SPELL: SPATIAL PROMPTING WITH CHAIN-OF-THOUGHT FOR ZERO-SHOT LEARNING IN SPATIAL TRANSCRIPTOMICS

Sumeer A. Khan[1,2+] , Xabier Martinez de Morentin [1+,*], Abdel Rahman Alsabbagh[3], Vincenzo Lagani[1,2], Robert Lehmann[1], Mahmoud Zahran[3], Narsis A. Kiani[4], David Gomez-Cabrero[1], Jesper Tegnér[1,3,5,6] †
[1]Biological and Environmental Science and Engineering Division (BESE)
King Abdullah University of Science and Technology (KAUST)
Thuwal,Saudi Arabia
{sumeer.khan,jesper.tegner}@kaust.edu.sa
[4]Algorithmic Dynamic Lab, Department of Oncology and Pathology
Karolinska Institute, Stockholm, Sweden
{narsis.kiani}@ki.se

ABSTRACT

Zero-shot learning (ZSL) for cell-type classification in spatially resolved transcriptomics remains underexplored, particularly when integrating spatial context with marker gene semantics. Here, we introduce SPELL (Spatial Prompt-Enhanced Zero-Shot Learning), combining graph autoencoder (GAE)-derived spatial embeddings with chain-of-thought (CoT) prompting for zero-shot classification. SPELL uses a spatial k-nearest neighbor graph to encode local cellular neighborhoods and generates interpretable prompts that integrate marker gene expression and the spatial embedding norms. We evaluated SPELL across five state-of-the-art zero-shot LLM classifiers on MERFISH, MIBI-TOF, and Stereo-seq datasets for cell-type classification. Guided by only expression values and spatial context, the two BART models solved the classification task surprisingly well (distilbart-mnli-12-1i 64% accuracy on the MERFISH, bart-large-mnli achieved 52% accuracy on MIBI-TOF dataset). Interestingly, removing the spatial context from the CoT prompt revealed a significant performance drop (20 – 24 % drop in accuracy), underscoring spatial information's critical role in zero-shot learning. Our work bridges spatial omics with LLM reasoning,enabling flexible adaptation and offering robust cell-type classification across diverse datasets without task-specific fine-tuning while maintaining biological interpretability.

1 INTRODUCTION

Spatial transcriptomics Ståhl et al. (2016) has transformed our understanding of tissue architecture by preserving spatial context while profiling gene expression at single-cell resolution. Techniques such as multiplexed error-robust fluorescence in situ hybridization (MERFISH) enable simultaneous imaging of hundreds of RNA species, thus capturing the cellular organization within intact tissues Chen et al. (2015); Moffitt et al. (2018). Recent progress in natural language processing - exemplified by zero-shot classification using large pretrained models enabled effective categorization of text into arbitrary labels without task-specific training,Brown et al. (2020); Wei et al. (2022). Zero-shot learning (ZSL) traditionally involves classifying unseen classes by transferring knowledge from seen classes via semantic embeddings.

[*]S.A.K and X.M.M contributed equally

[†]SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence[2],CEMSE KAUST[3], Unit of Computational Medicine, Department of Medicine, Center for Molecular Medicine, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden[5],
Science for Life Laboratory, Tomtebodavägen, Solna, Sweden.[6]

In this work, we mitigate the gap between these two lines of investigation. Our SPELL method integrates marker gene profiles with local spatial context derived from a graph autoencoder (GAE)-based embedding of a spatial k-nearest neighbor graph. We convert these biological and spatial features into a human-interpretable chain-of-thought (CoT) prompt used for zero-shot cell-type classification, as shown in Figure 1. Unlike methods treating gene expression and spatial coordinates as independent numerical features, SPELL encodes spatial relationships into human-interpretable chain-of-thought (CoT) prompts for zero-shot cell-type classification. SPELL adopts the large language model (LLM) zero-shot framework, where the model classifies cells into known classes without fine-tuning, using prompts enriched with training-derived marker genes. This differs from classical zero-shot learning, which typically involves classifying entirely unseen classes. This approach leverages labeled data for marker gene identification, common in spatial transcriptomics where annotated datasets guide analysis of new samples. We further demonstrate the importance of spatial context by comparing the entire model with an ablated version that uses only marker gene summaries in the prompt. Moreover, excluding the spatial context leads to a marked decrease in performance. We evaluated SPELL on MERFISH and MIBI-TOF datasets and extended our experiments to a Stereo-seq Drosophila spatial transcriptomics dataset, demonstrating the broad applicability of SPELL.

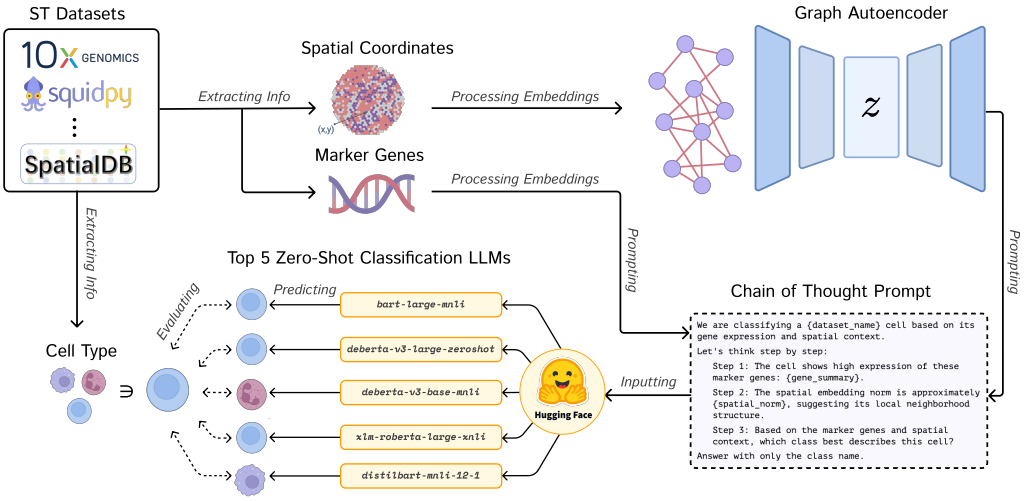

Figure 1. Integrated Workflow for Zero-Shot Classification in Spatial Transcriptomics: SPELL comprises three primary components: (i) extraction of marker genes, (ii) construction and embedding of a spatial graph using a Graph AutoEncoder (GAE), and (iii) generation of chain-of-thought prompts for zero-shot classification.

2 METHODS

2.1 DATASETS

The preprocessed MERFISH (Moffitt et al., 2018), MIBI-TOF(Hartmann et al., 2021) datasets (as curated in (Palla et al., 2022)) and the Stereo-seq Drosophila dataset (Qiu et al., 2024) were used to evaluate the zero-shot classification performance of various models within our SPELL framework. The MERFISH dataset comprises 12 consecutive slices from the mouse hypothalamic preoptic region, where each slice provides gene expression values, pre-classified cell types, and two-dimensional spatial coordinates, while as MIBI-TOF dataset provides single-cell metabolic profiles, phenotypes, and spatial organization of CD8+ T cells and colorectal carcinoma. In contrast, the Drosophila dataset provides a three-dimensional high-resolution gene expression map, capturing spatial and temporal dynamics during key developmental stages. This rich dataset illustrates the spatial heterogeneity and dynamic transcriptomic landscape observed during embryonic and larval development, making it an ideal test case for assessing the generalizability and robustness of our framework in a non-mammalian system. We selected a representative slice from each dataset for our analysis, as shown in Appendix A Figure 3.

2.2 SPELL FRAMEWORK

Our SPELL framework, illustrated in Figure 1, uses five zero-shot classification models to classify cell types in spatial transcriptomics. These models are sourced from the Hugging Face model repository, Hugging (2025), Appendix B.

2.2.1 MARKER GENE EXTRACTION

Using single cell spatial transcriptomic data (MERFISH) we first confirm that cell-level metadata includes preliminary cell-type annotations. We perform differential expression analysis using the Wilcoxon method to rank genes by their specificity to each cell class, employing the Scanpy pipeline. The top-ranking marker genes for each class are stored in a marker dictionary. This step ensures that the most discriminative molecular signatures are utilized to construct our natural language prompts.

2.2.2 SPATIAL GRAPH CONSTRUCTION AND EMBEDDING

We construct a k-nearest neighbor (k-NN) graph from cell spatial coordinates, connecting each cell to its 15 closest neighbors to form an undirected adjacency matrix that captures local spatial relationships. Gene expression profiles are processed via principal component analysis (PCA), retaining 20 principal components, and combined with the spatial graph to create a PyTorch Geometric Data object.

A Graph AutoEncoder (GAE) with two Graph Convolutional Networks (GCNs) is trained to reconstruct the graph structure using edge reconstruction loss. The resulting 20-dimensional latent embeddings encode each cell's local neighborhood structure, integrating spatial proximity and gene expression patterns. The norm of each embedding vector quantifies the cell's spatial context (e.g., density of neighbors), which is included in the chain-of-thought prompt to provide the large language model (LLM) with additional discriminative information beyond gene expression. This approach bridges spatial transcriptomics and natural language processing by translating complex spatial information into a numerical feature that the LLM can process as part of the prompt.

2.2.3 CHAIN-OF-THOUGHT PROMPT GENERATION AND ZERO-SHOT CLASSIFICATION

SPELL introduces chain-of-thought (CoT) prompting to leverage molecular and spatial information for each cell. Our zero-shot approach uses pre-trained LLMs to classify cells into known classes without fine-tuning, relying on prompts enriched with training-derived marker genes.The marker genes serve as domain knowledge in the prompt, akin to providing a dictionary of terms in a text classification task.The CoT prompt (Figure 1) comprises:

```
A summary of the top marker genes.
The norm of the spatial embedding quantifies the magnitude of
the spatial context,which may reflect how densely connected or
isolated a cell is within its neighborhood.
A structured reasoning sequence that proposes the likely cell type.
```

These prompts are then processed through a zero-shot classification pipeline, employing various models from our selection. Each model receives candidate cell-type labels and produces predictions with associated confidence scores. Even though LLMs are not explicitly trained on spatial data, they can treat the spatial norm as an additional feature or contextual cue within the prompt.

2.3 EVALUATION AND VISUALIZATION

We assessed the performance of the models using a test dataset comprising 100 cells. Key performance indicators included overall accuracy and F1 scores. To illustrate class-specific performance, we used confusion matrices. Furthermore, we conducted comparative analyses between the complete pipeline and a variant without spatial context, enabling a deeper understanding of the impact of spatial data on model effectiveness.

3 RESULTS AND DISCUSSION

Our evaluation focused on five distinct models within the SPELL framework, tested against the MERFISH, MIBI-TOF and Stereo-seq datasets. These models leverage marker gene data and spatial context through Chain-of-Thought (CoT) prompting. As shown in Figure 2 and Appendix A Figure 4, `distilbart-mnli-12-1` and `bart-large-mnli` achieved the highest accuracies on MERFISH and MIBI-TOF respectively, highlighting their capacity to handle complex classification tasks with notable efficiency. In contrast, on the Stereo-seq dataset, `distilbart-mnli-12-1i` and `bart-large-mnli` exhibited lower performance, highlighting how platform differences and added complexity can affect zero-shot classification as shown in Table 1.

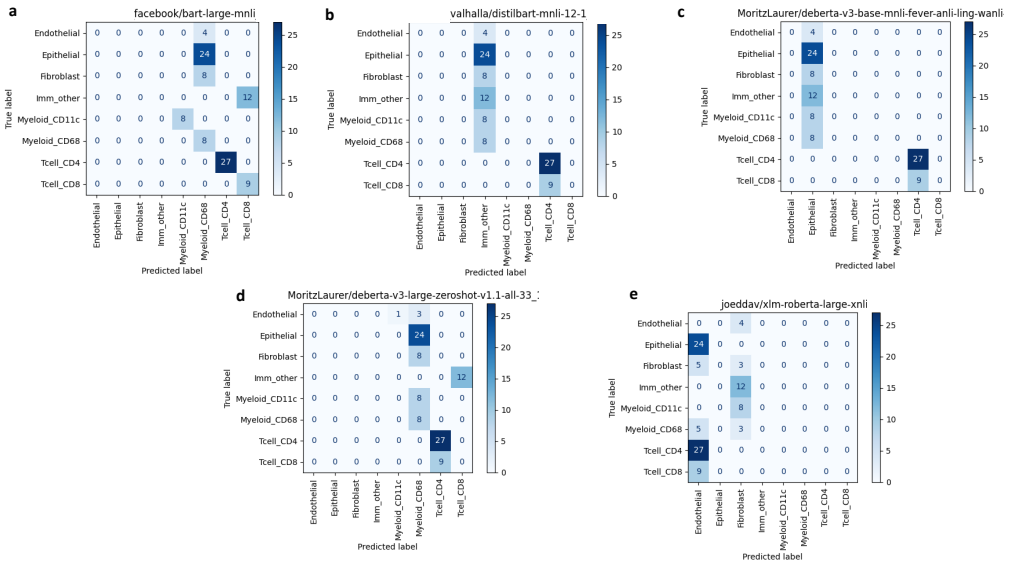

Figure 2. Comparative performance of zero-shot classification models on MIBI-TOF data, illustrating class-wise accuracy for cell types.

Table 1 highlights the strong performance of advanced architectures like BART, which leverages extensive natural language inference pretraining and rich contextual representations. Both `distilbart-mnli-12-1` and `bart-large-mnli` achieve high accuracy on MERFISH and MIBI-TOF data - with the distilled model retaining much of the critical pre-trained knowledge despite its reduced size - yet both struggle on the Stereo-Seq Drosophila dataset. This discrepancy underscores how different technical platforms present distinct classification challenges: MERFISH data may be less noisy and more uniform, while Drosophila data incorporate a temporal component that adds complexity. These findings suggest that future work must explicitly account for these platform-driven biases when deploying zero-shot models, potentially through hybrid architectures that dynamically adapt to dataset topology.

3.1 ABLATION STUDY

To assess the contribution of spatial context, we ablated the model such that the CoT prompt was generated solely using the marker gene summary (i.e., omitting the spatial embedding norm and related narrative, Appendix A). Classification performance was then compared between the full model and the ablated variant. As detailed in the Appendix A Table 2, the results demonstrate a significant decline in performance without the spatial context, highlighting its critical role in enhancing classification accuracy, thus confirming the essential role of spatial context, aligning with biological principles of microenvironment influence.

Table 1. Accuracy and F1 Scores for Zero-Shot Classification Models on MERFISH, MIBI-TOF, and Stereo-Seq Data

Model	MERFISH		MIBI-TOF		Stereo-Seq	
	Accuracy	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**
distilbart-mnli-12-1	**0.640**	**0.558**	0.390	0.269	**0.210**	**0.080**
bart-large-mnli	0.010	0.002	**0.520**	**0.429**	0.120	0.026
deberta-v3-base-mnli-fever-anli-ling-wanli-binary	0.150	0.039	0.510	0.362	0.080	0.012
deberta-v3-large-zeroshot-v1.1-all-33	0.040	0.003	0.350	0.253	0.020	0.001
xlm-roberta-large-xnli	0.160	0.075	0.030	0.013	0.070	0.009

4 CONCLUSION

This study introduces SPELL, a novel framework for zero-shot cell type classification in spatial transcriptomics, leveraging spatial prompting to integrate molecular and spatial data into structured natural language prompts. By harnessing the reasoning capabilities of pre-trained large language models (LLMs) without task-specific fine-tuning, SPELL uses chain-of-thought prompts enriched with training-derived marker genes and spatial embedding norms to classify cell types. We validated our methodology across MERFISH, MIBI-TOF, and Stereo-seq datasets, attaining good accuracy and F1 scores with interpretable reasoning steps. Ablation studies confirm that spatial context significantly boosts classification accuracy, highlighting its critical role. Future work will focus on integrating advanced spatial features, optimizing marker selection, and validating across diverse tissues and platforms. By minimizing the need for extensive labeled datasets, SPELL provides a scalable and interpretable solution for cell type classification in spatial transcriptomics.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Rna imaging. spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Felix J Hartmann, Dunja Mrdjen, Erin McCaffrey, David R Glass, Noah F Greenwald, Anusha Bharadwaj, Zumana Khair, Sanne GS Verberk, Alex Baranski, Reema Baskar, et al. Single-cell metabolic profiling of human cytotoxic t cells. *Nature biotechnology*, 39(2):186–197, 2021.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. corr abs/2111.09543 (2021). *arXiv preprint arXiv:2111.09543*, 2021.

Hugging. Hugging face: The ai community building the future, 2025. URL `https://huggingface.co/models?pipeline_tag=zero-shot-classification`. Accessed: 2025-02-12.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. Building efficient universal classifiers with natural language inference. *arXiv preprint arXiv:2312.17543*, 2023.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

J. R. Moffitt, D. Bambah-Mukku, S. W. Eichhorn, E. Vaughn, K. Shekhar, J. D. Perez, N. D. Rubinstein, J. Hao, A. Regev, C. Dulac, and X. Zhuang. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416):eaau5324, 2018.

Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L Ibarra, Olle Holmberg, Isaac Virshup, et al. Squidpy: a scalable framework for spatial omics analysis. *Nature methods*, 19(2):171–178, 2022.

Xiaojie Qiu, Daniel Y Zhu, Yifan Lu, Jiajun Yao, Zehua Jing, Kyung Hoi Min, Mengnan Cheng, Hailin Pan, Lulu Zuo, Samuel King, et al. Spatiotemporal modeling of molecular holograms. *Cell*, 187(26):7351–7373, 2024.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Peter L. Ståhl, Fredrik Salmén, Svante Vickovic, Anna Lundmark, Juan F. Navarro, Joakim Magnusson, and Joakim Lundeberg. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ethan Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

A Appendix

Prompt including spatial information.

```
prompt = (
    "We are classifying a MERFISH cell based on its gene expression and spatial conte
    "Let's think step by step:\n"
    f"Step 1: The cell shows high expression of these marker genes: {gene_summary}.\n
    f"Step 2: The spatial embedding norm is approximately {spatial_norm:.2f}, suggest
    "Step 3: Based on the marker genes and spatial context, which class best describe
    "Answer with only the class name."
)
```

Prompt with spatial information excluded.

```
prompt = (
    "We are classifying a MERFISH cell based on its gene expression and spatial conte
    "Let's think step by step:\n"
    f"Step 1: The cell shows high expression of these marker genes: {gene_summary}.\n
    "Step 2: Based on the marker genes,  which class best describes this cell?\n"
    "Answer with only the class name."
)
```

B Description of Zero-Shot Classification Models

DISTILBART-MNLI-12-1 (Lewis et al., 2019; Sanh et al., 2019)

Size & Architecture: A distilled version of BART, offering a smaller, more efficient model while preserving key capabilities.

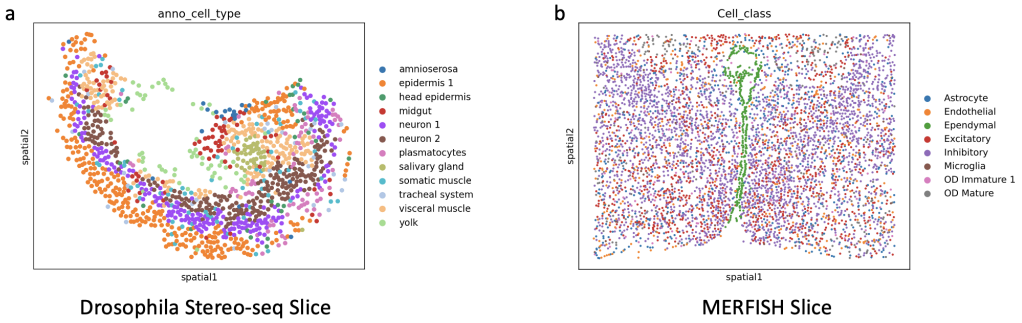

Figure 3. Representative slices from the Drosophila and MERFISH datasets used for evaluation in our SPELL framework. a)Representative slice from the Drosophila Stereo-seq dataset. b) MERFISH slice from the mouse hypothalamic preoptic region

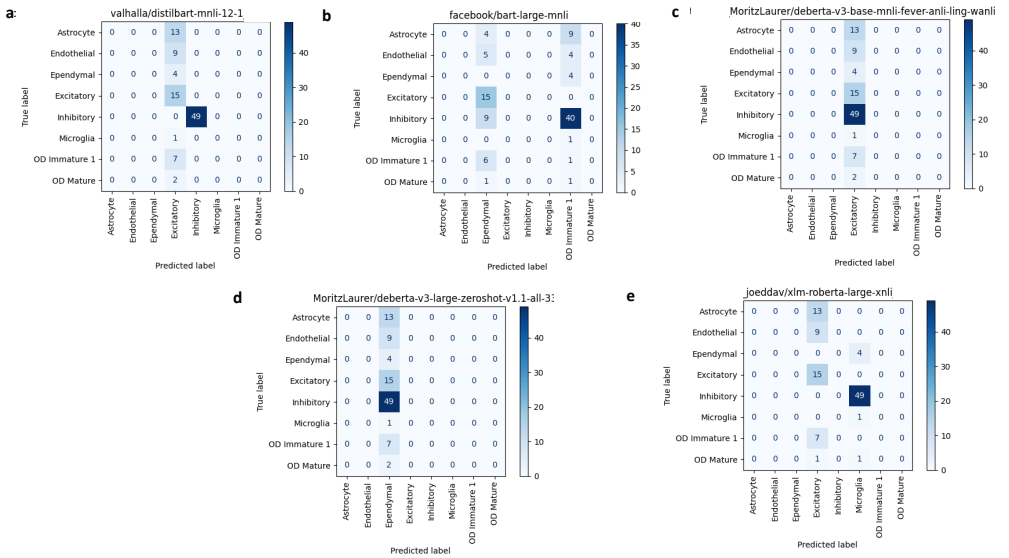

Figure 4. Comparative performance of zero-shot classification models on MERFISH Data. The figure displays confusion matrices for five different zero-shot learning models applied to classify cell types in MERFISH data. Each panel (a-e) corresponds to a model, illustrating the classification accuracy across various cell types

Training: Fine-tuned on the MNLI dataset for natural language inference (NLI) tasks.
Prior Performance: Matches 97% of BART-large's accuracy on MNLI with 40% faster inference .
Selection Rationale: Its efficiency and strong semantic reasoning make it attractive for rapid inference.

BART-LARGE-MNLI (LEWIS ET AL., 2019)

Size & Architecture: The full-scale BART model with a large number of parameters((406M parameters) and deep contextual representations.
Training: Pre-trained as a denoising autoencoder, then fine-tuned on MNLI.
Prior Performance: Achieves 90% + accuracy on MNLI and robust zero-shot transfer.
Selection Rationale: Gold-standard baseline for English zero-shot tasks..

Table 2. Accuracy and F1 scores for zero-shot classification models on MERFISH data excluding spatial embedding information in CoT prompt

Model Identifier	Accuracy	F1-Score
distilbart-mnli-12-1	0.490	0.338
xlm-roberta-large-xnli	0.150	0.039
deberta-v3-large-zeroshot-v1.1-all	0.020	0.003
deberta-v3-base-mnli-fever-anli-ling-wanli-binary	0.130	0.037
Bart-large-mnli	0.040	0.003

DEBERTA-V3-BASE-MNLI-FEVER-ANLI-LING-WANLI-BINARY (HE ET AL., 2020; 2021; LAURER ET AL., 2023)

Size & Architecture: DeBERTa-v3-base (183M parameters) with disentangled attention and enhanced mask decoder.
Training: Fine-tuned on MNLI, FEVER, ANLI, and LingWANLI reformulated as binary NLI tasks (entailment vs. non-entailment).
Prior Performance: Achieved state-of-the-art results on several NLI benchmarks.
Selection Rationale: Its diverse fine-tuning captures broad inference cues suitable for complex classification tasks.

DEBERTA-V3-LARGE-ZEROSHOT-V1.1-ALL-33 (HE ET AL., 2020; 2021; LAURER ET AL., 2023)

Size & Architecture: The large version of DeBERTa, with increased parameters for deeper semantic understanding.
Training: Fine-tuned on 33 datasets (e.g., MNLI, ANLI, SciTail) converted to NLI format for broad zero-shot generalization.
Prior Performance: Exhibits superior zero-shot accuracy, albeit at a higher computational cost.
Selection Rationale: Its enhanced capacity and reasoning abilities make it ideal for challenging datasets.

XLM-ROBERTA-LARGE-XNLI (CONNEAU ET AL., 2019)

Size & Architecture: XLM-RoBERTa-large (550M parameters), pre-trained on 100+ languages.
Training: Fine-tuned on XNLI (cross-lingual MNLI extension) for multilingual NLI.
Prior Performance: Demonstrated robustness and high zero-shot performance in multilingual contexts.
Selection Rationale: Its multilingual training enhances generalization and robustness across diverse domains.