iScience, Volume *26*

Supplemental information

Generative pretraining from large-scale transcriptomes for single-cell deciphering

Hongru Shen, Jilei Liu, Jiani Hu, Xilin Shen, Chao Zhang, Dan Wu, Mengyao Feng, Meng Yang, Yang Li, Yichen Yang, Wei Wang, Qiang Zhang, Jilong Yang, Kexin Chen, and Xiangchun Li

Supplementary Materials

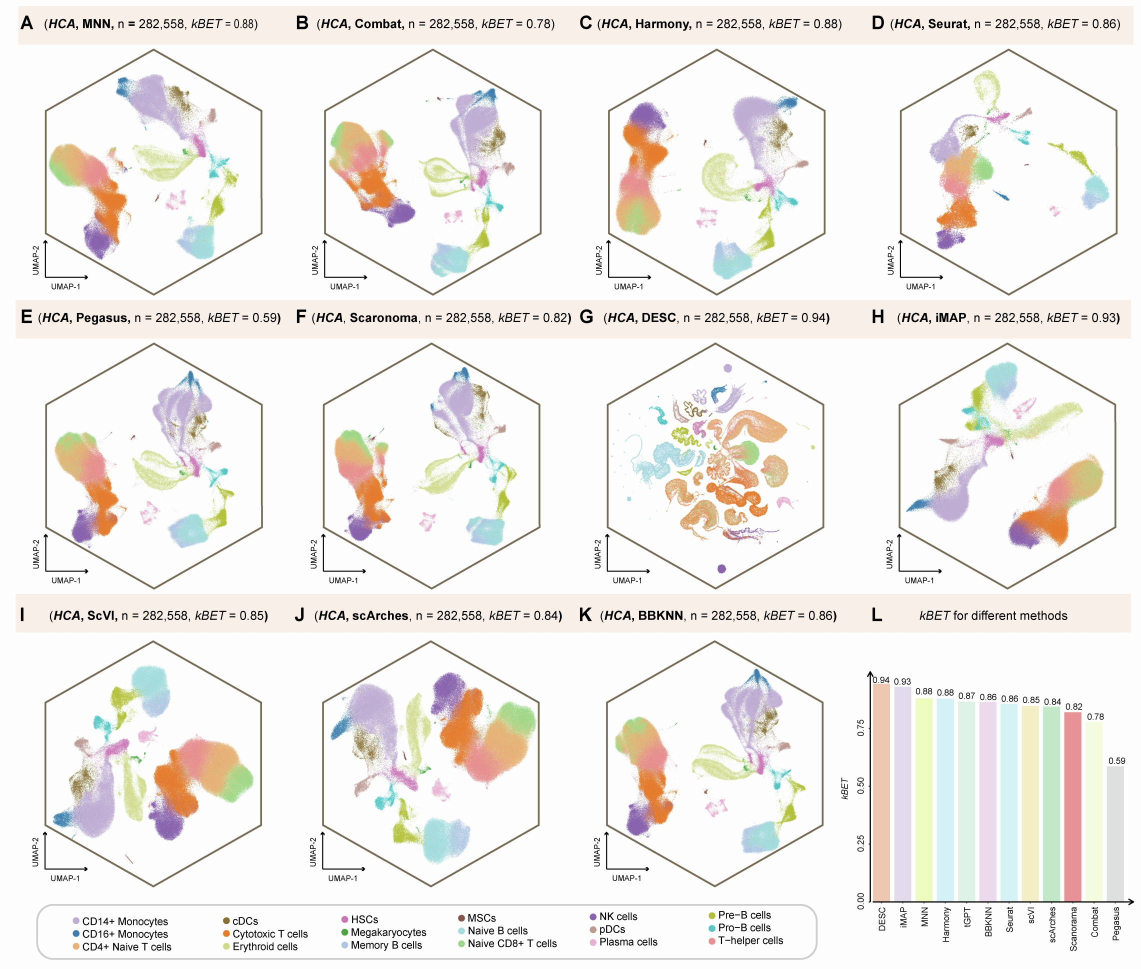

Figure S1. The UMAP visualization plots of different batch-correction methods on the HCA dataset (A to K) and *kBET* acceptance rate (L), related to Figure 2.

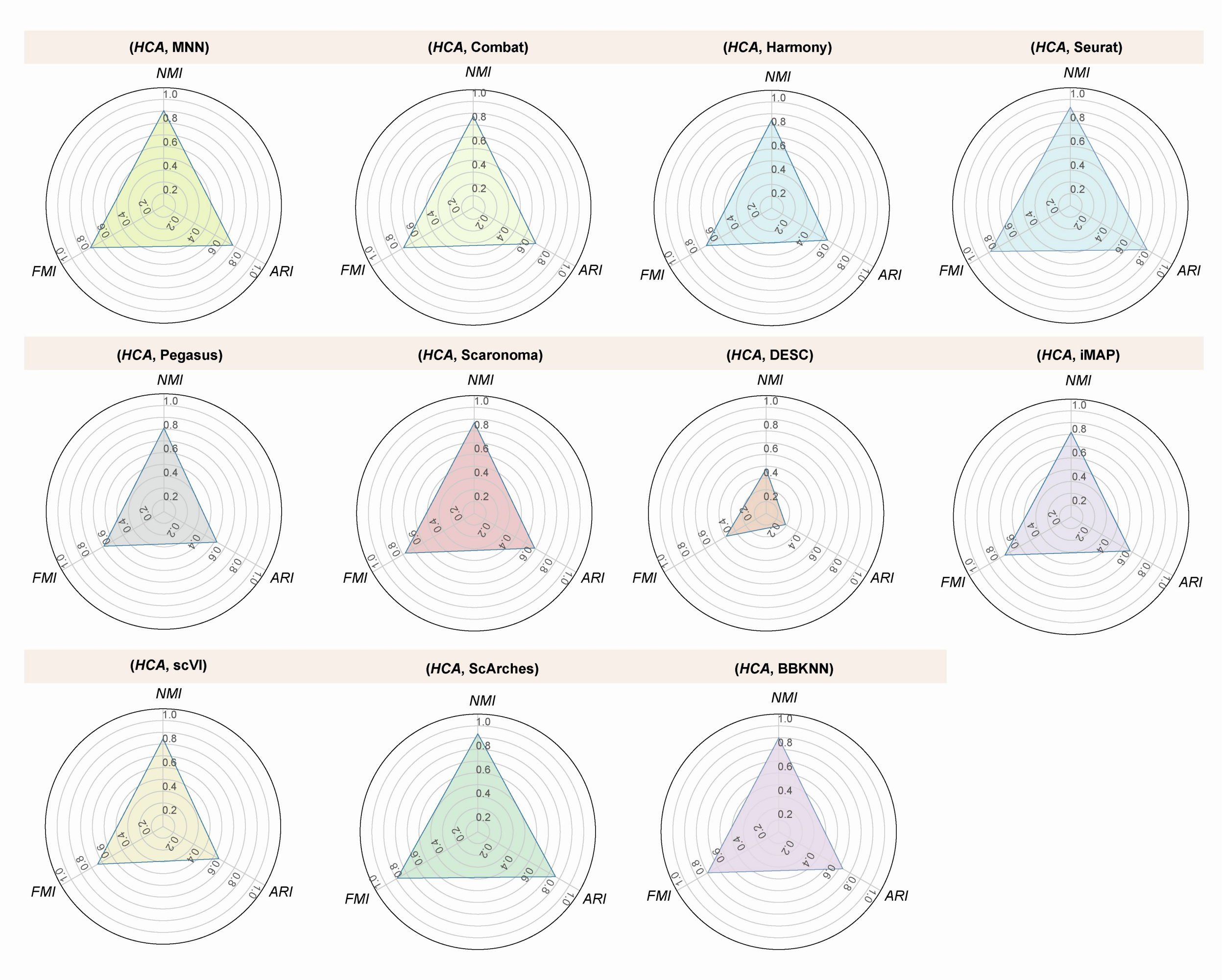

Figure S 2 . Radar charts illustrating the best clustering metrics for different batch-correction methods obtained from grid search on the *HCA* dataset, related to Figure 2. *ARI*, Adjusted Rand Index; *NMI*, Normalized Mutual information; *FMI*, Fowlkes-Mallows Index.

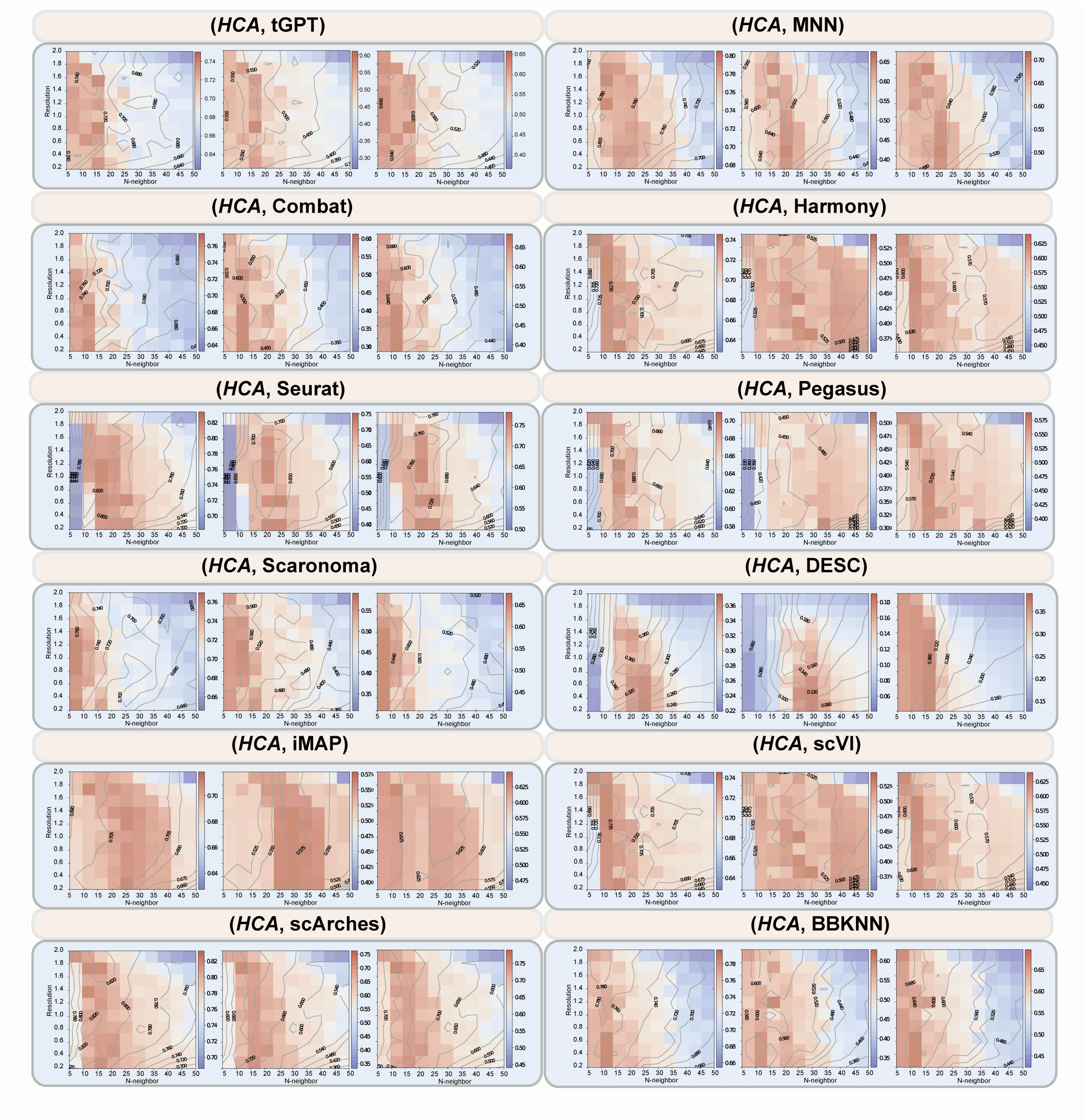

Figure S3. The clustering performance with grid search for resolution and number of neighbors for different batch-correction methods on the *HCA* dataset, related to Figure 2. Contour maps depict different cluster metrics (i.e. *NMI*, *ARI* and *FMI*) with respect to different values of *Resolution* and *N-neighbors*.

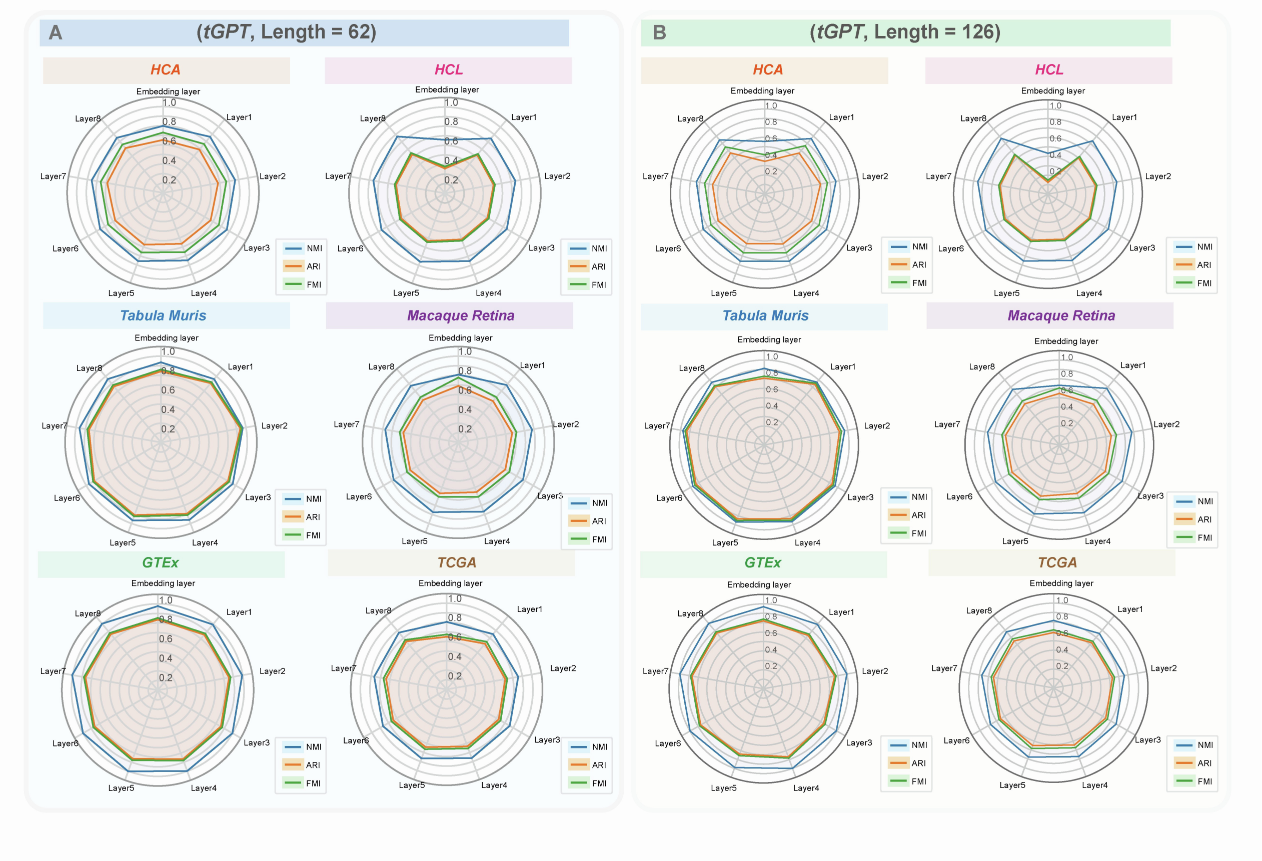

Figure S4. Radar charts illustrating the clustering performance achieved by feature representations extracted from different layers of *tGPT* for the top 62 (A) and 126 (B) expressing genes on *HCA*, *HCL*, *Tabula Muris*, *Macaque Retina*, *GTEx*, *and TCGA* datasets, related to Figure 2.

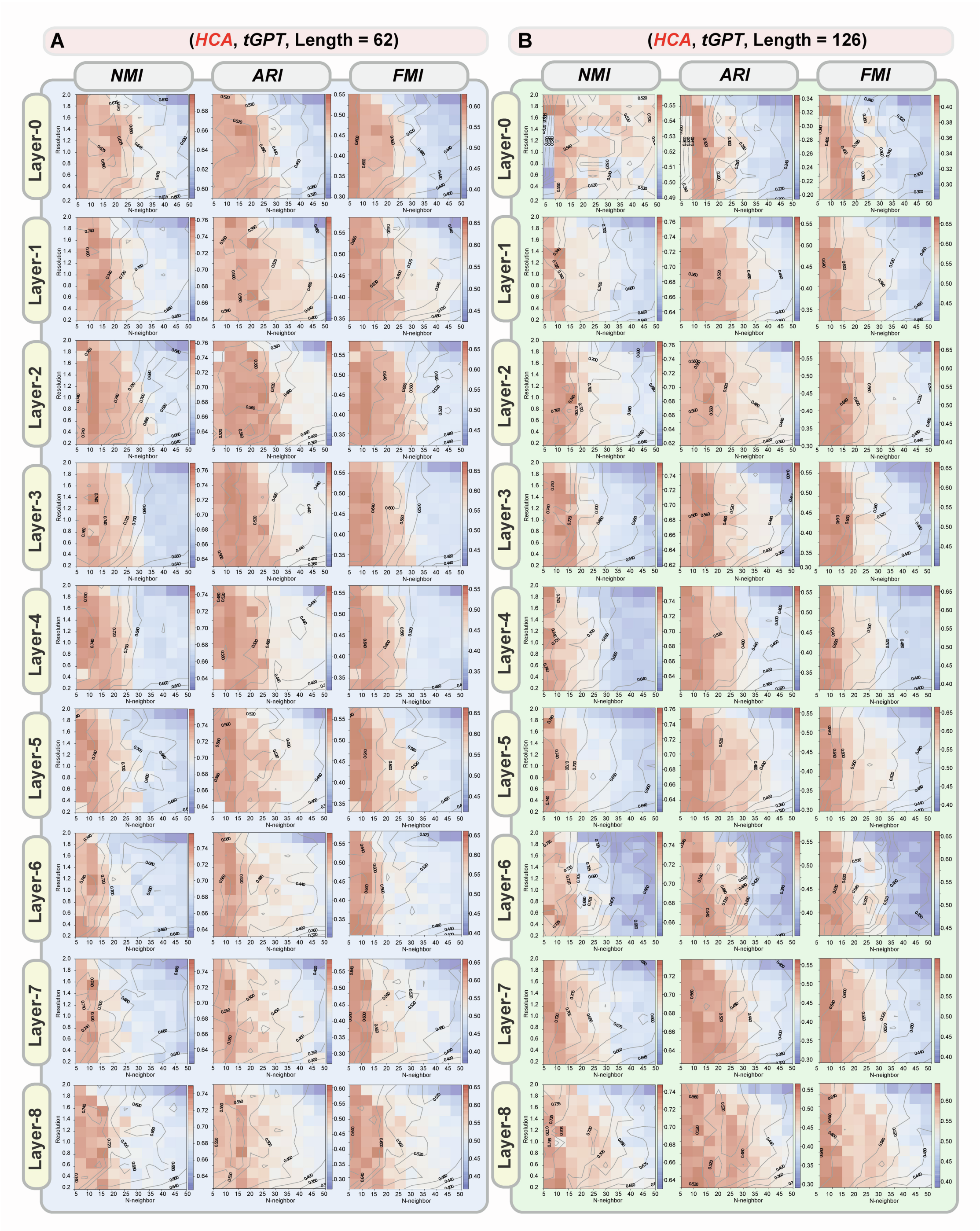

Figure S5. The clustering performance with grid search for resolution and number of neighbors for the top 62 (A) and 126 (B) expressing genes among feature representations extracted from different layers on the *HCA* dataset, related to Figure 2. Contour maps depict different cluster metrics (i.e. *NMI*, *ARI* and *FMI*) with respect to different values of *Resolution* and *N-neighbors*.

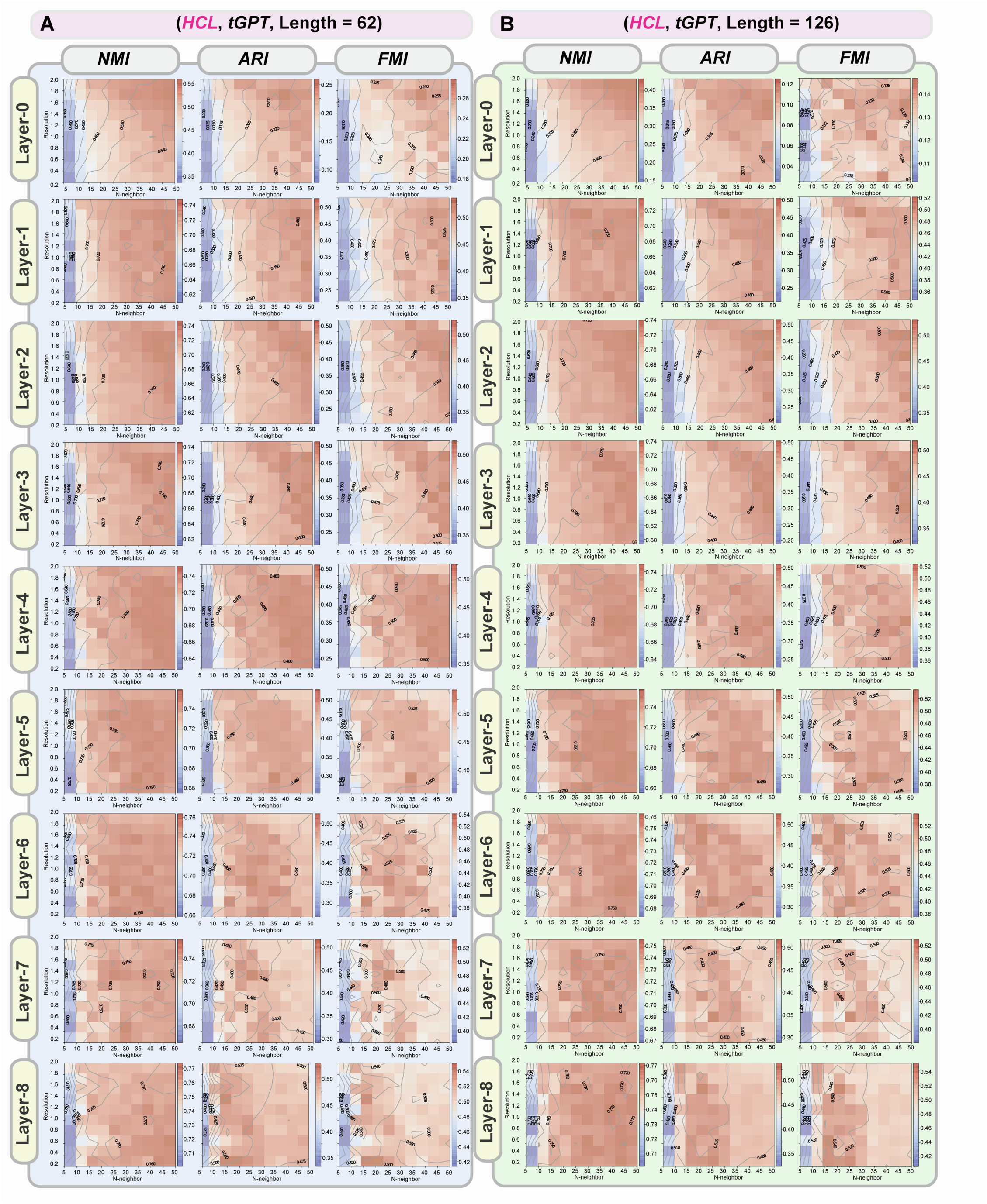

Figure S6. The clustering performance with grid search for resolution and number of neighbors for the top 62 (A) and 126 (B) expressing genes among feature representations extracted from different layers on the *HCL* dataset, related to Figure 2. Contour maps depict different cluster metrics (i.e. *NMI*, *ARI* and *FMI*) with respect to different values of *Resolution* and *N-neighbors*.

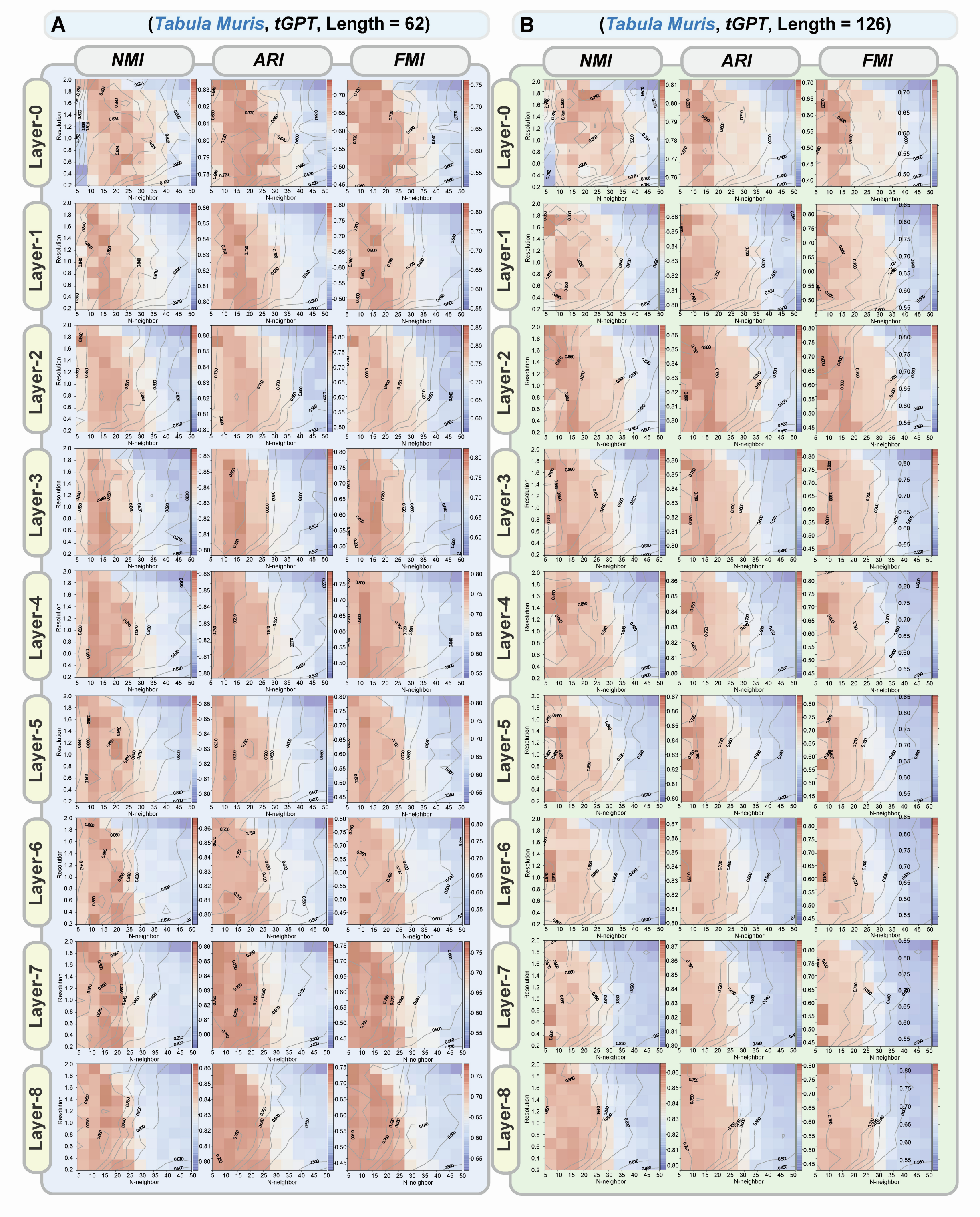

Figure S7. The clustering performance with grid search for resolution and number of neighbors for the top 62 (A) and 126 (B) expressing genes among feature representations extracted from different layers on the *Tabula Muris* dataset, related to Figure 2. Contour maps depict different cluster metrics (i.e. *NMI*, *ARI* and *FMI*) with respect to different values of *Resolution* and *N-neighbors*.

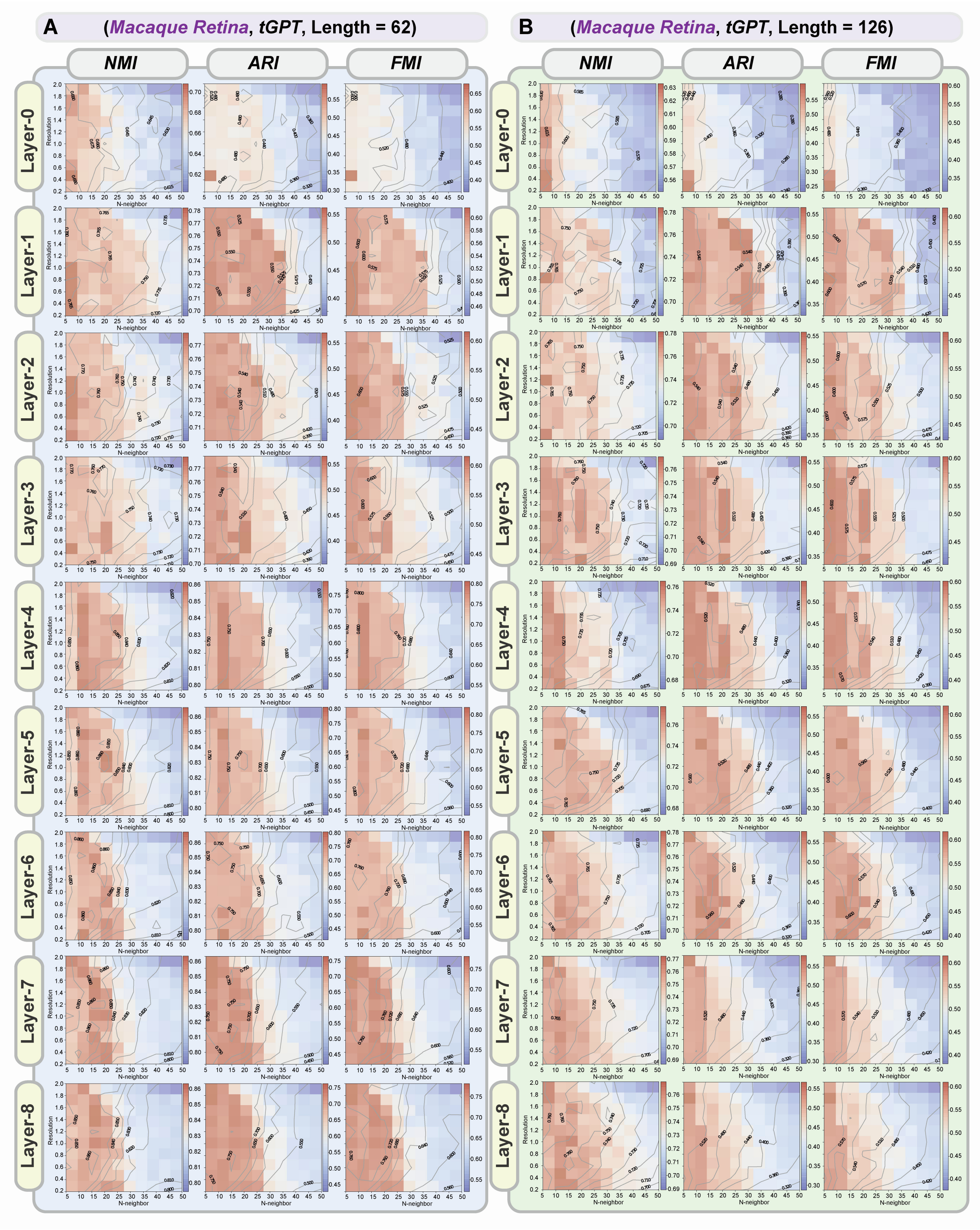

Figure S8. The clustering performance with grid search for resolution and number of neighbors for the top 62 (A) and 126 (B) expressing genes among feature representations extracted from different layers on the *Macaque Retina* dataset, related to Figure 2. Contour maps depict different cluster metrics (i.e. *NMI*, *ARI* and *FMI*) with respect to different values of *Resolution* and *N-neighbors*.

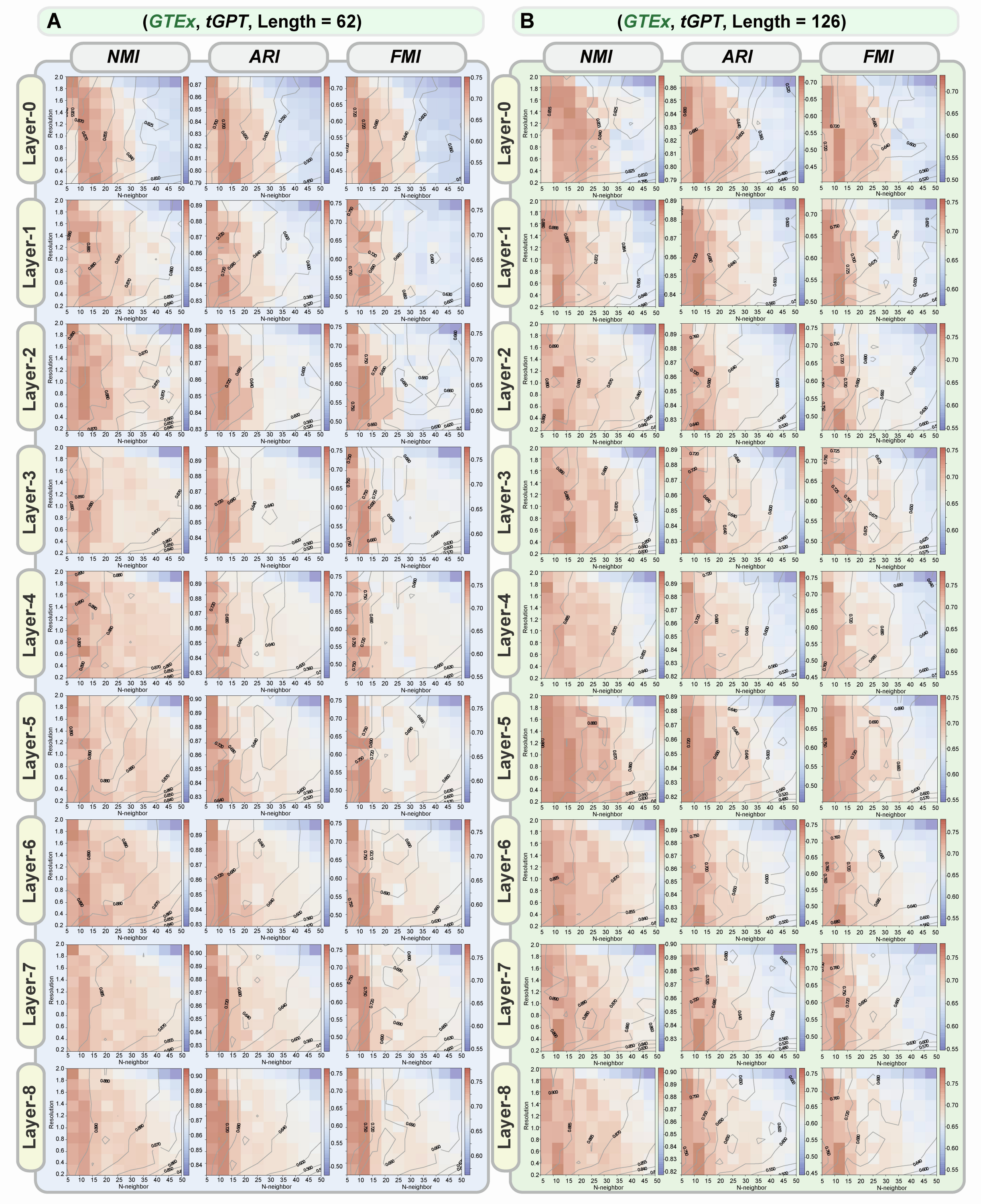

Figure S9. The clustering performance with grid search for resolution and number of neighbors for the top 62 (A) and 126 (B) expressing genes among feature representations extracted from different layers on the *GTEx* dataset, related to Figure 2. Contour maps depict different cluster metrics (i.e. *NMI*, *ARI* and *FMI*) with respect to different values of *Resolution* and *N-neighbors*.

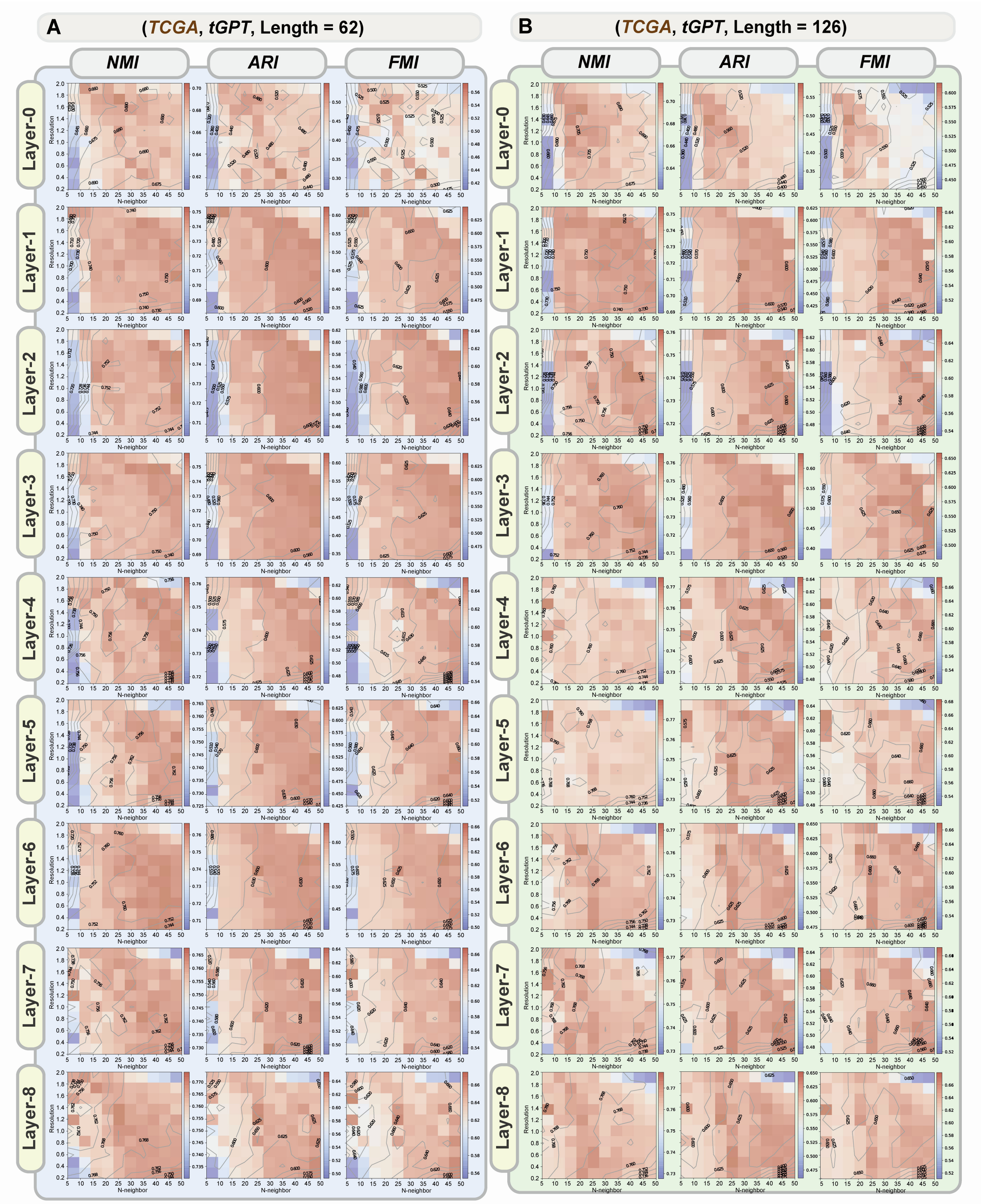

Figure S10. The clustering performance with grid search for resolution and number of neighbors for the top 62 (A) and 126 (B) expressing genes among feature representations extracted from different layers on the *TCGA* dataset, related to Figure 2. Contour maps depict different cluster metrics (i.e. *NMI*, *ARI* and *FMI*) with respect to different values of *Resolution* and *N-neighbors*.

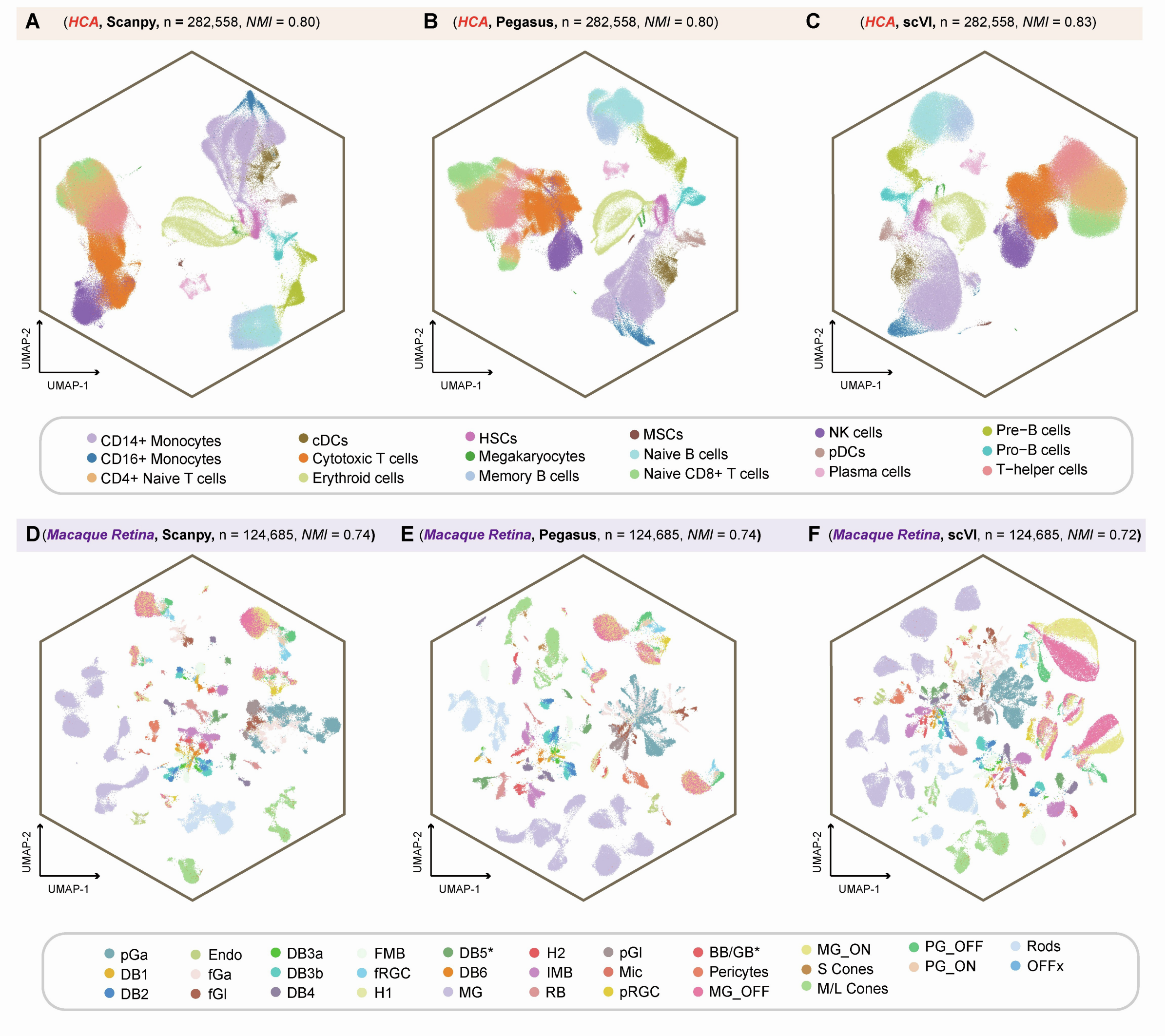

Figure S11. UMAP visualization on different datasets obtained from different methods, related to Figure 2.

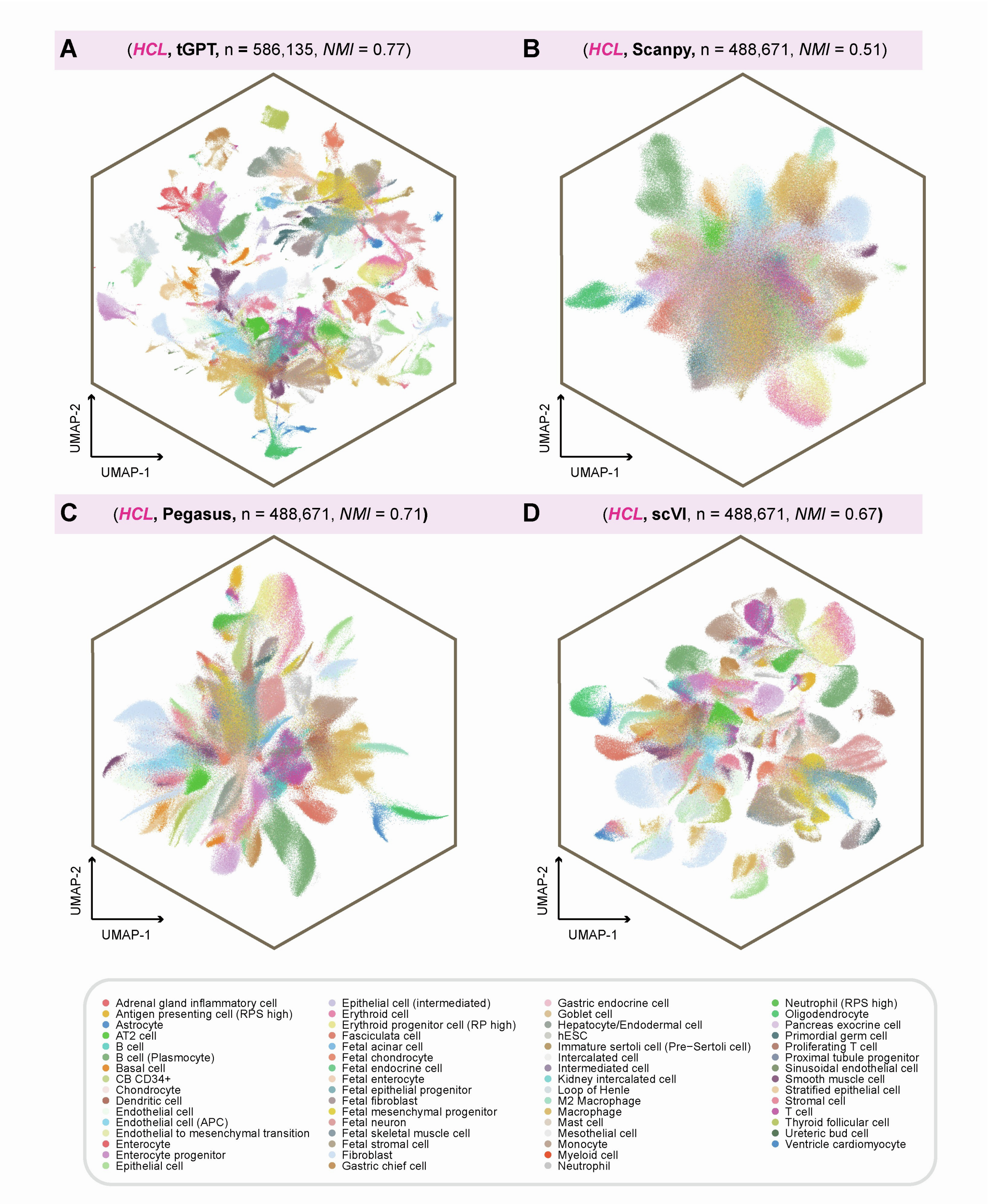

Figure S12. The full annotation of UMAP visualization of different methods on the *HCL* dataset, related to Figure 2. The *NMI* metric and annotation of cells are shown.

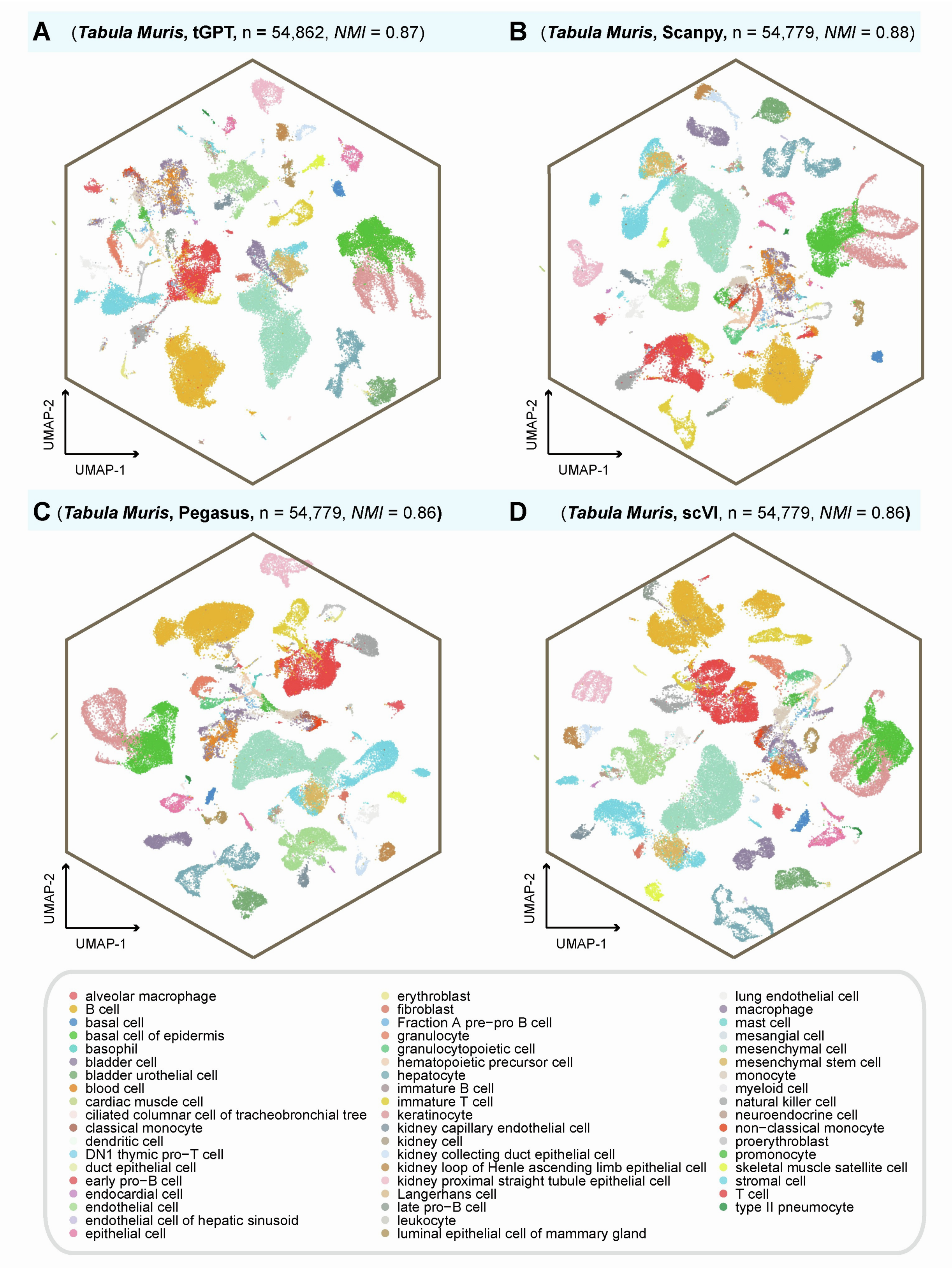

Figure S13. The full annotation of UMAP visualization of different methods on the *Tabula Muris* dataset, related to Figure 2. The *NMI* metric and annotation of cells are shown.

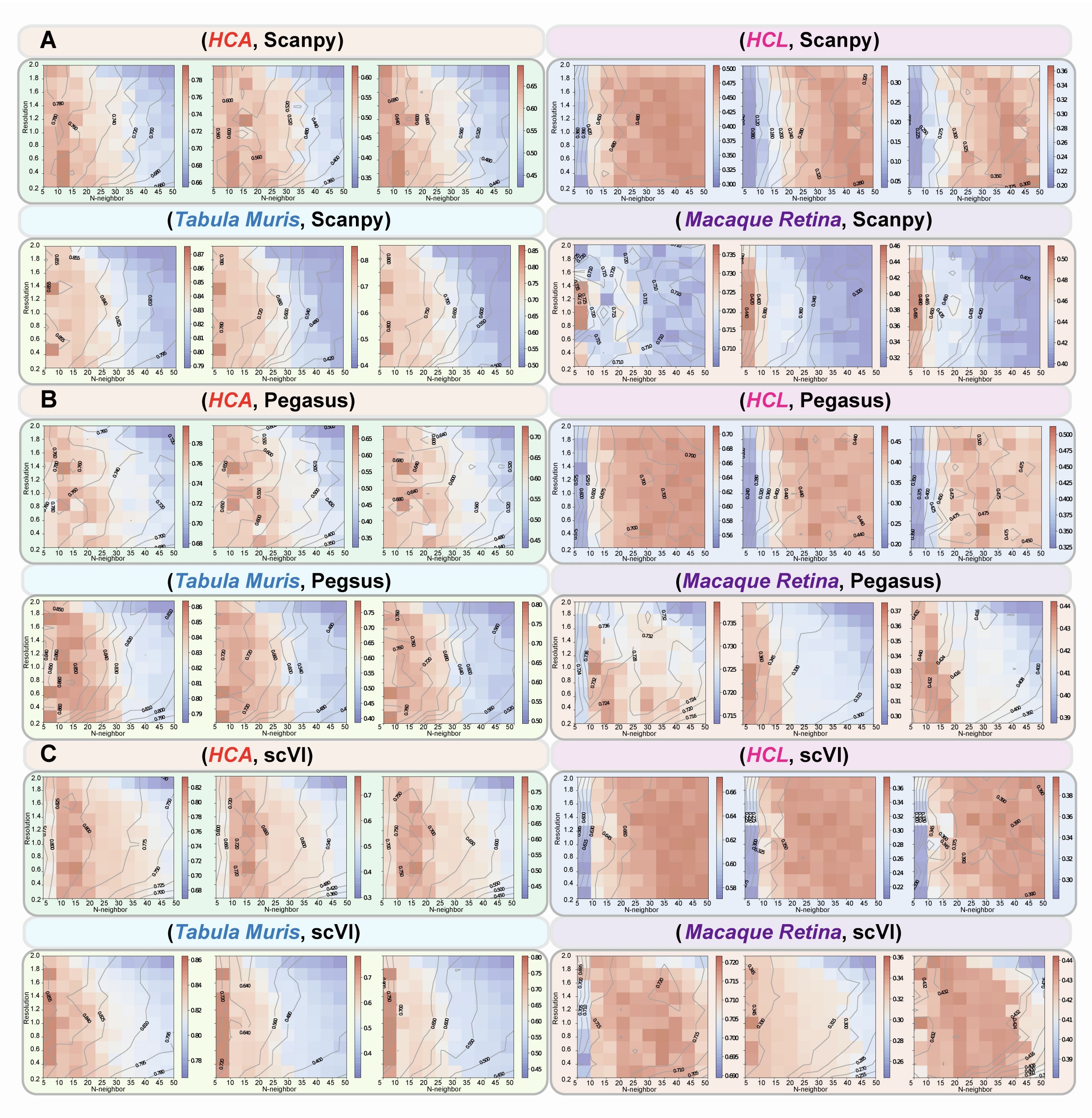

Figure S14. The clustering performance with grid search for resolution and number of neighbors for *Scanpy (A), Pegasus (B),* and *scVI (C)* on the *HCA*, *HCL*, *Tabula Muris* and *Macaque Retina* dataset, related to Figure 2. Contour maps depict different cluster metrics (i.e. *NMI*, *ARI* and *FMI*) with respect to different values of *Resolution* and *N-neighbors*.

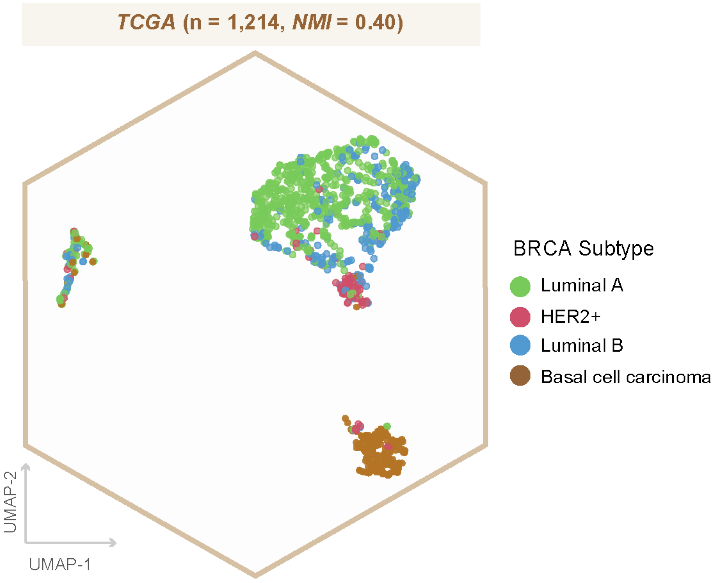

Figure S15. The UMAP visualization plots of *tGPT* for molecular subtypes of BRCA from the *TCGA* datasets, related to Figure 2. The NMI metric and annotation of cells are shown.

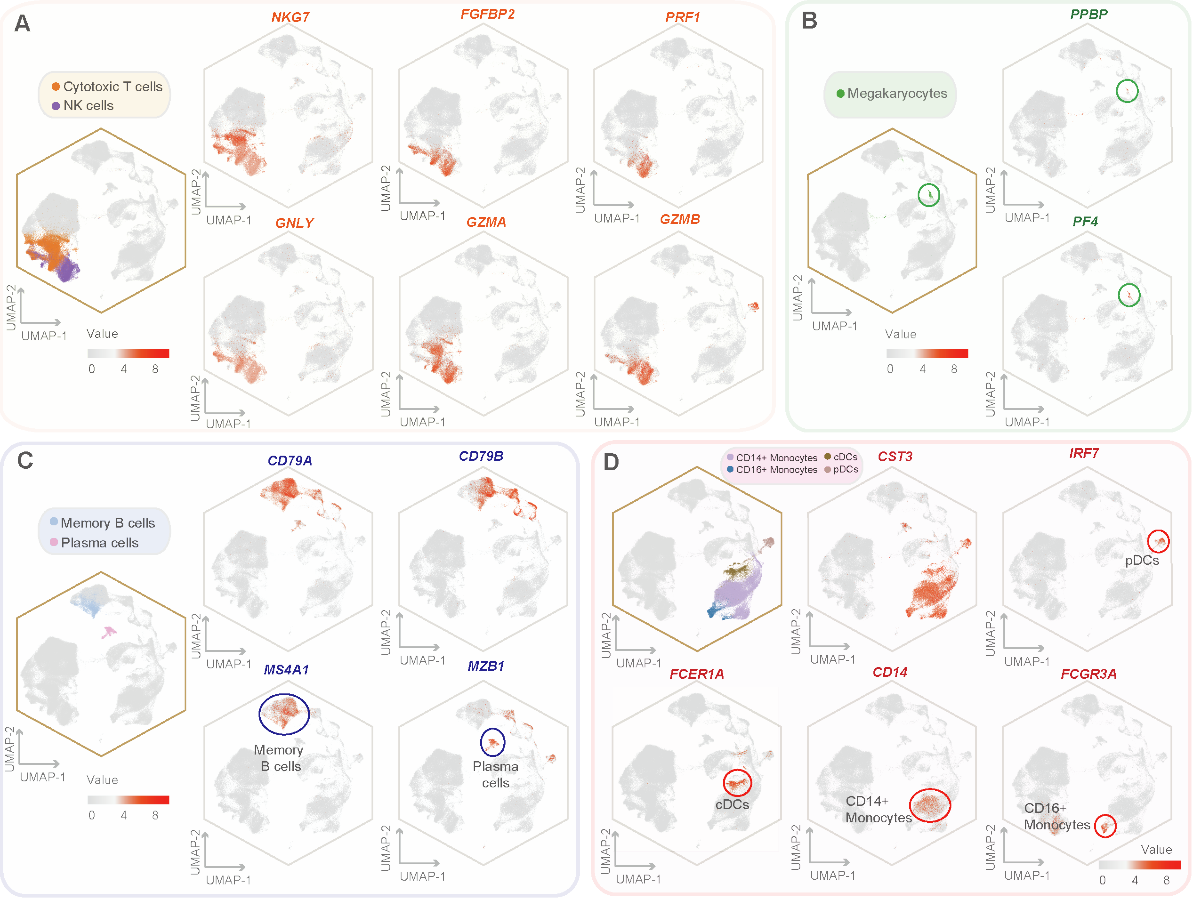

Figure S16. Distinct features of different cell types from the *HCA* dataset learned by *tGPT*, related to Figure 3. Scatter plots illustrating the distribution of attribution scores for different cell type specific genes across different cell types.

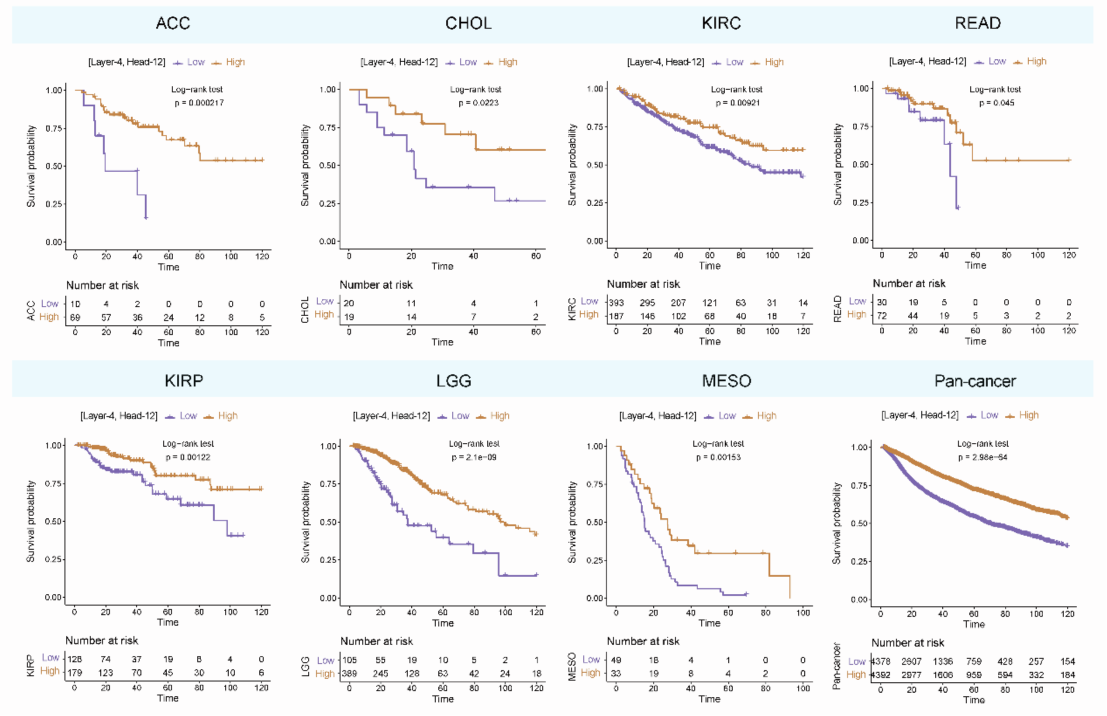

Figure S17. The survival curves of the attention head related to overall survival across multiple cancer types, related to Figure 5. ACC, Adrenocortical carcinoma; CHOL, Cholangiocarcinoma; KIRC, Kidney renal clear cell carcinoma; READ, Rectum adenocarcinoma; KIRP, Kidney renal papillary cell carcinoma; LGG, Brain Lower Grade Glioma; MESO, Mesothelioma.

Table S2. The annotated cell labels on the different datasets, related to Figure 2.

HCA	HCL	Tabula Muris	Macaque Retina	GTEx	TCGA
Naive B cell	Neutrophil	myeloid cell	fGa	Adipose	ACC
CD14+ Monocyte	Stromal cell	alveolar macrophage	fGl	Tissue	BLCA
T-helper cell	Fibroblast	B cell	DB3b	Adrenal	NA
Pre-B cells	Monocyte	natural killer cell	FMB	Gland	DLBC
Naive CD8+ T	Macrophage	T cell	IMB	Blood	UCEC
cell	Antigen presenting cell	lung endothelial cell	DB5*	Vessel	SKCM
Cytotoxic T cells	(RPS high)	stromal cell	DB4	Bladder	HNSC
Pro-B cell	Mast cell	non-classical	DB2	Brain	PRAD
CD4+ naive T	Sinusoidal endothelial cell	monocyte	DB1	Breast	KIRP
cell	T cell	leukocyte	BB/GB*	Blood	PAAD
NK cells	B cell	classical monocyte	RB	Skin	SARC
Erythroid cells	Dendritic cell	ciliated columnar cell	DB6	Cervix Uteri	CESC
cDCs	M2 Macrophage	of tracheobronchial	OFFx	Colon	COAD
Megakaryocyte	Epithelial cell	tree	DB3a	Esophagus	LUSC
Memory B cell	B cell (Plasmocyte)	type II pneumocyte	H1	Fallopian	READ
Plasma cell	Intercalated cell	mast cell	H2	Tube	KIRC
pDCs	Loop of Henle	monocyte	MG	Heart	LIHC
CD16+ Monocyte	Erythroid progenitor cell	granulocytopoietic	Pericytes	Kidney	BRCA
HSCs	(RP high)	cell	Endo	Liver	OV
MSCs	Fetal epithelial progenitor	promonocyte	Mic	Lung	UCS

	Ureteric bud cell Endothelial cell Endothelial cell (APC) Smooth muscle cell hESC Stratified epithelial cell Proximal tubule progenitor Fetal enterocyte Myeloid cell Proliferating T cell Endothelial cell (endothelial to mesenchymal transition) Enterocyte progenitor Enterocyte Fetal stromal cell Erythroid cell Hepatocyte/Endodermal cell Fetal mesenchymal progenitor Neutrophil (RPS high) Fetal neuron Fetal Neuron Fetal endocrine cell AT2 cell Basal cell Epithelial cell (intermediated) Chondrocyte CB CD34+ Fetal chondrocyte Intermediated cell Gastric endocrine cell Primordial germ cell Oligodendrocyte Astrocyte Fasciculata cell Immature sertoli cell (Pre-Sertoli cell) Fetal fibroblast Fetal skeletal muscle cell Fetal acinar cell Mesothelial cell Goblet cell Ventricle cardiomyocyte Kidney intercalated cell Thyroid follicular cell Adrenal gland	granulocyte erythroblast hematopoietic precursor cell proerythroblast late pro-B cell basophil macrophage early pro-B cell immature B cell Fraction A pre-pro B cell basal cell of epidermis keratinocyte Langerhans cell dendritic cell endothelial cell fibroblast endocardial cell cardiac muscle cell mesenchymal cell epithelial cell blood cell neuroendocrine cell bladder cell bladder urothelial cell luminal epithelial cell of mammary gland basal cell kidney capillary endothelial cell mesangial cell kidney cell kidney collecting duct epithelial cell kidney proximal straight tubule epithelial cell kidney loop of Henle ascending limb epithelial cell immature T cell DN1 thymic pro-T cell hepatocyte duct epithelial cell	M/L Cones S Cones Rods MG_OFF PG_OFF PG_ON MG_ON fRGC pGa pGl pRGC	Salivary Gland Muscle Nerve Ovary Pancreas Pituitary Prostate Small Intestine Spleen Stomach Testis Thyroid Uterus Vagina	GBM KICH THCA LGG LUAD MESO PCPG TGCT UVM THYM CHOL ESCA STAD

	inflammatory cell Pancreas exocrine cell Gastric chief cell	endothelial cell of hepatic sinusoid mesenchymal stem cell skeletal muscle satellite cell			

Table S3. Running time of different methods on the four datasets, related to Figure 2.

Dataset Method	Runtime (Seconds)			
	HCA (n = 282,558)	*HCL* (n = 586,135)	*Tabula Muris* (n = 54,862)	*Macaque Retina* (n = 124, 965)
tGPT	2513.5	2996.9	326.5	576.4
Scanpy	2352.2	3318.2	237.4	494.2
Pegasus	1575.8	1794.1	198.1	398.8
scVI	2849.5	3170.0	2084.0	2613.0

Table S4. The generative metrics of *tGPT* on the four datasets, related to Figure 2.

Metrics Dataset	*BLEU*
HCA	0.77
HCL	0.69
Tabula Muris	0.76
Macaque Retina	0.75