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 10, 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 mechanisms of cellular heterogeneity. Deciphering the effect of external stimuli (perturbation) at the cellular level, a field referred to as perturbation modeling [4], plays a crucial role in biomedicine and drug discovery. With the recent surge of data generation, machine learning methods aim to understand the effect of perturbations and to extrapolate on unseen events.

An overview of the models on perturbation modeling can be found on this study [3]. 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 [7] 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 Current single-cell perturbation modeling methods

In the literature body, there are several approaches for predicting single-cell perturbation responses. To compare our multi-task method, we have chosen the models of scGen [6], scPreGAN [8], scButterfly [1], and scVIDR [5].

scGen projects the gene expression profile to a probabilistic latent space with a VAE. Then the perturbation effect is modeled with a vector that represents the difference between the control and perturbed gene expression projections in the latent space.

scVIDR builds upon scGen, complementing the architecture with cell type specific knowledge. It can predict the response of multiple chemical perturbations on a dose-dependent use case. scVIDR is one of the instances that is leveraging the data of multiple perturbations to improve the predictions, based on a multi-task approach.

scPreGAN is based on a GAN and autoencoder setup. The study aims to decouple the perturbation effect from the latent space, and to apply it to the decoder stage.

scButterfly, although hasn't not been primarily designed for perturbation modeling, its cross modal architecture with dual aligned VAEs can also be used for perturbation objectives. Instead of cross predicting one omic from another one, the perturbed and control gene expressions can be considered as modalities.

4 Method

Multi-task learning is a machine learning paradigm and its core idea is that training a model to solve multiple tasks can be more effective than training separate models for each specific task [9]. A joint architecture that shares knowledge between the tasks can capture underlying

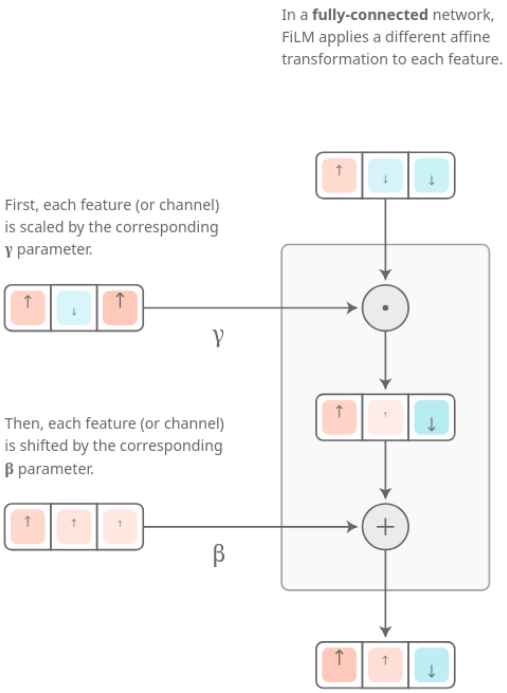


Figure 1: Illustration of the feature-wise transformation [2]

dynamics leading to an enhanced generalization performance. The relationship of the tasks determines the positive or negative transfer to each other and the overall effectiveness of the paradigm.

Defining as a task the prediction of the gene expression given a perturbation, we will explore designing a model that can predict gene expressions given multiple perturbations and its performance compared to single perturbation approaches.

One of the key problems of deep learning methods is the data demand. Another benefit of multi-task learning is the combination of data from multiple sources of informations, especially in perturbation modeling where the data is limited for a specific number of perturbations.

To integrate the tasks, we have explored the application of feature-wise transformations [2]. For this kind of transformation, we have:

$$\text{FiLM(x)} = \gamma(z) \odot x + \beta(z)$$

, where γ, and β are learnable parameters generated by a network that represent a condition, z is the condition, such as a vector that indicates the task, and x is the input.

This particular technique is referred to as conditional affine transformation (a combination of multiplicative and additive conditioning) that shifts and scales the input element-wise. It is efficient in terms of scaling and parameters compared to multi-head architectures, where each task has its dedicated network to generate the output of the task.

In our approach, we have attempted to decouple the perturbation effect, creating a perturbation free latent space, while taking control of the perturbation response with the conditioning vector. Thus, our architecture is based on an autoencoder, and the conditioning of the task, which in our case is the type of perturbation, will be delegated via FiLM layers fused in the decoder (`MultiTaskAutoencoder`). The γ and β are designed to be different for each

fusion. We have experimented with a few variations of this, keeping the architecture of the decoder with the inclusion of film layers consistent. Thus, we have the following models:

- The `NultiTakAae` is based on an adversarial autoencoder. The discriminator aims to differentiate between control and perturbed gene expressions, and the encoder is trained to fool the discriminator.
- The `MultiTaskAaeGaussian` is based on an adversarial autoencoder, that aims to create a latent space that follows the gaussian distribution. The discriminator aims to differentiate between samples of the gene expressions and the gaussian distribution.
- add the rest

5 Evaluation

To evaluate the models, we have used the count of DEGs, along with the R^2 of all the DEGs and the top 100 most variable ones. To complement the evaluation, based on scPerturb, we have calculated a set of five distance metrics, to capture the wholeness of the differences between the expected and predicted perturbed gene expressions.

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 three different seeds 1, 2, 19193, and the metrics have been averaged across experiments.

6 Results

6.1 Knowledge transfer

which tasks and why they are important?

6.2 TODO

- batch effect
- interpretability
- explainability
- integration with multiple omics

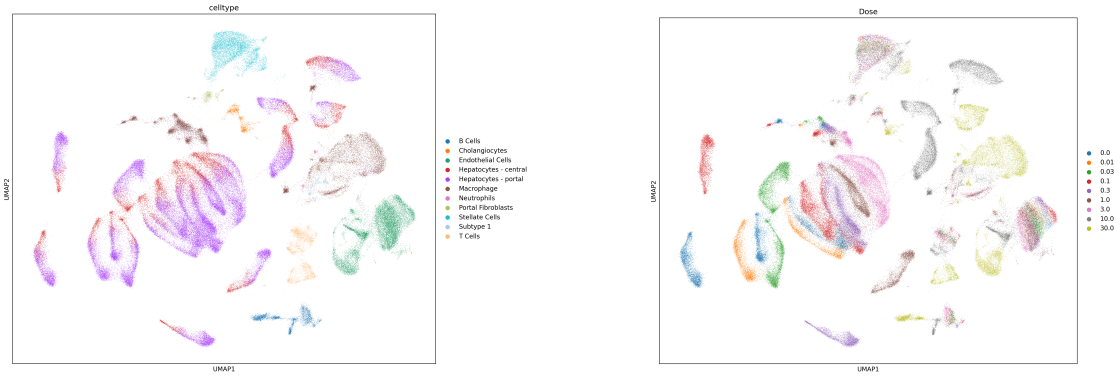
7 Conclusions

8 Future work

9 Benchmarking

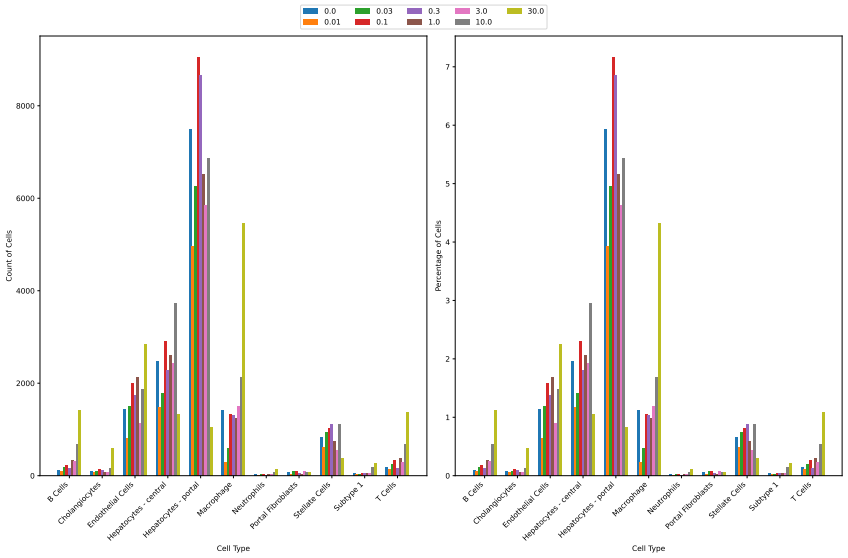
10 Datasets

10.1 Nault et al. 2022

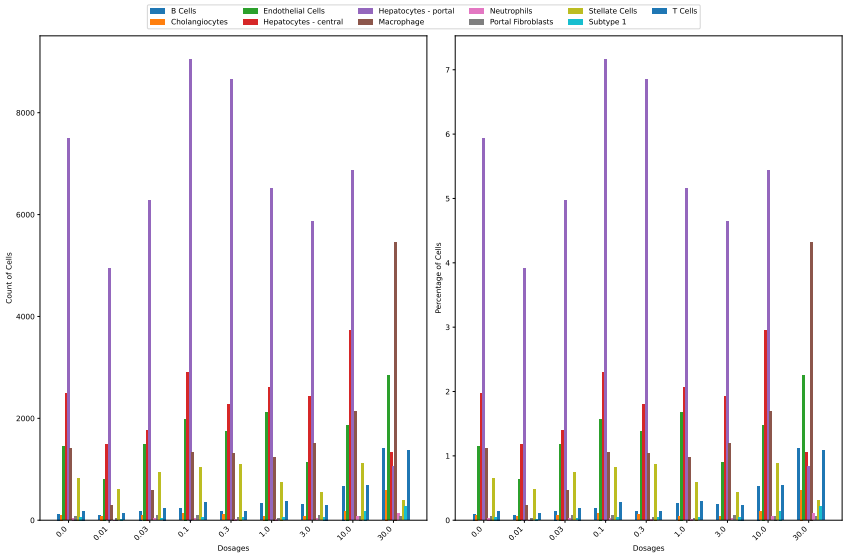


(a) (b)

.



(c)



(d)

Figure 2: Nault overview

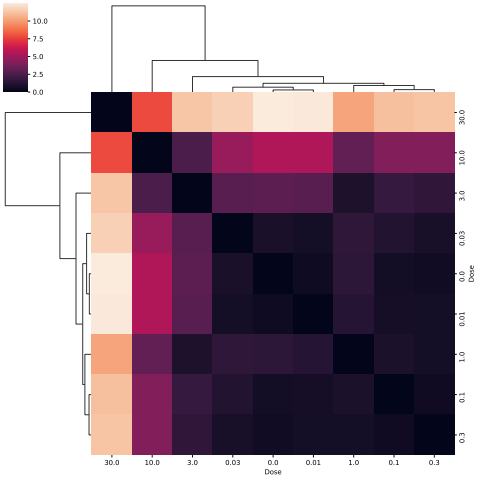


Figure 3: E-distance

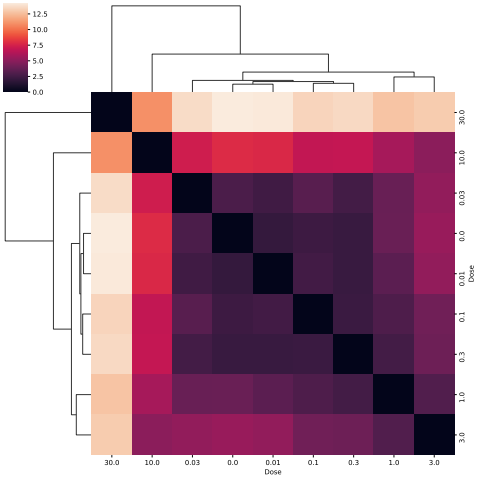


Figure 4: Euclidean

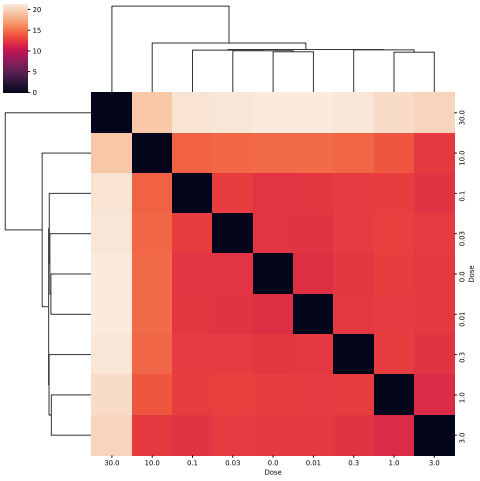


Figure 5: Mean pairwise

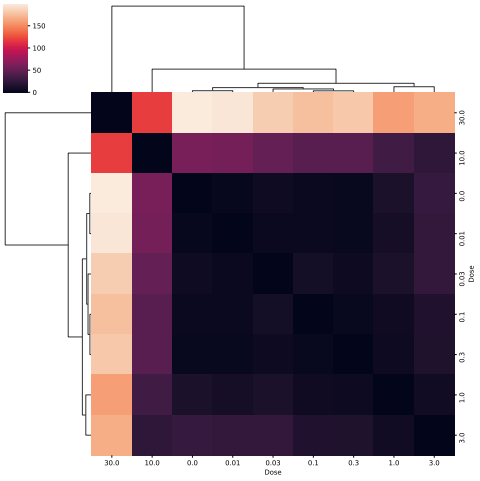


Figure 6: MMD

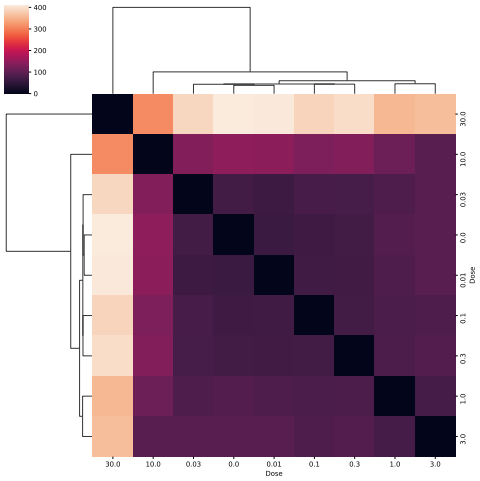


Figure 7: Wasserstein

Figure 8: Distance metrics across all cell types per dosage

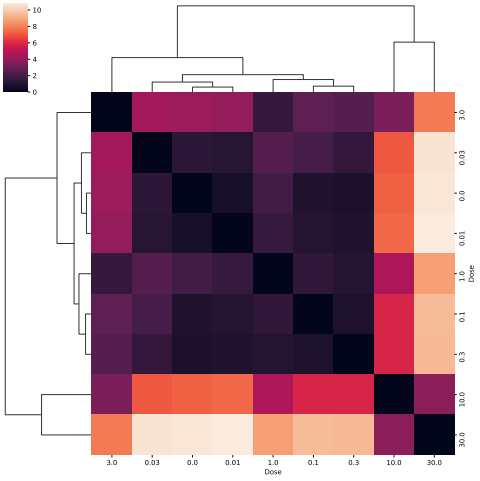


Figure 9: E-distance

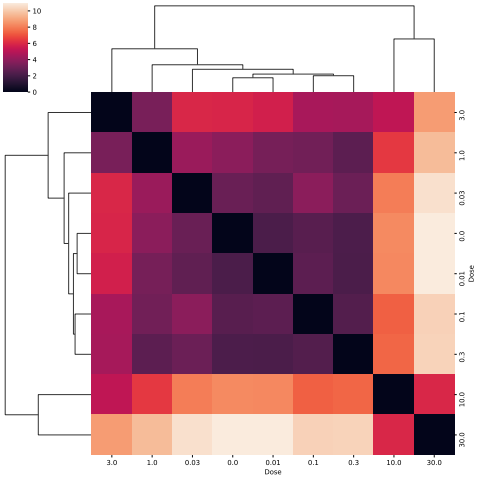


Figure 10: Euclidean

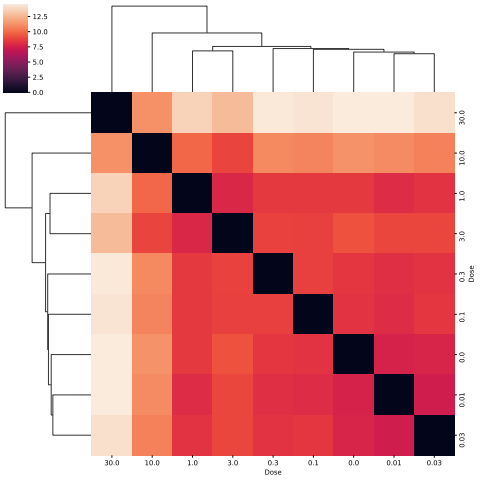


Figure 11: Mean pairwise

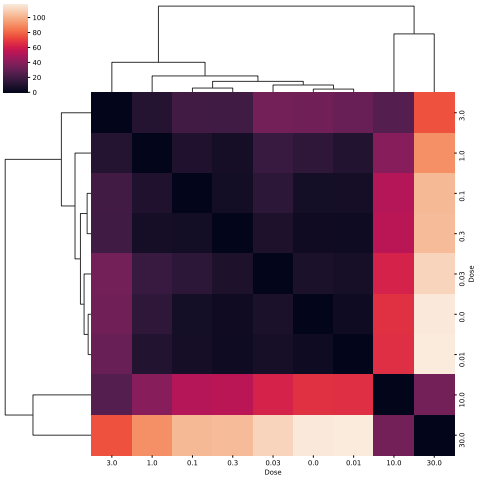


Figure 12: MMD

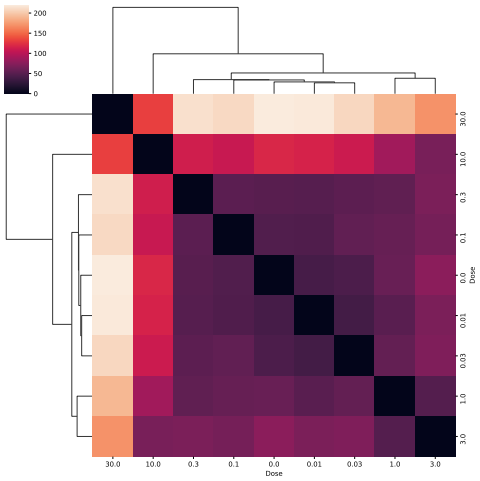


Figure 13: Wasserstein

Figure 14: Distance metrics for cell type Hepatocytes - portal per dosage

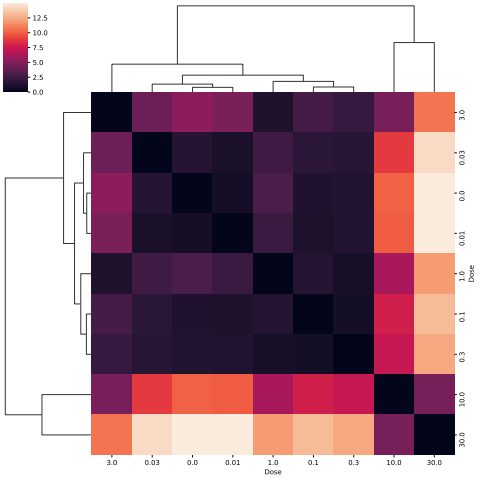


Figure 15: E-distance

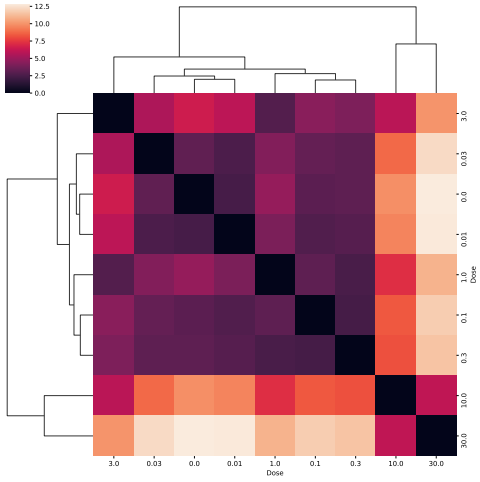


Figure 16: Euclidean

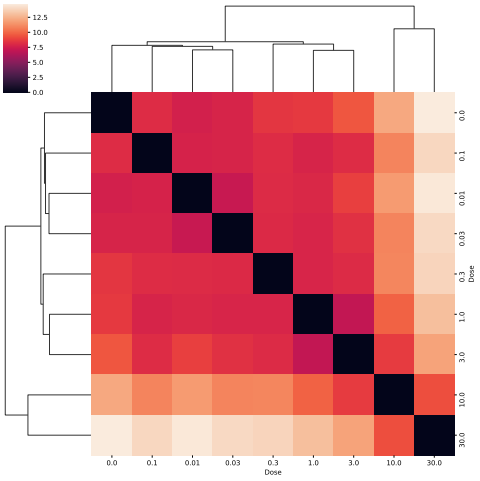


Figure 17: Mean pairwise

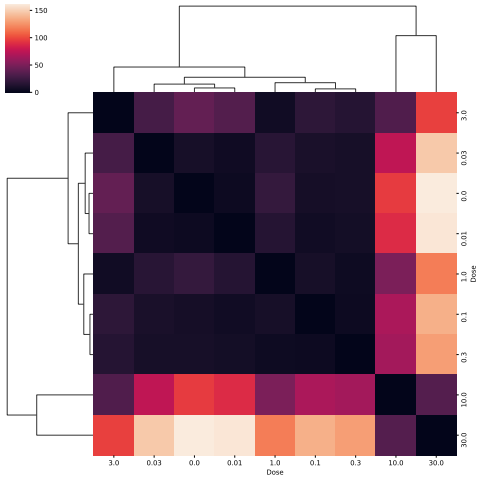


Figure 18: MMD

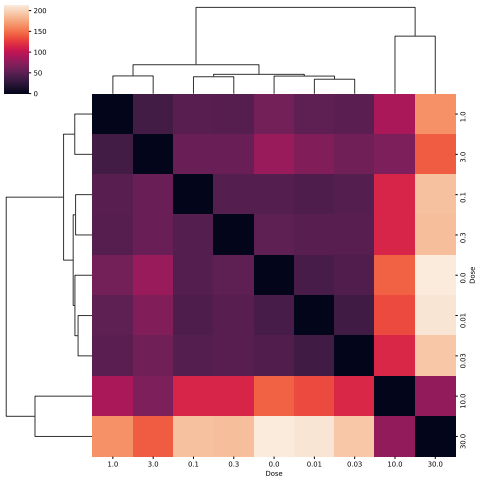


Figure 19: Wasserstein

Figure 20: Distance metrics for cell type Hepatocytes - central per dosage

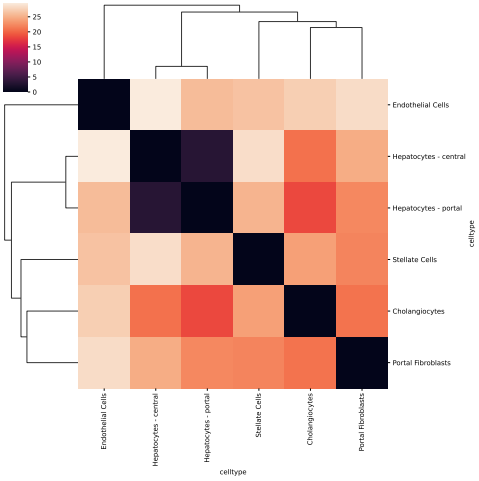


Figure 21: E-distance

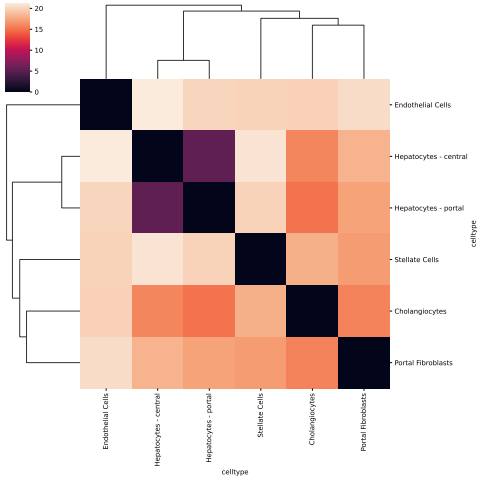


Figure 22: Euclidean

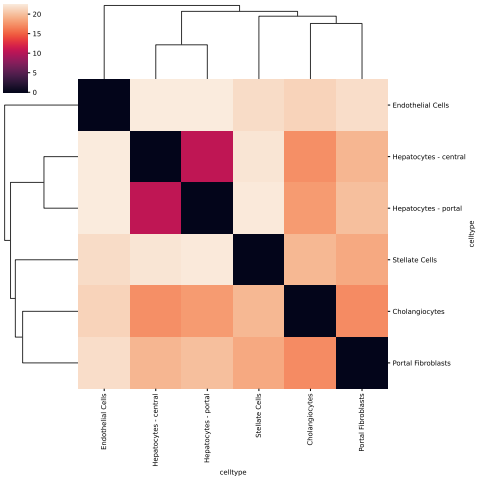


Figure 23: Mean pairwise

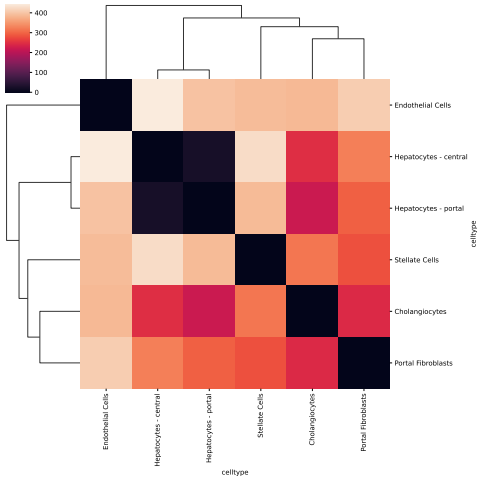


Figure 24: MMD

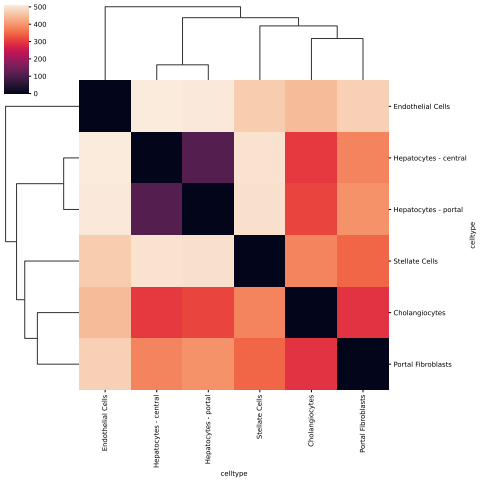


Figure 25: Wasserstein

Figure 26: Distance metrics for dosage highest 30 $\mu g/kg$ per cell type

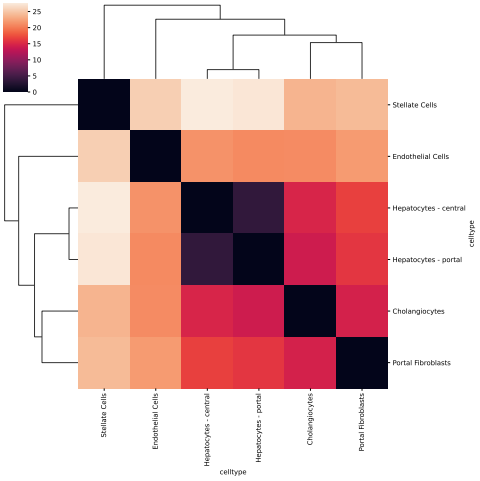


Figure 27: E-distance

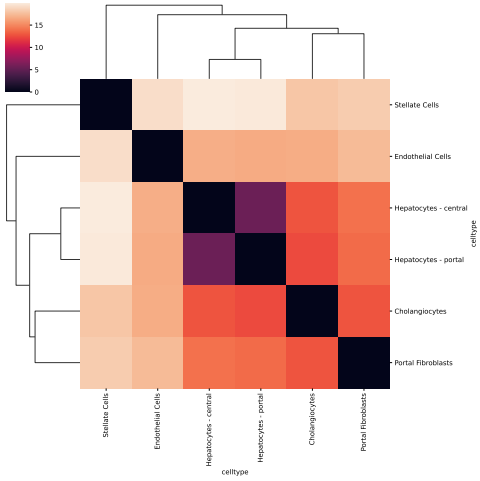


Figure 28: Euclidean

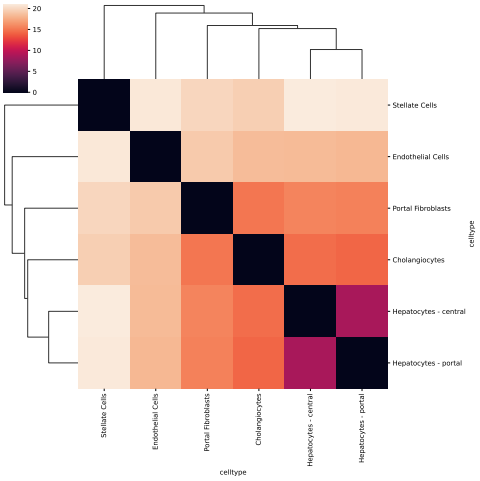


Figure 29: Mean pairwise

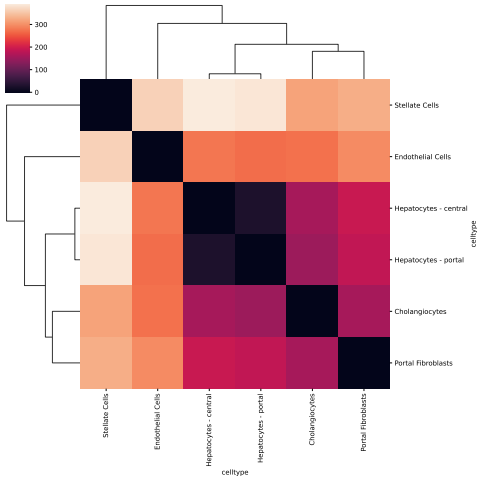


Figure 30: MMD

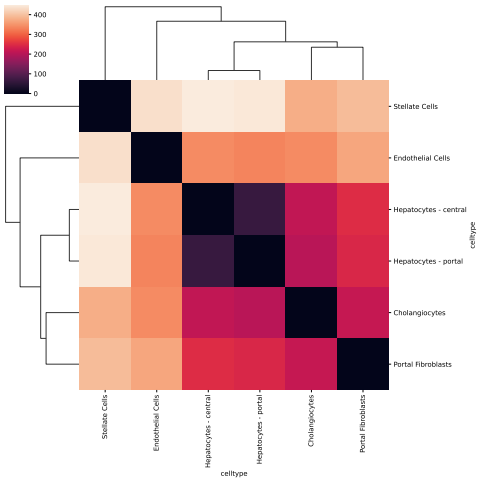


Figure 31: Wasserstein

Figure 32: Distance metrics for lowest dosage 0.01 $\mu g/kg$ per cell type

10.2 PBMC dataset

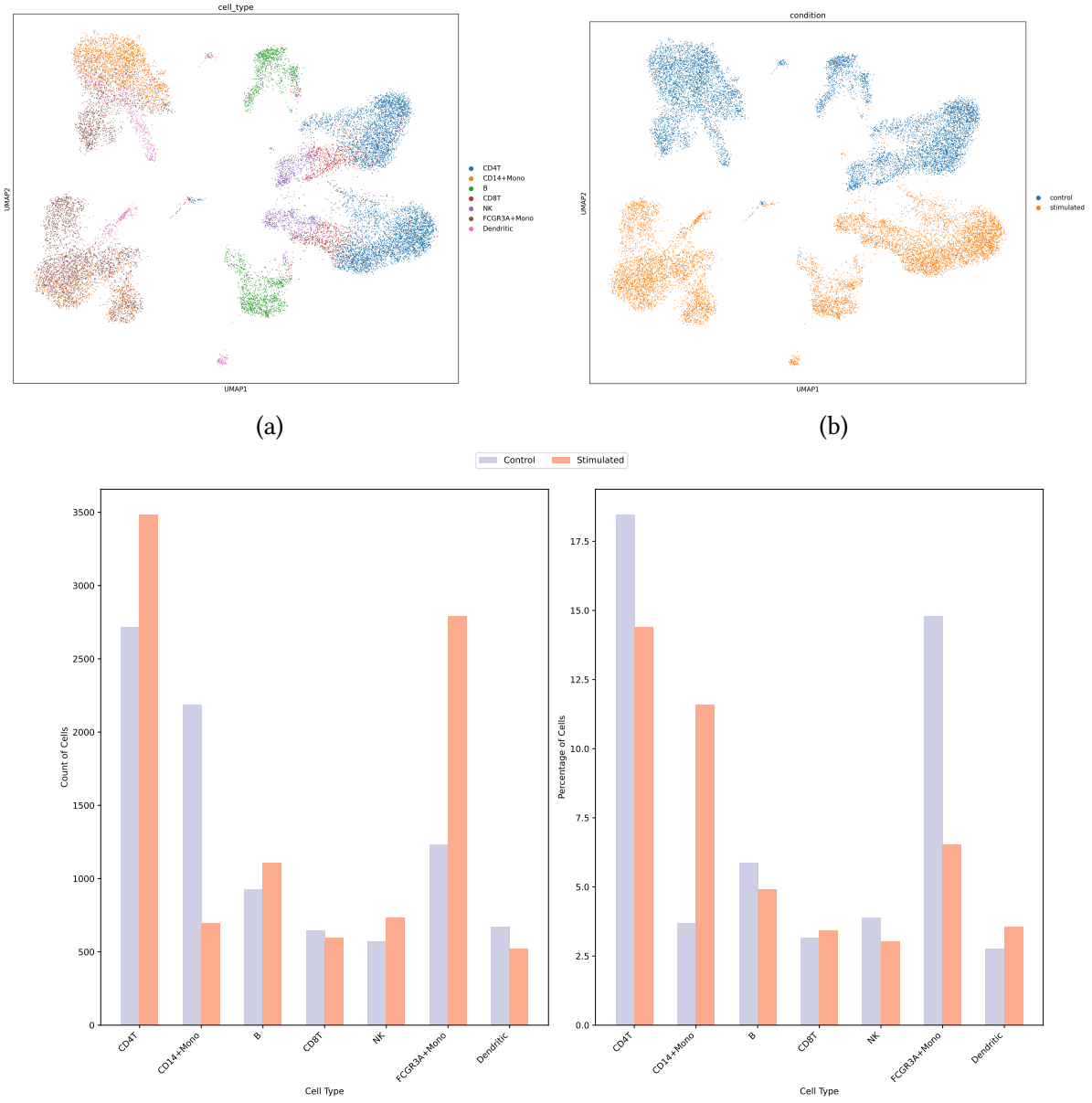


Figure 33: PBMC overview

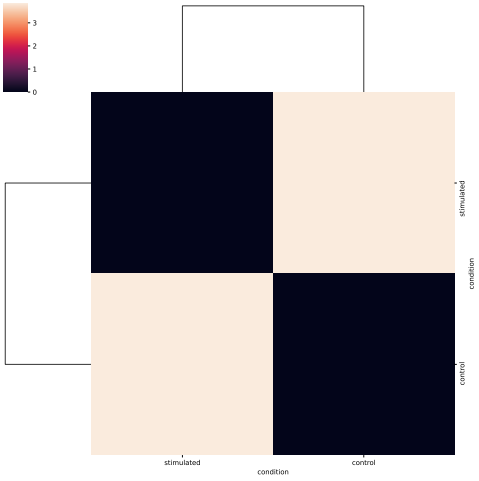


Figure 34: E-distance

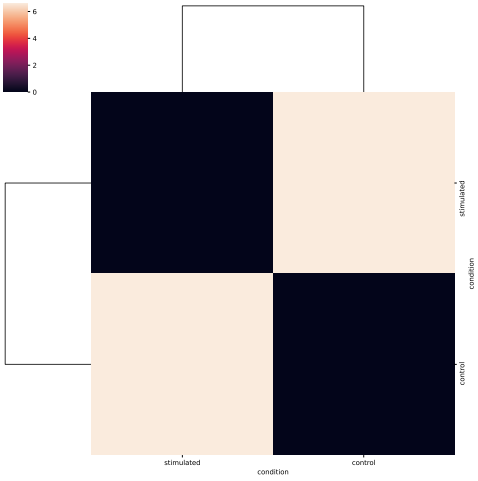


Figure 35: Euclidean

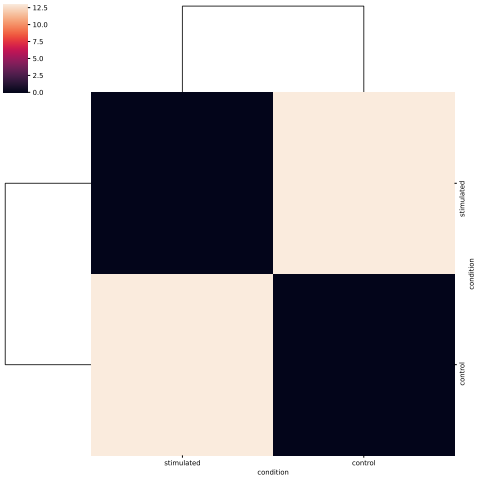


Figure 36: Mean pairwise

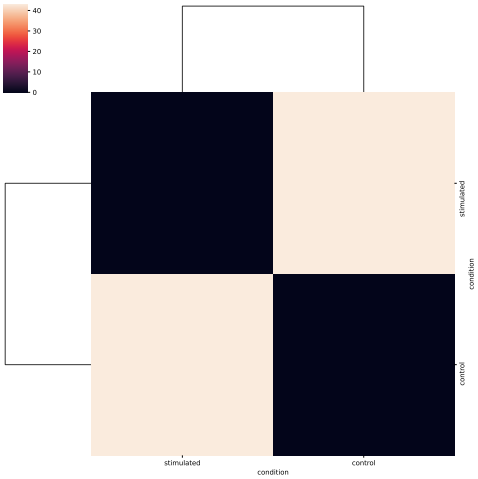


Figure 37: MMD

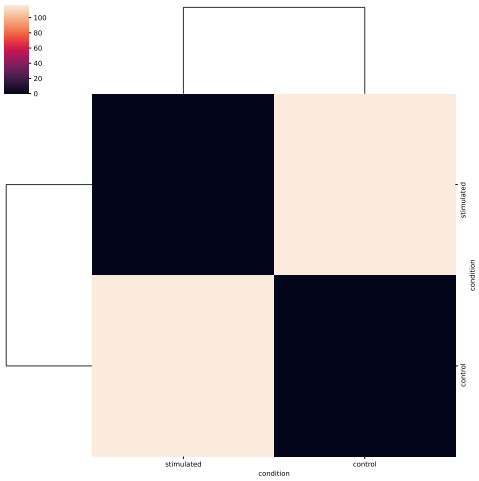


Figure 38: Wasserstein

Figure 39: Distance metrics per condition

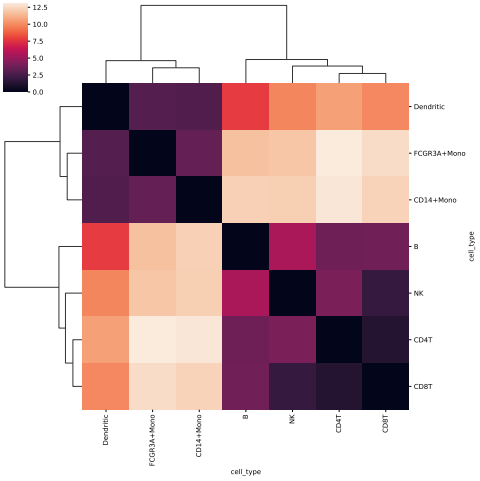


Figure 40: E-distance

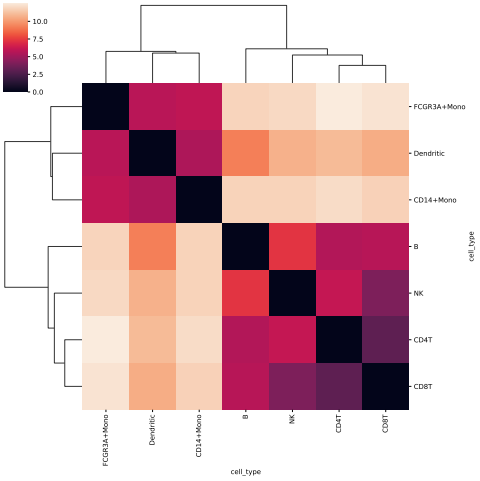


Figure 41: Euclidean

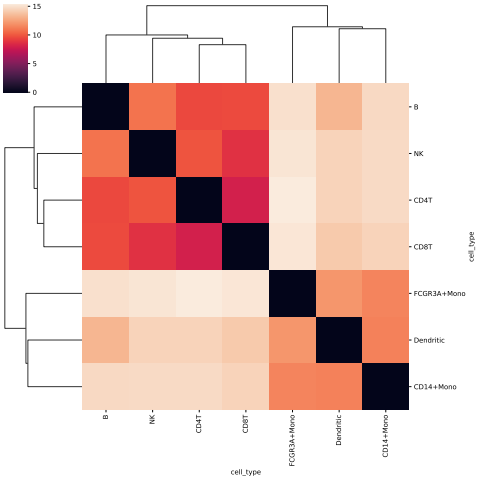


Figure 42: Mean pairwise

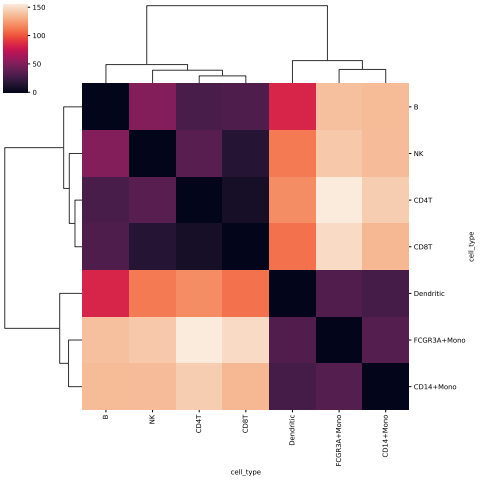


Figure 43: MMD

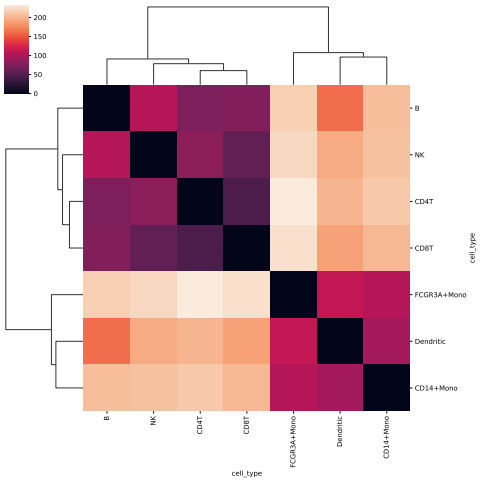
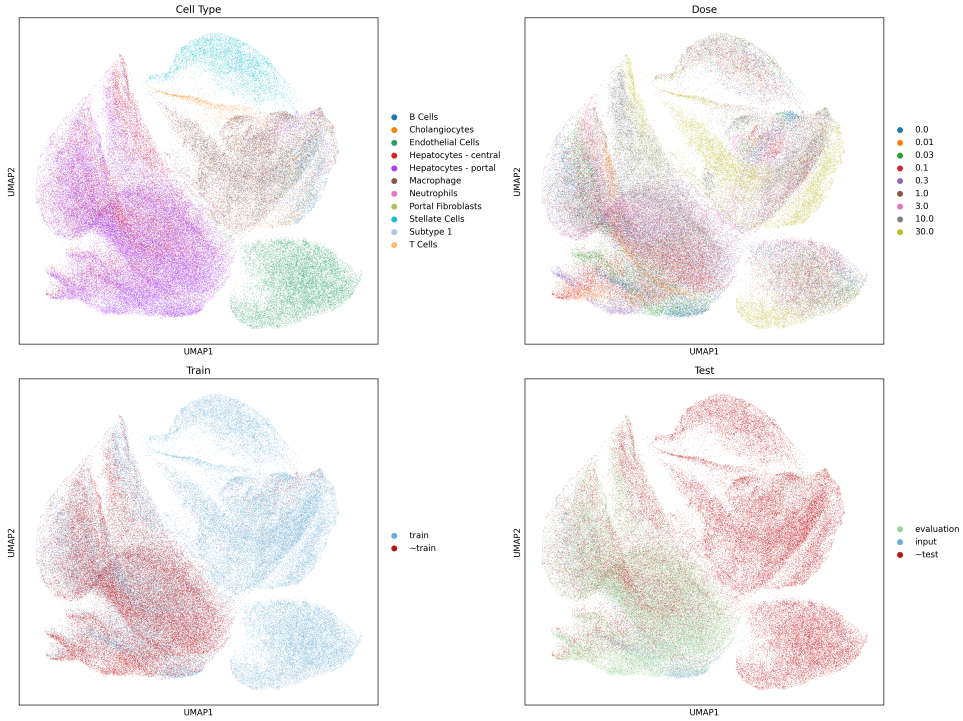


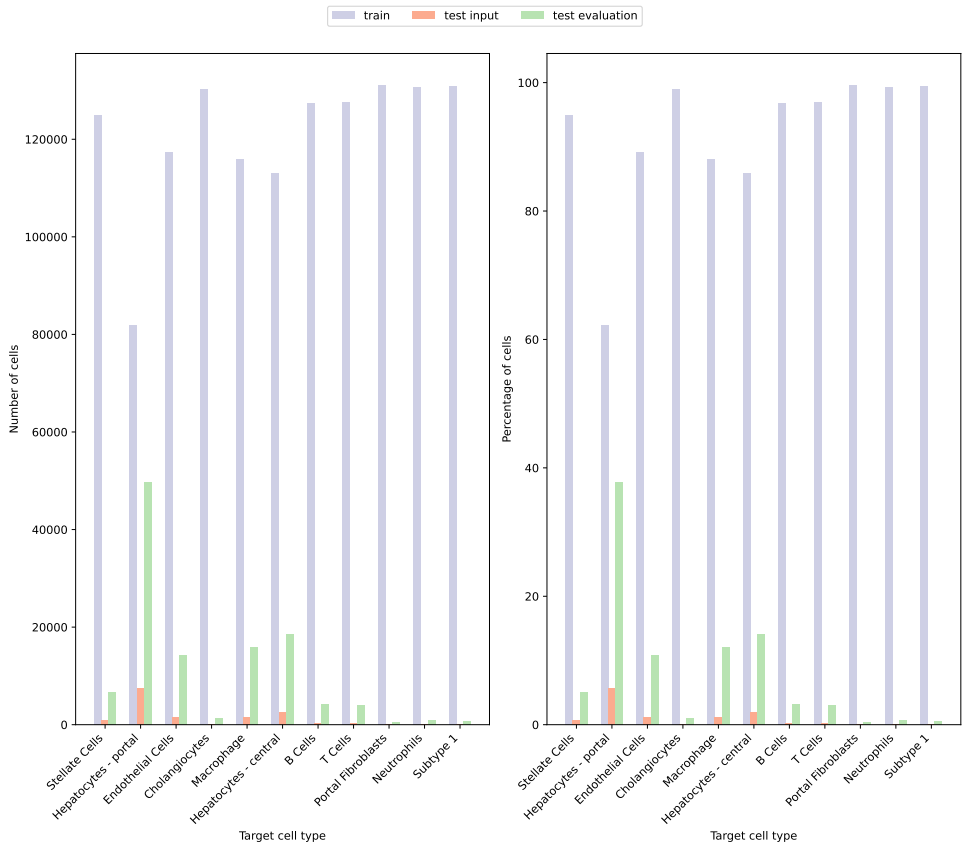
Figure 44: Wasserstein

Figure 45: Distance metrics per cell type

11 Nault all cell types evaluation

11.1 Multiple doses





11.2 Single dose

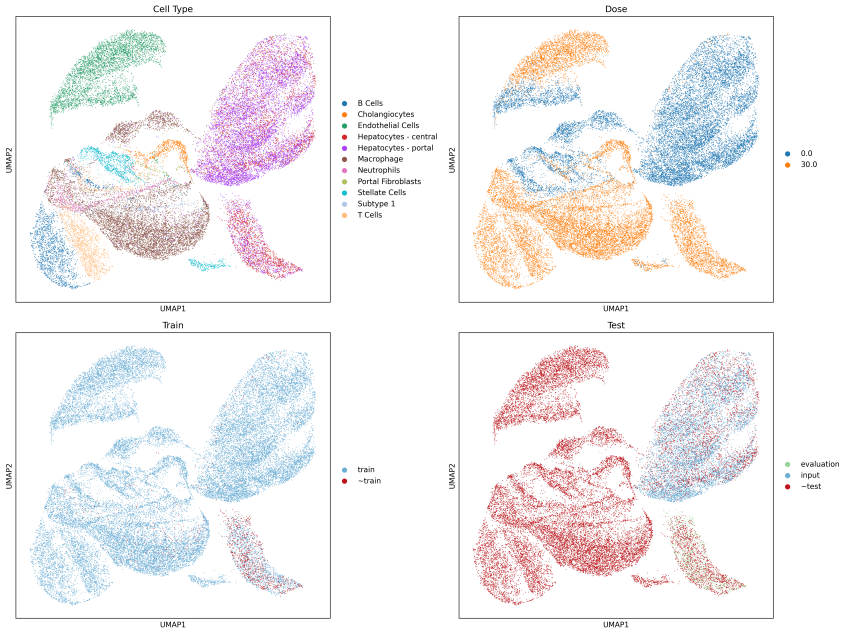


Figure 46: Example of $30\mu g/kg$

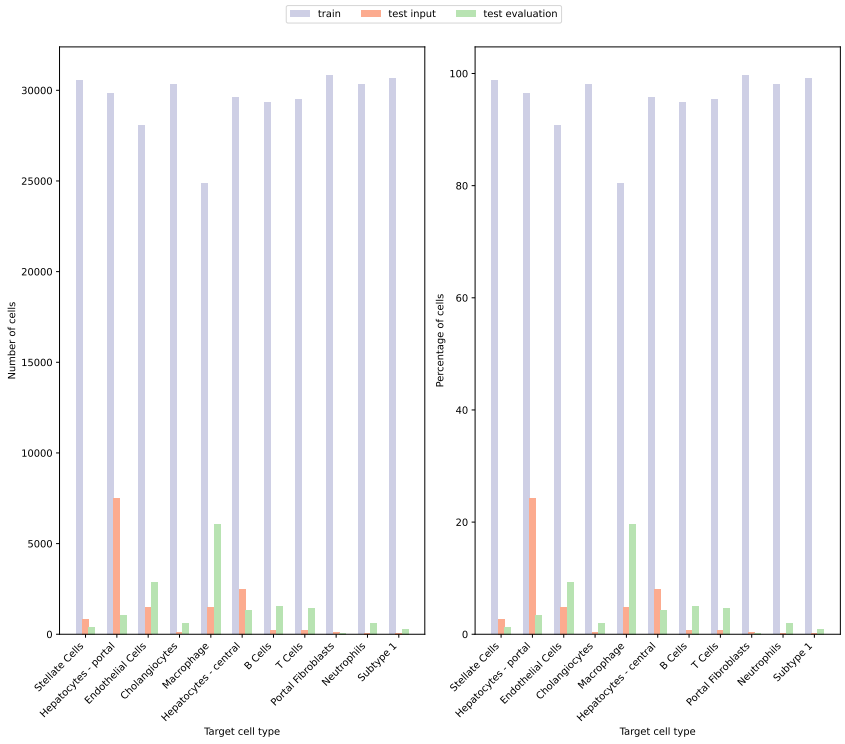


Figure 47: Number of cells per cell type for $30\mu g/kg$

11.3 Comparison

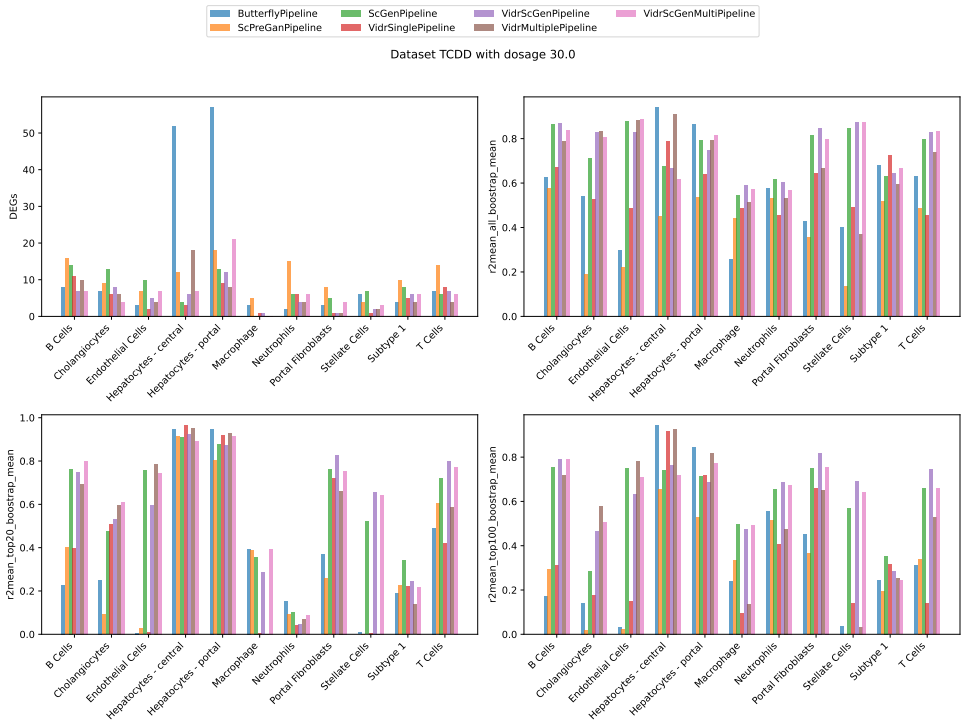


Figure 48: Baseline metrics for highest dosage $30\mu g/kg$ across cell types

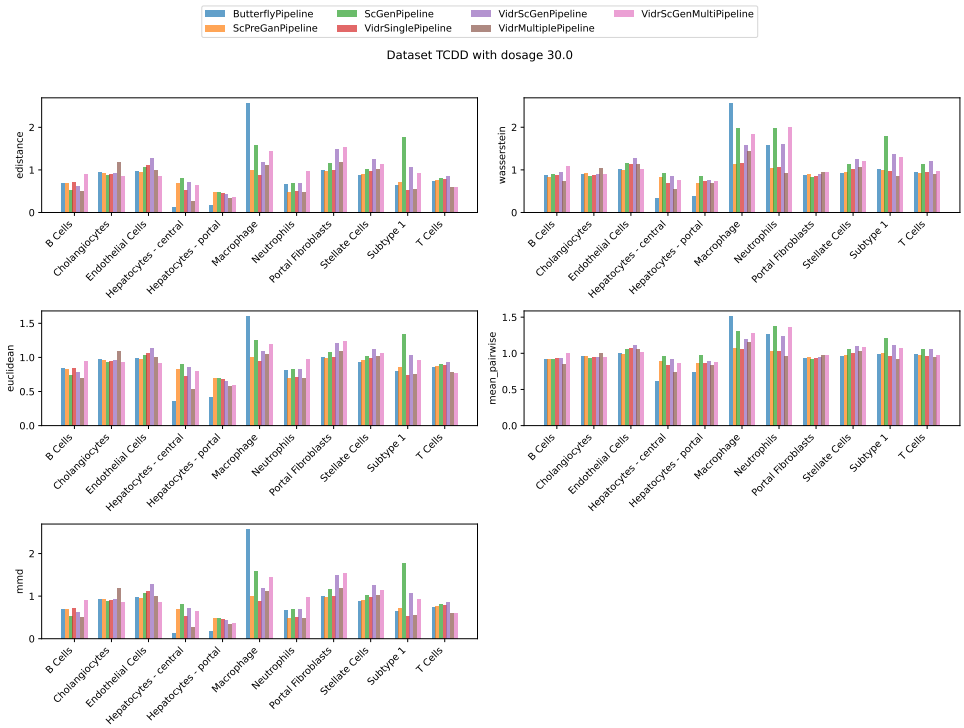


Figure 49: Distance metrics for highest dosage $30\mu g/kg$ across cell types

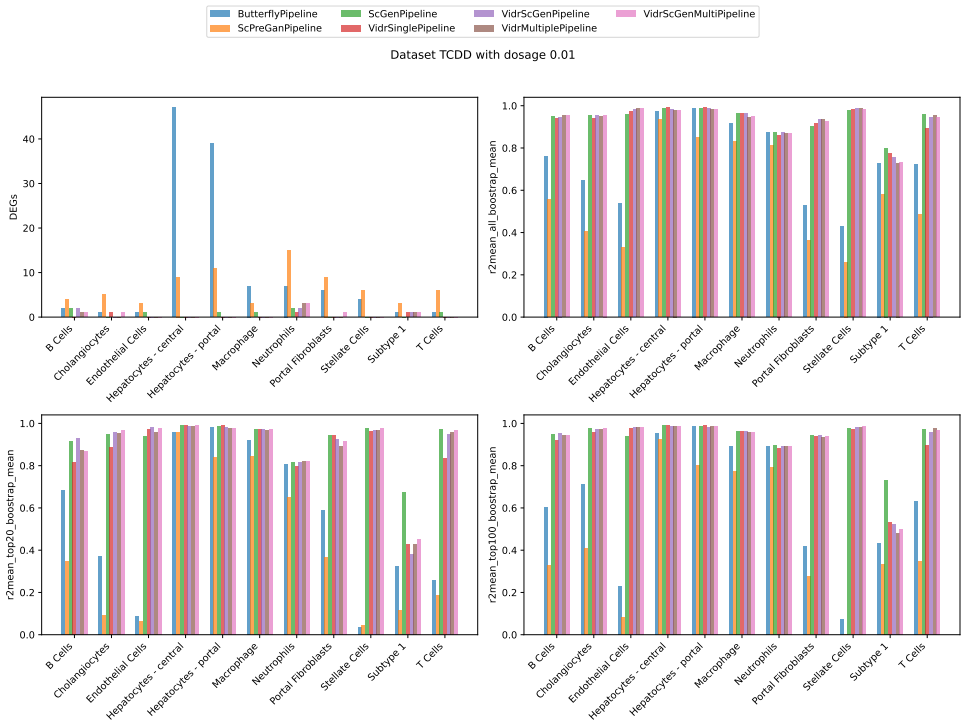


Figure 50: Baseline metrics for lowest dosage $0.01 \mu g/kg$ across cell types

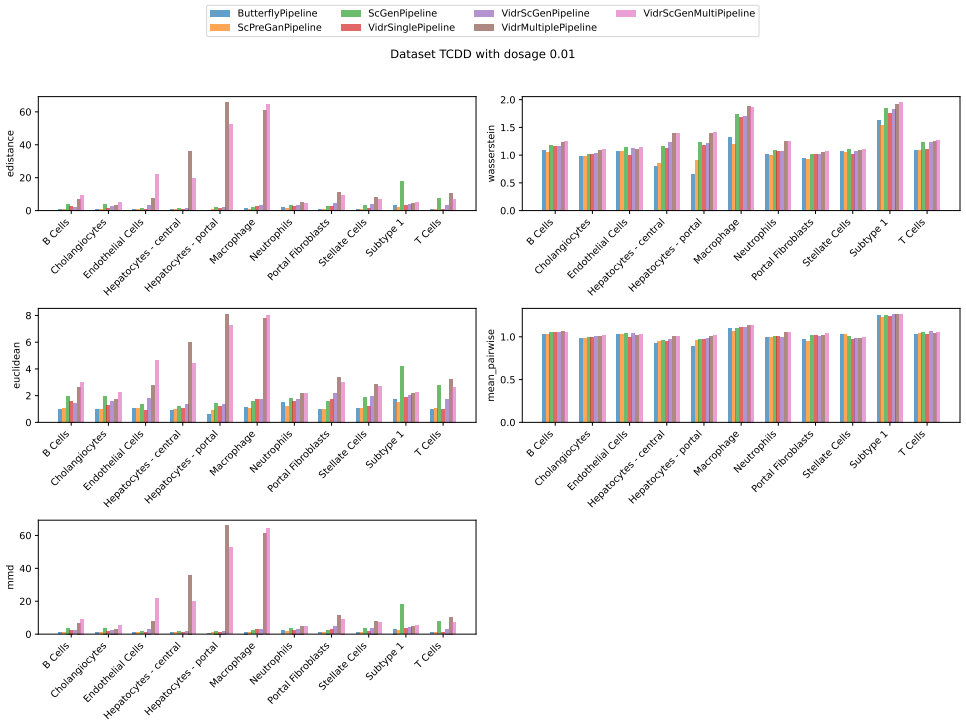


Figure 51: Distance metrics for lowest dosage $0.01 \mu g/kg$ across cell types

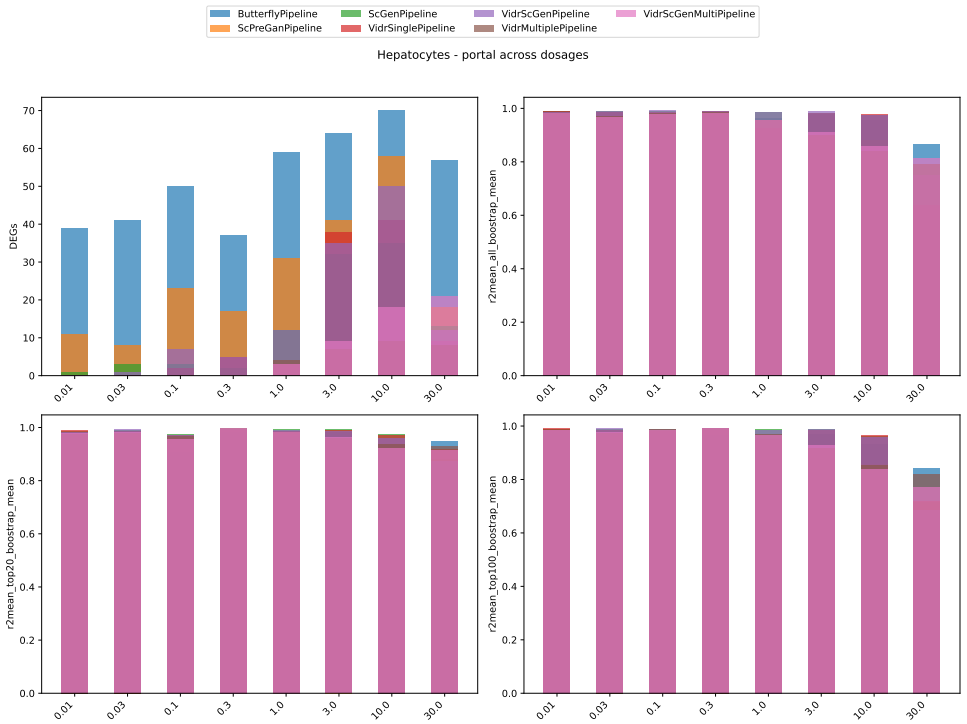


Figure 52: Baseline metrics for Hepatocytes - portal across dosages

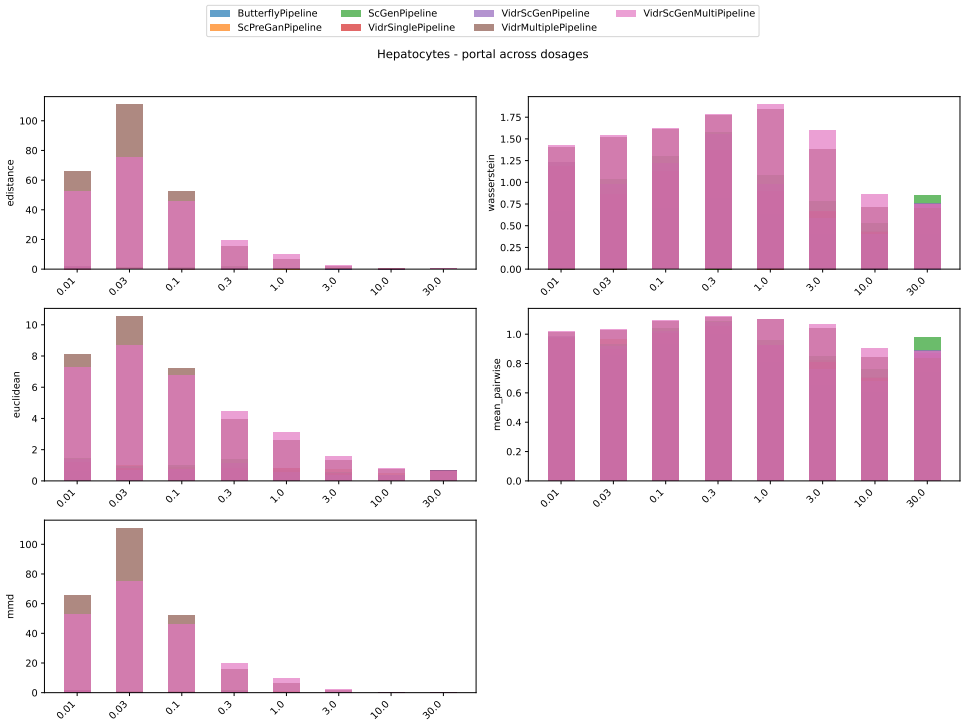


Figure 53: Distance metrics for Hepatocytes - portal across dosages

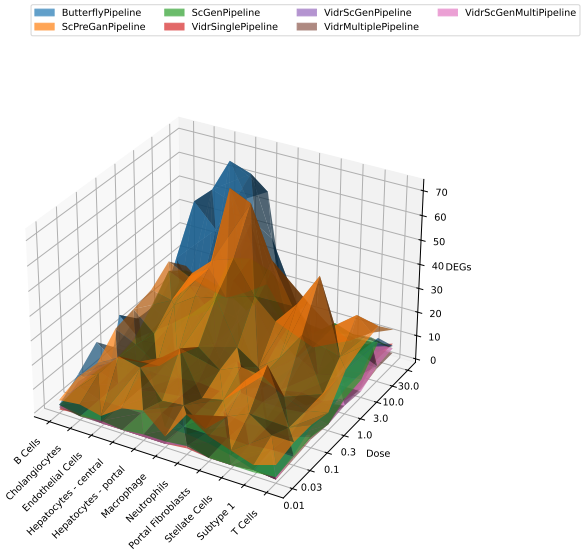


Figure 54

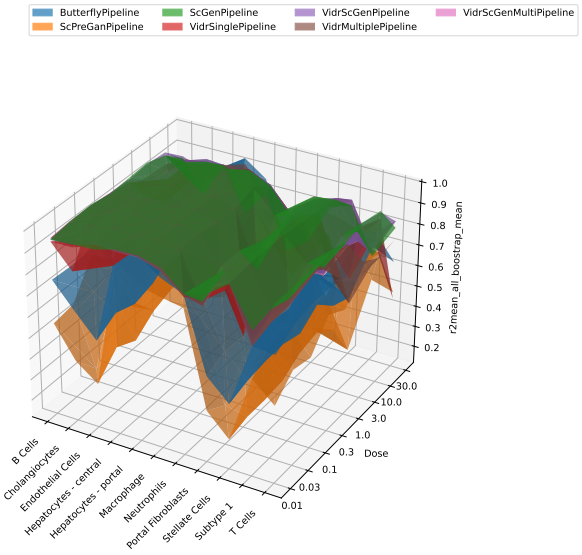


Figure 55

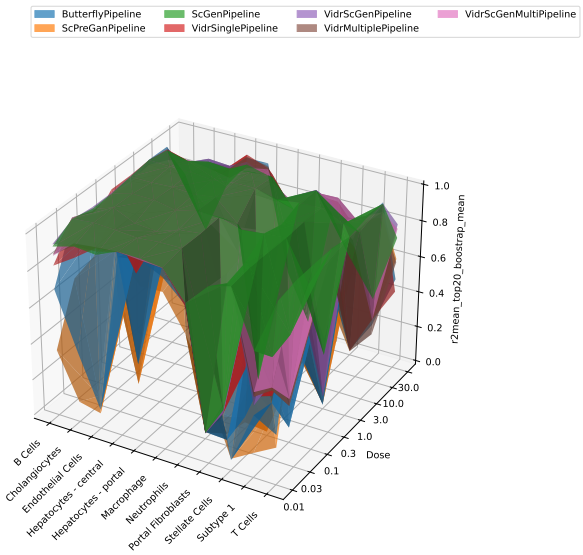


Figure 56

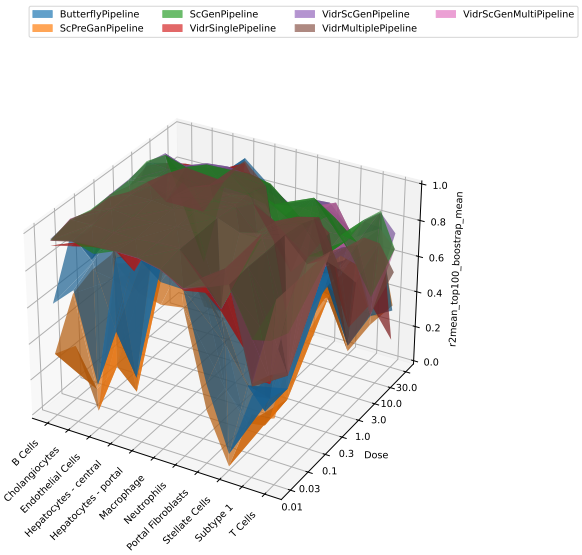


Figure 57

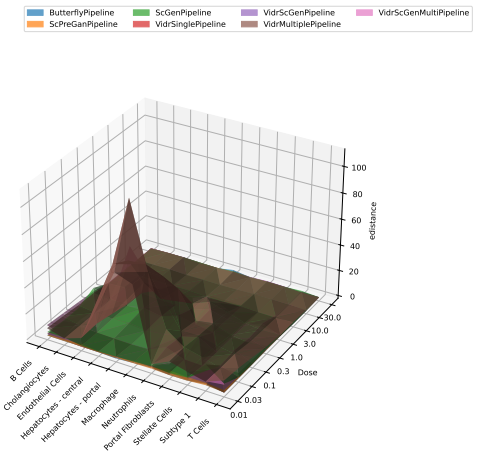


Figure 58: E-distance

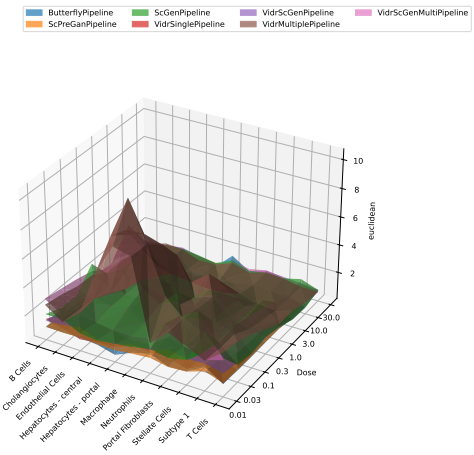


Figure 59: Euclidean

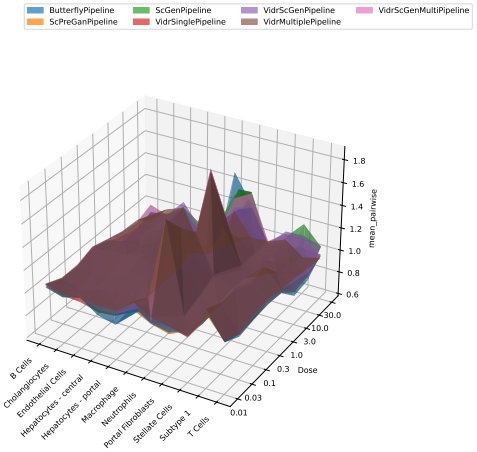


Figure 60: Mean pairwise

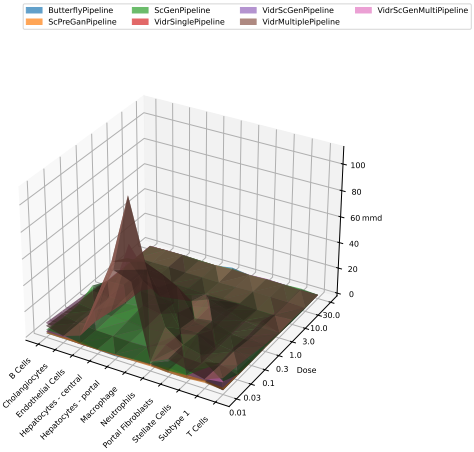


Figure 61: MMD

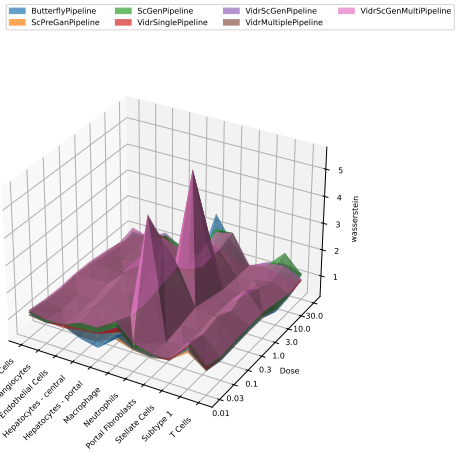
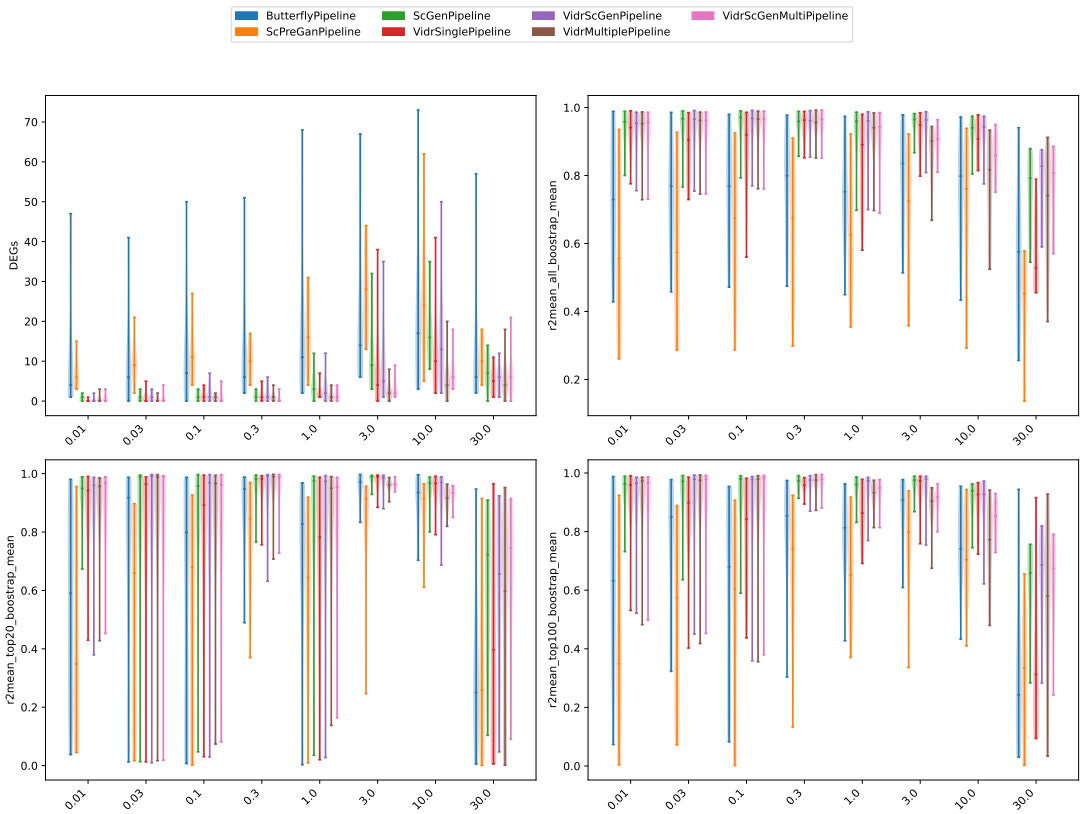
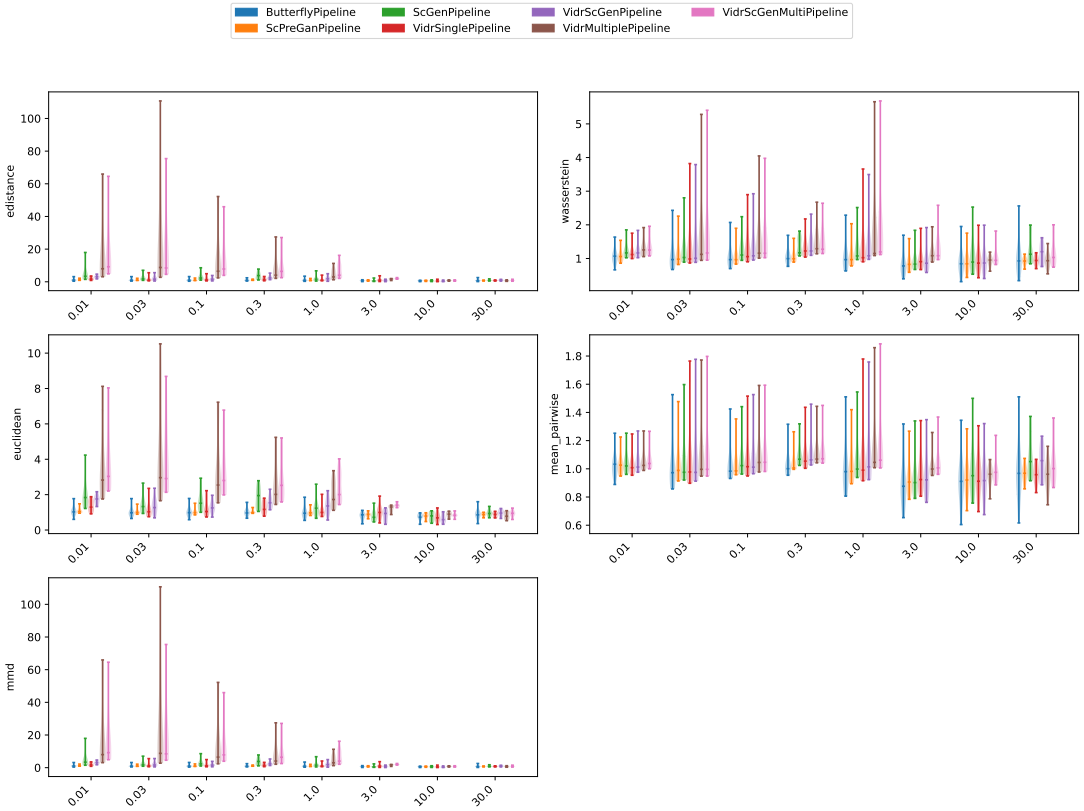
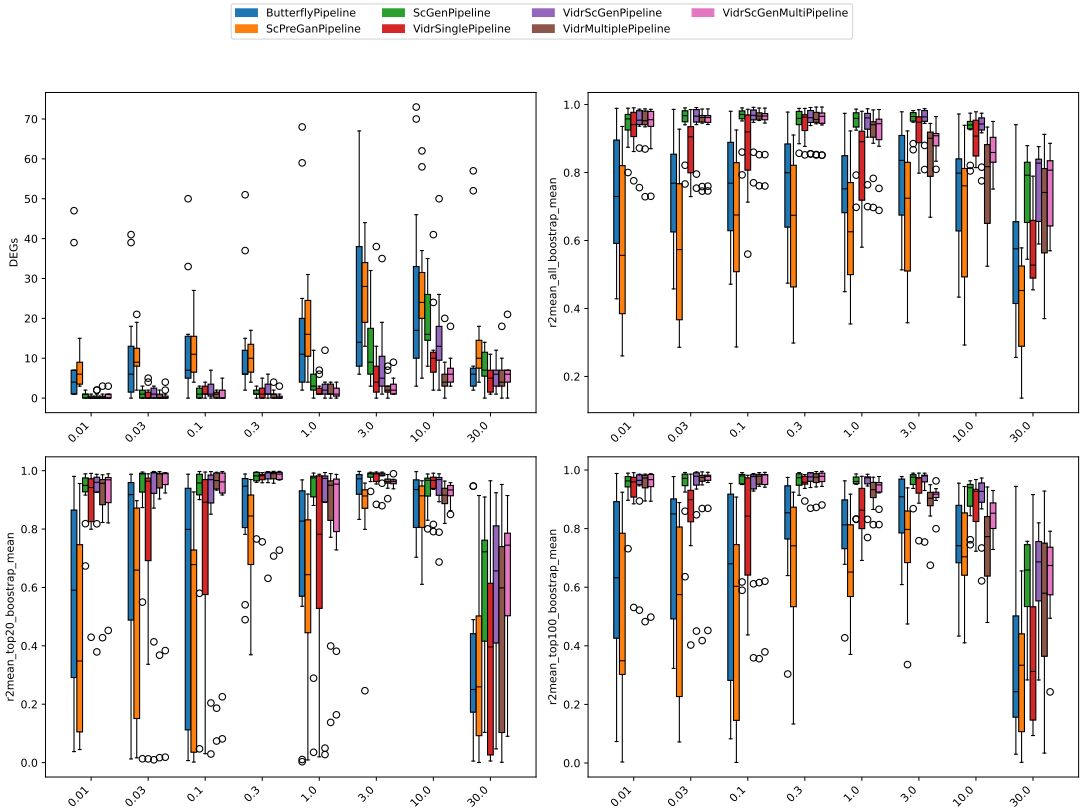


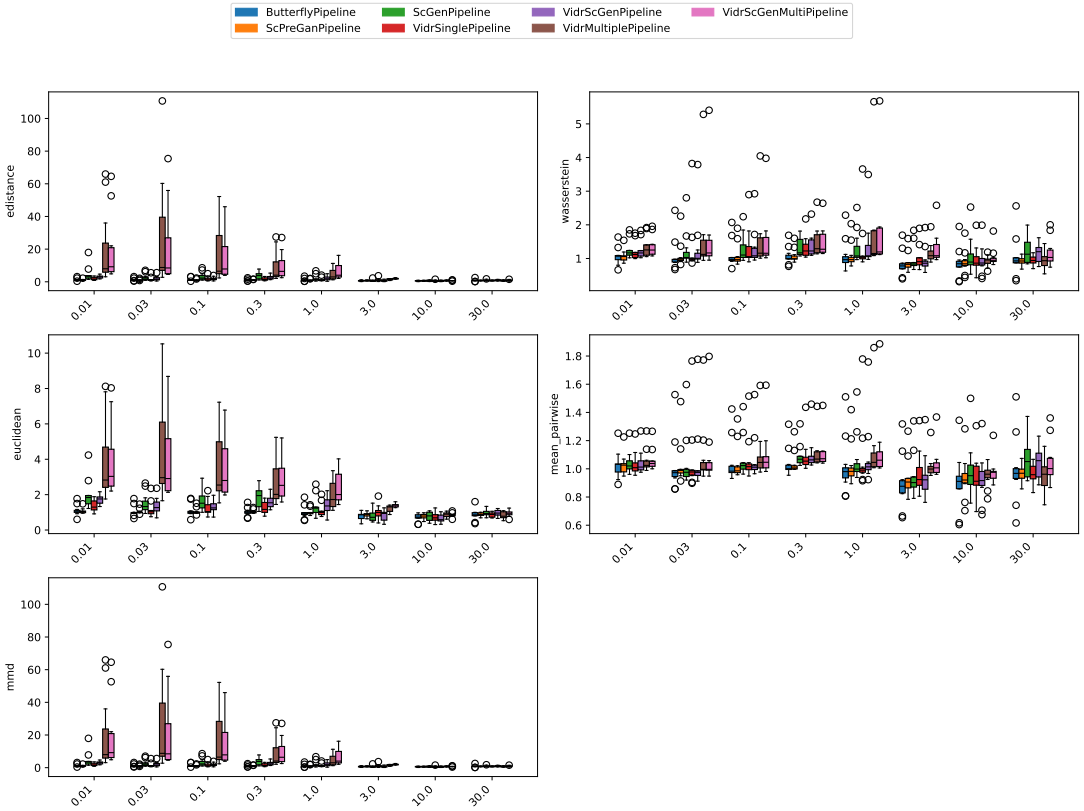
Figure 62: Wasserstein

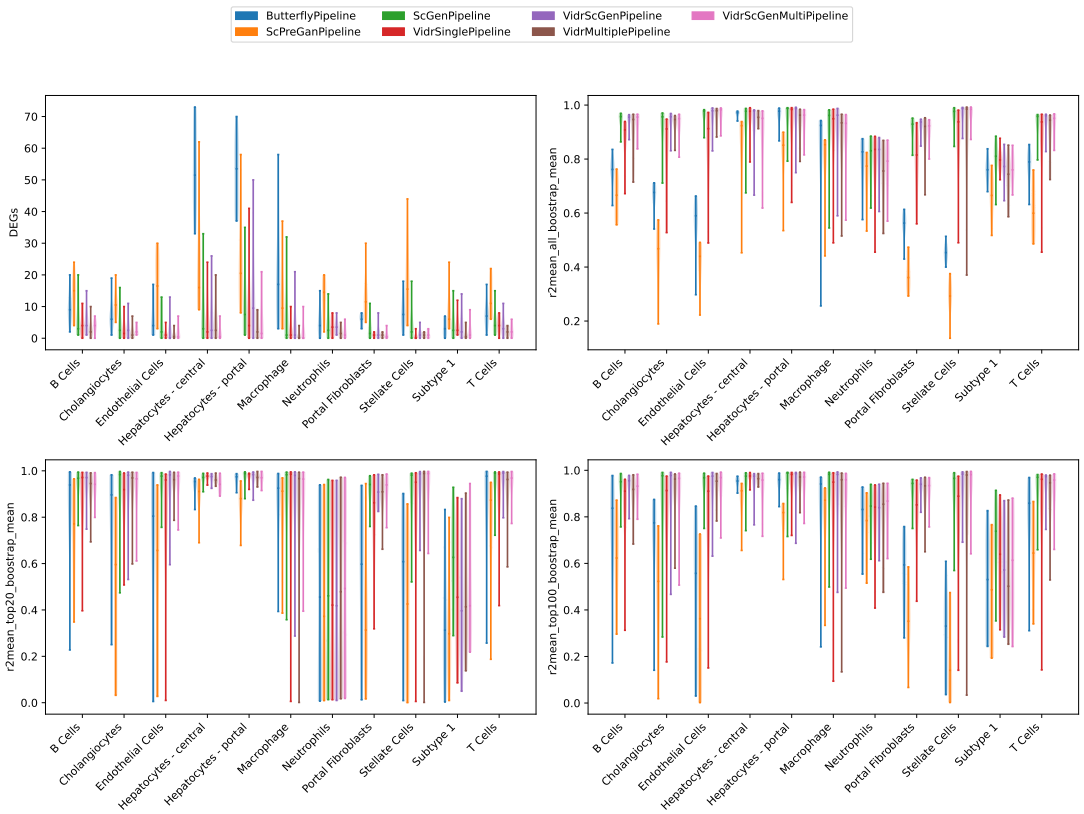
Figure 63: Distance metrics per cell type

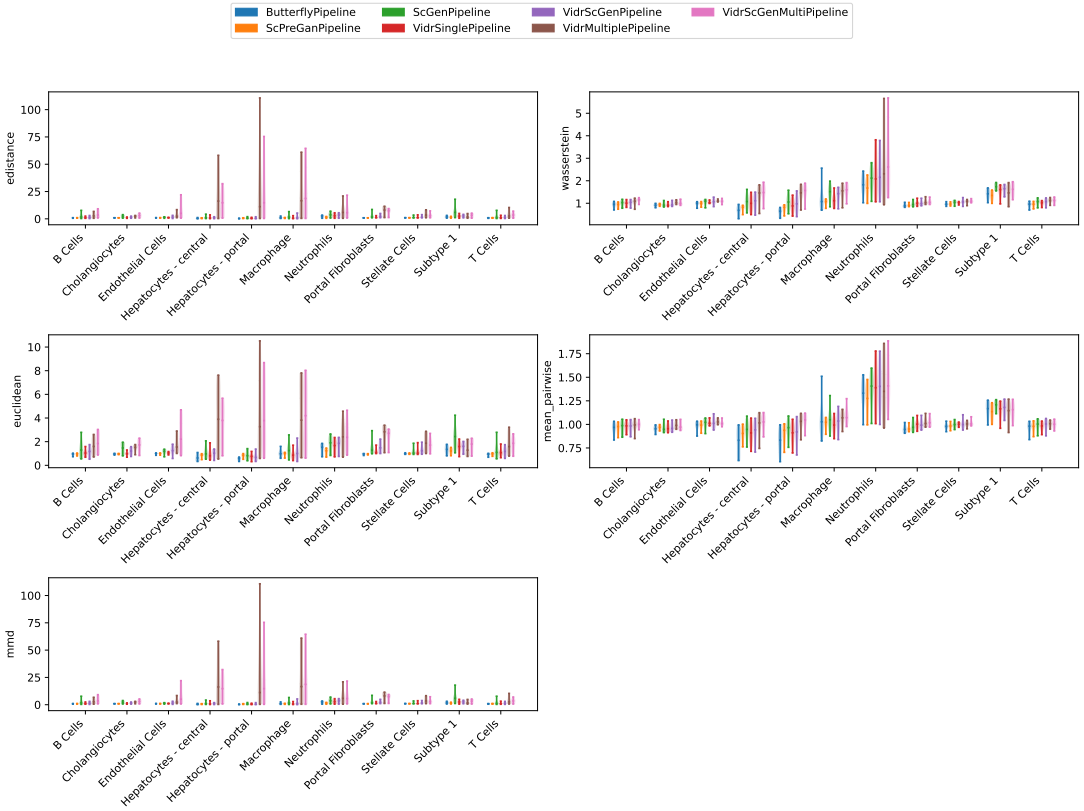

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0

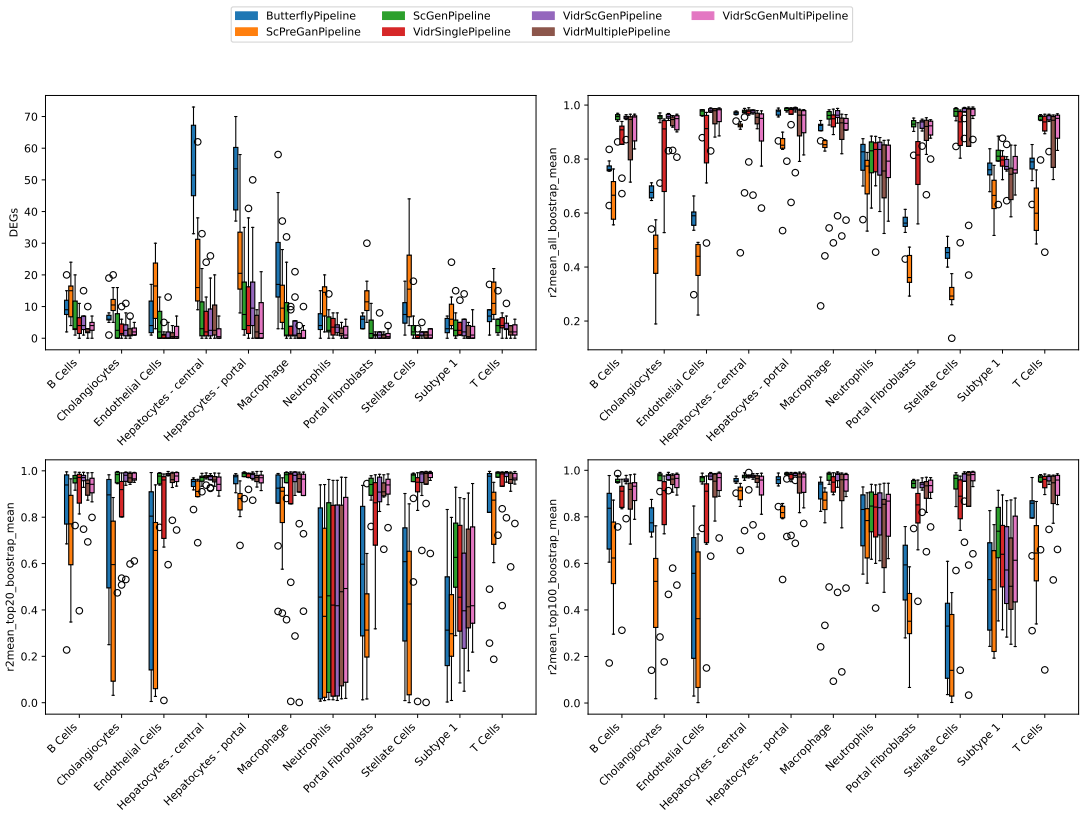

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
edistance
wasserstein
euclidean
mean_pairwise
mmd
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0

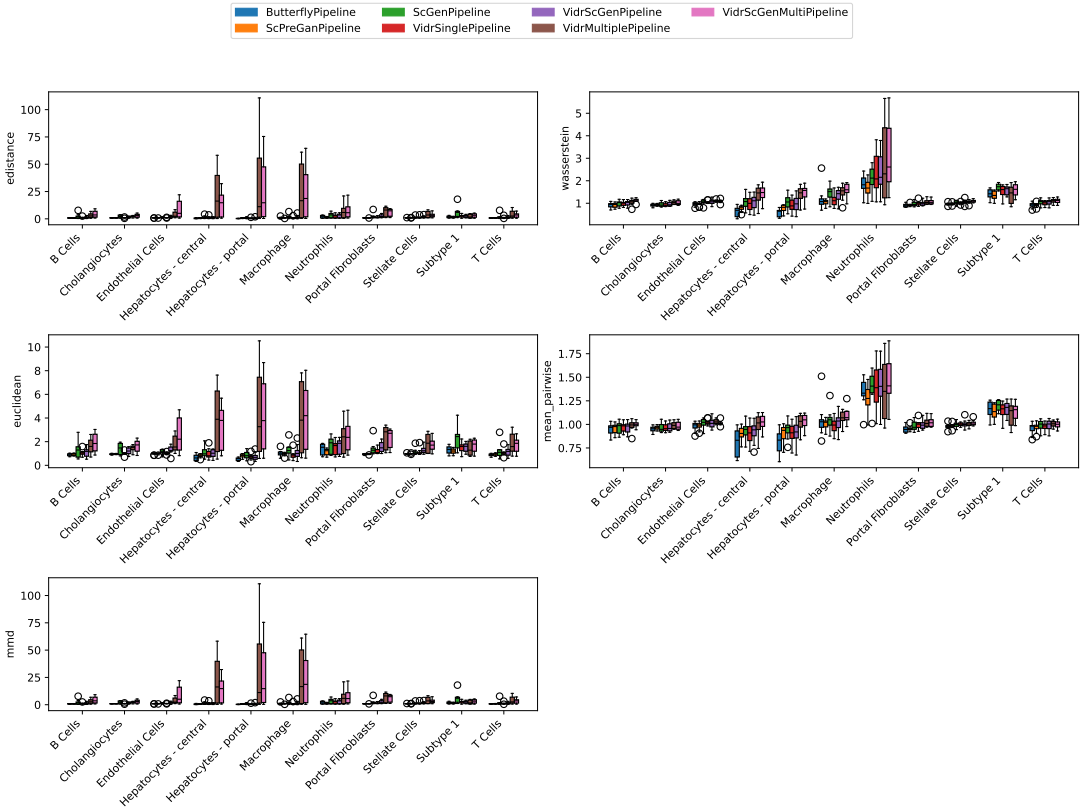

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0


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

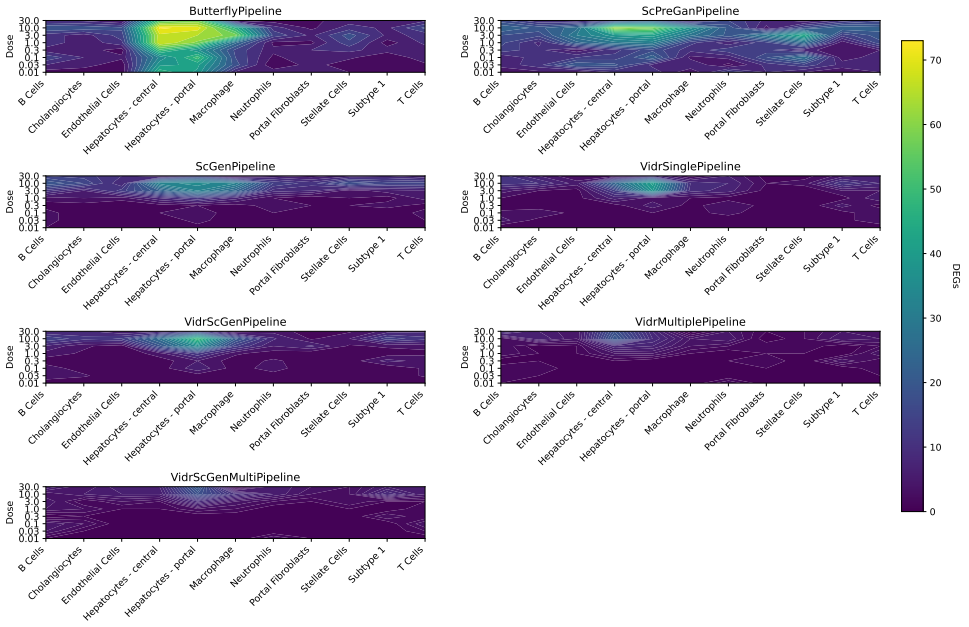


Figure 64: DEGs

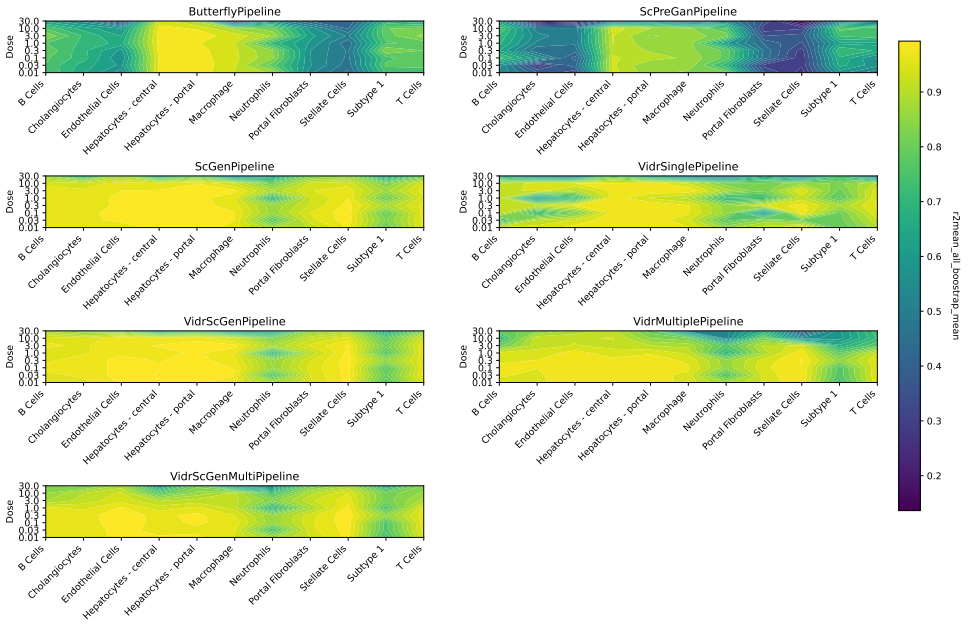


Figure 65: r2 HVGs

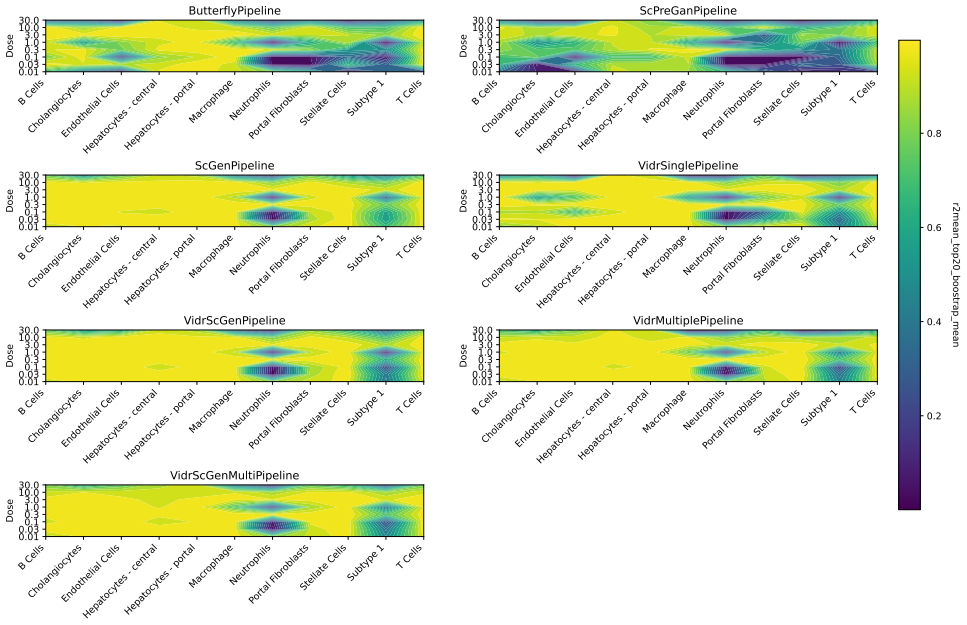


Figure 66: r2 top 20

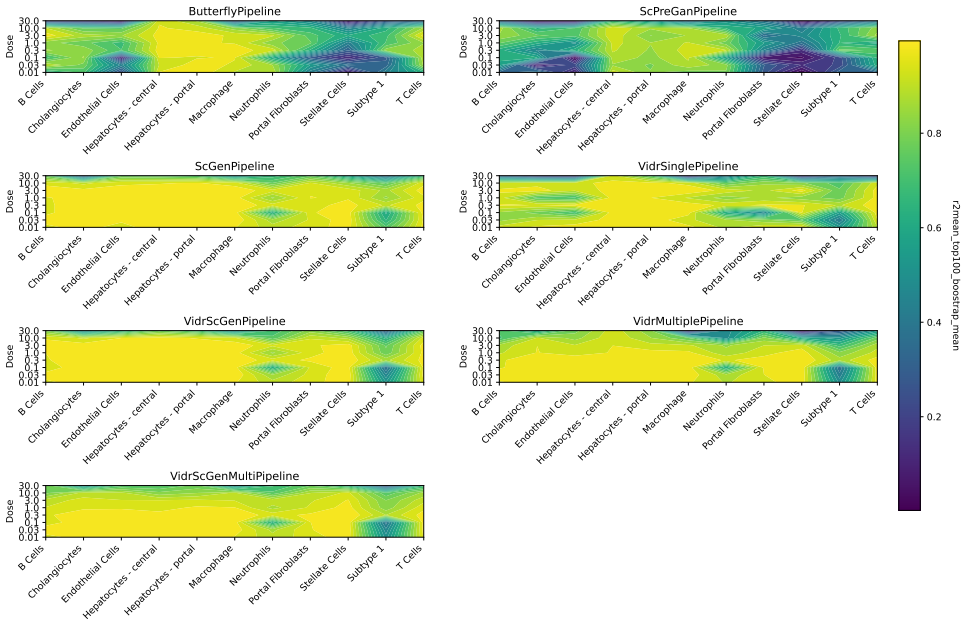


Figure 67: r2 top 100

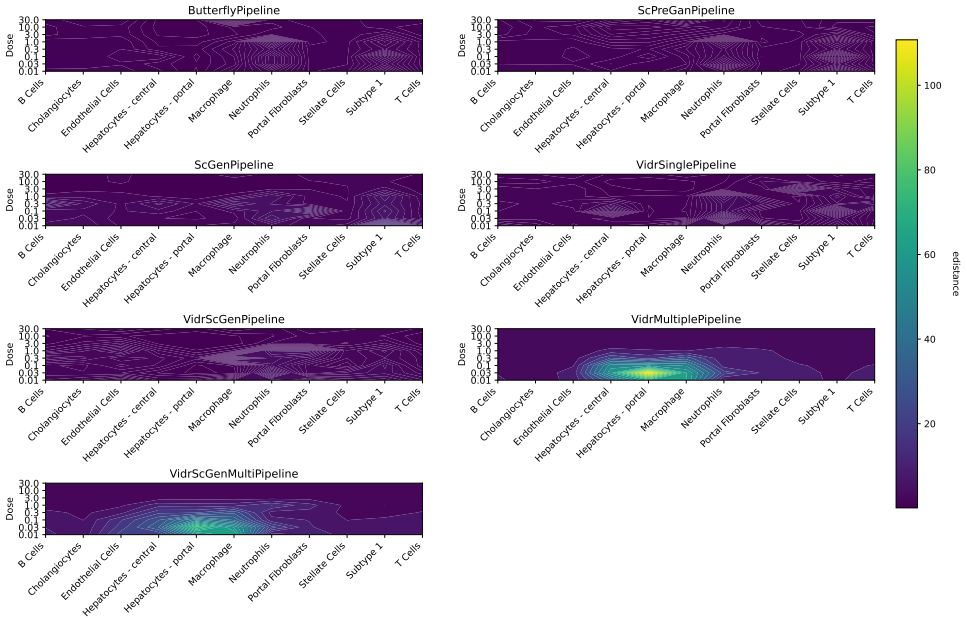


Figure 68: E-distance

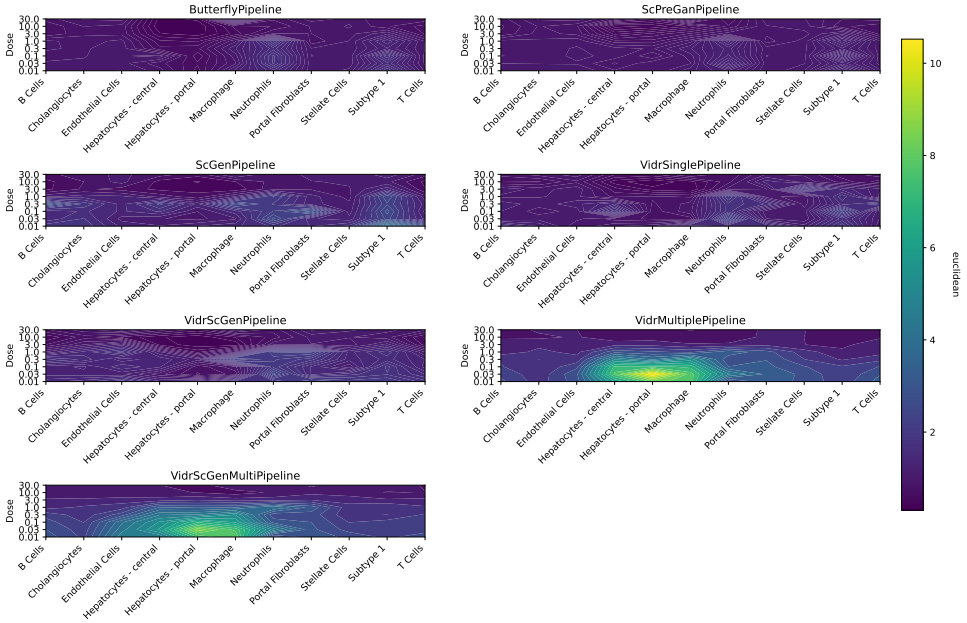


Figure 69: Euclidean

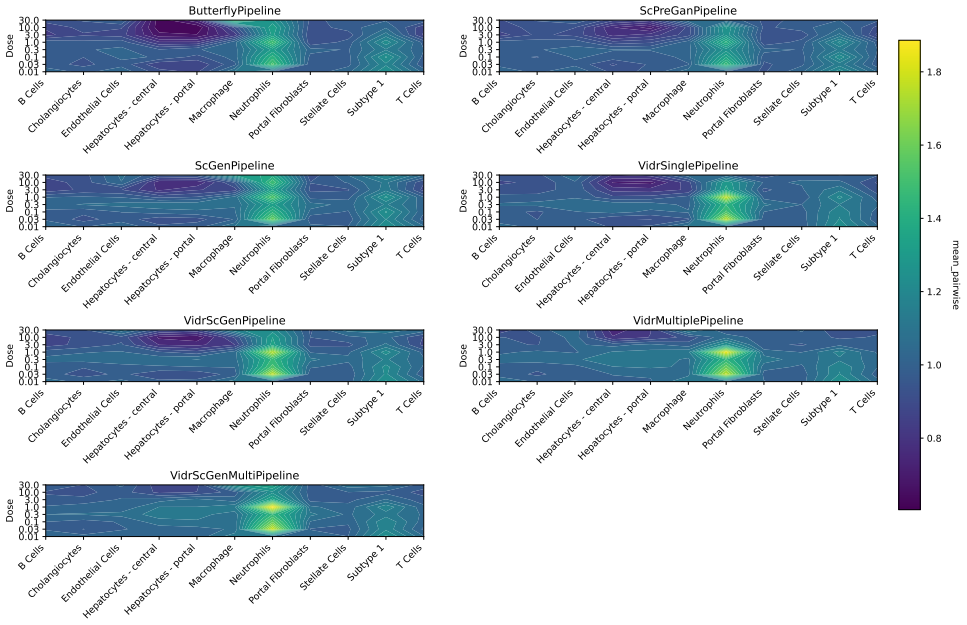


Figure 70: Mean pairwise

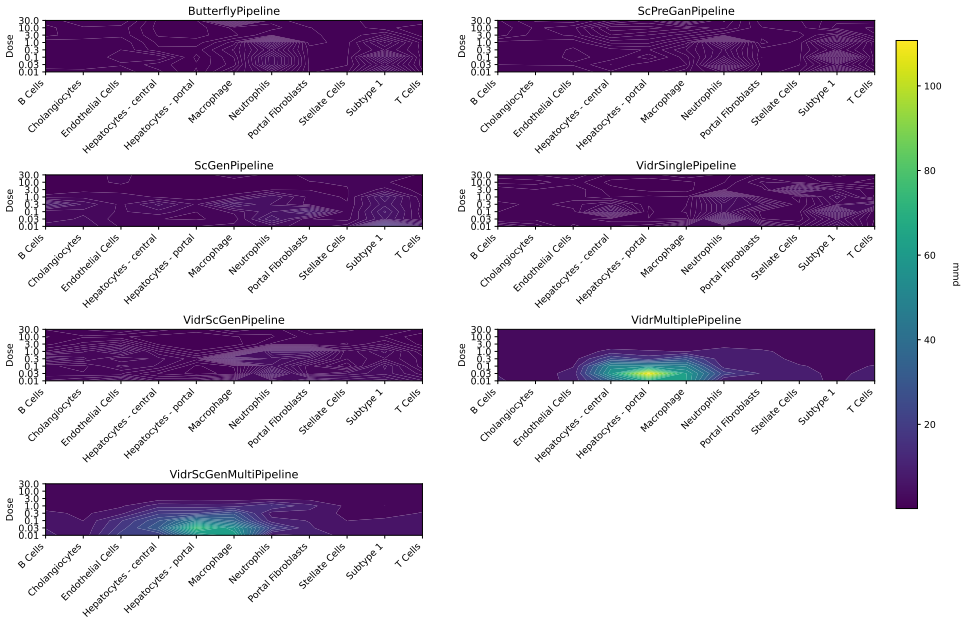


Figure 71: MMD

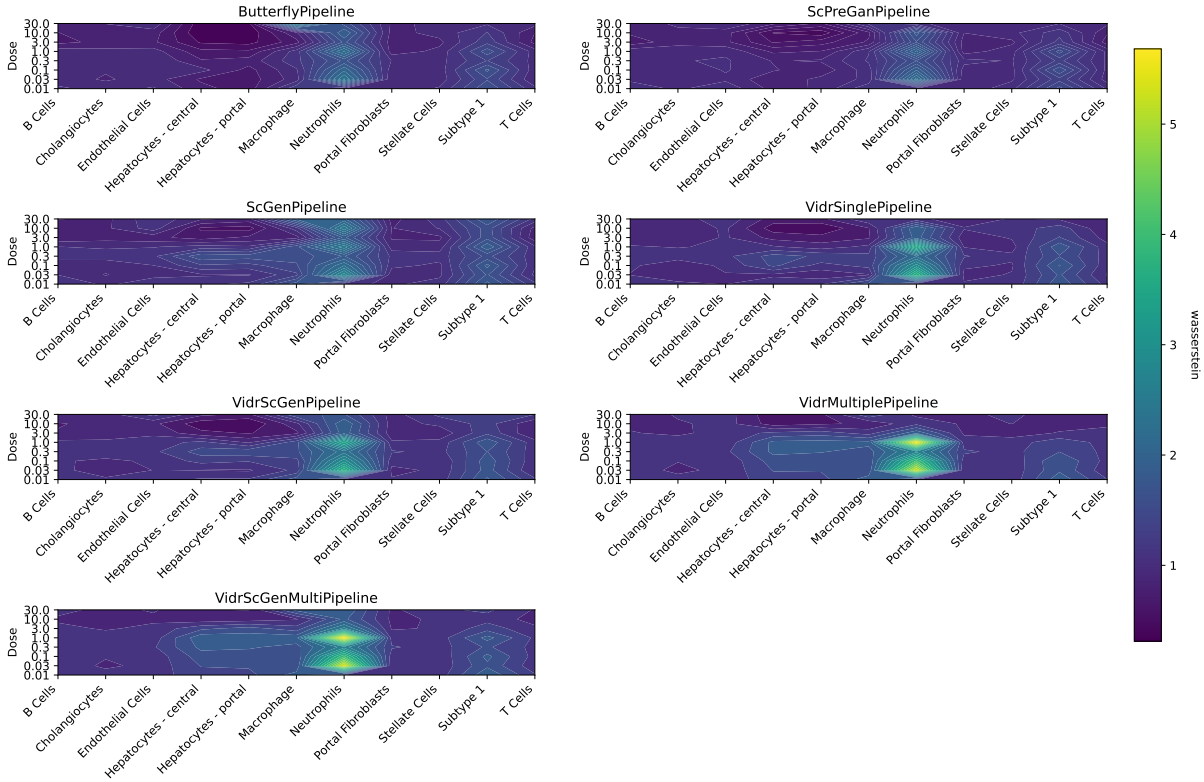
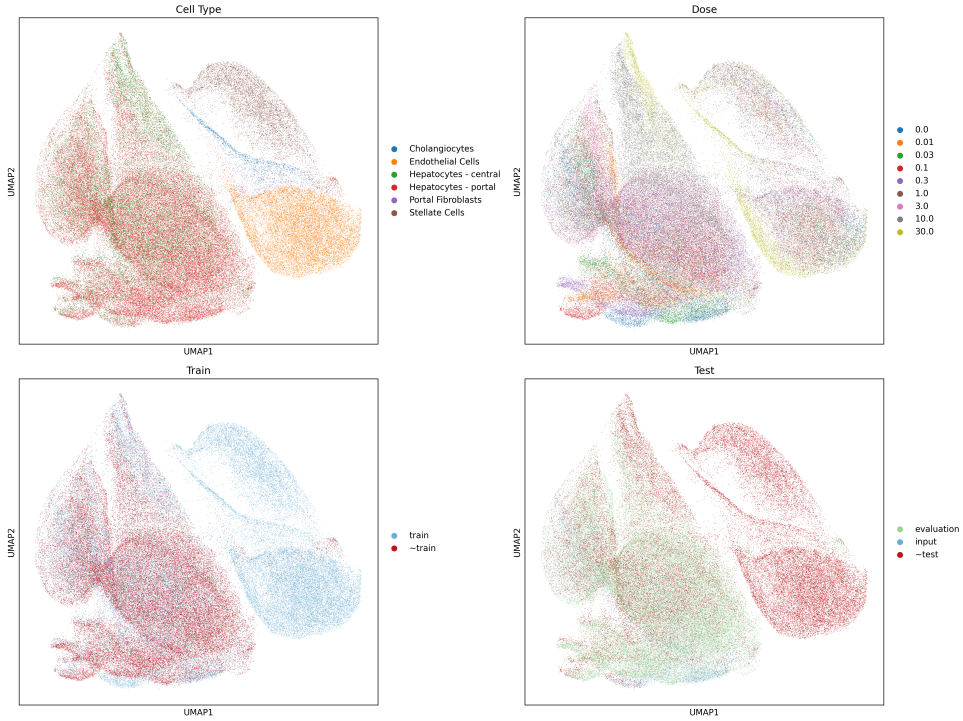
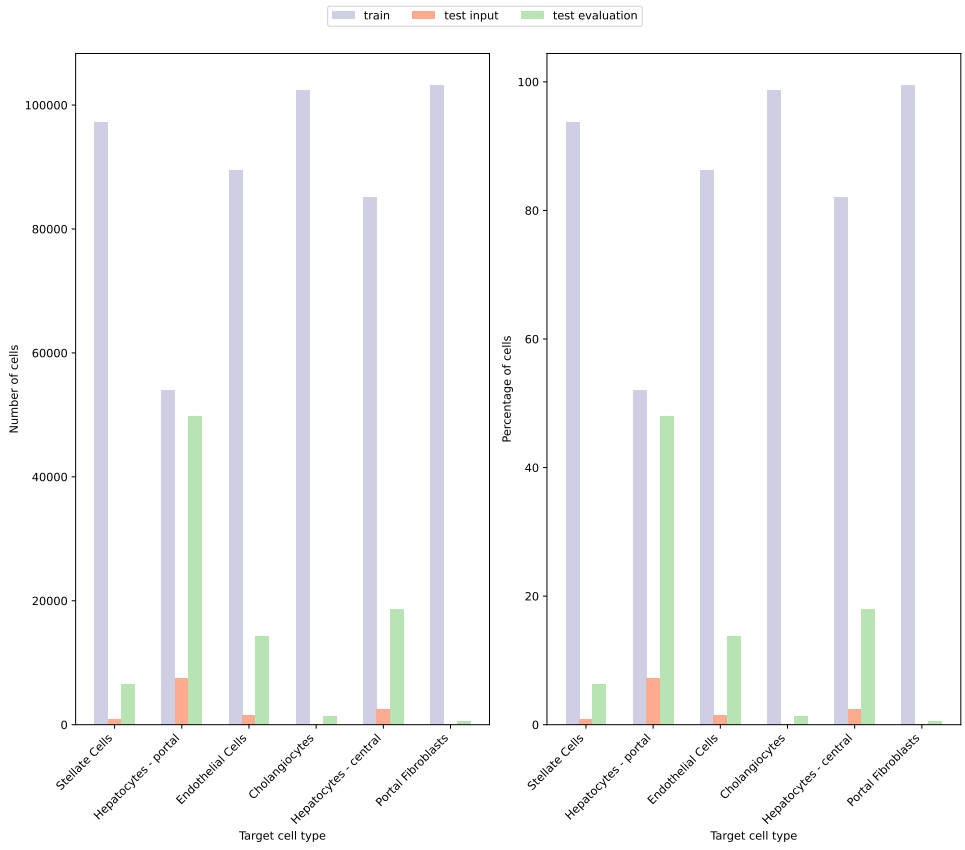


Figure 72: Wasserstein

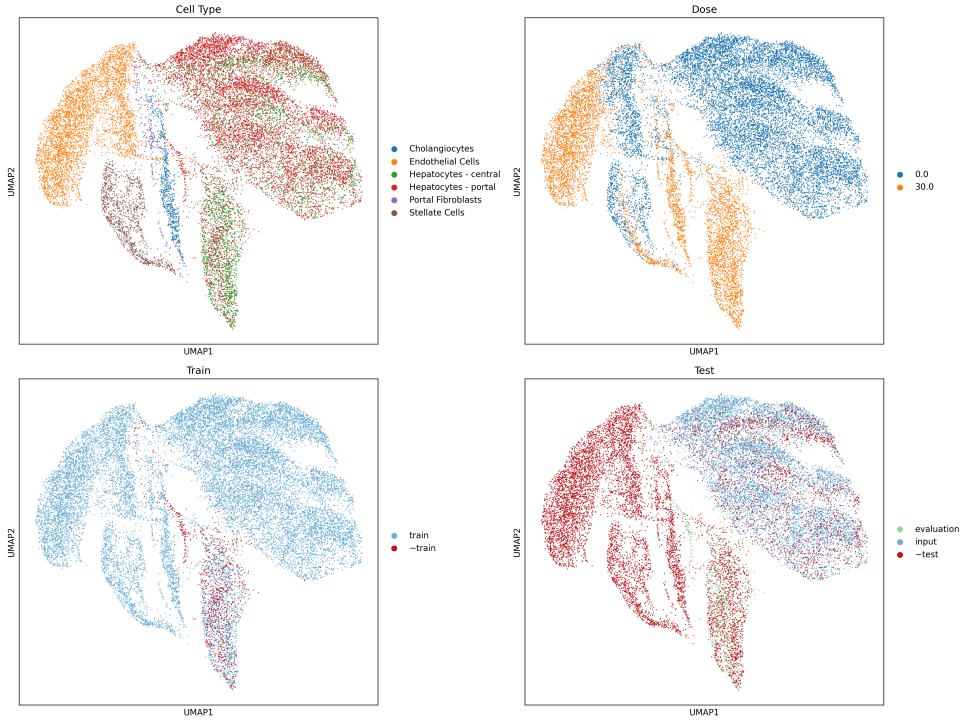
12 Nault liver cell types evaluation

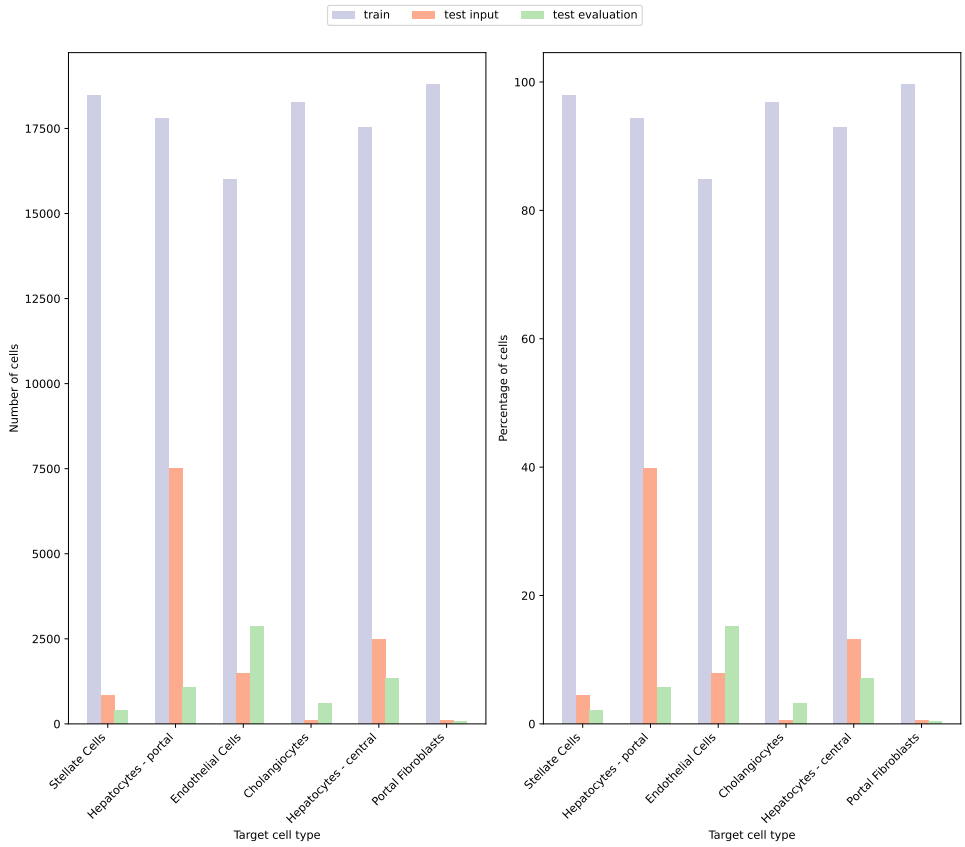
12.1 Multiple doses





12.2 Single dose 30 $\mu g/kg$





12.3 Comparison

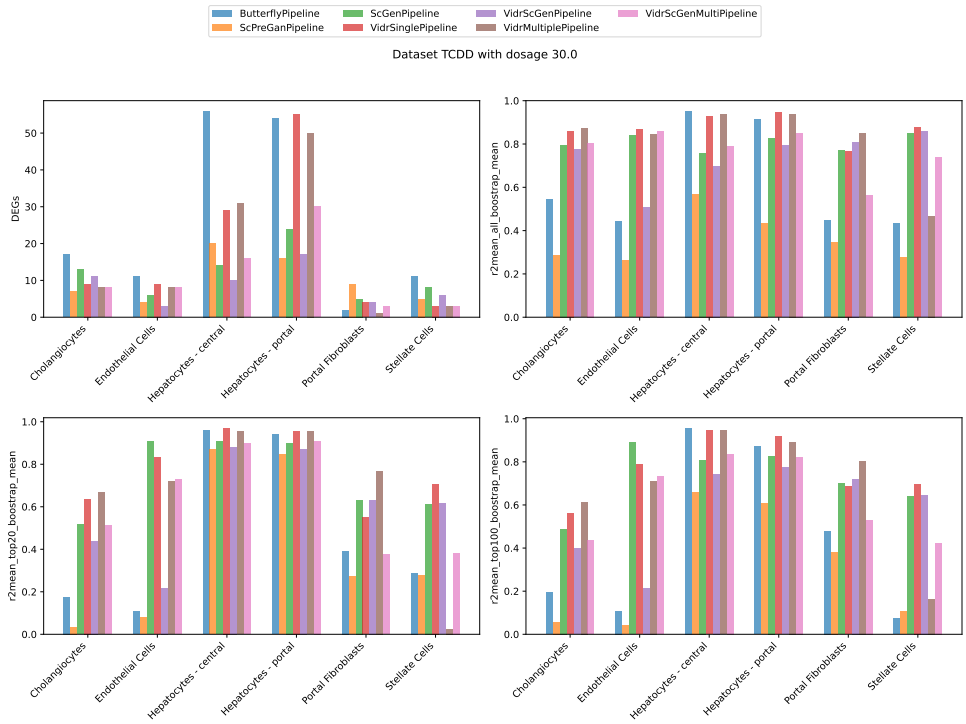


Figure 73: Baseline metrics for highest dosage $30\mu g/kg$ across cell types

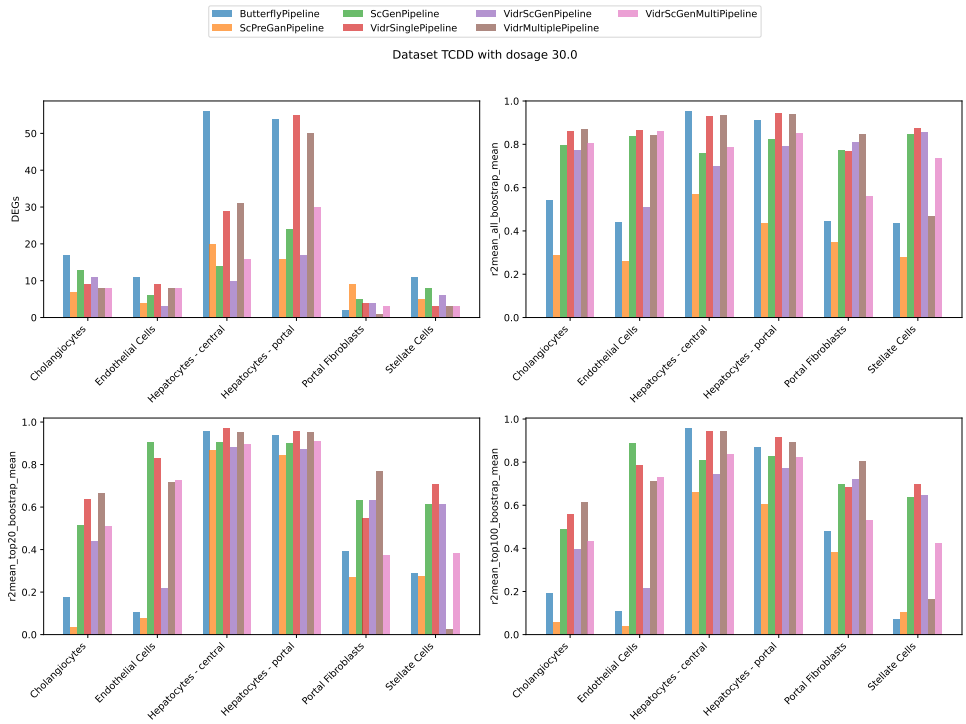


Figure 74: Distance metrics for highest dosage $30\mu g/kg$ across cell types

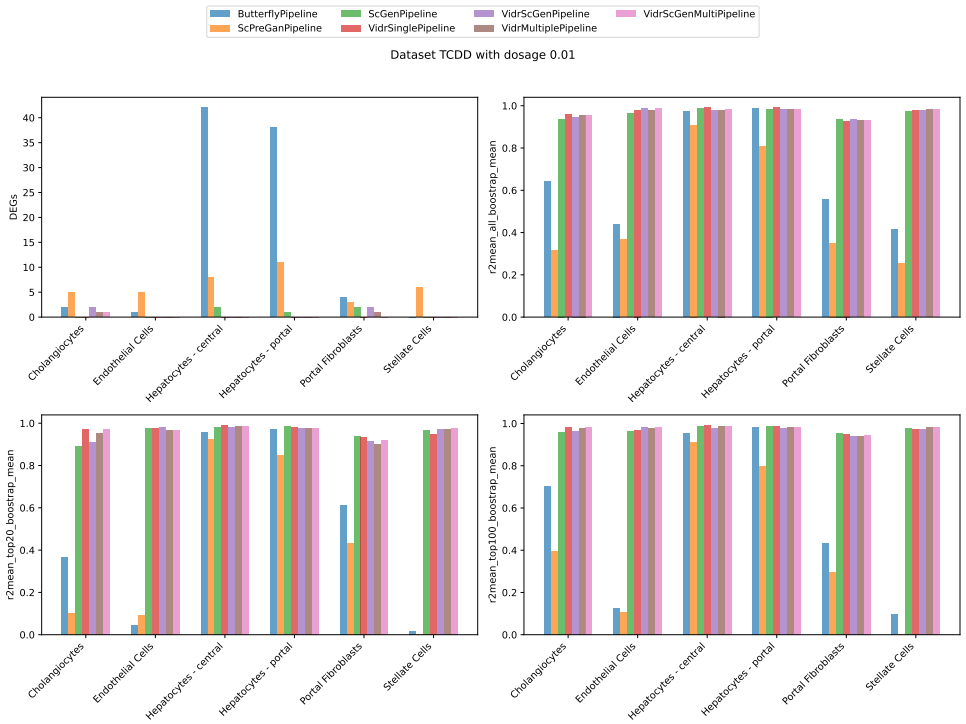


Figure 75: Baseline metrics for highest dosage $0.1\mu g/kg$ across cell types

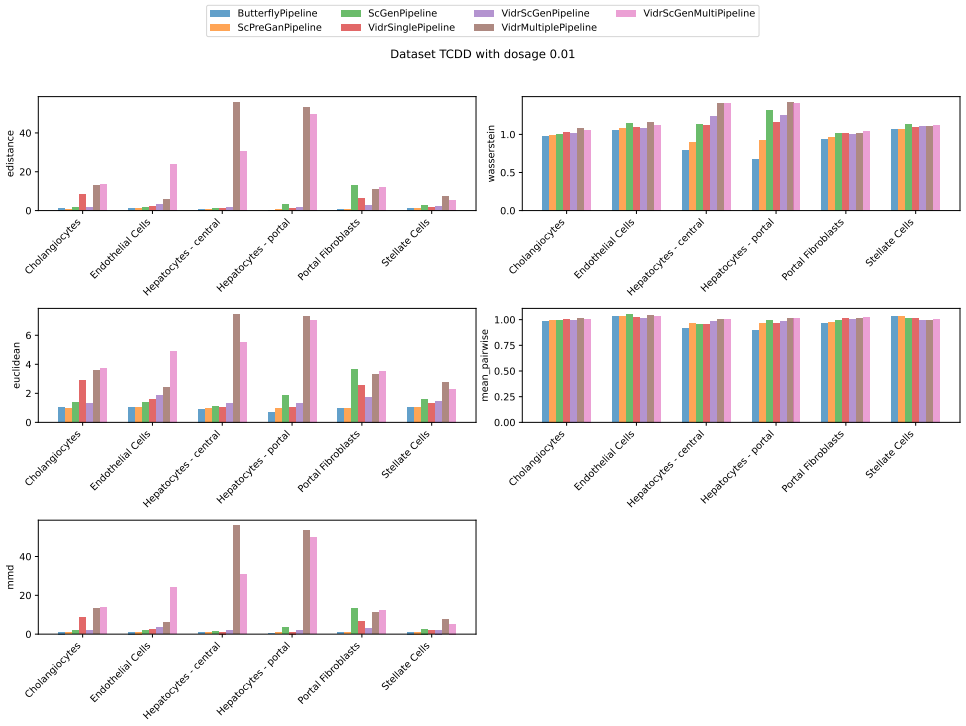


Figure 76: Distance metrics for highest dosage $0.1\mu g/kg$ across cell types

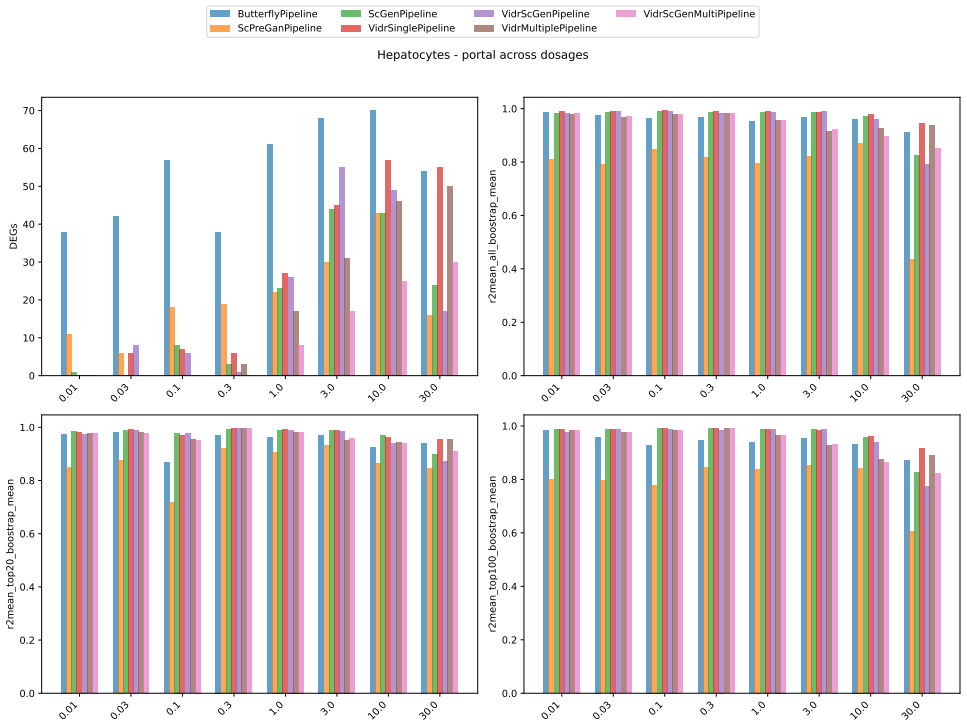


Figure 77: Baseline metrics for Hepatocytes - portal across dosages

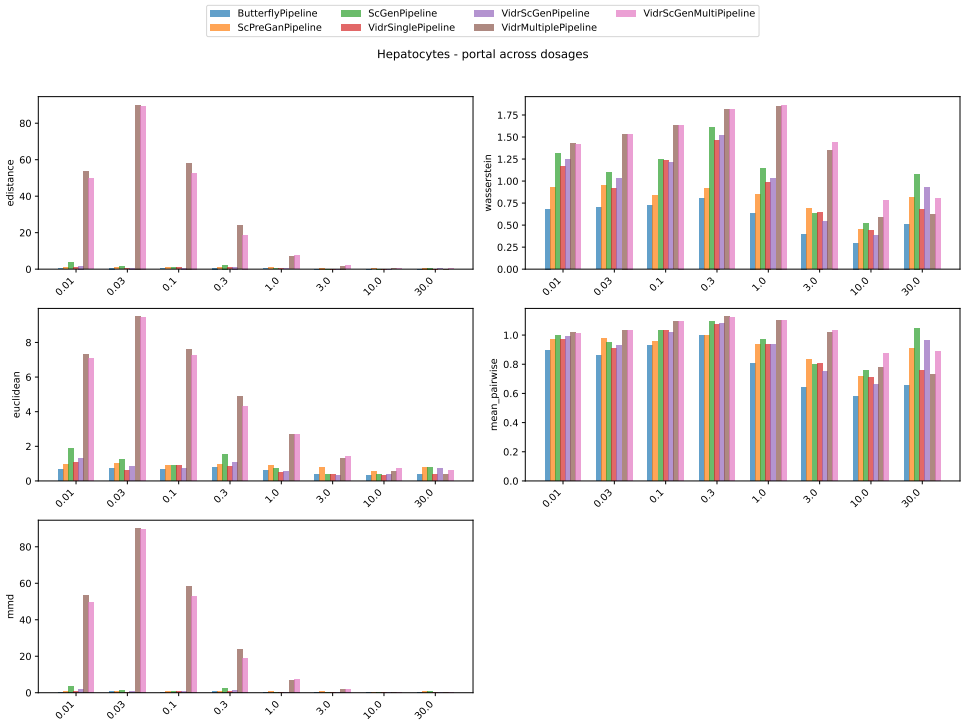


Figure 78: Distance metrics for Hepatocytes - portal across dosages

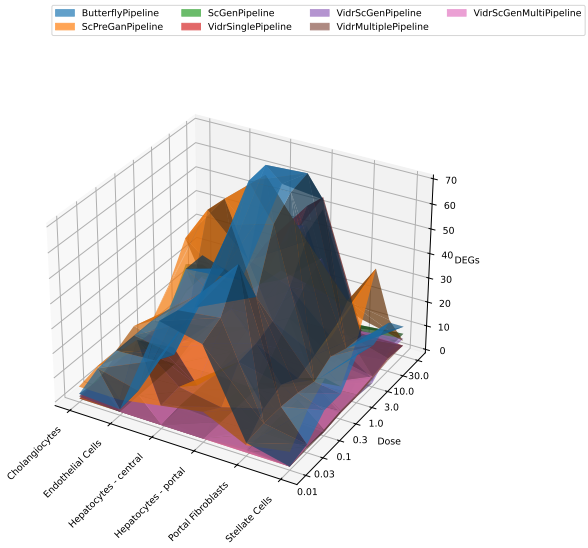


Figure 79

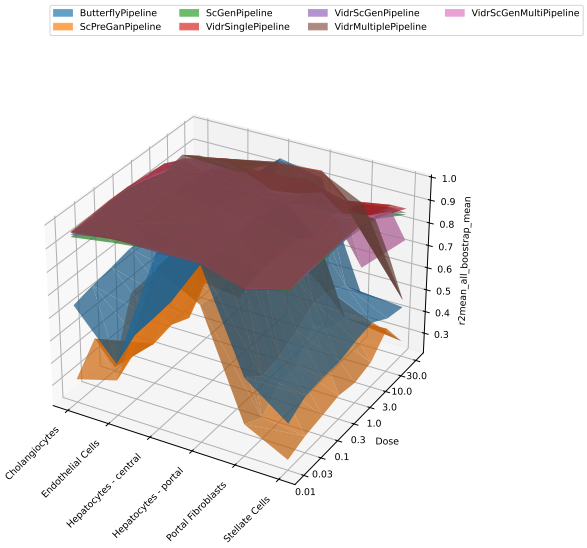


Figure 80

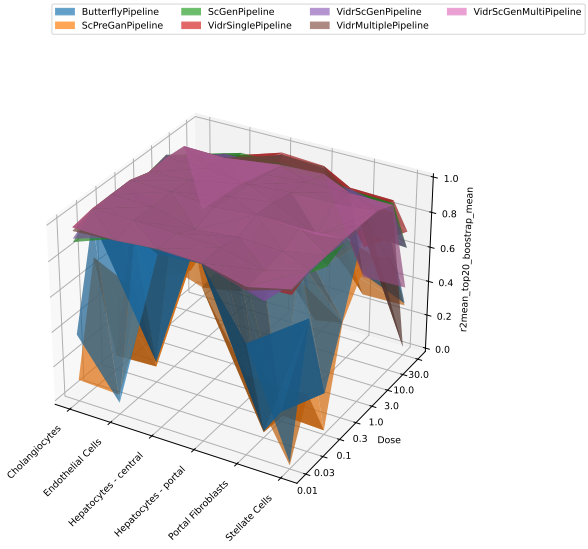


Figure 81

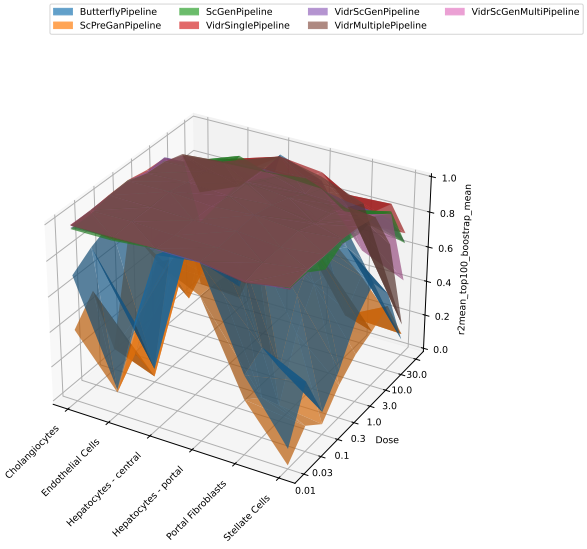


Figure 82

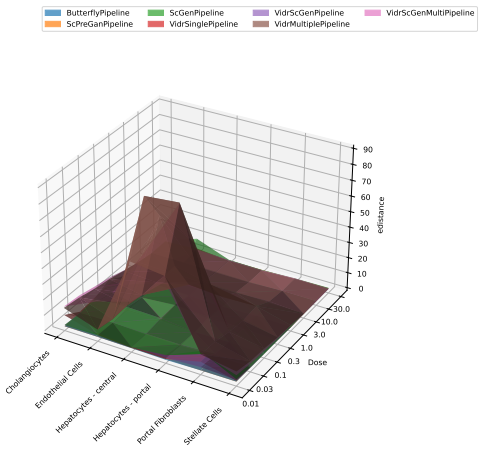


Figure 83: E-distance

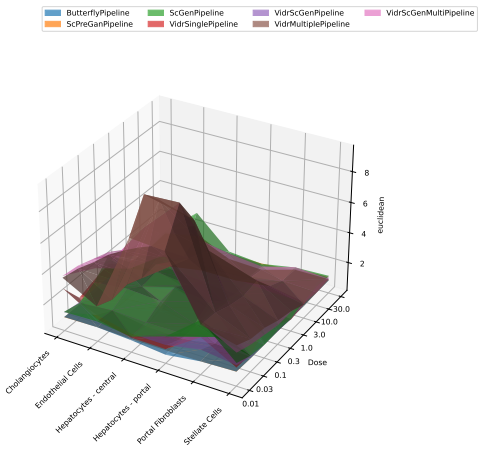


Figure 84: Euclidean

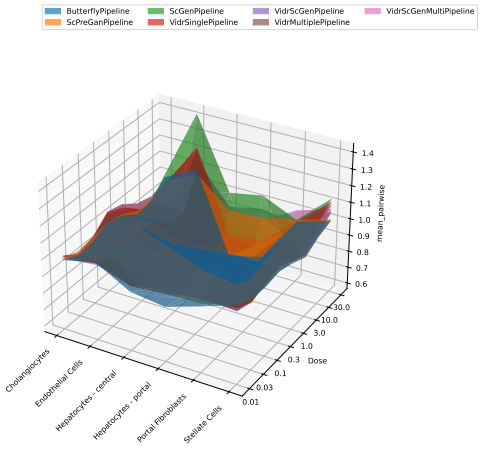


Figure 85: Mean pairwise

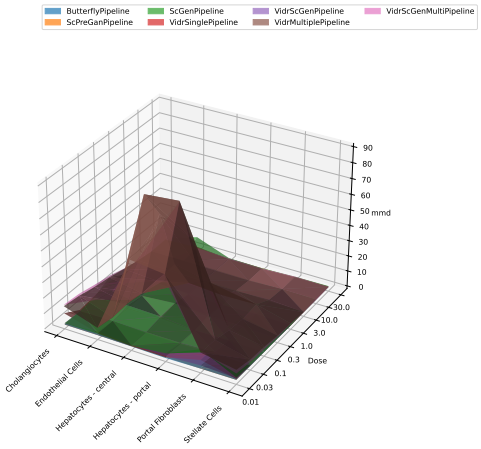


Figure 86: MMD

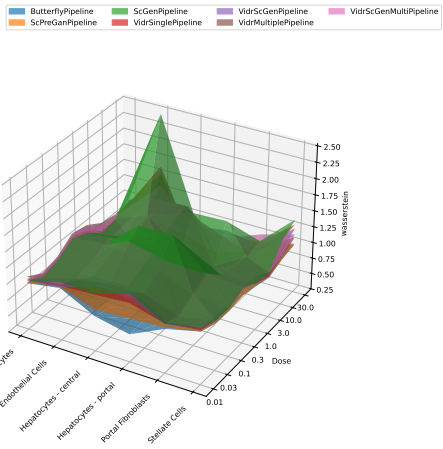
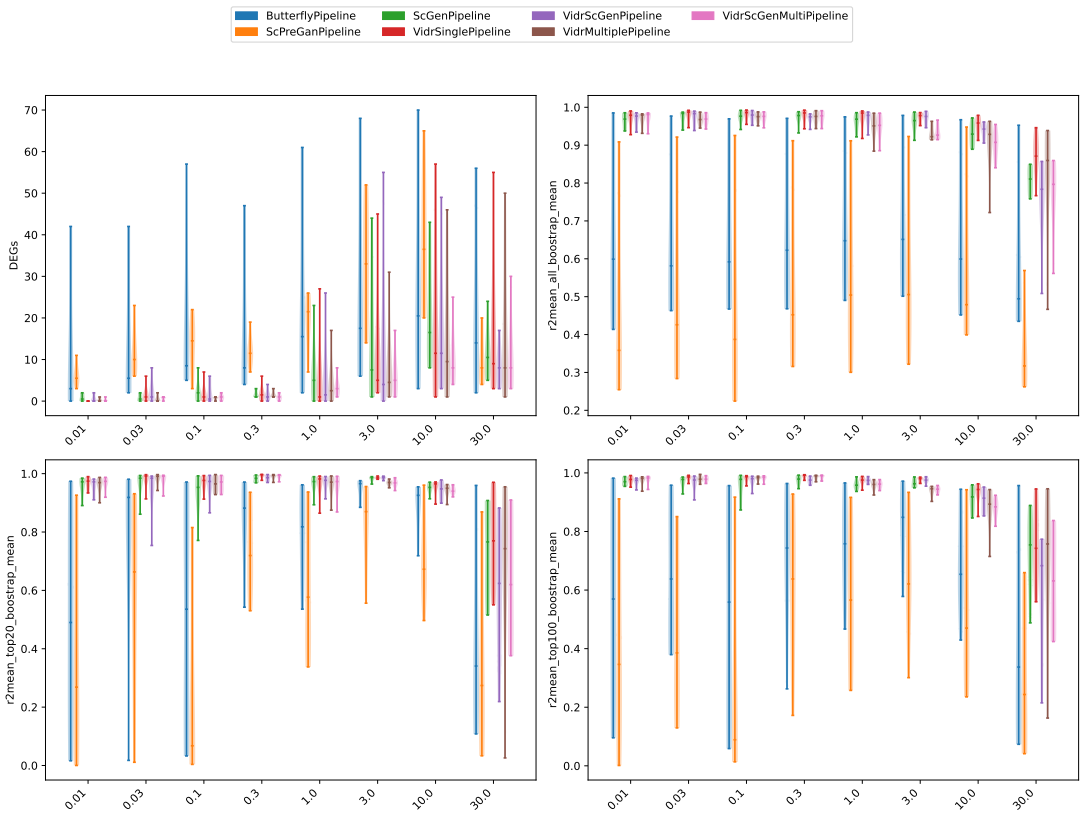
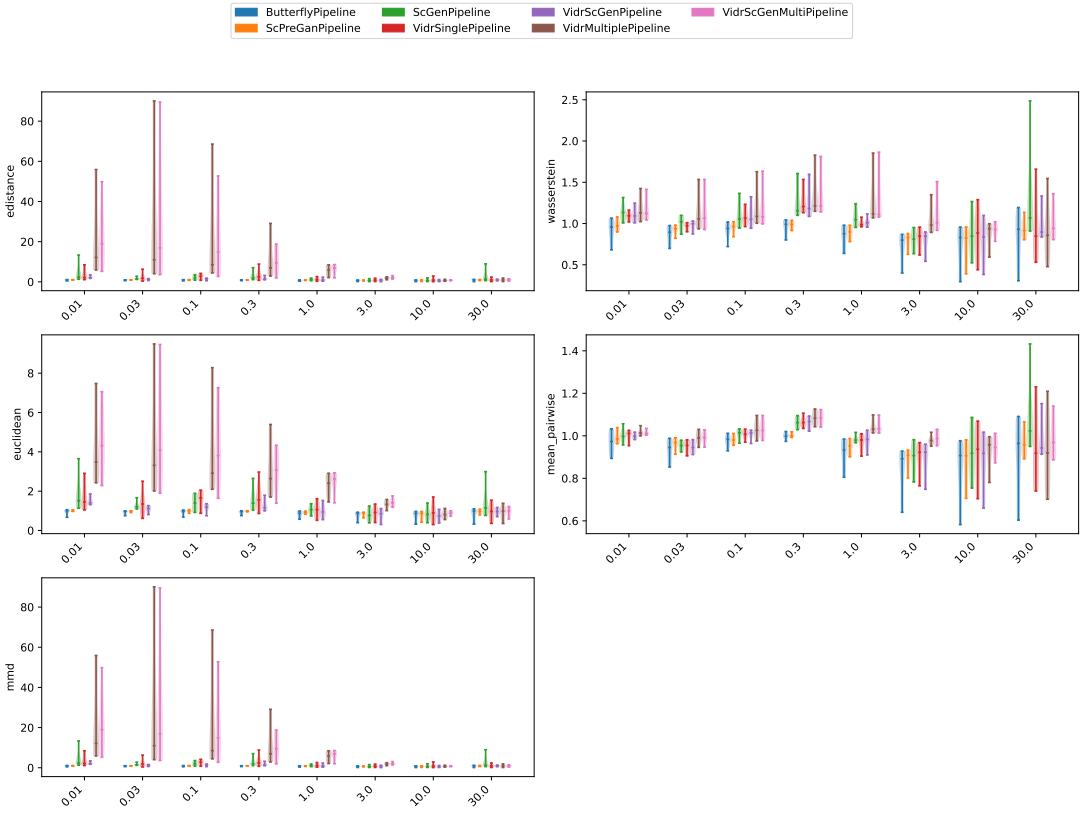
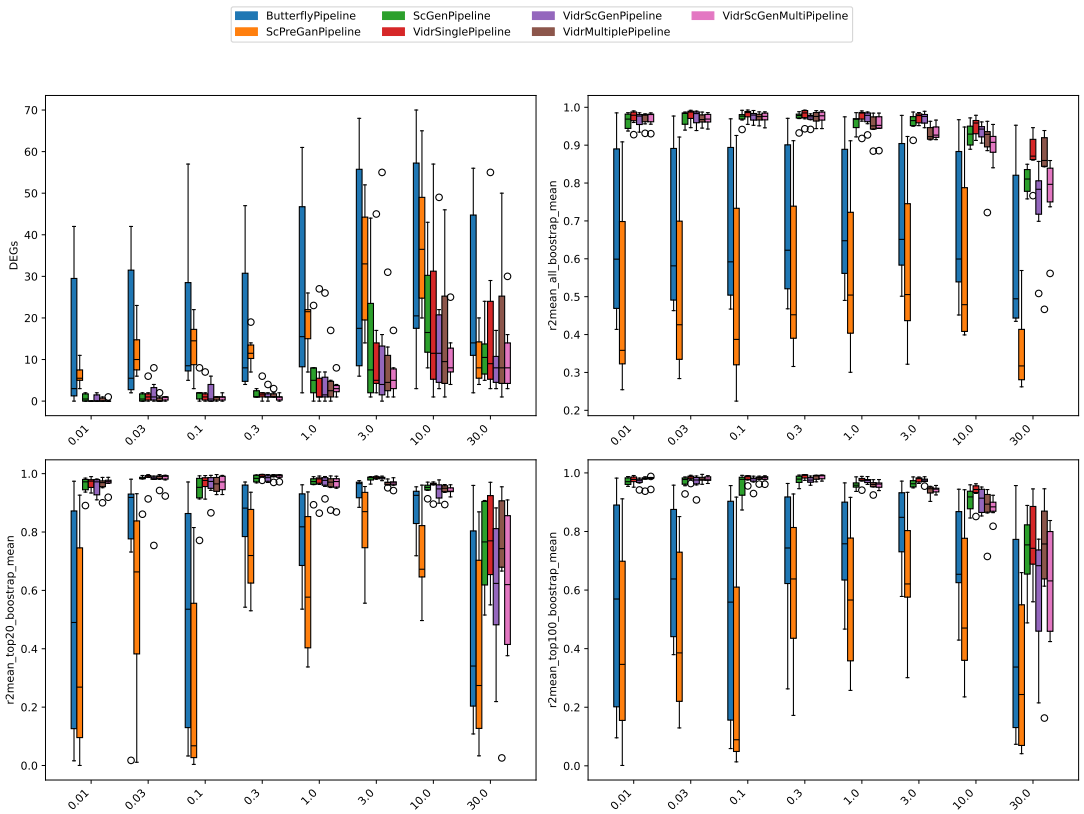


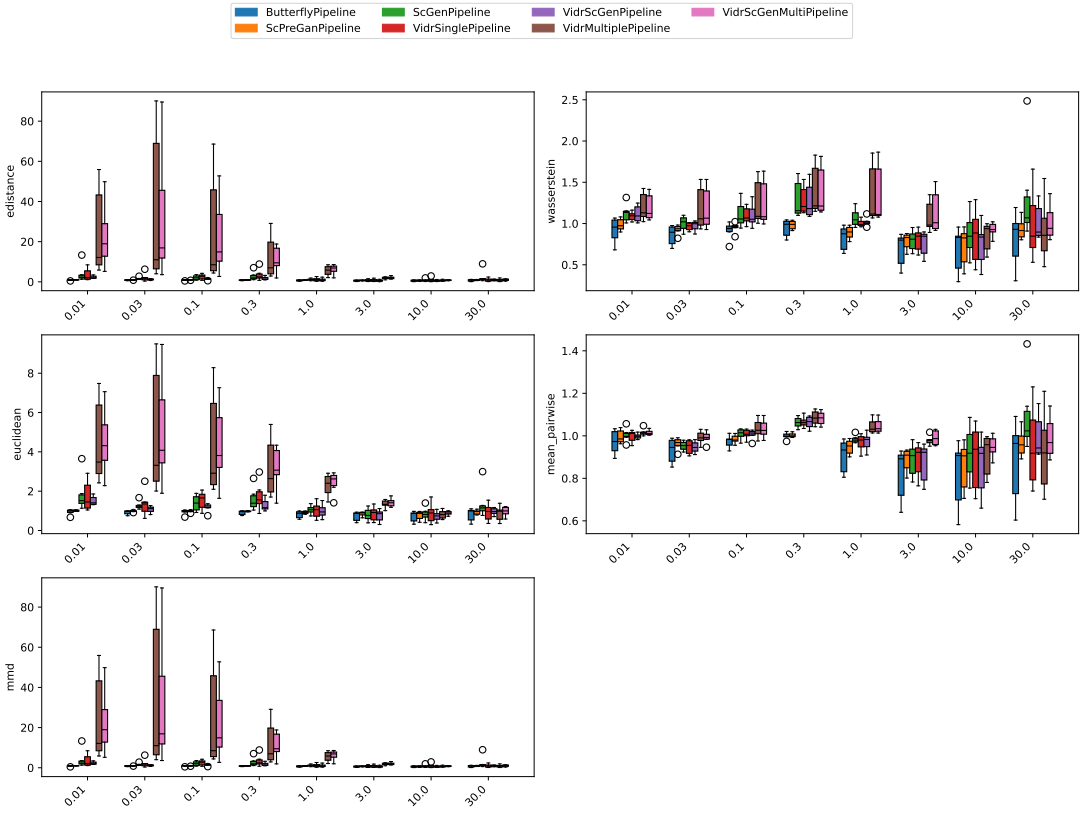
Figure 87: Wasserstein

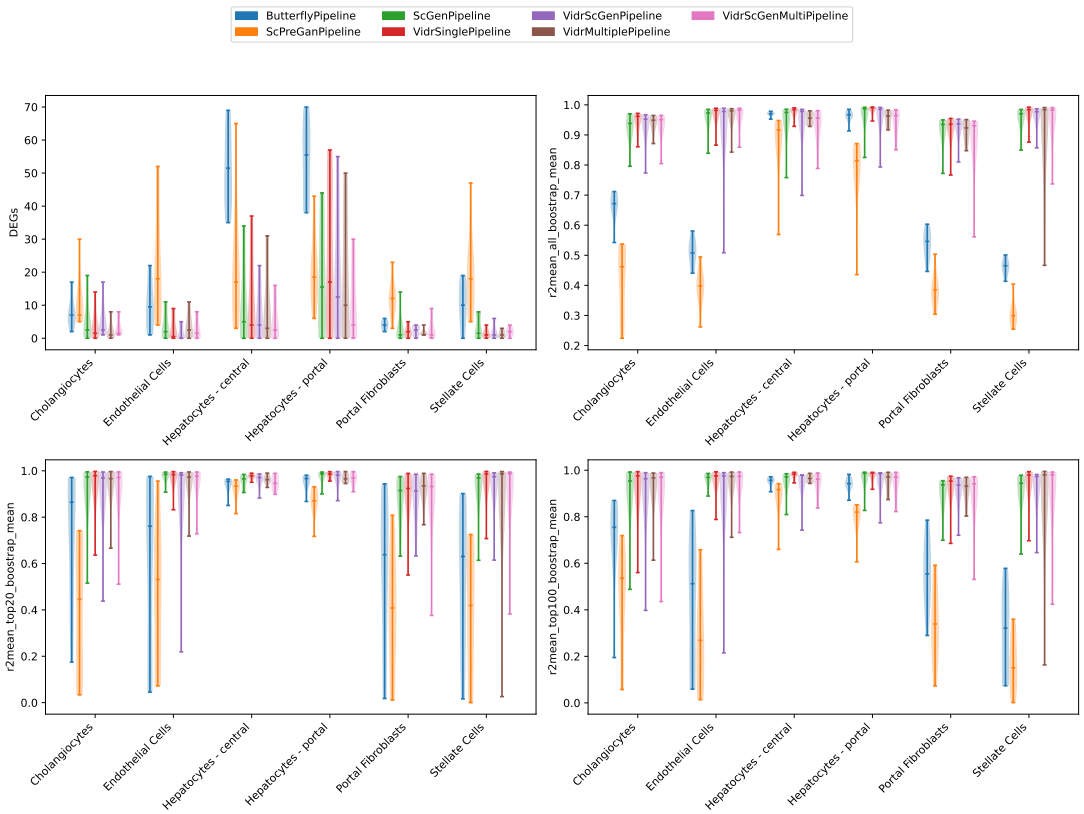
Figure 88: Distance metrics per cell type

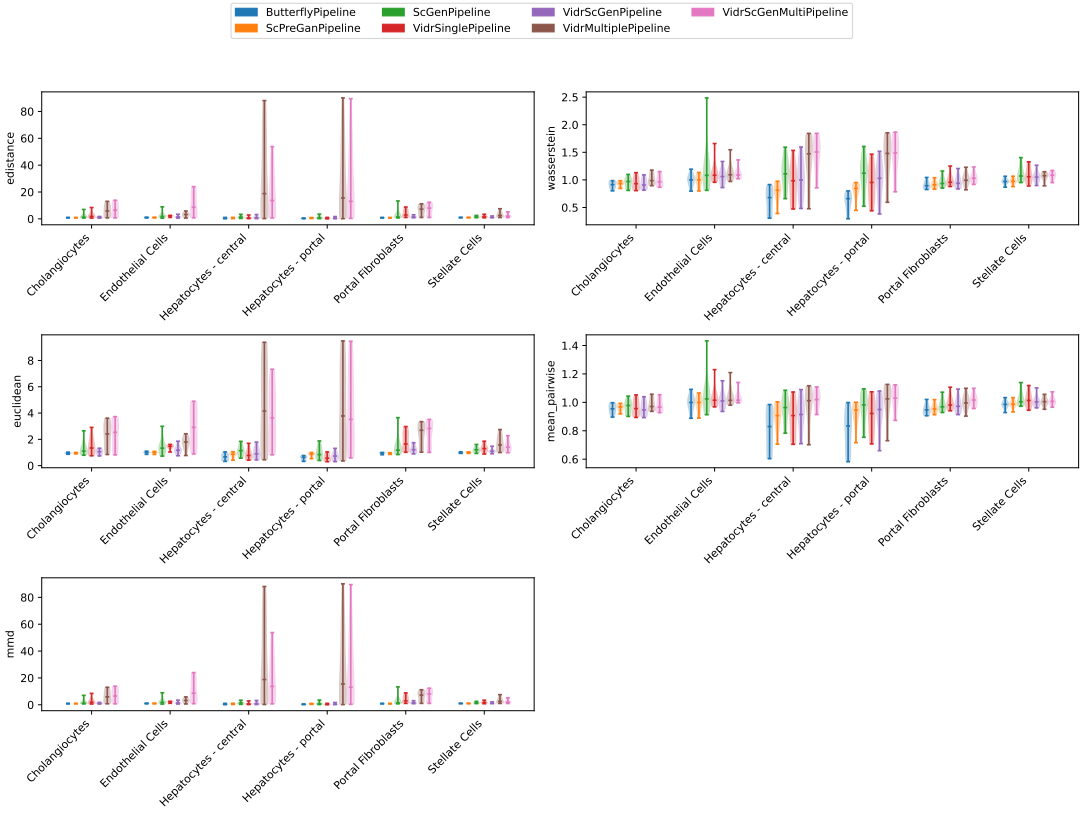

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0

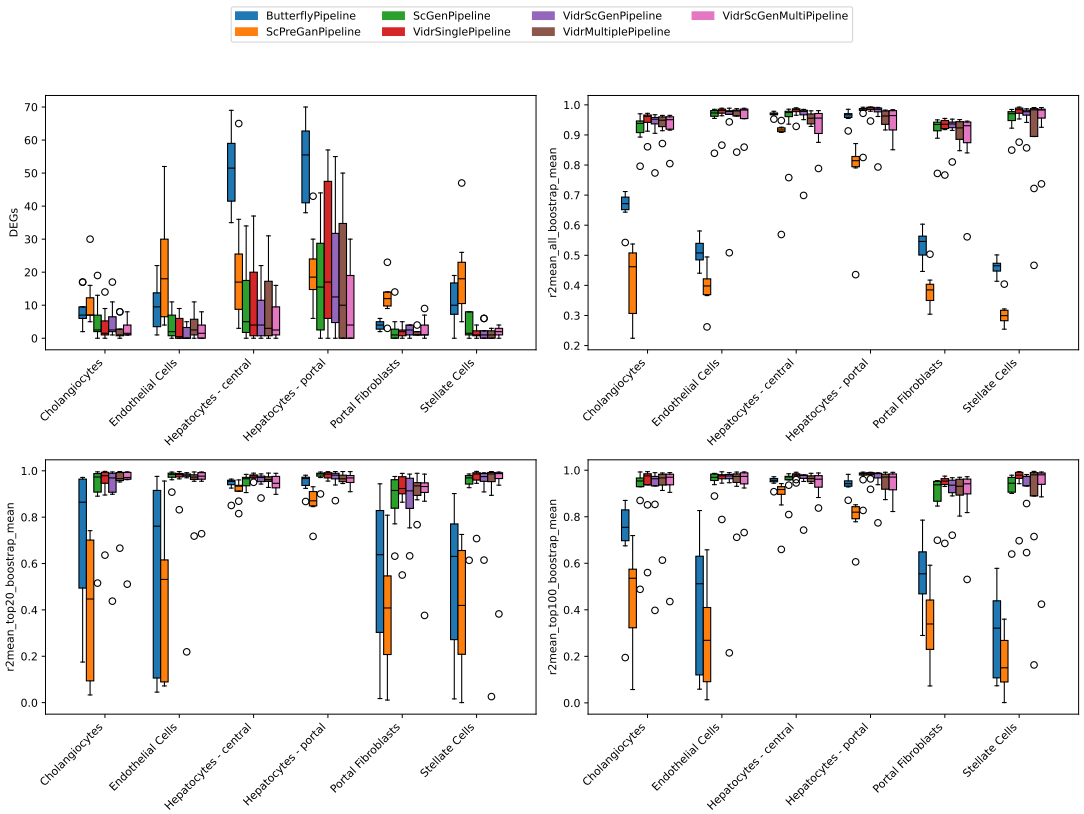

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
edistance
wasserstein
euclidean
mean_pairwise
mmd
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0

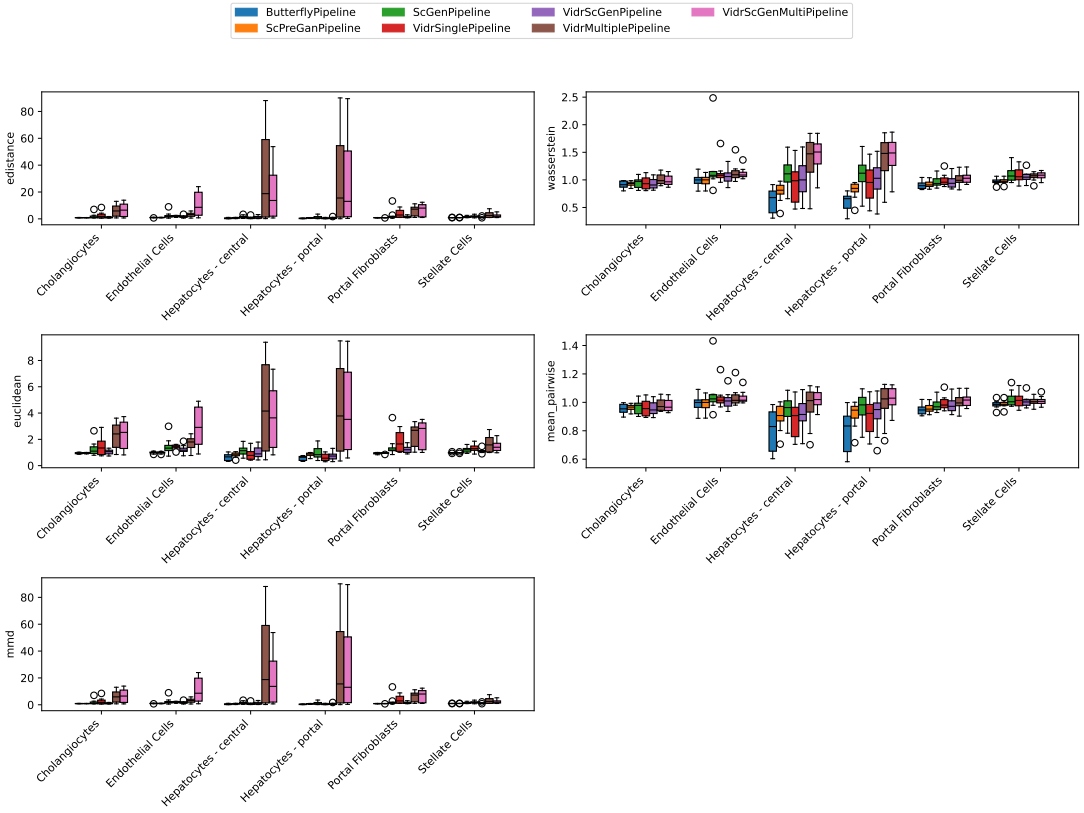

ButterflyPipeline
ScPreGanPipeline
ScGenPipeline
VidrSinglePipeline
VidrScGenPipeline
VidrMultiplePipeline
VidrScGenMultiPipeline
DEGs
r2mean_all_boostrap_mean
r2mean_top20_boostrap_mean
r2mean_top100_boostrap_mean
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0


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
VidrSingleLineage
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
VidrSingleLineage
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

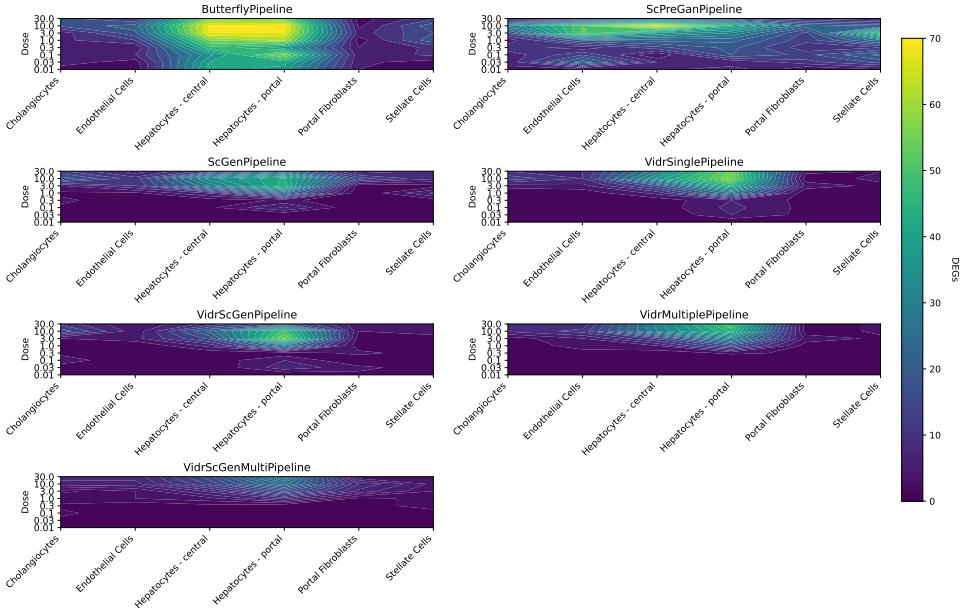


Figure 89: DEGs

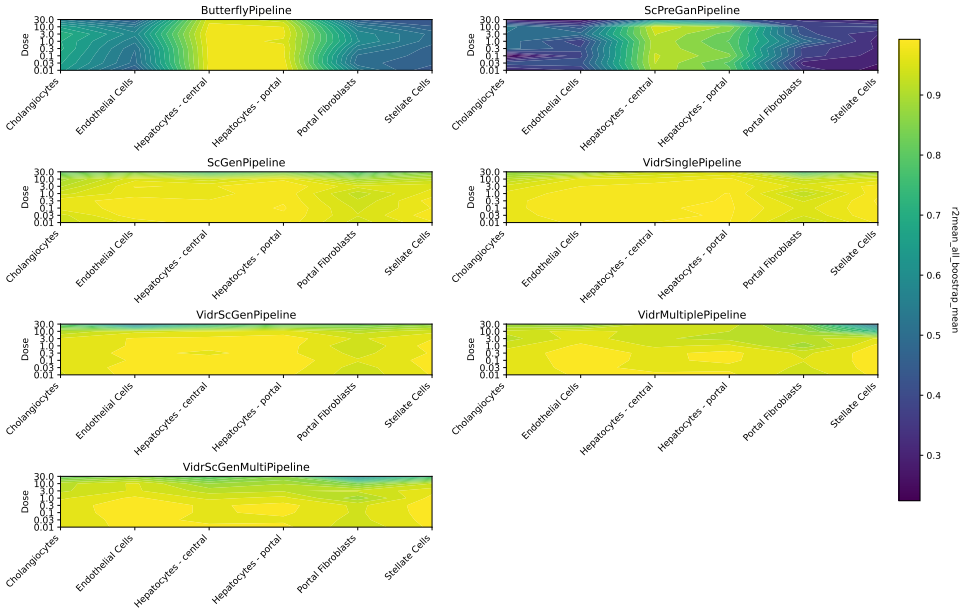


Figure 90: r2 HVGs

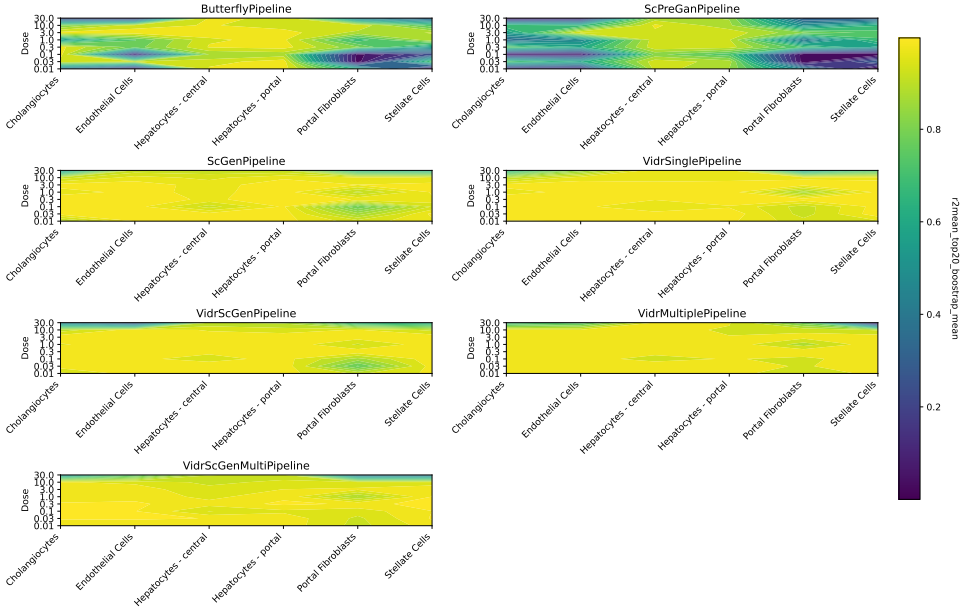


Figure 91: r2 top 20

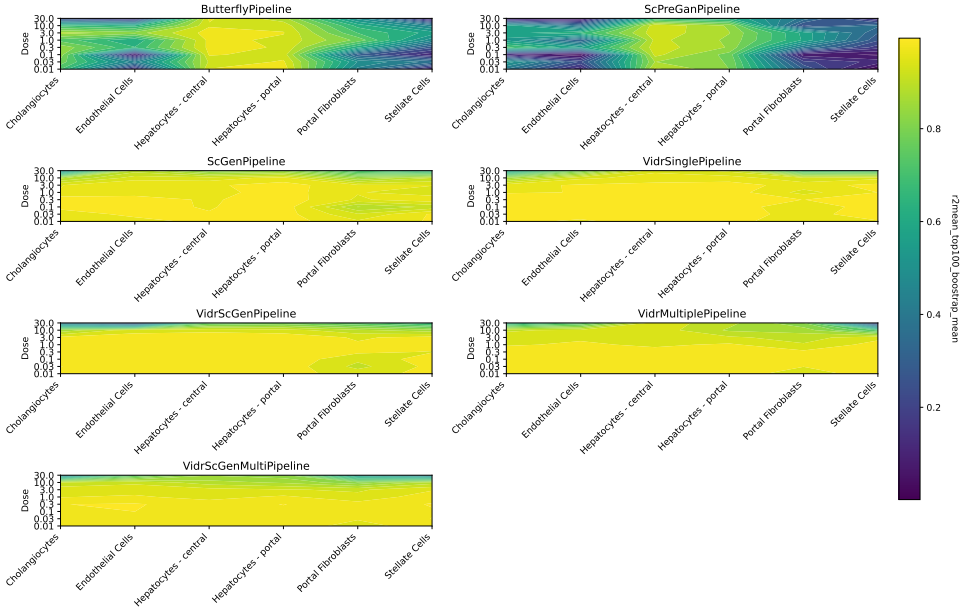


Figure 92: r2 top 100

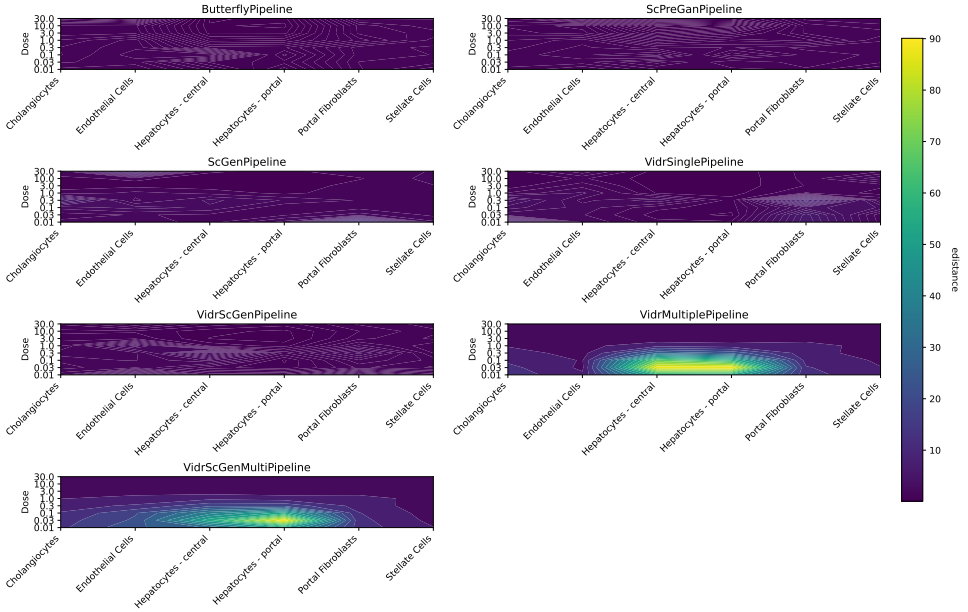
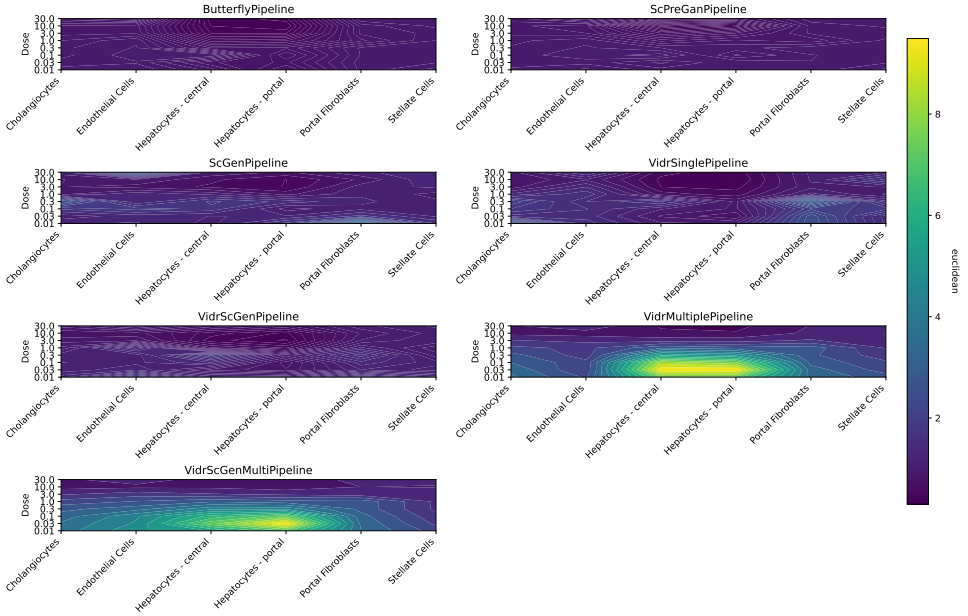


Figure 93: E-distance



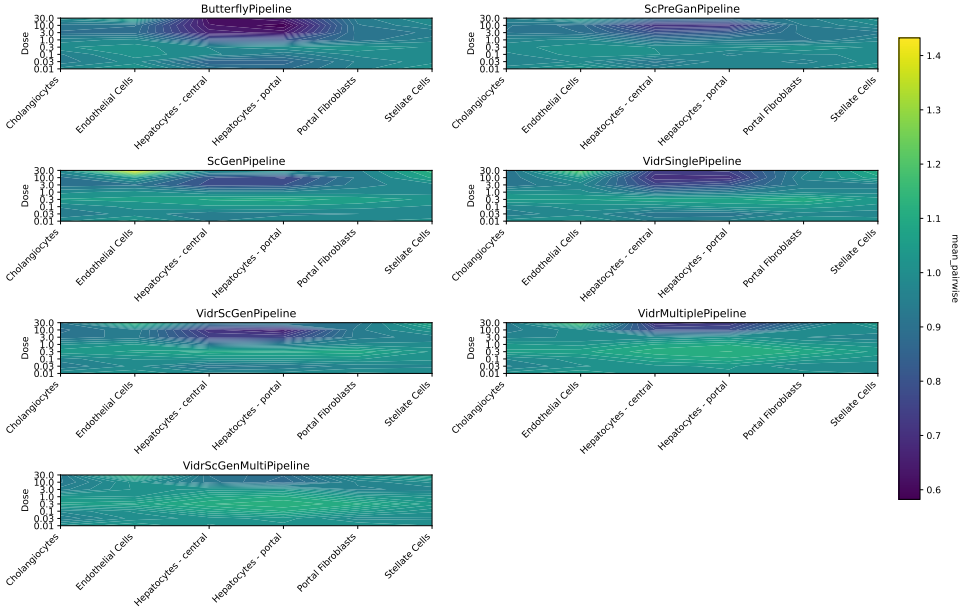


Figure 94: Mean pairwise

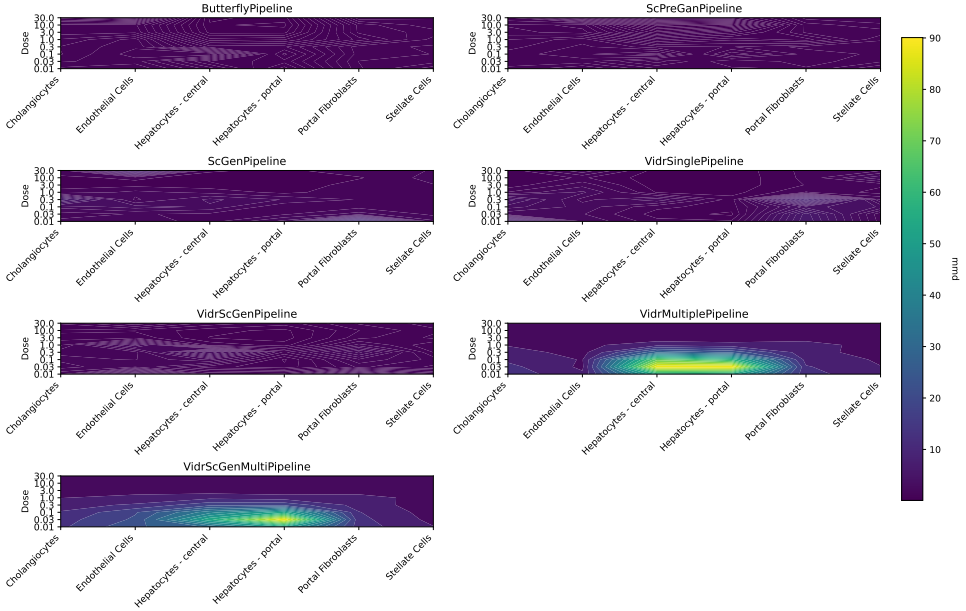


Figure 95: MMD

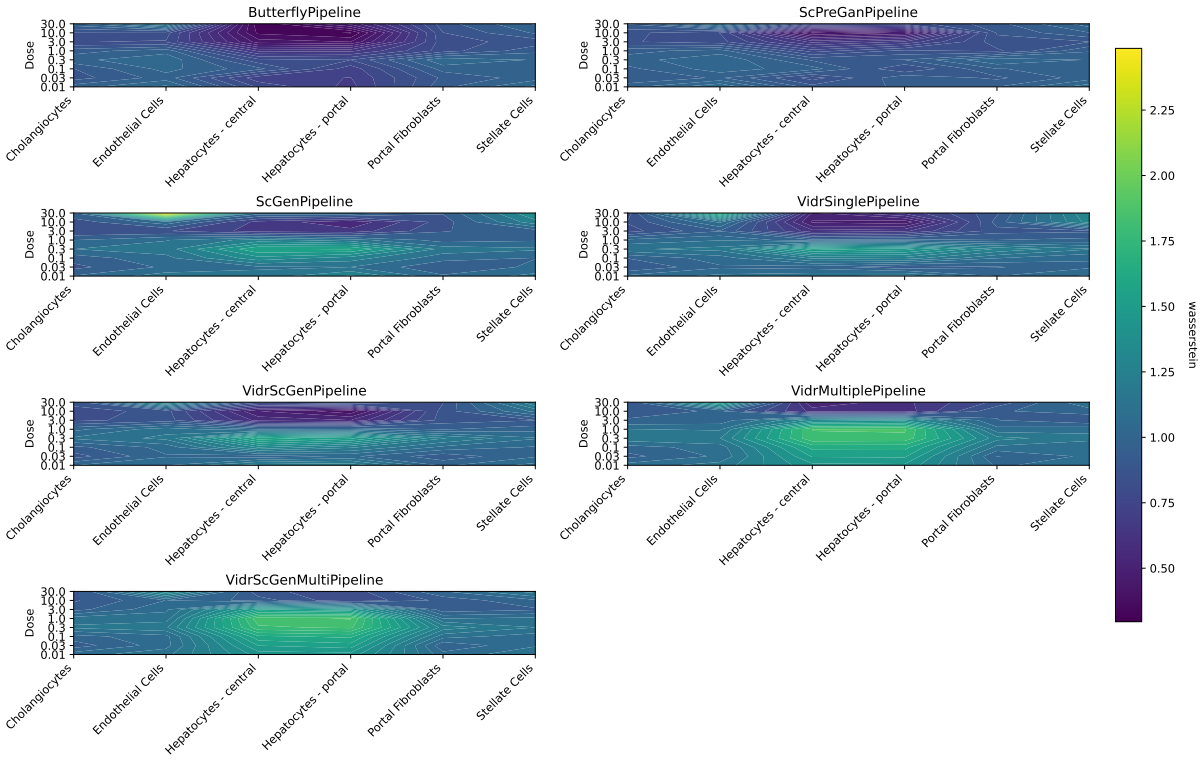
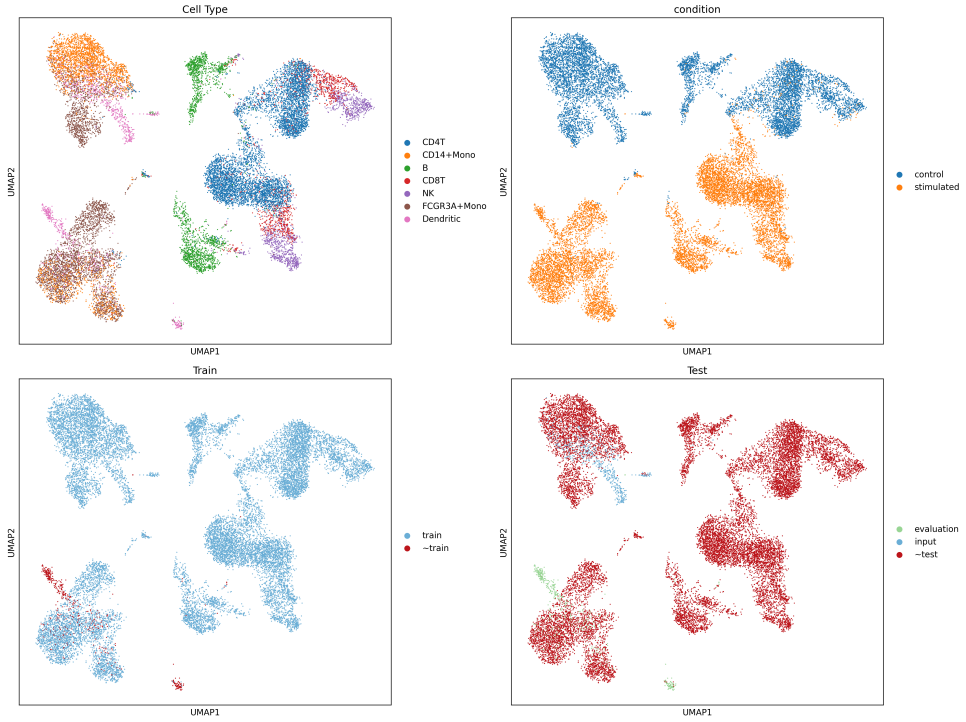


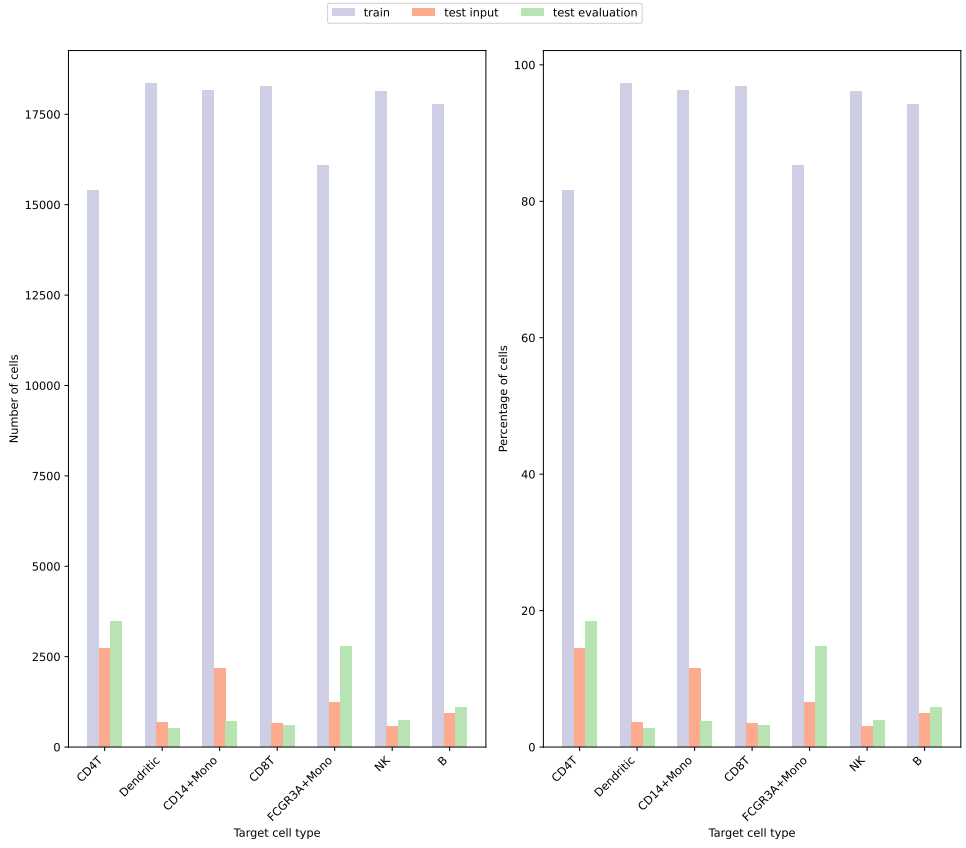
Figure 96: Wasserstein

12.4 Παρατηρήσεις

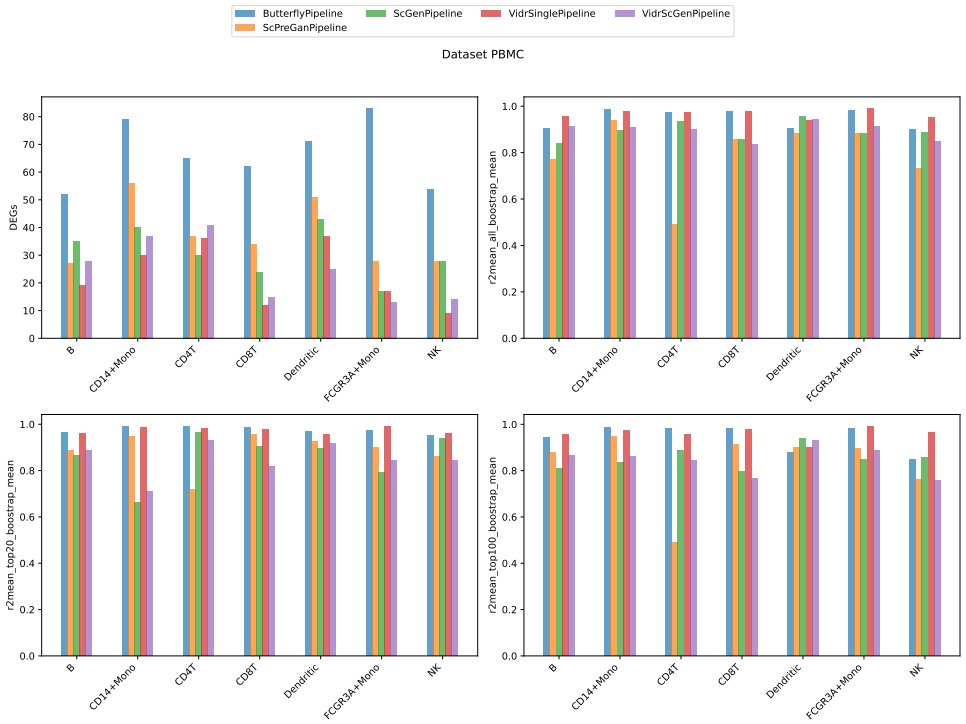
- Το scButterfly και το scPreGan έχουν παρόμοια συμπεριφορά στις μετρικές και εμφανίζουν μεγάλη διακύμανση κατά μήκος των τύπων των κυττάρων και των δόσεων.
- Τα μοντέλα που έχουν ως βάση την αρχιτεκτονική του scGen (scVIDR, και οι παραλλαγές του), VAR και post-processing στο latent space, έχουν την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

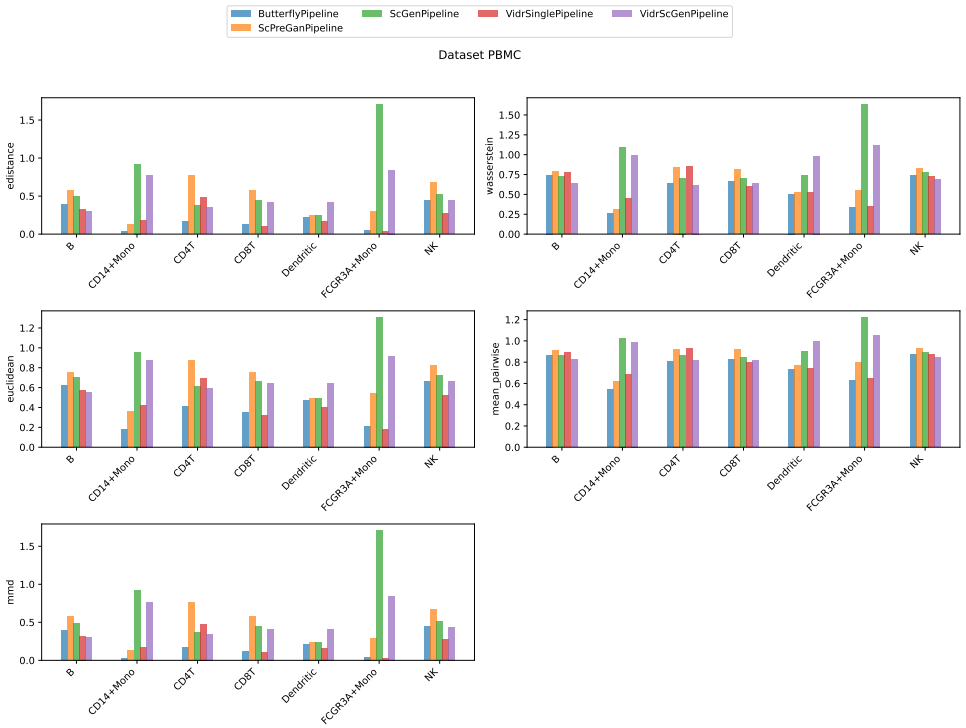
13 PBMC


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


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

13.1 Comparison





Α Ακρωνύμια και συντομογραφίες

LAN Local Area Network

References

[1] Yichuan Cao, Xiamiao Zhao, Songming Tang, Qun Jiang, Sijie Li, Siyu Li, and Shengquan Chen. scButterfly: A versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. 15(1):2973.

[2] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 2018. https://distill.pub/2018/feature-wise-transformations.

[3] 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.

[4] 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.

[5] Omar Kana, Rance Nault, David Filipovic, Daniel Marri, Tim Zacharewski, and Sudin Bhattacharya. Generative modeling of single-cell gene expression for dose-dependent chemical perturbations. 4(8):100817.

[6] Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. 16(8):715–721.

[7] Xin Tang, Jiawei Zhang, Yichun He, Xinhe Zhang, Zuwan Lin, Sebastian Partarrieu, Emma Bou Hanna, Zhaolin Ren, Hao Shen, Yuhong Yang, Xiao Wang, Na Li, Jie Ding, and Jia Liu. Explainable multi-task learning for multi-modality biological data analysis. 14(1):2546.

[8] Xiajie Wei, Jiayi Dong, and Fei Wang. scPreGAN, a deep generative model for predicting the response of single-cell expression to perturbation. 38(13):3377–3384.

[9] Yu Zhang and Qiang Yang. A Survey on Multi-Task Learning.