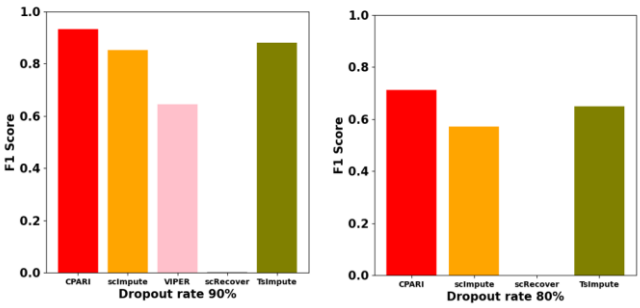
Supplementary to “CPARI: A Novel Approach Combining Cell Partitioning with Absolute and Relative Imputation to Address Dropout in Single-Cell RNA-seq Data”

Yi Zhang[1,2], Yin Wang [1,2,*], Xinyuan Liu[1,2], Xi Feng[1,2]

1. School of Computer Science and Engineering, Guilin University of Technology, Guilin 541004, China.
2. Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin541004, China.

* To whom correspondence should be addressed. Email: wangyin@glut.edu.cn

1. Supplementary Figures

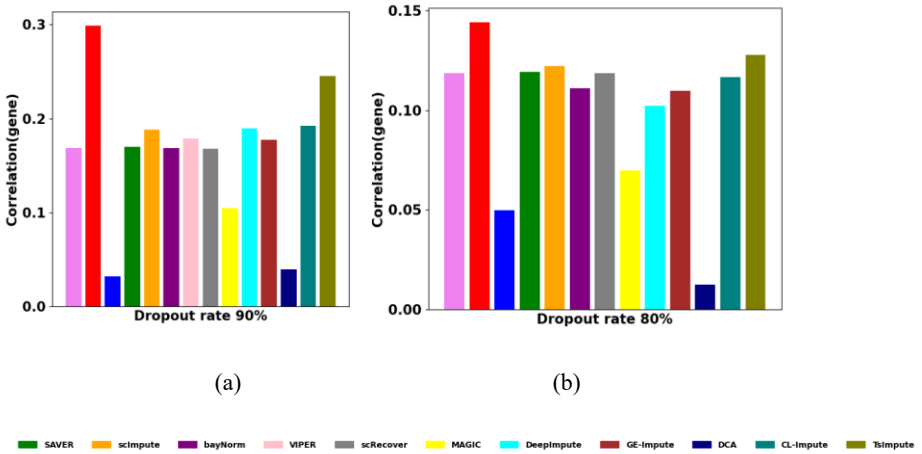


(a) (b)

Figure S1: Mean F1 Score for dropout datasets.

(a): Mean standard F1 Score for Dataset1*- Dataset3* (dropout rate 90%).

(b) Mean standard F1 Score for Dataset4*- Dataset6* (dropout rate 80%).

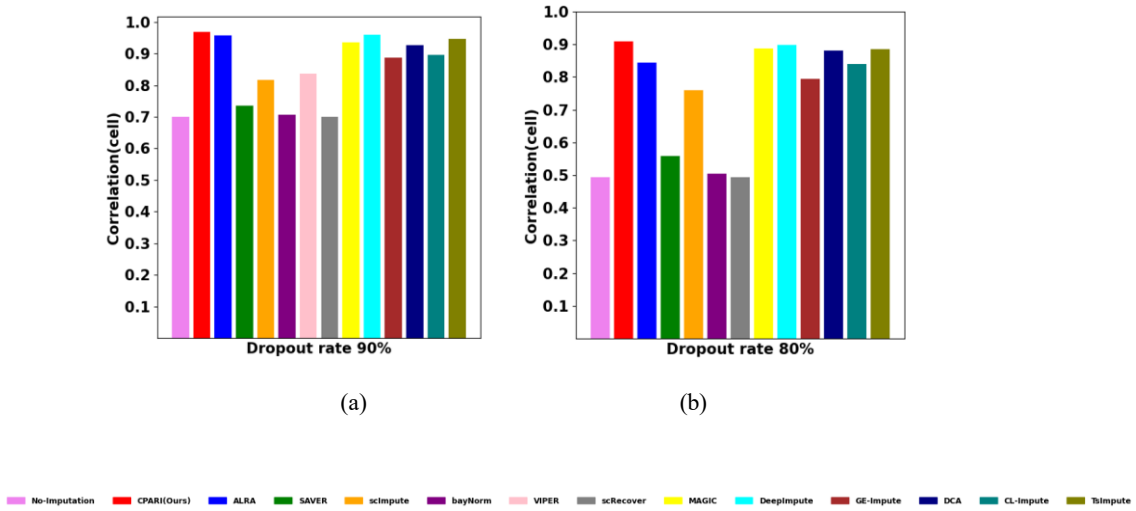


(a) (b)

Figure S2: Mean Correlatifor (gene) for dropout datasets.

(a): Mean standard Correlation (gene) for dropout Dataset1*- Dataset3* (dropout rate 90%).

(b) Mean standard Correlation (gene) for dropout Dataset4*- Dataset6* (dropout rate 80%).

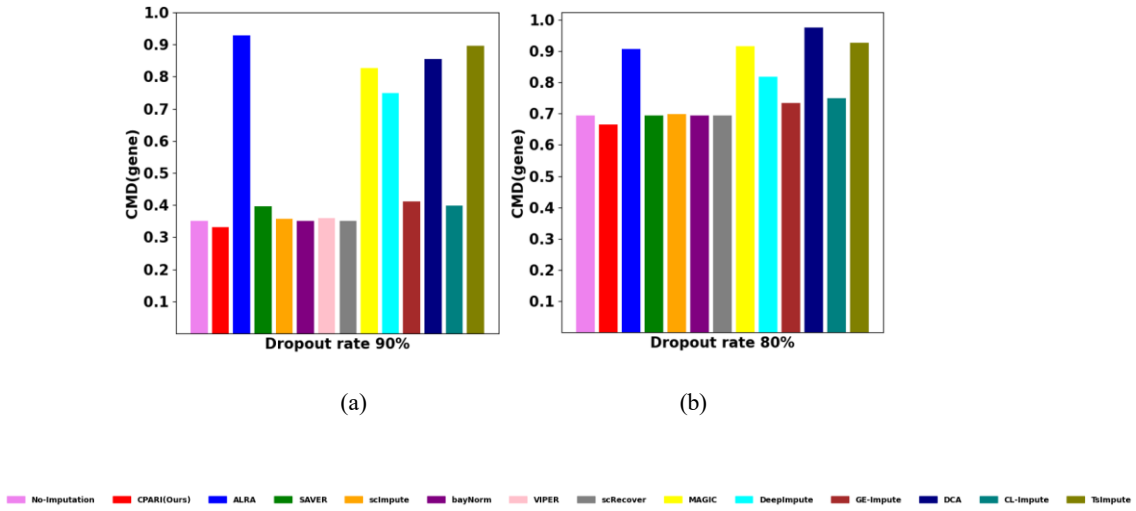


(a) (b)

Figure S3: Mean Correlation (cell) for dropout datasets.

(a): Mean standard Correlation (cell) for dropout Dataset1*- Dataset3* (dropout rate 90%).

(b) Mean standard Correlation (cell) for dropout Dataset4*- Dataset6* (dropout rate 80%).



(a) (b)

Figure S4: Mean CMD (gene) for dropout datasets.

(a): Mean standard CMD (gene) for dropout Dataset1*- Dataset3* (dropout rate 90%).

(b) Mean standard CMD (gene) for dropout Dataset4*- Dataset6* (dropout rate 80%).

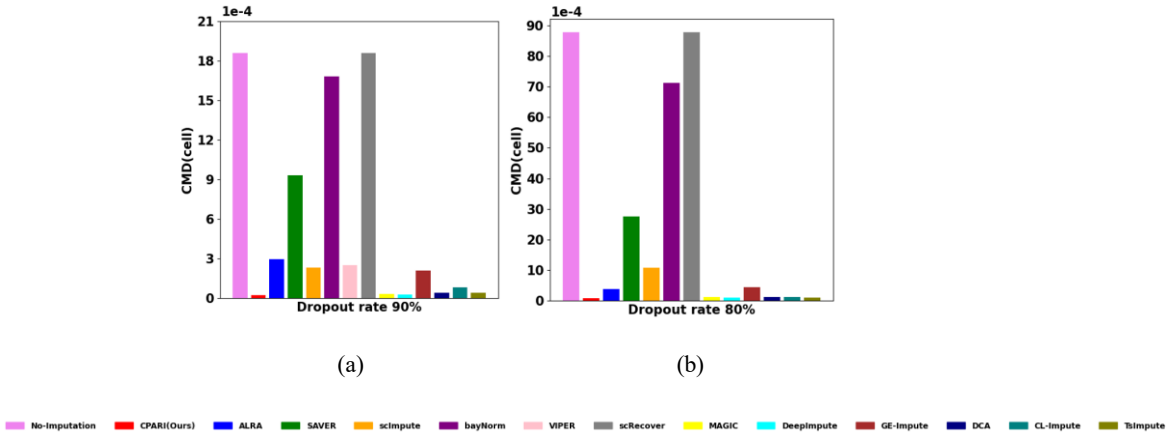


Figure S5: Mean CMD (cell) for dropout datasets.

(a): Mean standard CMD (cell) for dropout Dataset1*- Dataset3* (dropout rate 90%).

(b) Mean standard CMD (cell) for dropout Dataset4*- Dataset6* (dropout rate 80%).

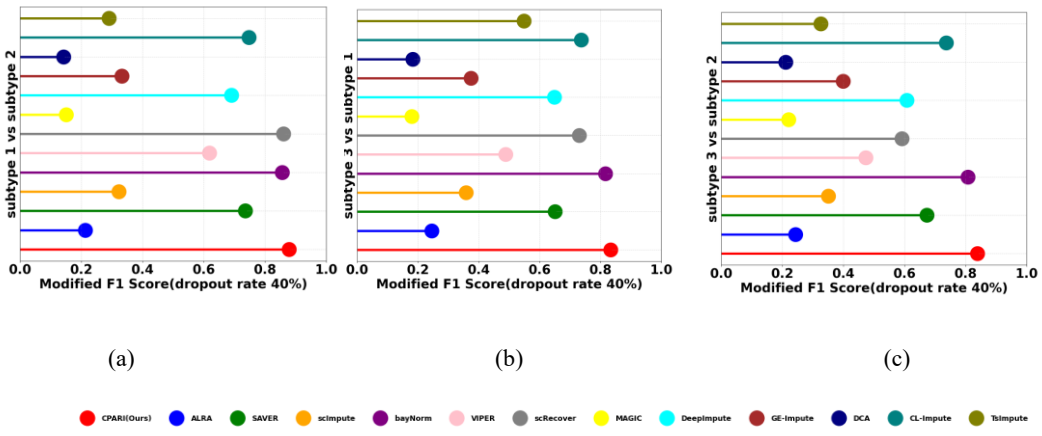


Figure S6: Mean F1 scores of three cell types relative to each other across dropout datasets 1-3 (with 40% dropout rate).

(a): Mean F1 Score of the cell type 2 vs cell type 1.

(b): Mean F1 Score of the cell type 3 vs cell type 1.

(c): Mean F1 Score of the cell type 3 vs cell type 2.

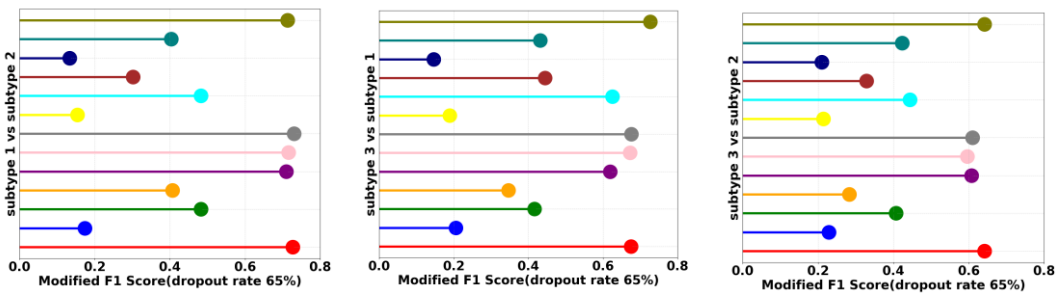


Figure S7: Mean F1 Score values of three cell types relative to each other across dropout datasets 1-3 (with 65% dropout rate).

(a): Mean F1 Score of the cell type 2 vs cell type 1.

(b): Mean F1 Score of the cell type 3 vs cell type 1.

(c): Mean F1 Score of the cell type 3 vs cell type 2.

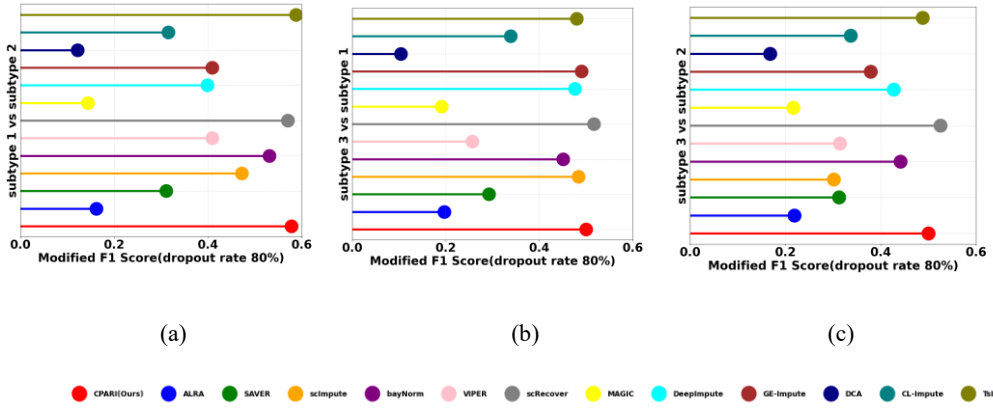
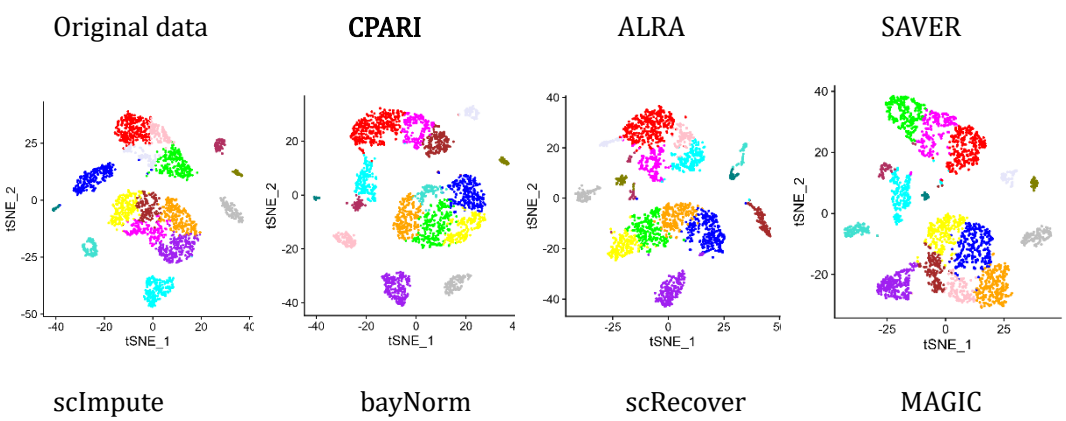


Figure S8: Mean F1 Score of the three types of cells in relation to each other across dropout datasets 1-3 (with 80% dropout rate).

(a): Mean F1 Score of the cell type 2 vs cell type 1.

(b): Mean F1 Score of the cell type 3 vs cell type 1.

(c): Mean F1 Score of the cell type 3 vs cell type 2.



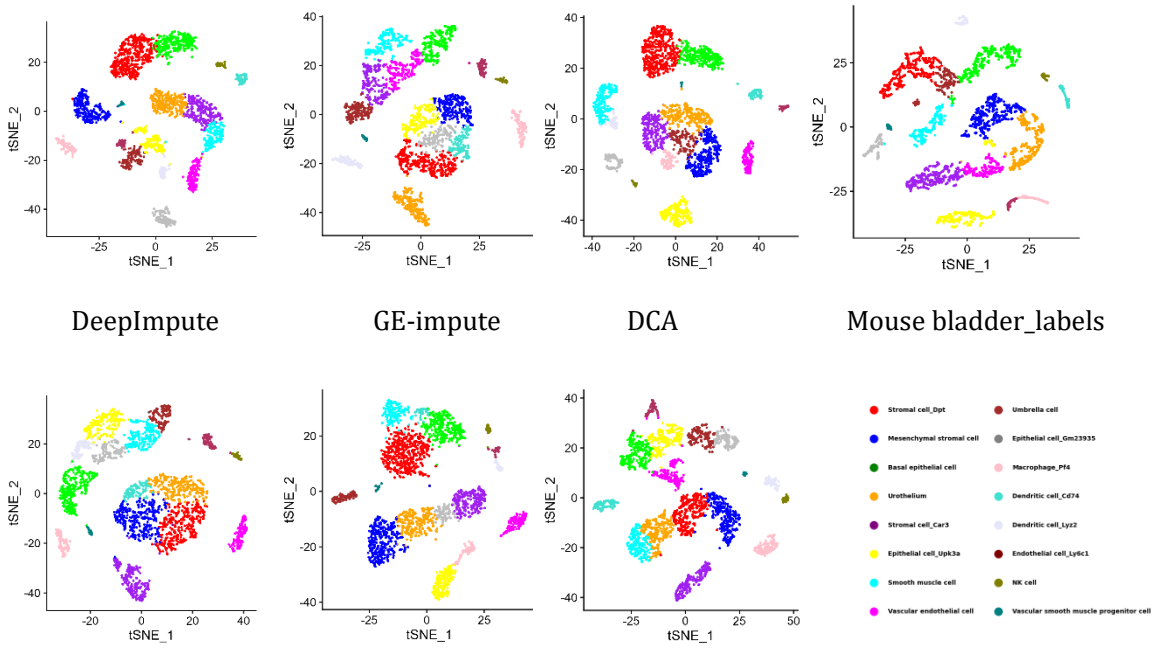


Figure S9: T-SNE visualization of imputed feature matrices for the Mouse bladder dataset.

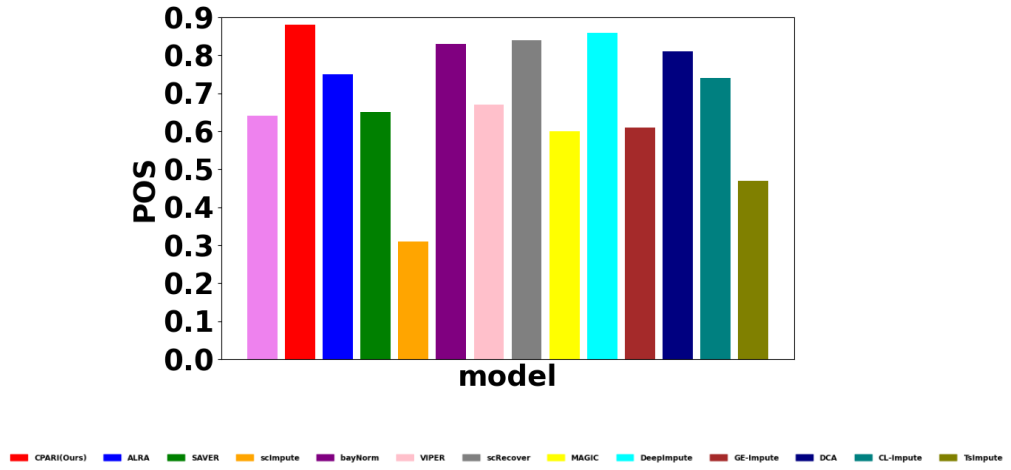


Figure S10: POS index of reconstructed cell pseudotime. A higher POS index indicates a more accurate reconstruction of the pseudotime.

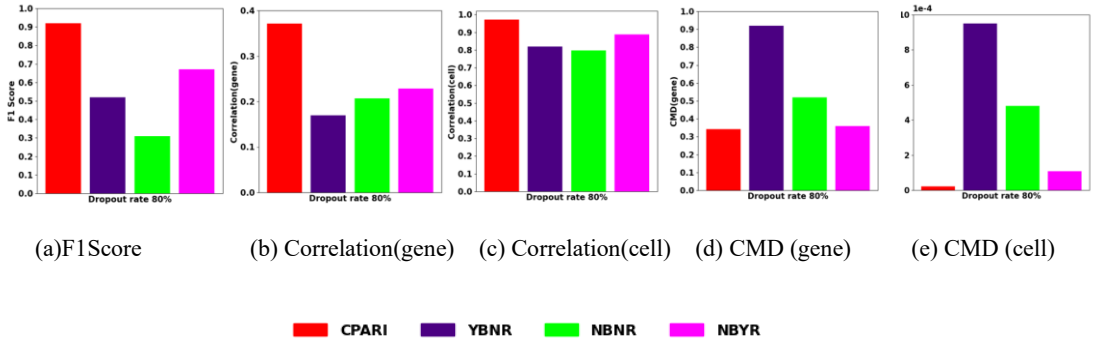


Figure S11: Performance comparison between CPARI and its variants.

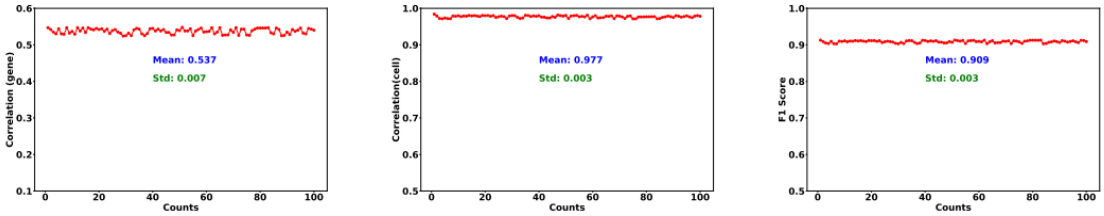
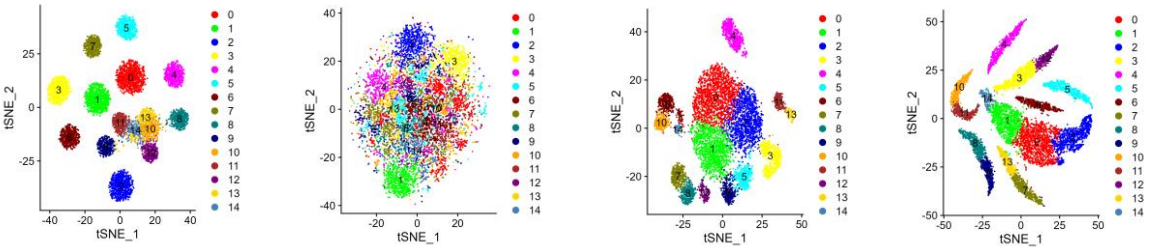


Figure S12. CPARI robustness evaluation. To evaluate the robustness of the CPARI model, 100 simulated datasets with a 65% dropout rate were generated. The mean and standard deviation of the gene correlation coefficient (Correlation(gene)), cell correlation coefficient (Correlation (cell)), and the ability to identify dropout zeros (F1 Score) were calculated for both the complete dataset and the dataset imputed using CPARI.

Complete data Dropout data CPARI (Ours) DeepImpute



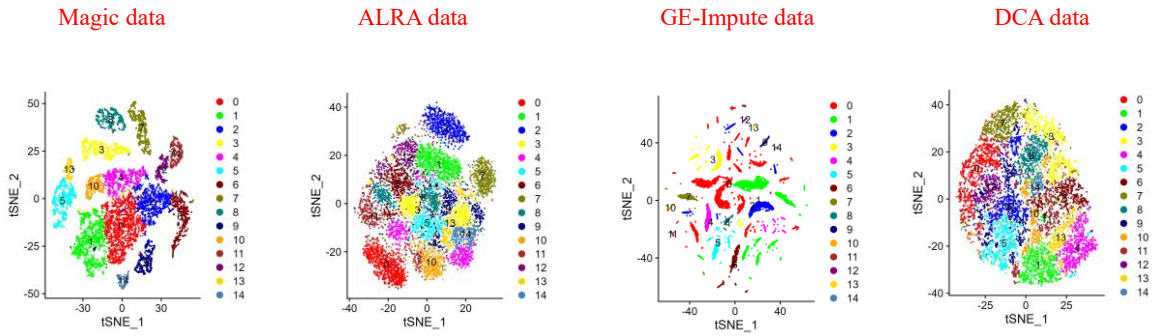


Figure S13: Clustering visualization of imputed data. This figure presents the results of clustering visualization applied to the imputed data. A subset of 20,000 genes and 30,000 cells was selected for analysis, and t-SNE was employed for dimensionality reduction.

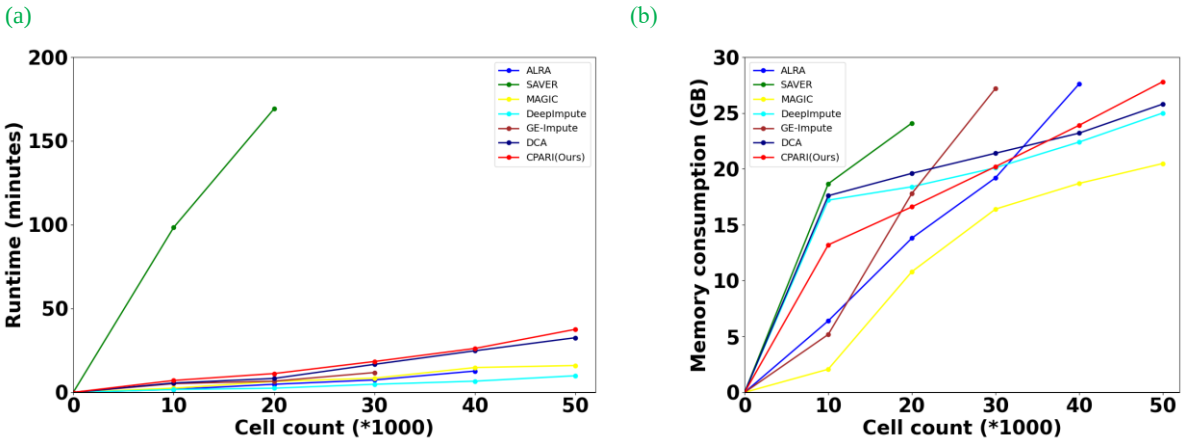


Figure S14: Computational efficiency evaluation. This figure presents the runtime and memory consumption of various imputation models, including CPARI. The analysis focuses on the computational efficiency of each model, taking into account CPARI's optimization to avoid redundant file reading and writing.

2. Supplementary Tables

Table S1. Mean Cophenetic of imputation methods on dropout datasets.

Methods	Cophenetic
Dropout rate 30% (Dataset1*- Dataset3*)	**0.9961**
Dropout rate 40% (Dataset1*- Dataset3*)	**0.9824**
Dropout rate 50% (Dataset1*- Dataset3*)	**0.9994**
Dropout rate 65% (Dataset1*- Dataset3*)	**0.9992**
Dropout rate 80% (Dataset1*- Dataset3*)	**0.9986**
Dropout rate 90% (Dataset1*- Dataset3*)	**0.9973**
Dropout rate 80% (Dataset4*- Dataset6*)	**0.9949**
PBMC	**0.9980**
Worm neuron cells	**0.9991**
Mouse bladder cells	**0.9987**
LPS	**0.9984**

Table S2. Mean F1 Score of imputation methods on dropout datasets.

Methods	Dropout Rate:30%	Dropout Rate:40%	Dropout Rate50%	Dropout Rate:65%	Dropout Rate:80%
CPARI	**0.7489**	**0.8175**	**0.8739**	**0.9089**	**0.9267**
ALRA	0.6219	0.7359	0.8321	0.8862	0.9127
SAVER	0.6026	0.7122	0.8029	0.8556	0.8865
scImpute	0.6484	0.7623	0.8551	0.9009	0.8928
bayNorm	0.6158	0.7229	0.8137	0.8674	0.8942
VIPER	0.7090	0.8118	0.8750	0.8598	0.7461
scRecover	0.5669	0.4674	0.2305	0.0700	0.0063
MAGIC	0.6063	0.7181	0.8085	0.8618	0.8918
DeepImpute	0.5921	0.7058	0.7954	0.8432	0.8721
GE-Impute	0.7315	0.7896	0.8018	0.7608	0.6640
DCA	0.6058	0.7175	0.8048	0.8594	0.8890
CL–Impute	0.6841	0.7085	0.7124	0.6629	0.5901
TsImpute	0.6224	0.7292	0.8253	0.8679	0.8797

Table S3. F1 Score comparison for dropout datasets.

Methods	F1 Score (dropout rate 90%)				F1 Score (dropout rate 80%)			
	Dataset1*	Dataset2*	Dataset3*	Mean	Dataset4*	Dataset5*	Dataset6*	Mean
CPARI	0.9356	0.9341	0.9287	**0.9328**	0.7159	0.7134	0.7085	**0.7126**
scImpute	0.8549	0.8594	0.8459	0.8534	0.5695	0.5756	0.5691	0.5714
VIPER	0.6455	0.6514	0.6402	0.6457	*	*	*	*
scRecover	0.0024	0.0027	0.0018	0.0023	0.0000	0.0000	0.0000	0.0000
TsImpute	0.8887	0.8827	0.8707	0.8807	0.6471	0.6543	0.6471	0.6495

*: No results for VIPER within 24 hours

Table S4. Mean Correlation (gene) of imputation methods on dropout datasets.

Methods	Dropout Rate:30%	Dropout Rate:40%	Dropout Rate50%	Dropout Rate:65%	Dropout Rate:80%
Original	0.700	0.574	0.427	0.304	0.207
CPARI	**0.875**	**0.802**	**0.691**	**0.539**	**0.371**
ALRA	0.098	0.086	0.074	0.058	0.045
SAVER	0.740	0.609	0.447	0.312	0.210
scImpute	0.857	0.762	0.580	0.379	0.240
bayNorm	0.733	0.595	0.435	0.306	0.208
VIPER	0.842	0.682	0.568	0.372	0.23
scRecover	0.713	0.577	0.428	0.304	0.207
MAGIC	0.285	0.250	0.202	0.156	0.120
DeepImpute	0.732	0.680	0.585	0.428	0.282
GE-Impute	0.827	0.752	0.597	0.400	0.233
DCA	0.147	0.125	0.094	0.071	0.048
CL–Impute	0.839	0.758	0.604	0.412	0.252
TsImpute	0.851	0.784	0.657	0.506	0.299

Table S5. Gene-level correlation analysis for dropout datasets.

Methods	Correlation-gene (dropout rate 90%)				Correlation-gene (dropout rate 80%)			
	Dataset1*	Dataset2*	Dataset3*	Mean	Dataset4*	Dataset5*	Dataset6*	Mean
Original	0.1808	0.1651	0.1587	0.1682	0.1306	0.1174	0.1078	0.1186
CPARI	0.3092	0.2974	0.2901	**0.2989**	0.1539	0.1431	0.1356	**0.1442**
ALRA	0.0405	0.0294	0.0252	0.0317	0.0616	0.0475	0.0403	0.0498
SAVER	0.1767	0.1682	0.1654	0.1701	0.1303	0.1174	0.1102	0.1193
scImpute	0.2008	0.1851	0.1793	0.1884	0.1328	0.1203	0.1135	0.1222
bayNorm	0.1815	0.1647	0.1596	0.1686	0.1211	0.1101	0.1024	0.1112
VIPER	1922	0.1742	0.1694	0.1786	*	*	*	*
scRecover	0.1805	0.1647	0.1591	0.1681	0.1290	0.1174	0.1094	0.1186
MAGIC	0.1126	0.1012	0.0991	0.1043	0.0806	0.0676	0.0612	0.0698
DeepImpute	0.1998	0.1863	0.1821	0.1894	0.1128	0.1005	0.0936	0.1023
GE-Impute	0.1883	0.1745	0.1697	0.1775	0.1203	0.1079	0.1012	0.1098
DCA	0.0503	0.0371	0.0314	0.0396	0.0229	0.0114	0.0032	0.0125
CL–Impute	0.2024	0.1894	0.1854	0.1924	0.1293	0.1124	0.1087	0.1168
TsImpute	0.2570	0.2402	0.2384	0.2452	0.1376	0.1254	0.1201	0.1277

*: No results for VIPER within 24 hours.

Table S6. Mean Correlation (cell) of imputation methods on dropout datasets.

Methods	Dropout Rate:30%	Dropout Rate:40%	Dropout Rate50%	Dropout Rate:65%	Dropout Rate:80%
Original	0.971	0.946	0.898	0.831	0.748
CPARI	**0.995**	**0.991**	**0.986**	**0.979**	**0.972**
ALRA	0.963	0.962	0.960	0.956	0.95
SAVER	0.983	0.967	0.929	0.865	0.782

scImpute	0.993	0.987	0.969	0.929	0.858
bayNorm	0.982	0.961	0.911	0.841	0.754
VIPER	0.991	0.737	0.970	0.932	0.873
scRecover	0.973	0.947	0.898	0.831	0.748
MAGIC	0.969	0.969	0.967	0.961	0.947
DeepImpute	0.989	0.986	0.981	0.973	0.966
GE-Impute	0.991	0.987	0.977	0.955	0.914
DCA	0.969	0.968	0.965	0.957	0.940
CL–Impute	0.991	0.980	0.975	0.953	0.918
TsImpute	0.992	0.986	0.980	0.963	0.953

Table S6*. Mean PCC of imputation methods on dropout datasets.

Methods	Dropout Rate:30%	Dropout Rate:40%	Dropout Rate50%	Dropout Rate:65%	Dropout Rate:80%
Original	0.9707	0.9462	0.8982	0.8309	0.7473
CPARI	**0.9942**	**0.9917**	**0.9865**	**0.9764**	**0.9713**
ALRA	0.9571	0.9560	0.9547	0.9522	0.9481
SAVER	0.9928	0.9674	0.9286	0.8651	0.7820
scImpute	0.9928	0.9866	0.9690	0.9288	0.8580
bayNorm	0.9818	0.9606	0.9114	0.8406	0.7541
VIPER	0.9927	0.9869	0.9696	0.9318	0.8730
scRecover	0.9729	0.9466	0.8981	0.8309	0.7463
MAGIC	0.9696	0.9692	0.9673	0.9612	0.9465
DeepImpute	0.9882	0.9854	0.9804	0.9727	0.9654
GE-Impute	0.9912	0.9871	0.9765	0.9547	0.9133
DCA	0.9684	0.9674	0.9646	0.9567	0.9389
CL–Impute	0.9916	0.9867	0.9752	0.9537	0.9181
TsImpute	0.9934	0.9898	0.9829	0.9741	0.9640

Table S7. Cell-level correlation analysis for dropout datasets.

Methods	Correlation-cell (dropout rate 90%)				Correlation-cell (dropout rate 80%)			
	Dataset1*	Dataset2*	Dataset3*	Mean	Dataset4*	Dataset5*	Dataset6*	Mean
Original	0.7116	0.6984	0.6915	0.7005	0.5032	0.4954	0.4801	0.4929
CPARI	0.9806	0.9675	0.9571	**0.9684**	0.9191	0.9115	0.8985	**0.9097**
ALRA	0.9714	0.9576	0.9456	0.9582	0.8560	0.8452	0.8329	0.8447
SAVER	0.7499	0.7340	0.7223	0.7354	0.5690	0.5591	0.5456	0.5579
scImpute	0.8315	0.8141	0.8042	0.8166	0.7594	0.7615	0.7582	0.7597
bayNorm	0.7227	0.7014	0.6939	0.7060	0.5114	0.5084	0.4928	0.5042
VIPER	0.8517	0.8315	0.8257	0.8363	*	*	*	*
scRecover	0.7103	0.6984	0.6925	0.7004	0.4994	0.4981	0.4812	0.4929
MAGIC	0.9515	0.9324	0.9232	0.9357	0.8980	0.8901	0.8765	0.8882
DeepImpute	0.9749	0.9548	0.9512	0.9603	0.9067	0.9015	0.8861	0.8981
GE-Impute	0.9073	0.8815	0.8731	0.8873	0.8019	0.7981	0.7832	0.7944
DCA	0.9419	0.9241	0.9156	0.9272	0.8886	0.8867	0.8701	0.8818

CL–Impute	0.9042	0.8954	0.8857	0.8951	0.8499	0.8426	0.8284	0.8403
TsImpute	0.9505	0.9512	0.9351	0.9456	0.8935	0.8891	0.8754	0.8860

*: No results for VIPER within 24 hours

Table S8. Mean Error of imputation methods on dropout datasets.

Methods	Dropout Rate:30%	Dropout Rate:40%	Dropout Rate50%	Dropout Rate:65%	Dropout Rate:80%
Original	0.0770	0.1531	0.3287	0.6170	0.9940
CPARI	**0.0135**	**0.0277**	**0.0366**	**0.0556**	**0.0769**
ALRA	0.1005	0.1035	0.1072	0.1133	0.1217
SAVER	0.0411	0.0891	0.2252	0.4872	0.8590
scImpute	0.0171	0.0346	0.0953	0.2681	0.6163
bayNorm	0.0452	0.1087	0.2817	0.5773	0.9641
VIPER	0.0169	0.0307	0.0821	0.2238	0.4933
scRecover	0.0675	0.1487	0.3278	0.6168	0.9940
MAGIC	0.0698	0.0731	0.0898	0.1405	0.2485
DeepImpute	0.0352	0.0440	0.0579	0.0817	0.0966
GE-Impute	0.0203	0.0301	0.0606	0.1423	0.3302
DCA	0.0728	0.0775	0.0977	0.1568	0.2800
CL–Impute	0.0194	0.0319	0.0670	0.1502	0.3178
TsImpute	0.0153	0.0295	0.0394	0.0705	0.0861

Table S9. Mean CMD (cell) of imputation methods on dropout datasets.

Methods	Dropout Rate:30%	Dropout Rate:40%	Dropout Rate50%	Dropout Rate:65%	Dropout Rate:80%
Original	4.171e-05	9.133e-05	2.111e-04	4.573e-04	1.094e-03
CPARI	**2.424e-06**	**4.739e-06**	**1.121e-05**	**1.679e-05**	**2.225e-05**
ALRA	1.24e-04	1.097e-04	1.214e-04	2.069e-04	3.467e-04
SAVER	1.888e-05	2.525e-05	5.497e-05	1.700e-04	5.175e-04
scImpute	1.804e-05	2.365e-05	3.534e-05	7.5574e-05	1.174e-04
bayNorm	1.809e-05	2.950e-05	1.074e-04	3.394e-04	9.525e-04
VIPER	5.667e-06	2.216e-04	2.715e-05	6.934e-05	1.679e-04
scRecover	3.716e-05	8.848e-05	2.117e-04	4.577e-04	1.095e-03
MAGIC	2.95e-05	2.946e-05	2.955e-05	3.029e-05	3.153e-05
DeepImpute	4.74e-06	8.676e-06	1.787e-05	2.785e-05	4.174e-05
GE-Impute	2.892e-06	6.585e-06	1.978e-05	5.111e-05	1.311e-04
DCA	2.965e-05	3.062e-05	3.2185e-05	3.565e-05	4.030e-05
CL–Impute	2.259e-06	5.618e-06	1.426e-05	2.113e-05	3.153e-05
TsImpute	5.591e-06	9.655e-06	1.766e-05	2.612e-05	5.447e-05

Table S10. Mean CMD (gene) of imputation methods on dropout datasets.

Methods	Dropout Rate:30%	Dropout Rate:40%	Dropout Rate50%	Dropout Rate:65%	Dropout Rate:80%
Original	0.170	0.229	0.288	0.326	0.345

CPARI	**0.077**	**0.125**	**0.188**	**0.238**	**0.343**
ALRA	0.909	0.908	0.904	0.895	0.886
SAVER	0.142	0.205	0.275	0.32	0.412
scImpute	0.090	0.145	0.269	0.339	0.354
bayNorm	0.151	0.217	0.283	0.324	0.345
VIPER	0.096	0.804	0.229	0.306	0.348
scRecover	0.167	0.229	0.289	0.326	0.345
MAGIC	0.748	0.778	0.819	0.84	0.833
DeepImpute	0.654	0.683	0.736	0.763	0.753
GE-Impute	0.134	0.175	0.258	0.347	0.385
DCA	0.737	0.761	0.797	0.821	0.841
CL–Impute	0.093	0.166	0.225	0.312	0.365
TsImpute	0.382	0.426	0.522	0.298	0.853

Table S11. Gene-level CMD analysis for dropout datasets.

Methods	CMD (gene) (dropout rate 90%)				CMD (gene) (dropout rate 80%)			
	Dataset1*	Dataset2*	Dataset3*	Mean	Dataset4*	Dataset5*	Dataset6*	Mean
Original	0.3401	0.3487	0.3654	0.3514	0.6801	0.6943	0.7058	0.6934
CPARI	0.3185	0.3276	0.3451	**0.3304**	0.6532	0.6648	0.6755	**0.6645**
ALRA	0.9183	0.9291	0.9402	0.9292	0.8944	0.9064	0.9175	0.9061
SAVER	0.3823	0.3972	0.4115	0.3970	0.6811	0.6946	0.7051	0.6936
scImpute	0.3425	0.3658	0.3624	0.3569	0.6865	0.7010	0.7116	0.6997
bayNorm	0.3395	0.3443	0.3701	0.3513	0.6826	0.6953	0.7062	0.6947
VIPER	0.3481	0.3569	0.3753	0.3601	*	*	*	*
scRecover	0.3402	0.3509	0.3634	0.3515	0.6813	0.6940	0.7049	0.6934
MAGIC	0.8124	0.8255	0.8398	0.8259	0.9027	0.9154	0.9272	0.9151
DeepImpute	0.7376	0.7494	0.7624	0.7498	0.8046	0.8196	0.8298	0.8180
GE-Impute	0.4002	0.3307	0.5027	0.4112	0.7204	0.7345	0.7459	0.7336
DCA	0.8427	0.8569	0.8675	0.8557	0.9628	0.9768	0.9881	0.9759
CL–Impute	0.3813	0.4011	0.4092	0.3972	0.7285	0.7584	0.7634	0.7501
TsImpute	0.8829	0.8991	0.9042	0.8954	0.9200	0.9286	0.9342	0.9276

*: No results for VIPER within 24 hours

Table S12. Cell-level CMD analysis for dropout datasets..

Methods	CMD (cell) (dropout rate 90%)				CMD (cell) (dropout rate 80%)			
	Dataset1*	Dataset2*	Dataset3*	Mean	Dataset4*	Dataset5*	Dataset6*	Mean
Original	1.18e-03	2.26e-03	2.14e-03	1.86e-03	8.31e-03	8.89e-03	9.14e-03	8.78e-03
CPARI	1.61e-05	2.70e-05	2.74e-05	**2.35e-05**	6.55e-05	7.34e-05	7.65e-05	**7.18e-05**
ALRA	2.22e-04	3.38e-04	3.30e-04	2.97e-04	3.23e-04	4.04e-04	4.10e-04	3.79e-04
SAVER	8.42e-04	9.76e-04	9.75e-04	9.31e-04	2.16e-03	3.07e-03	3.05e-03	2.76e-03
scImpute	1.49e-04	2.80e-04	2.61e-04	2.30e-04	4.1e-04	1.22e-03	1.61e-03	1.08e-03
bayNorm	7.9e-04	2.26e-03	1.99e-03	1.68e-03	6.32e-03	7.38e-03	7.69e-03	7.13e-03
VIPER	1.74e-04	2.91e-04	2.88e-04	2.51e-04	*	*	*	*
scRecover	1.13e-03	2.10e-03	2.35e-03	1.86e-03	8.14e-03	8.97e-03	9.23e-03	8.78e-03

MAGIC	2.43e-05	3.63e-05	3.69e-05	3.25e-05	5.8e-05	1.23e-04	1.61e-04	1.14e-04
DeepImpute	1.70e-05	2.85e-05	2.98e-05	2.51e-05	8.86e-05	9.56e-05	9.96e-05	9.46e-05
GE-Impute	1.42e-04	2.39e-04	2.52e-04	2.11e-04	3.89e-04	4.49e-04	4.85e-04	4.41e-04
DCA	3.69e-05	4.49e-04	4.75e-05	4.31e-05	6.7e-05	1.25e-04	1.74e-04	1.22e-04
CL–Impute	7.825e-05	8.192e-05	8.547e-05	8.188e-05	4.10e-05	1.27e-04	1.65e-04	1.11e-04
TsImpute	3.298e-05	3.901e-05	4.402e-05	3.867e-05	8.3e-05	1.19e-04	1.71e-04	1.03e-04

*: No results for VIPER within 24 hours

Table S13. Clustering performance evaluation.

Methods	PBMC		Mouse bladder		Worm neuron cells	
	NMI	ARI	NMI	ARI	NMI	ARI
Original	0.79	0.79	0.74	0.57	0.59	0.33
CPARI	**0.83**	**0.84**	**0.84**	**0.66**	**0.73**	**0.49**
ALRA	0.74	0.70	0.68	0.43	0.39	0.17
SAVER	0.78	0.78	0.73	0.53	0.56	0.31
scImpute	0.67	0.57	0.65	0.41	0.52	0.32
bayNorm	0.74	0.69	0.73	0.55	0.32	0.10
scRecover	0.76	0.76	0.67	0.40	0.46	0.27
MAGIC	0.70	0.66	0.70	0.59	0.55	0.30
DeepImpute	0.65	0.53	0.56	0.32	0.19	0.03
GE-Impute	0.75	0.73	0.68	0.45	0.46	0.29
DCA	0.80	0.80	0.75	0.52	0.64	0.42
CL-Impute	0.79	0.78	0.74	0.61	0.50	0.27
TsImpute	0.76	0.77	0.69	0.60	0.49	0.28
NBNR	0.62	0.50	0.68	0.46	0.43	0.22
NBYR	0.81	0.83	0.70	0.47	0.60	0.38

Table S14. Inferred trajectories analysis.

Model	Advantages	Disadvantages
Original data	The trajectory attempts to capture the progression from hour 1 to hour 6.	The trajectory is not very smooth, and the clusters for different time points are not well-separated, indicating noisy data.
CPARI	**Clear separation between clusters, especially between hour 1 and hour 6. The trajectory is more defined and follows a smoother path.**	**Some clusters, such as hour 4 (green), are slightly overlapping with hour 6 (purple).**
ALRA	The trajectory attempts to capture the progression from hour 1 to hour 6.	The trajectory is less defined and the clusters for different time points are not well-separated.
SAVER	The trajectory attempts to capture the progression from hour 1 to hour 6.	The trajectory is less defined and the clusters for different time points are not well-separated.

scImpute	the clusters for different time points are well-separated.	The trajectory is wrong and follows a rougher path.
bayNorm	Distinct separation between time points, especially between hour 1 and hour 6. The trajectory shows a clear progression.	Some overlaps between clusters, particularly between hour 4 (green) and hour 6 (purple).
scRecover	The trajectory attempts to capture the progression from hour 1 to hour 6 and the trajectory is smooth	Some overlaps between clusters, particularly between hour 4 (green) and hour 6 (purple).
MAGIC	The trajectory attempts to capture the progression from hour 1 to hour 6.	Some overlaps between clusters, particularly between hour 4 (green) and hour 6 (purple).
DeepImpute	The trajectory is smooth and captures the transition between time points well.	Some overlaps between clusters, particularly between hour 4 (green) and hour 6 (purple).
GE-impute	Displays a smooth trajectory with distinct clusters.	The trajectory is less defined.
DCA	Shows clear separation between clusters and a well-defined trajectory.	Some overlaps between clusters, particularly between hour 4 (green) and hour 6 (purple).
CL-IMpute	The trajectory is smooth and captures the transition between time points.	Some overlaps between clusters, particularly between hour 4 (green) and hour 6 (purple).
TsImpute	the clusters for different time points are well-separated.	The trajectory is wrong and follows a rougher path.

Table S15. POS index for each imputation method applied to the dataset.

Methods	POS
Original	0.64
CPARI	**0.88**
ALRA	0.75
SAVER	0.65
scImpute	0.31
bayNorm	0.83
VIPER	0.67
scRecover	0.84
MAGIC	0.60
DeepImpute	0.86
GE-Impute	0.61
DCA	0.81

CL-Impute	0.74
TsImpute	0.47

Table S16. Correlation between the true and imputed (dropout) data for a dataset with a 65% dropout rate, comprising 20,000 genes and 30,000 cells.

Methods	Correlation(gene)	Correlation(cell)
CPARI	**0.4269**	**0.9221**
ALRA	0.0360	0.7053
SAVER	*	*
scImpute	*	*
bayNorm	*	*
VIPER	*	*
scRecover	*	*
MAGIC	0.2075	0.9029
DeepImpute	0.3763	0.9089
GE-Impute	0.3129	0.5984
DCA	0.0261	0.8891
CL–Impute	*	*
TsImpute	*	*

*: The returned result either exceeds 24 hours or causes a memory overflow.

3. Supplementary evaluation metrics

To conduct a thorough evaluation of the accuracy of dropout zero identification, it is essential to utilize simulated datasets encompassing both complete datasets and dropout datasets. We delineate two distinct categories of zero expression values:

- Real biological zero: An expression value of zero observed in both the complete dataset and the corresponding dropout dataset.
- Dropout zero: An expression value of zero observed in the dropout dataset, but a non-zero value observed in the corresponding complete dataset.

3.1 Standard F1 Score

To assess a model's effectiveness in classification tasks, especially those with imbalanced datasets, F1 Score is a widely used metric . It considers both precision and recall, providing a balanced evaluation of the model's performance.

- True Positive (TP): Correctly imputed dropout zeros.
- True Negative (TN): Real biological zeros left un-imputed (correctly classified).
- False Positive (FP): Real biological zeros mistakenly imputed.
- False Negative (FN): Dropout zeros that the model failed to impute.

The F1 score is calculated as follows:

Standard F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Where:

- **Precision** is the proportion of correctly imputed dropout zeros among all imputed values (including false positives), calculated as:

$$\text{Precision} = TP / (TP + FP)$$

- **Recall** is the proportion of dropout zeros correctly imputed by the model compared to the total number of actual dropout zeros, calculated as:

$$\text{Recall} = TP / (TP + FN)$$

Additionally, **accuracy** can be defined as:

$$\text{Accuracy} = (TP+TN)/ (TP+TN+FP+FN)$$

Accuracy measures the overall proportion of correctly processed zeros, including both dropout zeros that were correctly imputed and real biological zeros that were correctly left un-imputed.

3.2 Recovery data metrics

To quantitatively assess the recovery of missing biosignals in scRNA-seq data, we employed three metrics:

- **Correlation (gene):** The Pearson correlation coefficients (PCC) between the imputed gene expression values and the complete gene expression values. It measures the linear relationship between the gene expression levels across samples in the imputed data and the complete data.
- **Correlation (cell)**: The Pearson correlation coefficients (PCC) between the imputed cellular gene expression values and the complete cellular gene expression values.
- **Error**: The mean squared error (MSE) between the imputed data and the complete data.

The formulations of the three metrics are as follows:

.

$$\text{Correlation(cell)} = \frac{\sum_{i=1}^{m}\left(X(g_i,c_j)-u(X(c_j))\right)\left(Y(g_i,c_j)-u(Y(c_j))\right)}{\sqrt{\sum_{i=1}^{m}\left(X(g_i,c_j)-u(X(c_j))\right)^2\sum_{j=1}^{n}\left(Y(g_i,c_j)-Y(X(c_j))\right)^2}}$$

$$\text{Error} = \frac{1}{n\,\text{m}}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X(g_i,c_j)-Y(g_i,c_j)\right)^2$$

where $X(g_i,c_j)$ represents the imputed expression value of gene g_i in cell c_j within the dropout dataset X, $Y(g_i,c_j)$ represents the expression value of gene g_i in cell c_j within the complete dataset Y, $u(X(g_i)$ represents the mean value of gene g_i across all cells within the dropout dataset X. Here, m represents the number of genes, and n represents the number of cells.

3.3 Consistency metrics

To assess the dissimilarity between correlation matrices derived from complete and dropout datasets, we employed the Gene Correlation Matrix Distance (**CMD (gene)**) and the Cell Correlation Matrix Distance (**CMD (cell)**). These metrics quantify the difference in the correlation structures between the two types of data.

$$\text{CMD(gene)} = \frac{trace(R1\ R2)}{norm(R1)\ norm(R2)}$$

$$\text{CMD(cell)} = \frac{trace(R3\ R4)}{norm(R3)\ norm(R4)}$$

where R1 denotes the gene correlation matrix computed for the complete dataset, R2 denotes the gene correlation matrix computed for the dropout dataset, R3 denotes the cell correlation matrix computed for the complete dataset and R4 denotes the cell correlation matrix computed for the complete dataset. Here, the function $trace(\bullet)$ calculates the trace of the two correlation matrices, and $norm(\bullet)$ represents the Frobenius norm. Similarly, we obtained CMD (cell).

3.4 Modified F1 Score

The complete dataset is divided into distinct subtypes based on the subtype variable. To identify differentially expressed genes (DEGs) between subtypes, we employed the Wilcoxon rank-sum test [1]. Multiple testing correction, specifically False Discovery Rate (FDR) correction, was applied to control for the number of false positives. Genes with a corrected p-value below 0.05 were considered DEGs and served as the gold standard for subsequent analyses [2]. After imputing missing data in the dropout dataset, DEGs were identified using the same methodology. To evaluate the imputation quality, precision and recall were calculated:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

where True Positives (TP) represent the DEGs in the imputed dataset that are also present in the gold standard set, False Positives (FP) denote the DEGs in the imputed dataset that are not present in the gold standard set, and False Negatives (FN) indicate the DEGs in the gold standard set that are not identified in the imputed dataset. The modified

F1 Score, a composite performance metric, integrates both Precision and Recall and is expressed as:

Modified F1 Score = 2 * (Precision* Recall) / (Precision+ Recall)

3.5 Clustering metrics

Analysis of cellular heterogeneity is also one of the main applications of scRNA-seq data, which supports biologists in understanding tissue formation and disease development [3, 4]. For the original single-cell data matrix $X \in R^{m \times n}$, Adjusted Rand Index (ARI) [3] for the given real label set L = $\{l_1, l_2, l_i \ldots l_n\}$ and the predicted label set U =$\{u_1, u_2, u_j \ldots u_n\}$ is calculated as:

$$\boldsymbol{ARI} = \frac{\sum_{i,j}\binom{n_{i,j}}{2} - \left[\sum_{\mathrm{i}}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}\right]/\binom{n}{2}}{\left[\sum_{\mathrm{i}}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}\right]/2 - \left[\sum_{\mathrm{i}}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}\right]/\binom{n}{2}}$$

where $n_{i,j}$ is the number of pairs of elements that are in the same cluster in L and in the same cluster in U, a_i is the number of elements in cluster l_i in L, b_j is the number of elements in cluster u_j in U, and the symbol (.) denotes the binomial coefficient.

Normalized Mutual Information (NMI) [5] measures the mutual dependence between two clustering results while taking into account the differences in their sizes, calculated as:

$$\boldsymbol{NMI} = \frac{MI(L,U)}{\sqrt{H(L)H(U)}}$$

$$MI(L,U) = \sum_{\mathrm{i}}\sum_{j} P(l_i \cap u_j)\, log\, \frac{P(l_i \cap u_j)}{P(l_i)P(u_j)}$$

$$H(L) = -\sum_{\mathrm{i}} P(l_i)\,\mathrm{log}P(l_i)$$

$$H(U) = -\sum_{j} P(u_j)\,\mathrm{log}P(u_j)$$

where $P(l_i \cap u_j)$ represents the probability that a sample belongs to both cluster l_i in L and cluster u_j in U. $P(l_i)$ and $P(u_j)$ are the probabilities of cluster l_i and cluster u_j, respectively.

3.6 Trajectory inference metrics

Trajectory inference is a crucial downstream analysis task for scRNA–seq data [6]. To evaluate CPARI's ability to facilitate accurate trajectory inference, we utilized SCORPIUS [7] on the LPS dataset, which comprises cells sampled across four time points (hour 1, 2, 4, and 6). SCORPIUS was used to reconstruct cell trajectories for the imputed data generated by each method. The Pseudo-Time Ordering Score (POS) metric was employed to assess the consistency between the inferred cell pseudo-time and the actual temporal progression. The **POS** metric has been demonstrated as a reliable indicator for evaluating the effectiveness of cell trajectory analysis methods [8, 9].

By comparing the POS scores obtained with CPARI-imputed data to those obtained with other imputation methods, we can assess CPARI's impact on the accuracy of trajectory inference.

References

[1] Datta S, Satten GA. Rank-sum tests for clustered data. J A m S t a t Assoc 2005; 100:908–15.

[2] Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289-300.

[3] Zheng R, Li M, Liang Z, et al SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation. Bioinformatics 2019;35: 3642–50.

[4] Liang Z, Li M, Zheng R, et al SSRE: cell type detection based on sparse subspace representation and similarity enhancement.Genomics Proteomics Bioinformatics 2021;19:282–91.

[5] Yan X, Zheng R, Li M. GLOBE: a contrastive learning-based framework for integrating single-cell transcriptome datasets. Brief Bioinform 2022;23: bbac311.

[6] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al The single-cell transcriptional landscape of mammalian organogenesis. Nature, 566(7745):496–502, 2019.

[7] Cannoodt, Robrecht, et al. "SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development." Cold Spring Harbor Laboratory (2016).

[8] Zhicheng Ji and Hongkai Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. Nucleic acids research, 44(13): e117–e117, 2016.

[9] Yanglan Gan, Ning Li, Cheng Guo, Guobing Zou, Jihong Guan, and Shuigeng Zhou. Tic2d: trajectory inference from single-cell rnaseq data using consensus clustering. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(4):2512–2522, 2021.