pmVAE: Learning Interpretable Single-Cell Representations with Pathway Modules

Stefan G Stark [*1] Gilles Gut [*1 2] Gunnar Rätsch [1] Natalie R Davidson [1]

Abstract

Deep learning techniques have revolutionized the field of computational biology, however it is often difficult to assign biological meaning to their results. To improve interpretability, methods have incorporated biological priors, like pathway definitions, directly into the learning task. However, due to the correlated and redundant structure of pathways, it is difficult to determine an appropriate computational representation.
Here, we present **pathway module Variational Autoencoder** (pmVAE). Our method utilizes pathway information by restricting the structure of our VAE to mirror gene-pathway memberships. Its architecture is composed of a set of subnetworks, refered to as pathway modules, that learn interpretable multi-dimensional latent representations by factorizing the latent space according to pathway gene sets. We directly address correlations between pathways by balancing a module-specific local loss and a global reconstruction loss. We demonstrate that these representations are directly interpretable and reveal underlying biology, such as perturbation effects and cell type interactions. We compare pmVAE against two other state-of-the-art methods on a single-cell RNA-seq case-control dataset, and show that our representations are both more discriminative and specific in detecting the perturbed pathways.

Availability and implementation: https://github.com/ratschlab/pmvae

1. Introduction

Biological interpretation of high-throughput experiments is often laborious and difficult to fully automate. This issue is intensified for single-cell experiments since they are high-dimensional and have multiple interacting factors, such as cell-type specific drug effects. Pathways are one way to help attach biological meaning to computational results and conceptually disentangle the multiple factors that could be driving observed differences between samples.

A natural way to identify which pathways are altered in a dataset is to correlate the learned parameters against external pathway or clinical data to explain the latent components. While this approach has proven fruitful (Dincer et al., 2018; Kompa & Coker, 2020; Tan et al., 2014; Way & Greene, 2018; Way et al., 2020), it requires careful analysis to identify what each component is capturing, especially since all features are likely not fully disentangled (Locatello et al., 2019).

An alternative approach is to integrate prior information from biology, such as pathway gene sets, to constrain the model of interest (Buettner et al., 2017; Fortelny & Bock, 2020; Kuenzi et al., 2020; Ma et al., 2018; Mao et al., 2019; Rybakov et al., 2020; Svensson et al., 2020). In this approach, prior information both regularizes the model as well as increases its interpretability. While this strategy has successfully identified overarching pathway changes, it does not directly address an essential aspect of pathway definitions – they are highly correlated and overlapping. The overlap issue is significant when identifying the most precisely augmented pathway from a perturbation, not just large sets of highly correlated pathways.

An additional shortcoming of each pathway encoding scheme is that a single pathway has a unidimensional representation. Many higher-level pathways (e.g., the immune system) will contain possibly disparate signals from more specific and independent pathways (e.g., T-cell and B-cell signaling) and require a richer representation.

In this paper we present **pathway module VAE** (pmVAE), an unsupervised method to learn instantly interpretable and multidimensional pathway representations. Through incorporating pathways, defined as a bag-of-genes, pmVAE

[*] Equal contribution [1] Department of Computer Science, Zürich, Switzerland [2] Department of Biology, Zürich, Switzerland. Correspondence to: Natalie R Davidson <natalie.davidson@cuanschutz.edu>, Gunnar Rätsch <gunnar.ratsch@ratschlab.org>.

The 2021 ICML Workshop on Computational Biology.

constructs a pathway factorized latent space that directly addresses the problem of overlapping pathway definitions. Through pmVAE's resultant pathway-specific multidimensional representations, one can immediately determine cell-type-specific perturbation effects for any pathway of interest.

2. Methods

pmVAE extends the VAE framework (Kingma & Welling, 2013). VAEs are probabilistic models that learn compressed representations of high dimensional data. They consist of two sets of functions, an encoder and a decoder, often parameterized by neural networks, with parameter sets θ and ϕ respectively. These models learn distributions over low-dimensional latent variables, $\mathbf{z}$, referred to as embeddings or latent representations, from high dimensional input data, $\mathbf{x}$, by approximating the posterior over latent representations and maximizing a lower bound on the log-likelihood of the data. For a Gaussian likelihood $p(x|z)$, this is equivalent to minimizing

$$||\hat{\mathbf{x}} - \mathbf{x}||^2 + \beta \cdot \mathbb{KL}(q(z|x;\theta)||p(z))$$

where $\mathbb{KL}$ is the Kullback-Leibler divergence which regularizes the complexity of the embedding distribution. To make this optimization tractable, the posterior over latent representations, $q(z|x)$ is often approximated with an isotropic Gaussian distribution and the prior over latent representations, $p(z)$ is chosen to be a standard Gaussian.

2.1. Pathway modules produce pathway specific representations

The pathway modules within pmVAE construct a latent space factorized by pathways. A graphical representation of the model is shown in Figure 1. Given a set of K pathways, each represented as a set of genes, pmVAE consists of K pathway modules, which each behave as a VAE constrained to the set of genes that participate in its pathway. The outputs of these modules are then combined to reconstruct the expression vector of a single cell.

Let N be the number of total genes, N_p be the number of genes in pathway p, $\mathbf{x}^{(p)}$ be the expression of the genes participating in pathway p and let θ_p and ϕ_p be the parameters of the encoder and decoder within the pathway p module. Then the pathway p module encodes $\mathbf{x}^{(p)}$ into a pathway-specific embedding $\mathbf{z}^{(p)}$, which is then decoded into the reconstruction vector $\hat{\mathbf{x}}^{(p)}$. A global embedding vector, $\mathbf{z}$ is obtained by concatenation over all local embeddings provided by the pathway modules, i.e. $q(z|x) = \prod_p q(z^{(p)}|x^{(p)})$ and a global reconstruction is obtained by summing over the local reconstructions provided by each module, achieved in practice by connecting the outputs of each module to the set of genes participating in the pathway (see Figure 1).

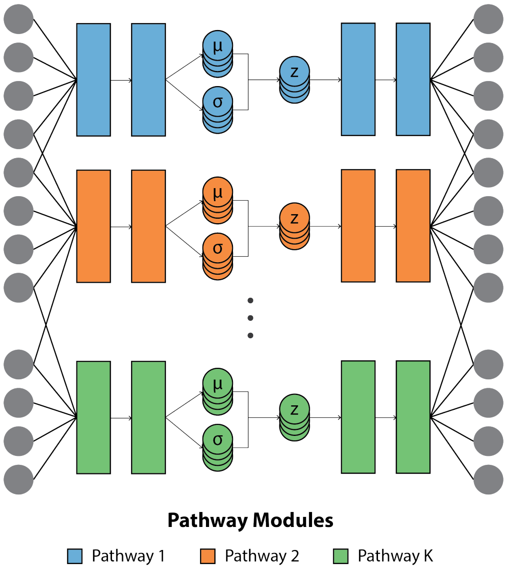

Figure 1. pmVAE is a variational autoencoder for expression data that constructs an interpretable latent space factorized by pathway gene sets. These pathway modules encode and decode the genes contained in their gene sets, forming a latent space for each pathway. A global reconstruction is achieved by summing over all pathway module outputs and a custom training procedure is implemented to address optimization challenges caused by overlapping pathways.

pmVAE minimizes the loss function:

$$||\hat{\mathbf{x}}-\mathbf{x}||^2+\frac{1}{K}\sum_p \frac{N}{N_p}||\hat{\mathbf{x}}^{(p)}-\mathbf{x}^{(p)}||^2+\beta\cdot\mathbb{KL}(q(z|x)||p(z))$$

which consists of the usual global reconstruction and $\mathbb{KL}$ terms (the first and last of Equation 2.1), but also introduces a set of local reconstruction terms (middle).

Our method attempts to balance the benefits of both local and global loss terms. Local reconstruction terms force each module to reconstruct the genes in its pathway independently. However, this approach will find all the modules explaining any variability in the data, even if another module better describes it. pmVAE attempts to explain the data with the most concise set of relevant modules to explain the data. While a global loss will provide the most concise data representation, it causes degenerate optimization problems, since pathway definitions are inherently redundant. For example, if an upstream (or otherwise overlapping) pathway module explains the signal of a targeted pathway, it will remove the signal of all downstream pathways. This leads to an *XOR*-type behavior (Fortelny & Bock, 2020). Through balancing both global and local loss, we resolve this *XOR*-type behavior and identify the most concise set of possibly overlapping pathways.

We compute the local reconstruction terms in practice by performing an additional K gradient steps, each computed on the parameters of exactly one module. This process is similar to an extreme dropout regularization technique (Srivastava et al., 2014), where all but one module is dropped out in the forward pass. To prevent the model from favoring large pathways, each local reconstruction term is weighted by pathway size relative to the global gene size $\frac{N}{N_p}$.

3. Results

To provide clear, quantitative comparisons against other methods, we consider a perturbation dataset with known effects, provided by Kang et al. (2018). This dataset is composed of 13,576 peripheral blood mononuclear cells from eight lupus patients, with and without Interferon-β stimulation. Using pathways as defined by Reactome v4 (Fabregat et al., 2018), the pathway *IFN-α/β Signaling* is a direct target of this perturbation. *Interferon Signaling* and *Cytokine Signaling in the Immune System* pathways both contain all genes in *IFN-α/β Signaling* plus additional child pathway genes. Therefore, the parent pathways are highly correlated with the perturbation, but contain less specific signal than the target pathway. *Anti viral mechanism by IFN-stimulated genes* is a downstream signaling target. Its gene set is not a subset of *IFN-α/β Signaling* (Jaccard similarity 0.075, after preprocessing).

We compare against two factor analysis models, f-scLVM (Buettner et al., 2017) and Interpetable AE (Rybakov et al., 2020) which constrain factors to gene set memberships [1]. To demonstrate pmVAE's ability to remove redundant pathway signals, we include an independent module pmVAE variant where the global loss term is omitted.

3.1. pmVAE identifies the most relevant perturbed pathways

We analyze the discriminative power of each method's pathway scores to differentiate perturbed and control cells and expect the targeted pathway scores to be most discriminative and all other pathways to have limited or no power. To quantify this, we learn a logistic regression model to predict stimulation status using the trained embeddings and compute the accuracy of this model on unseen test data.

pmVAE correctly identifies the perturbed pathway, *IFN-α/β Signaling* (accuracy: 0.951) as the most discriminative pathway. All other methods, except f-scLVM, also find the targeted pathway scores to be discriminative. The contrast between pmVAE and the comparator methods becomes apparent when we consider the two upstream pathways *Interferon signaling* and *Cytokine signaling in the immune system*. pmVAE does not find the upstream pathway scores as discriminative because the directly targeted pathways already explain the variability across the cells. In contrast, the independent VAE and interpretable AE find the upstream pathways as highly discriminative of perturbation status; f-scLVM only finds the most upstream pathway to be discriminative. Figure 2 shows the redundant pathways that independent VAE identifies as discriminative, but pmVAE does not. This finding confirms that pmVAE identifies non-redundant pathway representations by explaining the perturbation effect with only the most specific pathways.

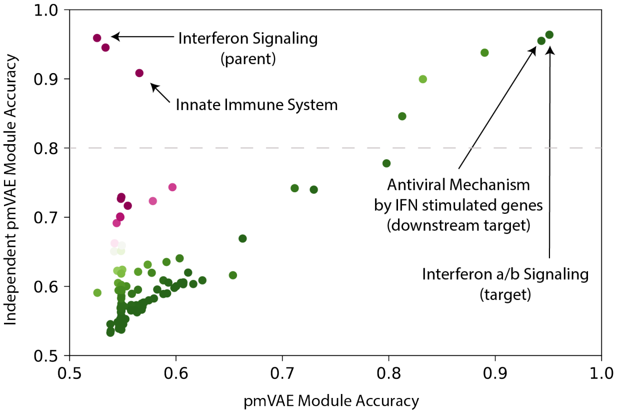

Figure 2. Accuracy of pmVAE modules against its Independent pmVAE variant to classify perturbation status. Each point is a single pathway module. Off-diagonal points are colored purple and indicate that these modules are significantly more discriminative in the independent variant. Modules above the dashed line are analyzed in Table 1

pmVAE also finds three additional pathways to be discriminative of the cell's perturbation state that is neither up- nor downstream of the targeted pathways. We believe that these biological processes are also discriminative because cell-type and perturbation status are entangled signals within our dataset. Kang et al. (2018) show empirically that the Interferon β stimulation has a cell-type-specific effect. We visualize this effect in our cell embeddings in the next section.

3.2. Multidimensional pathway embeddings capture cell-type-specific effects

In Section 3.1, we quantitatively demonstrated that pmVAE best identified the targetted pathway of interest in two distinct datasets. This section demonstrates that our representations capture additional relevant signals crucial to the

[1] We extended f-scLVM and Interpretable AE to use four-dimensional representations for completeness, but found similar or worse performance. Results are omitted due to space.

	pmVAE	Independent pmVAE	Interpretable AE	fscLVM
Interferon Alpha Beta Signaling	**0.951**	**0.964**	0.821	0.435
Antiviral Mechanism By Ifn Stimulated Genes	**0.943**	**0.955**	0.741	0.532
Rig I Mda5 Mediated Induction Of Ifn Alpha Beta...	0.890	**0.938**	0.820	0.530
Immune System	0.832	**0.900**	0.473	0.540
Interferon Gamma Signaling	0.813	0.846	0.558	0.540
Innate Immune System	0.566	**0.908**	0.614	0.529
Cytokine Signaling In Immune System	0.534	**0.945**	**0.961**	0.819
Interferon Signaling	0.526	**0.959**	**0.920**	0.506

Table 1. Accuracy of perturbation status prediction using pathway representations from selected relevant pathways (above dashed line in Figure 2). Pathways are sorted by the discriminative ability of pmVAE and highly discriminative pathways (> 0.9 accuracy) are bolded.

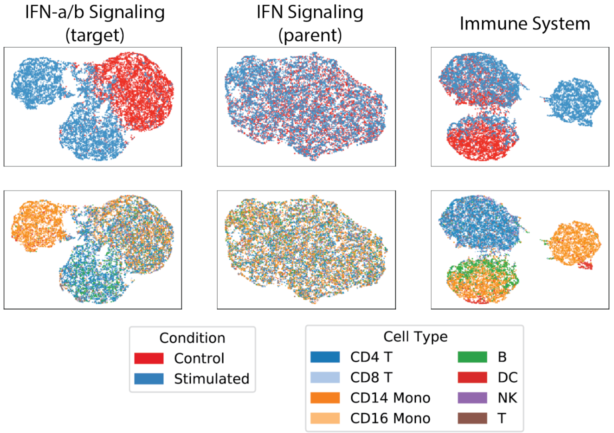

Figure 3. UMAP projections computed on selected pmVAE module embeddings (columns). Embeddings are colored by perturbation status (first row) and cell type (second row) and capture pathway-specific effects. Three modules are selected: 1) target, *IFN-α/β Signaling*, 2) parent *Interferon Signaling*, and 3) *Immune System*.

underlying biological process. In contrast to interpretable AE and f-scLVM, pmVAE provides a per-pathway multidimensional representation. This representation enables us to interrogate pathway-specific biological features, such as cell-type-specific effects.

The 4d embeddings of selected modules are visualized using UMAP projections (McInnes et al., 2018), shown in Figure 3. The targeted *IFN-α/β Signaling* module captures the known cell-type specific effects of IFN-β stimulation (Henig et al., 2013; van Boxel-Dezaire et al., 2010); cell types cluster strongly after stimulation, but are well mixed before. pmVAE removes the signal within the *Interferon Signaling* module, since it is already explained by the *IFN-α/β Signaling* subset. Finally, we show that the *Immune System* captures orthogonal cell-type effects of immune cells. However, monocytes retain some perturbation signal, which shows that we do not achieve a full disentanglement of effects.

4. Discussion

In this paper, we presented pmVAE, a method to learn highly specific, interpretable, multidimensional pathway representations. By incorporating pathway membership into the architectural design, pmVAE constructs a latent space factorized by pathway. This design enables direct association between the resulting pathway scores and clinically relevant features.

We have empirically shown on real scRNA-Seq datasets that pmVAE outperforms the two independent competitive models and individually trained VAEs. Our method is robust even when pathways are highly overlapping or correlated, a complication innate to most pathway structures. Furthermore, we demonstrate that our multidimensional pathway representations also capture cell-type, a pathway-specific biological signal. This enables quick interrogation of possible cell-type-specific perturbations.

Due in part to their hierarchical nature, redundancies between overlapping pathways result in degenerate solutions that make optimization challenging. pmVAE addresses this by enforcing independence relationships between pathway modules by introducing local reconstruction terms for each module's loss function. However, this independence ignores known pathway-pathway interactions arising from signaling effects. Learning pathway-factorized representations that explicitly model these effects, for example, by incorporating known signaling interactions into the architecture (Fortelny & Bock, 2020; Kipf & Welling, 2016; Ma et al., 2018; Scarselli et al., 2008) is an exciting direction of future work.

While we validated our method using a scRNA-seq data, we believe that our approach could work on other modalities individually or jointly, such as bulk RNA-seq, CyTOF, DIA mass spectrometry, etc. Since these technologies have limited direct feature correspondences, the integrative analysis of them is challenging (Irmisch et al., 2020; Lähnemann et al., 2020), yet the pathway structure underlying these feature sets is shared. Therefore, pathway-factorized latent representations, like those learned by pmVAE, could be used to more easily integrate (Cao et al., 2020; Demetci et al., 2020; Liu et al., 2019; Stark et al., 2020) these technologies.

References

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., and Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.*, 18(1):212, November 2017.

Cao, K., Bai, X., Hong, Y., and Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, 36(Supplement_1):i48–i56, 2020.

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv*, 2020.

Dincer, A. B., Celik, S., Hiranuma, N., and Lee, S.-I. Deepprofile: Deep learning of cancer molecular profiles for precision medicine. *bioRxiv*, pp. 278739, 2018.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2018.

Fortelny, N. and Bock, C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome Biol.*, 21(1):190, August 2020.

Henig, N., Avidan, N., Mandel, I., Staun-Ram, E., Ginzburg, E., Paperna, T., Pinter, R. Y., and Miller, A. Interferon-beta induces distinct gene expression response patterns in human monocytes versus t cells. *PloS one*, 8(4):e62366, 2013.

Irmisch, A., Bonilla, X., Chevrier, S., Lehmann, K.-V., Singer, F., Toussaint, N., Esposito, C., Mena, J., Milani, E. S., Casanova, R., et al. The tumor profiler study: integrated, multi-omic, functional tumor profiling for clinical decision support. *medRxiv*, 2020.

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C. M., et al. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89, 2018.

Kingma, D. P. and Welling, M. Auto-Encoding variational bayes. *arXiv*, December 2013.

Kipf, T. N. and Welling, M. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

Kompa, B. and Coker, B. Learning a latent space of highly multidimensional cancer data. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 25, pp. 379–390. World Scientific, 2020.

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., Ma, J., and Ideker, T. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*, 38(5):672–684.e6, November 2020.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

Liu, J., Huang, Y., Singh, R., Vert, J.-P., and Noble, W. S. Jointly embedding multiple single-cell omics measurements. *BioRxiv*, pp. 644310, 2019.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., and Ideker, T. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods*, 15 (4):290–298, April 2018.

Mao, W., Zaslavsky, E., Hartmann, B. M., Sealfon, S. C., and Chikina, M. Pathway-level information extractor (PLIER) for gene expression data. *Nat. Methods*, 16(7):607–610, July 2019.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. February 2018.

Rybakov, S., Lotfollahi, M., Theis, F. J., and Wolf, F. A. Learning interpretable latent autoencoder representations with annotations of feature sets. *bioRxiv*, 2020.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Stark, S. G., Ficek, J., Locatello, F., Bonilla, X., Chevrier, S., Singer, F., Rätsch, G., and Lehmann, K.-V. Scim: universal single-cell matching with unpaired feature sets. *Bioinformatics*, 36:i919–i927, 2020.

Svensson, V., Gayoso, A., Yosef, N., and Pachter, L. Interpretable factor models of single-cell rna-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, 2020.

Tan, J., Ung, M., Cheng, C., and Greene, C. S. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. In *Pacific Symposium on Biocomputing Co-Chairs*, pp. 132–143. World Scientific, 2014.

van Boxel-Dezaire, A. H., Zula, J. A., Xu, Y., Ransohoff, R. M., Jacobberger, J. W., and , G. R. Major differences in the responses of primary human leukocyte subsets to ifn-β. *The Journal of Immunology*, 185(10):5888–5899, 2010.

Way, G. P. and Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac Symp Biocomput*, 2018.

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