Learning More Effective Cell Representations Efficiently

Jason Xiaotian Dou
University of Pittsburgh
jasondpku@gmail.com

Minxue Jia
University of Pittsburgh
minxue.jia@pitt.edu

Nika Zaslavsky
Carnegie Mellon University
nzaslavsky@cmu.edu

Runxue Bao
University of Pittsburgh
runxue.bao@pitt.edu

Shiyi Zhang
Carnegie Mellon University
shiyiz@andrew.cmu.edu

Ke Ni
University of Pittsburgh
ken67@pitt.edu

Paul Pu Liang
Carnegie Mellon University
paul.liangpu@gmail.com

Haiyi Mao
University of Pittsburgh
ham112@pitt.edu

Zhi-hong Mao
University of Pittsburgh
zhm4@pitt.edu

Abstract

Capturing similarity among cells is the core of many tasks in single-cell transcriptomics, such as the identification of cell types and cell states. This problem can be formulated in a paradigm called metric learning. Metric learning aims to learn data embeddings (feature vectors) in a way that reduces the distance between feature vectors corresponding to cells belonging to the same cell type and increases the distance between the feature vectors corresponding to different cell types. Deep metric learning on the other hand uses neural networks to automatically learn discriminative features from the cells and then compute the metric. The (deep) metric learning approaches have been successfully applied to computational biology tasks like similar cell identification and synthesis of heterogeneous single-cell modalities. We identify two computational challenges: precise distance measurement between cells and scalability over a large amount of data in the applications of (deep) metric learning. And then we propose our solutions: optimal transport and coreset optimization. Optimal transport has the potential to measure cell similarity more effectively and coreset optimization is promising to train representation learning models more efficiently. Empirical studies in image retrieval and clustering tasks show the promise of the proposed approaches. We propose to further explore the applicability of our methods to cell representation learning.

1 Introduction

The success of machine learning algorithms largely depends on data representation. Metric learning learns data embeddings and feature vectors in a way that reduces the distance between feature vectors corresponding to objects belonging to the same class and increases the distance between the feature vectors corresponding to different classes. Deep metric learning, on the other hand, uses neural networks to automatically learn discriminative features from the objects and then compute the metric. (Deep) metric learning falls into the broad umbrella of representation learning, whose quest for representation learning is motivating the design of more powerful representation [3, 13].

These representation learning methods have achieved great successes in biology applications [40, 43, 19, 23, 35, 56, 23, 50, 1, 11, 30, 14, 46, 47]. For example, Schema [40] uses a principled metric

36th Conference on Neural Information Processing Systems (NeurIPS 2022) LMRL workshop.

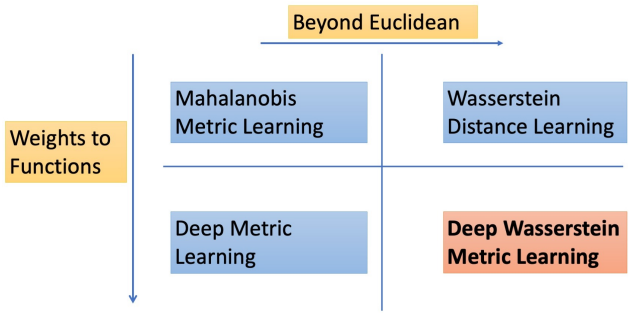

Figure 1: Overview of the Deep Wasserstein Metric Learning Framework

learning strategy to identify informative features in a modality to synthesize disparate modalities into a single coherent interpretation. It is used to infer cell types by integrating gene expression and chromatin accessibility data. Specifically, [53] presents an approach for integrating different modalities by learning a probabilistic coupling among them using autoencoders to map to a shared latent space. These methods can complement start of art single cell analytics tool such as Dynamo [38] to gain new insights into dynamic biological processes.

The deep metric learning framework SCimilarity (Invited Talk "Design for Inference in Drug Discovery and Development" by Aviv Regev, ICML 2022), which employs a standard triplet loss design, has achieved impressive performance in identifying similar cells in a massive collection of scRNA-Seq datasets. The results can help answer questions like in which tissues and diseases we find fibrotic macrophage-like cells.

[15] proposes the Deep Wasserstein Metric Learning Framework, as shown in Figure 1, which conducts multiple steps of adjustments over the original metric learning framework and achieves improved performance in image retrieval and clustering tasks. In the rest of the proposal, we introduce the adjustments accordingly and propose empirical studies for single-cell applications.

2 Methods

2.1 Representation Learning: Metric Learning and Deep Metric Learning

Representation learning is a class of machine learning approaches that allow a system to discover the representations required for feature detection or classification from raw data [38, 53, 18, 24, 17, 10, 26, 12, 55, 52, 42] The requirement for manual feature engineering is reduced by allowing a machine to learn the features and apply them to a given activity. Metric learning and deep metric learning, specifically, focus on similarity-based approaches to learning the representations. Thus the similarity measurement becomes very important. Previous work [7] studies similarity measurement in gene expression. Metric learning has only limited capability to capture non-linearity in the data, while deep metric learning captures non-linear features better by learning the non-linear transformation. The most widely used loss functions for deep metric learning are the contrastive loss and the triplet loss, both use euclidean distance to measure the distance between objects. A more comprehensive illustration of so-called "ranking-based" loss functions are summarized in Figure 2. Given an image pair, the contrastive loss minimizes their distance in the embedding space if their classes are the same, and separates them a fixed margin away otherwise. The triplet loss takes triplets of anchor, positive, and negative images, and enforces the distance between the anchor and the positive to be smaller than that between the anchor and the negative. The formation of contrastive loss is as the following. We first have embedding pairs $\mathcal{P}$, which is sampled from a minibatch of size b. The pair contains an anchor ϕ_a from class y_a and either a positive ϕ_p with $y_a = y_p$ or a negative ϕ_n from a different class, $y_a \neq y_n$. The distance function we utilize is the standard Euclidean distance $d_e(x, y) = \|x - y\|_2$. Then the network ϕ is trained to minimize:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{b} \sum_{(i,j)\in\mathcal{P}}^{b} \mathbb{I}_{y_i=y_j} d_e\left(\phi_i, \phi_j\right) + \mathbb{I}_{y_i\neq y_j} \left[\gamma - d_e\left(\phi_i, \phi_j\right)\right]_+ \tag{1}$$

Triplets extend the contrastive formulation by providing a triplets $\mathcal{T}$ sampled from a mini-batch:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{b} \sum_{\substack{(a,p,n)\in\mathcal{T} \\ y_a = y_p \neq y_n}}^{b} \left[d_e\left(\phi_a, \phi_p\right) - d_e\left(\phi_a, \phi_n\right) + \gamma \right]_+ \tag{2}$$

In the following, we present two adjustments that can improve (deep) metric learning's performances, one is optimal transport [9], and one is coreset optimization [31].

2.2 Optimal Transport

Optimal transport (OT) is the general problem of moving one distribution of mass to another as efficiently as possible [41, 15, 34, 49]. Optimal transport has been used tremendously in computational biology [39, 37, 44, 45]. For example, [39] uses scRNA-seq data collected across a time course to infer how these probability distributions evolve over time, by using the mathematical approach of optimal transport.

Wasserstein distance provides the mathematical tool to measure distances between functions, histograms, or more general objects in the optimal transport problem. Wasserstein distance is also called Earth Mover's Distance, which is employed to develop PhEMD (Phenotypic Earth Mover's Distance) [6], which is used to embed the space of drug perturbations on the basis of the drugs' effects on cell populations. Wasserstein distance-based loss functions have shown superior performance in learning tasks [20]. Thus a new set of loss functions are proposed to replace the Euclidean distance with Wasserstein distance in original contrastive loss and triplet loss by defining

$$d_w(x, y) = W_1(x, y) \tag{3}$$

The new Wasserstein-contrastive (wcontrastive) loss and Wasserstein-triplet (wtriplet) loss can be formulated as [15]:

$$\mathcal{L}_{\text{wcontrastive}} = \frac{1}{b} \sum_{(i,j)\in\mathcal{P}}^{b} \mathbb{I}_{y_i = y_j} d_w\left(\phi_i, \phi_j\right) + \mathbb{I}_{y_i \neq y_j} \left[\gamma - d_w\left(\phi_i, \phi_j\right)\right]_+ , \tag{4}$$

$$\mathcal{L}_{\text{wtriplet}} = \frac{1}{b} \sum_{\substack{(a,p,n)\in\mathcal{T} \\ y_a = y_p \neq y_n}}^{b} \left[d_w\left(\phi_a, \phi_p\right) - d_w\left(\phi_a, \phi_n\right) + \gamma \right]_+ . \tag{5}$$

We propose to apply the two new loss functions to similar cell identification and synthesis of heterogeneous modalities applications. In the similar cell identification task, the loss functions above can impose a discriminative constraint on the feature embedding to improve the similarity measurement [51].

2.3 Coreset Optimization

Coreset optimization is about data-efficient methods to find subsets of massive data that can generalize to the full data when trained on. In other words, a coreset is a subset of the original training set that is representative to train machine learning models [15, 54, 32, 28]. More specifically, Wasserstein measure coreset [8], is an extension of coresets that takes into account continuous data distribution and generalization. Recently coreset has been successfully applied to the purification of single-cell transcriptomics data [33]. It focuses on alleviating potential replicate-specific biases within single-cell datasets. The key is to select a "representative" subset (coreset) of cells from areas of the single-cell landscape where multiple replicates are represented. The approach [33] takes is solving the exemplar clustering problem, which minimizes the sum of pairwise dissimilarities between cells in the coreset and the rest of cells. We follow the idea in [33] but use Wasserstein distance to replace the Gaussian kernel to define similarity between cells. Here $r|V|$ is the exemplar cells from the groundset V and S is the targeted coreset.

$$S^* \in \arg\max \underset{|S| \leq r|V|}{S \subseteq V} \sum_{x\in V} \max_{y\in S} d_w(x, y) \tag{6}$$

[40] processes data from a Slide-seq replicate (three modalities with 20823 transcriptomes * 17607 genes) in 34 mins. [31] demonstrates a specific coreset optimization algorithm CRAIG can achieve

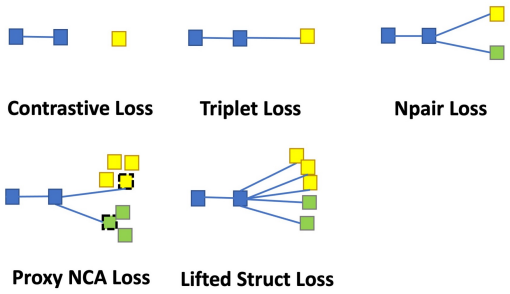

Figure 2: Illustration of different ranking-based loss functions. Different colors (blue, yellow, green) represent different classes. For simplicity, only 3 classes are shown. The left-most blue square is an anchor (query). In Contrastive loss, the anchor is compared with only one positive example. In Triplet loss, the anchor is compared with only one negative example and one positive example. In Npair, ProxyNCA, and Lifted Struct losses, one positive example and multiple negative classes are incorporated. Npair loss randomly selects one example per negative class. ProxyNCA loss pushes the anchor away from negative proxies instead of negative examples. The proxy is class-level and can represent any instance in the corresponding class. Lifted Struct loss uses all examples from all negative classes [16, 27].

the average speedup of 3x for similar loss residual and error rate. So we expect for the single cell synthesis task in [40], we can reduce data processing time from 34 mins to 11 mins. Furthremore, feature-efficient methods [1, 2, 48] can be applied to remove irrelevant variables during the training of coreset optimization algorithms.

3 Experiments Design

The experiments conducted on various datasets have demonstrated that optimal transport and coreset optimization can achieve superior performance on image retrieval and clustering tasks [15, 16]. In the following we lay out empirical studies to explore the applicability in building cell representations. The specific detail of computational studies is under investitation.

3.1 Cell Similarity Identification

For the cell similarity identification task, we plan to build on the datasets PBMC [25] and SLN-all [22] which are included in the phenomenal scvi-tools [21]. The PBMC dataset is measured with CITE-seq (161764 cells). The SLN-all dataset contains Immune cells from the murine spleen and lymph nodes.

3.2 Multimodal Integration

Regarding the multimodal integration challenge, we plan to follow the setup in [29] to apply the methods on multiple single-cell datasets including sci-CAR cell line [4], SNARE-seq cell line [5], and 10X Multiome T-cell depleted bone marrow [36] to validate the methods' effectiveness and develop new computational and biological insights from the downstream tasks.

4 Discussion

In this essay, we propose to apply two computational methods: optimal transport and coreset optimization, which are successfully demonstrated usability in image representation learning [15, 16], to cell representation learning with applications in cell similarity identification and multimodal integration. We will report results and insights in empirical studies in follow-up research.

References

[1] R. Bao, B. Gu, and H. Huang. Fast oscar and owl regression via safe screening rules. In *International Conference on Machine Learning*, pages 653–663. PMLR, 2020.

[2] R. Bao, B. Gu, and H. Huang. An accelerated doubly stochastic gradient method with faster explicit model identification. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 57–66, 2022.

[3] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives, 2012.

[4] J. Cao, D. A. Cusanovich, V. Ramani, D. Aghamirzaie, H. A. Pliner, A. J. Hill, R. M. Daza, J. L. McFaline-Figueroa, J. S. Packer, L. Christiansen, F. J. Steemers, A. C. Adey, C. Trapnell, and J. Shendure. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, 2018.

[5] S. Chen, B. B. Lake, and K. Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12):1452–1457.

[6] W. S. Chen, N. Zivanovic, D. van Dijk, G. Wolf, B. Bodenmiller, and S. Krishnaswamy. Embedding the single-cell experimental variable state space to reveal manifold structure of drug perturbation effects in breast cancer. *bioRxiv*, page 455436, 2018.

[7] M. Chikina. *Devising effective similarity measures and learning algorithms for the study of metazoan gene expression*. Princeton University, 2011.

[8] S. Claici, A. Genevay, and J. Solomon. Wasserstein measure coresets, 2020.

[9] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[10] X. Ding, L. Zhao, and L. Akoglu. Hyperparameter sensitivity in deep outlier detection: Analysis and a scalable hyper-ensemble solution. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[11] J. X. Dou. Impartial redistricting: a markov chain approach to the "gerrymandering problem". *arXiv preprint arXiv:1711.04618*, 2017.

[12] J. X. Dou and R. Bao. Clinical decision system using machine learning and deep learning: a survey.

[13] J. X. Dou, M. Jia, R. Bao, and H. H. Mao. Enhance 'similar'cell identification through optimal transport.

[14] J. X. Dou, M. Liu, H. Muneer, and A. Schlussel. What words do we use to lie?: Word choice in deceptive messages. *ArXiv*, abs/1710.00273, 2017.

[15] J. X. Dou, L. Luo, and R. M. Yang. An optimal transport approach to deep metric learning (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12935–12936, 2022.

[16] J. X. Dou, A. Q. Pan, R. Bao, H. H. Mao, L. Luo, and Z. Mao. Sampling through the lens of sequential decision making. *arXiv preprint arXiv:2208.08056*, 2022.

[17] J. X. Dou, N. Sun, and X. Zou. "draw my topics": Find desired topics fast from large scale of corpus. *ArXiv*, abs/1602.01428, 2016.

[18] Y. Ektefaie, G. Dasoulas, A. Noori, M. Farhat, and M. Zitnik. Geometric multimodal representation learning, 2022.

[19] M. Flores, Z. Liu, T. Zhang, M. M. Hasib, Y.-C. Chiu, Z. Ye, K. Paniagua, S. Jo, J. Zhang, S.-J. Gao, et al. Deep learning tackles single-cell analysis—a survey of deep learning for scrna-seq analysis. *Briefings in Bioinformatics*, 23(1):bbab531, 2022.

[20] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28, 2015.

[21] A. Gayoso, R. Lopez, G. Xing, P. Boyeau, V. Valiollah Pour Amiri, J. Hong, K. Wu, M. Jayasuriya, E. Mehlman, M. Langevin, Y. Liu, J. Samaran, G. Misrachi, A. Nazaret, O. Clivio, C. Xu, T. Ashuach, M. Gabitto, M. Lotfollahi, V. Svensson, E. da Veiga Beltrame, V. Kleshchevnikov, C. Talavera-López, L. Pachter, F. J. Theis, A. Streets, M. I. Jordan, J. Regier, and N. Yosef. A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, Feb 2022.

[22] A. Gayoso, Z. Steier, R. Lopez, J. Regier, K. L. Nazor, A. Streets, and N. Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature methods*, 18(3):272–282, 2021.

[23] S. Grebinoski, Q. Zhang, A. R. Cillo, S. Manne, H. Xiao, E. A. Brunazzi, T. Tabib, C. Cardello, C. G. Lian, G. F. Murphy, et al. Autoreactive cd8+ t cells are restrained by an exhaustion-like program that is maintained by lag3. *Nature Immunology*, 23(6):868–877, 2022.

[24] W. L. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[25] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[26] S. He, Y. Wang, S. Han, S. Zou, and F. Miao. A robust and constrained multi-agent reinforcement learning framework for electric vehicle amod systems. *arXiv preprint arXiv:2209.08230*, 2022.

[27] S. Kim, D. Kim, M. Cho, and S. Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247, 2020.

[28] T. Lin, Z. Zheng, E. Y. Chen, M. Cuturi, and M. I. Jordan. On projection robust optimal transport: Sample complexity and model misspecification. *arXiv preprint arXiv:2006.12301*, 2020.

[29] H. Mao, M. Jia, J. X. Dou, H. Zhang, and P. V. Benos. Coem: Cross-modal embedding for metacell identification. *arXiv preprint arXiv:2207.07734*, 2022.

[30] H. Mao, H. Liu, J. X. Dou, and P. V. Benos. Towards cross-modal causal structure and representation learning. In *Machine Learning for Health*, 2022.

[31] B. Mirzasoleiman, J. Bilmes, and J. Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning (ICML)*, July 2020.

[32] B. Mirzasoleiman, K. Cao, and J. Leskovec. Coresets for robust training of deep neural networks against noisy labels. *Advances in Neural Information Processing Systems*, 33, 2020.

[33] R. Pálovics, T. Wyss-Coray, and B. Mirzasoleiman. Purification of single-cell transcriptomics data with coreset selection.

[34] N. Papadakis. *Optimal transport for image processing*. PhD thesis, 2015.

[35] S. Paudel, B. E. Warner, R. Wang, J. Adams-Haduch, A. S. Reznik, J. X. Dou, Y. Huang, Y.-T. Gao, W.-P. Koh, A. Bäckerholm, J.-M. Yuan, and K. H. Y. Shair. Serological profiling using an Epstein-Barr virus mammalian expression library identifies EBNA1 IgA as a pre-diagnostic marker for nasopharyngeal carcinoma. *Clinical Cancer Research*, 09 2022. CCR-22-1600.

[36] S. Persad, Z.-N. Choo, C. Dien, I. Masilionis, R. Chaligné, T. Nawy, C. C. Brown, I. Pe'er, M. Setty, and D. Pe'er. Seacells: Inference of transcriptional and epigenomic cellular states from single-cell genomics data. *bioRxiv*, 2022.

[37] N. Prasad, K. Yang, and C. Uhler. Optimal transport using gans for lineage tracing, 2020.

[38] X. Qiu, Y. Zhang, J. D. Martin-Rufino, C. Weng, S. Hosseinzadeh, D. Yang, A. N. Pogson, M. Y. Hein, K. H. J. Min, L. Wang, et al. Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711, 2022.

[39] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[40] R. Singh, B. L. Hie, A. Narayan, and B. Berger. Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities. *Genome biology*, 22(1):1–24, 2021.

[41] R. Singh, J. S. S. Li, S. G. Tattikota, Y. Liu, J. Xu, Y. Hu, N. Perrimon, and B. Berger. Optimal transport analysis of single-cell transcriptomics directs hypotheses prioritization and validation. *bioRxiv*, 2022.

[42] J. X. Tan and T. Finkel. A phosphoinositide signalling pathway mediates rapid lysosomal repair. *Nature*, 609(7928):815–821, 2022.

[43] A. Tanay and A. Regev. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637):331–338, 2017.

[44] A. Tong, J. Huang, G. Wolf, D. Van Dijk, and S. Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pages 9526–9536. PMLR, 2020.

[45] A. Y. Tong, G. Huguet, A. Natik, K. MacDonald, M. Kuchroo, R. Coifman, G. Wolf, and S. Krishnaswamy. Diffusion earth mover's distance and distribution embeddings. In *International Conference on Machine Learning*, pages 10336–10346. PMLR, 2021.

[46] L. C. Wu, J. X. Dou, D. Sleator, A. M. Frieze, and D. Miller. Impartial redistricting: A markov chain approach. *ArXiv*, abs/1510.03247, 2015.

[47] L. Xiong, K. Tian, Y. Li, W. Ning, X. Gao, and Q. C. Zhang. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nature communications*, 13(1):1–17, 2022.

[48] A. Xu and H. Huang. Detached error feedback for distributed SGD with random sparsification. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24550–24575. PMLR, 17–23 Jul 2022.

[49] H. Xu, J. Liu, D. Luo, and L. Carin. Representing graphs via gromov-wasserstein factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[50] Q. Xu, Y. Yang, X. Zhang, and J. J. Cai. Association of pyroptosis and severeness of covid-19 as revealed by integrated single-cell transcriptome data analysis. *ImmunoInformatics*, 6:100013, 2022.

[51] J. Yan, E. Yang, C. Deng, and H. Huang. Metricformer: A unified perspective of correlation exploring in similarity learning. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[52] H. Yang and J. X. Tan. The pitt pathway: Keeping lysosomes young. *Clinical and Translational Medicine*, 12(10), 2022.

[53] K. D. Yang, A. Belyaeva, S. Venkatachalapathy, K. Damodaran, A. Katcoff, A. Radhakrishnan, G. Shivashankar, and C. Uhler. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature communications*, 12(1):1–10, 2021.

[54] Y. Yang, T. Y. Liu, and B. Mirzasoleiman. Not all poisons are created equal: Robust training against data poisoning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25154–25165. PMLR, 17–23 Jul 2022.

[55] T.-H. Zhang, M. M. Hasib, Y.-C. Chiu, Z.-F. Han, Y.-F. Jin, M. Flores, Y. Chen, and Y. Huang. Transformer for gene expression modeling (t-gem): An interpretable deep learning model for gene expression-based phenotype predictions. *Cancers*, 14(19):4763, 2022.

[56] H. Zhong, S. Liu, F. Cao, Y. Zhao, J. Zhou, F. Tang, Z. Peng, Y. Li, S. Xu, C. Wang, et al. Dissecting tumor antigens and immune subtypes of glioma to develop mrna vaccine. *Frontiers in immunology*, 12, 2021.