Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τηλεπικοινωνιών

Multi-task learning in perturbation modeling

Διπλωματική Εργασία
του
Θεόδωρου Κατζάλη

Επιβλέπων: Όνομα Επίθετο
Καθηγητής Α.Π.Θ.

April 3, 2025

Περιεχόμενα

1 Abstract

With the recent advancements in single cell technology and the large scale perturbation datasets, the field of perturbation modeling has created an opportunity for a wide variety of computational methods to be leveraged to harness its potential. Multi-task learning is one of the methods that has been left unexplored in this field. In this study we aim to bridge this gap unraveling the potential of multi-task learning in single cell perturbation modeling.

2 Introduction

The complexity of biological systems have imposed a challenge to capture the underlying mechanism of cellular heterogeneity. Understanding the effect of external stimuli (perturbations) to the cell level, a field named as perturbation modeling [2], has a significant impact in biomedicine and drug discovery. With the recent surge of data generation, machine learning methods aim to understand the effect of perturbations, given a limited number of perturbation experiments.

An overview of the models on perturbation modeling can be found on this study [1]. One of the main objectives is the out-of-distribution detection, which is the focal point of our study. The task is about predicting the perturbation response of the omics signature of cells with a specific cell type, while having observed the perturbation response of other cell types.

UnitedNet is a multi-task framework that has shown its potential in multi-omics tasks such as cross modal prediction and cell type classification. We aim to extend this approach to perturbation modeling.

3 Method

A short intro of multi-task and the rejection of a multi-head architecture. Intro to film layers and explanation of the method.

the usage of film layers

4 Results and discussion

In the literature body, there are several approaches for predicting single cell perturbation responses. To compare our multi-task model, we have chosen the models of scGen, scButterfly, scPregan, and scVIDR.

scGen used a vae that captures the perturbation response using vectors in the perturbation space

scbutterfly with a vae had shown its potential on the perturbation modeling applied on the pbmc dataset. It is based on vae architecture with these characteristics.

scPreGAN.

scVIDR.

To quantify the objective of the task, we have used a multi-faceted benchmarking suite, relying on distance metrics along with the number of differentially expressed genes (DEGs) and the r2 of all genes and the most highly variable genes.

We have tested the models on two datasets, one where human peripheral blood mononuclear cells have been stimulated by IFN-b interferon, and a multi-perturbation dataset, where

liver cells have been stimulated by multiple doses of tetrachlorodibenzo-p-dioxin (TCDD) in vivo.

Regarding the single perturbation response models, the scGen, scButterfly, scPreGAN, in the multi-perturbation dataset of ten dosages, we have trained a dedicated model for each dosage.

To address the randomness of the models, we have performed the experiments three times, with thres different seeds 1, 2, 19193, and the metrics have been averaged across experiments.

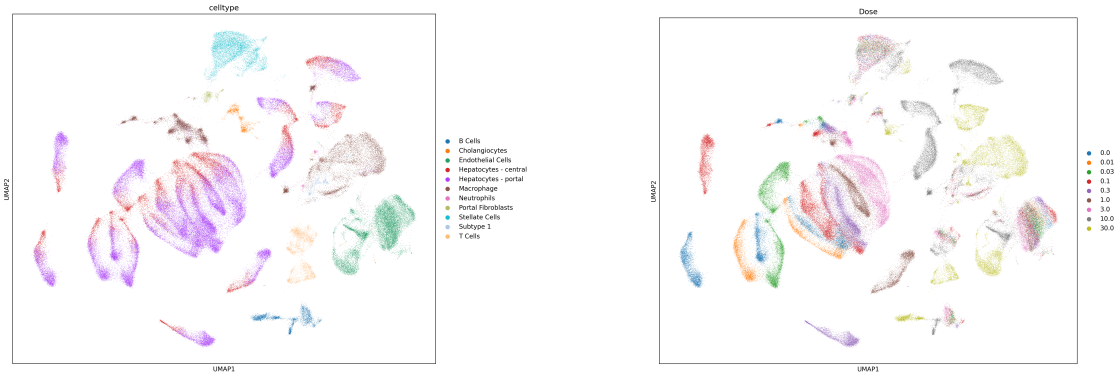
5 Conclusions

6 Future work

7 Benchmarking

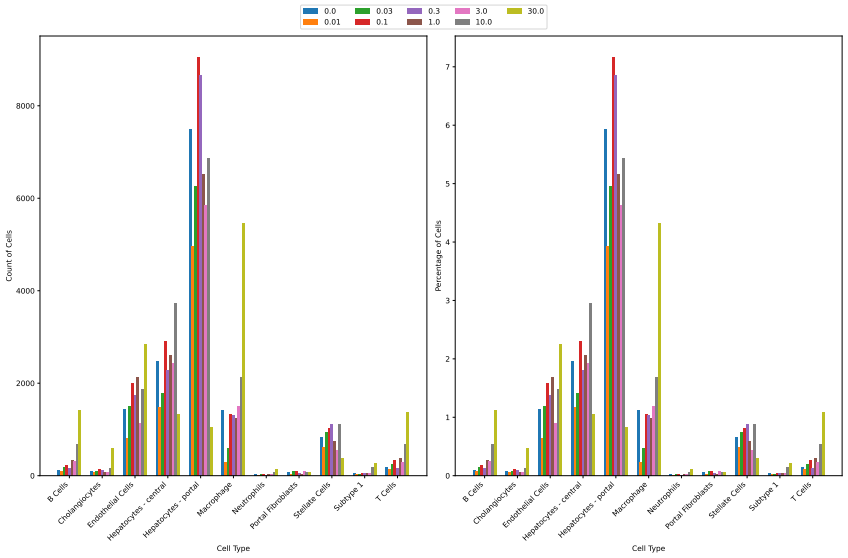
8 Datasets

8.1 Nault et al. 2022

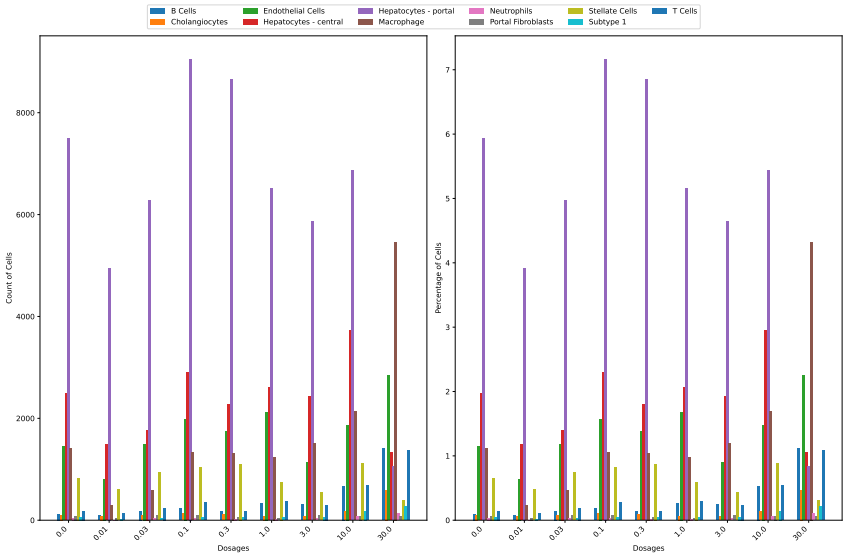


(a) (b)

.



(c)



(d)

Figure 1: Nault overview

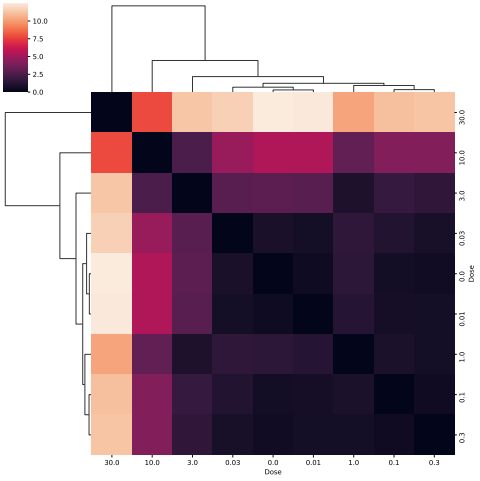


Figure 2: E-distance

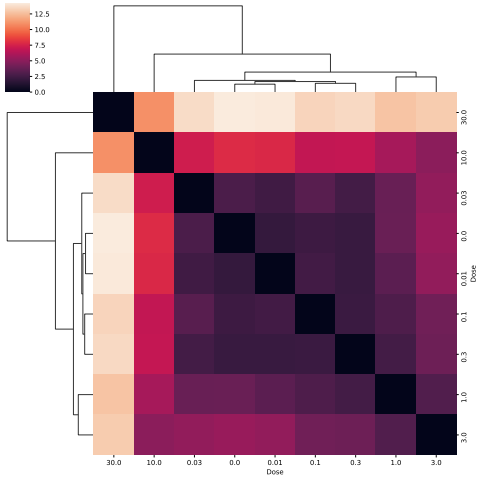


Figure 3: Euclidean

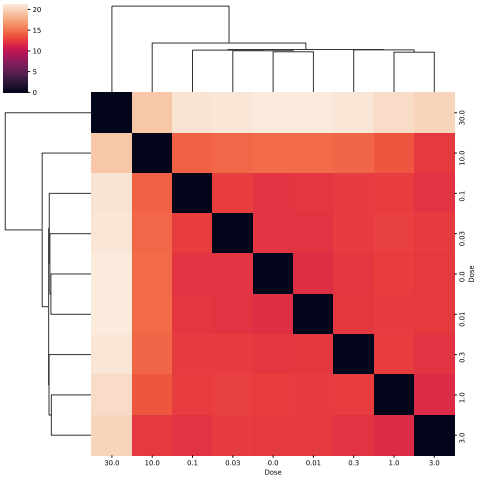


Figure 4: Mean pairwise

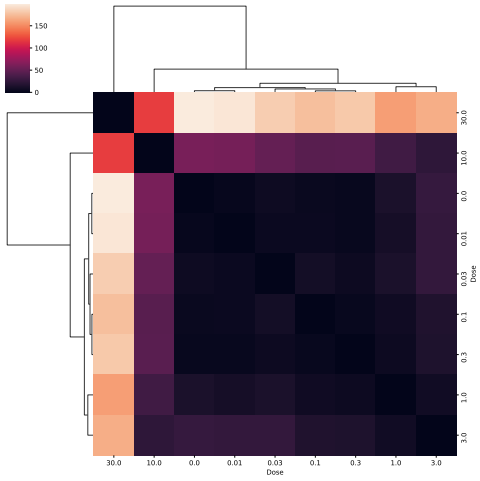


Figure 5: MMD

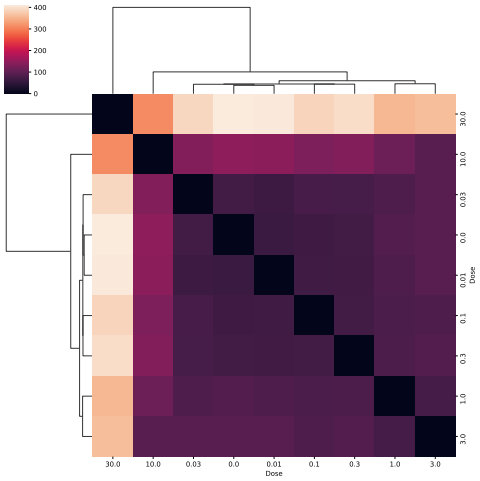


Figure 6: Wasserstein

Figure 7: Distance metrics across all cell types per dosage

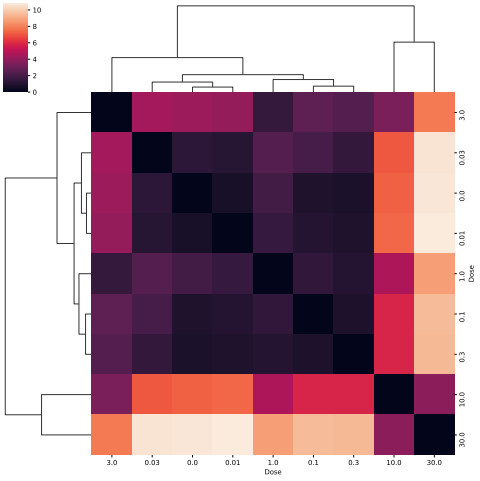


Figure 8: E-distance

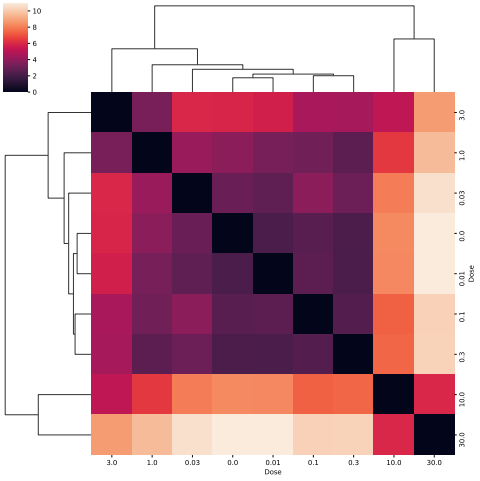


Figure 9: Euclidean

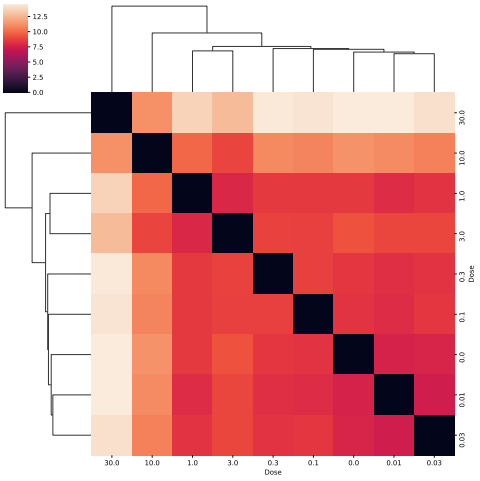


Figure 10: Mean pairwise

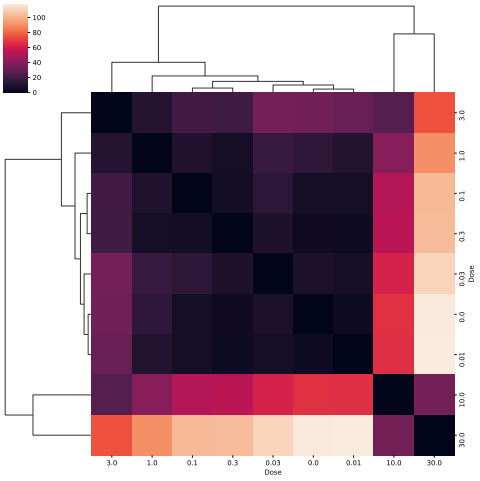


Figure 11: MMD

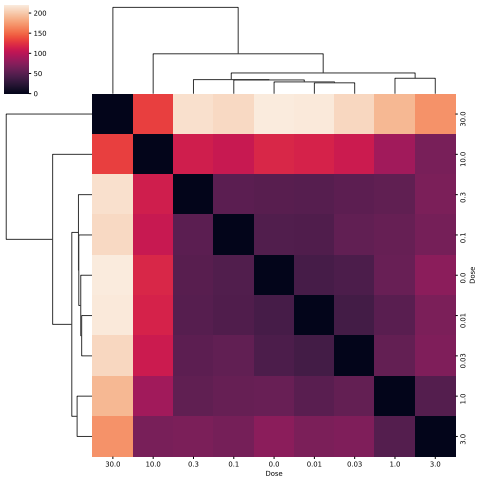


Figure 12: Wasserstein

Figure 13: Distance metrics for cell type Hepatocytes - portal per dosage

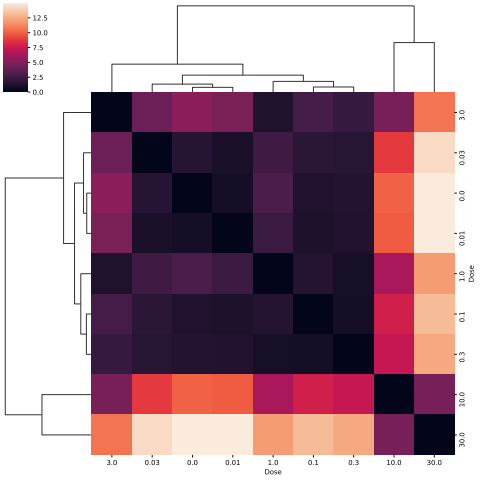


Figure 14: E-distance

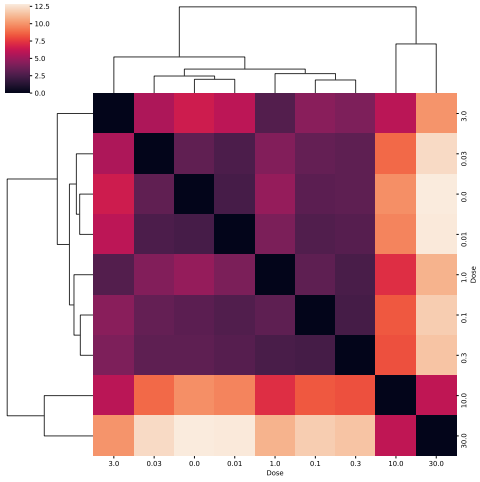


Figure 15: Euclidean

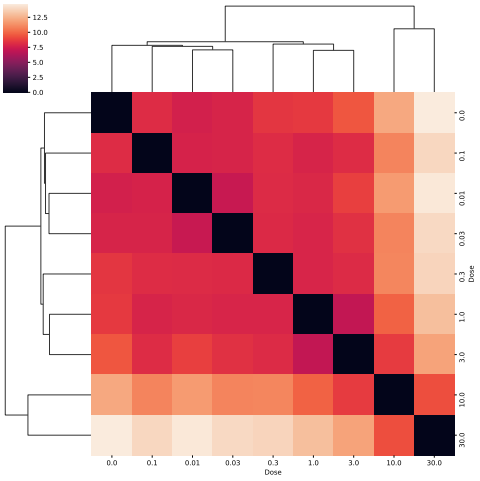


Figure 16: Mean pairwise

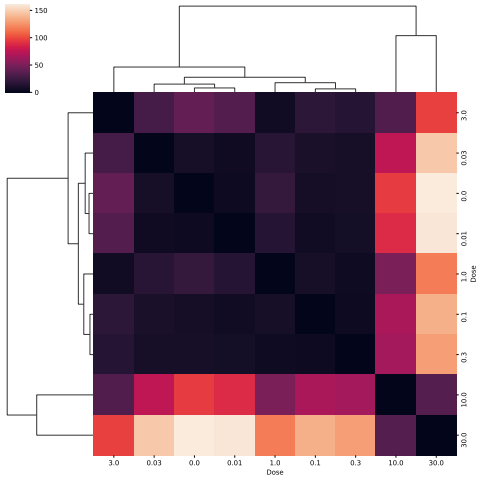


Figure 17: MMD

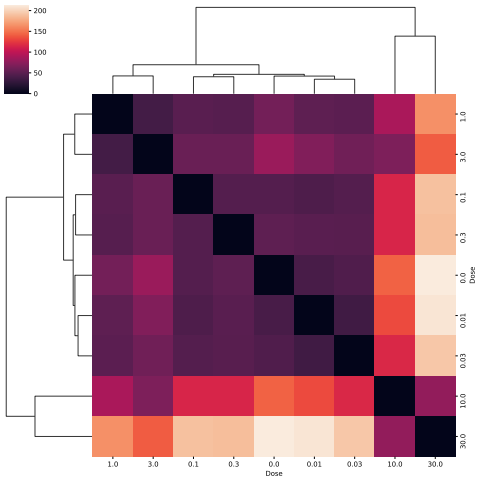


Figure 18: Wasserstein

Figure 19: Distance metrics for cell type Hepatocytes - central per dosage

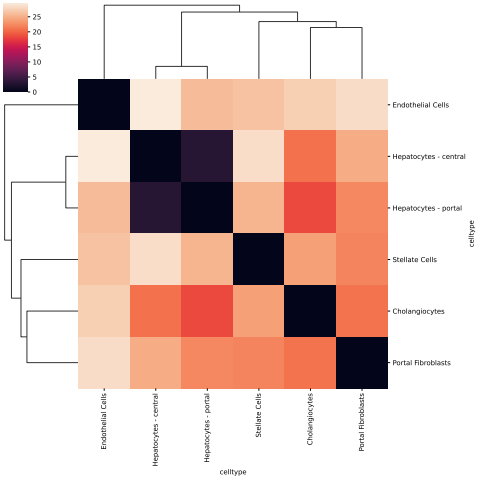


Figure 20: E-distance

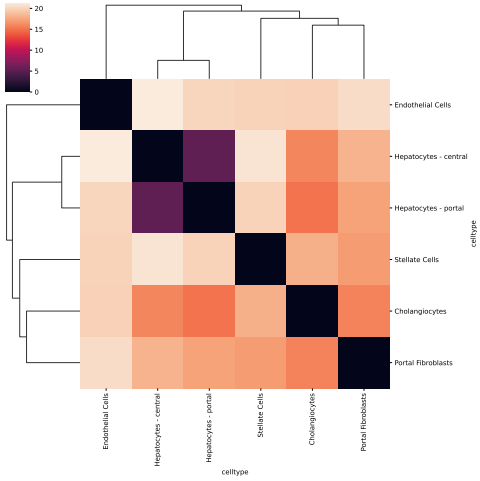


Figure 21: Euclidean

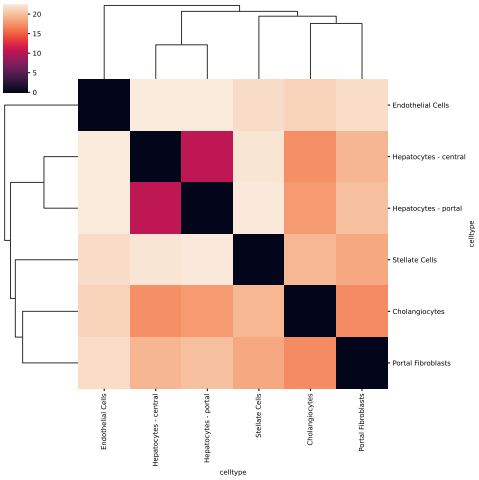


Figure 22: Mean pairwise

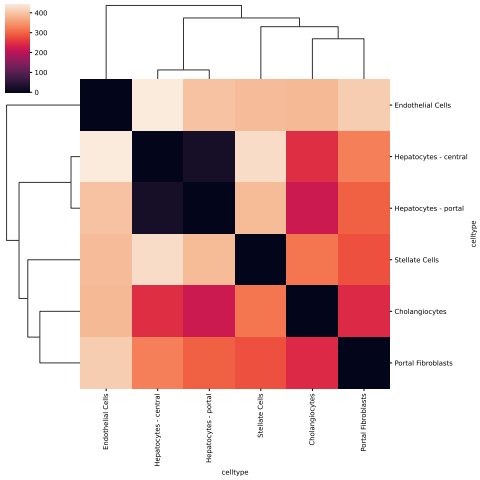


Figure 23: MMD

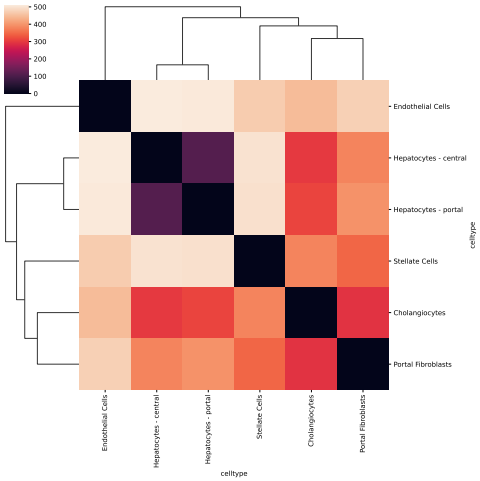


Figure 24: Wasserstein

Figure 25: Distance metrics for dosage highest 30 $\mu g/kg$ per cell type

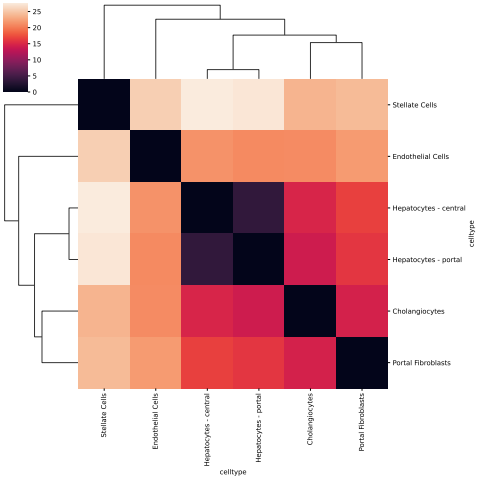


Figure 26: E-distance

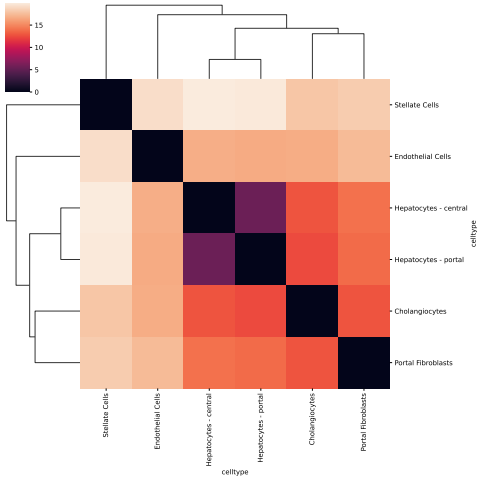


Figure 27: Euclidean

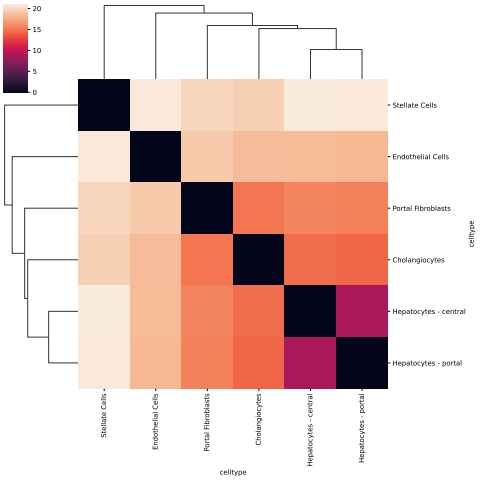


Figure 28: Mean pairwise

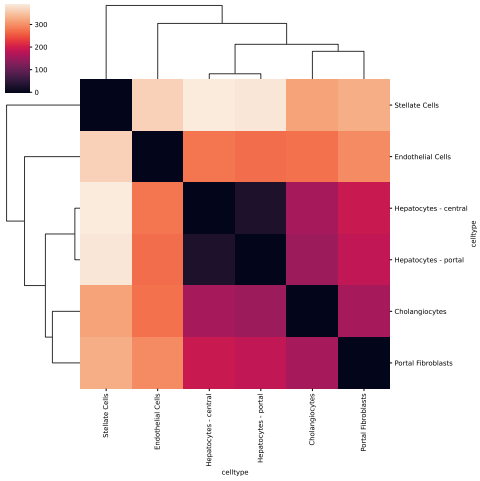


Figure 29: MMD

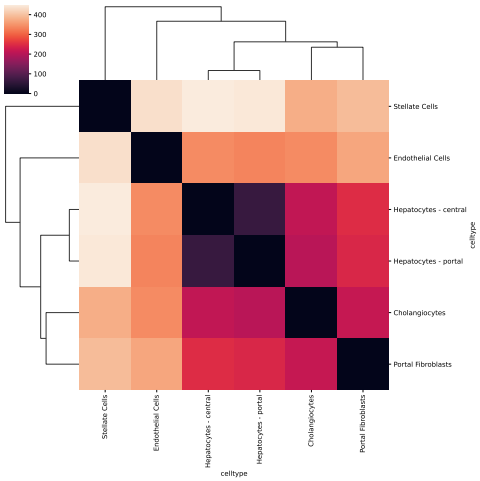
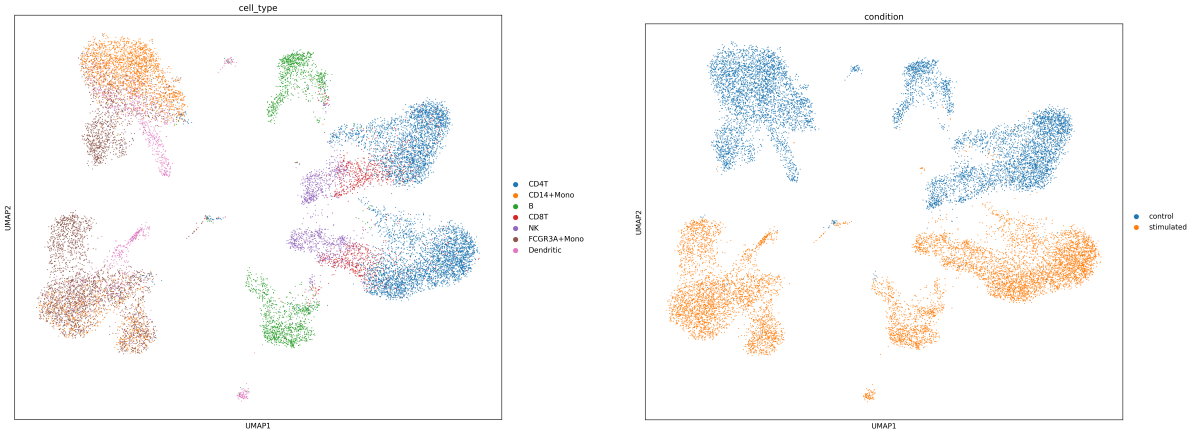


Figure 30: Wasserstein

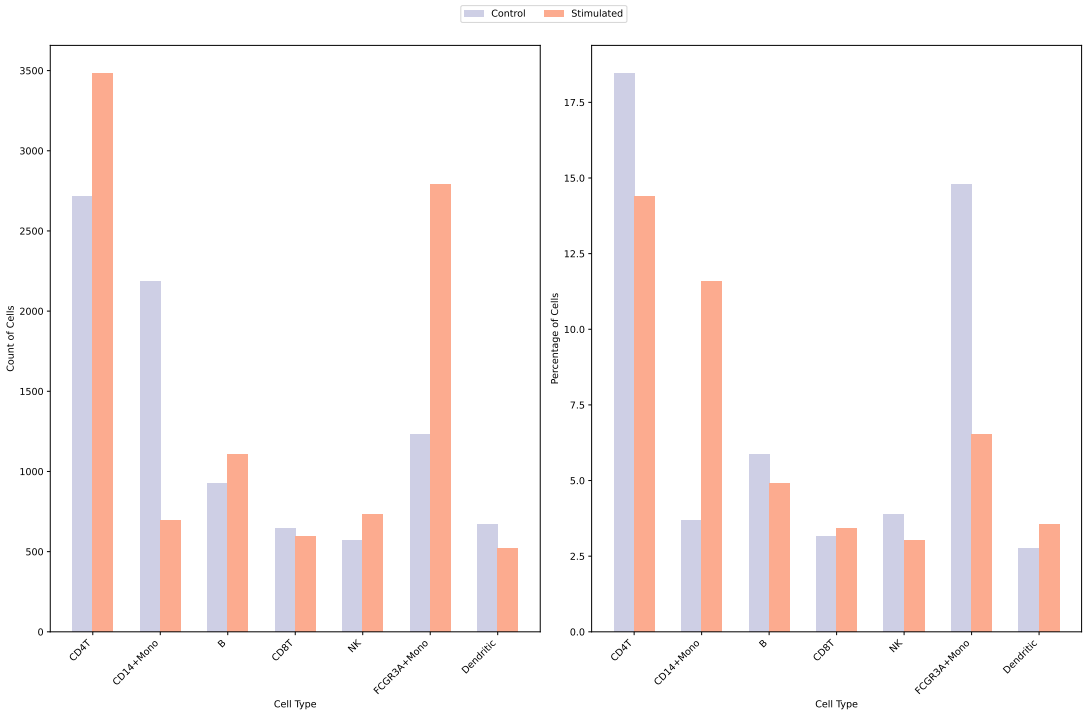
Figure 31: Distance metrics for lowest dosage 0.01 $\mu g/kg$ per cell type

8.2 PBMC dataset



(a)

(b)



(c)

Figure 32: PBMC overview

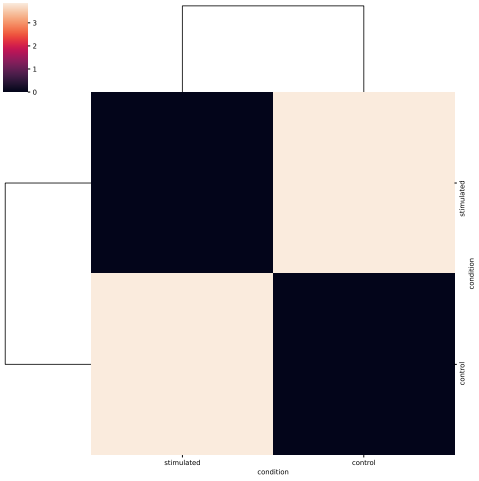


Figure 33: E-distance

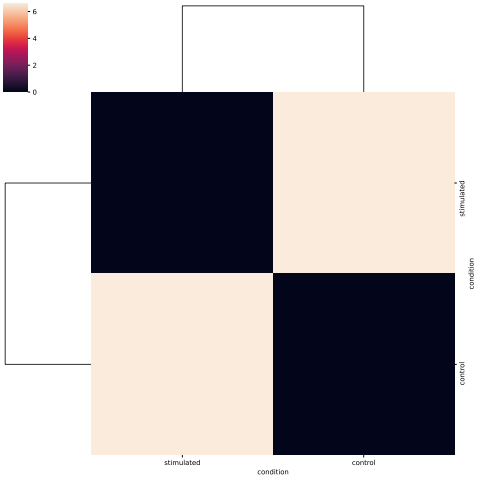


Figure 34: Euclidean

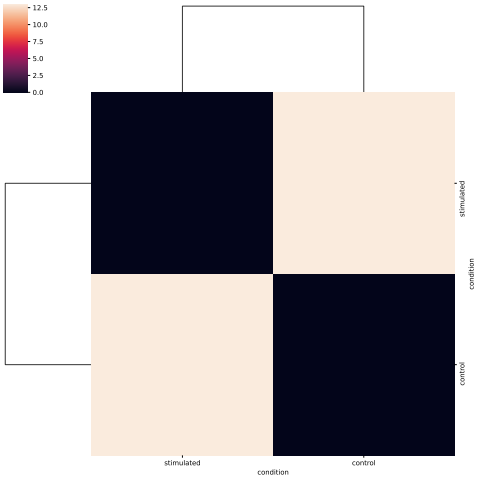


Figure 35: Mean pairwise

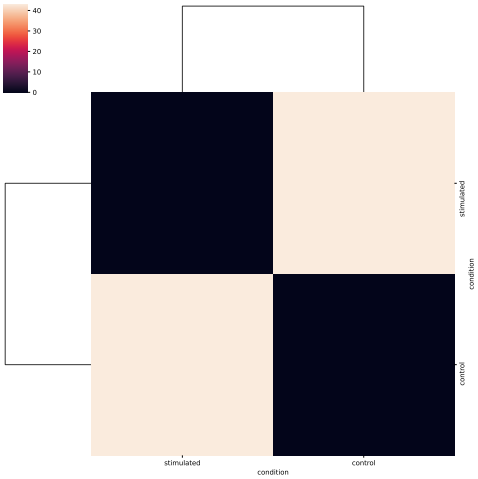


Figure 36: MMD

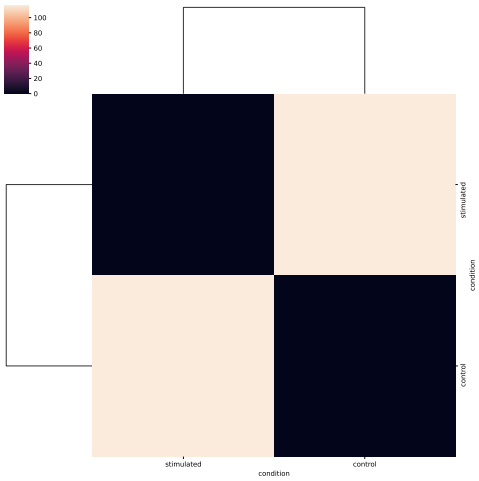


Figure 37: Wasserstein

Figure 38: Distance metrics per condition

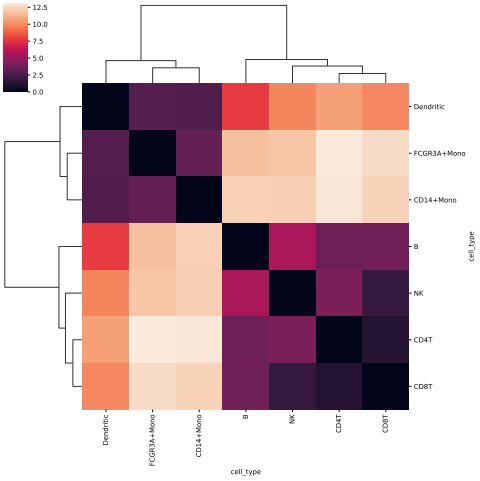


Figure 39: E-distance

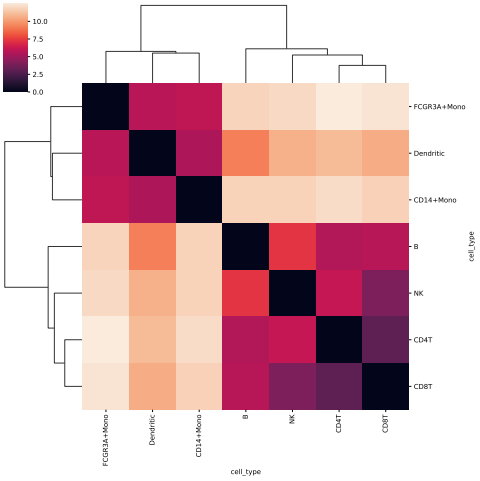


Figure 40: Euclidean

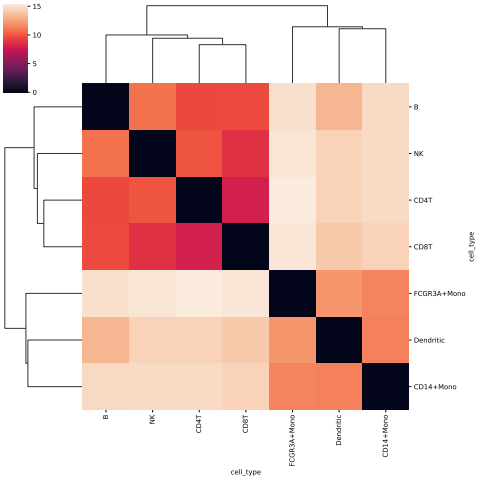


Figure 41: Mean pairwise

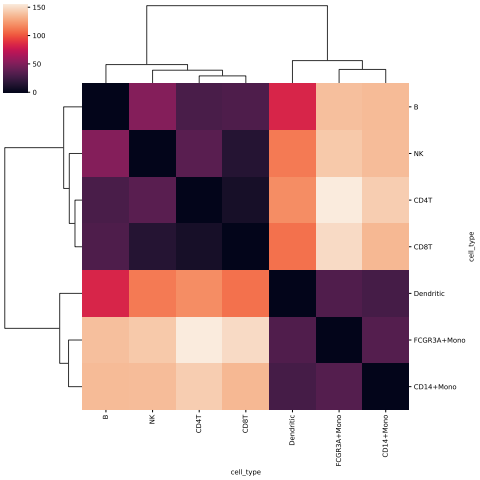


Figure 42: MMD

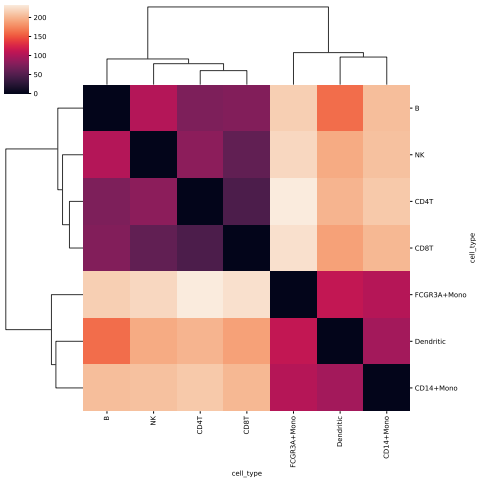
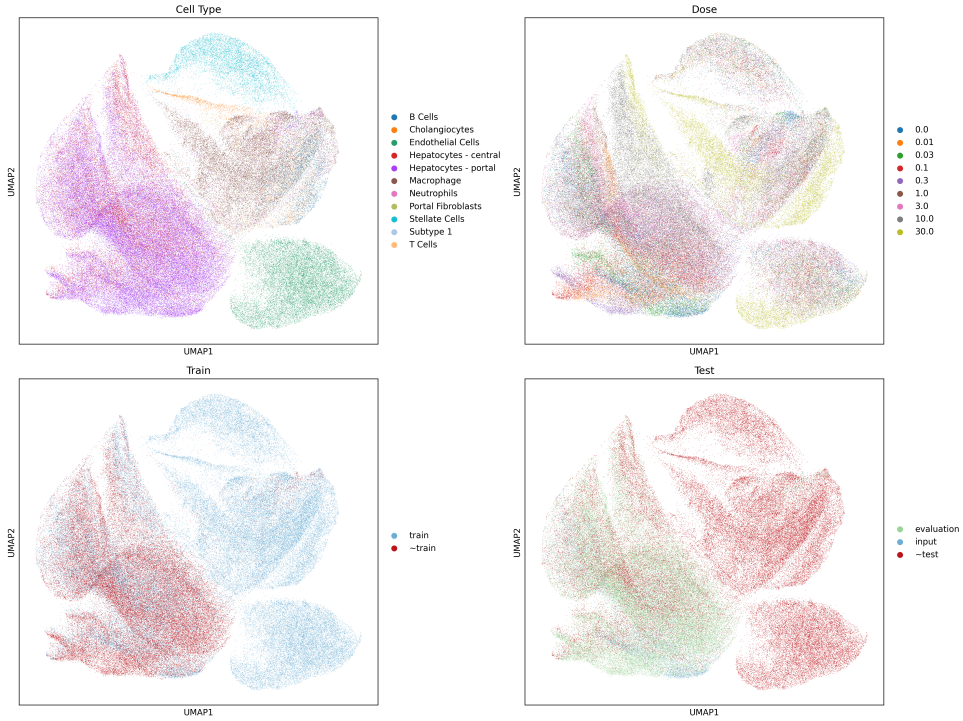


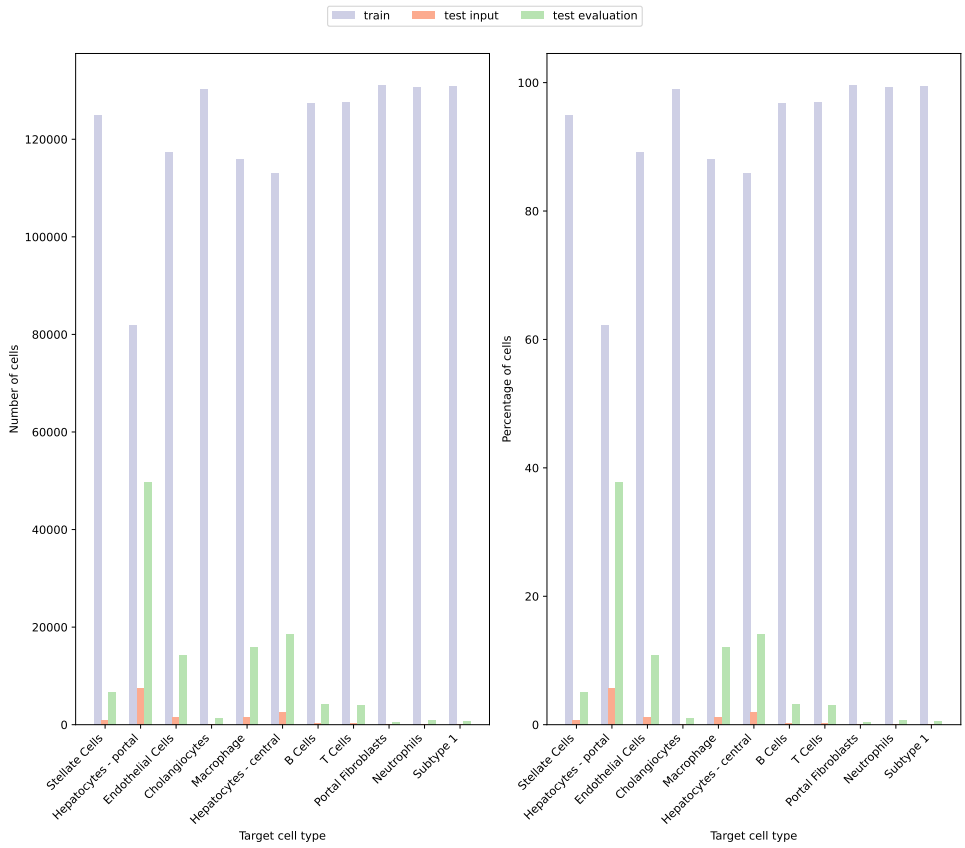
Figure 43: Wasserstein

Figure 44: Distance metrics per cell type

9 Nault all cell types evaluation

9.1 Multiple doses





9.2 Single dose

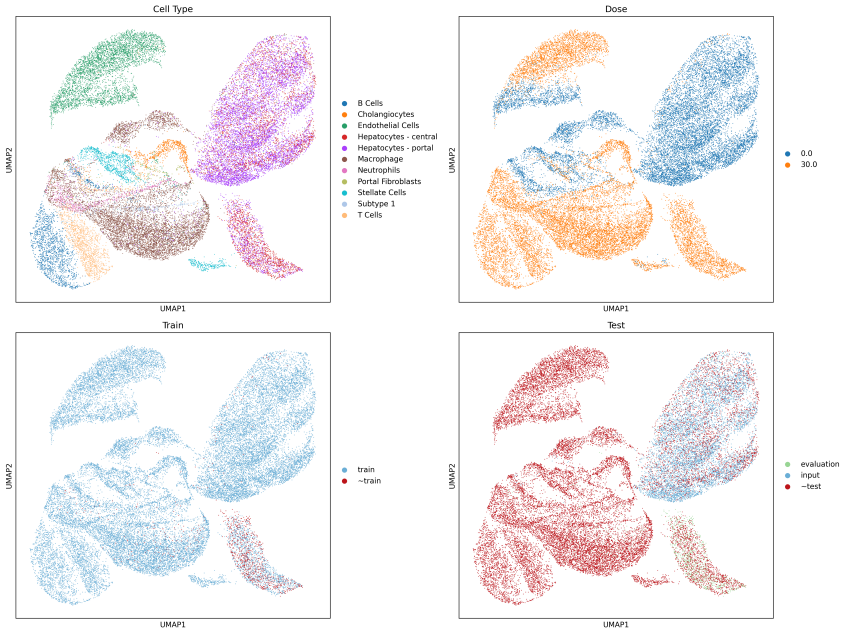


Figure 45: Example of $30\mu g/kg$

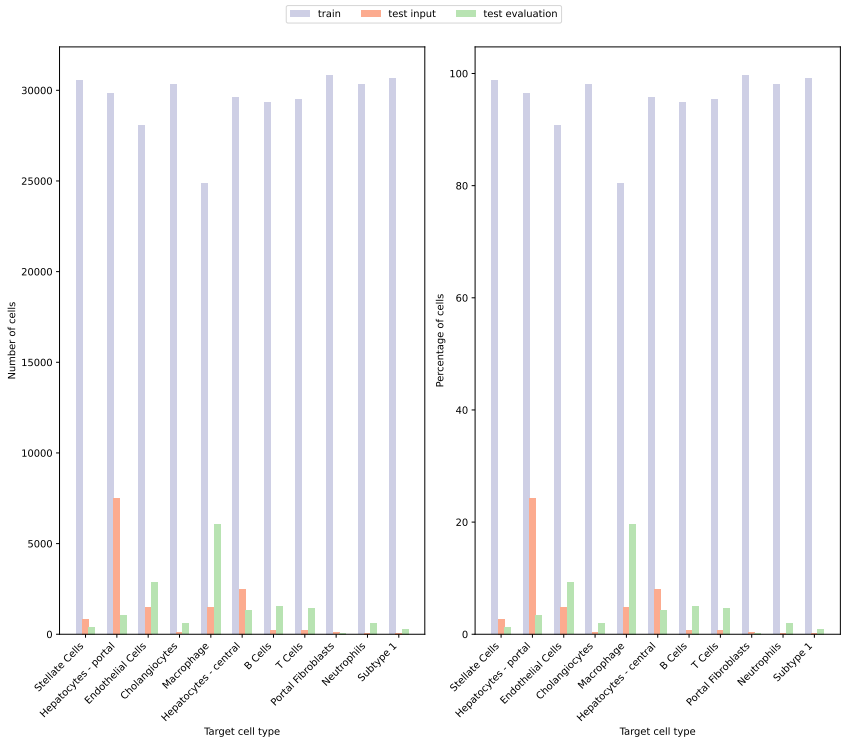


Figure 46: Number of cells per cell type for $30\mu g/kg$

9.3 Comparison

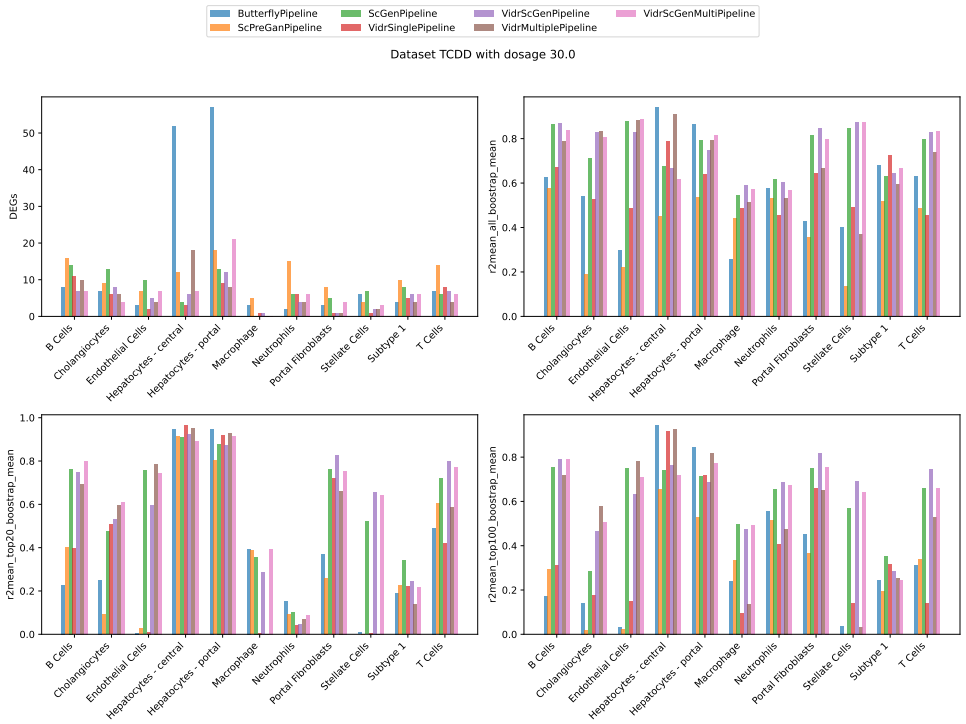


Figure 47: Baseline metrics for highest dosage $30\mu g/kg$ across cell types

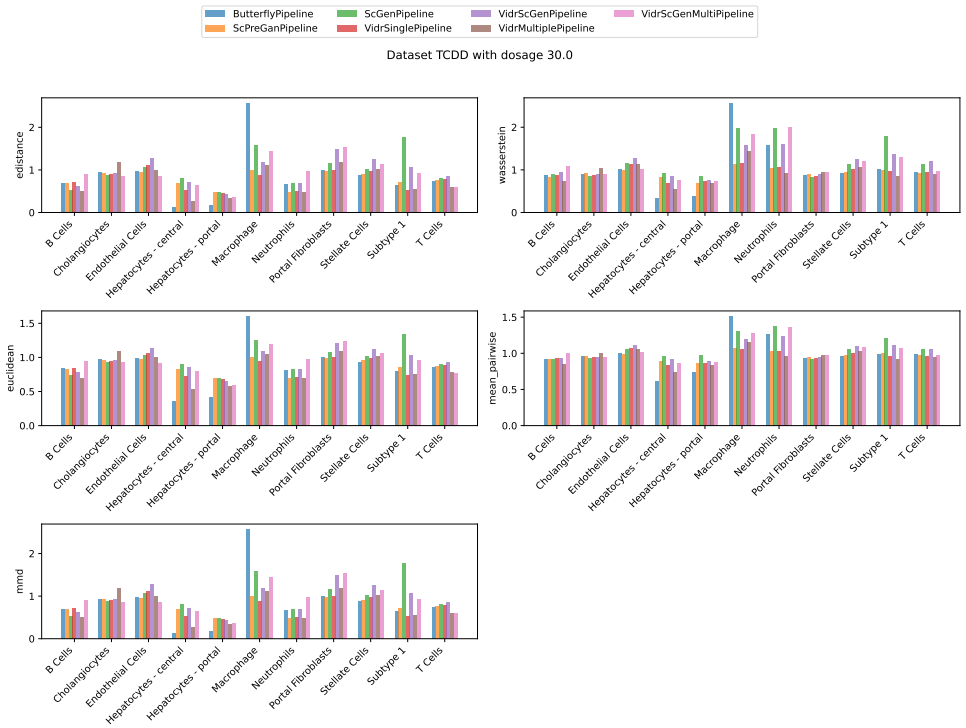


Figure 48: Distance metrics for highest dosage $30\mu g/kg$ across cell types

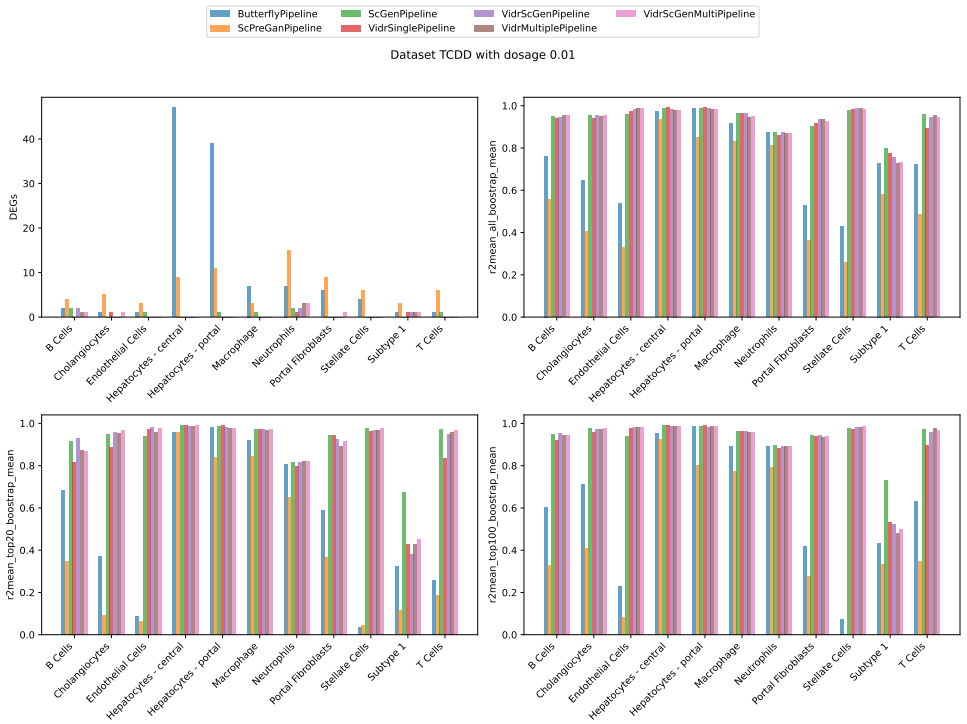


Figure 49: Baseline metrics for lowest dosage $0.01\mu g/kg$ across cell types

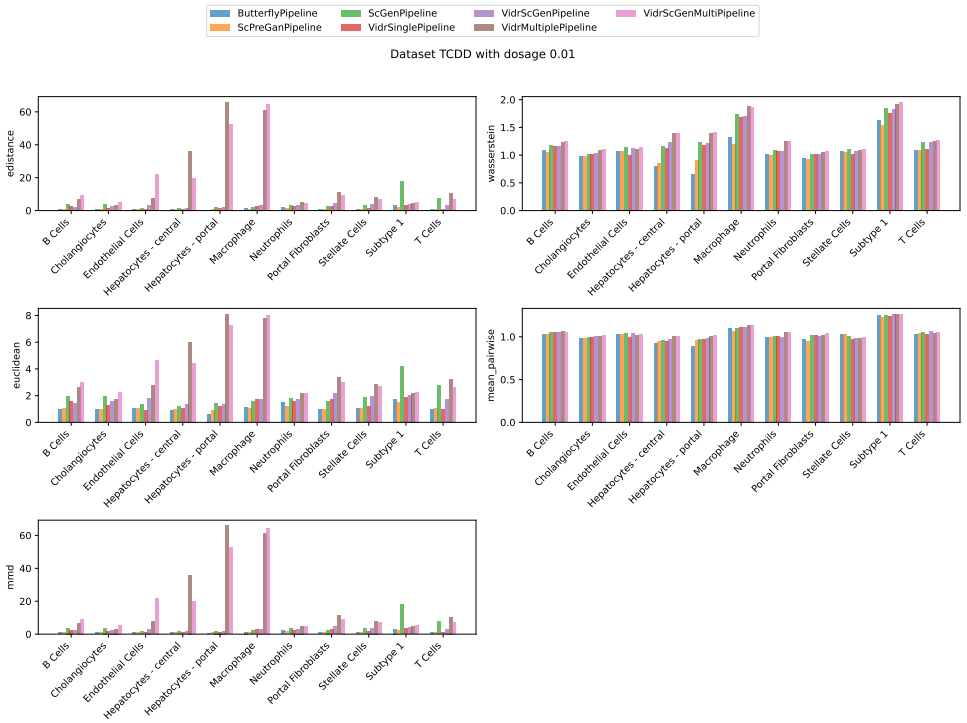


Figure 50: Distance metrics for lowest dosage $0.01\mu g/kg$ across cell types

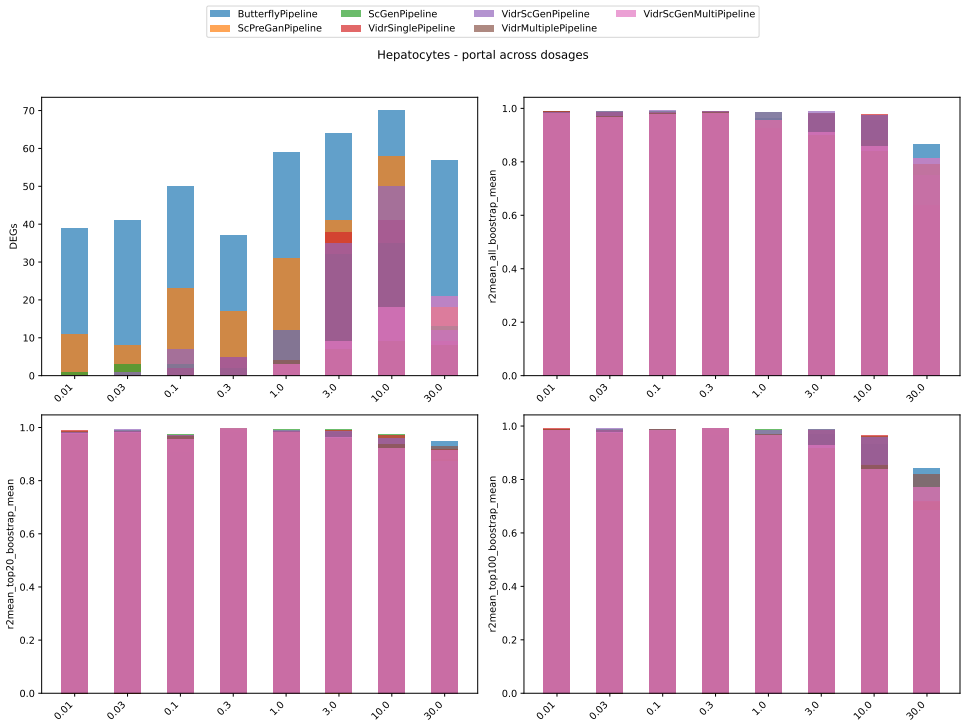


Figure 51: Baseline metrics for Hepatocytes - portal across dosages

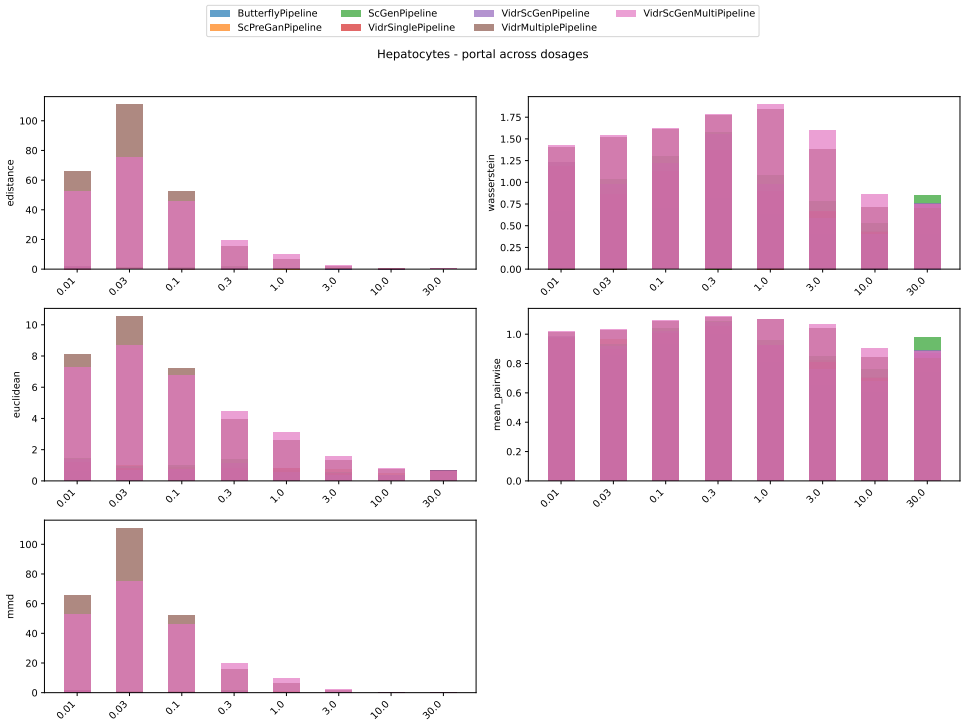


Figure 52: Distance metrics for Hepatocytes - portal across dosages

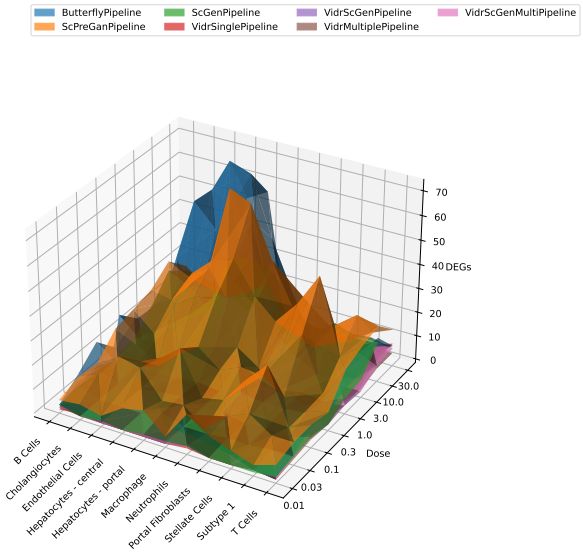


Figure 53

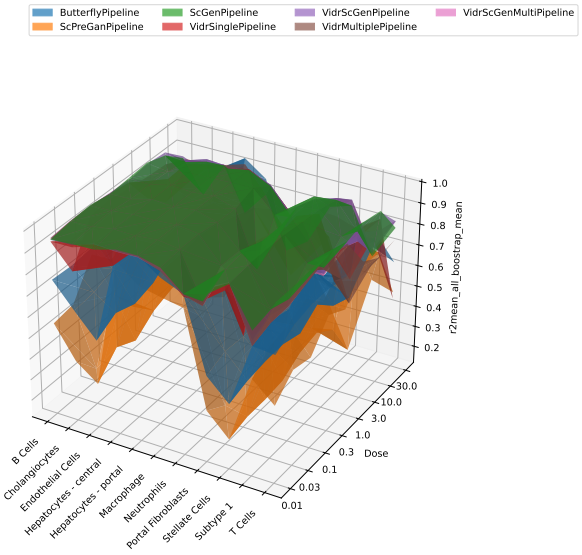


Figure 54

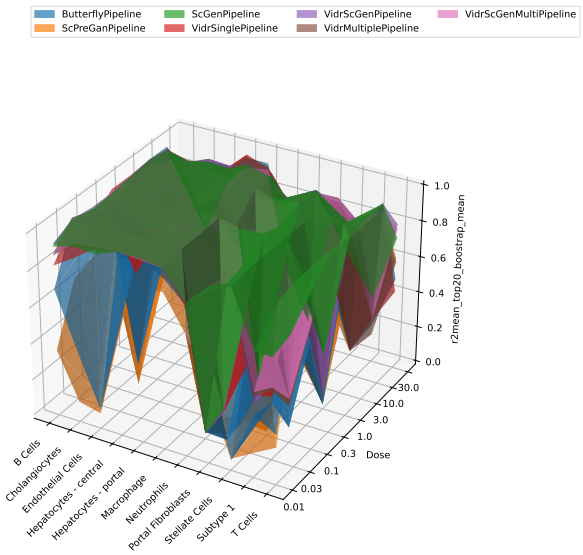


Figure 55

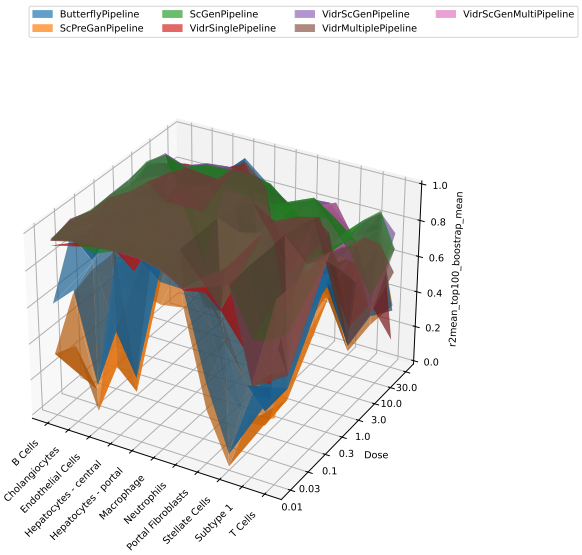


Figure 56

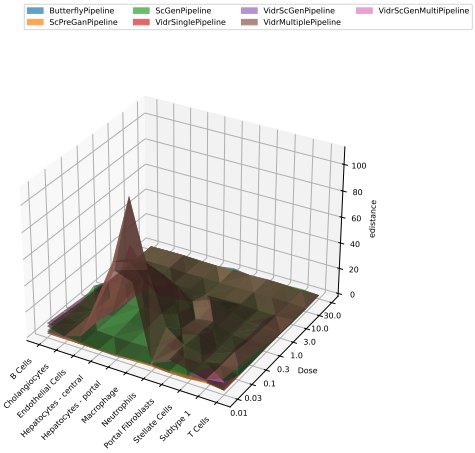


Figure 57: E-distance

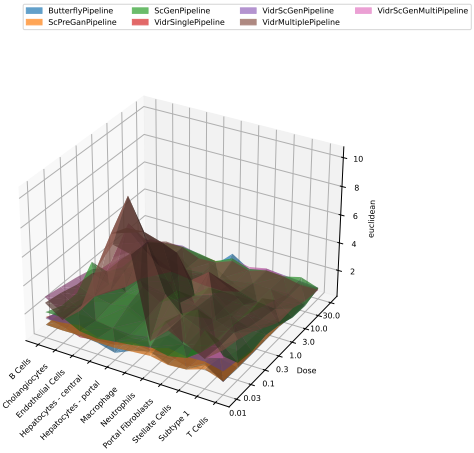


Figure 58: Euclidean

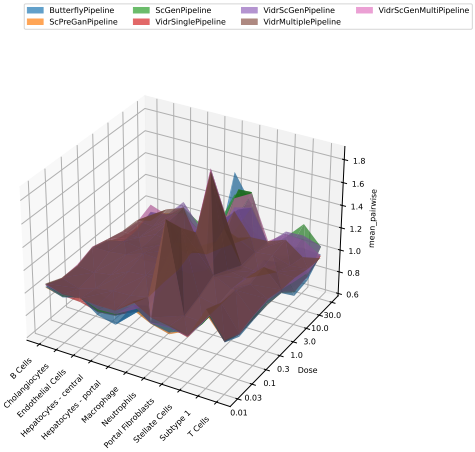


Figure 59: Mean pairwise

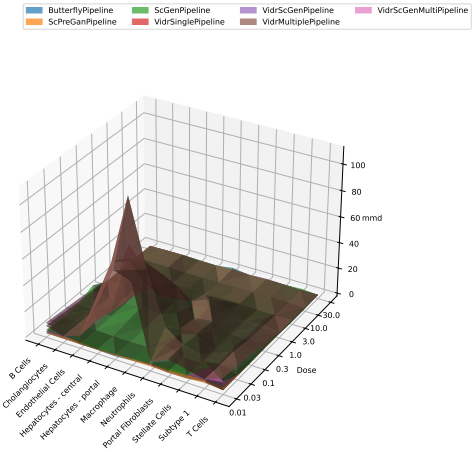


Figure 60: MMD

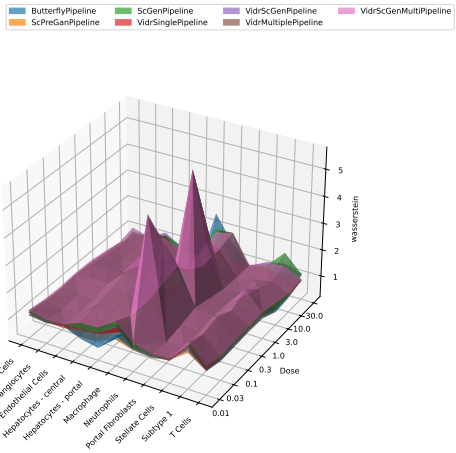
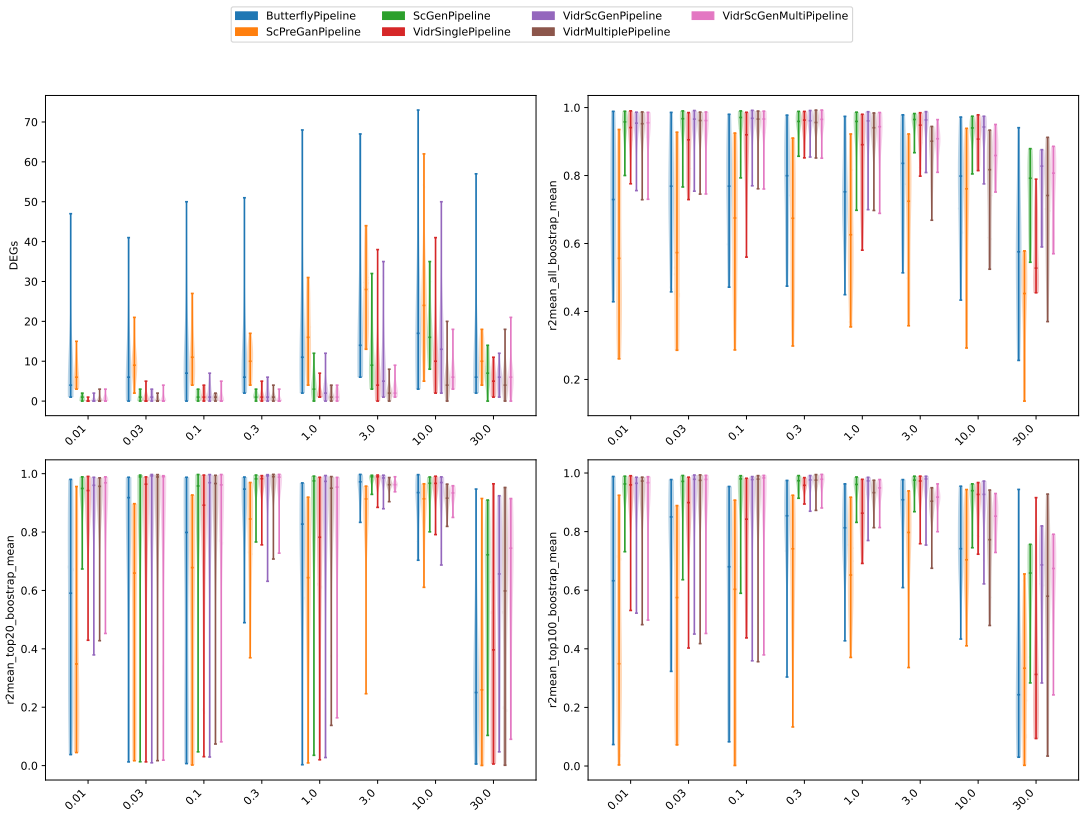
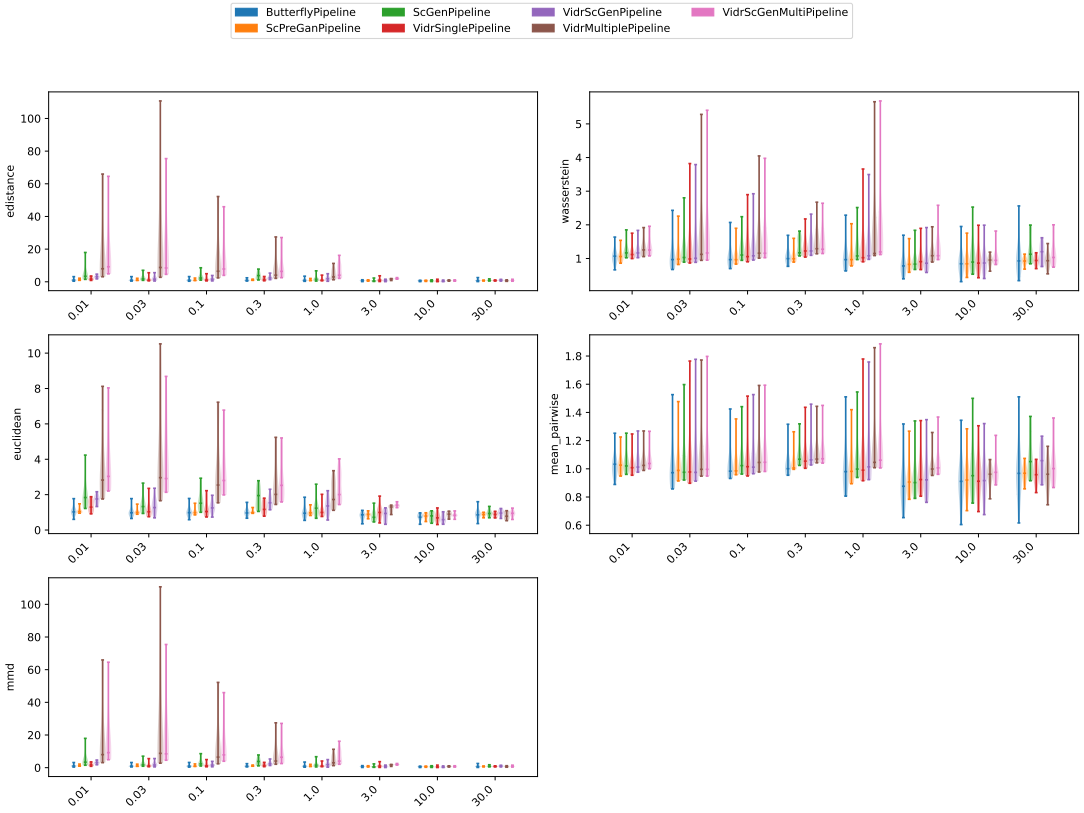
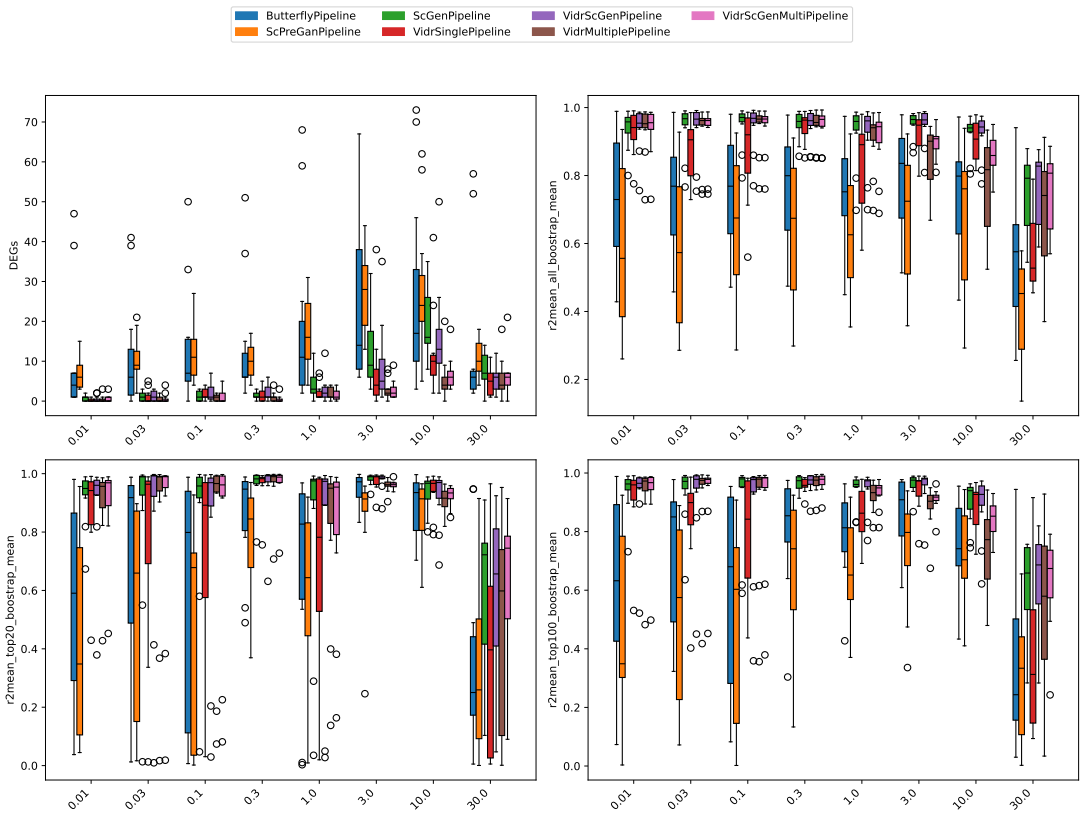


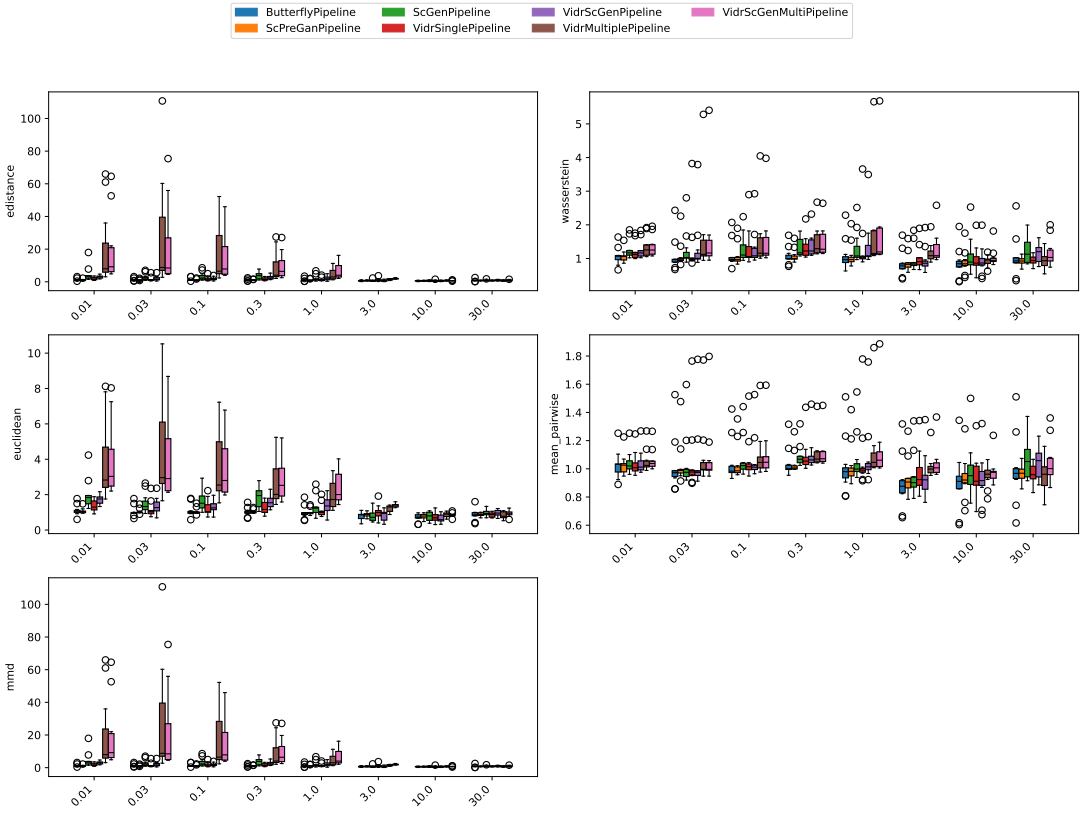
Figure 61: Wasserstein

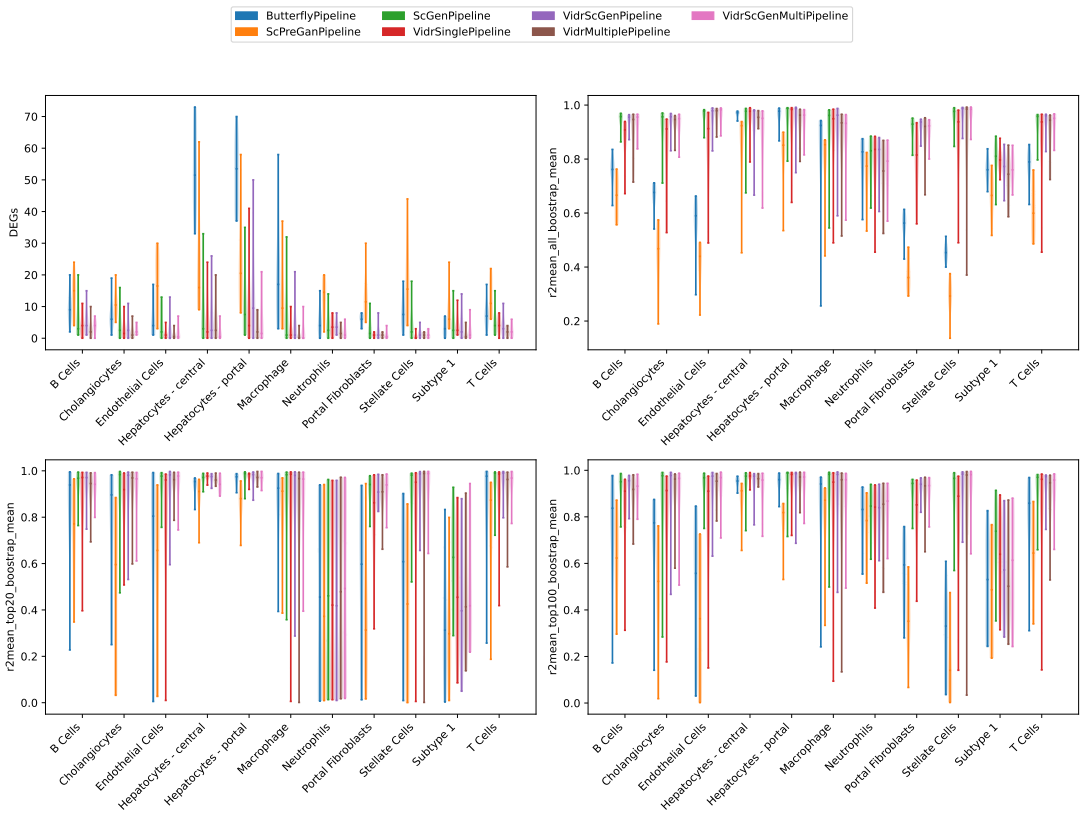
Figure 62: Distance metrics per cell type

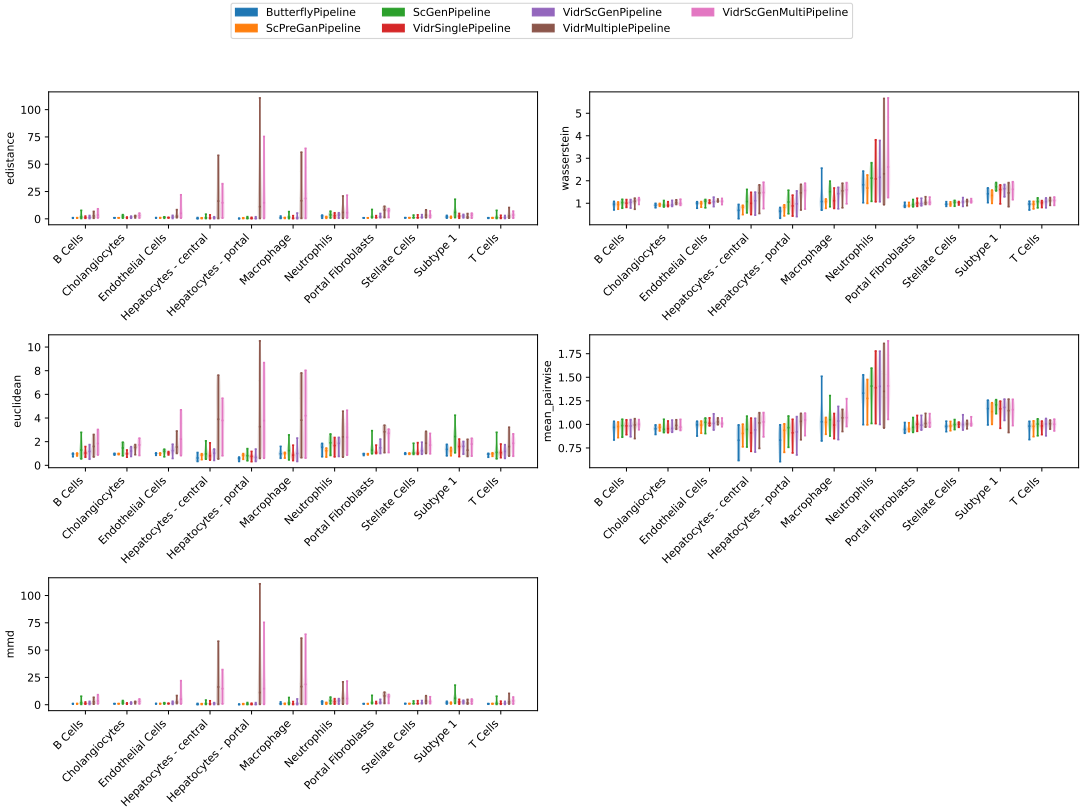

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0

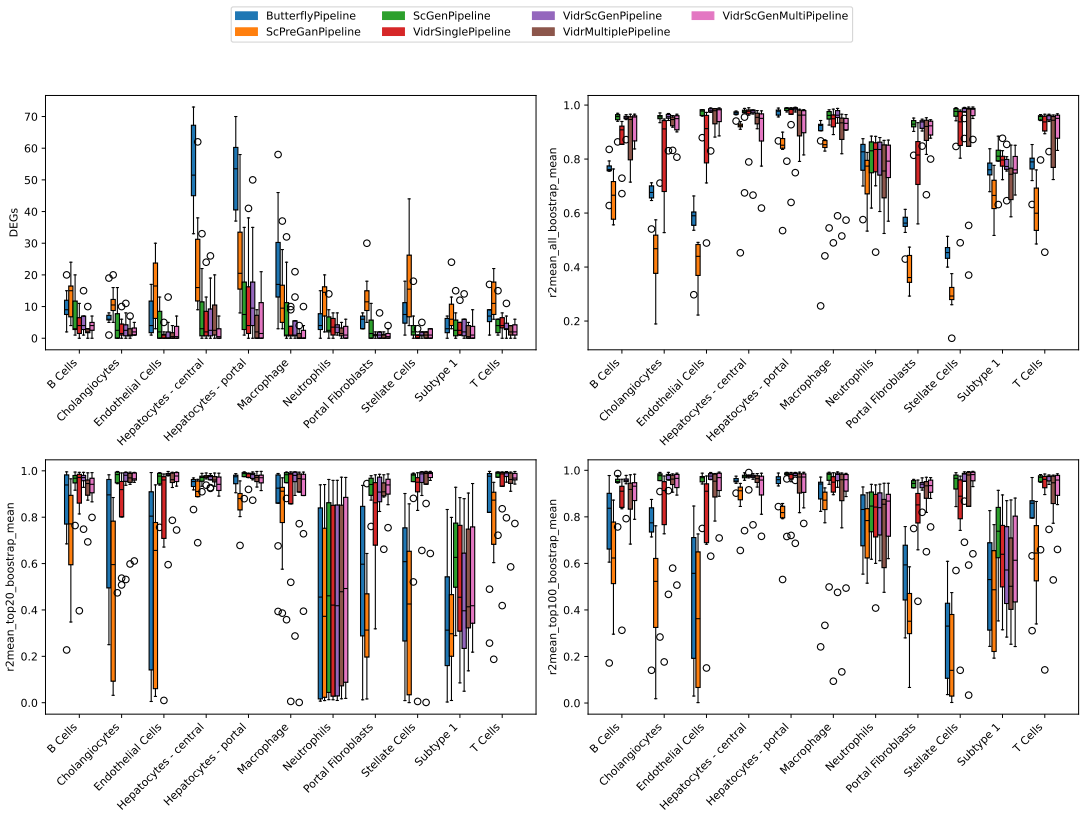

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
edistance
wasserstein
euclidean
mean_pairwise
mmd
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0

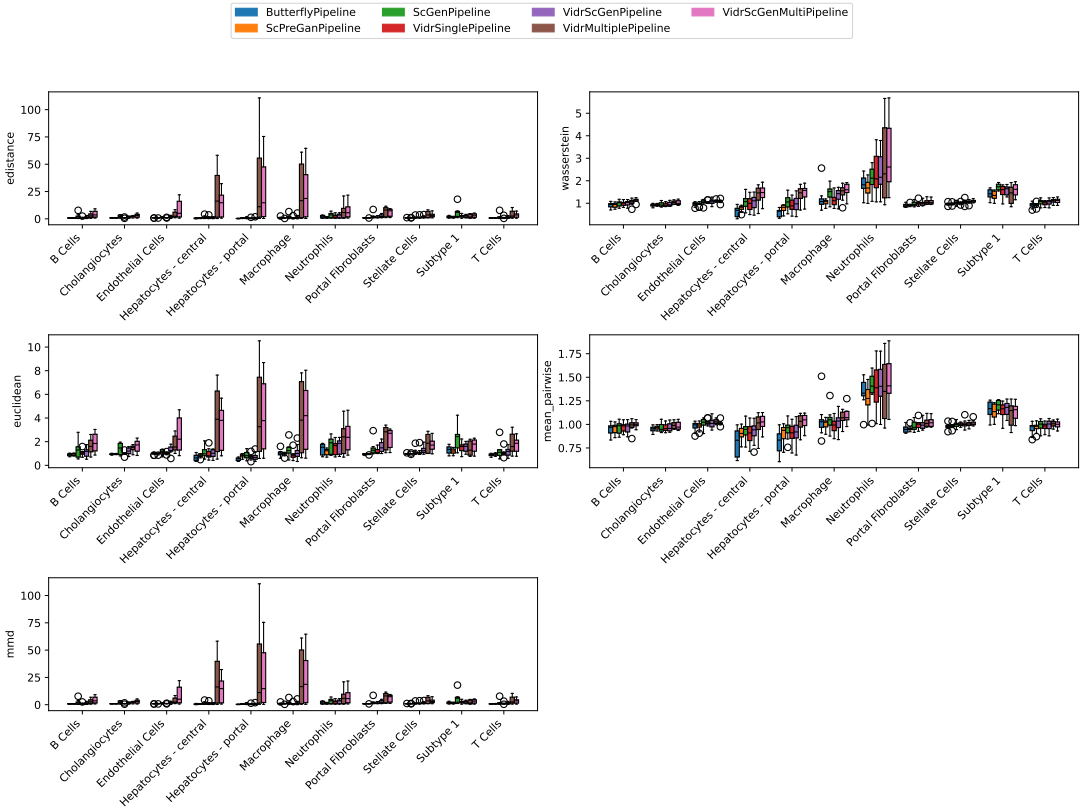

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0


ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
edistance
wasserstein
euclidean
mean_pairwise
mmd
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0


ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSingleCellPipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
B Cells
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Macrophage
Neutrophils
Portal Fibroblasts
Stellate Cells
Subtype 1
T Cells


ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSingleCellPipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
edistance
wasserstein
euclidean
mean_pairwise
mmd
B Cells
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Macrophage
Neutrophils
Portal Fibroblasts
Stellate Cells
Subtype 1
T Cells


ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSingleCellPipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
B Cells
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Macrophage
Neutrophils
Portal Fibroblasts
Stellate Cells
Subtype 1
T Cells


ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSingleCellPipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
edistance
wasserstein
euclidean
mean_pairwise
mmd
B Cells
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Macrophage
Neutrophils
Portal Fibroblasts
Stellate Cells
Subtype 1
T Cells

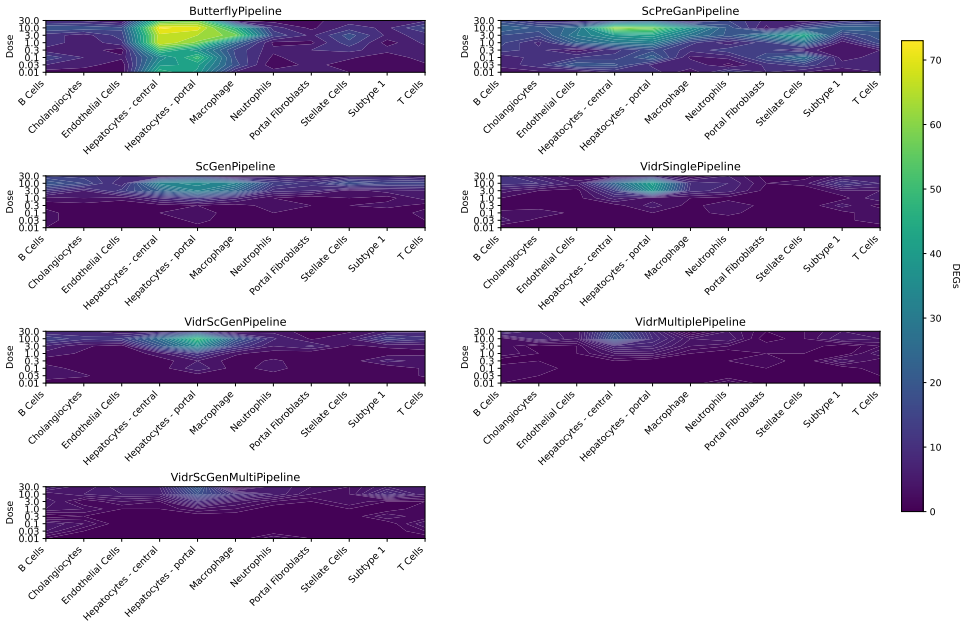


Figure 63: DEGs

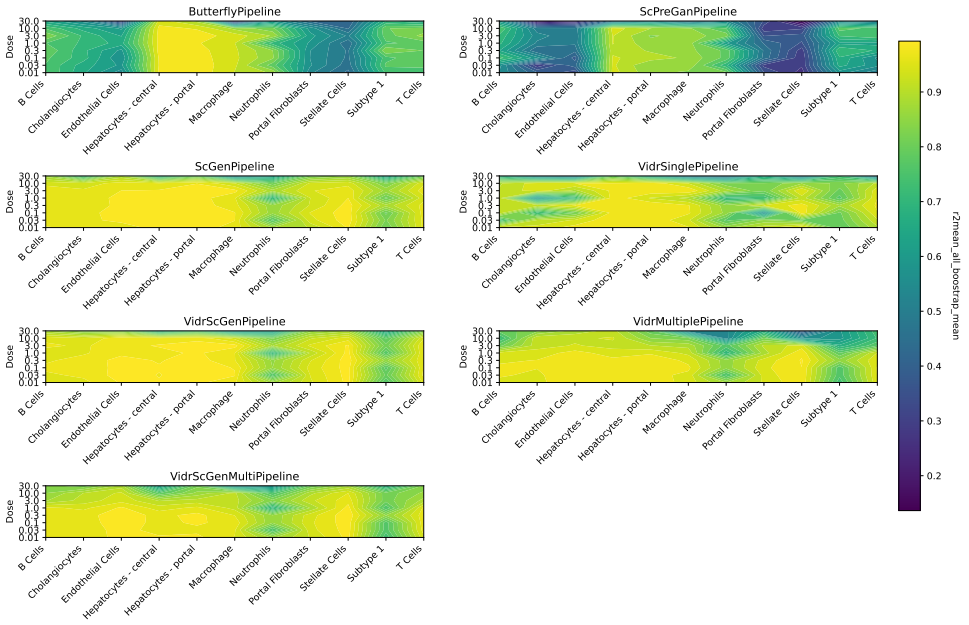


Figure 64: r2 HVGs

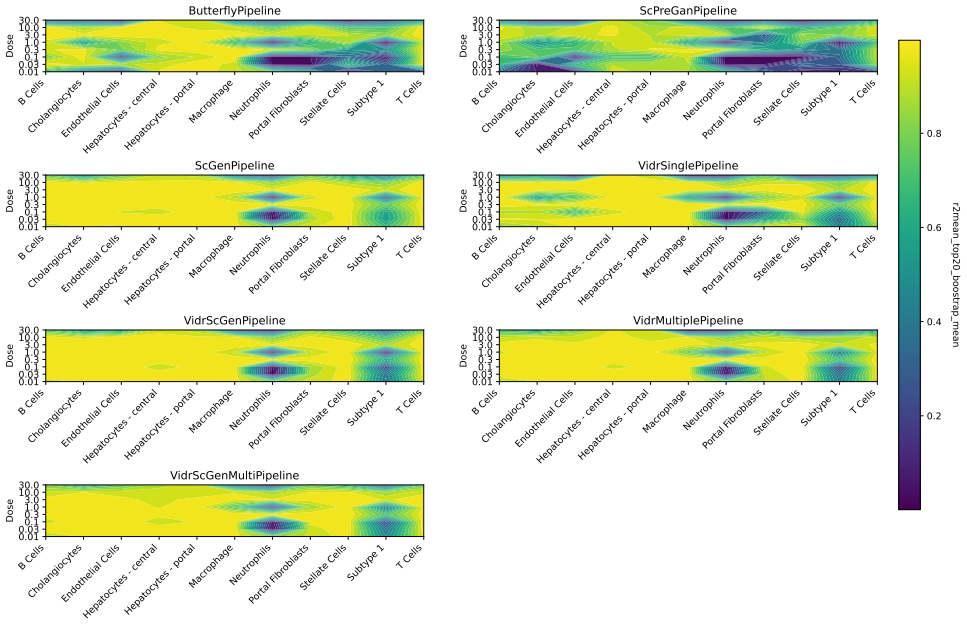


Figure 65: r2 top 20

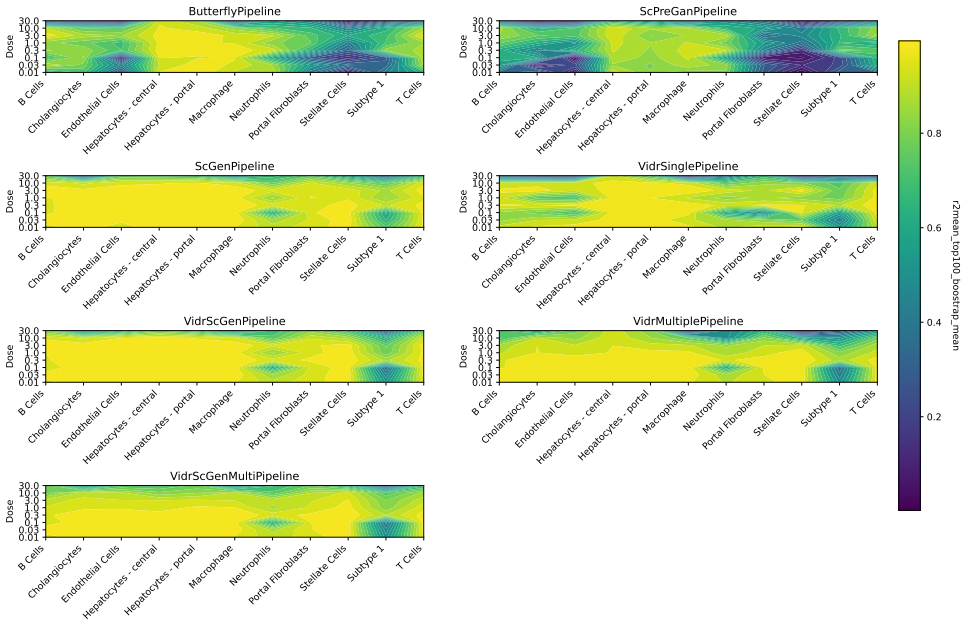


Figure 66: r2 top 100

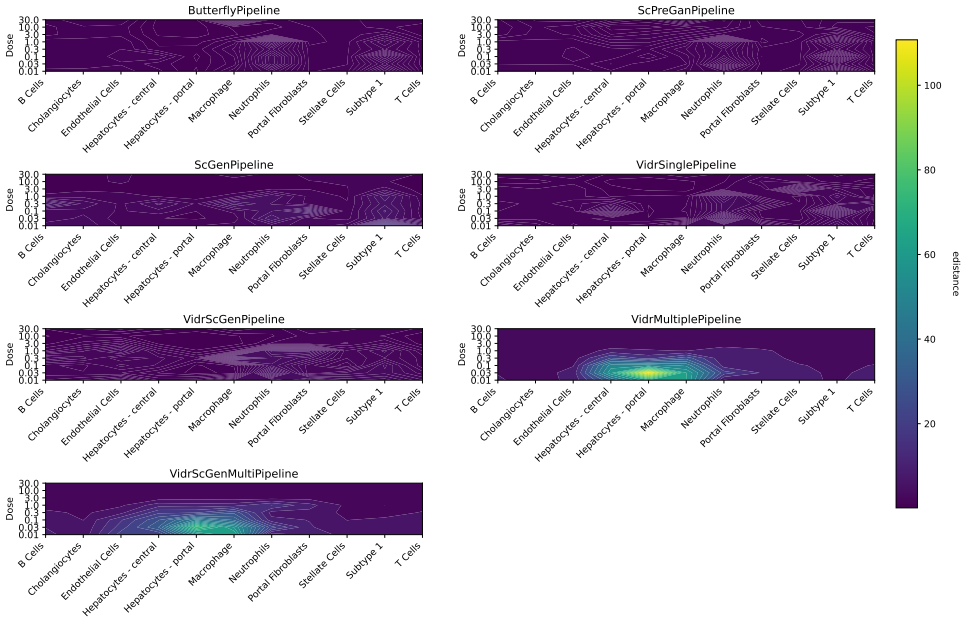


Figure 67: E-distance

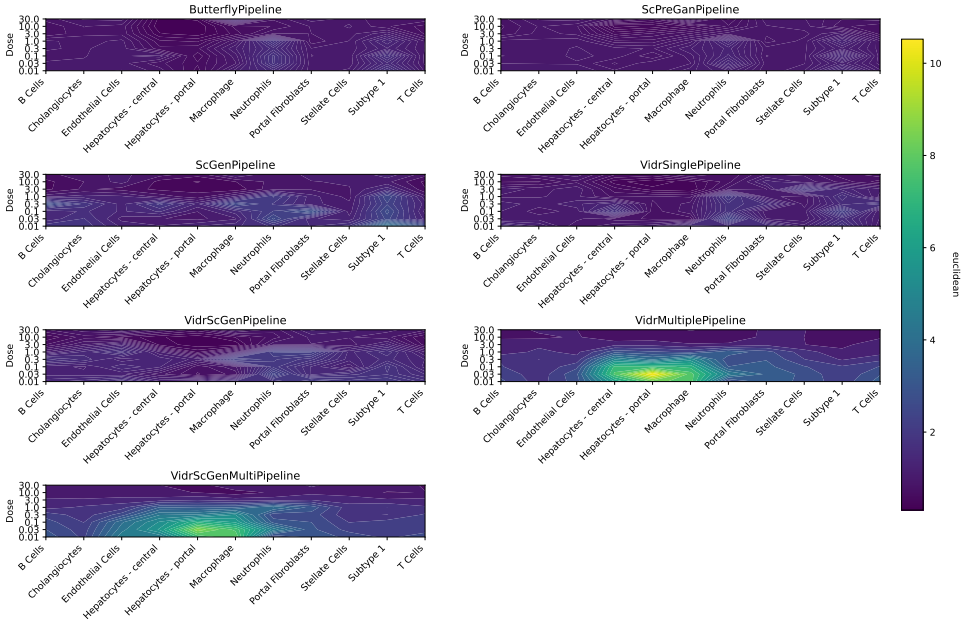


Figure 68: Euclidean

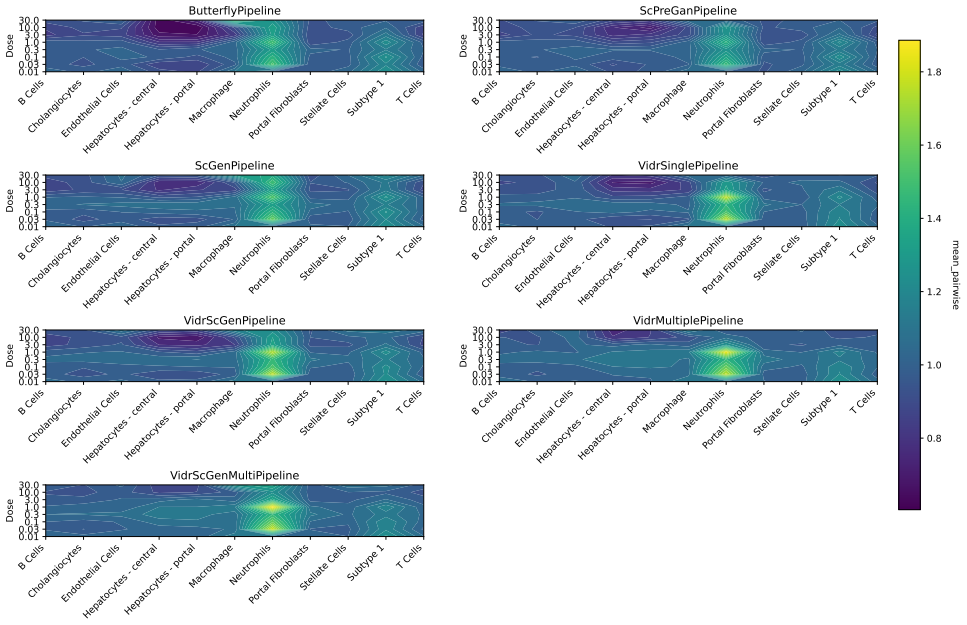


Figure 69: Mean pairwise

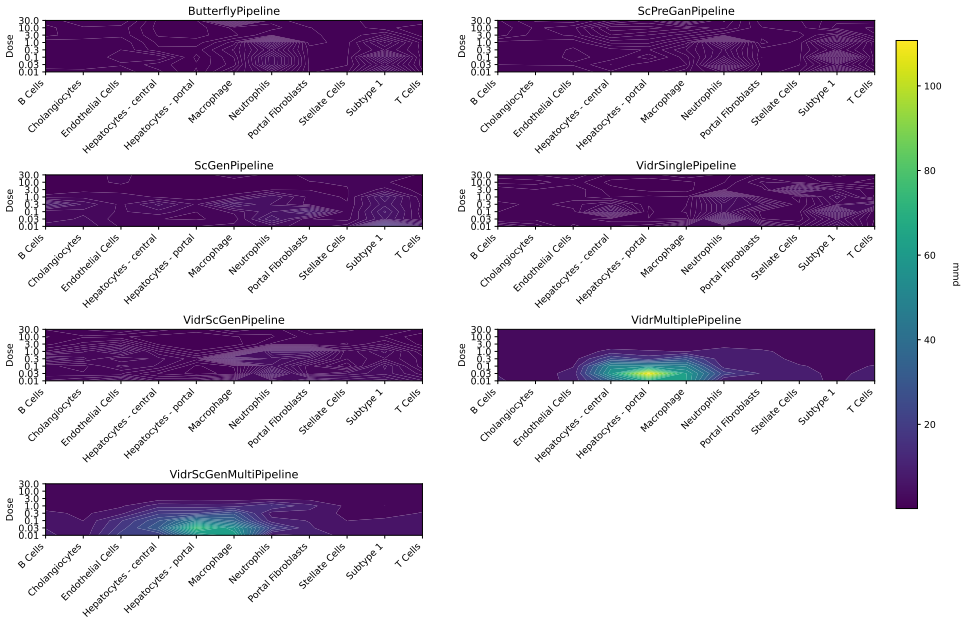


Figure 70: MMD

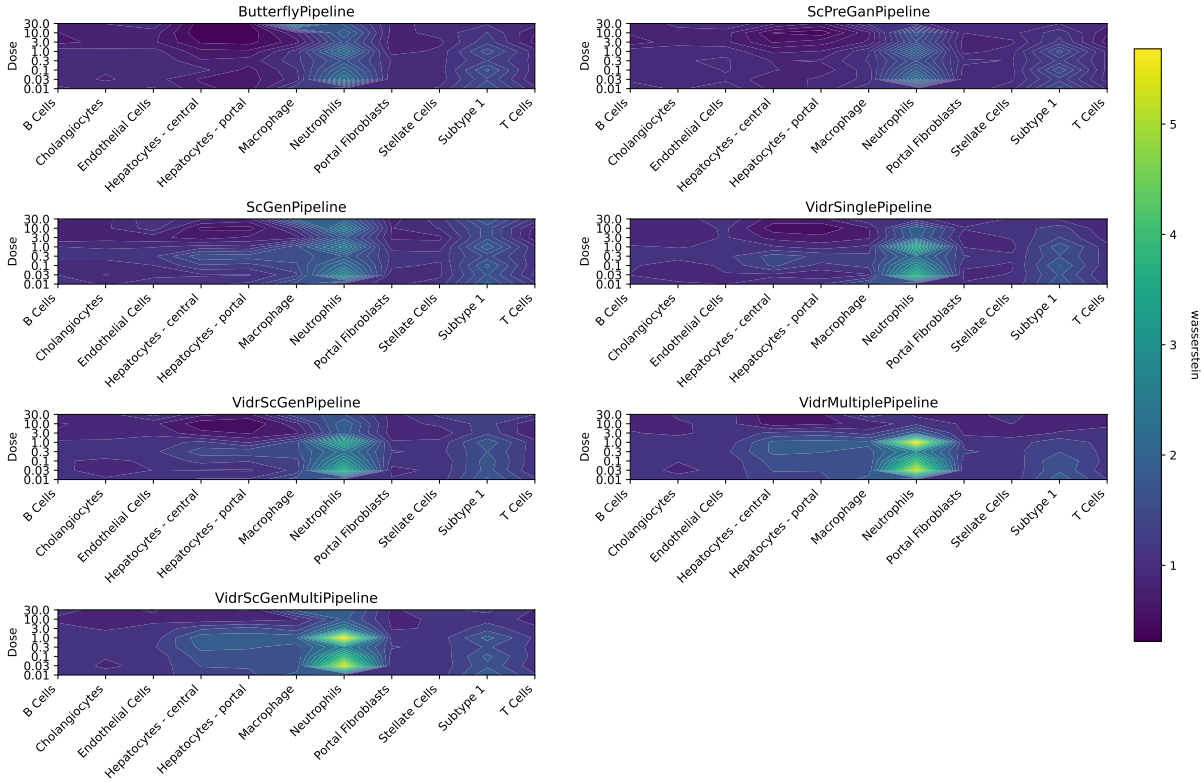
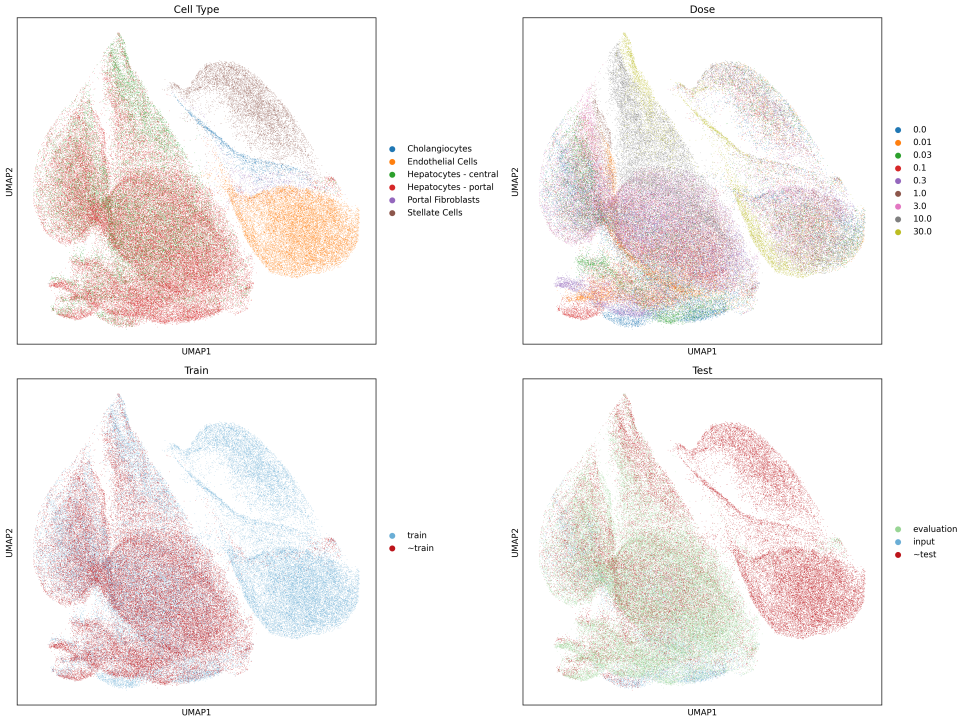
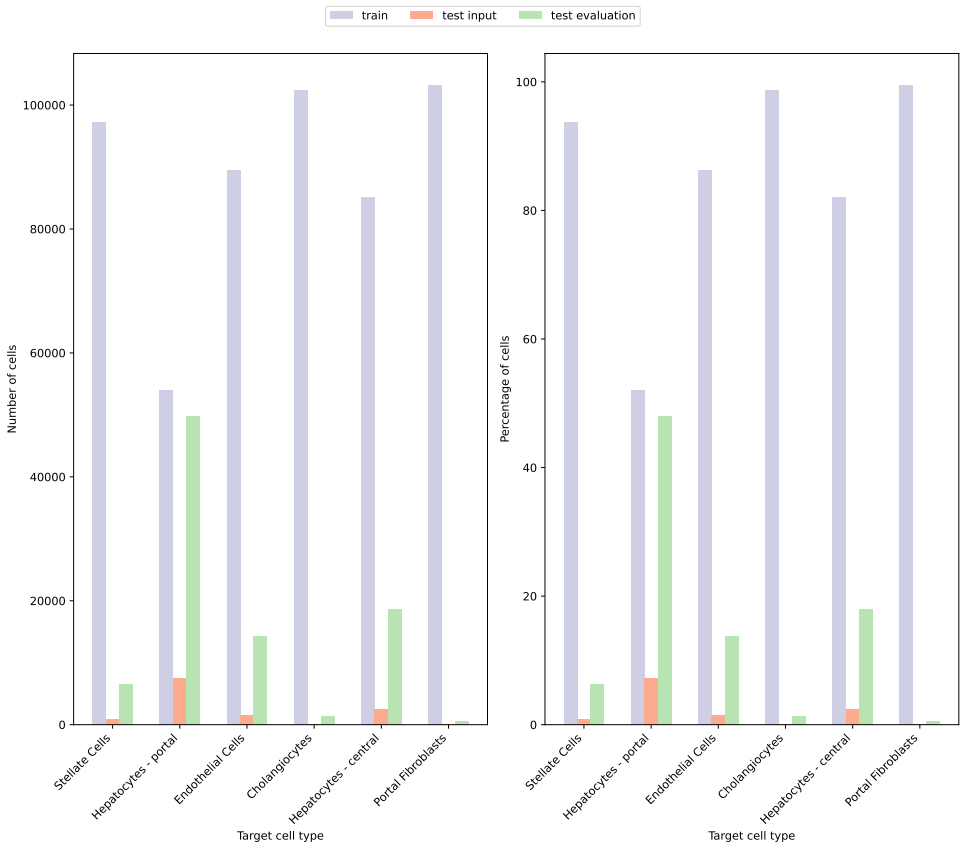


Figure 71: Wasserstein

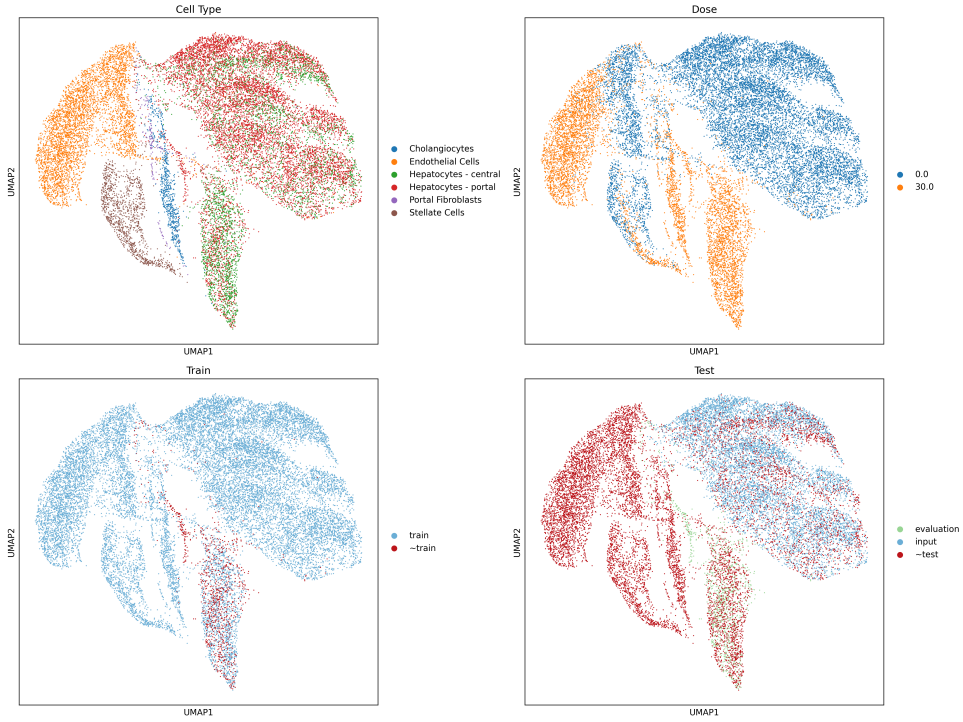
10 Nault liver cell types evaluation

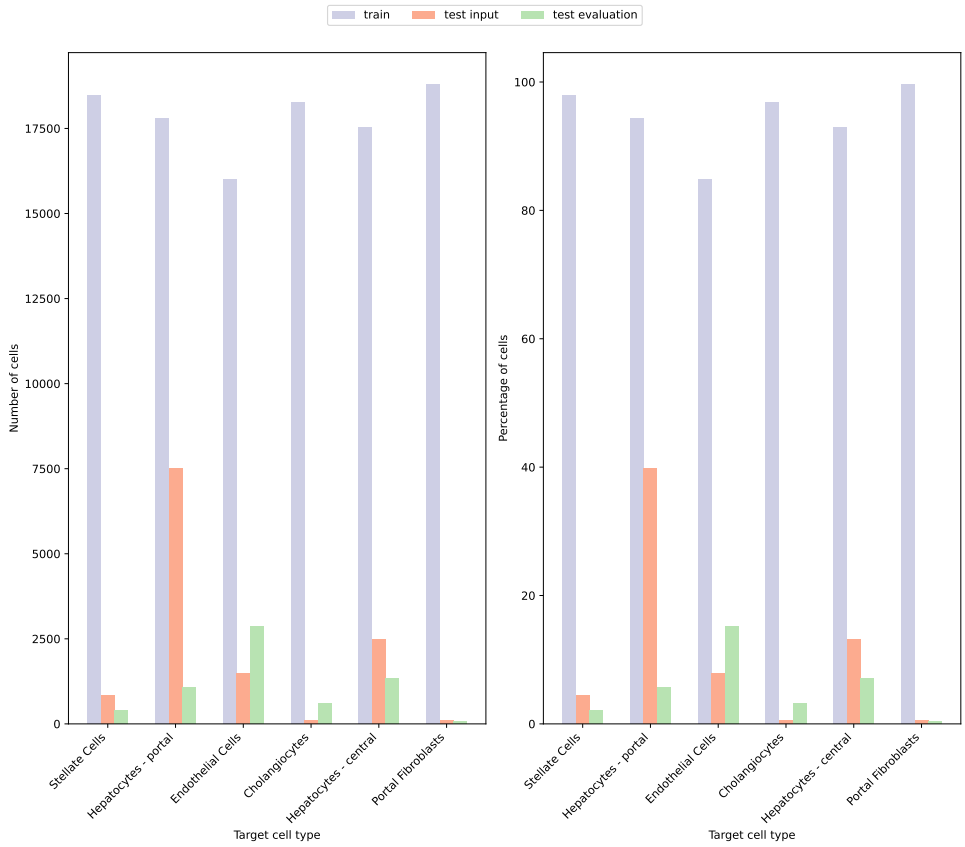
10.1 Multiple doses





10.2 Single dose 30 $\mu g/kg$





10.3 Comparison

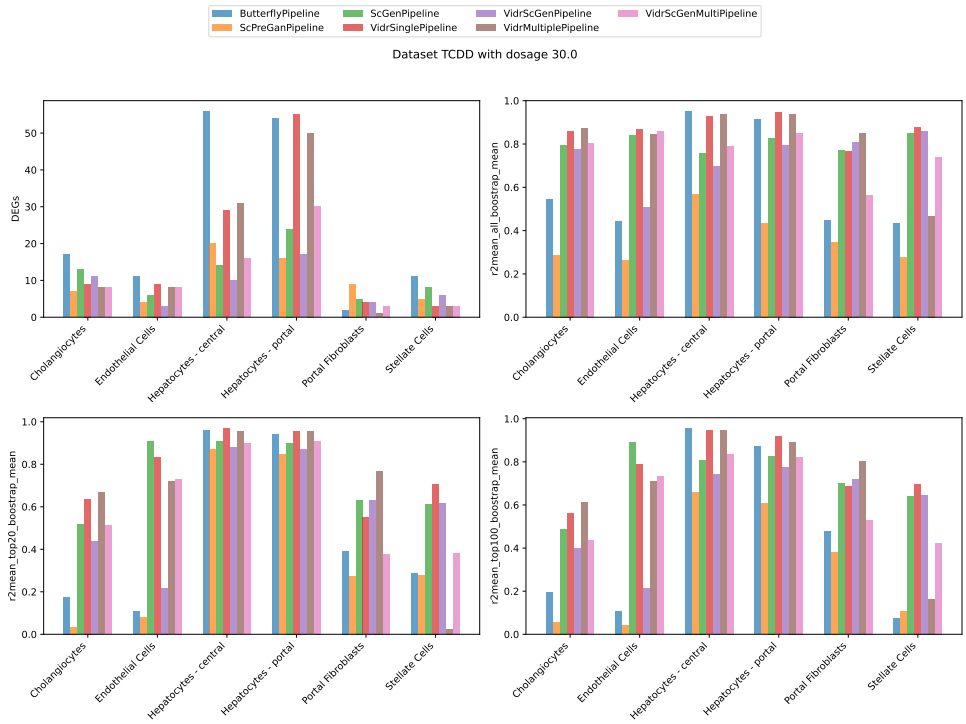


Figure 72: Baseline metrics for highest dosage $30\mu g/kg$ across cell types

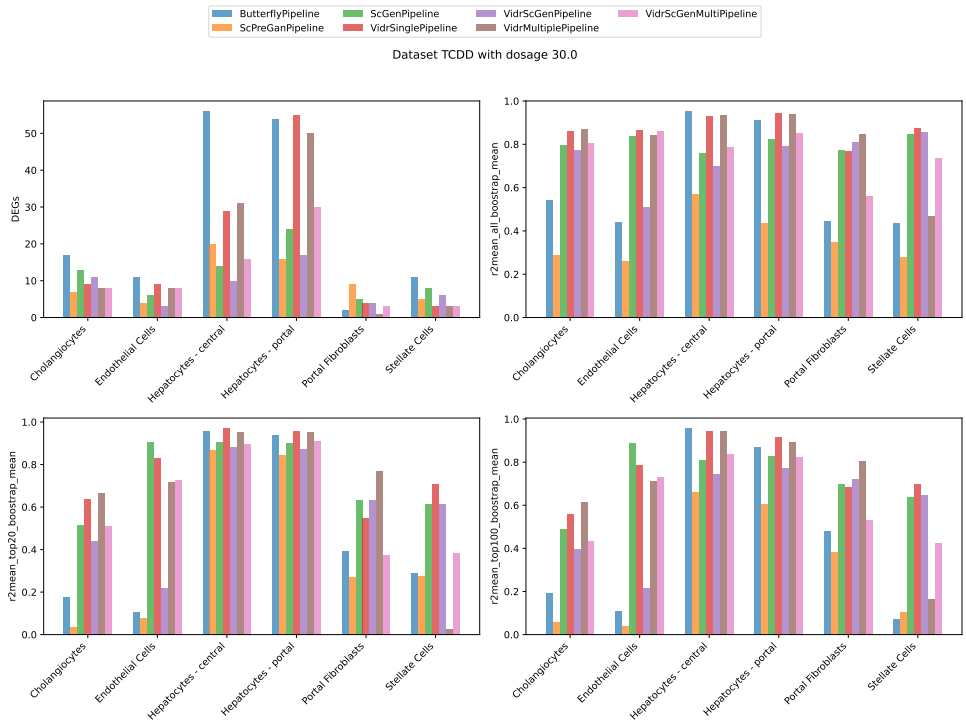


Figure 73: Distance metrics for highest dosage $30\mu g/kg$ across cell types

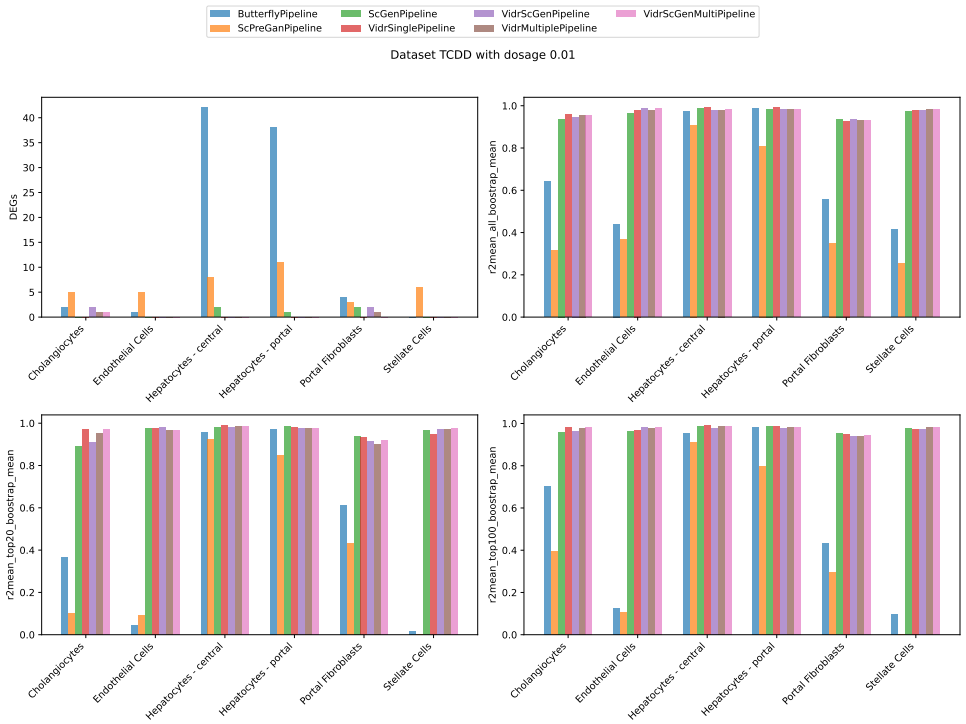


Figure 74: Baseline metrics for highest dosage $0.1\mu g/kg$ across cell types

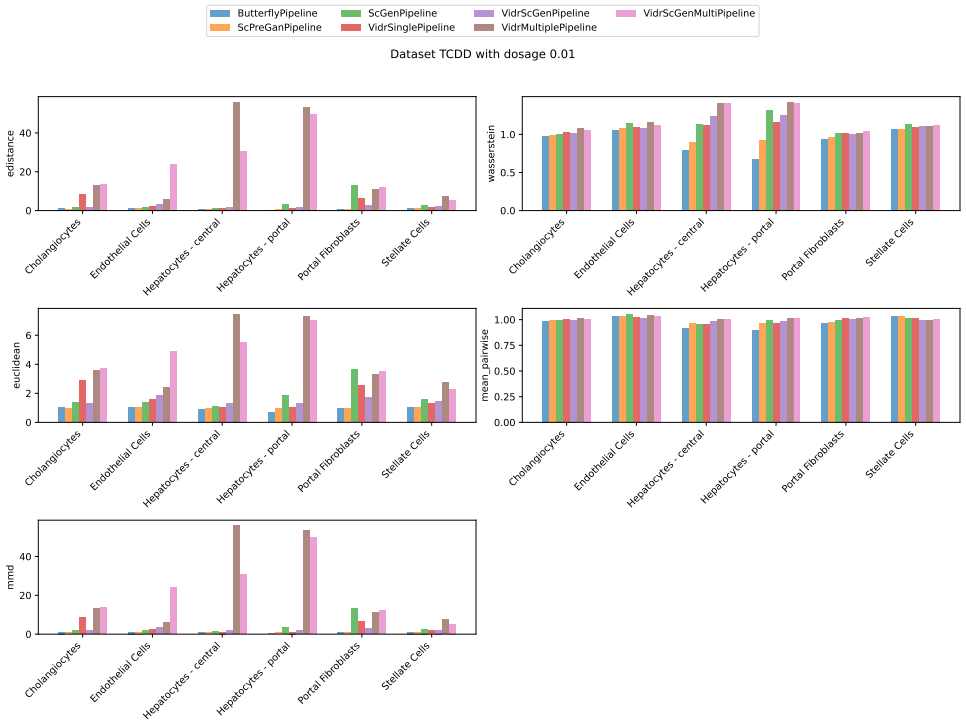


Figure 75: Distance metrics for highest dosage $0.1\mu g/kg$ across cell types

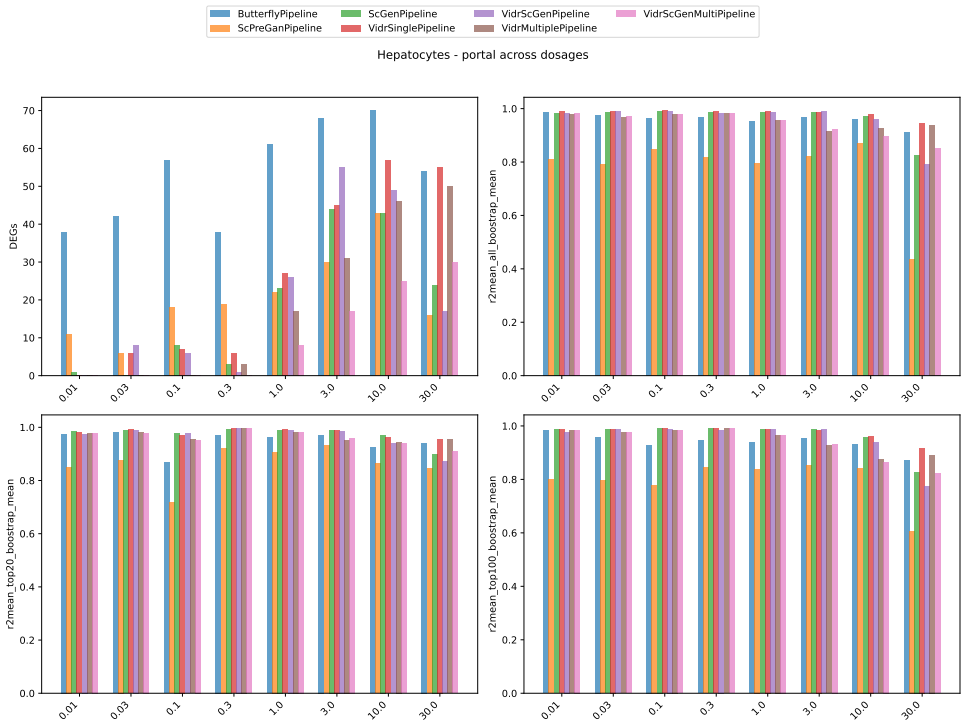


Figure 76: Baseline metrics for Hepatocytes - portal across dosages

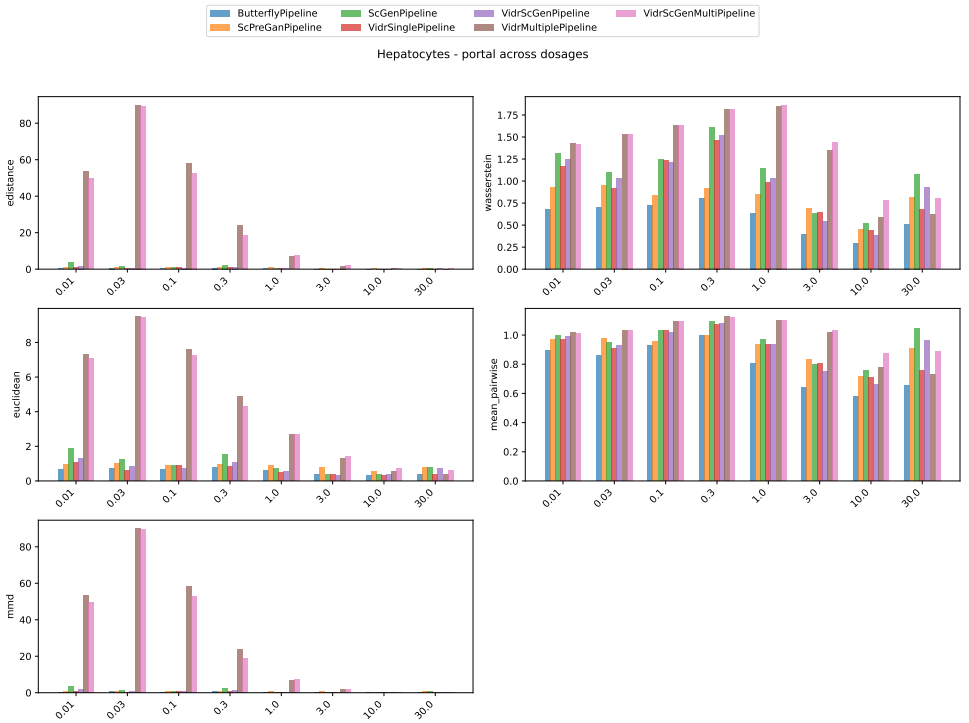


Figure 77: Distance metrics for Hepatocytes - portal across dosages

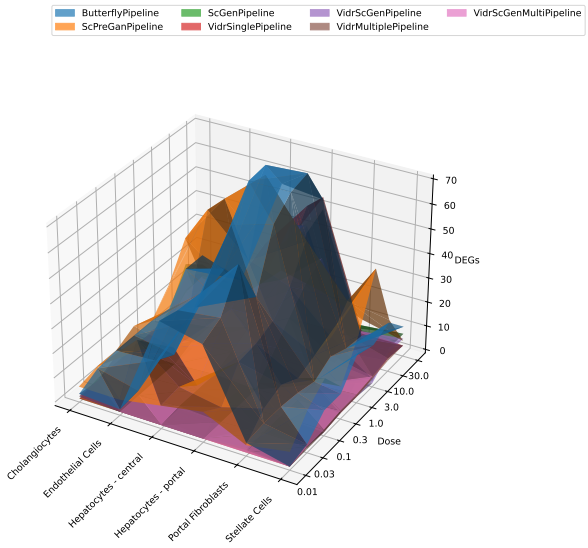


Figure 78

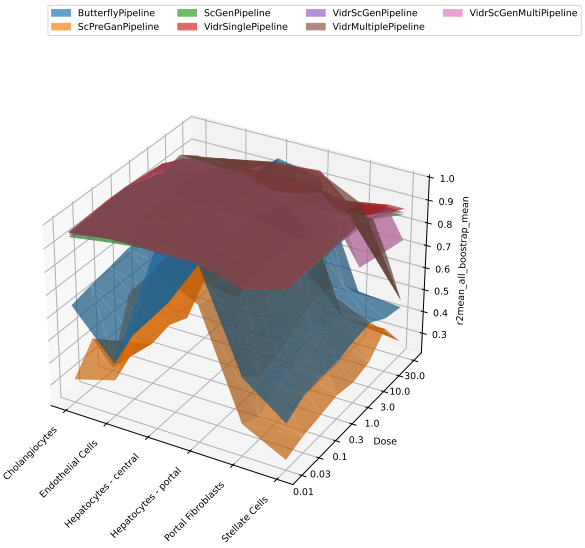


Figure 79

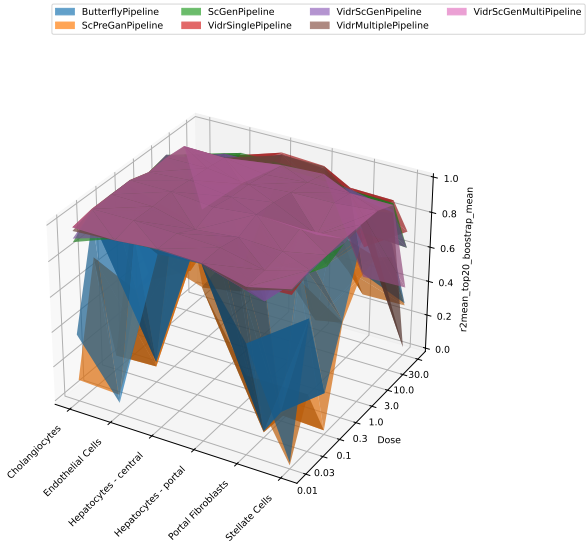


Figure 80

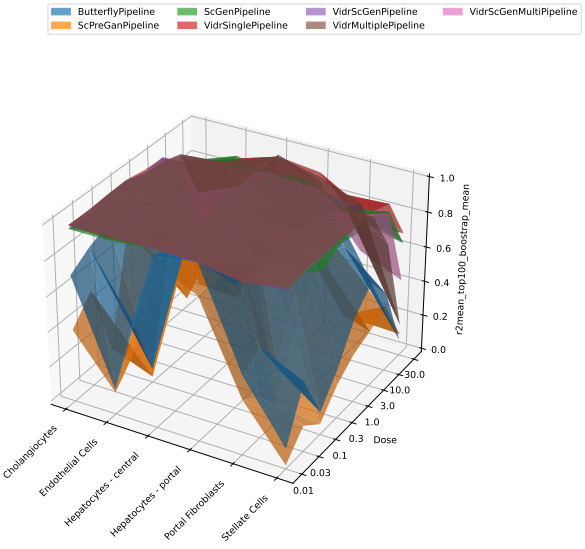


Figure 81

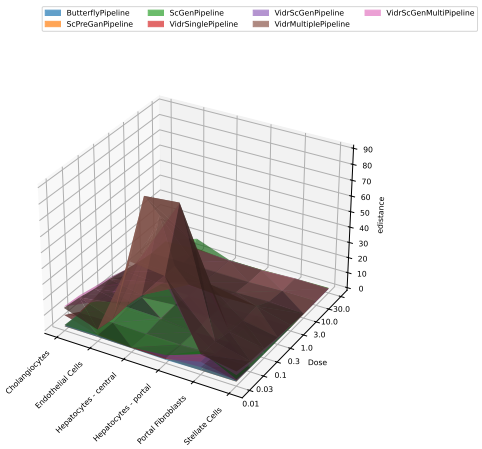


Figure 82: E-distance

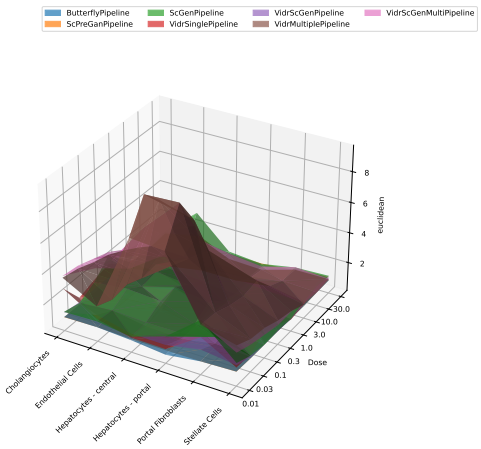


Figure 83: Euclidean

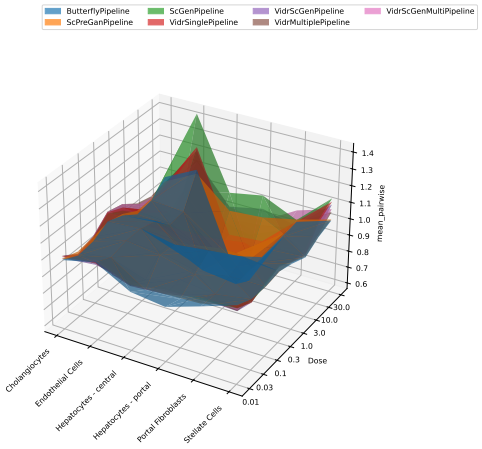


Figure 84: Mean pairwise

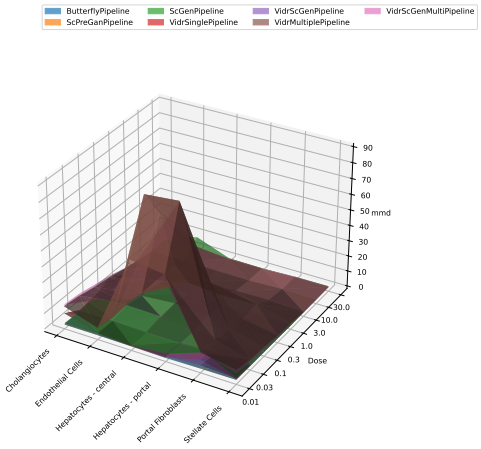


Figure 85: MMD

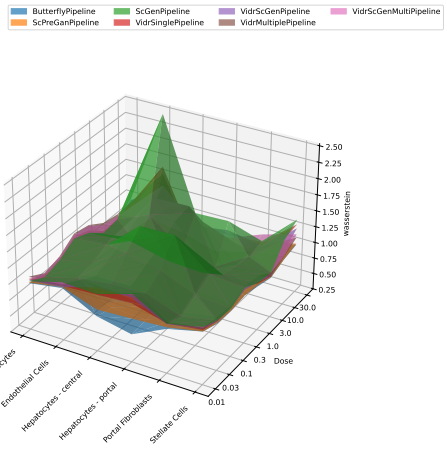
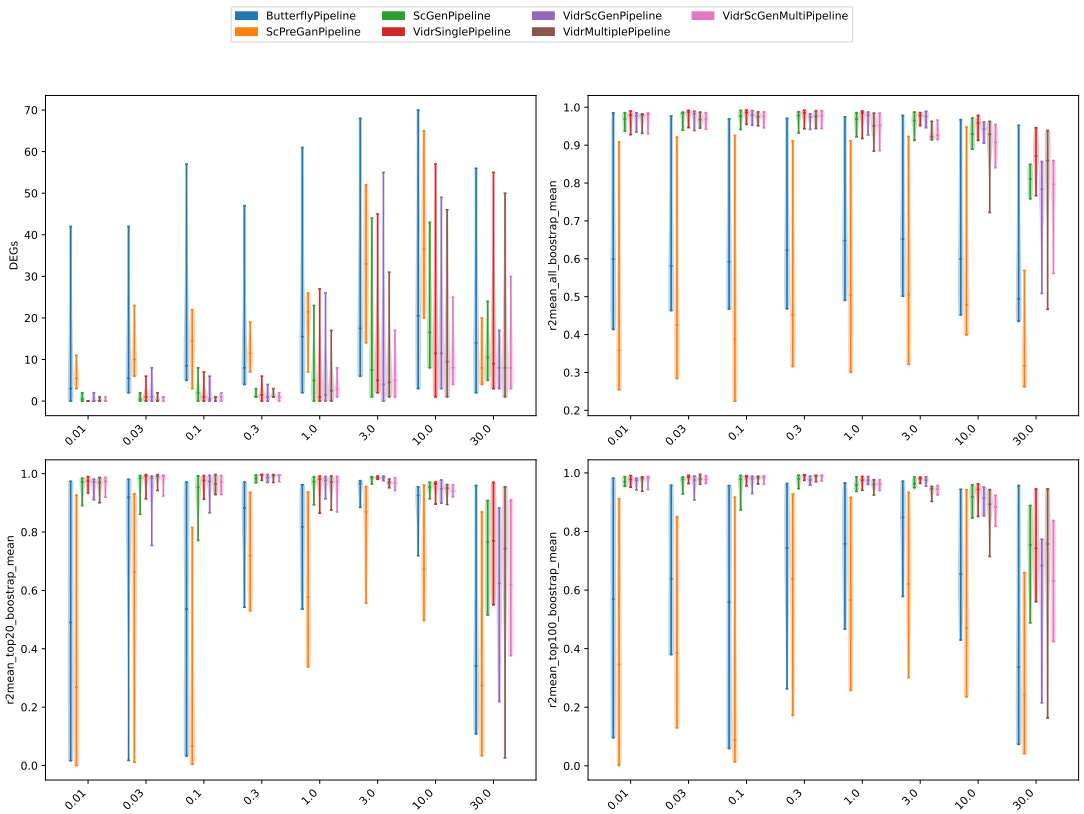
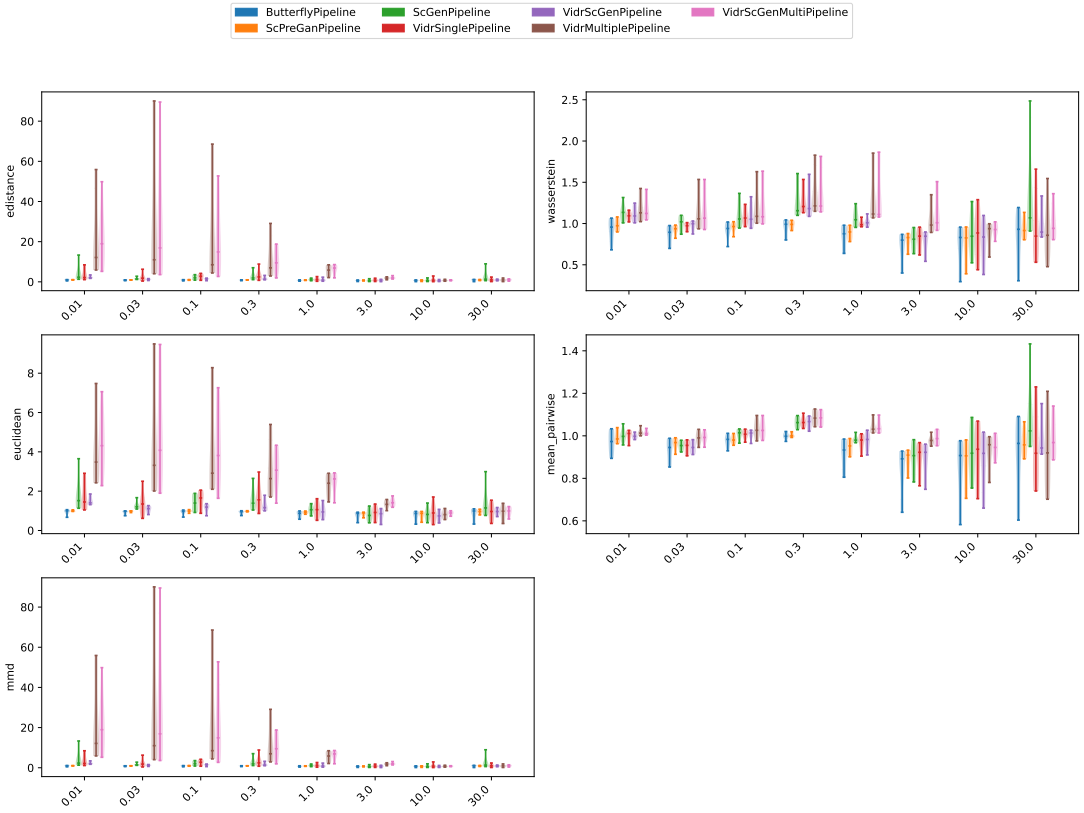
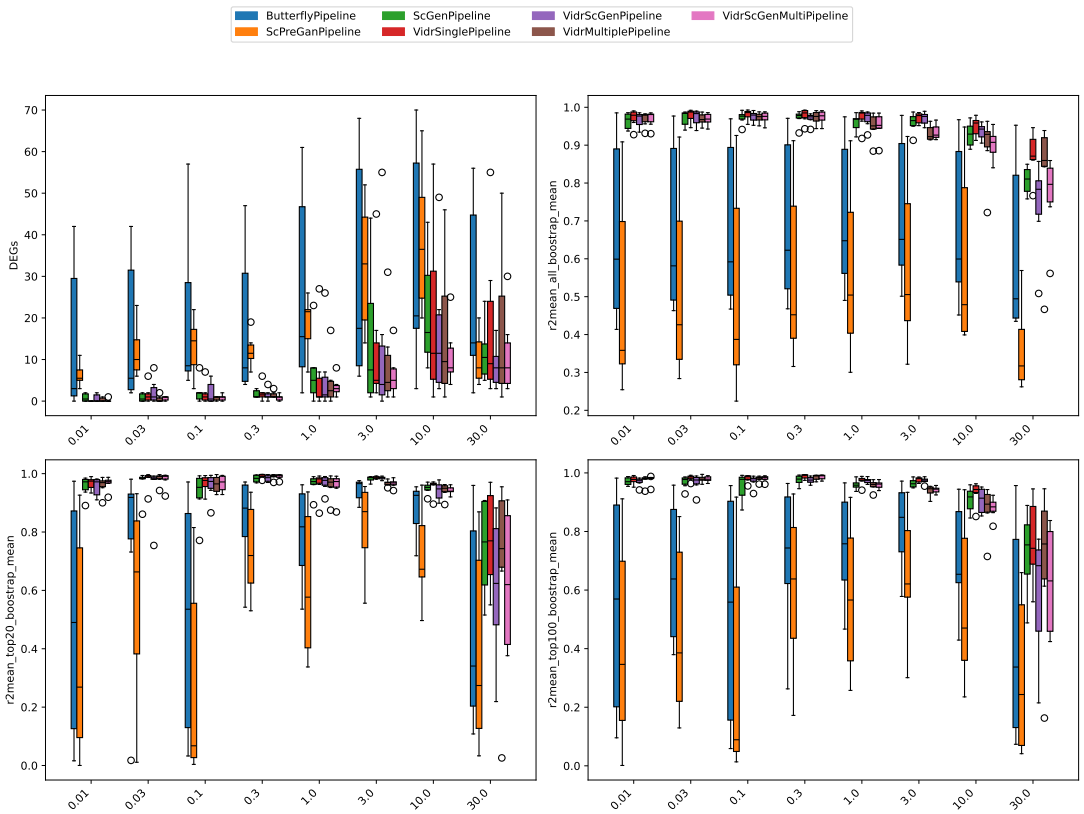


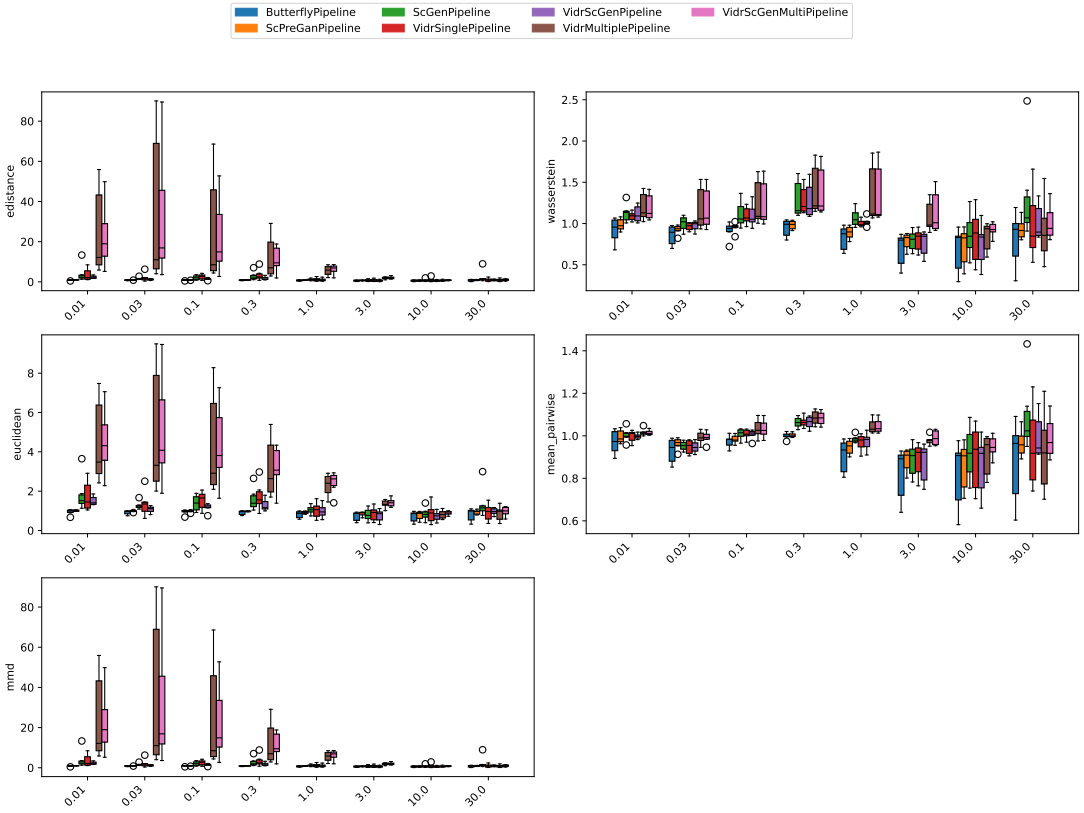
Figure 86: Wasserstein

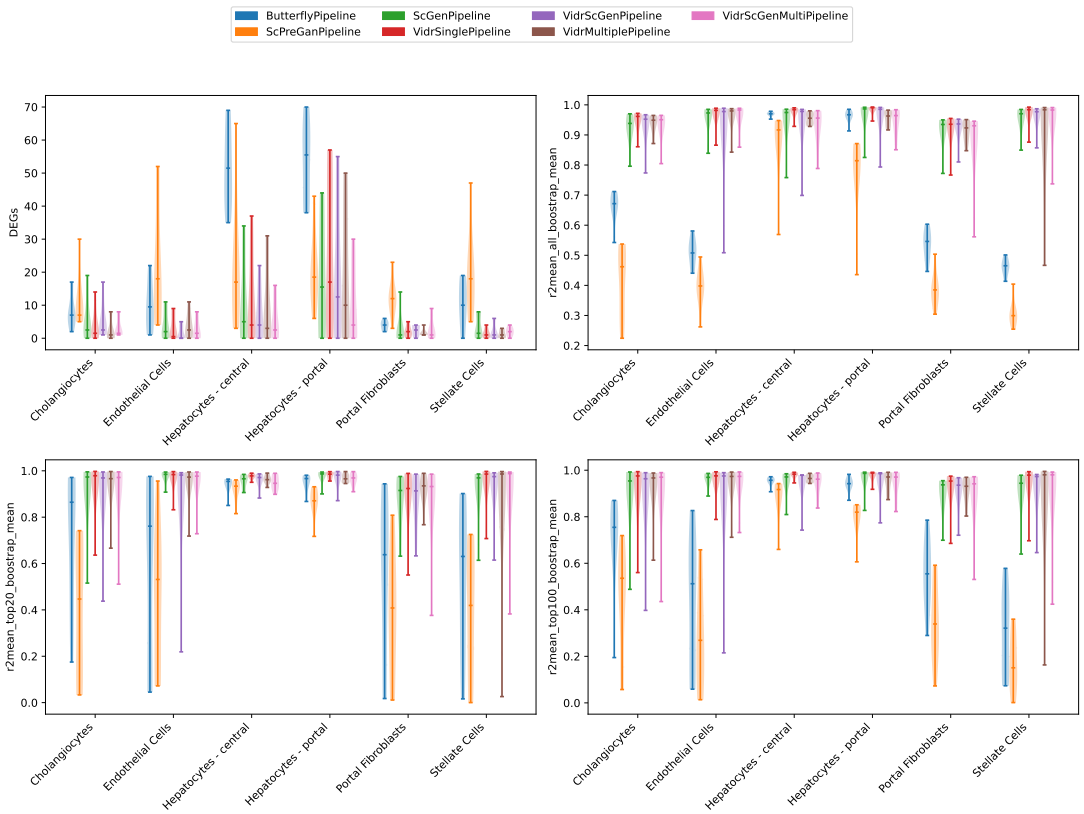
Figure 87: Distance metrics per cell type

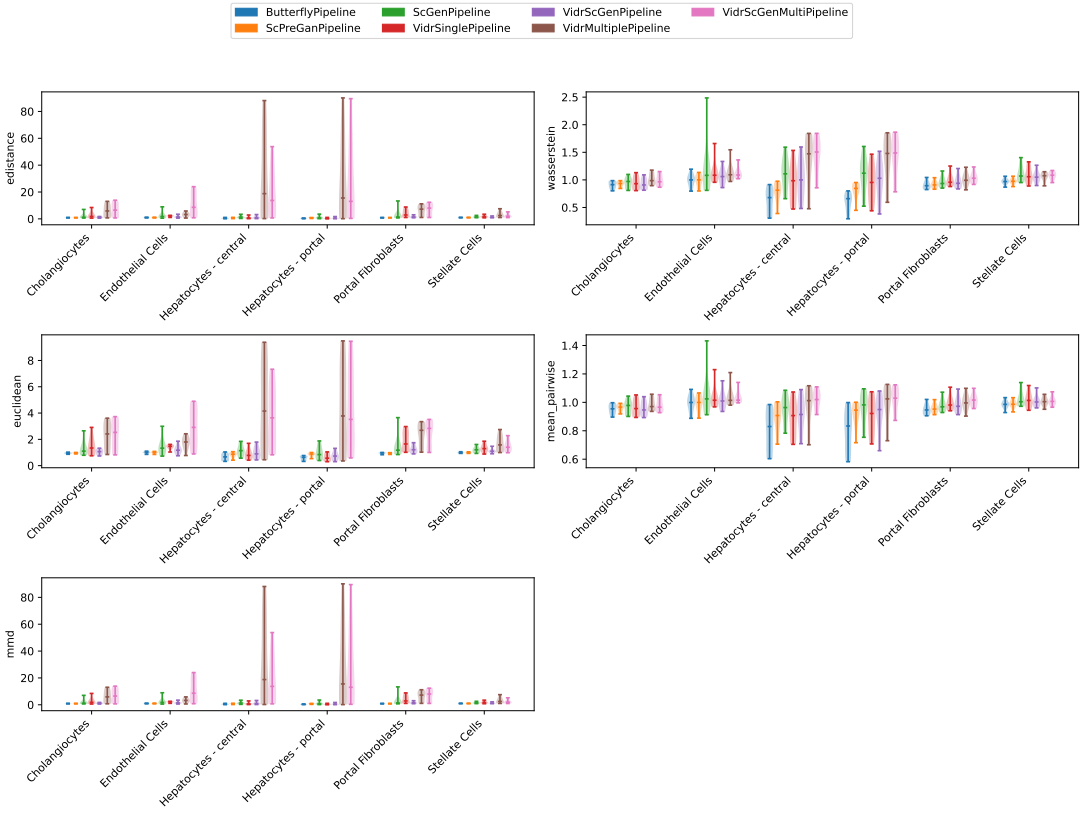

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0

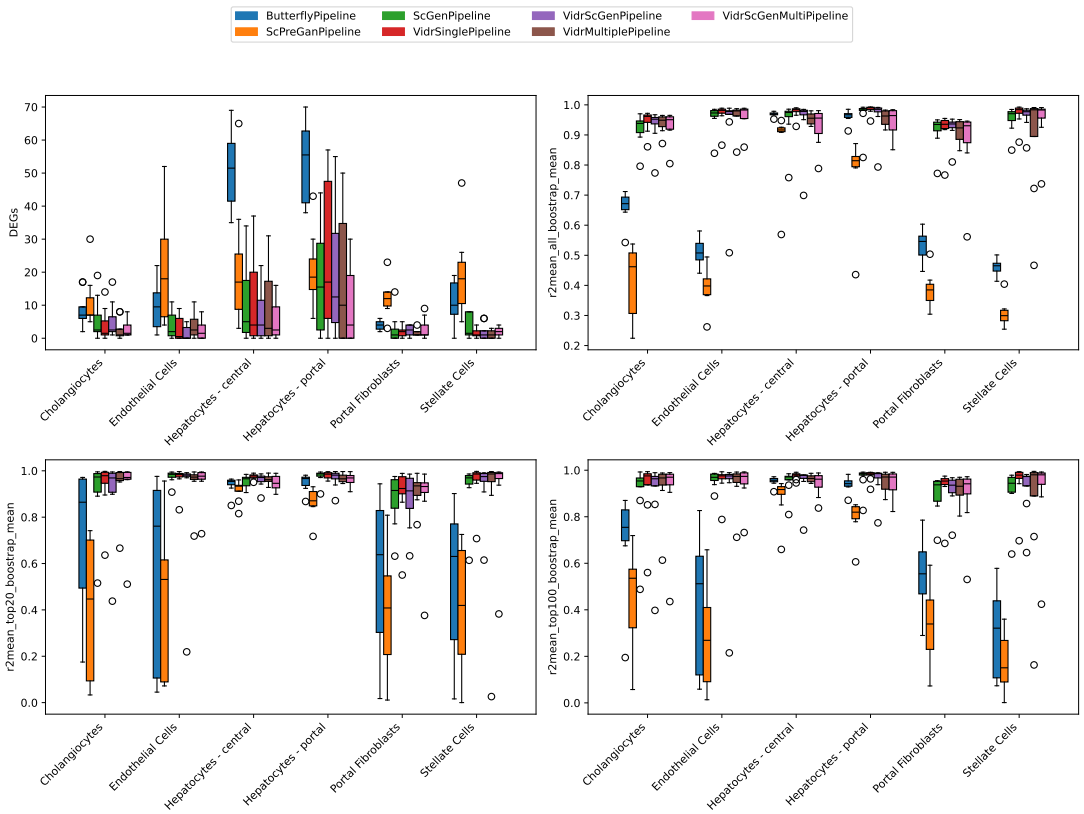

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
edistance
wasserstein
euclidean
mean_pairwise
mmd
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0

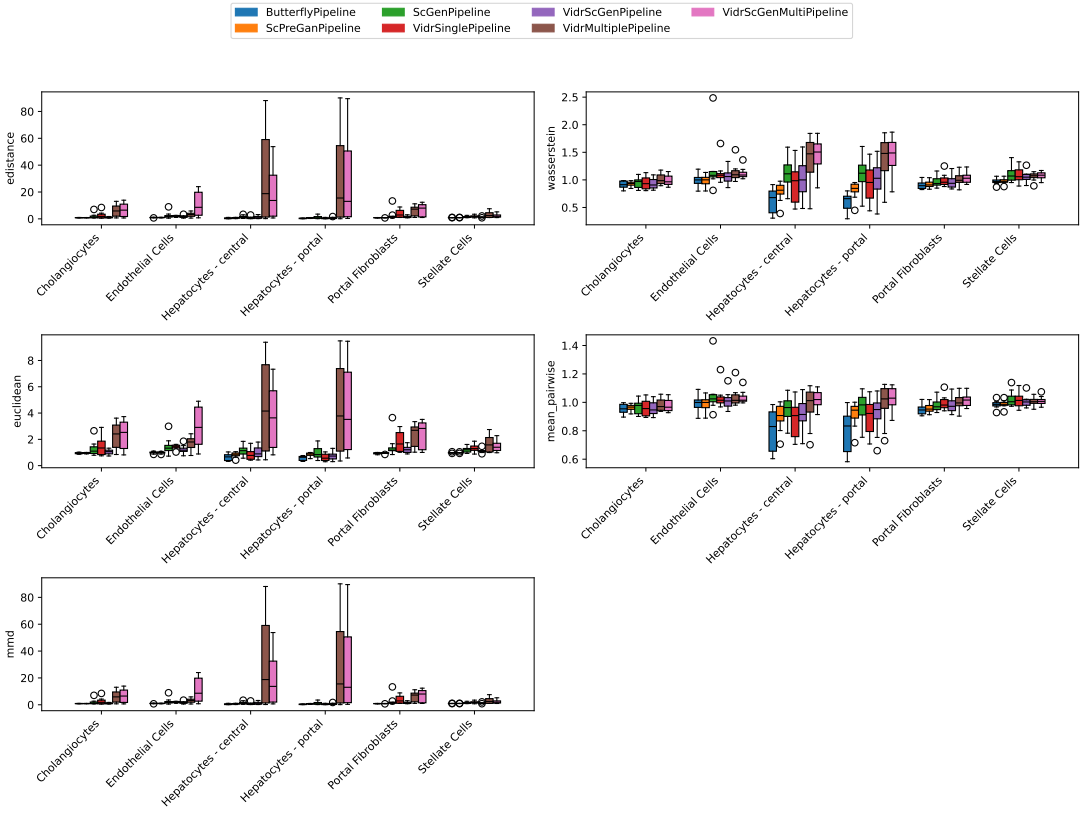

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0


ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
edistance
wasserstein
euclidean
mean_pairwise
mmd
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0


ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSingleCellPipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Portal Fibroblasts
Stellate Cells


ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSingleCellPipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
edistance
wasserstein
euclidean
mean_pairwise
mmd
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Portal Fibroblasts
Stellate Cells


ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Portal Fibroblasts
Stellate Cells


ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
edistance
wasserstein
euclidean
mean_pairwise
mmd
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Portal Fibroblasts
Stellate Cells

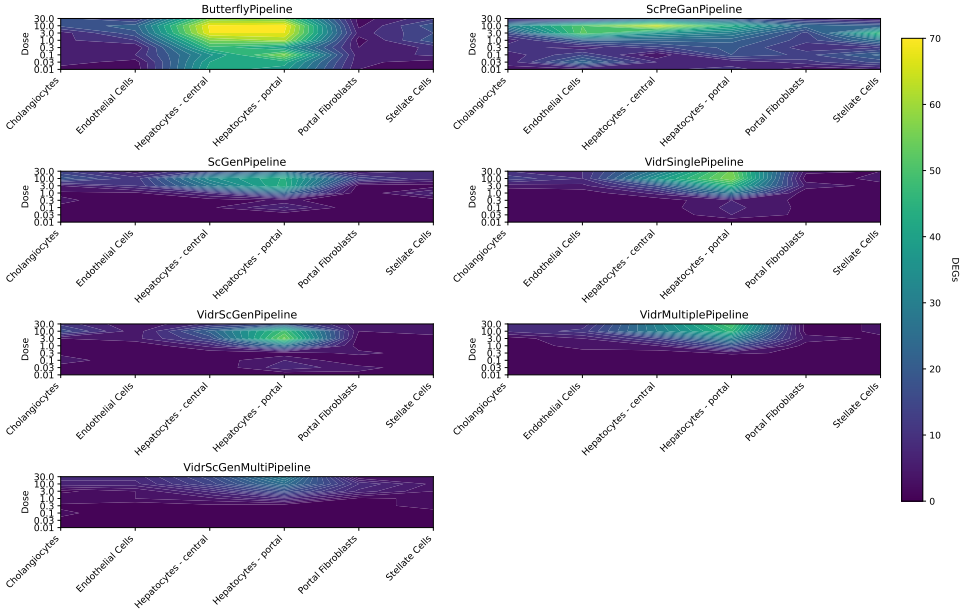


Figure 88: DEGs

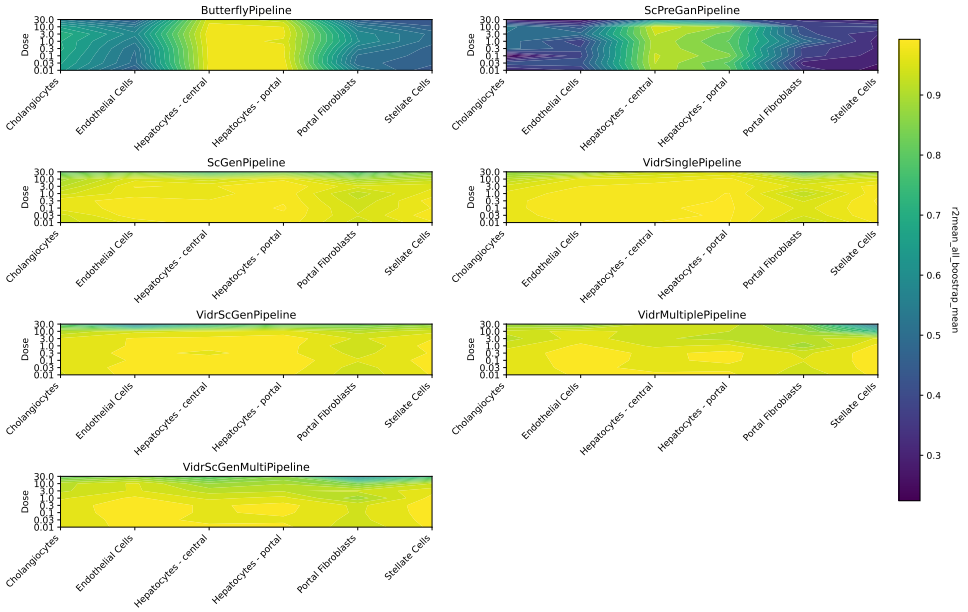


Figure 89: r2 HVGs

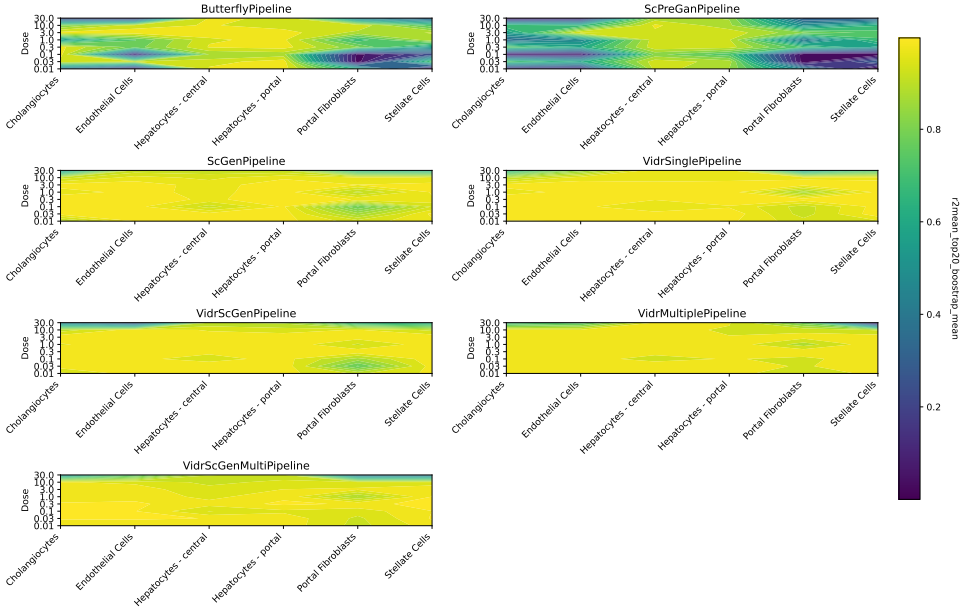


Figure 90: r2 top 20

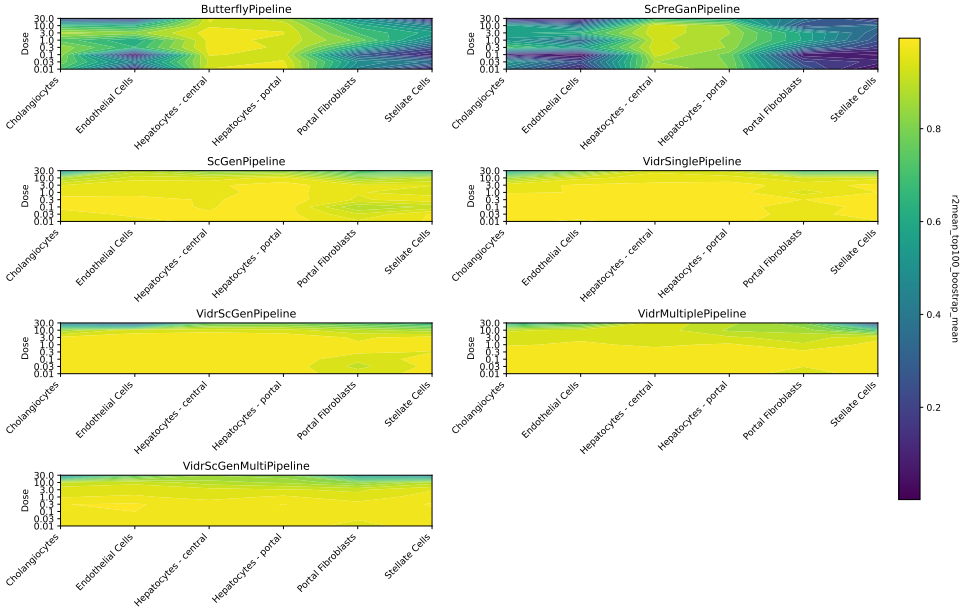


Figure 91: r2 top 100

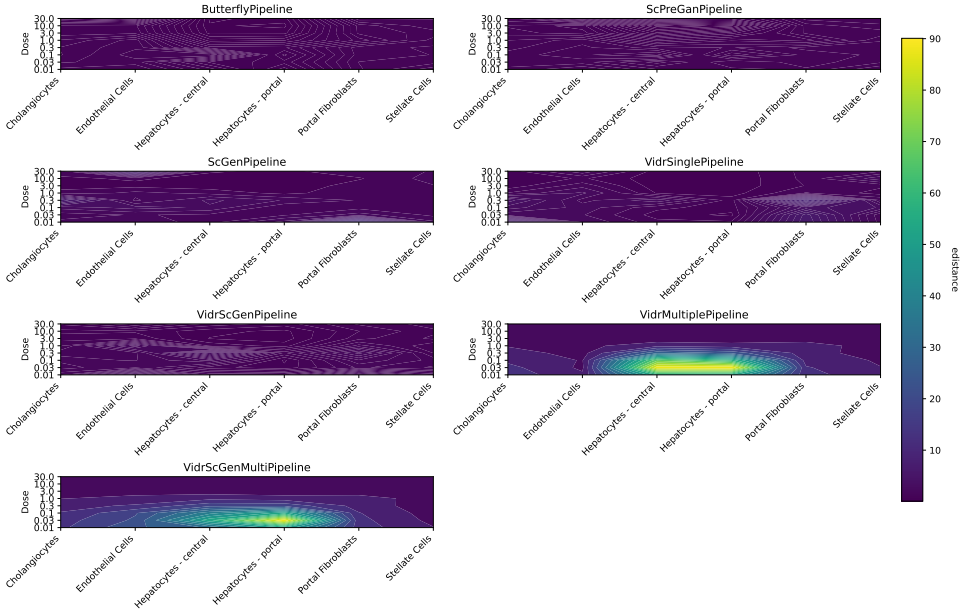
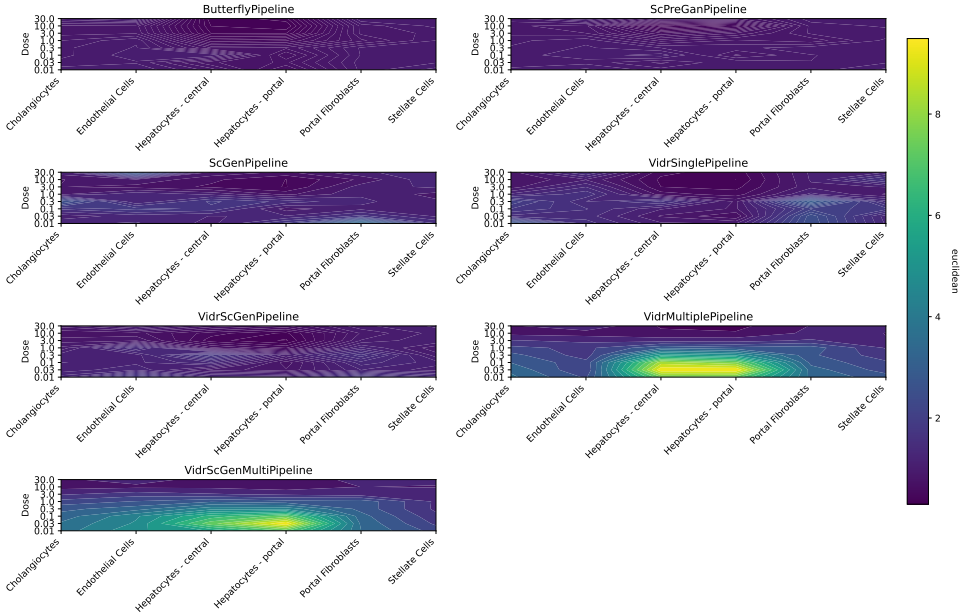


Figure 92: E-distance



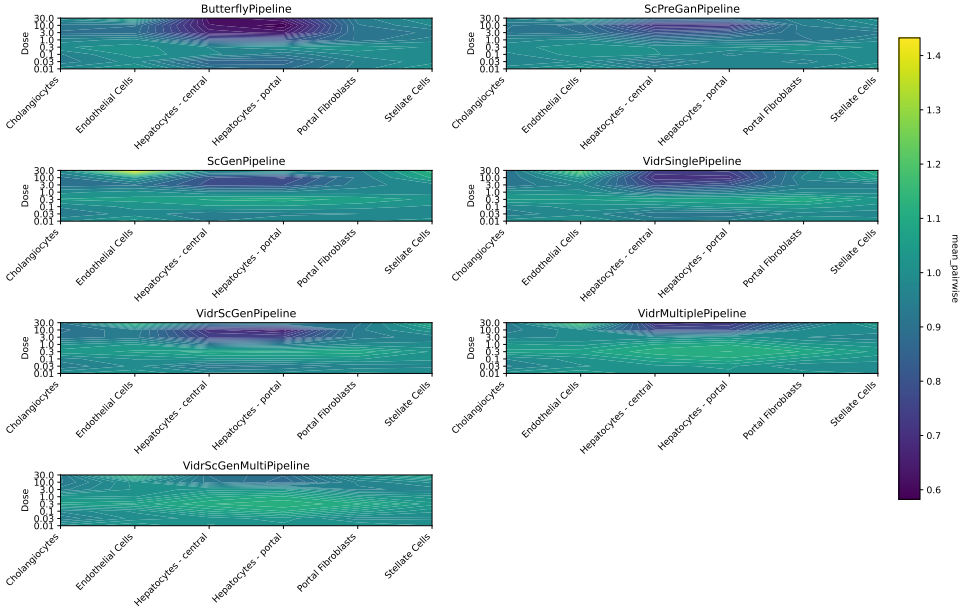


Figure 93: Mean pairwise

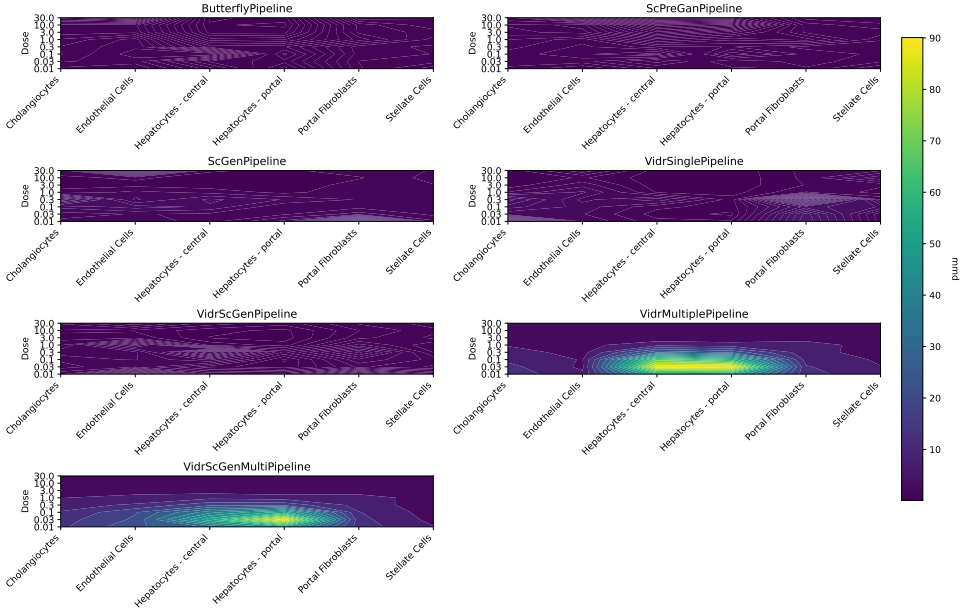


Figure 94: MMD

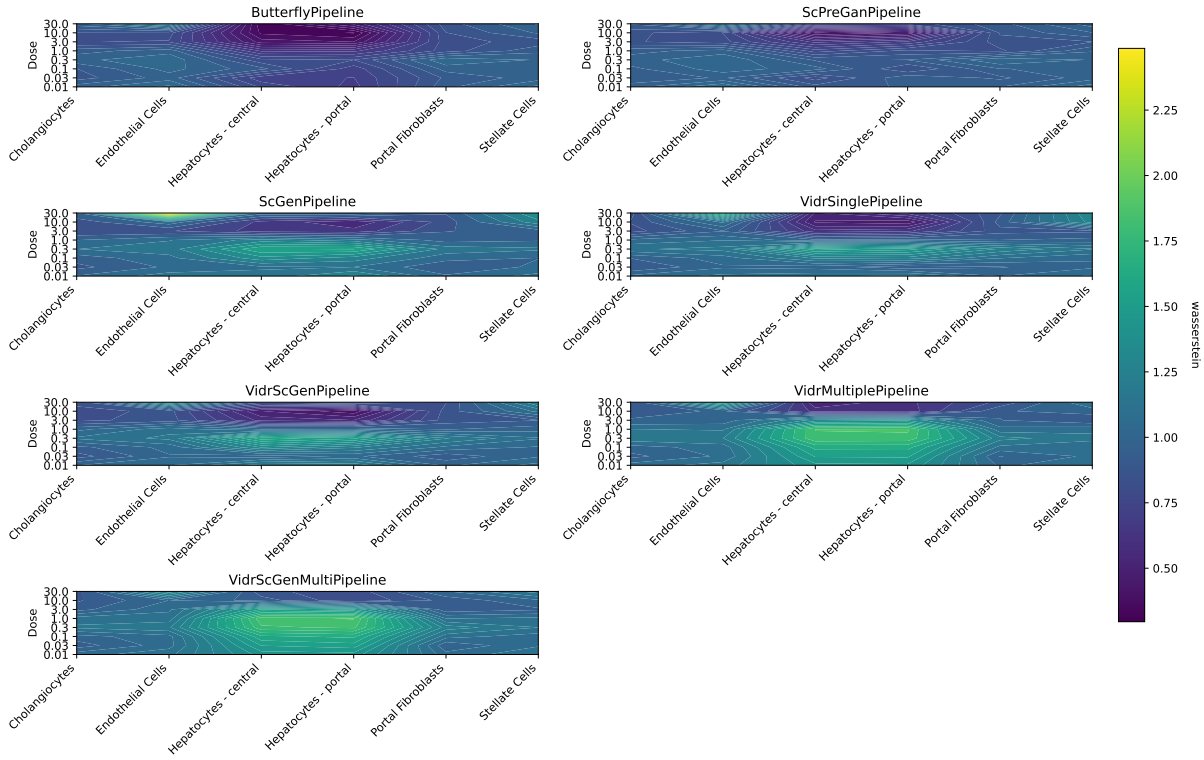
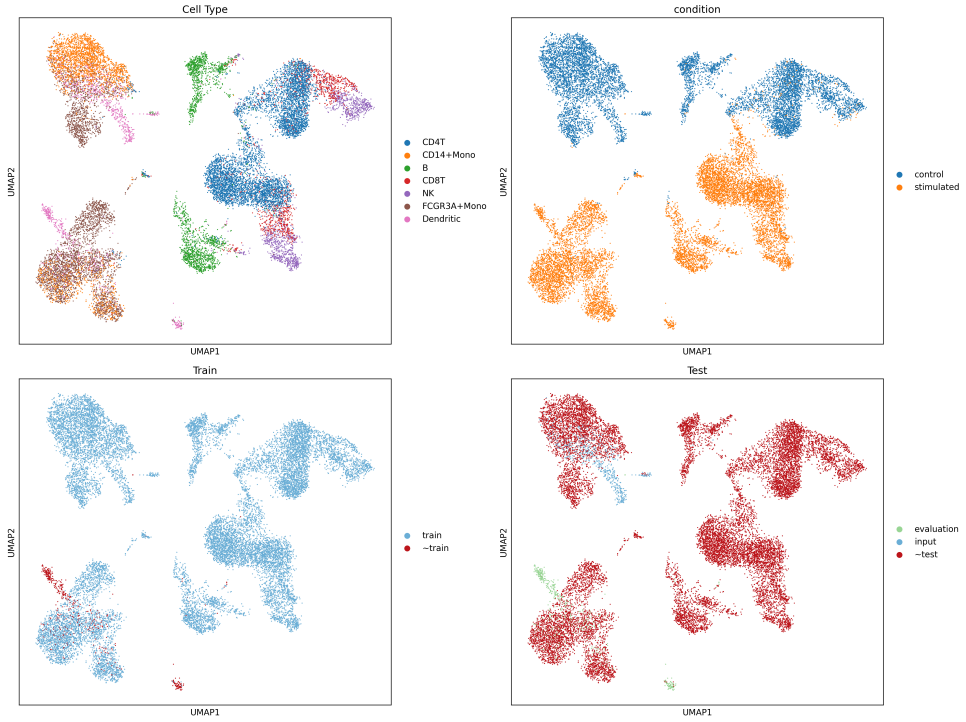


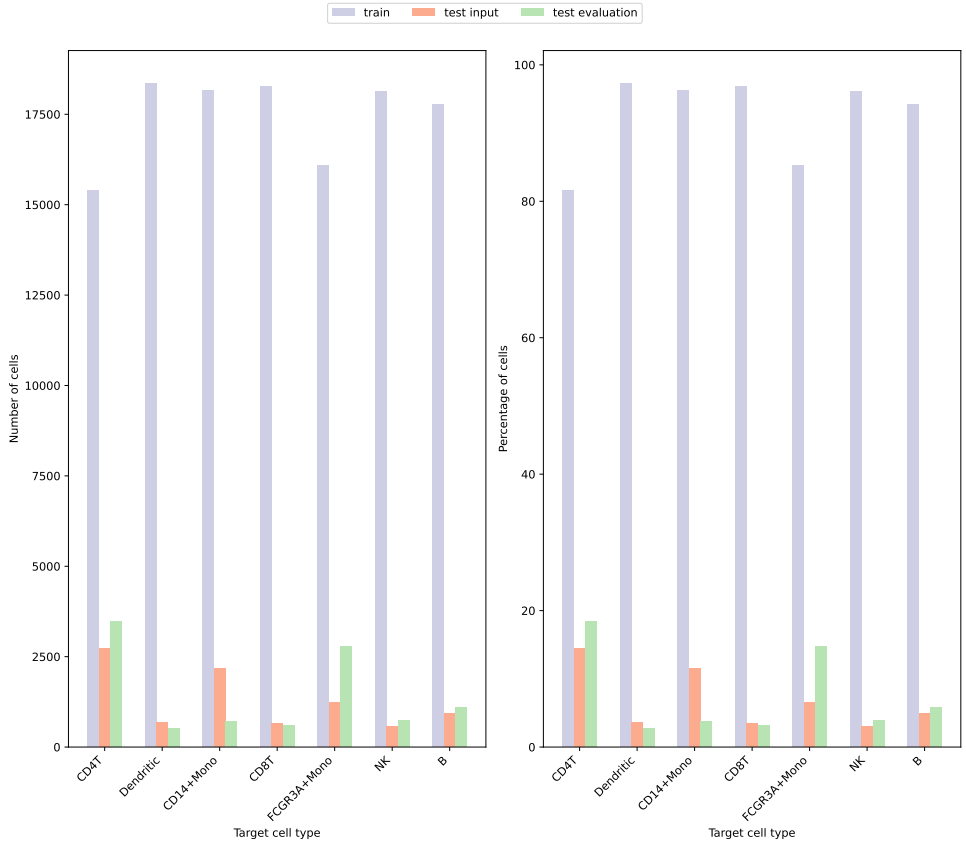
Figure 95: Wasserstein

10.4 Παρατηρήσεις

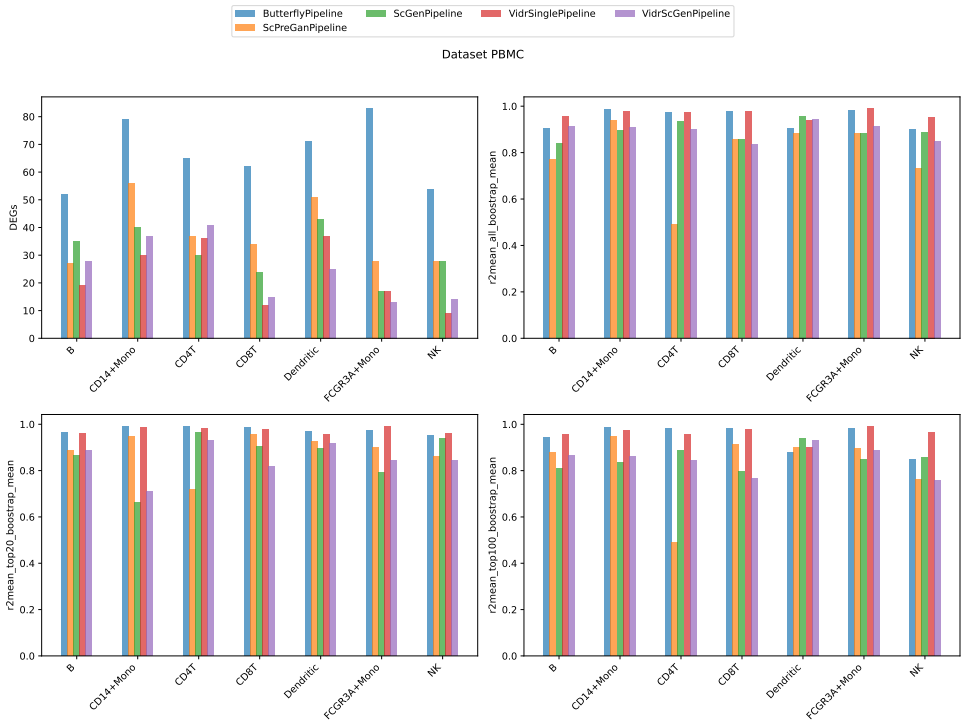
- Το scButterfly και το scPreGan έχουν παρόμοια συμπεριφορά στις μετρικές και εμφανίζουν μεγάλη διακύμανση κατά μήκος των τύπων των κυττάρων και των δόσεων.
- Τα μοντέλα που έχουν ως βάση την αρχιτεκτονική του scGen (scVIDR, και οι παραλλαγές του), VAR και post-processing στο latent space, έχουν την υψηλότερη και πιο σταθερή απόδοση σε μετρικές του R^2, ωστόσο υστερούν στην καταμέτρηση των κοινών διαφοροποιήσιμων γονιδίων έκφρασης (DEGs).

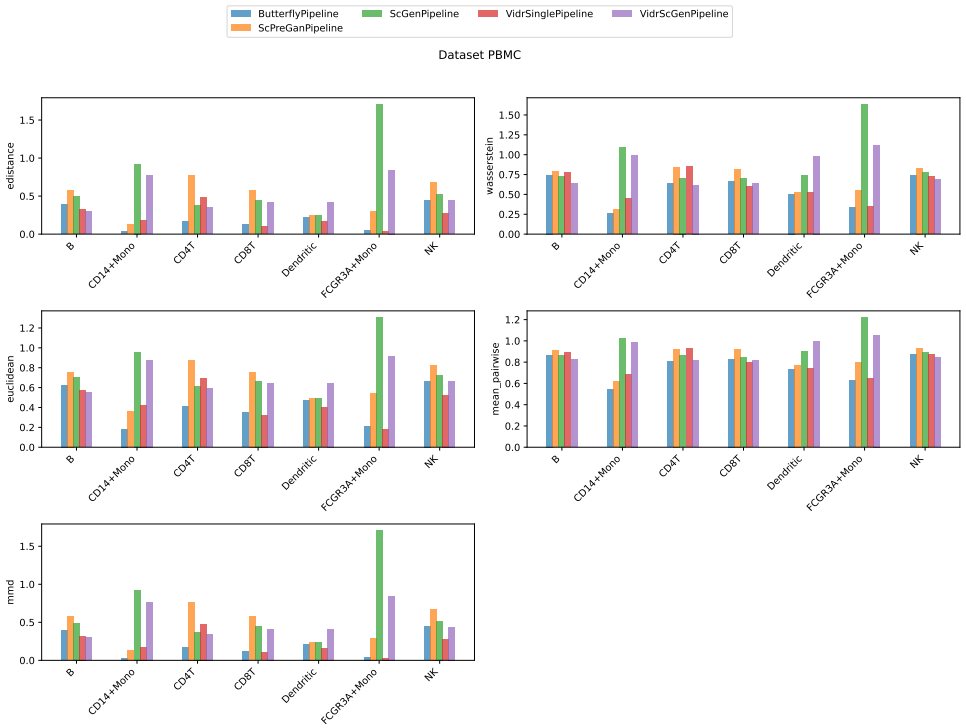
11 PBMC


Cell Type
UMAP2
UMAP1
CD4T
CD14+Mono
B
CD8T
NK
FCGR3A+Mono
Dendritic
condition
control
stimulated
Train
train
~train
Test
evaluation
input
~test


train
test input
test evaluation
Number of cells
Percentage of cells
Target cell type
CD4T
Dendritic
CD14+Mono
CD8T
FCGR3A+Mono
NK
B

11.1 Comparison





Α Ακρωνύμια και συντομογραφίες

LAN Local Area Network

References

[1] George I. Gavriilidis, Vasileios Vasileiou, Aspasia Orfanou, Naveed Ishaque, and Fotis Psomopoulos. A mini-review on perturbation modelling across single-cell omic modalities. 23:1886–1896.

[2] Yuge Ji, Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6):522–537, June 2021.