Dimension Reduction for Data with Heterogeneous Missingness Supplementary Material

Yurong Ling[1] **Zijing Liu**[2] **Jing-Hao Xue**[1]

[1]Department of Statistical Science, University College London, London, UK
[2]Department of Brain Sciences, Imperial College London, London, UK

S.1 DATASET PRE-PROCESSING

For the pre-processing of the scRNA-seq datasets, genes (features) observed in less than 2 cells (observations) are first discarded, followed by a log2 transformation with pseudo 1 count added. For the other real datasets, we keep all features obtained from the original publications.

S.2 DATA AVAILABILITY

- Pollen: cells in this dataset are defined by the human cell lines. We directly download the TPM values from `https://s3.amazonaws.com/scrnaseq-public-datasets/manual-data/pollen/NBT_hiseq_linear_tpm_values.txt`.
- Deng: this dataset was used to study monoallelic expression at the single cell level. The data are available in the Gene Expression Omnibus (GEO) database under the accession number GSE45719 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45719`).
- Treutlein: this dataset contains cell populations from direct reprogramming from fibroblast to neuron (MEF, day 2, 5, and 22). The data are available in the GEO database under the accession number GSE67310 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67310`).
- Koh: this dataset comprises 9 different cell types that include H7 hESCs, H7-derived anterior primitive streak populations, H7-derived mid primitive streak populations, H7-derived lateral mesoderm, H7-derived FACS-purified GARP+ cardiac mesoderm, H7-derived FACS-purified DLL1+ paraxial mesoderm populations, H7-derived day 3 early somite progenitor populations, H7-derived dermomyotome populations, and H7-derived FACS-purified PDGFRα+ sclerotome populations. RNA was extracted from either whole cell populations or, alternatively, cell subsets purified by fluorescence activated cell sorting (FACS). We download the processed data from conquer, a repository of consistently processed, analysis-ready public scRNA-seq datasets Soneson and Robinson [2018] (`http://imlspenticton.uzh.ch:3838/conquer/`).
- Usoskin: this dataset consists of 11 types of mouse lumbar DRG (dorsal root ganglion). We download the normalized data from `http://linnarssonlab.org/drg/` (External resource table 1).
- Kumar: this dataset contains three populations of mouse embryonic stem cells. The data are available in the GEO database under the accession number GSE60749 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60749`).
- Olivetti faces: this dataset is loaded using a sklearn function in Python (`https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_olivetti_faces.html`).
- fashion MNIST: this dataset is loaded using Keras (`https://keras.io/api/datasets/fashion_mnist/`).
- wine: this dataset is loaded using a sklearn function in Python (`https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html?highlight=wine%20dataset`).

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

S.3 ESTIMATION OF MISSINGNESS PROBABILITIES

Suppose that the probability of non-missingness for the i-th cell and the s-th gene is given by $p_{is} = c_i g_s$, where c_i and g_s account for the influence from the i-th cell and s-th gene, respectively.

By using the method of moments, we get

$$\begin{aligned} \mathbb{E}\left[\sum_{s=1}^{D} h_{is}\right] &= c_i \sum_{s=1}^{D} g_s = M_{i\cdot}, \\ \mathbb{E}\left[\sum_{i=1}^{N} h_{is}\right] &= g_s \sum_{i=1}^{N} c_i = M_{\cdot s}, \end{aligned}$$

where $M_{i\cdot} = \sum_{s=1}^{D} M_{is}$, $M_{\cdot s} = \sum_{i=1}^{N} M_{is}$, M is the indicator matrix with 1 representing non-missingness event, and $s = 1, \dots, D$; $i = 1, \dots, N$. It is straightforward to verify that $\bar{c}_i = \frac{M_{i\cdot}}{N} / \sqrt{\frac{m}{DN}}$ and $\bar{g}_s = \frac{M_{\cdot s}}{D} / \sqrt{\frac{m}{DN}}$, where $m = \sum_i M_{i\cdot} = \sum_s M_{\cdot s}$, satisfy the above equations. However, it is likely that $\bar{c}_i \bar{g}_s > 1$. Hence, we normalise the estimator by $\bar{c}_i = \frac{M_{i\cdot}}{N\sqrt{n_c}}$ and $\bar{g}_s = \frac{M_{\cdot s}}{D\sqrt{n_c}}$, where $n_c = \max_{i\in[N], s\in[D]}(\frac{M_{i\cdot} M_{\cdot s}}{DN}, \frac{m}{DN})$.

S.4 IMPLEMENTATION DETAILS

There are several hyperparameters in the procedure of distinguishing between dropouts and biological non-expression. We use k-means clustering, a key ingredient in analysing scRNA-seq data, and spectral clustering, accounting for the nonlinearity in the data. For each clustering method, a wide range of cluster numbers are assigned: $(4, 6, 8, 10, 12)$. The kernel coefficient in the spectral clustering is set to be the average distance to the 7-th nearest neighbour. All other parameters are used as default.

A zero count would be deemed to be biological non-expression if the proportion of similar cells showing zero expression in the same gene exceeds the threshold 85%. In practice, we found that this threshold should be higher than 0.5, and we would get better results when it ranges from 0.7 to 0.95.

For each DR benchmark or its variant when applied to the scRNA-seq datasets, we replicate the procedure of first performing DR and then applying k-means 20 times. Each time when performing the k-means algorithm on the extracted low-dimensional components, the number of repeats of k-means starting with different centroid initialisations is set as 30 for more reliable results.

For the simulated datasets, we first sample subsets of features, followed by performing DR techniques. Last, we run the K-means algorithms on the low-dimensional data points with 30 different centroid initialisations and obtain the most reliable clustering results in terms of inertia.

All clustering methods and ice (IterativeImputer) are implemented with scikit-learn v0.22.1 of Python [Pedregosa et al., 2011]. For GPLVM, tSNE and UMAP, we use the GPflow package v1.3.0 [Matthews et al., 2017], the scikit-learn package v0.21.3 [Pedregosa et al., 2011] and the umap-learn package v0.3.10 [McInnes et al., 2018], with default settings, respectively. For softImpute, we use the fancyimpute package v0.5.5 [Rubinsteyn and Feldman, 2016].

S.5 PROOFS

S.5.1 PROOF FOR PROPOSITION 1

For $i \neq j$, based on the law of total expectation, we get

$$\begin{aligned} \mathbb{E}\left[y_{is}y_{js}\right] &= \mathbb{E}\left[\mathbb{E}\left[y_{is}y_{js} \mid h_{is}, h_{js}\right]\right] \\ &= \mathbb{E}\left[y_{is}y_{js} \mid h_{is} = 1 \,\text{and}\, h_{js} = 1\right] p_{is}p_{js} + \mathbb{E}\left[y_{is}y_{js} \mid h_{is} = 0 \,\text{and}\, h_{js} = 0\right] (1 - p_{is})(1 - p_{js}) \\ &+ \mathbb{E}\left[y_{is}y_{js} \mid h_{is} = 1 \,\text{and}\, h_{js} = 0\right] (1 - p_{js}) p_{is} + \mathbb{E}\left[y_{is}y_{js} \mid h_{is} = 0 \,\text{and}\, h_{js} = 1\right] (1 - p_{is}) p_{js} \\ &= \mathbb{E}\left[y_{is}y_{js} \mid h_{is} = 1\right] p_{is}p_{js} = p_{is}p_{js}K_{ij}, \end{aligned}$$

$$
\begin{aligned}
\mathrm{Var}\left[y_{is}y_{js}\right] &= \mathbb{E}\left[y_{is}^2 y_{js}^2\right] - \mathbb{E}\left[y_{is}y_{js}\right]^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[y_{is}^2 y_{js}^2 \mid h_{is}, h_{js}\right]\right] - \mathbb{E}\left[y_{is}y_{js}\right]^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[y_{is}^2 y_{js}^2 \mid h_{is}, h_{js}\right]\right] - p_{is}^2 p_{js}^2 K_{ij}^2 \\
&= \left(K_{ii}K_{jj} + 2K_{ij}^2\right) p_{is}p_{js} - p_{is}^2 p_{js}^2 K_{ij}^2 \\
&= p_{is}p_{js}K_{ii}K_{jj} + K_{ij}^2\left(2p_{is}p_{js} - p_{is}^2 p_{js}^2\right).
\end{aligned}
$$

For $i = j$, we get

$$
\begin{aligned}
\mathbb{E}\left[y_{is}^2\right] &= \mathbb{E}\left[\mathbb{E}\left[y_{is}^2 \mid h_{is}\right]\right] \\
&= \mathbb{E}\left[y_{is}^2 \mid h_{is} = 1\right] p_{is} + \mathbb{E}\left[y_{is}^2 \mid h_{is} = 0\right](1 - p_{is}) \\
&= \mathbb{E}\left[y_{is}^2 \mid h_{is} = 1\right] p_{is} = p_{is}K_{ii},
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}\left[y_{is}^2\right] &= \mathbb{E}\left[y_{is}^4\right] - \mathbb{E}\left[y_{is}^2\right]^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[y_{is}^4 \mid h_{is}\right]\right] - \mathbb{E}\left[y_{is}^2\right]^2 \\
&= 3K_{ii}^2 p_{is} - K_{ii}^2 p_{is}^2 = K_{ii}^2\left(3p_{is} - p_{is}^2\right).
\end{aligned}
$$

S.5.2 PROOF FOR PROPOSITION 2

Based on Proposition 1, it is straightforward to get that the elements in the unbiased estimator $\tilde{G}$ of K are given by

$$
\begin{cases}
\tilde{G}_{ij} = \frac{G_{ij}}{\sum_{s=1}^D p_{is}p_{js}}, \text{ for } i \neq j; \\
\tilde{G}_{ii} = \frac{G_{ii}}{\sum_{s=1}^D p_{is}},
\end{cases}
$$

and the corresponding variances are given by

$$
\begin{cases}
\mathrm{Var}\left[\tilde{G}_{ij}\right] = \frac{K_{ii}K_{jj}\sum_{s=1}^D p_{is}p_{js} + K_{ij}^2\sum_{s=1}^D (2p_{is}p_{js} - p_{is}^2 p_{js}^2)}{\left(\sum_{s=1}^D p_{is}p_{js}\right)^2}, \text{ for } i \neq j; \\
\mathrm{Var}\left[\tilde{G}_{ii}\right] = \frac{K_{ii}^2\sum_{s=1}^D p_{is}(3 - p_{is})}{\left(\sum_{s=1}^D p_{is}\right)^2}.
\end{cases}
$$

We first consider the bounds for $\mathrm{Var}\left[\tilde{G}_{ij}\right]$, for $i \neq j$. $\mathrm{Var}\left[\tilde{G}_{ij}\right]$ can be re-written as

$$
\begin{aligned}
\mathrm{Var}\left[\tilde{G}_{ij}\right] &= \frac{K_{ii}K_{jj}\sum_{s=1}^D p_{is}p_{js} + K_{ij}^2\sum_{s=1}^D (2p_{is}p_{js} - p_{is}^2 p_{js}^2)}{\left(\sum_{s=1}^D p_{is}p_{js}\right)^2} \\
&= \frac{K_{ii}K_{jj}}{\bar{p}_{ij}D} + \frac{K_{ij}^2}{D}\frac{\sum_{s=1}^D\left(\frac{2p_{is}p_{js}}{D} - \frac{p_{is}^2 p_{js}^2}{D}\right)}{\left(\frac{\sum_{s=1}^D p_{is}p_{js}}{D}\right)^2} \\
&= \frac{K_{ii}K_{jj}}{\bar{p}_{ij}D} + \frac{K_{ij}^2}{D}\frac{\left(2\bar{p}_{ij} - \frac{\sum_{s=1}^D p_{is}^2 p_{js}^2}{D}\right)}{\bar{p}_{ij}^2},
\end{aligned}
$$

where $0 < \bar{p}_{ij} = \frac{1}{D}\sum_{s=1}^D p_{is}p_{js} \leq 1$. Note that $2\bar{p}_{ij} - \frac{\sum_{s=1}^D p_{is}^2 p_{js}^2}{D} \geq \bar{p}_{ij}$, We thus get the lower bound for $\mathrm{Var}\left[\tilde{G}_{ij}\right]$:

$$
\mathrm{Var}\left[\tilde{G}_{ij}\right] \geq \frac{K_{ii}K_{jj}}{\bar{p}_{ij}D} + \frac{K_{ij}^2}{D}\frac{\bar{p}_{ij}}{\bar{p}_{ij}^2} = \frac{K_{ii}K_{jj} + K_{ij}^2}{D\bar{p}_{ij}}
$$

Meanwhile, since $f(x) = 2x - x^2$ (with $E[f(x)] \leq f[E(x)]$) is a strictly concave function, we have

$$\begin{aligned}
\mathrm{Var}\left[\tilde{G}_{ij}\right] &= \frac{K_{ii}K_{jj}}{\bar{p}_{ij}D} + K_{ij}^2 D \frac{\sum_{s=1}^{D}(\frac{2p_{is}p_{js}}{D} - \frac{p_{is}^2 p_{js}^2}{D})}{(\sum_{s=1}^{D} p_{is}p_{js})^2} \\
&\leq \frac{K_{ii}K_{jj}}{\bar{p}_{ij}D} + K_{ij}^2 D \frac{\frac{2\sum_{s=1}^{D} p_{is}p_{js}}{D} - \left(\frac{\sum_{s=1}^{D} p_{is}p_{js}}{D}\right)^2}{\left(\sum_{s=1}^{D} p_{is}p_{js}\right)^2} \\
&= \frac{K_{ii}K_{jj}}{\bar{p}_{ij}D} + \frac{K_{ij}^2}{D}\left(\frac{2}{\bar{p}_{ij}} - 1\right).
\end{aligned}$$

Next, we consider the bounds for $\mathrm{Var}\left[\tilde{G}_{ii}\right]$. By using the fact that $\sum_{s=1}^{D} p_{is}^2 \leq \sum_{s=1}^{D} p_{is}$ we get

$$\mathrm{Var}\left[\tilde{G}_{ii}\right] = \frac{3K_{ii}^2}{D\bar{p}_i} - \frac{K_{ii}^2 \sum_{s=1}^{D} p_{is}^2}{(\sum_{s=1}^{D} p_{is})^2} \geq \frac{2K_{ii}^2}{D\bar{p}_i},$$

where $\bar{p}_i = \frac{1}{D}\sum_{s=1}^{D} p_{is}$. Again, since since $g(x) = 3x - x^2$ is a strictly concave function, we have

$$\begin{aligned}
\mathrm{Var}\left[\tilde{G}_{ii}\right] &= \frac{K_{ii}^2 \sum_{s=1}^{D}(3p_{is} - p_{is}^2)}{(\sum_{s=1}^{D} p_{is})^2} \\
&\leq K_{ii}^2 D \frac{\frac{3\sum_{s=1}^{D} p_{is}}{D} - \left(\frac{\sum_{s=1}^{D} p_{is}}{D}\right)^2}{\left(\sum_{s=1}^{D} p_{is}\right)^2} \\
&= \frac{K_{ii}^2}{D}\left(\frac{3}{\bar{p}_i} - 1\right).
\end{aligned}$$

S.5.3 PROOF FOR PROPOSITION 3

First, we consider the limiting distribution of $Z_{ij,D}$ for $i \neq j$. We verify that for any $\epsilon > 0$,

$$\lim_{D\to\infty} \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} \mathrm{E}\left[(x_{ij,s} - \mu_{ij,s})^2 \mathbb{1}_{\{|x_{ij,s}-\mu_{ij,s}| > \epsilon S_{ij,D}\}}\right] = 0,$$

where $S_{ij,D}^2 = \sum\limits_{s=1}^{D} \mathrm{Var}(x_{ij,s}) = K_{ii}K_{jj} \sum\limits_{s=1}^{D} p_{is}p_{js} + K_{ij}^2 \sum\limits_{s=1}^{D}(2p_{is}p_{js} - p_{is}^2 p_{js}^2)$, and $\mathbb{1}_{\{\cdot\}}$ is the indicator function. Based on the law of total expectation, we get

$$\begin{aligned}
\frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} \mathrm{E}\left[(x_{ij,s} - \mu_{ij,s})^2 \mathbb{1}_{\{|x_{ij,s}-\mu_{ij,s}| > \epsilon S_{ij,D}\}}\right] &= \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D}(1 - p_{is}p_{js})\mu_{ij,s}^2 \mathbb{1}_{\{|\mu_{ij,s}| > \epsilon S_{ij,D}\}} \\
+ \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} p_{is}p_{js} \mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js} - \mu_{ij,s})^2 \mathbb{1}_{\{|\tilde{y}_{is}\tilde{y}_{js}-\mu_{ij,s}| > \epsilon S_{ij,D}\}}\right]
\end{aligned} \tag{S1}$$

Since $\sum\limits_{s=1}^{D} p_{is}p_{js} \asymp D$, there exist constants $0 < m < M < \infty$, and an integer n_0 such that

$$m < \frac{\sum_{s=1}^{D} p_{is}p_{js}}{D} < M, \text{ for all } D > n_0.$$

Furthermore, we obtain

$$\begin{aligned}\frac{p_{is}^2 p_{js}^2 K_{ij}^2}{S_{ij,D}^2} &\le \frac{p_{is}^2 p_{js}^2 K_{ij}^2}{\left(K_{ii}K_{jj}+K_{ij}^2\right)\sum_{s=1} p_{is}p_{js}} \\ &< \frac{p_{is}^2 p_{js}^2 K_{ij}^2}{\left(K_{ii}K_{jj}+K_{ij}^2\right) mD} \le \frac{1}{mD\left(\frac{K_{ii}K_{jj}}{K_{ij}^2}+1\right)},\ for\ D > n_0.\end{aligned}$$

Let $n = \max\left\{\frac{1}{\epsilon^2 m\left(\frac{K_{ii}K_{jj}}{K_{ij}^2}+1\right)}, n_0\right\}$, we get $\frac{p_{is}^2 p_{js}^2 K_{ij}^2}{S_{ij,D}^2} < \epsilon^2,$ for $D > n$. It is hence clear that

$$\lim_{D\to\infty} \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} (1 - p_{is}p_{js}) \mu_{ij,s}^2 \mathbb{1}_{\{|\mu_{ij,s}| > \epsilon S_{ij,D}\}} = 0. \tag{S2}$$

Without loss of generality, suppose that $K_{ij} \ge 0$, with $0 < p_{is} \le 1$ for any i and s, we get $\epsilon S_{ij,D} + p_{is}p_{js}K_{ij} > \epsilon S_{ij,D}$ and $p_{is}p_{js}K_{ij} - \epsilon S_{ij,D} \le K_{ij} - \epsilon S_{ij,D}$. Since $(\tilde{y}_{is}\tilde{y}_{js} - \mu_{ij,s})^2 \ge 0$, we obtain that

$$\begin{aligned} 0 &\le \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} p_{is}p_{js} \mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js} - \mu_{ij,s})^2 \mathbb{1}_{\{|\tilde{y}_{is}\tilde{y}_{js} - \mu_{ij,s}| > \epsilon S_{ij,D}\}}\right] \\ &\le \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} p_{is}p_{js} \mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js} - \mu_{ij,s})^2 \left(\mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}} + \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} < K_{ij} - \epsilon S_{ij,D}\}}\right)\right] \end{aligned} \tag{S3}$$

To investigate the limit of $\frac{1}{S_{ij,D}^2} \sum\limits_{s=1}^{D} p_{is}p_{js} \mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js} - \mu_{ij,s})^2 \mathbb{1}_{\{|\tilde{y}_{is}\tilde{y}_{js} - \mu_{ij,s}| > \epsilon S_{ij,D}\}}\right]$in Equation (3), we first investigate the limit of

$$\begin{aligned} &\frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} p_{is}p_{js} \mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js} - \mu_{ij,s})^2 \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right] \\ &= \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} p_{is}p_{js} \mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js})^2 \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right] + \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} p_{is}^3 p_{js}^3 K_{ij}^2 \mathrm{E}\left[\mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right] \\ &- \frac{2}{S_{ij,D}^2} \sum_{s=1}^{D} p_{is}p_{js} K_{ij} \mathrm{E}\left[\tilde{y}_{is}\tilde{y}_{js} \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right]. \end{aligned} \tag{S4}$$

The limit of above equation can be obtained by showing that the upper bound for each term goes to 0. Based on the fact that $\tilde{y}_{i1}\tilde{y}_{j1},\ \tilde{y}_{i2}\tilde{y}_{j2},\ \ldots$ are iid, the upper bounds for terms in Equation 4 are respectively given by

$$\begin{aligned} 0 &\le \frac{\left(\sum_{s=1}^{D} p_{is}p_{js}\right) \mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js})^2 \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right]}{S_{ij,D}^2} \le \frac{\mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js})^2 \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right]}{K_{ii}K_{jj} + K_{ij}^2}, \\ 0 &\le \frac{K_{ij}^2 \left(\sum_{s=1}^{D} p_{is}^3 p_{js}^3\right) \mathrm{E}\left[\mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right]}{S_{ij,D}^2} \le K_{ij}^2 \frac{\mathrm{E}\left[\mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right]}{K_{ii}K_{jj} + K_{ij}^2}, \\ 0 &\le \frac{K_{ij} \left(\sum_{s=1}^{D} p_{is}^2 p_{js}^2\right) \mathrm{E}\left[\tilde{y}_{is}\tilde{y}_{js} \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right]}{S_{ij,D}^2} \le K_{ij} \frac{\mathrm{E}\left[\tilde{y}_{is}\tilde{y}_{js} \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right]}{K_{ii}K_{jj} + K_{ij}^2}. \end{aligned} \tag{S5}$$

Since $\left|\tilde{y}_{is}^2\tilde{y}_{js}^2 \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right| \le \tilde{y}_{is}^2\tilde{y}_{js}^2$, $\left|\tilde{y}_{is}\tilde{y}_{js} \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right| \le |\tilde{y}_{is}\tilde{y}_{js}|$, and

$$\mathrm{E}\left[\tilde{y}_{is}^2\tilde{y}_{js}^2\right], \mathrm{E}\left[|\tilde{y}_{is}\tilde{y}_{js}|\right] < \infty,$$

the dominated convergence theorem implies that

$$\lim_{D\to\infty} \frac{\mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js})^2 \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right]}{K_{ii}K_{jj} + K_{ij}^2} = 0,\ \lim_{D\to\infty} K_{ij} \frac{\mathrm{E}\left[\tilde{y}_{is}\tilde{y}_{js} \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js} > \epsilon S_{ij,D}\}}\right]}{K_{ii}K_{jj} + K_{ij}^2} = 0.$$

Moreover, since $\mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js}>\epsilon S_{ij,D}\}} \xrightarrow{P} 0$, it is clear that $K_{ij}^2 \frac{\mathrm{E}\left[\mathbb{1}_{\left\{\tilde{y}_{is}\tilde{y}_{js}>\epsilon S_{ij,D}\right\}}\right]}{K_{ii}K_{jj}+K_{ij}^2}$ goes to 0 based on the bounded convergence theorem. Thus, the limits of all upper bounds shown in Equation (5) are 0 and we get

$$\lim_{D\to\infty} \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} p_{is}p_{js}\mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js} - \mu_{ij,s})^2 \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js}>\epsilon S_{ij,D}\}}\right] = 0.$$

Analogously, we can obtain

$$\lim_{D\to\infty} \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} p_{is}p_{js}\mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js} - \mu_{ij,s})^2 \mathbb{1}_{\{\tilde{y}_{is}\tilde{y}_{js}<K_{ij}-\epsilon S_{ij,D}\}}\right] = 0.$$

Hence, by taking the limits of both sides of inequality provided in Equation 3, we get

$$\lim_{D\to\infty} \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} p_{is}p_{js}\mathrm{E}\left[(\tilde{y}_{is}\tilde{y}_{js} - \mu_{ij,s})^2 \mathbb{1}_{\{|\tilde{y}_{is}\tilde{y}_{js}-\mu_{ij,s}|>\epsilon S_{ij,D}\}}\right] = 0. \tag{S6}$$

Furthermore, by taking the limits of both sides of Equation (1), we get

$$\lim_{D\to\infty} \frac{1}{S_{ij,D}^2} \sum_{s=1}^{D} \mathrm{E}\left[(x_{ij,s} - \mu_{ij,s})^2 \mathbb{1}_{\{|x_{ij,s}-\mu_{ij,s}|>\epsilon S_{ij,D}\}}\right] = 0,$$

based on the limiting results provided in Equation (6) and Equation (2). The Lindeberg's condition is therefore satisfied and $Z_{ij,D} \xrightarrow{dist.} \mathcal{N}(0,1)$ for $i \neq j$ [Lehmann, 2004]. For $i = j$, $Z_{ij,D} \xrightarrow{dist.} \mathcal{N}(0,1)$ can be proved by following the same logic as that for showing $Z_{ij,D} \xrightarrow{dist.} \mathcal{N}(0,1)$ for $i \neq j$.

S.5.4 PROOF FOR COROLLARY 1

For $i \neq j$, $\tilde{G}_{ij} = \frac{Z_{ij,D}S_{ij,D}+\sum_{s=1}^{D}\mu_{ij,s}}{\sum_{s=1}^{D}p_{is}p_{js}} = Z_{ij,D}\frac{S_{ij,D}}{\sum_{s=1}^{D}p_{is}p_{js}} + \frac{\sum_{s=1}^{D}\mu_{ij,D}}{\sum_{s=1}^{D}p_{is}p_{js}}$. Since

$$\frac{S_{ij,D}^2}{\left(\sum_{s=1}^{D}p_{is}p_{js}\right)^2} \leq \frac{K_{ii}K_{jj}}{\sum_{s=1}^{D}p_{is}p_{js}} + K_{ij}^2\left(\frac{2}{\sum_{s=1}^{D}p_{is}p_{js}} - \frac{1}{D}\right) \to 0,$$

as $D \to \infty$;

$$\frac{\sum_{s=1}^{D}\mu_{ij,s}}{\sum_{s=1}^{D}p_{is}p_{js}} = K_{ij},$$

by the Slutsky's theorem, we obtain that $\tilde{G}_{ij} \xrightarrow{P} K_{ij}$ for $i \neq j$. Analogously, we get $\tilde{G}_{ii} \xrightarrow{P} K_{ii}$.

S.5.5 PROOF FOR COROLLARY 2

It is straightforward to get the Corollary by using the Slutsky's theorem and $\frac{G}{D} = \frac{G}{\sum_{s=1}^{D}p_{is}p_{js}} \frac{\sum_{s=1}^{D}p_{is}p_{js}}{D}$.

S.6 MORE EXPERIMENTAL RESULTS

S.6.1 VISUALISATION OF PCA RESULTS ON THE REAL DATASETS

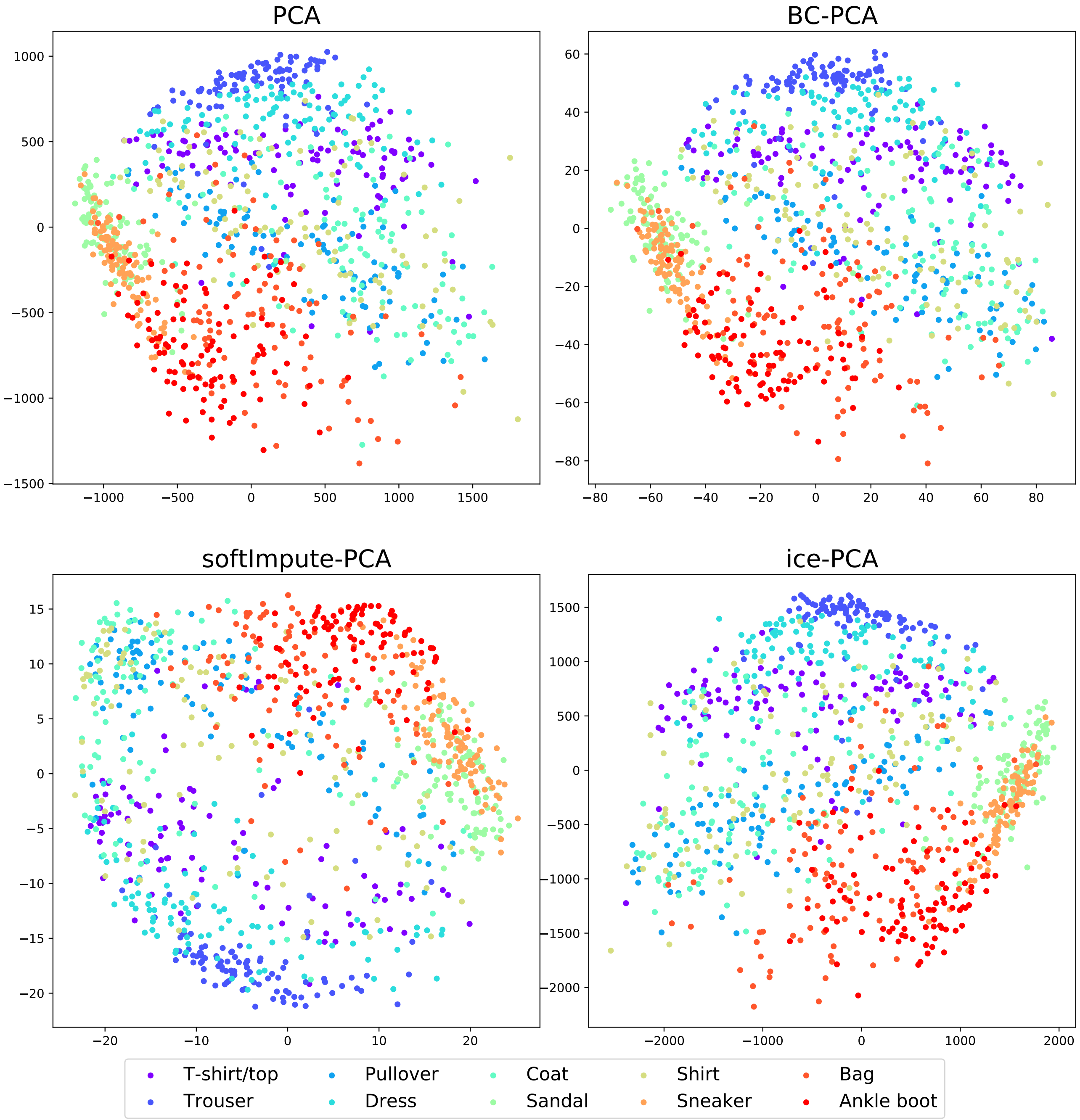


Figure S1: Visualisation of the fashion MNIST dataset obtained by PCA and its variants integrated with the bias correction or imputations.

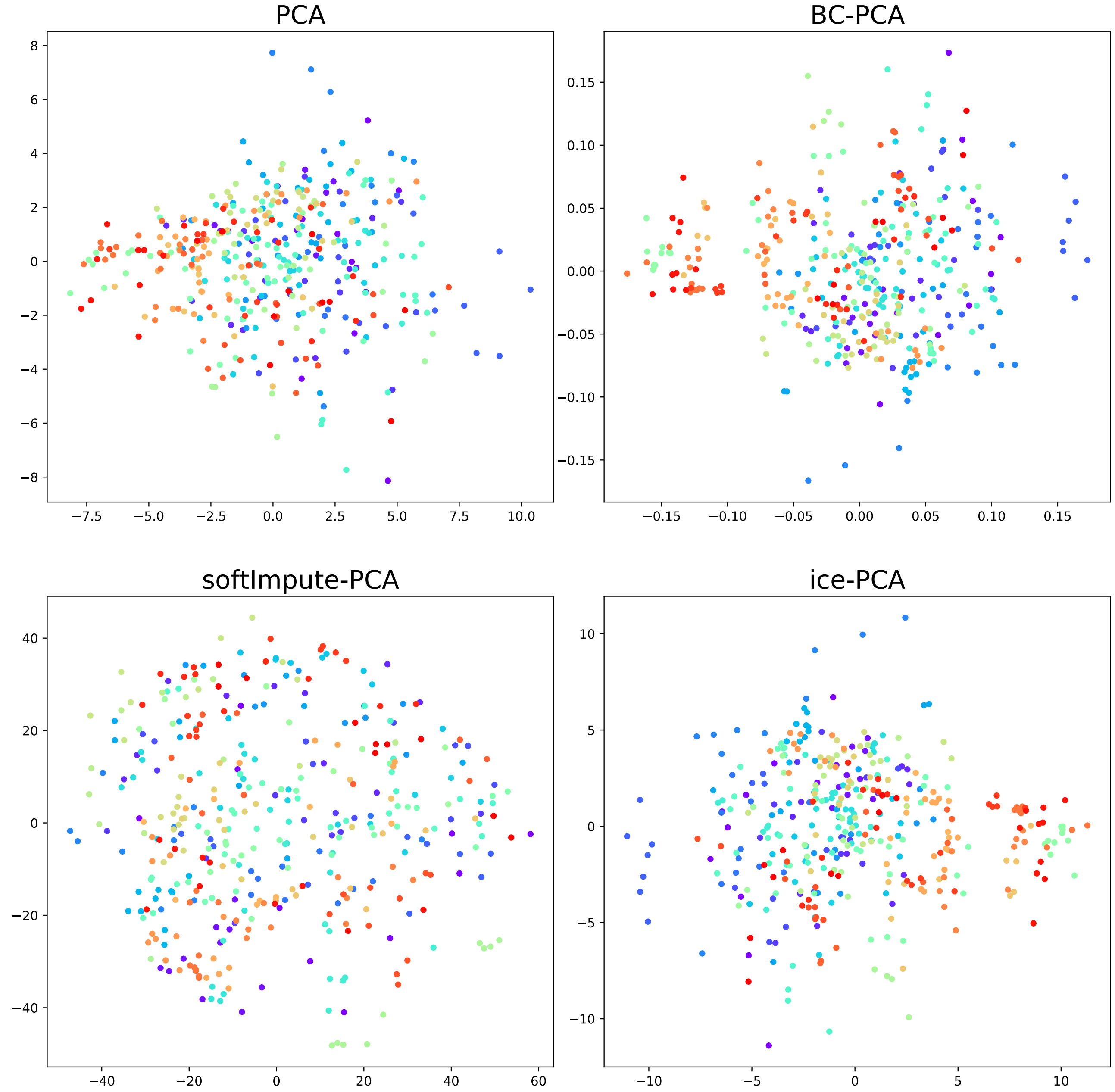


Figure S2: Visualisation of the Olivetti faces dataset obtained by PCA and its variants integrated with the bias correction or imputations. Different colours represent face images of different persons, and there are 40 persons in total.

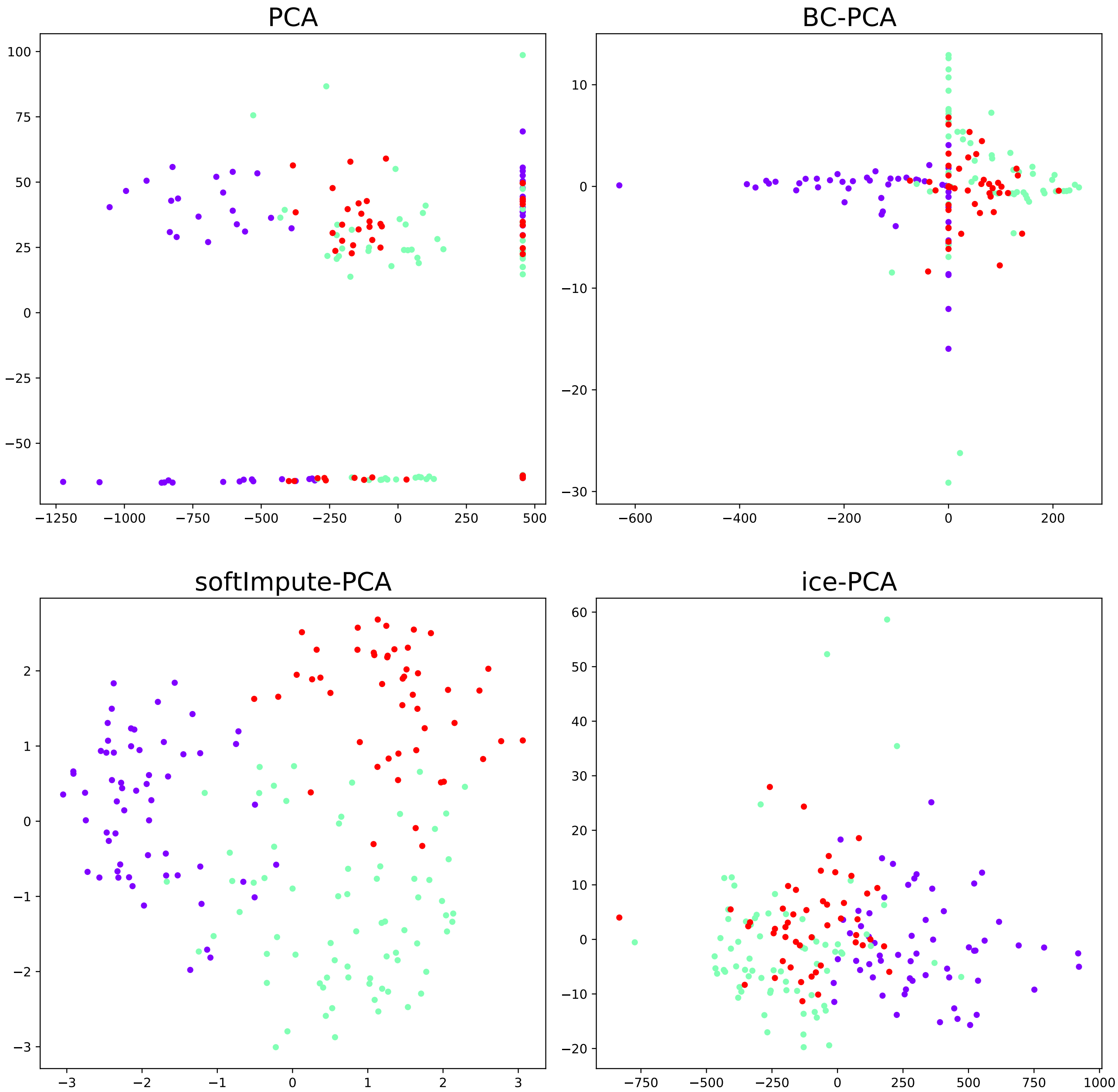


Figure S3: Visualisation of the wine dataset obtained by PCA and its variants integrated with the bias correction or imputations. Different colours represent different classes of wine.

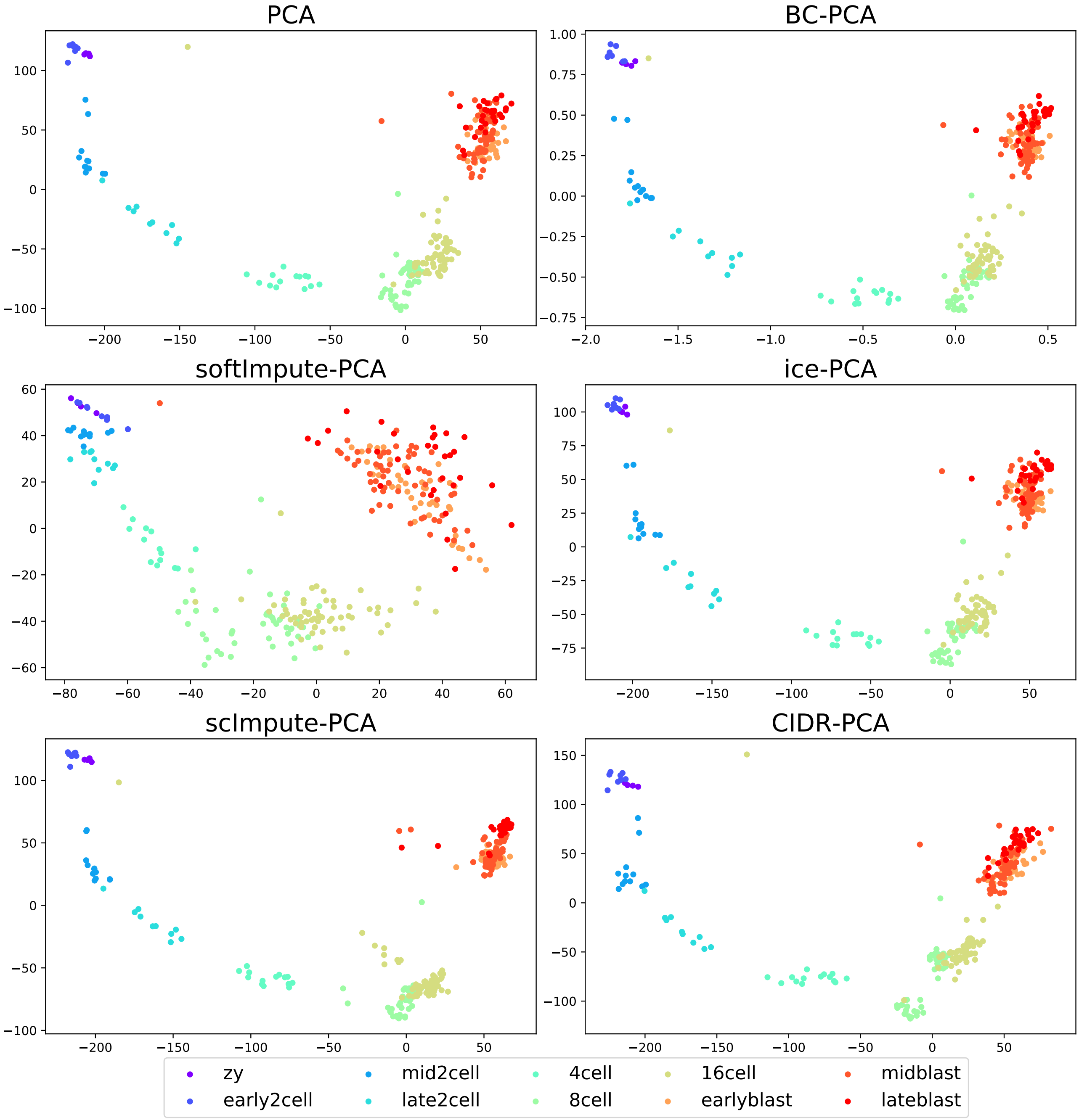


Figure S4: Visualisation of the Deng dataset obtained by PCA and its variants integrated with the bias correction or imputations.

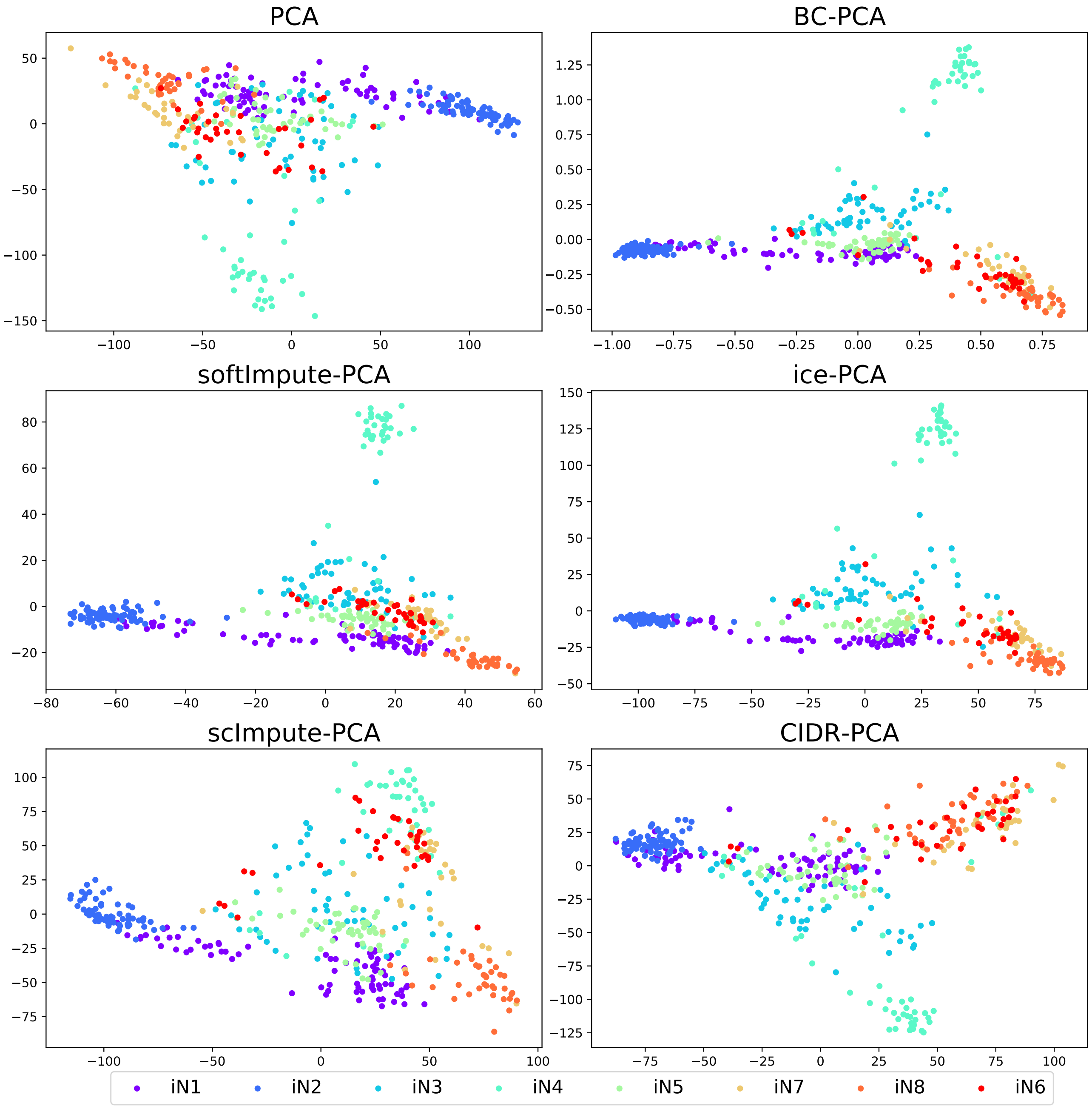


Figure S5: Visualisation of the Treutlein dataset obtained by PCA and its variants integrated with the bias correction or imputations.

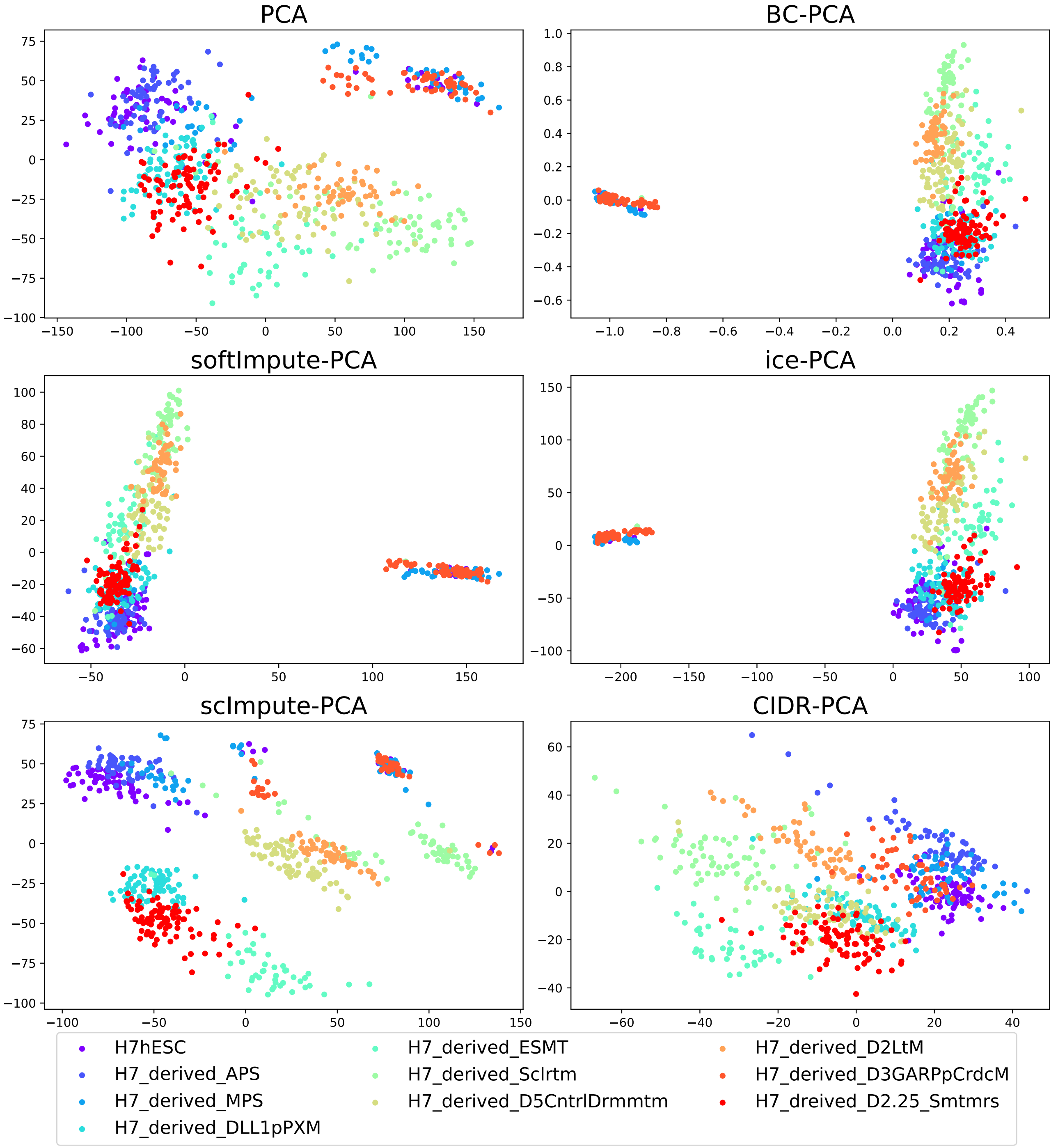


Figure S6: Visualisation of the Koh dataset obtained by PCA and its variants integrated with the bias correction or imputations.

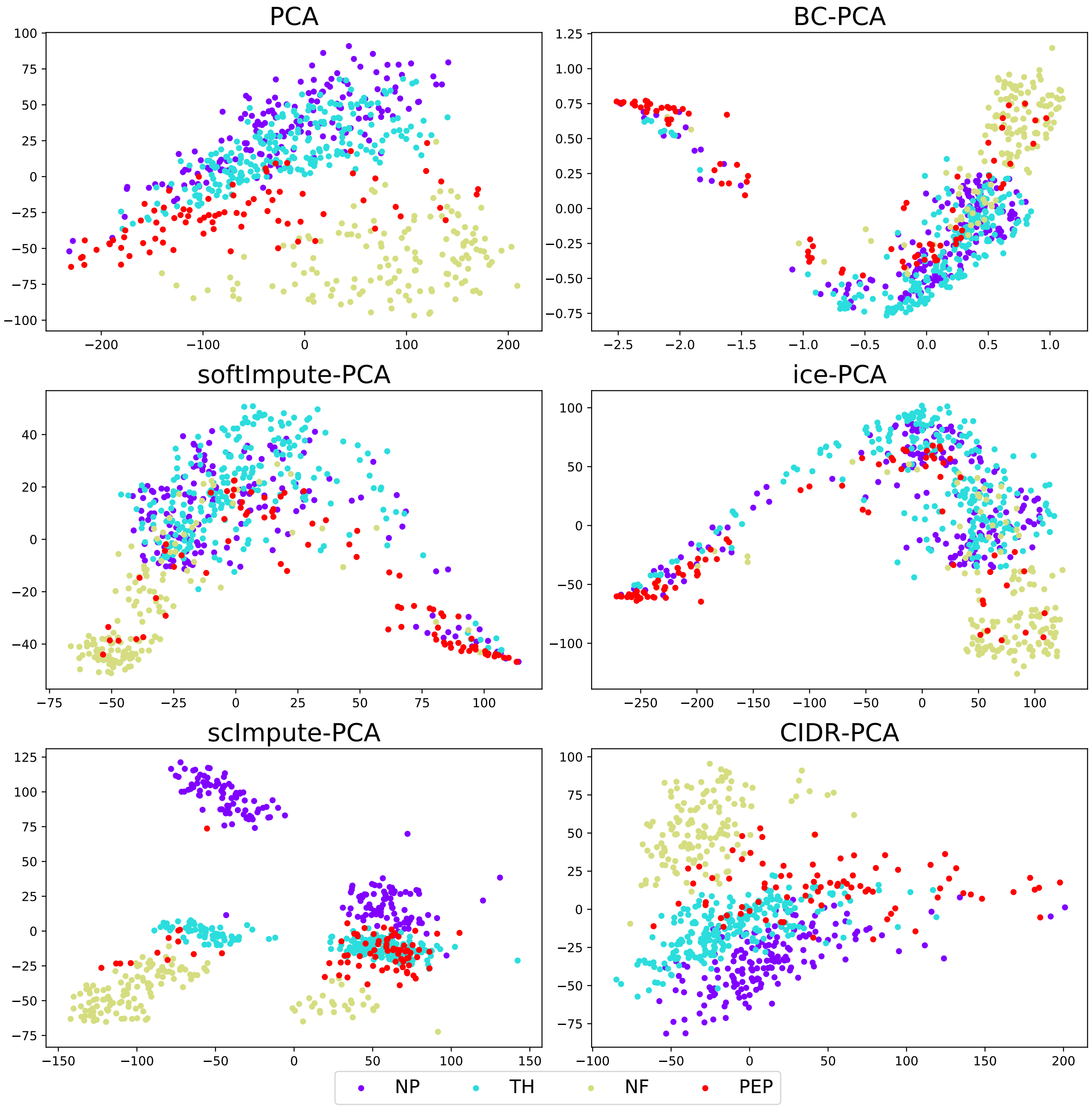


Figure S7: Visualisation of the Usoskin dataset obtained by PCA and its variants integrated with the bias correction or imputations.

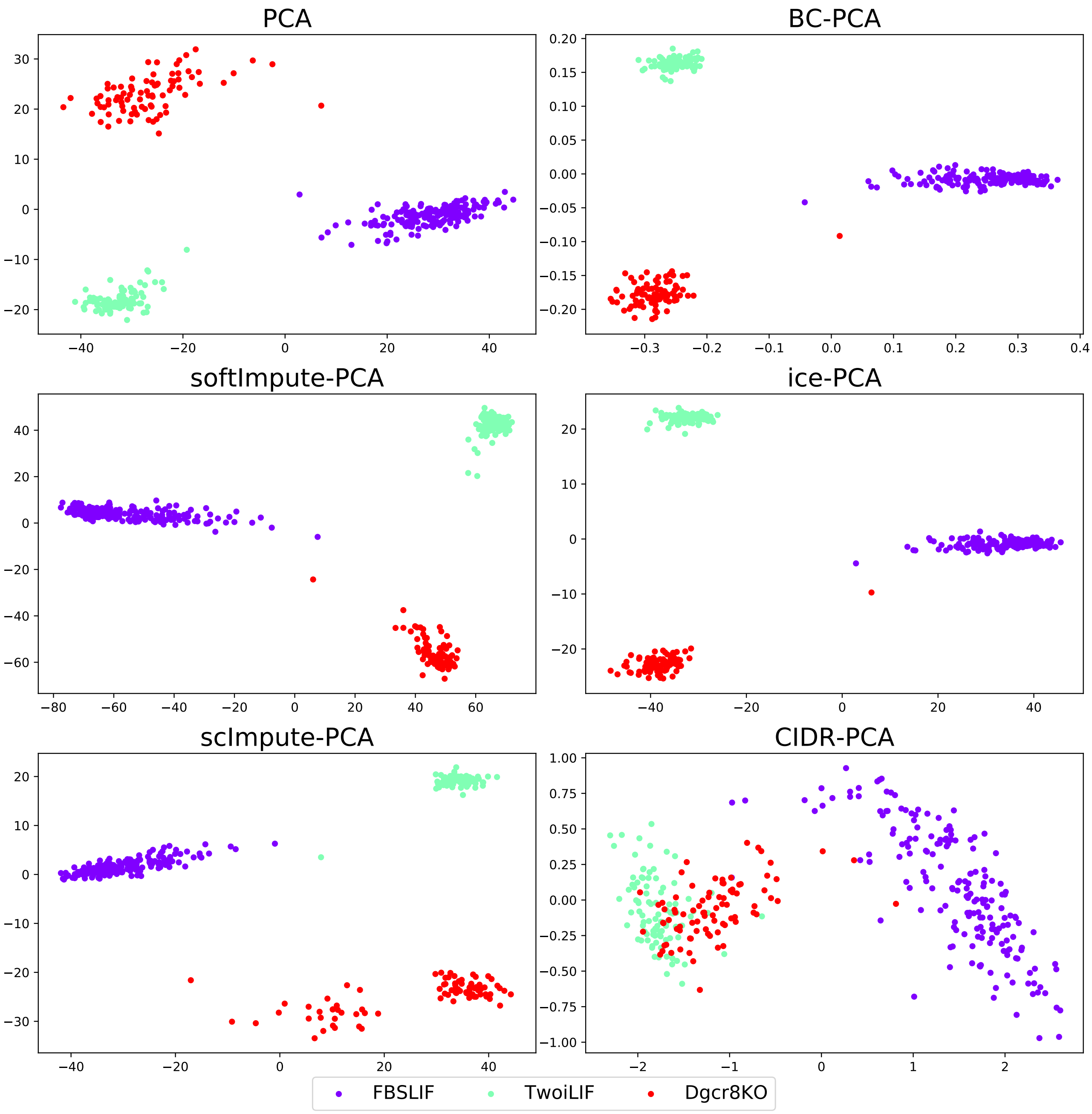


Figure S8: Visualisation of the Kumar dataset obtained by PCA and its variants integrated with the bias correction or imputations.

S.6.2 VISUALISATION OF GPLVM RESULTS ON THE REAL DATASETS

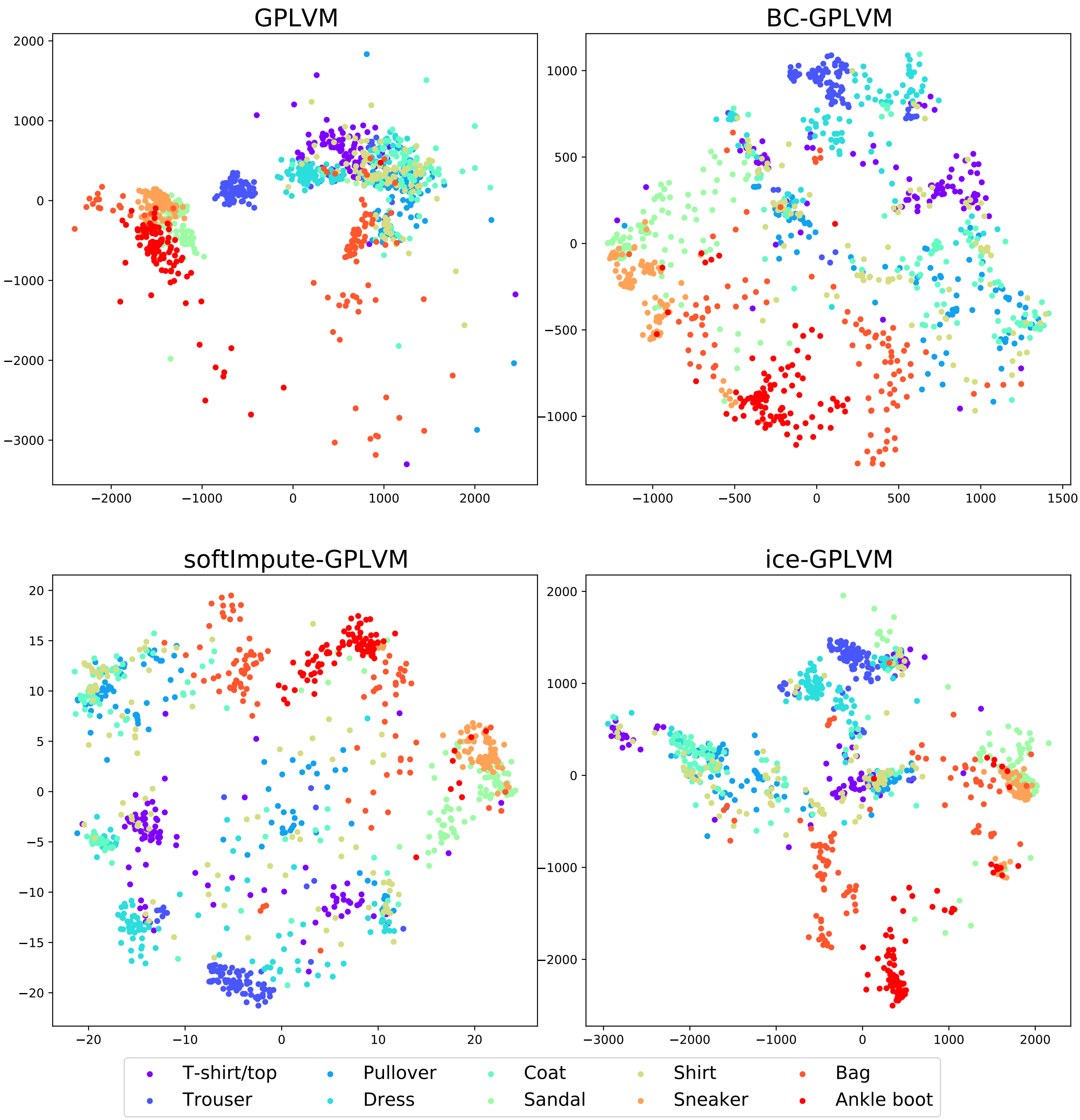


Figure S9: Visualisation of the fashion MNIST dataset obtained by GPLVM and its variants integrated with the bias correction or imputations.

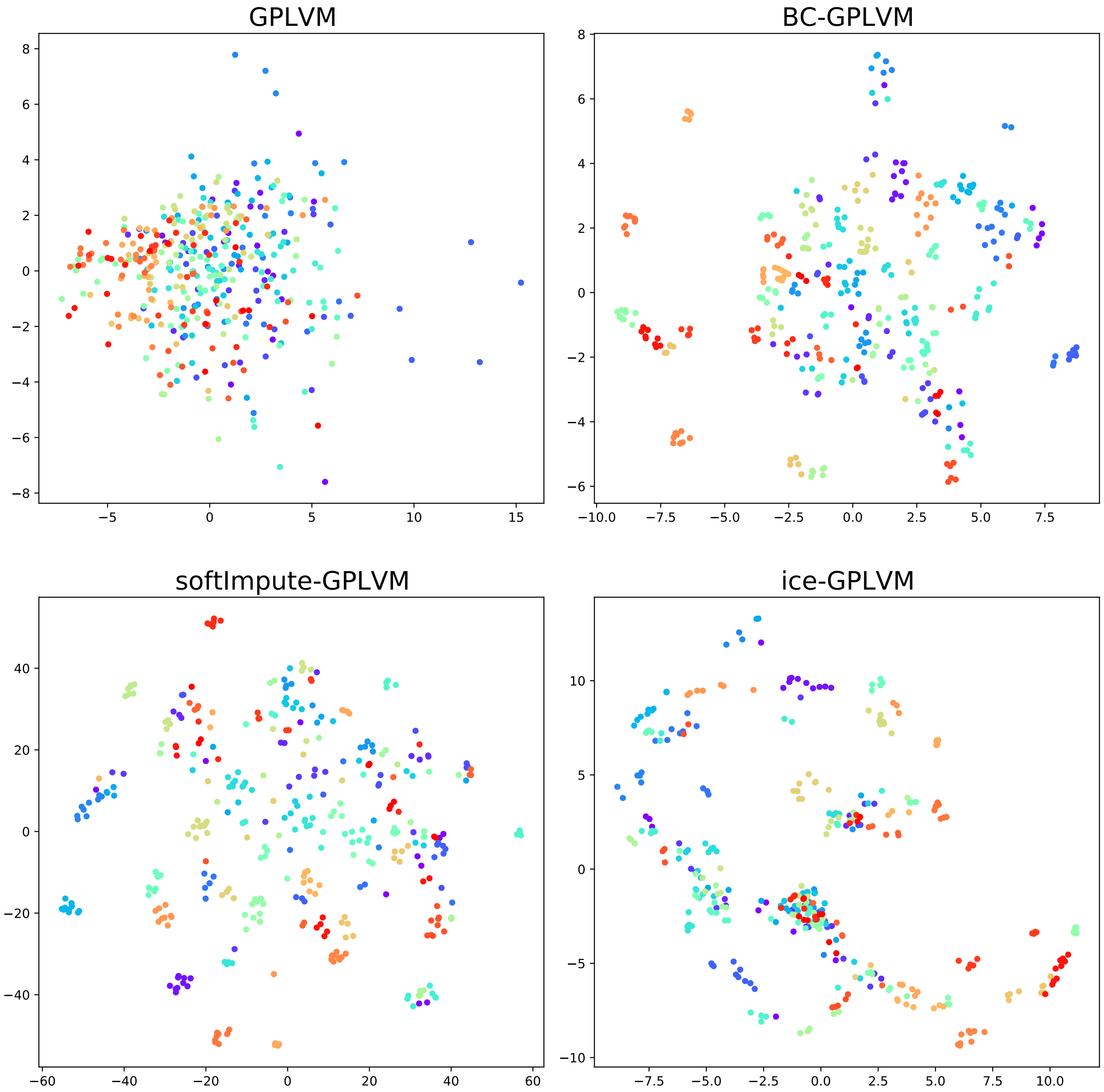


Figure S10: Visualisation of the Olivetti faces dataset obtained by GPLVM and its variants integrated with the bias correction or imputations.

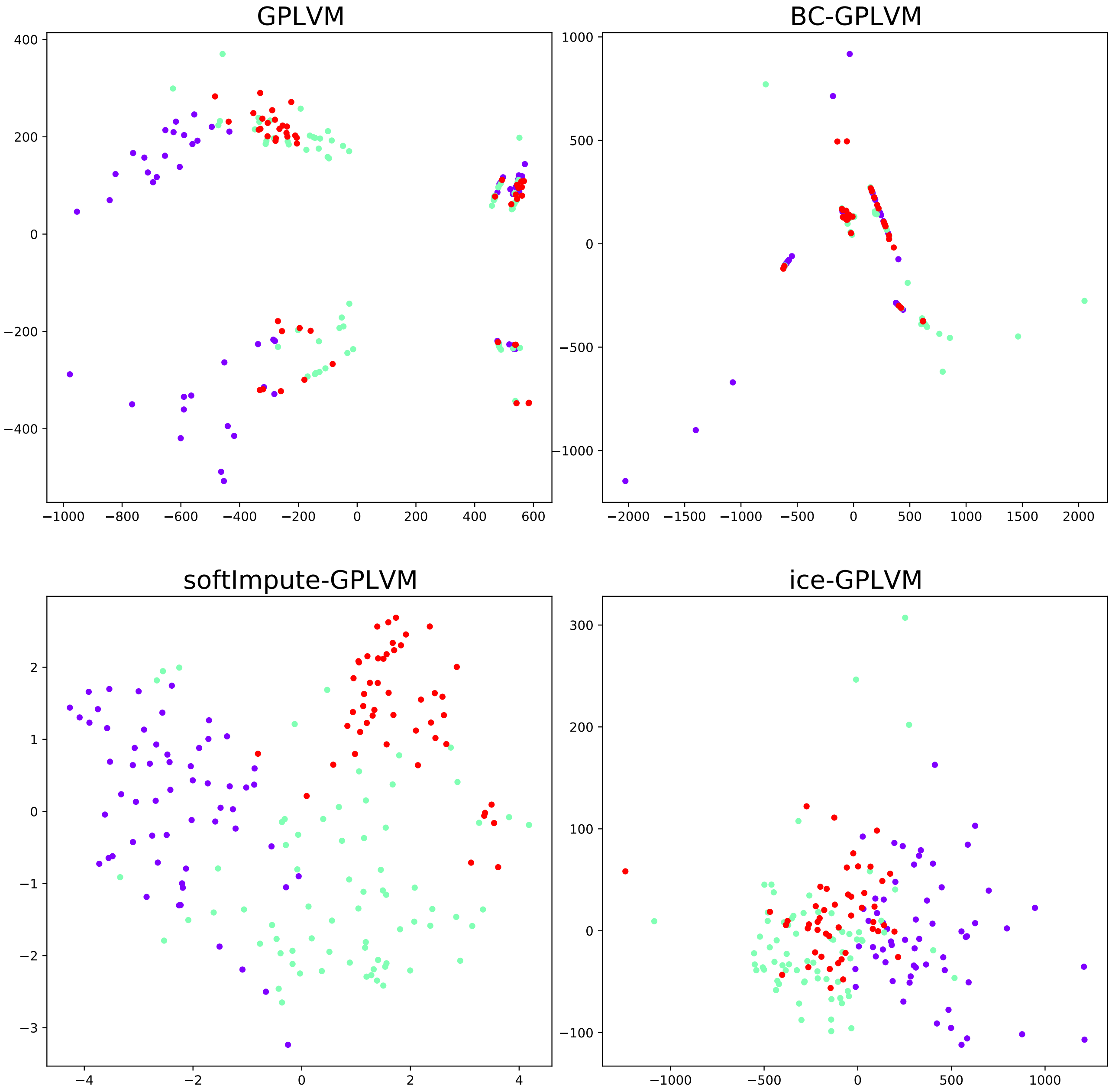


Figure S11: Visualisation of the wine dataset obtained by GPLVM and its variants integrated with the bias correction or imputations.

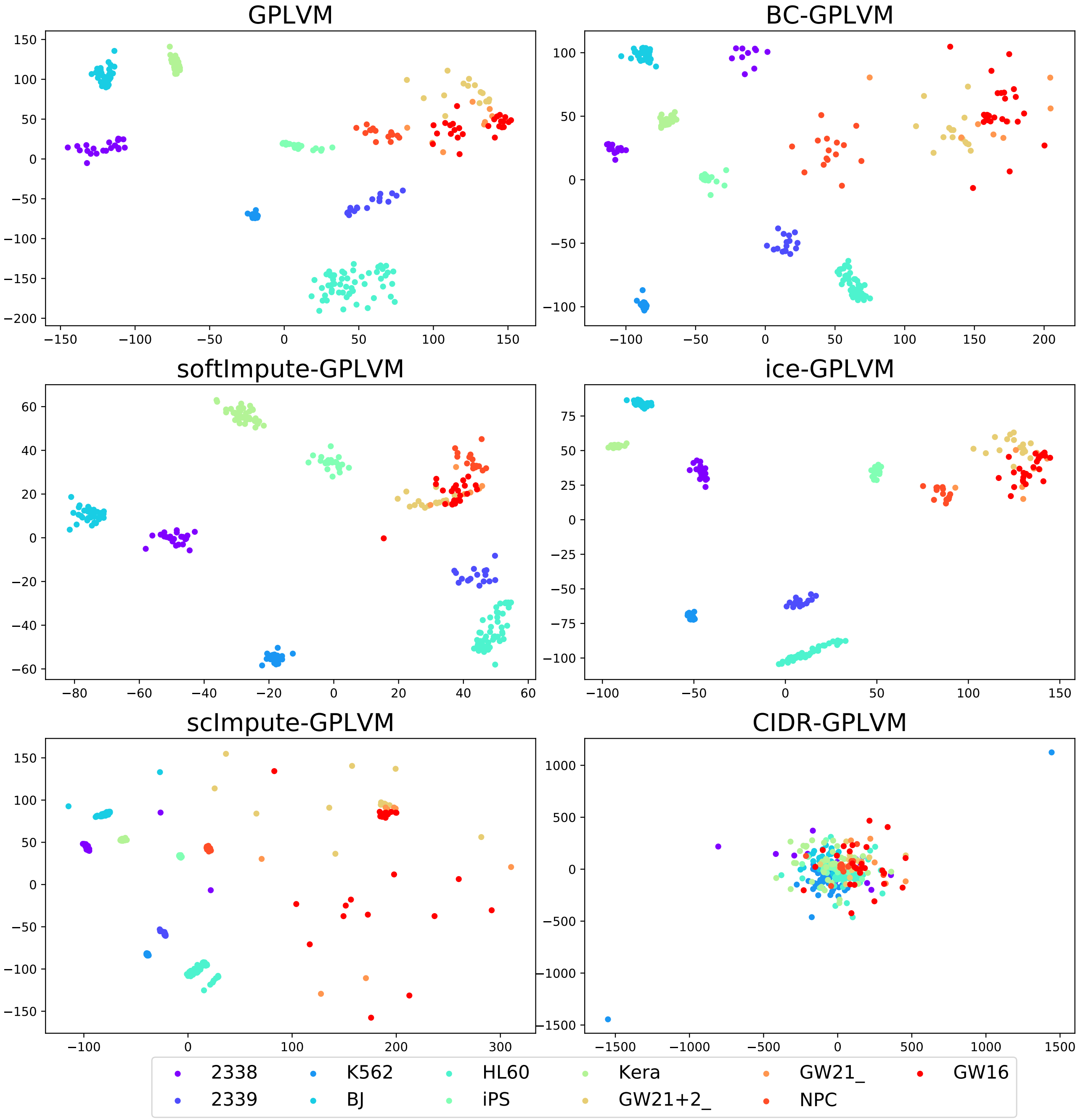


Figure S12: Visualisation of the Pollen dataset obtained by GPLVM and its variants integrated with the bias correction or imputations.

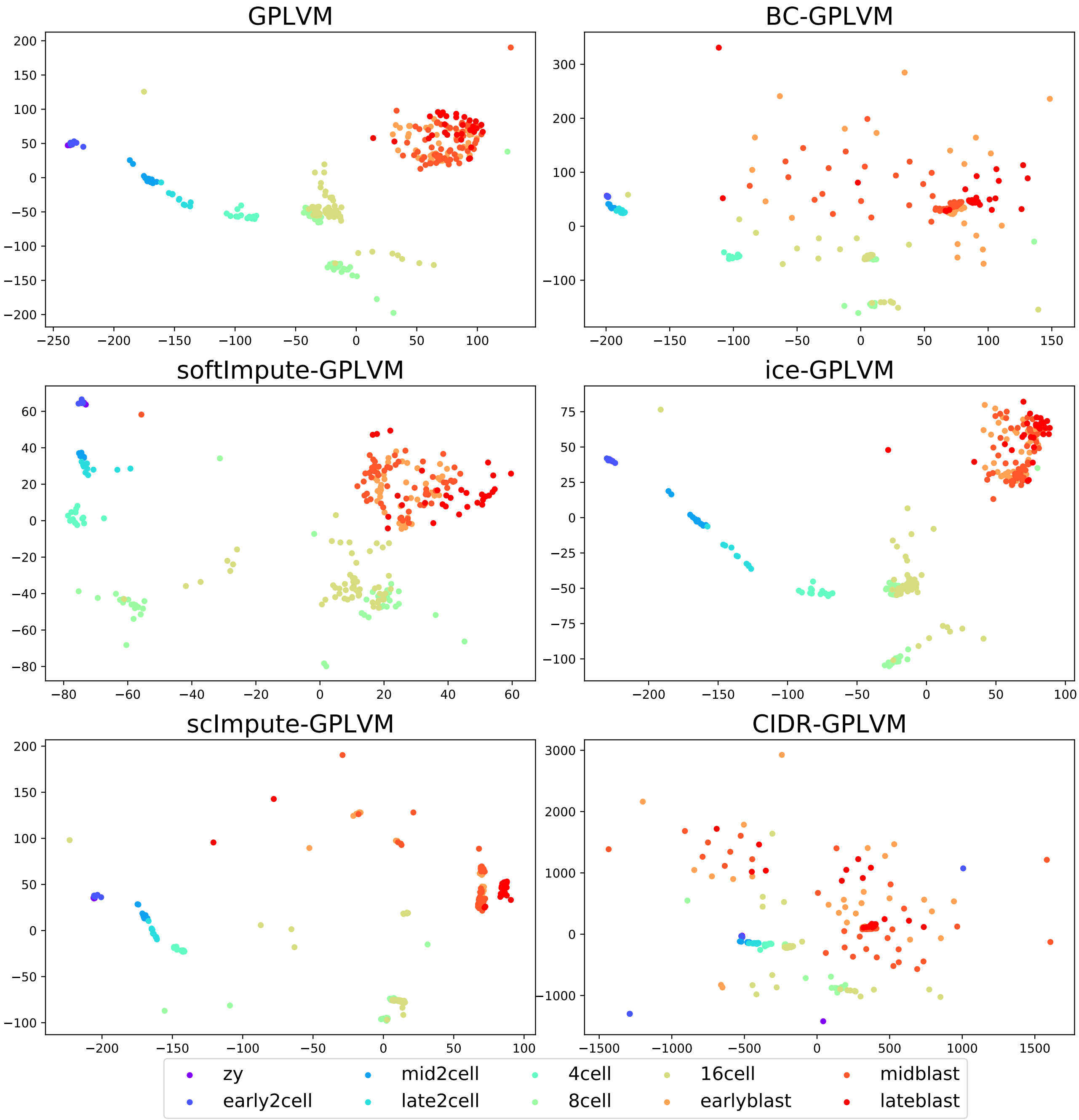


Figure S13: Visualisation of the Deng dataset obtained by GPLVM and its variants integrated with the bias correction or imputations.

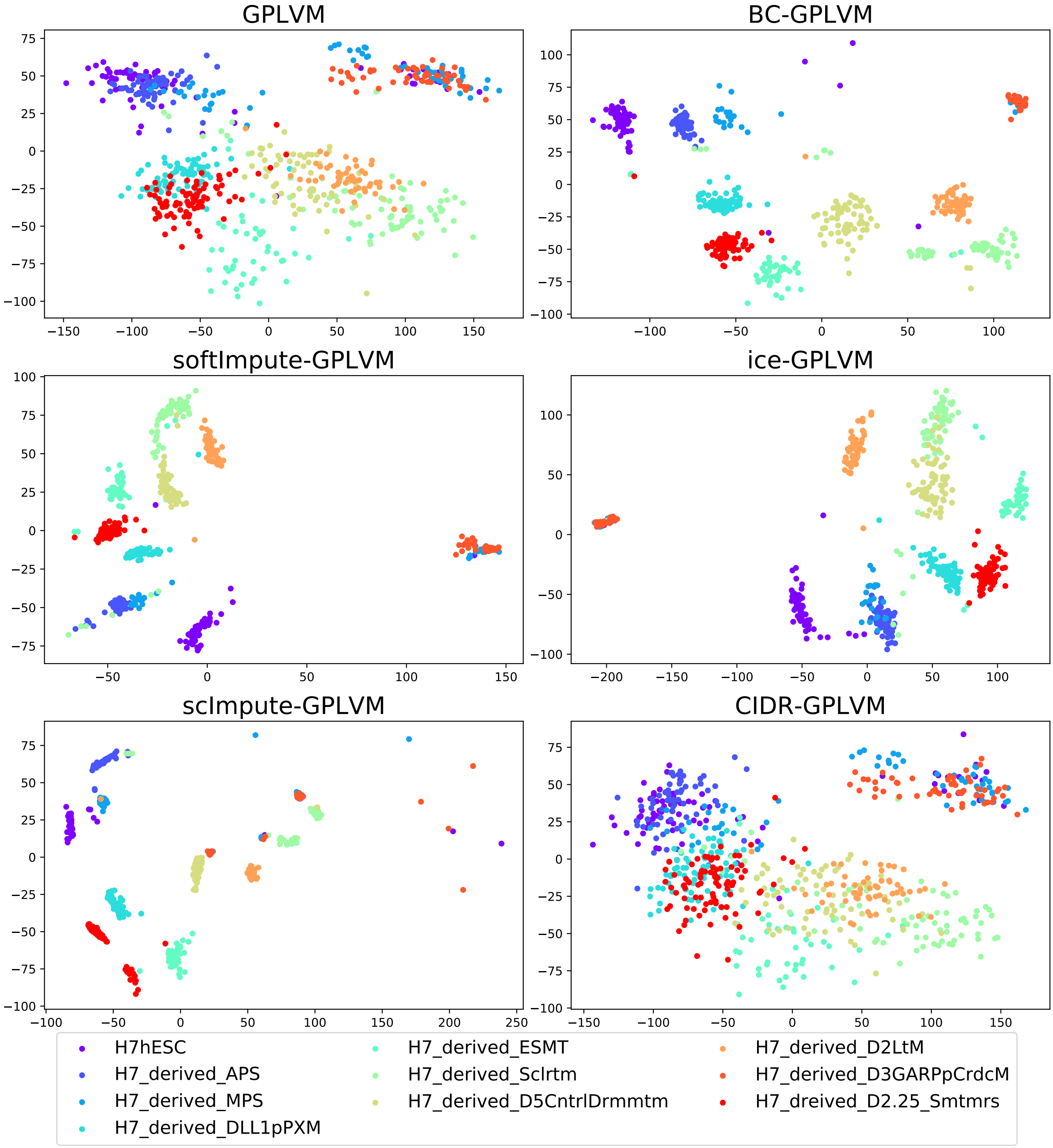


Figure S14: Visualisation of the Koh dataset obtained by GPLVM and its variants integrated with the bias correction or imputations.

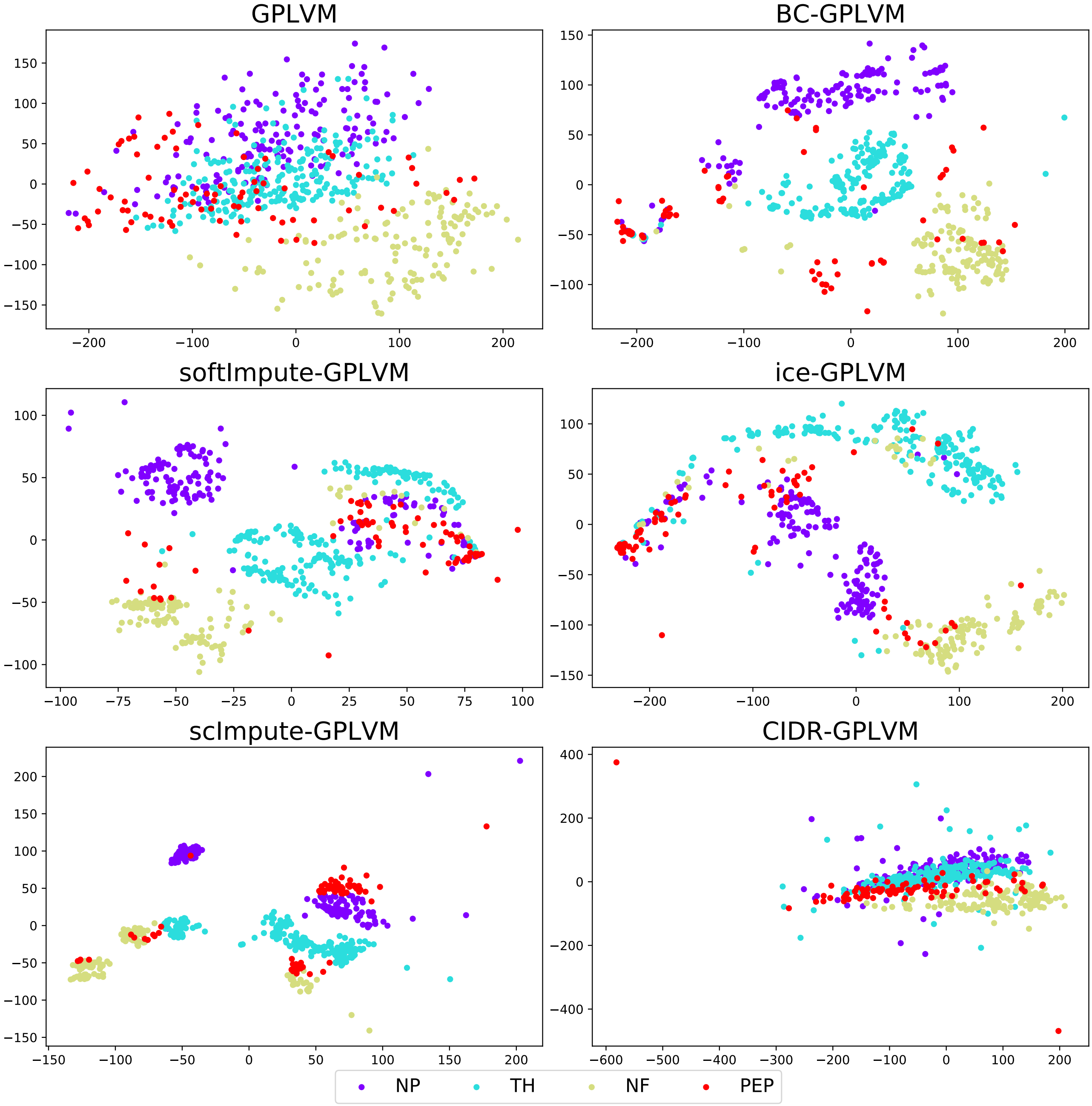


Figure S15: Visualisation of the Usoskin dataset obtained by GPLVM and its variants integrated with the bias correction or imputations.

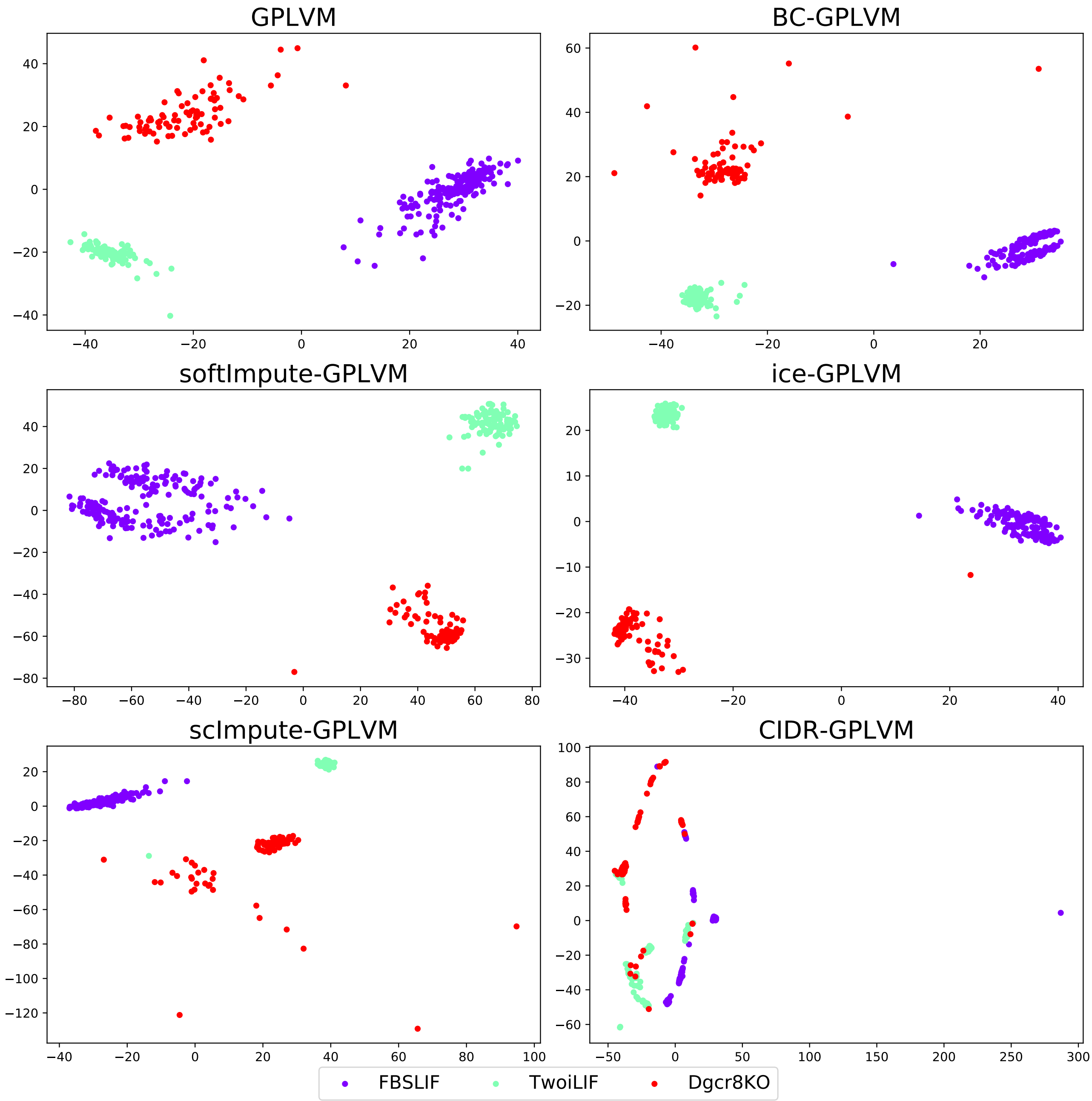


Figure S16: Visualisation of the Kumar dataset obtained by GPLVM and its variants integrated with the bias correction or imputations.

S.6.3 VISUALISATION OF TSNE RESULTS ON THE REAL DATASETS

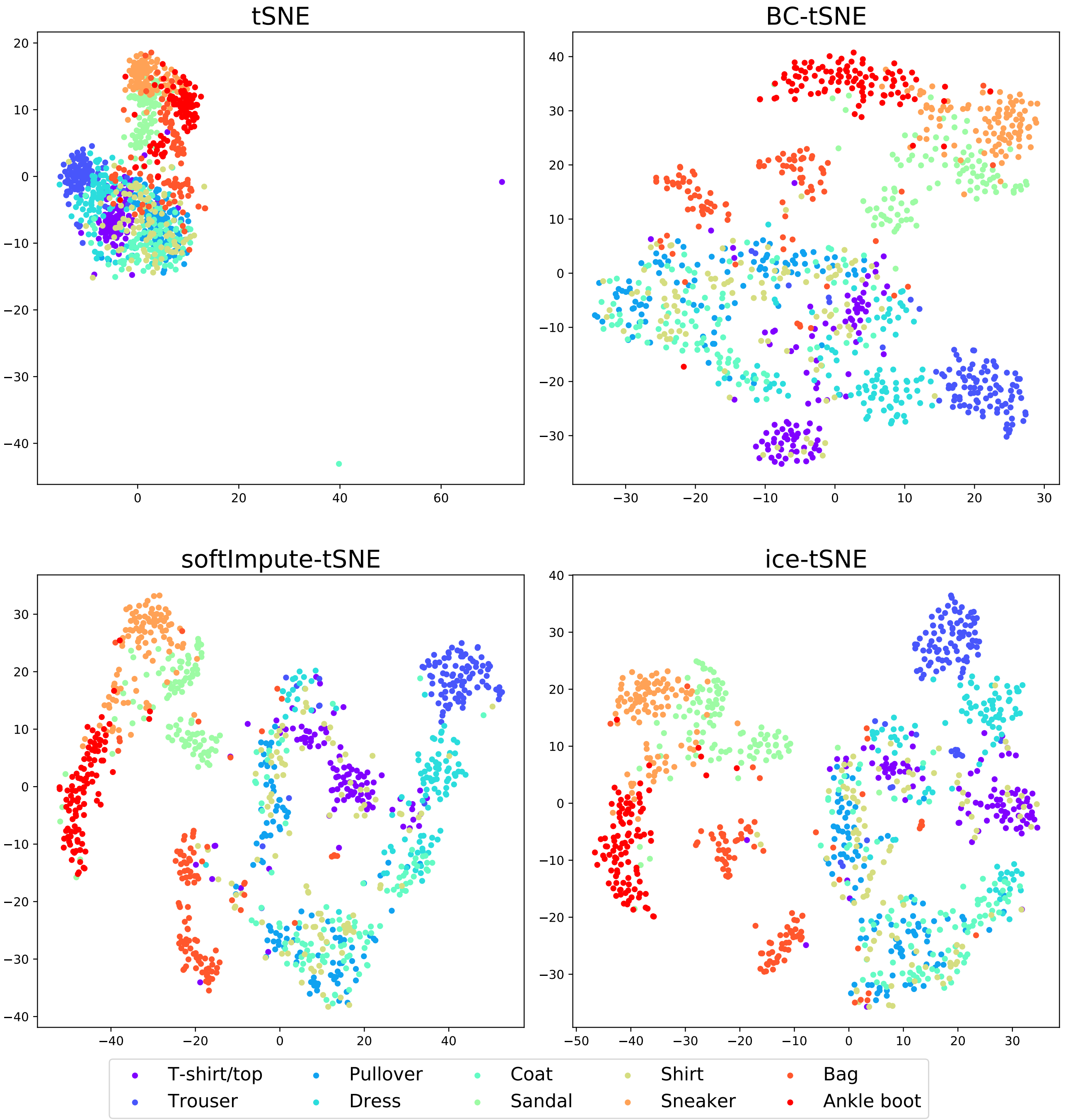


Figure S17: Visualisation of the fashion MNIST dataset obtained by tSNE and its variants integrated with the bias correction or imputations.

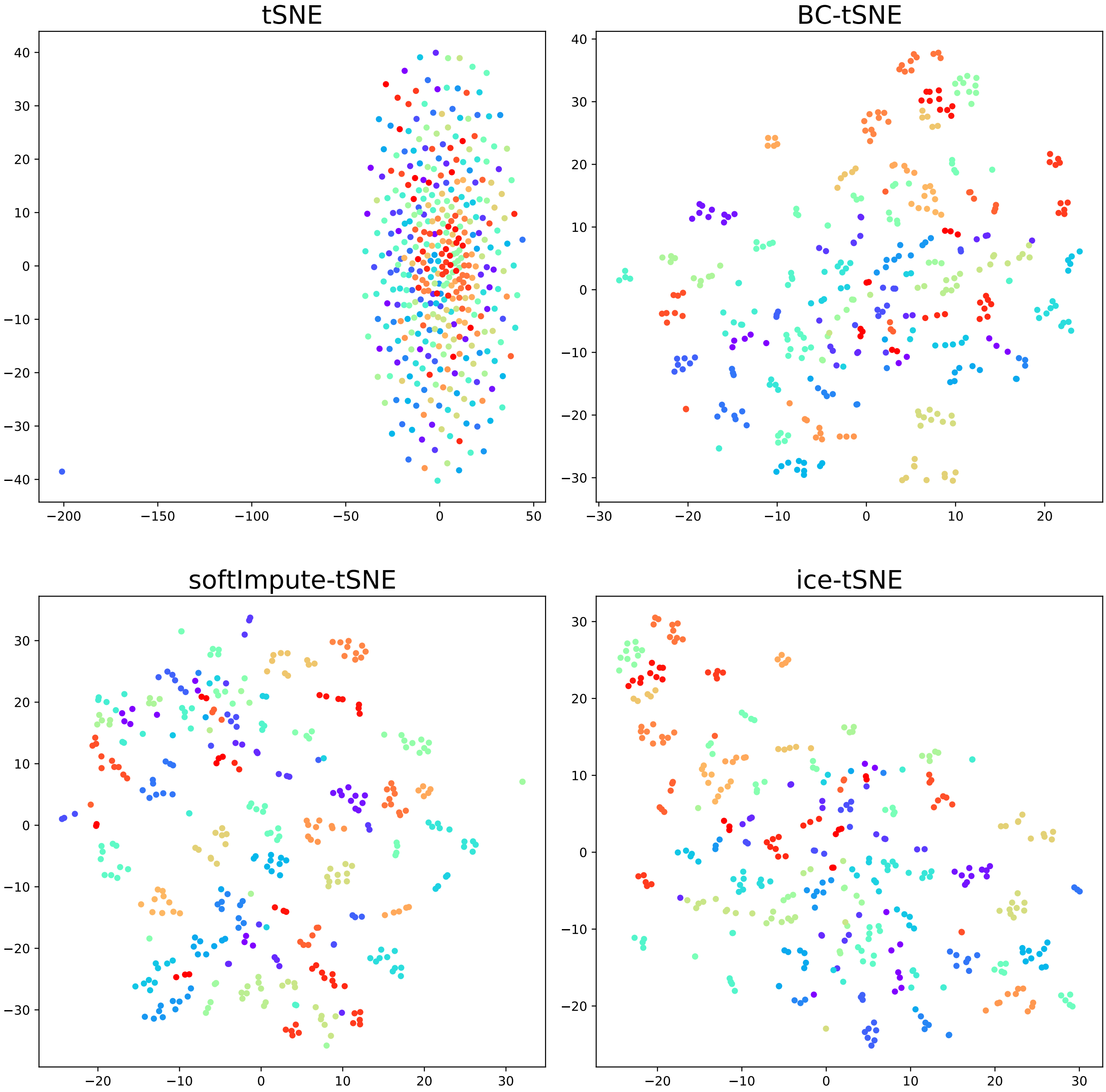


Figure S18: Visualisation of the Olivetti faces dataset obtained by tSNE and its variants integrated with the bias correction or imputations.

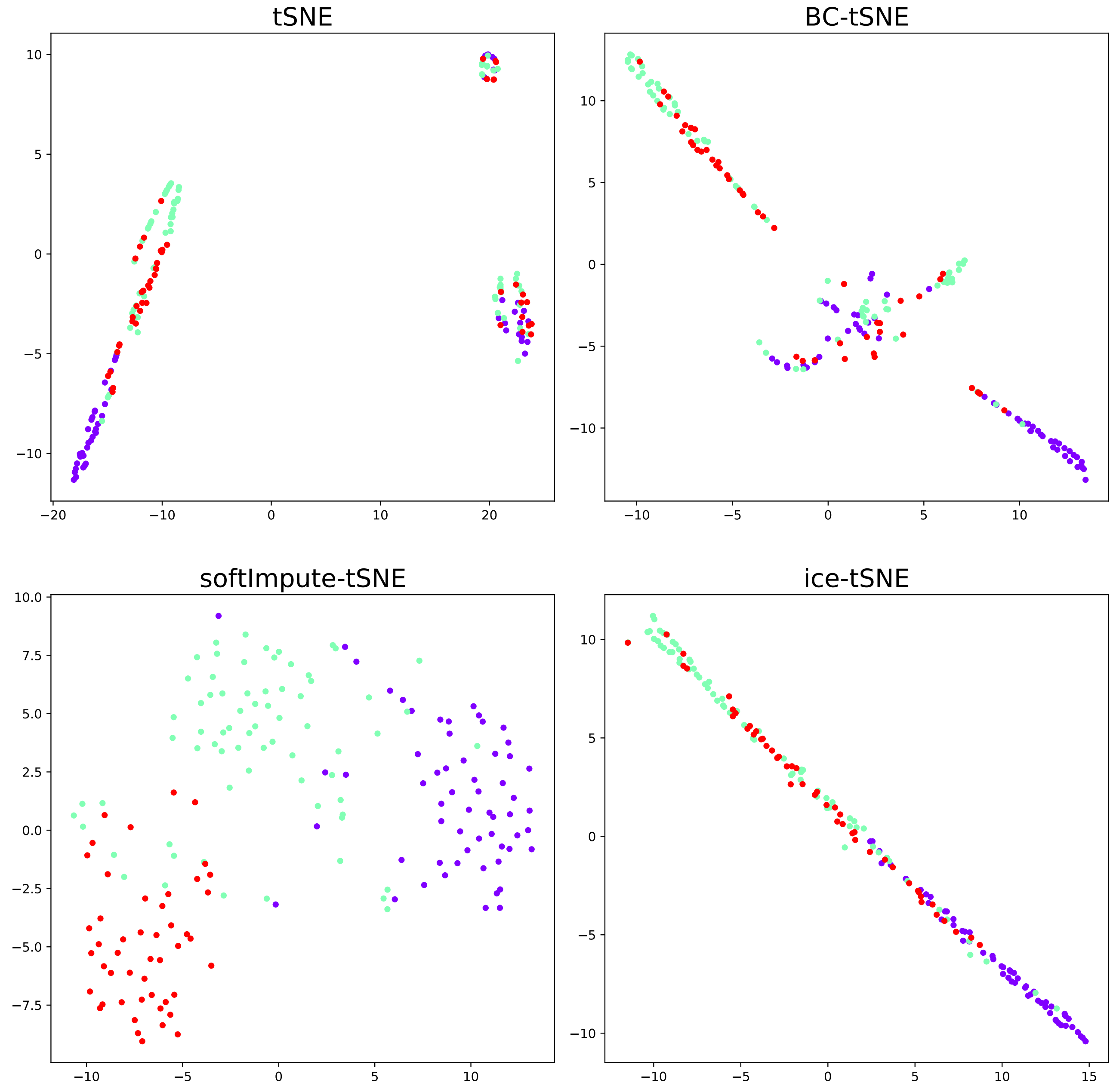


Figure S19: Visualisation of the wine dataset obtained by tSNE and its variants integrated with the bias correction or imputations.

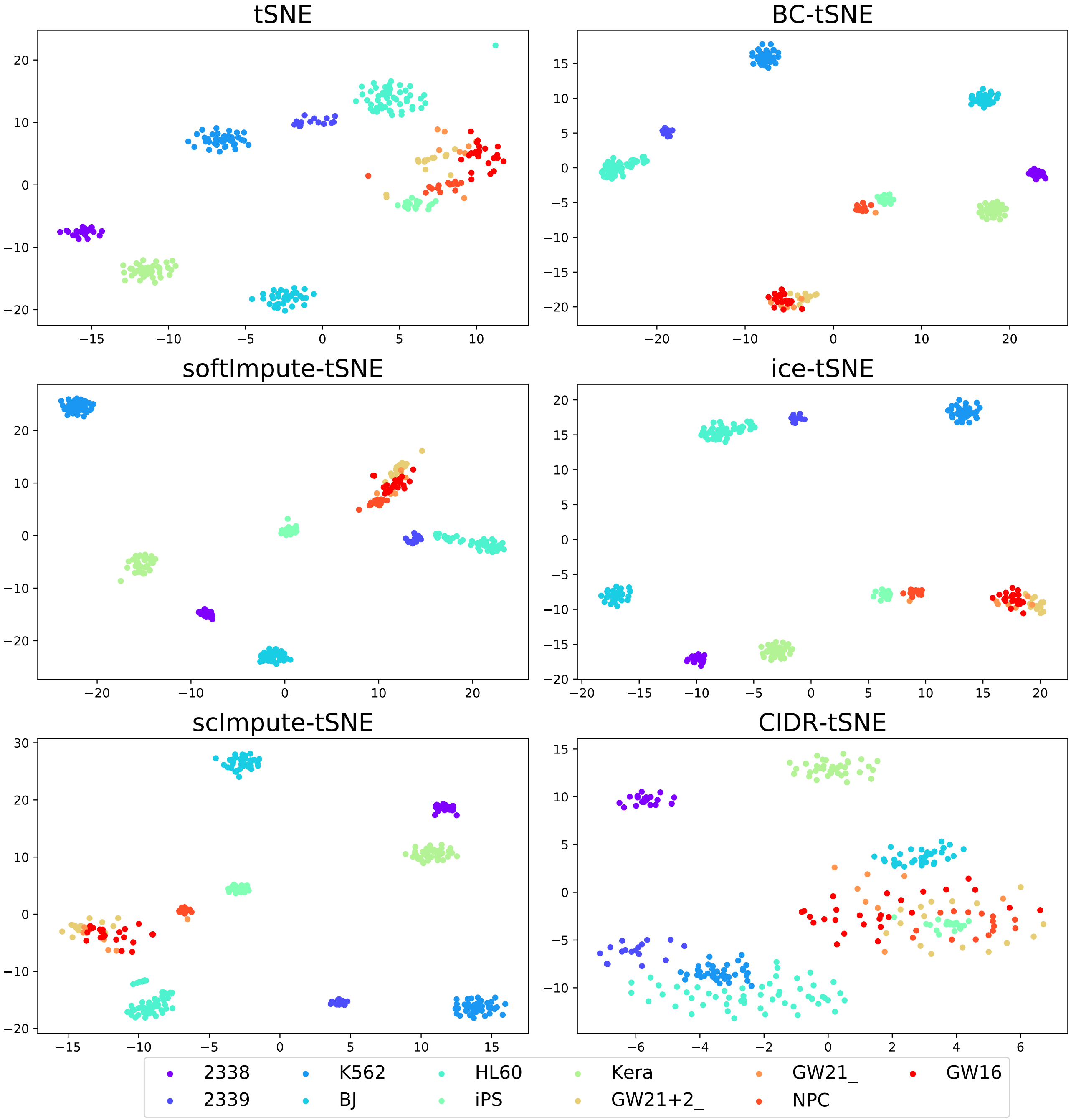


Figure S20: Visualisation of the Pollen dataset obtained by tSNE and its variants integrated with the bias correction or imputations.

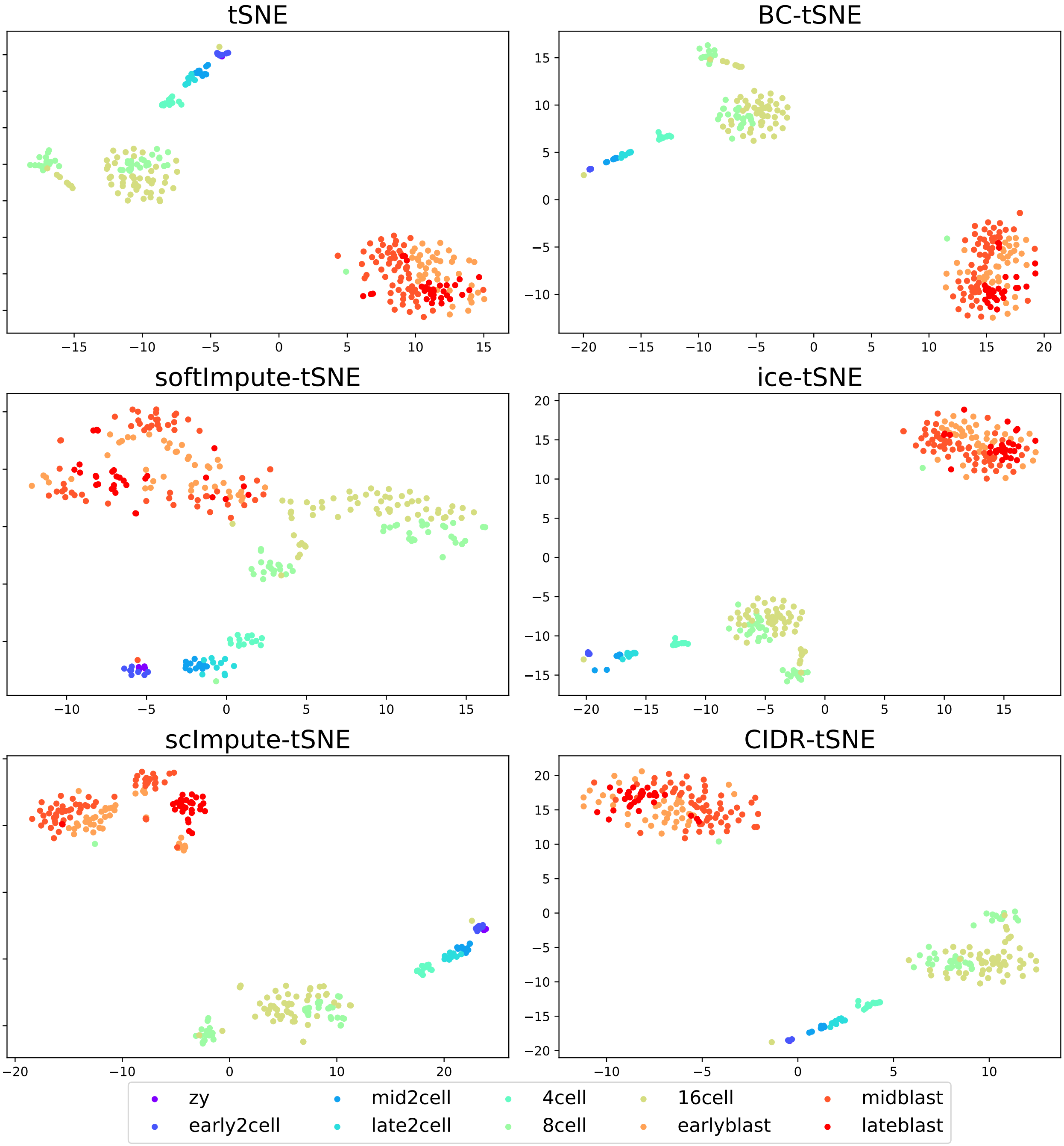


Figure S21: Visualisation of the Deng dataset obtained by tSNE and its variants integrated with the bias correction or imputations.

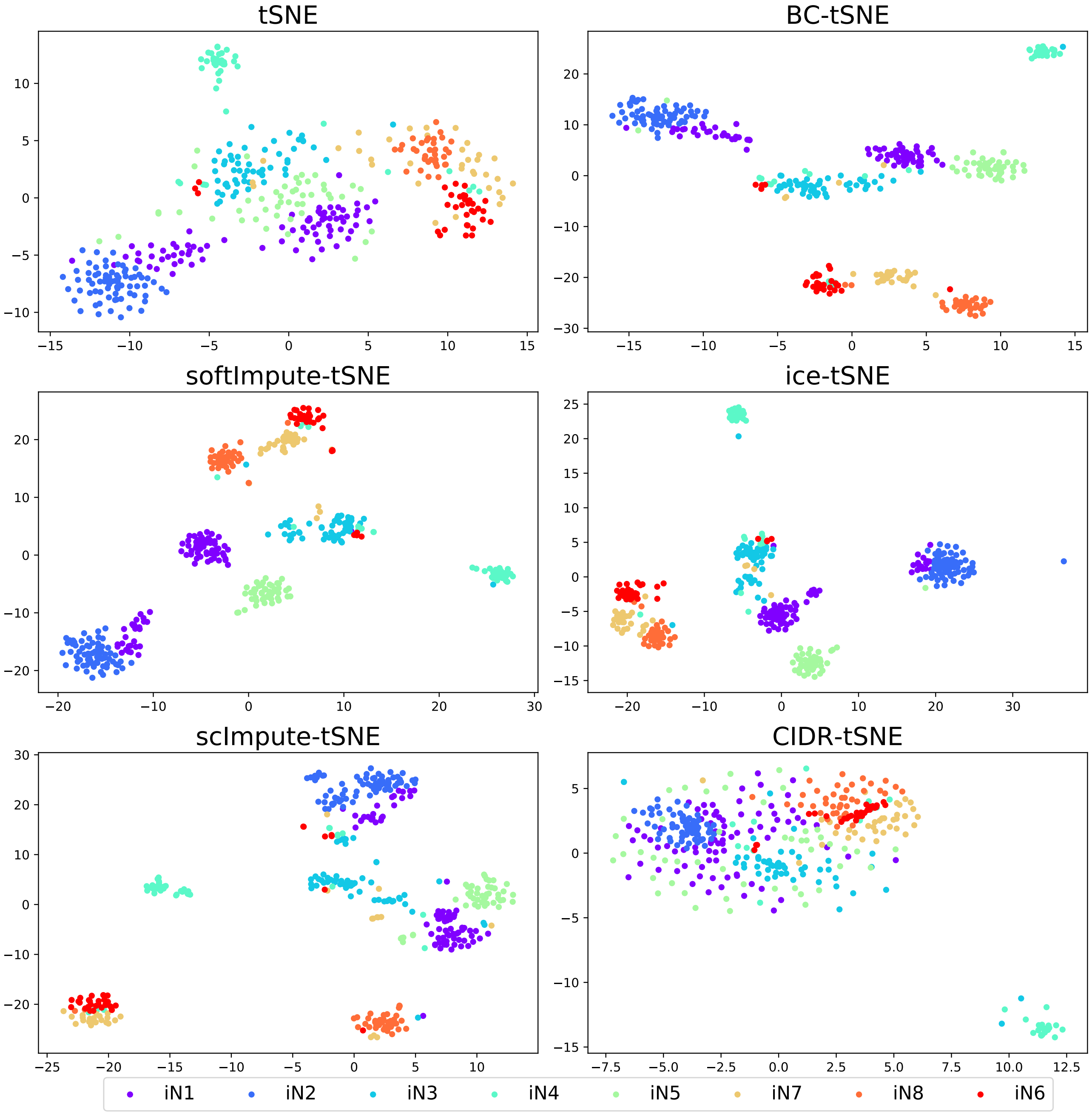


Figure S22: Visualisation of the Treutlein dataset obtained by tSNE and its variants integrated with the bias correction or imputations.

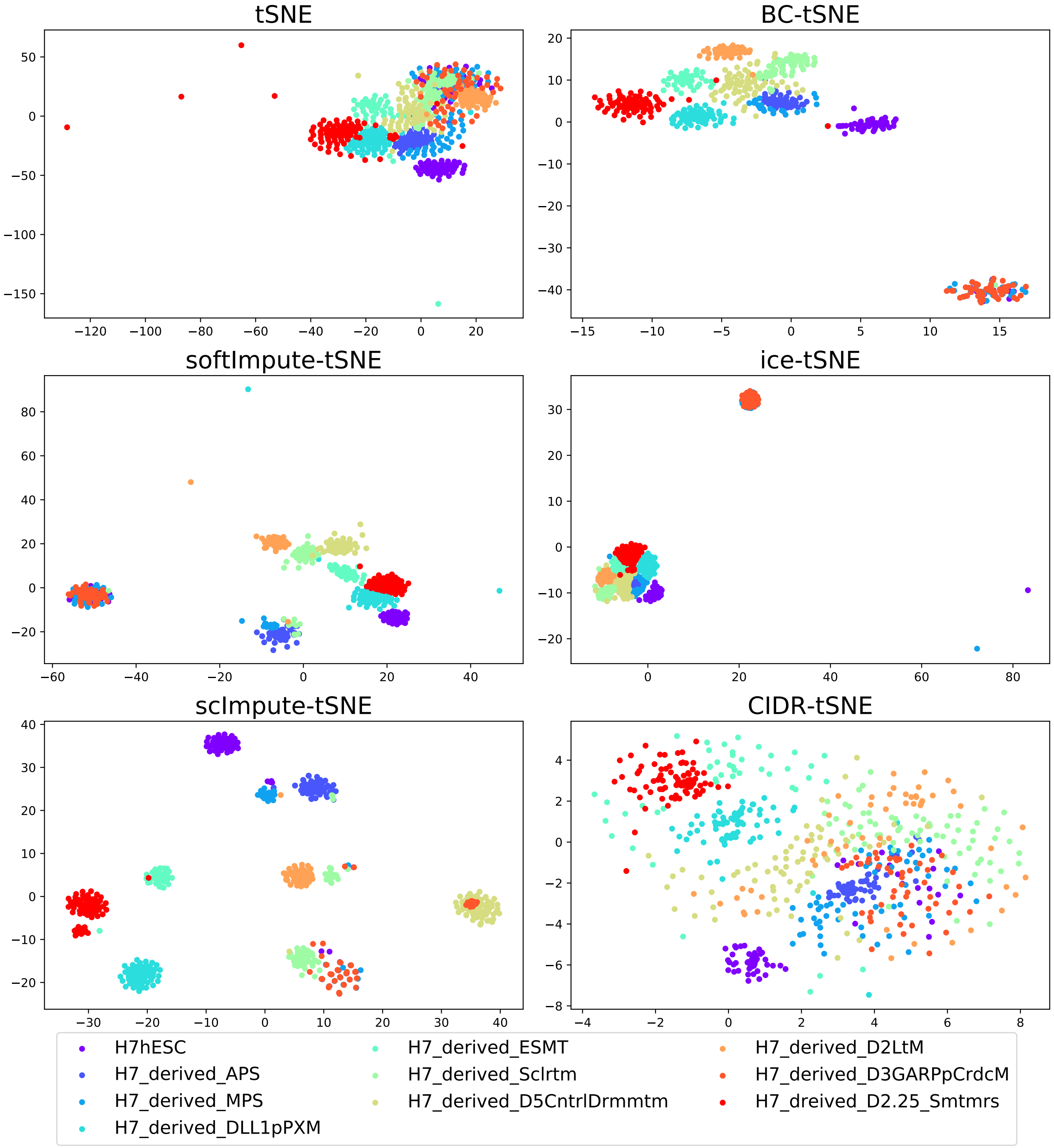


Figure S23: Visualisation of the Koh dataset obtained by tSNE and its variants integrated with the bias correction or imputations.

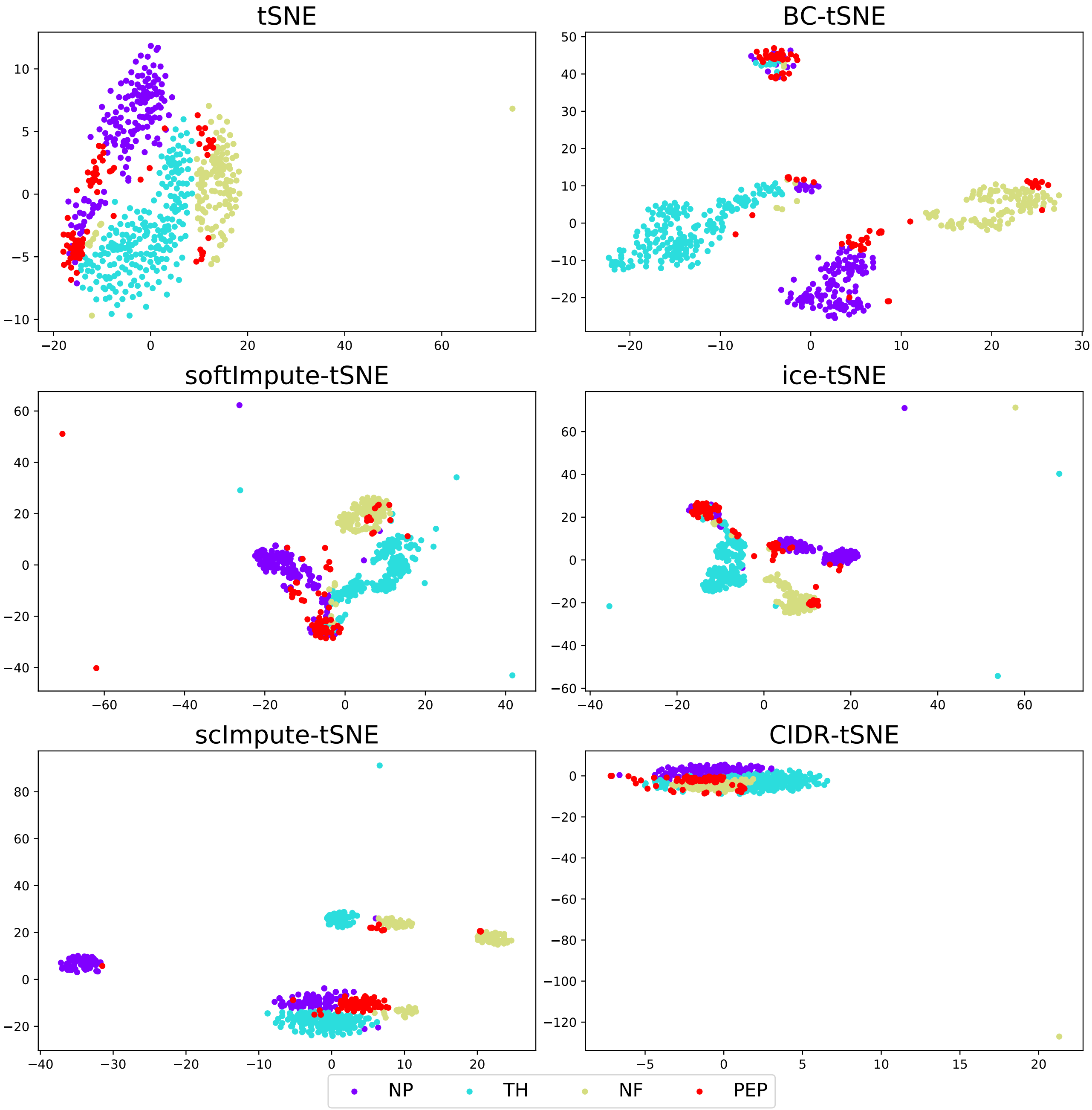


Figure S24: Visualisation of the Usoskin dataset obtained by tSNE and its variants integrated with the bias correction or imputations.

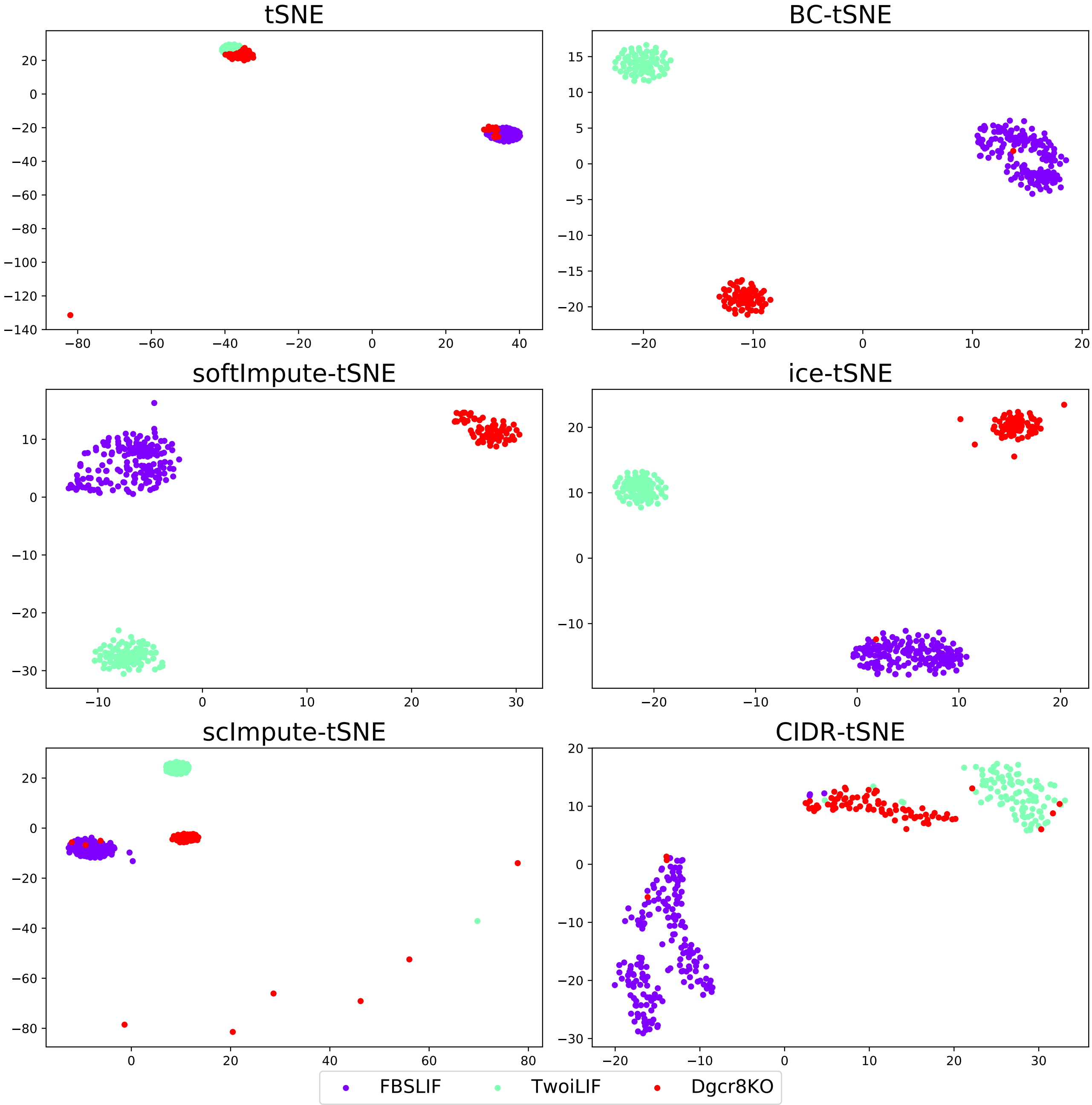


Figure S25: Visualisation of the Kumar dataset obtained by tSNE and its variants integrated with the bias correction or imputations.

S.6.4 VISUALISATION OF UMAP RESULTS ON THE REAL DATASETS

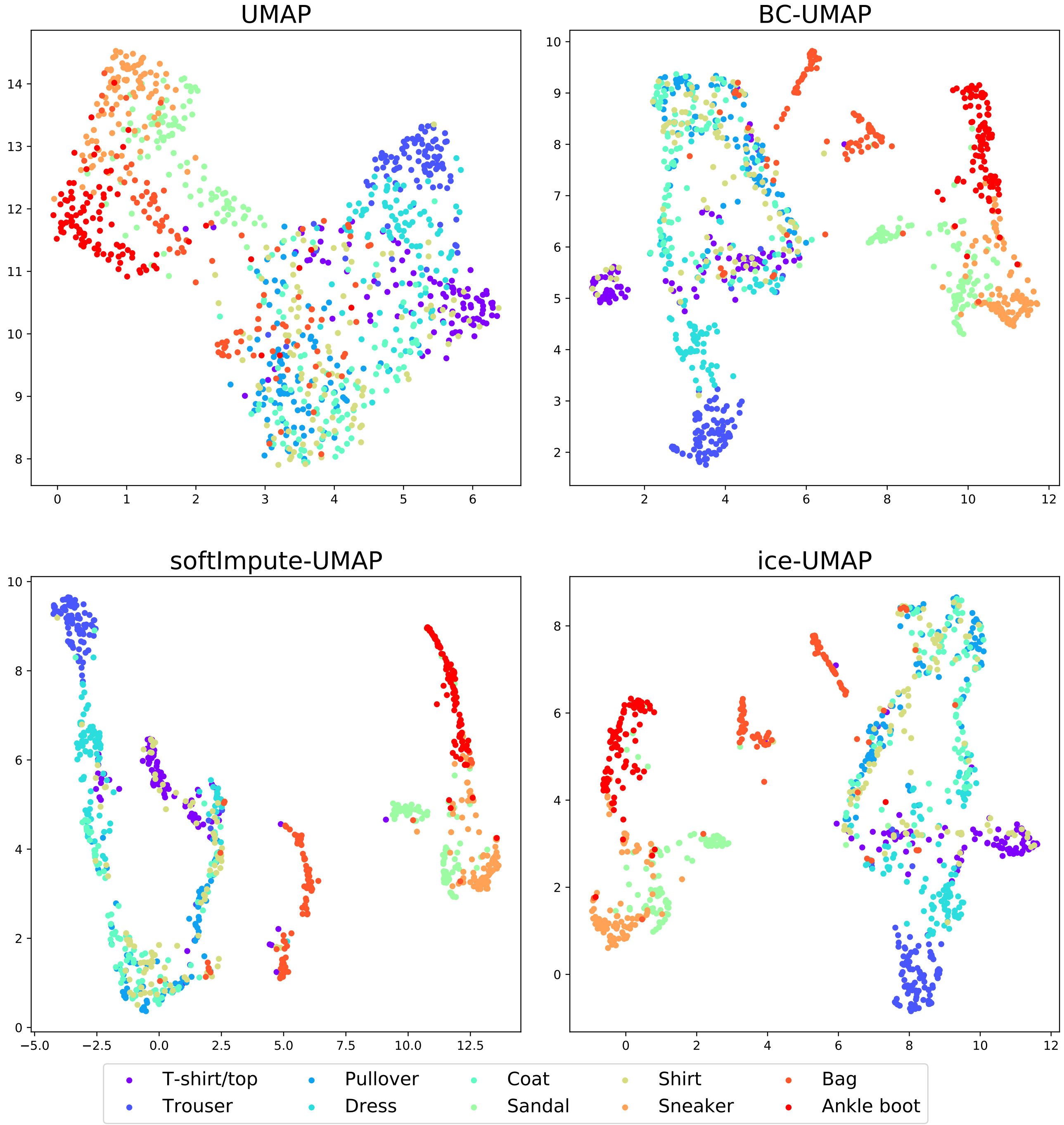


Figure S26: Visualisation of the fashion MNIST dataset obtained by UMAP and its variants integrated with the bias correction or imputations.

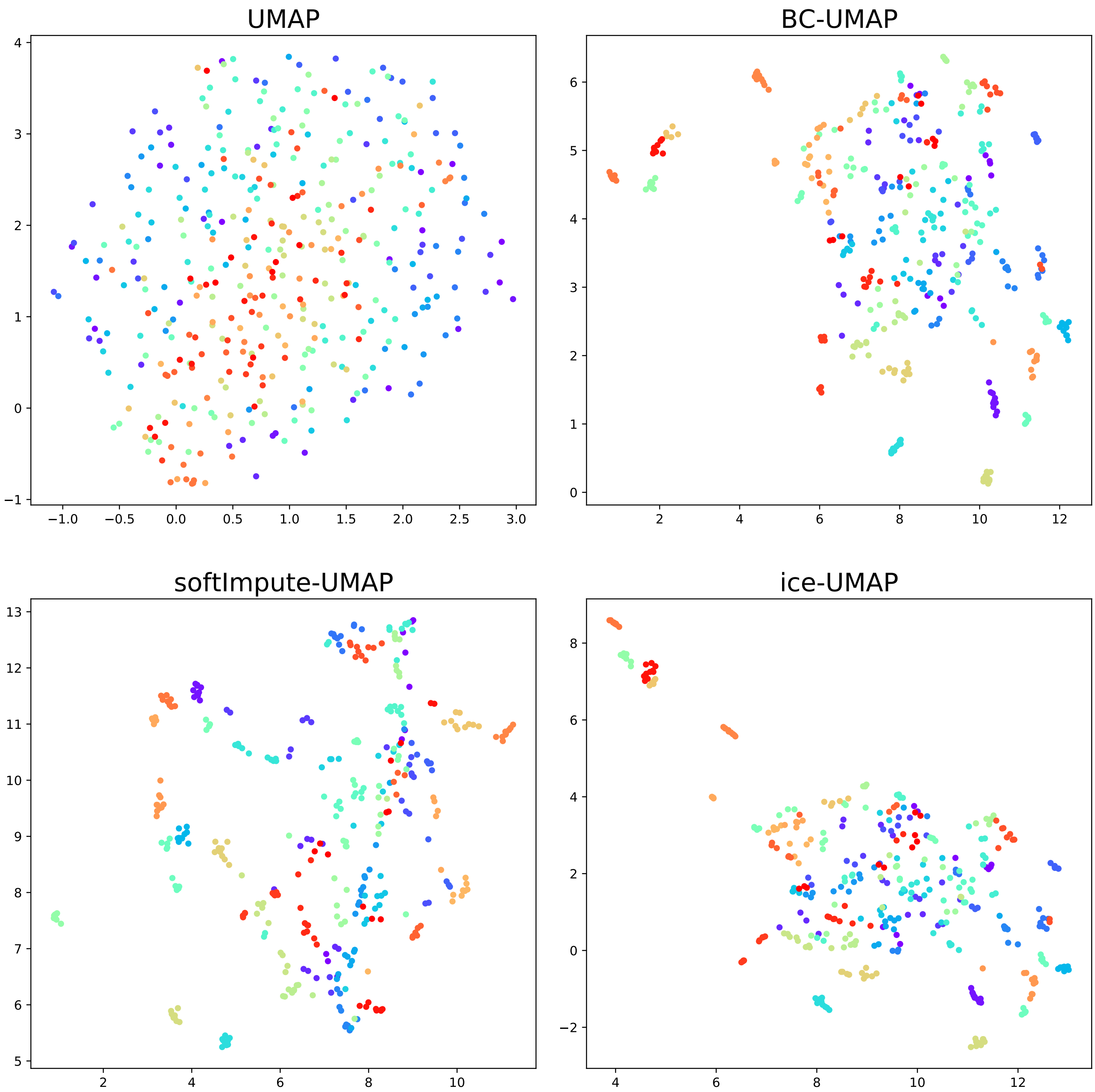


Figure S27: Visualisation of the Olivetti faces dataset obtained by UMAP and its variants integrated with the bias correction or imputations.

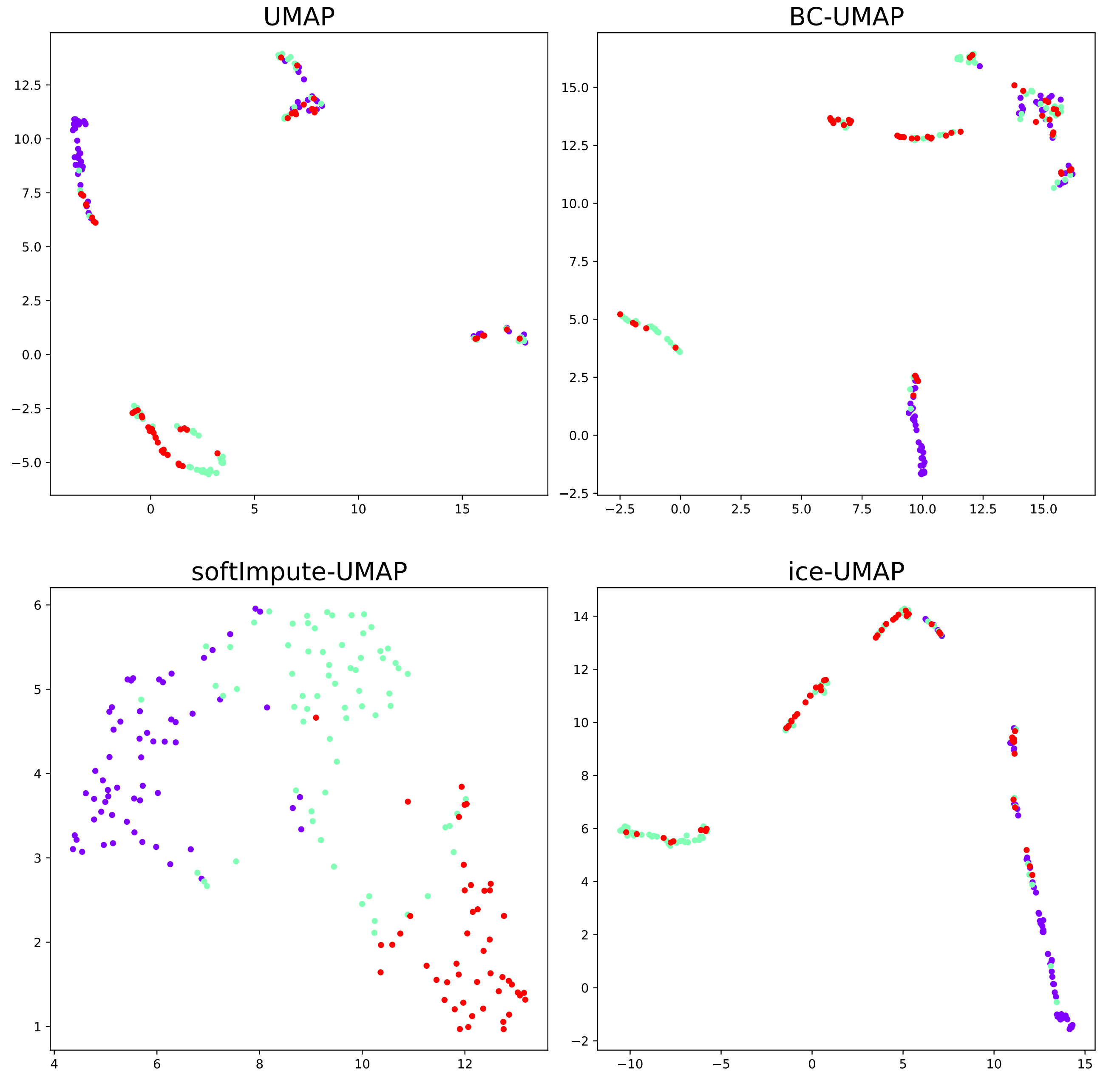


Figure S28: Visualisation of the wine dataset obtained by UMAP and its variants integrated with the bias correction or imputations.

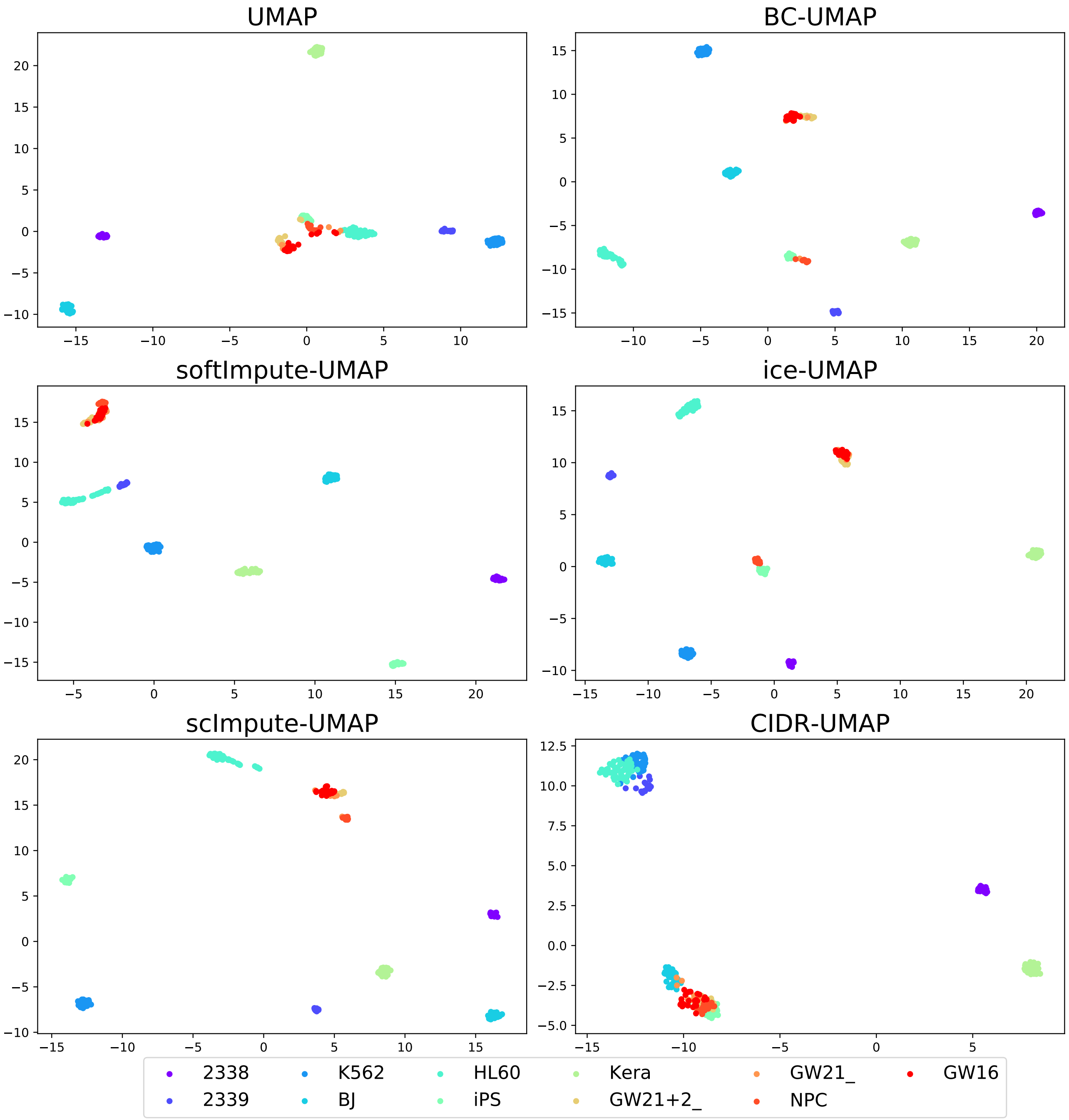


Figure S29: Visualisation of the Pollen dataset obtained by UMAP and its variants integrated with the bias correction or imputations.

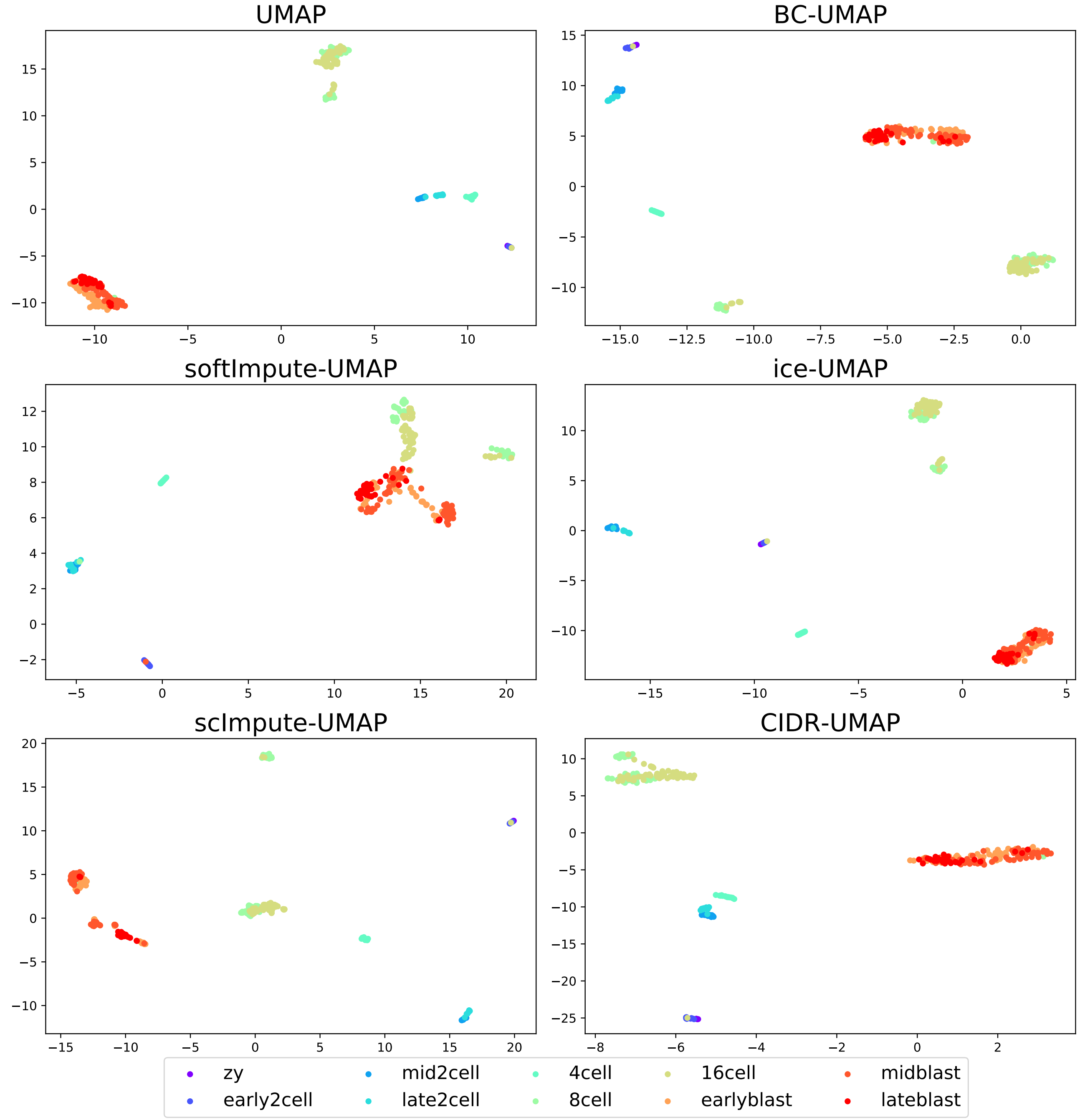


Figure S30: Visualisation of the Deng dataset obtained by UMAP and its variants integrated with the bias correction or imputations.

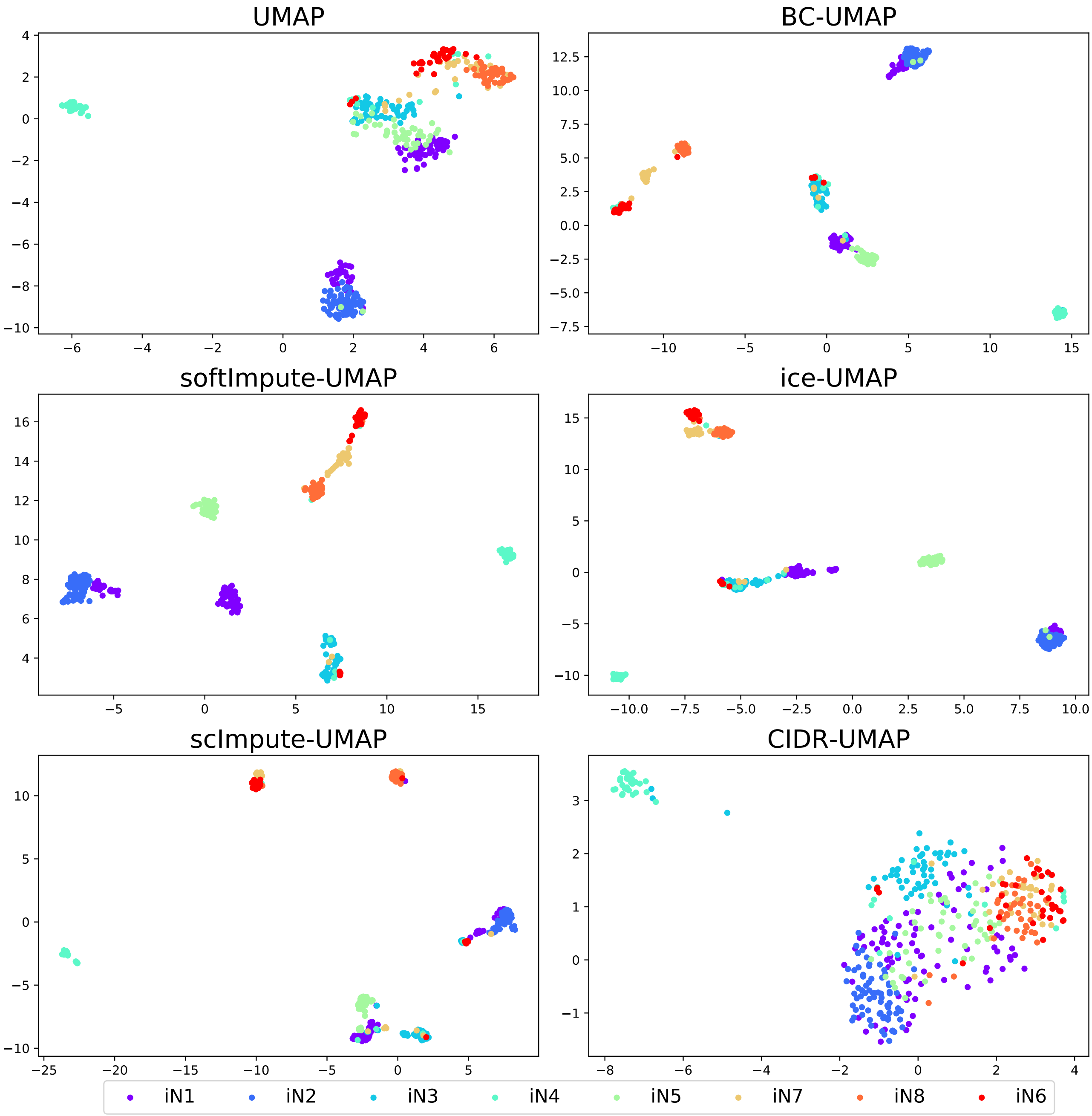


Figure S31: Visualisation of the Treutlein dataset obtained by UMAP and its variants integrated with the bias correction or imputations.

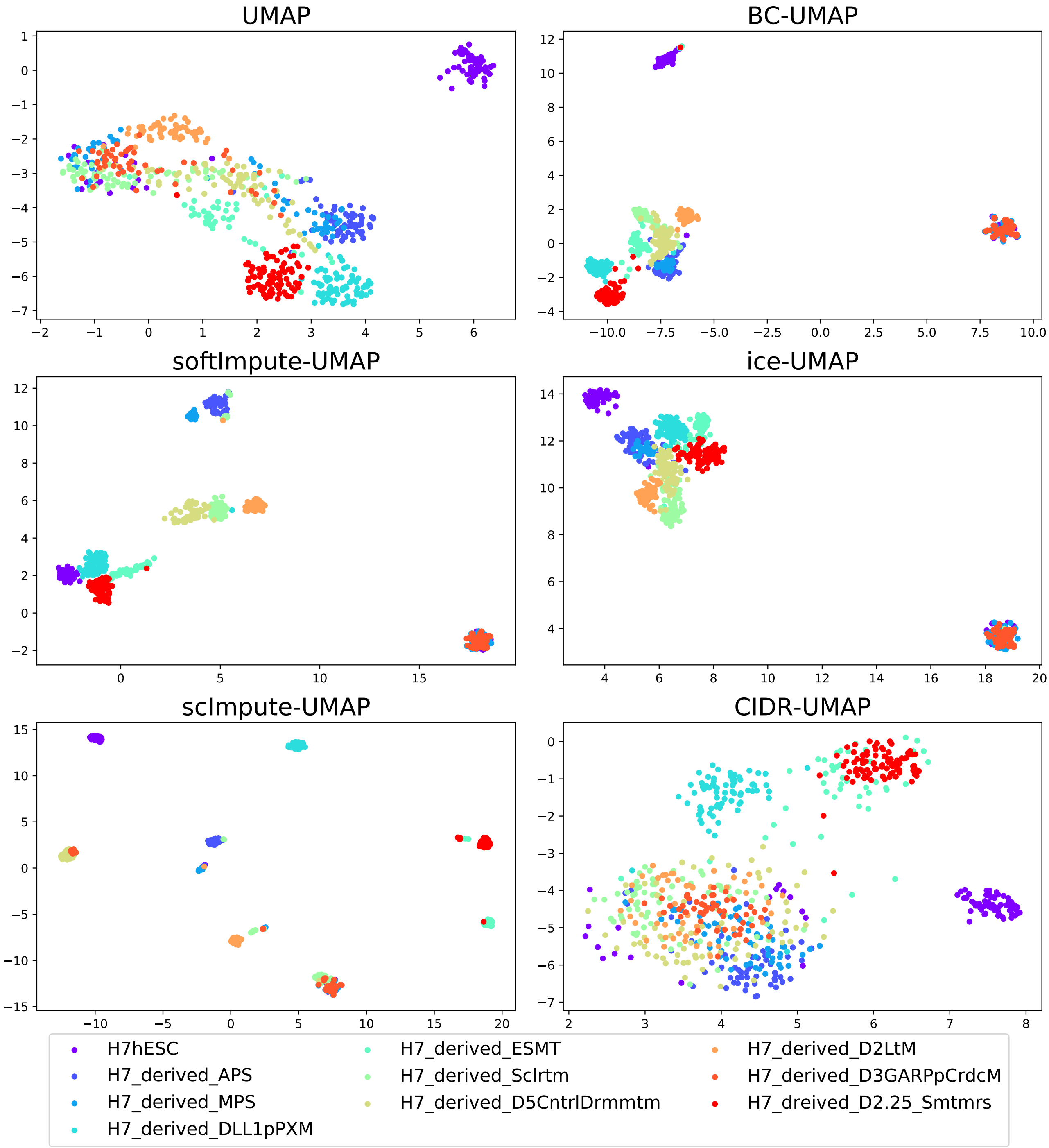


Figure S32: Visualisation of the Koh dataset obtained by UMAP and its variants integrated with the bias correction or imputations.

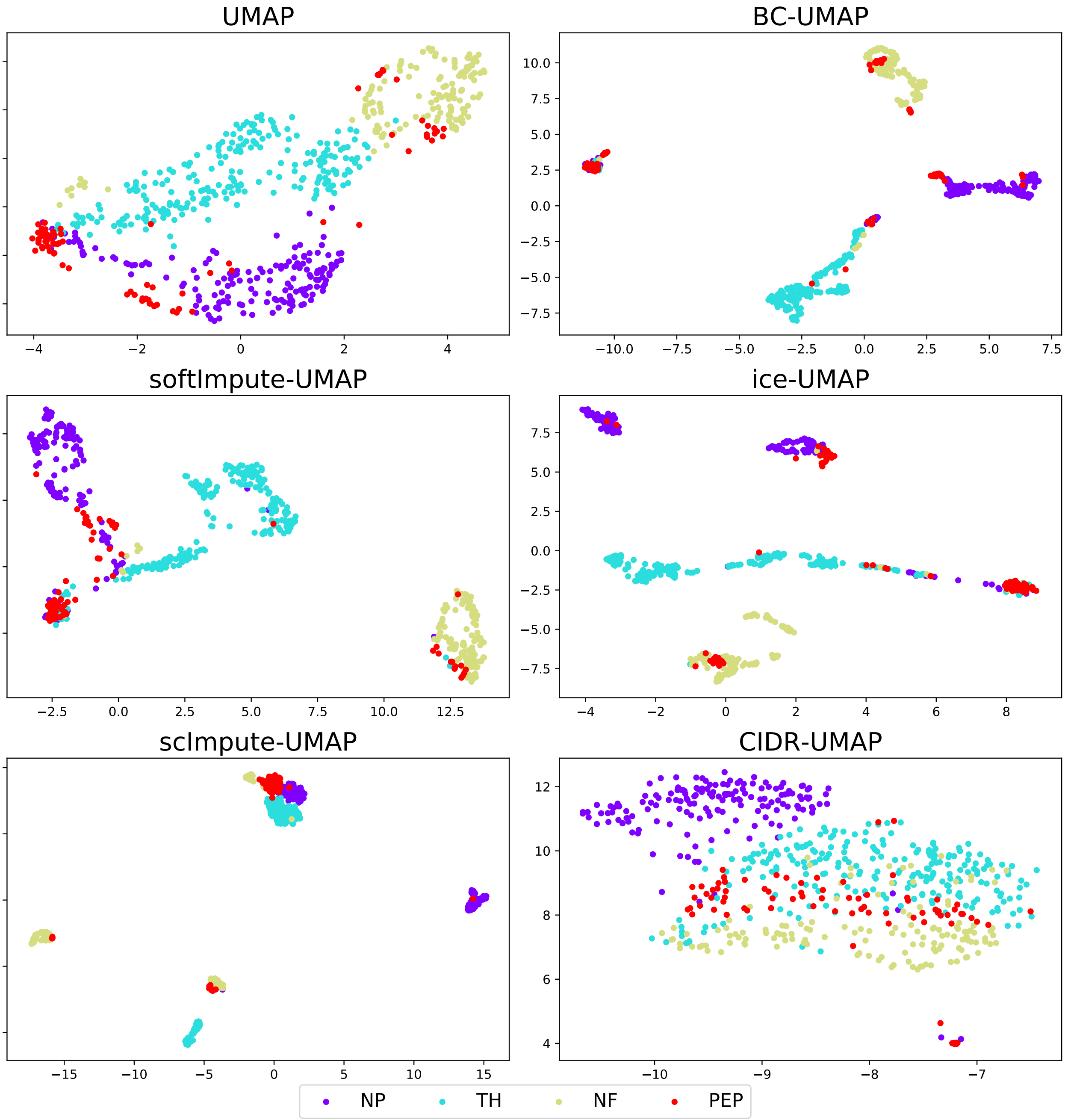


Figure S33: Visualisation of the Usoskin dataset obtained by UMAP and its variants integrated with the bias correction or imputations.

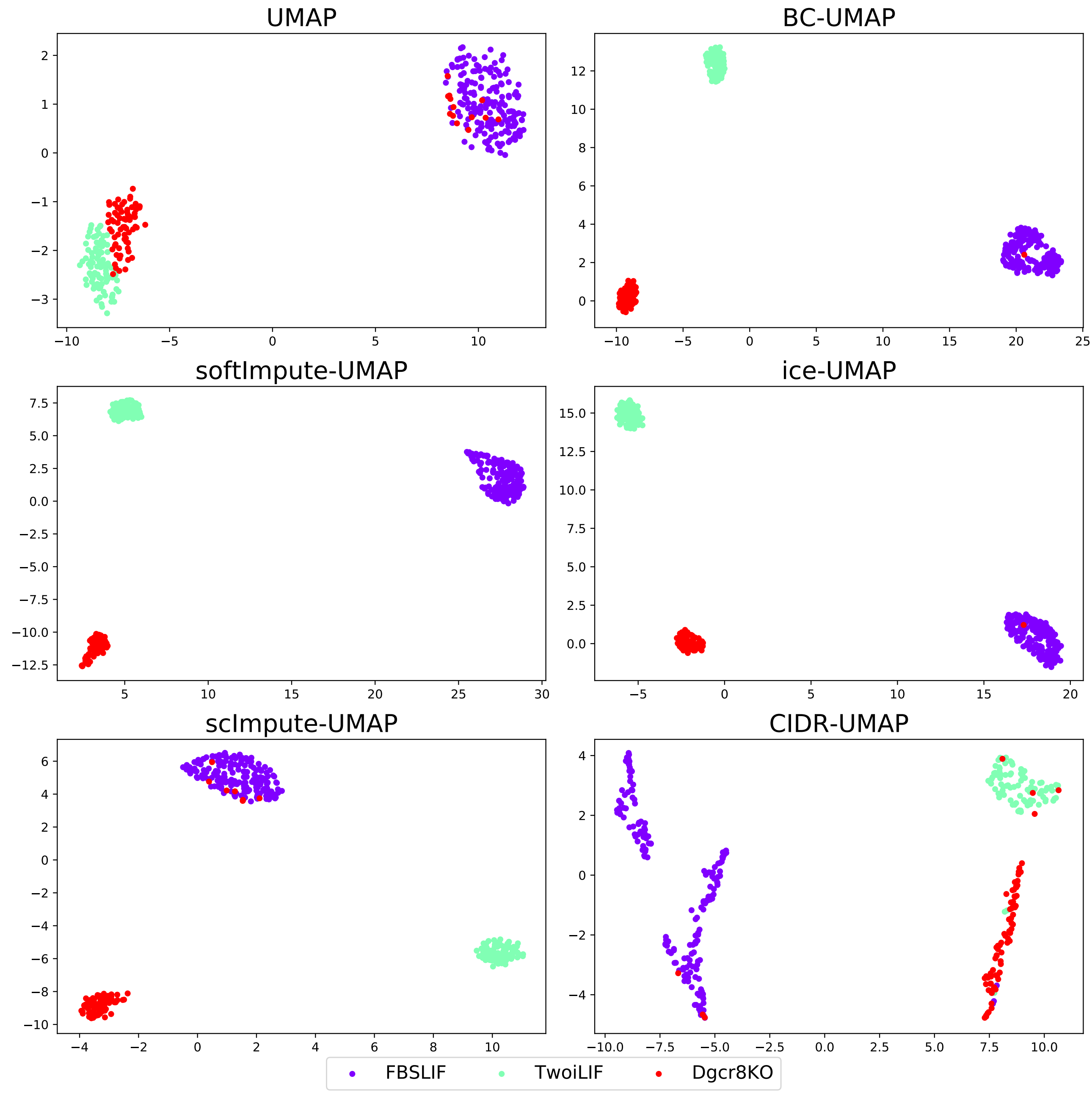
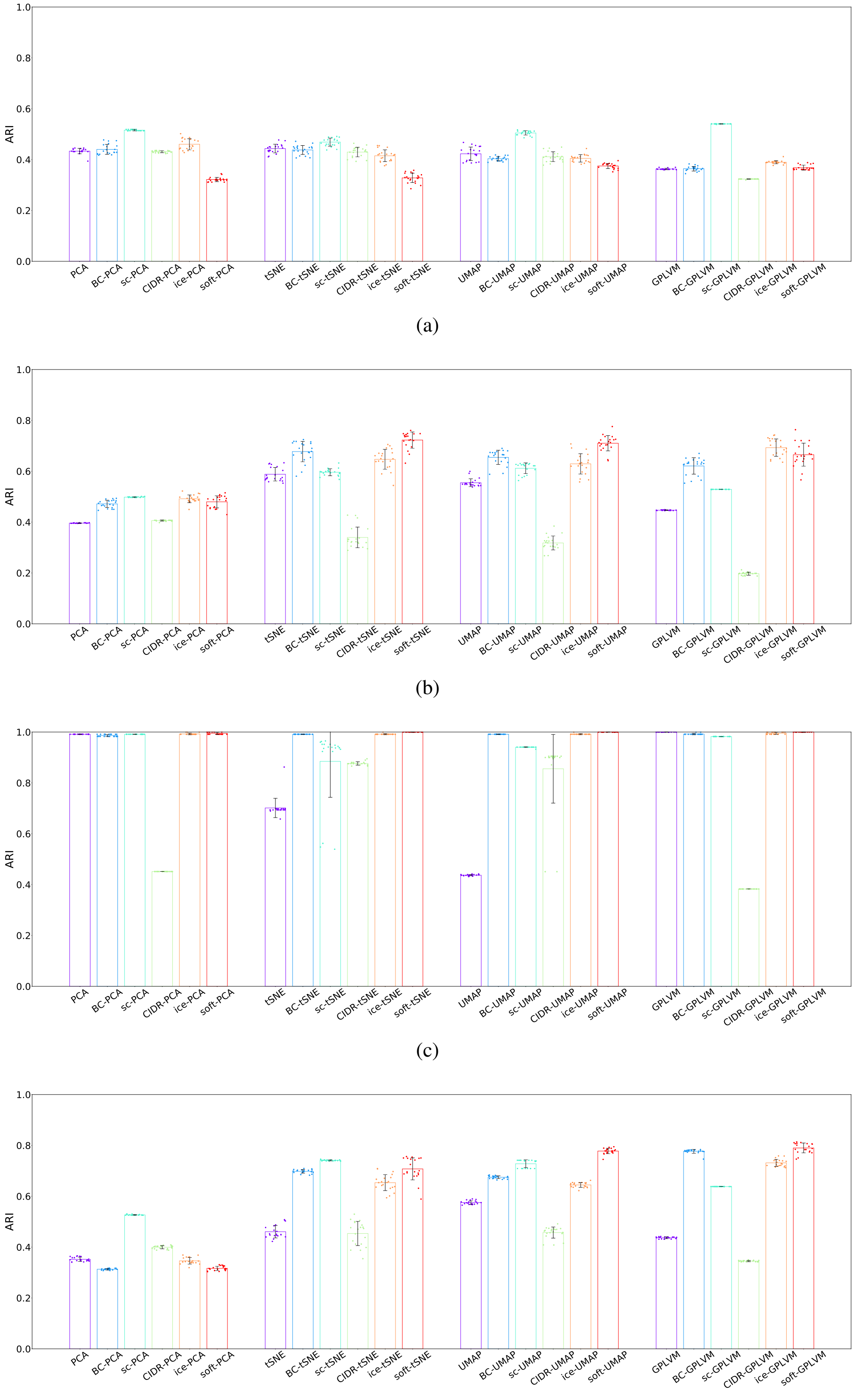


Figure S34: Visualisation of the Kumar dataset obtained by UMAP and its variants integrated with the bias correction or imputations.

S.6.5 K-MEANS RESULTS ON THE REAL DATASETS

S.6.5.1 Results on the scRNA-seq datasets



(a)

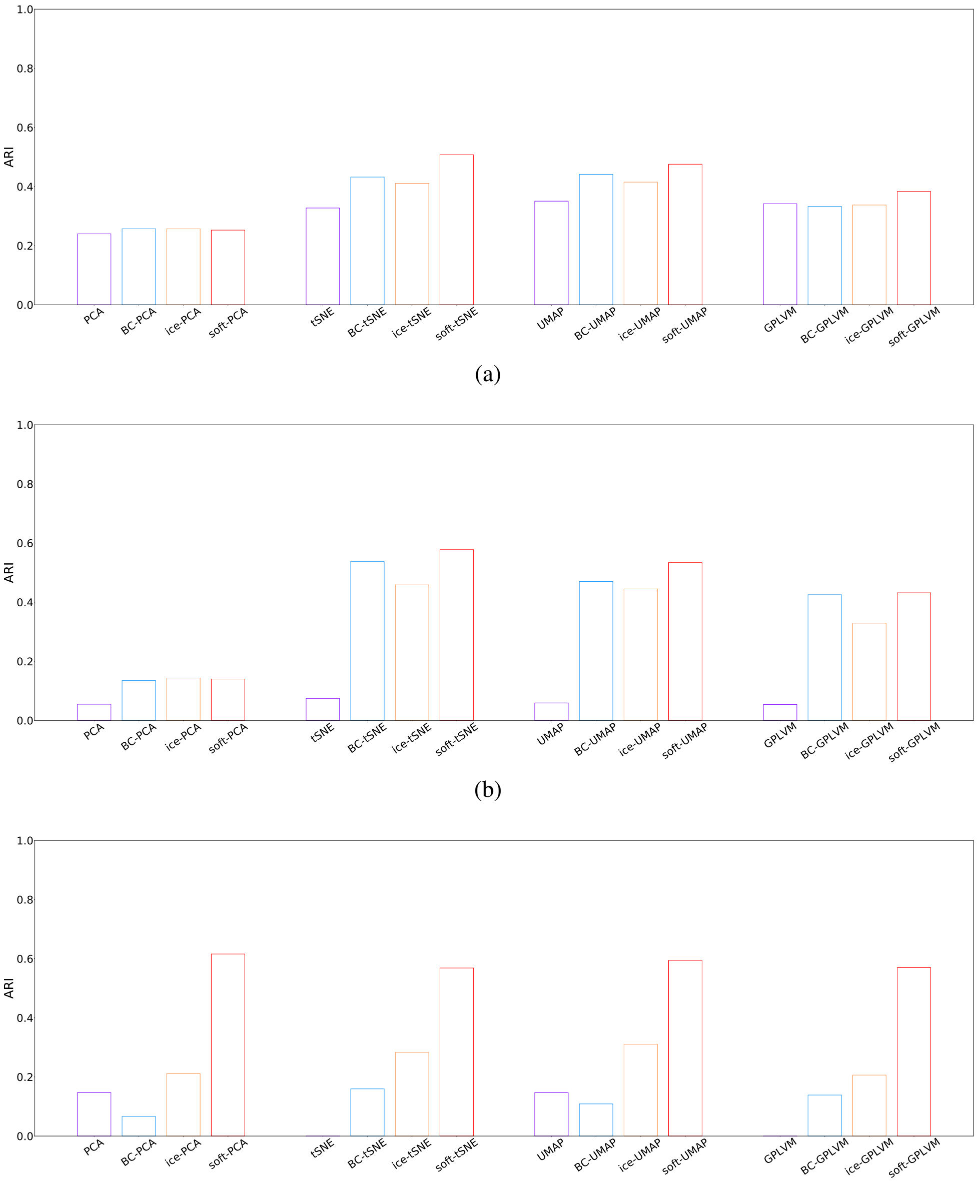
(b)

(c)

(d)

Figure S35: ARI of k-means with different DR approaches and their variants on the scRNA-seq datasets (a) Deng, (b) Treutlein, (c) Kumar, and (d) Koh.

S.6.5.2 Results on the other datasets



(a)

(b)

(c)

Figure S36: ARI of k-means with different DR approaches and their variants on the real datasets (a) fashion MNIST, (b) Olivetti faces, and (c) wine.

References

Erich Leo Lehmann. *Elements of large-sample theory*. Springer Science & Business Media, 2004.

Alexander G. de G. Matthews, Mark van der Wilk, Tom Nickson, Keisuke. Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6, 2017.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

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

Alex Rubinsteyn and Sergey Feldman. fancyimpute: An imputation library for python, 2016. URL `https://github.com/iskandr/fancyimpute`.

Charlotte Soneson and Mark D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, 2018.