

Genes

Conditional out-of-distribution generation for unpaired data using transfer VAE

Mohammad Lotfollahi^{1,2}, Mohsen Naghipourfar¹, Fabian J. Theis^{1,2,3,*} and F. Alexander Wolf^{1,*}

¹Institute of Computational Biology, Helmholtz Center Munich, Neuherberg, Germany, ²School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany and ³Department of Mathematics, Technische Universität München, Munich, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: While generative models have shown great success in sampling high-dimensional samples conditional on low-dimensional descriptors (stroke thickness in MNIST, hair color in CelebA, speaker identity in WaveNet), their generation out-of-distribution poses fundamental problems due to the difficulty of learning compact joint distribution across conditions. The canonical example of the conditional variational autoencoder (CVAE), for instance, does not explicitly relate conditions during training and, hence, has no explicit incentive of learning such a compact representation.

Results: We overcome the limitation of the CVAE by matching distributions across conditions using maximum mean discrepancy in the decoder layer that follows the bottleneck. This introduces a strong regularization both for reconstructing samples within the same condition and for transforming samples across conditions, resulting in much improved generalization. As this amounts to solving a style-transfer problem, we refer to the model as *transfer VAE* (trVAE). Benchmarking trVAE on high-dimensional image and single-cell RNA-seq, we demonstrate higher robustness and higher accuracy than existing approaches. We also show qualitatively improved predictions by tackling previously problematic minority classes and multiple conditions in the context of cellular perturbation response to treatment and disease based on high-dimensional single-cell gene expression data. For generic tasks, we improve Pearson correlations of high-dimensional estimated means and variances with their ground truths from 0.89 to 0.97 and 0.75 to 0.87, respectively. We further demonstrate that trVAE learns cell-type-specific responses after perturbation and improves the prediction of most cell-type-specific genes by 65%.

Availability and implementation: The trVAE implementation is available via github.com/theislab/trvae. The results of this article can be reproduced via github.com/theislab/trvae_reproducibility.

Contact: fabian.theis@helmholtz-muenchen.de or alex.wolf@helmholtz-muenchen.de

1 Introduction

The task of generating high-dimensional samples x conditional on a latent random vector z and a categorical variable s has established solutions (Mirza and Osindero, 2014; Ren *et al.*, 2016). The situation becomes more complicated if the support of z is divided into domains d that come with different meanings: say $d \in \{\text{cat}, \text{dog}\}$ and one is interested in out-of-distribution (OOD) generation of samples x in a domain and condition (d, s) that are not part of the training data. Now, predicting how a given brown dog would look like with black fur becomes an OOD problem if the training data does not have observations of black dogs. To still have a chance of solving it, we assume training data with brown dogs, and brown *and* black cats. In an application with higher relevance, there is strong interest in how untreated humans ($s = 0, d = 0$) respond to drug treatment ($s = 1$) based on training data from human in vitro ($d = 1$) and in vivo mouse ($d = 2$) experiments. Hence, the target domain of interest ($d = 0$) does not offer training data for $s = 1$, but only for $s = 0$.

In this article, we suggest to address the challenge of generating samples OOD by regularizing the joint distribution across the categorical variable s using maximum mean discrepancy (MMD) in the framework of a conditional variational autoencoder (CVAE) (Sohn *et al.*, 2015). This produces a more compact representation of a cross-condition distribution that would otherwise display high variance in the standard CVAE. We will show that this leads to more accurate OOD prediction. MMD has proven successful in a variety of tasks. In particular, matching distributions with MMD in variational autoencoders (VAEs) (Kingma and Welling, 2013) has been suggested for unsupervised domain adaptation (Louizos *et al.*, 2015) or for learning statistically independent latent dimensions (Lopez *et al.*, 2018b). In supervised domain adaptation approaches, MMD-based regularization has been shown to be a viable strategy of learning label-predictive features that are stripped off of domain-specific information (Long *et al.*, 2015; Tzeng *et al.*, 2014). In these instances, however, MMD was used at the bottleneck layer, where it leads to different properties.

Matching distributions across perturbed and control populations has also been studied in the context of causal inference (Johansson *et al.*, 2016), albeit not in the context of generative modeling and OOD generation. Johansson *et al.* (2016) showed how to improve counterfactual inference by learning representations that enforce similarity between perturbed and control using a linear discrepancy measure, mentioning MMD as an alternative metric.

In further related work, the OOD generation problem was addressed via hard-coded latent space vector arithmetics (Lotfollahi *et al.*, 2019) and histogram matching (Amodio *et al.*, 2018). The approach of this article, however, introduces a data-driven end-to-end approach, which does not involve hard-coded elements and generalizes to more than one condition. We hope this work further stimulates the recent success of generative models in single-cell biology (Eraslan *et al.*, 2019; Lopez *et al.*, 2018a).

2 Materials and methods

2.1 Variational autoencoder

The motivation of the VAE (Kingma and Welling, 2013) is to provide a neural-network based parametrization for maximizing the likelihood

$$p_\theta(X|S) = \int p_\theta(X|Z, S) p_\theta(Z|S) dZ, \quad (1)$$

where X denotes a high-dimensional random variable, S a random variable representing conditions, θ the model parameters and $p_\theta(X|Z, S)$ the generative distribution that decodes Z into X . Here and in the following, we adapt the notation of Lopez *et al.* (2018b) while adapting the presentation of Doersch (2016).

To assign probability mass to values of Z that are likely to produce actually observed values of X , one introduces an encoding distribution q_ϕ , which can be related to p_θ via

$$\begin{aligned} \log p_\theta(X|S) - (q_\phi(Z|X, S) \| p_\theta(Z|X, S)) \\ = \mathbb{E}_{q_\phi(Z|X, S)} [\log p_\theta(X|Z, S)] - (q_\phi(Z|X, S) \| p_\theta(Z|S)). \end{aligned}$$

The right hand side of this equation provides the cost function ℓ_{VAE} for optimizing neural-network based parametrizations of p_θ and q_ϕ . The left-hand side describes the likelihood subtracted by an error term.

The case in which $S \neq \phi$ is referred to as the CVAE (Sohn *et al.*, 2015), and a straight-forward extension of the original framework (Kingma and Welling, 2013), which treated $S \equiv \phi$.

2.2 Maximum-mean discrepancy

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, \mathcal{X} a separable metric space, $x : \Omega \rightarrow \mathcal{X}$ a random variable and $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ a continuous, bounded, positive semi-definite kernel with a corresponding reproducing kernel Hilbert space (RKHS) \mathcal{H} . Consider the kernel-based estimate of a distance between two distributions p and q over the random variables X and X' . Such a distance, defined via the canonical distance between their \mathcal{H} -embeddings, is called the MMD (Gretton *et al.*, 2012) and denoted $\ell_{\text{MMD}}(p, q)$, with an explicit expression:

$$\begin{aligned} \ell_{\text{MMD}}(X, X') = \frac{1}{n_0^2} \sum_{n,m} k(x_n, x_m) + \frac{1}{n_1^2} \sum_{n,m} k(x'_n, x'_m) \\ - \frac{2}{n_0 n_1} \sum_{n,m} k(x_n, x'_m), \end{aligned} \quad (2)$$

where the sums run over the number of samples n_0 and n_1 for x and x' , respectively. Asymptotically, for a universal kernel such as the Gaussian kernel $k(x, x') = e^{-\gamma \|x-x'\|^2}$, $\ell_{\text{MMD}}(X, X')$ is 0 if and only if $p \equiv q$. For the implementation, we use multi-scale RBF kernels defined as:

$$k(x, x') = \sum_{i=1}^l k(x, x'; \gamma_i) \quad (3)$$

where $k(x, x', \gamma_i) = e^{-\gamma_i \|x-x'\|^2}$ and γ_i is a hyper-parameter.

Addressing the domain adaptation problem, the ‘Variational Fair Autoencoder’ (VFAE) (Louizos *et al.*, 2015) uses MMD to match latent distributions $q_\phi(Z|s=0)$ and $q_\phi(Z|s=1)$ —where s denotes a domain—by adapting the standard VAE cost function ℓ_{VAE} according to

$$\begin{aligned} \mathcal{L}_{\text{VFAE}}(\phi, \theta; X, X', S, S') = \mathcal{L}_{\text{VAE}}(\phi, \theta; X, S) \\ + \mathcal{L}_{\text{VAE}}(\phi, \theta; X', S') \\ - \beta \ell_{\text{MMD}}(Z_{s=0}, Z'_{s=1}), \end{aligned} \quad (4)$$

where X and X' are two high-dimensional observations with their respective conditions S and S' .

In contrast to GANs (Goodfellow *et al.*, 2014) whose training procedure is notoriously hard due to the minmax optimization problem, training models using MMD or Wasserstein distance metrics is comparatively simple (Arjovsky *et al.*, 2017; Dziugaite *et al.*, 2015a; Li *et al.*, 2015) as only a direct minimization of a single loss is involved. It has been shown that MMD-based GANs have some advantages over Wasserstein GANs resulting in a simpler and faster-training algorithm with matching performance (Bińkowski *et al.*, 2018). This motivated us to choose MMD as a metric for implementing distribution matching as a regularization of a CVAE.

2.3 Transfer VAE

Let us adapt the following notation for the transformation within a standard CVAE: High-dimensional observations x and a scalar or low-dimensional condition s are transformed using f (encoder, corresponding to distribution q_ϕ) and g (decoder, corresponding to distribution p_θ), which are parametrized by weight-sharing neural networks, and give rise to predictors \hat{z} , \hat{y} and \hat{x} :

$$\hat{z} = f(x, s) \quad (5a)$$

$$\hat{y} = g_1(\hat{z}, s) \quad (5b)$$

$$\hat{x} = g_2(\hat{y}) \quad (5c)$$

where we distinguished the first (g_1) and the remaining layers (g_2) of the decoder $g = g_2 \circ g_1$ (Fig. 1). While z formally depends on s , it is commonly empirically observed $Z \perp\!\!\!\perp S$, that is, the representation z is disentangled from the condition information s . By contrast, the original representation typically strongly covaries with S : $X \not\perp\!\!\!\perp S$. The observation can be explained by admitting that an efficient z -representation, suitable for minimizing reconstruction and regularization losses, should be as free as possible from information about s . Information about s is directly and explicitly available to the decoder Equation 5b, and hence, there is an incentive to optimize the parameters of f to *only* explain the variation in x that is *not* explained by s . Experiments below demonstrate that indeed, MMD regularization on the *bottleneck layer* z does not improve performance.

However, even if z is completely free of variation from s , the y representation has a strong s component, $Y \not\perp\!\!\!\perp S$, which leads to a separation of $y_{s=1}$ and $y_{s=0}$ into different regions of their support \mathcal{Y} . In the standard CVAE, without any regularization of this y representation, a highly varying, non-compact distribution emerges across different values of s (Fig. 2). To compactify the distribution so that it displays only subtle, controlled differences, we impose MMD Equation 2 in the first layer of the decoder (Fig. 1). We assume that modeling y in the same region of the support of \mathcal{Y} across s forces learning common features across s where possible. The more of these common features are learned, the more accurately the transformation task will be performed and the higher are chances of successful OOD generation. Using one of the benchmark datasets introduced, below, we qualitatively illustrate the effect (Fig. 2).

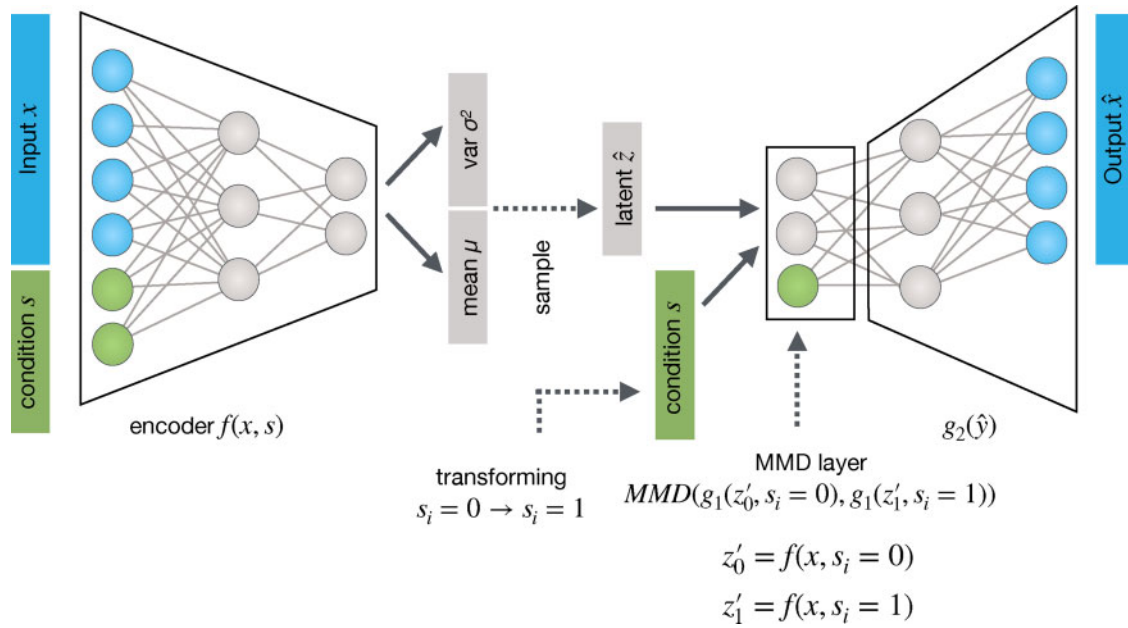


Fig. 1. Transfer VAE (trVAE) is an MMD-regularized conditional VAE. It receives randomized batches of data (x) and condition (s) as input during training, stratified for approximately equal proportions of s . In contrast to a standard CVAE, we regularize the effect of s on the representation obtained after the first-layer $g_1(\hat{z}, s)$ of the decoder g . During prediction time, we transform batches of the source condition $x_{s=0}$ to the target condition $x_{s=1}$ by encoding $\hat{z}_0 = f(x_0, s = 0)$ and decoding $g(\hat{z}_0, s = 1)$

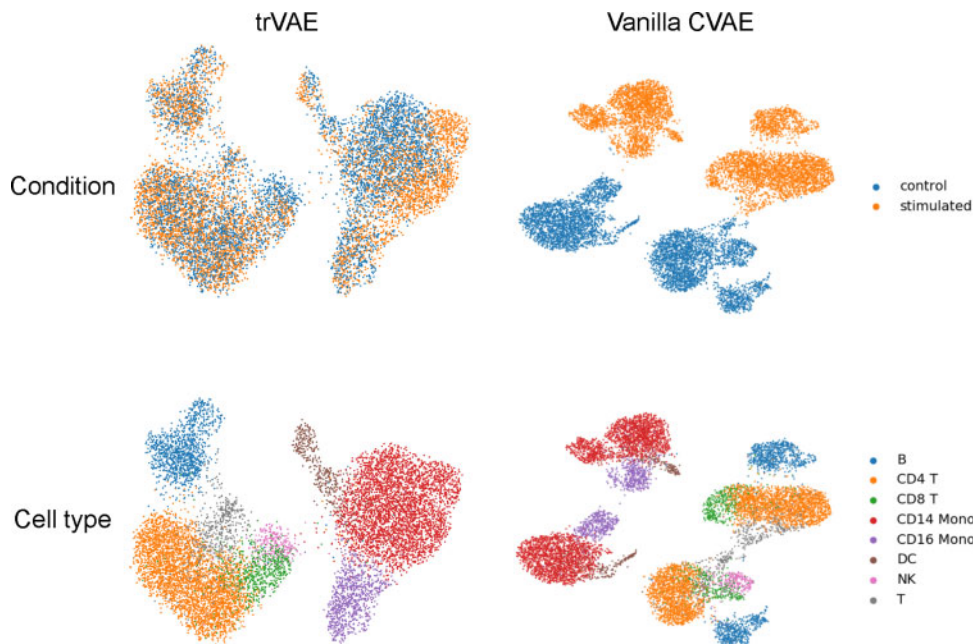


Fig. 2. Comparison of representations for MMD-layer in trVAE and the corresponding layer in the standard CVAE using UMAP (McInnes et al., 2018). The MMD regularization incentivizes the model to learn condition-invariant features resulting in a more compact representation. The figure shows the qualitative effect for the ‘PBMC data’ introduced in experiments section. Both representations show the same number of samples

During training time, all samples are passed to the model with their corresponding condition labels (x_s, s). At prediction time, we pass ($x_{s=0}, s = 0$) to the encoder f to obtain the latent representation $\hat{z}_{s=0}$. In the decoder g , we pass ($\hat{z}_{s=0}, s = 1$) and through that, let the model transform data to $\hat{x}_{s=1}$.

The cost function of trVAE derives directly from the standard CVAE cost function, as introduced in the backgrounds section,

$$\mathcal{L}_{\text{CVAE}}(\phi, \theta; X, S, \alpha, \eta) = \eta \mathbb{E}_{q_\theta(Z|X, S)} \log(p_\phi(X|Z, S)) - \alpha D_{\text{KL}}(q_\theta(Z|X, S) || p_\phi(Z|X, S)). \quad (6)$$

Consistent with the above, let $\hat{y}_{s=0} = g_1(f(x, s = 0), s = 0)$ and $\hat{y}_{s=1} = g_1(f(x', s = 1), s = 1)$. Through duplicating the cost function for X' and adding an MMD term, the loss of trVAE becomes:

$$\begin{aligned} \mathcal{L}_{\text{trVAE}}(\phi, \theta; X, X', S, S', \alpha, \eta, \beta) = & \mathcal{L}_{\text{CVAE}}(\phi, \theta; X, S, \alpha, \eta) \\ & + \mathcal{L}_{\text{CVAE}}(\phi, \theta; X', S', \alpha, \eta) \\ & - \beta \ell_{\text{MMD}}(\hat{Y}_{s=0}, \hat{Y}_{s'=1}). \end{aligned} \quad (7)$$

4 Results

We demonstrate the advantages of an MMD-regularized first layer of the decoder by benchmarking versus a variety of existing methods and alternatives:

- Standard CVAE (Sohn *et al.*, 2015)
- CVAE with MMD on bottleneck (MMD-CVAE), similar to VFAE (Louizos *et al.*, 2015)
- MMD-regularized autoencoder (Amodio *et al.*, 2019; Dziugaite *et al.*, 2015b)
- CycleGAN (Zhu *et al.*, 2017)
- scGen, a VAE combined with vector arithmetics (Lotfollahi *et al.*, 2019)
- scVI, a CVAE with a negative binomial output distribution (Lopez *et al.*, 2018a).

First, we demonstrate trVAE's basic OOD style transfer capacity on two established image datasets, on a qualitative level. We then address quantitative comparisons of challenging benchmarks with clear ground truth, predicting the effects of biological perturbation based on high-dimensional structured data. We used convolutional layers for imaging examples in section and fully connected layers for single-cell gene expression datasets in sections and. The optimal hyper-parameters for each application were chosen by using a parameter grid-search for each model.

4.1 MNIST and CelebA style transformation

Here, we use Morpho-MNIST (Castro *et al.*, 2018), which contains 60 000 images each of 'normal' and 'transformed' digits, which are drawn with a thinner and thicker stroke. For training, we used all normal-stroke data. Hence, the training data covers all domains ($d \in \{0, 1, 2, \dots, 9\}$) in the normal stroke condition ($s=0$). In the transformed conditions (thin and thick strokes, $s \in \{1, 2\}$), we only kept domains $d \in \{1, 3, 6, 7\}$.

We train a convolutional trVAE in which we first encode the stroke width via two fully connected layers with 128 and 784 features, respectively. Next, we reshape the 784-dimensional into $28 \times 28 \times 1$ images and add them as another channel in the image. Such trained trVAE faithfully transforms digits of normal stroke to digits of thin and thicker stroke to the OOD domains (Fig. 3).

Next, we apply trVAE to CelebA (Liu *et al.*, 2015), which contains 202 599 images of celebrity faces with 40 binary attributes for each image. We focus on the task of learning a transformation that turns a non-smiling face into a smiling face. We kept the smiling (s) and gender (d) attributes and trained the model with images from both smiling and non-smiling men but only with non-smiling women.

In this case, we trained a deep convolutional trVAE with a U-Net-like architecture (Ronneberger *et al.*, 2015). We encoded the binary condition labels as in the Morpho-MNIST example and fed them as an additional channel in the input.

Predicting OOD, trVAE successfully transforms non-smiling faces of women to smiling faces while preserving most aspects of the original image (Fig. 4). In addition to showing the model's capacity to handle more complex data, this example demonstrates the flexibility of the model adapting to well-known architectures like U-Net in the field.

4.2 Infection response

Accurately modeling cell response to perturbations is a key question in computational biology. Recently, neural network models have been proposed for OOD predictions of high-dimensional tabular

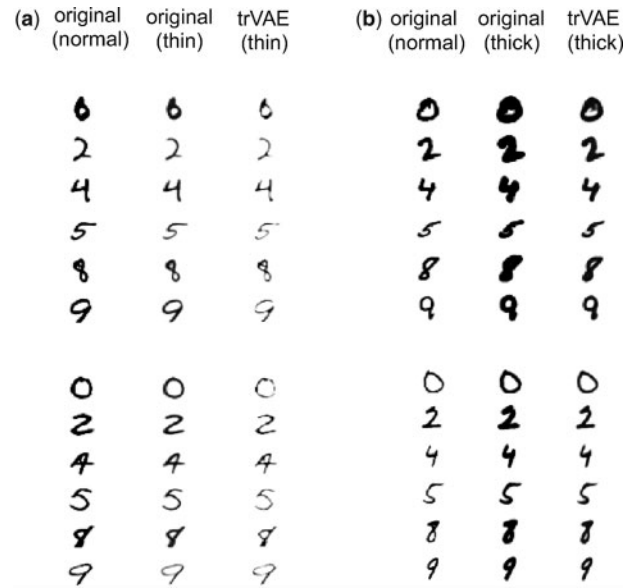


Fig. 3. OOD style transfer for Morpho-MNIST dataset containing normal, thin and thick digits. trVAE successfully transforms normal digits to thin (a) and thick (b) for digits not seen during training (OOD)

data that quantifies gene expression of single-cells (Amodio *et al.*, 2018; Lotfollahi *et al.*, 2019). However, these models are not trained on the task relying instead on hard-coded transformations and cannot handle more than two conditions.

We evaluate trVAE on a single-cell gene expression dataset that characterizes the gut (Haber *et al.*, 2017) after Salmonella or Heligmosomoides polygyrus (H. poly) infections, respectively. For this, we closely follow the benchmark as introduced in Lotfollahi *et al.* (2019). The dataset contains eight different cell types in four conditions: control or healthy cells ($n=3240$), H.Poly infection after three days (H.Poly.Day3, $n=2121$), H.poly infection after 10 days (H.Poly.Day10, $n=2711$) and salmonella infection ($n=1770$) (Fig. 5a). The normalized gene expression data has 1000 dimensions corresponding to 1000 genes. Since three of the benchmark models are only able to handle two conditions, we only included the control and H.Poly.Day10 conditions for model comparisons. In this setting, we hold out Tuft infected cells for training and validation, as these constitute the hardest case for OOD generalization (least shared features, few training data).

Figure 5b and c shows trVAE accurately predicts the mean and variance for high-dimensional gene expression in Tuft cells. We compared the distribution of *Defa24*, the gene with the highest change after H.poly infection in Tuft cells, which shows trVAE provides better estimates for mean and variance compared to other models. Moreover, trVAE outperforms other models also when quantifying the correlation of the predicted 1000 dimensional x with its ground truth (Fig. 5e). In particular, we note that the MMD regularization on the *bottleneck layer* of the CVAE does not improve performance, as argued above.

In contrast to existing approaches, trVAE can handle multiple perturbations at the same time. To illustrate this, we performed another experiment by training eight different models holding out each of the eight cell types from all three conditions. trVAE accurately predicts all cell types across different perturbations (Fig. 5f). The ability to handle multiple perturbations enables analysis and prediction for large drug screening studies.

4.3 Stimulation response

Similar to modeling infection response as above, we benchmark on another single-cell gene expression dataset consisting of 7217 IFN- β stimulated and 6359 control peripheral blood mononuclear cells (PBMCs) from eight different human Lupus patients

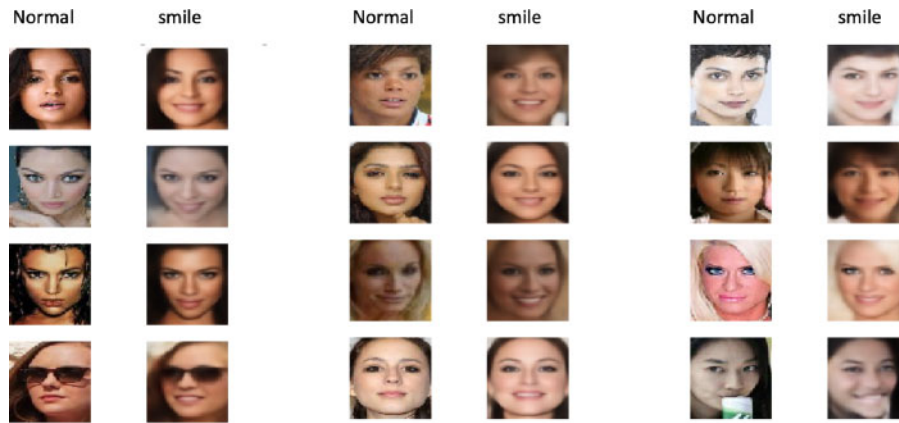


Fig. 4. CelebA dataset with images in two conditions: celebrities without a smile and with a smile on their face. trVAE successfully adds a smile on faces of women without a smile despite these samples completely lacking from the training data (OOD). The training data only comprises non-smiling women and smiling and non-smiling men

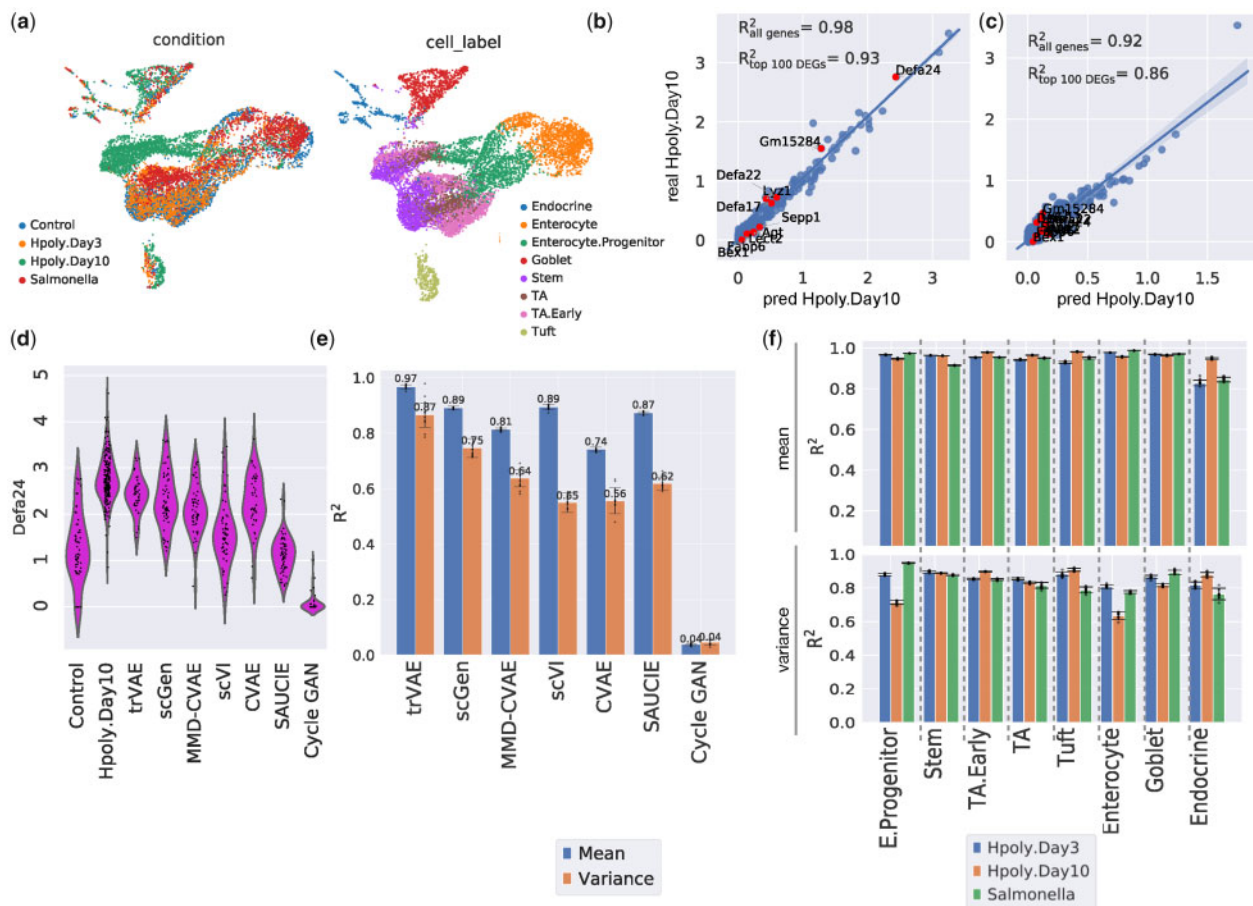


Fig. 5. (a) UMAP visualization of conditions and cell type for gut cells. (b and c) Mean and variance expression of 1000 genes comparing trVAE-predicted and real infected Tuft cells together with the top 10 differentially expressed genes highlighted in red (R^2 denotes Pearson correlation between ground truth and predicted values). (d) Distribution of *Defa24*: the top response gene to H.poly.Day10 infection between control, predicted and real stimulated cells for different models. Vertical axis: expression distribution for *Defa24*. Horizontal axis: control, real and predicted distribution by different models. (e) Comparison of Pearson's R^2 values for mean and variance gene expression between real and predicted cells for different models. Center values show the mean of R^2 values estimated using $n = 100$ random subsamples for the prediction of each model and error bars depict standard deviation. (f) Comparison of R^2 values for mean and variance gene expression between real and predicted cells by trVAE for the eight different cell types and three conditions. Center values show the mean of R^2 values estimated using $n = 100$ random subsamples for each cell type and error bars depict standard deviation

(Kang et al., 2018). The stimulation with IFN- β induces dramatic changes in the transcriptional profiles of immune cells, which causes big shifts between control and stimulated cells (Fig. 6a). We studied the OOD prediction of natural killer (NK) cells held out during the training of the model.

trVAE accurately predicts mean (Fig. 6b) and variance (Fig. 6c) for all genes in the held out NK cells. In particular, genes strongly responding to IFN- β (highlighted in red in Fig. 6b and c) are well captured. An effect of applying IFN- β is an increase in ISG15 for NK cells, which the model never sees during training. trVAE

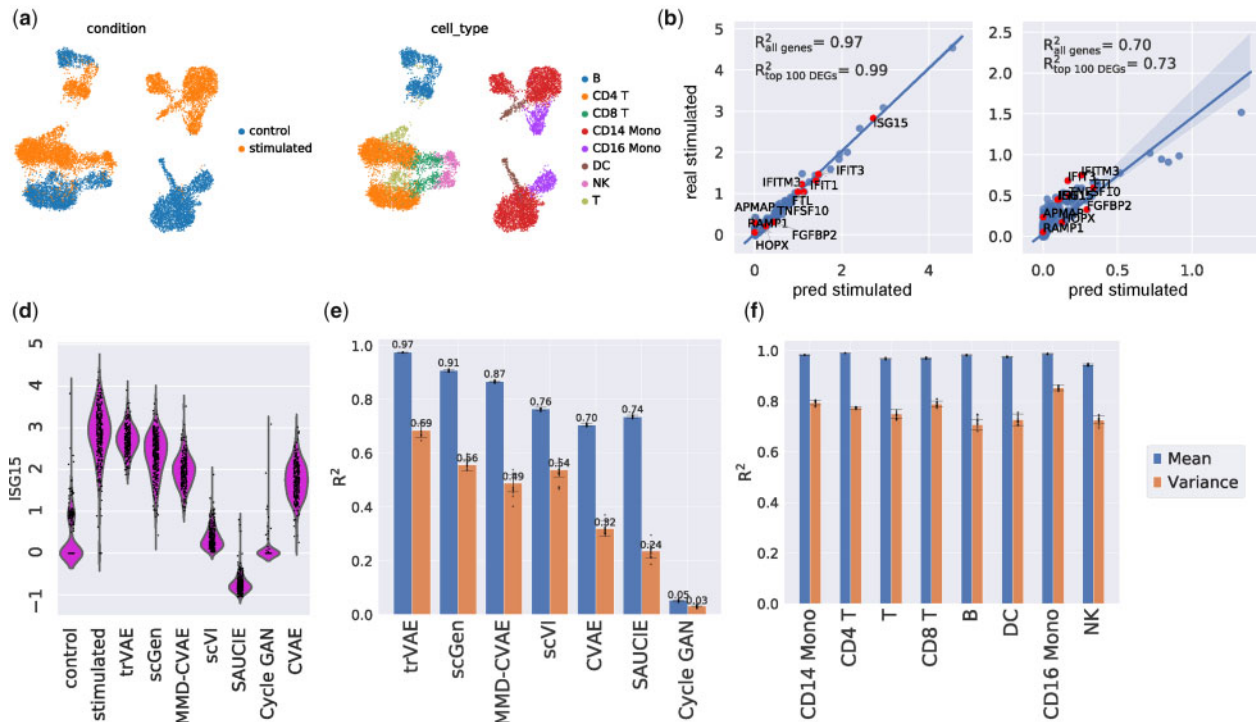


Fig. 6. (a) UMAP visualization of peripheral blood mononuclear cells (PBMCs). (b and c) Mean and variance per 2000 dimensions between trVAE-predicted and real natural killer cells (NK) together with the top 10 differentially expressed genes highlighted in red. (d) Distribution of *ISG15*: the most strongly changing gene after IFN- β perturbation among control, real and predicted stimulated cells for different models. Vertical axis: expression distribution for *ISG15*. Horizontal axis: control, real and predicted distribution by different models. (e) Comparison of R^2 values for mean and variance gene expression between real and predicted cells for different models. Center values show the mean of R^2 values estimated using $n = 100$ random subsamples for the prediction of each model and error bars depict standard deviation. (f) Comparison of R^2 values for mean and variance gene expression between real and predicted cells by trVAE for eight different cells in the study. Center values show the mean of R^2 values estimated using $n = 100$ random subsamples for each cell type and error bars depict standard deviation

predicts this change by increasing the expression of *ISG15* as observed in real NK cells (Fig. 6d). A cycle GAN and an MMD-regularized auto-encoder (SAUCIE) and other models yield less accurate results than our model. Comparing the correlation of predicted mean and variance of gene expression for all dimensions of the data, we find trVAE performs best (Fig. 6e). To demonstrate the generality of our method we trained seven other models, removing stimulated cells for each of seven different cell types in the study. Our model robustly predicted all other seven cell types (Fig. 6f).

The specificity of perturbation responses of cells depends on many factors leading to changes in gene expression levels that are either shared across all types or specific to some. Predicting both groups of responses is necessary to address questions such as which cell types are most responsive to a perturbation, and successful drug dose prediction (Hu *et al.*, 2020; Srivatsan *et al.*, 2020).

trVAE can capture specific responses after IFN- β when any of the cell types is absent from training and afterward predicted. To demonstrate this, we scored the specificity of differentially expressed genes (DEGs) after IFN- β stimulation using a median-based score (see Supplementary Methods). trVAE successfully predicts top 10 most cell-type-specific responding genes (Fig. 7a). Specifically, our model predicted the up-regulation of *CCL8*, a *CD14-Mono* specific response gene after IFN- β . As another example, trVAE not only predicted the up-regulation of *ISG15* as a shared response gene but also captured the specific expression pattern of this gene across different cell types. Next, we compared our approach with the state-of-the-art model (scGen) for this task using the top 250 most cell-type-specific DEGs. Our model improves the mean error on the first and the second top 50 specific DEGs by 65% and 44%, respectively (Fig. 7b and c). Further comparison demonstrated that trVAE not only outperforms scGen but also all other benchmarked methods (Supplementary Figs S1 and S2).

5 Discussion

By arguing that the standard CVAE yields representations in the first layer following the bottleneck that vary strongly across categorical conditions, we introduced an MMD regularization that forces these representations to be similar across conditions. The resulting model (trVAE) outperforms existing modeling approaches on benchmark and real-world datasets.

Within the bottleneck layer, CVAEs already display a well-controlled behavior and regularization does not improve performance. Further regularization at later layers might be beneficial but is numerically costly and unstable as representations become high-dimensional. However, we have not yet systematically investigated this and leave it for future studies.

We have evaluated the predictive power of trVAE by leaving out one cell type and trying to predict it in cases in which the training data contains cell types that are rather similar to the targeted OOD cells (Lotfollahi *et al.*, 2019). Further evaluation is needed when OOD samples are very different from the training data. Also, further studies are required to understand the uncertainty quantification inherent to the probabilistic nature of the model. Finally, we note that architectures related to Gaussian mixture VAEs or GANs may be considered as alternatives to the MMD regularization.

The ability to analyze and predict multiple perturbations allow trVAE to be applied to experiments with many biological conditions. Specifically, recent advances in massive single-cell compounds screening (Srivatsan *et al.*, 2020) provide great potential to exploit our model for further experimental design and the study of interaction effects among different drugs. Future conceptual investigations concern establishing connections to causal-inference-inspired models beyond (Johansson *et al.*, 2016) such as CEVAE (Louizos *et al.*, 2017), establishing further that faithful modeling of an

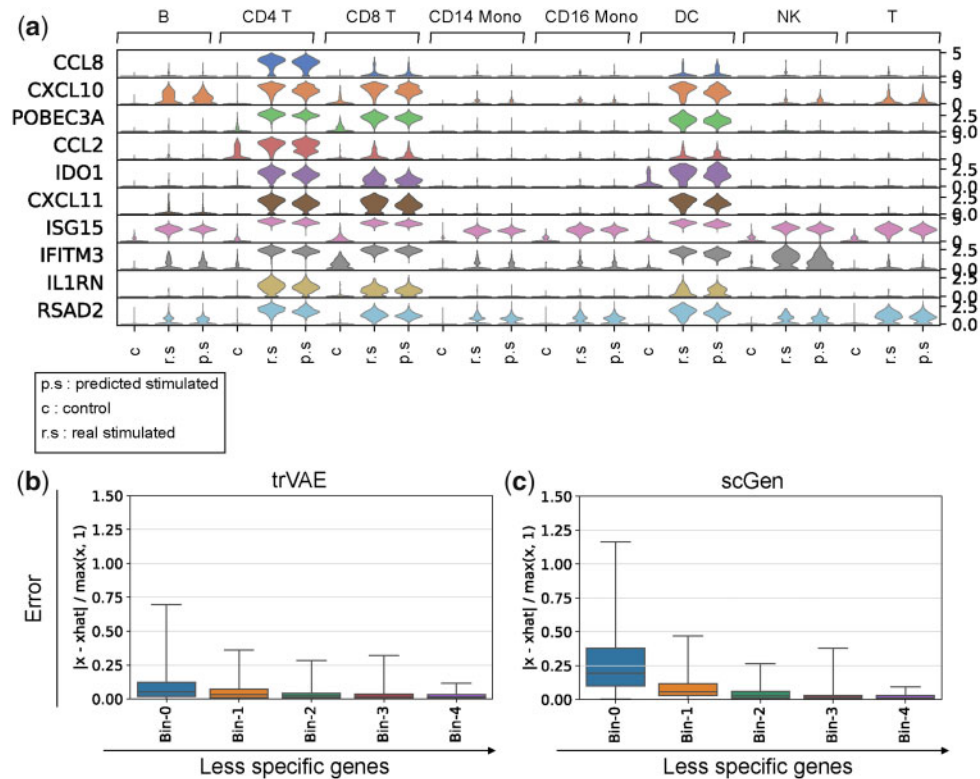


Fig. 7. (a) Violin plot for top 10 specific response genes after IFN- β stimulation out of 250 DEGs according to the gene specificity score across control (c), real stimulated (r.s) and predicted stimulated (p.s) for different cell types. Vertical axis: expression distribution for top specific genes. Horizontal axis: control, real and predicted distribution by trVAE for different cell types. (b and c) Box plots of top 500 DEGs ordered by the gene specificity score. Each bin is composed of 50 genes and each point in the bin shows the error between average expression of that gene within a cell type and average prediction by trVAE and scGen for that cell type. In total, each boxplot has been derived from 50 (number of genes) \times 8 (number of cell types) points ($n = 400$). Box plots indicate the median (center lines), interquartile range (hinges) and whiskers represents min and max values

interventional distribution can be re-framed as successful perturbation effect prediction across domains.

Acknowledgements

The authors are grateful to Anna Klimovskaia for pointing us to the reference of Johansson *et al.* (2016) and to Romain Lopez for pointing us to problems with the background section on the VAE in the arXiv preprint version of this manuscript.

Financial Support: This work was supported by BMBF grant nos. 01IS18036A and 01IS18053A, by the German Research Foundation within the Collaborative Research Center 1243, Subproject A17, by the Helmholtz Association (Incubator grant sparse2big, grant no. ZT-I-0007) and by the Chan Zuckerberg Initiative DAF (advised fund of Silicon Valley Community Foundation, no. 182835). M.L. acknowledges financial support from the Joachim Herz Stiftung.

Conflict of Interest: F.J.T. reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and ownership interest in Cellarity, Inc. F.A.W. reports being a full-time employee of Cellarity Inc., and ownership interest in Cellarity, Inc.

Data availability

All of the data sets analyzed in this manuscript are public and published in other papers. We have referenced them in the manuscript and they are downloadable at github.com/theislab/trvae_reproducibility.

References

Amodio, M. *et al.* (2018) Out-of-sample extrapolation with neuron editing. *arXiv:1805.12198*.

- Amodio, M. *et al.* (2019) Exploring single-cell data with deep multitasking neural networks. *Nat. Methods*, **16**, 1139–1145.
- Arjovsky, M. *et al.* (2017) Wasserstein generative adversarial networks. In: Precup, D. and Teh, Y.W. (eds.) *Proceedings of the 34th International Conference on Machine Learning, Volume 70 of Proceedings of Machine Learning Research*. PMLR, International Convention Centre, Sydney, Australia, pp. 214–223.
- Bińkowski, M. *et al.* (2018) Demystifying mmd gans. *arXiv:1801.01401*.
- Castro, D.C. *et al.* (2018) Morpho-MNIST: quantitative assessment and diagnostics for representation learning.
- Doersch, C. (2016) Tutorial on variational autoencoders. *arXiv:1606.05908*.
- Dziugaite, G.K. *et al.* (2015a) Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI'15*, AUAI Press, Arlington, Virginia, United States. pages 258–267.
- Dziugaite, G.K. *et al.* (2015b) Training generative neural networks via maximum mean discrepancy optimization. *arXiv:1505.03906*.
- Eraslan, G. *et al.* (2019) Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, **10**, 390.
- Goodfellow, I. *et al.* (2014) Generative adversarial nets. In: *Advances in Neural Information Processing Systems, Palais des Congrès de Montréal, Montréal, Canada*, pp. 2672–2680.
- Gretton, A. *et al.* (2012) A kernel two-sample test. *J. Mach. Learn. Res.*, **13**, 723–773.
- Haber, A.L. *et al.* (2017) A single-cell survey of the small intestinal epithelium. *Nature*, **551**, 333–339.
- Hu, Y. *et al.* (2020) Single-cell transcriptome mapping identifies common and cell-type specific genes affected by acute delta9-tetrahydrocannabinol in humans. *Sci. Rep.*, **10**, 3450.
- Johansson, F. *et al.* (2016) Learning representations for counterfactual inference. In: *International Conference on Machine Learning*, PMLR, vol. **48**, pp. 3020–3029.
- Kang, H.M. *et al.* (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.*, **36**, 89–94.

- Kingma,D.P. and Welling,M. (2013) Auto-encoding variational bayes. *arXiv:1312.6114*.
- Li,Y. *et al.* (2015) Generative moment matching networks. In: *International Conference on Machine Learning*, PMLR, pp. 1718–1727.
- Liu,Z. *et al.* (2015) Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
- Long,M. *et al.* (2015) Learning transferable features with deep adaptation networks. *arXiv:1502.02791*.
- Lopez,R. *et al.* (2018a) Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, **15**, 1053–1058.
- Lopez,R. *et al.* (2018b) Information constraints on auto-encoding variational bayes. In: *Advances in Neural Information Processing Systems*, pp. 6114–6125.
- Lotfollahi,M. *et al.* (2019) scGen predicts single-cell perturbation responses. *Nat. Methods*, **16**, 715–721.
- Louizos,C. *et al.* (2015) The variational fair autoencoder. *arXiv:1511.00830*.
- Louizos,C. *et al.* (2017) Causal effect inference with deep latent-variable models. In: *Advances in Neural Information Processing Systems*, Curran Associates, Inc., pp. 6446–6456.
- McInnes,L. *et al.* (2018) UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv:1802.03426*.
- Mirza,M. and Osindero, S. (2014) Conditional generative adversarial nets. *arXiv:1411.1784*.
- Ren,Y. *et al.* (2016) Conditional generative moment-matching networks. In: *Advances in Neural Information Processing Systems*, pp. 2928–2936. <http://papers.nips.cc/paper/6254-conditional-generative-moment-matching-networks>.
- Ronneberger,O. *et al.* (2015) U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, pp. 234–241.
- Sohn,K. *et al.* (2015) Learning structured output representation using deep conditional generative models. In: *Advances in Neural Information Processing Systems*, Vol. 28, Curran Associates, Inc., pp. 3483–3491.
- Srivatsan,S.R. *et al.* (2020) Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, **367**, 45–51.
- Tzeng,E. *et al.* (2014) Deep domain confusion: Maximizing for domain invariance. *arXiv:1412.3474*.
- Zhu,J.-Y. *et al.* (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *IEEE International Conference on Computer Vision (ICCV)*, ICCV.