Clustering of Omic Data Using Semi-Supervised Transfer Learning for Gaussian Mixture Models via Natural-Gradient Variational Inference: Method and Applications to Bulk and Single-Cell Transcriptomics

Qiran Jia
Jesse A. Goodrich
David V. Conti

Division of Biostatistics and Health Data Science,
Department of Population and Public Health Sciences,
Keck School of Medicine, University of Southern California

Abstract

Recent advances in high-throughput technologies have enabled observational studies to collect high-dimensional omic data. However, such data, often measured on small sample sizes, pose challenges to model-based clustering approaches such as Gaussian Mixture Models. Existing methods often fail to generalize due to model instability under complex mixture patterns. To overcome these limitations, we propose a natural-gradient variational inference framework for Gaussian mixture models named Praxis-BGM that incorporates informative priors—cluster-specific means, covariances, and structural connectivity—from large-scale reference data with known cluster or class labels to enable semi-supervised transfer learning. We derive natural-gradient updates that integrate prior knowledge, leveraging the Variational Online Newton algorithm. We also perform feature selection for clustering using Bayes Factors. Implemented using the JAX library for accelerator-oriented computation, Praxis-BGM is computationally efficient and scalable. We demonstrate the effectiveness of Praxis-BGM in extensive simulations and with two real-world applications: bulk transcriptomic datasets for breast cancer subtyping (the Cancer Genome Atlas Breast Invasive Carcinoma and the Molecular Taxonomy of Breast Cancer International Consortium), and transferring cell-type annotations between single-cell transcriptomic datasets produced by different single-cell RNA-seq technologies in a human pancreas study. Even when priors are partially mismatched with the target data, Praxis-BGM enhances semi-supervised clustering accuracy and biological interpretability.

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

1 Introduction

Omics data, such as genomics, metabolomics, proteomics, and transcriptomics, 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 or variations in disease risk (Mo et al. 2013). Unsupervised clustering methods, such as Gaussian Mixture Models (GMMs), infer such as underlying patterns in the data 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 downstream inference (Zhao et al. 2024, Mo et al. 2013, Rappoport & Shamir 2018). However, complex correlations, high dimensionality, and limited sample sizes (HDLSS) pose substantial challenges for statistical analysis (Rappoport & Shamir 2018).

Statistical transfer learning addresses these challenges by leveraging related source domains to improve the performance of modeling and inference on a target (Pan & Yang 2010). Common strategies include sample transfer, feature transformation into a shared latent space, parameter or coefficient sharing with debiasing, and relational transfer that borrows graph or structural information (Pan & Yang 2010, Zhuang et al. 2019). A prominent line of work develops transfer learning methods for high-dimensional generalized linear models (GLMs) (Li et al. 2022, Tian & Feng 2023). Recent advances have also developed transfer-learning frameworks for GMMs that jointly model related clustering tasks in a multi-task learning fashion (Tian et al. 2024) and distributed GMMs that share parameters rather than raw data across decentralized sites (Wang et al. 2023) to address the HDLSS challenge. While effective, these methods cannot exploit scenarios in which the source domain has already been characterized by curated unsupervised clusters that are annotated with interpretable class or group labels, as required for semi-supervised transfer learning (Zhuang et al. 2019). In this context, Bayesian modeling offers a natural vehicle for

knowledge transfer across data domains by encoding source-domain knowledge (and broader biological information) into informative priors (Suder et al. 2023). Unlike classification-oriented transfer learning approaches relying on predefined labels in the reference domain and making deterministic predictions on the target domain (Xu et al. 2021, Cai et al. 2025), Bayesian inference with GMMs fundamentally updates the prior model and cluster inclusion probabilities to encourage the refinement of latent structures after observing the target domain. These considerations motivate a scalable semi-supervised Bayesian GMM that transfers an interpretable and predefined cluster structure from the source to the target data.

Building on the recent advancements in natural gradient variational inference (NGVI) for the mixture of multivariate Gaussians (Khan & Nielsen 2018, Lin et al. 2019), we propose a method called Praxis-BGM, which is a semi-supervised Bayesian transfer learning method for GMMs, offering an alternative to multi-task (i.e., unsupervised) transfer learning frameworks. In this setting, the reference domain is assumed to have known curated cluster labels and ideally to be annotated for biological interpretation (e.g., cancer subtypes). The target domain remains unlabeled. Such a situation may be applicable when using large biobank data as the reference for a smaller, more unique target study. To the best of our knowledge, no existing transfer-learning approaches for Gaussian mixture models (GMMs) have been integrated with natural-gradient variational inference (NGVI), nor has a formal framework been proposed for semi-supervised transfer learning in GMMs under labeled reference datasets. This provides a principled transfer mechanism that complements existing GLM-based strategies (Li et al. 2022). We further implemented a Bayes-factor–based feature selection procedure to identify features that are most responsible for cluster separation, combining evidence from both the source and the target datasets. We applied the method to two transcriptomic datasets of breast cancer subtyping, using the subtype-specific structure

in the reference dataset to inform clustering in the target dataset and thereby enable subtype interpretation of the target clusters. 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 evaluation of Praxis-BGM's performance.

The rest of the paper is organized as follows. Section 2 introduces the proposed method by explaining its model formulation and variational inference framework. Section 3 focuses on simulations to evaluate the contribution of each prior component, while Sections 4 and 5 apply the method to two real data applications and benchmark its performance against existing approaches. Finally, Section 6 concludes with a discussion of our findings and directions for future research.

2 Methods

2.1 Variational Bayesian Transfer of Gaussian Mixture Models for Semi-supervised Learning

Let $\mathcal{D} = \{D_n : n = 1, \ldots, N\}$ with $D_n \in \mathbb{R}^P$ denote the observed (target) mixture data with P features. Let $z \in \mathbb{R}^P$ denote a continuous latent variable representing the feature space of an observation drawn from the GMM. From a Bayesian perspective, $p(z \mid \mathcal{D})$ represents the posterior distribution of z after observing the dataset $\mathcal{D}$, which is defined as,

$$p(z \mid \mathcal{D}) \;=\; \sum_{j=1}^{J} \theta_j \, \mathcal{N}_P\Big(\mu_j,\, \Sigma_j\Big) \tag{1}$$

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

We can leverage a source dataset $\mathcal{D}_0$, defined in the same feature-space as $\mathcal{D}$ but sampled from a related population, for Bayesian transfer learning (Suder et al. 2023). By Bayes rule, we can easily obtain

$$p(z \mid \mathcal{D}, \mathcal{D}_0) \;=\; \frac{p(\mathcal{D} \mid z, \mathcal{D}_0)\, p(z \mid \mathcal{D}_0)}{p(\mathcal{D} \mid \mathcal{D}_0)}. \tag{2}$$

The denominator, $p(\mathcal{D} \mid \mathcal{D}_0) = \int p(\mathcal{D} \mid z', \mathcal{D}_0)\, p(z' \mid \mathcal{D}_0)\, dz'$, is intractable, so it is impossible to compute the exact posterior. Thus, we first estimate $p(z \mid \mathcal{D}_0)$ and then employ the variational inference framework introduced in Appendix A.1 to approximate the posterior $p(z \mid \mathcal{D}, \mathcal{D}_0)$ given $p(z \mid \mathcal{D}_0)$. To summarize, we place a prior on z, parameterized by informative $\mu_{0j}, \Sigma_{0j}, \theta_{oj}$, representing the entire latent distribution of the GMM learned from the source data $\mathcal{D}_0$, and then infer the posterior of z after observing the target data $\mathcal{D}$.

When the source dataset $\mathcal{D}_0$ includes cluster labels $\mathbf{X}_0$, we can directly use cluster-specific summary statistics to derive $\mu_{0j}, \Sigma_{0j}, \theta_{oj}$. In the absence of labels, a Bayesian GMM with a Normal-Inverse-Wishart (NIW) prior can be used to estimate the mixture distribution. The priors derived from the source data $\mathcal{D}_0$, obtained either way, can be represented as

$$p(z \mid \mathcal{D}_0) \;=\; \sum_{j=1}^{J} \theta_{0j}\, \mathcal{N}_P\Big(\mu_{0j},\, \Sigma_{0j}^{(A)}\Big), \tag{3}$$

where $\mu_{0j} \in \mathbb{R}^P$ and $\Sigma_{0j} \in \mathbb{S}_{++}^P$ are the source–domain mean and covariance for the multivariate Gaussian component j, and θ_{0j} are the corresponding prior weights. The prior cluster-specific variance-covariance matrices in (3) are adjusted by additional prior knowledge to encourage sparsity as

$$\Sigma_{0j}^{(A)} \;=\; A_j \odot \Sigma_{0j}, \tag{4}$$

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 for genes or metabolites. The matrix A_j not only shapes the prior covariance via a Hadamard mask as

shown in (4), but also determines the number of free covariance parameters in posterior estimation.

2.2 Prior-Augmented Natural–Gradient Variational Inference

Given informative priors derived from the source data $p(z \mid \mathcal{D}_0)$ defined in Equation (3) , referred to as $p_\omega(z)$ hereafter for simplicity, we approximate the posterior $p(z \mid \mathcal{D}, \mathcal{D}_0)$ with a surrogate distribution $q_\phi(z)$ using the natural–gradient variational inference framework (Khan & Nielsen 2018, Lin et al. 2019). Leveraging the variational online Newton (VON) algorithm, we extended it with component–specific informative priors. Given priors $p_\omega(z)$ and target data $\mathcal{D}$, we seek a variational distribution that maximizes the ELBO,

$$\mathcal{L}(q_\phi) = \mathbb{E}_{q_\phi(z)}\Big[\sum_{i=1}^{N} \log p(D_i \mid z)\Big] - D_{\mathrm{KL}}\big(q_\phi(z) \,\|\, p_\omega(z)\big). \tag{5}$$

Equivalently, we may express $\mathcal{L}(q_\phi) = \mathbb{E}_{q_\phi(z|\mathcal{D})}\Big[-\ h(z)\Big]$, so $h(z) = -\sum_{n=1}^{N} \log p(D_n \mid z) + \log \frac{q_\phi(z|\mathcal{D})}{p_\omega(z)}$. Under the expectation of the variational distribution q_ϕ, $h(z)$ is the loss function to be minimized. VON uses stochastic samples to estimate the expectation under $q_\phi(z)$ for natural-gradient updates. At iteration t, we first draw a component and then draw a sample conditioned on the component,

$$v \ \sim\ \mathrm{Cat}(\boldsymbol{\theta}^{(t)}), z_0 \ \sim\ q_\phi^{(t)}(z \mid j = v), \tag{6}$$

which serves as the anchor point for natural-gradient updates. With variational $q_\phi(z) = \sum_{j=1}^{J} \theta_j \mathcal{N}(z \mid \mu_j, \Sigma_j)$ and with an informative mixture prior, $p_\omega(z) = \sum_{k=1}^{J} \theta_{0k} \mathcal{N}(z \mid \mu_{0k}, \Sigma_{0k})$, define

$$h(z_0) = \log\left(\sum_{j=1}^{J} \theta_j \mathcal{N}(z_0 \mid \mu_j, \Sigma_j)\right) - \log\left(\sum_{k=1}^{J} \theta_{0k} \mathcal{N}(z_0 \mid \mu_{0k}, \Sigma_{0k})\right) - \sum_{n=1}^{N} \log p(D_n \mid z_0). \tag{7}$$

The stochastic gradient and Hessian of $h(z_0)$ w.r.t. z_0 are denoted as $\nabla_z h(z_0)$ and $\nabla_z^2 h(z_0)$, and we use $\boldsymbol{g}(z_0)$ and $\boldsymbol{H}(z_0)$ to denote the gradient and Hessian contributions from the

data log-likelihood evaluated at z_0. Incorporating these into the closed-form VON step of Lin et al. (2019):

$$
\begin{aligned}
\boldsymbol{\Sigma}_j^{(t+1)\,-1} &= \boldsymbol{\Sigma}_j^{(t)\,-1} \;+\; \alpha\,\delta_j(z_0)\,\nabla_z^2 h(z_0) \\
&= \boldsymbol{\Sigma}_j^{(t)\,-1} \;+\; \alpha\,\delta_j(z_0)\Big(\boldsymbol{H}(z_0) + \boldsymbol{\Sigma}_{0j}^{-1} - \boldsymbol{\Sigma}_j^{(t)\,-1}\Big)\,, \\
\boldsymbol{\mu}_j^{(t+1)} &= \boldsymbol{\mu}_j^{(t)} \;-\; \alpha\,\delta_j(z_0)\,\boldsymbol{\Sigma}_j^{(t+1)}\,\nabla_z h(z_0) \\
&= \boldsymbol{\mu}_j^{(t)} \;-\; \alpha\,\delta_j(z_0)\,\boldsymbol{\Sigma}_j^{(t+1)}\Big(\boldsymbol{g}(z_0) + \boldsymbol{\Lambda}_{0j}\Big(\boldsymbol{\mu}_j^{(t)} - \boldsymbol{\mu}_{0j}\Big)\Big), \\
\log\frac{\theta_j^{(t+1)}}{\theta_J^{(t+1)}} &= \log\frac{\theta_j^{(t)}}{\theta_J^{(t)}} \;-\; \alpha\,\Big(\delta_j(z_0) - \delta_J(z_0)\Big)\,h(z_0), \qquad j = 1,\ldots,J-1,
\end{aligned}
\tag{8}
$$

where

$$
\delta_j(z_0) = \frac{\mathcal{N}\Big(z_0 \mid \mu_j^{(t)}, \Sigma_j^{(t)}\Big)}{\sum_{\ell=1}^{J} \theta_\ell^{(t)}\,\mathcal{N}\Big(z_0 \mid \mu_\ell^{(t)}, \Sigma_\ell^{(t)}\Big)}. \tag{9}
$$

With a single anchor sample z_0 drawn, the stochastic VON updates are unbiased for updating ϕ under z by Bonnet's and Price's theorems (Opper & Archambeau 2009). We also show that the update of $\boldsymbol{\Sigma}_j^{(t+1)\,-1}$ could be performed in the Cholesky space in Appendix A.2. We summarize Praxis-BGM (prior-augmented NGVI for GMMs) in Algorithm 1. Praxis-BGM is implemented in JAX to exploit accelerator-friendly automatic differentiation and just-in-time compilation for scalable and numerically stable NGVI updates (Bradbury et al. 2018). In practice, the gradient and Hessian updates can also be computed in mini-batches within each iteration to further scale to large data.

2.3 Clustering-Driven Feature Selection via Bayes Factors

To evaluate the evidence of which individual feature p contributes to clustering via the mixture, we compare its marginal posterior Gaussian distribution to an empirical null. Let $\mathcal{D}_p = \{D_{ip}\}_{i=1}^N$ denote the observed values of feature p. The empirical null is the univariate Gaussian representing no mixture structure. We assume $\mathcal{D}_p$ is standardized so that $\hat{\mu}_p = 0$ and $\hat{\sigma}_p^2 = 1$.

Algorithm 1 Praxis-BGM: Prior-augmented NGVI for Bayesian Gaussian Mixture Model

Data: Target data $\mathcal{D} = \{D_n\}_{n=1}^{N}$; source data $\mathcal{D}_0$ with labels $\boldsymbol{X}_0$ define parameters $\omega_0 = \Big(\boldsymbol{\theta}_0, \{\mu_{0j}, \Sigma_{0j}\}_{j=1}^{J}\Big)$ for the prior $p_\omega(z)$, with elements optionally set to `None`; structural sparsity $\{A_j\}$.

Input: Stepsize α, epochs T, covariance scaling factor $c > 0$ to encode prior belief, jitter $\tau = 10^{-9}$.

Working prior parameter $\omega_0^\star$ construction: For each component j:

$$\mu_{0j}^\star \leftarrow \begin{cases} \mu_{0j}, & \text{if } \mu_{0j} \text{ given,} \\ \mathbf{0}, & \text{otherwise,} \end{cases} \qquad \Sigma_{0j}^\star \leftarrow \begin{cases} \Big(A_j \odot \Sigma_{0j}\Big)c + \tau I_P, & \text{if } \Sigma_{0j} \text{ given,} \\ c\, I, & \text{otherwise,} \end{cases}$$

and $\boldsymbol{\theta}_0^\star \leftarrow \boldsymbol{\theta}_0$ if given, else $J^{-1}\mathbf{1}_J$.

Initialization of variational ϕ: *If μ_{0j} is provided:* $\mu_j \leftarrow \mu_{0j}^\star$, $\Sigma_j \leftarrow \Sigma_{0j}^\star$, $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0^\star$, $\rho \leftarrow \log(\theta/\theta_J)$; *else:* fit `BayesiangGaussianMixture` model with full covariance type on $\mathcal{D}$, set $\mu_j, \Sigma_j, \boldsymbol{\theta} \leftarrow J^{-1}\mathbf{1}$, and ρ based on posterior results.

for $t = 0, \ldots, T-1$ **do**

  Sample an anchor z_0 from the current ϕ via Eq. (6) and evaluate $\delta_j(z_0)$ via Eq. (9), $h(z_0)$ via Eq. (7).

  for $j = 1, \ldots, J$ **do**

    Following Eq. (8), update Σ_j^{-1}, μ_j, and ρ_j.

  end

  $\boldsymbol{\theta} \leftarrow \mathrm{softmax}\Big([\rho_1, \ldots, \rho_{J-1}, 0]\Big)$

end

Result: Parameters $\phi = (\boldsymbol{\theta}, \{\mu_j, \Sigma_j\}_{j=1}^{J})$ for the posterior $p_\phi(z)$, yielding cluster inclusion probabilities $\gamma_{nj} = \Pr(z_n = j \mid D_n, \phi)$, and feature-specific Bayes factors BF_p as in (10) for clustering-driven feature selection.

We use Bayes factor (BF) for this comparison. For feature p, the posterior mixture

model assigns cluster responsibilities $\theta_j \in [0,1]$ with $\sum_{j=1}^{J} \theta_j = 1$, and univariate posterior parameters μ_{jp} and $\sigma_{jp}^2 = \Sigma_{j,pp}$. Let $\ell(x \mid m, v)$ denote the univariate Gaussian log-likelihood for value x with mean m and variance v. The BF is defined as

$$\mathrm{BF}_p = \exp\left\{ \sum_{i=1}^{N} \log\left[\frac{\sum_{j=1}^{J} \theta_j \, \exp\{\ell(D_{ip} \mid \mu_{jp}, \sigma_{jp}^2)\}}{\exp\{\ell(D_{ip} \mid \hat{\mu}_p, \hat{\sigma}_p^2)\}} \right] \right\}. \tag{10}$$

Interpretation of BF_p follows the scale of Jeffreys (1961), with categories later popularized by Kass & Raftery (1995).

3 Simulation

To evaluate the effectiveness of leveraging informative priors by the proposed method under varying conditions, we generated data from a Gaussian mixture model with J = 4 components, where each component has an identity covariance matrix unless otherwise specified. Let $\mathcal{C} \subset \{1, \ldots, P\}$ denote the set of causal features and $\mathcal{N} = \{1, \ldots, P\} \setminus \mathcal{C}$ the non-causal features. For each $p \in \mathcal{C}$ and cluster j, $\mu_{j,p} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \nu)$; for $p \in \mathcal{N}$, $\mu_{j,p} = 0$ for all j. Unless otherwise stated, we set $\nu = 0.09$ for a low signal-to-noise (SNR) scenario, $\theta_j = 1/J$, $\mathbf{\Sigma}_j = \mathbf{I}_P \in \mathbb{R}^{P \times P}$, $P = 100$, $\mathcal{C} = 40$, and sample size $N = 100$ to generate the target data. Once the ground-truth GMM parameters for the target data are sampled, we perturbed these parameters to generate the source data with a sample size of $N_0 = 1000$, which are then used to derive the informative priors for Praxis-BGM and to train the benchmark classification-based methods.

Across four components of informative priors, we compared eight methods: (1) Linear discriminant analysis (LDA); (2) multinomial regression; (3) Extreme Gradient Boosting (XGBoost), a popular machine learning classification method that leverage a large collection of weak decision trees (Chen & Guestrin 2016); (4) K-Nearest Neighbors (KNN) classifier, another machine learning classification method that predicts based on the majority class of a

data point's finite nearest neighbors; and (5) Praxis-BGM with an informative prior derived from the source data. Additionally, we benchmarked these approaches with unsupervised clustering methods that do not use prior information: (6) Bayesian Gaussian Mixture Models (BGM), a VI-based GMM method using Normal-Invert-Wishart Prior without prior information (fit with J fixed) ; (7) the Leiden algorithm (Traag et al. 2019), a graph-based community detection method applied to a k-nearest neighbor graph, which does not require J; and (8) Praxis-BGM with no informative prior. We applied LDA, multinomial regression, KNN, and BGM implemented in *scikit-learn* with default settings (Pedregosa et al. 2011). Performance was quantified using the Adjusted Rand Index (ARI), which ranges from -1 to 1 and measures concordance between estimated clusters and the simulated truth. Each scenario was replicated 500 times.

3.1 Cluster-specific Means Prior

We extensively evaluated the effect of incorporating informative priors for cluster means derived from varying source data misspecifications under the isotropic mixture of multivariate Gaussians. The mean parameter to generate the source data $\boldsymbol{\mu}_j^{(0)} \in \mathbb{R}^P$ were perturbed by adding Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.3^2)$ with varying proportion $\rho \in \{0\%, 10\%, 30\%, 50\%, 70\%, 100\%\}$ of features. After generating the source datasets and labels for each ρ level, we computed cluster-specific means and used them as priors for Praxis-BGM to transfer knowledge to the target; we also trained classification methods on the source data and applied them to the target for prediction.

Figures 1(a) illustrate clustering performance under different levels of noise or perturbation in the source data informing cluster-specific mean priors. The x-axis groups represent increasing levels of perturbation ρ, corresponding to greater divergence from the target data distribution, along with unsupervised baseline methods. Praxis-BGM with mean priors

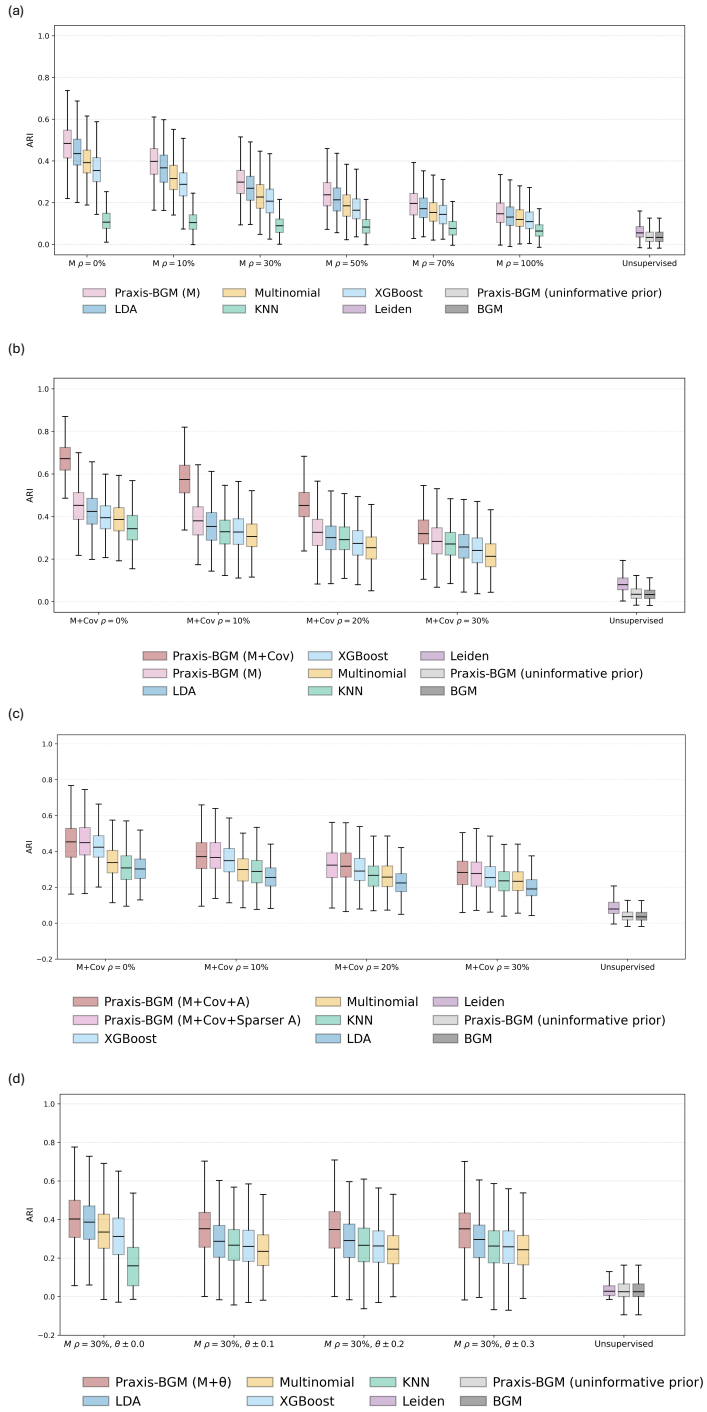

Figure 1: Simulation results for evaluating the effectiveness of the four types of priors for Praxis-BGM on clustering performance under different scenarios of source-data misspecification. The y-axis shows the ARI (agreement with ground-truth clusters), while the x-axis represents increasing levels of misspecification and baseline methods. **(a)** Informative priors on cluster means; **(b)** Informative priors on means and covariance structures; **(c)** Priors on the sparsity-imposing matrix; **(d)** Priors on cluster weights.

has the best clustering accuracy as measured by ARI when the source data is unperturbed (Praxis-BGM (M), M $\rho = 0\%$). LDA is the second-best method in the unperturbed source

data. As ρ increases, we observed a decreasing trend in ARI for Praxis-BGM and other classification methods. However, Praxis-BGM still achieves a better ARI performance compared to competing methods within the same ρ group, demonstrating the ability of Praxis-BGM to cluster in the presence of heterogeneity across data domains through the capturing of uncertainty within the variational Bayes framework. All methods leveraging source data outperform unsupervised alternatives (Leiden, BGM, and Praxis-BGM with an uninformative prior) even under heavily perturbed source data (M $\rho = 100\%$). This phenomenon can be attributed to the fact that, even when the distributions of all features are perturbed, stochastic variation sampled from a Gaussian distribution with mean 0 may still preserve fragments of the original cluster structure. Praxis-BGM with an uninformative prior performs comparably to BGM, but both fall short of the performance achieved by Leiden, reflecting the advantage of a graph-based clustering method for low SNR mixture data. Overall, these results suggest that priors or information derived from the source data substantially improve clustering performance for hard-to-cluster mixture data when source data and the target data are similar or moderately distorted, and Praxis-BGM works best for transferring such knowledge.

3.2 Cluster-specific Variance-Covariance Prior

We next evaluated the impact of incorporating priors for both cluster means and variance-covariance matrices. Let $\mathbf{I}_P \in \mathbb{R}^{P\times P}$ denote the identity matrix and $B_j \subset \{1,\ldots,P\}$ be an index set corresponding to a contiguous block of features of size $\lfloor 0.1P \rfloor$, chosen independently for each cluster j. Define the block-indicator matrix $\mathbf{E}_j \in \{0,1\}^{P\times P}$ as $(\mathbf{E}j)uv = 1$ if $u \in B_j$ and $v \in B_j$, and 0 otherwise. The ground-truth covariance for the cluster j is then $\mathbf{\Sigma}_j = \mathbf{I}_P + 0.3\,\mathbf{E}_j$. This construction yields covariance patterns that differ across clusters. Similar to the perturbations applied to $\boldsymbol{\mu}_j^{(0)}$ in Section 3.1 to generate the source data, we vary the perturbation level as $\rho \in \{0\%, 10\%, 30\%\}$ for both $\boldsymbol{\mu}_j^{(0)}$ and $\mathbf{\Sigma}_j^{(0)}$.

After generating the source data and its labels, we extracted cluster-specific means and covariances as priors for Praxis-BGM and trained classification models on the source for use on the target data.

Figure 1(b) shows clustering performance when informative priors of varying qualities derived from the corresponding source data were specified either for the cluster mean prior $\boldsymbol{\mu}^{(0)}$ alone or jointly for the cluster means prior $\boldsymbol{\mu}^{(0)}$ and covariance matrix prior $\boldsymbol{\Sigma}^{(0)}$, in the setting where the ground-truth clusters exhibit distinct covariance structures. While Praxis-BGM with accurate mean priors $\boldsymbol{\mu}^{(0)}$ (Praxis-BGM (M), M + Cov $\rho = 0\%$) already achieves a high ARI, incorporating accurate covariance priors $\boldsymbol{\Sigma}^{(0)}$ (Praxis-BGM (M+Cov)) further increases the ARI, confirming the complementary value of cluster-specific variance-covariance priors across all ρ groups. As ρ increases, ARI decreases for all methods that learn from the source data, while Praxis-BGM (M+Cov) consistently performs the best within the same source data group. However, variance–covariance priors exhibit lower robustness compared to the mean prior as the performance gap between Praxis-BGM (M+Cov) and Praxis-BGM (M) narrows as ρ increases. In summary, when clustering is influenced not only by differences in cluster means but also by variance–covariance structures, incorporating additional prior information in $\boldsymbol{\Sigma}^{(0)}$ can further improve clustering performance.

3.3 Sparsity–imposing matrix Prior

We evaluated the impact of incorporating a sparse feature–feature connectivity prior, denoted by $\mathbf{A} \in \{0,1\}^{P \times P}$, into the proposed Praxis-BGM model in addition to priors on the cluster-specific means and covariance matrices. $\mathbf{A}$ encodes prior knowledge about independence between features, with $A_{uv} = 0$ indicating that features u and v are assumed *a priori* uncorrelated. We simulated $\mathbf{A}$ by randomly choosing $\left\lfloor \alpha \dfrac{P(P-1)}{2} \right\rfloor$ pairs from the upper-triangular indices (u, v), where $1 \leq u < v \leq P$, and setting those off-diagonal entries

to one, with $\alpha = 0.05$.

The mask entries are then defined by $A_{uv} = 1$ if $u = v$; $A_{uv} = 1$ if $(u, v) \in \mathcal{S}$ or $(v, u) \in \mathcal{S}$; and $A_{uv} = 0$ otherwise. Similar to Section 3.2, we simulated four sets of source data with increasing ρ representing a decrease in the quality of the source data. The simulated $\mathbf{A}$ was applied elementwise to the sampled cluster covariance matrices to produce masked covariance matrices $\mathbf{\Sigma}_j^{(0)} = \mathbf{\Sigma}_j^{0*} \odot \mathbf{A}$ for target data generation. In contrast, the source data were generated directly from the unadjusted $\mathbf{\Sigma}_j^{0*}$. This allows for additional independence information from the target data to only be specified through the $\mathbf{A}$ prior for Praxis-BGM. Besides the benchmarks and Praxis-BGM analysis with both mean $\boldsymbol{\mu}_j^{(0)}$ and covariance priors $\mathbf{\Sigma}_j^{(0)}$, we further evaluated scenarios in which the sparse connectivity mask $\mathbf{A}$ was incorporated into the mean and covariance priors, using either the exact $\mathbf{A}$ or a modified sparser $\mathbf{A}'$ in which half of the off–diagonal 1s were randomly set to zero to induce greater sparsity. The latter setting represents potential inaccuracies in the prior specification of $\mathbf{A}$, a scenario commonly encountered in real data.

Figure 1 (c) demonstrates the added effect of $\mathbf{A}$ alongside $\boldsymbol{\mu}^{(0)}$ and $\mathbf{\Sigma}^{(0)}$. As expected, using the true $\mathbf{A}$ in combination with accurate $\boldsymbol{\mu}^{(0)}$ and $\mathbf{\Sigma}^{(0)}$ yields the highest ARI (Praxis-BGM (M+Cov+A), M + Cov $\rho = 0\%$). When $\mathbf{A}$ was replaced by its sparser variant $\mathbf{A}'$ (Praxis-BGM (M+Cov+Sparser A)), the ARI is on par with using the true $\mathbf{A}$, suggesting the method is robust to moderate misspecification of sparser connectivity patterns. Classification-based approaches are inferior to Praxis-BGM analyses incorporating informative priors from the same source data, with unsupervised methods performing the poorest overall. When the source data lacks additional independence information and instead contains noisy, misleading correlation structures as ρ increases, we did not observe an obvious decrease in clustering performance. This is likely due to $\boldsymbol{\mu}^{(0)}$ and $\mathbf{\Sigma}^{(0)}$ from the source data becoming less informative. The results highlight that incorporating a

well-specified $\mathbf{A}$ can substantially enhance clustering accuracy, particularly when the raw covariance prior includes superfluous structure that does not reflect the observed data.

3.4 Cluster Weight Prior

This simulation evaluates the effectiveness of cluster weight prior, $\boldsymbol{\theta}^{(0)}$. Under the same isotropic setting as in Simulation 3.1, we fixed the source data means at $\rho = 0.3$ and generated imbalanced cluster weights for the target data by sampling $\boldsymbol{\theta} \sim \mathrm{Dir}(0.4)$. We perturbed the true weights with uniform noise sampled from $[-0.1, 0.1]$, $[-0.2, 0.2]$, and $[-0.3, 0.3]$ and applied a softmax transformation to obtain the source data weights.

Figure 1 (d) reports results for the informative weight prior scenario along with moderately misspecified mean priors. Praxis-BGM with mean and true weight priors (Praxis-BGM (M+θ, M $\rho = 30\%$, $\theta \pm 0.0$) achieves the highest ARI, outperforming other classification methods. Across different source datasets, the perturbation level of the mean prior $\boldsymbol{\mu}_j^{(0)}$ remained fixed at $\rho = 0.3$ while the perturbation level of the weights $\boldsymbol{\theta}^{(0)}$ increased, representing increasingly inaccurate weight priors. Here, we did not observe a substantial decrease in ARI performance for either Praxis-BGM or the classification-based methods, indicating that while correctly specified weight priors can improve performance, moderately misspecified prior weights do not significantly impact performance.

Additional simulations were performed to evaluate Praxis-BGM under different SNR settings and various forms of source–target misalignment for prior misspecification. These are presented in Appendix A.3.1 and A.3.2. Furthermore, we assessed the proposed cluster-driving feature selection procedure in terms of Bayes factor–based performance and computational efficiency in Appendices A.3.3 and A.3.4.

4 Application: Semi-Supervised Transfer Learning Between Two Breast Cancer Transcriptomic Datasets

The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) are two breast cancer omics studies (Network 2012, Curtis et al. 2012). For its breast cancer samples, TCGA-BRCA includes detailed clinical information, survival outcomes, and RNA sequencing data measured via Agilent Custom 244 K, covering around 20k genes at isoform-level resolution. In addition, for a subset of 522 subjects, PAM50 intrinsic subtypes (Basal-like, HER2-enriched, Luminal A, Luminal B (HER2+), Normal-like) are characterized by gene-expression data (Network 2012). METABRIC includes a cohort of breast cancer patients profiled with transcriptomics using the Illumina HumanHT-12 v3 BeadChip microarray, also quantifying approximately 20k genes (Curtis et al. 2012). The study has long-term follow-up for survival data and clinical hormone receptor status. Although METABRIC does not provide official PAM50 subtype labels, clinical hormone receptor status offers a reliable surrogate, as it is strongly correlated with molecular subtypes (Parker et al. 2009, Sørlie et al. 2003, Kreike et al. 2007). In particular, the triple-negative group is widely accepted as a clinical proxy for the Basal-like subtype, with 80–90% of triple-negative tumors classified as Basal-like based on gene expression profiles (Kreike et al. 2007). Accordingly, we assigned METABRIC subjects a subtype label based on this criterion, and Figure 2(a) visualizes the prevalence of breast cancer subtypes for both datasets.

For our analysis, METABRIC serves as the labeled source data - we derive subtype-specific priors from clinical hormone-receptor status and incorporate them as priors for Praxis-BGM using the TCGA-BRCA data as the target dataset. We hypothesized that incorporating these priors would yield TCGA-BRCA clusters with greater concordance to PAM50 subtypes and stronger prognostic value than clustering on observed data alone or classification models

trained on the source and applied to predict directly to the target data.

4.0.0.1 Analysis Method After removing subjects with complete missing transcriptomics or clinical data, we ended up with 1,980 METABRIC subjects and 1,060 TCGA-BRCA subjects. Using the set of genes common to both datasets, we identified the top 2,000 most variable genes based on the reference data and subsetted both datasets with these 2,000 genes. For METABRIC, we approximated PAM50 subtype labels by mapping standardized clinical hormone receptor status to corresponding subtypes. Specifically, ER+/HER2 cases were approximated as Luminal A, ER+/HER2+ as Luminal B (HER2+), ER/PR/HER2+ as HER2-enriched, and ER/PR/HER2 as Triple-negative to approximate basal-like. For TCGA-BRCA, we used the available PAM50 subtype labels from the metadata as ground-truth labels to validate the performance of Praxis-BGM; however, 538 subjects (50.8%) are missing this annotation.

To construct informative subtype–specific priors from METABRIC, for each of the four available subtypes, estimated the subtype-specific empirical mean and variance-covariance matrices across the top 2,000 most variable genes. These empirical estimates served as prior means and covariances for the corresponding components of the Praxis–BGM model. We set the learning rate parameter β to 10^{-4} and used default settings for all other hyperparameters in the Praxis-BGM model. The clusters estimated by Praxis-BGM were compared against baseline methods, including BGM, GMM, and predictive classification models: multinomial regression, random forest, and XGBoost trained on source data. The supervised classification models for multinomial regression and XGBoost were first tuned for optimal hyperparameters using cross-validation. To assess the effect of prior information on the clinical relevance of clusters identified by Praxis-BGM, we evaluated the association between the estimated clusters and PAM50 subtypes via the Chi-square test in the TCGA-BRCA dataset and performed survival analysis using a Cox proportional hazards model, incorporating the

estimated clusters as predictors while adjusting for age. We further leveraged the feature selection functionality of Praxis-BGM to compute Bayes factors for all genes to identify genes driving the clustering for subtypes in breast cancer. For the set of identified genes, we performed Gene Ontology (GO) pathway overrepresentation analysis using the Gene Set Enrichment Analysis (GSEA) approach implemented in the `GSEApy` Python package (Fang et al. 2022).

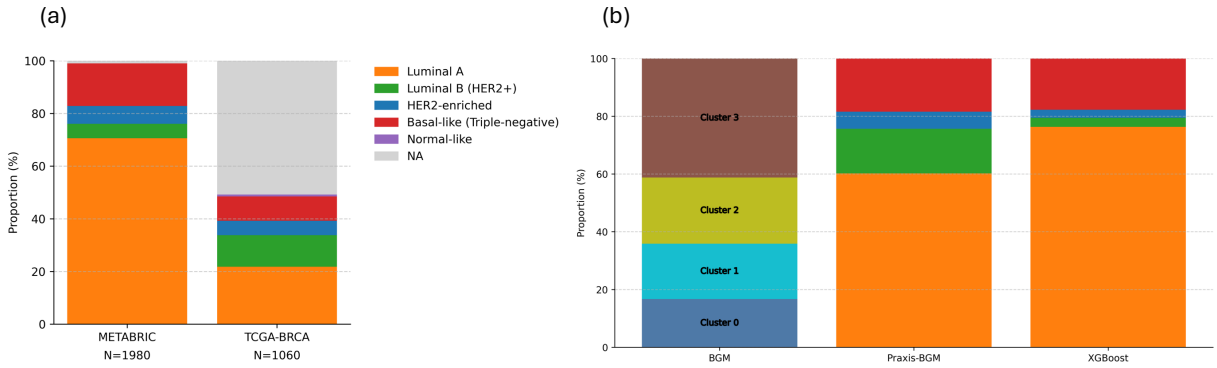

Figure 2: Subtype distributions of the source and target datasets, along with model-based estimated clusters(subtypes) for the target data. Panel (a) shows the annotated subtypes served as ground-truth for METABRIC and TCGA-BRCA, respectively. "NA" indicates that ground-truth subtype annotation was unavailable for those samples. Panel (b) shows the label distributions estimated by BGM, Praxis–BGM, and XGBoost for all subjects in TCGA-BRCA. Praxis–BGM and XGBoost map estimated cluster/class to subtypes, whereas BGM does not and is therefore denoted as clusters 1–4.

4.0.0.2 Results Table 1 presents the results (χ^2, p-value, and ARI) of benchmarking clustering and classification methods on TCGA RNA-seq data using subtype association. GMM and BGM produced identical clustering solutions, each showing a strong association with PAM50 molecular subtypes among the 522 TCGA-BRCA subjects with available labels ($\chi^2 = 441.50$, P-value $= 6.03 \times 10^{-87}$, ARI $= 0.239$). Thus, even in the absence of subtype-specific priors, the intrinsic transcriptomic structure of TCGA-BRCA samples strongly

indicates molecular subtypes. Similarly, supervised classification models of multinomial regression, random forest, and XGBoost showed significant association with the ground-truth subtypes: multinomial regression ($\chi^2 = 499.84$, P-value $= 2.39 \times 10^{-99}$, ARI $= 0.298$) and XGBoost ($\chi^2 = 595.17$, P-value $= 1.14 \times 10^{-119}$, ARI $= 0.367$). Compared with unsupervised clustering methods, multinomial regression and XGBoost achieved stronger concordance, which suggests that reference data can improve subtype resolution beyond clustering on observed data alone. Finally, by incorporating cell-type-specific mean priors derived from METABRIC, the clusters estimated by Praxis-BGM showed substantially higher concordance with PAM50 subtypes ($\chi^2 = 747.34$, P-value $= 3.20 \times 10^{-152}$, ARI $=$ 0.471).

Panel (a) Figure 3 visualizes the cluster/class distribution in the Principal Component (PC) space for labels predicted by BGM, XGBoost, and Praxis-BGM. In the PCA embedding, BGM clusters distribute across both PC1 and PC2 and reflect unconstrained use of global variance. By contrast, Praxis-BGM shows separation almost exclusively along PC2, which illustrates how the semi-supervised priors steer component means toward subtype-consistent structure and suppress variation along unrelated axes. For XGBoost, possibly due to domain shift, only a small number of subjects were predicted to be Luminal B (HER2+) and HER2-enriched, but separation is also roughly along PC2.

We compared three Cox proportional hazards models to assess the association with survival for clusters estimated by Praxis-BGM, baseline BGM, and XGBoost, respectively. As shown in Panel (b) Figure 3, no individual cluster reached statistical significance in the BGM model and the XGBoost model. However, Praxis-BGM had a significant association for the Luminal B (HER2+) cluster defined using prior information with an increased hazard ratio compared to Luminal A (HR $= 2.36$, P-value $= 1.34 \times 10^{-5}$). In addition, compared with Luminal A, the HER2-enriched and triple-negative clusters estimated by Praxis-BGM both

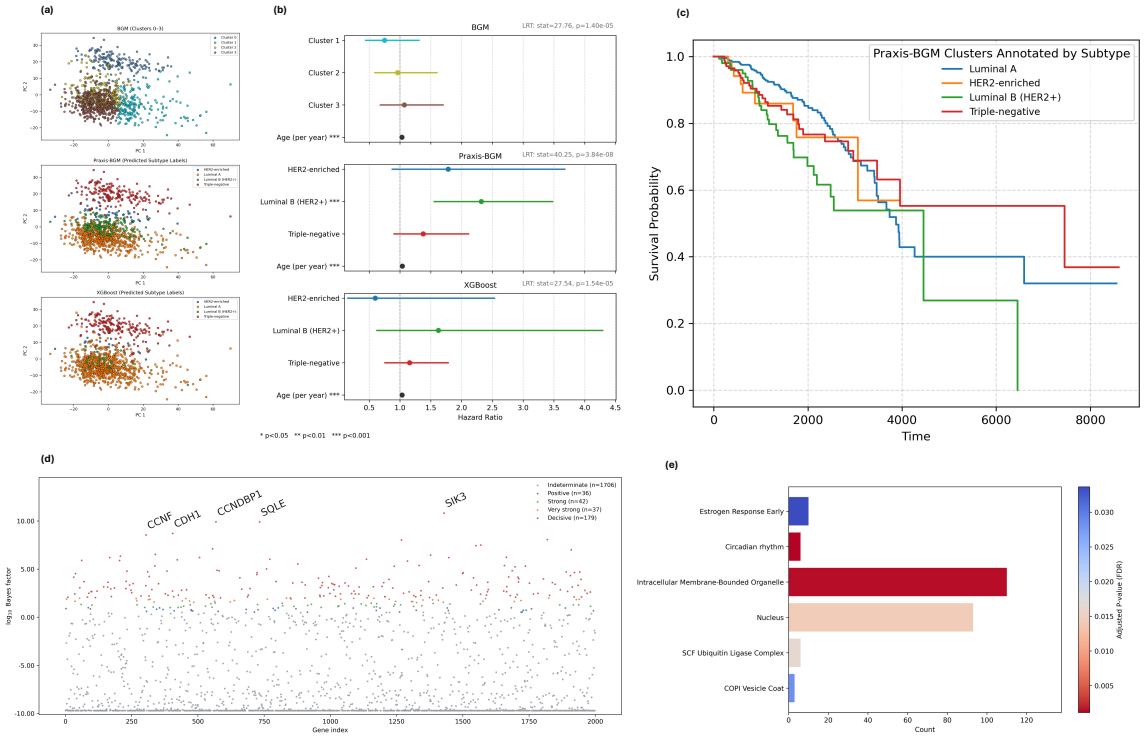

Figure 3: (a) PC embeddings colored by predicted cluster/label assignments from BGM, Praxis-BGM, and XGBoost; (b) hazard ratio estimates from three separate Cox models, each using cluster/label assignments from one method (BGM, Praxis-BGM, or XGBoost) as the predictor, adjusted for age; (c) Kaplan Meier curve comparing survival across Praxis-BGM clusters annotated by subtypes; (d) Bayes factor distribution across genes, with the top five genes annotated. Jeffreys' decision framework was applied to categorize each gene based on its Bayes factor; and (e) Pathway enrichment analysis of the non-intermediate genes identified in panel (d).

Table 1: Benchmarking clustering methods on TCGA-BRCA RNA-seq data by assessing the association between estimated clusters and ground-truth subtypes using the Chi-square test, restricted to the subset of samples with available subtype annotations.

Method	χ^2 **Stat**	**P-Value**	**ARI**	**Time (s)**	**Hardware**
Praxis-BGM	**747.34**	$\mathbf{3.20 \times 10^{-152}}$	**0.471**	19.46	v2-8 TPU
GMM/BGM	441.50	6.03×10^{-87}	0.239	3.16/3.82	CPU
MultinomialReg	499.84	2.39×10^{-99}	0.298	4.47	CPU
XGBoost	595.17	1.14×10^{-119}	0.367	42.17	CPU

had P-values below 0.1, though not statistically significant. We also visualized the survival across Praxis-BGM clusters using Kaplan Meier curves in Panel (c) Figure 3. By contrast, neither the BGM model nor the XGBoost model identified any cluster with a P-value below this threshold. Overall, the Praxis-BGM model demonstrated a stronger model fit based on the likelihood ratio test ($\chi^2 = 40.25$, P-value $= 3.84 \times 10^{-8}$) in comparison to the LRT for BGM ($\chi^2 = 27.76$, P-value $= 1.40 \times 10^{-5}$) and XGBoost ($\chi^2 = 27.54$, P-value $= 1.54 \times 10^{-5}$). The survival analysis from Praxis-BGM estimated clusters is consistent with previous clinical findings showing that patients with Luminal B subtypes have a worse prognosis than those with Luminal A subtypes (Lammers et al. 2025).

To identify genes driving breast cancer subtype differentiation, we computed BFs for all genes. Panel (d) of Figure 3 displays the distribution of $\log_{10}$ Bayes factors, with the five most noteworthy genes (CCNF, CDH1, CCNDBP1, SQLE, and SIK3) indicated. Some of these genes are strongly implicated in prior studies of breast cancer biology and mechanisms. For example, germline mutations of CDH1 has been found to be strongly associated with invasive lobular carcinoma (Girardi et al. 2022), and SIK3 has recently been identified as a tumor promoter enhancing proliferation and therapeutic resistance (Sorrentino et al.

2022). Panel (e) of Figure 3 shows the pathway overrepresentation analysis of the non-intermediate genes. After FDR adjustment, six pathways were significant, with the number of genes contributing to each pathway on the x-axis and color denoting the adjusted P-value. Interestingly, the early activation of the estrogen response pathway is a well-established signature in the biology of breast cancer (Perou et al. 2000). These results suggest that integrating strong priors from the source data via Praxis-BGM not only leads to more clinically relevant clusters but also enhances the biological interpretability and downstream utility.

5 Application: Label Transfer Across Platforms in Pancreatic scRNA-seq Data

When analyzing single-cell RNA sequencing (scRNA-seq) data, a common step is to annotate the cell type of each estimated cluster of cells. The increased availability of annotated reference datasets for different tissues has facilitated the development of various label transfer tools (Pasquini et al. 2021). A state-of-the-art method for label transfer is scANVI (Xu et al. 2021), a semi-supervised extension of scVI (Lopez et al. 2018) that leverages one or more labeled source (also referred to as reference in the field) datasets to predict cell types in unlabeled target (also referred to as query) datasets. The single-cell analysis framework for Python, `Scanpy` (Hao et al. 2021), also provides built-in functionality to integrate the PC embeddings of a query dataset with those of a reference dataset and then annotate. In addition, label transfer can also be achieved by applying heterogeneity correction and integration methods such as ComBat-seq (Zhang et al. 2020) and Scanorama (Hie et al. 2024), followed by training a classifier (e.g., KNN) on the shared corrected embeddings of the reference and then using it to predict labels in the query data. This task of label transfer for scRNA-seq data naturally aligns with the objective of Praxis-BGM. Thus, we applied

Praxis-BGM to perform clustering guided by prior cell-type-specific information. We used cluster means derived from annotated data in the shared embedding space using the human pancreas single-cell transcriptomic atlas, a publicly available dataset comprising 16,382 cells and annotations with 14 cell types, across 4 technology platforms: inDrop (droplet-based), CEL-Seq (plate-based), Smart-seq (full-length, plate-based), and Fluidigm C1 (microfluidic, full-length) (Luecken et al. 2020). Since these platforms pose challenges for label transfer across batches, this dataset has become a widely used benchmark for evaluating scRNA-seq data integration and label transfer methods (Luecken et al. 2022).

5.0.0.1 Analysis Method Quality control filtering was applied to exclude low-quality cells by removing zero-count entries, followed by the selection of 2,000 highly variable genes (HVGs) on the reference dataset using `Scanpy`. Principal component analysis (PCA) was then performed on the normalized and scaled pancreas scRNA-seq data, and the top 100 components were selected. For label transfer, we defined cells measured from the platform of inDrop as the reference dataset (52.3%), while cells from the other platforms combined were treated as the query dataset (47.7%). We evaluated label transfer performance on the query data using expert-derived cell type annotations that were available for both the reference and query datasets. The ground-truth labels in the reference data were treated as known for label transfer, while the ground-truth labels in the query data were used only for validation. These annotations were generated by integrating clustering results and assigning cell types based on known marker genes (Luecken et al. 2020). Fig. 4(a) shows the ground-truth cell type proportions and count distributions for the reference and query datasets. Specifically, we computed cell-type–specific mean priors in PC space and used them to guide clustering with Praxis-BGM on the query dataset. We benchmarked Praxis-BGM against label-transfer methods (scANVI and `Scanpy`'s `ingest`) and integration-based approaches (ComBat-seq and Scanorama), the latter followed by KNN classification (K = 15 neighbors). In particular,

while Praxis-BGM and scANVI were both applied to the query subset leveraging patterns in the reference dataset, they incorporate the information in distinct ways. Praxis-BGM uses cell–type–specific priors in the corrected PC space to guide clustering of the query data, whereas scANVI learns a shared latent embedding across both the reference and query datasets and then performs label prediction. For the evaluation of computational time, we excluded the cost of computing PCs or the scVI latent representation and measured only the time required for heterogeneity correction and cluster inference on the query data.

5.0.0.2 Results

Table 2: Benchmarking label transfer methods on Pancreas single-cell transcriptomic data against expert annotations.

Method	**ARI**	**Accuracy**	**Time (s)**	**Hardware**	**Embedding**
Praxis-BGM	**0.957**	**0.975**	4.2	TPU	PCs
ComBat-seq + KNN	0.933	0.953	3.4	CPU	PCs
scANVI	0.918	0.927	31.0	GPU	scVI
Scanorama + KNN	0.913	0.916	15.5	CPU	PCs
Scanpy-ingest	0.895	0.914	10.8	CPU	PCs

Table 2 presents the results of benchmarking 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 = 97.5% 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 0.24 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.

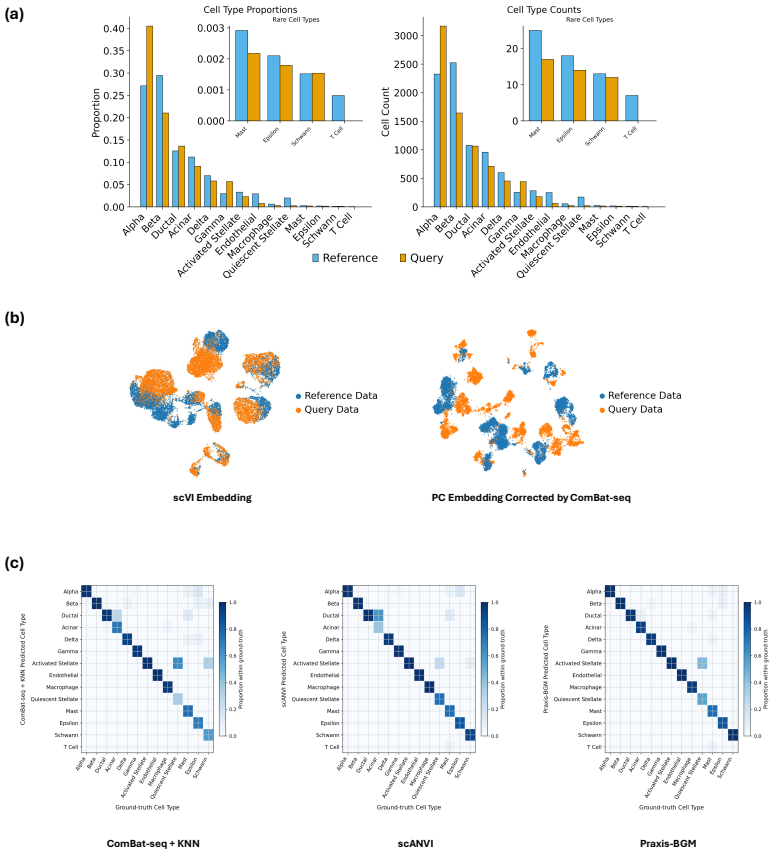

Figure 4: (a) Ground-truth cell type proportions (left) and count (right) distributions for the reference and query datasets. (b) UMAPs using PCs and scVI representations, colored by reference–query datasets. ComBat-seq was applied to PCs to adjust for heterogeneity across datasets, while scVI adjusts for it intrinsically. (c) Confusion heatmap comparing predicted cell types to ground-truth annotations of the cell types in the query data for ComBat-seq + KNN, scANVI, and Praxis-BGM, with the intensity of the color indicating the proportion within the ground-truth cell type.

Praxis-BGM and scANVI relied on different embeddings to perform the label transfer (PCs for Praxis-BGM and scVI representations for scANVI). UMAP visualizations based on the

embeddings provide further insight into the distinct characteristics. Even after Combat-seq batch effect adjustment, for the UMAP of PCs, when coloring cells by reference vs. query datasets, we observed pronounced heterogeneity, leading to strong separation between reference and query cells (Fig. 4(b)). In comparison, the scVI latent representation showed improved mingling, albeit without achieving full mixing. Therefore, Praxis-BGM has the potential advantage of not requiring fully homogeneous reference and query domains. The prior-regularized VI clustering itself can adapt across any domains. Figure 4(c) provides the confusion heatmap comparing predicted cell types to ground-truth annotations of the cell types in the query data. Notably, 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. Only Praxis-BGM accurately assigned the rare cell type of Schwann. In addition, we also evaluated a scenario in which the reference dataset contained fewer rare cell types than the query dataset (Appendix A.4). Under this mismatch, Praxis-BGM still outperformed the benchmarking 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.

6 Discussion

In this paper, we introduced Praxis-BGM, a semi-supervised Gaussian mixture model clustering method based on natural-gradient variational inference that facilitates transfer learning. The source data are assumed to have curated cluster labels, which can be obtained by fitting an unsupervised Gaussian mixture model with an appropriate model selection procedure and subsequently refining the labels through biological interpretation. Relying on the labeled source data, Praxis-BGM provides a highly flexible framework for incorporating prior information through a prior GMM model estimated from source data and updating the prior model with the target data. One or a combination of cluster-specific priors can be specified on the means (prior belief in differences in expression levels across clusters), variance–covariance structures (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 $\boldsymbol{A}$ can be incorporated to encode feature connectivity further. This allows for priors such as those informed by pathway and ontology databases. These prior components jointly 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 subtype-associated genes. 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. Its performance is comparable to deep learning transfer learning approaches such as scANVI,

while maintaining a more interpretable Bayesian GMM framework and having computational advantages.

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 (Pan & Yang 2010). 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 such as Trans-Lasso (Li et al. 2022), well-established strategies exist to select the set of informative source samples. For example, under a Bayesian framework, Wycoff et al. (2025) proposed 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. In addition, feature selection strategies based on shrinkage priors (e.g., modeling the cluster-specific means using the double exponential (Laplace) distribution) could be incorporated into the natural gradient variational inference framework of Praxis-BGM to further suppress noise and enhance robustness.

We applied Praxis-BGM to transcriptomics data and single-cell RNA sequence data. Praxis-BGM has the flexibility to be applied to many omic data types, assuming that the data is properly processed and ready for general clustering techniques. In addition, it is becoming

more common for studies to obtain multiple omic data types on the same individuals, or multi-omic data. Here, popular multi-omics clustering methods such as iCluster often apply integrative factor analysis across multiple omics layers to capture shared variation first, then perform clustering in the latent factor space to combine information from all omics layers (Mo et al. 2013, Rappoport & Shamir 2018). Although Praxis-BGM does not explicitly model multiple omics layers, both the informative priors and observed data could be embedded into a shared low-dimensional space through latent factor analysis (e.g., via Joint and Individual Variation Explained (JIVE) (Lock et al. 2013)), allowing prior construction and clustering to occur in an integrative latent space to leverage more useful information across layers. Thus, extending Praxis-BGM to integrate multi-omics data within a structured joint modeling framework could facilitate more comprehensive biological insights by jointly leveraging information across source and target domains and across multiple omics layers.

Data Availability Statement

All data used in this study are publicly available. The METABRIC dataset (Curtis et al. 2012) was obtained from `cBioPortal` (Cerami et al. 2012), and the TCGA-BRCA dataset (Network 2012) was downloaded via the `curatedTCGAData` package (Ramos et al. 2025). The single-cell RNA-seq pancreas data were aggregated by Luecken et al. (2020) from Gene Expression Omnibus (GEO) GSE81076, GSE85241, GSE86469, GSE84133, GSE81608 , and E-MTAB-5061 (ArrayExpress).

Acknowledgment

All authors contributed to the conception, design, analysis, and writing of this manuscript and approved the final version submitted. The authors declare no conflicts of interest.

This research was supported by the National Institutes of Health grants P01CA196569, U01CA261339, P30CA014089, P30ES007048, U01CA164973, P42ES036506, and R01DK14083.

References

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Wanderman-Milne, S. & Zhang, Q. (2018), 'JAX: Autograd and XLA for high-performance machine learning research', https://github.com/google/jax.

Cai, T., Li, M. & Liu, M. (2025), 'Semi-supervised triply robust inductive transfer learning', *Journal of the American Statistical Association* **120**(550), 1037–1047.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E. et al. (2012), 'The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data', *Cancer discovery* **2**(5), 401–404.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)', Association for Computing Machinery, New York, NY, USA, pp. 785–794.

URL: *https://doi.org/10.1145/2939672.2939785*

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y. et al. (2012), 'The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups', *Nature* **486**(7403), 346–352.

Fang, Z., Liu, X. & Peltz, G. (2022), 'Gseapy: a comprehensive package for performing

gene set enrichment analysis in python', *Bioinformatics* .

URL: *https://doi.org/10.1093/bioinformatics/btac757*

Girardi, A., Magnoni, F., Vicini, E., Kouloura, A., La Vecchia, C., Veronesi, P. & Corso, G. (2022), 'Cdh1 germline mutations in families with hereditary lobular breast cancer', *European Journal of Cancer Prevention* **31**(3), 274–278.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zagar, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. B., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P. & Satija, R. (2021), 'Integrated analysis of multimodal single-cell data', *Cell* .

URL: *https://doi.org/10.1016/j.cell.2021.04.048*

Hie, B. L., Kim, S., Rando, T. A., Bryson, B. & Berger, B. (2024), 'Scanorama: integrating large and diverse single-cell transcriptomic datasets', *Nature Protocols* **19**(8), 2283–2297.

URL: *https://doi.org/10.1038/s41596-024-00991-3*

Jeffreys, H. (1961), *Theory of Probability*, 3rd edn, Oxford University Press.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**(430), 773–795.

Khan, M. E. & Nielsen, D. (2018), 'Fast yet simple natural-gradient descent for variational inference in complex models', *arXiv preprint arXiv:1807.04489* .

URL: *https://arxiv.org/abs/1807.04489*

Kreike, B., van Kouwenhove, M., Horlings, H., Weigelt, B., Peterse, H., Bartelink, H. & van de Vijver, M. J. (2007), 'Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas', *Breast Cancer Research* **9**(5), R65.

Lammers, S. W. M., Geurts, S. M. E., Hermans, K. E. P. E., Kooreman, L. F. S., Swinkels, A. C. P., Smorenburg, C. H., van der Sangen, M. J. C., Kroep, J. R., Honkoop, A. H., van den Berkmortel, F. W. P. J., de Roos, W. K., Linn, S. C., Imholz, A. L. T., Vriens, I. J. H. & Tjan-Heijnen, V. C. G. (2025), 'The prognostic and predictive value of the luminal-like subtype in hormone receptor-positive breast cancer: an analysis of the data trial', *ESMO Open* **10**(2), 104154.

Li, S., Cai, T. T. & Li, H. (2022), '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.

Lin, W., Khan, M. E. & Schmidt, M. (2019), Fast and simple natural-gradient variational inference with mixture of exponential-family approximations, *in* 'Proceedings of the 36th International Conference on Machine Learning (ICML)', Vol. 97 of *Proceedings of Machine Learning Research*, pp. 1–10.

Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. (2013), 'Joint and individual variation explained (jive) for integrated analysis of multiple data types', *Annals of Applied Statistics* **7**(1), 523–542.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. (2018), 'Deep generative modeling for single-cell transcriptomics', *Nature Methods* **15**, 1053–1058.
URL: *https://doi.org/10.1038/s41592-018-0229-2*

Luecken, M., Büttner, M., Danese, A., Interlandi, M., Müller, M., Strobl, D. et al. (2020), 'Benchmarking atlas-level data integration in single-cell genomics - integration task datasets'.
URL: *https://doi.org/10.6084/m9.figshare.12420968.v8*

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F.,

Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M. & Theis, F. J. (2022), 'Benchmarking atlas-level data integration in single-cell genomics', *Nature Methods* **19**(1), 41–50.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M. & Shen, R. (2013), 'Pattern discovery and cancer gene identification in integrated cancer genomic data', *Proceedings of the National Academy of Sciences* **110**(11), 4245–4250.

Network, C. G. A. (2012), 'Comprehensive molecular portraits of human breast tumours', *Nature* **490**(7418), 61–70.

Opper, M. & Archambeau, C. (2009), 'The variational gaussian approximation revisited', *Neural Computation* **21**(3), 786–792.

Pan, S. J. & Yang, Q. (2010), 'A survey on transfer learning', *IEEE Transactions on Knowledge and Data Engineering* **22**(10), 1345–1359.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z. et al. (2009), 'Supervised risk predictor of breast cancer based on intrinsic subtypes', *Journal of Clinical Oncology* **27**(8), 1160–1167.

Pasquini, G., Arias, J. E. R., Schäfer, P. & Busskamp, V. (2021), 'Automated methods for cell type annotation on scrna-seq data', *Computational and Structural Biotechnology Journal* **19**, 961–969. Open Access under CC BY-NC-ND license.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011), 'Scikit-learn: Machine learning in python', *Journal of Machine Learning Research* **12**, 2825–2830.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergamenschikov, A., Williams,

C., Zhu, S., Lønning, P. E., Børresen-Dale, A.-L., Brown, P. O. & Botstein, D. (2000), 'Molecular portraits of human breast tumours', *Nature* **406**, 747–752.

Ramos, M., Waldron, L., Schiffer, L., Geistlinger, L., Obenchain, V. & Morgan, M. (2025), *curatedTCGAData: Curated Data from The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects*. R package version 1.28.1.
URL: *https://bioconductor.org/packages/curatedTCGAData*

Rappoport, N. & Shamir, R. (2018), 'Multi-omic and multi-view clustering algorithms: review and cancer benchmark', *Nucleic acids research* **46**(20), 10546–10562.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S. et al. (2003), 'Repeated observation of breast tumor subtypes in independent gene expression data sets', *Proceedings of the National Academy of Sciences* **100**(14), 8418–8423.

Sorrentino, A., Menevse, A. N., Michels, T., Volpin, V., Durst, F. C. & et al. (2022), 'Salt-inducible kinase 3 protects tumor cells from cytotoxic t-cell attack by promoting TNF-induced NF-κB activation', *Journal for ImmunoTherapy of Cancer* **10**(5), e004258.
URL: *https://doi.org/10.1136/jitc-2021-004258*

Suder, P. M., Xu, J. & Dunson, D. B. (2023), 'Bayesian transfer learning'.
URL: *https://arxiv.org/abs/2312.13484*

Tian, Y. & Feng, Y. (2023), 'Transfer learning under high-dimensional generalized linear models', *Journal of the American Statistical Association* **118**(544), 2684–2697.

Tian, Y., Weng, H., Xia, L. & Feng, Y. (2024), 'Robust unsupervised multi-task and transfer learning on gaussian mixture models', *arXiv preprint arXiv:2209.15224* . Version 4, August 2024.
URL: *https://arxiv.org/abs/2209.15224*

Traag, V. A., Waltman, L. & van Eck, N. J. (2019), 'From louvain to leiden: guaranteeing well-connected communities', *Scientific Reports* **9**(1), 5233.

Wang, R., Li, M., Zhang, W., Liu, Y., Zhou, X. & Yang, L. (2023), 'Transfer-learning-based gaussian mixture model for distributed clustering', *IEEE Transactions on Cybernetics* **53**(11), 7058–7070.

Wycoff, N., Arab, A. & Singh, L. O. (2025), 'Formal bayesian transfer learning via the total risk prior', *arXiv preprint arXiv:2507.23768* .

URL: *https://arxiv.org/abs/2507.23768v1*

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I. & Yosef, N. (2021), 'Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models', *Molecular Systems Biology* **17**, e9620.

Zhang, Y., Parmigiani, G. & Johnson, W. E. (2020), 'Combat-seq: batch effect adjustment for rna-seq count data', *NAR Genomics and Bioinformatics* **2**(3), lqaa078.

URL: *https://doi.org/10.1093/nargab/lqaa078*

Zhao, Y., Jia, Q., Goodrich, J., Darst, B. & Conti, D. V. (2024), 'An extension of latent unknown clustering integrating multi-omics data (lucid) incorporating incomplete omics data', *Bioinformatics Advances* **4**(1), vbae123.

URL: *https://doi.org/10.1093/bioadv/vbae123*

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. & He, Q. (2019), 'A comprehensive survey on transfer learning', *CoRR* **abs/1911.02685**.

URL: *http://arxiv.org/abs/1911.02685*