

# CASTLE: Cell Annotation in Spatial Transcriptomics through Learned Embedding Clustering

WeiQi (Grace) Li   Yun-Kuei Lin   Alina Hyk  
Oregon State University

liweiq, linyunk, hyka@oregonstate.edu

## Abstract

*Cell type annotation in Spatial Transcriptomics datasets is a critical step for downstream genomic analysis. However, existing methods fail to take full advantage of the rich information. They also rely heavily on external datasets and frame the problem as a classification task, which overlook intrinsic dataset structure and limit discovery of novel cell types. We propose CASTLE, a clustering framework that integrates foundational models for multi-modal feature extraction and uses Deep Embedding Clustering to jointly optimize representation learning and clustering. Our results suggest encoded multi-modal features can increase the separability of among cell groups, providing key insights into the role of foundational model embeddings in enhancing cluster separation. We also demonstrate the superiority of Stacked Denoising AutoEncoder (SDAE) for fusing multi-modal features comparing to conventional clustering such as PCA and UMAP.*

## 1. Introduction

Spatial transcriptomics (ST) captures gene expression with spatial resolution, enabling deeper insights into tissue structure, development, and disease. A key first step in analyzing ST data is cell type annotation, which supports downstream studies such as cell to cell communication mechanisms and spatial gene regulation. Even small errors in cell type annotation can propagate through analyses, leading to false discoveries and wasted research efforts.

Cell type annotation remains challenging for ST datasets due to its multimodal nature – combining gene expression, spatial coordinates, and histological images– and lack of high-quality ground truth annotations. While each modality contains biologically relevant information, existing methods often fail to integrate them effectively. For example, SpaGCN—an algorithm proposed for identifying spatial domain within tissue samples using ST datasets—have demonstrated effective use of multi-modality dataset can lead to

further meaningful biological insights [6]. SpatialID, a deep learning based cell type annotation method, has shown the spatial information can boost classification accuracy by 8% [8].

In addition, most existing methods requires an annotated single cell RNA-seq dataset to transfer labels to ST dataset based on feature similarity. While this approach reduces manual effort of assigning cell types to each clustered cell group, it introduces limitations: annotation accuracy is highly dependent on the quality and compatibility of the reference, and novel or rare cell types—particularly relevant in fields like neuron science and developmental biology—may be missed entirely [5].

In this work, we present CASTLE, a clustering framework designed to address multimodal integration challenges and dependency on reference dataset. We leverage foundational models to better capture the cell type distinction contained in each modality, and use deep embedding clustering (DEC) algorithm to fuse information together and facilitate separation of clusters. Although we haven’t achieved clear state of art performance so far, the experiments have generated meaningful insights that foundational model embeddings lead to better cluster separation. In addition, Stacked Denoising AutoEncoder (SDAE) as a fusion strategy does outperform classic dimension reduction methods.

## 2. Related Works

Despite being a relatively recent technology, spatial transcriptomics has inspired the development of numerous cell type annotation methods ranging from statistical models to deep learning techniques.

Statistical approaches typically operate solely on gene expression data and require reference dataset. These models often assume the gene count  $Y_{i,g}$  of gene  $g$  in cell  $i$  follows a distribution such as Poisson, where the expected value is determined by cell type. Bayesian inference is then applied to map cells to cell types based on prior knowledge of gene expression profiles from reference datasets [3, 7].

Deep learning methods introduce more flexibility by in-

corporating additional data types and complex model architectures. From instance, SpatialID uses reference dataset to pretrain a feedforward neuron network (FNN) for cell classification. This model generates pseudo-labels for ST cells, which are then used to supervise a graph based autoencoder trained on spatial coordinates and gene expression data. While powerful, such methods still inherit limitations from reference dependency and underutilize available multimodal information.

Our approach addresses these gaps through three innovations. First, we use foundational models to extract semantically rich features from each modality. Second we fuse these features using a Stacked Denoising AutoEncoder (SDAE) to enhance representation quality. Third, we reframe the cell type annotation as an unsupervised clustering problem using the Deep Embedding Clustering (DEC) framework. This allows for label-free cell grouping and the discovery of novel cell populations, especially in under-annotated or complex tissue contexts.

### 3. Methods

#### 3.1. Dataset

This paper uses Xenium breast cancer microenvironment dataset containing H&E tissue image and spatially resolved expressions for 313 relevant genes [2]. The two pieces of breast tissues in the dataset contain 161,000 and 118,000 cells respectively. Three modalities of information are extracted at single cell level from the dataset: cell coordinate, gene expression, and cell image. Cell (x, y) coordinates are given in micrometers ( $\mu m$ ) in the frame of DAPI tissue image where cell nuclei are stained. Since the absolute location of each cell are converted into relative spatial information in the feature embedding stage, we make no further scale alignment to cell coordinates. We then extract the counts of each gene present in corresponding cells. Lastly, we crop 32x32 cell images centered at each cell from H&E images through affine transformation aligned coordinates. An overview of the preprocessing and entire clustering pipeline is illustrated in Figure 1.

The dataset includes ground truth labels for each cell; however, the documentation lacks details on how these labels were obtained [1]. In reality, they were derived using a clustering algorithm followed by manual annotation of clusters to cell types [4]. This reliance on raw gene expression and clustering renders the labels potentially unreliable, complicating their use as a benchmark as discussed in our experimental results.

#### 3.2. Feature Embedding

Most existing methods treat gene expression vector – i.e. counts of various genes – either as direct model input or reduce its dimensionality using conventional tech-

niques before feeding the resulting latent embedding into downstream models. To the best of our knowledge, this is the first work to utilize context-aware foundational model, Geneformer, to encode gene expression profiles for cell type annotation task.

Since there is currently no comprehensive study evaluating different modality-specific encoding strategies, we implemented and systematically compared several embeddings. For gene expression, we evaluated (1) raw gene expression values, (2) base version of Geneformer, (3) a fine-tuned Geneformer for cell classification task released by the original authors. The fine-tuning was conducted on a limited dataset, only four cell types and 38,000 cells [9]. Based on our experiment, the performance of fine-tuned Geneformer considerably deteriorates compare to base version. For spatial coordinates, we compared (1) mapping embeddings and (2) sinusoidal positional embeddings. Although the mapping embedding lacks inherent spatial information, its inclusion—originally unintended—provides a useful contrast that aids in interpreting evaluation metrics, as discussed in the conclusion. For cell images, we resize them to 224x224 pixels through interpolation and extracted features using a pretrained DINOv2 model.

#### 3.3. Deep Embedding Clustering

##### 3.3.1 Latent Space Learning

We found that deep embedding clustering (DEC) model fails to learn any meaningful group separation when the encoder cannot already capture informative representations. When initialized randomly, DEC attempts to separate clusters from nearly random noise, resulting in poor separation and unstable soft clustering entropy. To address this, we first pretrain an AutoEncoder (AE) using reconstruction loss. To further enhance performance, we have also experimented with Stacked Denoising AutoEncoder (SDAE) [10].

We begin by concatenating the best-performing embeddings from all three modalities and project them into a latent space using  $AE_{encoder}$ , denoted as  $f_{\theta} : X \rightarrow Z$ , where  $\theta$  is learnable. The input dimension is kept smaller than that of  $Z$  to mitigate the curse of dimensionality. Once a meaningful latent representation is learned, the encoder is passed to the DEC module for clustering.

##### 3.3.2 Clustering

DEC defines training objective as a KL divergence loss between the soft assignments  $q_i$  and auxiliary distribution  $p_i$  [11].

$$L = \text{KL}(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (1)$$

Soft assignment  $q_i$  is a Student's t-distribution to measure the similarity.  $q_{ij}$  represents the probability of assigning

sample  $i$  to cluster  $j$  (i.e., a soft assignment).

$$q_{ij} = \frac{(1 + \|z_i - \mu_j\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{j'} (1 + \|z_i - \mu_{j'}\|^2/\alpha)^{-\frac{\alpha+1}{2}}} \quad (2)$$

Auxiliary distribution  $p_i$  is a target distribution that is defined to improve cluster purity and assignment confidence, computed by  $q_{ij}^2$  and then normalizing by frequency per cluster.

$$p_{ij} = \frac{q_{ij}^2/f_j}{\sum_{j'} q_{ij'}^2/f_{j'}}$$

$f_j = \sum_i q_{ij}$  are soft cluster frequencies.

## 4. Experiment Results

The accuracy of state-of-art methods ranges from 60% to above 90% depending on the tissue sample [8]. Our method does not immediately reach comparable performance, prompting a deeper examination of both the problem and our approach, rather than a direct benchmark against existing methods on this dataset. There are two major complexities underlie our task.

### 4.1. Evaluation Metrics

First and foremost, most existing studies implicitly assume the authenticity of the ground truth labels [4, 7, 8]. However, these labels are typically derived from clustering algorithms followed by manual annotation—assigning each cluster to specific cell types [1, 4]. The ground truth is then naturally biased towards the features and clustering originally used. While cell type annotation is a fundamental task in genomics, no spatial transcriptomics dataset has been curated with universally accepted ground truth labels, complicating the evaluation of model performance [4].

Given this complication, we use two ground-truth dependent metrics—majority vote accuracy and label entropy—alongside one unsupervised metric, the silhouette score. To make the ground truth based metrics applicable, we set the number of clusters to the number of ground truth labels in our experiments.

#### 4.1.1 Majority Vote Accuracy

Let  $C$  be the set of predicted clusters,  $Y$  be the set of ground truth labels,  $N$  be the total number of cells,  $c_i$  be the predicted cluster label of sample  $i$ ,  $y_i$  be the ground truth label of sample  $i$ . In calculation of majority vote accuracy, we first assign each predicted cluster to the ground truth label that constitutes the majority within that cluster.

$$\text{map}(c) = \arg \max_{y \in Y} |\{i : c_i = c \text{ and } y_i = y\}| \quad (3)$$

We then evaluate accuracy by measuring the agreement between these mapped labels and the true labels across all cells.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\text{map}(c_i) = y_i\} \quad (4)$$

#### 4.1.2 Label Entropy

For each cluster  $c \in C$ , define the proportion of label  $y \in Y$  in cluster  $c$  as:

$$p_{c,y} = \frac{|\{i : c_i = c \text{ and } y_i = y\}|}{|\{i : c_i = c\}|} \quad (5)$$

The entropy of cluster  $c$  is:

$$H(c) = - \sum_{y \in Y} p_{c,y} \log p_{c,y} \quad (6)$$

The overall label entropy is the average entropy across all clusters, weighted by cluster size.

$$\text{Label Entropy} = \sum_{c \in C} \frac{|\{i : c_i = c\}|}{n} H(c) \quad (7)$$

#### 4.1.3 Silhouette Score

Let  $x_i$  be a data point assigned to cluster  $C_i$ . Define:

- $a(i)$ : the average distance from  $x_i$  to all other points in the same cluster  $C_i$ .
- $b(i)$ : the minimum average distance from  $x_i$  to all points in any other cluster  $C_j$ , where  $j \neq i$ .

Then the silhouette score for sample  $x_i$  is:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (8)$$

The overall silhouette score is the mean across all samples.

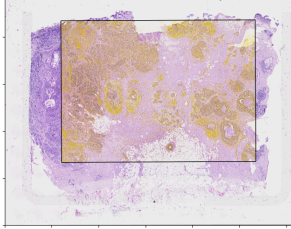
## 4.2. Experiment Setup

In addition, our method deviates in multiple aspects compared to existing deep learning based cell type annotation methods [8, 12], and thus it introduces multiple hypotheses. To identify opportunity for improvement, we explicitly assess whether each hypothesis contributes positively to the final outcome:

- Hypothesis 1: The context-aware foundational model Geneformer better captures cell-type distinctions than raw gene expression or conventional dimensionality reduction methods (PCA and UMAP).

**a. Preprocessing:** Extract gene expression, coordinates and cell image at cell level.

**a1. Crop cell images:** centered at each cell's aligned coordinate, crop a 32x32 cell images. The small black boxes below represent one crop of cell image.



**a2. Interpolate:** We interpolate the image to 224x224 images.

Original 32x32 Image



Interpolated 224x224 Image

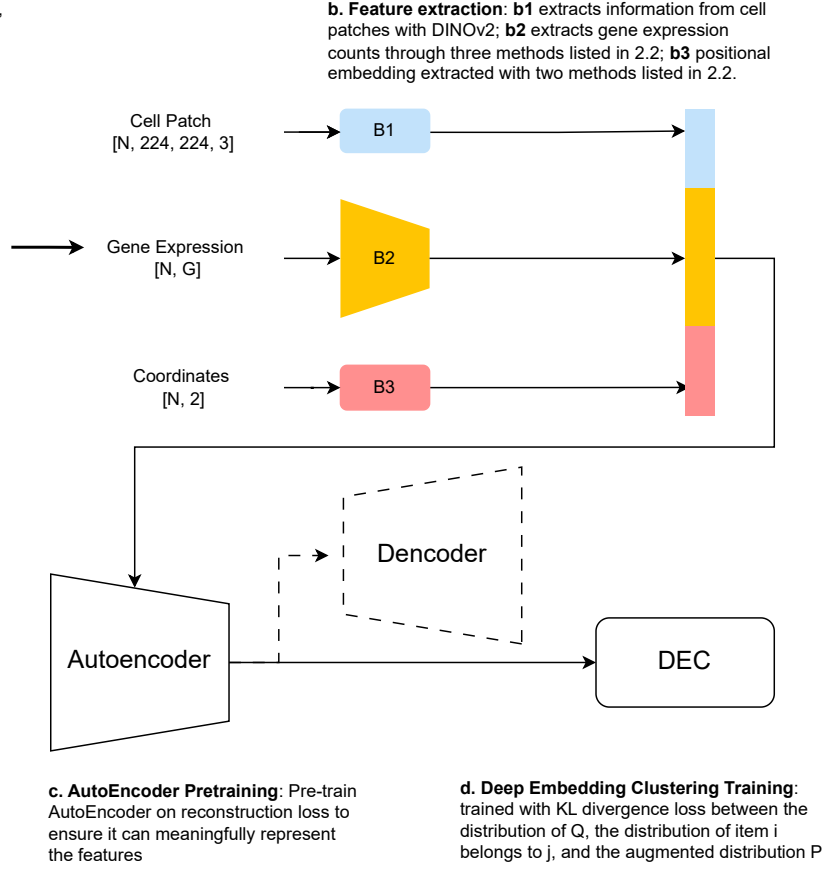


Figure 1. Overview of clustering pipeline: the workflow consists of four main stages: (a) preprocessing cell images including cropping and interpolation, extract gene expression and cell coordinates at cell level, (b) modality-specific feature extraction (c) unsupervised pretraining of AutoEncoder to learn latent representations, and (d) deep embedding clustering using the pretrained encoder.

- Hypothesis 2: Incorporating cell images and spatial coordinates provides additional discriminative power of cell type annotation.
- Hypothesis 3: the pretrained encoder enhances informative features rather than introducing noise.
- Hypothesis 4: the DEC model outperforms conventional clustering algorithms in this specific context.

#### 4.2.1 Evaluating Feature Representation

It is observed that the optimal DEC hyperparameters vary substantially across different embeddings, likely due to differences in their dimensionality. Thus, DEC model's performance cannot reliably reflect the quality of feature representation without also accounting for the quality of hyperpa-

rameter tuning. To isolate the effectiveness of each embedding from the influence of DEC-specific tuning, we evaluate the embeddings independently using conventional dimensionality reduction methods (PCA and UMAP) following by standard clustering algorithms (K-means and Louvain). As shown in Table 1, the evaluation was performed for both single-modality embeddings and their combinations.

#### 4.2.2 Evaluating the DEC Framework

In order to draw some conclusion regarding hypothesis 3 and 4, we introduce simplifying assumptions to reduce the complexity of evaluation. Specifically, for assessing DEC clustering performance, we assume the ground truth label is reliable. In addition, we use only the concatenation of the embedding with highest accuracy from each single modal-

Metric	Clustering	Features	Gene Expression			Coordinates		Image	Multi-modal		
			Raw	Geneformer	GF-Tuned	MapEmb	Cosine	DINOv2	A+B	A+C	A+B+C
Accuracy (%)	K-means	PCA	64.60	46.70	29.49	23.40	23.40	38.60	65.71	65.14	65.47
		UMAP	67.10	47.80	29.20	23.70	23.71	39.00	67.32	66.45	66.07
	Louvain	PCA	64.40	45.80	29.18	23.30	23.93	37.70	61.14	66.54	64.82
		UMAP	<b>69.30</b>	48.80	30.16	<b>25.00</b>	24.60	38.10	<b>70.40</b>	67.59	66.21
Silhouette (%)	K-means	PCA	12.70	15.40	31.97	2.30	9.54	7.00	12.98	11.87	11.23
		UMAP	36.10	<b>49.10</b>	43.91	9.00	28.72	26.60	36.54	<b>37.55</b>	36.22
	Louvain	PCA	11.20	15.70	26.23	0.10	4.50	6.30	9.95	9.10	10.38
		UMAP	31.40	39.20	38.08	6.20	<b>30.32</b>	20.30	32.51	32.03	32.20
Entropy	K-means	PCA	9.98	10.22	10.40	10.87	11.19	9.88	9.86	9.85	9.99
		UMAP	9.75	10.23	10.49	10.76	11.03	10.05	9.76	9.81	9.95
	Louvain	PCA	10.02	10.22	10.62	11.15	10.77	10.62	10.09	9.73	10.06
		UMAP	<b>9.61</b>	10.03	10.34	10.73	<b>10.52</b>	10.32	<b>9.62</b>	9.58	9.65

Table 1. Clustering performance across individual modalities—Gene Expression (A), Coordinates (B), and Image (C)—and their combinations. Geneformer Fine-Tuned (GF-Tuned) denotes a model fine-tuned on cell classification. In multi-modal representations, A+B signifies concatenation of gene expression and coordinate features; A+C gene expression and image; A+B+C all three. All features are reduced to a 32-dimensional latent space.

Clustering	Latent Emb	Accuracy
DEC	AE	67.80
	SDAE	66.76
Kmeans	PCA	65.47
	UMAP	66.07
	AE	67.90
	SDAE	<b>70.24</b>
Louvain	PCA	66.28
	UMAP	66.21
	AE	65.76
	SDAE	67.30

Table 2. The fine-tuned hyperparameters include hidden layer sizes, pretraining learning rate, SDAE noise factor, DEC learning rate, scheduler gamma, and Student’s  $t$ -distribution parameter  $\alpha$ .

ity. We normalize the representation of cells by each feature and then compare the performance of DEC model using autoencoder and denoising autoencoder to conventional dimension reduction and clustering methods.

## 5. Conclusion and Discussion

Unfortunately, the ground truth label was derived by clustering raw gene expression, and therefore inherently biased towards that representation. As a result, we cannot draw a definite conclusion about the superiority of different embeddings for cell type annotation. However, experiment 1 does lead to compelling hypothesis that, to our knowledge, has not been discussed in prior literature.

While Geneformer embeddings yield lower accuracy

compared to raw gene expression (48.80% vs. 63.90%, Table 1), they achieve substantially higher silhouette scores (Silhouette score 49.10% vs. 36.10%, Table 1), suggesting strong cluster separation. The interpretability of silhouette score is indirectly supported by the contrast between cosine and map embedding of spatial coordinates (30.32% vs. 6.20%, Table 2). The map embedding is essentially random, while cosine embedding encodes meaningful spatial relationships. The high silhouette score of cosine embedding reflects this distinction, although it may capture spatial patterns rather than cell type distinction. This raises an important question: does the Geneformer embedding separate biologically meaningful cell clusters, or does it simply enhances cluster separability without biological relevance? While answering this question is beyond the time frame of the current study, further investigation into the biological identity of Geneformer derived clusters would be a valuable direction for future research.

In addition, hypothesis 2 doesn’t yield a conclusive outcome. As shown in Table 1, it appears that the multi-modal representations do offer slight improvement in accuracy over gene expression alone, but they perform worse in other two metrics. However, the observed difference in accuracy and entropy may fall within margin of error. Further more, the lower silhouette scores observed for multi-modal embeddings can’t definitively reject the hypothesis. The reduction in silhouette might be attributed to compression of multi-modal representation from much higher dimensionality to 32-dimensional latent space. Conventional dimensionality reduction methods like PCA and UMAP may not adequately preserve the structure of high-dimensional, sparse inputs, potentially distorting the cluster separation.

We can observe that SDAE has a clear advantage over

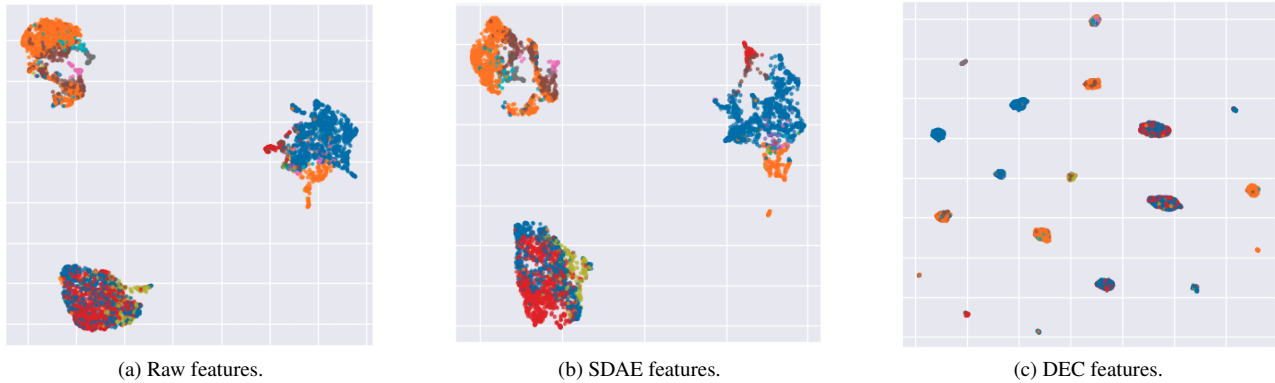


Figure 2. Comparison of UMAP 2D projection of representations of cells at different stage in the pipeline: (a) concatenated features including raw gene expression, spatial embedding and image embedding; (b) SDAE embeddings (c) DEC encoder embeddings. The color of the scatters represent ground truth labels.

any classic dimension reduction techniques. However, despite that DEC training has converged and clusters have far wider separations as shown in Figure 2c, the result of DEC doesn’t align well with the provided ground truth. It appears that SDAE embeddings can separate red and blue better than concatenated features, but DEC later mixes them into same clusters. This might be attributed to the highly imbalanced nature of cell datasets. During DEC training, the centroids end up being positioned around the dominant classes. As a result, samples from minority classes often assign to incorrect clusters, leading to cluster that do not accurately reflect the true distribution of each class.

## References

- [1] 10x Genomics. Xenium in situ preview dataset: Human breast. <https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast>, 2023. Document Number CG000689 Rev A, 10x Genomics, (2023, September 8). **2, 3**
- [2] 10x Genomics. Human breast tissue, xenium in situ preview dataset, 2025. Accessed June 12, 2025. **2**
- [3] Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandrina Goeva, Evan Z. Macosko, Fei Chen, and Rafael A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4):517–526, Apr. 2022. Epub2021Feb18. **1**
- [4] Jiarui Cheng, Xiaohui Jin, Gordon K. Smyth, and Yunshun Chen. Benchmarking cell type annotation methods for 10x xenium spatial transcriptomics data. *BMC Bioinformatics*, 26(1):22, 2025. **2, 3**
- [5] Lucie C. Gaspard-Boulin, Luca Gortana, Thomas Walter, Emmanuel Barillot, and Florence M. G. Cavalli. Cell-type deconvolution methods for spatial transcriptomics. *Nature Reviews Genetics*, 2025. PMID: 40369312. **1**
- [6] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J. Irwin, Edward B. Lee, Russell T. Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11):1342–1351, Nov. 2021. **1**
- [7] Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W. King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, Mika SarkinJain, Jun Sung Park, Lauma Ramona, Elizabeth Tuck, Anna Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, and Omer Ali Bayraktar. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 40(5):661–671, 2022. **1, 3**
- [8] Rongbo Shen, Ziqi Zhang, Lingxiao Zhang, Yu Li, and Cheng Li. Spatial-id: A cell typing method for spatially resolved transcriptomics via transfer learning and spatial embedding. *Nature Communications*, 13(1):7673, 2022. **1, 3**
- [9] Christina V. Theodoris et al. A foundation model for gene expression programs. *Nature*, 617(7960):311–319, 2023. **2**
- [10] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, volume 307 of *ACM International Conference Proceeding Series*, pages 1096–1103. ACM, 2008. **2**
- [11] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning (ICML)*, pages 478–487, 2016. **2**
- [12] Roxana Zahedi, Reza Ghamsari, Ahmadreza Argha, Calum Macphillamy, Amin Beheshti, Roohallah Alizadehsani, Nigel H. Lovell, Mohammad Lotfollahi, and Hamid Alinejad-Rokny. Deep learning in spatially resolved transcriptomics: A comprehensive technical view. *Briefings in Bioinformatics*, 25(2):1–19, 2024. **3**