Supplementary Material for Enhancing Gene Expression Prediction from Histology Images with Spatial Transcriptomics Completion

Gabriel Mejia*[0000−0003−4382−6390], Daniela Ruiz*[0000−0001−6636−1173], Paula Cárdenas[0009−0005−1185−548X], Leonardo Manrique[0009−0008−9428−6009], Daniela Vega[0009−0002−9731−7591], and Pablo Arbelaez[0000−0001−5244−2407]

[1] Center for Research and Formation in Artificial Intelligence
[2] Universidad de los Andes, Colombia
{gm.mejia,da.ruizl1,p.cardenasg,dl.manrique,
d.vegaa,pa.arbelaez}@uniandes.edu.co

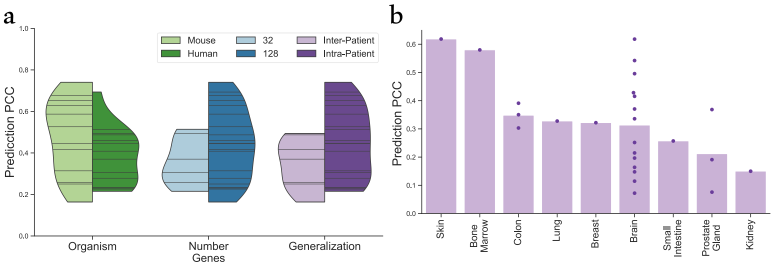

Fig. 1: (a) Violin plots illustrate the variation in key characteristics of the datasets, such as organism, number of genes, and generalization task. The data depicted represent the best prediction PCC achieved for each dataset within the SpaRED collection. (b) Bar charts display the prediction performance, measured by PCC, for each type of tissue analyzed in SpaRED. The dots represent the best prediction PCC achieved for each dataset, while the bars indicate the average PCC across each type of tissue.

Table 1: Detailed overview of the SpaRED datasets, showcasing the generalization task, the number of genes analyzed, the abbreviation for each dataset, the organism studied, the tissue or disease, and the number of slides, patients and spots. Additionally, the table provides the percentages of missing data after processing, labeled as 'Processed Missing Data'. The 'Corrupt Spots' column indicates the percentage of spots with at least one corrupted value for each dataset. Meanwhile, 'Raw Missing Data' represents the percentage of missing data before processing.

Generalization	**Genes**	**Abbreviation/ Access**	**Organism**	**Tissue/ Disease**	**Slides**	**Patients**	**Spots**	**Processed Missing Data**	**Corrupt Spots**	**Raw Missing Data**
Inter-Patient	128	VMB [20]	Mouse	Brain	14	4	43804	28%	100%	89%
		MMBR [12]	Mouse	Brain	8	2	34583	20%	100%	79%
		MHPC [12]	Human	Prostate cancer	4	2	15684	21%	100%	79%
		PMI [15]	Mouse	Intestine	2	2	6234	7%	99%	79%
	32	VLMB [21]	Mouse	Brain	5	2	12202	29%	100%	95%
		MHSI [12]	Human	Small intestine	4	2	10474	21%	100%	92%
		MHCP2 [12]	Human	Colon	2	2	7101	24%	100%	97%
Intra-Patient	128	EHPCP2[5]	Human	Prostate cancer	10	1	24465	37%	100%	92%
		EHPCP1 [5]	Human	Prostate cancer	7	1	20987	34%	100%	92%
		MMBP2 [12]	Mouse	Brain	4	1	17353	25%	100%	88%
		MMBP1 [12]	Mouse	Brain	4	1	17243	11%	100%	70%
		AHSCC [1]	Human	Squamous cell carcinoma	4	1	10374	27%	100%	94%
		10XGHB [18]	Human	Brain	2	1	9882	23%	100%	89%
		FMBC [6]	Mouse	Brain - Coronal	2	1	9132	24%	100%	80%
		10GHBC [18]	Human	Breast cancer	2	1	7785	12%	100%	79%
		MMBO [12]	Mouse	Bone	4	1	7184	17%	100%	87%
		10XGMBSP [18]	Mouse	Brain sagittal posterior	2	1	6644	21%	100%	79%
		MHPBTP1 [12]	Human	Pediatric brain tumor	4	1	5937	21%	100%	73%
		10XGMBC [18]	Mouse	Brain coronal	2	1	5709	18%	100%	80%
		10XGMBSA [18]	Mouse	Brain sagittal anterior	2	1	5520	15%	100%	75%
		FMOB [6]	Mouse	Brain	2	1	2938	10%	100%	67%
		VLO [21]	Human	Lung organoids	4	1	1832	26%	100%	90%
		VKO [21]	Human	Kidney organoids	3	1	1355	33%	100%	92%
	32	VHS [20]	Human	Striatium	4	1	19033	30%	100%	97%
		MHPBTP2 [12]	Human	Pediatric brain tumor	2	1	3163	30%	100%	97%
		MHCP1 [12]	Human	Colon	2	1	2225	16%	100%	90%

Table 2: The matrix illustrates the statistically significant differences in MSE among all models across all datasets. A Dunn test with a 5% significance level was used to identify these differences. Differences that are statistically significant between models are highlighted in **bold**.

	ShuffleNet	**STNet**	**EGN**	**EGGN**	**HisToGene**	**ViT-B**	**VIT-B+Δ**	**SEPAL**	**BLEEP**
ShuffleNet	-	1	1	1	**3.01e-4**	1	0.6028	1	**0.0023**
STNet		-	1	1	**2.03e-8**	1	1	1	**3.02e-7**
EGN			-	1	**1.21e-3**	1	0.2465	0.861	**8.16e-3**
EGGN				-	**8.75e-5**	1	1	1	**7.37e-4**
HisToGene					-	**2.20e-5**	**3.17e-10**	**6.24e-9**	1
ViT-B						-	1	1	**2.09e-4**
VIT-B+Δ							-	1	**5.92e-9**
SEPAL								-	**9.91e-8**
BLEEP									-