Dimension Reduction for High-dimensional Small Counts with KL Divergence Supplementary Material

Yurong Ling[1] **Jing-Hao Xue**[1]

[1]Department of Statistical Science, University College London, London, UK

S.1 PROOFS

In this section, we provide the proofs of Proposition 1 and Corollary 2 presented in the main paper.

S.1.1 PROOF OF PROPOSITION 1

Proof 1 *First note that $R_D\left(F_x, F_y\right)$ converges to $\frac{\mathbb{E}[D^2(x,y)]}{\mathbb{E}[D^2(x,\tilde{x})]+\mathbb{E}[D^2(y,\tilde{y})]}$ in probability, as $p \to \infty$, based on the weak law of large numbers and the Slutsky's theorem. Now we have*

$$\begin{aligned}\mathbb{E}[D^2(x,y)] &= \mathrm{Cov}\left(f(x)-f(y), g(x)-g(y)\right) + \left[\mathrm{E}f(x)-\mathrm{E}f(y)\right]\left[\mathrm{E}g(x)-\mathrm{E}g(y)\right] \\ &= \mathrm{Cov}\left(f(x), g(x)\right) + \mathrm{Cov}\left(f(y), g(y)\right) + \left[\mathrm{E}f(x)-\mathrm{E}f(y)\right]\left[\mathrm{E}g(x)-\mathrm{E}g(y)\right],\end{aligned} \tag{S1}$$

and

$$\begin{aligned}\mathbb{E}[D^2(x,\tilde{x})] + \mathbb{E}[D^2(y,\tilde{y})] &= 2\mathrm{E}\left[f(x)g(x)\right] - 2\mathrm{E}f(x)\mathrm{E}g(x) + 2\mathrm{E}\left[f(y)g(y)\right] - 2\mathrm{E}f(y)\mathrm{E}g(y) \\ &= 2\mathrm{Cov}\left(f(x), g(x)\right) + 2\mathrm{Cov}\left(f(y), g(y)\right).\end{aligned} \tag{S2}$$

Combining results from Equation (S1) *and Equation* (S2)*, we obtain*

$$\frac{\mathbb{E}[D^2(x,y)]}{\mathbb{E}[D^2(x,\tilde{x})] + \mathbb{E}[D^2(y,\tilde{y})]} = \frac{1}{2} + \frac{1}{2}\frac{\left[\mathrm{E}f(x)-\mathrm{E}f(y)\right]\left[\mathrm{E}g(x)-\mathrm{E}g(y)\right]}{\mathrm{Cov}\left(f(x), g(x)\right) + \mathrm{Cov}\left(f(y), g(y)\right)},$$

which completes the proof.

S.1.2 PROOF OF COROLLARY 2

Proof 2 *First note that $\lim_{\mu_x \to 0} PMF(x=0) = 1$ for non-negative random variable x and $\lim_{\mu_x \to 0} \mathrm{Var}\left(x\right) = 0$. Further, we obtain $\lim_{\mu_x \to 0} \mathrm{E}\left[g(x)\right] = g(0)$ and $\lim_{\mu_x \to 0} \mathrm{Var}\left[g(x)\right] = 0$. Now consider the limiting difference between c_g and c_E:*

$$\begin{aligned}\lim_{\mu_x \to 0} c_g - c_E &= \lim_{\mu_x \to 0}\left[\frac{1}{2}\frac{\left[\mathrm{E}g(x)-\mathrm{E}g(y)\right]^2}{\mathrm{Var}\left[g(x)\right] + \mathrm{Var}\left[g(y)\right]} - \frac{1}{2}\frac{\left[\mathrm{E}(x)-\mathrm{E}(y)\right]^2}{\mathrm{Var}\left(x\right) + \mathrm{Var}\left(y\right)}\right] \\ &= \frac{1}{2}\left[\frac{\left[g(0)-\mathrm{E}g(y)\right]^2}{\mathrm{Var}\left[g(y)\right]} - \frac{\mathrm{E}^2\left(y\right)}{\mathrm{Var}\left(y\right)}\right].\end{aligned}$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

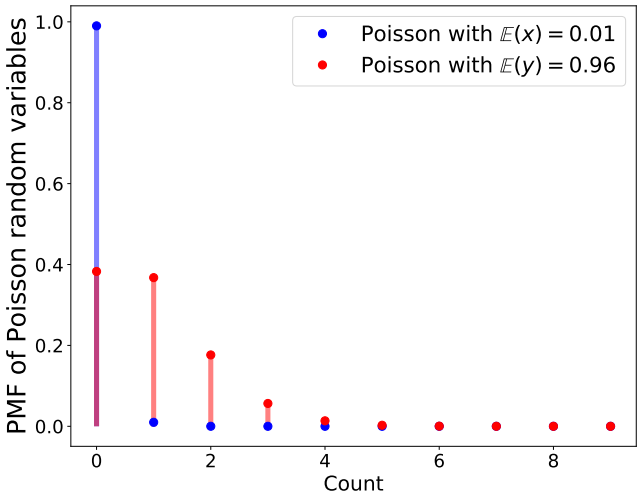

Figure S1: PMFs of two Poisson distributions.

Table S1: Parameter spaces for F_y.

Distributions	Parameter 1	Parameter 2
Poisson	means: $[0.05, 1]$ #samples: 50	variances: the same as the space of means
NB	r (# failures): $(\frac{3}{4}, 5]$; #samples: 50	p (probability of success): $[0.05, 0.2]$; #samples: 10

S.2 AN EXAMPLE SHOWING THE EUCLIDEAN DISTANCE PERFORMS BADLY ON SMALL COUNTS

Suppose we have two Poisson distributions with mean values 0.01 and 0.96, respectively. Although both distributions produce small counts, the PMFs of these two distributions, provided in Figure S1, display disparate patterns: the mass of one distribution concentrates in 0, while the frequent values of the other one lie in $\{0, 1, 2\}$. It is highly likely that $R_E(F_x, F_y) < 1$ in the space of high dimension according to Corollary 1 due to the mean-variance dependency, indicating that D_E cannot distinguish these two distributions well.

S.3 SIMULATION DETAILS

To obtain $\hat{R}(F_x, F_y)$, we first generate high-dimensional samples in S_X and S_Y with each coordinate i.i.d., and then calculate $\hat{R}(F_x, F_y)$ with different measures. As we compared D_E with the Euclidean distances of the transformed data, we let F_x be a count distribution that frequently generates zeros. Specifically, F_x is either $\text{Pois}(0.05)$ or $\text{NB}(r, 0.05)$ that has the same r as that of F_y. Note that the probability of getting a zero from $\text{Pois}(0.05)$ is around 0.951 and that from $\text{NB}(r, 0.05)$ is 0.95^r. The parameter spaces for F_y are provided in Table S1. Note that the parameters for simulating NB and Poisson distributions are selected in an evenly spaced way from the intervals provided in Table S1.

With the mean value increasing, the proportion of zeros produced from a Poisson distribution decreases to 0. We hence bound the mean value from above by 1 to generate small counts and the expected fractions of zeros in the Poisson-distributed count data lie in the range $[0.368, 0.905]$. When r approaches infinity, the NB distribution approximates an equi-dispersed Poisson distribution. Thus, the upper bound of r for simulating NB distributions should be small in order to generate overdispersed distributions. When simulating data from NB distributions, the values of r for both F_x and F_y are the same, and we use the true value of r for calculating D_{NB} and D_{asin}. Since $r < \frac{3}{4}$ results in the numerical problem when calculating D_{asin}, we only take values that are larger than $\frac{3}{4}$ for r . Again, we restrict the range of p in NB distributions with the aim of simulating small counts. The proportion of zeros in the simulated NB data ranges from 0.363 to 0.963. For Poisson distributions, we set r to 1000 in D_{asin} due to the approximation of NB distributions to Poisson distributions when r is large. The dimension for each data point is set to be 5000 which is high enough to obtain a reliable estimation of the constant that $R(F_x, F_y)$ converges to. Although the convergence of $R(F_x, F_y)$ holds regardless of the number of samples in each distribution, small

values of n_x and n_y possibly result in the meaningless coordinates, in the form of vectors of zeros. Therefore, we set n_x and n_y to 1000 to avoid such cases.

S.4 EVALUATION DETAILS

When applying tSNE with the proposed measures to the datasets, we replace the Euclidean distance with the proposed ones to characterise the dissimilarities between data points for the construction of conditional probabilities. When performing PCA with the proposed measures, we first get a pseudo Gram matrix by double-centring the corresponding pairwise dissimilarity matrix and then get the low-dimensional components by eigen-decomposing the pseudo Gram matrix. We empirically found that making the pseudo Gram matrix positive semi-definite would produce better results when combined with GPLVM. Thus, we first eigen-decompose the pseudo Gram matrix and keep all the eigenvectors with non-negative eigenvalues. We then feed the modified pseudo Gram matrix $Q\Lambda Q^T$ into the GPLVM, where Q is the matrix whose columns are the kept eigenvectors and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues.

S.5 MORE VISUALIZATION RESULTS

S.5.1 VISUALIZATION OF GPLVM RESULTS

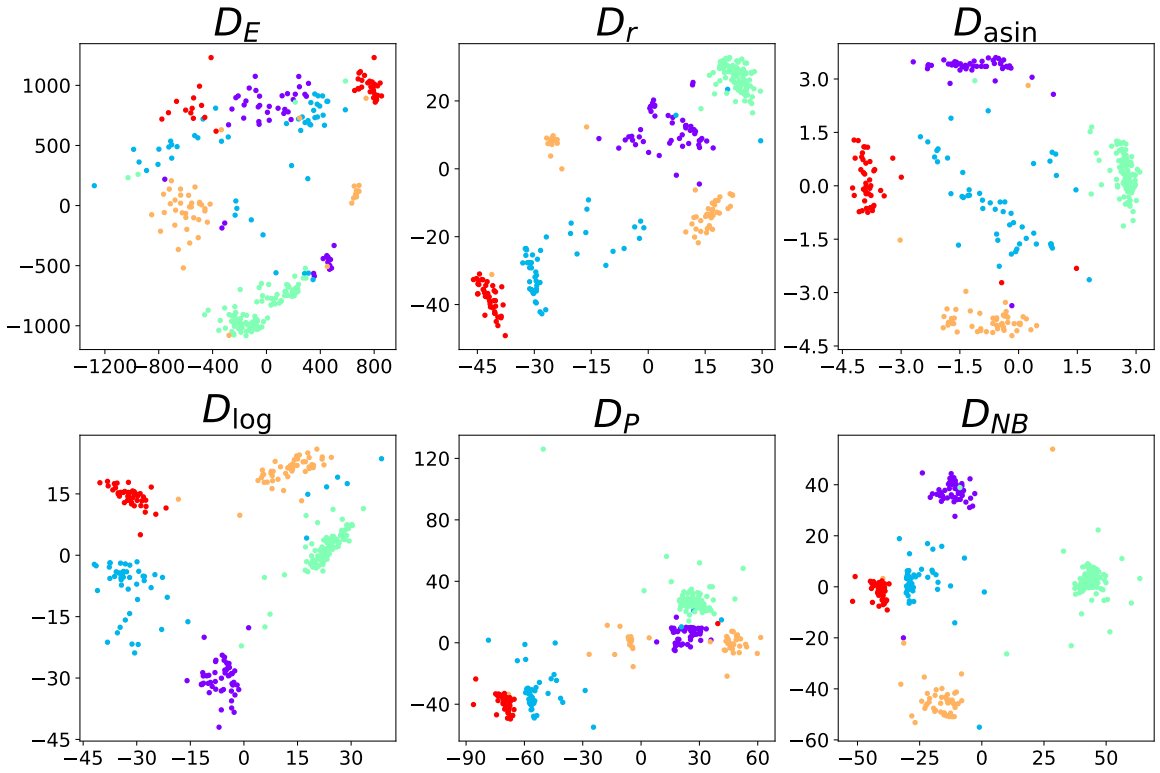

Figure S2: Visualization of the sc-CEL-seq2-5cl-p1 dataset obtained by GPLVMs with different measures.

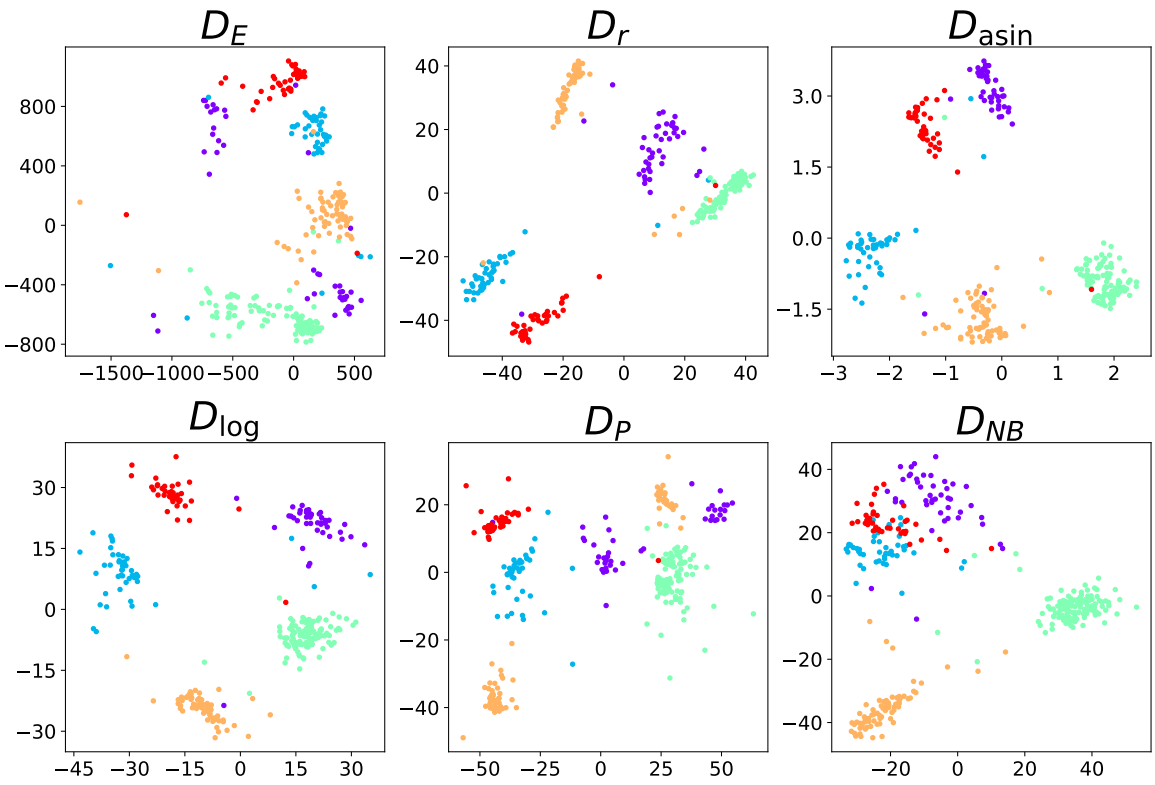

Figure S3: Visualization of the sc-CEL-seq2-5cl-p2 dataset obtained by GPLVMs with different measures.

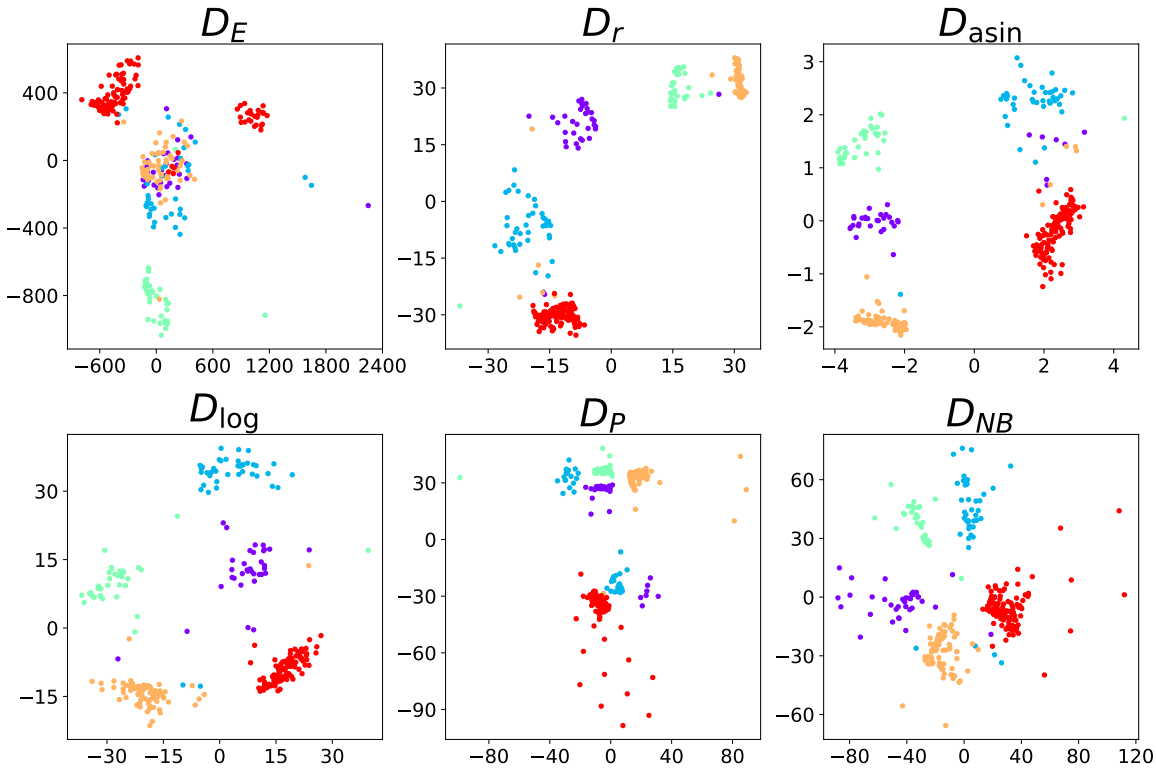

Figure S4: Visualization of the sc-CEL-seq2-5cl-p3 dataset obtained by GPLVMs with different measures.

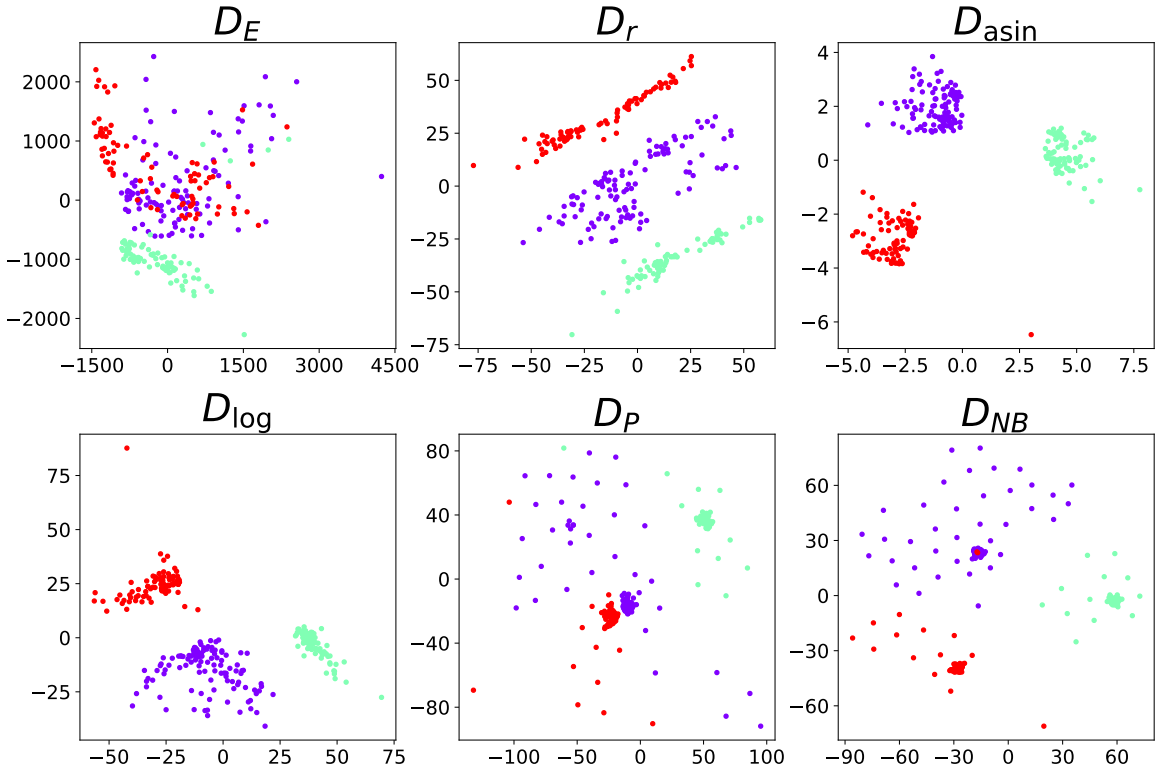

Figure S5: Visualization of the sc-CEL-seq2 dataset obtained by GPLVMs with different measures.

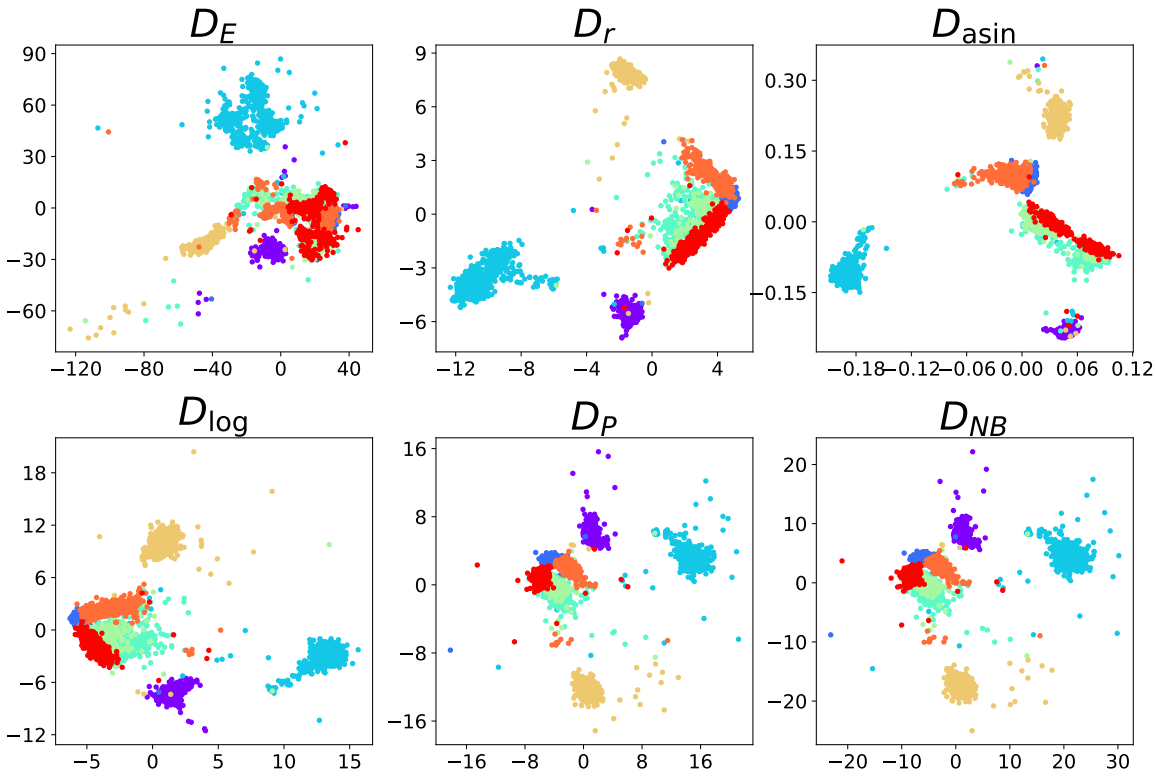

Figure S6: Visualization of the Zheng8eq dataset obtained by GPLVMs with different measures.

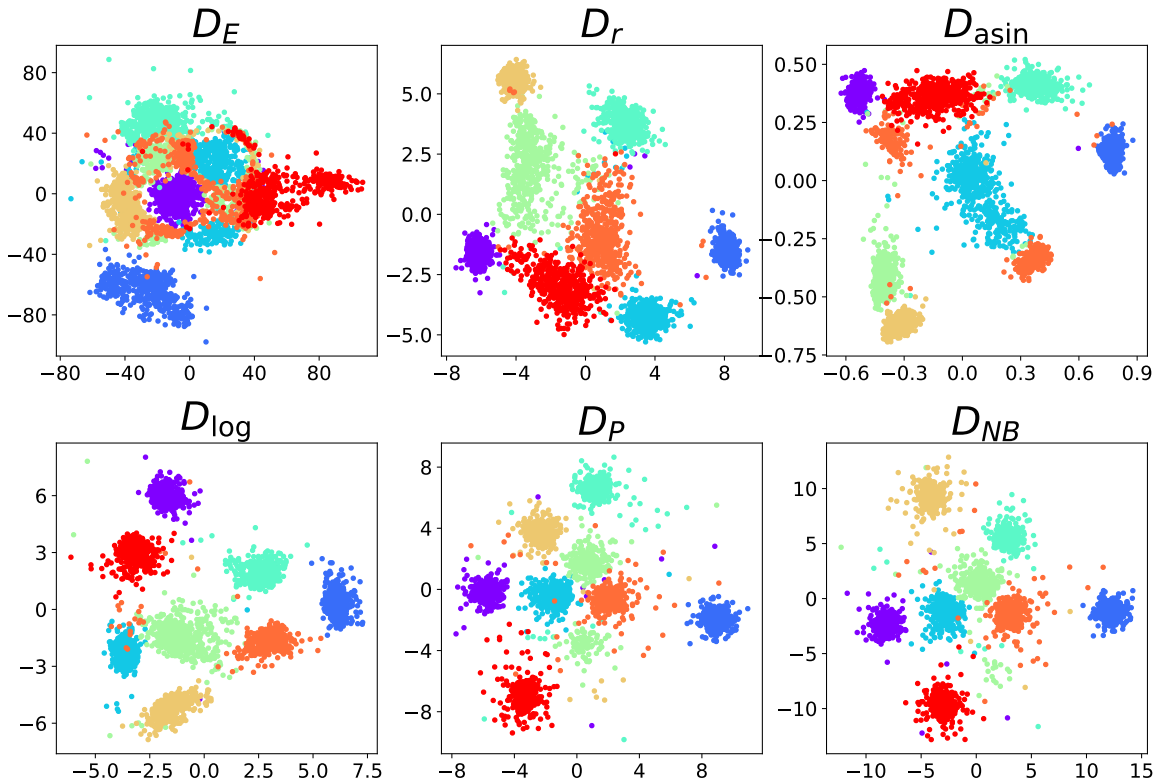

Figure S7: Visualization of the sim-Zheng8eq dataset obtained by GPLVMs with different measures.

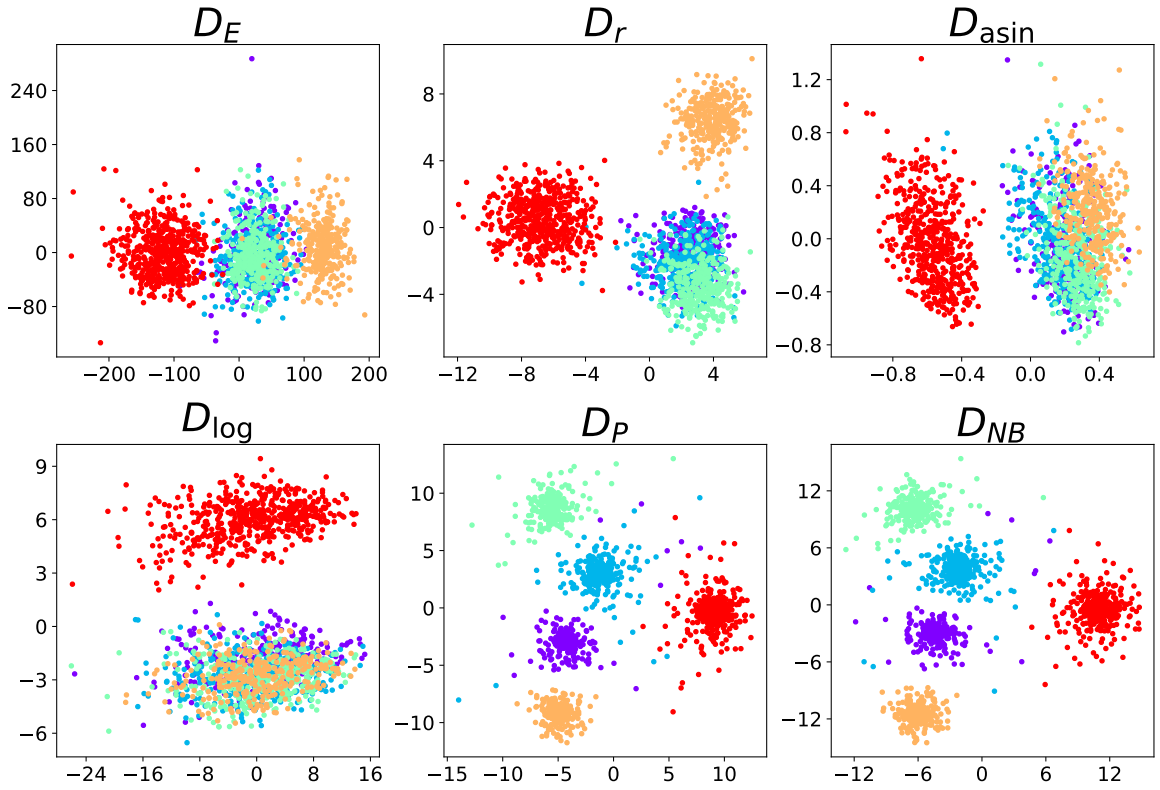

Figure S8: Visualization of the sim-manno-ESCs dataset obtained by GPLVMs with different measures.

S.5.2 VISUALIZATION OF TSNE RESULTS

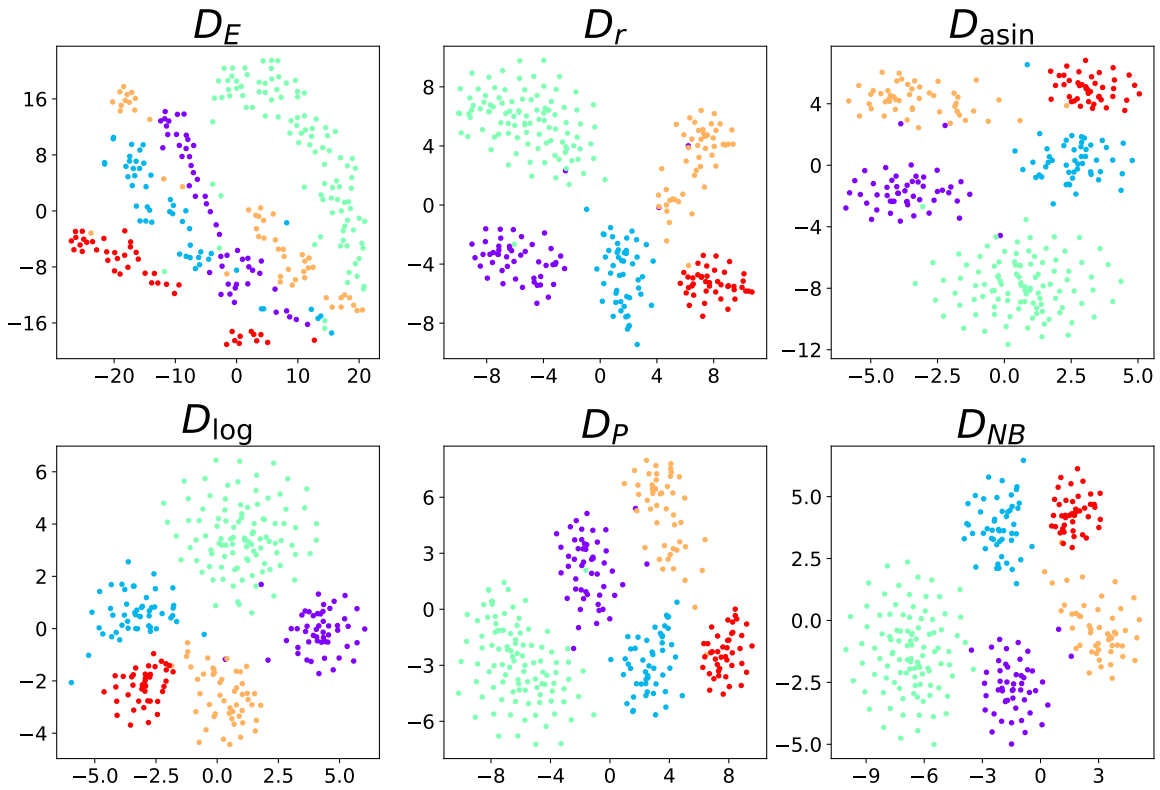

Figure S9: Visualization of the sc-CEL-seq2-5cl-p1 dataset obtained by tSNE with different measures.

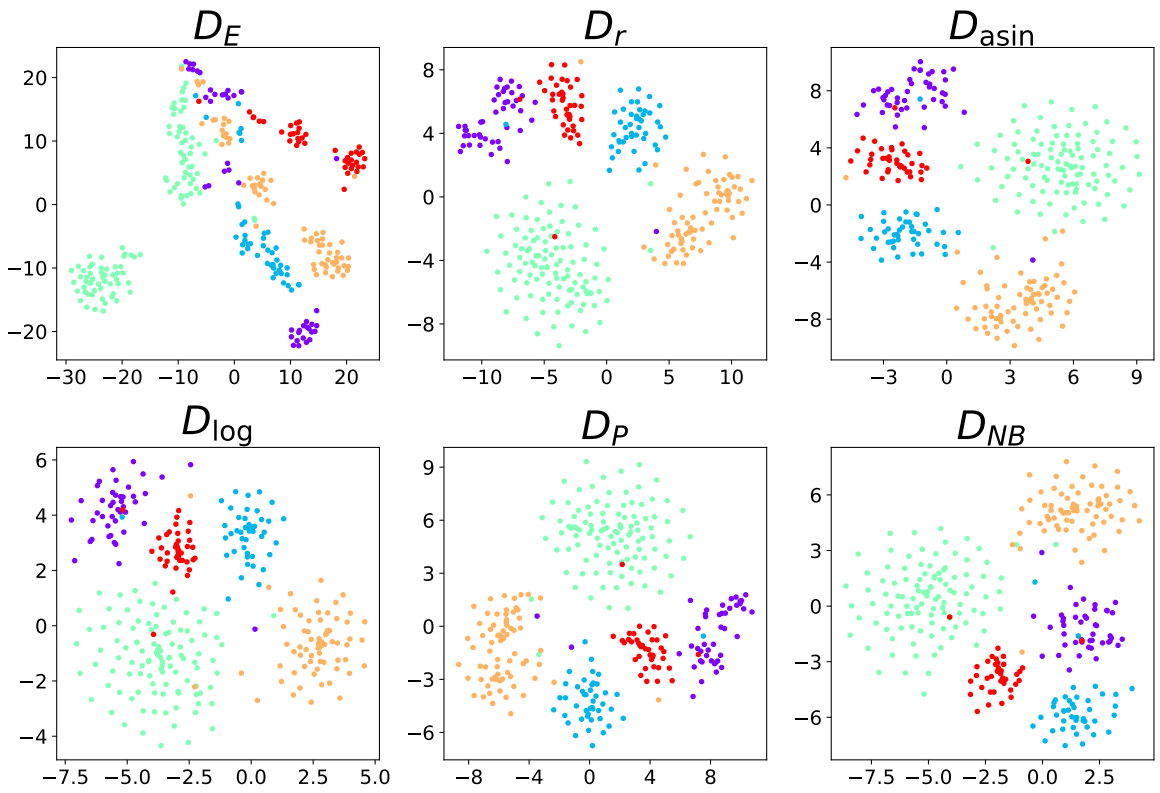

Figure S10: Visualization of the sc-CEL-seq2-5cl-p2 dataset obtained by tSNE with different measures.

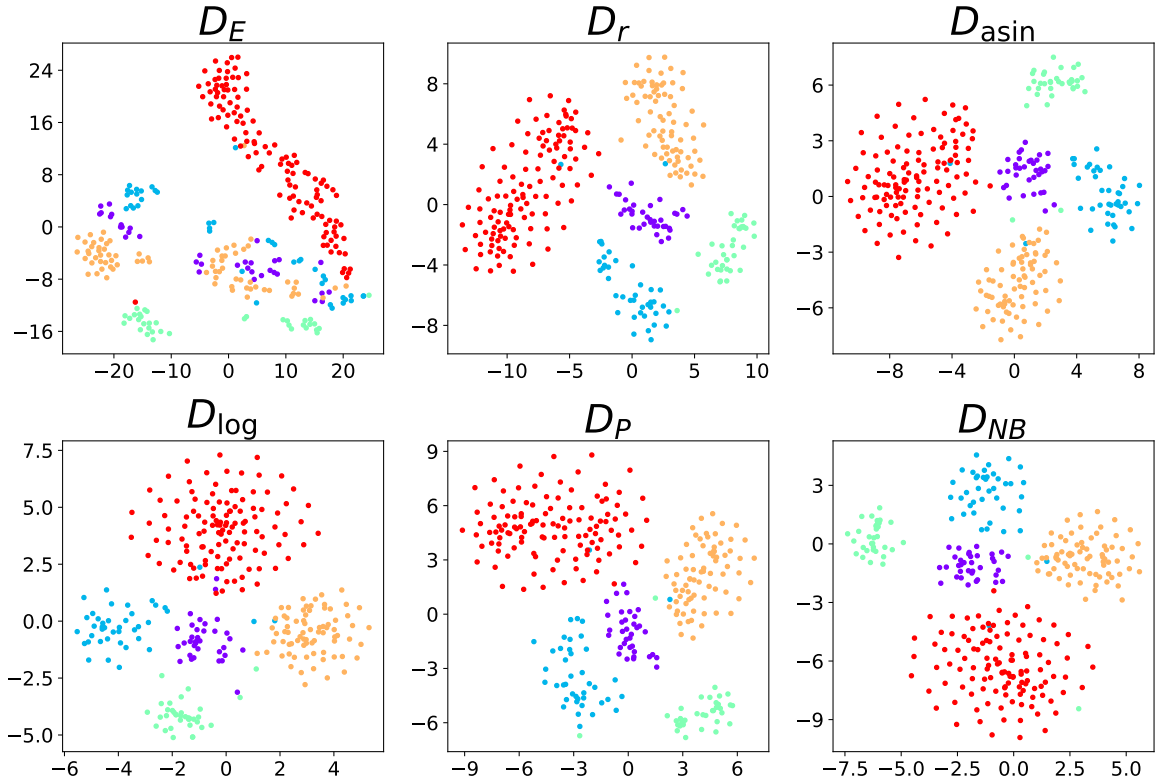

Figure S11: Visualization of the sc-CEL-seq2-5cl-p3 dataset obtained by tSNE with different measures.

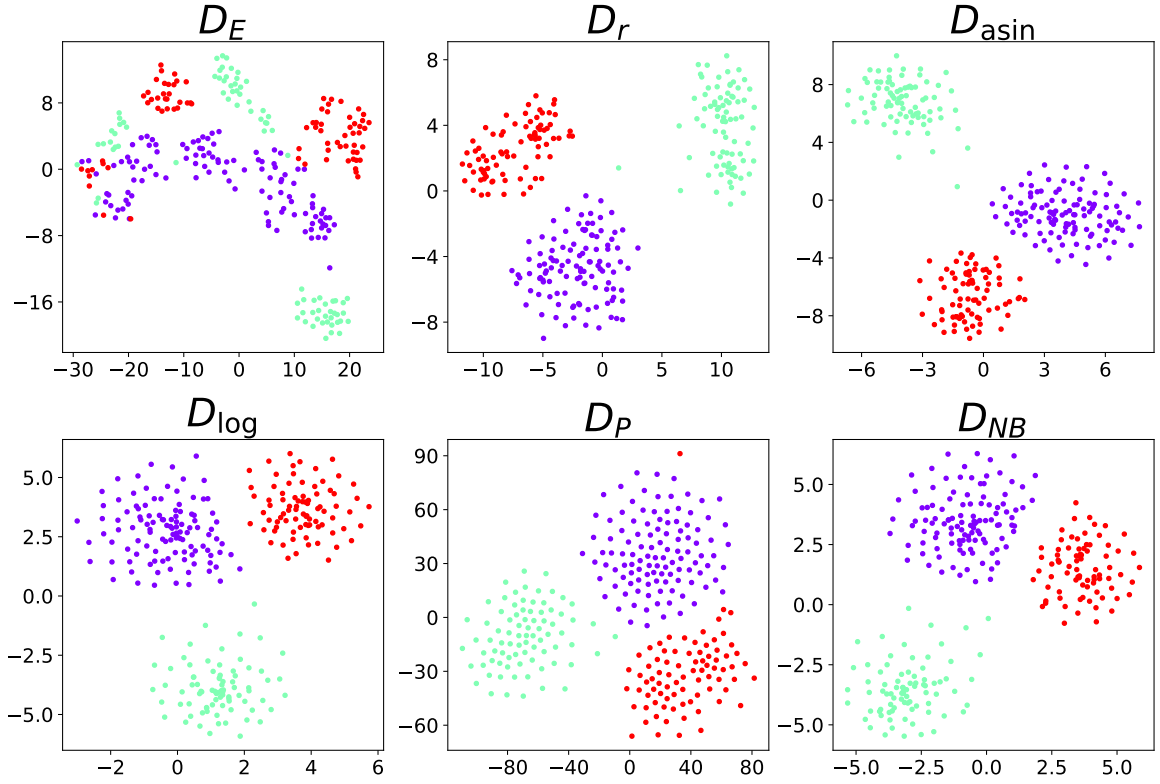

Figure S12: Visualization of the sc-CEL-seq2 dataset obtained by tSNE with different measures.

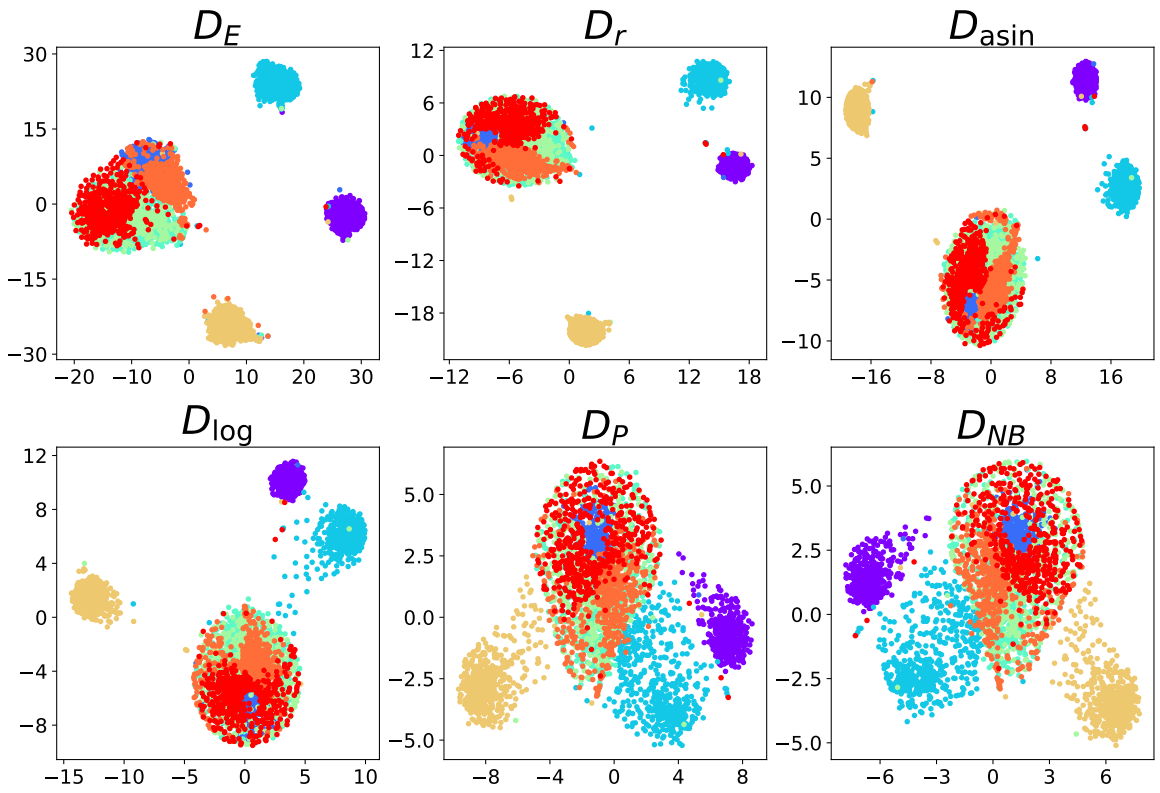

Figure S13: Visualization of the Zheng8eq dataset obtained by tSNE with different measures.

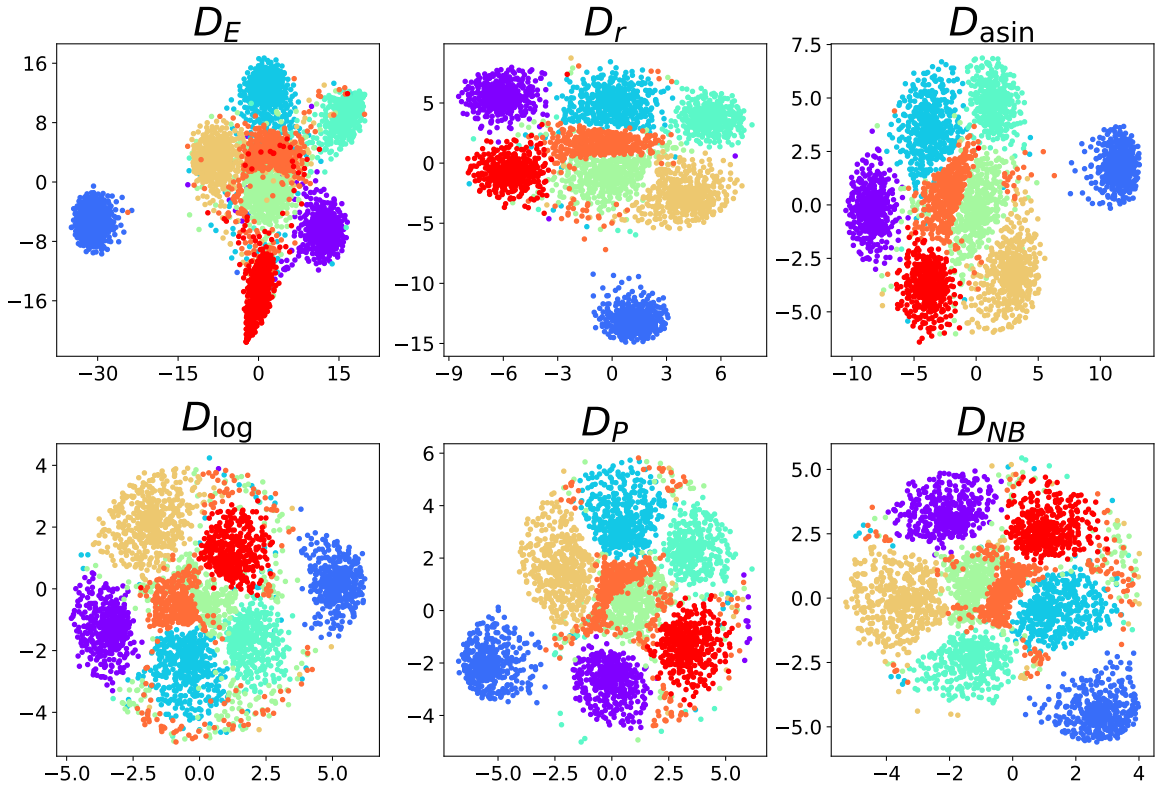

Figure S14: Visualization of the sim-Zheng8eq dataset obtained by tSNE with different measures.

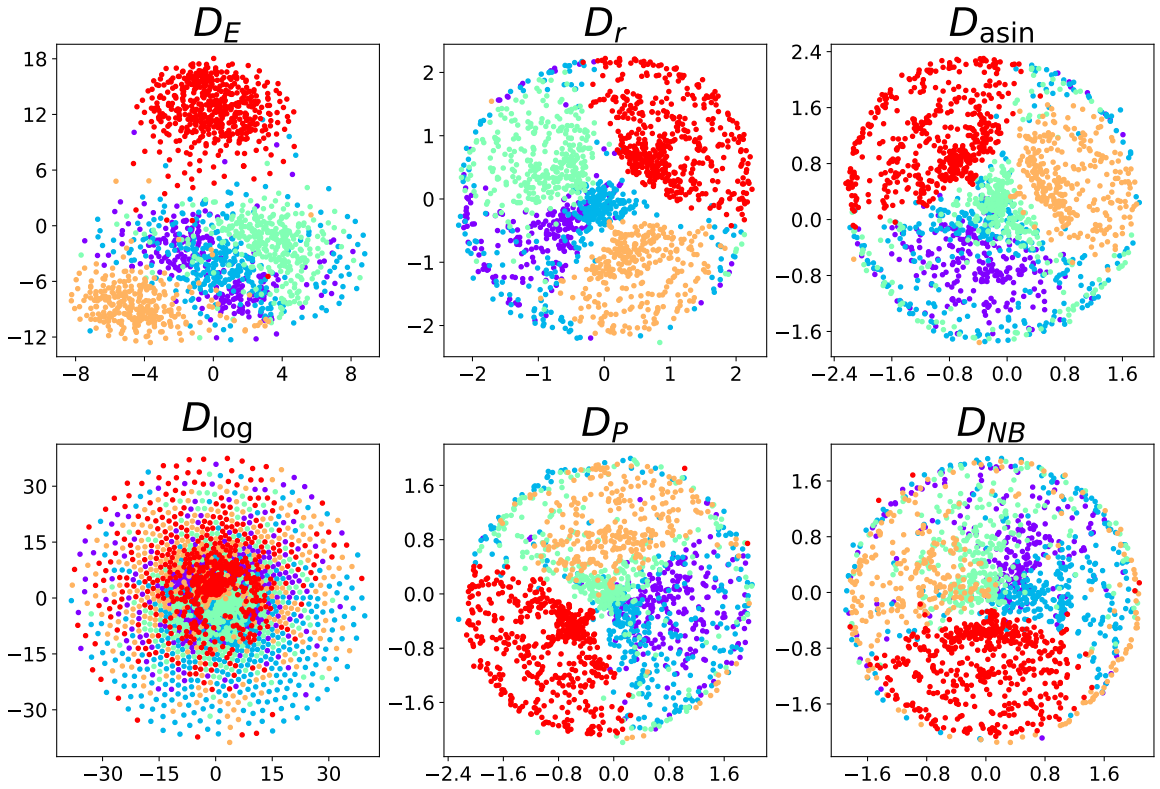

Figure S15: Visualization of the sim-manno-ESCs dataset obtained by tSNE with different measures.

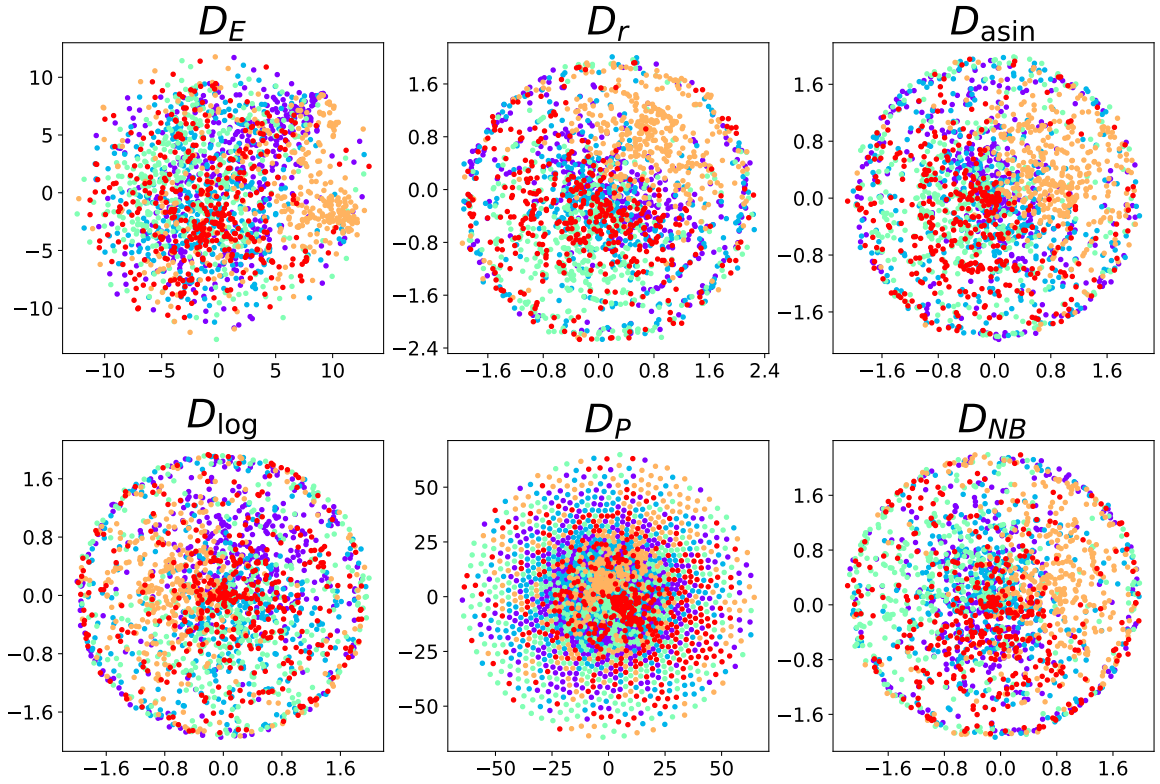

Figure S16: Visualization of the sim-manno-vm dataset obtained by tSNE with different measures.

S.5.3 VISUALIZATION OF PCA RESULTS

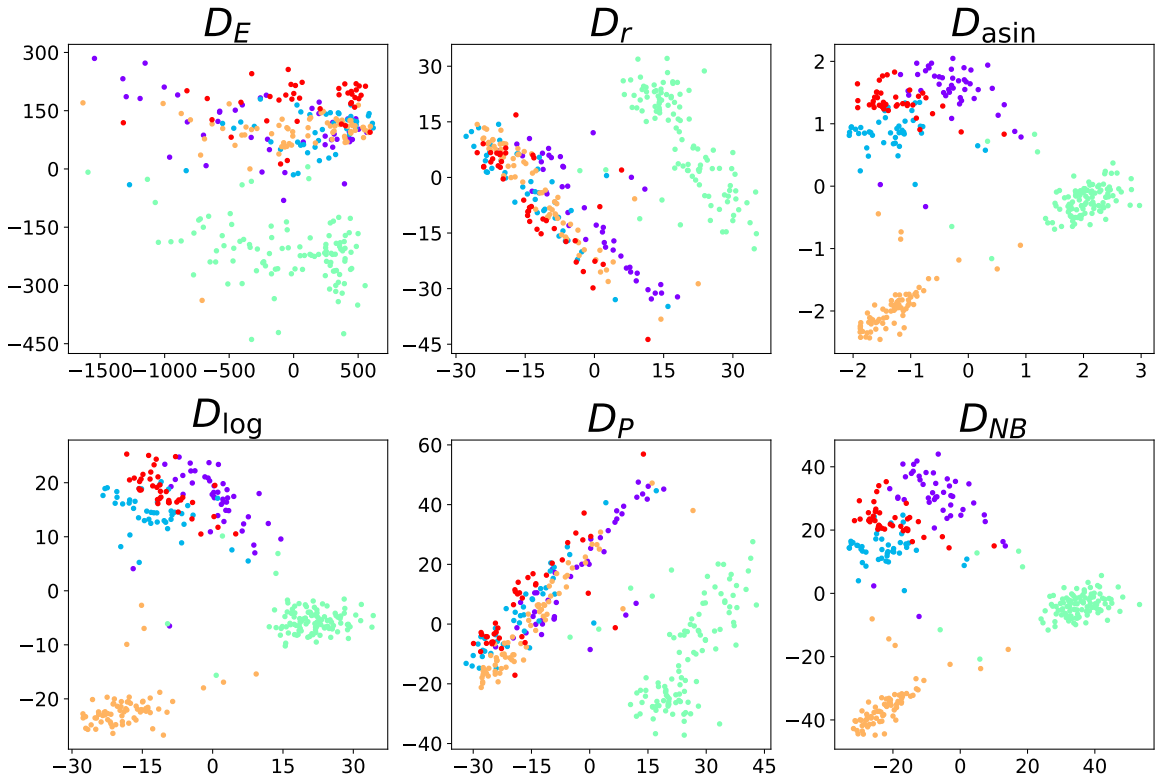

Figure S17: Visualization of the sc-CEL-seq2-5cl-p2 dataset obtained by PCA with different measures.

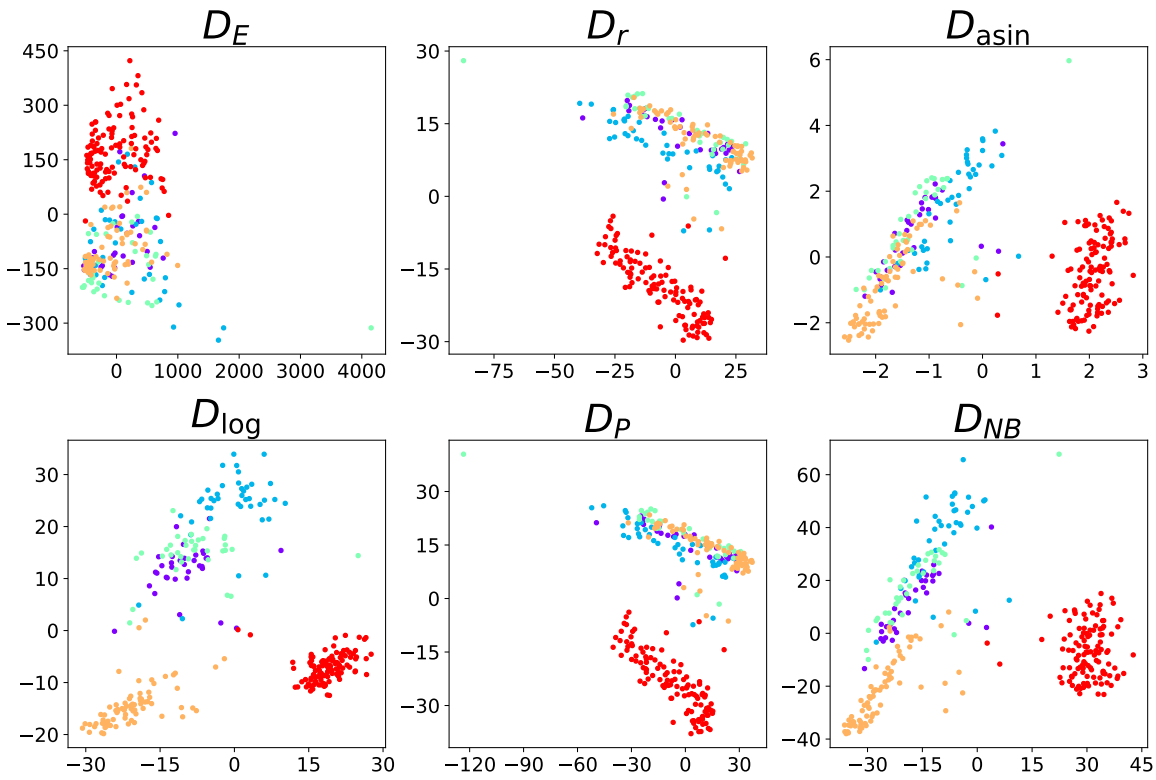

Figure S18: Visualization of the sc-CEL-seq2-5cl-p3 dataset obtained by PCA with different measures.

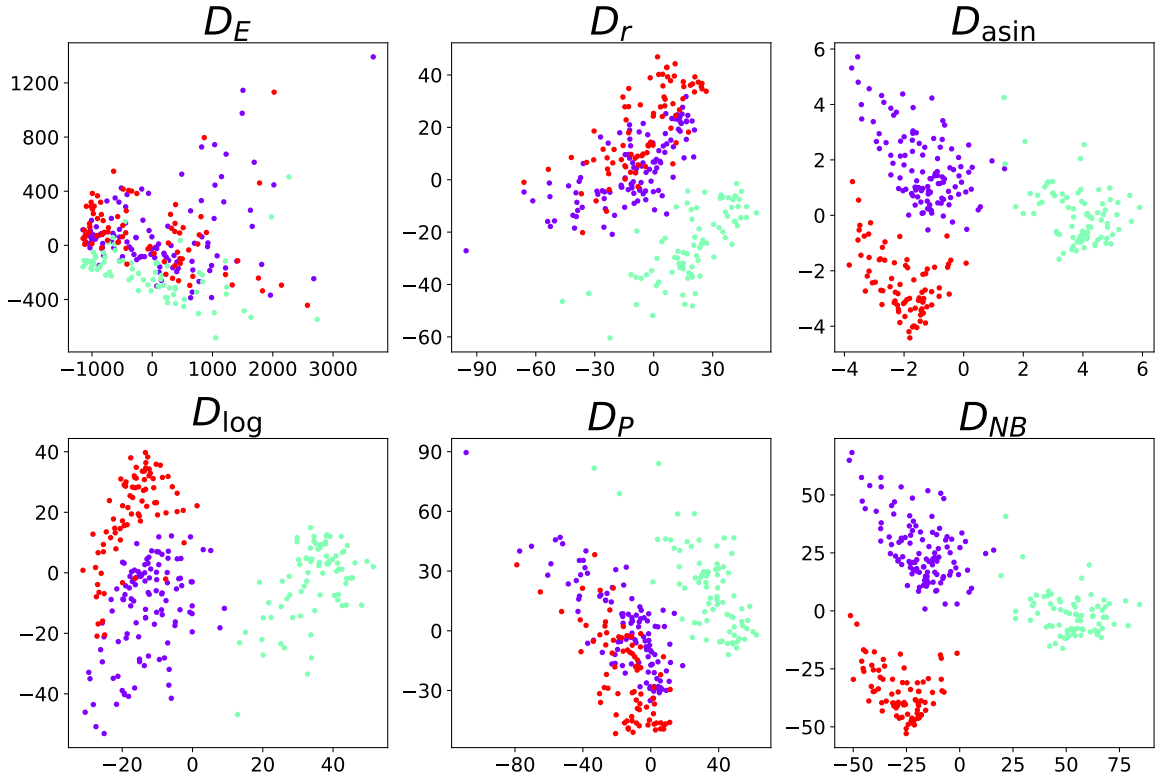

Figure S19: Visualization of the sc-CEL-seq2 dataset obtained by PCA with different measures.

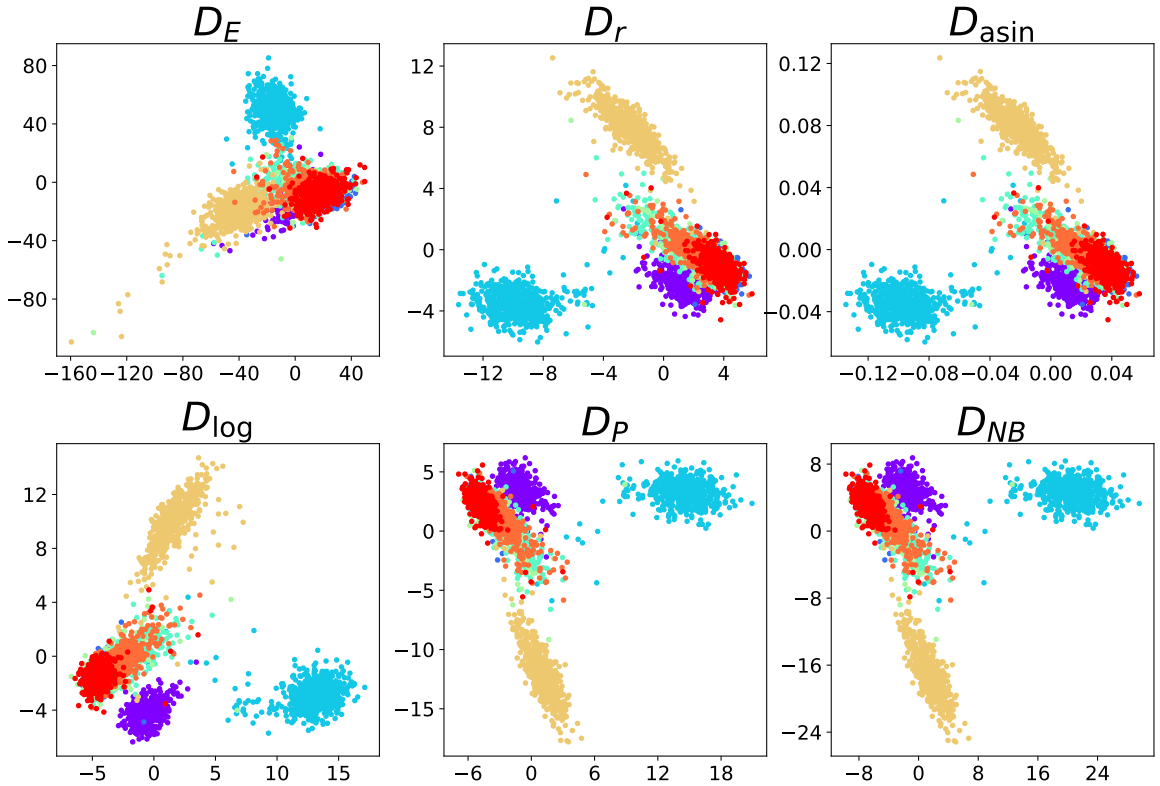

Figure S20: Visualization of the Zheng8eq dataset obtained by PCA with different measures.

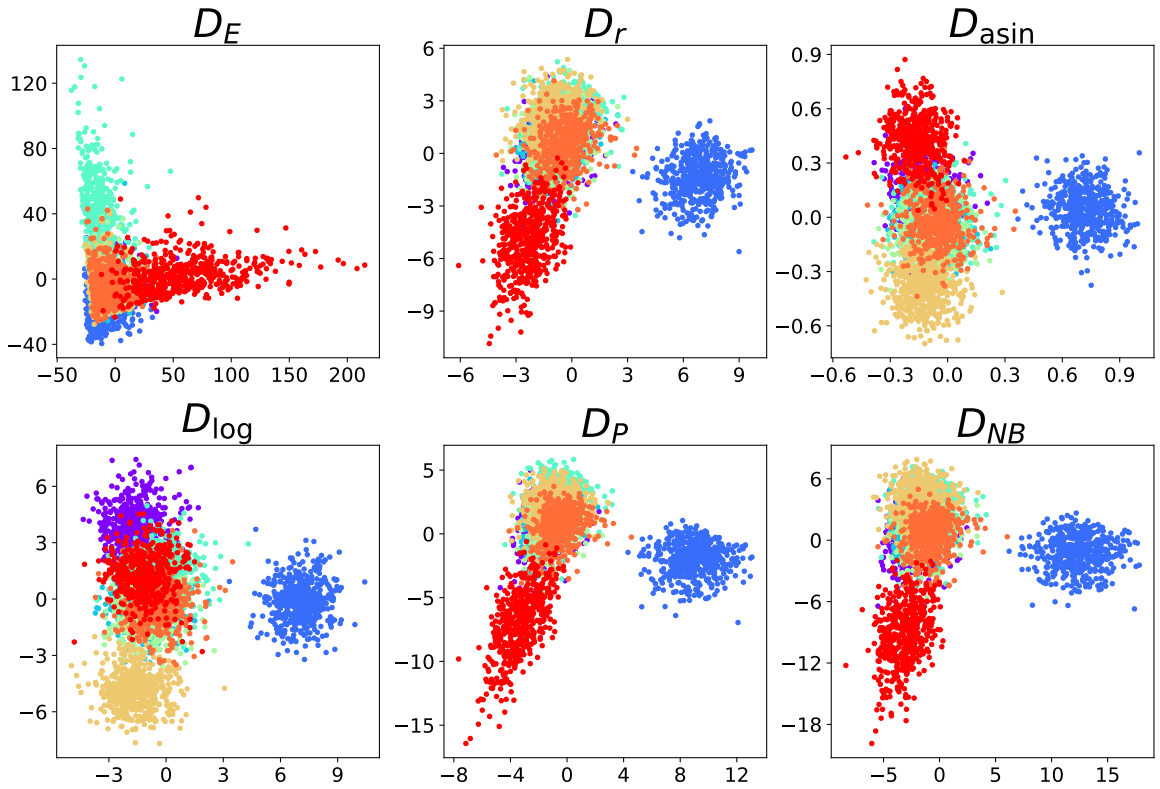

Figure S21: Visualization of the sim-Zheng8eq dataset obtained by PCA with different measures.

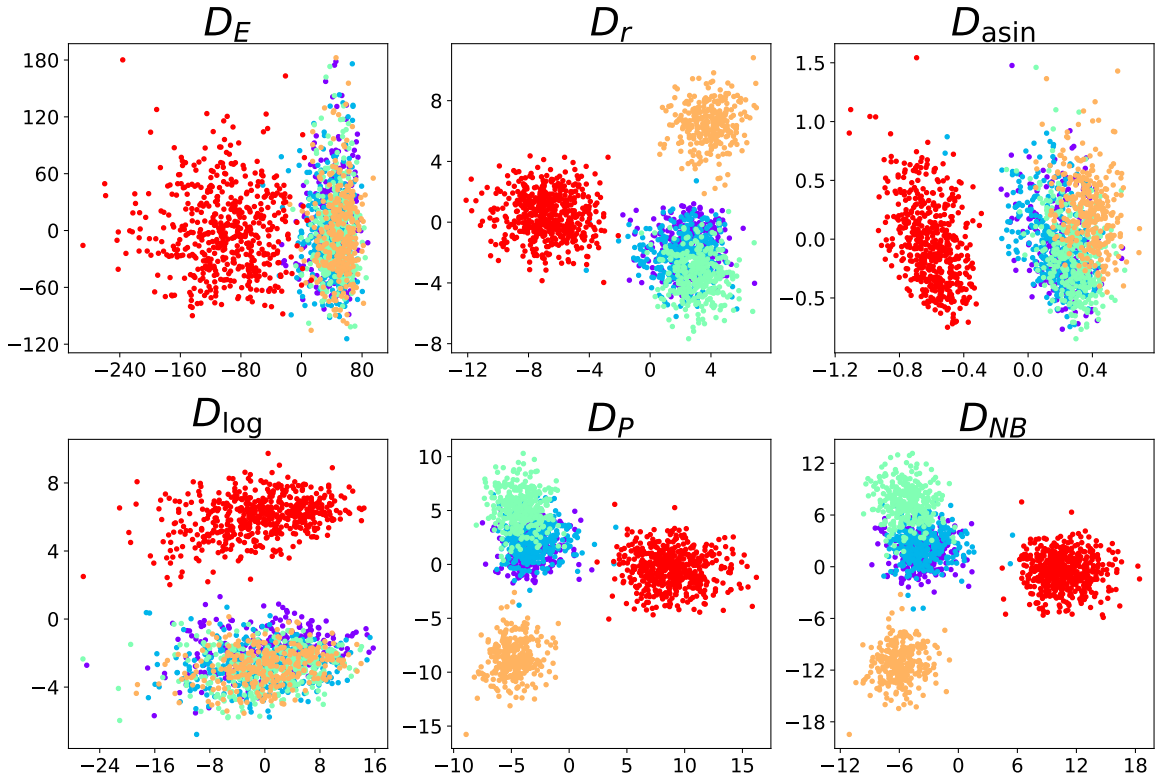

Figure S22: Visualization of the sim-manno-ESCs dataset obtained by PCA with different measures.

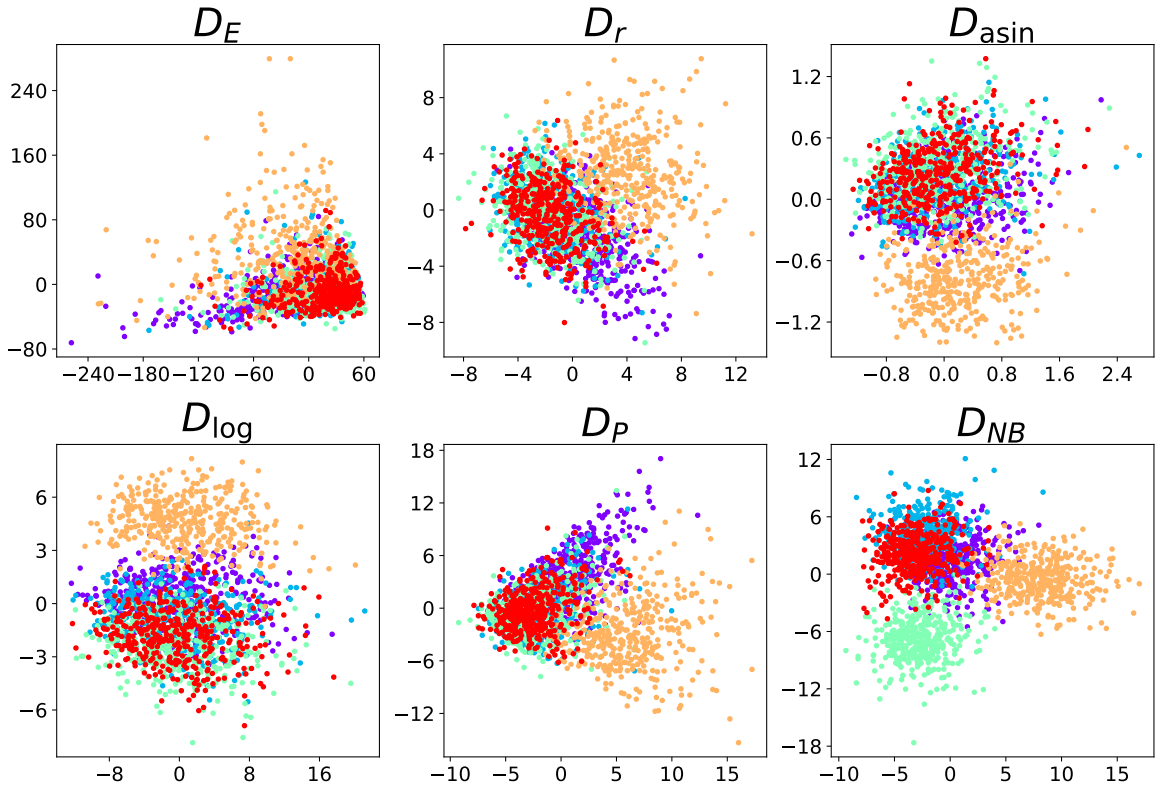

Figure S23: Visualization of the sim-manno-vm dataset obtained by PCA with different measures.

S.6 EXPERIMENTAL RESULTS OF GPCA AND NMF

S.6.1 VISUALIZATION

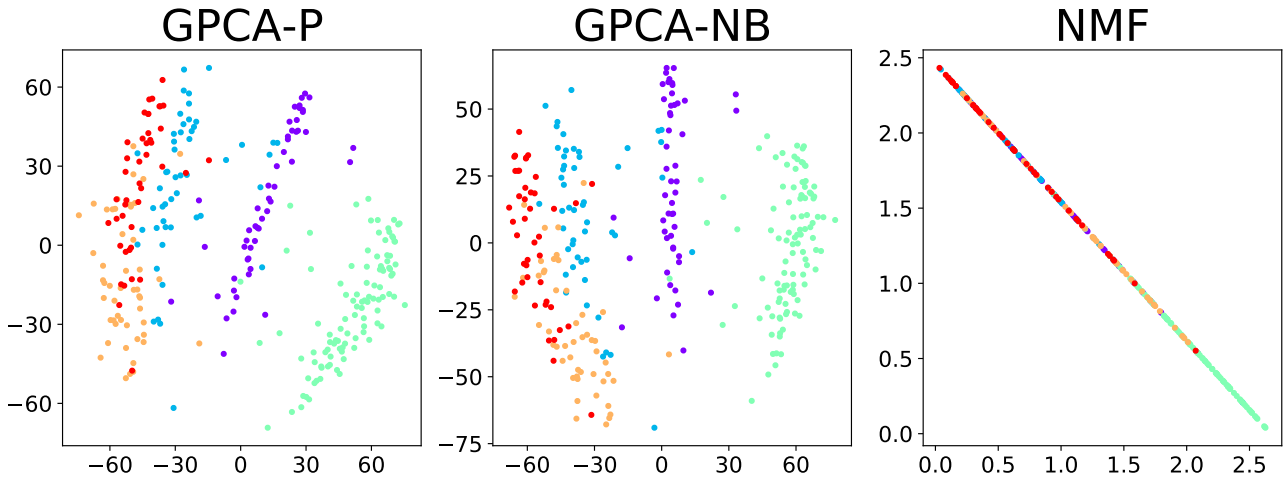

Figure S24: Visualization of the sc-CEL-seq2-5cl-p1 dataset obtained by GPCA and NMF.

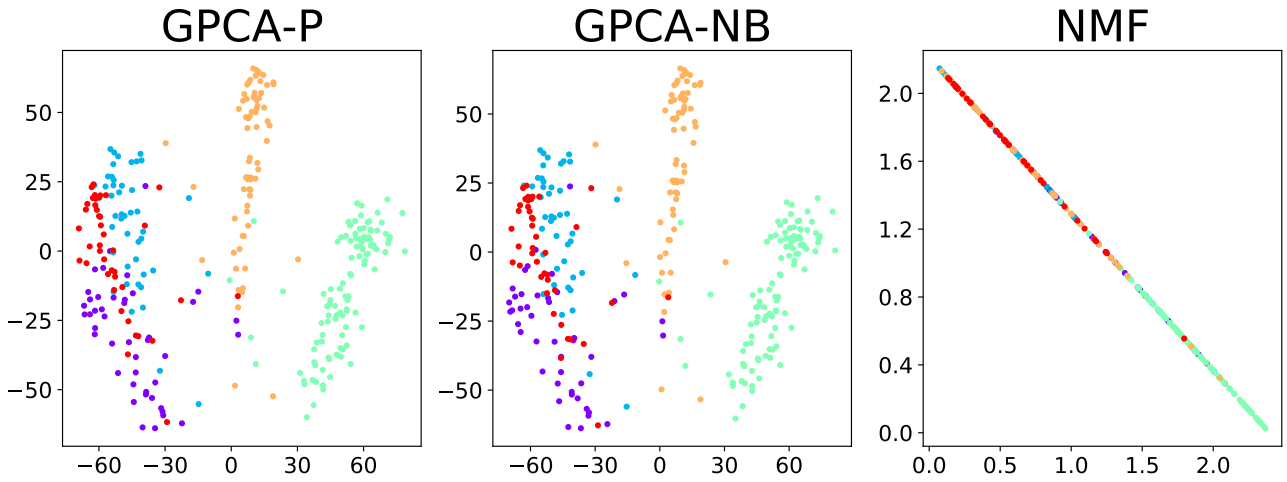

Figure S25: Visualization of the sc-CEL-seq2-5cl-p2 dataset obtained by GPCA and NMF.

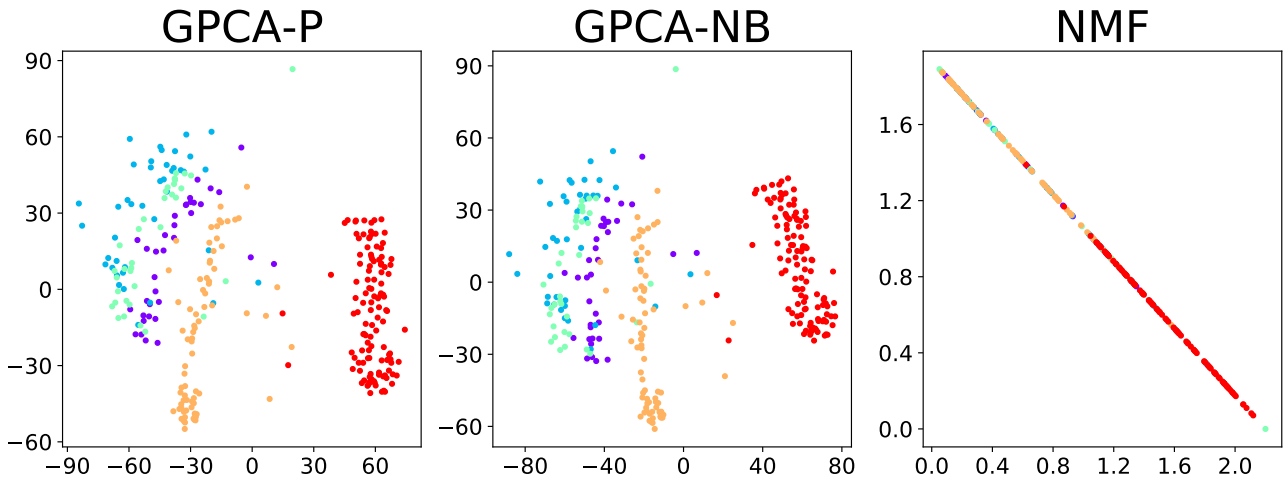

Figure S26: Visualization of the sc-CEL-seq2-5cl-p3 dataset obtained by GPCA and NMF.

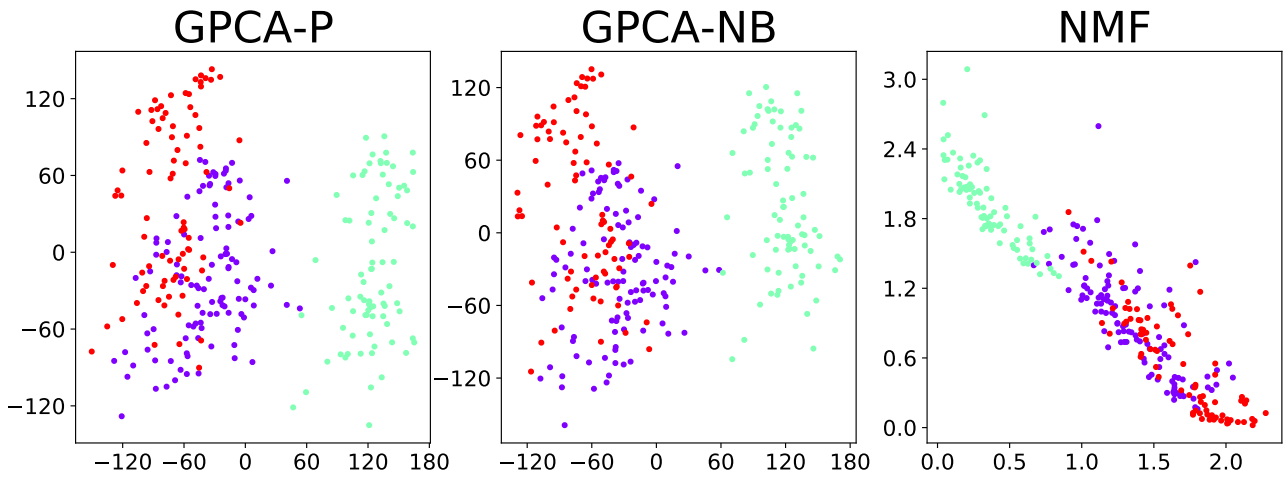

Figure S27: Visualization of the sc-CEL-seq2 dataset obtained by GPCA and NMF.

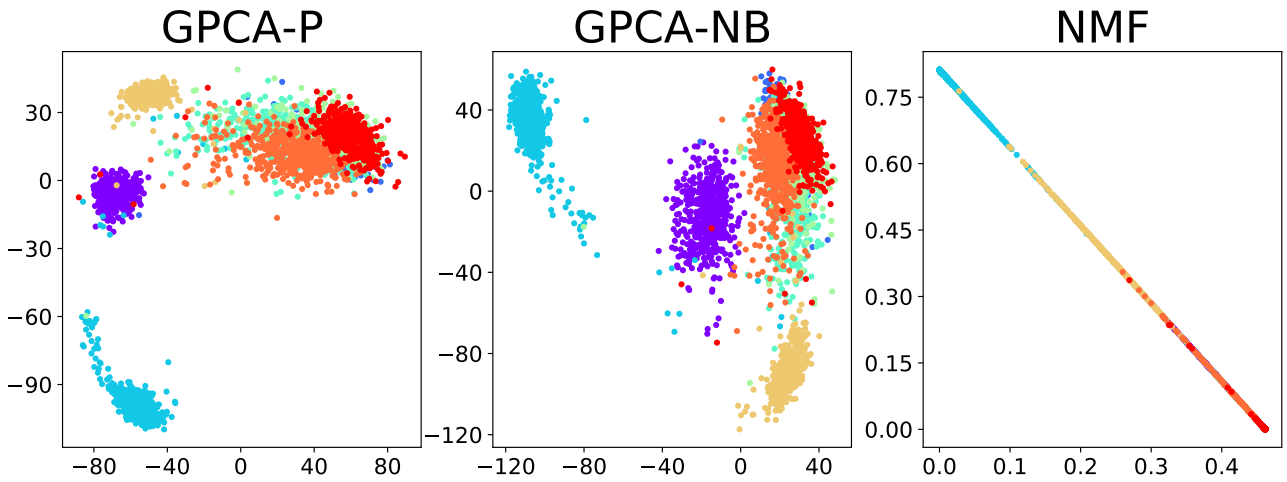

Figure S28: Visualization of the Zheng8eq dataset obtained by GPCA and NMF.

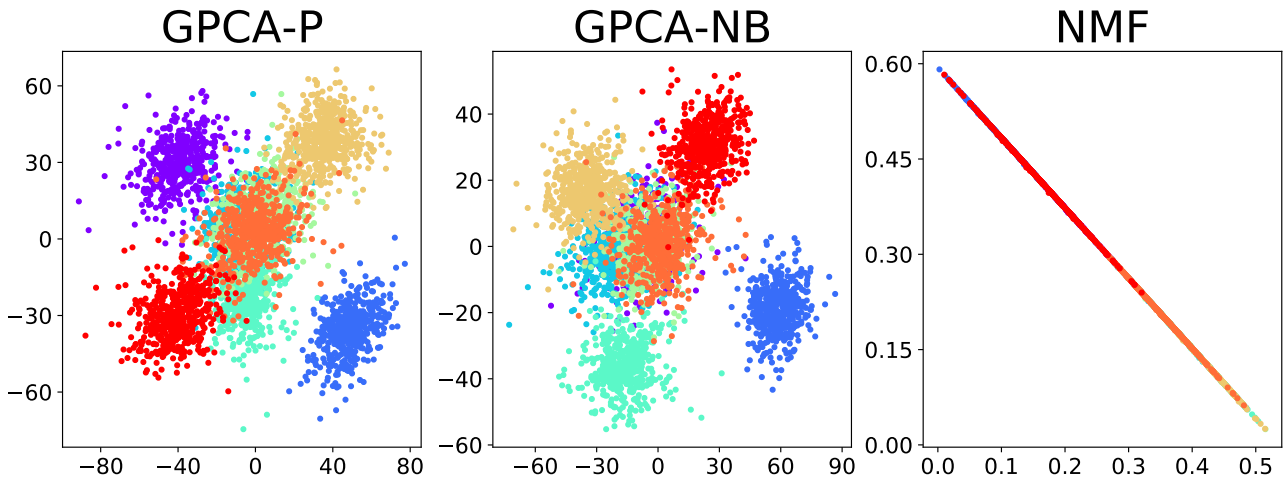

Figure S29: Visualization of the sim-Zheng8eq dataset obtained by GPCA and NMF.

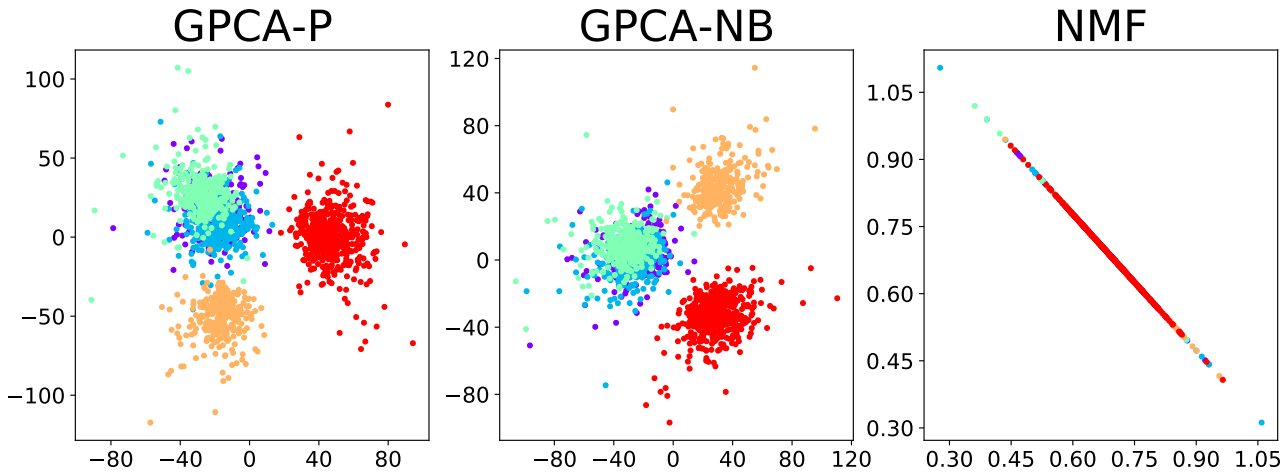

Figure S30: Visualization of the sim-manno-ESCs dataset obtained by GPCA and NMF.

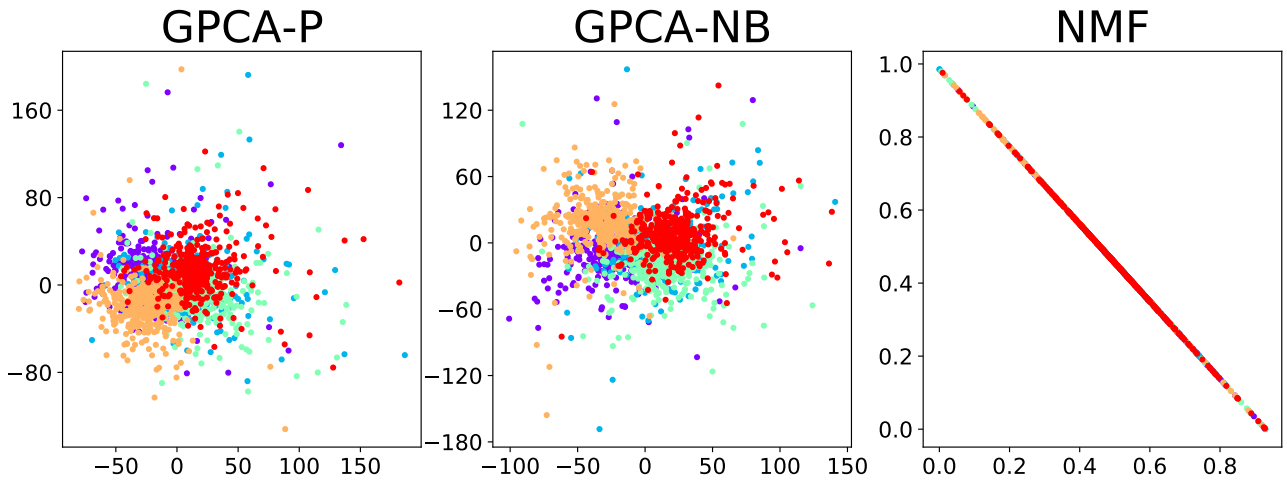

Figure S31: Visualization of the sim-manno-vm dataset obtained by GPCA and NMF.

S.6.2 CLUSTERING RESULTS

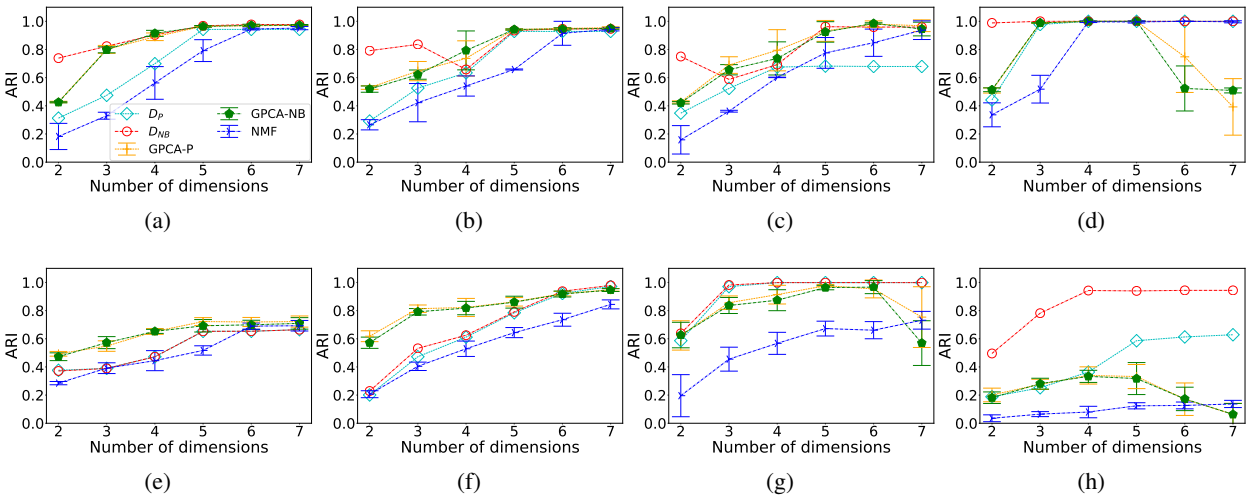

Figure S32: ARI of k-means with GPCA-P and GPCA-NB in comparison with that with PCA and the proposed measures on the following datasets: (a) sc-CEL-seq2-5cl-p1, (b) sc-CEL-seq2-5cl-p2, (c) sc-CEL-seq2-5cl-p3, (d) sc-CEL-seq2, (e) Zheng8eq, (f) sim-Zheng8eq, (g) sim-manno-ESCs, and (h) sim-manno-vm.

In this section, GPCA and NMF, which handle original non-negative data, are compared with the PCAs with D_{NB}. GPCA assumes observed data follow either Poisson or NB distributions, denoted GPCA-P and GPCA-NB, respectively. First, we assess the visualization produced by GPCA-P, GPCA-NB and NMF (Figure S24-S31) and compare them with those obtained by the PCA with D_{NB}. It is observed that the PCA with D_{NB} display more well-grouped data than GPCA-P, GPCA-NB and NMF on most datasets except the Zheng8eq dataset (Figure S28) and sim-Zheng8eq dataset (Figure S29). For these two datasets, GPCA-P and GPCA-NB distinguish the groups existing in the data more clearly compared with the PCA with D_{NB}.

Secondly, we compare the proposed measures, which are input into PCA, with GPCA-P, GPCA-NB and NMF, according to the clustering results provided in Figure S32. On the sc-CEL-seq2 dataset and the sim-manno-vm dataset, the PCA with D_{NB} not only outperforms GPCAs and NMF by a large margin when dimension is 2 but also achieves the highest ARI value with the lowest dimension. The value of ARI obtained by D_{NB} is much higher than those by GPCA-P, GPCA-NB and NMF when dimension equals 2 on the sc-CEL-seq2-5cl-p1 dataset. For the sim-manno-ESCs dataset, the ARI values of GPCAs dramatically decrease as the dimension increases from 6 to 7, while those obtained by the PCA with D_{NB} and D_P are higher and more stable. Although the ARI values of D_{NB} do not rise as the dimension grows from 2 to 4 on the sc-CEL-seq2-5cl-p1 dataset and the sc-CEL-seq2-5cl-p3 dataset, the PCA with D_{NB} outperforms GPCAs and NMF by a wide margin in the 2D space. For the Zheng8eq dataset and the sim-Zheng8eq dataset, GPCA-P accomplishes the highest value with the lowest dimension. It is found that the results of GPCA-P and GPCA-NB could be highly variable.

To sum up, the application of D_{NB} often results in better visualization and clustering results when integrated into PCA compared with GPCA-P, GPCA-NB and NMF.