Doublet Detection in Droplet Based Sequencing Data by Masked Gaussian Mixture Model

HANZHANG LIU[1], Zeyuan Johnson Chen[2,3], Eran Halperin[3], Sriram Sankararaman[2,3,4]

Machine Learning and Genomics Lab, Department of Computer Science, University of California, Los Angeles

Contact: helenwolfie@g.ucla.edu

Abstract

- **Multiplets** in droplet-based single-cell sequencing (droplets containing ≥2 cells) introduce spurious signals that confound downstream analyses.
- **Existing detection methods** rely on generating doublets & are irrespective of the underlying cell-type identity.
- **mGMM** uses an EM algorithm & adapts its optimization process to recover the true singlet distribution in epigenomic and transcriptomic datasets.
- **Outperforms** existing methods across a range of doublet rates (PRAUC ≥ 0.80; **P = 0.012**).

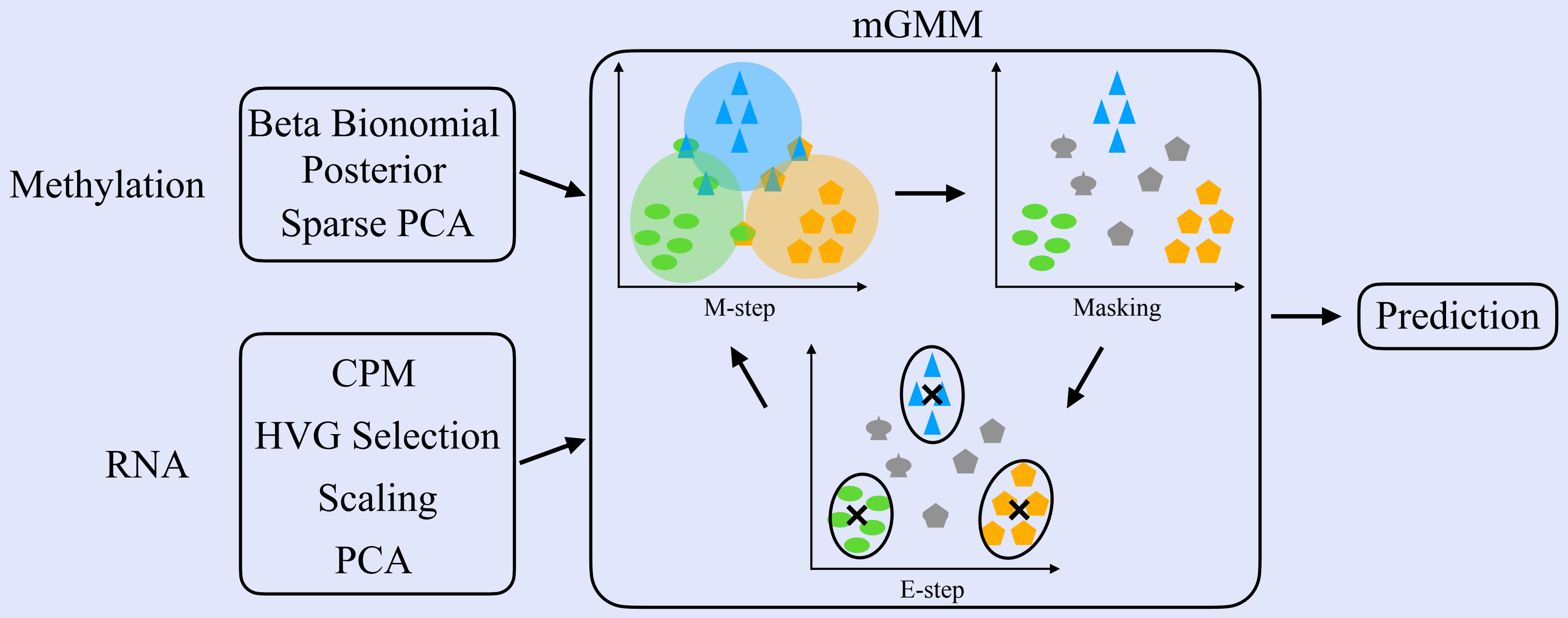
Method

E-step: Compute the responsibilities (posterior probability) $\gamma_{nk} \propto \pi_k N(x_n \mid \mu_k, \Sigma_k)$. Build an augmented K×K proportion matrix Π that collapses singlets and *homotypic* doublets on the diagonal, and assigns *heterotypic* fractions off-diagonal. For each droplet, compute two summary statistics: the LLR_n, and the overall log-likelihood $\log P(x_n)$.

Masking: Calculate the *heterotypic* doublet rate $d\_adj = d\,(1 - \Sigma_k\, \pi_k^2)$. Droplets in the bottom d_adj quantile of LLR_n are considered to be plausible **Multiplets** under the current iteration. Droplets in the bottom outlier rate (q) quantile of $\log P(x_n)$ are identified as plausible **Outliers**. Both groups are subsequently masked from the model's parameter updates in the M-step.

M-step: Refit mixture weights, means, and covariances using only unmasked droplets. Iterate E/M with masking until log-likelihood stabilizes or reaches a max iteration cap.

Method Overview:



$$\underset{\Pi}{\begin{pmatrix} 0.168 & 0.012 & 0.02 \\ 0.012 & 0.258 & 0.03 \\ 0.02 & 0.03 & 0.45 \end{pmatrix}} = \underset{d}{0.2} \underset{\pi\pi^\top}{\begin{pmatrix} 0.04 & 0.06 & 0.10 \\ 0.06 & 0.09 & 0.15 \\ 0.10 & 0.15 & 0.25 \end{pmatrix}} + \underset{1 - d}{0.8} \underset{\mathrm{diag}(\pi)}{\begin{pmatrix} 0.2 & & \\ & 0.3 & \\ & & 0.5 \end{pmatrix}}$$

Log Likelihood Ratio Test:

$$LL_s = max_k\ \mathbb{P}(x_n, z_n = k|\theta)$$

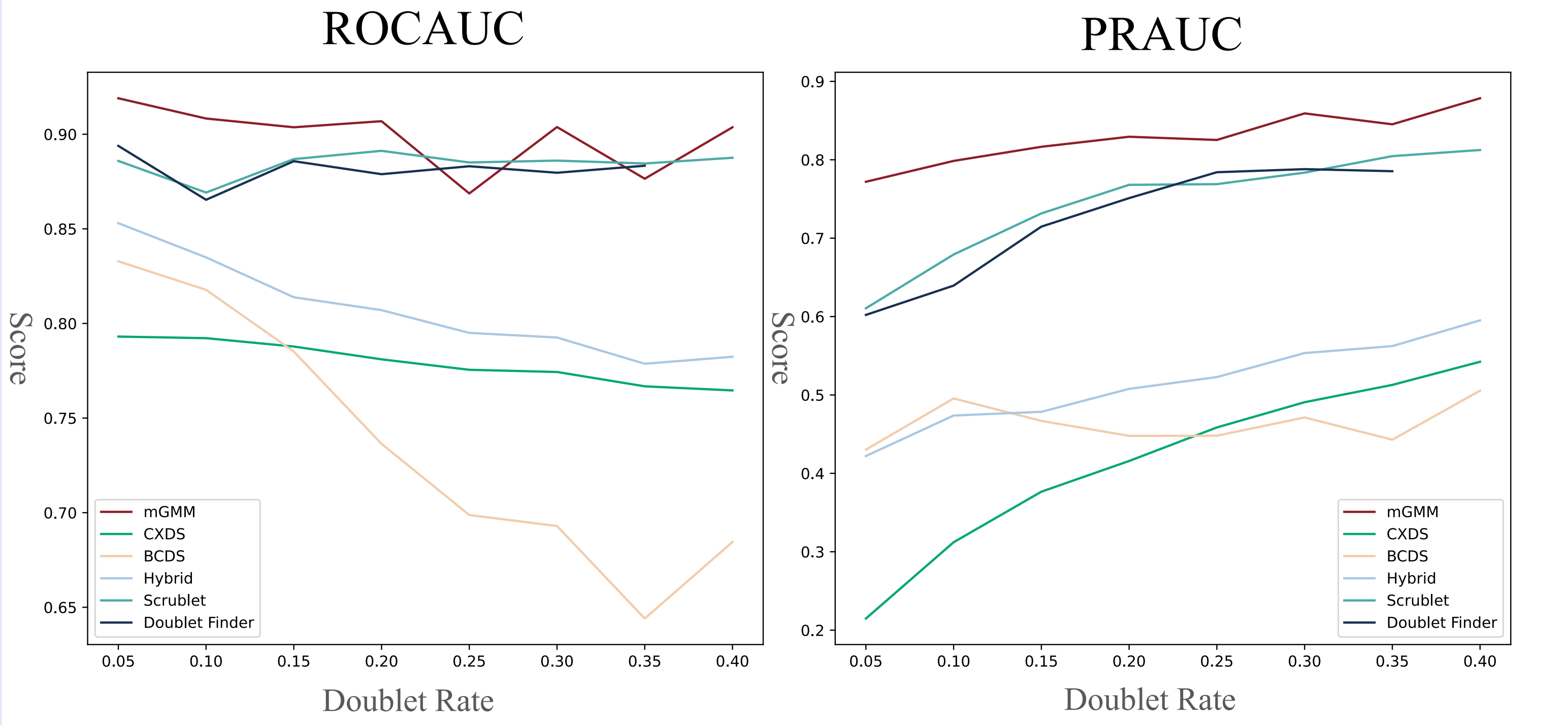
$$LL_d = max_{k_1<k_2}\ \mathbb{P}(x_n, z_n = (k_1, k_2)|\theta)$$

$$LLR = \frac{LL_s}{max(LL_d, LL_s)}$$

Results

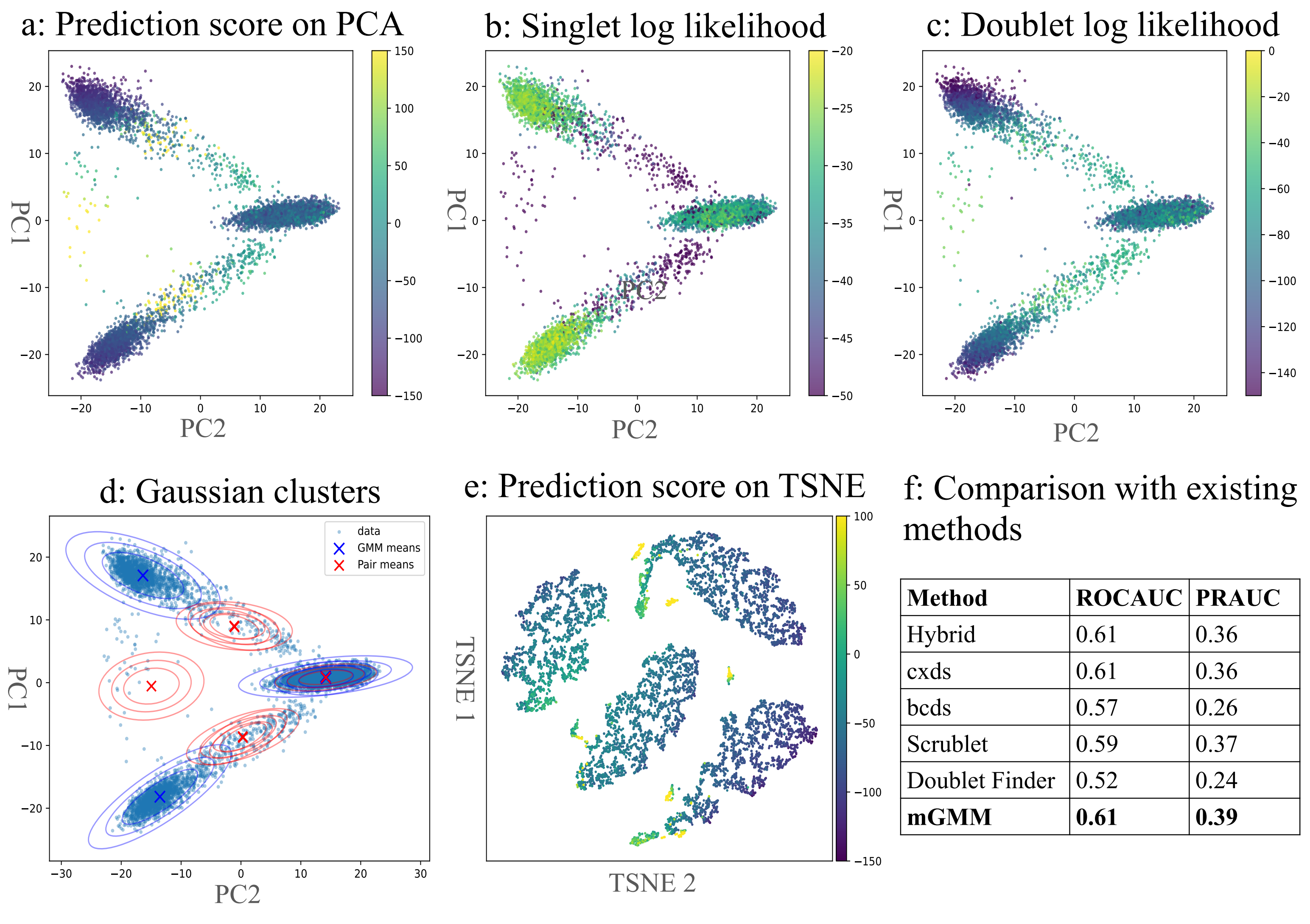
Epigenomic datasets:

We applied mGMM to the human prefrontal cortex (PFC) single-nucleus methylation dataset from Lee *et al.,* in which nuclei underwent fluorescence-activated nuclei sorting (FANS) to remove multiplex. Doublets were simulated at rates ranging from 5% to 40% of the original number of singlets. We compared mGMM with Doublet Finder (McGinnis et al.), Scrublet (Wolock et al.), Hybrid, CXDS, and BCDS (Bais & Kostka). mGMM consistently achieved high performance in recovering simulated doublets across all simulation settings, with ROCAUC ≥ 0.85 and PRAUC ≥ 0.75. Compared to the second-best method, mGMM yielded a statistically significant improvement in PRAUC (P = 0.012). Notably, it is the only method that performs well in the lower doublet rates regime (0.05 ~ 0.15).



Transcriptomic datasets:

We evaluated mGMM on the cline-ch dataset (Stoeckius et al.), comprising four human cell lines (HEK, K562, KG1, THP1) with multiplets annotated via cell hashing.



a: Prediction score on PCA

b: Singlet log likelihood

c: Doublet log likelihood

d: Gaussian clusters

e: Prediction score on TSNE

f: Comparison with existing methods

Method	ROCAUC	PRAUC
Hybrid	0.61	0.36
cxds	0.61	0.36
bcds	0.57	0.26
Scrublet	0.59	0.37
Doublet Finder	0.52	0.24
mGMM	**0.61**	**0.39**

a-c: Visualization of all droplets in PC space. Droplets are colored by predicted scores in **a**, log likelihood of singlet in **b**, and log likelihood of doublet in **c**. (d) Gaussian clusters identified by mGMM (blue contours indicate the ellipsoid encapsulating 1 to 2 SD for singlets, red contours indicate that of doublets), (e) Similar to **a**, except on t-SNE embedding space (f) Comparing the performance of mGMM with other doublet detection methods evaluated using ROCAUC and PRAUC.

Data Processing

Simulation

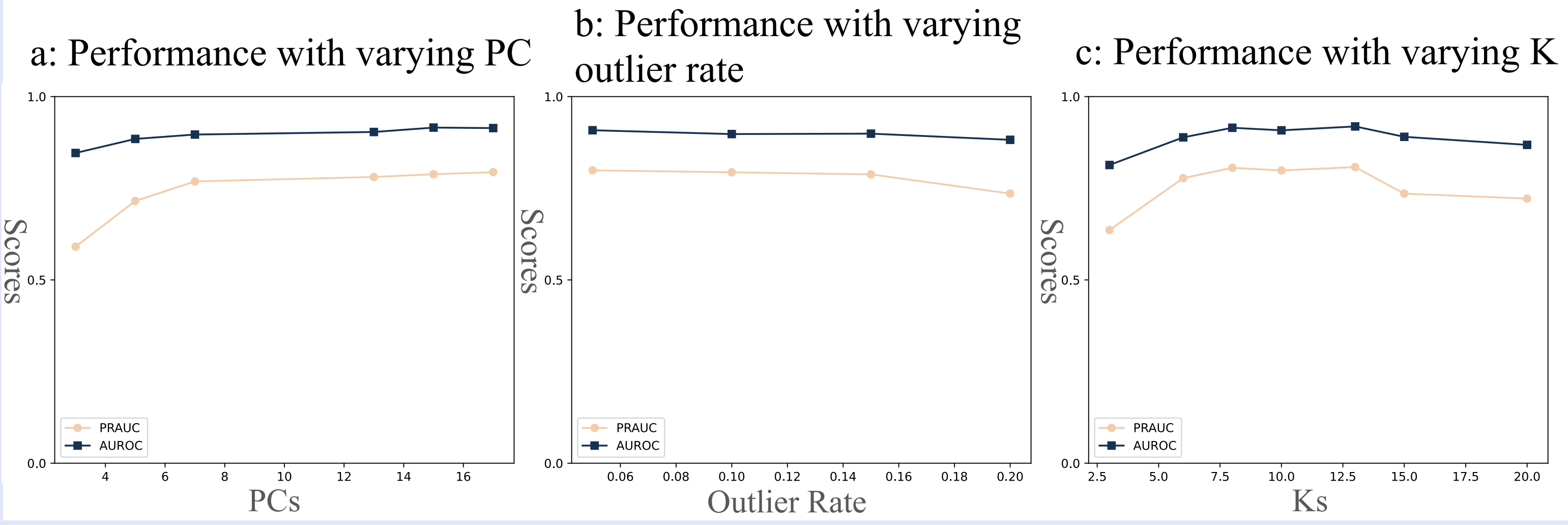
Multiplets were generated using the following procedure:
1. Randomly select two singlet cells (C1 and C2).
2. Aggregate the total counts and methylated counts of C1 and C2 to M1.
3. Append M1 to the dataset as a simulated multiplet
4. Compute the methylation ratio for all cells as the fraction of counts that are methylated

Sparse PCA

Sparse PCA is used in place of HVG selection. Previous work has shown that features (e.g., CpG sites, chromosome bins) better reconstructed through low-rank approximation capture greater cell-type heterogeneity in **methylation data profiled by arrays** and are therefore beneficial for methods primarily focused on detecting cell-type-level doublets.

Discussion

mGMM detects and masks outliers and doublets, thereby removing their influence in each iteration of the EM process. The method performs optimally when clusters are well-separated in low-dimensional space and can be approximated by multivariate Gaussian mixtures, a condition commonly satisfied in methylation data from solid tissue but less so in expression data from liquid tissue. It also demonstrates robust performance across a range of hyperparameter settings, including the number of clusters (K), principal components (PC), and outlier rate. The current framework assumes a linear projection of the data, which places doublets midway between two well-formed clusters in the embedding space. However, such an assumption may not hold when the preprocessing step includes non-linear transformations. Finally, similar to competing methods that solely rely on transcriptomic/epigenomic profiles, mGMM does not have the capacity to distinguish homotypic multiplets from singlets.



References

Wolock et al. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Bioinformatics. 2020. **Lee et al.** Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nat Methods. 2019. **McGinnis et al.** DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Systems. 2019. **Rahmani et al.**, Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. Nat Methods. 2016. **Bais & Kostka.** scds: computational annotation of doublets in single-cell RNA sequencing data. Bioinformatics. 2020. **Stoeckius et al.** Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 2018.