evoCancerGPT: Generating Cancer Progression Using Single-Cell RNA Sequencing Data

Xi Wang*, Simona Cristea*

*Department of Data Science, Dana-Farber Cancer Institute

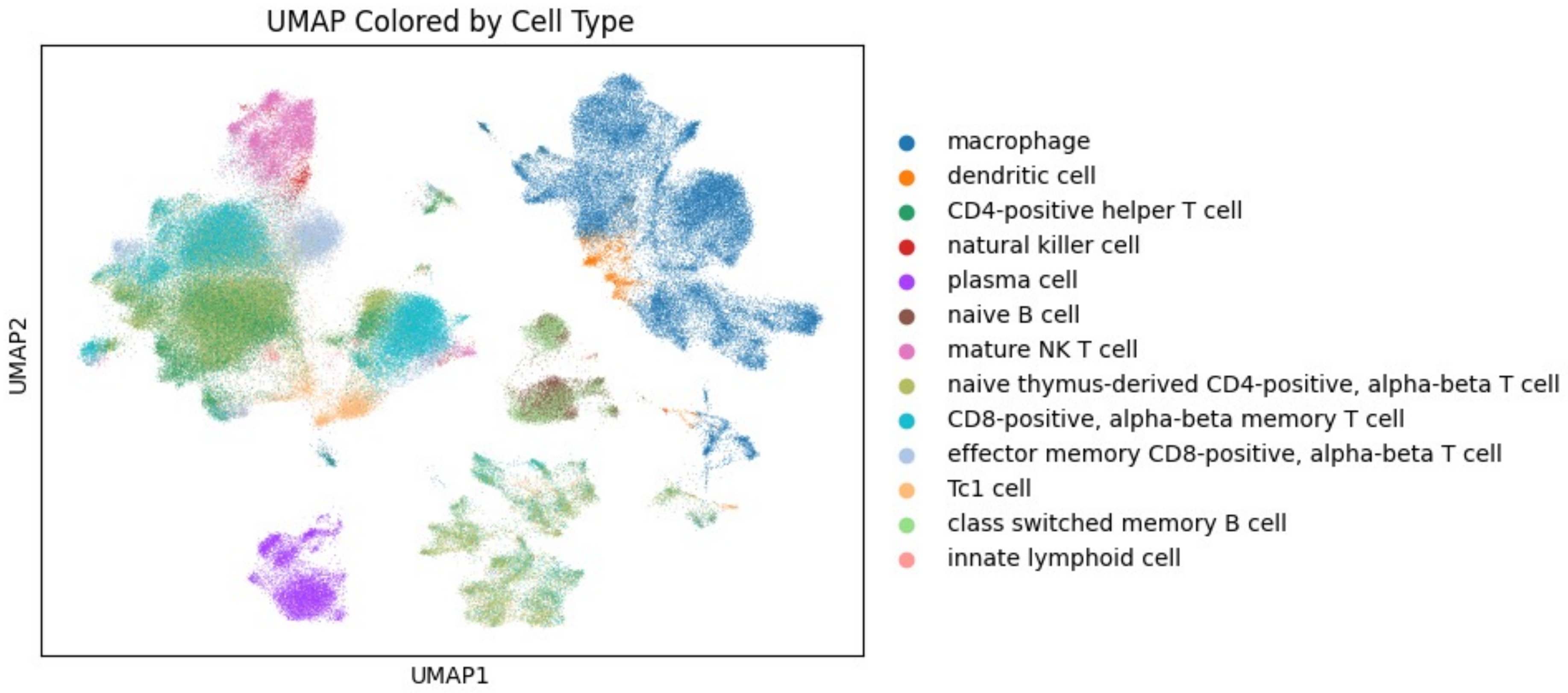
Dana-Farber Cancer Institute

Background

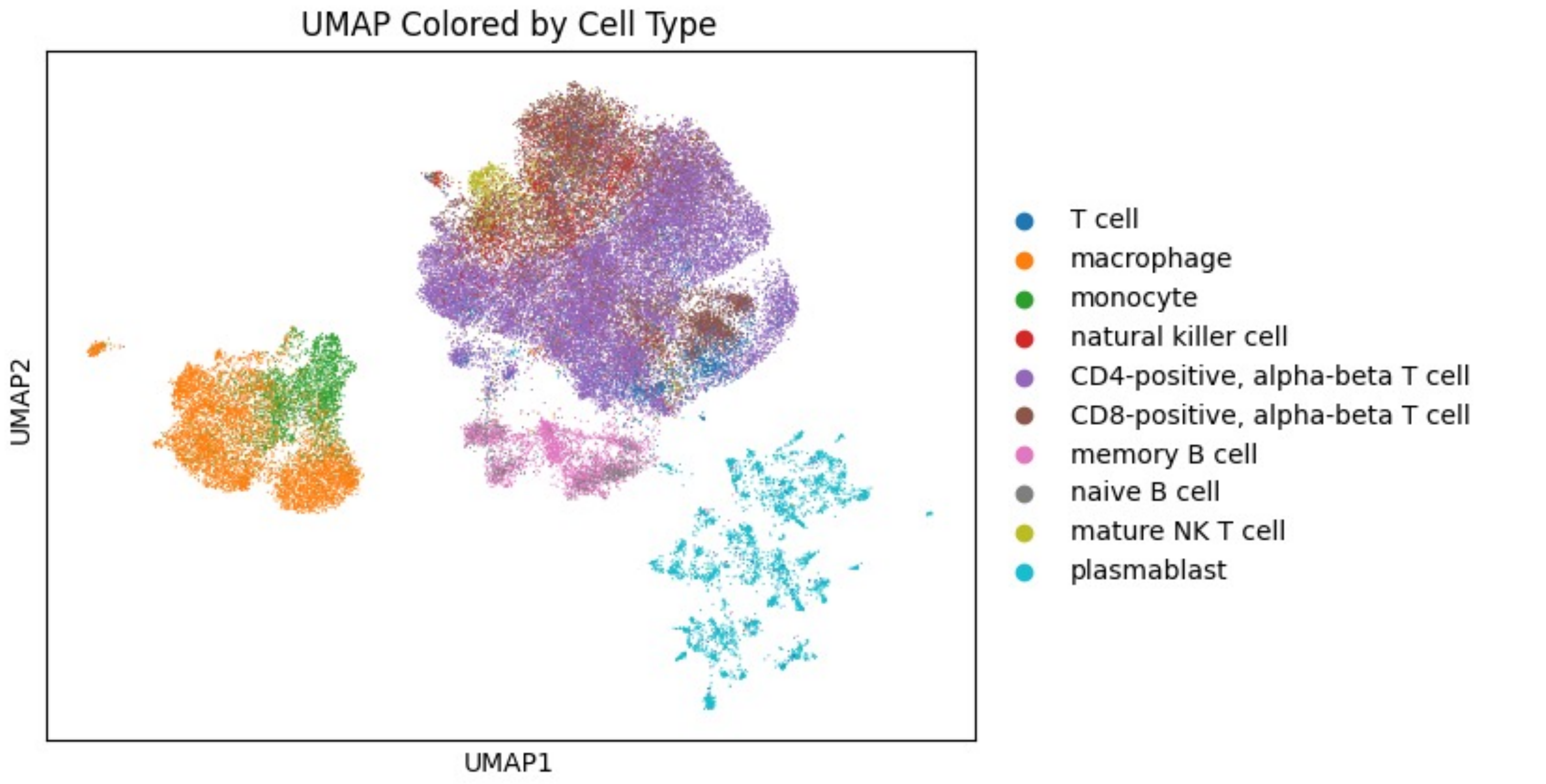
- Cancer evolution is driven by complex changes in gene expression as cells transition across normal, precursor disease, primary, and metastatic stages.
- Single-cell RNA sequencing has provided insights into how the transcriptomes of tumors evolve during tumorigenesis, but whether the existing knowledge can be used to reliably learn and generate tumor evolution remains unknown.
- Goal: We propose **evoCancerGPT**, a foundation model trained on a large corpus of normal and cancer cells arranged in pseudotime. evoCancerGPT can capture key transitions in cancer evolution, highlighting critical shifts in gene regulatory networks or programs driving cancer progression.

Data

- 2,112,065 cells from normal healthy human breast samples[1] from CZ CellXGene[2].



- 100,064 Cells from primary human breast cancer samples[3] from CZ CellXGene[2] .



Methods

Sentence Constructions: 5000 sentences of length 1000 from randomized sampling, ranked by pseudotime, with each cell as a token from 1 of the 3 major cell types.

Normal Datasets

	Cell N_1	Cell N_2	…	Cell N_N
g_1	1.3	0.84	…	3.4
g_2	0.0	0.2	…	0.04
…	…	…	…	…
g_{2000}	4.21	1.02	5.1	0.0

Primary Datasets

	Cell P_1	Cell P_2	…	Cell P_M
g'_1	0.0	3.09	…	0.0
g'_2	0.04	0.	…	6.56
…	…	…	…	…
g'_{2000}	1.45	0.16	0.0	0.0

Sentences

$$S_k : [N_1, N_2, \dots, N_i, P_1, \dots, P_j], \quad k \in \{1, \dots, 5000\}$$

$$\text{where } i \sim U(1, 1000), \quad i + j = 1000$$

Ranked by psuedotime:

$$T_{N_1} < T_{N_2} < \cdots < T_{N_N} \ll T_{P_1} < T_{P_2} < \cdots < T_{P_M}$$

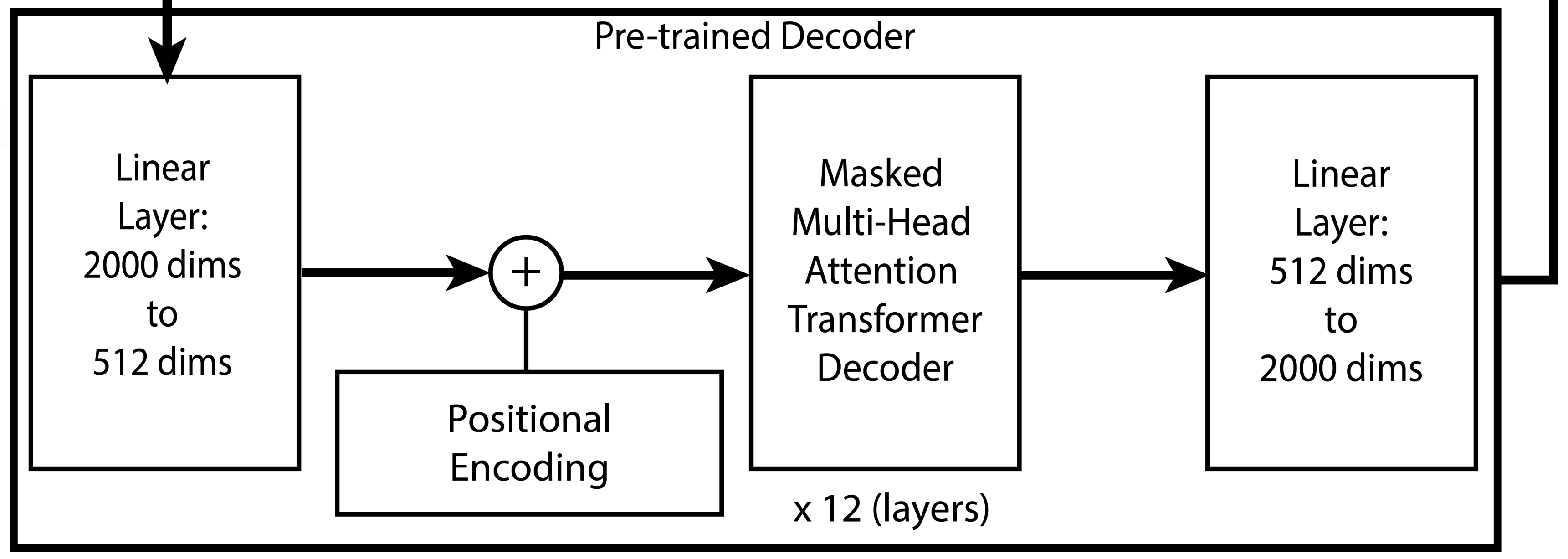
$$\text{where } N = 2,112,065 \quad \text{and} \quad M = 100,064$$

Tokenizer: CellType | N_1 | N_2 | ··· | N_i | P_1 | P_2 | ··· | P_j

categorical: immune, fibroblast, or epithelial_malignant

continuous expression value

continuous expression value


Pre-trained Decoder
Linear Layer: 2000 dims to 512 dims
+
Positional Encoding
Masked Multi-Head Attention Transformer Decoder
Linear Layer: 512 dims to 2000 dims
x 12 (layers)

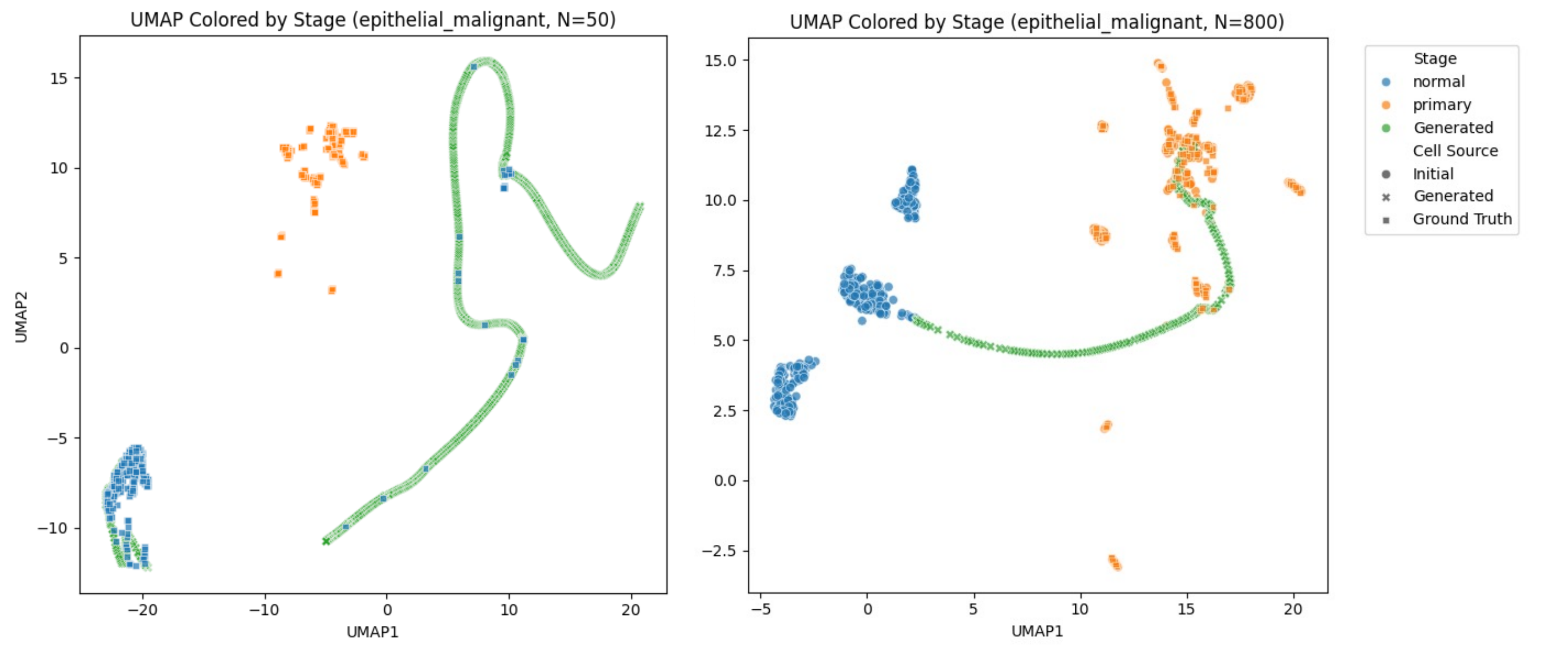
Zero shot generation: next cell (token) prediction

evoCancerGPT is a transformer-based decoder-only generative model that enables the simulation of future cancer cell states and tumor progression using GenAI, with the potential to improve our understanding of tumor progression and identify novel biomarkers, contributing to more personalized cancer care.

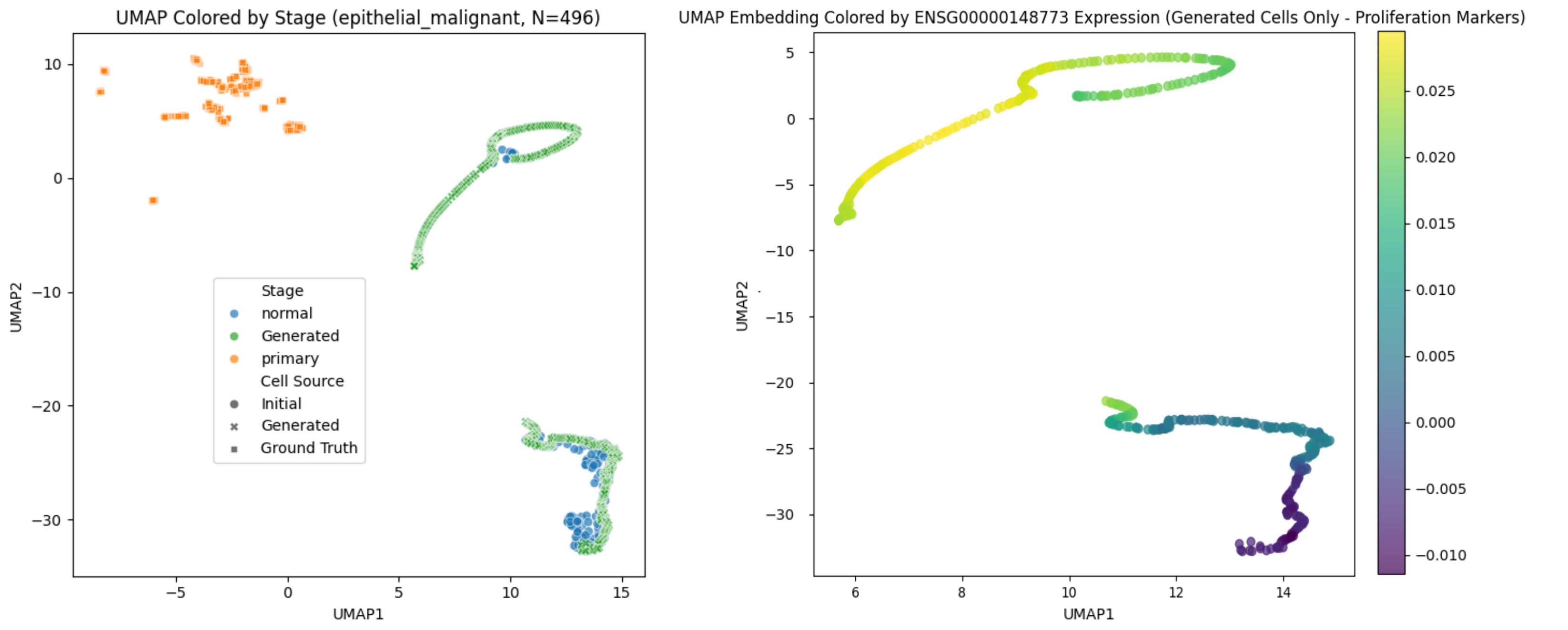
This model integrates the continuous gene expression data of each cell to create a comprehensive representation of a cell token.

Results

- Given a variable number of normal or/and primary cells, evoCancerGPT1-44M can generate cancer progression. The model performs better with longer input sequence, yet sometimes "hallucinates" intermediate cell states.



- Although evoCancerGPT1-44M is not yet able to reconstruct randomly sampled primary cells given only normal cells as input, it can reconstruct pseudotime analysis along key proliferation marker: ex. MKI67/KI-67/ENSG00000148773.



Discussions

- evoCancerGPT is powerful in capturing cell state dynamics during tumor progression via pseudotime.
- However, while randomized sampling for sentence constructions help contribute to sample size and reduce sentence length, it can also bias the trained model or input samples.
- On-going work: evoCancerGPT2
  - Dramatically increase the number of sentences for training.
  - Include metastatic datasets.
  - Expand to pan-cancer analysis.
  - Incorporate meta info for each cell.
  - Improve sampling strategies.

References:

1. Reed, A.D., Pensa, S., Steif, A. *et al.* A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. *Nat Genet* **56**, 652–662 (2024). https://doi.org/10.1038/s41588-024-01688-9
2. CZI Single-Celll Biology Program, Abdulla, S., Aevermann, B. *et al*. CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *bioRxiv*, (2023). https://doi.org/10.1101/2023.10.30.563174
3. Wu, S.Z., Al-Eryani, G., Roden, D.L. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* **53**, 1334–1347 (2021). https://doi.org/10.1038/s41588-021-00911-1