A PRETRAINED SCVI MODEL FOR 60,000 DRUG PERTURBATION EXPERIMENTS IN 100 MILLIONS CELLS

Valentine Svensson
Vevo Therapeutics
South San Francisco, CA 94080, USA
{valentine}@vevo.ai

ABSTRACT

We present a pre-trained SCVI model for the Tahoe-100M single-cell dataset, enabling large-scale single-cell analyses on systems with limited GPU memory. By compressing expression profiles from over 95 million cells into a 42 GB "minified" file plus a 1 GB model, this approach preserves essential biological signals while remaining practical for routine exploratory tasks. The openly available model supports downstream analyses such as differential expression and method development, without requiring access to the entire raw dataset.

Through technological advances, single-cell genomics remains in an exponential growth phase in terms of scaling the number of observations in a single study (Svensson et al., 2020). Data set sizes have increased to the point where exploring raw data requires extremely powerful computational resources.

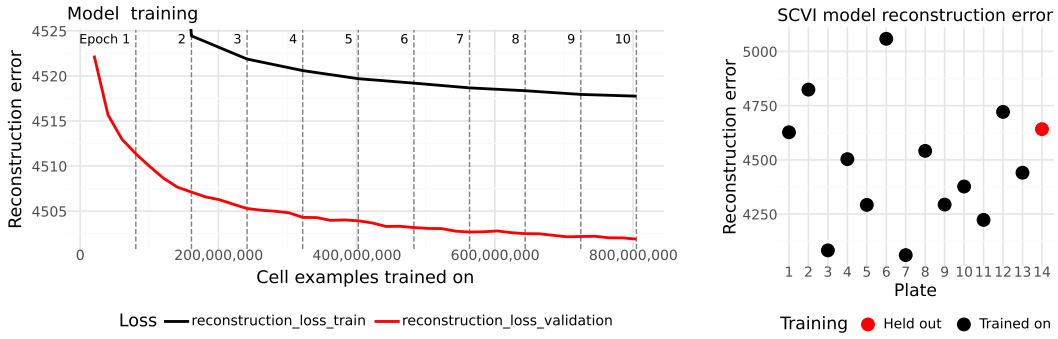
We have produced the largest single-cell RNA-sequencing (scRNA-seq) data set to date, encompassing 105,609,549 single cells from 58,045 experimental conditions corresponding to varying drug treatments and controls across 50 different cancer cell lines (Zhang et al., 2025). Here we describe training and making openly available an SCVI model with encoded representations of the Tahoe-100M data for the community to use for downstream exploratory analysis and methods development.

The UMI (unique molecular identifier) counts of all 62,710 genes and metadata for all the cells use 1.1 TB of storage when stored in H5AD format (Virshup et al., 2021), with a similar amount of memory required if the full dataset was to be loaded into memory. With gz compression, the storage requirement can be pushed down to 375 GB, though substantially reducing read speed as the data is decompressed upon read (about 5×, depending on CPU performance), and would still require the full 1.1TB of RAM after being read into memory. For exploratory analysis, the data needs to be loaded in 'backed' mode where expression values are left on disk and only read when needed. While this strategy greatly reduces the memory required to work with the dataset, lazily reading UMI counts from disk substantially slows down calculations for exploratory analysis and other downstream tasks such as differential expression analysis.

To ease exploratory analysis, we fitted an SCVI model (Lopez et al., 2018; Gayoso et al., 2022) to the data with the goal of accurately representing expression of all genes in all cells in 'minified' format, where only variational parameters for latent representations are stored (Ergen et al., 2024).

The Tahoe-100M dataset was filtered to ensure particularly high quality training data, retaining 95,624,334 cells. The Tahoe-100M dataset was generated as 14 plates, where plate 14 replicates the treatment conditions in plate 6. Plate 14 was held out from training to be used for out-of-sample validation and model criticism. The data from plates 1 through 13 left 89,423,257 cells for model fitting. Using the train () method for the SCVI class, the dataset was further randomly divided into 90% training and 10% validation during training.

Based on prior experience fitting SCVI models to smaller datasets, this initial version of the model was set up with the default structure of 1 hidden layer of 128 units width, with 10-dimensional latent representations of the cells. Since our goal is to accurately represent observed gene expression, the model was not set up to perform batch integration. The model was trained with a batch size of 128, default learning rate and other training parameters. The model was trained for a total 10 epochs over



(a) (b)

Figure 1: Tahoe-100M-SCVI-v1 training and evaluation.

the filtered Tahoe-100M data, allowing the model to see over 800 million examples of single-cell gene expression profiles. The training validation reconstruction error curve shows no sign of overfitting, and does not appear to have reached convergence after training finished, suggesting simply training the model longer will improve it (Figure 1a).

Each of the 95,624,334 filtered cells (including cells from the held-out replication plate) were encoded as 10-dimensional mean and variance parameters. The encoded dataset requires 42GB of storage and memory, including cell metadata. Cell embeddings alone are estimated to use 8GB storage and memory. To use the encoded ('minified') data, the SCVI model parameters are needed, which requires 964 MB storage and GPU memory or RAM. The small model size enables the encoded data to be decoded into gene expression and explored even on computers with older-generation GPUs that have limited memory.

Average reconstruction error (negative log likelihood of the observed UMI counts given the latent representations) was evaluated for all plates separately, including the held-out plate 14. Plate 14 has a similar reconstruction error to the plates used for training, further indicating the model has not overfit to the training data (Figure 1b).

The model and minified .h5ad file for the Tahoe-100M data are freely available on Hugging Face at `https://huggingface.co/vevotx/Tahoe-100M-SCVI-v1` and can be loaded into a Python session directly using `scvi.hub` functions from the `scvi-tools` package (Ergen et al., 2024).

This model is an initial version. Future work will include optimization for more accurate decoding of expression levels and generation of UMI counts. One important aspect of this optimization, is a cost-benefit analysis of representation dimensionality. Higher dimensionality will facilitate more accurate decoding of gene expression, but at the cost of larger memory requirements. What is the marginal gain in practical, observable, accuracy in moving from 10-dimensional representations (8 GB) to a 64-dimensional representation (51 GB)? One interesting direction would be the use of Matryoshka representations, where downstream analysts can decide on the cost-benefit value of higher dimensionality depending on the accuracy needed for the task Kusupati et al. (2022).

Another avenue of future work is quantitative accuracy metrics for reconstruction of gene expression and generation of UMI counts. The average reconstruction error of the model, defined as the negative log likelihood of the observed UMI counts given the representation vectors through the generative SCVI model, across the full Tahoe-100M dataset is 4,509.95. This value carries little intuition to a user without comparison to alternative models. Since SCVI is a generative model, posterior predictive distributions of UMI counts can be used to assess the calibration of the model (Gelman et al., 1996). For example, using the held-out validation plate, we find that observed UMI counts for (cell, gene) pairs are within the 95% confidence interval of posterior predictive distributions generated by the model 97.7% of the time. However, the observed UMI count for a (cell, gene) pair is 0 in 97.4% of cases in the data, and so a very naive 'generative model' which only produces 0-values would score 97.4% on the calibration metric. Further work on evaluation metrics is needed.

Nevertheless, this work provides a means for researchers to explore gene expression across all 62,710 genes in nearly 60,000 experimental conditions, at the single-cell level in nearly 100 million cells, while only needing 42 GB storage and memory, and 1 GB GPU memory.

MEANINGFULNESS STATEMENT

Living organisms react to chemical stimuli through highly complex and fine-grained transcriptional regulation. Answering the question of which of the over 62,000 genes can be regulated by which chemicals will allow us to understand a major and controllable aspect of the operation of life. The data we have generated to answer this question is massive in scale. By training a generative model of this data, the information in the massive data in a compressed representation can be analyzed to answer questions about transcriptional responses to drug with fewer resources at lower cost.

REFERENCES

Can Ergen, Valeh Valiollah Pour Amiri, Martin Kim, Aaron Streets, Adam Gayoso, and Nir Yosef. Scvi-hub: an actionable repository for model-driven single cell analysis. *bioRxiv*, pp. 2024.03.01.582887, March 2024.

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J Theis, Aaron Streets, Michael I Jordan, Jeffrey Regier, and Nir Yosef. A python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.*, 40(2):163–166, February 2022.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. POSTERIOR PREDICTIVE ASSESSMENT OF MODEL FITNESS VIA REALIZED DISCREPANCIES. *Stat. Sin.*, 6(4):733–760, 1996.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. *arXiv [cs.LG]*, May 2022.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 15(12):1053–1058, December 2018.

Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. A curated database reveals trends in single-cell transcriptomics. *Database*, 2020, November 2020.

Isaac Virshup, Sergei Rybakov, Fabian J Theis, Philipp Angerer, and F Alexander Wolf. anndata: Annotated data. *bioRxiv*, December 2021.

Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, Vuong Tran, Joseph Pangallo, Efthymia Papalexi, Ajay Sapre, Hoai Nguyen, Oliver Sanderson, Maria Nigos, Olivia Kaplan, Sarah Schroeder, Bryan Hariadi, Simone Marrujo, Crina Curca Alec Salvino, Guillermo Gallareta Olivares, Ryan Koehler, Gary Geiss, Alexander Rosenberg, Charles Roco, Daniele Merico, Nima Alidoust, Hani Goodarzi, and Johnny Yu. *Tahoe-100M*: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, pp. 2025.02.20.639398, February 2025.