A Experiment Details

A.1 scQuery Dataset

The studies from which the data are collected for our scQuery dataset are shown as below:

Training Set: ERP017366, ERP021445, ERP022096, ERP022251, ERP022289, ERP022298, ERP022654, GSE101487, GSE101984, GSE102159, GSE102332, GSE102346, GSE102455, GSE102456, GSE103268, GSE103892, GSE104156, GSE104396, GSE105054, GSE106447, GSE106472, GSE106663, GSE107115, GSE107122, GSE108097, GSE109774, GSE109796, GSE112033, GSE115070, GSE22182, GSE29087, GSE33979, GSE38198, GSE42268, GSE42704, GSE42706, GSE45719, GSE52564, GSE57403, GSE57609, GSE59114, GSE60066, GSE61346, GSE61844, GSE62952, GSE64959, GSE64960, GSE65160, GSE66343, GSE66390, GSE66582, GSE68981, GSE69761, GSE69970, GSE70713, GSE71982, GSE72852, GSE72854, GSE72855, GSE72856, GSE75454, GSE75659, GSE75804, GSE76381, GSE77113, GSE78140, GSE78471, GSE78521, GSE79306, GSE79374, GSE79380, GSE79578, GSE79812, GSE80155, GSE80168, GSE80280, GSE80483, GSE81275, GSE84498, GSE87375, GSE89468, GSE90697, GSE90822, GSE90824, GSE90860, GSE92707, GSE93524, GSE94333, GSE94389, GSE94579, GSE98048, GSE98664, GSE98816, GSE98969, GSE98971, GSE99235, GSE99701, GSE99786, GSE99866.

Validation Set: ERP013319, ERP022703, GSE100120, GSE102163, GSE103267, GSE107053, GSE107527, GSE112642, GSE113043, GSE44183, GSE57393, GSE59127, GSE65924, GSE68770, GSE71585, GSE71794, GSE71802, GSE74534, GSE78401, GSE78510, GSE79108, GSE85234, GSE85627, GSE93421, GSE94388, GSE96981.

Test Set: ERP022293, GSE102827, GSE106471, GSE107740, GSE107909, GSE108291, GSE108478, GSE32190, GSE39522, GSE39523, GSE57249, GSE57391, GSE59129, GSE60749, GSE65525, GSE65970, GSE67120, GSE68769, GSE75790, GSE75901, GSE77705, GSE78045, GSE82174, GSE84324, GSE86479, GSE87631, GSE89900, GSE90047, GSE96986, GSE99058.

A.2 Training Details

For the scQuery, we find that the optimal value of λ is 0.02 and $m = 17$. For other datasets we use $m = 1$ except for Pancreas 3 to which we apply $m = 3$. For Pancreas 3 we use $\lambda = 0.1$ and Pancreas 1 and 2 we use $\lambda = 0.2$. For all other datasets we use $\lambda = 1$. The models are implemented in PyTorch and trained on a machine with GeForce GTX 980 Ti and 32 GB memory. As for the ScQuery dataset, it takes around 9 and 19 seconds to train NN and scDGN for each epoch respectively.

A.3 Seurat Pancreas Datasets

The cell type compositions in our pancreas datasets are shown as below:

	CelSeq	CelSeq2	SMART-Seq2	Fluidigm
Pancreas 1	gamma, ductal, mast, endothelial, beta, quiescent_stellate, macrophage	gamma, ductal, mast, epsilon, alpha, delta, acinar	macrophage, activated_stellate, schwann, epsilon, alpha, delta, acinar	All
Pancreas 2	macrophage, activated_stellate, schwann, epsilon, alpha, delta, acinar	gamma, ductal, mast, endothelial, beta, quiescent_stellate, macrophage	gamma, ductal, mast, epsilon, alpha, delta, acinar	All
Pancreas 3	gamma, ductal, mast, epsilon, alpha, delta, acinar	macrophage, activated_stellate, schwann, epsilon, alpha, delta, acinar	gamma, ductal, mast, endothelial, beta, quiescent_stellate, macrophage	All
Pancreas 4	gamma, ductal, mast, epsilon, alpha, delta, acinar	gamma, ductal, mast, endothelial, beta, quiescent_stellate, macrophage	macrophage, activated_stellate, schwann, epsilon, alpha, delta, acinar	All
Pancreas 5	macrophage, activated_stellate, schwann, epsilon, alpha, delta, acinar	gamma, ductal, mast, epsilon, alpha, delta, acinar	gamma, ductal, mast, endothelial, beta, quiescent_stellate, macrophage	All
Pancreas 6	gamma, ductal, mast, endothelial, beta, quiescent_stellate, macrophage	macrophage, activated_stellate, schwann, epsilon, alpha, delta, acinar	gamma, ductal, mast, epsilon, alpha, delta, acinar	All

B Per Cell-type Performance

Fig. B.1: Test Accuracy of each model on different cell types from scquery dataset. The darker color represents the better performance. Note that the cell types that are not contained in the test dataset are not shown in this Table. The value is calculated by 10 experiments with different initializations for NN-based model to alleviate the randomness. The average is weighted by the number of the test samples. Note that scDGN performs the best on average.

# train	# test	Cell Type	NN	CaSTLe	MNN	scVI	Seurat	scDGN
946	91	hematopoietic stem cell	0.2516	0.736264	0.042879	0.3275	0	0.3593
14	401	type B pancreatic cell	0.08	0.014963	0	0	/	0.0688
128	11	B cell	0	1	0	0.5	0	0
106	40	blastoderm cell	0.89	0.45	0	0.2275	0	0.715
956	50	neuron	0.514	0	0.076078	0.059	0	0.534
525	45	hematopoietic cell	0.0289	0	0.016993	0.0056	0	0
3	516	embryonic cell	0	0	0	0	/	0
3202	1105	embryonic stem cell	0.3707	0.245249	0.143479	0.1867	0.3705	0.4396
1	49	mesenchymal cell	0	0	0	0	/	0
3598	11	kidney cell	0.1364	0	0	0.5	0	0.2182
13	423	testis	0	0	0	0	/	0
6	6	tissue	0	0	0	0	/	0
152	6	epithelium	0	0	0	0	0	0
542	40	embryo	0.19	0	0.139216	0.3125	0	0.305
6531	251	brain	0.8705	0.729084	0.906588	0.9442	0.1608	0.7988
287	13	pancreas	0.0231	0	0	0	0	0.0308
1825	14	cortex	0.1643	0	0	0.1	0	0.0714
882	8	midbrain	0	0	0	0	0	0
1719	145	lung	0.9779	0.089655	0.953888	0.1145	0.0295	0.9766
256	2	heart left ventricle	0	0	0	0.05	0	0
2876	8	spleen	0.0875	0.25	0	0.0812	0	0.025
1676	529	liver	0.1414	0.035917	0.368917	0.1167	0.0877	0.1989
2888	6	thymus	0.8667	0	0	0.1333	0	0.4333
29132	3770	Average	0.255	0.156	0.2	0.257	0.144	0.286

Fig. B.2: Test Accuracy of each model on different cell types from PBMC dataset.

# train	# test	Cell Type	NN	CaSTLe	MNN	scVI	Seurat	scDGN
4562	288	B	0.9757	0.954861	0.9781	0.975	0.9851	0.9812
4347	602	CD14+monocyte	0.9816	0.988372	0.9822	0.9498	0.9701	0.9867
692	102	CD16+monocyte	0.798	0.745098	0.7961	0.7127	0.7353	0.8294
6749	550	CD4+T	0.8911	0.883636	0.8975	0.7882	0.8833	0.9036
7518	1174	CytotoxicT	0.7886	0.818569	0.7797	0.7485	0.7381	0.794
356	55	Dendritic	0.7509	0.618182	0.7727	0.5909	0.6727	0.7455
186	29	Megakaryocyte	1	1	1	1	1	1
1399	166	Naturalkiller	0.6886	0.674699	0.6771	0.609	0.5982	0.6723
134	26	Plasmacytoiddendritic	0.9615	0.807692	0.9615	0.8462	0.9154	0.9615
25943	2992	Average	0.861	0.865	0.859	0.808	0.83	0.868

Fig. B.3: Test Accuracy of each model on different cell types from pancreas1 dataset.

# train	# test	Cell Type	NN	CaSTLe	MNN	scVI	Seurat	scDGN
7	1	gamma	0	1	0.2	0.9	0	0
771	36	ductal	0.925	0.833333	0.9472	0.95	0.9917	0.9917
8	3	mast	0	0	0.1333	0.5	0.2667	0
18	18	endothelial	0.2778	0.444444	0.4	0.9056	0.9389	0.2778
191	239	beta	0.7054	0.330544	0.8	0.9079	0.8586	0.9042
161	258	quiescent_stellate	0.714	0.794574	0.776	0.8484	0.8279	0.8678
5	1	macrophage	0.9	1	0.7	1	0	0.3
274	21	activated_stellate	0.9952	1	0.9905	0.8524	0.8381	0.9952
4	5	schwann	0	0	0.02	0.36	0	0
145	16	epsilon	1	0.5625	1	0.6562	1	1
330	25	alpha	0.924	0.64	0.896	0.384	0.16	0.864
22	1	delta	1	1	1	1	0.9	0.3
42	14	acinar	0.4929	0.428571	0.4929	0.8643	0.65	0.4571
1978	638	Average	0.72	0.591	0.785	0.855	0.812	0.856

Fig. B.4: Test Accuracy of each model on different cell types from pancreas2 dataset.

# train	# test	Cell Type	NN	CaSTLe	MNN	scVI	Seurat	scDGN
330	25	gamma	0.832	0.56	0.82	0.716	0.724	0.794
462	21	ductal	1	1	1	1	1	1
323	18	mast	0.95	0.833333	0.9556	0.9833	0.8833	0.8833
21	14	endothelial	0.4786	0.428571	0.4786	0.75	0.0286	0.5
6	3	beta	0	0	0.0333	0	0.3333	0.0333
15	1	quiescent_stellate	1	1	1	1	0	0.9
5	1	macrophage	0.6	1	0.6	0.7	0.1	0.4
1	1	activated_stellate	0.5	1	0.4	1	0	0
327	36	schwann	0.7806	0.805556	0.7639	0.8556	0.1667	0.866
3	5	epsilon	0.16	0	0.18	0.34	1	0
1199	239	alpha	0.8695	0.648536	0.8833	0.777	0.9063	0.9381
74	16	delta	0.975	0.625	0.9687	0.9	1	0.9812
469	258	acinar	0.9647	0.910853	0.9725	0.8	0.8725	0.9667
3235	638	Average	0.891	0.764	0.899	0.852	0.825	0.918

Fig. B.5: Test Accuracy of each model on different cell types from pancreas3 dataset.

# train	# test	Cell Type	NN	CaSTLe	MNN	scVI	Seurat	scDGN
7	1	gamma	0	1	0	0	0	0
771	36	ductal	1	0.888889	1	0.9806	0.8639	1
8	3	mast	0.1333	0	0.0667	0.3333	0	0.0667
18	18	endothelial	0	0.777778	0	0.2722	0.2389	0
191	239	beta	0.4895	0.987448	0.5515	0.8025	0.6837	0.7042
161	258	quiescent_stellate	0.4814	0.965116	0.5574	0.3938	0.8605	0.5806
5	1	macrophage	0	0	0	1	0	0
274	21	activated_stellate	0.9952	0.952381	1	0.8429	0.7095	1
4	5	schwann	0.18	0	0.18	0.46	0	0
145	16	epsilon	0.9937	0.3125	1	0.9437	1	1
330	25	alpha	0.94	0.04	0.944	0.824	0.556	0.928
22	1	delta	1	1	1	1	0	1
42	14	acinar	0.5857	0.285714	0.5429	0.7643	0.4429	0.5429
1978	638	Average	0.545	0.722	0.564	0.651	0.751	0.663

Fig. B.6: Test Accuracy of each model on different cell types from pancreas4 dataset.

# train	# test	Cell Type	NN	CaSTLe	MNN	scVI	Seurat	scDGN
253	25	gamma	0.64	0.48	0.636	0.616	0.692	0.664
502	21	ductal	0.9619	1	0.9571	0.9524	1	0.9524
128	18	mast	0.6333	0.666667	0.6333	0.9444	0.9444	0.7111
5	14	endothelial	0.0643	0.214286	0.0643	0.3429	0	0
1	3	beta	0	0	0	0	0	0
1	1	quiescent_stellate	0	0	0	0	0	0
9	1	macrophage	0.5	1	0.4	0.8	0	0
6	1	activated_stellate	0	1	0	0	0	0
444	36	schwann	0.8972	0.916667	0.9028	0.9694	0.4361	0.9139
6	5	epsilon	0.08	0	0.1	0.56	0.4	0.02
1851	239	alpha	0.9954	0.983264	0.9958	0.9837	0.9326	0.9879
145	16	delta	1	0.6875	1	1	1	1
753	258	acinar	0.9938	0.984496	0.9934	0.943	0.9624	0.993
4104	638	Average	0.927	0.914	0.928	0.925	0.881	0.925

Fig. B.7: Test Accuracy of each model on different cell types from pancreas5 dataset.

# train	# test	Cell Type	NN	CaSTLe	MNN	scVI	Seurat	scDGN
18	1	gamma	1	1	1	1	0	0.2
702	36	ductal	0.9222	0.888889	0.9278	0.9583	0.9083	0.9194
13	3	mast	0.2333	0	0.2	0.6667	0.6667	0.0333
213	18	endothelial	0.9389	0.777778	0.9444	0.95	0.95	0.9167
1008	239	beta	0.9854	0.987448	0.9828	0.9414	0.9222	0.9619
308	258	quiescent_stellate	0.9922	0.965116	0.9926	0.7888	0.9349	0.9907
9	1	macrophage	0.3	0	0.1	0.9	0.1	0
228	21	activated_stellate	0.9524	0.952381	0.9524	0.8952	0.281	0.9714
1	5	schwann	0	0	0	0.02	0	0
109	16	epsilon	0.9187	0.3125	0.9062	0.8375	1	0.975
253	25	alpha	0.352	0.04	0.46	0.604	0.368	0.468
16	1	delta	0.5	1	0.7	1	0	0.7
26	14	acinar	0.3429	0.285714	0.3571	0.5429	0.3643	0.3857
2904	638	Average	0.928	0.882	0.932	0.895	0.865	0.923

Fig. B.8: Test Accuracy of each model on different cell types from pancreas6 dataset.

# train	# test	Cell Type	NN	CaSTLe	MNN	scVI	Seurat	scDGN
330	25	gamma	0.768	0.68	0.772	0.74	0.84	0.824
462	21	ductal	1	1	1	0.9714	1	0.9952
323	18	mast	0.9556	0.833333	0.95	0.9889	1	0.9722
21	14	endothelial	0.3643	0.142857	0.3857	0.5286	0.1214	0.3857
6	3	beta	0	0.333333	0	0	0	0.0333
15	1	quiescent_stellate	1	1	1	1	0.2	1
5	1	macrophage	0	0	0.1	0	0	0
1	1	activated_stellate	0	0	0	0	0	0
327	36	schwann	0.8944	0.75	0.9028	0.8889	0.5083	0.9583
3	5	epsilon	0.06	0	0.06	0.36	0.38	0
1199	239	alpha	0.9824	0.979079	0.9812	0.9774	0.9682	0.9791
74	16	delta	1	0.75	1	0.825	0.975	1
469	258	acinar	0.9922	0.988372	0.9942	0.8678	0.9628	0.9922
3235	638	Average	0.944	0.917	0.946	0.893	0.907	0.95

C Representation Visualization

C.1 scQuery

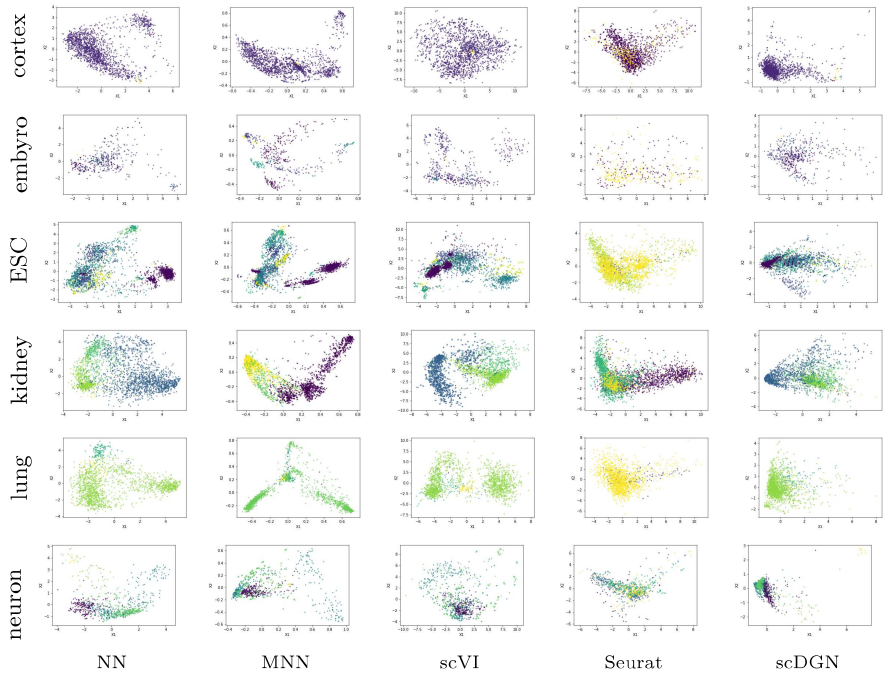


Fig. C.9: PCA visualization of the representations learned by different models on scQuery dataset for certain cell types. The colors are used to distinguish the batches.

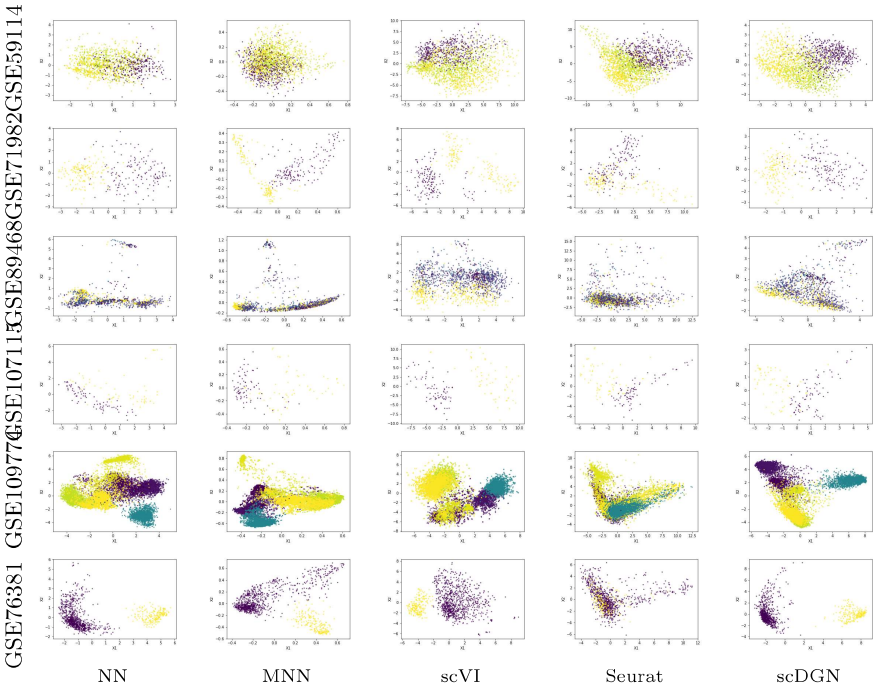
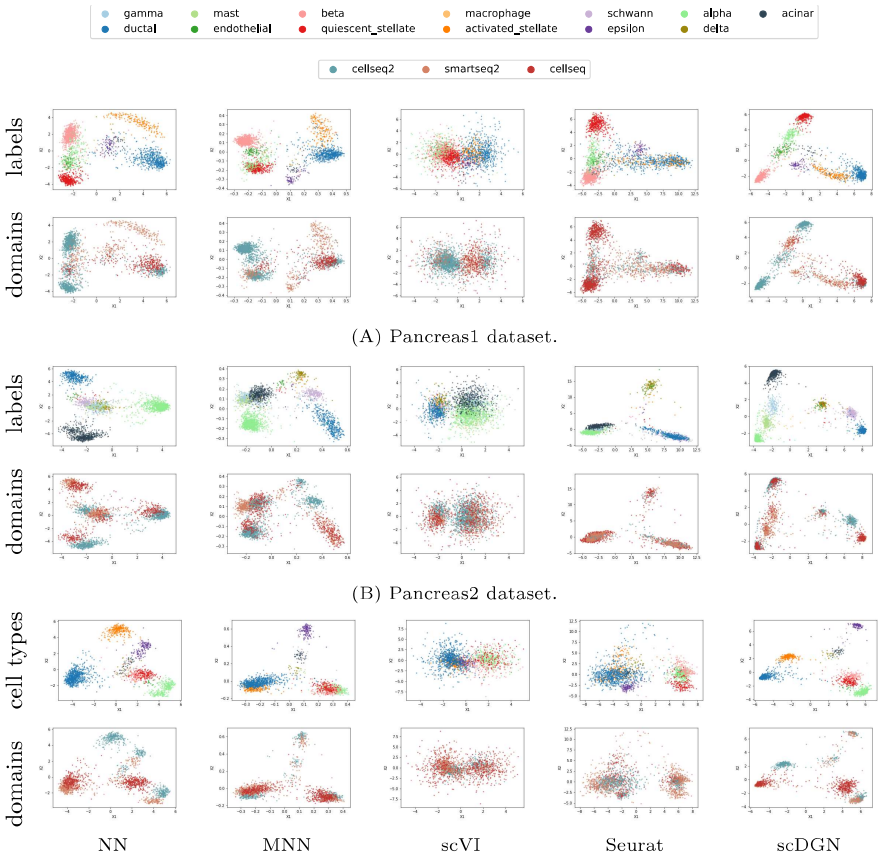


Fig. C.10: PCA visualization of the representations learned by different models on scQuery dataset for certain batches. The colors are used to distinguish the cell types.

C.2 Full Pancreas Datasets

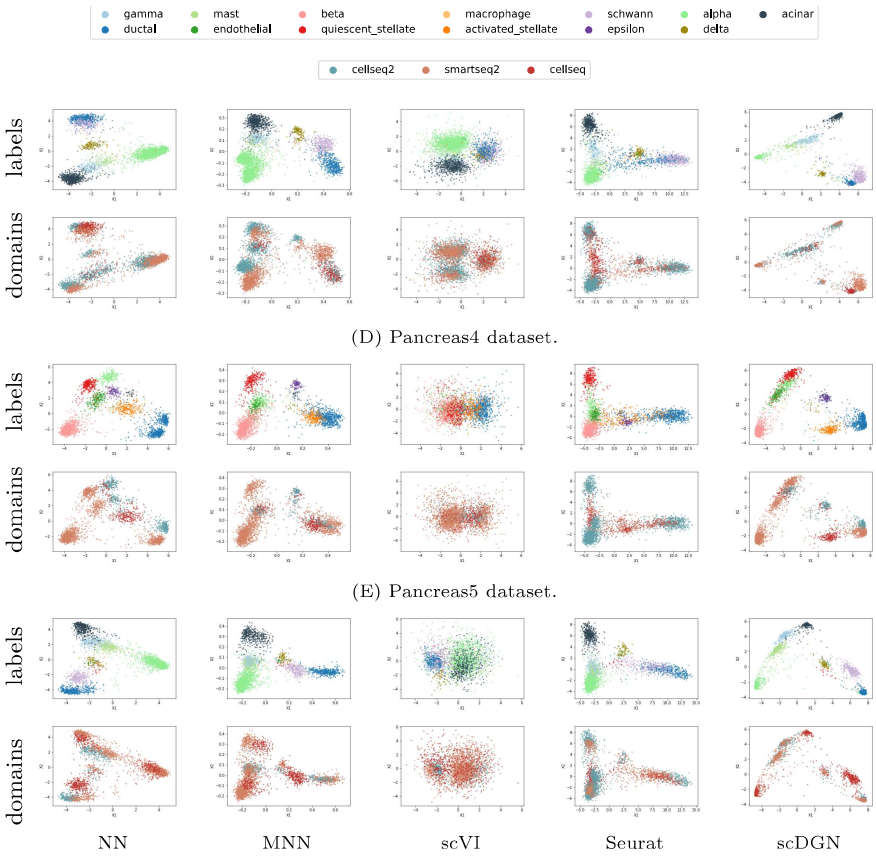


(A) Pancreas1 dataset.

(B) Pancreas2 dataset.

(C) Pancreas3 dataset.

Fig. C.11: PCA visualization of the representations learned by different models on the whole Pancreas datasets.



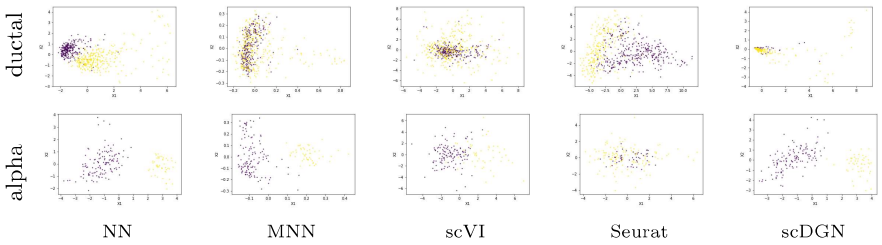
(D) Pancreas4 dataset.

(E) Pancreas5 dataset.

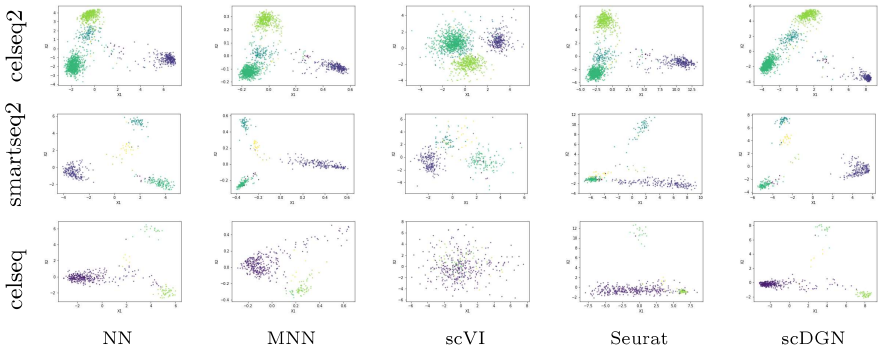
(F) Pancreas6 dataset.

Fig. C.12: PCA visualization of the representations learned by different models on the Pancreas datasets.

C.3 Pancreas1



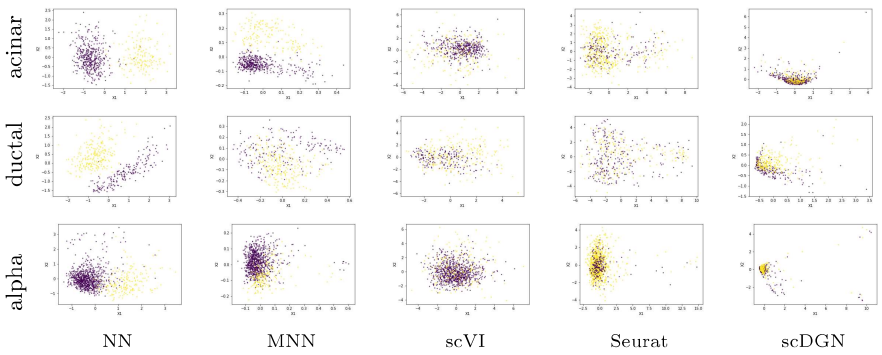
(A) Visualization of representation for certain cell types. The colors are used to distinguish the batches.



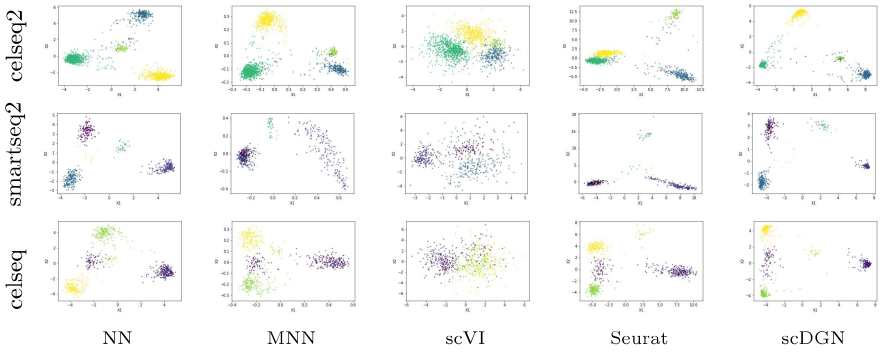
(B) Visualization of representation for certain batches. The colors are used to distinguish the cell types.

Fig. C.13: PCA visualization of the representations for certain cell types and domains on Pancreas1 dataset.

C.4 Pancreas2



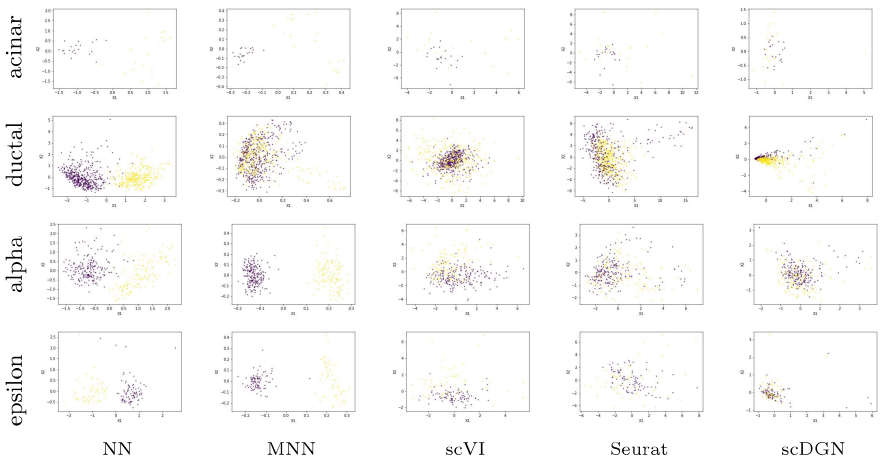
(A) Visualization of representation for certain cell types. The colors are used to distinguish the batches.



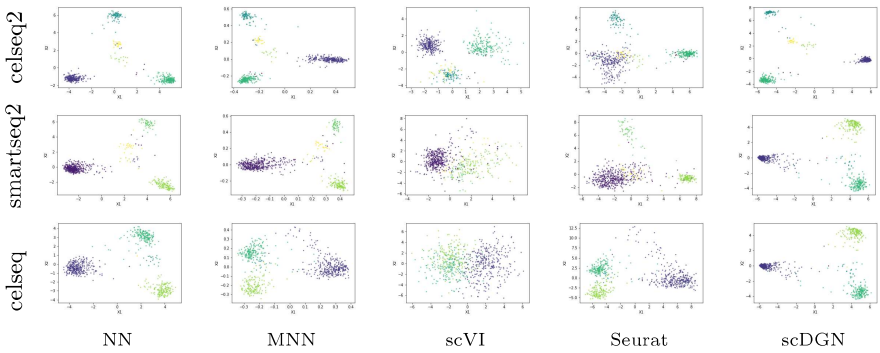
(B) Visualization of representation for certain batches. The colors are used to distinguish the cell types.

Fig. C.14: PCA visualization of the representations for certain cell types and domains on Pancreas2 dataset.

C.5 Pancreas3



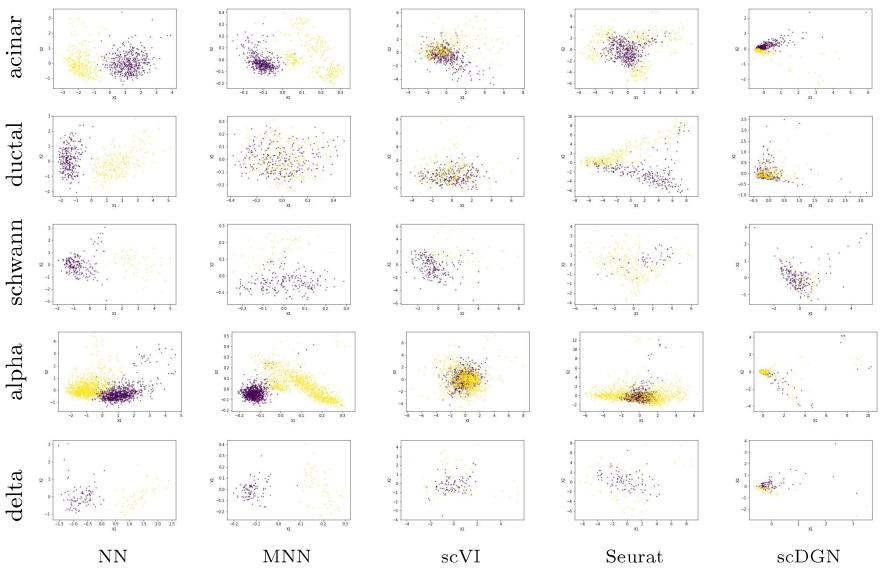
(A) Visualization of representation for certain cell types. The colors are used to distinguish the batches.



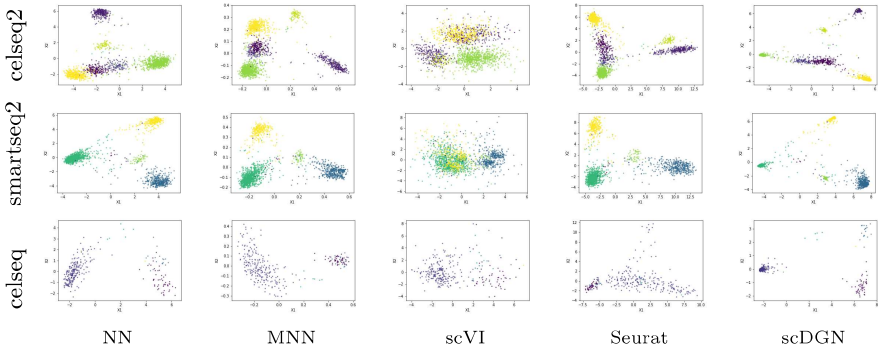
(B) Visualization of representation for certain batches. The colors are used to distinguish the cell types.

Fig. C.15: PCA visualization of the representations for certain cell types and domains on Pancreas3 dataset.

C.6 Pancreas4



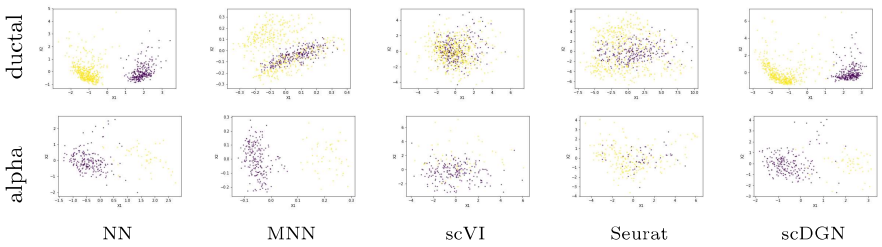
(A) Visualization of representation for certain cell types. The colors are used to distinguish the batches.



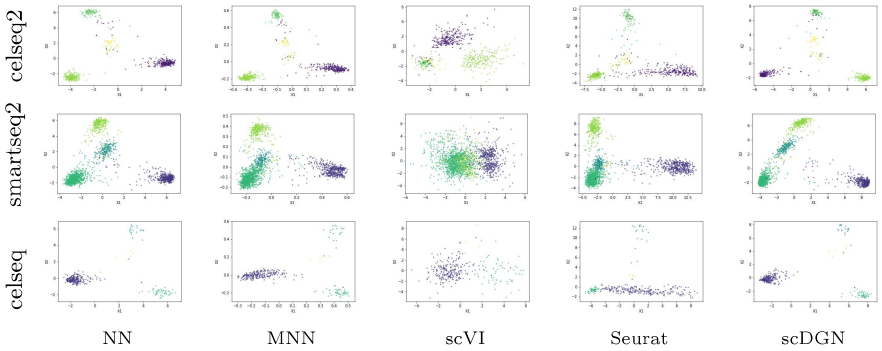
(B) Visualization of representation for certain batches. The colors are used to distinguish the cell types.

Fig. C.16: PCA visualization of the representations for certain cell types and domains on Pancreas4 dataset.

C.7 Pancreas5



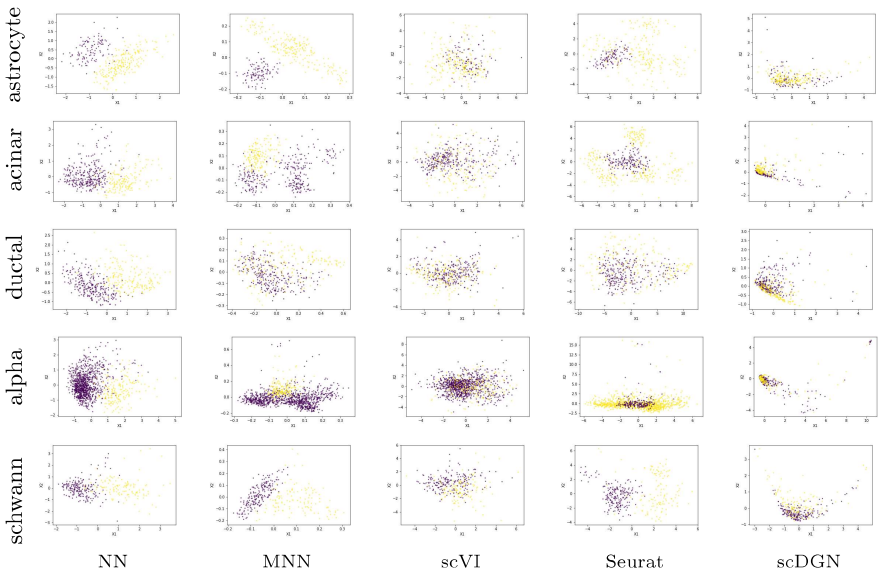
(A) Visualization of representation for certain cell types. The colors are used to distinguish the batches.



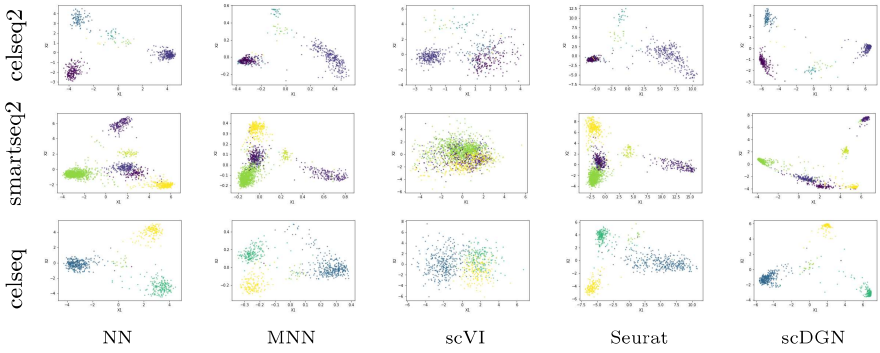
(B) Visualization of representation for certain batches. The colors are used to distinguish the cell types.

Fig. C.17: PCA visualization of the representations for certain cell types and domains on Pancreas5 dataset.

C.8 Pancreas6



(A) Visualization of representation for certain cell types. The colors are used to distinguish the batches.



(B) Visualization of representation for certain batches. The colors are used to distinguish the cell types.

Fig. C.18: PCA visualization of the representations for certain cell types and domains on Pancreas6 dataset.

D Key gene analysis

D.1 Top 100 Genes for Liver Cell Type[2]

- **NN:** *ubb*, *ins2*, *apoe*, *hba-a1*, *actb*, *b2m*, *hba-a2*, *mup11*, *cd74*, *dnase1l3*, *h2-d1*, *mup22*, *tuba1a*, *tmsb4x*, *mup14*, *selenop*, *aw112010*, *ins1*, *cst3*, *mup7*, *iapp*, *ccdc152*, *ftl1*, *mup13*, *mup10*, *chl1*, *hbb-bt*, *clec4g*, *igfbp7*, *nrep*, *mup19*, *sh3bgrl3*, *hbb-bs*, *mup1*, *mup18*, *h2-ab1*, *mup16*, *mup12*, *rpl18a*, *c1qb*, *map1lc3b*, *mup15*, *ccl5*, *hbb-y*, *rps29*, *ctsd*, *ly86*, *ifitm2*, *mup9*, *h3f3b*, *mup8*, *nkg7*, *tmsb10*, *mt1*, *tagln2*, *ttr*, *gstm1*, *hba-x*, *atpif1*, *lyz2*, *apoc1*, *lgals1*, *mup2*, *stmn2*, *apoa2*, *plp1*, *calm1*, *gpx3*, *prdx1*, *igfbp4*, *prl3d1*, *c1qa*, *apod*, *tma7*, *h2-aa*, *cmss1*, *ptma*, *cd79a*, *gapdh*, *plpp1*, *sst*, *abi1*, *dennd1b*, *cfl1*, *rgs1*, *gclm*, *pitpnc1*, *dppa3*, *fcgr2b*, *serp1*, *rpl8*, *tyrobp*, *sumo1*, *zfp976*, *rplp1*, *gm10591*, *gm21541*, *rpsa*, *s100a8*, *pcnp*.
- **scDGN:** *rps29*, *apoe*, *mup11*, *ins2*, *mup22*, *dnase1l3*, *selenop*, *mup10*, *b2m*, *gcg*, *mup7*, *h2-d1*, *mup14*, *tmsb10*, *ccdc152*, *clec4g*, *apoa2*, *hba-a1*, *ttr*, *mup13*, *myl7*, *mgp*, *mup18*, *mup16*, *mt1*, *mup15*, *npm1*, *mup19*, *hba-a2*, *aw112010*, *igfbp7*, *chl1*, *mup1*, *actg1*, *apoc1*, *ins1*, *ly6e*, *mup9*, *mup8*, *ptma*, *mup12*, *mup2*, *cst3*, *fabp3*, *btg1*, *iapp*, *rpl35*, *apoa1*, *h2-k1*, *cmss1*, *mup17*, *lgals1*, *tmsb4x*, *stmn1*, *gm13304*, *ptp4a2*, *prdx1*, *gm21541*, *hmga1*, *snap25*, *set*, *plp1*, *ccl4*, *fabp4*, *trim30a*, *gstp1*, *gm10591*, *ubc*, *scgb1a1*, *resp18*, *sumo1*, *fabp1*, *nrep*, *h2-q7*, *npy*, *itm2b*, *hspe1*, *car2*, *sub1*, *slc25a5*, *h2afz*, *ywhah*, *ccl21b*, *pomc*, *rpl41*, *cbx3*, *ctsd*, *rps27rt*, *laptm5*, *chchd2*, *s100a8*, *actc1*, *hba-x*, *hbb-bt*, *myl4*, *eif1*, *gpihbp1*, *sod1*, *gabarapl2*, *calm1*.

[2] The results in text format are also available at: *https://github.com/SongweiGe/scDGN/blob/master/supplementary materials/top_genes.txt*

D.2 GO Analysis Results for the Other Cell Types

Table 5: GO analysis results of the genes with respect to embryonic stem cell that are only recognized by scDGN.

term_name	term_id	p_{adj}	$-\log_{10} p_{adj}$
SRP-dependent cotranslational protein targeting to membrane	GO:0006614	1.8109E-06	5.74210612
cotranslational protein targeting to membrane	GO:0006613	2.59394E-06	5.58603936
protein targeting to ER	GO:0045047	4.44101E-06	5.35251780
establishment of protein localization to endoplasmic reticulum	GO:0072599	5.72221E-06	5.24243629
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	GO:0000184	9.79366E-06	5.00905515
protein localization to endoplasmic reticulum	GO:0070972	2.19947E-05	4.65768238

Table 6: GO analysis results of the genes with respect to hematopoietic stem cell that are only recognized by scDGN.

term_name	term_id	p_{adj}	$-\log_{10} p_{adj}$
protein targeting to ER	GO:0045047	0.041444862	1.382529302
cotranslational protein targeting to membrane	GO:0006613	0.026817903	1.571575182
SRP-dependent cotranslational protein targeting to membrane	GO:0006614	0.020030591	1.698306246
protein localization to endoplasmic reticulum	GO:0070972	0.013637519	1.865264616
translational initiation	GO:0006413	0.001555052	2.808254977