ImmuneNet: Composition-Aware Quantification of Adaptive Lymphocytes in High-Grade Serous Ovarian Cancer

Jaelyn S. Liang
Moores Cancer Center
University of California, San Diego
San Diego, CA 92093
jsl047@health.ucsd.edu

Abstract

Bulk RNA-seq mixes signals from tumor, stromal, and immune cells, which obscures the adaptive lymphocyte readouts needed for prognosis and treatment design. In high-grade serous ovarian cancer, recovering T and B cell fractions is particularly difficult due to marker leakage from malignant programs, cohort and platform mismatch between references and targets, and miscalibrated predictions. We present ImmuneNet, a context-aware estimator of T and B fractions from bulk RNA-seq. The pipeline builds a tumor-matched single-cell reference, curates a compact marker panel by removing genes with measured malignant leakage, and trains on pseudo-bulk mixtures that cover the composition simplex. The model uses branched, gene-wise attention to aggregate lineage evidence and a parallel ridge-regularized linear component on the same standardized panel; a fixed blend and post hoc calibration (temperature scaling and isotonic regression) yields well-calibrated fraction estimates. Evaluation emphasizes composition-aware criteria alongside standard correlations. On held-out mixtures, ImmuneNet attains Spearman 0.967 (T) and 0.991 (B), RMSE 0.049 (T) and 0.032 (B), and the lowest total variation distance 0.047 over (T,B,O), outperforming strong linear and PLS baselines on joint composition error while remaining robust and data efficient. On TCGA-OV, ImmuneNet correlates strongly with tumor-matched T and B module proxies. These design choices produce a practical and reliable deconvolution approach tailored to ovarian cancer and suitable for downstream analyses on bulk cohorts.

1 Introduction

Quantifying adaptive lymphocytes from bulk RNA-seq transcriptomes is challenging, and immune composition is prognostic across cancers [1]. In high-grade serous ovarian cancer (HGSOC), tumor-infiltrating $CD8^+$ T cells and B cell rich tertiary lymphoid structures associate with favorable outcomes [2,3], consistent with pan cancer immune subtypes [4]. Previous approaches include signature based linear models, single cell informed references, and deep networks trained on pseudo bulk mixtures [5,6,7,8,9]. In practice, marker leakage, cohort or platform mismatch, oversized marker sets, simplex-agnostic evaluation, and miscalibration remain common [10–12,20], and are pronounced in ovarian cancer [12]. Bulk RNA-seq is broadly available in retrospective cohorts and clinical archives, which enables immediate reuse without new tissue processing. In contrast, other methods, such as cytometry, typically require fresh tissue and site-specific workflows. This practicality motivates bulk-based deconvolution at cohort scale, particularly when harmonized resources such as TCGA-OV are accessible through platforms like Xena [17].

The NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development (AI4D3 2025), San Diego, California, USA, 2025.

We introduce **ImmuneNet**, a context aware model that estimates the fractions of T and B from the bulk RNA sequencing using a compact curated marker panel. The pipeline builds a tumor matched single cell reference [13,14], filters canonical markers by measured leakage in malignant epithelial cells, and trains on pseudo-bulk mixtures that cover the composition simplex. A branched architecture with gene wise attention aggregates lineage evidence, while a parallel ridge regularized linear component on the same standardized panel provides a low variance signal; predictions blend the two with a weight chosen on a calibration split. Post hoc calibration with temperature scaling and isotonic regression improves reliability [15,16]. External bulk data are drawn from TCGA-OV via the Xena platform for downstream analyses [17].

Our design emphasizes four elements: (i) a tumor-matched leakage audit that yields a compact, context safe marker panel [13,14]; (ii) explicitly composition-aware training and reporting with total variation distance and CLR-RMSE along with rank and absolute errors [20]; (iii) a hybrid attention–linear estimator that favors stability and interpretability; and (iv) explicit calibration using temperature scaling and isotonic regression to improve reliability [15,16]. The evaluation includes harder mixtures sampling challenging regions of the composition simplex, robustness checks via marker dropout and single-gene rescaling, and external validation on TCGA-OV. Baselines span signature-based linear, single-cell-derived linear, and latent-variable models.

2 Background

2.1 Deconvolution Methods

Methods fall into three families: signature based linear models that use predefined gene sets and reference profiles [5,6], single cell informed approaches that model cross subject variability from scRNA-seq [7,8], and deep learning trained on pseudo bulk mixtures [9]. Comparative studies document sensitivity to reference choice, platform effects, and evaluation protocols [10,11,12]. Classic ridge type regularization stabilizes linear models under correlated genes [18,19].

2.2 Key challenges and resources

Two recurring issues in solid tumors, pronounced in ovarian cancer, are marker leakage from malignant or stromal programs and cohort or platform shift between reference and target, both of which degrade accuracy and calibration [10,11,12]. Additional concerns include honoring the simplex in evaluation, avoiding overly large marker sets, and producing calibrated probabilities [15,16,20]. Tumor matched single cell atlases support realistic pseudo bulk generation and marker audits [13,14], and harmonized cohorts such as TCGA-OV via UCSC Xena enable external assessment [17].

Evaluation and calibration considerations. To respect the simplex structure and avoid misleading improvements on a single lineage, evaluation includes total variation distance and CLR-RMSE alongside correlations and absolute errors [20]. Because external cohorts seldom provide ground-truth fractions, validation uses lineage proxy scores defined as the first principal component of curated T and B modules derived from tumor-matched single-cell references, and we report both raw and partial correlations that condition on the complementary lineage [13,14,17]. Reliability is summarized with calibration curves and pre- versus post-calibration residuals, and temperature scaling with isotonic regression is applied to align predicted and true compositions [15,16].

3 Methodology

Problem setup. Given a bulk RNA-seq vector over G genes, we estimate the fractions of T and B cells (p_T, p_B) with $p_O = 1 - p_T - p_B$ for all other cells. The model outputs a fraction vector $(\hat{p}_T, \hat{p}_B, \hat{p}_O)$ on the simplex and is trained on synthetic mixtures with known compositions using a log-likelihood loss so the predicted fractions are well calibrated.

Data sources and preprocessing. We construct a tumor-matched single-cell reference for high-grade serous ovarian cancer by merging GSE146026 and GSE217517, intersecting genes, and harmonizing labels into T, B, malignant epithelial, myeloid, fibroblast, and endothelial compart-

ments [13,14]. External bulk data are TCGA-OV from UCSC Xena, processed to CPM or TPM, log transformed, and standardized per gene using statistics from the training split [17].

Context-safe marker curation. Starting from canonical immune markers, we screen each candidate for malignant leakage in the single-cell reference, defined as the fraction of malignant epithelial cells with nonzero expression. Genes with leakage above 5% are removed. This keeps signal specific to immune lineages in the tumor context and limits spillover from epithelial programs. The final panel contains eight genes: T $\{$CD3D, CD3E, CD2, CD8A, CD4$\}$ and B $\{$MS4A1, CD79A, CD79B$\}$. An NK panel is held out as a validation-only covariate.

Pseudo-bulk synthesis. To obtain mixtures with known composition, we sample target fractions over the simplex, draw cells from the single-cell pools accordingly, sum counts within lineages, apply modest library-size jitter and noise, then transform to $\log_2(x{+}1)$. This yields inputs restricted to the curated panel and aligned targets (p_T, p_B, p_O). Mixtures are replicated per composition to reduce sampling variance, with a stratified split into calibration and test compositions.

Model architecture: ImmuneNet. ImmuneNet processes the T and B panels in two branches. Each branch embeds genes, applies lightweight self-attention across genes to aggregate marker evidence, and produces a lineage embedding. The branch embeddings are fused and passed to a small head that outputs (T, B, O) fractions. A parallel linear component with ℓ_2 regularization on the same standardized panel provides a stable low-variance signal [18,19]. A data-driven blend between the neural and linear outputs is chosen on the calibration split to minimize average error on (p_T, p_B). The design emphasizes parsimony, context awareness through the curated panel, and interpretability via attention over genes.

Training and regularization. We train on pseudo-bulk mixtures with an 80/20 train–validation split. Optimization uses AdamW with cosine annealing and early stopping [21,22]. Inputs are standardized using training statistics and the same transform is reused for validation and inference. All model variants share splits, scaling, and seeds.

Calibration. Post hoc calibration aligns predicted fractions with target composition. We consider temperature scaling on logits and per-lineage isotonic regression, fitted on the calibration split and then fixed for the test split [15,16].

4 Experimental Evaluation

Datasets. We synthesize pseudo-bulk mixtures from a tumor-matched HGSOC single-cell reference by merging GSE146026 and GSE217517, intersecting genes, and harmonizing cell labels [13,14]. Compositions (p_T, p_B, p_O) are drawn from a simplex-covering grid and from Dirichlet samples, then replicated to reduce sampling variance. A stratified split separates calibration and test compositions. For external assessment we use TCGA-OV bulk RNA-seq from UCSC Xena, processed to $\log_2(\mathrm{TPM}{+}1)$ or $\log_2(\mathrm{CPM}{+}1)$ with per-gene standardization inherited from training [17]. Unless noted, inputs are restricted to the curated T and B marker panels.

Metrics. We quantify accuracy with rank correlations (Spearman, Pearson) on (p_T, p_B), absolute errors (RMSE, MAE), concordance correlation coefficient for mean and variance agreement [23], and composition-aware criteria on (T, B, O), including total variation distance and CLR-RMSE in Aitchison geometry [20]. For TCGA-OV we compute independent lineage proxies as PC1 of tumor-matched T and B modules and report raw and partial Spearman that conditions on the complementary lineage [13,14,17]. Calibration quality is summarized with reliability curves and pre versus post calibration residuals [15,16].

Baselines and variants. We compare ImmuneNet to matched-input baselines: ridge regression [18,19], elastic net [24], partial least squares (PLS) [25], and a NNLS-signature fit that operates on lineage signatures [26]. Ablations include a neural-only variant and a no-attention variant with the same twin-branch MLPs. Robustness checks perturb inputs via feature masking, additive Gaussian noise, and single-gene scaling at fixed intensities. Baseline families include signature-based linear

models, single-cell informed linear methods, latent-variable approaches, and deep models trained on pseudo-bulk mixtures [5–9]. All baselines share inputs, scaling, splits, and seeds for comparability.

Compute resources. All experiments were executed in a Jupyter notebook on a single CPU on a laptop-class machine with roughly 16 GB RAM. End-to-end runs (pseudo-bulk synthesis, training, baselines, calibration, and plotting) complete within about an hour; ImmuneNet training finishes in minutes and inference is milliseconds per sample with peak memory under 2 GB.

5 Results

Headline performance on known-composition mixtures. In mixtures held out with known (p_T, p_B, p_O), ImmuneNet attains Spearman 0.967 (T) and 0.991 (B), RMSE 0.049 (T) and 0.032 (B), MAE 0.033 (T) and 0.024 (B), CCC 0.950 (T) and 0.986 (B), and TVD 0.047 on (T, B, O) [23]. Linear baselines are competitive for B-only RMSE but have higher composition-aware error. PLS reaches the highest T-lineage rank correlation while trailing on absolute and composition-aware metrics. NNLS-signature serves as a stress test baseline and performs markedly worse. Trends were stable under marker masking and single-gene rescaling at fixed intensities.

Model	Spearman$_T$	Spearman$_B$	RMSE$_T$	RMSE$_B$	MAE$_T$	MAE$_B$	CCC$_T$	CCC$_B$	TVD
ImmuneNet	0.967	**0.991**	**0.049**	0.032	**0.033**	**0.024**	**0.950**	0.986	**0.047**
Ridge [18,19]	0.959	0.990	0.056	0.030	0.045	0.025	0.931	**0.988**	0.056
ElasticNet [24]	0.972	0.990	0.052	**0.029**	0.041	**0.024**	0.941	**0.988**	0.053
PLS [25]	**0.981**	0.984	0.055	0.047	0.046	0.037	0.932	0.967	0.063
NNLS-signature [26]	0.953	0.988	0.429	0.210	0.424	0.197	0.181	0.540	0.424

Table 1: **Known-composition mixtures.** Metrics computed on held-out compositions after averaging replicates per composition. TVD is average total variation distance on (T, B, O), CCC is the concordance correlation coefficient [23]. Best values per column are in bold.

Statistical Significance. We estimate confidence intervals by bootstrapping composition identities with replacement (1,000 resamples) while keeping calibration and hyperparameters fixed.

Metric	Estimate	68% CI	95% CI
T Spearman	0.967	[0.962, 0.973]	[0.955, 0.978]
T Pearson	0.936	[0.932, 0.940]	[0.927, 0.944]
B Spearman	0.991	[0.990, 0.992]	[0.989, 0.992]
B Pearson	0.982	[0.981, 0.983]	[0.979, 0.984]

Table 2: Bootstrap confidence intervals for correlations on known-composition mixtures.

External validation on TCGA-OV. ImmuneNet correlates strongly with independent lineage proxies constructed as PC1 of curated T and B modules in TCGA-OV [13,14,17]: raw Spearman 0.931 (T) and 0.927 (B), raw Pearson 0.901 (T) and 0.868 (B), and partial Spearman 0.835 (T) and 0.918 (B) when conditioning on the complementary lineage.

	Raw Spearman	Raw Pearson	Partial Spearman
T lineage	0.931	0.901	0.835
B lineage	0.927	0.868	0.918

Table 3: **External validation on TCGA-OV.** Correlation of ImmuneNet predictions with PC1 proxy scores for T and B modules; partial Spearman conditions on the complementary lineage [13,14,17].

Interpretability signals. Attention concentrates on expected markers, with the T branch focusing on CD3D, CD3E, CD4, and CD8A and the B branch on MS4A1, CD79A, and CD79B, matching the curated panel, supporting biological plausibility.

6 Conclusion

We introduce ImmuneNet, a context-aware estimator of T and B fractions from bulk RNA-seq built on a tumor-matched single-cell reference, a compact marker panel, and post hoc calibration. Context-safe marker curation removes genes with malignant leakage, gene-wise attention focuses signal on biologically plausible markers, and evaluation emphasizes composition-aware criteria. On held-out mixtures ImmuneNet outperforms strong linear and PLS baselines while remaining interpretable. Calibration reduced misalignment between predicted and target compositions, which was reflected in improved reliability curves [15,16]. On TCGA-OV, correlations with tumor-matched lineage proxies supported external validity of the estimates [13,14,17].

Broader impact. Accurate immune fraction estimates from bulk data can support retrospective analyses, enable stratification in translational studies, and help prioritize samples for deeper profiling. ImmuneNet achieves this by combining tumor-matched marker curation with calibrated, composition-aware predictions, producing reliable estimates at scale on bulk cohorts, widening access to immune phenotyping without new wet-lab data. Its compact, interpretable design eases deployment across biobanks and multi-site studies, where inferred compositions can surface batch artifacts, standardize cohort integration, and inform trial design.

Limitations. Our study estimates two lineages (T and B), which limits collinearity and simplifies inference. Extending to finer immune subsets will require stricter marker audits to prevent cross-lineage overlap. Pseudo-bulk training approximates bulk artifacts but may underrepresent extremes such as very high tumor purity or strong copy-number changes.

Future work. We will extend ImmuneNet beyond HGSOC by building tumor-matched references for melanoma, NSCLC, TNBC, colorectal cancer, and glioma, and evaluate cross-context transfer with light recalibration. Target deployments include immunotherapy trials, pathology screening to triage samples for spatial or single-cell profiling, and a brief protocol for laboratories to generate local, context-matched marker panels from limited single-cell data.

References

[1] Fridman, W.H., Pagès, F., Sautès-Fridman, C. & Galon, J. (2012) The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer* **12**(4):298–306.

[2] Sato, E. *et al.* (2005) Intraepithelial $CD8^+$ tumor-infiltrating lymphocytes and a high $CD8^+$/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer. *Proceedings of the National Academy of Sciences USA* **102**(51):18538–18543.

[3] Kasikova, L. *et al.* (2024) Tertiary lymphoid structures and B cells determine clinically relevant T cell phenotypes in ovarian cancer. *Nature Communications* **15**:2528.

[4] Thorsson, V., *et al.* (2018) The immune landscape of cancer. *Immunity* **48**(4):812–830.e14.

[5] Newman, A.M., Liu, C.L., Green, M.R., *et al.* (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods* **12**(5):453–457.

[6] Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D.E. & Gfeller, D. (2017) Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**:e26476.

[7] Wang, X., Park, J., Susztak, K., Zhang, N.R. & Li, M. (2019) Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications* **10**:380.

[8] Jew, B., *et al.* (2020) Accurate estimation of cell composition in bulk expression through robust integration of single-cell information (Bisque). *Nature Communications* **11**:1971.

[9] Menden, K., Marouf, M., Oller, S., *et al.* (2020) Deep learning–based cell composition analysis from tissue expression profiles (Scaden). *Science Advances* **6**(30):eaba2619.

[10] Avila Cobos, F., Alquicira-Hernandez, J., Powell, J.E., Mestdagh, P. & De Preter, K. (2020) Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications* **11**:5650.

[11] Jin, H., *et al.* (2021) A benchmark for RNA-seq deconvolution analysis under multiple confounding factors. *Genome Biology* **22**:86.

[12] Hippen, A.A., Omran, D.K., Weber, L.M., *et al.* (2023) Performance of computational algorithms to deconvolve heterogeneous bulk ovarian tumor tissue depends on experimental factors. *Genome Biology* **24**:239.

[13] Izar, B., Tirosh, I., Stover, E.H., *et al.* (2020) A single-cell landscape of high-grade serous ovarian cancer. *Nature Medicine* **26**(8):1271–1279.

[14] GEO Series GSE217517 (2022) Paired bulk and single-cell RNA-seq on high-grade serous ovarian cancer samples. *NCBI Gene Expression Omnibus*.

[15] Guo, C., Pleiss, G., Sun, Y. & Weinberger, K.Q. (2017) On calibration of modern neural networks. In *Proc. ICML*, pp. 1321–1330.

[16] Zadrozny, B. & Elkan, C. (2002) Transforming classifier scores into accurate multiclass probability estimates. In *Proc. KDD*, pp. 694–699.

[17] Goldman, M.J., *et al.* (2020) Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology* **38**(6):675–678.

[18] Hoerl, A.E. & Kennard, R.W. (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1):55–67.

[19] Tikhonov, A.N. & Arsenin, V.Y. (1977) *Solutions of Ill-Posed Problems*. Winston & Sons, Washington, DC.

[20] Aitchison, J. (1982) The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B* **44**(2):139–177.

[21] Loshchilov, I. & Hutter, F. (2019) Decoupled weight decay regularization. In *Proc. ICLR*.

[22] Loshchilov, I. & Hutter, F. (2017) SGDR: Stochastic gradient descent with warm restarts. In *Proc. ICLR*.

[23] Lin, L.I.-K. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**(1):255–268.

[24] Zou, H. & Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67**(2):301–320.

[25] Wold, S., Sjöström, M. & Eriksson, L. (2001) PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**(2):109–130.

[26] Lawson, C. & Hanson, R. (1974) *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ.