Praxis-BGM: Clustering of Omics Data Using Semi-Supervised Transfer Learning for Gaussian Mixture Models via Natural-Gradient Variational Inference

Qiran Jia[1,*], Jesse A. Goodrich[2], David V. Conti[1,3]

[1]Division of Biostatistics and Health Data Science, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California
[2]Division of Environmental Health, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California
[3]Department of Biostatistics and Informatics, Colorado School of Public Health

[*]Correspondence: qiranjia@usc.edu

Abstract

Motivation: High-dimensional omics data are typically measured on limited sample sizes, which challenges model-based clustering methods such as Gaussian mixture models, often leading to instability and poor generalization under complex mixture structures. To address these limitations, we developed Praxis-BGM, a natural-gradient variational inference framework for Gaussian mixture models that enables semi-supervised transfer learning by incorporating an informative prior Gaussian mixture model derived from large-scale reference data with robust cluster structures. This prior can encode cluster-specific means, covariance structures, and structural connectivity patterns, and is updated using the target data with variational inference to improve clustering in small-sample settings.
Results: We derived natural-gradient updates for standard parameters and assess feature-level contributions to posterior clustering via Bayes Factors. Implemented in Python library JAX for accelerator-oriented computation, Praxis-BGM is computationally efficient and scalable. Across extensive simulations and two real-world applications—breast cancer bulk transcriptomics for subtype recovery and single-cell transcriptomics for cross-platform label transfer—Praxis-BGM improves posterior clustering performance, stability, and biological interpretability, even when priors are partially mismatched.
Availability and Implementation: Praxis-BGM is freely available at `https://github.com/ContiLab-usc/Praxis-BGM`, and an archival version is available on Zenodo at `https://doi.org/10.5281/zenodo.19657680`.
Contact: qiranjia@usc.edu
Supplementary Information: Supplementary materials are available with the manuscript submission.

Keywords: Bayesian Clustering, Statistical Transfer Learning, Variational Inference, Gaussian Mixture Model, Omics

1 Introduction

Omics data, such as transcriptomics, metabolomics, and proteomics, offer potential insight into human health and disease. Since omics data captures an enormous number of molecular parameters, the inherent patterns and complexities within these measurements can reveal distinct disease mechanisms, variations in disease risk, or heterogeneous survival probabilities [25, 33]. Unsupervised clustering methods, such as Gaussian Mixture Models (GMMs), infer underlying patterns and assign data points to latent clusters, playing a crucial role in omics analyses for data exploration, disease risk stratification, subtype discovery, cell-type identification, and other downstream inferences [12, 25]. However, complex correlations, high dimensionality, and limited sample sizes (HDLSS) in many cohort studies pose substantial challenges for statistical analysis [25, 33, 37, 43].

As a burgeoning concept, statistical transfer learning offers an effective strategy for addressing these challenges by leveraging a related source domain to improve modeling and inference in a target domain, especially in generalized linear model (GLM)-based settings [19, 27, 34]. Existing transfer-learning extensions of GMMs have mainly focused on unsupervised settings, in which all available datasets are jointly clustered. One example is a multi-task GMM framework that jointly learns multiple related GMMs by leveraging shared parameter structure across datasets and improving robustness to outlier datasets [36]. Distributed transfer GMMs, in contrast, learn node-specific clustering models while exchanging model information across decentralized sites to improve clustering without sharing raw data [38]. While useful, these approaches do not directly address the setting in which the source domain is well characterized by robust clusters, often with interpretable annotations, and a reference GMM can therefore be reliably fitted or constructed, such as when a large biobank or consortium dataset with cluster annotations is used to guide clustering in a smaller, more specialized clinical study. Bayesian inference provides a natural mechanism for transfer learning in this setting by encoding source-domain knowledge into informative priors [34, 39]. Accordingly, we propose Praxis-BGM, a semi-supervised Bayesian transfer-learning method for GMMs, in which the source-domain-derived GMM is used as an informative prior for inference in the target domain. Building on recent advances in natural gradient variational inference (NGVI) for mixtures of minimal exponential-family distributions [18, 21], Praxis-BGM effectively updates this prior model with target-domain data, allowing the source-informed prior model to regularize the posterior while still adapting to the observed target domain.

Figure 2 provides an overview of the Praxis-BGM workflow, illustrating prior construction from a source domain and NGVI-based update in the target domain. In this setting, the source domain is assumed to contain robust cluster structures or known curated cluster labels, and ideally to be annotated with meaningful biological interpretations (e.g., cancer subtypes). To the best of our knowledge, no existing Bayesian framework has been proposed for semi-supervised transfer learning in GMMs under labeled reference datasets. This provides a principled transfer mechanism that complements existing GLM-based transfer learning strategies [19]. We also showed that priors can be constructed not only from a source dataset in the same feature space with different samples, but also from an auxiliary omic layer measured on the same samples, thereby enabling multi-omics in-serial integration, as explored in previous studies but within different modeling frameworks [14, 43].

We applied Praxis-BGM to breast cancer subtyping across transcriptomic datasets to transfer

subtype structure from a reference cohort to a target cohort, and subsequently performed in-serial integrative clustering using Praxis-BGM from RNA-seq to miRNA within the same subjects to refine molecular subtypes. We also applied the method to the task of label transfer across single-cell RNA-seq technologies (i.e., sequencing platforms), such as inDrop (droplet-based) and CEL-seq, using pancreatic single-cell RNA-seq data, where transfer and inference were performed in the latent embedding space. In both applications, ground-truth labels were either fully or partially available for the query data, allowing for a direct validation of Praxis-BGM's superior performance in transferring prior knowledge and guiding clustering.

The remainder of the paper is structured as follows. Section 2 details the proposed method and optimization with the NGVI framework. Section 3 introduces the real-world data and outlines the experimental settings for applications. Section 4 reports the results, demonstrating the strong and consistent advantages of Praxis-BGM. We also present extensive simulation results evaluating the effectiveness of Praxis-BGM in transferring cluster structure via individual components of informative priors, as well as jointly, across diverse signal-to-noise regimes and source–target discrepancies in Supplementary Data S.1 and S.2. Overall, these simulation results demonstrate that Praxis-BGM achieves efficient, robust and consistently superior performance relative to competing methods, even under challenging and partially misspecified conditions. Finally, Section 5 concludes with a discussion of our findings and future research directions.

2 Methods

Let $z \in \mathbb{R}^P$ be a random variable following a GMM, i.e., a mixture of multivariate Gaussian distributions, with J components. Its marginal density is

$$p(z) = \sum_{j=1}^{J} \theta_j \, \mathcal{N}_P\big(z \mid \mu_j, \Sigma_j\big), \text{ such that } \sum_{j=1}^{J} \theta_j = 1, \tag{2.1}$$

where θ_j is the mixing weight for component j, $\mu_j \in \mathbb{R}^P$ is the corresponding mean vector, and $\Sigma_j \in \mathbb{S}^P_{++}$ is the corresponding symmetric positive definite (SPD) covariance matrix.

2.1 Priors from Source Data in the Same Feature Space as the Target Data

Suppose a source dataset $\mathcal{D}_0$, measured in the same P-dimensional feature space as the target dataset $\mathcal{D}$, is available from a related population. We leverage $\mathcal{D}_0$ to construct informative priors for z. This yields a Bayesian transfer-learning formulation in which the source data inform the prior distribution $p(z \mid \mathcal{D}_0)$, and the target data $\mathcal{D}$ updates this prior through Bayes' rule [34]. Specifically,

$$p(z \mid \mathcal{D}, \mathcal{D}_0) \propto p(\mathcal{D} \mid z)\, p(z \mid \mathcal{D}_0). \tag{2.2}$$

The normalizing constant $p(\mathcal{D} \mid \mathcal{D}_0) = \int p(\mathcal{D} \mid z)\, p(z \mid \mathcal{D}_0)\, dz$ is generally intractable. Therefore, in a two-stage procedure, we first estimate the source-informed prior distribution $p(z \mid \mathcal{D}_0)$ and then employ variational inference to approximate the posterior $p(z \mid \mathcal{D}, \mathcal{D}_0)$ by maximizing a tractable lower bound on the log marginal likelihood, namely the evidence lower bound (ELBO) [4].

Specifically, we parameterize the prior $p(z \mid \mathcal{D}_0)$ as a GMM with parameters $\omega = \{\theta_{0j}, \mu_{0j}, \Sigma_{0j}\}_{j=1}^{J}$, representing the distribution learned from $\mathcal{D}_0$. When $\mathcal{D}_0$ includes cluster labels $\mathbf{X}_0$, we can directly use cluster-specific summary statistics or discriminant analysis to derive ω. In the absence of labels, a standard GMM can be used to estimate the prior mixture distribution, with the optimal number of clusters selected using standard criteria such as the Bayesian Information Criterion (BIC) [12, 37]. It is often useful to inflate the variance of the source-informed prior, as the distribution of z may differ substantially across domains, to improve robustness [34]. Specifically, for each component j, we replace Σ_{0j} by $c\,\Sigma_{0j}$ with $c > 1$. The prior cluster-specific covariance matrices Σ_{0j} can optionally be adjusted by additional prior knowledge to encourage sparsity as

$$\Sigma_{0j}^{(A)} \;=\; A_j \odot \Sigma_{0j}, \tag{2.3}$$

where the sparsity-imposing matrix $A_j \in \{0,1\}^{P\times P}$ is symmetric with unit diagonal $\big(A_{j,pp} = 1,\ \forall p\big)$, and is derived from grouping information, such as that available from existing pathway databases to encode element-wise structural modulation [35]. The matrix A_j shapes the prior covariance matrix via a Hadamard mask as shown in Equation (2.3).

2.2 Priors from an Auxiliary Omic Layer Measured on the Same Samples as the Target Data

We also allow for constructing informative priors from an auxiliary or multi-omic layer collected on the same samples as $\mathcal{D}$. Let $\mathcal{D}^{(O)}$ denote an alternative omic layer (e.g., DNA methylation) measured on the same subjects who provide the target data $\mathcal{D}$ (e.g., transcriptomics). When cluster or class labels $\mathbf{X}^{(O)}$ are available or can be reliably inferred from $\mathcal{D}^{(O)}$, we construct a prior distribution $p(z \mid \mathcal{D}^{(O)})$ parameterized by $\omega = \{\theta_{0j}, \mu_{0j}, \Sigma_{0j}\}_{j=1}^{J}$, where the parameters are obtained from discriminant analysis or empirical class-conditional summary statistics in the target feature space [12]. This enables using a biologically related but distinct omic layer, measured on the same subjects, to derive the prior distribution for the latent variable z representing upstream cluster-structure signals, in line with in-serial integration strategies for multi-omics data [14, 43].

2.3 Variational Inference with Informative Priors

Given informative priors derived from either the source data $\mathcal{D}_0$ or the auxiliary omic layer $\mathcal{D}^{(O)}$, denoted collectively by $p_\omega(z)$ for simplicity, we approximate the posterior with a surrogate finite mixture of multivariate Gaussian distributions $q_\phi(z)$ with $\phi = \{\theta_j, \mu_j, \Sigma_j\}_{j=1}^{J}$ using a NGVI framework [18, 21]. Given priors $p_\omega(z)$ and target data $\mathcal{D}$, we seek a variational distribution that maximizes the ELBO,

$$\mathcal{L}(q_\phi) = \mathbb{E}_{q_\phi(z)}\Big[\sum_{i=1}^{N} \log p(D_i \mid z)\Big] - D_{\mathrm{KL}}\big(q_\phi(z) \,\|\, p_\omega(z)\big). \tag{2.4}$$

Maximizing $\mathcal{L}(q_\phi)$ encourages $q_\phi(z)$ to explain the target mixture data $\mathcal{D}$ well, while remaining regularized toward the prior $p_\omega(z)$. In general, NGVI is a first-order optimization method for maximizing $\mathcal{L}(q_\phi)$ of minimal exponential-family (EF) distributions that replaces the standard Euclidean gradient with the natural gradient, thereby accounting for the information geometry of

the natural parameter space [18]. Natural gradient yields parameterization invariance and often significantly improves convergence efficiency and stability [1, 18]. Specifically, NGVI updates the natural parameter $\boldsymbol{\lambda}_z$ by scaling the standard Euclidean gradient with the inverse Fisher information matrix (FIM): $\boldsymbol{\lambda}_z^{(t+1)} = \boldsymbol{\lambda}_z^{(t)} + \beta F^{-1}(\boldsymbol{\lambda}_z^{(t)})\, \nabla_{\boldsymbol{\lambda}_z} \mathcal{L}(\boldsymbol{\lambda}_z^{(t)})$, where β is the step size. Direct computation and inversion of the FIM is computationally expensive in high-dimensional settings. For minimal EF variational distributions, one can avoid this explicit evaluation by working in the expectation-parameter space $\boldsymbol{m}_z$, using the duality between natural and expectation parameters [18]. Under this duality, the natural gradient for $\boldsymbol{\lambda}_z$ can thus be expressed as the ordinary gradient with respect to $\boldsymbol{m}_z$: $F^{-1}(\boldsymbol{\lambda}_z^{(t)})\, \nabla_{\boldsymbol{\lambda}_z} \mathcal{L}(\boldsymbol{\lambda}_z^{(t)}) = \nabla_{\boldsymbol{m}_z} \mathcal{L}_*(\boldsymbol{m}_z^{(t)})$, where $\mathcal{L}_*$ denotes the ELBO expressed in the expectation-parameter space. $\nabla_{\boldsymbol{m}_z} \mathcal{L}_*(\boldsymbol{m}_z^{(t)})$ can be obtained in terms of gradients with respect to the standard parameters (e.g., $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for Guassian) by using the chain rule, which in turn yields the corresponding natural-gradient update for standard parameters via the variational online Newton (VON) algorithm [21]. NGVI has been extended beyond minimal EF to mixtures of minimal EF distributions, including finite mixtures of multivariate Gaussians (i.e., GMMs), under a minimal conditional EF representation [21, 24].

We apply NGVI for GMMs with informative priors, $p_\omega(z)$, to facilitate Bayesian transfer learning. We define $\mathcal{L}(q_\phi) = \mathbb{E}_{q_\phi(z|\mathcal{D})}\big[-h(z)\big]$, and $h(z) = -\sum_{n=1}^{N} \log p(D_n \mid z) + \log \frac{q_\phi(z|\mathcal{D})}{p_\omega(z)}$. The VON algorithm is used to efficiently evaluate $h(z)$ by drawing a stochastic sample under $q_\phi(z)$ and then use it to compute natural gradients for standard parameters ϕ of GMMs [18, 21]. Specifically, at iteration t, we first draw a component v and then draw a sample conditioned on the component,

$$v \sim \mathrm{Cat}(\boldsymbol{\theta}^{(t)}), z_0 \sim q_\phi^{(t)}(z \mid j = v), \tag{2.5}$$

which serves as the anchor point for natural-gradient updates. We then define

$$\begin{aligned} h(z_0) = \log &\left(\sum_{j=1}^{J} \theta_j \, \mathcal{N}(z_0 \mid \mu_j, \Sigma_j) \right) \\ &- \log \left(\sum_{j=1}^{J} \theta_{0j} \, \mathcal{N}(z_0 \mid \mu_{0j}, \Sigma_{0j}) \right) \\ &- \sum_{n=1}^{N} \log p(D_n \mid z_0). \end{aligned} \tag{2.6}$$

The stochastic gradient and Hessian of $h(z_0)$ with respect to z_0, denoted by $\nabla_z h(z_0)$ and $\nabla_z^2 h(z_0)$, respectively, are key quantities in the derived NGVI updates for standard parameters ϕ, obtained using Bonnet's and Price's theorems as shown in [21]:

$$\boldsymbol{\Sigma}_j^{-1\,(t+1)} = \boldsymbol{\Sigma}_j^{-1\,(t)} + \beta \, \delta_j(z_0) \, \nabla_z^2 h(z_0), \tag{2.7}$$

$$\boldsymbol{\mu}_j^{(t+1)} = \boldsymbol{\mu}_j^{(t)} - \beta \, \delta_j(z_0) \, \boldsymbol{\Sigma}_j^{(t+1)} \, \nabla_z h(z_0), \tag{2.8}$$

$$\begin{aligned} \log \frac{\theta_j^{(t+1)}}{\theta_J^{(t+1)}} = \log \frac{\theta_j^{(t)}}{\theta_J^{(t)}} - \beta \left(\delta_j(z_0) - \delta_J(z_0) \right) h(z_0), \\ j = 1, \ldots, J-1. \end{aligned} \tag{2.9}$$

where

$$\delta_j(z_0) = \frac{\mathcal{N}\big(z_0 \mid \mu_j^{(t)}, \Sigma_j^{(t)}\big)}{\sum_{\ell=1}^{J} \theta_\ell^{(t)} \mathcal{N}\big(z_0 \mid \mu_\ell^{(t)}, \Sigma_\ell^{(t)}\big)}. \tag{2.10}$$

We summarize Praxis-BGM (prior-augmented NGVI for GMMs) in Algorithm 1. We initialize the variational parameters ϕ using the prior parameters ω to facilitate alignment between the target-domain cluster labels and the prior structure, thereby enhancing interpretability. Praxis-BGM is implemented in Python using JAX, enabling accelerator-friendly computation on GPUs and TPUs, as well as just-in-time compilation for scalable and numerically stable NGVI [5].

Algorithm 1. Praxis-BGM: Prior-augmented Natural Gradient Variational Inference for Gaussian Mixture Models

Inputs. Target data $\mathcal{D} = \{D_n\}_{n=1}^N$; source data $\mathcal{D}_0$ or $\mathcal{D}^{(O)}$ with labels $\mathbf{X}_0$ or $\mathbf{X}^{(O)}$ defining prior parameters $\omega = (\boldsymbol{\theta}_0, \{\mu_{0j}, \Sigma_{0j}\}_{j=1}^J)$, where $\{\Sigma_{0j}\}$ and $\boldsymbol{\theta}_0$ may be specified only if $\{\mu_{0j}\}$ is provided; structural sparsity masks $\{A_j\}$; stepsize β; number of epochs T; covariance-scaling constant c; jitter $\tau = 10^{-9}$.

Procedure.

1. Construct working prior parameters $\omega^\star$ by setting, for each $j = 1, \ldots, J$,

$$\mu_{0j}^\star = \begin{cases} \mu_{0j}, & \text{if provided} \\ \mathbf{0}_P, & \text{otherwise,} \end{cases} \qquad \Sigma_{0j}^\star = \begin{cases} (A_j \odot \Sigma_{0j})c + \tau I_P, & \text{if provided} \\ cI_P, & \text{otherwise,} \end{cases}$$

   and

$$\boldsymbol{\theta}_0^\star = \begin{cases} \boldsymbol{\theta}_0, & \text{if provided} \\ J^{-1}\mathbf{1}_J, & \text{otherwise.} \end{cases}$$

2. Initialize variational parameters ϕ. If prior means $\{\mu_{0j}\}$ are provided, set $\mu_j \leftarrow \mu_{0j}^\star$, $\Sigma_j \leftarrow \Sigma_{0j}^\star$, $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0^\star$, and $\rho_j \leftarrow \log(\theta_j/\theta_J)$. Otherwise, fit an optimal *BayesianGaussianMixture* model using [30] on $\mathcal{D}$ to obtain initial μ_j and Σ_j, then set $\boldsymbol{\theta} \leftarrow J^{-1}\mathbf{1}_J$ and initialize ρ_j accordingly.
3. For $t = 0, \ldots, T-1$, sample anchor z_0 from q_ϕ using Eq. (2.5); compute $\delta_j(z_0)$ from Eq. (2.10) and the global term $h(z_0)$ from Eq. (2.6); then, for each $j = 1, \ldots, J$, update Σ_j^{-1}, μ_j, and ρ_j using the VON updates in Eq. (2.7), Eq. (2.8), and Eq. (2.9), respectively. Finally, set $\boldsymbol{\theta} \leftarrow \mathrm{softmax}\big([\rho_1, \ldots, \rho_{J-1}, 0]\big)$.

Output. Posterior parameters ϕ and cluster responsibilities γ_{nj}.

3 Evaluation with Real-world Data

We demonstrated the effectiveness of Praxis-BGM in transferring prior cluster structural information to improve clustering performance and enhance clinical relevance compared with standard clustering and classification methods across two real-world data applications.

3.1 Breast Cancer Bulk Transcriptomics

The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) are two breast cancer genomic cohorts [9, 26], each profiling the transcriptomic expression of $\sim$ 20,000 genes (bulk RNA-seq data) and accompanied by clinical and survival information. TCGA-BRCA includes PAM50 subtype labels for 500 subjects, whereas METABRIC provides hormone-receptor status, a well-established surrogate strongly correlated with intrinsic subtypes [28, 31]. For METABRIC, we constructed surrogate subtype labels by mapping clinical receptor status to their closest intrinsic subtype counterparts. Specifically, ER+/HER2− cases were categorized as Luminal A-like, ER+/HER2+ as Luminal B-like, ER−/PR−/HER2+ as HER2-enriched-like, and ER−/PR−/HER2− as Triple-negative, used here as a surrogate for Basal-like subtype given their strong overlap [31]. These labels are interpreted as approximate clinical proxies rather than definitive molecular subtype assignments, and we used them to derive related but not fully accurate informative priors for transfer-learning-based clustering. For TCGA-BRCA, we utilized the RNA-based PAM50 labels available as ground-truth, noting that 560 subjects (52.83%) in the expanded cohort lack these formal molecular annotations.

The RNA-seq data of METABRIC was used as the labeled source dataset to derive priors for Praxis-BGM, with the RNA-seq data of TCGA-BRCA serving as the unlabeled target dataset. We hypothesized that incorporating informative priors in METABRIC would yield clusters in TCGA-BRCA that better align with PAM50 subtypes and exhibit stronger prognostic relevance than clustering or supervised prediction across datasets alone. After filtering incomplete records, 1,980 METABRIC and 1,060 TCGA-BRCA samples remained. Using $\log_2$ transformed gene expressions shared across datasets, we selected the top 1,000 most variable genes based on METABRIC based on an elbow-based selection method on the ranked gene variance plot to remove low-information genes. We then applied standardization and subset both datasets accordingly. For each METABRIC subtype, we computed empirical means and covariance matrices to form component-specific priors for Praxis-BGM.

For performance evaluation, Praxis-BGM clusters were compared with the unsupervised Bayesian Gaussian Mixture Models (BGM) [37], a VI-based GMM method with uninformative priors (fit with $J = 4$), and supervised classifiers trained on the source data and applied to the target data (multinomial logistic regression, Linear Discriminant Analysis (LDA) [12], and Extreme Gradient Boosting (XGBoost) [7] tuned via cross-validation). Estimated cluster labels in TCGA-BRCA were evaluated by Adjusted Rand Index (ARI) and accuracy with PAM50 labels, and were further evaluated as predictors in separate age-adjusted Cox proportional hazards models for five-year survival. Bayes factors (BFs) for feature importance in driving breast cancer subtype differentiation (method detailed in Supplementary Data S.3.1) were computed to identify subtype-driven genes, followed by functional enrichment analysis using Gene Ontology (GO) Biological Process [2, 8] and Human Molecular Signatures Database (MSigDB) Hallmark gene sets [20] via `GSEApy` [11].

We next propagated the estimated RNA-seq Praxis-BGM clusters as priors for clustering of the microRNA (miRNA) layer in the same TCGA-BRCA cohort, enabling a second Bayesian update of the GMM while achieving in-serial multi-omics integration [14, 43]. Biologically, miRNA captures regulatory signals not fully reflected by RNA-seq expression alone and may provide complementary information for breast cancer subtyping [3]. This in-serial analysis follows Section 2.2 and the

analysis method is detailed in Supplementary Data S.3.4.

3.2 Cross-platform Label Transfer in Pancreatic scRNA-seq

When analyzing single-cell RNA sequencing (scRNA-seq) data, a key initial task is to annotate cell types for clusters of cells. The increasing availability of annotated reference datasets has led to numerous label-transfer methods [29]. Among them, scANVI [40], a semi-supervised extension of scVI [22], is a state-of-the-art approach that jointly models labeled reference and unlabeled query data. The `Scanpy` framework [15] also provides K-Nearest Neighbors(KNN)-based label-transfer functionality by integrating PC embeddings across datasets, called `ingest`, and heterogeneity-correction methods, such as ComBat-seq [42] and Scanorama [16], can be paired with KNN classifiers for label transfer.

Label transfer naturally aligns with the goal of Praxis-BGM, which performs clustering guided by cell-type-specific priors. We used the human pancreas scRNA-seq atlas assembled by [23], a benchmark dataset of 16,382 cells annotated into 14 cell types across four platforms (inDrop, CEL-Seq, Smart-seq, Fluidigm C1), chosen for its heterogeneous batch structure. After quality control to remove low-quality cells (zero-count entries), we applied $\log_2$ transformation, selected the top 2,000 highly variable genes (HVGs) using `Scanpy`, standardized them, performed PCA, and retained the top 100 principal components. Cells from inDrop (52.3%) served as the reference dataset, while cells from the remaining platforms (47.7%) formed the query dataset. Ground-truth annotations were available for all cells and used only for evaluation on the query subset. Figure 4(a) shows the cell-type distributions across reference and query sets.

For Praxis-BGM, we computed cell-type-specific mean priors in the PC space corrected by ComBat-seq from the reference data and applied them to guide clustering on the query data. We benchmarked Praxis-BGM against scANVI, `Scanpy`'s `ingest` function, and integration-based methods (ComBat-seq and Scanorama) followed by KNN classification. Notably, Praxis-BGM and scANVI use the reference information differently: Praxis-BGM incorporates priors directly into cluster inference on the query set, whereas scANVI learns a shared latent embedding for both datasets before label prediction. For computational time comparisons, we measured only the cost of heterogeneity correction and inference, excluding embedding computation.

4 Results

4.1 Praxis-BGM Enhances Breast Cancer Subtype Recovery and Clinical Relevance in TCGA-BRCA

Our results show that, by leveraging informative priors derived from METABRIC, Praxis-BGM improved both subtype concordance and clinical interpretability relative to unsupervised clustering and supervised classification approaches applied to TCGA-BRCA RNA-seq data.

Table 1 presents the ARI and accuracy results for concordance with ground-truth PAM50 subtype labels for Praxis-BGM and other benchmark methods on TCGA-BRCA RNA-seq data. BGM clusters captured partial concordance with the PAM50 molecular subtypes (ARI = 0.392), indicating that even without subtype-specific priors, the intrinsic transcriptomic structure of TCGA-BRCA

Table 1: Benchmarking clustering and classification methods on TCGA-BRCA RNA-seq data by assessing the ARI and accuracy with ground-truth PAM50 subtype labels, restricted to the 500 subset of samples with available subtype annotations, and by evaluating prognostic value using the likelihood-ratio test (LRT) statistic and corresponding p-value from age-adjusted Cox proportional hazards models for 5-year survival, and reporting runtime with the corresponding hardware.

Method	**ARI**	**Accuracy**	**LRT Stat (p-value)**	**Time (s) / Hardware**
Praxis-BGM	**0.430**	**0.714**	**57.787** (8.460×10^{-12})	6.04 /V6e TPU
BGM	0.392	-	45.121 (3.752×10^{-9})	**1.77**/CPU
Multinomial Regression	0.381	0.654	31.711 (2.191×10^{-6})	2.24/CPU
LDA	0.379	0.650	33.661 (8.744×10^{-7})	11.08/CPU
XGBoost	0.351	0.638	30.556 (3.771×10^{-6})	11.82/CPU

samples reflects moderate underlying subtype-related patterns. For classification benchmark methods, multinomial regression and LDA produced similar and better subtype concordance (ARI = 0.381 and 0.379, accuracy = 0.654 and 0.650 respectively). XGBoost yielded the lowest ARI (0.351) and accuracy (0.638) among the reported supervised classification methods, suggesting potential overfitting despite the use of 10-fold cross-validation for hyperparameter tuning. Among all methods, Praxis-BGM achieved the strongest overall performance (ARI = 0.430; accuracy = 0.714). We then visualized the 5-year survival across Praxis-BGM clusters using Kaplan Meier curves in Panel (a) Figure 3. The Kaplan–Meier curves show clear survival heterogeneity across the inferred Praxis-BGM clusters. The posterior Praxis-BGM cluster corresponding to Luminal A has the best prognosis, whereas the cluster corresponding to Luminal B (HER2+) has the poorest survival with the steepest decline. Overall, Praxis-BGM clusters yield the strongest survival stratification under the age-adjusted Cox proportional hazards model for 5-year survival (LRT statistic = 57.79, p-value = 8.460×10^{-12}). The Cox model using BGM clusters is likewise significant (LRT statistic = 45.121, p-value = 3.752×10^{-9}) but less pronounced, whereas all three classification methods show weaker significance in their corresponding Cox models. Taken together, these findings show that Praxis-BGM clusters not only recover established subtype structure, but also provide superior prognostic stratification beyond benchmark unsupervised and supervised methods through Bayesian transfer learning. As for runtime, although BGM achieved the fastest CPU runtime, Praxis-BGM remained computationally efficient and benefited from TPU acceleration.

To identify genes driving breast cancer subtype differentiation, we computed BFs for all genes and identified 39 genes showing at least substantial evidence based on Jeffreys' scale [17]. To better interpret the identified clustering-drive genes, Panel (b) of Figure 3 shows the results of the pathway over-representation analysis for genes with at least substantial evidence (BF $> 10^{0.5}$). After FDR adjustment, seven pathways were significant. Interestingly, the two estrogen response pathways found in MSigDB are well-established signatures in the progression of breast cancer [41]. Among GO Biological Process, the three miRNA-related significant pathways are also biologically relevant to breast cancer, as aberrant miRNA expression has been implicated in tumor initiation, progression, and therapeutic resistance [3]. These results suggest that integrating informative priors from METABRIC through Praxis-BGM not only yields more clinically relevant clusters, but also improves biological interpretability and downstream analysis, further motivating the in-serial integration of the miRNA layer in the target data.

For the in-serial analysis with miRNA data in TCGA-BRCA, leveraging the posterior RNA-seq Praxis-BGM clusters as priors, the posterior miRNA Praxis-BGM clusters were largely consistent with the RNA-based Praxis-BGM clusters while providing refined stratification that improved alignment with PAM50 labels among subjects with available annotations and miRNA data. Specifically, clustering agreement measured by ARI increased from 0.417 to 0.478. To explore this increase, in Panel (c) of Figure 3 we presents the subtype-specific concordance of the RNA- and in-serial miRNA-based Praxis-BGM clusters with PAM50 ground-truth labels. Notably, the Praxis-BGM clusters updated with the miRNA layer correctly assigned 9 additional subjects to the Luminal B subtype compared with the RNA-based Praxis clusters. These refined posterior miRNA Praxis-BGM clusters incorporate priors from both the external source dataset METABRIC and the RNA-seq layer of TCGA-BRCA, providing an illustrative example of multi-omics integration through sequential Bayesian updating.

4.2 Praxis-BGM Outperforms Existing Label Transfer Methods on Cross-platform scRNA-seq

Benchmarking on cross-platform scRNA-seq data revealed that Praxis-BGM provides superior label transfer accuracy and robustness compared with widely used alternatives.

Table 2: Benchmarking label transfer methods on Pancreas single-cell transcriptomic data against expert annotations, and reporting runtime with the corresponding hardware platform.

Method	**ARI**	**Accuracy**	**Time (s)**	**Hardware**	**Embedding**
Praxis-BGM	**0.957**	**0.975**	**2.94**	T4 GPU	PCs
ComBat-seq + KNN	0.933	0.953	3.4	CPU	PCs
scANVI	0.912	0.921	33.0	T4 GPU	scVI
Scanorama + KNN	0.913	0.916	15.5	CPU	PCs
Scanpy-ingest	0.896	0.915	10.3	CPU	PCs

Table 2 presents the results of benchmark label transfer methods on the query subset. Praxis-BGM and ComBat-seq + KNN substantially outperformed other methods in terms of ARI and accuracy, while also being computationally faster. Notably, Praxis-BGM achieved the highest clustering accuracy and ARI (accuracy = 0.975 and ARI = 0.957), better than those of ComBat-seq + KNN and scANVI. Because ComBat-seq + KNN uses the same corrected PC embeddings as Praxis-BGM, the 2.2% improvement in accuracy and 0.024 improvement in ARI further highlight the performance advantage of Praxis-BGM over the deterministic classification method of KNN. `Scanpy-ingest` and Scanorama + KNN yielded inferior ARI and accuracy, and required longer computation times than Praxis-BGM did. Among all methods, scANVI was the most computationally intensive as it is neural-network-based.

Praxis-BGM and scANVI relied on different embeddings to perform the label transfer (PCs for Praxis-BGM and scVI representations for scANVI). The improved performance of Praxis-BGM compared with other KNN-based methods that also use PC embeddings suggests a potential advantage of Praxis-BGM in better adapting to settings where the reference and query domains are not fully homogeneous leveraging variational inference. Figure 4(b) provides the confusion heatmap comparing predicted cell types to ground-truth annotations of the cell types in the query data. No-

tably, the annotated reference subset contained 7 T cells while the query subset did not (Fig. 4(a)). ComBat-seq + KNN and scANVI exhibited pronounced misassignment of Acinar cells to Ductal, while Praxis-BGM did not. All methods showed some degree of confusion between Quiescent and Activated Stellate cells, likely due to their similarity, and all performed somewhat less accurately on the rarer cell types of Mast and Epsilon. Praxis-BGM and scANVI accurately assigned the rare cell type of Schwann, while ComBat-seq + KNN did not. In addition, we conducted an additional sensitivity analysis investigating whether KNN-based benchmark label-transfer methods would benefit from using a smaller and more targeted set of source-derived differentially expressed genes (DEGs) in Supplementary Data S.4.1, rather than the original workflow based on 2,000 HVGs. Overall, more aggressive DEG restriction modestly improved KNN-based benchmark methods, while Praxis-BGM remained the best-performing approach across all methods. We also evaluated a scenario in which the reference dataset contained fewer rare cell types than the query dataset (Supplementary Data S.4.2). Under this mismatch, Praxis-BGM still outperformed the benchmark methods.

This application demonstrates that Praxis-BGM can leverage cell-type-informative priors to achieve accurate and computationally efficient label transfer between heterogeneous single-cell datasets, making it a practical and scalable alternative to existing methods. Looking ahead, as emerging technologies (e.g., in situ spatial transcriptomics) enable more comprehensive characterization of rare cell types, increasingly comprehensive reference datasets can be used for Praxis-BGM. Capable of addressing platform heterogeneity, Praxis-BGM is also especially useful for transferring well-annotated scRNA-seq labels to spatial RNA-seq data, where direct cell-type identification remains challenging.

5 Discussion

In this paper, we introduced Praxis-BGM, a semi-supervised GMM clustering method based on NGVI that facilitates transfer learning. The source data are assumed to have curated cluster labels, which, if not already available, can be obtained by fitting an unsupervised GMM with an appropriate model selection procedure and subsequently annotating the labels through biological interpretation. Relying on the source data, Praxis-BGM provides a highly flexible framework for incorporating prior information through a prior GMM estimated and updating the prior model with the target data through variational inference. One or any combination of cluster-specific parameters in the prior GMM can be specified, including the mean vector (prior belief in differences in expression levels across clusters), covariance matrix (prior belief in differences in variability and correlation patterns across clusters), and weights (prior belief in differences in cluster prevalence). In addition, overall or cluster-specific sparsity-inducing matrices informed from pathway and ontology databases can be incorporated to encode feature connectivity further. These prior components define a prior GMM and regularize the posterior estimation of the GMM on the target data. We demonstrated the superior performance of Praxis-BGM over competing deterministic classification-based methods in both simulation studies and real-world applications, as well as the advantage of incorporating prior knowledge over purely unsupervised clustering. In the context of breast cancer, Praxis-BGM successfully transferred subtype-specific component structures from METABRIC to TCGA-BRCA, yielding significant improvements in biological relevance and clinical utility while also enabling the discovery of cluster-associated genes. We further showed that Praxis-BGM can

flexibly integrate multi-omic layers for in-serial analysis by using cluster-specific priors derived from a preceding omic layer to inform the analysis of the next layer. We also demonstrated that the application of Praxis-BGM extends naturally to label transfer in single-cell transcriptomics to effectively leverage annotated reference datasets to cluster and annotate query cells.

When multiple omics layers are available, the choice of which layer to prioritize for transfer learning, and whether a downstream in-serial update with additional omics layers is beneficial, depends mainly on two considerations: (i) biological relevance to the scientific question and study objective, and (ii) data availability and completeness. For example, in the breast cancer bulk transcriptomics application, our primary objective and main validation metric were to recover PAM50 labels in the target data. Accordingly, the transcriptomic layer is naturally the most informative layer for this objective, whereas the miRNA layer may offer complementary refinement for explorative analysis. By contrast, other omics layers, such as copy number variation (CNV), may not be prioritized for this specific goal. In practice, it is also often reasonable to begin with the layer measured in the largest and most complete set of subjects. More broadly, Praxis-BGM is a highly flexible approach, as priors may be derived from an external source dataset in the same feature space, from another omics layer measured in the same cohort, or from a sequential combination of both, as illustrated in the breast cancer bulk transcriptomics application. The most appropriate choice of modeling strategy depends on the research objective, the underlying biology, and the structure and availability of the source and target data. Moreover, when the source dataset contains high-quality integrative cluster labels estimated from multi-omics data, Praxis-BGM can still leverage that information to perform transfer learning using a single omics layer shared between the source and target datasets; the target dataset does not need to contain all omics layers available in the source data. An important direction for future work is the development of a parallel multi-omics transfer-learning framework that jointly integrates multiple omics layers while allowing layer-specific priors to be derived from potentially distinct source datasets.

One caveat about statistical transfer learning is the risk of negative transfer, which could occur when the knowledge transferred from the source data reduces the performance of modeling and inference in the target data [27]. Inappropriate priors from unsuitable source data for Praxis-BGM can also potentially lead to negatively biased posteriors. Our Variational Bayes framework seeks to mitigate this potential negative impact by reconciling discrepancies between priors and data through variational inference. Simulation studies demonstrate that Praxis-BGM maintains strong performance even under severely mis-specified priors. Future work could focus on enhancing robustness by learning the reliability of priors derived from multiple sources and selecting the most informative source datasets or samples through data-driven methods. For supervised statistical transfer learning methods, well-established strategies exist to select the set of informative source samples [19]. [39] introduced a prior-misalignment-risk minimizer conditional on source parameters that estimates the inclusion probability of each source dataset through a full Bayesian framework. A similar mechanism could be incorporated into the Praxis-BGM framework to reduce the risk of negative transfer.

6 Availability of data

The METABRIC breast cancer dataset [9] was accessed through `cBioPortal` [6, 10, 13], and the TCGA-BRCA dataset [26] was obtained from the R package `curatedTCGAData` [32]. The pancreas scRNA-seq benchmark data followed the article [23] and were assembled from the public GEO and ArrayExpress datasets listed in that article's supplementary materials.

7 Competing interests

No competing interest is declared.

8 Author contributions statement

All authors contributed to the conception and development of the methodology, performed the experiments and data analyses, and jointly wrote and reviewed the manuscript.

9 Funding

This study was supported by the National Institutes of Health grants P01CA196569, U01CA261339, P30CA014089, P30ES007048, U01CA164973, P42ES036506, R01DK14083, and 5U01HG013288 (LEON).

References

[1] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[2] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, J. Michael Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, Julie C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[3] Gloria Bertoli, Claudia Cava, and Isabella Castiglioni. Micrornas: New biomarkers for diagnosis, prognosis, therapy prediction and therapeutic tools for breast cancer. *Theranostics*, 5(10):1122–1143, 2015.

[4] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[5] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. Jax: composable transformations of python+numpy programs, 2018.

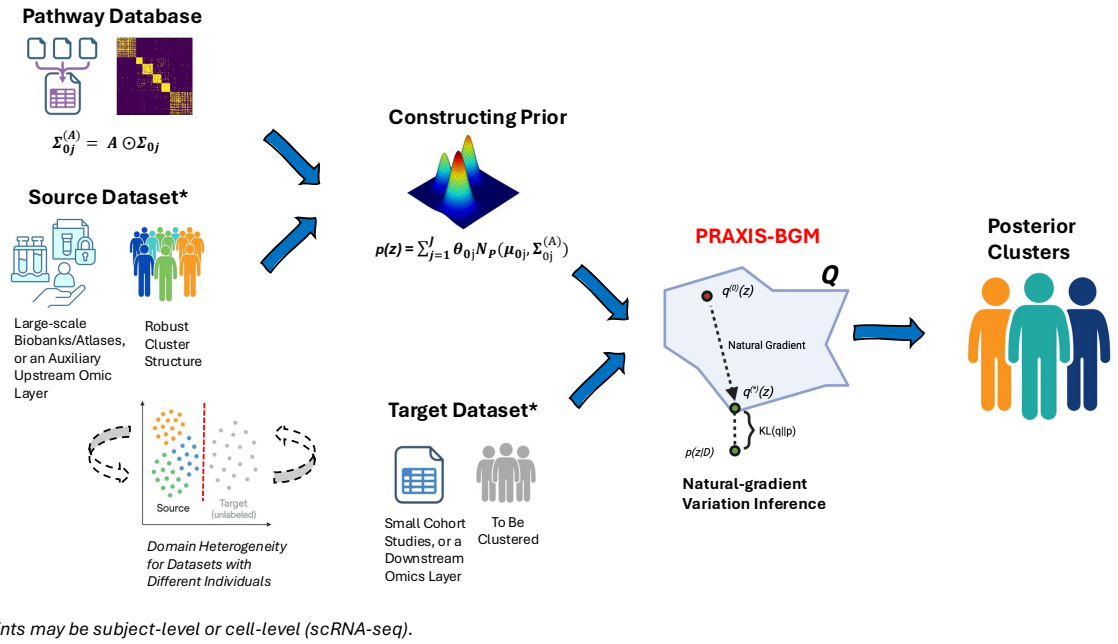

Figure 2: Overview of the Praxis-BGM framework for prior construction and natural-gradient variational inference-based Bayesian transfer learning for Gaussian Mixture Models, where the source of prior information may originate either from a different dataset measured on different samples (e.g., large-scale biobanks or atlases) or from an informative upstream omic layer measured on the same samples (e.g., from RNA-seq to miRNA).

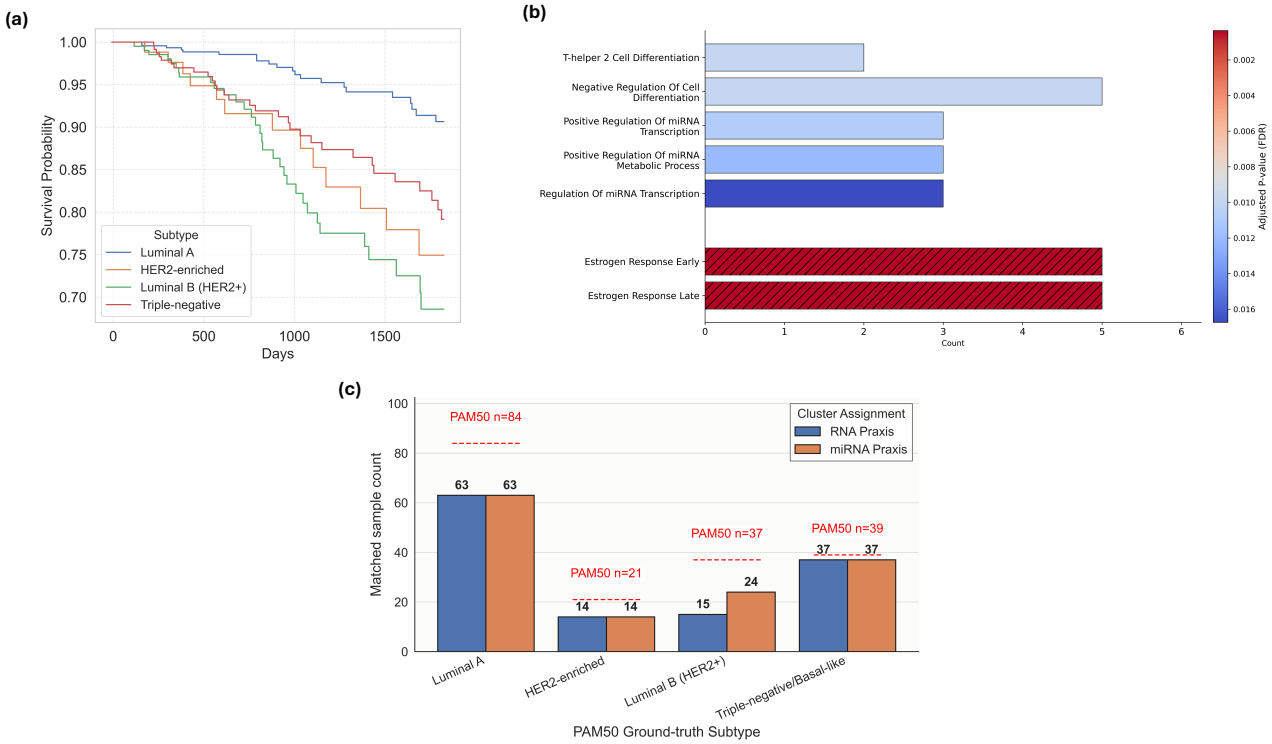

Figure 3: (a) Kaplan–Meier curve comparing survival across Praxis-BGM clusters annotated by subtypes; (b) Pathway enrichment analysis of the non-indeterminate genes (BF $> 10^{0.5}$). The solid bars represent pathways from the Gene Ontology (GO) Biological Process database, whereas the hatched bars represent Human Molecular Signatures Database (MSigDB) Hallmark gene sets; (c) Subtype-specific concordance of RNA- and miRNA-based Praxis-BGM clusters with PAM50 ground-truth labels.

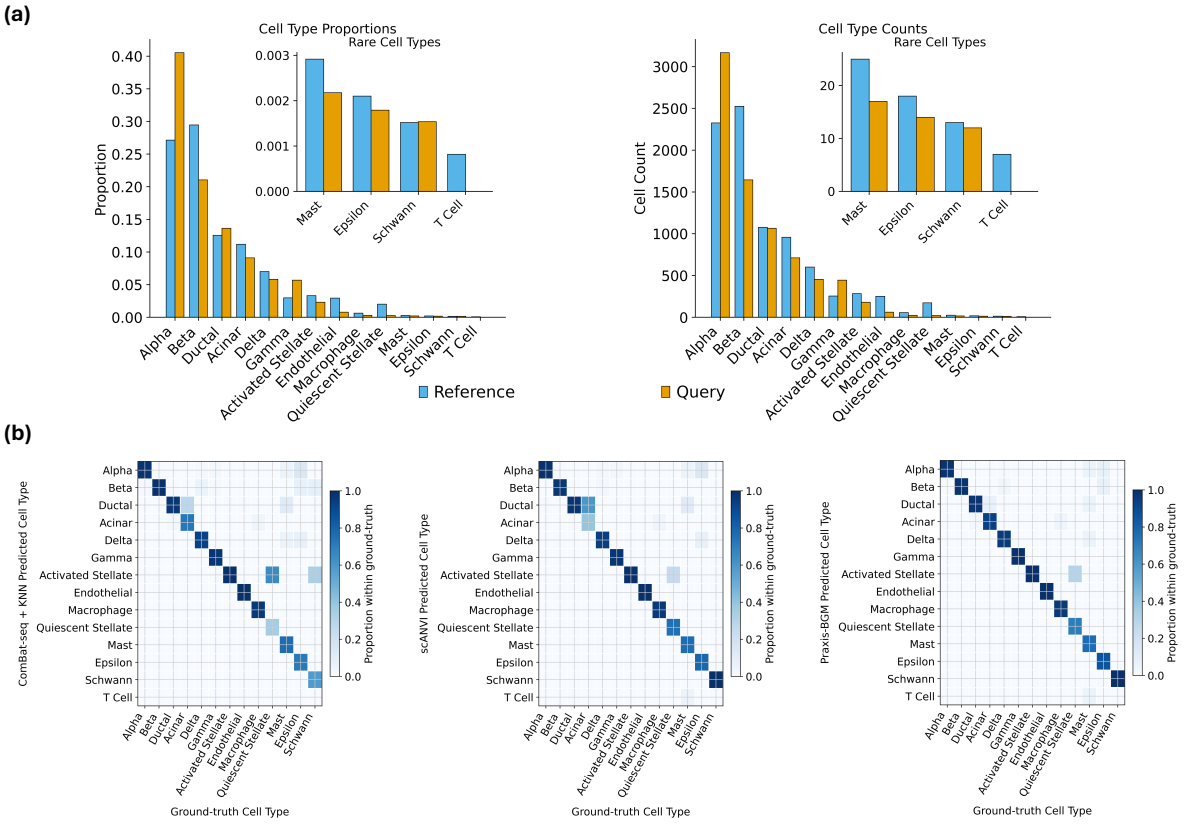

Figure 4: (a) Ground-truth cell type proportions (left) and count (right) distributions for the reference and query datasets. (b) Confusion heatmap comparing predicted cell types to ground-truth annotations in the query data for ComBat-seq + KNN, scANVI, and Praxis-BGM, with color intensity representing the proportion within each ground-truth cell type.

[6] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, S. Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P. Goldberg, Chris Sander, and Nikolaus Schultz. The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012.

[7] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[8] The Gene Ontology Consortium. The gene ontology knowledgebase in 2026. *Nucleic Acids Research*, 54(D1):D1779–D1792, 01 2026.

[9] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[10] Ino de Bruijn, Ritika Kundra, Brooke Mastrogiacomo, Thinh Ngoc Tran, et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR project GENIE biopharma collaborative in cBioPortal. *Cancer Research*, 83(23):3861–3867, 2023.

[11] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 01 2023.

[12] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[13] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269):pl1, 2013.

[14] Jesse A. Goodrich, H. Wang, Q. Jia, N. Stratakis, Y. Zhao, David V. Conti, et al. Integrating multi-omics with environmental data for precision health: A novel analytic framework and case study on prenatal mercury induced childhood fatty liver disease. *Environment International*, 190:108930, 2024.

[15] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021.

[16] Brian L. Hie, S. Kim, Thomas A. Rando, Bryan Bryson, and Bonnie Berger. Scanorama: integrating large and diverse single-cell transcriptomic datasets. *Nature Protocols*, 19(8):2283–2297, August 2024.

[17] Harold Jeffreys. *Theory of Probability*. Oxford University Press, 3rd edition, 1961.

[18] Mohammad Emtiyaz Khan and Didrik Nielsen. Fast yet simple natural-gradient descent for variational inference in complex models. In *2018 International Symposium on Information Theory and Its Applications (ISITA)*, pages 31–35. IEEE, 2018.

[19] Sai Li, T. Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 02 2022.

[20] Arthur Liberzon, Chad Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database (msigdb) hallmark gene set collection. *Cell Systems*, 1(6):417–425, December 2015.

[21] Wu Lin, Mohammad Emtiyaz Khan, and Mark Schmidt. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3992–4002. PMLR, 2019.

[22] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018.

[23] Malte D. Luecken, Maren Büttner, Karan Chaichoompu, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022.

[24] Farzaneh Mahdisoltani. Natural gradient variational inference with gaussian mixture models. *arXiv preprint arXiv:2111.08002*, 2021.

[25] Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.

[26] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[27] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[28] Joel S Parker, Michael Mullins, Maggie CU Cheang, Susan Leung, David Voduc, Thomas Vickery, Sherri Davies, Christian Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, 2009.

[29] Giovanni Pasquini, Jesus Eduardo Rojo Arias, Patrick Schäfer, and Volker Busskamp. Automated methods for cell type annotation on scrna-seq data. *Computational and Structural Biotechnology Journal*, 19:961–969, 2021. Open Access under CC BY-NC-ND license.

[30] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] Charles M. Perou, Therese Sørlie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Christian A. Rees, Jonathan R. Pollack, David T. Ross, Hilde Johnsen, Lars A. Akslen, Øystein Fluge, Anna Pergamenschikov, Christian Williams, Shuji X. Zhu, Per Eystein Lønning, Anne-Lise Børresen-Dale, Patrick O. Brown, and David Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.

[32] Marcel Ramos, Ludwig Geistlinger, Sehyun Oh, Lucas Schiffer, Rimsha Azhar, Hanish Kodali, Ino de Bruijn, Jianjiong Gao, Vincent J. Carey, Martin Morgan, and Levi Waldron. Multiomic integration of public oncology databases in bioconductor. *JCO Clinical Cancer Informatics*, 1(4):958–971, 2020. PMID: 33119407.

[33] Nimrod Rappoport and Ron Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids research*, 46(20):10546–10562, 2018.

[34] Piotr M. Suder, Jason Xu, and David B. Dunson. Bayesian transfer learning. *Statistical Science*, 40(3):477–494, 2025.

[35] Duncan C. Thomas, David V. Conti, James Baurley, Frederik Nijhout, Michael Reed, and Cornelia M. Ulrich. Use of pathway information in molecular epidemiology. *Human Genomics*, 4(1):21–42, 2009.

[36] Ye Tian, Haolei Weng, Lucy Xia, and Yang Feng. Robust unsupervised multi-task and transfer learning on gaussian mixture models. *arXiv preprint arXiv:2209.15224*, 2024. Version 4, August 2024.

[37] Sinead Wade. Bayesian cluster analysis. *Philosophical Transactions of the Royal Society A*, 381(2247):20220149, 2023.

[38] Rongrong Wang, Shiyuan Han, Jin Zhou, Yuehui Chen, Lin Wang, Tao Du, Ke Ji, Ya-ou Zhao, and Kun Zhang. Transfer-learning-based gaussian mixture model for distributed clustering. *IEEE Transactions on Cybernetics*, 53(11):7058–7070, 2023.

[39] Nathan Wycoff, Ali Arab, and Lisa O. Singh. Formal bayesian transfer learning via the total risk prior. *arXiv preprint arXiv:2507.23768*, 2025.

[40] Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, 2021.

[41] Shunchao Yan, Jiale Ji, Zhijie Zhang, Murshid Imam, Hong Chen, Duo Zhang, and Jinpeng Wang. Targeting the crosstalk between estrogen receptors and membrane growth factor receptors in breast cancer treatment: Advances and opportunities. *Biomedicine & Pharmacotherapy*, 175:116615, 2024.

[42] Yuqing Zhang, Giovanni Parmigiani, and W Evan Johnson. Combat-seq: batch effect adjustment for rna-seq count data. *NAR Genomics and Bioinformatics*, 2(3):lqaa078, September 2020.

[43] Yinqi Zhao, Qiran Jia, Jesse A. Goodrich, and David V. Conti. Lucidus: An r package for implementing latent unknown clustering by integrating multi-omics data (lucid) with phenotypic traits. *The R Journal*, 16(2):4–27, 2025.