Helical Coding challenge

In-silico perturbation predictions on the example of ALS

Dr. Maren Büttner, 6/11/2025

Task 1: Design an In-Silico Perturbation Workflow

Input: Anndata object

Model: Pre-trained single-cell foundation model on the basis of GeneFormer

Workflow:

- Process anndata object (count data) with helical Python package
- In-silico perturbation for knockdown and knockup experiments:
  1. Compute embeddings of gene expression matrix to obtain original embeddings
  2. Multiply gene expression values for selected gene(s) by a factor c (>1 for knockup and <1 for knockdown) or set to 0 for full knockout
  3. Compute embedding of perturbed data matrix

Task 2: Apply Perturbations to Disease-Specific Genes

Input: anndata object (count data) from ALS patients and healthy controls (brain)

Experiments:

1. Single-gene perturbation

2. Pathway perturbation

Knockup with c = 2 all on healthy controls on all cell types.

Previously reported genes linked to ALS: SOD1, ANXA11, ARPP21, CAV1, C21ORF2, CCNF, DNAJC7, GLT8D1, KIF5A, NEK1, SPTLC1, **TIA1, TARDBP, FUS, C9orf72** and WDR7 (bold genes are part of the RNA metabolism and considered part of the same pathway and used to exemplify the analysis)

References: Wang et al. Frontiers in Neuroscience (2023), Pineda et al. Cell (2024), Marlin et al. BJP (2023), Flores et al. Nature Communications (2025)

Task 3: Interpret the Embedding Space

- **Metrics and diagnostic:**

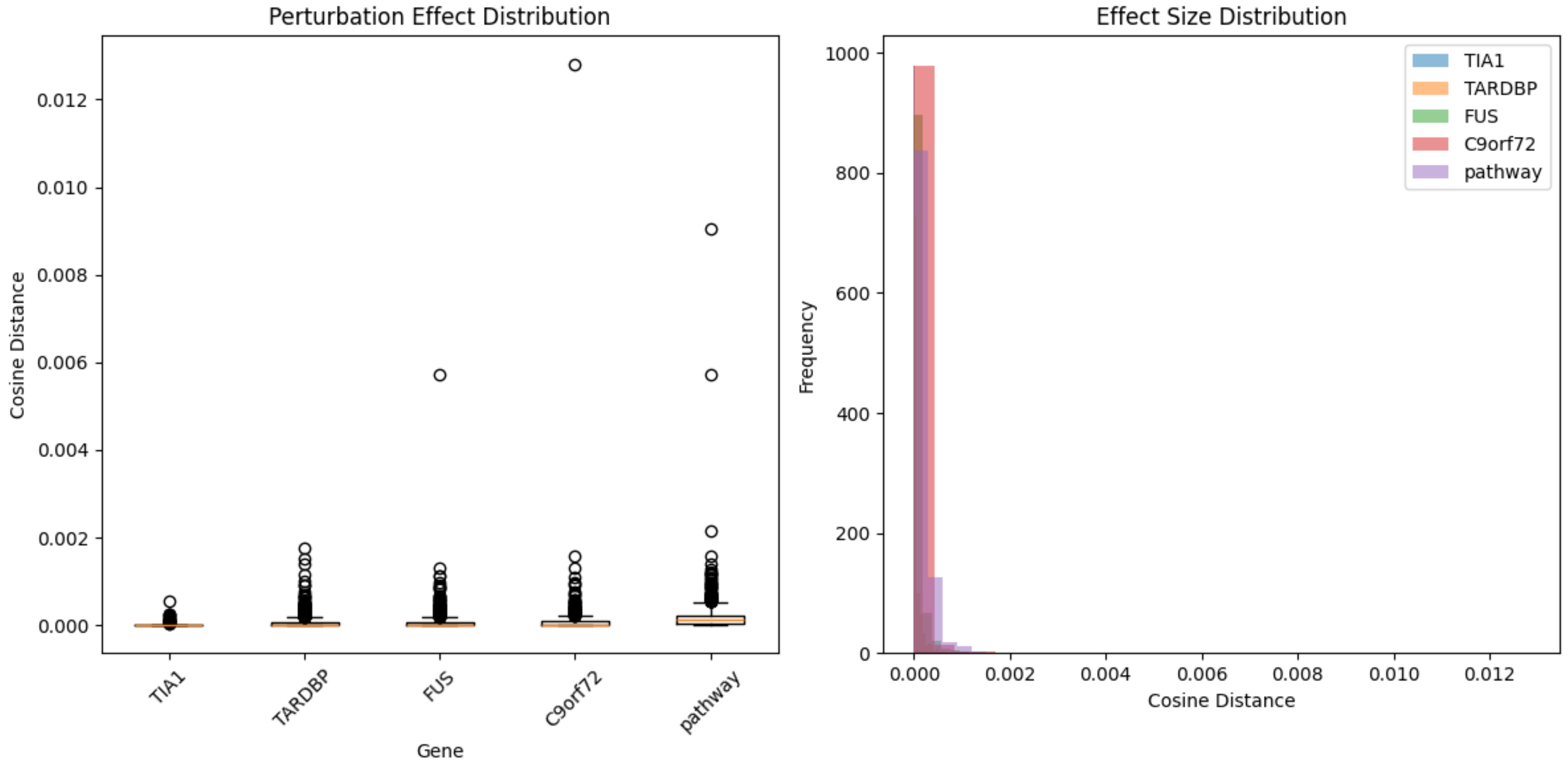
  1. Effect size distribution as the distance of perturbed cells to original cells in embedding space (cosine similarity or Euclidean distance) per perturbation (overall and by cell type)

  2. Visualisation of original and perturbed cells on a UMAP

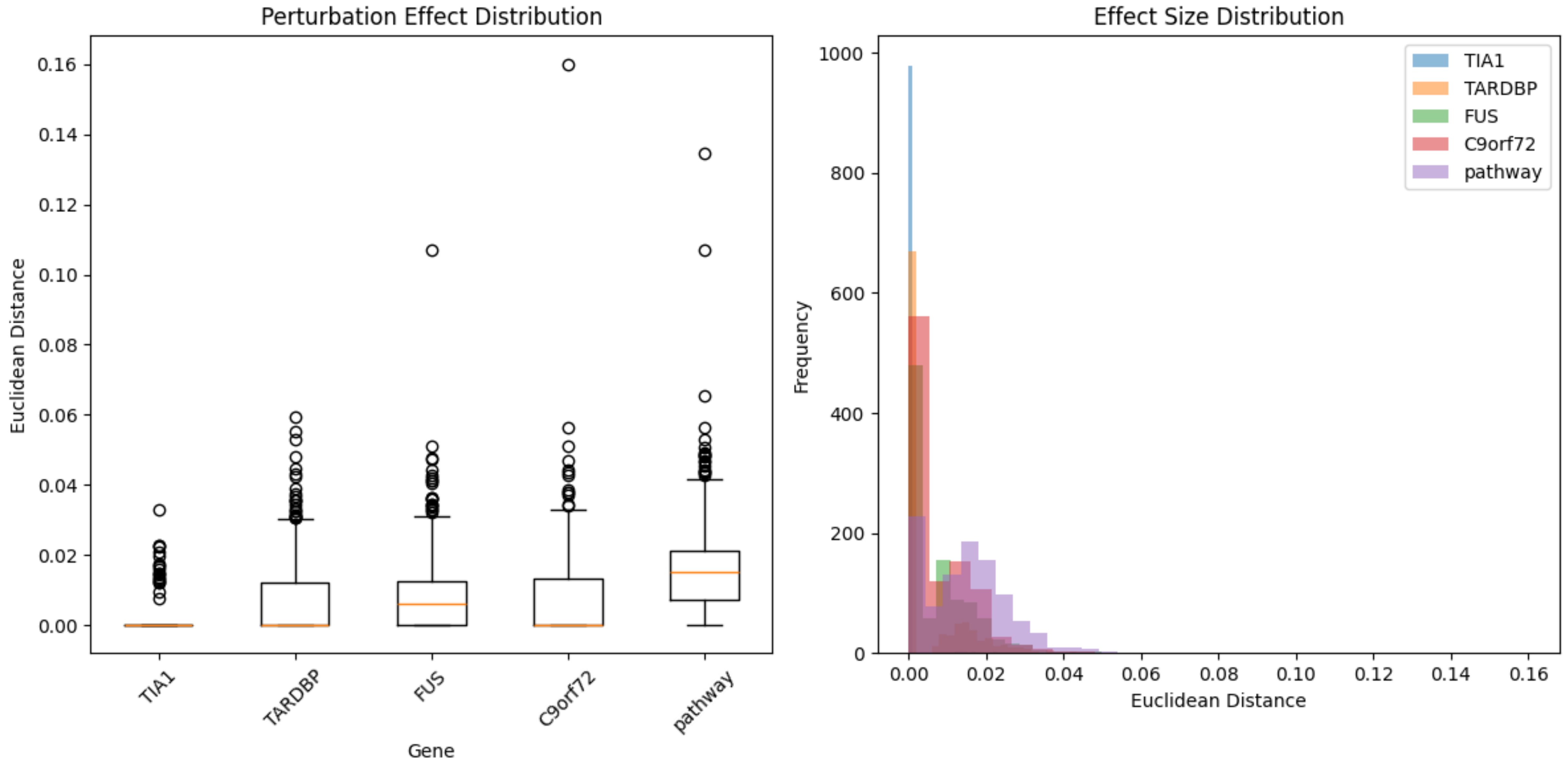
  3. Neighbourhood mixing analysis using a chi-square test per cell

Task 3: Interpret the embedding space

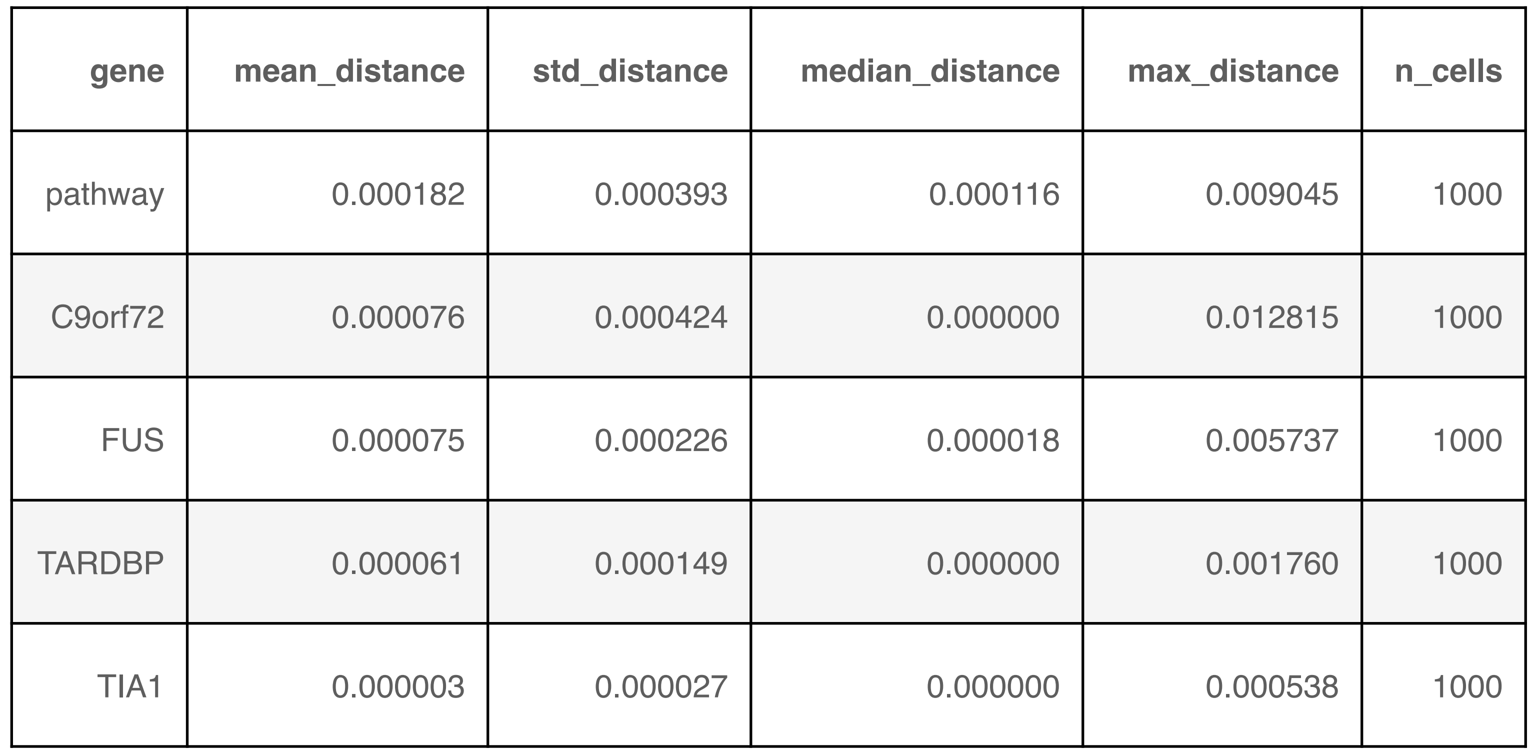
Metric: cosine similarity



Metric: Euclidean distance



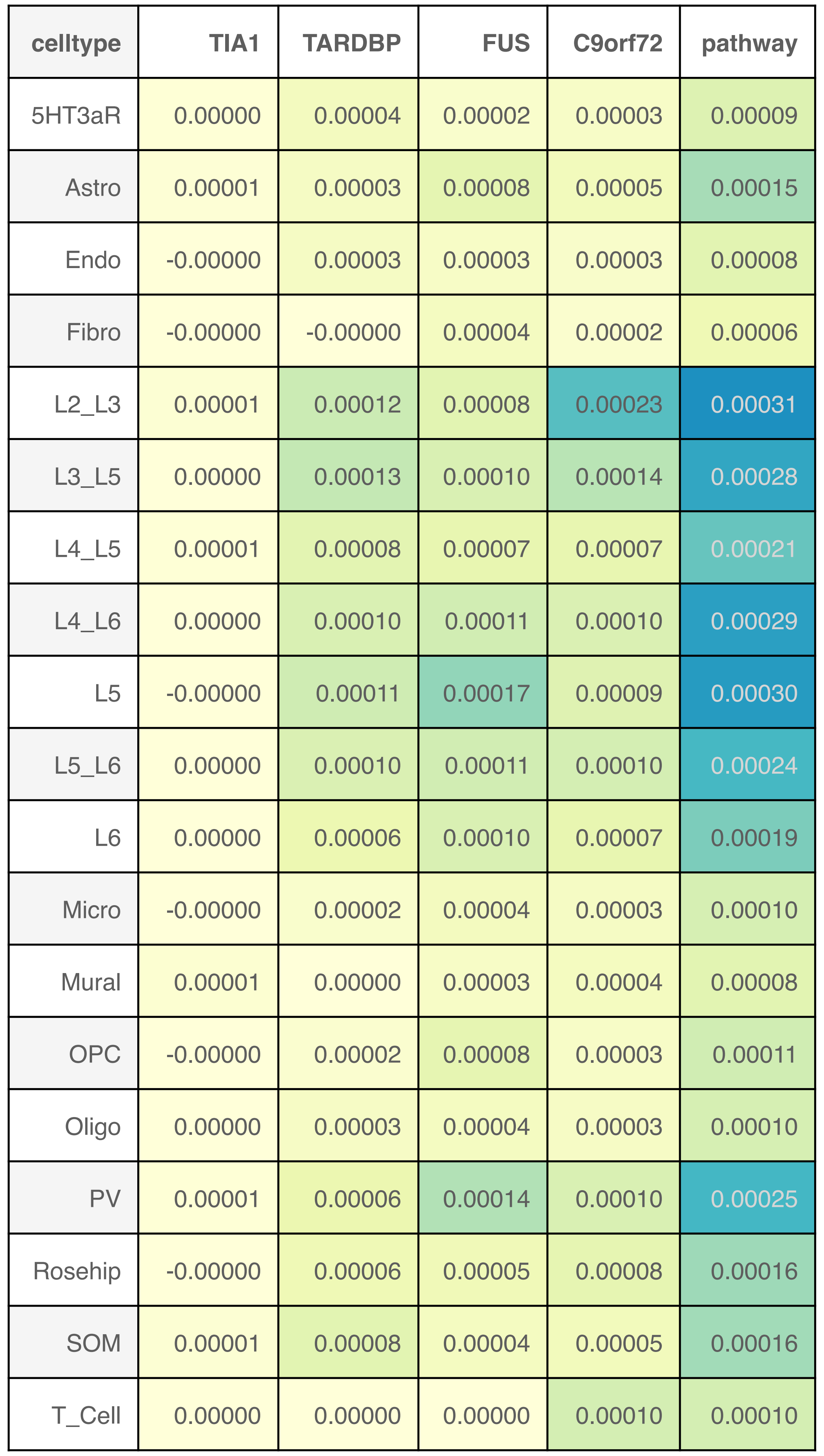
Ranking of perturbations:



gene	mean_distance	std_distance	median_distance	max_distance	n_cells
pathway	0.000182	0.000393	0.000116	0.009045	1000
C9orf72	0.000076	0.000424	0.000000	0.012815	1000
FUS	0.000075	0.000226	0.000018	0.005737	1000
TARDBP	0.000061	0.000149	0.000000	0.001760	1000
TIA1	0.000003	0.000027	0.000000	0.000538	1000

Task 3: Interpret the embedding space

Distance by cell type:

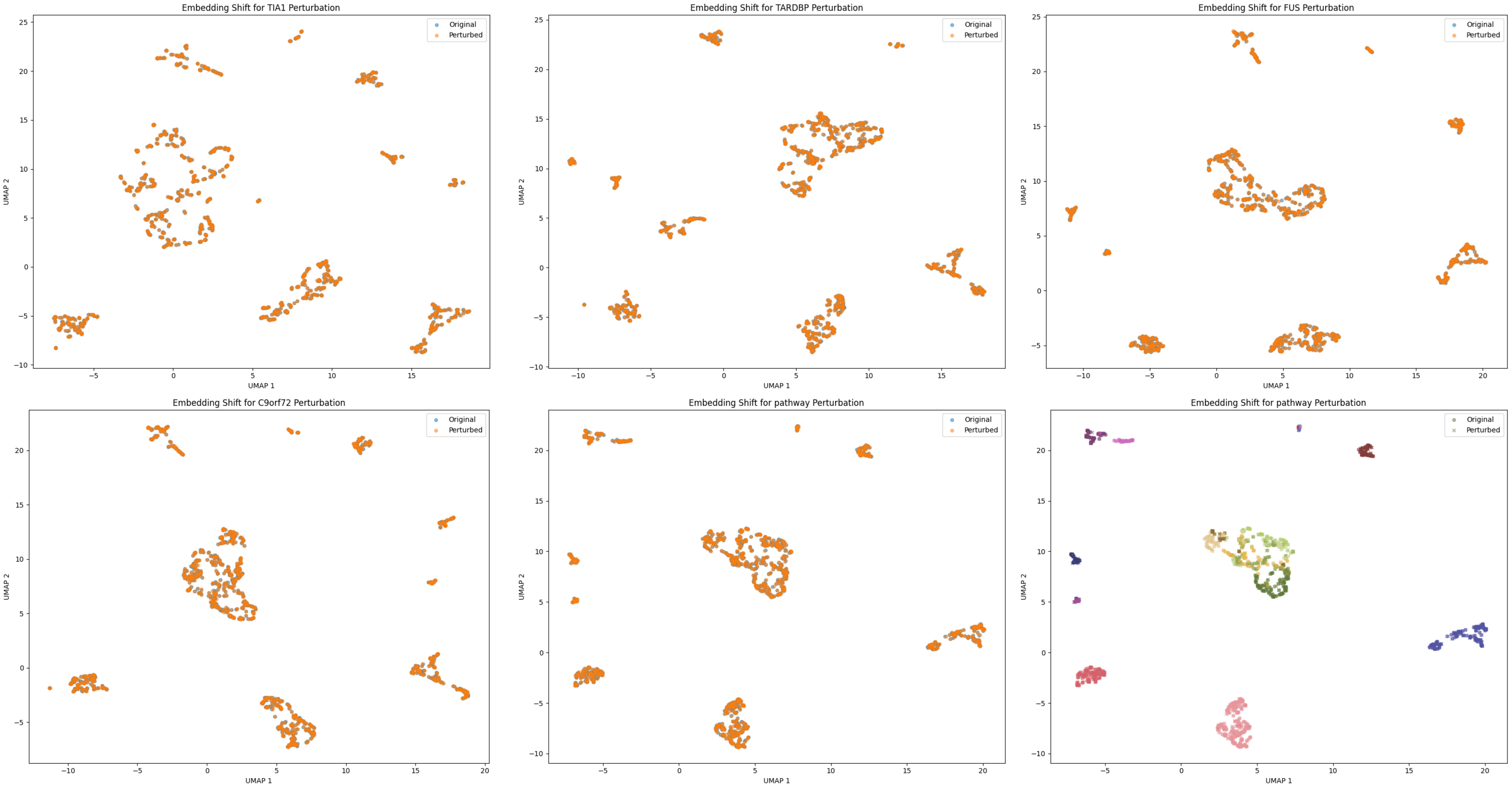


celltype	TIA1	TARDBP	FUS	C9orf72	pathway
5HT3aR	0.00000	0.00004	0.00002	0.00003	0.00009
Astro	0.00001	0.00003	0.00008	0.00005	0.00015
Endo	-0.00000	0.00003	0.00003	0.00003	0.00008
Fibro	-0.00000	-0.00000	0.00004	0.00002	0.00006
L2_L3	0.00001	0.00012	0.00008	0.00023	0.00031
L3_L5	0.00000	0.00013	0.00010	0.00014	0.00028
L4_L5	0.00001	0.00008	0.00007	0.00007	0.00021
L4_L6	0.00000	0.00010	0.00011	0.00010	0.00029
L5	-0.00000	0.00011	0.00017	0.00009	0.00030
L5_L6	0.00000	0.00010	0.00011	0.00010	0.00024
L6	0.00000	0.00006	0.00010	0.00007	0.00019
Micro	-0.00000	0.00002	0.00004	0.00003	0.00010
Mural	0.00001	0.00000	0.00003	0.00004	0.00008
OPC	-0.00000	0.00002	0.00008	0.00003	0.00011
Oligo	0.00000	0.00003	0.00004	0.00003	0.00010
PV	0.00001	0.00006	0.00014	0.00010	0.00025
Rosehip	-0.00000	0.00006	0.00005	0.00008	0.00016
SOM	0.00001	0.00008	0.00004	0.00005	0.00016
T_Cell	0.00000	0.00000	0.00000	0.00010	0.00010

Cell type specific impact of the knockup perturbation in healthy cells:

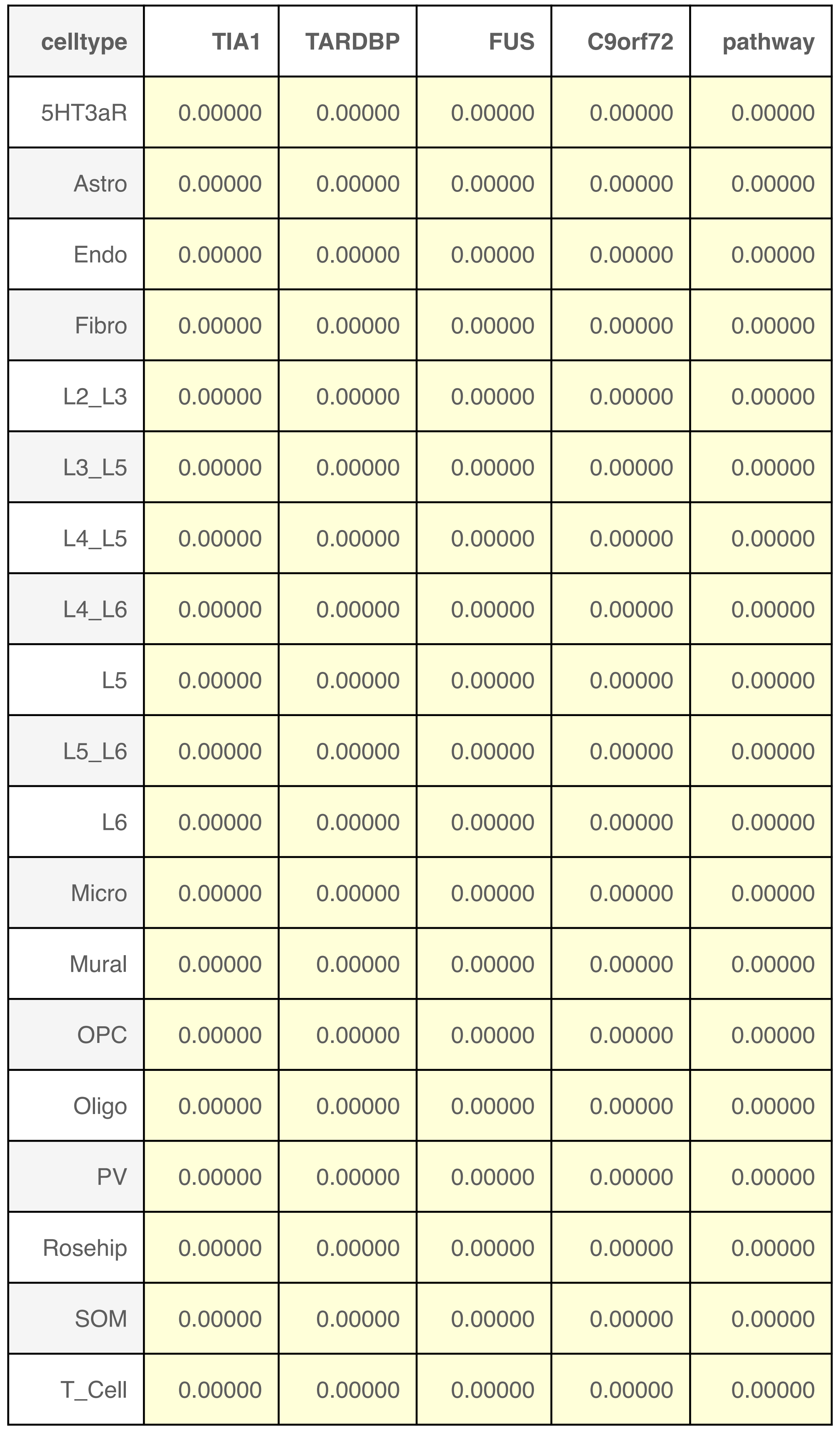
- *TARDBP* and *C9orf72* affect L2_L3 and L3_L5 upper motor neurons (UMNs) the strongest
- *TIA1* did not seem to affect any of the cell types in particular
- FUS over-expression affects L5 UMNs and PV inhibitory neurons the strongest
- perturbation of all four RNA metabolism genes predicts the strongest changes in L2 - L6 UMNs as well as neocortical inhibitory neurons (PV, Rosehip and SOM) and a slightly smaller change in astrocytes

Task 3: Interpret the embedding space



Task 3: Interpret the embedding space

Percentage of cells in biased neighbourhoods:



celltype	TIA1	TARDBP	FUS	C9orf72	pathway
5HT3aR	0.00000	0.00000	0.00000	0.00000	0.00000
Astro	0.00000	0.00000	0.00000	0.00000	0.00000
Endo	0.00000	0.00000	0.00000	0.00000	0.00000
Fibro	0.00000	0.00000	0.00000	0.00000	0.00000
L2_L3	0.00000	0.00000	0.00000	0.00000	0.00000
L3_L5	0.00000	0.00000	0.00000	0.00000	0.00000
L4_L5	0.00000	0.00000	0.00000	0.00000	0.00000
L4_L6	0.00000	0.00000	0.00000	0.00000	0.00000
L5	0.00000	0.00000	0.00000	0.00000	0.00000
L5_L6	0.00000	0.00000	0.00000	0.00000	0.00000
L6	0.00000	0.00000	0.00000	0.00000	0.00000
Micro	0.00000	0.00000	0.00000	0.00000	0.00000
Mural	0.00000	0.00000	0.00000	0.00000	0.00000
OPC	0.00000	0.00000	0.00000	0.00000	0.00000
Oligo	0.00000	0.00000	0.00000	0.00000	0.00000
PV	0.00000	0.00000	0.00000	0.00000	0.00000
Rosehip	0.00000	0.00000	0.00000	0.00000	0.00000
SOM	0.00000	0.00000	0.00000	0.00000	0.00000
T_Cell	0.00000	0.00000	0.00000	0.00000	0.00000

Test if the neighbourhoods deviate significantly from the expected 50:50 ratio at significance level alpha < 0.05 and number of nearest neighbours k=15:

- None of the neighbourhoods over-represent the original nor the perturbed condition

Task 3: Interpret the embedding space

Conclusions:

- In-silico perturbation simulation of RNA metabolism genes led to a change in specific neurons, which play a role in the disease progression of ALS
- Pathway perturbation showed expectedly the strongest change
- Effects overall seem very small, not visible on a UMAP embedding; recommend to increase knockup factor
- Strongest individual genes: C9orf72, TARDBP and FUS
- Affected cell types:
  - L2_L3, L3_L5 and L5 UMNs in single-gene perturbations
  - L2 - L6 UMNs, neocortical inhibitory neurons (PV, Rosehip and SOM) and astrocytes in pathway perturbation

Task 4: Prioritize potential drug target genes

- **Experiment:**

  Single-gene perturbation (knockdown c=0.5) in cells from ALS patients

- Metrics and diagnostic similar to task 3

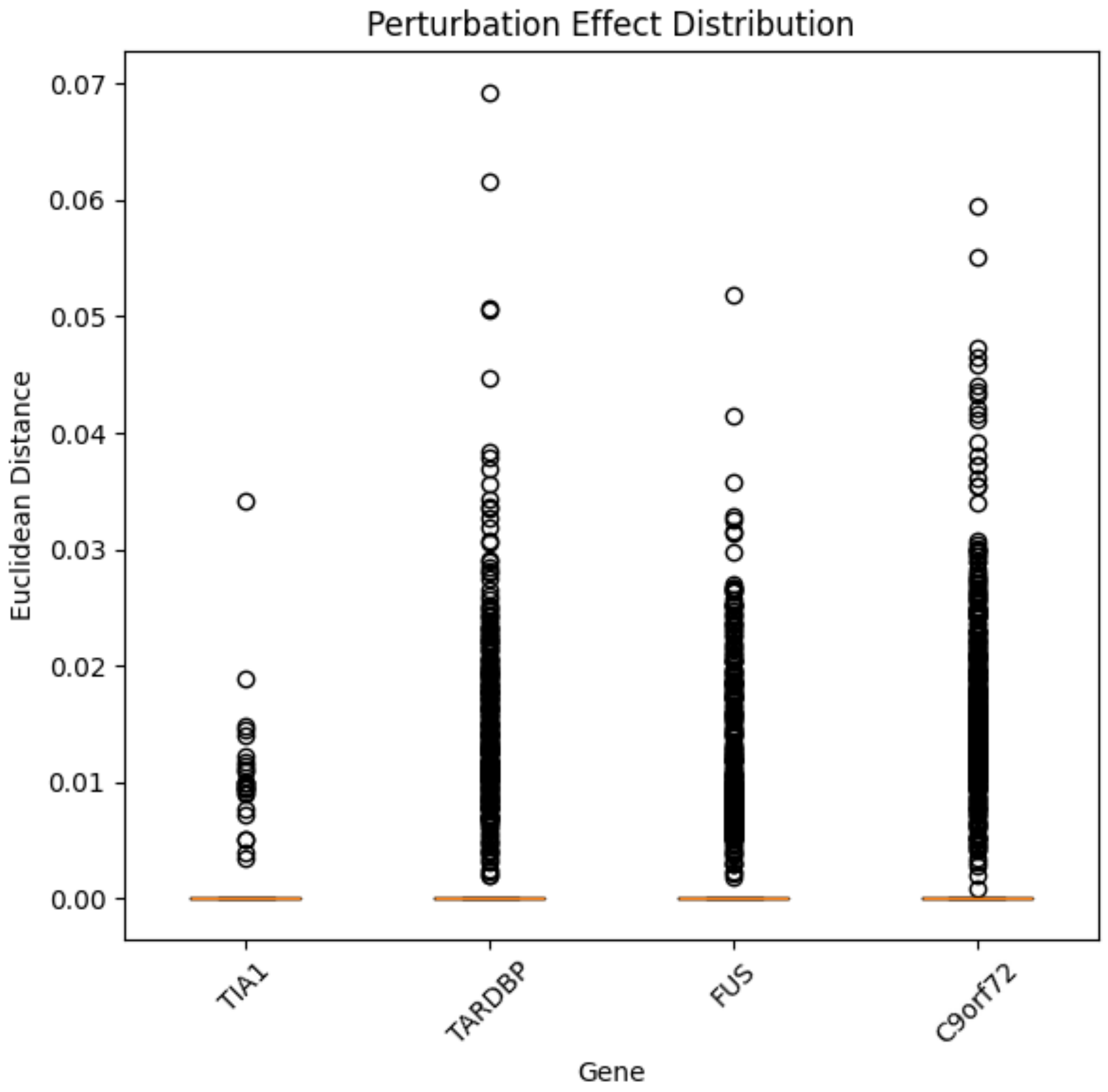
- Tested genes: *TIA1, TARDBP, FUS, C9orf72*

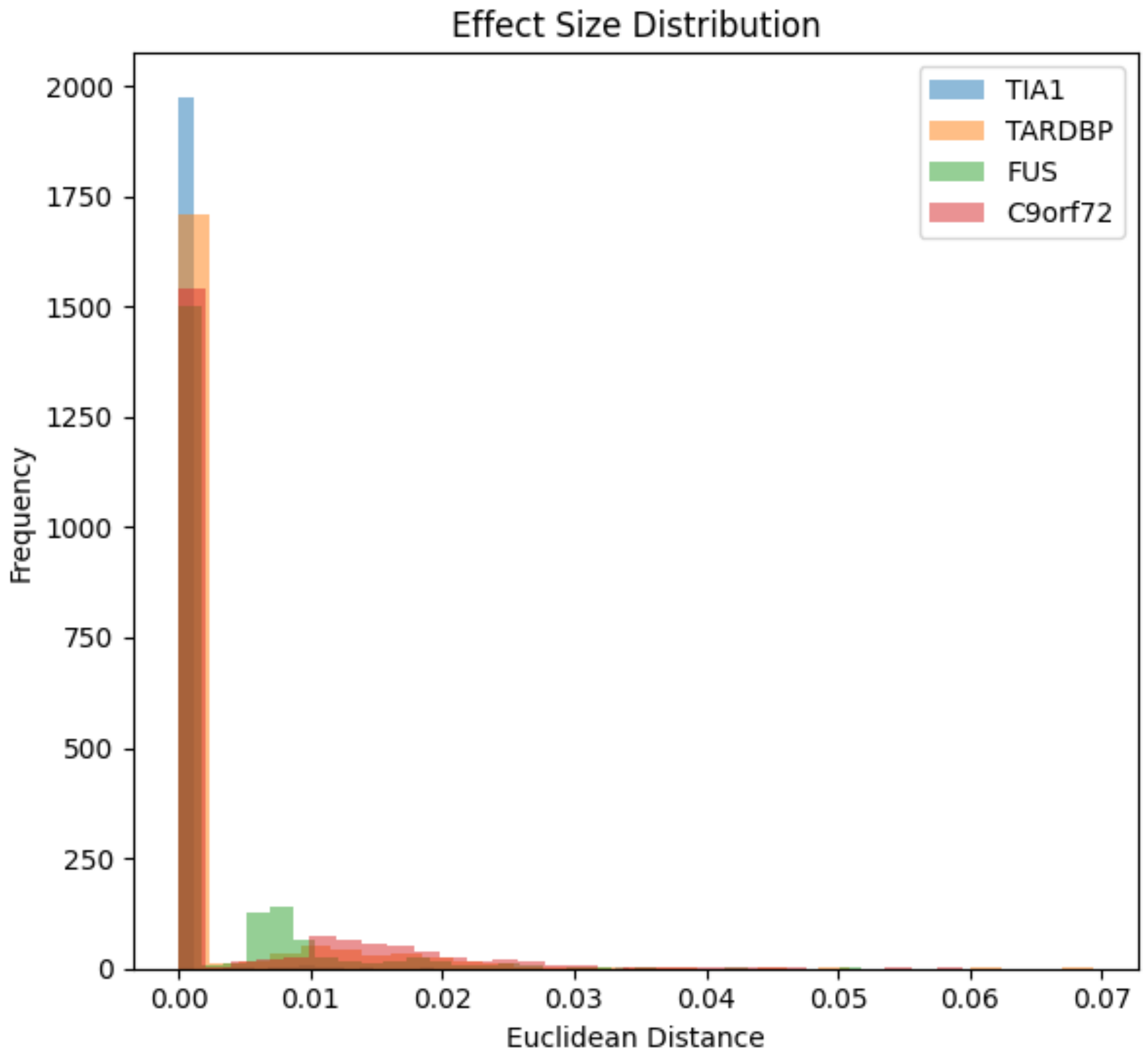
- **Conclusions:**

  - *C9orf72* and *TARDBP* showed strongest perturbation effect in diseased cells; but effect size too small to move them considerably towards the healthy state

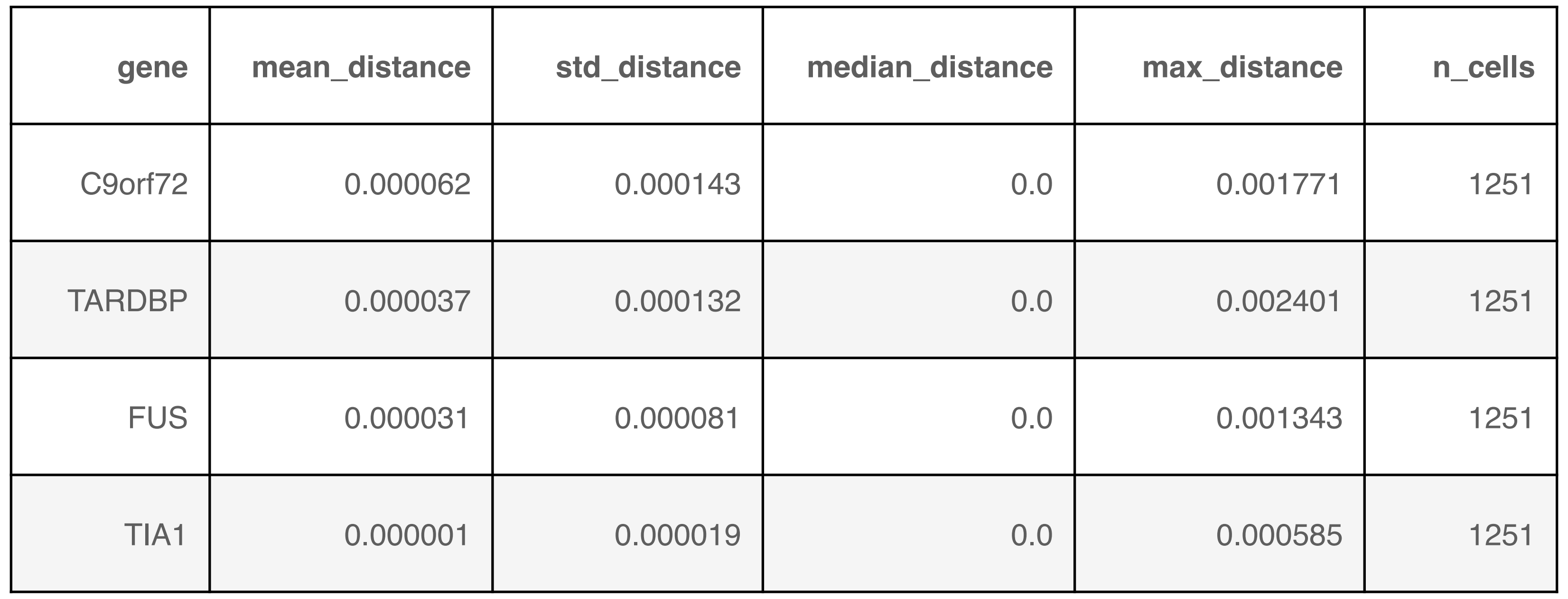
Task 4: Prioritize potential drug target genes

Metric: Euclidean distance





Ranking of effect sizes:

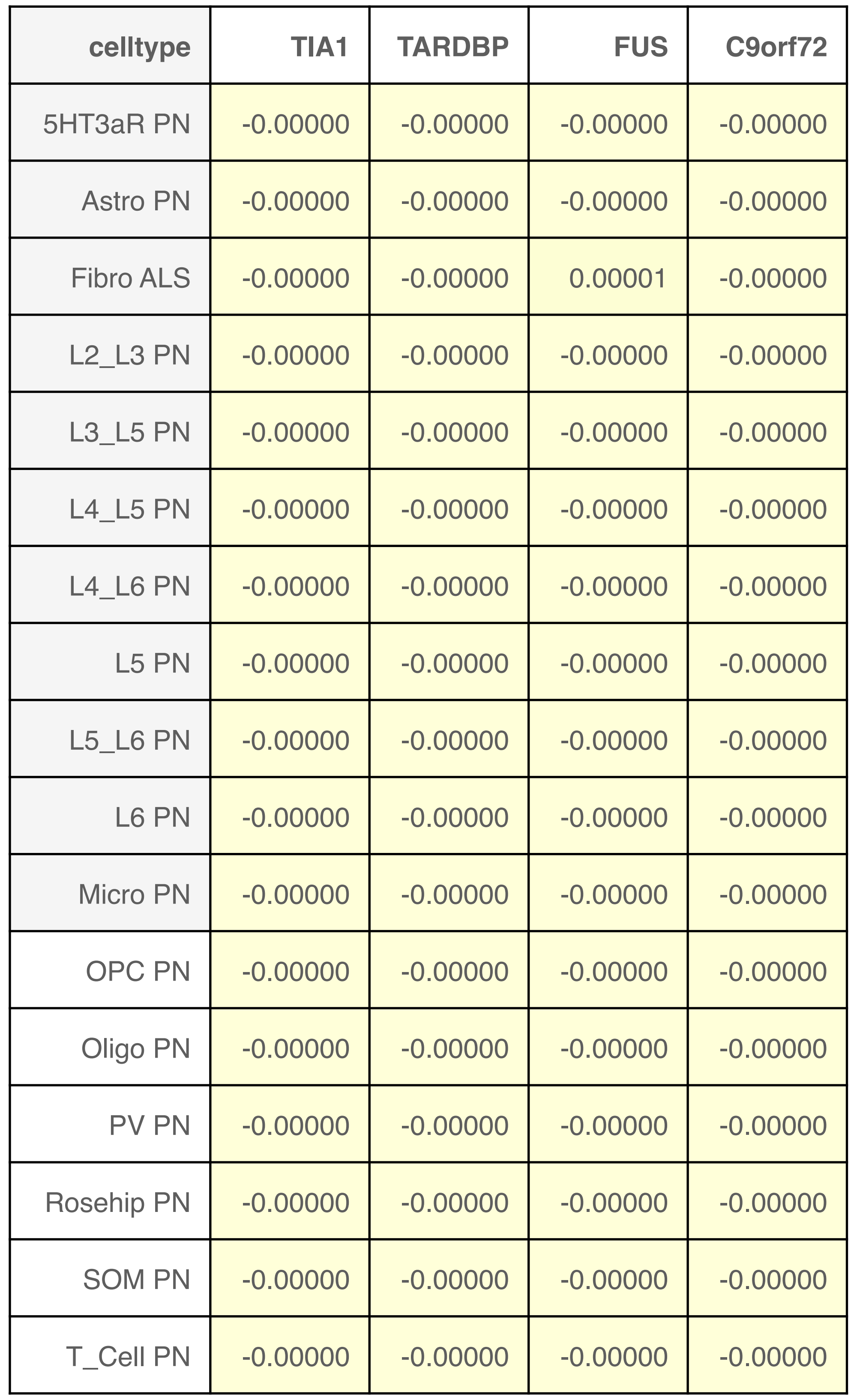


gene	mean_distance	std_distance	median_distance	max_distance	n_cells
C9orf72	0.000062	0.000143	0.0	0.001771	1251
TARDBP	0.000037	0.000132	0.0	0.002401	1251
FUS	0.000031	0.000081	0.0	0.001343	1251
TIA1	0.000001	0.000019	0.0	0.000585	1251

Task 4: Prioritize potential drug target genes

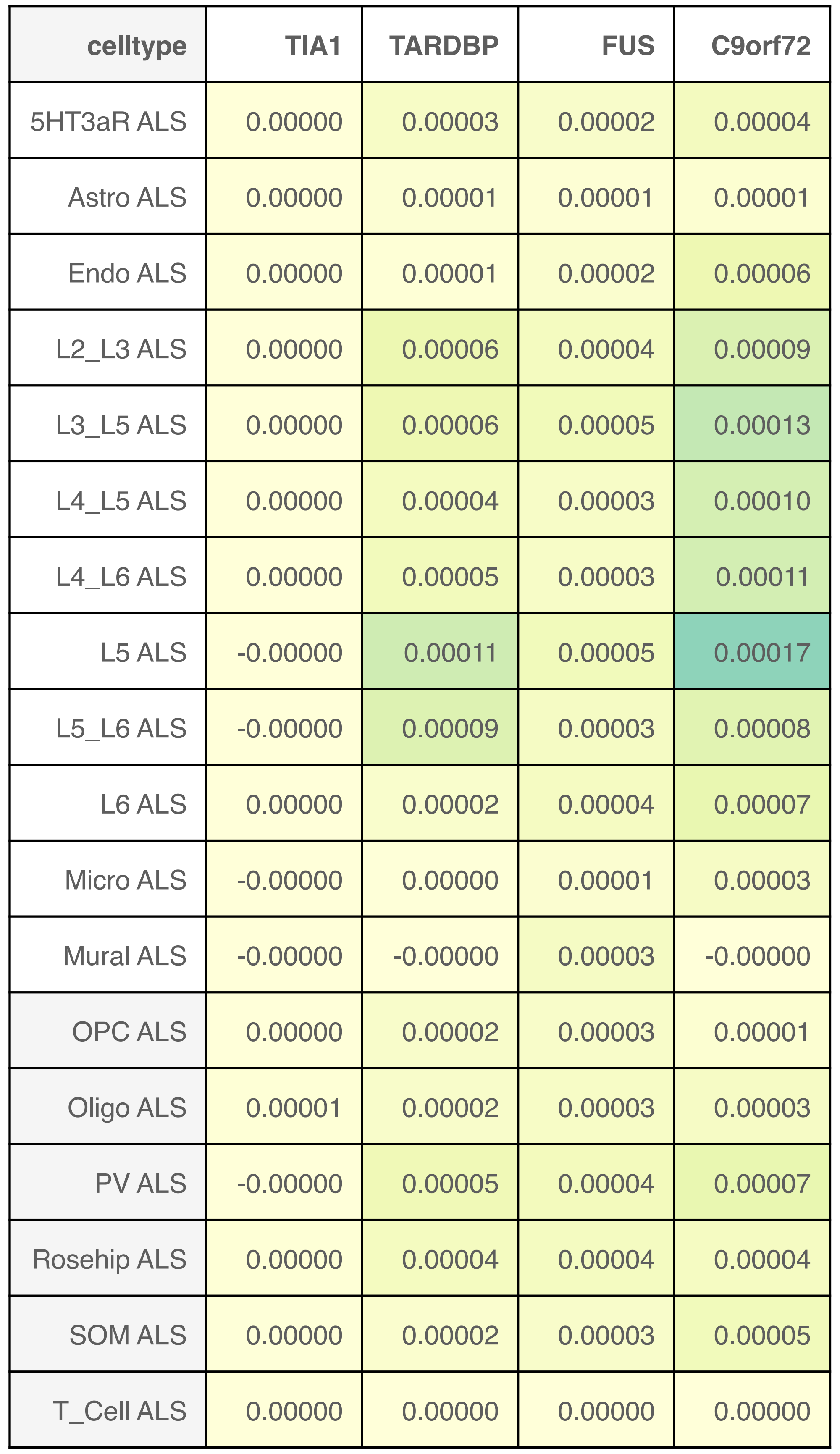
Distance to perturbed condition by cell type and condition:

Healthy controls



celltype	TIA1	TARDBP	FUS	C9orf72
5HT3aR PN	-0.00000	-0.00000	-0.00000	-0.00000
Astro PN	-0.00000	-0.00000	-0.00000	-0.00000
Fibro ALS	-0.00000	-0.00000	0.00001	-0.00000
L2_L3 PN	-0.00000	-0.00000	-0.00000	-0.00000
L3_L5 PN	-0.00000	-0.00000	-0.00000	-0.00000
L4_L5 PN	-0.00000	-0.00000	-0.00000	-0.00000
L4_L6 PN	-0.00000	-0.00000	-0.00000	-0.00000
L5 PN	-0.00000	-0.00000	-0.00000	-0.00000
L5_L6 PN	-0.00000	-0.00000	-0.00000	-0.00000
L6 PN	-0.00000	-0.00000	-0.00000	-0.00000
Micro PN	-0.00000	-0.00000	-0.00000	-0.00000
OPC PN	-0.00000	-0.00000	-0.00000	-0.00000
Oligo PN	-0.00000	-0.00000	-0.00000	-0.00000
PV PN	-0.00000	-0.00000	-0.00000	-0.00000
Rosehip PN	-0.00000	-0.00000	-0.00000	-0.00000
SOM PN	-0.00000	-0.00000	-0.00000	-0.00000
T_Cell PN	-0.00000	-0.00000	-0.00000	-0.00000

ALS



celltype	TIA1	TARDBP	FUS	C9orf72
5HT3aR ALS	0.00000	0.00003	0.00002	0.00004
Astro ALS	0.00000	0.00001	0.00001	0.00001
Endo ALS	0.00000	0.00001	0.00002	0.00006
L2_L3 ALS	0.00000	0.00006	0.00004	0.00009
L3_L5 ALS	0.00000	0.00006	0.00005	0.00013
L4_L5 ALS	0.00000	0.00004	0.00003	0.00010
L4_L6 ALS	0.00000	0.00005	0.00003	0.00011
L5 ALS	-0.00000	0.00011	0.00005	0.00017
L5_L6 ALS	-0.00000	0.00009	0.00003	0.00008
L6 ALS	0.00000	0.00002	0.00004	0.00007
Micro ALS	-0.00000	0.00000	0.00001	0.00003
Mural ALS	-0.00000	-0.00000	0.00003	-0.00000
OPC ALS	0.00000	0.00002	0.00003	0.00001
Oligo ALS	0.00001	0.00002	0.00003	0.00003
PV ALS	-0.00000	0.00005	0.00004	0.00007
Rosehip ALS	0.00000	0.00004	0.00004	0.00004
SOM ALS	0.00000	0.00002	0.00003	0.00005
T_Cell ALS	0.00000	0.00000	0.00000	0.00000

- No change in healthy controls expected because not being perturbed
- Strongest effect by *C9orf72* and *TARDBP* in L2 to L6 UMNs (strongest in L5 UMN)

Task 4: Prioritize potential drug target genes

Overall mixing score:

Gene	Mixing score
TIA1	0,933
TARDBP	0,932
FUS	0,931
C9orf72	0,929

Mixing score m denotes fraction of cells in a 50:50 neighbourhood (0<=m<=1)

- m>0.8 weak effect and m<0.5 strong effect
- Mixing with original (and healthy cells) is very high for all perturbations.
- No further analysis which perturbation resembles healthy state the most.