Principled feature attribution for unsupervised gene expression analysis

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

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington

[2]Medical Scientist Training Program, University of Washington

[3]Recursion Pharmaceuticals

[4]Department of Pathology, University of Washington

[5]Department of Genome Sciences, University of Washington

[*]Corresponding: suinlee@cs.washington.edu

Abstract

As interest in unsupervised deep learning models for the analysis of gene expression data has grown, an increasing number of methods have been developed to make these deep learning models more interpretable. These methods can be separated into two groups: (1) *post hoc* analyses of black box models through feature attribution methods and (2) approaches to build inherently interpretable models through biologically-constrained architectures. In this work, we argue that these approaches are not mutually exclusive, but can in fact be usefully combined. We propose a novel unsupervised pathway attribution method, which better identifies major sources of transcriptomic variation than prior methods when combined with biologically-constrained neural network models. We demonstrate how principled feature attributions aid in the analysis of a variety of single cell datasets. Finally, we apply our approach to a large dataset of post-mortem brain samples from patients with Alzheimer's disease, and show that it identifies Mitochondrial Respiratory Complex I as an important factor in this disease.

1 Background

As technologies to measure gene expression in biological samples have advanced over the last several decades, the tools and methods to extract biological meaning from these high-dimensional measurements have developed in parallel [1]. The increasing number of available samples in large transcriptomic compendia has enabled deep learning as a viable option for reference mapping [2], data integration [3], dimensionality reduction and visualization [4]. In particular, *unsupervised* machine learning approaches to solve these problems have become popular. For instance, tools like scVI use variational autoencoders (VAEs) to help analyze single cell RNA-seq data [5]. Similarly, variational autoencoders have been used to learn low-dimensional embeddings of bulk cancer RNA-seq data, improving the prediction of drug response [6]. Deterministic autoencoders have also been used in transcriptomic analyses, including the application of denoising autoencoders to large *Pseudomonas aeruginosa* and yeast datasets [7, 8]. While these deep learning approaches all are capable of modeling complex systems with high fidelity, they unfortunately share the drawback of lacking interpretability. Their latent embeddings do not have inherent biological meaning, and further methods are required to understand mechanisms captured by these models [9].

Two competing trends have emerged to "open the black box" and elucidate the biological mechanisms learned by these models, and have so far been framed as orthogonal approaches. The first trend could be described as *post hoc* interpretability, and aims to identify which genes were the most important contributors to each latent variable learned by an autoencoder. A variety of methods have been used to do this interpretation, and range from the use of feature attribution methods [10–12] to the direct analysis of weights in shallower autoencoders. For instance, Dincer et al. [13] apply the feature attribution method Integrated Gradients [12] to the latent variables of VAEs to identify the most important genes for each latent dimension, while Way et al. [14] measure the magnitude of each input weight to shallow, non-linear autoencoders. Likewise, in single cell analysis, Svensson et al. [15] modify a popular VAE architecture to have a linear decoder, allowing enrichment tests to be run using the magnitude of weights in that linear decoder. Despite the successes of these approaches, these types of analyses may be very difficult as autoencoders are known to learn entangled representations, meaning that each latent variable may capture multiple biological processes [16]. Furthermore, generating feature attributions for dozens or hundreds of latent variables may be computationally inefficient.

Another recent trend could be described as *biologically-constrained modeling*, and aims to create models with latent spaces that are inherently interpretable. These models 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). These biologically-constrained networks have been used in a supervised setting to improve the prediction of survival or treatment resistance from cancer gene expression or mutational status [17, 18], and to improve Genome Wide Association studies by aggregating the effects of single nucleotide polymorphisms (SNPs) into SNP sets [19]. In particular, a variety of recent works have proposed using biologicaly-constrained autoencoders to model gene expression data [20–22].

This paper aims to demonstrate that these two trends in biological interpretability are not mutually exclusive, and that principled attribution methods improve the analysis of interpretable, unsupervised models of gene expression analysis by quantifying the importance of pathways. We first propose a novel, fully-unsupervised attribution method and demonstrate how it can be used to identify important pathways in transcriptomic datasets when combined with biologically-constrained autoencoders. This allows for *fully unsupervised* analysis of gene expression data. We then show how existing, feature-level attribution approaches still provide useful information for annotated, unsupervised models. Finally, we apply our approach to a large transcriptomic compendium of post-mortem brain data from patients with Alzheimer's disease and demonstrate how our approach can identify Mitochondrial Respiratory Complex I as a potential drug target for this disease.

2 Results

2.1 Overview of PAUSE approach

Our approach, **p**rincipled **a**ttribution for **un**supervised gene **e**xpression analysis (PAUSE), aims to improve the utility of "interpretable" deep autoencoder models using techniques from the area of feature attribution. By "interpretable" autoencoder model, we refer to any model that learns a latent representation with dimensions that correspond to biological pathways or functions (Fig. 1a). This encompasses models trained with special regularization [20], as well as models with biologically-defined architectures [22]. We focus on two types of attributions for these models: *pathway-level* attribution (Fig. 1b) and *gene-level* attribution (Fig. 1c).

While interpretable autoencoder models have latent nodes corresponding to biological pathways, there is no clear-cut way to identify which pathways are the most important in a dataset. For Principal Components Analysis (PCA), a classical approach to unsupervised gene expression analysis, the eigenvalues of each principal component indicate how much variance in the original expression space is explained by that component. While autoencoders learn latent variables (or distributions, in the case of variational autoencoders), there is no obvious correspondence that reveals how much variance is explained by each latent dimension. 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 eigenvalue in PCA, in the sense that this attribution value shows how much variance in the original gene expression space is explained by each latent pathway. By posing the reduction in reconstruction error as the reward to be allocated in a cooperative game in which the pathways are the players (see Methods for more details), solution concepts like the Shapley value [23] or Aumann-Shapley value [24] can be used to provide **pathway attributions**.

Additionally, although the latent factors of biologically-constrained autoencoders are more interpretable than the latent factors of standard autoencoders, in the sense that they represent the expression only of genes within their annotated pathway, these latent variables are often not fully interpretable on their own. For example, the pathway module variational autoencoder (pmVAE) architecture [22] learns multiple latent variables for each pathway, which helps to encode more biologically relevant information. However, this approach provides no way for understanding the differences between the different nodes within a given pathway module. Are these latent variables sub-pathways supported by different genes with different expression patterns? Furthermore, if there are important pathways in an expression dataset that are not represented in the knowledge base used to define the model architecture, can we still identify these pathways? We propose generating **gene attributions**, again using feature attribution techniques based on the Shapley or Aumann-Shapley value, to identify the genes contributing to important learned pathway embeddings, or to identify the genes contributing to densely connected auxiliary pathways.

2.2 Pathway attributions accurately identify major sources of variation

To empirically validate that the pathways identified by our pathway attributions correspond to the most important sources of variation in real gene expression datasets, we adapt two benchmarks from the feature attribution literature [25]. The first is an *imputation* benchmark, and 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. The better an attribution method ranks important pathways, the more quickly the reconstruction error will increase. The second benchmark is a *retrain* benchmark, and measures the extent to which pathways identified as important can reconstruct gene expression space when used to train a new model. Attributions that do a better job of ranking pathways will decrease the reconstruction error more quickly.

Previous approaches to identifying important pathways in autoencoder models have been limited in a variety of ways. For example, previous papers have used supervised metrics like logistic regression accuracy or Bayes factors to
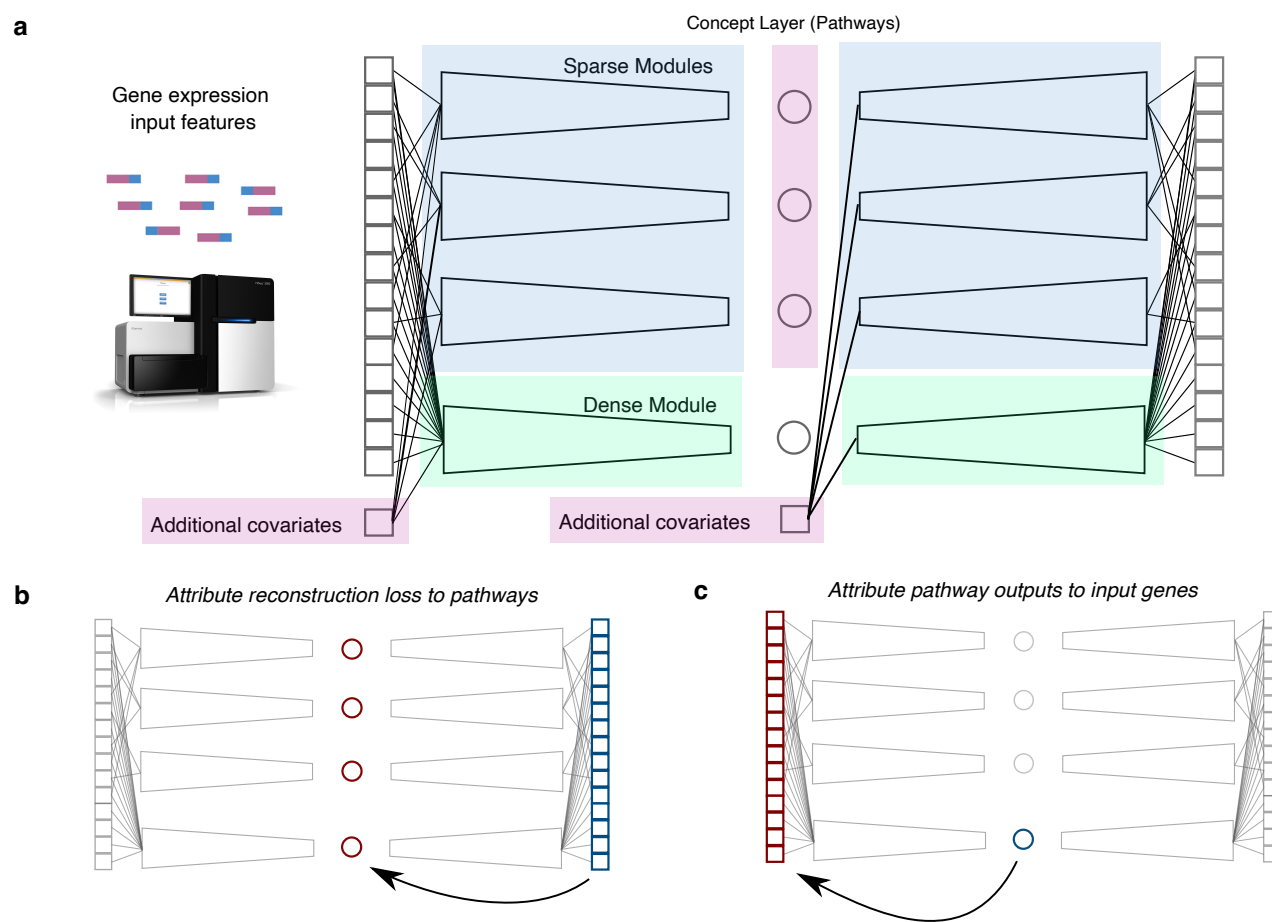
Fig. 1 | Biologically structured networks combine with game-theoretic attributions to create more interpretable unsupervised models. While inherently interpretable models and post-hoc interpretability methods like feature attribution have traditionally been framed as orthogonal approaches to creating biologically meaningful models, these approaches are in fact mutually beneficial. Our approach, **p**rincipled **a**ttribution for **u**n**s**upervised gene **e**xpression analysis (PAUSE), combines biologically-constrained autoencoders with principled attributions to improve the unsupervised analysis of gene expression data. **a**, Biologically-constrained "interpretable" autoencoders use prior knowledge to define sparse connections in a deep autoencoder, such that latent variables correspond to the activity of biological pathways. **b**, We propose a novel unsupervised pathway attribution method to identify the most important pathways for a given dataset, leading to an intuitive quantification corresponding to the amount of variance explained by the pathway. **c**, Gene attributions increase the interpretability of latent pathways, identifying distinct sub-pathways and finding biological pathways not present in the prior constraints.
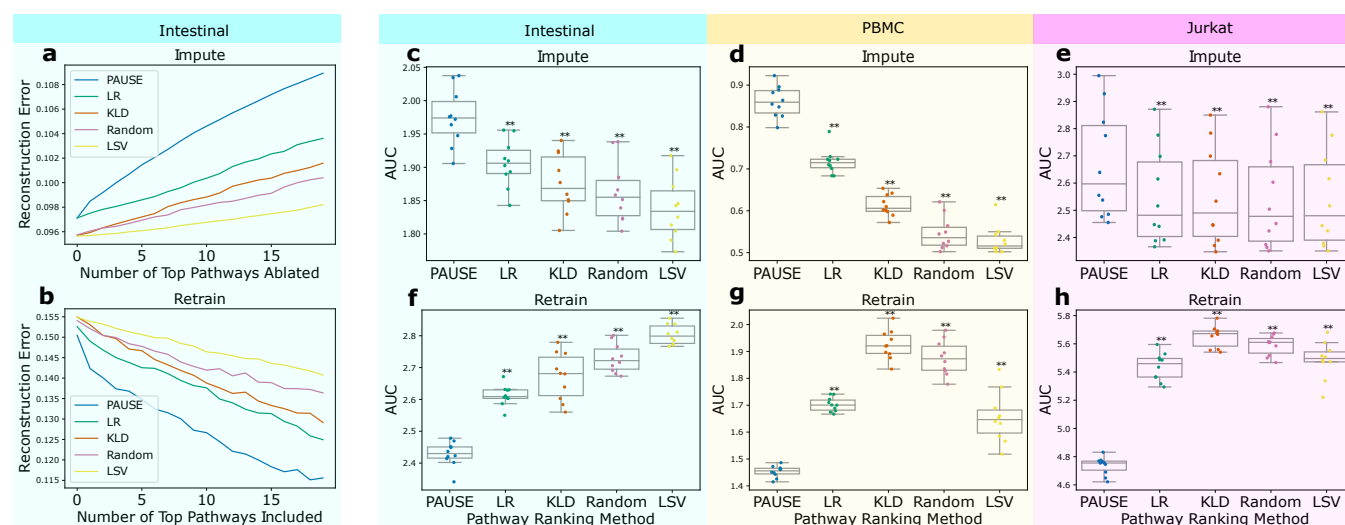
Fig. 2 | PAUSE pathway attributions accurrately identify the major sources of variation in single cell datasets. To verify that our novel pathway attribution method ("PAUSE" in the plots above) is capable of identifying the major sources of transcriptomic variation across a variety of datasets, we apply two benchmarks of pathway identification. The first (**a**), termed our *impute* benchmark, measures how much the reconstruction error of a biologically-constrained autoencoder model increases as important pathways are replaced with an uniformative, imputed baseline. Better methods will increase the error faster, leading to a larger area under the curve (AUC). The second benchmark (**b**), termed our *retrain* benchmark, measures how well a model can reconstruct the observed expression when retrained using only the most important pathways. Better methods will decrease the error faster, leading to a smaller AUC. The AUCs for both the impute (**c-e**) and retrain (**f-h**) benchmarks are shown for ten separate train/test splits across three separate single cell gene expression datasets (intestinal cells, peripheral blood monocytes, and Jurkat cells). In each experiment, the PAUSE loss attribution method significantly outperforms other methods. The other methods shown here are logistic regression score (LR), Kullback–Leibler divergence (KLD), random ranking, and latent space variance (LSV). The boxes in **c-h** mark the quartiles (25th, 50th, and 75th percentiles) of the distribution, while the whiskers extend to show the minimum and maximum of the distribution (excluding outliers).
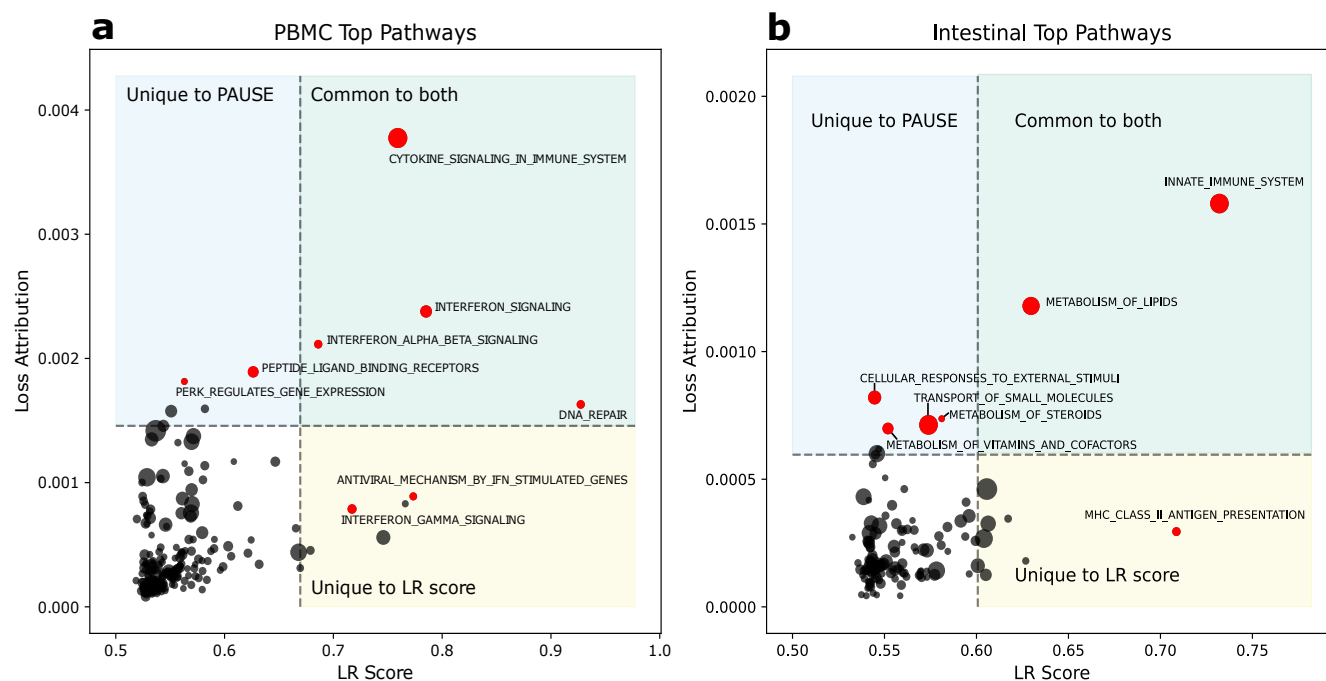
Fig. 3 | PAUSE loss attribution reveals biologically relevant pathways that overlap with pathways found using a supervised approach. Comparison of top pathways found using PAUSE loss attribution (y-axis, unsupervised) and logistic regression (LR) score (x-axis, supervised) in PBMC (**a**) and Intestinal (**b**) datasets. Gene sets in the blue rectangles are those ranked in the top 10 only by PAUSE, those in the yellow rectangles are ranked in the top 10 only by LR score, and those in the green rectangles are ranked in the top 10 by both methods. The size of the dots indicates gene set size.

identify important pathways [20, 22]. While these approaches have been used to successfully identify important pathways, calculating these metrics depends on having labeled data, which means that this analysis is not truly unsupervised. In terms of fully unsupervised approaches, Higgins et al. [16] have proposed finding important latent factors of variation by measuring the Kullback–Leibler (KL) divergence between the learned latent distribution in VAEs and the prior distributions. While this approach does not require labeled data, it can only be applied to VAE models, and can not be applied to standard, deterministic autoencoders. Another fully unsupervised approach to ranking important pathways examines the L2 norm of the weights connecting each latent pathway to the reconstruction output; however, this metric is obviously limited to models with linear decoders [21]. Unlike all of these approaches, our proposed pathway attribution (see Methods) is both fully unsupervised, meaning it requires no labeled data, and model agnostic, meaning it can be applied to any model regardless of architecture or implementation details.

The three datasets used in the two benchmarks were selected because they have labels corresponding to known perturbations applied to the cells, which allowed us to compare the information gleaned from our fully unsupervised analysis to that highlighted by supervised methods. These datasets have also been extensively analyzed in previous, related work, and include a dataset of peripheral blood monocytes (PBMC) stimulated with interferon-β [26], a dataset of intestinal epithelial cells stimulated with *Salmonella enterica* and *Heligmosomoides polygyrus* [27], and a dataset of Jurkat cells stimulated with a combination of anti-CD3/anti-CD28 antibodies to induce T cell activation [28] (see Methods for more details, including data preprocessing). The particular model architecture trained for this benchmark was a pathway module VAE (pmVAE), which is a sparse variational autoencoder model with deep, non-linear encoders and decoders [22].

Across both benchmarks and all three datasets, our pathway attribution method consistently identified the most important sources of dataset-wide transcriptomic variation better than previous approaches (two-sided Wilcoxon signed rank test, p=1.95×10^{-3}, statistic=0.0). Remarkably, we also found that our pathway loss attributions outperformed directly examining the L2 norm of the decoder weights in an architecture with a linear decoder (see Supplementary Fig. 1). In conclusion, our benchmark demonstrates that our pathway attribution method is able to accurately identify the major sources of transcriptomic variation across different datasets and autoencoder architectures, showing the benefit of using prinicpled attributions in conjunction with biologically-constrained models.

2.3 Pathway attributions identify biologically relevant pathways

After demonstrating in the previous section that our pathway attributions were capable of accurately identifying the major sources of transcriptomic variation in three scRNA-seq datasets, 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 (Fig. 3).

We first considered a biologically-constrained VAE (pmVAE, see Methods) trained on a dataset of peripheral blood mononuclear cells (PBMCs), where the PBMCs were either untreated as controls or stimulated with interferon-β [26]. The latent nodes of this model were defined using Reactome pathways [29]. Because this is a well-studied dataset with a known ground truth perturbation, we know that pathways related to interferon-β (IFN-β) stimulation should be important, and because we have control/stimulated labels for the cells, we can 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 to differentiate stimulated and controlled cells [22].

We find that our fully unsupervised pathway attributions are able to identify many of the same pathways as supervised approaches, without requiring access to labeled data (See Fig. 3a). For example, we see that the Reactome pathways "Cytokine Signaling in Immune system," "IFN Signaling," and "IFN alpha/beta signaling," are in the list of top ten pathways identified by both approaches. More interestingly, we can also examine where the pathway rankings are discordant. For example, "IFN Gamma Signaling" is a top pathway according to the supervised approach, but not according to our unsupervised rankings. This demonstrates why our unsupervised approach can complement the supervised approach even when labels are present. Although certain genes are regulated by both type I and type II IFNs, other genes are expected to be selectively regulated by either type I or type II IFNs [30]. Looking at our unsupervised attributions shows that while there may be significant enough differences in expression of genes in the "IFN Gamma Signaling" pathway to differentiate stimulated and unstimulated cells, this pathway is not one of the major sources of variation in the dataset, at least when compared to pathways more directly related to IFN-β signaling. Therefore, in addition to being useful when labeled data is not present, our unsupervised attributions can provide additional utility even when labels are present by explicitly quantifying the amount of variance in the observed data explained by each pathway.

We next considered a VAE (again with latent modules defined by Reactome pathways) trained on a dataset of mouse intestinal epithelial cells where control cells were untreated, and stimulated cells had been exposed to the parasitic roundworm *Heligmosomoides polygyrus* [27]. Again, there is substantial overlap in the top pathways identified by our unsupervised approach and supervised metrics (Figure 3b). For example, the Reactome pathway "Innate Immune Response" is the most important pathway according to both supervised and unsupervised attributions. Next, we again can look to see where these approaches are discordant in their rankings.

Examining the top ten pathways according to supervised attributions, we see that the pathway "MHC Class II Antigen Presentation" is a top pathway, despite not being highly ranked by our unsupervised metric (Figure 3b). This pathway is particularly biologically plausible as an important process in this dataset, as the MHC Class II complex is constitutively expressed in the upper villi of the small intestine, indicating a potential role as non-professional antigen presenting cells for intestinal epithelial cells [31]. Previous work has shown that in both epithelial cells and in other cell types, expression of the MHC Class II complex is regulated in response to pathogen exposure [32, 33]. The fact that this pathway is highlighted by supervised attributions, but not by our unsupervised approach, shows that these two approaches (supervised and unsupervised) should be considered complementary. While our PAUSE attributions can help identify the pathways responsible for explaining the largest quantities of variation in a given dataset, supervised attributions can help identify specific differences in expression between known groups.

2.4 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. For example, previous work has analysed the expression of Jurkat cells, an immortalized human T lymphocyte cell line, when stimulated with anti-CD3 and anti-CD28 antibodies to induce T cell activation [28]. This prior analysis trained a pathway module VAE (pmVAE) to reconstruct the expression of these cells, and found that having multiple nodes in each pathway module led to embeddings that were more discriminative for T cell activation [22]. This raises the question of whether multiple nodes may be necessary in order to distinguish different processes of the biology of T cell activation.

We therefore generated *gene attributions* to understand the expression programs represented by the multiple nodes present within given pathway modules. The two most important genes for each of the four latent nodes in the TCR signaling module were PTPRC and PTPN22 (Supplementary Fig. 2). PTPRC enocdes a tyrosine phosphatase commonly known as CD45 antigen, while PTPN22 codes for another protein tyrosine phosphatase commonly known as PEP. These genes are known to play an important, albeit complex, role in proximal TCR signaling, regulating the activity of Src family
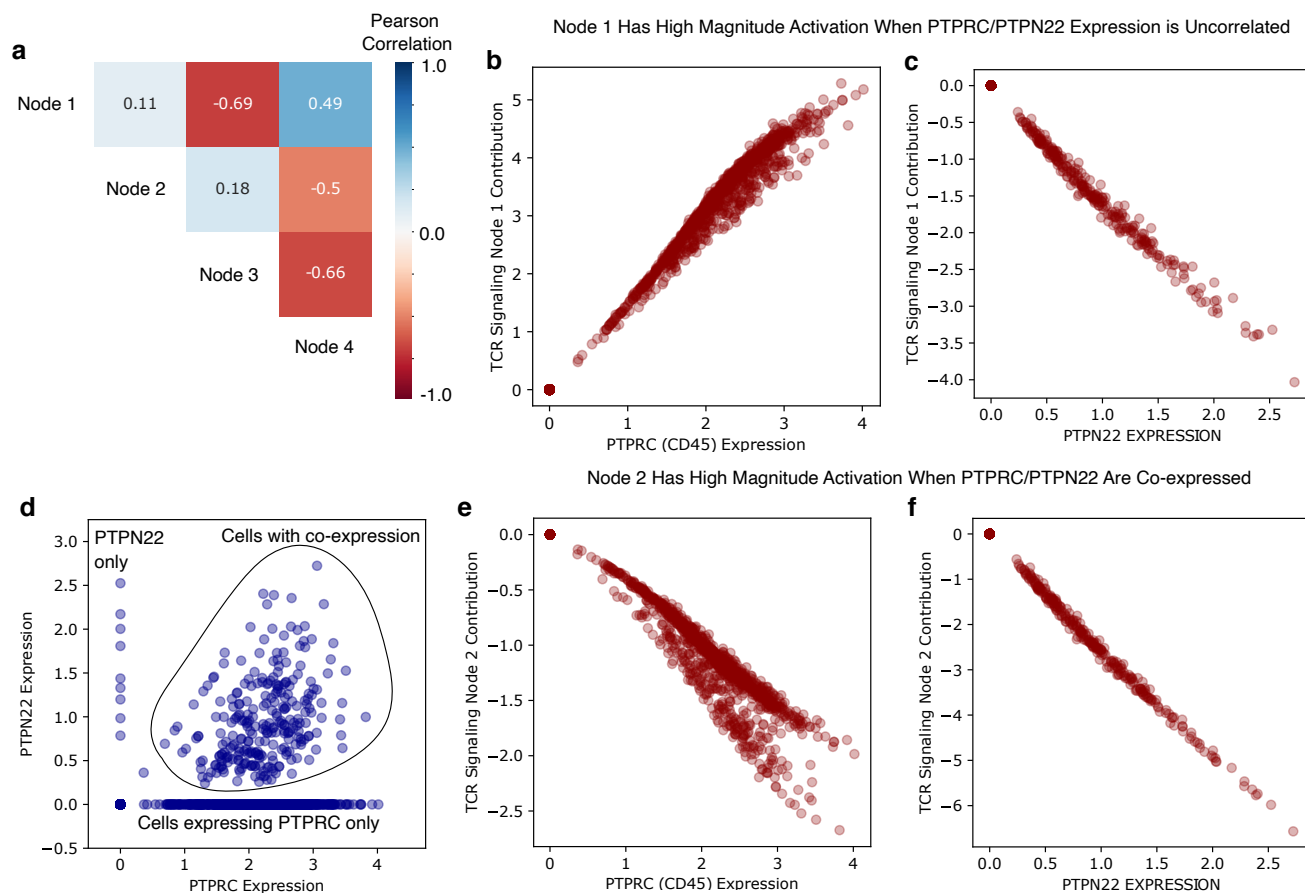
Fig. 4 | Gene attributions increase latent variable interpretability and differentiate TCR signaling pathway expression programs. Gene level attributions help to gain a deeper understanding of the expression programs represented by the latent variables of a biologically-constrained autoencoder trained on a dataset of Jurkat cells stimulated with anti-CD3 and anti-CD28 antibodies. **a,** Pairwise correlations between the four latent variables in the pathway module corresponding to T cell receptor signaling. **b-c** Gene attribution dependence plots for the two most important genes for TCR Signaling Latent Node 1. **e-f** Gene attribution dependence plots for the two most important genes for TCR Signaling Latent Node 2. **d** Jurkat cells plotted by their PTPRC and PTPN22 expression levels display distinct sub-populations, including cells where PTPRC and PTPN22 are co-expressed, and cells where only one or the other is expressed.

kinases downstream of T cell receptor activation [34].

Looking at the pairwise correlations between the activations of each of the four latent nodes in the TCR Signaling pathway module over all cells (see Fig. 4a), we see that while many of the pairs of nodes have a high degree of correlation or anti-correlation (potentially indicating redundancy between these nodes), Node 1 and Node 2 have a relatively low correlation (Pearson's R = 0.11). We can plot gene attributions for the top two genes for Node 1 (Fig. 4b,c) and Node 2 (Fig. 4e,f) in order to understand what precisely is being represented by the different nodes of the pathway module.

We see that Node 1's output is high when PTPRC's expression is high (Fig. 4b), and low when PTPN22 expression is high (Fig. 4c). This means that in cells where PTPRC and PTPN22 are co-expressed, the contributions of these two genes will cancel out and the latent node's activation will have a low magnitude. Node 2's output, in contrast, is highly negative when PTPRC expression is high (Fig. 4e) and highly negative when PTPN22 expression is high (Fig. 4f). Directly examining the expression values of PTPRC and PTPN22 across the cells in this dataset, we see that there are distinct populations of cells where PTPRC and PTPNN2 are coexpressed, and other populations where they are not (Fig. 4d). This example represents a common situation in biology – the genes in a particular annotated pathway are not perfectly correlated and co-expressed as a single block across all cell types. Rather, the pathway comprises a variety of distinct subprograms. The utility of gene attributions is in allowing multiple nodes to be modeled for each pathway in a fully deep and non-linear manner. After modeling, gene attributions can then be used to help us understand what the specific sub-processes a model has learned to represent within a particular pathway.

In addition to differentiating sub-processes within an annotated pathway module, another use for gene attributions is identifying biological processes learned by unannotated, densely-connected latent nodes. In addition to latent nodes corresponding to annotated biological pathways, densely connected nodes or modules can be jointly modeled to capture novel biology. In order to verify that densely connected nodes could identify biological expression programs not represented in prior knowledge bases, we modeled a dataset where we know the ground truth perturbations (PBMCs stimulated with IFN-β). We omitted the particular pathways that should have been most relevant for the perturbation, but included a densely-connected module with several nodes capable of representing expression from all genes. When we used gene attributions to identify the important expression contributors for these densely-connected, unannotated nodes, we found that they corresponded to the expected ground truth biology (Supplementary Fig. 3). Specifically, the 50 most important genes for each densely connected node were significantly enriched for multiple pathways corresponding to biological ground truth (Supplementary Fig. 4). As a control experiment, we repeated this procedure, but instead of omitting pathways that correspond to known biological ground truth, we omitted pathways that we expect to be unrelated to the given perturbation. As an example, we look at removing the Reactome pathway 'Regulation of PLK1 Activity at G2/M Transition', which is involved in regulating a number of cellular proteins during mitosis [29]. We then again used gene attributions to identify the most important expression contributors to the densely-connected nodes (Supplementary Fig. 5). In this case, we found no significant pathway enrichments for the top 50 genes for any of the auxiliary nodes. This control experiment shows that densely-connected nodes can specifically represent unknown, novel sources of biological variation, rather than all missing pathway information. Gene attributions allow these densely-connected nodes to be inspected to understand biological signal not represented in prior knowledge.

In conclusion, while pathway attributions help give a high-level view of important processes in gene expression datasets, gene attributions help elucidate particular mechanisms in greater detail. Furthermore, these attributions can help identify novel biology in cases where the prior knowledge used to define the biologically-constrained model is mismatched with the expression programs present in the dataset being modeled.

2.5 PAUSE identifies mitochondrial oxidative phosphorylation as an important process in Alzheimer's brain samples

After testing the PAUSE approach by applying it to a variety of single cell datasets with known perturbations and well-characterized downstream effects, we wanted to use our PAUSE approach to gain insight into the biology of Alzheimer's disease (AD). AD is the most common form of dementia, and accounts for an estimated 60–70% of worldwide cases [35]. Clinically, AD manifests with memory loss and cognitive decline, while neuropathologically AD is associated with amyloid-β (Aβ) plaques and abnormal tau tangles in the brain [36]. We therefore applied PAUSE to a large collection of postmortem brain RNA-sequencing measurements from patients with AD, assembled by the AMP-AD (Accelerating Medicines Partnership Alzheimer's Disease) consortium. This dataset includes brain samples from a variety of sources, including the Religious Orders Study/Memory and Aging Project (ROSMAP) [37], Adult Changes in Thought (ACT) [38], Mount Sinai Brain Bank (MSBB), and Harvard Brain Tissue Resource Center (HBTRC) studies (see Methods for more detail).

Because of the diversity of data sources combined in this dataset, 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 (Figure 5b). To control for these dataset effects, we therefore modified the pmVAE architecture to be a *conditional* pathway module VAE (cpmVAE) (Figure 1a, Figure 5a). In addition to the set
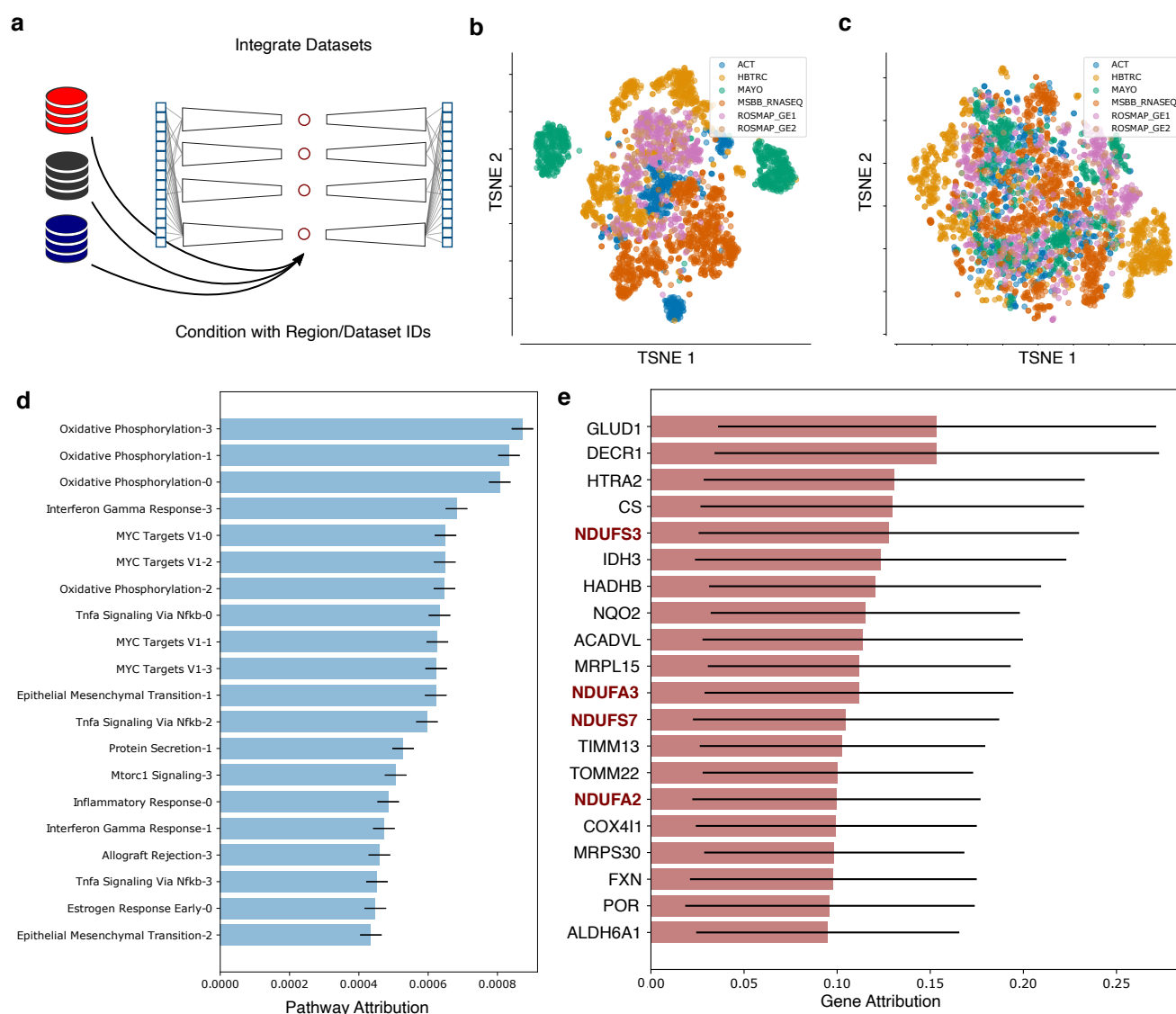
Fig. 5 | PAUSE analysis highlights the importance of Mitochondrial Respiratory Complex I in Alzheimer's disease. a, A conditional pathway module VAE (cpmVAE) is trained to integrate human post-mortem brain tissue samples from multiple Alzheimer's disease (AD) studies. **b**, TSNE plot of the latent space of a pathway module VAE (pmVAE) trained *without* batch effect correction by conditioning on study ID. **c**, TSNE plot of the latent space of a cpmVAE trained *with* batch effect correction by conditioning on study ID. **d**, Most important pathway latent variables in cpmVAE model according to PAUSE pathway loss attribution. Error bars indicate the standard deviation over local attributions **e**, Most important gene expression contributors for top pathway (Oxidative Phosphorylation Node 3). Gene names bolded and in red encode proteins in Mitochondrial Respiratory Complex I. Error bars indicate the standard deviation over local attribution magnitudes.

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 (like data source identity). By explicitly encoding the batch effects in these variables, the model is free to learn biological information in the latent space. When our cpmVAE model is trained on the data, we see that there is less separation in the latent space on the basis of source dataset (Figure 5c, Supplementary Fig. 6). To assess the quality of the embedding learned by our cpmVAE model and further "sanity check" our approach, we showed that our cpmVAE model reconstructed the original gene expression space more accurately than a standard pmVAE model, had less clustering by dataset source, and was more predictive of the neuropathological phenotype information we had for a small subset of samples than the original gene expression space (Supplementary Fig. 6).

When we generate PAUSE pathway attributions to interrogate the latent space learned by our cpmVAE model, we can see that the single pathway explaining the most observed variation in the biological latent space is the Hallmark Oxidative Phosphorylation pathway (Fig. 5d). Oxidative phosphorylation in the mitochondria is one of the major metabolic processes responsible for generating the ATP necessary for neuronal function, and dysregulation and differential expression of this pathway has previously been linked to AD [39]. 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 [40], these changes in expression of oxidative phosphorylation genes likely play an important role in the pathophysiology of AD [41]. Furthermore, the expression of this pathway varies not only between patients with AD and without AD, but also varies greatly across AD patients, suggesting that this pathway may be related to the heterogeneity of clinical phenotypes observed in AD patients [42].

After identifying several nodes related to oxidative phosphorylation as the top general processes in this dataset, we used gene attributions to further interrogate the specific function represented in the top pathway node, which is Oxidative Phosphorylation Node 3. We can visualize the individual gene attributions for this node using a summary plot (Fig. 5e). 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 [43].

2.6 Experimental validation of PAUSE analysis identifies mitochondrial Complex I as a potential AD therapeutic target

Unlike the single cell datasets with clear ground truth perturbations and well-studied downstream effects examined in the prior analyses, the biology underlying the differences in gene expression in real-world samples from patients with AD is less clearly characterized. Therefore, in order to verify that the genes and pathways identified by PAUSE are biologically-relevant, we could not reference a known ground truth, and had to experimentally validate our findings.

To gain insight into the biological relevance of the genes identified by our PAUSE analysis of the Alzheimer's brain expression data, we used the nematode *C. elegans*, a well-established animal model of Aβ proteotoxicity [44]. We conducted experiments with GMC101, a transgenic worm line displaying an age-associated aggregation of human Aβ1-42 peptide resulting in rapid onset of age-associated paralysis. A stringent reciprocal best hits (RBH) approach (BLAST e-value $\leq 10^{-30}$; details in Methods) was used to identify nematode orthologs for human genes.

We used RNAi via bacterial feeding to reduce expression of 13 *C. elegans* genes encoding homologs of human Complex I proteins in animals expressing toxic Aβ. This resulted in a striking delay in paralysis relative to control animals treated with empty vector (EV) RNAi (Figure 6) (details in Methods). Suppression of paralysis by knockdown of Complex I genes was comparable, or often exceeded, the suppression of paralysis by our positive control, RNAi knockdown of the insulin-like receptor gene daf-2 (Figure 6), one of the strongest known suppressors of Aβ toxicity in worms [45].

3 Discussion

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, as confirmed by known ground-truth perturbations in the single cell analyses and by experimental validation in the AD dataset. Unlike existing methods, our pathway attributions do not depend on having labeled data to identify important pathways. This enables fully unsupervised data exploration, and can complement supervised analyses when labels are present. 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.

An important limitation of our unsupervised pathway attributions is that they 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 (see Fig. 3b). We emphasize that our unsupervised attribution method measures a different quantity than supervised approaches, and should be considered a complementary approach rather than a replacement in applications where labeled data are present.
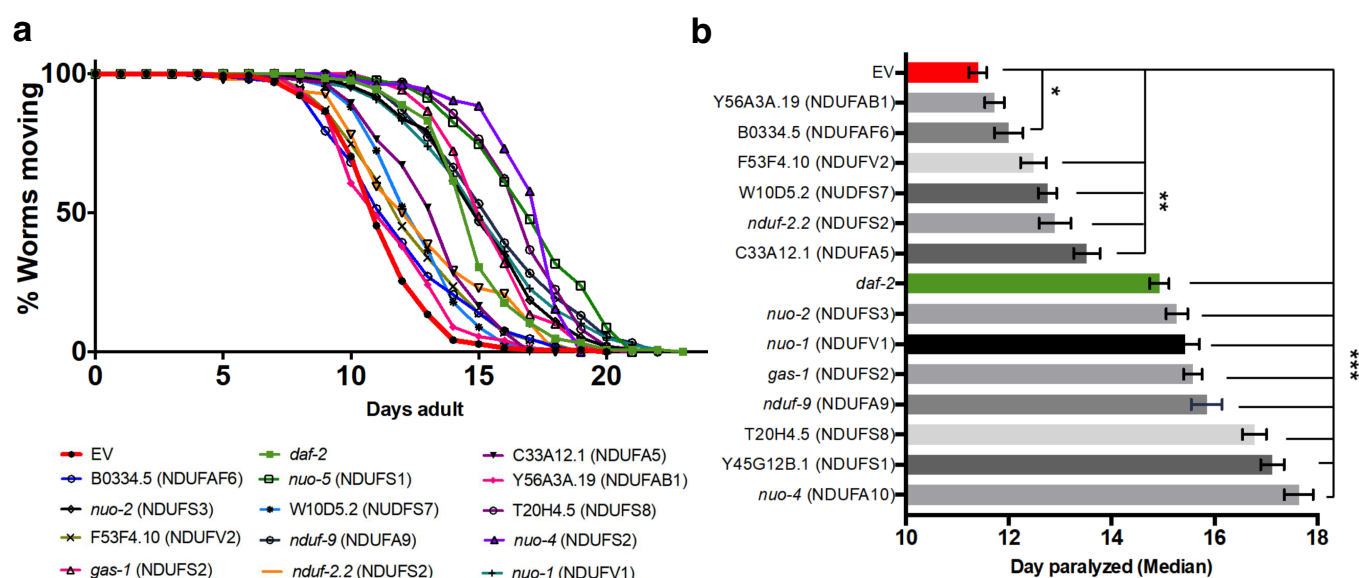
Fig. 6 | *C. elegans* Aβ proteotoxicity assay validates importance of mitochondrial Complex I genes. a, Paralysis curves for the reciprocal best orthologs of human Complex I genes. All tested RNAi conditions significantly suppress paralysis. **b,** The same data as in **a** plotted as the median day 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.

An assumption of our approach is that pathway databases provide good representations of the underlying biology of the system being modeled, which is an assumption shared by any method that incorporates pathway information into the modeling architecture. We note, however, that our incorporation of gene attributions specifically helps to alleviate this problem. Our gene attributions, like the pathway attributions, can be applied to arbitrarily deep networks, and allow for the biological interpretation of models with multiple latent variables per pathway. These attributions also allow the user to understand what biology is captured by the model in densely connected nodes that are not associated with a particular pathway.

Importantly, PAUSE was able to identify mitochondrial Complex I as a critical factor mediating Aβ-related pathophysiology in AD, pointing to the promise of mild inhibition of Complex I as a potential new paradigm for developing future AD therapeutics. This finding is supported by prior data indicating that mild inhibition of Complex I reduced Aβ and tau levels in mouse models of familial AD [46], that diets rich in capsaicin — a natural product that can inhibit Complex I — are associated with lower serum Aβ levels in the elderly [47], and that capsaicin reduced AD-associated changes in the hippocampus of rats with type 2 diabetes [48].

4 Conclusions

We anticipate that our approach of combining principled attributions with interpretable model architectures will prove to be a broadly useful strategy in domains beyond gene expression. Rather than viewing attribution methods or biologically-meaningful representations as individual panaceas for machine learning interpretability, we hope that researchers will view these approaches as composable tools, each addressing unique needs. For example, principled feature attribution methods are currently applied to computer vision models, but most of these methods generate attributions at the level of pixels, which are not the most human-interpretable features and struggle to represent higher-level concepts [49, 50]. On the other hand, generative models like GANs can more clearly illustrate complex patterns differentiating domains, but lack the rigorous quantification of feature attribution methods [49, 51]. We believe future work that is able to rigorously quantify contributions of human-interpretable concepts in deep learning models will be useful in domains like microscopy, computer vision, and digital pathology.

Additionally, while biological pathways represent one useful representation, they will not necessarily be the only meaningful representation for all data types. For example, in the area of computer vision, researchers have proposed using additional labeled data to train medical image models with representations encoding high-level clinical concepts, such as the presence or absence of bone spurs in x-ray images of knees for osteoarthritis severity determination [52]. Future work defining other useful conceptual representations will be helpful in biological applications, while finding methods such as the sparsely-constrained architectures used in this manuscript to learn representations without the need for costly additional labels will be useful in other domains.

5 Acknowledgements

6 Methods

6.1 Model Architectures

Biologically interpretable autoencoders are unsupervised deep learning models with latent variables corresponding to known biological pathways or other gene groupings. Methods to create biologically interpretable autoencoders can be grouped into two broad strategies.

The first involves directly modifying the architecture of a network with sparse, masked connections. This approach is taken by methods such as VEGA (VAE Enhanced by Gene Annotations) [21], which masks the weights in the linear decoder of a variational autoencoder (VAE) so that each latent variable is only connected to genes in the pathway represented by that latent variable. The second involves a "soft" modification of the network weights using weight regularization during training. For example, the expiMap method (explainable programmable mapper), which is also a VAE with a linear decoder, allows all connections from all latent variables in its decoder to potentially have non-zero weight, but adds an L1 regularization penalty specifically to the weights corresponding to genes not present in the original pathway definitions [20].

In our experiments, we primarily used a *fully deep* architecture similar to that proposed in the pathway module variational autoencoder (pmVAE) approach [22]. This autoencoder network's encoder and decoder *both* have multiple non-linear layers. To better describe these networks, we can start with a description of a typical, unmodified VAE [53]. 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 $\mathbf{z}$ from high-dimensional data $\mathbf{x}$, such as gene expression data. These networks are trained by stochastic gradient descent to maximize the variational lower bound on the likelihood of the data:

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

In order to learn a low-dimensional latent space $\mathbf{z}$ that corresponds to biological pathways, 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. These modules may have multiple hidden layers in both the encoder and decoder network, and may have multiple latent variables associated with 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. Each subsequent layer is masked with a binary *separation mask*, which is a block diagonal matrix ensuring that non-zero connections only occur within pathway modules. In addition to the sparse masking modifications made to the encoder and decoder architectures, the pmVAE method also alters the training objective to add a *local reconstruction loss*:

$$\mathcal{L}_{\text{recon}}^{(\text{local})} = \frac{1}{K} \sum_p \frac{N}{N_p} \|\hat{x}^{(p)} - x^{(p)}\|_2^2, \tag{2}$$

where $\hat{x}^{(p)}$ is the reconstructed expression of the genes in pathway (p), $x^{(p)}$ is the observed expression of the genes in pathway (p), $\frac{N}{N_p}$ is a weighting term to weight each module's local reconstruction loss dependent on the number of genes in a particular module N_p compared to the total number of genes N, and K is the total number of modules considered.

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), and therefore maximize the *conditional* variational lower bound objective:

$$\log p_\theta(x|c) \geq \mathbb{E}_{q(z|x,c;\theta)}[\log p(x|c,z;\psi)] - \mathbb{KL}(q(z|x,c;\theta)||p(z|c)) = -\mathcal{L}_{\text{ELBO}}^c. \tag{3}$$

These condition labels are passed to each pathway module in the encoder, and again to each pathway module in the decoder. In order to ensure that the learned embedding is fully independent of the unwanted sources of variation encoded in the conditional labels, we add a regularization term to the loss based on the Hilbert Schmidt independence criterion (HSIC) between the latent embedding $\mathbf{z}$ and the dataset labels $\mathbf{c}$ [54, 55]. Therefore, our total loss function optimized for the cpmVAE architecture is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ELBO}}^c + \mathcal{L}_{\text{recon}}^{(\text{local})} + \mathcal{L}_{\text{HSIC}}. \tag{4}$$

We implemented our cpmVAE model (and the pmVAE models) using the PyTorch deep learning library [56]. For both the cpmVAE and pmVAE models, we included BatchNorm1D layers following all but the final layers of both the encoder and decoder networks. We optimized the networks using an Adam optimizer with an initial learning rate of 0.001, and decreased the learning rate by a factor of 10 following each epoch if there was not a decrease in global reconstruction error on a held out validation set. All models were trained for 200 epochs, and the model checkpoint with the lowest reconstruction error on a held out validation set was used for downstream analysis. For single cell benchmark experiments, each pathway module had a hidden layer of 12 nodes, followed by a pathway latent space of 1 node. For all other experiments, each pathway module had a hidden layer of 12 nodes, followed by a pathway latent space of 4 nodes.

For the cpmVAE experiments, the HSIC loss was multiplied by a factor of 1×10^6, to ensure this term was around equal magnitude with the other loss terms.

Finally, for our experiments involving models with linear decoders (see Supplementary Fig. 1), we used the Interpretable Autoencoder architecture proposed by Rybakov et al. [57]. This model is similar to the expiMap method [20], but is a deterministic autoencoder rather than a VAE. We trained these networks using the code from the intercode repository (https://github.com/theislab/intercode), and for both experiments the network was trained using the hyperparameters described in (https://github.com/theislab/intercode/blob/main/notebooks/intercode-api-Kang18.ipynb): 80 epochs with a batch size of 62 samples, a learning rate of 0.001, and the regularization hyperparameters of $\lambda_0 = 0.1$, $\lambda_1 = 0.93$, $\lambda_2 = 0.0$, $\lambda_3 = 0.57$.

6.2 Attributions

Feature attribution is one of the largest and most well-studied classes of methods for machine learning interpretability. Methods in this class are based on concepts from cooperative game theory, like the Shapley value [23] and Aumann-Shapley value [24]. By framing the output of a complex, black box model as a game, and the input features as players in that game, these concepts can be used to understand which input features of a model most impacted that model's predictions. For example, the feature attribution method SHAP defines a cooperative game to be the expected value of a machine learning model evaluated on a sample of interest, conditional on the features present in each coalition, and then uses the Shapley value to allocate credit to those features [10]. Another feature attribution method, Integrated Gradients, calculates an Aumann-Shapley value by integrating the gradient of the model's output with respect to its input features, and can be used when machine learning models are differentiable, such as neural network models [12].

We first propose a novel attribution method to understand which latent factors are important in unsupervised neural networks, based on Integrated Gradients [12] and the Aumann-Shapley value [24]. However, rather than attributing the output of a supervised machine learning model to its input features, we attribute the reconstruction loss of an autoencoder model to its latent features. To make our formulation general enough to encompass both variational and standard autoencoders, we define an encoder network $f : \mathbb{R}^d \mapsto \mathbb{R}^h$ and a decoder network $g : \mathbb{R}^h \mapsto \mathbb{R}^d$. For a decoder, we can define the reconstruction error, $\ell(z) = \|x - g(z)\|_2^2$, as a function of a latent input z, which is either a sample drawn from the distribution parameterized by the encoder network given an input sample x in the case of a variational autoencoder, or simply the deterministic embedding of the encoder network in the case of a standard autoencoder. To quantify the contribution of a particular latent node i (which in the case of an interpretable autoencoder, corresponds to a biological pathway), we can therefore apply the following formula to get a local pathway importance value:

$$\phi_i^{\mathrm{pathway}}(z) = (z_i - z_i') \times \int_{\alpha=0}^1 \frac{\partial \ell(z' + \alpha(z - z'))}{\partial z_i} d\alpha, \tag{5}$$

where z is the value of the learned pathway embedding for a particular sample, z' is a baseline value for that embedding, such as a vector of all 0s or the mean of that embedding's value over the dataset, and α is a scalar representing the distance on the straight-line path being integrated between the baseline and the actual value of the sample.

These local attributions have a variety of desirable properties [12], such as *completeness*, meaning that the attributions for each pathway node sum to the difference between the reconstruction error of the model at the input z (when the pathway information is "present") and the baseline z' (when the pathway information is "absent"). Completeness is important in that it gives the attributions a natural scale: each pathway's attribution represents the amount of variance explained by the model for the original sample that can be credited to that pathway. These local attributions are also *implementation invariant*, meaning the attributions are always identical for two functionally equivalent networks. This is important, in that it ensures the attributions reflect real functional differences between networks, rather than relying on artifacts based on particularities of the model employed.

To go from local pathway importance (the contribution of a particular pathway for a particular sample) to global pathway importance (the contribution of a pathway over an entire dataset), we can take the expected value of the local attributions over the samples in the original data:

$$\Phi_i^{\mathrm{pathway}} = \mathbb{E}_{z \sim \mathcal{D}}[\phi_i^{\mathrm{pathway}}(z)]. \tag{6}$$

Since these attributions can be calculated so efficiently, rather than sampling we calculate the local attributions for every point in the dataset. Because the local attributions allocate the reduction in the sample-level reconstruction error to each pathway, the global attributions allocate the reduction in the dataset-level reconstruction error to each pathway. Since the mean squared error is proportional to the variance in the original expression space explained by the model, allocating the reconstruction error to each pathway tells us how much variance is explained by each pathway.

For our gene-level attributions, we want to understand which genes are important contributors to each of the latent pathway nodes. For a particular node in the latent space k that we wish to explain, we define the function $f_k(x) : \mathbb{R}^d \mapsto \mathbb{R}$ as the kth node in the encoder network's latent space. We can therefore quantify gene j's contribution to pathway node k by applying the Integrated Gradients formula to this function to get:

$$\phi_j^{\mathrm{gene,k}}(x) = (x_j - x_j') \times \int_{\alpha=0}^1 \frac{\partial f_k(x' + \alpha(x - x'))}{\partial x_j} d\alpha, \tag{7}$$

where x is the observed expression for a particular sample, x' is a baseline value for gene expression, such as a vector of all 0s or the average gene expression over the dataset, and α is a scalar representing the distance on the straight-line path being integrated between the baseline and the actual value of the sample.

To summarize local gene attributions into global gene attributions, we take the average of the magnitude of the local attributions over the samples in the original data:

$$\Phi_j^{\text{gene,k}} = \mathbb{E}_{x \sim \mathcal{D}}[|\phi_j^{\text{gene,k}}(x)|]. \tag{8}$$

It is necessary to average the magnitudes rather than the raw values for gene attributions to avoid cancellation effects from attributions with different signs [25]. This was not a problem for pathway attributions, as all attributions should have the same sign, as adding more pathway information should never *increase* the loss.

All attributions were calculated using the Path Explain repository (https://github.com/suinleelab/path_explain), which is a Python library for explaining feature importances and feature interactions in deep neural networks using path attribution methods [58–60]. All attributions were generated using the "attributions" method of the PyTorch explainer; the argument "num_samples" was set to 200, while the argument "use_expectation" was set to "FALSE."

6.3 Pathway gene set definition

We use pathway modules defined by gene set annotations from the Reactome database for all single cell experiments [29]. This database contains a total of 674 gene sets (with a median gene set size of 27 genes). The number of gene sets included for a given model depends on the dataset. For all single cell datasets, we only include gene sets that contain a minimum of 13 genes from the given dataset. This yields a total of 200 gene sets for the PBMC INF-β dataset, 129 gene sets for the Intestinal Epithelium dataset, and 341 gene sets for the Jurkat Anti-CD3/Anti-CD28 dataset.

For the bulk Alzheimer's brain expression expermients, the pathway module architecture of the networks was defined using the Hallmark gene sets from MSigDB. These pathways are highly curated to "summarize and represent specific well-defined biological states or processes and display coherent expression" [61].

6.4 Single cell expression datasets

PBMC INF-β
This dataset, from Kang et al. [26], contains human peripheral blood mononuclear cells (PBMCs) from eight patients with Lupus who were either treated with INF-β or left untreated as a control. We followed the same preprocessing steps taken in prior analysis of this dataset [22, 57], which consists of library size normalization, removal of low variance genes and log transformation. After preprocessing, the final dataset we use contains 13,576 samples with 979 genes. The dataset contains 6359 control and 7217 stimulated cells.

Intestinal Epithelium
This dataset, from Haber et al. [27], profiles the response of mouse small intestinal epithelial cells to pathogen exposures, specifically *Salmonella enterica* and *H. polygyrus*. Here, we include healthy control cells and cells 10 days after being infected with *H. polygyrus*. Preprocessing was the same as that used by Weinberger and Lin [62], and involved retaining only the top 2000 most highly variable genes followed by library size normalization and log transformation. After preprocessing, the final dataset we use contains 5951 samples with 2000 genes. The dataset contains 3240 control and 2711 *H.polygyrus*-stimulated cells.

Jurkat Anti-CD3/Anti-CD28
This dataset, from Datlinger et al. [28], contains human Jurkat cells (immortalized T-lymphoctyes) that were either starved or stimulated with anti-CD3 and anti-CD28 antibodies. Preprocessing follows the approach described in Gut et al. [22], who filter out genes and cells with low expression, perform library size normalization and log transformation, and filter out low variance genes. After preprocessing, the final dataset we use contains 1288 samples with 2139 genes. This dataset contains 607 control and 681 stimulated cells.

6.5 Bulk brain expression datasets

For experiments with AD, we used data from the ROSMAP, ACT, HBTRC, MAYO, and MSBB studies. Around half of the people in each cohort had been diagnosed with dementia by the time of death. ROSMAP RNA-Seq data and MSBB RNA-Seq data were made available by Sage Bionetworks on the AMP-AD Knowledge Portal23 with Synapse IDs syn3505732 and syn7391833, respectively. The ACT RNA-Seq data20 was collected by the Allen Institute for Brain Science, Kaiser Permanente Washington Health Research Institute (KPWHRI), and the University of Washington (UW), and it was made available with Synapse ID syn5759376. Mayo Clinic Brain Bank data was made available at Synapse ID (syn5550404; https://doi.org/10.1038/sdata.2016.89), and the Harvard Brain Tissue Resource Center (HBTRC) study was made available at the following Synapse ID (syn3159435). In each study, we used the protein-coding genes that have a nonzero RNA-Seq read count in at least one-third of the samples. Overlapping these genes across the three studies resulted in 16,252 genes which we used in our experiments. We used normalized and log-transformed RNA-Seq read counts for all datasets. For the subset of samples with neuropathology data, MSBB neuropathology data was made available

by the AMP-AD Knowledge Portal of Sage Bionetworks through synapse.org with Synapse ID syn6101474. We accessed neuropathology data from the ROSMAP and ACT studies through data use agreements.

6.6 Benchmark experiments

To benchmark how well our PAUSE pathway attributions identified important factors of latent variation explaining the observed gene expression measurements, we adapted two benchmarks of feature importance from the feature attribution literature [10]. The first benchmark, our *impute* benchmark, measures how much the reconstruction error of a model increases when the pathways identified as important by an attribution measure are removed from a model. After training a pmVAE model [22] (12 nodes in a single hidden layer, and a single latent node for each module) on each of the 3 single cell RNA-seq datasets, pathways were ranked by one of five methods: (1) PAUSE attributions; (2) LR score, the accuracy of a logistic regression model trained to classify the ground truth perturbation using the latent pathway information [22, 57]; (3) KL Divergence between the learned posterior distribution and the isotropic gaussian prior; (4) Random; and (5) the magnitude of the learned latent space scale parameter for each pathway. After attaining a pathway ranking for each method, pathway information is removed by masking the real latent embeddings with a constant value. The increase in reconstruction error over the dataset is then measured after each pathway is removed. Attribution methods that do a better job of identifying important pathways will increase the reconstruction error more quickly, leading to a larger area under the curve. These curves were summarized using the scikit-learn metrics auc function [63]. The second benchmark, which we termed a *retrain* benchmark, follows the same initial steps as the impute benchmark (train models and rank pathways), but then trains new models using only modules corresponding to the top pathways. In this benchmark, methods that perform better will decrease the reconstruction error more quickly as pathway modules are added. Finally, to benchmark the models with linear decoders, we followed the same steps as the impute benchmark, but trained the intercode models as described above in the section on Model Architectures, rather than pmVAE models.

6.7 Sanity checks for bulk expression cpmVAE model

Before interpreting the learned embedding for our conditional pathway module VAE (cpmVAE) model trained on the AD dataset, we wanted to perform several sanity checks to ensure the model was sufficiently plausible and reliable. We first wanted to ensure that our proposed architecture (cpmVAE) was able to reconstruct the data with at least the same accuracy as the previously proposed, non-conditional pmVAE model. We therefore randomly split the data into a train set (75% of samples) and a test set (25% of samples), trained both a pmVAE and a cpmVAE on the training set for 50 epochs, and measured the reconstruction error on the held out test set for each model. We repeated this procedure a total of 10 times with different random train/test splits, and compared the distribution of test reconstruction errors for the two models (Supplementary Fig. 6a). We found that the distribution of test reconstruction errors attained with the cpmVAE model was lower than the distribution of test reconstruction errors attained by the previously proposed pmVAE model (Wilcoxon rank-sums test statistic = 1.96, p = 0.049).

Next, we wanted to ensure that the dataset integration was successful, and that the conditioning and HSIC regularization indeed led to a pathway embedding that encoded less information about the data sources than the standard pmVAE embedding. To measure how clustered cells were according to dataset source in the latent space, we first embedded the pathway latent spaces of the cpmVAE model and the pmVAE model into 2 dimensions using the scikit-learn implementation of TSNE with the perplexity parameter set to 40.0 [63, 64]. We then applied k-means clustering to the latent space (with k set to the number of true data sources), and computed the adjusted Rand index (ARI) of the clustering [65]. When we compare the distribution of ARIs for the pmVAE embedding to the distribution of ARIs for the cpmVAE embedding (Supplementary Fig. 6b), we see that the embedding learned by the cpmVAE model has significantly less clustering according to dataset source (Wilcoxon rank-sums test statistic = 3.78, p = 1.57×10^{-4}).

Finally, after demonstrating that the cpmVAE model removed unwanted sources of variation from its pathway embedding, we wanted to ensure that our cpmVAE embedding still retained biological information. For a small subset of the brain expression samples (664 samples), associated amyloid-β protein density measurements via immunohistochemistry (Aβ IHC) were available. To ensure that biological information related to neuropathological phenotype was not being lost by our model, we wanted to measure how well Aβ IHC could be predicted from the latent space, as compared to the original full transcriptomic representation. Our hypothesis was that if the biologically relevant information in the latent space was being preserved, prediction of Aβ density should be no worse when using the cpmVAE embedding than when using the original gene space embedding. For 20 random train/test splits of the data, we compared the Spearman correlation between predicted and actual Aβ density for linear models trained on the full gene expression measurements to the Spearman correlation between predicted and actual Aβ density for linear models trained only on the cpmVAE pathway embeddings (Supplementary Fig. 6c). All linear models were trained using the scikit-learn RidgeCV function [63], with a hyperparameter search for alpha values ranging from 1e-3 to 1e2. We see that the cpmVAE embedding features actually lead to significantly more predictive models of Aβ density than the original gene expression features (Wilcoxon rank-sums test statistic = 2.81, p = 4.90×10^{-3}).

6.8 Identification of *C. elegans* homologs

To enable biological testing of the human genes identified using our computational analysis, we obtained the Reciprocal Best Hits (RBHs) between human and *C. elegans*. We first identified all unique protein sequences for each potential marker gene using the biomaRt R package available on CRAN [66, 67]. Then, we used the NCBI BLAST tool to identify the *C. elegans* orthologs for each complete human protein query sequence [68, 69]. We downloaded the *C. elegans* protein sequences from wormbase.org/species/c_elegans (release WS266). We took into account only the protein pairs mapped from human to *C. elegans* with a BLAST e-value smaller than 10^{-30}. For each *C. elegans* isoform, we identified the corresponding human genes, again using the NCBI BLAST tool, and used only the orthologs that achieved a BLAST e-value smaller than 10^{-30}. This process resulted in high-confidence RBHs for us to test in *C. elegans*.

6.9 *C. elegans* strain, cultivation, and RNAi treatment

Standard procedures for *C. elegans* strain maintenance and manipulation were used, as previously described [70, 71]. All experiments were performed using the GMC101 strain expressing the human Aβ1-42 peptide under the unc-54 promoter [72]. Experimental worm populations of GMC101 animals were obtained from the Caenorhabditis Genetics Center (CGC) and cultivated on NGM plates with OP50 *E. coli* at 15C [72, 73]. Care was taken to ensure that the animals were never starved and the plates remained free of contamination.

The gene-specific RNAi clones were obtained from the commercial Ahringer or Vidal *C. elegans* RNAi-feeding libraries (BioScience, Nottingham, UK). Each bacterial clone was struck-out onto LB plates containing carbenicillin (50 ug/ml) and tetracycline (10 ug/ml). Single colonies were then seeded into 5 ml LB + carbenicillin (50 ug/ml) and tetracycline (10 ug/ml) for growth overnight on a 37C rotator. 100 ul of each overnight culture was then inoculated into 10 ml of LB containing carbenicillin (50 ug/ml) and tetracycline (10 ug/ml) and IPTG (5mM) and incubated on a 37C rotator for 4 hours. Each bacterial growth was then centrifuged at 3500 X G for 25 minutes, decanted, and the pellet resuspended in 0.5 ml of LB containing carbenicillin (50 ug/ml), tetracycline (10 ug/ml), and IPTG (5 mM). To verify that the RNAi plasmid DNA contained the expected gene target, each RNAi clone was purified and assessed through PCR (polymerase chain reaction) with sequence-specific primers or through Sanger sequencing.

6.10 Nematode paralysis assays

Paralysis assays were performed by visually inspecting recordings of the animals daily to determine if they were capable of normal locomotion or if they were paralyzed and unable to transit the agar plate. A custom-built robotic system, the WormBot [74], was equipped with a digital camera and was used to obtain images of individual wells of a 12 well-plate at 10-minute intervals over the entire course of the experiment. Each well contained 30-40 individual GMC101 animals expressing Aβ. Using a custom-built web interface that enabled manual annotation of serial images from each plate, the age at which each animal stopped moving could be easily determined. We applied this system to the transgenic Aβ model line GMC101 to determine the time of paralysis onset for each individual animal across all RNAi experiments. Statistical significance of mean paralysis time-points between RNAi conditions was determined by a weighted log-rank test [75, 76]. All key results were independently verified using standard manual microdissection methodologies.

Prior to loading on the experimental plates, animal populations were propagated on high-growth plates seeded with NA22 *E. coli*. Worm populations were developmentally synchronized by hypochlorite treatment, and the remaining eggs were deposited on unseeded plates overnight. Synchronized larval stage 1 animals were washed off unseeded plates and moved onto standard *C. elegans* RNAi plates containing carbenicillin (50 mg/ml), tetracycline (10 mg/ml), and IPTG (5 mM) 48h at 20C. These developmentally synchronized, late larval stage 4-populations were then washed and transferred to their respective RNAi conditions on 12-well plates. We used standard RNAi conditions plus FuDR (100 ug/ml) to prevent progeny and nystatin (200 mg/ml) to prevent fungal growth [70]. Each RNAi condition was tested in 2-3 wells as technical replicates. At least three biological replicates, each started on different weeks, were conducted for each RNAi clone. Animals were maintained at ambient room temperature (generally 22-24C) over the course of the paralysis assay.

6.11 Code Availability

Code to reproduce results and figures is available at: https://github.com/suinleelab/PAUSE.
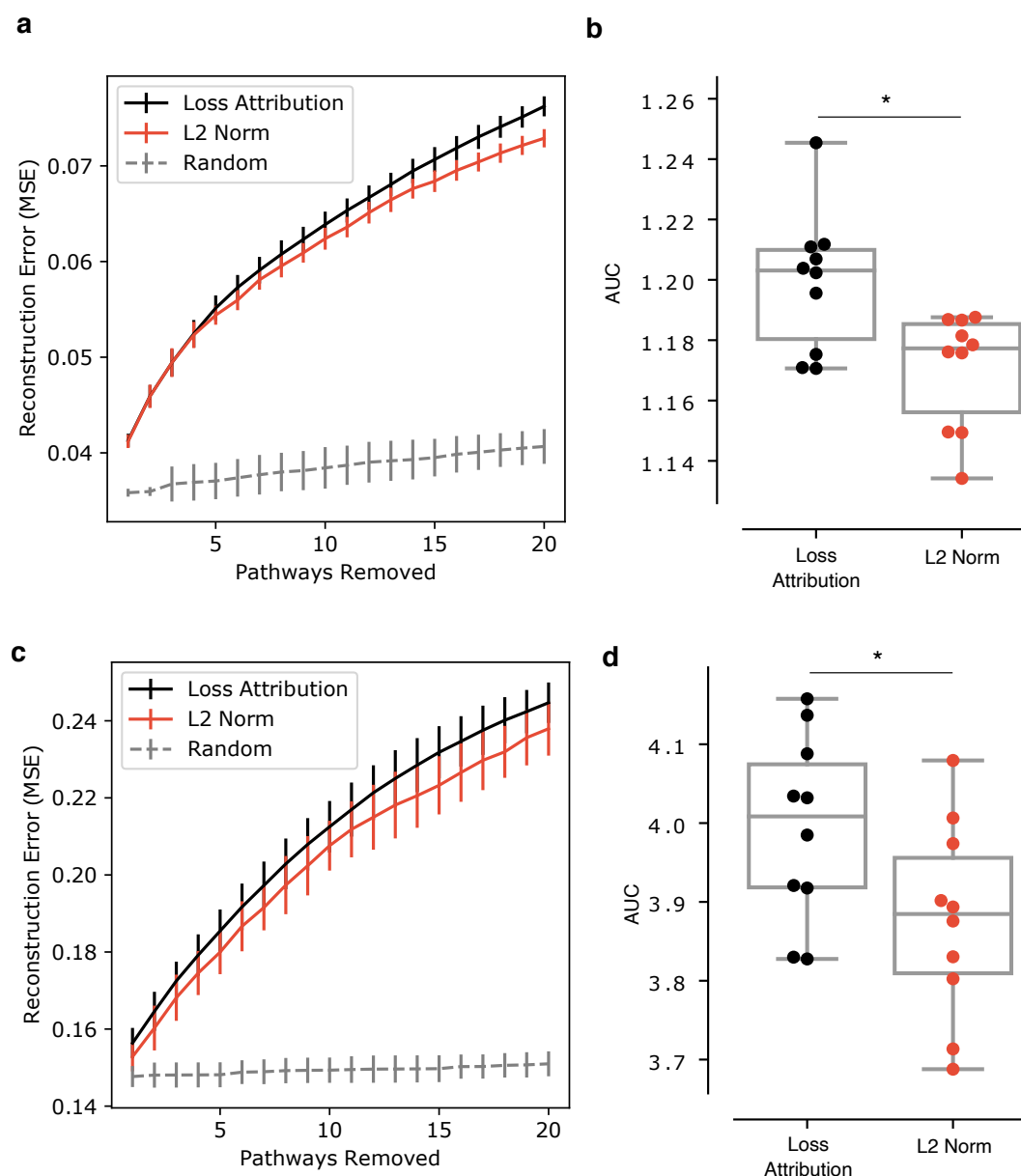
References

1. Zappia, L. & Theis, F. J. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome biology* **22,** 1–18 (2021).

2. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology* **40,** 121–130 (2022).

3. Dincer, A. B., Janizek, J. D. & Lee, S.-I. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* **36,** i573–i582 (2020).

4. Amodio, M. *et al.* Exploring single-cell data with deep multitasking neural networks. *Nature methods* **16,** 1139–1145 (2019).

5. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15,** 1053–1058 (2018).

6. Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B. & Goldenberg, A. Dr. VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics* **35,** 3743–3751 (2019).

7. Tan, J., Hammond, J. H., Hogan, D. A., Greene, C. S. & Gilbert, J. A. ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions. *mSystems* **1,** e00025–15 (2016).

8. Tan, J. *et al.* Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. *Cell systems* **5,** 63–71 (2017).

9. Way, G. P. & Greene, C. S. Discovering pathway and cell type signatures in transcriptomic compendia with machine learning. *Annual Review of Biomedical Data Science* **2,** 1–17 (2019).

10. Lundberg, S. M. & Lee, S.-I. *A unified approach to interpreting model predictions* in *Proceedings of the 31st international conference on neural information processing systems* (2017), 4768–4777.

11. Shrikumar, A., Greenside, P. & Kundaje, A. *Learning important features through propagating activation differences* in *International Conference on Machine Learning* (2017), 3145–3153.

12. Sundararajan, M., Taly, A. & Yan, Q. *Axiomatic attribution for deep networks* in *International Conference on Machine Learning* (2017), 3319–3328.

13. Dincer, A. B., Celik, S., Hiranuma, N. & Lee, S.-I. DeepProfile: Deep learning of cancer molecular profiles for precision medicine. *BioRxiv,* 278739 (2018).

14. Way, G. P., Zietz, M., Rubinetti, V., Himmelstein, D. S. & Greene, C. S. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome biology* **21,** 1–27 (2020).

15. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* **36,** 3418–3421 (2020).

16. Higgins, I. *et al.* *beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework* in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (OpenReview.net, 2017). https://openreview.net/forum?id=Sy2fzU9gl.

17. Hao, J., Kim, Y., Kim, T.-K. & Kang, M. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC bioinformatics* **19,** 1–13 (2018).

18. Elmarakeby, H. A. *et al.* Biologically informed deep neural network for prostate cancer discovery. *Nature* **598,** 348–352 (2021).

19. Demetci, P. *et al.* Multi-scale inference of genetic trait architecture using biologically annotated neural networks. *PLoS genetics* **17,** e1009754 (2021).

20. Lotfollahi, M. *et al.* Biologically informed deep learning to infer gene program activity in single cells. *bioRxiv* (2022).

21. Seninge, L., Anastopoulos, I., Ding, H. & Stuart, J. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nature communications* **12,** 1–9 (2021).

22. Gut, G., Stark, S. G., Rätsch, G. & Davidson, N. R. PmVAE: Learning interpretable single-cell representations with pathway modules. *bioRxiv* (2021).

23. Shapley Ll, S. A value for n-person games. *Contributions to the Theory of Games II, Annals of Mathematical Studies* **28** (1953).

24. Aumann, R. J. & Shapley, L. S. *Values of non-atomic games* (Princeton University Press, 2015).

25. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* **2,** 56–67 (2020).

26. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature biotechnology* **36,** 89–94 (2018).

27. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551,** 333–339 (2017).
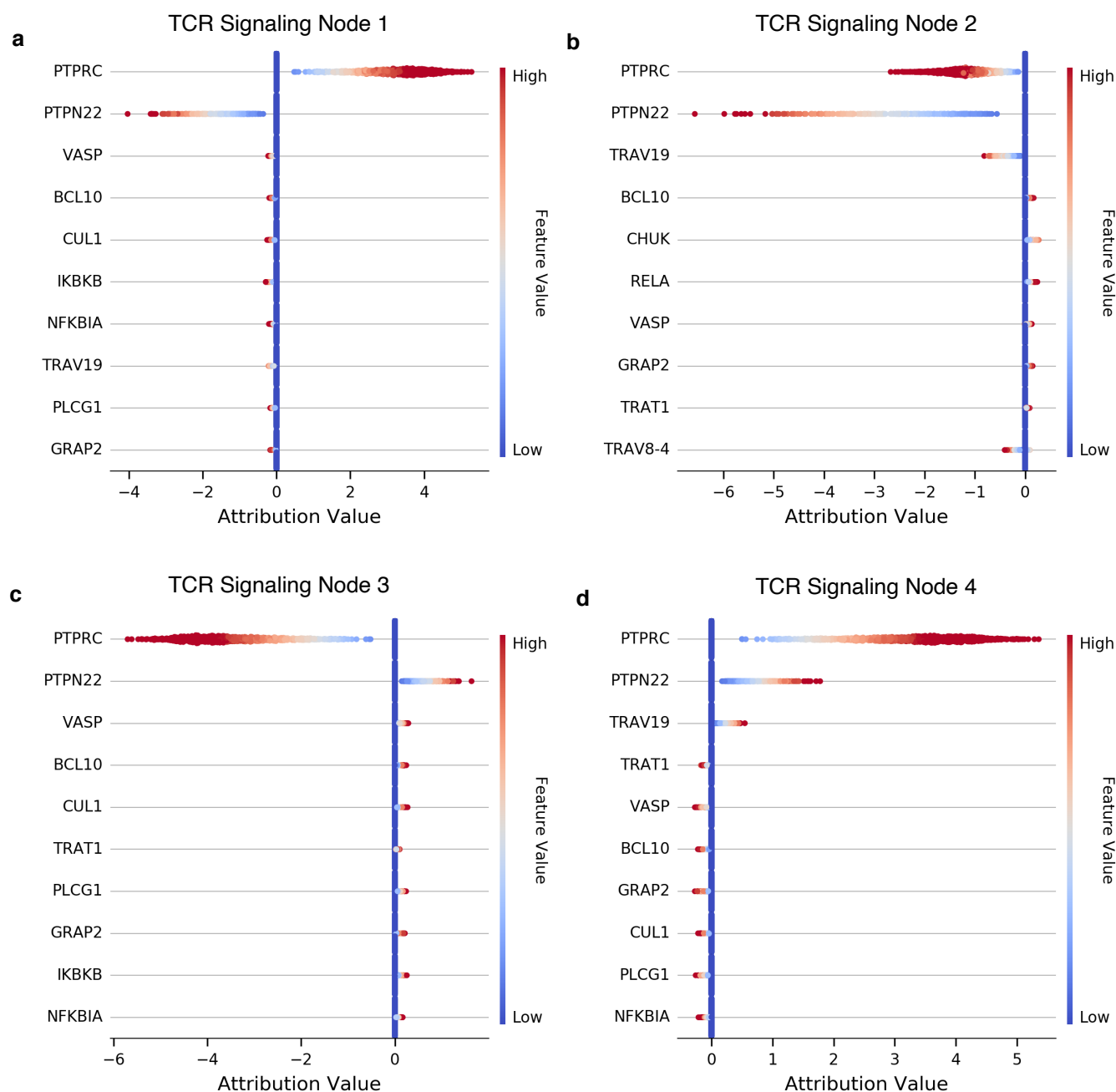
28. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods* **14.** Number: 3 Publisher: Nature Publishing Group, 297–301. ISSN: 1548-7105. https://www.nature.com/articles/nmeth.4177 (2022) (Mar. 2017).

29. Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic acids research* **46,** D649–D655 (2018).

30. Platanias, L. C. Mechanisms of type-I-and type-II-interferon-mediated signalling. *Nature Reviews Immunology* **5,** 375–386 (2005).

31. Wosen, J. E., Mukhopadhyay, D., Macaubas, C. & Mellins, E. D. Epithelial MHC class II expression and its role in antigen presentation in the gastrointestinal and respiratory tracts. *Frontiers in immunology,* 2144 (2018).

32. Forbester, J. L. *et al.* Interaction of Salmonella enterica serovar Typhimurium with intestinal organoids derived from human induced pluripotent stem cells. *Infection and immunity* **83,** 2926–2934 (2015).

33. Roche, P. A. & Furuta, K. The ins and outs of MHC class II-mediated antigen processing and presentation. *Nature Reviews Immunology* **15,** 203–216 (2015).

34. Hermiston, M. L., Zikherman, J. & Zhu, J. W. CD45, CD148, and Lyp/Pep: critical phosphatases regulating Src family kinase signaling networks in immune cells. *Immunological reviews* **228,** 288–311 (2009).

35. Association, A. *et al.* 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia* **14,** 367–429 (2018).

36. Leng, F. & Edison, P. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? *Nature Reviews Neurology* **17,** 157–172 (2021).

37. De Jager, P. L. *et al.* A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Scientific data* **5,** 1–13 (2018).

38. J Montine, T., A Sonnen, J., S Montine, K., K Crane, P. & B Larson, E. Adult Changes in Thought study: dementia is an individually varying convergent syndrome with prevalent clinically silent diseases that may be modified by some commonly used therapeutics. *Current Alzheimer Research* **9,** 718–723 (2012).

39. Wang, W., Zhao, F., Ma, X., Perry, G. & Zhu, X. Mitochondria dysfunction in the pathogenesis of Alzheimer's disease: Recent advances. *Molecular Neurodegeneration* **15,** 1–22 (2020).

40. Rhein, V. *et al.* Amyloid-β and tau synergistically impair the oxidative phosphorylation system in triple transgenic Alzheimer's disease mice. *Proceedings of the National Academy of Sciences* **106,** 20057–20062 (2009).

41. Shoffner, J. M. Oxidative phosphorylation defects and Alzheimer's disease. *Neurogenetics* **1,** 13–19 (1997).

42. Manczak, M., Park, B. S., Jung, Y. & Reddy, P. H. Differential expression of oxidative phosphorylation genes in patients with Alzheimer's disease. *Neuromolecular medicine* **5,** 147–162 (2004).

43. Sharma, L. K., Lu, J. & Bai, Y. Mitochondrial respiratory complex I: structure, function and implication in human diseases. *Current medicinal chemistry* **16,** 1266–1277 (2009).

44. Chen, X., Barclay, J. W., Burgoyne, R. D. & Morgan, A. Using C. elegans to discover therapeutic compounds for ageing-associated neurodegenerative diseases. *Chemistry Central Journal* **9,** 1–20 (2015).

45. Florez-McClure, M. L., Hohsfield, L. A., Fonte, G., Bealor, M. T. & Link, C. D. Decreased insulin-receptor signaling promotes the autophagic degradation of β-amyloid peptide in C. elegans. *Autophagy* **3,** 569–580 (2007).

46. Zhang, L. *et al.* Modulation of mitochondrial complex I activity averts cognitive decline in multiple animal models of familial Alzheimer's disease. *EBioMedicine* **2,** 294–305 (2015).

47. Liu, C.-H. *et al.* The associations between a capsaicin-rich diet and blood amyloid-β levels and cognitive function. *Journal of Alzheimer's Disease* **52,** 1081–1088 (2016).

48. Xu, W. *et al.* Capsaicin reduces Alzheimer-associated tau changes in the hippocampus of type 2 diabetes rats. *PloS one* **12,** e0172477 (2017).

49. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* **3,** 610–619 (2021).

50. Adebayo, J., Muelly, M., Abelson, H. & Kim, B. *Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation* in *International Conference on Learning Representations* (2021).

51. Singla, S., Pollack, B., Chen, J. & Batmanghelich, K. Explanation by progressive exaggeration. *arXiv preprint arXiv:1911.00483* (2019).

52. Koh, P. W. *et al. Concept bottleneck models* in *International Conference on Machine Learning* (2020), 5338–5348.

53. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

54. Gretton, A., Bousquet, O., Smola, A. & Schölkopf, B. *Measuring statistical dependence with Hilbert-Schmidt norms* in *International conference on algorithmic learning theory* (2005), 63–77.

55. Lopez, R., Regier, J., Jordan, M. I. & Yosef, N. Information constraints on auto-encoding variational bayes. *Advances in Neural Information Processing Systems* **31** (2018).

56. Paszke, A. *et al.* in *Advances in Neural Information Processing Systems 32* 8024–8035 (Curran Associates, Inc., 2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

57. Rybakov, S., Lotfollahi, M., Theis, F. J. & Wolf, F. A. *Learning interpretable latent autoencoder representations with annotations of feature sets* Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article (Dec. 3, 2020), 2020.12.02.401182. https://www.biorxiv.org/content/10.1101/2020.12.02.401182v1 (2022).

58. Janizek, J. D., Sturmfels, P. & Lee, S.-I. Explaining Explanations: Axiomatic Feature Interactions for Deep Networks. *Journal of Machine Learning Research* **22,** 1–54. http://jmlr.org/papers/v22/20-1223.html (2021).

59. Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M. & Lee, S.-I. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence* **3,** 620–631 (2021).

60. Sturmfels, P., Lundberg, S. & Lee, S.-I. Visualizing the impact of feature attribution baselines. *Distill* **5,** e22 (2020).

61. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550. ISSN: 0027-8424, 1091-6490. http://www.pnas.org/cgi/doi/10.1073/pnas.0506580102 (2021) (Oct. 25, 2005).

62. Weinberger, E., Lin, C. & Lee, S.-I. Isolating salient variations of interest in single-cell transcriptomic data with contrastiveVI. *bioRxiv* (2021).

63. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

64. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9** (2008).

65. Rand, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* **66,** 846–850 (1971).

66. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21,** 3439–3440 (2005).

67. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols* **4,** 1184–1191 (2009).

68. Wheeler, D. L. *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* **36,** D13–D21 (2007).

69. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215,** 403–410 (1990).

70. Sutphin, G. L. & Kaeberlein, M. Measuring Caenorhabditis elegans life span on solid media. *JoVE (Journal of Visualized Experiments),* e1152 (2009).

71. Steinkraus, K. A. *et al.* Dietary restriction suppresses proteotoxicity and enhances longevity by an hsf-1-dependent mechanism in Caenorhabditis elegans. *Aging cell* **7,** 394–404 (2008).

72. McColl, G. *et al.* Utility of an improved model of amyloid-beta (Aβ1-42) toxicity in Caenorhabditis elegansfor drug screening for Alzheimer's disease. *Molecular neurodegeneration* **7,** 1–9 (2012).

73. Brenner, S. The genetics of Caenorhabditis elegans. *Genetics* **77,** 71–94 (1974).

74. Pitt, J. N. *et al.* WormBot, an open-source robotics platform for survival and behavior analysis in C. elegans. *GeroScience* **41,** 961–973 (2019).

75. Mantel, N. *et al.* Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* **50,** 163–170 (1966).

76. Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34,** 187–202 (1972).

77. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49,** D605–D612. ISSN: 1362-4962 (D1 Jan. 8, 2021).
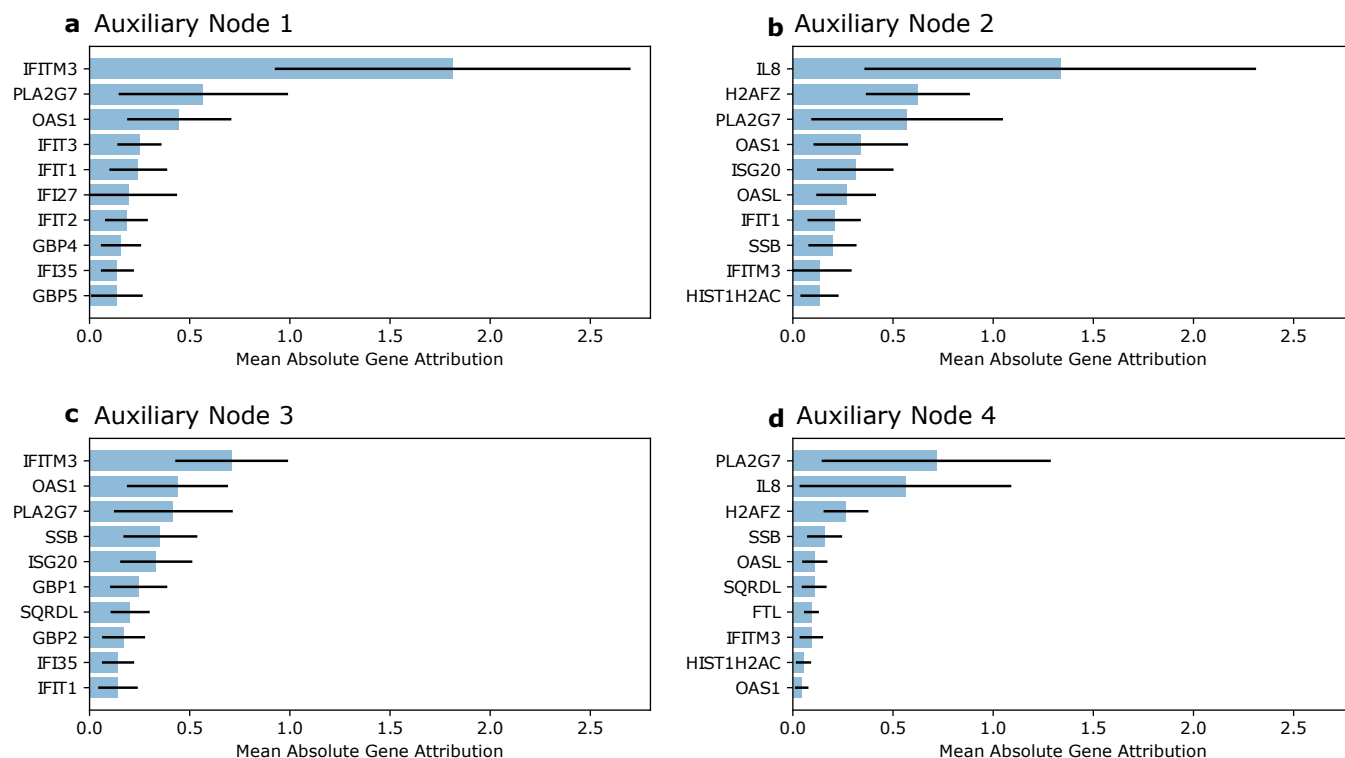
Supplementary Data

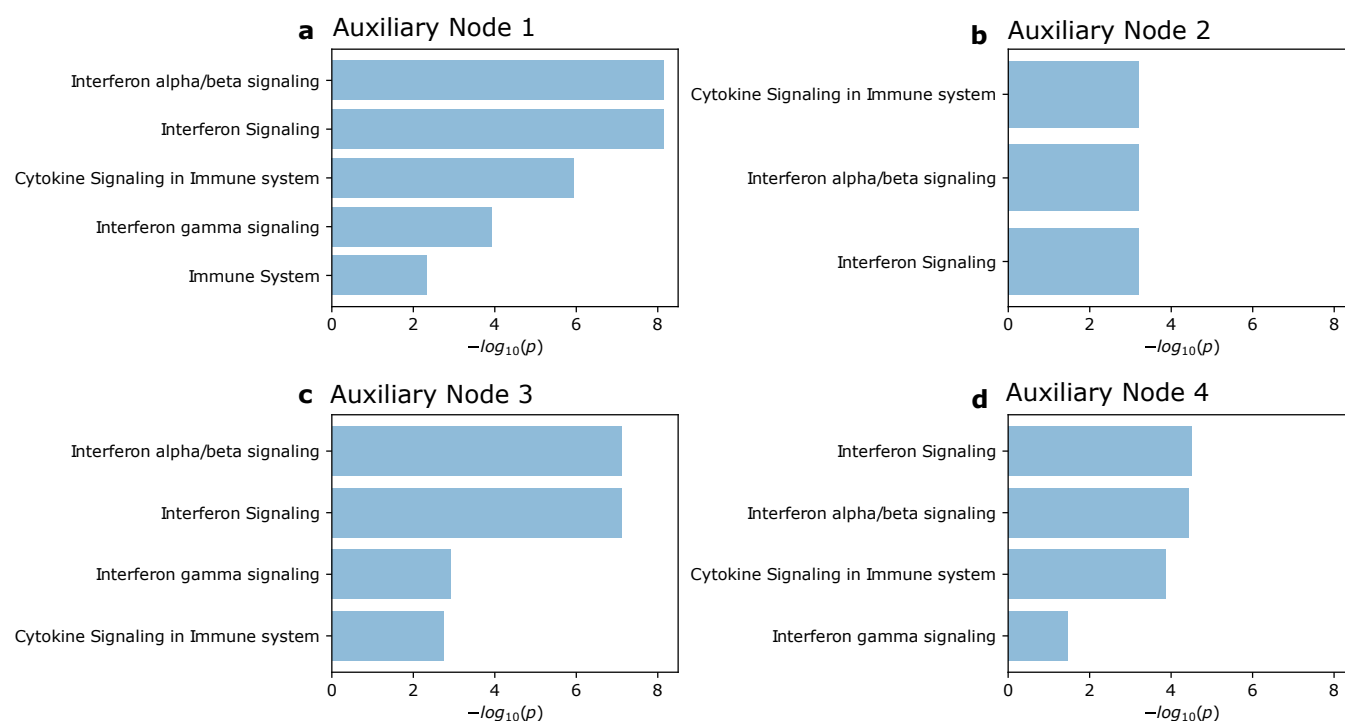**Supplementary Fig. 1 | Impute benchmark for VAE architecture with a linear decoder. a**, As top pathways ranked by our PAUSE Loss Attribution, ranked by the L2 norm of the weights in the decoder, or ranked randomly are removed by mean imputation, the reconstruction error (MSE) of a VAE with a linear decoder trained on the PBMC dataset [26] increases. Error bars indicate the standard deviation when the benchmark is re-run over 10 independent randomizations of the dataset into train/val/test splits. **b**, Summary of the area under the imputation curves in **a** for loss attribution and L2 norm. The boxes mark the quartiles (25th, 50th, and 75th percentiles) of the distribution, while the whiskers extend to show the minimum and maximum of the distribution (excluding outliers). Loss attribution identifies important pathways significantly better than the L2 norm (Wilcoxon rank-sums test statistic = 2.19, p = 0.028). **c,d**, These subpanels show the same information as **a,b**, where this time the benchmark (again using a VAE with a linear decoder) is run on the dataset of stimulated Jurkat cells [28]. Again, PAUSE identifies important pathways significantly better than the L2 norm (Wilcoxon rank-sums test statistic = 1.97, p = 0.049).

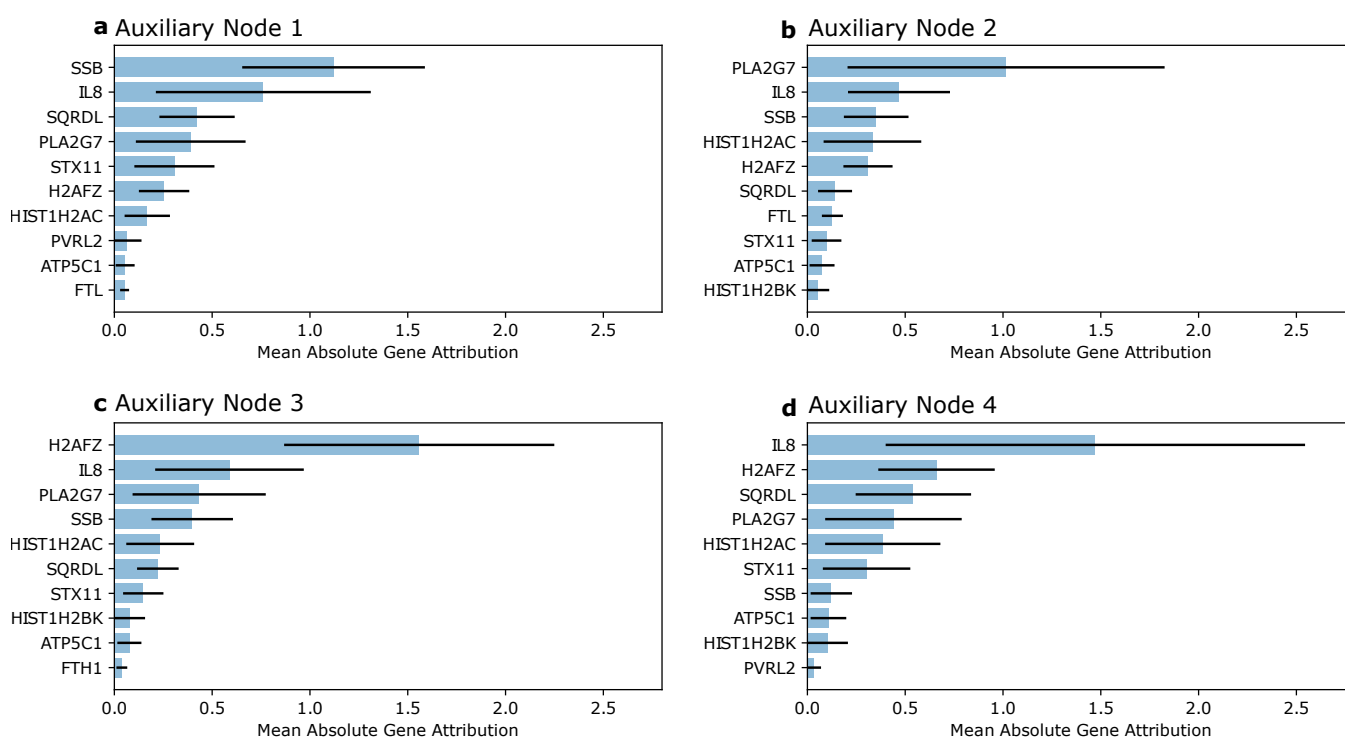**Supplementary Fig. 2 | Summary plot of top genes for TCR Signaling Module.** The four summary plots show swarmplots of the distributions of gene attributions for the four nodes in the TCR Signaling module for a pmVAE model trained on the single cell dataset of CD3/CD28-stimulated Jurkat cells. Each point is a single sample, the x-axis of each plot shows the attribution for that gene for that sample, and the color encoding shows whether the expression level of that gene is high or low in the sample.
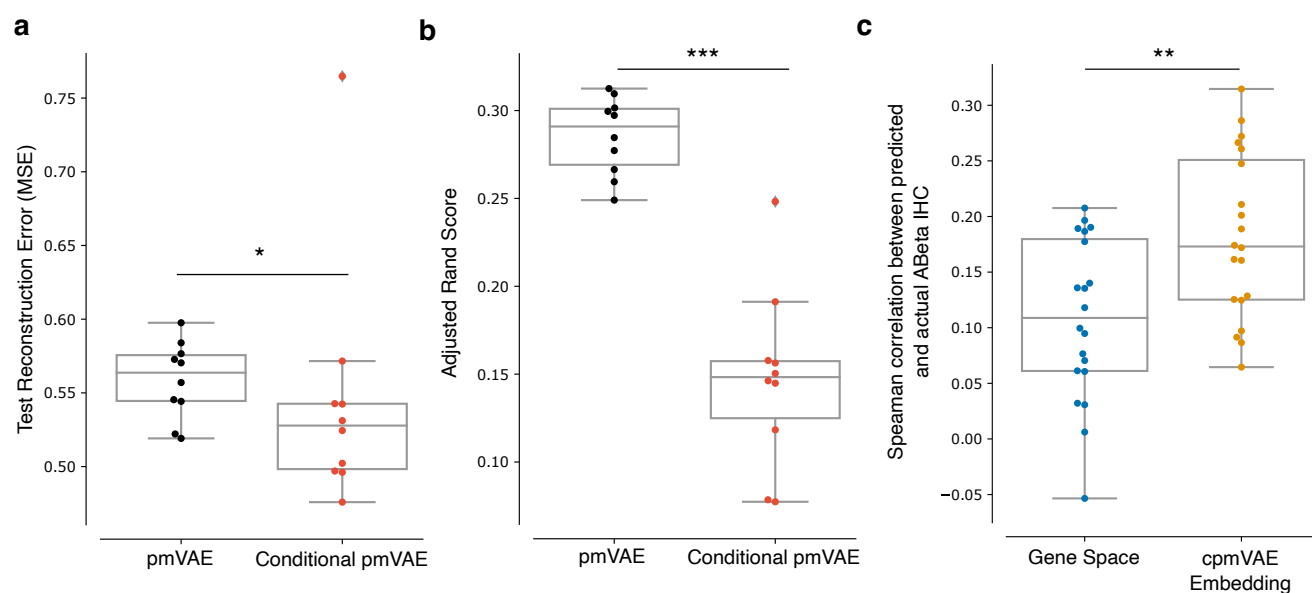
Supplementary Fig. 3 | Inteferon-related genes are identified as top genes in auxiliary nodes when Inteferon-related pathways are not included as pathway modules In this experiment, we trained the network without main Interferon-related pathways: ['REACTOME ANTIVIRAL MECHANISM BY IFN STIMULATED GENES', 'REACTOME DDX58 IFIH1 MEDIATED INDUCTION OF INTERFERON ALPHA BETA', 'REACTOME INTERFERON GAMMA SIGNALING', 'REACTOME INTERFERON ALPHA BETA SIGNALING', 'REACTOME INTERFERON SIGNALING'] for the PBMC single cell dataset. We then examined the densely connected auxiliary pathway, which is comprised of 4 pathway nodes. Shown here are the top 10 most important genes by mean absolute gene attribution value for each of the four nodes, with error bars representing standard deviation.
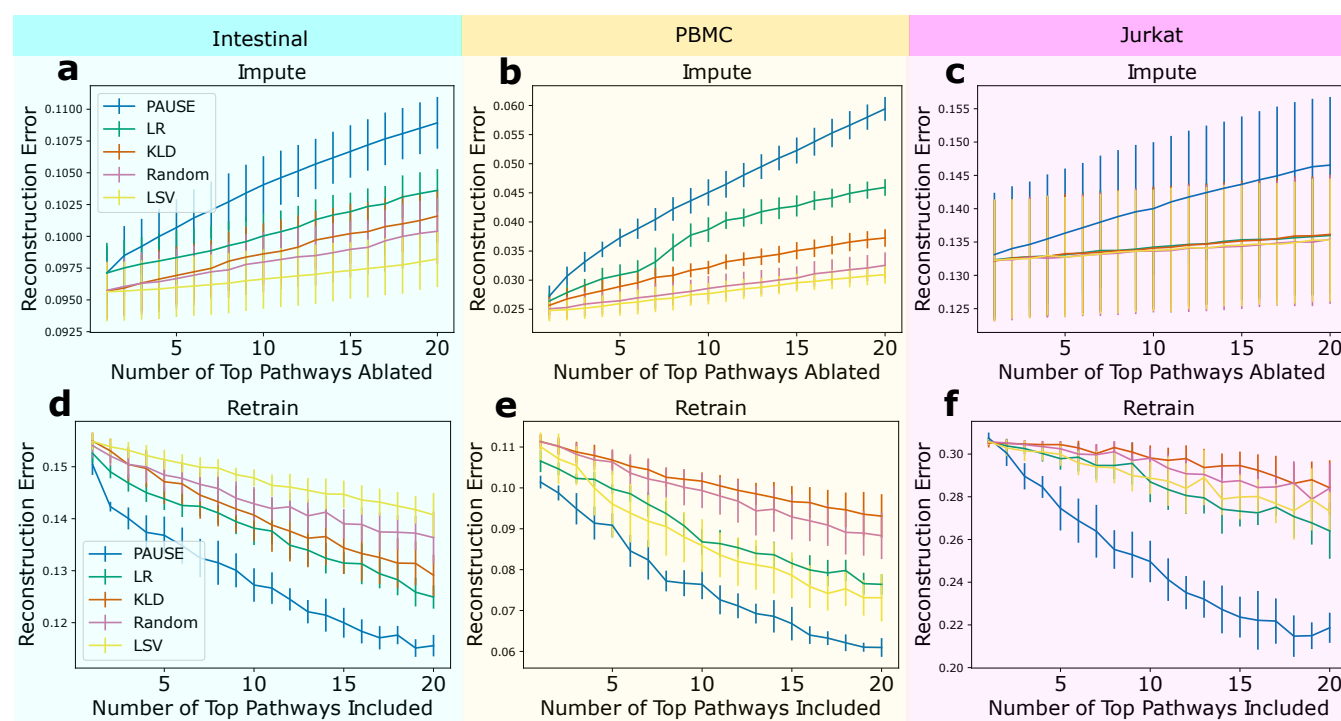
Supplementary Fig. 4 | Interferon-related pathway enrichments are found in auxiliary node top genes when Interferon-related pathways are not included as pathway modules. For each of the four auxiliary nodes (when the network is trained without Interferon pathway information), we generated gene-level attributions to obtain the top 50 most important genes for each node. We then used the STRING database [77] to analyze Reactome pathway enrichments for these top genes for each node compared to all genes in the dataset. In plots **a-d**, we show $-log_{10}$ p values for pathways with $p < 0.05$ when corrected for multiple testing using the Benjamini–Hochberg procedure.

Supplementary Fig. 5 | Genes from the pathway 'Regulation of PLK1 Activity at G2/M Transition' are not identified as top genes in auxiliary nodes when this pathway is not included as a pathway module. In this experiment, we trained the network without the pathway 'REACTOME REGULATION OF PK1 ACTIVITY AT G2/M TRANSITION' for the PBMC single cell dataset. We then examined the densely connected auxiliary pathway, which is comprised of 4 pathway nodes. Shown here are the top 10 most important genes by mean absolute gene attribution value for each of the four nodes, with error bars representing standard deviation.

Supplementary Fig. 6 | Sanity checks ensure biological validity of cpmVAE model on AD data. a, Comparison of reconstruction error attained by pmVAE model and our conditional pmVAE (cpmVAE) model on the AD dataset. For 10 different random train/test splits, we trained both a pmVAE and a cpmVAE on the training set for 50 epochs, measured the reconstruction on the held out test set for each model, and compared the distribution of test reconstruction errors for the two models. The distribution of test reconstruction errors attained with the cpmVAE model is significantly lower than the distribution of test reconstruction errors attained by the previously proposed pmVAE model (Wilcoxon rank-sums test statistic = 1.96, p = 0.049). **b**, Comparison of adjusted Rand indices (ARIs) of latent space clusters by dataset source between pmVAE and cpmVAE models. The distribution of ARIs for the pmVAE embeddings were significantly larger than the distribution of ARIs for the cpmVAE embeddings, indicating that the embedding learned by the cpmVAE model has significantly less clustering according to dataset source (Wilcoxon rank-sums test statistic = 3.78, p = 1.57×10^{-4}). **c**, For 20 random train/test splits of the AD data, we compared the Spearman correlation between predicted and actual Aβ density for linear models trained on the full gene expression measurements to the Spearman correlation between predicted and actual Aβ density for linear models trained only on the cpmVAE pathway embeddings. The cpmVAE embedding features lead to significantly more predictive models of Aβ density than the original gene expression features (Wilcoxon rank-sums test statistic = 2.81, p = 4.90×10^{-3}). In each plot, the boxes mark the quartiles (25th, 50th, and 75th percentiles) of the distribution, while the whiskers extend to show the minimum and maximum of the distribution (excluding outliers).

Supplementary Fig. 7 | Full plots of impute and retrain benchmarks for single cell datasets. Reconstruction error is shown for impute (top row) and retrain (bottom row) benchmarks for Intestinal **(a,d)**, PBMC **(b,e)** and Jurkat **(c,f)** single cell datasets. Bars represent standard deviation. Besides PAUSE loss attribution, the other methods shown here are logistic regression score (LR), Kullback–Leibler divergence (KLD), random ranking, and latent space variance (LSV).