nature methods

ARTICLES

https://doi.org/10.1038/s41592-019-0494-8

scGen predicts single-cell perturbation responses

Mohammad Lotfollahi[1,2], F. Alexander Wolf[1*] and Fabian J. Theis[1,2,3*]

Accurately modeling cellular response to perturbations is a central goal of computational biology. While such modeling has been based on statistical, mechanistic and machine learning models in specific settings, no generalization of predictions to phenomena absent from training data (out-of-sample) has yet been demonstrated. Here, we present scGen (https://github.com/theislab/scgen), a model combining variational autoencoders and latent space vector arithmetics for high-dimensional single-cell gene expression data. We show that scGen accurately models perturbation and infection response of cells across cell types, studies and species. In particular, we demonstrate that scGen learns cell-type and species-specific responses implying that it captures features that distinguish responding from non-responding genes and cells. With the upcoming availability of large-scale atlases of organs in a healthy state, we envision scGen to become a tool for experimental design through in silico screening of perturbation response in the context of disease and drug treatment.

Single-cell transcriptomics has become an established tool for the unbiased profiling of complex and heterogeneous systems[1,2]. The generated data sets are typically used for explaining phenotypes through cellular composition and dynamics. Of particular interest are the dynamics of single cells in response to perturbations, be it to dose[3], treatment[4,5] or the knockout of genes[6–8]. Although advances in single-cell differential expression analysis[9,10] have enabled the identification of genes associated with a perturbation, generative modeling of perturbation response takes a step further in that it enables the generation of data in silico. The ability to generate data that cover phenomena not seen during training is particularly challenging and referred to as out-of-sample prediction.

While dynamic mechanistic models have been suggested for predicting low-dimensional quantities that characterize cellular response[11,12], such as a scalar measure of proliferation, they face fundamental problems. These models cannot easily be formulated in a data-driven way and require temporal resolution of the experimental data. Due to the typically small number of time points available, parameters are often hard to identify. Resorting to linear statistical models for modeling perturbation response[6,8] leads to low predictive power for the complicated non-linear effects that single-cell data display. In contrast, neural network models do not face these limits.

Recently, such models have been suggested for the analysis of single-cell RNA sequencing (scRNA-seq) data[13–17]. In particular, generative adversarial networks (GANs) have been proposed for simulating single-cell differentiation through so-called latent space interpolation[16]. While providing an interesting alternative to established pseudotemporal ordering algorithms[18], this analysis does not demonstrate the capability of GANs for out-of-sample prediction. The use of GANs for the harder task of out-of-sample prediction is hindered by fundamental difficulties: (1) GANs are hard to train for structured high-dimensional data, leading to high-variance predictions with large errors in extrapolation, and (2) GANs do not allow for the direct mapping of a gene expression vector x on a latent space vector z, making it difficult or impossible to generate a cell with a set of desired properties. In addition, for structured data, GANs have not yet shown advantages over the simpler variational autoencoders (VAE)[19] (Methods).

To overcome these problems, we built scGen, which is based on a VAE combined with vector arithmetics, with an architecture adapted for scRNA-seq data. scGen enables predictions of dose and infection response of cells for phenomena absent from training data across cell types, studies and species. In a broad benchmark, it outperforms other potential modeling approaches, such as linear methods, conditional variational autoencoders (CVAE)[20] and style transfer GANs. The benchmark of several generative neural network models should present a valuable resource for the community showing opportunities and limitations for such models when applied to scRNA-seq data. scGen is based on Tensorflow[21] and on the single-cell analysis toolkit Scanpy[22].

Results

scGen accurately predicts single-cell perturbation response out-of-sample. High-dimensional scRNA-seq data is typically assumed to be well parametrized by a low-dimensional manifold arising from the constraints of the underlying gene regulatory networks. Current analysis algorithms mostly focus on characterizing the manifold using graph-based techniques[23,24] in the space spanned by a few principal components. More recently, the manifold has been modeled using neural networks[13–17]. As in other application fields[25,26], in the latent spaces of these models, the manifolds display astonishingly simple properties, such as approximately linear axes of variation for latent variables explaining a major part of the variability in the data. Hence, linear extrapolations of the low-dimensional manifold could in principle capture variability related to perturbation and other covariates (Supplementary Note 1 and Supplemental Fig. 1).

Let every cell i with expression profile x_i be characterized by a variable p_i, which represents a discrete attribute across the whole manifold, such as perturbation, species or batch. To start with, we assume only two conditions 0 (unperturbed) and 1 (perturbed). Let us further consider the conditional distribution $P(x_i|z_i, p_i)$, which assumes that each cell x_i comes from a low-dimensional representation z_i in condition p_i. We use a VAE to model $P(x_i|z_i, p_i)$ in its dependence on z_i and vector arithmetics in the latent space of VAE to model the dependence on p_i (Fig. 1).

Equipped with this, consider a typical extrapolation problem. Assume cell type A exists in the training data only in the unperturbed

[1]Helmholtz Zentrum München – German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany. [2]School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany. [3]Department of Mathematics, Technical University of Munich, Munich, Germany. *e-mail: alex.wolf@helmholtz-muenchen.de; fabian.theis@helmholtz-muenchen.de

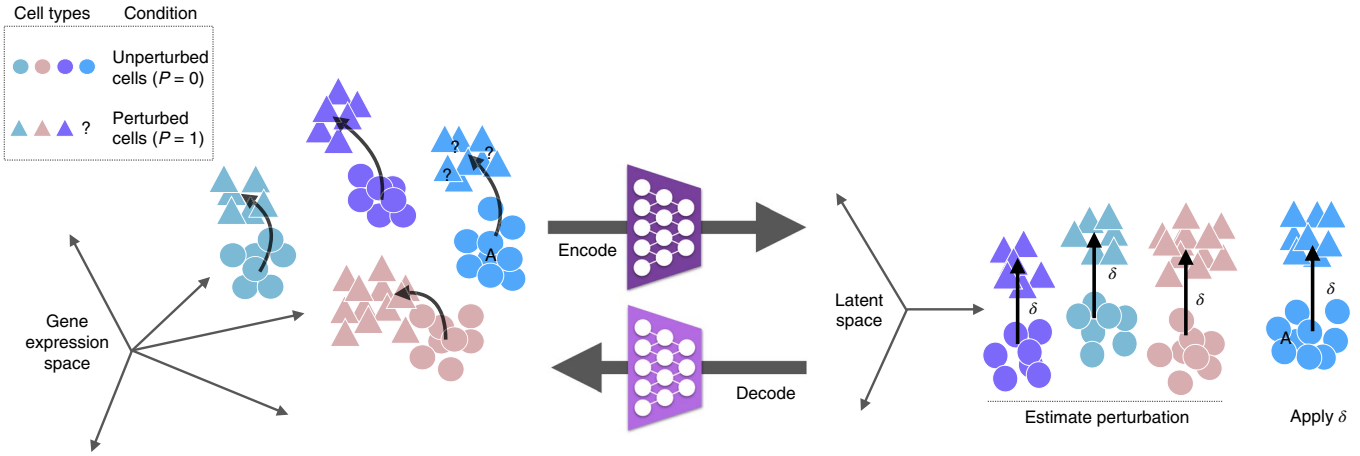

Fig. 1 | scGen, a method to predict single-cell perturbation response. Given a set of observed cell types in control and stimulation, we aim to predict the perturbation response of a new cell type *A* (blue) by training a model that learns to generalize the response of the cells in the training set. Within scGen, the model is a variational autoencoder, and the predictions are obtained using vector arithmetics in the latent space of the autoencoder. Specifically, we project gene expression measurements into a latent space using an encoder network and obtain a vector δ that represents the difference between perturbed and unperturbed cells from the training set in latent space. Using δ, unperturbed cells of type *A* are linearly extrapolated in latent space. The decoder network then maps the linear latent space predictions to highly non-linear predictions in gene expression space.

($P=0$) condition. From that, we predict the latent representation of perturbed cells ($P=1$) of cell type *A* using $\hat{z}_{i,A,p=1} = z_{i,A,p=0} + \delta$, where $z_{i,A,p=0}$ and $\hat{z}_{i,A,p=1}$ denote the latent representation of cells with cell type *A* in conditions $P=0$ and $P=1$, respectively, and δ, is the difference vector of means between cells in the training set in condition 0 and 1(Methods). From the latent space, scGen maps predicted cells to high-dimensional gene expression space using the generator network estimated while training the VAE.

To demonstrate the performance of scGen, we apply it to published human peripheral blood mononuclear cells (PBMCs) stimulated with interferon (IFN-β)[3] (Methods). As a first test, we study the predictions for stimulated CD4-T cells that are held out during training (Fig. 2a). Compared with the real data, the prediction of mean expression by scGen correlates well with the ground truth across all genes (Fig. 2b), in particular, those strongly responding to IFN-β and hence most differentially expressed (labeled genes in Fig. 2b and inset 'top 100 DEGs'). To evaluate generality, we trained six other models holding out each of the six major cell types present in the study. Figure 2c shows that our model accurately predicts all other cell types (average $R^2=0.948$ and $R^2=0.936$ for all and the top 100 differentially expressed genes (DEGs), respectively). Moreover, the distribution of the strongest regulated IFN-β response gene *ISG15* as predicted by scGen not only provides a good estimate for the mean but well predicts the full distribution (Fig. 2d, all genes in Supplementary Fig. 2a).

scGen outperforms alternative modeling approaches. Aside from scGen, we studied further natural candidates for modeling a conditional distribution that is able to capture the perturbation response. We benchmark scGen against four of these candidates, including two generative neural networks and two linear models. The first of these models is the conditional variational autoencoder (CVAE)[20] (Supplementary Note 2 and Supplementary Fig. 3a), which has recently been adapted to preprocessing, batch-correcting and differential testing of single-cell data[13]. However, it has not been shown to be a viable approach for out-of-sample predictions, even though, formally, it readily admits the generation of samples from different conditions. The second class of models are style transfer GAN (Supplementary Note 3 and Supplementary Fig. 3b), which are commonly used for unsupervised image-to-image translation[27,28]. In our implementation, such a model is directly trained for the task of transferring cells from one condition to another. The adversarial training is highly flexible and does not require an assumption of linearity in a latent space. In contrast to other propositions for mapping biological manifolds using GANs[29], style transfer GANs are able to handle unpaired data, a necessity for their applicability to scRNA-seq data. We also tested ordinary GANs combined with vector arithmetics similar to Ghahramani et al.[16]. However, for the fundamental problems outlined above, we were not able to produce any meaningful out-of-sample predictions using this setup. In addition to the non-linear generative models, we tested simpler linear approaches based on vector arithmetics in gene expression space and the latent space of principal component analyses (PCA). Applying the competing models to the PBMC data set, we observe that all other models fail to predict the distribution of *ISG15* (all genes in Supplementary Fig. 2), in stark contrast to the performance by scGen (Fig. 2d). The predictions from the CVAE and the style transfer GAN are less accurate compared to the predictions of scGen and linear models even yield incorrect negative values (Fig. 2e, Supplementary Fig. 2 and Supplementary Note 4).

A likely reason for why CVAE fails to provide more accurate out-of-sample predictions is that it disentangles perturbation information from its latent space representation z in the bottleneck layer. Hence, the layer does not capture non-trivial patterns linking perturbation to cell type. A likely reason why the style transfer GAN is incapable of achieving the task is its attempt to match two high-dimensional distributions with much more complex models involved than in the case of scGen, which are notoriously more difficult to train. Some of these arguments can be better understood when inspecting the latent space distribution embeddings of the generative models. As the CVAE completely strips off all perturbation-variation, its latent space embedding does not allow perturbed cells to be distinguished from unperturbed cells (Supplementary Fig. 4a). In contrast to CVAE representations, the scGen (VAE) latent space representation captures both information for condition and cell type (Supplementary Fig. 4c), reflecting that non-trivial patterns across condition and cell type variability are stored in the bottleneck layer. Hyperparameters (Supplementary Note 5) and architectures are reported in Supplementary Tables 1 (scGen), 2 (style transfer GAN) and 3 (CVAE).

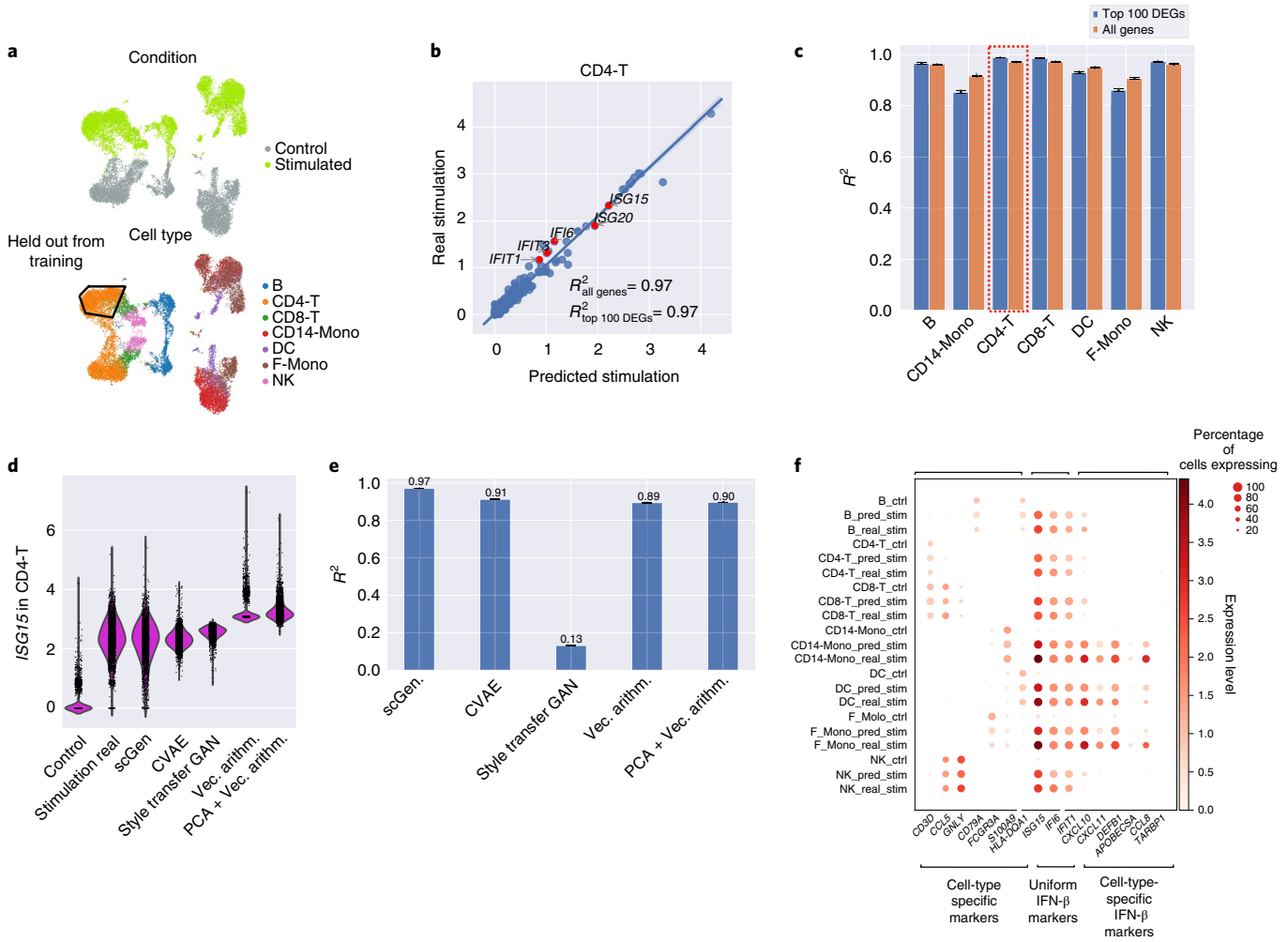

Fig. 2 | scGen accurately predicts single-cell perturbation response out-of-sample. a, UMAP visualization[37] of the distributions of conditions, cell type and data split for the prediction of IFN-β stimulated CD4-T cells from PBMCs in Kang et al.[3] (n = 18,868). **b**, Mean gene expression of 6,998 genes between scGen predicted and real stimulated CD4-T cells together with the top five upregulated DEGs (R^2 denotes squared Pearson correlation between ground truth and predicted values). **c**, Comparison of R^2 values for mean gene expression between real and predicted cells for the seven different cell types in the study. Center values show the mean of R^2 values estimated using n = 100 random subsampling for each cell type and error bars depict s.d. **d**, Distribution of *ISG15*: the top uniform response gene to IFN-β[32] between control (n = 2,437), predicted (n = 2,437) and real stimulated (n = 3,127) cells of scGen when compared with other potential prediction models. Vertical axis: expression distribution for *ISG15*. Horizontal axis: control, real and predicted distribution by different models. **e**, Similar comparison of R^2 values to predict unseen CD4-T stimulated cells. Center values show the mean of R^2 values estimated using n = 100 random subsampling for each cell type and error bars depict s.d. **f**, Dot plot for comparing control, real and predicted stimulation in predictions on the seven cell types from Kang et al.[3].

scGen predicts response shared among cell types and cell-type-specific response. Depending on shared or individual receptors, signaling pathways and regulatory networks, the perturbation response of a group of cells may result in expression level changes that are shared across all cell types or unique to only some. Predicting both types of responses is essential for understanding mechanisms involved in disease progression as well as adequate drug dose predictions[30,31]. scGen is able to capture both types of responses after stimulation by IFN-β when any of the cell types in the data is held out during training and subsequently predicted (Fig. 2f). For this, we use previously reported marker genes[32] of three different kinds: cell-type-specific markers independent of the perturbation such as *CD79A* for B cells, perturbation response specific genes like *ISG15*, *IFI6* and *IFIT1* expressed in all cell types, and genes of cell type-specific responses to the perturbation such as *APOBEC3A* for DC cells. Across the seven different held out perturbed cell types present in the data of Kang et al.[3], scGen consistently makes good predictions not only of unperturbed and shared perturbation effects but also for cell type-specific ones. These findings not only hold for these few selected marker genes but for the top 10 most cell-type-specific responding genes and to the top 500 DEGs between stimulated and control cells (Supplementary Fig. 5a,b and Supplementary Note 6). The linear model, by contrast, fails to capture cell-type-specific differential expression patterns (Supplementary Fig. 5c,d).

scGen robustly predicts response of intestinal epithelial cells to infection. We evaluate scGen's predictive performance for two data sets from Haber et al.[4] (Methods) using the same network architecture as for the data of Kang et al.[3]. These data sets consist of intestinal epithelial cells after *Salmonella* or *Heligmosomoides polygyrus* (*H. poly*) infections, respectively. scGen shows good performance for early transit-amplifying (TA.early) cells after infection with *H. poly* and *Salmonella* (Fig. 3a,b), predicting each condition with high precision ($R^2_{\text{all genes}}$ = 0.98 and $R^2_{\text{all genes}}$ = 0.98, respectively). Figure 3c,d depicts similar analyses for both data sets and all occurring cell types—as before, the predicted ones are held out during

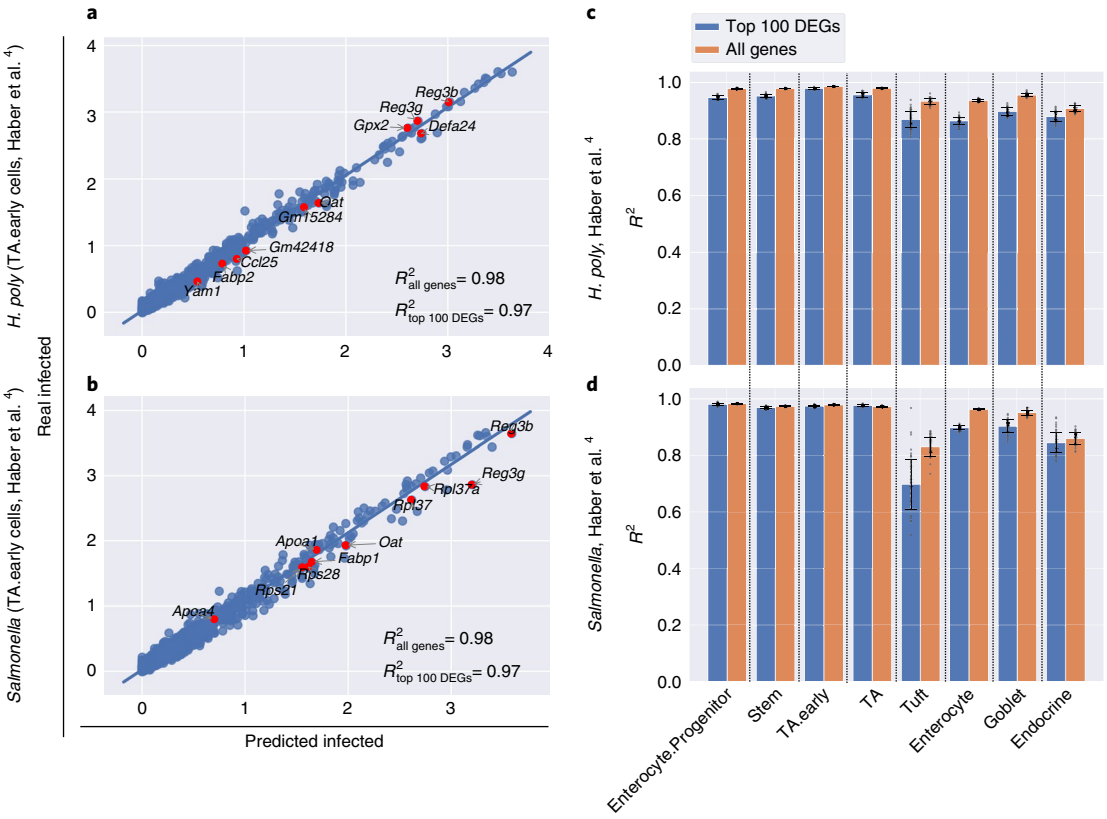

Fig. 3 | scGen models infection response in two data sets of intestinal epithelial cells. a,**b**, Prediction of early transit-amplifying (TA.early) cells from two different small intestine data sets from Haber et al.[4] infected with *Salmonella* ($n = 5{,}010$) and helminth *H. poly* ($n = 5{,}951$) after 2 and 10 days, respectively. The mean gene expression of 7,000 genes between infected and predicted cells for different cell types shows how scGen transforms control to predicted perturbed cells in a way that the expression of the top five upregulated and downregulated differentially expressed genes are similar to real infected cells. R^2 denotes squared Pearson correlation between ground truth and predicted values. **c**,**d**, Comparison of R^2 values for mean gene expression between real and predicted cells for all the cell types in two different data sets illustrates that scGen performs well for all cell types in different scenarios. Center values show the mean of R^2 values estimated using $n = 100$ random subsampling for each cell type and error bars depict s.d.

training—indicating that scGen's prediction accuracy is robust across most cell types. Again, we show that these results generalize to the top 10 most cell-type-specific responding genes out of 500 DEGs (Supplementary Fig. 6).

To understand when scGen starts to fail at making meaningful predictions, we trained it on the PBMC data of Kang et al.[3], but now with more than one cell type held out. This study shows that predictions by scGen are robust when holding out several dissimilar cell types (Supplementary Fig. 7a,b) but start failing when training on data that only contains information about the response of one highly dissimilar cell type (see CD4-T predictions in Supplementary Fig. 7c).

Finally, similar to that shown for the differentiation of epidermal cells, we cannot only generate fully responding cell populations but also intermediary cell states between two conditions. Here, we do this for IFN-β stimulation and *Salmonella* infection (Supplementary Note 7 and Supplementary Fig. 8).

scGen enables cross-study predictions. To be applicable to broad cell atlases such as the Human Cell Atlas[33], scGen needs to be robust against batch effects and generalize across different studies. To achieve this, we consider a scenario with two studies: study A, where cells have been observed in two biological conditions, for example, control and stimulation, and study B with the same setting as study A but only in control conditions. By jointly encoding the two data sets, scGen provides a model for predicting the perturbation for study B (Fig. 4) by estimating the study effect as the linear perturbation in the latent space. To demonstrate this, we use as study A the PBMC data set from Kang et al.[3] and as study B another PBMC study consisting of 2,623 cells that are available only in the control condition (Zheng et al.[34]). After training the model on data from study A, we use the trained model to predict how the PBMCs in study B would respond to stimulation with IFN-β.

As a first sanity check, we show that *ISG15* is also expressed in the prediction of stimulated cells based on the Zheng et al.[34] study (Fig. 3b). This observation holds for all other differential genes associated with the stimulation, which we show for *FCGR3A*+-Monocytes (F-Mono) (Fig. 3c): The predicted stimulated F-Mono cells correlate more strongly with the control cells in their study than with stimulated cells from study A while still expressing DEGs known from study A. Similarly, predictions for other cell types yield a higher correlation than the direct comparison with study A (Fig. 3d).

scGen predicts single-cell perturbation across species. In addition to learning the variation between two conditions, for example. health and disease for a species, scGen can be used to predict across species. We trained a model on a scRNA-seq data set by Hagai et al.[5] comprised of bone marrow-derived mononuclear phagocytes from mouse, rat, rabbit and pig perturbed with lipopolysaccharide (LPS) for six hours. Similar to previously, we held out the LPS perturbed rat cells from the training data (Fig. 5a).

In contrast to previous scenarios, two global axes of variation now exist in the latent space associated with species and stimulation.

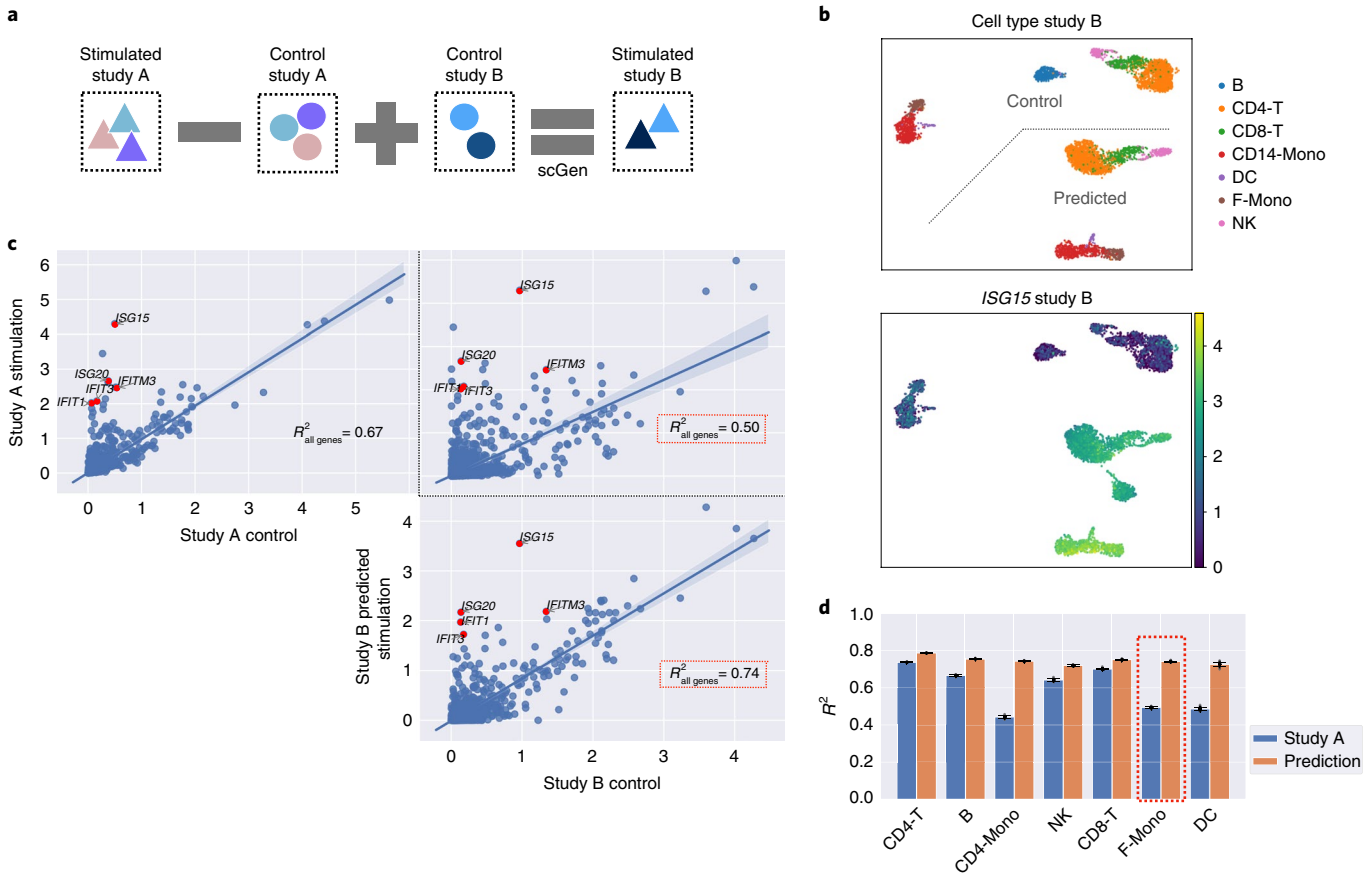

Fig. 4 | scGen accurately predicts single-cell perturbation across different studies. a, scGen can be used to translate the effect of a stimulation trained in study A to how stimulated cells would look in study B, given a control sample set. **b**, UMAP visualization of cell types for control and predicted stimulated cells ($n = 5{,}246$) for study B (Zheng et al.[34]) in two conditions where *ISG15*, the top IFN-β response gene, is only expressed in stimulated cells. Colour scale indicates expression level of *ISG15*. **c**, Average expression between control and stimulated F-Mono cells from study A (upper left), control from study B and stimulated cells from study A (upper right) and control from study B and predicted stimulated cells for study B (lower right). Red points denote top five DEGs for F-Mono cells after stimulation in study A. R^2 denotes squared Pearson correlation. Shaded lines depict 95% confidence interval for the regression estimate. The regression line is shown in blue. **d**, Comparison of R^2 values highlighted in panel **c** for F-Mono and all other cell types. Center values show the mean of R^2 values estimated using $n = 100$ random subsampling for each cell type and error bars depict s.d.

Based on this, we have two latent difference vectors:δ_{LPS}, which encodes the variation between control and LPS cells, and $\delta_{species}$, which accounts for differences between species. Next, we predict LPS rat cells using $z_{i,\text{rat,LPS}} = \frac{1}{2}\left(z_{i,\text{mouse,LPS}} + \delta_{\text{species}} + z_{i,\text{rat,control}} + \delta_{\text{LPS}}\right)$ (Fig. 5b). This equation takes an average of the two alternative ways of reaching LPS perturbed rat cells (Fig. 5a). All other predictions along the major linear axes of variation also yield plausible results for stimulated rat cells (Supplementary Fig. 9).

In addition to the species-conserved response of a few upregulated genes, such as *Ccl3* and *Ccl4*, cells also display species-specific responses. For example, *Il1a* is highly upregulated in all species except rat. Strikingly, scGen correctly identifies rat cells as non-responding with this gene. Only the fraction of cells expressing *Il1a* increases at a low expression level (Fig. 5c). Based on these early demonstrations, we hope to predict cellular responses to treatment in humans based on data from untreated humans and treated animal models.

Discussion

By adequately encoding the original expression space in a latent space, scGen achieves simple near-to-linear mappings for highly non-linear sources of variation in the original data, which explain a large portion of the variability in the data associated with, for instance, perturbation, species or batch. This allows the use of scGen in several contexts including perturbation prediction response for unseen phenomena across cell types, studies and species, for interpolating cells between conditions. Moreover, using the cell type labels from studies, scGen is able to correct for batch effects, performing equally well as state-of-art methods (Supplementary Note 8 and Supplementary Fig. 10–12).

While we showed proof-of-concept for in silico predictions of cell type and species-specific cellular response, in the present work, scGen has been trained on relatively small data sets, which only reflect subsets of biological and transcriptional variability. Although we demonstrated the predictive power of scGen in these settings, a trained model cannot be expected to be predictive beyond the domain of the training data. To gain confidence in predictions, one needs to make realistic estimates for prediction errors by holding out parts of the data with known ground truth that are representative of the task. It is important to realize that such a procedure arises naturally when applying scGen in an alternating iteration of experiments, and retraining is based on new data and in silico prediction. By design, such strategies are expected to yield highly performing models for specific systems and perturbations of interest. It is evident that such strategies could readily exploit the upcoming availability of large-scale atlases of organs in a healthy state, such as the Human Cell Atlas[33].

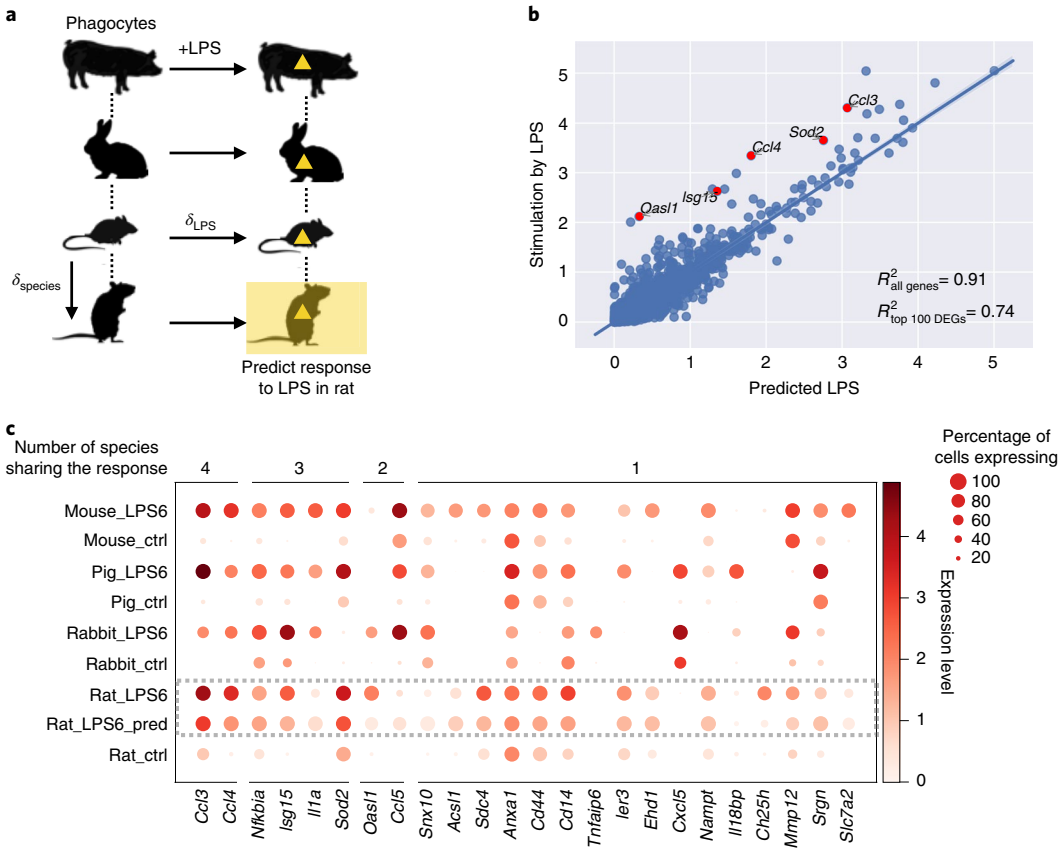

Fig. 5 | scGen predicts perturbation response across different species. a, Prediction of unseen LPS perturbed rat phagocytes on control and stimulated scRNA-seq from mouse, rabbit and pig by Hagai et al.[5] ($n = 77{,}642$). **b**, Mean gene expression of 6,619 one-to-one orthologs between species for predicted LPS perturbed rat cells plotted against real LPS perturbed cells, whereas highlighted points represent the top five DEGs after LPS stimulation in the real data. R^2 denotes squared Pearson correlation between ground truth and predicted values. **c**, Dot plot of top 10 DEGs after LPS stimulation in each species, with numbers indicating how many species have those responsive genes among their top 10 DEGs.

We have demonstrated that scGen is able to learn cell type and species-specific response. To enable this, the model needs to capture features that distinguish weakly from strongly responding genes and cells. Building biological interpretations of these features, for instance, along the lines of Ghahramani et al.[16] or Way and Greene[35], could help in understanding the differences between cells that respond to certain drugs and cells that do not respond, which is often crucial for understanding patient response to drugs[36].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0494-8.

Received: 7 January 2019; Accepted: 17 June 2019;
Published online: 29 July 2019

References

1. Stubbington, M. J. T., Rozenblatt-Rosen, O., Regev, A. & Teichmann, S. A. Single-cell transcriptomics to explore the immune system in health and disease. *Science* **358**, 58–63 (2017).
2. Angerer, P. et al. Single cells make big data: New challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* **4**, 85–91 (2017).
3. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
4. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
5. Hagai, T. et al. Gene expression variability across cells and species shapes innate immunity. *Nature* **563**, 197–202 (2018).
6. Dixit, A. et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
7. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882 (2016).
8. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
9. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
10. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* **11**, e1004333 (2015).
11. Fröhlich, F. et al. Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. *Cell Syst.* **7**, 567–579.e6 (2018).
12. Choi, K., Hellerstein, J., Wiley, S. & Sauro, H. M. Inferring reaction networks using perturbation data. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/351767v1 (2018).
13. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
14. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
15. Ding, J., Condon, A. & Shah, S. P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **9**, 2002 (2018).
16. Ghahramani, A., Watt, F. M. & Luscombe, N. M. Generative adversarial networks simulate gene expression and predict perturbations in single cells. Preprint at *bioRxiv* https://doi.org/10.1101/262501 (2018).
17. Marouf, M. et al. Realistic in silico generation and augmentation of single cell RNA-seq data using generative adversarial neural networks. Preprint at *bioRxiv* https://doi.org/10.1101/390153 (2018).

18. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
19. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at arXiv https://arxiv.org/abs/1312.6114 (2013).
20. Sohn, K., Lee, H. & Yan, X. in *Advances in Neural Information Processing Systems 28* (eds Cortes, C. et al.) 3483–3491 (Curran Associates, Inc., 2015).
21. Abadi, M. et al. Tensorflow: a system for large-scale machine learning. Preprint at https://arxiv.org/abs/1605.08695v2 (2016).
22. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
23. Bendall, S. C. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
24. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
25. Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. Preprint at *arXiv* https://arxiv.org/abs/1511.06434 (2015).
26. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint *arXiv* https://arxiv.org/abs/1301.3781 (2013).
27. Liu, M.-Y. & Tuzel, O. in *Advances in Neural Information Processing Systems 29* (eds Lee, D. D. et al.) 469–477 (Curran Associates, Inc., 2016).
28. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision* (2017).
29. Amodio, M. & Krishnaswamy, S. MAGAN: Aligning biological manifolds. In *Proceedings of the 35th International Conference on Machine Learning* Vol. 80 (eds Dy, J. & Krause, A.) 215–223 (PMLR, Stockholmsmässan, Stockholm, 2018).
30. Clift, M. J. D. et al. A novel technique to determine the cell type specific response within an in vitro co-culture model via multi-colour flow cytometry. *Sci. Rep.* **7**, 434 (2017).
31. Schubert, M. et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* **9**, 20 (2018).
32. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
33. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
34. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
35. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput.* **23**, 80–91 (2018).
36. Smillie, C. S. et al. Rewiring of the cellular and inter-cellular landscape of the human colon during ulcerative colitis. Preprint at *bioRxiv* https://doi.org/10.1101/455451 (2018).
37. McInnes, L. & Healy, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* https://arxiv.org/abs/1802.03426v2 (2018).

Acknowledgements

We are grateful to all members of the Theis lab, in particular, D.S. Fischer for early comments on predicting across species. M.L. is grateful for valuable feedback from L.Haghverdi regarding batch effect removal. F.A.W. acknowledges discussions with N. Stranski on responding and non-responding cells and support by the Helmholtz Postdoc Program, Initiative and Networking Fund of the Helmholtz Association. 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).

Author contributions

M.L. performed the research, implemented the models and analyzed the data. F.A.W. conceived the project with contributions from M.L. and F.J.T. F.A.W. and F.J.T. supervised the research. All authors wrote the manuscript.

Competing interests

F.J.T. reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and ownership interest in Cellarity, Inc.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0494-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to F.A.W. or F.J.T.

Peer review information: Nicole Rusk was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Variational autoencoders. A variational autoencoder is a neural network consisting of an encoder and a decoder similar to classical autoencoders. Unlike classical autoencoders, however, VAEs are able to generate new data points. The mathematics underlying VAEs also differs from that of classical autoencoders. The difference is that the model maximizes the likelihood of each sample x_i (more accurately, maximizes the log evidence sum of log likelihoods of all x_i) In the training set under a generative process as formulated in equation (1):

$$P(x_i|\theta) = \int P(x_i|z_i;\theta)P(z_i|\theta)\mathrm{d}z_i \tag{1}$$

where θ is a model parameter that in our model corresponds to a neural network with its learnable parameters and z_i is a latent variable. The key idea of a VAE is to sample latent variables z_i that are likely to produce x_i and using those to compute $P(x_i|\theta)$ (ref. [38]). We approximate the posterior distribution $P(z_i|x_i,\theta)$ using the variational distribution $Q(z_i|x_i,\phi)$, which is modeled by a neural network with parameter ϕ, called the inference network (the encoder). Next, we need a distance measure between the true posterior $P(z_i|x_i,\theta)$ and the variational distribution. To compute such a distance, we use the Kullback–Leibler (KL) divergence between $Q(z_i|x_i,\phi)$ and $P(z_i|x_i,\theta)$, which yields:

$$\begin{aligned}&\mathrm{KL}(Q(z_i|x_I,\phi)||P(z_i|x_i,\theta))\\&= E_{Q(z_i|x_i,\phi)}[Q(z_i|x_i,\phi) - P(z_i|x_i,\theta)]\end{aligned} \tag{2}$$

Now, we can derive both $P(x_i|\theta)$ and $P(x_i|z_i,\theta)$ by applying Bayes' rule to $P(z_i|x_i,\theta)$, which results in:

$$\begin{aligned}\mathrm{KL}(\mathrm{Q}(\mathrm{z_i}|\mathrm{x_i},\phi) \parallel \mathrm{P}(\mathrm{z_i}|\mathrm{x_i},\theta)) = \; & E_{Q(z_i|x_i,\phi)}[\log\mathrm{Q}(\mathrm{z_i}|\mathrm{x_i},\phi) - \log\mathrm{P}(\mathrm{z_i}|\theta)\\ & -\log\mathrm{P}(\mathrm{x_i}|\mathrm{z_i},\theta)] + \log P(x_i|\theta)\end{aligned} \tag{3}$$

Finally, by rearranging some terms and exploiting the definition of KL divergence we have:

$$\begin{aligned}&\log\mathrm{P}(\mathrm{x_i}|\theta) - \mathrm{KL}(\mathrm{Q}(\mathrm{z_i}|\mathrm{x_i},\phi) \parallel \mathrm{P}(\mathrm{z_i}|\mathrm{x_i},\theta))\\&= E_{Q(z_i|x_i,\phi)}[\log\mathrm{P}(\mathrm{x_i}|\mathrm{z_i},\theta)] - \mathrm{KL}[Q(z_i|x_i,\phi) \parallel P(z_i|\theta)]\end{aligned} \tag{4}$$

On the left-hand side of equation (4), we have the log-likelihood of the data denoted by $\log P(x_i|\theta)$ and an error term that depends on the capacity of the model. This term ensures that Q is as complex as P and assuming a high capacity model for $Q(z_i|x_i,\phi)$, this term will be zero[38]. Therefore, we will directly optimize $\log\mathrm{P}(\mathrm{x_i}|\theta)$:

$$E_{Q(z_i|x_i,\phi)}[\log\mathrm{P}(\mathrm{x_i}|\mathrm{z_i},\theta)] - \mathrm{KL}[\mathrm{Q}(\mathrm{z_i}|\mathrm{x_i},\phi)||\mathrm{P}(\mathrm{z_i}|\theta)] \tag{5}$$

Equation (4) and (5) are also known as the evidence lower bound (ELBO). To maximize the equation (5), we choose the variational distribution $Q(z_i|x_i,\phi)$ to be a multivariate Gaussian $Q(z_i|x_i) = N(z_i;\mu_\phi(x_i),\Sigma_\phi(x_i))$ where $\mu_\phi(x_i)$ and $\Sigma_\phi(x_i)$ are implemented with the encoder neural network and $\Sigma_\phi(x_i)$ is constrained to be a diagonal matrix. The KL term in equation (5) can be computed analytically since both prior ($P(z_i|\theta)$) and posterior ($Q(z_i|x_i,\phi)$ are multivariate Gaussian distributions. The integration for the first term in equation (5) has no closed-form and we need Monte Carlo integration to estimate it. We can sample $\mathrm{Q}(z_i|x_i,\phi)$ L times and directly use stochastic gradient descent to optimize equation (6) as the loss function for every training point x_i from data set D:

$$\mathrm{Loss}(x_i) = \frac{1}{L}\sum_{l=1}^{L}\log P(x_i|z_{i,l},\theta) - \alpha\mathrm{KL}[Q(z_i|x_i,\phi)||P(z_i|\theta)] \tag{6}$$

where the hyperparameter (α) controls how much the KL divergence loss contributes to learning. However, the first term in equation (6) only depends on the parameters of P, without reference to the parameters of variational distribution Q. Therefore, it has no gradient with respect to ϕ to be backpropagated. To address this, the reparameterization trick[19] has been proposed. This trick works by first sampling from $\epsilon \sim N(0, \mathrm{I})$ and then computing $z_i = \mu_\phi(x_i) + \Sigma_\phi^{\frac{1}{2}}(x)\times\epsilon$. Thus, we can use gradient-based algorithms to optimize equation (6).

δ vector estimation. To estimate δ, first, we extracted all cells for each condition. Next, for each cell type, we upsampled the cell type sizes to be equal to the maximum cell type size for that condition. To further remove the population size bias, we randomly downsampled the condition with a higher sample size to match the sample size of the other condition. Finally, we estimated the difference vector by calculating $\delta = \mathrm{avg}(z_{\mathrm{condition}=1}) - \mathrm{avg}(z_{\mathrm{condition}=0})$ where $z_{\mathrm{condition}=0}$ and $z_{\mathrm{condition}=1}$ denote the latent representation of cells in each condition, respectively.

Data sets and preprocessing. Kang et al.[3] included two groups of control and stimulated PBMCs. We annotated cell types by extracting an average of the top 20 cluster genes from each of eight identified cell types in PMBCs[34]. Next, the Spearman correlation between each single cell and all eight cluster averages was calculated, and each cell was assigned to the cell type for which it had a maximum correlation. After identifying cell types, megakaryocyte cells were removed from the data set due to the high uncertainty of the assigned labels. Next, the data set was filtered for cells with a minimum of 500 expressed genes and genes that were expressed in five cells at least. Moreover, we normalized counts per cell, and the top 6,998 highly variable genes were selected. Finally, we log-transformed the data to facilitate a smoother training procedure. The final data include 18,868 cells. Count matrices are available with accession number GSE96583.

The Haber et al.[4] data set contained epithelial cell responses to pathogen infection (accession number GSE92332). In this data set, the responses of intestinal epithelial cells to *Salmonella* and parasitic helminth *H. poly* were investigated. These data included three different conditions: 1,770 *Salmonella*–infected cells; 2,711 cells 10 d after *H. poly* infection and a group of 3,240 control cells. Each set of data was normalized per cell and log-transformed and the top 7,000 highly variable genes were selected.

The PBMC data set from Zheng et al.[34] was obtained from http://cf.10xgenomics.com/samples/cellexp/1.1.0/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz. After filtering cells, the data were merged with filtered PBMCs from Kang et al.[3]. The megakaryocyte cells were removed from the smaller data set. Next, the data were normalized, and we selected the top 7,000 highly variable genes. The merged data set was log-transformed and cells from Kang et al.[3] ($n = 16{,}893$) were used for training the model. The remaining 2,623 cells from Zheng et al.[34] were used for the prediction.

Pancreatic data sets ($n = 14{,}693$) were downloaded from ftp://ngs.sanger.ac.uk/production/teichmann/BBKNN/objects-pancreas.zip. Comparisons to other batch correction methods were performed similar to previously[39] with 50 principal components. The data were already preprocessed and directly used for training the model.

Mouse cell atlas data ($n = 114{,}600$) were obtained from ftp://ngs.sanger.ac.uk/production/teichmann/BBKNN/MouseAtlas.zip. The data were already preprocessed and directly used for training the model.

The LPS data set[5] (accession id E-MTAB-6754) was obtained from https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6754/?query=tzachi+hagai. The data were further filtered for cells, normalized and log-transformed. We used BiomaRt (v.84) to find ENSEMBL IDs of one-to-one orthologs in the other three species with mouse. In total 6,619 genes were selected from all species for training the model. The final data include 77,642 cells.

Statistics. All the differential tests to extract DEGs were performed using Scanpy's *rank_genes_groups* function with Wilcoxon as the method parameter. Error bars were computed by randomly resampling 80% of the data with replacement 100 times and recomputing Pearson R^2 for each resampled data. The interval represents the mean of R^2 values plus/minus the standard deviation of those 100 R^2 values. We used the mean of 100 R^2 values for the magnitude of each bar. All the R^2 values were calculated by squaring the *rvalue* output of the *scipy.stats.linregress* function and denote squared Pearson correlation.

Evaluation. *Silhouette width.* We calculated the silhouette width based on the first 50 PCs of the corrected data or the latent space of the algorithm if it did not return corrected data. The silhouette coefficient for cell i is defined as: $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, where $a(i)$ and $b(i)$ indicate the mean intra-cluster distance and the mean nearest-cluster distance for sample i, respectively. Instead of cluster labels, batch labels can be used to assess batch correction methods. We used the *silhouette_score* function from scikit-learn[40] to calculate the average silhouette width over all samples.

Cosine similarity. The *cosine_similarity* function from scikit-learn was used to compute cosine similarity. This function computes the similarity as the normalized dot product of X and Y defined as: $\mathrm{cosine\_similarity}(X,Y) = \frac{\langle X,Y\rangle}{\|X\|\|Y\|}$

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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 https://github.com/theislab/scgen-reproducibility.

Code availability

The software is available at https://github.com/theislab/scgen. The code to reproduce the results of the papers is also available at https://github.com/theislab/scgen-reproducibility.

References

38. Doersch, C. Tutorial on variational autoencoders. Preprint at *arXiv* https://arxiv.org/abs/1606.05908 (2016).
39. Park, J.-E., Polanski, K., Meyer, K. & Teichmann, S. A. Fast batch alignment of single cell transcriptomes unifies multiple mouse cell atlases into an integrated landscape. Preprint at *bioRxiv* https://doi.org/10.1101/397042 (2018).
40. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

nature research

Corresponding author(s): F. Alexander Wolf, Fabian J. Theis

Last updated by author(s): Jun 5, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect the data.
Data analysis	https://github.com/theislab/scGen_reproducibility

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All of the datasets analyzed in this manuscript are public and published in other papers. We referenced them in the manuscript and they are downloadable at https://github.com/theislab/scGen_reproducibility

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No experiments in study
Data exclusions	No experiments in study
Replication	No experiments in study
Randomization	No experiments in study
Blinding	No experiments in study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging