis. And since... So the first trajectory inference method is proposed in 2014 and since then it has become a very popular tool that are used in analyzing single-cell RNA sequencing data. And in this study, it calculates, it summarizes the number of single-cell RNA sequences studies that are published per month. And you can see that in recent years, more than half of the published single-cell RNA sequencing studies
will have some investments of the pseudotime and trajectories in the cell population that they are investigating. And there has also been a lot of methods for trajectory inference. And in this, there is a comprehensive benchmarking paper, recently in “Nature Biotech” and it has summarized 70 different trajectory methods. And in your paper they have compared about 45 different trajectory inference methods from different aspects. So you may wonder, since there are so many trajectory inference methods that are already there, why do we still want to develop, a new trajectory inference method? So the first point is that, although we have 70 different methods, many trajectory inference methods, they are assuming a specific type of the trajectory structure. So many methods only work for a sound developmental process. If you consider the methods that can work for... They have the flexibility if you work for a variety of the trajectory structures then we don’t have that many methods that are available. And another concern that I have is that most methods, these trajectory inference methods, do not have explicit statistical models. So what I mean is that,
though people are kind of clear about what’s the biological signal that we want to find in the trajectory inference, it is actually, many methods are actually pretty vague about from the aspect of like for the single-cell data matrix, what can be the definition of the trajectory that they want to infer. So, and how that they are generating, and how the data it can be modeled and generated with the trajectory structure. So as the statistician, I think it would be beneficial, if we have a model-based trajectory inference approach, so that we can better understand the profit, how good our estimations are and have some certain qualification of the trajectories or slow times that we infer. And the third point is that there is also a growing need, to efficiently align trajectories or do a joint analysis from multiple single-cell RNA sequencing datasets. As the single-cell RNA sequencing datasets are expanding, there has already been a lot of studies for datasets, they are for the same tissue or for the same cell type. And it will be... (clears throat) And we can learn a better picture of, on this, the biological process in the tissue or for the cell time,
if we can use all available datasets.

And so there’s a strong need, an increasing need to do this joint trajectory analysis for multiple datasets.

So because of these reasons, we develop a new statistical framework and a new method, and we call it VITAE, which is short for variational inference for trajectory by autoencoders. And it is a model-based trajectory inference approach. So our model starts with a definition of the trajectory backbone. So we use a graph to define the trajectory backbone. So we start with a complete graph G, well the vertices are the distinct cell states and cell type and an edge denotes a possible transition between two cell states and or cell types. And then we can define a cell position on the graph which is a vector, which is a landscape vector. And it’s... A K is the number of vertices on the graph, in the graph. So if a cell is exactly belongs to one cell state or cell type, then it is on cell vertex. And if the cell is experiencing a transition between two cell states or cell types, then we denote it as on the edge between two vertices. And then we can define the trajectory backbone as a subgraph of G.
So we only include an edge or vertex if we really observe cells that are on the edge. So though there are many possible transitions between the cell types or cell states, We believe, we include a transition only when we do observe cells that are experiencing such transitions. So though there can be many edges in our complete graph, on the sub graph, it’s a sparse success of the other edges that are possible on the graph. And a benefit of the above definition, is that we can allow any types of the trajectory structure. So it can be either a linear structure, a bifurcate chain or a tree-like structure or a cycle, it completely depends on how the data shows. And so we allow, we want the data to automatically determine the other, the trajectory structure or topology of the underlying dynamic process. And we can also define the pseudotime for each cell. I have not written down the exact definition here but the idea is that we first need a root vertex. So a root vertex is the start of this dynamic process. And it can be given by the user, depending on looking at the marker genes or other side biological information. And later we will also... will also show you that for some datasets, we can automatically determine the root vertex. And with a given root vertex,
then the graph becomes a directed graph.
And we had defined the pseudotime of the cell as the shortest path from the root to a specific cell along the trajectory, along the trajectory backbone.
So this graph defines the trajectory structure.

And the next step, is that we want to link the trajectory structure with the data generation model.

So the single-cell RNA sequencing data matrix is typically a high dimensional matrix because for each cell, we typically observe tens of thousands of genes and there are also complicated dependency relationships among the genes.

And what we assume, is that we assume that these dependencies across genes, can be explained by a latent variables, $Z(i)$ in a low dimensional space.

And we assume that these latent variables, are following our normal distributions and they also have the graph structure.

So $U$ here, are the positions of the vertices on the graph in this low dimensional space, which is a linear combination of these vertices, depending on, of the positions of the vertices, depending on the cell’s graphic position on the graph.

what I want to emphasize here is one point, that we assume a non-linear marking.
from the latent space to the high dimensional observed data,
because we think that though in the low dimensional space,
we can represent the trajectory as these linear lines,
it is very likely a manifold on the observed data.
So this non-linear mapping,
can map this linear lines to hertz
in the high dimensional space.
And now to consider,
to account for the confounding covariates,
such as the data source or cell cycle,
we also allow this non-linear mapping
to depend on this covariates.
And here we are...
Because the observed data count,
we assume it follows an active binomial distribution,
and L(i) here should be known library size.
Sorry for the typo. It should be known library sizes.
And CRJ, and the CRG, the dispersion parameter
are unknown parameters.
And so in this, in the current model,
the unknown parameters we have,
are these cell, the vertex positions,
the cell positions on the graph,
the non-linear mapping
and this unknown dispersion parameters.
So we have a lot of parameters.
So to further simplify our estimation,
we assume that there are a mixture prior
on the cell positions.

So it’s a very tactical idea.

So we assume that first the cell,

there are some latent variables, Ci for each cell.

And so Ci determines which edge or vertex a cell chooses.

So the cell has some probability to choose a specific edge

or a specific vertex,

and if it chooses an edge,

then, it then eventually choose the location of,

the relative location on the edge,

So that’s what becomes a mixture prior on this diagram two.

And,

for these non-linear mapping functions,

we’re including non-linear mappings.

We model these F(G(i)) functions by a neural network.

So, Oh, do you?

So our parameters now, our known parameters now are,

this U the positions of the vertices

this PI, the probability of each vertex and edge,

and the non-linear mappings,

which are the waste in the neural network

this dispersion parameters.

And we...

I space these parameters by combining our mixture model

with a variational autoencoder.

So the variational autoencoder.

So the autoencoder has been a very popular model
So what it can do is, is that it can, can have some non-linear mapping of the observed data to a low-dimensional space and we want the low-dimensional space to best, recover our observed time rational data. And here, we also have such a task. We have the low-dimensional space and we want to best to recover our observed data. And what’s different is that we also have a prior on the latent space, because we have the prior, so we use the variational autoencoder model. So the classical variational autoencoder in deep learning, we’ll assume that the latent space, has the standard normal distribution as the prior. And here we just modify it so that we have the... So that the latent space have the mixture prior that are assumed in our previous mixture models. And we use the same approach as the variational autoencoder, the variational path which is, though the, to approximate the posteriors of the latent space. So though, because of the complicated priors and non-linear mappings, this prior, the posterior of the latent space conditional on the observed data and the confounding covariates, it can be complicated, we approximate it by our normal distributions and the mean and the variances of the normal distributions which are functions of the observed data Y(i)
and the covariance $X_i$, are also non-linear functions and we model them by the neural network and that’s the encoder. So the decoder is the nominal mapping function $F(G(i))$, mapping the latent space to the observed data. And encoder are neural networks that are approximating the posteriors of the latent space.

Hi, Jingshu, can I ask a very quick question? If I understand correctly, up to now you have not used the time information, or you have considered to include the time information in the covariate $X$?

So we will not use time information in our trajectory inference. If I have time, I may have a last slide which is a review of the literature into data inference. The most commonly used and best performing methods, they will tend to not use the time information because there is... Though the time information is typically correlated with developmental like, timing of the cells, but because at each collectional time, is a mixture of cells at different environmental time. So it’s a big, complicated relation and some methods use that information but many methods do not use that.
And so our approach is the methods that do not use the, the collection time information. And, we use it only when we decide which vertex is the root. And I will show that later.

So our... So the last function is, is composed of three parts. The first part is this likelihood based reconstruction loss. So this evaluates how goods are our latent spaces to, to reconstruct the high-dimensional observed data. And the second part is the KL divergence between the posterior distribution and the prior. And you can think of it as a regularization term. So here we make the, we add the training parameter beta, and in practice we said, beta are larger than one, so that we can encourage the posterior, the regularization, so that the posteriors of the I will, are more likely to tend to align along the edges and vertices. And that’s the idea that has been used in deep learning which is called the beta.

And the third term,
it’s a term for adjusting for the covariance. So the covariance...
So this covariance * covers. So we want our latent space C to be kind of, be correlated with the covariance, While...
So we, certain...
So we want to maximize the reconstruction of the data by only by the covariates.
And setting the tuning parameter alpha larger than zero,
we can help decorrelate Z(i) from Xi.
So another art in our training is that, we need a good internalization of the graph.
So specifically we need to determine, how many vertices there are,
and also the positions of the vertices in the low-dimensional space.
That’s not an easy job.
And if we just randomly,
because our final graph depend on,
the total number of vertices that we set at the beginning.
So how we pretrain the model to return it’s initial value?
To get the initial values of the unknown parameters,
is that we first trained with beta equal to zero,
so that we don’t make use of any,
these prior distributions of the I.
So it’s only...
We’re only looking at the reconstruction loss from the likelihood of the data.
So it’s like the normal, the classical autoencoder.
And from that we can get some initial estimate of $Z(i)$,
The latent space variables.
And then we perform clustering on $Z(i)$,
and we let the clustering algorithm,
to automatically determine the number of clusters
and we use that as the number of vertices.
And we also use the cluster centers
as the initialization, as the initial values of $U$.
So that’s the main part,
the key ideas in our pre-training start,
so that we can have a good initial addition to,
for the start of the training.
And another trick that we have taken,
is that in practice, sorry,
the best performing existing trajectory inference methods.
They will attempt.
So they are typically very fast.
And in order to have comparable computational costs
of these methods,
we also have accelerated version of our algorithm
which is a simply to reduce the input,
the dimension of the input space,
so we can replace $Y(i)$,
the high-dimensional vector of the gene expressions
with its principal components.
Now, principal component, principal scores,
which is a low-dimensional vector $L$.
We, by default we will take the first 64 dimensions
for the principal scores.
And so we replace the elected binomial distribution by a normal gaussian distribution assumption of these principal scores. And as you will see later in our evaluations with real and synthetic data, we see that we actually have comparable performance with our previous likelihood, with our standard likelihood based methods for this accelerated version. So after the final step is that after the training the autoencoder, we have approximated distributions, posterior distributions of the latent space and also the cell positions that and which vertex, the vertex or the posterior distribution of \( C_i \), well \( C_i^* \) is which vertex or edge the cell is from. And we need to use those information, to determine the trajectory backbone and to project each cell on our inferred trajectory backbone. So how we do that is, first we calculate an edge score. So this edge score is... So we have different scores for an edge, and that is determined on looking at the posteriors of cells. How many cells from the posterior distribution? How many cells choose to lie on that specific edge? If there are a lot of cells then that means it’s very likely that edge exist. If there are very few cells
then very likely that edge should not be, the edge that is included in the trajectory backbone. And the denominator is that we want to, give a relatively high fair score for the edges of that connecting to small clusters, to small cell types, such as we want to also capture the transition between two rare cell types. So that’s why we have this regularization, waiting by with, in the denominator. And we include an edge in the trajectory backbone if it’s edge score is larger than some*. And when we have an inferred trajectory backbone, then the next step is, we want to project the cells on the inferred trajectory, and we do it by looking at the... Based on the posterior, approximated posterior distributions of the cell positions that *.

We want to find the closest point on the inferred trajectory for this cell based on the posterior distributions. And the distance, this expectation we can also use as a evaluation of the, some uncertainty quantification of this projection or the cell positions. And the third thing we need to, because our final results is some kind of directed graph. So we need to determine the root vertex. So the root vertex can be either given by the user, or if as I feel I asked,
for some datasets like the data at the beginning of my talk, the cells are collected in the time series, and we can make use of that time series to determine the root vertex. The rough idea is that for each vertex, we can calculate a fraction time score, which is an average of the cells or projected on the edge that connects to the vertex, depending on the distance from the cell to the vertex. And so we can have some vertex collection time, collection time score for each vertex. And we choose the root vertex as the vertex that has the smallest collection time score. And with the roots and with our inferred trajectory, it’s straightforward to calculate the pseudotimes for each score. So that’s the whole process of our model-based methods for trajectory inference. And now let’s look at some benchmarking results. So our benchmarking includes both some real datasets and some synthetic datasets. And we follow the... Most of the benchmarking follows the same framework as this well known benchmarking paper in the “Nature of Biotech” in 2019. And our datasets are selected as a subset of the datasets that they have tried and our criteria is that these datasets,
maybe have enough number of cells not too few cells.
And we wanted to cover different types of topologies.
And this is the benchmarking results.
So the columns, sorry.
So the rows are the different datasets
that I have mentioned.
And we compare five different methods.
We have, would come first compare two versions
of our approach.
Vitae one as the original elected binomial likelihood base.
Vitae and accelerated version,
replacing the gene expression vectors
by principal scores.
Then we compare it with three different,
state of the arch trajectory inference methods.
The monocle series are always commonly used
in these single-cell RNA sequencing papers.
And two, I expect the performing,
trajectory inference methods in the benchmarking paper,
the PAGA and Slingshot.
And all these methods,
do not use this time information explicitly.
So it’s a fair comparison between these methods.
And for all the methods they are given to by those.
We give them the two numbers of clusters or the vertices to start from, and the two root vertices. And we, and each column is, we compare it's measurement, it's metric for the evaluation of the performance of each method. So the first two columns, are the matrix for recovery of the trajectory topology or the trajectory structure. And next two columns are the evaluation of the cell position, estimation accuracy. And the last metric is for evaluating the pseudotime accuracy. And a larger score means a better performance, and a lower score means like, a worse performance. So you can see that, our approach first is, our approach has much better performance in recovery of the trajectory topology. We also have some benefits of the cell position estimates, and because of both, we have a better performance in the pseudotime accuracy. And the other thing you can see is that our, our accelerated version have comparable, slightly worse but comparable performance, compared to the, our likelihood based vitae. And though it has a much quicker, much less computational cost. So finally, let's come back to the case study on mouse neocortex. So this is the,
the visualization of merging the raw data. And this is the performance of our methods. And for comparison, we compare is, another very popular use, data integration method called Seurat. So Seurat is the software, the most often used software, for single-cell RNA sequencing analysis. Their lab have different, have developed a series of computational tools for analyzing the single-cell RNA sequencing data. And this is from their integration methods. So you can see that both methods can, is able to integrate the both two datasets, but for some details, I think, because we are assuming this trajectory structure, we have a slightly better performance. For example, this group of cells are the layer one neurons, where the group of here, are here in Seurat. So you can see that because they come from, because the outer layer parts and the layer parts, come from two datasets. Because as I mentioned earlier in dataset B, they have collected cells from, at day 11. So this are, we can take a look of the collection days of each cell. So you can see that these cells, they are, the layer one parts come from day 11, And the rest parts is a mixture. of cells from both two datasets. And by the way they all belong to the layer one. So we know that they belong to layer one
by looking at the marker genes expression which I did not show here. So it’s because we encourage the cells to align together if they are along the address, if they are similar cells. And you can see here, that’s, so the two datasets, they have this interpolation of the pseudo, of the collection time. And you can see for example, for these projected cells, we can see this continuous positions, like alignments of the cells of different days from the most the dark is the cells from day ten. And the red ones are the cells from day 18 and even days are, are from dataset A, and odd days are from dataset B. So you can see that though they’re coming from two different sources, we can, we are able to align them in the right order. And, as another comparison. So we compare our estimation of shared trajectory, with another partisan approach which is we’re first to do data integration with Seurat and then we can use Slingshots, to perform trajectory inference on the integrated data. And you can see that this, we have a much cleaner trajectory structure. And we also have a comparable computational cost. So Seurat and Slingshots, they cannot be, they do not need regularization. And with one CPU, they, it takes about 12 minutes.
And for our accelerated VITAE, generating this figure, we have, we can, we take about three minutes on one GPU port at about 10 minutes on eight CPU cores which is, the eight CPU cores are like currently, like most of our laptops, but we’ll have such computational resources.

So we have comparable computation cost with this state of our methods. And in addition, because we are... Based on this approximated posterior distributions we also have some kind of uncertainty quantifications on the cell positions. For example, here, it shall say some parts of the cells, these cell positions along the trajectory are not very reliable. And that will help us to evaluate our, the estimate, how we think our estimate in pseudotime. And finally, this is showing some gene expression change along the pseudotime order. And, and we look at some top markers that are changing along the pseudotime order for some trajectories in the whole trajectory structure. And you can see, here we separately fish the curve for two datasets, but you can see that they overlap with each other quite well. And so that’s also an evidence showing that we can have a good,
A good performance in aligning the two datasets. So the take-home message is, first we perform this model-based trajectory inference, to understand cell dynamics. And our, the second is our methods. So our method is a model-based approach. We can combine the mixture prior model, Oh, sorry. We can combine the collected mixture structure for defining the trajectory structure with the variational autoencoders so that we can efficiently, efficiently solve the mixture model and have enough flexibility to fit the data well. And so our, trajectory inference approach features, the analysis of integrating multiple single-cell RNA sequencing datasets. And if you are anxious to know more details, we have our paper, a manuscript already available on bio archives and we also have our package codes on VITAE. And that’s all. Thank you. And if you have any questions, I’m happy to answer them. Thanks Jingshu for this excellent, excellent talk. I wonder whether the audience, have any questions for Jingshu. Okay. So Jingshu I have some maybe minor questions. I recall that in the model, you have this actual term encouraging X, the data explained by X the covariates,
the confounding covariates,

to orthogonal to the leading factor, right?

And then there is a penalty term alpha.

So I wonder,

what’s the motivation for you to including this term

And in practice, how should we have set alpha?

- So, so, so the thing is a bit tricky here

So we want to remove these confounding effects, right?

But the other fact is that,

these confounding effects, the X,

are not exactly orthogonal with Z,

because for instance for the two datasets that I have,

we cannot say that the signal is completely orthogonal

to each dataset they have come from

because there are two biological differences

between the two datasets.

So, so here,

the problem is not completely identifiable

but people do it in practice a lot.

So we want to kind of decorrelate Z and X to some extent

so that we can remove,

remove the batch effects that we do not want

but keep the two biological differences.

So I think the underlying assumption

the big assumption is that we are assuming

that the two biological differences are large enough

so that compared to the...

So we remove the smaller differences, the batch effects
but we can still pick the two biological differences, to some extent.
So there’s no guarantee that it will work, but in practice, it work on a lot of datasets. I think that’s...

So we will inherit this idea from this paper by Nancy Huang and her students. So I think in removing the batch effects. So I think the idea is that, we hope that it can work for a lot of datasets. And the reason that we want to have this penalty, is that if we don’t add any penalty, then, because this autoencoder is trained by this, by this stochastic gradient descent. So sometime it may not find the optimal global solution. So if we don’t encourage it, the X and Z to be decorrelated, it sometimes may not be able, it may give you a solution that is not, that’s the Z still are highly correlated with X and the batch effects are still there. So then this alpha, I think in practice we set it to be 0.0, so it’s a very small penalty so that we can put some, some kind of penalty to regularize that. Small amount, I guess you mentioned. - Yes, yes. And it may be that does not work, and then in practice you can choose another alpha and try it. and see if it gives you the best results that you want.
I guess, right.
So I guess my question is like, since it’s not entirely a supervised problem, like how, right.
So I’m not sure how to check, what is a good alpha in the sense, but if you tell me like a small alpha, well you’re going to be fine. Then I just take it to be small alpha.
- Yeah. I think the way you check it is that, for example the way we check it here is that, so sometimes we have some referenced cell types, so that, you know roughly what you are doing. So here, for example,
here this reference cell types, these are, these are used in the modeling approach, so these are for evaluation the performance. And for some datasets, you can’t, we can mark our genes and do some class points, so, you know, roughly like which though, (indistinct)
and for two datasets, you can see, like whether you can correctly merge the cell types that are shared among the datasets but keep the cell types that are unique to different cells. Then for our trajectory inference is slightly complicated because these cell types are not well separated.
So another way that we can check our performance, is that you can see here that we can correctly, for these projected cells we can correctly like,
order the wrong days in the right order.
So that we know that we keep
some biological meaningful signals there.
I think there still can be some bias. Yeah.
Okay, great. Thanks.
So, thanks Jingshu again for this excellent talk.
And if you have any question you want to ask Jingshu
that you cannot think about for now,
you can always email her offline
and if you want to use her software,
I think she’ll be more than happy to answer your question.
Yes. Yes.
(chuckles)
So I guess that’s all for today.
Thank you everyone joining.
Thank you Jingshu for being here,
and have a nice remaining day.