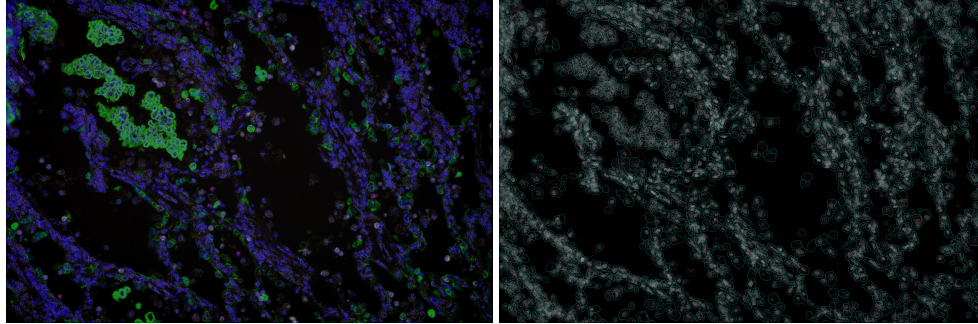


Appendix for CellPLM: Pre-training of Cell Language Model Beyond Single Cells

A Spatially-Resolved Transcriptomic Data

Recently, spatial transcriptomic technologies are developed to spatially resolve transcriptomics profiles [57, 58]. With spatial transcriptomics data, researchers can learn the spatial context of cells and cell clusters within a tissue [59]. The major technologies/platforms for spatial transcriptomics are Visium by 10x [57], GeoMx Digital Spatial Profiler (DSP) [58] by NanoString and CosMx Spatial Molecular Imager (SMI) by NanoString, MERFISH, Vizgen, Resolve, Rebus, and molecular cartography. 10x Visium does not profile at single-cell resolution, and while GeoMx DSP is capable of single-cell resolution through user-drawn profiling regions, the scalability is limited. The most recent platform, CosMx Spatial Molecular Imager (SMI) [60], can profile consistently at single-cell and even sub-cellular resolution. CosMx SMI follows much of the initial protocol as GeoMx DSP, with barcoding and ISH hybridization. However, the SMI instrument performs 16 cycles of automated cyclic readout, and in each cycle, the set of barcodes (readouts) are UV-cleaved and removed. These cycles of hybridization and imaging yield spatially resolved profiling of RNA and protein at single-cell ($\sim 10\mu m$) and subcellular ($\sim 1\mu m$) resolution. In this work, we use two published and one unpublished dataset produced by the CosMx platform. In order to obtain the cellular level gene expression, CellPose [61] software is applied to conduct cell segmentation.

To give a concrete example, we provide a sample field-of-view (FOV) in Fig. 5. Pre-selected types of RNA molecules are captured by the molecular imager, which are denoted as white dots in the figures. Colors in the first sub-figure indicate the protein molecules that are stained. These proteins contribute to the cell segmentation process, which results in the second sub-figure. The final output from the pipeline consists of the position of each cell and a cell-by-gene count matrix, which is produced by counting the number of RNA molecules within each cell. The difference between scRNA-seq and SRT data is further demonstrated in Fig. 6.



(a) Visualization of molecular image.

(b) Visualization of cell segmentation.

Figure 5: (a) A sample image of protein and RNA molecules. (b) A sample image of segmented cells.

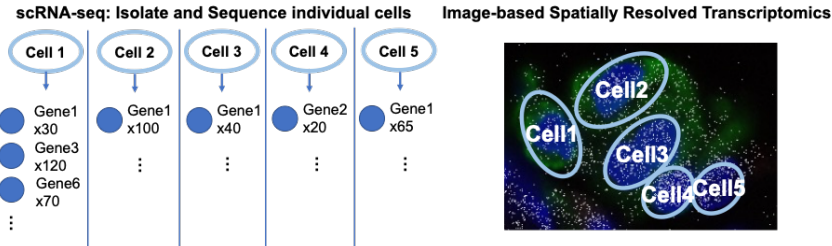


Figure 6: An illustration of the difference between scRNA-seq and SRT data.

B 2D Sinusoid Positional Encodings

Since 2D sinusoidal PE achieves a competitive performance and has a lower complexity on SRT data [16], in our transformer encoder, we generate a sinusoidal PE for cells in SRT data, formulated as:

$$\begin{aligned} \text{PE}_{(x,y,2i)} &= \sin\left(x/10000^{4i/d}\right), \text{PE}_{(x,y,2i+1)} = \cos\left(x/10000^{4i/d}\right), \\ \text{PE}_{(x,y,2j+d/2)} &= \sin\left(y/10000^{4j/d}\right), \text{PE}_{(x,y,2j+1+d/2)} = \cos\left(y/10000^{4j/d}\right), \end{aligned} \quad (10)$$

where d is the total dimension of positional encoding, $i, j \in [0, d/4)$ specify a specific feature dimension. Let $\tilde{\mathbf{C}} \in \mathcal{R}^{N \times 2}$ be a normalized coordinate matrix, where we normalize and truncate coordinates in \mathbf{C} to integers ranging in $[0, 100)$. x, y then refer to the spatial coordinates from $\tilde{\mathbf{C}}$, e.g., $x = \tilde{\mathbf{C}}_{t,0}$ and $y = \tilde{\mathbf{C}}_{t,1}$ for cell t . In this way, we generate a PE matrix $\mathbf{P} \in \mathcal{R}^{N \times d}$ for every cell in SRT data, where \mathbf{P}_i is the PE vector for cell i . Meanwhile, for scRNA-seq data, a randomly initialized d -dimensional vector p' is shared among all cells, which also results in a placeholder PE matrix \mathbf{P} .

C Broader Impact

Our method lies in an emerging and important application area, single-cell analysis. Especially, we leverage a novel type of single-cell data, Spatially Resolved Transcriptomics (SRT). SRT is a rapidly developing technology that allows scientists to map the gene expression of individual cells in their tissue environment. It combines traditional imaging techniques with transcriptome analysis to provide a spatially resolved, high-resolution view of gene expression in complex tissues. Essentially, single-cell technologies and SRT allow researchers to see where specific genes are being expressed within a tissue sample, which can help them better understand cellular interactions and the function of specific genes in complex biological systems.

We evaluate our method on various downstream tasks and the empirical results demonstrate the practical value of our method. Specifically, scRNA-seq Denoising improves the data quality of scRNA-seq data, which often suffer from technical artifacts and dropout events [19, 20], as well as significant batch effects between sequencing platforms and experiments [21, 22]. SRT imputation helps to obtain more precise cell state profiles for SRT data, while also resulting in more accurate integration and clustering between SRT data and scRNA-seq data. Perturbation prediction has great clinical value to aid in drug design and disease mechanism research.

While our work offers a significant contribution to the field of single-cell analysis, there are potential negative societal impacts that are important to consider: one of the primary potential negative societal impacts is privacy and data security. Single-cell analysis involves working with sensitive genetic information which, if mishandled, could lead to breaches in privacy and the misuse of personal data. Another potential negative impact is over-reliance on automated analysis. The complexity of single-cell data requires careful interpretation, and the risk of false-positive or false-negative results may be elevated due to computational errors or algorithmic biases. It is crucial to remember that these tools should serve as aids to human understanding and decision-making rather than replacements.

As single-cell technologies continue to evolve, it is critical that we continue to consider and address these broader societal impacts. Moving forward, it is crucial that our work is coupled with ongoing discussions on best practices in data management, privacy protection, and equitable access to technology. This includes strengthening collaborations with ethicists, policymakers, and regulatory bodies to navigate these complex issues.

D Pre-training Settings

D.1 Hyperparameter Settings

We pre-trained *CellPLM* model with 3 different parameter sizes: 10M, 20M and 40M, with the hyperparameters specified in Table 3. However, according to our preliminary experiments in Fig. 7, the performance does not significantly increase with the larger model. Therefore, we consider *CellPLM* 10M as an optimal model and conduct fine-tuning experiments based on this version.

	CellPLM-10M	CellPLM-20M	CellPLM-40M
encoder hidden dim	256	384	512
encoder layers	4	6	10
latent dimension	64	64	64
decoder hidden dim	128	192	256
decoder layers	2	2	2
model dropout	0.1	0.1	0.1
cell mask rate	0.5	0.5	0.5
gene mask rate	0.7	0.7	0.7
learning rate	2e-4	2e-4	2e-4
weight decay	1e-8	1e-8	1e-8
num of cluster (for GMM)	16	16	16

Table 3: Hyperparameters for pretrained *CellPLM* models of different sizes.

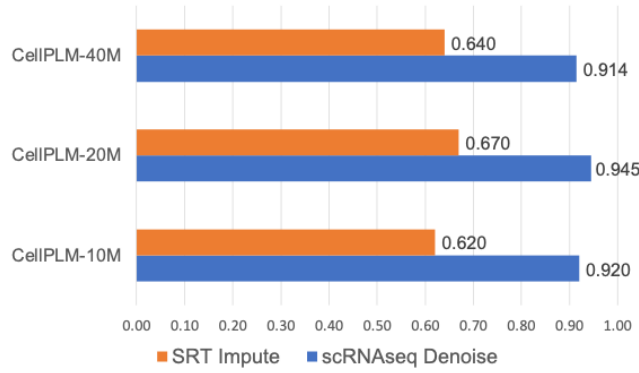


Figure 7: Comparisons of model sizes on zero-shot performances. scRNAseq denoise refers to RMSE performance on PBMC 5K dataset, SRT imputation refers to RMSE performance on Lung2 dataset.

411 D.2 Datasets for Pre-training

412 The dataset for pre-training contains 11.4 million cells from scRNA-seq and SRT data. scRNA-seq
413 data consist of 4.7 million cells from human tumor cell atlas (HTCA, <https://humantumoratlas.org/>),
414 1.4 million cells from human cell atlas (HCA, <https://www.humancellatlas.org/>), and
415 2.6 million cells from Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>).
416 All of them are public available data. A more detailed list will be disclosed in our GitHub repository.
417 Note that although our *CellPLM* is capable to handle various input feature sets, when we concatenated
418 these scRNA-seq datasets, we used inner join by default of Anndata package. As a result, all scRNA-
419 seq datasets only contain a 13,500 common gene set. We will address this issue and increase the size
420 of the gene set in future versions of *CellPLM*.

421 The SRT datasets we used are publicly available on Nanostring official website: [https://](https://nanostring.com/products/cosmx-spatial-molecular-imager/nsclc-ffpe-dataset/)
422 nanostring.com/products/cosmx-spatial-molecular-imager/nsclc-ffpe-dataset/,
423 where 2.7 million cells and 1,000 genes are measured.

424 E Additional Experimental Details

425 In this section, we provide more experimental details about fine-tuning, baselines, and evaluation
426 metrics under each downstream task.

427 E.1 scRNA-seq Denoising

428 **Downstream Task Datasets.** In scRNA-seq denoising task, we evaluate *CellPLM* on two datasets,
429 i.e., PBMC 5K and Jurkat from 10x Genomics [39]. It is worth noting that during the preprocessing
430 stage, we performed sub-setting on both datasets to ensure that all the genes were included in the gene

set of pre-training data. Additionally, any genes with zero counts were removed from the analysis. We list the statistics of them in Table 4.

Table 4: scRNA-seq denoising datasets

	5K PBMC	Jurkat
Number of genes	33,538	32,738
Number of cells	5,247	3,258
Num genes picked	7,197	7,618

Evaluation Metrics. Following the setting of scGNN [13], scGNN2.0 [40] and DeepImpute [41], we performed synthetic dropout simulation with missing at random (MAR) setting. While scGNN only considered a simple scenario, i.e., randomly flipped 10% of the non-zero entries to zeros, DeepImpute applied cell-wise mask with masking probability given by a multinomial distribution. Specifically, we adapted the setting from DeepImpute with exponential kernel. For cell i that contains at least 5 expressed genes, the probability that one non-zero count $x_{i,j}$ is masked during the training process is given by $\text{Exp}(0, 20)$:

$$p_{i,j} = \frac{1}{20} e^{-\frac{x}{20}},$$

$$q_{i,j} = \frac{p_{i,j}}{\sum_{j=0}^{J_i} p_{i,j}},$$

where J_i is the number of non-zero counts within cell i . We masked 10% of the non-zero counts according to $\{q_{i,j}\}_{j=0}^{J_i}$ and evaluate model performance on the masked entries. We calculate the root mean squared error (RMSE) and mean absolute error (MAE) between the predicted values and ground truth.

Baselines (1) DeepImpute [41] employed a strategy of dividing genes into subsets and constructing deep neural networks to impute scRNA-seq data. We implemented DeepImpute with default settings in DANCE [62] package. (2) scGNN2.0 [40] incorporated a feature autoencoder, a cluster autoencoder and a graph attention autoencoder for simultaneous imputation and clustering. scGNN2.0 is implemented by DANCE package with default settings. (3) GraphSCI [63] combined autoencoders with graph convolution networks among a gene-gene similarity graph. We accommodated the implementation of GraphSCI in DANCE package. (4) SAVER [42] leveraged Poisson LASSO regression to model the scRNA-seq counts with Poisson–gamma mixture. We utilized R package SAVER to illustrate the performance of it. (5) DCA [43] introduced an autoencoder framework based on zero inflated negative binomial (ZINB) distribution. We applied DCA to aforementioned datasets with its Python package. (6) MAGIC [44] utilized Markov affinity to capture gene-gene relationship and impute missing gene expression. We adapted its Python package to access the performance of it. (7) scImpute [45] developed a Gamma and Gaussian mixture model to identify dropout values. We revealed the performance of scImpute with its R package.

Fine-tuning. Since denoising task requires model to recover the gene expression matrix, we can directly get the zero shot performance of *CellPLM* by specifying the gene set of target dataset. Additionally, we fine-tuned *CellPLM* by replacing the pre-trained decoder with a MLP head and initializing encoder with pre-trained weights. Additionally, for methods require model selection on validation set, we performed another 10% simulation dropout and treat masked entries as validation set. The fine-tuned *CellPLM* was trained on MSE reconstruction loss, while the best model was selected by evaluating MSE on validation set.

E.2 Spatial Transcriptomic Imputation

Downstream Task Datasets. To evaluate spatial transcriptomic imputation models at single-cell resolution, we collected two samples from MERSCOPE FFPE Human Immuno-oncology Data [46]. Specifically, we chose "Lung cancer 2" and "Liver cancer 2" as our samples, and subsequently referred to them as "Lung2" and "Liver2" respectively. The Lung2 and Liver2 datasets were subsetted to align with the gene set of the pre-training data. Additionally, we removed the fields of view (FOVs) that contained fewer than 100 cells and retained only the first 100 FOVs from both datasets. Note that all baselines require reference scRNA-seq datasets to impute the unseen genes of SRT data, we

collected GSE131907 [64] and GSE151530 [65] for lung cancer and liver cancer, respectively. The statistics of all datasets are illustrated in Table 5.

Table 5: Spatial transcriptomic imputation datasets.

	Lung2	Liver2	GSE131907	GSE151530
Number of genes	500	500	29,634	18,667
Number of cells	836,739	598,141	208,506	56,721
Num genes picked	462	446	All	ALL
Num cells picked	40,114	20,629	All	All

Evaluation Metrics. Following the evaluation pipeline proposed by Avşar et al. [51], we selected target genes of SRT data with stratified sampling according to gene sparsity. Specifically, we grouped genes into four categories: low sparse, moderate sparse, high sparse, and very-high sparse. Empirically, the boundaries were defined as $[x < 75, 75 \leq x < 90, 90 \leq x < 95, 95 \leq x]$ to approximate the Gaussian mean and standard deviation slices. Subsequently, we randomly selected 25 genes from each sparsity group and remove them from training data. After training the models, we calculate the evaluation metrics on the target genes. Namely, we compute the root mean squared error (RMSE), Pearson’s correlation coefficient (PCC) and cosine similarity (Cosine) between the ground truth values and the corresponding imputed values in a gene-wise approach.

Baselines. (1) SpaGE [47] relied on domain adaptation to map scRNA-seq data onto SRT data and utilized a k -nearest-neighbor (k-NN) graph to predict unseen genes. We implemented SpaGE with default settings on both datasets. (2) stPlus [48] developed an autoencoder framework for learning cell embeddings and imputing SRT genes using a weighted k-NN approach. The performance of stPlus is accessed by its Python package. (3) gimVI [49] introduced a variational autoencoder based model with protocol-specific treatments on scRNA-seq data and SRT data. We applied the scvi-tools [66] Python package with default settings to evaluate the performance of gimVI. (4) Tangram [50] utilized a deep learning approach to learn the spatial alignment of scRNA-seq data based on a reference SRT dataset with consistent spatial maps. We evaluated Tangram with its Python package.

Fine-tuning. Similar to scRNA-seq denoising, the spatial transcriptomic imputation task requires the output of the model to be the gene expression. Thus, we directly fine-tune *CellPLM* on the pre-trained weights while specifying the input genes and target genes. The last two batches were hold out for validation.

E.3 Perturbation Prediction

Downstream Task Datasets. We included the Adamson Perturb-Seq dataset [54] for one-gene perturbations and the Norman Perturb-Seq dataset [55] for two-gene perturbations. We followed the preprocess pipeline of GEARS [53] and both datasets were then gene-wise subsetted to fit in the gene set of pre-training data. The statistics are summarized in Table 6.

Table 6: Perturbation prediction datasets.

	Adamson	Norman
Number of genes	5,060	5,045
Number of cells	68,603	91,205
Num genes picked	3,246	2,353
Num one-gene pert.	87	105
Num two-gene pert.	–	131

Evaluation Metrics. Following the setting of GEARS [53], we applied data split such that the testing perturbation are unseen during the training process. Specifically, For Adamson dataset, we randomly hold out 25% of the perturbations for testing and 10% of the perturbations within the training set for validation. For Norman dataset, two settings for two-gene perturbations are implemented for evaluation purpose: 1/2 unseen and 2/2 unseen. We excluded all two-gene combinations in which at least one of the individual genes involved in the combination belonged to the unseen set. Finally, we

508 evaluate the performance by calculating the root mean squared error (RMSE) between the predictions
509 and the true values within the testing set.

510 **Baselines.** (1) GEARS [53] utilized gene co-expression knowledge graph and Gene Ontology-derived
511 knowledge graph to model the influence of perturbations. We followed the recommended parameter
512 settings within its Python package to access the performance. (2) scGen [56] built a conditional
513 variational autoencoders and incorporated vector arithmetics to model phenomena response. We
514 implemented scGen with its Python package on both datasets.

515 **Fine-tuning.** For one perturbation, we set the input of perturbed genes to be -100 to mimic the
516 gene perturbation action. During the fine-tuning process, we substituted the original batch-aware
517 decoder with a simplified MLP decoder. Additionally, we initialized the remaining components of
518 *CellPLM* with pre-trained weights. The final model was chosen to be the best-performed model on
519 the validation set.

520 References

- 521 [1] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu,
522 Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani.
523 mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.
- 524 [2] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and
525 Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of
526 single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.
- 527 [3] Jing Gong, Minsheng Hao, Xin Zeng, Chiming Liu, Jianzhu Ma, Xingyi Cheng, Taifeng Wang,
528 Xuegong Zhang, and Le Song. xtrimogene: An efficient and scalable representation learner for
529 single-cell rna-seq data. *bioRxiv*, pages 2023–03, 2023.
- 530 [4] Hongru Shen, Jilei Liu, Jiani Hu, Xilin Shen, Chao Zhang, Dan Wu, Mengyao Feng, Meng
531 Yang, Yang Li, Yichen Yang, et al. Generative pretraining from large-scale transcriptomes for
532 single-cell deciphering. *iScience*, 2023.
- 533 [5] Haotian Cui, Chloe Wang, Hassaan Maan, and Bo Wang. scgpt: Towards building a foundation
534 model for single-cell multi-omics using generative ai. *bioRxiv*, pages 2023–04, 2023.
- 535 [6] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep
536 bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages
537 4171–4186, 2019.
- 538 [7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece
539 Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general
540 intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- 541 [8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,
542 Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information
543 processing systems*, 30, 2017.
- 544 [9] Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E Lewis. Deciphering cell–cell
545 interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88,
546 2021.
- 547 [10] Robrecht Cannoodt, Wouter Saelens, and Yvan Saeys. Computational methods for trajectory
548 inference from single-cell transcriptomics. *European journal of immunology*, 46(11):2496–2506,
549 2016.
- 550 [11] Dylan Molho, Jiayuan Ding, Zhaocheng Li, Hongzhi Wen, Wenzhuo Tang, Yixin Wang, Julian
551 Venegas, Wei Jin, Renming Liu, Runze Su, et al. Deep learning in single-cell analysis. *arXiv
552 preprint arXiv:2210.12385*, 2022.
- 553 [12] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth
554 Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell
555 transcriptomics. *BMC genomics*, 19:1–16, 2018.

- [13] Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. scgcn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications*, 12(1):1882, 2021.
- [14] Xin Shao, Chengyu Li, Haihong Yang, Xiaoyan Lu, Jie Liao, Jingyang Qian, Kai Wang, Junyun Cheng, Penghui Yang, Huajun Chen, et al. Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with spatalk. *Nature Communications*, 13(1):4429, 2022.
- [15] Junlin Xu, Jielin Xu, Yajie Meng, Changcheng Lu, Lijun Cai, Xiangxiang Zeng, Ruth Nussinov, and Feixiong Cheng. Graph embedding and gaussian mixture variational autoencoder network for end-to-end analysis of single-cell rna sequencing data. *Cell Reports Methods*, page 100382, 2023.
- [16] Hongzhi Wen, Wenzhuo Tang, Wei Jin, Jiayuan Ding, Renming Liu, Feng Shi, Yuying Xie, and Jiliang Tang. Single cells are spatial tokens: Transformers for spatial transcriptomic data imputation. *arXiv preprint arXiv:2302.03038*, 2023.
- [17] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association.
- [18] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *elife*, 6:e27041, 2017.
- [19] Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. Power analysis of single-cell rna-sequencing experiments. *Nature methods*, 14(4):381–387, 2017.
- [20] Peng Qiu. Embracing the dropouts in single-cell rna-seq analysis. *Nature communications*, 11(1):1169, 2020.
- [21] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21:1–32, 2020.
- [22] Ricard Argelaguet, Anna SE Cuomo, Oliver Stegle, and John C Marioni. Computational principles and challenges in single-cell data integration. *Nature biotechnology*, 39(10):1202–1215, 2021.
- [23] Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sørensen, Tune H Pers, and Ole Winther. scvae: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 2020.
- [24] Jing Jiang, Junlin Xu, Yuansheng Liu, Bosheng Song, Xiulan Guo, Xiangxiang Zeng, and Quan Zou. Dimensionality reduction and visualization of single-cell rna-seq data with an improved deep variational autoencoder. *Briefings in Bioinformatics*, page bbad152, 2023.
- [25] Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.
- [26] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.
- [27] Rui Hou, Elena Denisenko, Huan Ting Ong, Jordan A Ramiłowski, and Alistair RR Forrest. Predicting cell-to-cell communication networks using natmi. *Nature communications*, 11(1):5011, 2020.

- [28] S Jin, CF Guerrero-Juarez, L Zhang, I Chang, R Ramos, CH Kuan, P Myung, MV Plikus, and Q Nie. Inference and analysis of cell-cell communication using cellchat. *nat. commun.* 12, 1088, 2021.
- [29] Micha Sam Brickman Raredon, Taylor Sterling Adams, Yasir Suhail, Jonas Christian Schupp, Sergio Poli, Nir Neumark, Katherine L Leiby, Allison Marie Greaney, Yifan Yuan, Corey Horien, et al. Single-cell connectomic analysis of adult mammalian lungs. *Science advances*, 5(12):eaaw3851, 2019.
- [30] Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC genomics*, 20:7–15, 2019.
- [31] Ruochen Jiang, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome biology*, 23(1):1–24, 2022.
- [32] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.
- [33] Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology*, 17(1):e9620, 2021.
- [34] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.
- [35] Linxiao Yang, Ngai-Man Cheung, Jiaying Li, and Jun Fang. Deep clustering by gaussian mixture variational autoencoders with graph embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6440–6449, 2019.
- [36] Daniel Im Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. Denoising criterion for variational auto-encoding framework. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- [37] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- [38] Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637–640, 2014.
- [39] 10x genomics datasets. <https://support.10xgenomics.com/single-cell-gene-expression/datasets>.
- [40] Haocheng Gu, Hao Cheng, Anjun Ma, Yang Li, Juexin Wang, Dong Xu, and Qin Ma. scgcn 2.0: a graph neural network tool for imputation and clustering of single-cell rna-seq data. *Bioinformatics*, 38(23):5322–5325, 2022.
- [41] Cédric Arisdakessian, Olivier Poirion, Breck Yunits, Xun Zhu, and Lana X Garmire. Deepimpute: an accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. *Genome biology*, 20(1):1–14, 2019.
- [42] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, 15(7):539–542, 2018.
- [43] Gökçen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):390, 2019.
- [44] David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.

- [45] Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):997, 2018.
- [46] Merscope ffpe human immuno-oncology datasets. <https://info.vizgen.com/ffpe-showcase?submissionGuid=88ba0a44-26e2-47a2-8ee4-9118b9811fbf>.
- [47] Tamim Abdelaal, Soufiane Mourragui, Ahmed Mahfouz, and Marcel JT Reinders. Spage: spatial gene enhancement using scrna-seq. *Nucleic acids research*, 48(18):e107–e107, 2020.
- [48] Chen Shengquan, Zhang Boheng, Chen Xiaoyang, Zhang Xuegong, and Jiang Rui. stplus: a reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics*, 37(Supplement_1):i299–i307, 2021.
- [49] Romain Lopez, Achille Nazaret, Maxime Langevin, Jules Samaran, Jeffrey Regier, Michael I Jordan, and Nir Yosef. A joint model of unpaired data from scrna-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv preprint arXiv:1905.02269*, 2019.
- [50] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R Vanderburg, Åsa Segerstolpe, Meng Zhang, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature methods*, 18(11):1352–1362, 2021.
- [51] Gülben Avcı and Pinar Pir. A comparative performance evaluation of imputation methods in spatially resolved transcriptomics data. *Molecular Omics*, 2023.
- [52] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Aron, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.
- [53] Yusuf Roohani, Kexin Huang, and Jure Leskovec. Gears: Predicting transcriptional outcomes of novel multi-gene perturbations. *BioRxiv*, pages 2022–07, 2022.
- [54] Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016.
- [55] Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.
- [56] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.
- [57] Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.
- [58] Christopher R Merritt, Giang T Ong, Sarah E Church, Kristi Barker, Patrick Danaher, Gary Geiss, Margaret Hoang, Jaemyeong Jung, Yan Liang, Jill McKay-Fleisch, et al. Multiplex digital spatial profiling of proteins and rna in fixed tissue. *Nature Biotechnology*, 05 2020.
- [59] Darren J Burgess. Spatial transcriptomics coming of age. *Nature Reviews Genetics*, 20(6):317–317, 2019.
- [60] Shanshan He, Ruchir Bhatt, Carl Brown, Emily A. Brown, Derek L. Buhr, Kan Chantranuvatana, Patrick Danaher, Dwayne Dunaway, Ryan G. Garrison, Gary Geiss, Mark T. Gregory, Margaret L. Hoang, Rustem Khafizov, Emily E. Killingbeck, Dae Kim, Tae Kyung Kim, Youngmi Kim, Andrew Klock, Mithra Korukonda, Aleksandr Kutchma, Zachary R. Lewis, Yan Liang,

699 Jeffrey S. Nelson, Giang T. Ong, Evan P. Perillo, Joseph C. Phan, Tien Phan-Everson, Erin
700 Piazza, Tushar Rane, Zachary Reitz, Michael Rhodes, Alyssa Rosenbloom, David Ross, Hiromi
701 Sato, Aster W. Wardhani, Corey A. Williams-Wietzikoski, Lidan Wu, and Joseph M. Beechem.
702 High-plex multiomic analysis in ffpe at subcellular level by spatial molecular imaging. *bioRxiv*,
703 2022.

704 [61] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist
705 algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.

706 [62] Jiayuan Ding, Hongzhi Wen, Wenzhuo Tang, Renming Liu, Zhaocheng Li, Julian Venegas,
707 Runze Su, Dylan Molho, Wei Jin, Wangyang Zuo, et al. Dance: A deep learning library and
708 benchmark for single-cell analysis. *bioRxiv*, 2022.

709 [63] Jiahua Rao, Xiang Zhou, Yutong Lu, Huiying Zhao, and Yuedong Yang. Imputing single-cell
710 rna-seq data by combining graph convolution and autoencoder neural networks. *Isience*,
711 24(5):102393, 2021.

712 [64] Nayoung Kim, Hong Kwan Kim, Kyungjong Lee, Yourae Hong, Jong Ho Cho, Jung Won Choi,
713 Jung-Il Lee, Yeon-Lim Suh, Bo Mi Ku, Hye Hyeon Eum, et al. Single-cell rna sequencing
714 demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma.
715 *Nature communications*, 11(1):2285, 2020.

716 [65] Lichun Ma, Limin Wang, Subreen A Khatib, Ching-Wen Chang, Sophia Heinrich, Dana A
717 Dominguez, Marshonna Forgues, Julián Candia, Maria O Hernandez, Michael Kelly, et al.
718 Single-cell atlas of tumor cell evolution in response to therapy in hepatocellular carcinoma and
719 intrahepatic cholangiocarcinoma. *Journal of hepatology*, 75(6):1397–1408, 2021.

720 [66] Probabilistic models for single-cell omics data. <https://scvi-tools.org/>.