22 February 2024

PAUSE: principled feature attribution for unsupervised gene expression analysis

Joseph D. Janizek[1,2], Anna Spiro[1], Safiye Celik[3], Ben W. Blue[4], John C. Russell[4], Ting-I Lee[4], Matt Kaeberlin[4,5] and Su-In Lee[1*]

Chanhee Lee

Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea

서울대학교 통계학과
생물정보통계연구실
BIBS

Outline

- Background
- Methods
  - ✓ Model architectures
  - ✓ Attributions
  - ✓ Datasets
- Results
  - ✓ Overview of PAUSE approach (result 1)
  - ✓ Pathway attributions : Identify major sources of variation & biologically relevant pathways (result 2-3)
  - ✓ Gene attributions : Increase latent node interpretability (result 4)
  - ✓ Biological Findings & Experimental validation : Mitochondrial Complex I as a potential AD therapeutic target (result 5-6)
- Discussion & Conclusion

Background

Interpreting High-Dimensional NGS data using Deep learning

- Deep learning approaches are capable of modeling complex systems with high fidelity
- However, they unfortunately share the drawback of lacking interpretability
- High-dimensional NGS → Low-dimensional Data (Deterministic AE, Variational AE)
  - Latent embeddings do not have inherent biological meaning.
  - Further methods are required to understand mechanisms captured by these models

**AE : Auto Encoder*

Background

Opening the ‘black box’ : Two competing trends

- First trend : *post hoc interpretability*
  - Aims to identify genes that were most important contributors to each latent variable learned by an autoencoder. (Integrated Gradients [12], Way et al. [14], Svensson et al. [15])
  - May be very difficult as autoencoders are known to learn entangled representations, meaning that each latent variable may capture multiple biological processes

Background

Opening the ‘black box’ : Two competing trends

- Second trend (recent) *: biologically-constrained modeling*
  - Aims to create models with latent spaces that are inherently interpretable
  - Use prior information to define sparse connections between input nodes corresponding to genes, and latent nodes corresponding to biological pathways (or other pre-defined groups).
  - Variety of recent works have proposed using biologically-constrained autoencoders to model gene expression data [20–22]

Background

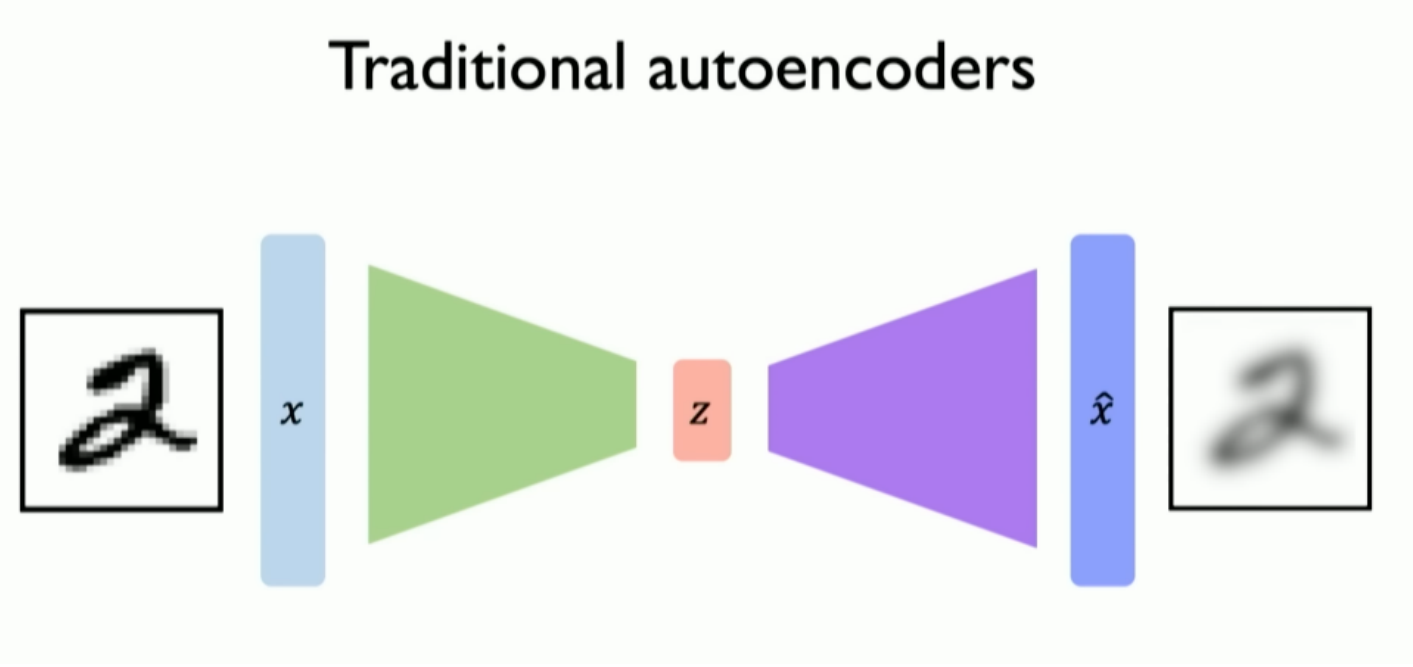
Aim

- Our paper aims to demonstrate that these two trends in biological interpretability are not mutually exclusive
- Principled attribution methods can improve the analysis of unsupervised models of gene expression analysis by quantifying the importance of pathways
- PAUSE: principled feature attribution for unsupervised gene expression analysis

Methods

Model architectures – Traditional Auto Encoder

- The weights in the encoder and decoder is fixed. (deterministic encoder and decoder)
- When input x_i is given, latent variable z_i and reconstructed data $\hat{x}_i$ always gets the same value



Loss function of AE : Reconstruction loss

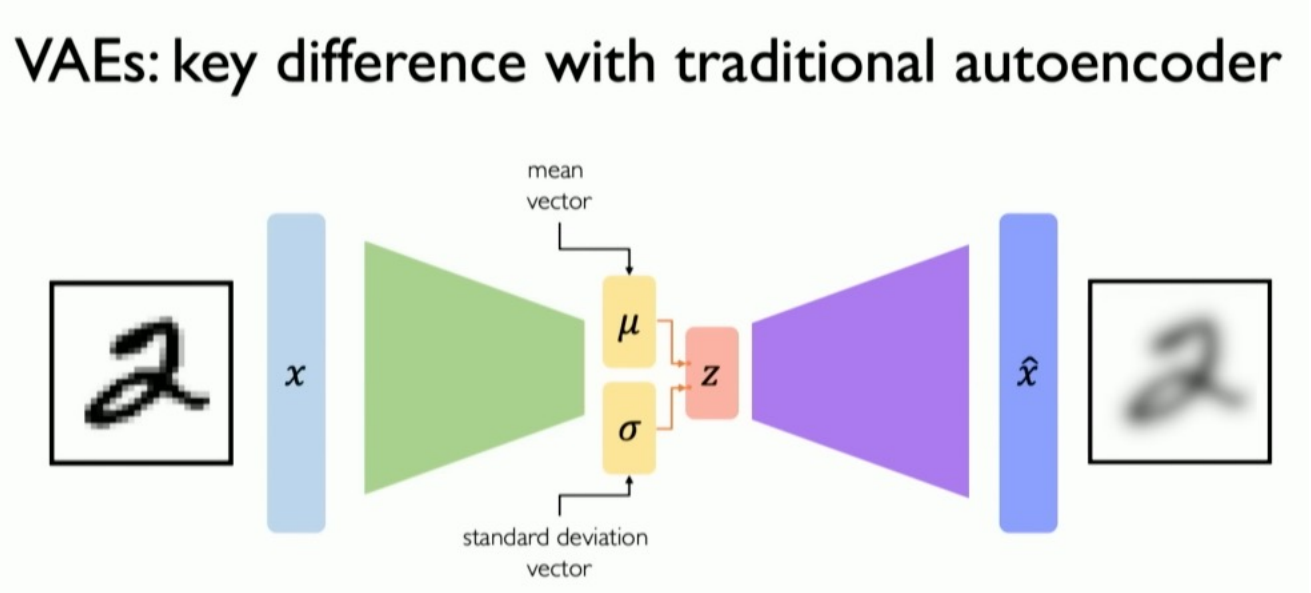
$$\mathcal{L}(x, \hat{x}) = \|x - \hat{x}\|^2$$

Loss function doesn't use any labels!!

Methods

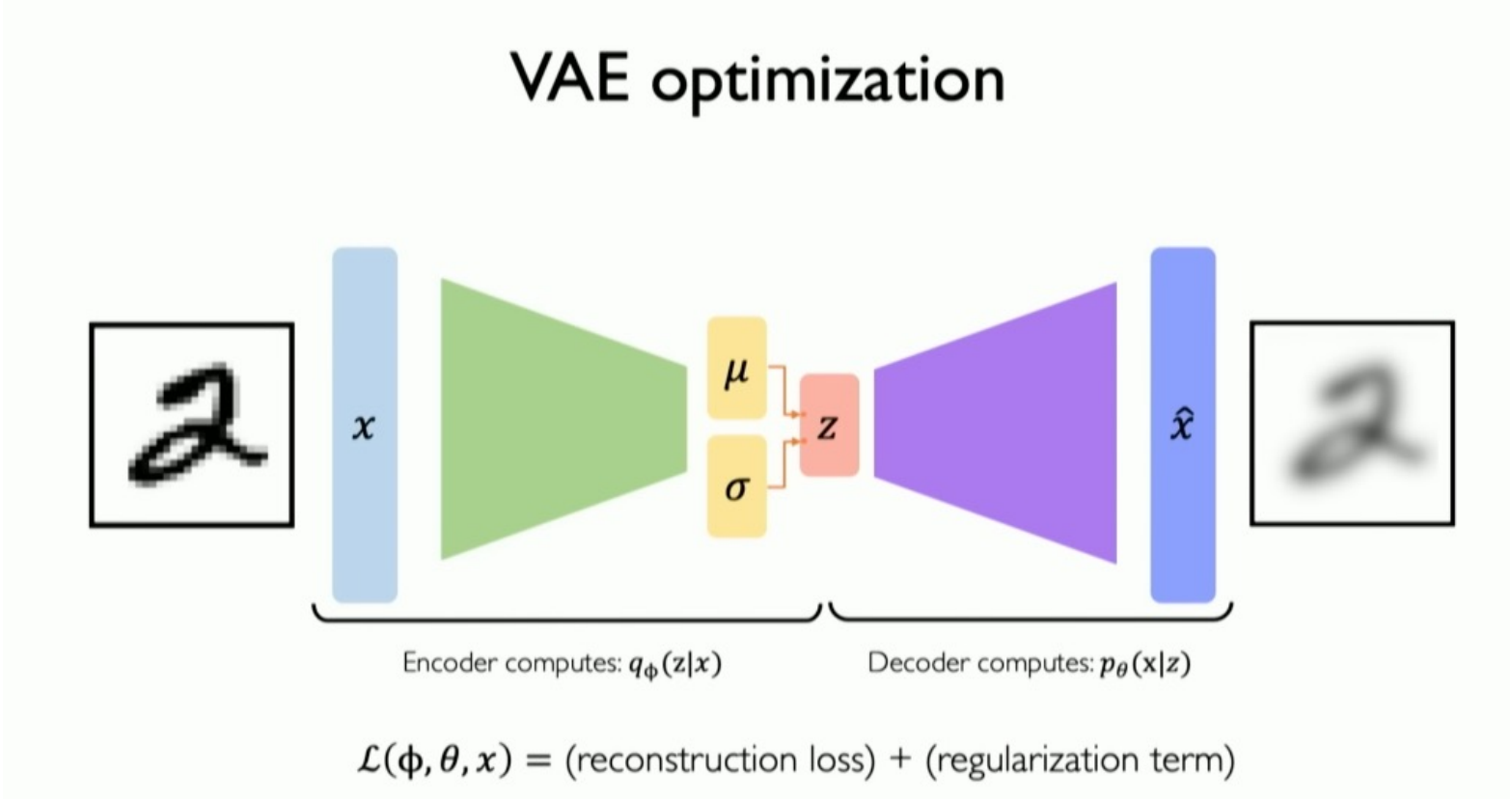
Model architectures – Variational Auto Encoder (VAE)

- VAE is probabilistic twist on autoencoders
- Sample from the mean and standard deviation to compute latent sample



Methods

Model architectures – Variational Auto Encoder (VAE)



$$\log p_\theta(x) \geq \mathbb{E}_{q(z|x;\theta)}[\log p(x|z;\psi)] - \mathbb{KL}(q(z|x;\theta)||p(z)) = -\mathcal{L}_{\text{ELBO}}. \quad (1)$$

Methods

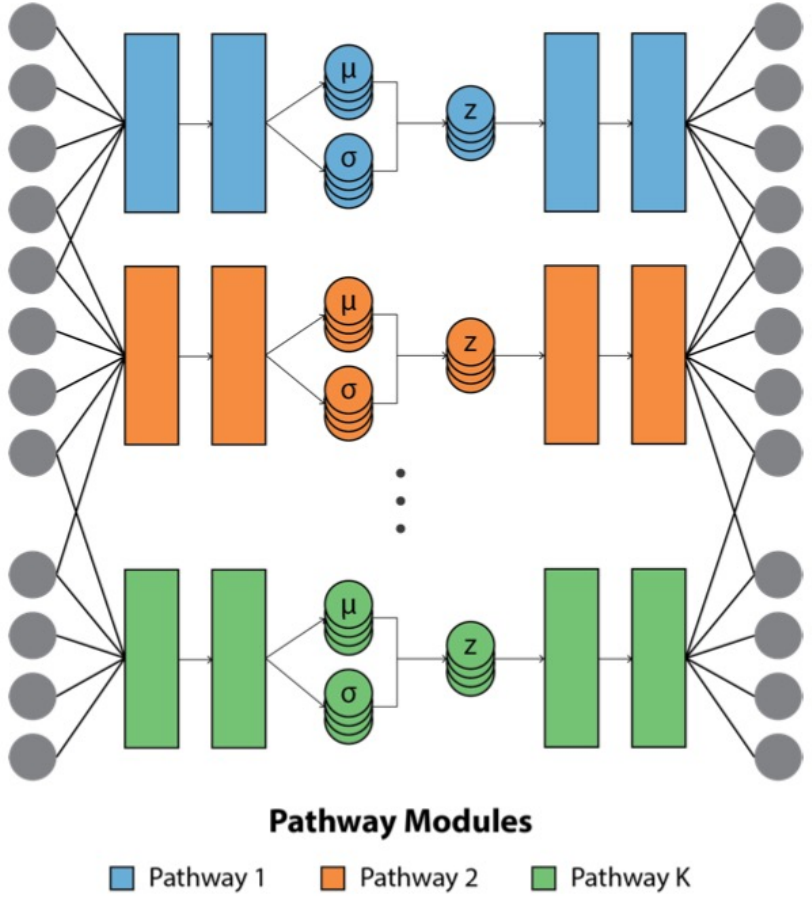
Model architectures – Variational Auto Encoder (VAE)

- Standard VAEs are a pair of neural networks, an encoder with parameters θ and a decoder with parameters ψ, optimized to learn a low-dimensional distribution over latent variables z from high-dimensional data x
- These networks are trained by stochastic gradient descent to maximize the variational lower bound on the likelihood of the data

Methods

Model architectures – pathway module VAE (pmVAE)

- the pmVAE approach uses sparse masked weight matrices to separate the weights of the encoder and decoder neural networks into non-interacting modules for each pathway.
- The network's first layer is masked with a binary assignment mask, which ensures that each gene is only connected with non-zero weight to the hidden nodes of the modules corresponding to its pathways.



Methods

Model architectures – conditional pathway module VAE (cpmVAE)

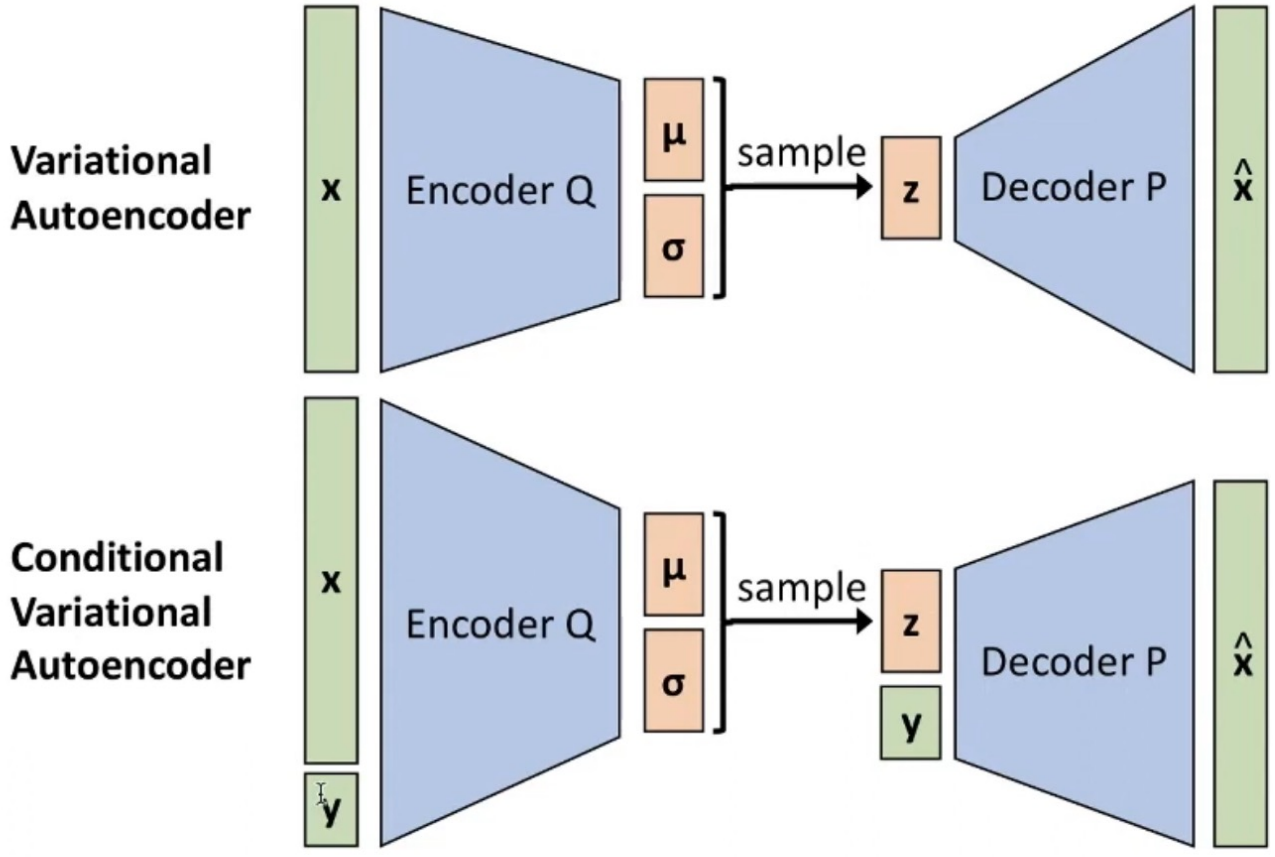
- In our experiments on bulk RNA-seq brain data, the brain gene expression samples were compiled from multiple data sources.
- In order to disentangle the batch effects, represented by a vector c, from the biological variation of interest, we wanted to train a conditional pathway module VAE (cpmVAE)
- maximize the conditional variational lower bound objective:

$$\log p_\theta(x|c) \geq \mathbb{E}_{q(z|x,c;\theta)}[\log p(x|c,z;\psi)] - \mathbb{KL}(q(z|x,c;\theta)||p(z|c)) = -\mathcal{L}^c_{\text{ELBO}}.$$

- These condition labels are passed to each pathway module in the encoder, and again to each pathway module in the decoder.

Methods

Model architectures – conditional pathway module VAE (cpmVAE)

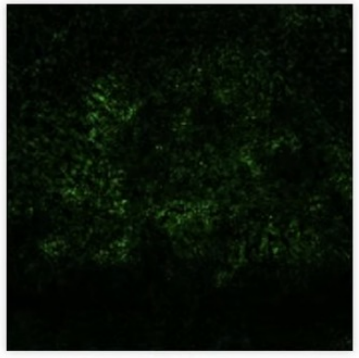


Methods

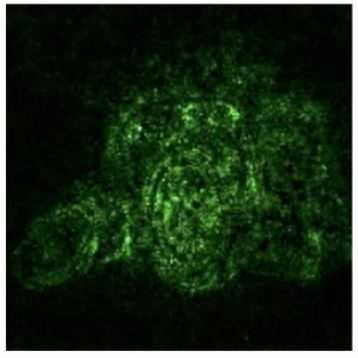
Attributions – Integrated Gradient

Why not just gradients? Saturation

Original Image
(Input)



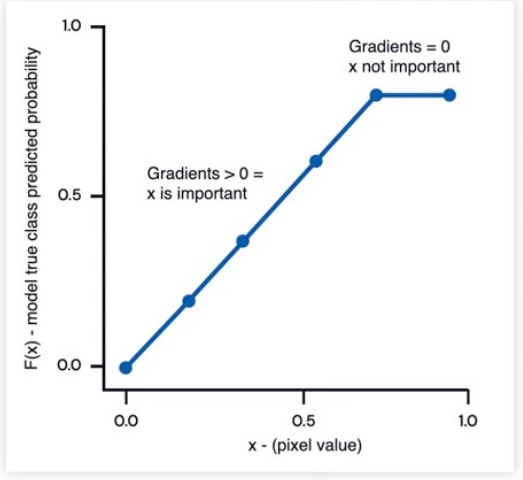
Vanilla
Gradients



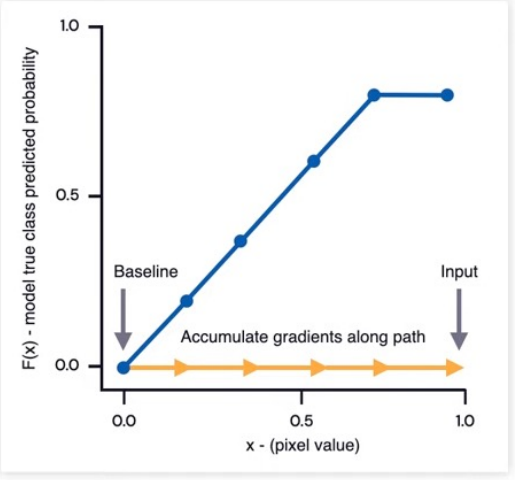
Integrated
Gradients

Intuition: How IG solves the gradient saturation problem

Gradients saturate over F(x)



IG intuition



Methods

Translating IG formulas to code

$$IntegratedGrads_i^{approx}(x) ::= \overbrace{(x_i - x'_i)}^{5.} \times \overbrace{\sum_{k=1}^{m}}^{4.} \frac{\overbrace{\overbrace{\partial F(x' + \overbrace{\frac{k}{m}}^{1.} \times (x - x'))}^{2.}}^{3.}}{\partial x_i} \times \overbrace{\frac{1}{m}}^{4.}$$

1. Generate alphas α
2. Generate interpolated images = $(x' + \frac{k}{m} \times (x - x'))$
3. Compute gradients between model F output predictions with respect to input features = $\frac{\partial F(interpolated\ path\ inputs)}{\partial x_i}$
4. Average integral approximation = $\sum_{k=1}^{m}$ gradients $\times \frac{1}{m}$
5. Scale integrated gradients with respect to original image = $(x_i - x'_i) \times$ integrated gradients
   The reason this step is necessary is to make sure that the attribution values accumulated across multiple interpolated images are in the same units and faithfully represent the pixel importances on the original image.

Methods

Attributions

- We propose a novel attribution method to understand which latent factors are important in unsupervised neural networks, based on Integrated Gradients and the Aumann-Shapley value
- To quantify the contribution of a particular latent node i

$$\phi_i^{\text{pathway}}(z) = (z_i - z_i') \times \int_{\alpha=0}^{1} \frac{\partial \ell(z' + \alpha(z - z'))}{\partial z_i} d\alpha,$$

- from local pathway importance to global pathway importance we can take the expected value of the local attributions over the samples in the original data:

$$\Phi_i^{\text{pathway}} = \mathbb{E}_{z \sim \mathcal{D}}[\phi_i^{\text{pathway}}(z)].$$

Methods

Attributions

- We propose a novel attribution method to understand which latent factors are important in unsupervised neural networks, based on Integrated Gradients and the Aumann-Shapley value
- To quantify the contribution of a particular gene node j for pathway k

$$\phi_j^{\text{gene,k}}(x) = (x_j - x_j') \times \int_{\alpha=0}^{1} \frac{\partial f_k(x' + \alpha(x - x'))}{\partial x_j} d\alpha,$$

- from local gene importance to global gene importance we can take the expected value of the local attributions over the samples in the original data:

$$\Phi_j^{\text{gene,k}} = \mathbb{E}_{x \sim \mathcal{D}}[|\phi_j^{\text{gene,k}}(x)|].$$

Methods

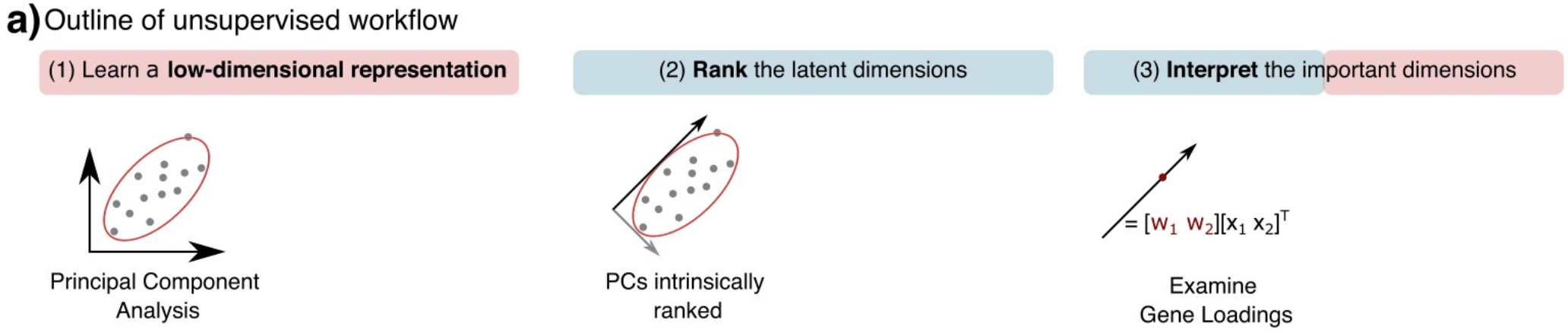
Datasets

- ***Peripheral blood mononuclear cells (PBMC) INF-β***
  - 13,576 samples with 979 genes. The dataset contains 6359 control and 7217 stimulated cells
- ***Intestinal epithelium***
  - 5,951 samples with 2,000 genes. The dataset contains 3240 control and 2711 *H.polygyrus*-stimulated cells
- ***Jurkat anti-CD3/anti-CD28***
  - 1288 samples with 2139 genes. This dataset contains 607 control and 681 stimulated cells
- ***Bulk brain expression datasets***
  - we used data from the ROSMAP, ACT, HBTRC, MAYO, and MSBB studies

Results

Overview of PAUSE approach

- It is first helpful to understand a representative unsupervised workflow
- Traditional Approach like PCA
  1. Learn a low-dimensional representation of gene expression data
  2. Rank the latent dimensions according to the amount of variance in the original data explained by each dimension
  3. Interpret the biological meaning of the most important dimensions.



Results

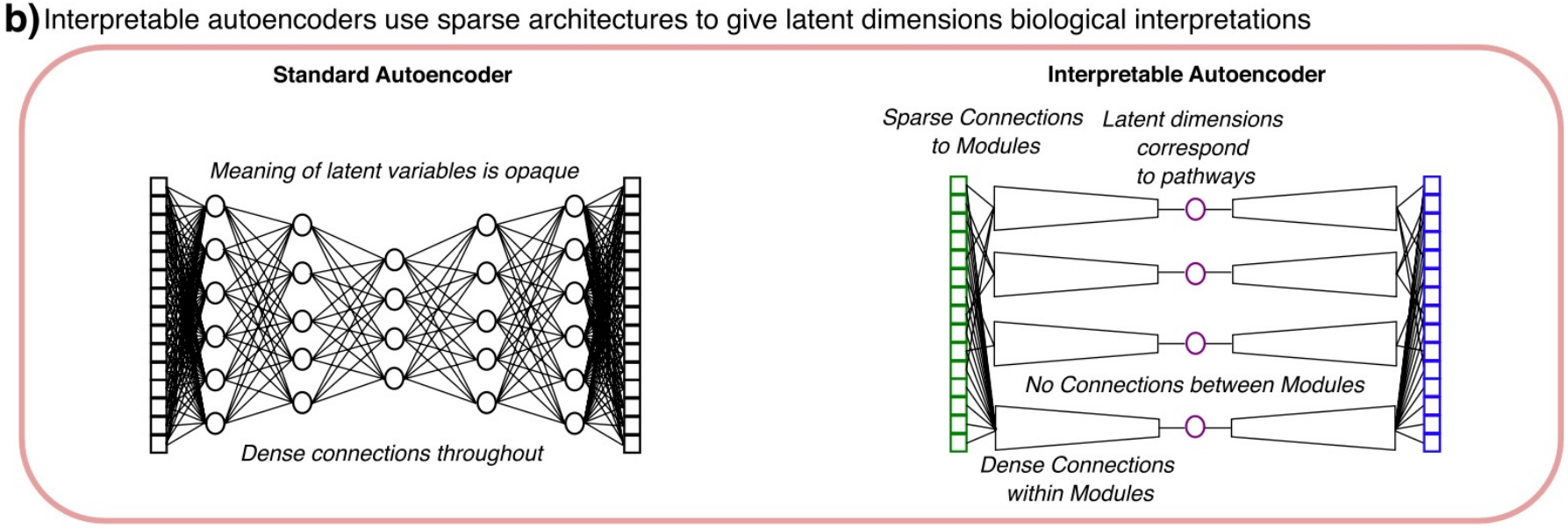
Overview of PAUSE approach

- While deep learning-based autoencoders are able to reconstruct gene expression with high fidelity, they fall short at steps 2. and 3. in the workflow described above.
- Interpretable Autoencoders : model that learns a latent representation with dimensions that correspond to biological pathways or functions.
  - improve step (3) for deep learning models by constraining the learned representation
  - hard constraint or soft constraint using regularization

Results

Overview of PAUSE approach

- While deep learning-based autoencoders are able to reconstruct gene expression with high fidelity, they fall short at steps 2. and 3. in the workflow described above.
- Interpretable Autoencoders : model that learns a latent representation with dimensions that correspond to biological pathways or functions.



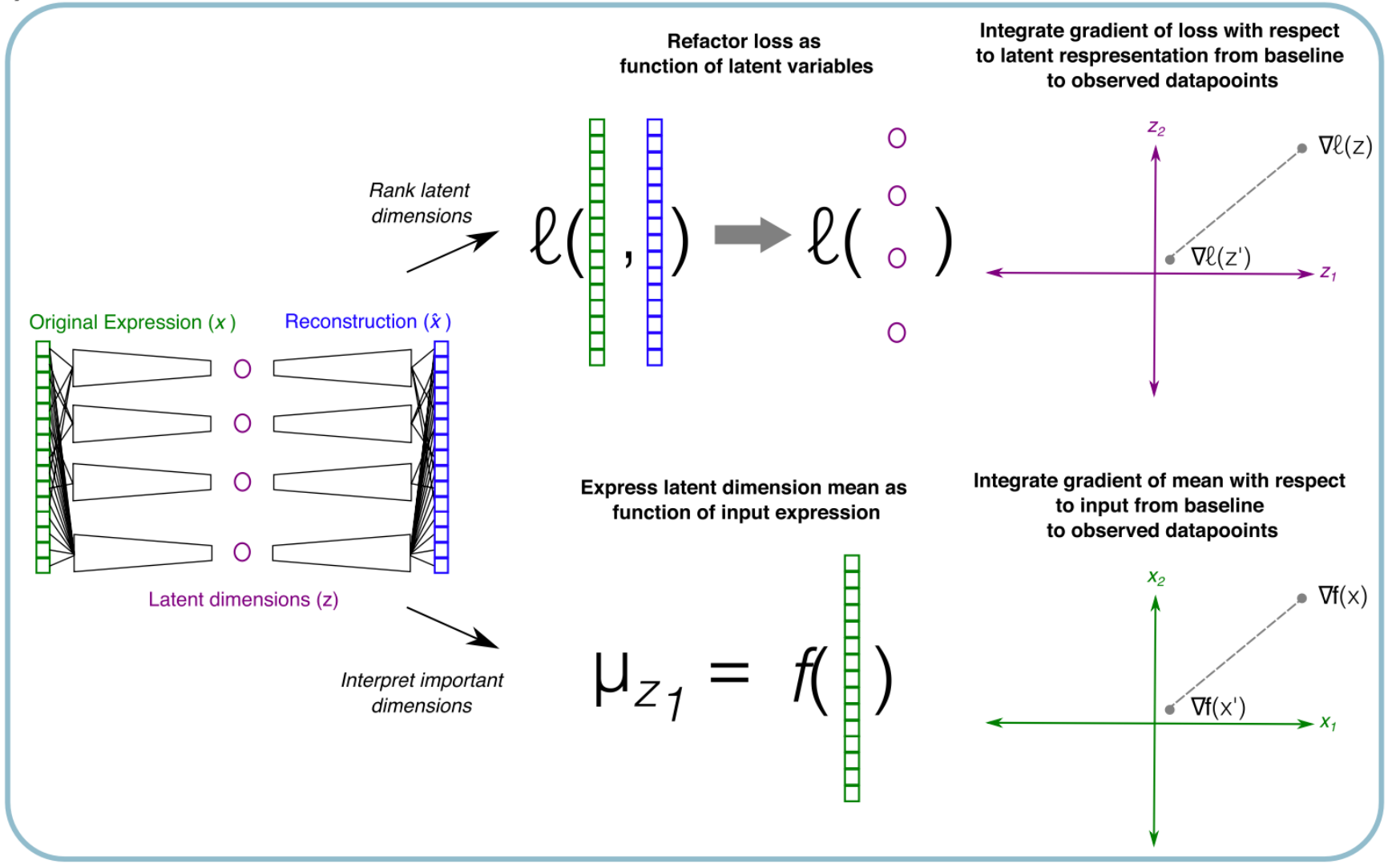
Results

Overview of PAUSE approach

- Interpretable autoencoder models have latent nodes corresponding to biological pathways, but there is no clear-cut way to identify which pathways are the most important in a dataset (step 2).
- Our approach, ***p***rincipled ***a***ttribution for ***u***n***s***upervised gene ***e***xpression analysis (PAUSE), aims to improve the utility of "interpretable" deep autoencoder models using techniques from the area of feature attribution
- Using approaches from game theory for credit allocation among players in cooperative games, we derive a novel pathway attribution that can be thought of analogously to the eigenvalues in PCA

Results

c) Principled attributions rank latent dimensions and aid in the interpretation of important dimensions



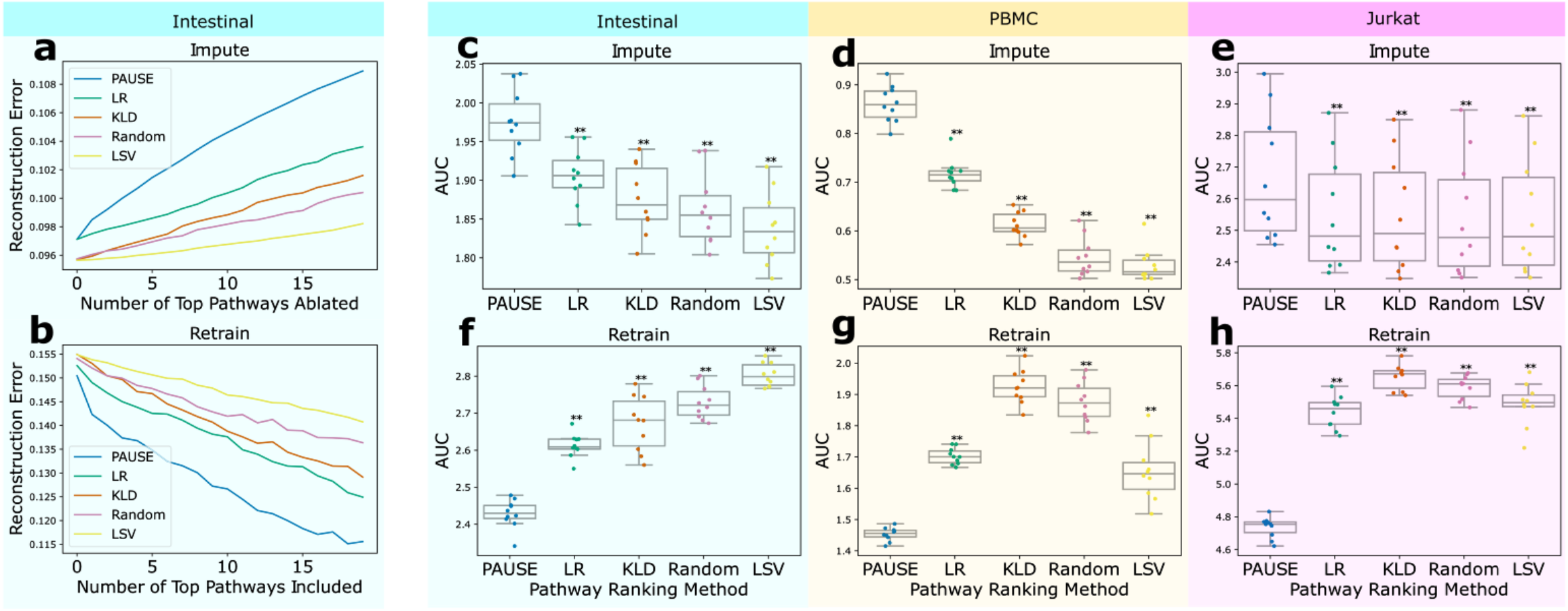
Results

Pathway attributions : Identify major sources of variation

- To empirically validate that the pathways identified, we adapt two benchmarks from the feature attribution literature
- ***imputation* benchmark** : measures the extent to which the reconstruction error of a trained biologically-constrained autoencoder model increases when the learned embeddings for each pathway are replaced with an uninformative mean imputation
- ***retrain* benchmark** : measures the extent to which pathways identified as important can reconstruct gene expression space when used to train a new model

Results

Pathway attributions : Identify major sources of variation



Results

Pathway attributions : Identify biologically relevant pathways

- We wanted to demonstrate that these major sources of transcriptomic variation corresponded to biologically interesting pathways.
- We therefore compared the top pathways found by our unsupervised approach to the top pathways according to a more conventional, supervised analysis

Results

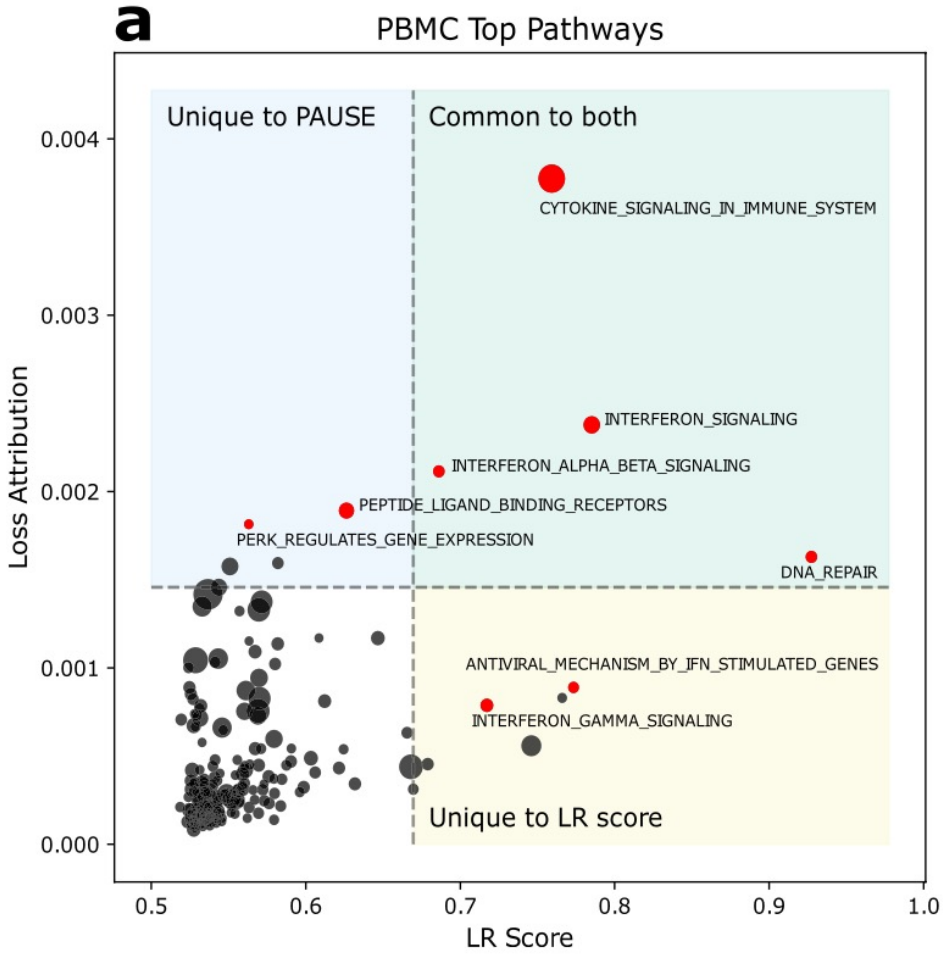
Pathway attributions : Identify biologically relevant pathways

- PBMC : untreated vs stimulated with interferon-β
  - pathways related to interferon-β (IFN-β) stimulation should be important
  - compare the pathways identified as important by our unsupervised analysis with the pathways identified by supervised analysis.
  - For a supervised metric of pathway importance, we considered the accuracy attained by a logistic regression model trained on the pathway latent nodes

Results

Identify biologically relevant pathways : PBMC

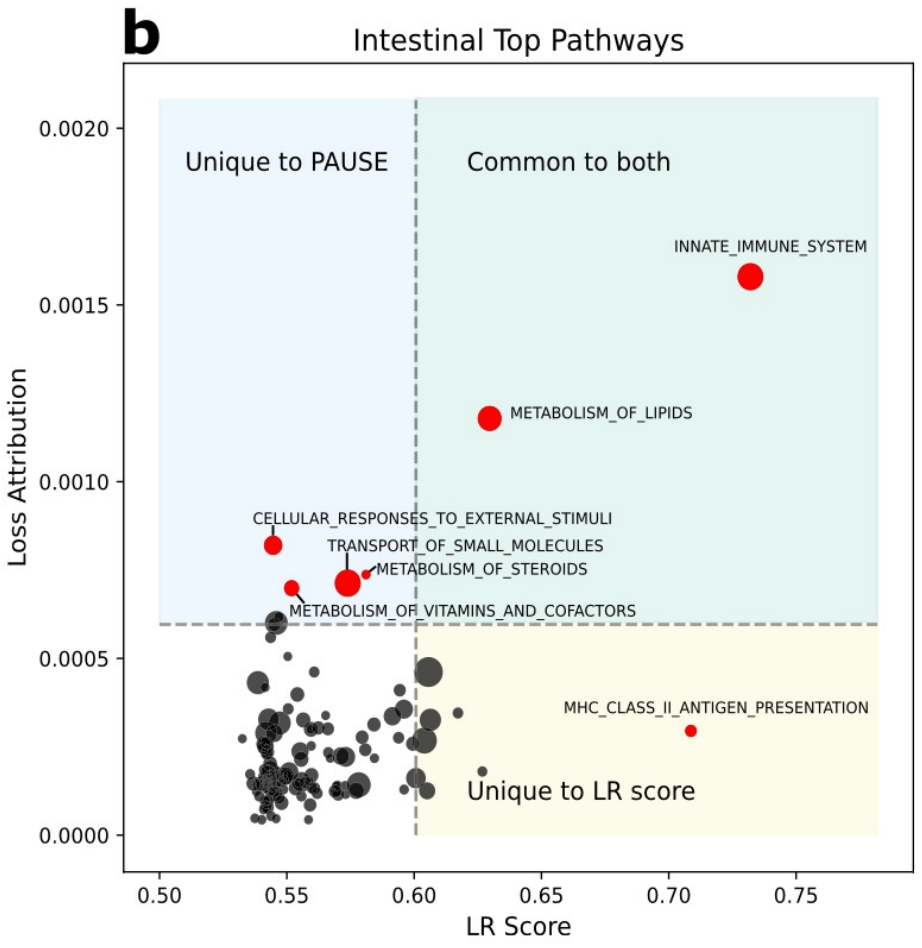
- We find that our fully unsupervised pathway attributions are able to identify many of the same pathways as supervised approaches
- More interestingly, we can also examine where the pathway rankings are discordant.



Results

Identify biologically relevant pathways : Intestinal

- mouse intestinal epithelial cells : untreated vs parasitic roundworm *Heligmosomoides polygyrus*
- Again, there is substantial overlap in the top pathways identified by our unsupervised approach and supervised metrics
- The fact that some pathways are highlighted by supervised attributions, but not by our unsupervised approach, shows that these two approaches should be considered complementary.

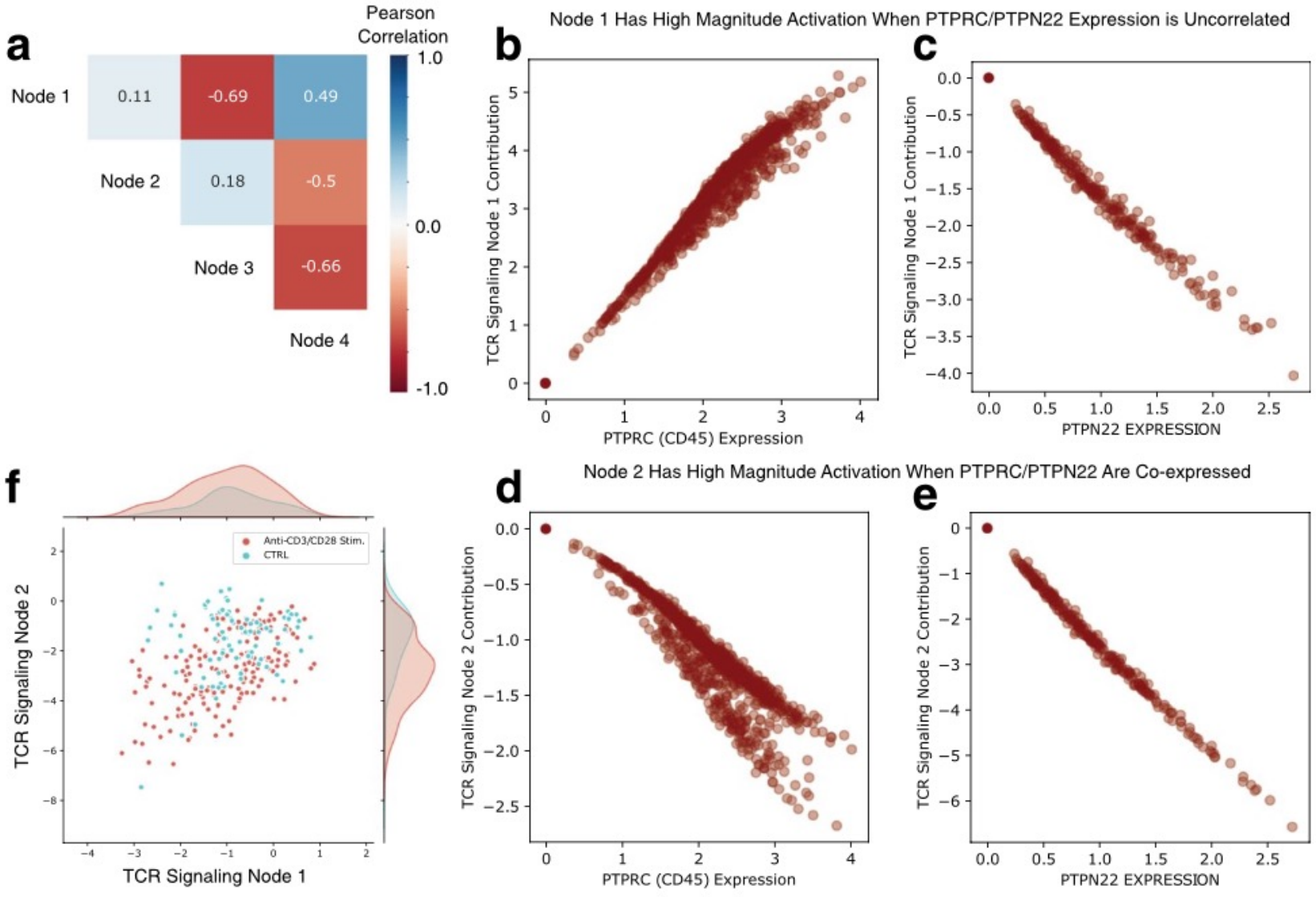


Results

Gene attributions : Increase latent node interpretability

- In addition to identifying important pathways, the interpretability of biologically-constrained models can be enhanced through the application of gene attribution values
- These gene attribution values help identify which gene expression values are important determinants of each learned pathway representation.
- We looked at the pairwise correlations between the activations of each of the four latent nodes in the TCR Signaling pathway module over all cells

Results


a
Pearson Correlation
1.0
0.0
-1.0
Node 1
Node 2
Node 3
Node 4
0.11
-0.69
0.49
0.18
-0.5
-0.66
b
Node 1 Has High Magnitude Activation When PTPRC/PTPN22 Expression is Uncorrelated
TCR Signaling Node 1 Contribution
PTPRC (CD45) Expression
c
TCR Signaling Node 1 Contribution
PTPN22 EXPRESSION
f
Anti-CD3/CD28 Stim.
CTRL
TCR Signaling Node 2
TCR Signaling Node 1
d
Node 2 Has High Magnitude Activation When PTPRC/PTPN22 Are Co-expressed
TCR Signaling Node 2 Contribution
PTPRC (CD45) Expression
e
TCR Signaling Node 2 Contribution
PTPN22 EXPRESSION

Results

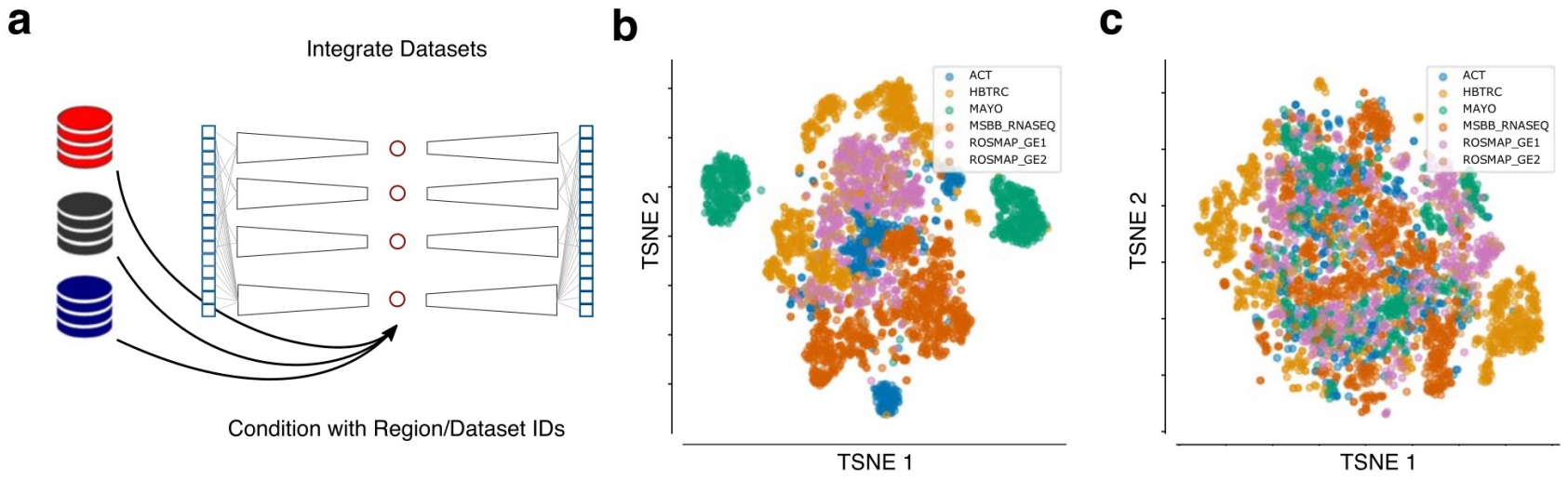
Biological Findings : mitochondrial oxidative phosphorylation in AD

- We wanted to use our PAUSE approach to gain insight into the biology of Alzheimer's disease (AD)
- Because of the diversity of data sources, it was essential to account for batch effects so that the embedding learned by our model represented true biological variation rather than technical artifacts
- When a standard pmVAE model is trained on the data, we see that the latent space separates samples predominantly according to the data source from which the samples were derived

Results

Biological Findings : mitochondrial oxidative phosphorylation in AD

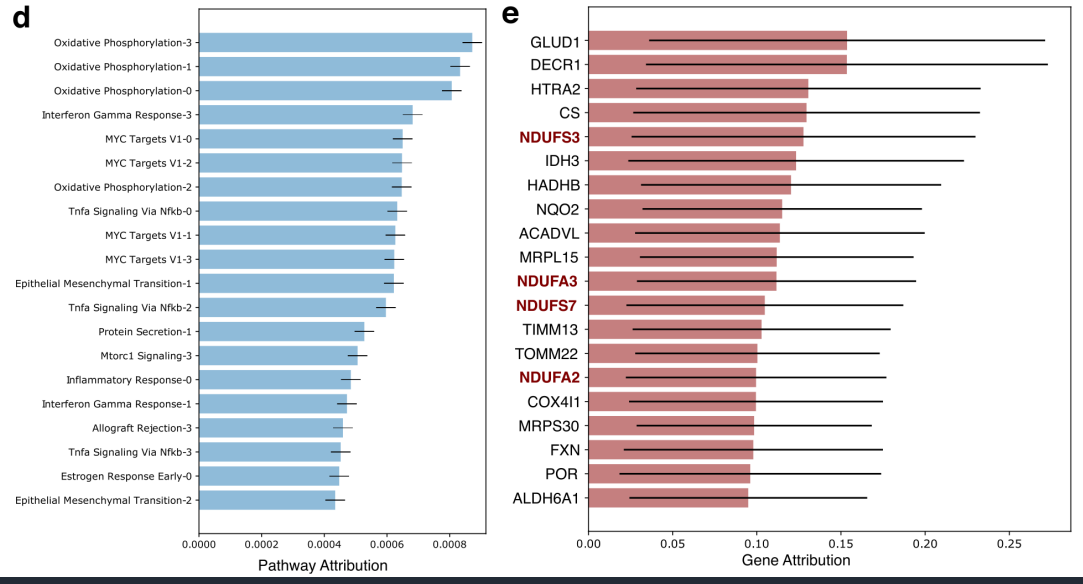
- To control for these dataset effects, we therefore modified the pmVAE architecture to be a conditional pathway module VAE (cpmVAE)
- In addition to the set of gene expression values corresponding to a particular pathway, each module takes as an additional input a vector of labels describing any unwanted sources of technical variation



Results

Biological Findings : mitochondrial oxidative phosphorylation in AD

- PAUSE pathway attributions showed Hallmark Oxidative Phosphorylation pathway was the single pathway explaining the most observed variation in the biological latent space
- When we look at the top 20 genes by gene attribution, we see four genes, NDUFS3, NDUFA3, NDUFS7, and NDUFA2, that are all part of mitochondrial respiratory Complex I, which is responsible for the oxidation of NADH in the mitochondria



Results

Biological Findings : mitochondrial oxidative phosphorylation in AD

- While primary defects in oxidative phosphorylation are not thought to be likely causes of AD, and mitochondrial dysfunction is in fact thought to be the result of Aβ and tau protein accumulation
- These changes in expression of oxidative phosphorylation genes likely play an important role in the pathophysiology of AD

Results

Experimental Validation: mitochondrial Complex I as a potential therapeutic target

- In order to verify that the genes and pathways identified by PAUSE are biologically-relevant, we had to experimentally validate our findings
- Used a special strain of the worm(C. elegans) that has been genetically modified to produce the human Aβ1-42 protein, which leads to symptoms similar to Alzheimer's in humans
- They then used a technique called RNA interference (RNAi) to "turn off" 13 worm genes that are similar to human genes involved in a cellular component called Complex I, which is part of the mitochondria
- Turning off these genes delayed the onset of paralysis in the worms, suggesting that these genes, and by extension, the human genes they're similar to, could play a role in the development of Alzheimer's disease.

Results

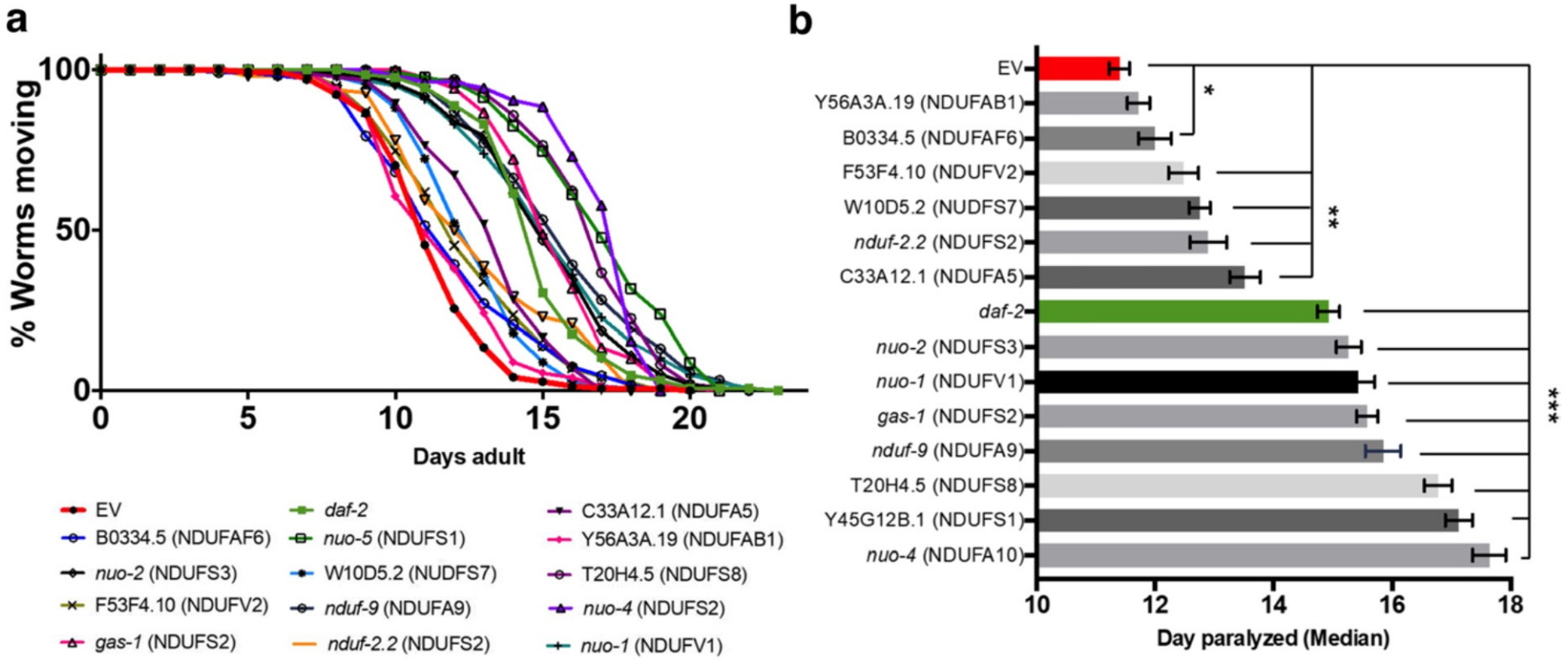


Fig. 6 *C. elegans* Aβ proteotoxicity assay validates importance of mitochondrial Complex I genes. **a** Paralysis curves for the reciprocal best orthologs of human Complex I genes. All tested RNAi conditions significantly suppress paralysis. **b** The same data as in **a** plotted as the median day of paralysis for each population of worms. Half of the conditions showed even stronger suppression than daf-2 RNAi conditions. Error bars indicate standard error of the mean across experiments. * p-value < 0.05. ** p-value < 0.01. *** p-value < 0.001

Discussion

Summary

- By combining principled attributions with pathway-based representations, our PAUSE approach enables the interpretable and fully unsupervised analysis of gene expression datasets
- In both the single cell and the bulk RNA-seq datasets analyzed, we found that the major sources of variation quantified by our pathway attribution approach corresponded to biologically-meaningful processes
- Confirmed by known ground-truth perturbations in the single cell analyses and by experimental validation in the AD dataset.

Discussion

Strengths

- Unlike existing methods, our pathway attributions do not depend on having labeled data to identify important pathways
- Furthermore, this approach can be applied to arbitrarily deep networks, unlike existing methods which depend on having a fully linear network or a linear decoder.

Discussion

Limitations

- Do not always capture all pathways that differ significantly between two groups of interest, if those pathways do not represent major sources of expression variation
- An assumption of our approach is that pathway databases provide good representations of the underlying biology of the system being modeled.

Conclusion

- Combining principled attributions with interpretable model architectures will prove to be a broadly useful strategy in domains beyond gene expression
- While biological pathways represent one useful representation, they will not necessarily be the only meaningful representation for all data types.
- In particular, translating interpretable approaches to multi-modal datasets will be an important future direction

References

- Gut G, Stark SG, Rätsch G, Davidson NR. PmVAE: Learning interpretable single-cell representations with pathway modules. bioRxiv. 2021.01.28.428664. https://doi.org/10.1101/2021.01.28.428664.
- **MIT 6.S191: Deep Generative Modeling**
- **Deep Learning Lecture 12.1 - Variational Autoencoder Extensions**
- **Introduction to Explainable AI (ML Tech Talks)**

Questions

💡 Thank you for listening. Feel free to ask any questions now 🙂