Cross-modal transfer learning for mapping bulk transcriptomes at cellular level

Aarthi Venkat [1 2] **Zhiwen Jiang** [* 3] **Daniel Marbach** [* 3] **Nir Hacohen** [# 1 4 5] **Marinka Zitnik** [# 2 6 1 7]

Abstract

Bulk transcriptomics is widely used in clinical research, yet existing methods struggle to extract single-cell-level structure from bulk measurements, which aggregate signals across heterogeneous cell populations. Here we introduce POPPY, a framework which uses ontology-based contrastive learning to align single-cell and bulk transcriptomic foundation models and construct cell-type-aware bulk patient embeddings. Trained on 1,458 single-cell tumor samples from the Curated Cancer Cell Atlas (3CA) and 1,286 bulk profiles derived from sorted cell populations, POPPY recovers cell type and gene program structure from bulk tumor transcriptomes without fine-tuning. Using a single-cell melanoma atlas, POPPY predicts immunotherapy response across six bulk melanoma cohorts and identifies individual cells associated with response in bulk tumors. These results demonstrate single-cell foundation models can be leveraged to build bulk embeddings that capture cellular biology, enabling interpretable patient stratification from bulk data.

1. Introduction

Despite their high resolution and potential to yield clinically actionable biomarkers, single-cell technologies are not typically used for clinical decision-making due to financial, conceptual, and computational constraints (Skinnider et al., 2025). Instead, bulk transcriptomic sequencing, which measures the average expression across entire samples rather than individual cells, is a more routine part of patient care and clinical research. Approaches that derive patient endotypes and predict phenotypes from bulk data alone, however, often exhibit poor generalizability across cohorts due to the reliance on predefined biomarker sets (Shen et al., 2025).

Foundation models for single-cell data, e.g. (Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024; Pearce et al., 2026), learn flexible representations of cellular states, but are not designed for bulk transcriptomic inputs. Conversely, foundation models trained on bulk data (Kang et al., 2025; Shen et al., 2025) do not incorporate single-cell-level data or supervision. An approach to construct a shared representation space bridging these modalities would uniquely enable the generalization of single-cell representations to the bulk setting, unlocking their utility in clinical workflows.

We present POPPY, a framework for transferring representations from single-cell foundation models to the bulk transcriptomic setting via ontology-aware contrastive learning. POPPY models cell type composition from single-cell, pseudobulk, and sorted bulk data as signals on the Cell Ontology graph (Diehl et al., 2016), then aligns cross-modal pairs to preserve pairwise similarities defined by optimal transport between signals. We further propose a sampling strategy to encourage patient alignment across cohorts. Trained with 1,458 single-cell tumor samples from the 3CA (Gavish et al., 2023; Tyler et al., 2025) and 1,286 bulk profiles constructed from sorted cell populations (Zaitsev et al., 2022), POPPY substantially improves cohort and assay integration, cell type proportion preservation, and gene program preservation over bulk-only encoders. Moreover, an optional cross-attention module enables fine-tuning for patient phenotype prediction by querying a single-cell reference atlas with the learned bulk patient embedding, producing per-cell attribution scores reflecting relevance to the prediction. In a case study of 330 bulk melanoma tumors across six cohorts, POPPY improved immunotherapy response prediction over the bulk-only encoder and the fine-tuned model without contrastive training, identifying T cells, NK cells, and memory B cells as associated with responder profiles, as shown in prior literature (Mellman et al., 2023). Together, these results support using POPPY to extend single-cell foundation models to cellular-level mapping of bulk transcriptomes.

*Equal contribution. #Co-senior authorship. [1]Broad Institute of MIT and Harvard, Cambridge, MA, USA [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA [3]Roche Pharma Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland [4]Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA [5]Krantz Family Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA [6]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Allston, MA, USA [7]Harvard Data Science Initiative, Cambridge, MA, USA. Correspondence to: Marinka Zitnik <marinka@hms.harvard.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

2. Methods

We outline the POPPY framework by detailing how we compute patient-patient similarity from cell type composition, then learn cross-modal representations (Figure 1).

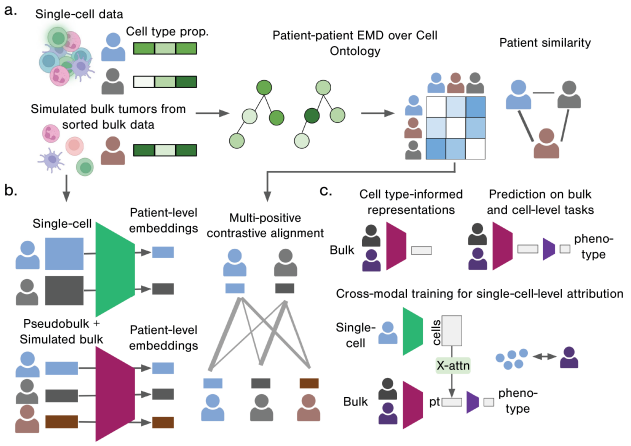

Figure 1. (a) From large-scale single-cell RNA-seq data and simulated bulk RNA-seq tumors constructed from sorted bulk data, we derive cell type composition vectors and represent them as patient-level signals on the Cell Ontology graph. Patient-patient similarity is then computed using graph signal optimal transport. (b) POPPY uses these similarity scores as soft supervision for contrastive alignment across modalities. (c) This framework enables downstream analysis of bulk RNA-seq data, including construction of cell type-informed representations, bulk and cell-level phenotype prediction, and single-cell-level attribution via cross-modal fine-tuning. Abbreviations: prop., proportion.

2.1. Patient similarity from Cell Ontology transport

We first define a patient-level similarity metric grounded in cell type composition to serve as supervision for bulk representation learning (Figure 1a). As high-quality cell type composition labels are limited for bulk samples, we construct surrogate bulk profiles in two ways: pseudobulk profiles, formed by aggregating counts across all cells per patient from single-cell datasets, and *in silico* bulk profiles derived by taking weighted averages of sorted populations.

We then define patient-level similarity based on cell types and gene programs derived from dataset annotations. These proportion vectors are represented as *graph signals* on the Cell Ontology (Section A.1), where graph signals are functions defined on the nodes of the graph. The distance between two patients is then computed as the Earth Mover's Distance (EMD), or optimal transport, between two signals over the graph. This endows the metric with biological structure and robustness to annotation differences: two samples whose compositions differ only in closely related cell types receive a lower distance than samples that differ across unrelated lineages.

Let $\mathbf{p}_i, \mathbf{p}_j \in \Delta^{K-1}$ denote the cell type composition vectors of samples i and j, where K is the number of cell types and Δ^{K-1} is the probability simplex (each vector is ℓ_1-normalized such that $\sum_k \mathbf{p}_{ik} = 1$). The entry $\mathbf{p}_{ik}$ gives the proportion of cells in sample i belonging to type k. Pairwise sample similarity is measured via the Earth Mover's Distance (EMD) and solved exactly (Bonneel et al., 2011):

$$\mathrm{EMD}(i,j) = \mathrm{EMD}(j,i) = \min_{\mathbf{T} \in \Pi(\mathbf{p}_i, \mathbf{p}_j)} \sum_{k,l} \mathbf{T}_{kl}\, d_{kl}, \qquad \Pi(\mathbf{p}_i, \mathbf{p}_j) = \left\{ \mathbf{T} \geq 0 : \mathbf{T}\mathbf{1} = \mathbf{p}_i,\ \mathbf{T}^\top \mathbf{1} = \mathbf{p}_j \right\}. \tag{1}$$

Π is the set of joint transport plans and $\mathbf{D} \in \mathbb{R}^{K \times K}_{\geq 0}$ is a ground metric with d_{kl} giving the cost of transporting mass from cell type k to cell type l. $\mathbf{D}$ is derived from the Cell Ontology, where the distance d_{kl} reflects the shortest path between types k and l in the ontology graph so that biologically proximate types incur lower transport cost than distantly related types. These pairwise values populate the similarity matrix $\mathbf{S}$ after conversion of distances to similarities with a graph diffusion kernel (Section A.2). $\mathbf{S}$ is used as a soft supervision target in the objective described below.

2.2. Overview of POPPY model

Given $\mathbf{S}$, we train cell-level encoder E_c and bulk-level encoder E_b to map single-cell and bulk profiles into a shared embedding space via relationships defined in $\mathbf{S}$ (Figure 1b). Notably, both encoders can be applied independently; bulk analysis does not require paired single-cell data.

2.2.1. CELL-LEVEL ENCODER

The single-cell encoder E_c maps a variable-size set of cells to a fixed-size embedding $\mathbf{z}_c \in \mathbb{R}^d$. Formally, for a given patient sample, let $\{\mathbf{x}_c^{(i)}\}_{i=1}^{N_c}$ denote the set of N_c cell vectors, each of dimension m_c (the number of measured genes), where we suppress the patient index for clarity. E_c first projects each cell into a unified latent space $\mathbb{R}^{d_c}$ using any pretrained single-cell model, then applies a learned scoring function $f : \mathbb{R}^{d_c} \to \mathbb{R}$ that assigns a scalar attention logit to each cell. The pooled representation is the attention-weighted sum, following (Ilse et al., 2018):

$$\mathbf{h}_c = \sum_{i=1}^{N_c} \alpha_i \mathbf{x}_c^{(i)}, \qquad \alpha_i = \frac{\exp(f(\mathbf{x}_c^{(i)}))}{\sum_{j=1}^{N_c} \exp(f(\mathbf{x}_c^{(j)}))}, \tag{2}$$

where $f(\mathbf{x}_c^{(i)}) = \mathbf{w}^\top \mathbf{x}_c^{(i)} + b$. The pooled vector $\mathbf{h}_c \in \mathbb{R}^{d_c}$ is passed through a two-layer FFN and ℓ_2-normalized, resulting in a final patient embedding $\mathbf{z}_c \in \mathbb{R}^d$.

2.2.2. BULK-LEVEL ENCODER

For each patient sample, the bulk encoder E_b maps a bulk expression measurement to a fixed-size patient embedding $\mathbf{z}_b \in \mathbb{R}^d$. Given a bulk measurement $\mathbf{x}_b \in \mathbb{R}^{m_b}$ of m_b

genes, a pretrained transcriptomic encoder first embeds each gene into a shared d_b-dimensional token space, producing matrix $\mathbf{H} \in \mathbb{R}^{m_b \times d_b}$. The token matrix is then pooled across genes using the attention-weighted sum of Equation 2, yielding pooled vector $\mathbf{h}_b \in \mathbb{R}^{d_b}$. Finally, $\mathbf{h}_b$ is passed through a two-layer FFN and ℓ_2-normalized, resulting in a final patient-level representation $\mathbf{z}_b \in \mathbb{R}^d$.

2.2.3. Multi-positive contrastive alignment

Let $\mathbf{Z}_c, \mathbf{Z}_b \in \mathbb{R}^{Nxd}$ denote row-stacked embeddings of N cell and bulk encoder outputs in a given batch, respectively. We form pairwise logit matrices scaled by temperature τ_{cl}:

$$\mathbf{L}^{\times} = \mathbf{Z}^c(\mathbf{Z}^b)^{\top}/\tau_{cl}, \mathbf{L}^c = \mathbf{Z}^c(\mathbf{Z}^c)^{\top}/\tau_{cl}, \mathbf{L}^b = \mathbf{Z}^b(\mathbf{Z}^b)^{\top}/\tau_{cl}, \tag{3}$$

where $\mathbf{L}^c$ and $\mathbf{L}^b$ have their diagonals masked to $-\infty$ to exclude self-similarity. Per-direction logit matrices are formed by concatenating the cross-modal and intra-modal blocks:

$$\tilde{\mathbf{L}}^{c\to b} = \left[\mathbf{L}^{\times} \mid \mathbf{L}^c\right], \qquad \tilde{\mathbf{L}}^{b\to c} = \left[(\mathbf{L}^{\times})^{\top} \mid \mathbf{L}^b\right]. \tag{4}$$

The label distribution $\mathbf{S} \in [0,1]^{N\times N}$ is the cell type-proportion-derived similarity matrix defined in Section 2.1. Each row is normalized to a valid probability distribution $\hat{\mathbf{S}}$, where $\hat{\mathbf{S}}_{ij} = \frac{\mathbf{S}_{ij}}{\sum_k \mathbf{S}_{ik}}$. A zero block of equal width is appended, i.e. $\tilde{\mathbf{S}} = \left[\hat{\mathbf{S}} \mid \mathbf{0}_{N\times N}\right]$, encouraging the model to align representations across modalities.

The loss in each direction is the KL divergence between the soft label distribution and the log-softmax predictive distribution. The final objective is the average:

$$\begin{aligned} \mathcal{L}^{c\to b} &= \mathrm{KL}\Big(\tilde{\mathbf{S}} \,\Big\|\, \mathrm{softmax}\Big(\tilde{\mathbf{L}}^{c\to b}\Big)\Big), \\ \mathcal{L}^{b\to c} &= \mathrm{KL}\Big(\tilde{\mathbf{S}} \,\Big\|\, \mathrm{softmax}\Big(\tilde{\mathbf{L}}^{b\to c}\Big)\Big) \\ \mathcal{L} &= \tfrac{1}{2}\big(\mathcal{L}^{c\to b} + \mathcal{L}^{b\to c}\big). \end{aligned} \tag{5}$$

2.2.4. Cross-cohort positive sampler

To encourage cross-cohort and cross-assay generalization, we introduce a structured sampling strategy. Positive candidates are restricted to samples from a different cohort with non-zero similarity under $\mathbf{S}$. For each anchor i, a cross-cohort positive p is sampled uniformly from the positive candidate set. Then, other samples are drawn without replacement from those strictly less similar to i than p, with preferential sampling to same-cohort samples.

2.2.5. Fine-tuning module

The POPPY multimodal framework can be utilized here to allow bulk samples to attend to different regions of the cellular state space for making patient-level predictions. In the setting where we have phenotype labels associated with each patient, POPPY is trained with an additional module in which each bulk patient sample attends over a fixed reference set of cell embeddings $\{\mathbf{x}_c^{(i)}\}_{i=1}^N$, constructed by encoding cells from a new tissue or disease-relevant atlas, or from the pretrained cell embedding space. Suppressing patient index (s) for clarity, a query is formed from the bulk embedding $\mathbf{q} = W_Q\mathbf{z}_b$, and a key from each atlas cell embedding $\mathbf{k}^{(i)} = W_K\mathbf{x}_c^{(i)}$. Attention weights are computed via scaled dot-product attention: $\alpha_i = \frac{\exp(\mathbf{q}^{\top}\mathbf{k}^{(i)}/\tau_{ft})}{\sum_j \exp(\mathbf{q}^{\top}\mathbf{k}^{(j)}/\tau_{ft})}$. Each cell additionally receives a scalar value score $v_i = W_V\mathbf{x}_c^{(i)}$. The final patient phenotype score is the attention-weighted sum of cell scores $\hat{y} = \sum_i \alpha_i v_i + b$, where b is a learned bias. Per-cell attention α_i is retained for post-hoc interpretation to decompose patient prediction into single cell contributions.

POPPY can be flexibly optimized for patient-level tasks. We evaluate POPPY in a two-class setting and use a binary cross-entropy loss $\mathcal{L} = -\frac{1}{S}\sum_s [y^{(s)} \log \sigma(\hat{y}^{(s)}) + (1 - y^{(s)}) \log(1 - \sigma(\hat{y}^{(s)}))]$, where S is the number of training samples, $y^{(s)} \in \{0,1\}$ is the true label, and $\hat{y}^{(s)}$ is the predicted logit for sample s.

3. Results

3.1. Datasets

We train POPPY on two sources of single-cell and bulk profiles. From the Curated Cancer Cell Atlas (3CA), we retain 2,732,789 cells from 1,631 patient samples across 85 cohorts and 15 cancer types, reserving 6 cohorts for evaluation. We additionally construct 2,286 simulated bulk profiles from sorted cell populations drawn from the Kassandra database (full details in Section A.3).

3.2. Pre-trained model details

POPPY uses frozen pretrained encoders for both modalities. For single-cell profiles, we use TranscriptFormer-Sapiens; for bulk profiles, we evaluate two initializations: BulkFormer, pretrained on bulk data (POPPY-BULKFORMER) and TranscriptFormer-Sapiens, pretrained on single-cell data (POPPY-TRANSCRIPTFORMER). We then compare POPPY to the respective base encoder to assess transfer in both directions: a bulk-pretrained model applied to single-cell-level tasks, and a single-cell-pretrained model applied to bulk transcriptomics.

3.3. Cross-modal alignment across cohorts

To evaluate the fidelity of cross-modal alignment, we constructed paired single-cell and pseudobulk profiles using the six cohorts held-out from 3CA and visualized both modalities (Becht et al., 2018) (Figure 2a). POPPY performs zero-shot alignment of single-cell and bulk representations per

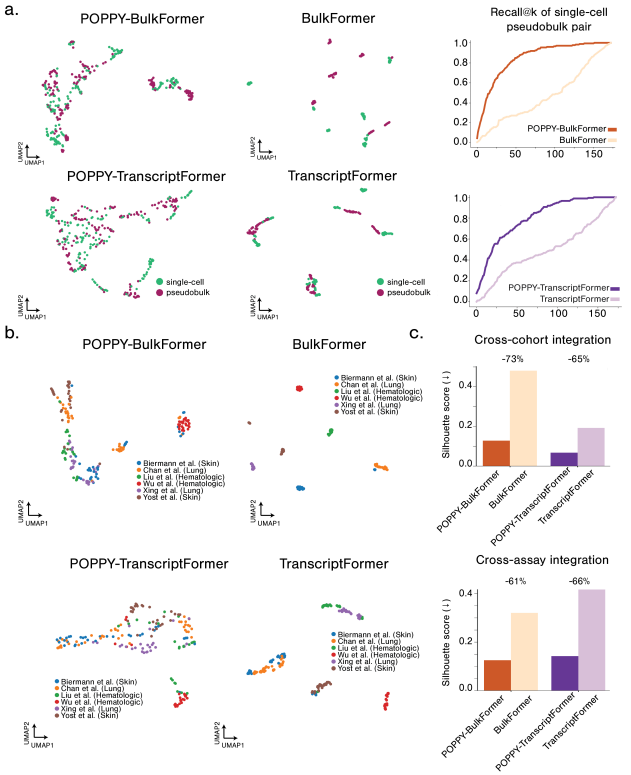

Figure 2. (a) POPPY improves fidelity of cross-modal alignment and (b) integration across cohorts and technologies over corresponding baseline bulk encoder.

sample, with improved recall@k for matched single-cell-pseudobulk pairs compared to without our alignment. We next assessed the utility of the embeddings from the bulk encoder alone. We demonstrate that, compared to baseline encoders, POPPY better integrates patient representations across cohorts and assays (Figure 2b, c), with up to 73% improvement in silhouette score.

3.4. Cell type proportion preservation

We next evaluated our pseudobulk sample embeddings for organization by cell type proportion, where we have ground truth from the corresponding single-cell dataset. Visualizing T cell proportion over the bulk embedding latent space clearly demonstrates our embeddings organize by cell type proportion better than the baseline embeddings (Figure 3a). To quantify this across all cell types, we compute the pairwise patient embedding distance and the ground truth pairwise patient cell type-based distance, then present the Spearman correlation between these distances (Figure 3a). These results show performance gains within cohorts for both cell type proportion and gene program proportion correlation (up to 117% increase); cross-cohort correlation improves even more dramatically (up to 714%), highlighting POPPY's ability to capture cross-cohort patient composition similarity.

We verify this result in true bulk transcriptomic data using 1,000 simulated tumors constructed by sampling sorted populations from a held out set, with cell type ratios drawn uniformly. Computing the correlation between true and embedding-based patient-patient distances demonstrates up to 120% increase compared to baseline (Figure 3b).

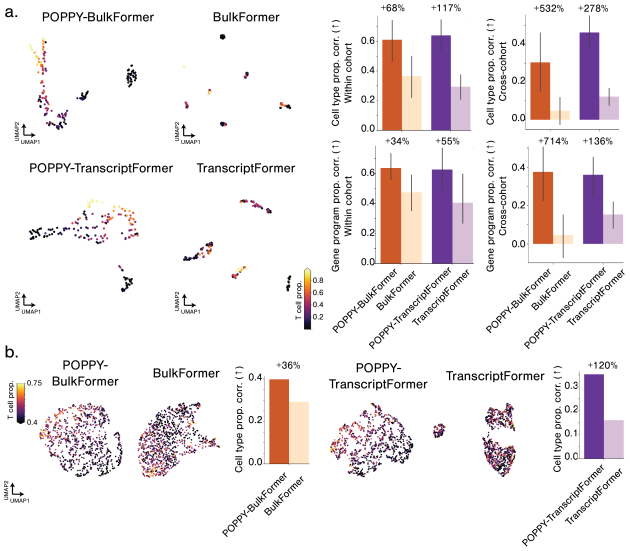

Figure 3. (a) POPPY preserves cell type and gene program proportion over baselines in held-out 3CA pseudobulk datasets, with high patient-patient similarity within and across-cohorts. (b) POPPY preserves cell type proportion in simulated tumors from held out sorted bulk profiles. Abbreviations: corr., Spearman correlation.

3.5. Single-cell-guided bulk phenotype prediction

We hypothesized POPPY could be fine-tuned to predict patient-level phenotypes and, through the addition of the cross-attention module, distinguish cells that are differentially attended to for the prediction. We evaluated POPPY on 330 pre-treatment bulk tumor samples from six melanoma cohorts (Van Allen et al., 2015; Gide et al., 2019; Hugo et al., 2016; Freeman et al., 2022; Riaz et al., 2017; Liu et al., 2019), where each bulk sample was derived from a patient responsive (R) or non-responsive (NR) to the administered immunotherapy. For a reference single-cell atlas, we used 120 melanoma tumors from 3CA, where 62 samples were from cohorts held out of the initial training stage. POPPY cell embeddings organize much more consistently by cell type than the TranscriptFormer embeddings for the same datasets (Figure 4a). We freeze the cell encoder and fine-tune a cross-attention module and bulk classifier head, allowing bulk embeddings to attend over cell embeddings for response prediction. Using a leave-one-cohort-out split, POPPY outperforms both baselines: BulkFormer with a classification head, and the cross-attention setup without prior cross-modal alignment (POPPY w/o stage 1) (Figure 4b).

Of the 120 single-cell patient samples, 45 had labels

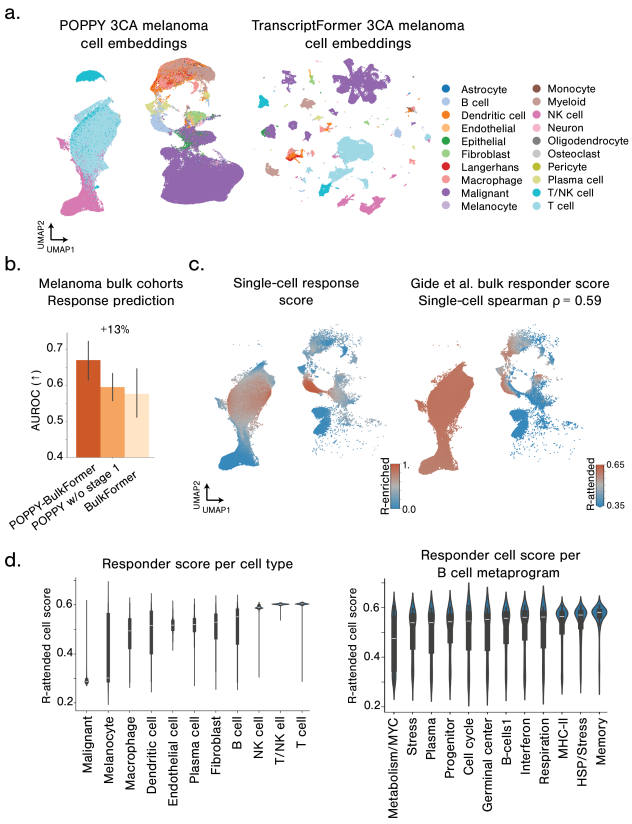

Figure 4. (a) 3CA melanoma single-cell embeddings show better cross-dataset cell type integration with POPPY than TranscriptFormer. (b) Leave-one-cohort-out classification of responders versus non-responders across six bulk melanoma cohorts shows improvement over baseline methods. (c) POPPY responder scores align with ground-truth response labels at single-cell resolution in Gide et al. (d) Responder score visualized at the level of individual cell types and gene programs.

corresponding to response status. We visualized cells from this subset and calculated an enrichment score for each individual cell (Burkhardt et al., 2021), where a cell with score >0.5 is enriched in responders, and a cell with score <0.5 is enriched in non-responders (Figure 4c). Then, for bulk samples from (Gide et al., 2019), we retrieved the per-cell attention scores and calculated a responder enrichment score. This correlated to the original single-cell dataset ($\rho = 0.59$), suggesting attention weights may reflect true biological associations. Aggregating per cell type, T cells and NK cells are most strongly associated with response, whereas malignant cells are associated with non-response, following literature on immune infiltration in the tumor microenvironment (Mellman et al., 2023). Within B cells, activated memory B cells, characterized by memory markers *CD27* and *CXCR3* and activation markers *CD86* and *IFI30*, showed strongest responder association, consistent with their role in anti-tumor immunity (Kim et al., 2021).

4. Conclusion

Recent advances in single-cell foundation models have enabled unique insights into cellular biology, but have shown limited progress in transfer learning across modalities (Xie et al., 2025). POPPY bridges this gap by constructing bulk transcriptomic representations informed by cell type composition. Through the cross-attention module, bulk patient embeddings can be decomposed as weighted combinations of single-cell reference profiles, enabling attribution of bulk phenotypes to individual cells. Together, these results suggest the structure captured by biological foundation models can be transferred to lower-resolution transcriptomic data, broadening their applicability across experimental contexts. Future work will demonstrate generalizability of the framework in diverse tissue, cell type, and disease contexts at larger sample scales.

Acknowledgements

We thank the reviewers for their insights and contributions to this work, and Y. Huang. B. Liu, and X. Lin for their helpful feedback. We gratefully acknowledge support from the NSF CAREER Award 2339524, ARPA-H Biomedical Data Fabric (BDF) Toolbox Program, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, GlaxoSmithKline Award, Roche Alliance with Distinguished Scientists (ROADS) Program, Sanofi iDEA-iTECH Award, Boehringer Ingelheim Award, Merck Award, Optum AI Research Collaboration Award, Pfizer Research, Gates Foundation (INV-079038), Chan Zuckerberg Initiative, Collaborative Center for XDP at Massachusetts General Hospital, John and Virginia Kaneb Fellowship at Harvard Medical School, Biswas Computational Biology Initiative in partnership with the Milken Institute, Harvard Medical School Dean's Innovation Fund for the Use of Artificial Intelligence, and the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. A.V. acknowledges support from the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard. N.H. was supported by the Mark Foundation Endeavor Award and the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation (AMRF).

Impact Statement

This paper presents work whose goal is to advance the field of machine learning with applications to computational biology and medicine. All datasets used in this work are publicly available and de-identified. POPPY is designed as a predictive and exploratory tool and careful usage is required for downstream biomedical research and application.

Conflict of Interest

D.M. and Z.J. are currently employed by F. Hoffmann-La Roche Ltd.

References

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37(1):38–44, December 2018. ISSN 1546-1696. doi: 10.1038/nbt.4314. URL http://dx.doi.org/10.1038/nbt.4314.

Bonneel, N., van de Panne, M., Paris, S., and Heidrich, W. Displacement interpolation using lagrangian mass transport. *ACM Trans. Graph.*, 30(6):1–12, December 2011.

Burkhardt, D. B., Stanley, J. S., Tong, A., Perdigoto, A. L., Gigante, S. A., Herold, K. C., Wolf, G., Giraldez, A. J., van Dijk, D., and Krishnaswamy, S. Quantifying the effect of experimental perturbations at single-cell resolution. *Nature Biotechnology*, 39 (5):619–629, February 2021. ISSN 1546-1696. doi: 10.1038/s41587-020-00803-5. URL http://dx.doi.org/10.1038/s41587-020-00803-5.

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. ISSN 1063-5203. doi: 10.1016/j.acha.2006.04.006. URL http://dx.doi.org/10.1016/j.acha.2006.04.006.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Diehl, A. D., Meehan, T. F., Bradford, Y. M., Brush, M. H., Dahdul, W. M., Dougall, D. S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., et al. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of biomedical semantics*, 7(1): 44, 2016.

Freeman, S. S., Sade-Feldman, M., Kim, J., Stewart, C., Gonye, A. L., Ravi, A., Arniella, M. B., Gushterova, I., LaSalle, T. J., Blaum, E. M., et al. Combined tumor and immune signals from genomes or transcriptomes predict outcomes of checkpoint inhibition in melanoma. *Cell Reports Medicine*, 3(2), 2022.

Gavish, A., Tyler, M., Greenwald, A. C., Hoefflin, R., Simkin, D., Tschernichovsky, R., Galili Darnell, N., Somech, E., Barbolin, C., Antman, T., et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature*, 618(7965):598–606, 2023.

Gide, T. N., Quek, C., Menzies, A. M., Tasker, A. T., Shang, P., Holst, J., Madore, J., Lim, S. Y., Velickovic, R., Wongchenko, M., et al. Distinct immune cell populations define response to anti-pd-1 monotherapy and anti-pd-1/anti-ctla-4 combined therapy. *Cancer cell*, 35 (2):238–255, 2019.

Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., and Song, L. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024.

Heimberg, G., Kuo, T., DePianto, D. J., Salem, O., Heigl, T., Diamant, N., Scalia, G., Biancalani, T., Turley, S. J., Rock, J. R., et al. A cell atlas foundation model for scalable search of similar human cells. *Nature*, 638(8052): 1085–1094, 2025.

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., et al. Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma. *Cell*, 165(1):35–44, 2016.

Ilse, M., Tomczak, J., and Welling, M. Attention-based deep multiple instance learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Kang, B., Fan, R., Yi, M., Cui, C., and Cui, Q. A large-scale foundation model for bulk transcriptomes. *BioRxiv*, pp. 2025–06, 2025.

Kim, S. S., Sumner, W. A., Miyauchi, S., Cohen, E. E., Califano, J. A., and Sharabi, A. B. Role of b cells in responses to checkpoint blockade immunotherapy and overall survival of cancer patients. *Clinical Cancer Research*, 27(22):6075–6082, 2021. ISSN 1557-3265. doi: 10.1158/1078-0432.ccr-21-0697. URL http://dx.doi.org/10.1158/1078-0432.CCR-21-0697.

Liu, D., Schilling, B., Liu, D., Sucker, A., Livingstone, E., Jerby-Arnon, L., Zimmer, L., Gutzmer, R., Satzger, I., Loquai, C., Grabbe, S., Vokes, N., Margolis, C. A., Conway, J., He, M. X., Elmarakeby, H., Dietlein, F., Miao, D., Tracy, A., Gogas, H., Goldinger, S. M., Utikal, J., Blank, C. U., Rauschenberg, R., von Bubnoff, D., Krackhardt, A., Weide, B., Haferkamp, S., Kiecker, F., Izar, B., Garraway, L., Regev, A., Flaherty, K., Paschen, A., Van Allen, E. M., and Schadendorf, D. Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nat. Med.*, 25(12):1916–1927, December 2019.

Mellman, I., Chen, D. S., Powles, T., and Turley, S. J. The cancer-immunity cycle: Indication, genotype, and immunotype. *Immunity*, 56(10):2188–2205, 2023.

Pearce, J. D., Simmonds, S. E., Mahmoudabadi, G., Krishnan, L., Palla, G., Istrate, A.-M., Tarashansky, A., Nelson, B., Valenzuela, O., Li, D., Quake, S. R., and Karaletsos, T. Transcriptformer: A generative cell atlas across 1.5 billion years of evolution. *Science*, May 2026. ISSN 1095-9203. doi: 10.1126/science.aec8514. URL http://dx.doi.org/10.1126/science.aec8514.

Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., Hodi, F. S., Martín-Algarra, S., Mandal, R., Sharfman, W. H., et al. Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell*, 171(4):934–949, 2017.

Shen, W., Nguyen, T. H., Li, M. M., Huang, Y., Moon, I., Nair, N., Marbach, D., and Zitnik, M. Generalizable ai predicts immunotherapy outcomes across cancers and treatments. *medRxiv*, 2025.

Skinnider, M. A., Courtine, G., Bloch, J., and Squair, J. W. A clinical road map for single-cell omics. *Cell*, 188(14): 3633–3647, July 2025.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Tyler, M., Gavish, A., Barbolin, C., Tschernichovsky, R., Hoefflin, R., Mints, M., Puram, S. V., and Tirosh, I. The curated cancer cell atlas provides a comprehensive characterization of tumors at single-cell resolution. *Nature Cancer*, 6(6):1088–1101, 2025.

Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Foppen, M. H. G., Goldinger, S. M., Utikal, J., Hassel, J. C., Weide, B., Kaehler, K. C., Loquai, C., Mohr, P., Gutzmer, R., Dummer, R., Gabriel, S., Wu, C. J., Schadendorf, D., and Garraway, L. A. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*, 350 (6257):207–211, October 2015.

Wang, S., Pisco, A. O., McGeever, A., Brbic, M., Zitnik, M., Darmanis, S., Leskovec, J., Karkanias, J., and Altman, R. B. Unifying single-cell annotations based on the cell ontology. *BioRxiv*, pp. 810234, 2019.

Xie, F., Zhao, B., Xu, S., Wang, Z., Moon, J. J., and Garmire, L. X. Overcoming barriers to the wide adoption of single-cell large language models in biomedical research. *Nature Biotechnology*, pp. 1–5, 2025.

Zaitsev, A., Chelushkin, M., Dyikanov, D., Cheremushkin, I., Shpak, B., Nomie, K., Zyrin, V., Nuzhdina, E., Lozinsky, Y., Zotova, A., et al. Precise reconstruction of the tme using bulk rna-seq and a machine learning algorithm trained on artificial transcriptomes. *Cancer cell*, 40(8): 879–894, 2022.

A. Appendix

A.1. Cell Ontology

The Cell Ontology (CL) is an ontology designed to offer a controlled vocabulary for cell types, proposed as a basis for consistently annotating large-scale single-cell atlases (Diehl et al., 2016). As the CL contains valuable information about the hierarchical relationship between cell types, it has been leveraged in machine learning models for classification of unseen cell types (Wang et al., 2019) and supervised metric learning (Heimberg et al., 2025).

Here, we first identify all cell types annotated in 3CA samples, then add 3CA-annotated metaprograms, or gene programs defined within cell types, as child nodes to the corresponding cell types. We additionally map cell types associated with sorted bulk populations to the ontology. For both sets of cell types, we retrieve the set of all ancestors where the root is the "cell" node (CL:0000000), and we prune the cell ontology to these nodes, resulting in 248 cell types total.

For each single-cell dataset, we retrieve the cell type annotations and, if gene programs are calculated, assign each cell to a gene program if its program loading is positive and exceeds 75th percentile across all cells. We then compute the proportion of cells belonging to each gene program and cell type and propagate these proportions hierarchically (such that the root node has a proportion of 1.0).

A.2. Similarity calculation from EMD

Raw pairwise distances d_{ij} computed from the cell-type EMD are not directly suitable as contrastive supervision targets: they are unbounded, may not reflect the manifold structure of the data, and assign non-zero similarity to all pairs regardless of their true relationship. We convert distances to similarities using a graph diffusion kernel which respects local neighborhood geometry and produces a sparse, globally consistent affinity matrix.

A weighted k-nearest-neighbor graph $G = (\mathcal{V}, \mathcal{E})$ is constructed over the training samples using the precomputed EMD distance matrix $\mathbf{D}$. Each node i is connected to its $k = 100$ nearest neighbors under $\mathbf{D}$, where restricting connections to the k-NN neighborhood enforces sparsity and ensures that similarity is propagated only through locally consistent paths. Edge weights are computed via an adaptive-bandwidth Gaussian kernel applied to the k-NN adjacency. For connected nodes i and j the kernel entry takes the form:

$$\mathbf{K}_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma_i \sigma_j}\right), \tag{6}$$

where σ_i is a locally adaptive bandwidth set to the distance from node i to its k-th nearest neighbor (Coifman & Lafon, 2006), normalizing for non-uniform sampling density. The kernel matrix $\mathbf{K}$ constructed on training samples is used directly as the similarity matrix $\mathbf{S}$ introduced in Section 2.2.3. Its entries are non-negative, concentrated on local neighborhoods (zero or near-zero for non-neighbors), and lie in $[0, 1]$, satisfying the requirements of the soft label normalization procedure.

A.3. Training datasets

A.3.1. Curated Cancer Cell Atlas (3CA)

To train encoders for bulk tumor transcriptome analysis, we leveraged the Curated Cancer Cell Atlas (3CA), curated across a large collection of scRNA-seq cancer datasets (Tyler et al., 2025; Gavish et al., 2023). We retain samples where raw scRNA-seq is available and prune cells to those that have cell type annotations and at least 1000 gene counts. Then, all samples with less than 10 cells are removed. After preprocessing, this preserves 2,732,789 cells from 1,631 patient samples across 85 cohorts, 16 single-cell technologies, and 15 high-level cancer types. Reserving 6 cohorts for evaluation (comprising 380,967 cells, 173 patient samples, and 3 cancer types), we construct a single-cell and pseudobulk dataset for each sample to train POPPY.

A.3.2. Sorted bulk populations

We additionally construct tumor profiles from bulk RNA-seq datasets in the Kassandra database, comprised of 9,056 bulk RNA-seq samples from 505 datasets of sorted cells, cancer cells and cell lines (Zaitsev et al., 2022). As in the original work, we resample datasets within each cell type to improve dataset variability and, for each simulated tumor, aggregate the expression of one cancer cell line and the average expression of nine non-cancer cell lines. Using 7,245 samples (and reserving the remaining for the test set), we construct simulated bulk tumors from a weighted average of different sorted

populations using the cancer model-based annotations from the original work. To learn a common manifold between the single-cell, pseudobulk, and bulk data, we construct simulated bulk tumors that are similar to the proportions of single-cell datasets in 3CA. For each training single-cell dataset with overlapping cell types with the sorted data, we construct a simulated bulk tumor with the same proportions for shared cell types, resulting in 1,286 *in silico* bulk profiles derived from sorted populations.