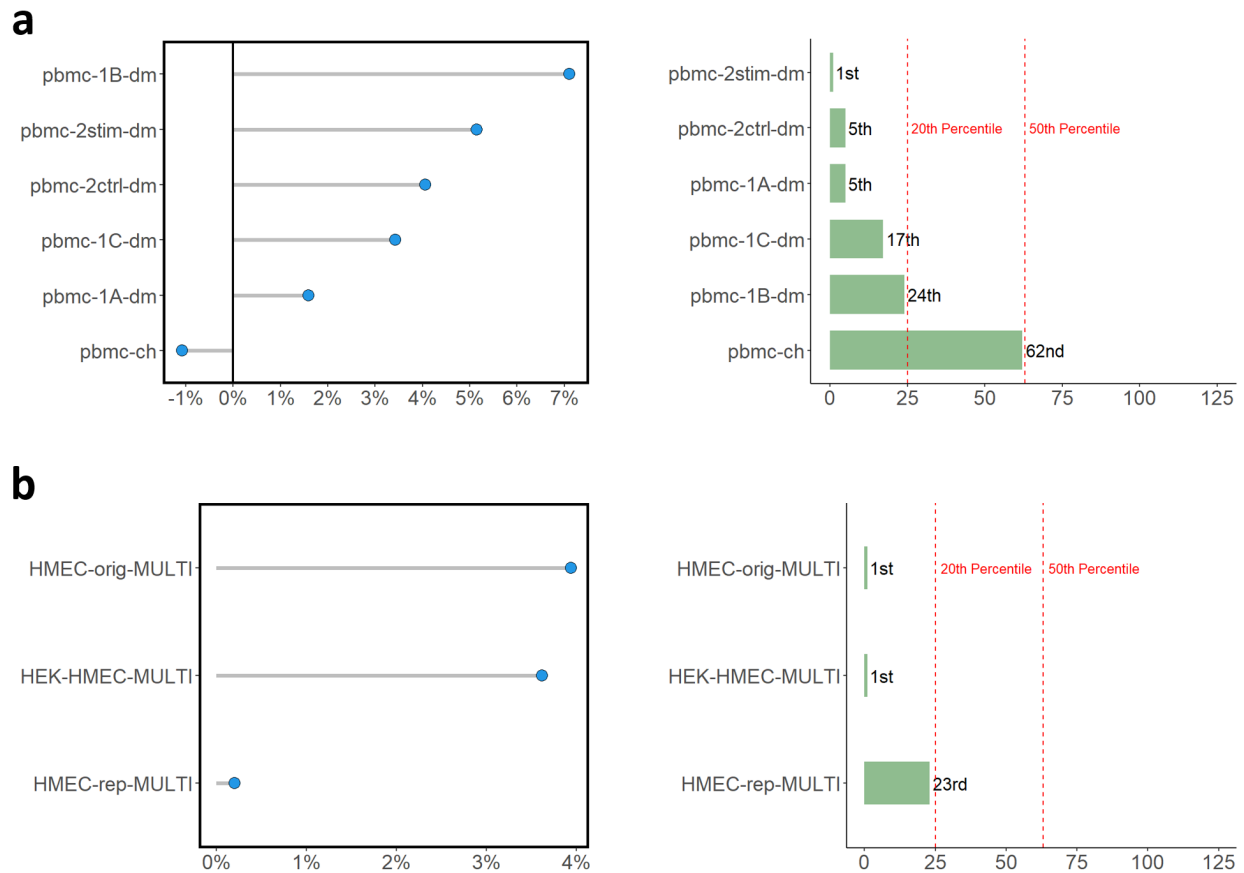


# Supplementary Materials for “Tuning hyperparameters of doublet-detection methods for single-cell RNA sequencing data”

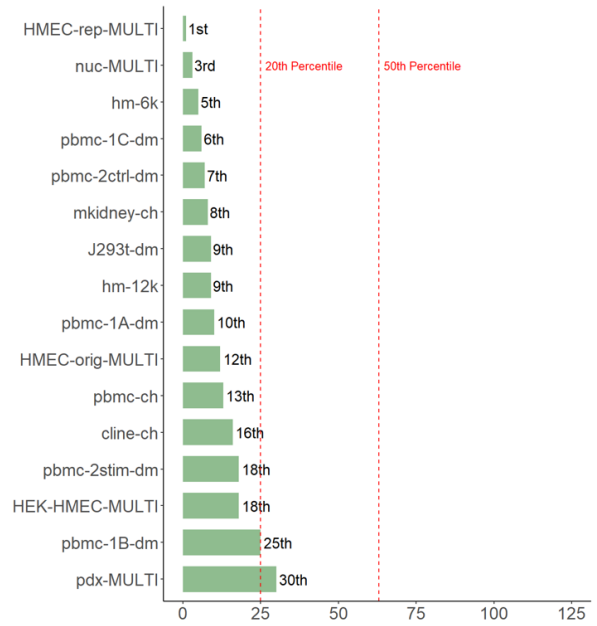
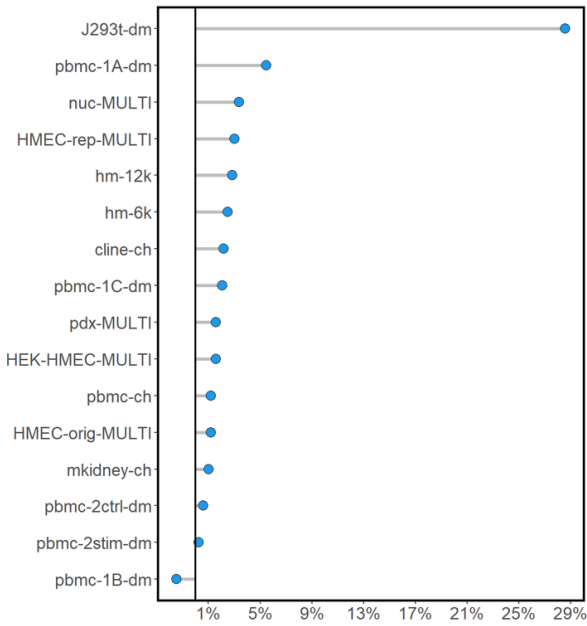
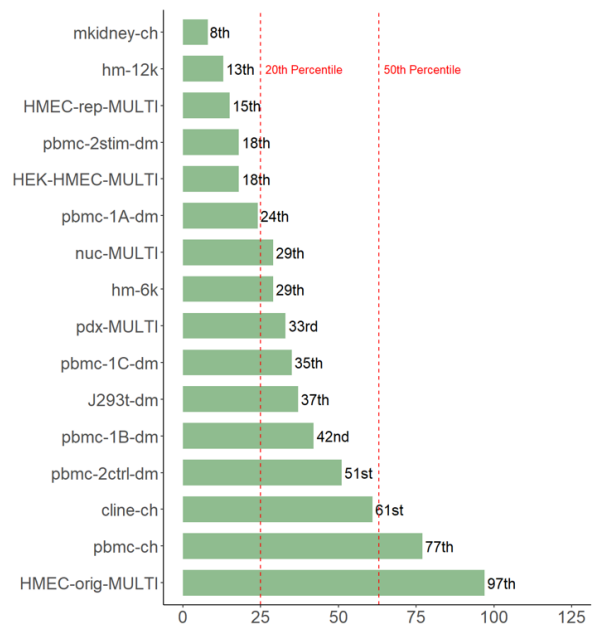
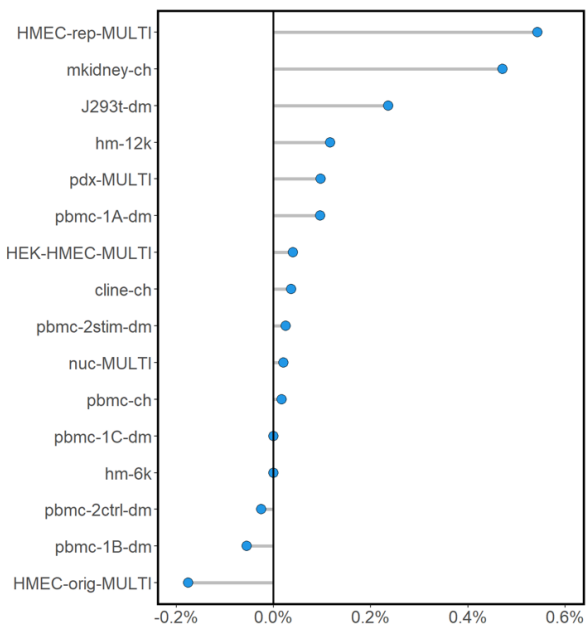
Nan Miles Xi <sup>a,\*</sup> and Angelos Vasilopoulos <sup>a</sup>

<sup>a</sup> Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL 60660, USA

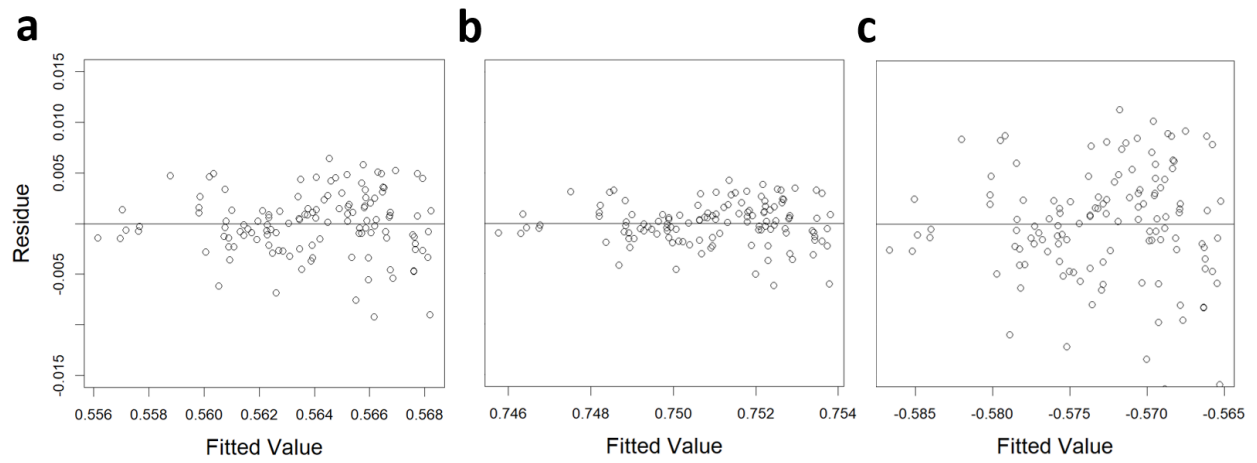
\* Correspondence: Nan Miles Xi ([mxil@luc.edu](mailto:mxil@luc.edu))



**Supplementary Figure S1.** AUPRC improvement and ranking with hyperparameters optimized on datasets of different biological conditions. **a**, Optimized on pbmc-related datasets. **b**, Optimized on HMEC-related datasets.

**a****b**

**Supplementary Figure S2.** Double-detection improvement and ranking in terms of different measurements. **a**, Measured by true positive rate. **b**, Measured by true negative rate. The hyperparameters are optimized on all 16 scRNA-seq datasets



**Supplementary Figure S3.** Residue plots of model (1) on datasets before and after transformation on the response variable. All three figures have the same residue scale. **a**, No transformation. **b**, Square-root-transformation on response variable AUPRC. **c**, Natural-log-transformation on response variable AURPC.

**Supplementary Table S1.** The significant hyperparameters in model (1) and their optimal values (rounding to the nearest integers) optimized on different biological conditions and accuracy metrics of doublet detection.

<b>Optimization subset</b>	<b>Optimization metric</b>	<b>Significant hyperparameter</b>	<b>Optimal value</b>
All	AUPRC	<i>pc, nf</i>	<i>pc = 19,</i> <i>nf = 1352</i>
pbmc-related	AUPRC	<i>pc</i>	<i>pc = 18</i>
HMEC-related	AUPRC	<i>nf</i>	<i>nf = 1520</i>
All	True positive rate	<i>depth</i>	<i>depth = 5</i>
All	True negative rate	<i>depth</i>	<i>depth = 5</i>

**Supplementary Table S2.** The AUPRC of doublet detection under optimal and default hyperparameters on pbmc-related datasets. The last column shows the highest AUPRC achieved by one of the 125 hyperparameter combinations. The highest AUPRC between the optimum and default of each dataset is underscored.

<b>Dataset</b>	<b>Optimum</b>	<b>Default</b>	<b>Maximum</b>
pbmc-1A-dm	<u>0.5528</u>	0.5441	0.5693
pbmc-1B-dm	<u>0.4440</u>	0.4145	0.4818
pbmc-1C-dm	<u>0.5941</u>	0.5744	0.6082
pbmc-2ctrl-dm	<u>0.7023</u>	0.6749	0.7088
pbmc-2stim-dm	<u>0.7111</u>	0.6763	0.7124
pbmc-ch	0.6402	<u>0.6472</u>	0.6520

**Supplementary Table S3.** The AUPRC of doublet detection under optimal and default hyperparameters on HMEC-related datasets. The last column shows the highest AUPRC achieved by one of the 125 hyperparameter combinations. The highest AUPRC between the optimum and default of each dataset is underscored.

<b>Dataset</b>	<b>Optimum</b>	<b>Default</b>	<b>Maximum</b>
HEK-HMEC-MULTI	<u>0.5005</u>	0.4830	0.4966
HEK-orig-MULTI	<u>0.5065</u>	0.4873	0.5054
HMEC-rep-MULTI	<u>0.5976</u>	0.5964	0.6020

**Supplementary Table S4.** The true positive rate of doublet detection under optimal and default hyperparameters. The last column shows the highest true positive rate achieved by one of the 125 hyperparameter combinations. The larger true positive rate between optimum and default in each dataset is underscored.

<b>Dataset</b>	<b>Optimum</b>	<b>Default</b>	<b>Maximum</b>
cline-ch	<u>0.3543</u>	0.3468	0.3604
HEK-HMEC-MULTI	<u>0.5297</u>	0.5215	0.5358
HEK-orig-MULTI	<u>0.8973</u>	0.8726	0.9233
hm-12k	<u>0.9591</u>	0.9357	0.9649
hm-6k	<u>0.5028</u>	0.4969	0.5121
HMEC-rep-MULTI	<u>0.5314</u>	0.5158	0.5314
J293t-dm	<u>0.2143</u>	0.1667	0.2381
mkidney-ch	<u>0.5875</u>	0.5816	0.5894
nuc-MULTI	<u>0.4505</u>	0.4358	0.4505
pbmc-1A-dm	<u>0.6417</u>	0.6083	0.6500
pbmc-1B-dm	0.5154	<u>0.5231</u>	0.5462
pbmc-1C-dm	<u>0.6266</u>	0.6139	0.6266
pbmc-2ctrl-dm	<u>0.7196</u>	0.7153	0.7240
pbmc-2stim-dm	<u>0.7192</u>	0.7174	0.7253
pbmc-ch	<u>0.6346</u>	0.6271	0.6369
pdx-MULTI	<u>0.4374</u>	0.4305	0.4487

**Supplementary Table S5.** The true negative rate of doublet detection under optimal and default hyperparameters. The last column shows the highest true negative rate achieved by one of the 125 hyperparameter combinations. The larger true negative rate between optimum and default in each dataset is underscored.

<b>Dataset</b>	<b>Optimum</b>	<b>Default</b>	<b>Maximum</b>
cline-ch	<u>0.8530</u>	0.8527	0.8558
HEK-HMEC-MULTI	<u>0.9774</u>	0.9770	0.9777
HEK-orig-MULTI	<u>0.9936</u>	0.9925	0.9955
hm-12k	<u>0.9986</u>	0.9986	0.9994
hm-6k	0.9199	<u>0.9216</u>	0.9239
HMEC-rep-MULTI	<u>0.7868</u>	0.7825	0.7895
J293t-dm	<u>0.9279</u>	0.9258	0.9323
mkidney-ch	<u>0.7546</u>	0.7511	0.7558
nuc-MULTI	<u>0.9479</u>	0.9477	0.9490
pbmc-1A-dm	<u>0.9865</u>	0.9855	0.9871
pbmc-1B-dm	0.9831	<u>0.9836</u>	0.9844
pbmc-1C-dm	<u>0.9756</u>	0.9756	0.9764
pbmc-2ctrl-dm	0.9629	<u>0.9631</u>	0.9643
pbmc-2stim-dm	<u>0.9629</u>	0.9626	0.9637
pbmc-ch	<u>0.9257</u>	0.9256	0.9276
pdx-MULTI	<u>0.9176</u>	0.9167	0.9194

**Supplementary Table S6.** The significant hyperparameters and their raw optimal values before and after transformation on response variable AUPRC.

<b>Transformation</b>	<b>Significant hyperparameter</b>	<b>Optimal value</b>
No transformation	<i>nf, pc</i>	nf = 1351.661, pc = 19.181
Square root	<i>nf, pc</i>	nf = 1352.087, pc = 19.189
Natural log	<i>nf, pc</i>	nf = 1352.509, pc = 19.196