Supplementary Data "Uncovering the Key Dimensions of High-Throughput Biomolecular Data using Deep Learning"

Shixiong Zhang[1], Xiangtao Li[1], Qiuzhen Lin[1], Jiecong Lin[1], and Ka-Chun Wong[1,*]

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong SAR.

[*]Corresponding author, E-mail: kc.w@cityu.edu.hk

TOP 10% GENES SELECT METHOD

We show an example of the calculation process for the first key dimension of central hidden layer. Step1, the central hidden layer $X_5^{10\times1}$ is derived from $X_4^{256\times1}$ multiplied by its weight matrix $W_4^{10\times256}$. The first row of $W_4^{10\times256}$ is contributed to the first key dimension. We select the top 10% ($256 \times 10\% = 25$) weights from it and save their indexes $Index_4$; Step 2, the hidden layer $X_4^{256\times1}$ is derived from $X_3^{640\times1}$ multiplied by its weight matrix $W_3^{256\times640}$ and then select the top 10% ($640 \times 10\% = 64$) weights from the rows of $W_3^{256\times640}$ according to $Index_4$ respectively and then save the ($64 \times 25 =$) 1,600 indexes $Index_3$. The $Index_3$ has duplicated indexes since 1,600 is larger than 640. We calculate the frequency of each index in $Index_3$ and sort it. We select the top 10% indexes with the top frequencies and then update the $Index_3$ that has fewer than 640 indexes (59 indexes for GSE60361); Step 3, the hidden layer $X_3^{640\times1}$ is derived from $X_2^{1,280\times1}$ multiplied by its weight matrix $W^{640\times1,280}$ and then select the top 10% ($1280 \times 10\% = 128$) weights from the rows of $W_2^{640\times1,280}$ according to $Index_3$ respectively and then save the ($128 \times 59 =$) 7,552 ($> 1,280$) indexes $Index_2$. We calculate the frequency of each index in $Index_2$ and sort it. We select the top 10% indexes with high frequency and then update the $Index_2$ (127 indexes for GSE60361); Step 4, the hidden layer $X_2^{1,280\times1}$ is derived from input layer $X_1^{19,972\times1}$ (for GSE60361) multiplied by its weight matrix $W_1^{1,280\times19,972}$ and then select top 10% ($19,972 \times 10\% = 1,997$) high weights

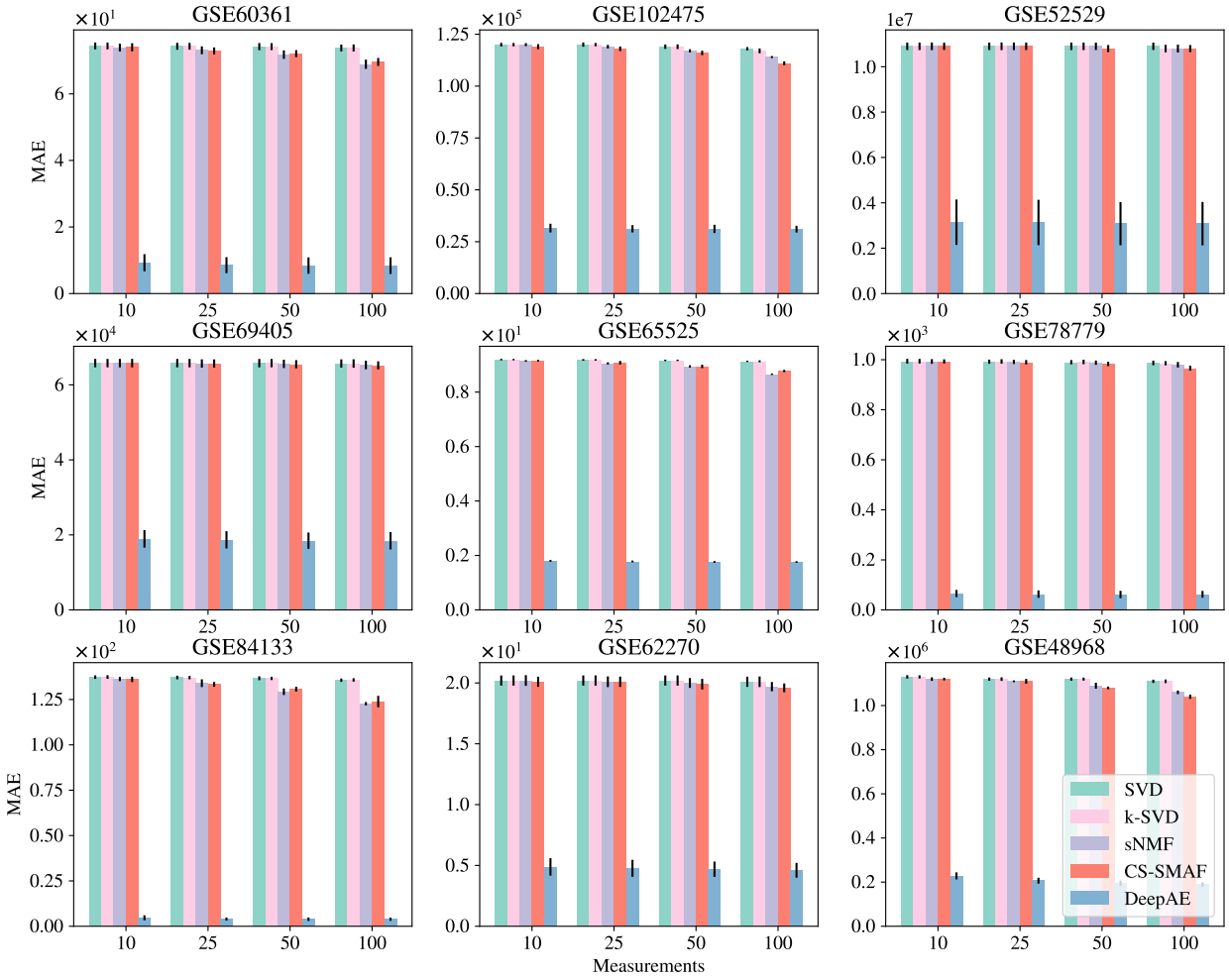


Fig. S 1. Performance comparisons of the proposed DeepAE and benchmark methods on nine single-cell RNA-seq datasets with distinct *measurement* (10, 25, 50, and 100) evaluated in *MAE* metric. Each bar height stands for the mean performance value across multiple runs, and the black line on the top of bar denotes the standard deviation; the Y-axis scale of sub-figure is different with each other. The benchmark methods include SVD, k-SVD, sNMF, and CS-SMAF.

from the rows of $W_1^{1,280\times19,972}$ according to $Index_2$ respectively and then save the $(1,997 \times 127 =)$ 253,619 $(> 19,972)$ indexes $Index_1$. We calculate the frequency of each index in $Index_1$ and sort it. Finally, we select the top 10% (1,997 for GSE60361) indexes from updated $Index_1$ according to the frequency, resulting in the top 10% genes selected.

Table S 1. Performance comparisons of four distinct DeepAE architectures on single-cell RNA-seq data evaluated by *PCC*, *EM*, and *MAE* with standard deviations. The best performance of architectures on each dataset is highlighted in bold.

Datasets	Metrics	DeepAE1	DeepAE2	DeepAE3	DeepAE4
GSE60361	*PCC*	0.8155 ± 0.0243	0.9192 ± 0.0218	0.9337 ± 0.0198	**0.9420 ± 0.0192**
	EM	3.81E+04 ± 2.19E+03	2.55E+04 ± 3.53E+03	2.32E+04 ± 3.47E+03	**2.16E+04 ± 3.64E+03**
	MAE	2.55E+01 ± 2.85E+00	1.16E+01 ± 3.08E+00	9.57E+00 ± 2.80E+00	**8.40E+00 ± 2.75E+00**
GSE71858	*PCC*	0.8551 ± 0.0369	0.9341 ± 0.0166	0.9439 ± 0.0120	**0.9531 ± 0.0110**
	EM	3.39E+05 ± 2.88E+04	2.35E+05 ± 3.20E+04	2.15E+05 ± 2.69E+04	**1.94E+05 ± 2.49E+04**
	MAE	1.26E+04 ± 2.11E+04	6.07E+04 ± 1.55E+04	5.07E+04 ± 1.24E+04	**4.13E+04 ± 1.10E+04**
GSE62270	*PCC*	0.8396 ± 0.0294	0.8686 ± 0.0132	0.8716 ± 0.0178	**0.8770 ± 0.0168**
	EM	6.28E+03 ± 7.52E+02	5.75E+03 ± 3.84E+02	5.63E+03 ± 4.53E+02	**5.50E+03 ± 4.29E+02**
	MAE	6.17E+00 ± 1.52E+00	5.12E+00 ± 6.92E-01	4.92E+00 ± 7.70E-01	**4.69E+00 ± 7.06E-01**
GSE48968	*PCC*	0.7977 ± 0.0095	0.8932 ± 0.0045	0.9035 ± 0.0052	**0.9097 ± 0.0059**
	EM	1.55E+06 ± 3.69E+04	1.16E+06 ± 2.68E+04	1.10E+06 ± 3.30E+04	**1.07E+06 ± 3.76E+04**
	MAE	4.09E+05 ± 1.94E+04	2.29E+05 ± 1.06E+04	2.07E+05 ± 1.23E+04	**1.95E+05 ± 1.36E+04**
GSE52529	*PCC*	0.8209 ± 0.1116	0.8650 ± 0.0262	0.8712 ± 0.0239	**0.8735 ± 0.0235**
	EM	8.14E+06 ± 2.08E+06	7.51E+06 ± 1.03E+06	7.34E+06 ± 1.26E+06	**7.09E+06 ± 1.20E+06**
	MAE	4.17E+06 ± 2.26E+06	3.46E+06 ± 1.21E+06	3.30E+06 ± 1.15E+06	**3.08E+06 ± 1.06E+06**
GSE77564	*PCC*	0.9004 ± 0.0140	0.9345 ± 0.0175	0.9538 ± 0.0125	**0.9703 ± 0.0070**
	EM	2.62E+05 ± 5.52E+04	1.55E+05 ± 6.73E+04	1.31E+05 ± 1.48E+04	**1.03E+05 ± 1.33E+04**
	MAE	1.97E+05 ± 7.20E+04	1.47E+04 ± 1.22E+04	4.79E+04 ± 4.54E+04	**2.97E+04 ± 7.55E+03**
GSE78779	*PCC*	0.9636 ± 0.0046	0.9683 ± 0.0059	0.9674 ± 0.0062	**0.9691 ± 0.0059**
	EM	1.23E+04 ± 1.37E+03	1.15E+04 ± 1.59E+03	1.16E+04 ± 1.58E+03	**1.12E+04 ± 1.45E+03**
	MAE	7.34E+01 ± 1.68E+01	6.46E+01 ± 1.86E+01	6.59E+01 ± 1.87E+01	**6.15E+01 ± 1.66E+01**
GSE69405	*PCC*	0.8214 ± 0.0206	**0.8526 ± 0.0159**	0.8471 ± 0.0189	0.8481 ± 0.0129
	EM	4.86E+05 ± 2.84E+04	**4.45E+05 ± 2.64E+04**	4.52E+05 ± 3.03E+04	4.51E+05 ± 3.12E+04
	MAE	2.14E+04 ± 2.52E+03	**1.80E+04 ± 2.08E+03**	1.86E+04 ± 2.41E+03	1.85E+04 ± 2.45E+03
GSE102475	*PCC*	0.8506 ± 0.0094	0.8474 ± 0.0096	0.8467 ± 0.0095	**0.8532 ± 0.0105**
	EM	1.83E+05 ± 5.73E+03	1.84E+05 ± 5.66E+03	1.85E+05 ± 5.61E+03	**1.81E+05 ± 6.37E+03**
	MAE	3.18E+04 ± 2.00E+03	3.23E+04 ± 2.00E+03	3.24E+04 ± 1.98E+03	**3.12E+04 ± 2.22E+03**

Table S 2. Performance comparisons of DeepAE4 and two deeper architectures (DeepAE5 and DeepAE6) on GSE84133 and GSE65525 evaluated by *PCC, EM, and MAE* with standard deviations. The best performance of architectures on each dataset is highlighted in bold.

Datasets	Metrics	DeepAE4	DeepAE5	DeepAE6
GSE84133	*PCC*	0.9860 ± 0.0040	**0.9887 ±0.0032**	0.9882 ± 0.0023
	EM	1.01E+04 ± 1.29E+03	**9.09E+03 ± 1.17E+03**	9.32E+03 ± 8.51E+02
	MAE	3.90E+00 ± 1.05E+00	**3.14E+00 ± 8.56E-01**	3.28E+00 ± 6.14E-01
	Running Time	**1445.55 secs.**	2542.53 secs.	2567.37 secs.
GSE65525	*PCC*	**0.8914 ± 0.0022**	0.8903 ± 0.0029	0.8879 ± 0.0029
	EM	**1.05E+04 ± 1.06E+02**	1.05E+04 ± 1.45E+02	1.06E+04 ± 1.34E+02
	MAE	**1.76E+00 ± 3.56E-02**	1.78E+00 ± 4.88E-02	1.82E+00 ± 4.57E-02
	Running Time	**2767.70 secs.**	4593.58 secs.	4631.89 secs.

Table S3. Performance comparisons of DeepAE and benchmark methods on single-cell RNA-seq data (*measurements* = 50) evaluated by *PCC*, *EM*, and *MAE* with standard deviations. The best performance of methods on each dataset is highlighted in bold. The benchmark methods include SVD, k-SVD, sNMF, and CS-SMAF.

Datasets	Metrics	SVD	k-SVD	sNMF	CS-SMAF	DeepAE
GSE60361	*PCC*	0.6731 ± 0.0575	0.8406 ± 0.0543	0.8949 ± 0.0124	0.8553 ± 0.0167	**0.9420 ± 0.0192**
	EM	6.50E+04 ± 4.75E+00	6.50E+04 ± 4.58E+02	6.40E+04 ± 5.69E+02	6.41E+04 ± 4.71E+02	**2.16E+04 ± 3.64E+03**
	MAE	7.42E+01 ± 1.08E+00	7.42E+01 ± 1.05E+00	7.18E+01 ± 1.28E+00	7.21E+01 ± 1.06E+00	**8.40E+00 ± 2.75E+00**
GSE84133	*PCC*	0.6571 ± 0.0609	0.8897 ± 0.0263	0.9632 ± 0.0076	0.9604 ± 0.0027	**0.9860 ± 0.0040**
	EM	5.88E+04 ± 2.35E+02	5.88E+04 ± 1.96E+02	5.72E+04 ± 4.01E+02	5.75E+04 ± 2.89E+02	**1.01E+04 ± 1.29E+03**
	MAE	1.37E+02 ± 1.09E+00	1.37E+02 ± 9.13E-01	1.29E+02 ± 1.81E+00	1.31E+02 ± 1.32E+00	**3.90E+00 ± 1.05E+00**
GSE62270	*PCC*	0.8227 ± 0.0469	0.8317 ± 0.0258	0.8289 ± 0.0164	0.8487 ± 0.0163	**0.8770 ± 0.0168**
	EM	1.44E+04 ± 1.18E+02	1.14E+04 ± 1.19E+02	1.14E+04 ± 1.14E+02	1.14E+04 ± 1.24E+02	**5.50E+03 ± 4.29E+02**
	MAE	2.02E+01 ± 4.15E-01	2.02E+01 ± 0.42E+00	2.00E+01 ± 4.02E-01	1.99E+01 ± 4.35E-01	**4.69E+00 ± 7.06E-01**
GSE48968	*PCC*	0.7829 ± 0.0407	0.8260 ± 0.0248	0.8549 ± 0.0113	0.8673 ± 0.0106	**0.9097 ± 0.0059**
	EM	2.57E+06 ± 8.26E+03	2.56E+06 ± 7.67E+03	2.53E+06 ± 1.58E+04	2.52E+06 ± 7.24E+03	**1.07E+06 ± 3.76E+04**
	MAE	1.12E+06 ± 6.69E+03	1.12E+06 ± 6.69E+03	1.09E+06 ± 1.36E+04	1.08E+06 ± 6.21E+03	**1.95E+05 ± 1.36E+04**
GSE52529	*PCC*	0.7840 ± 0.0368	0.7964 ± 0.0371	0.7058 ± 0.0950	0.7859 ± 0.0480	**0.8735 ± 0.0235**
	EM	1.35E+07 ± 1.03E+05	1.35E+07 ± 1.03E+05	1.35E+07 ± 1.02E+05	1.35E+07 ± 9.81E+04	**7.09E+06 ± 1.20E+06**
	MAE	1.09E+07 ± 1.65E+05	1.09E+07 ± 1.66E+05	1.09E+07 ± 1.63E+05	1.08E+07 ± 1.57E+05	**3.08E+06 ± 1.06E+06**
GSE65525	*PCC*	0.5922 ± 0.0444	0.8065 ± 0.0596	0.8663 ± 0.0121	0.8611 ± 0.0056	**0.8914 ± 0.0022**
	EM	2.33E+04 ± 3.46E+01	2.33E+04 ± 3.34E+01	2.30E+04 ± 5.51E+01	2.30E+04 ± 8.16E+01	**1.05E+04 ± 1.06E+02**
	MAE	9.17E+00 ± 2.72E-02	9.17E+05 ± 2.64E-02	8.94E+00 ± 4.28E-02	8.94E+00 ± 6.36E-02	**1.76E+04 ± 3.56E-202**
GSE78779	*PCC*	0.9563 ± 0.0242	0.9546 ± 0.0105	0.9615 ± 0.0042	0.8720 ± 0.0325	**0.9691 ± 0.0059**
	EM	4.53E+04 ± 1.94+02	4.54E+04 ± 2.08E+02	4.53E+04 ± 1.96E+02	4.52E+04 ± 2.00E+02	**1.12E+04 ± 1.45E+03**
	MAE	9.90E+02 ± 8.45E+00	9.91E+02 ± 9.08E+00	9.88E+02 ± 8.53E+00	9.83E+02 ± 8.71E+00	**6.15E+01 ± 1.66E+01**
GSE69405	*PCC*	0.8458 ± 0.0146	0.8299 ± 0.0196	0.8165 ± 0.0218	0.8334 ± 0.0263	**0.8481 ± 0.0.29**
	EM	8.52E+05 ± 7.31E+03	8.52E+05 ± 7.23E+03	8.51E+05 ± 7.24E+03	8.51E+05 ± 7.17E+03	**4.51E+05 ± 3.12E+04**
	MAE	6.58E+04 ± 1.12E+03	6.58E+04 ± 1.11E+03	6.56E+04 ± 1.11E+03	6.55E+04 ± 1.10E+03	**1.85E+04 ± 2.45E+03**
GSE102475	*PCC*	0.8233 ± 0.0154	0.8098 ± 0.0061	0.8058 ± 0.0110	0.8151 ± 0.0085	**0.8532 ± 0.0105**
	EM	3.54E+05 ± 1.56E+03	3.54E+05 ± 1.68E+03	3.50E+05 ± 1.77E+03	3.49E+05 ± 1.61E+03	**1.81E+05 ± 6.37E+03**
	MAE	1.19E+05 ± 1.05E+03	1.19E+05 ± 1.13E+03	1.17E+05 ± 7.81E+02	1.16E+05 ± 1.07E+03	**3.12E+04 ± 2.22E+03**

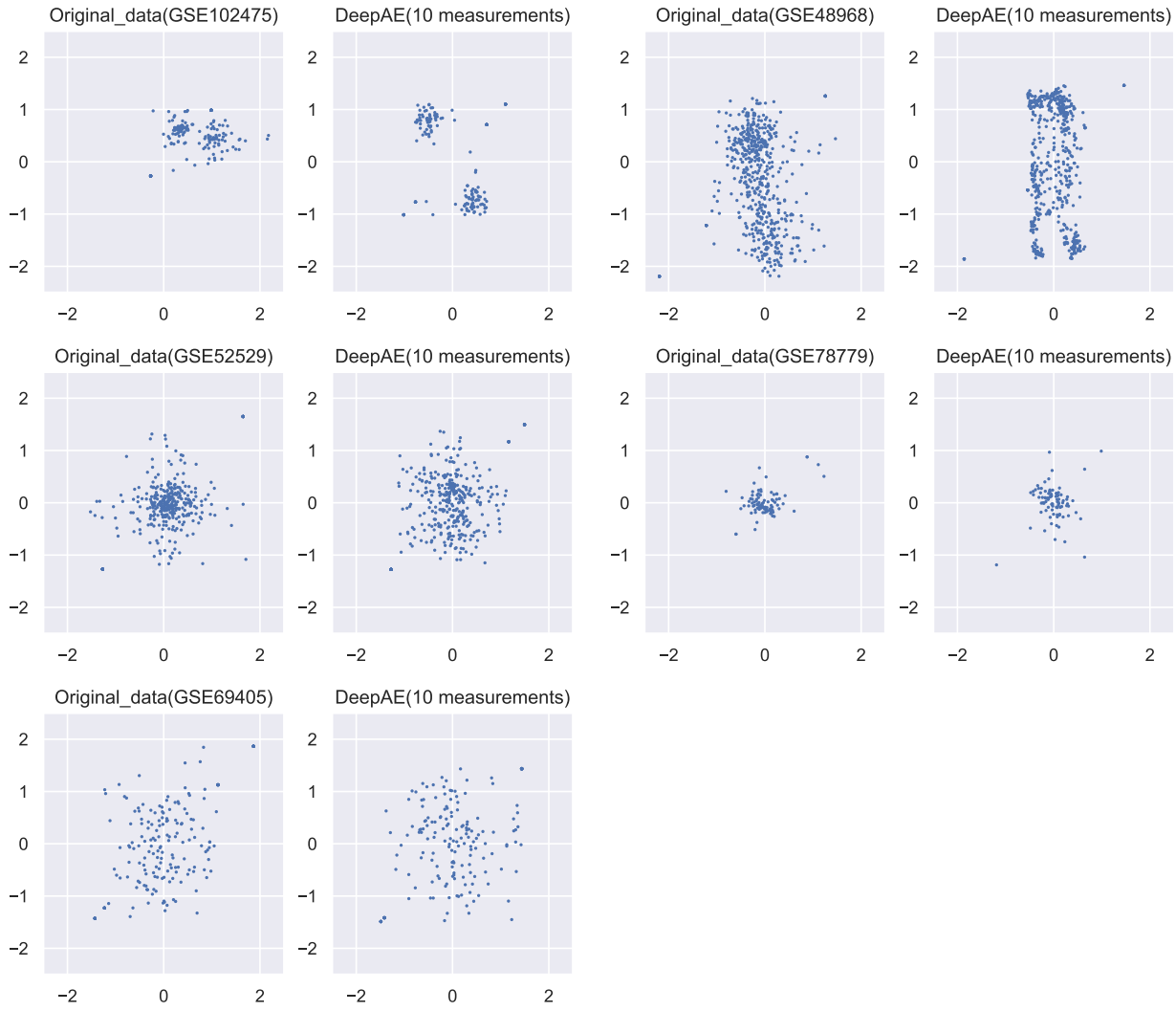
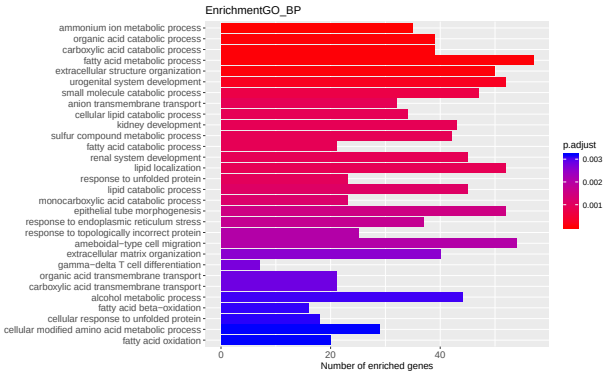
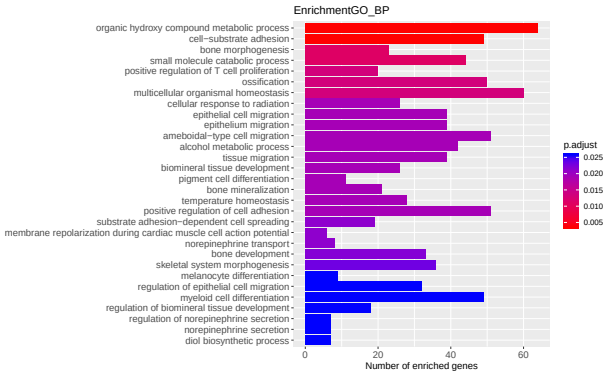


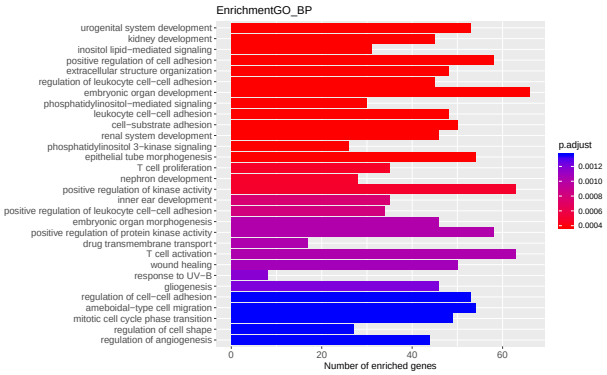
Fig. S 2. Key dimension (measurement) t-SNE 2D visualizations between the original data (~20,000 dimensions) and compressed data (10 dimensions) from the transcriptomic profiling datasets (GSE102475, GSE48968, GSE52529, GSE78779, and GSE69405).



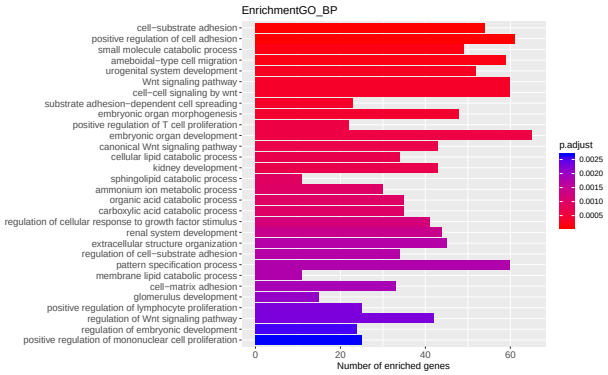
(a) First dimension in central hidden layer.



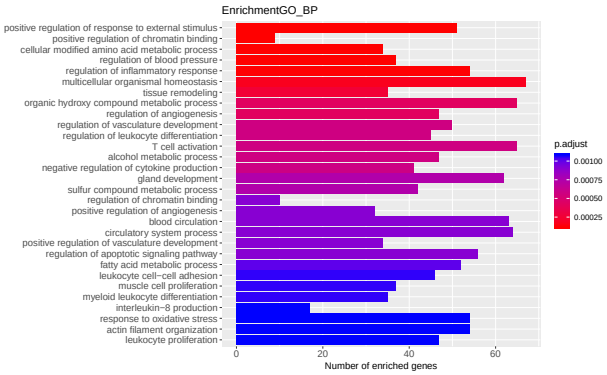
(b) Second dimension in central hidden layer.



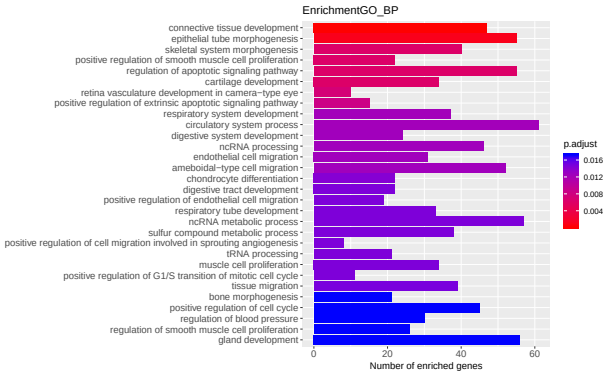
(c) Third dimension in central hidden layer.



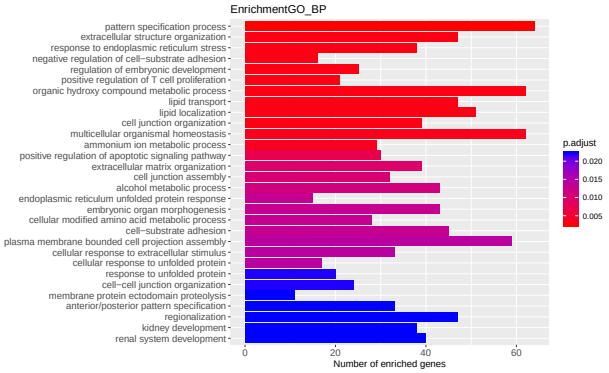
(d) Forth dimension in central hidden layer.



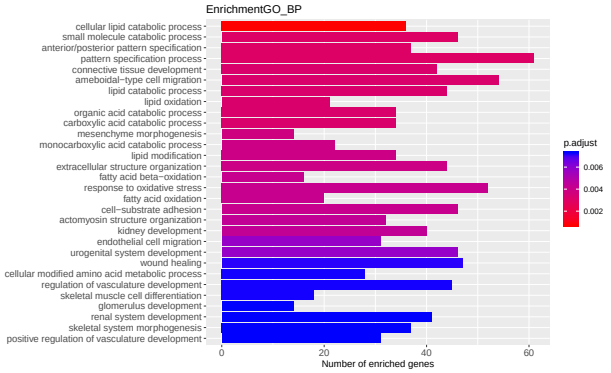
(e) Fifth dimension in central hidden layer.



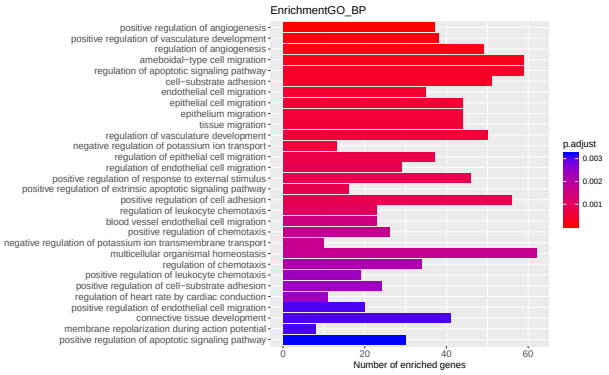
(f) Sixth dimension in central hidden layer.



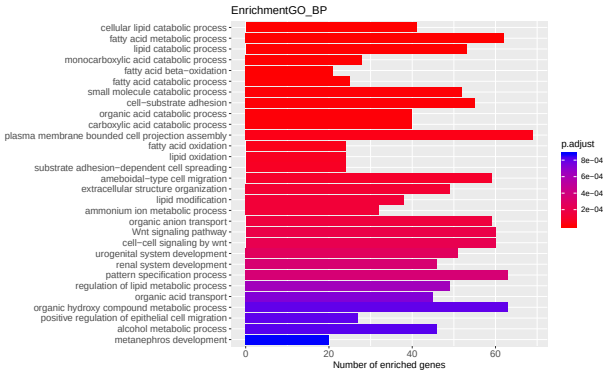
(g) Seventh dimension in central hidden layer.



(h) Eighth dimension in central hidden layer.

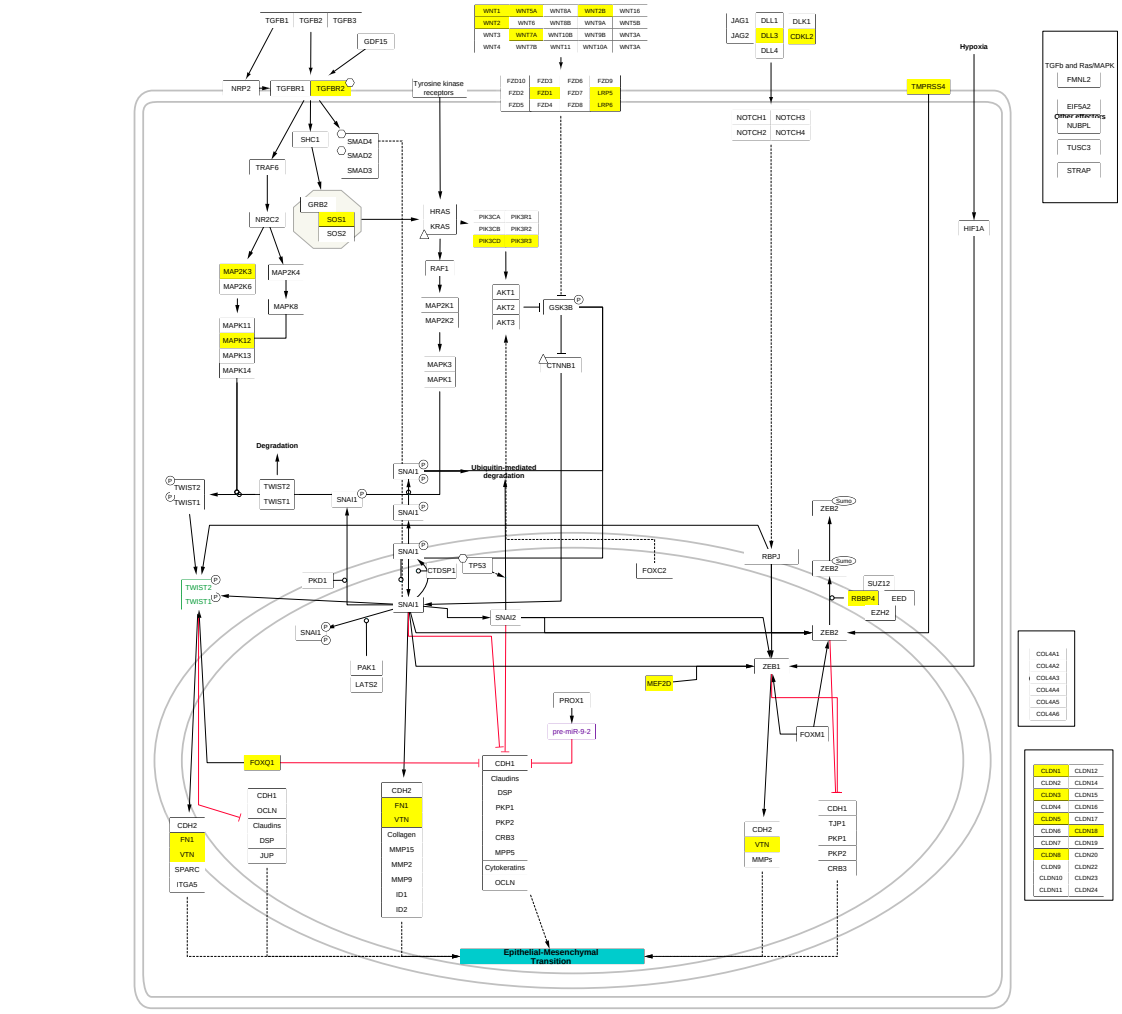


(i) Ninth dimension in central hidden layer.

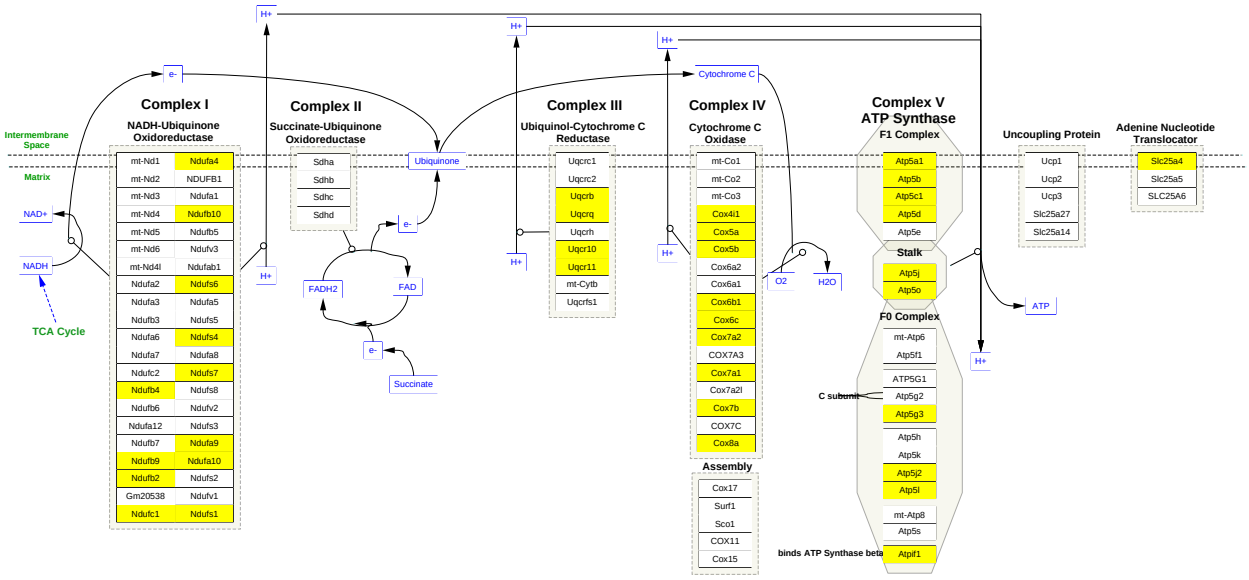


(j) Tenth dimension in central hidden layer.

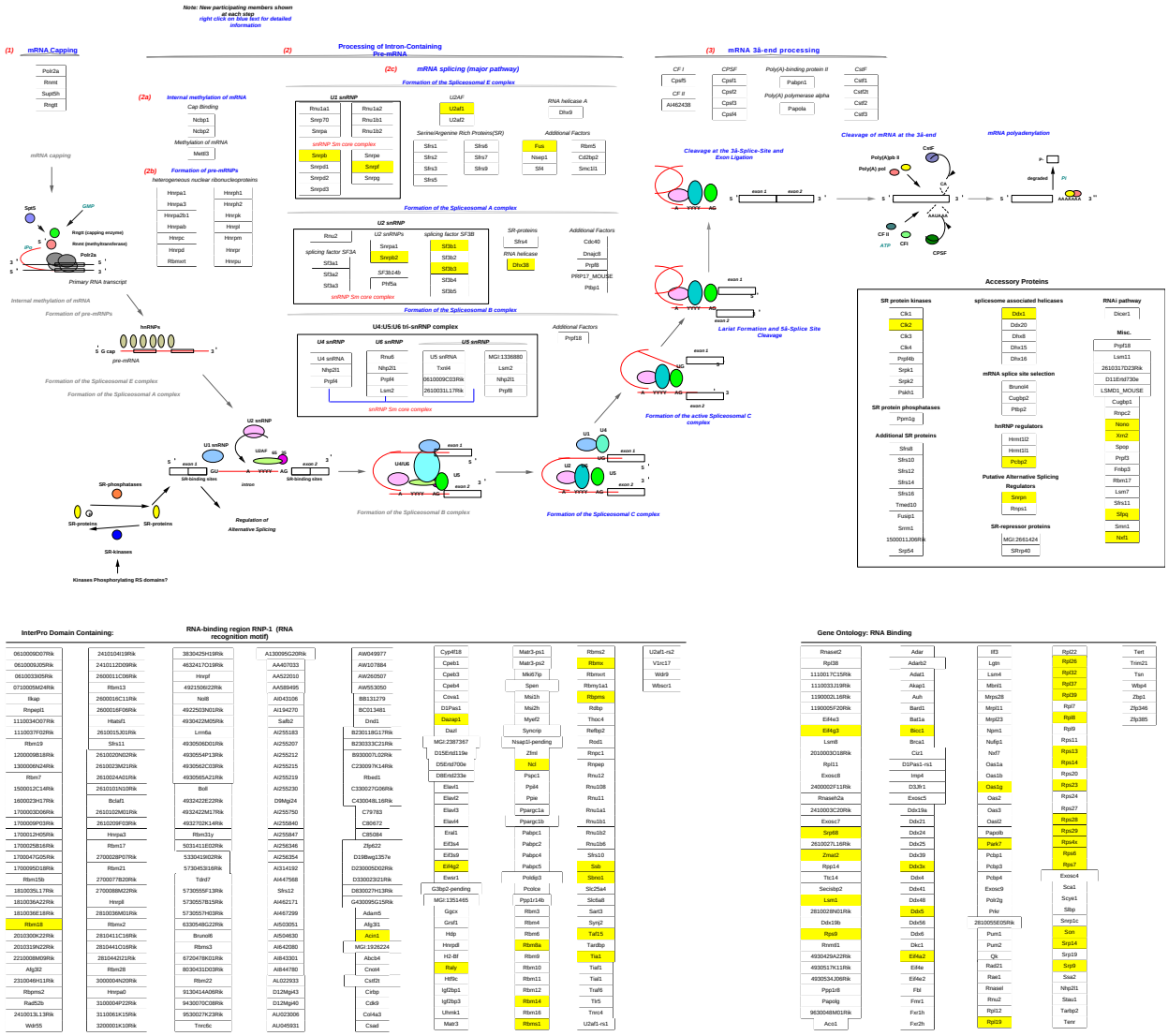
Fig. S3. GO (biological process) enrichment in the central hidden layer of DeepAE on GSE60361. All sub-figures demonstrate the top 30 categories of biological process ontology. The biological processes are ordered by p-values.



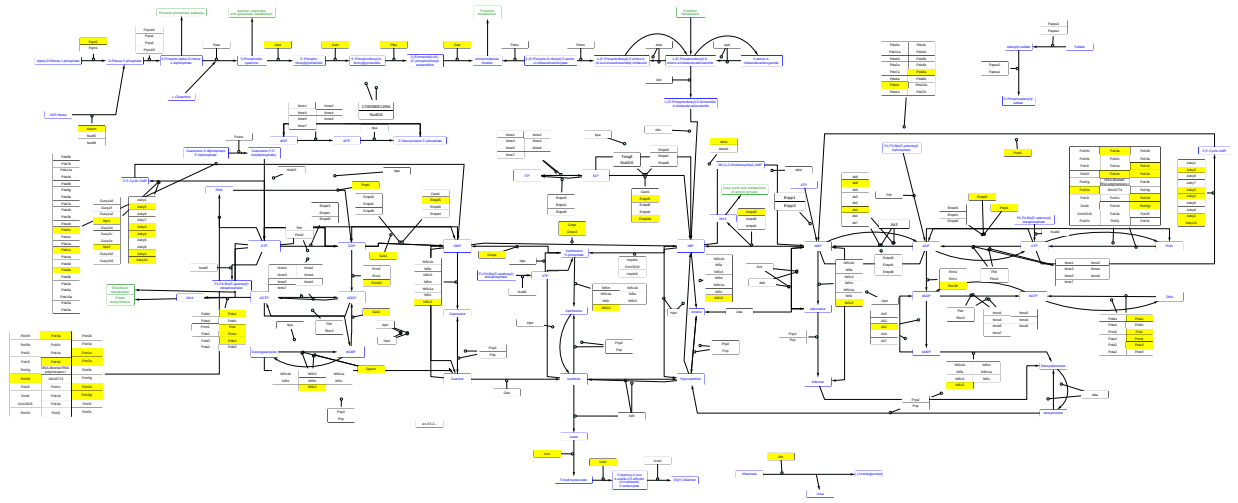
(a) Epithelial to mesenchymal transition in colorectal cancer (GSE69405).



(b) Electron transport chain (GSE60361).



(c) mRNA processing (GSE84133).



(d) Purine metabolism (GSE78779).

Fig. S4. WikiPathways founded from the central hidden layers trained on GSE69405, GSE60361, GSE77564, and GSE78779. In each pathways, the rectangle boxes represent the genes involved in the pathways, while the yellow boxes represent the genes corresponded to the central hidden layer.

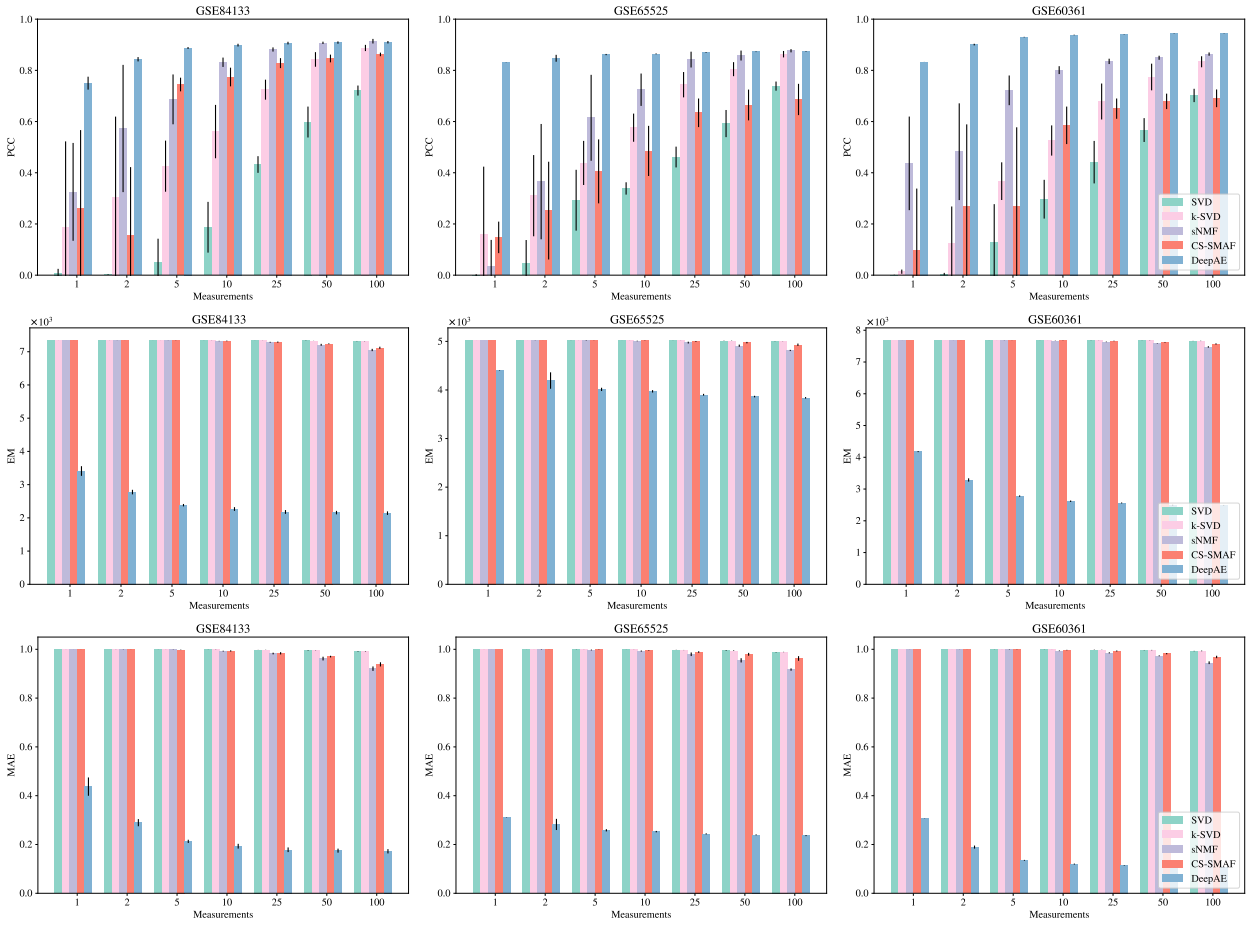


Fig. S5. Performance comparisons among the proposed DeepAE and benchmark methods on three single-cell RNA-seq datasets (GSE60361, GSE65525, and GSE84133) with distinct measurements (1, 2, 5, 10, 25, 50, and 100) evaluated in the PCC, EM, and MAE metrics. Each bar height denotes the mean performance value across multiple runs, and the black line on the top of bar denotes the standard deviation. The benchmark methods include SVD, k-SVD, sNMF, and CS-SMAF.

Table S4. Statistical results in DeepAE calculated with each benchmark method on nine transcriptomic profiling datasets in all metrics by *t* test. The benchmark methods include SVD, k-SVD, sNMF, and CS-SMAF.

Datasets	Metrics	(SVD, DeepAE)	(k-SVD, DeepAE)	(sNMF, DeepAE)	(CS-SMAF, DeepAE)
GSE60361	*PCC*	4.52E-06	2.16E-03	8.65E-04	3.16E-05
	EM	2.24E-09	2.23E-09	2.80E-09	2.63E-09
	MAE	1.46E-11	1.40E-11	2.40E-11	1.79E-11
GSE65525	*PCC*	1.87E-07	6.48E-03	9.61E-04	1.76E-06
	EM	3.06E-17	2.98E-17	6.34E-17	1.56E-16
	MAE	1.61E-18	1.46E-18	1.17E-17	1.01E-16
GSE62270	*PCC*	2.04E-02	5.44E-03	8.92E-04	1.33E-02
	EM	8.78E-10	8.83E-10	9.09E-10	1.01E-09
	MAE	5.43E-11	5.56E-11	5.46E-11	7.10E-11
GSE48968	*PCC*	6.27E-05	3.99E-05	4.91E-06	4.79E-05
	EM	1.72E-13	1.70E-13	3.30E-13	2.16E-13
	MAE	5.22E-15	4.68E-15	4.14E-14	6.64E-15
GSE52529	*PCC*	8.97E-04	2.18E-03	2.55E-03	7.77E-03
	EM	1.15E-06	1.15E-06	1.16E-06	1.17E-06
	MAE	1.07E-07	1.07E-07	1.08E-07	1.09E-07
GSE84133	*PCC*	1.04E-06	1.99E-05	1.77E-04	1.13E-06
	EM	2.50E-13	2.41E-13	4.16E-13	3.32E-13
	MAE	2.53E-16	1.30E-16	5.39E-15	8.63E-16
GSE78779	*PCC*	1.42E-01	1.34E-02	2.33E-02	1.78E-04
	EM	1.04E-11	1.04E-11	1.05E-11	1.11E-11
	MAE	2.31E-14	2.60E-14	2.40E-14	2.90E-14
GSE69405	*PCC*	4.18E-01	8.80E-02	1.98E-02	2.10E-01
	EM	1.44E-09	1.43E-09	1.46E-09	1.47E-09
	MAE	9.89E-11	9.72E-11	1.00E-10	1.00E-10
GSE102475	*PCC*	3.35E-03	2.22E-03	5.73E-05	1.56E-04
	EM	3.83E-12	3.94E-12	4.07E-12	5.16E-12
	MAE	3.35E-13	3.71E-13	2.96E-13	5.51E-13

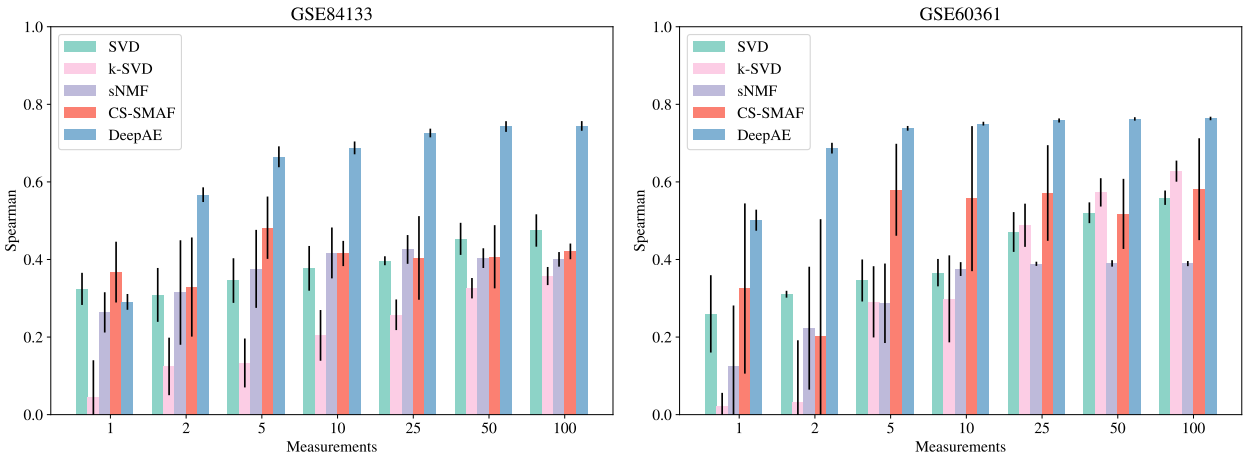


Fig. S6. Performance comparisons among the proposed DeepAE and benchmark methods on two single-cell RNA-seq datasets (GSE84133 and GSE60361) with distinct measurements (1, 2, 5, 10, 25, 50, and 100) evaluated in the Spearman Correlation Coefficient metrics. Each bar height denotes the mean performance value across multiple runs, and the black line on the top of bar denotes the standard deviation. The benchmark methods include SVD, k-SVD, sNMF, and CS-SMAF.

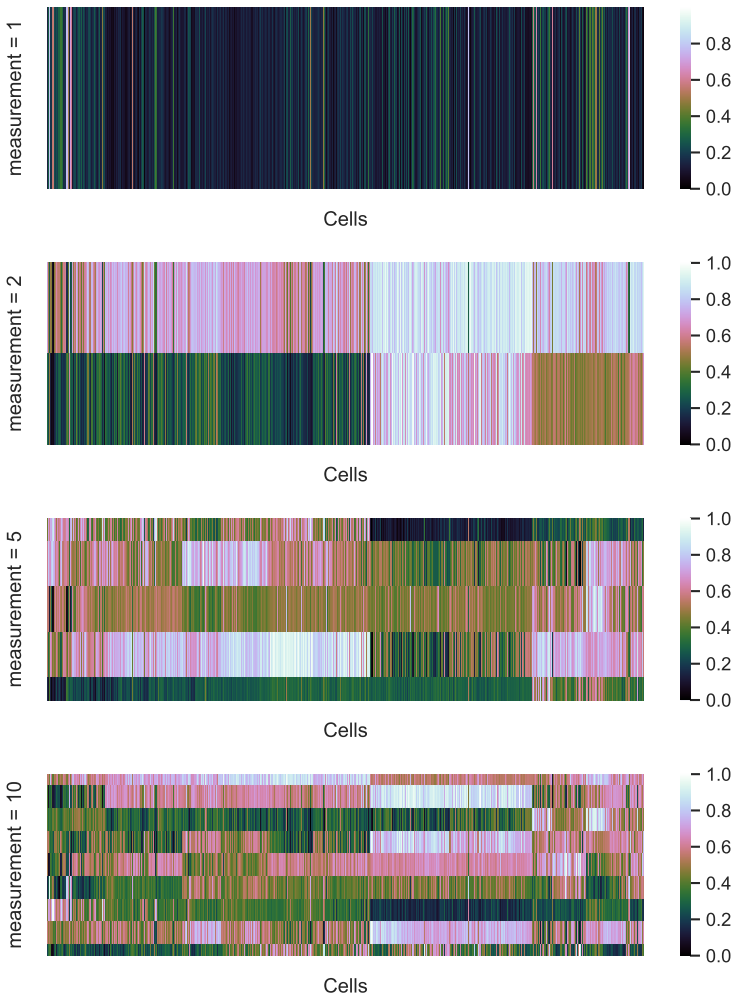


Fig. S7. Heatmaps of the central hidden layers (encoded data) of DeepAE with the measurement = 1, 2, 5, and 10 on GSE60361.