



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 2, 2025

Abstract

abstract

Abstract

abstract

Ευχαριστίες

Τίτλος διπλωματικής

Όνομα Επίθετο
empty@auth.gr

April 2, 2025

Περιεχόμενα

0.1	Abstract	2
0.2	Introduction	2
0.3	Baseline	2
0.4	Method	2
0.5	Results and discussion	2
0.6	Conclusions	2
0.7	Future work	2
1	Benchmarking	3
1.1	Datasets	4
1.1.1	Nault et al. 2022	4
1.1.2	PBMC dataset	11
1.2	Nault all cell types evaluation	14
1.2.1	Multiple doses	14
1.2.2	Single dose	15
1.2.3	Comparison	16
1.3	Nault liver cell types evaluation	30
1.3.1	Multiple doses	30
1.3.2	Single dose $30 \mu g/kg$	31
1.3.3	Comparison	32
1.3.4	Παρατηρήσεις	46
1.4	PBMC	47
1.4.1	Comparison	48
A	Ακρωνύμια και συντομογραφίες	49

0.1 Abstract

With the recent advancements in single cell technology and the large scale perturbation datasets, the field of perturbation modeling [1] 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.

0.2 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 (add citation), has a significant impact to the biomedicine sector and drug discovery. With the recent surge of data generation, machine learning methods have addressed to extrapolate on the perturbation atlas, given a limited number of perturbation experiments.

An overview of the models on perturbation modeling can be found on this study (cite the mini review). One of its main objectives is to predict the out-of-distribution samples, which is the focal point of our study. The task is about predicting the perturbation response of the transcriptome of a cell type, while having observed the perturbation responses of other cell types.

One of the key problems of deep learning methods is the data demand. We assume that the utilization of the data of multiple perturbations, under a multi-task context, to be beneficial to mitigate this. UnitedNet, an explainable multi-modal framework, had shown the potential of multi-task learning in a multi-omics dataset. We aim to extend this approach to perturbation modeling.

0.3 Baseline

We have chosen the models of scGen, scButterfly, scPregan, and scVIDR to compare our multi-task model.

scGen is considered a fundamental model to the perturbation community, thus used in our model repertoire. It uses 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.

0.4 Datasets

Datasets used along with preprocessing.

0.5 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

0.6 Results and discussion

To compare the models, a unified common ground is needed. To our best knowledge, a standatadized benchmarking apporach on perturbation modeling is lacking. We have attempted to review the models in a multi-faceted way, including dimensions metrics, along with the most used ones such as DEGs, and r² of all and the most highly variable genes.

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.

0.7 Conclusions

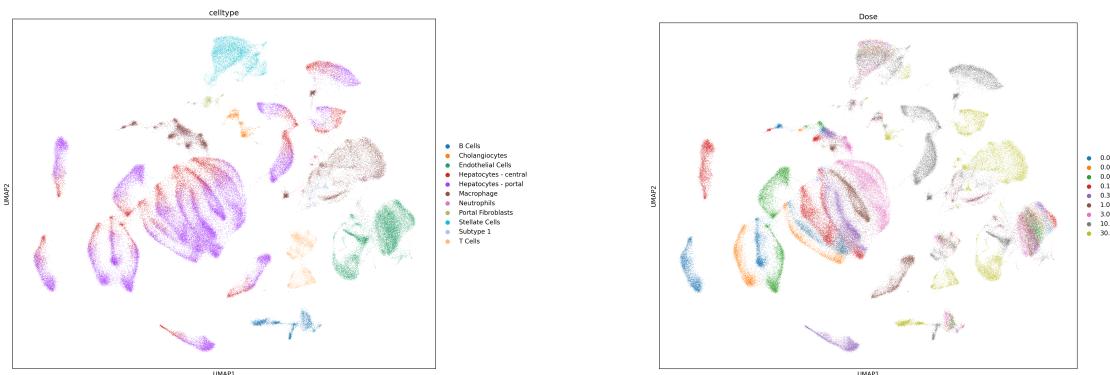
0.8 Future work

Chapter 1

Benchmarking

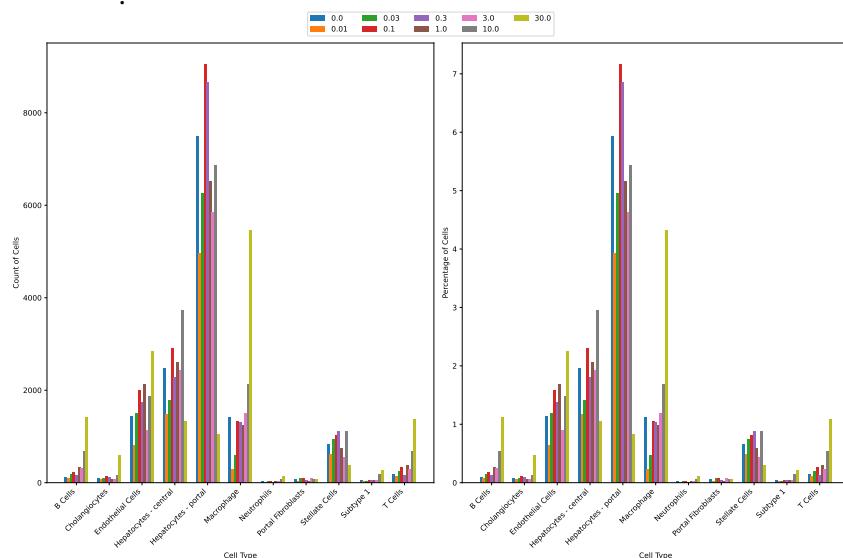
1.1 Datasets

1.1.1 Nault et al. 2022

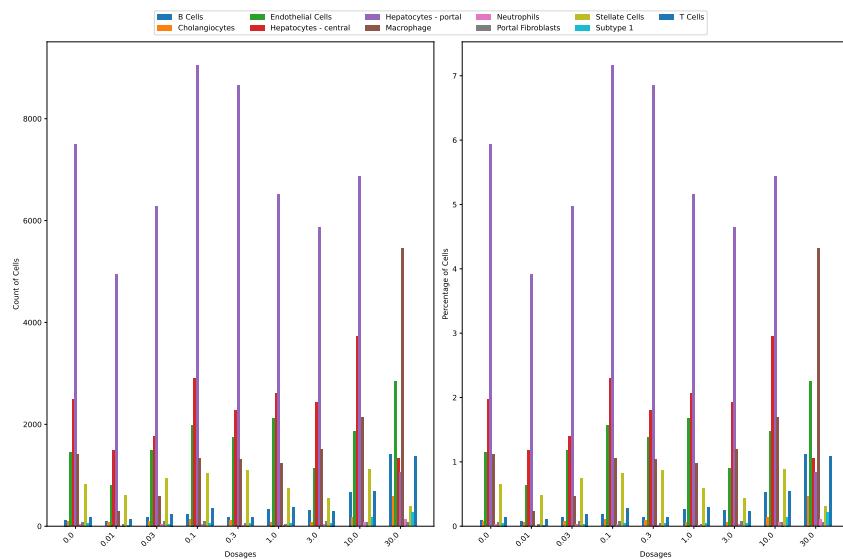


(a)

(b)



(c)



(d)

Figure 1.1: Nault overview

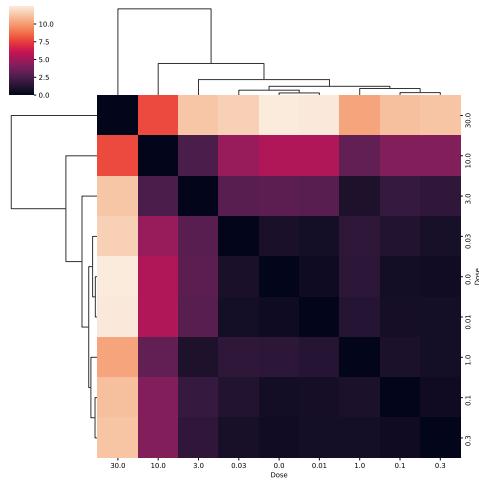


Figure 1.2: E-distance

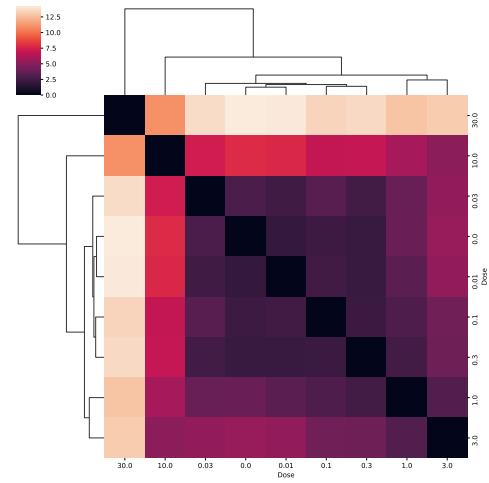


Figure 1.3: Euclidean

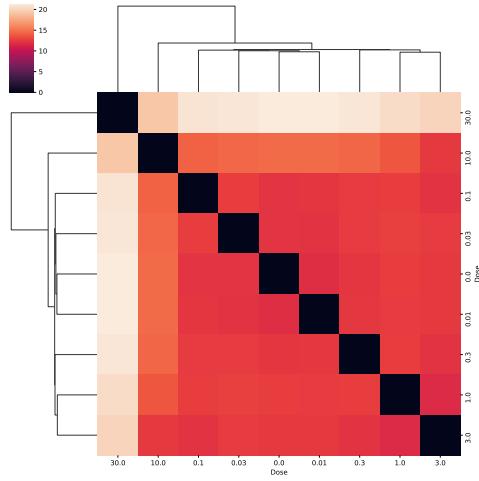


Figure 1.4: Mean pairwise

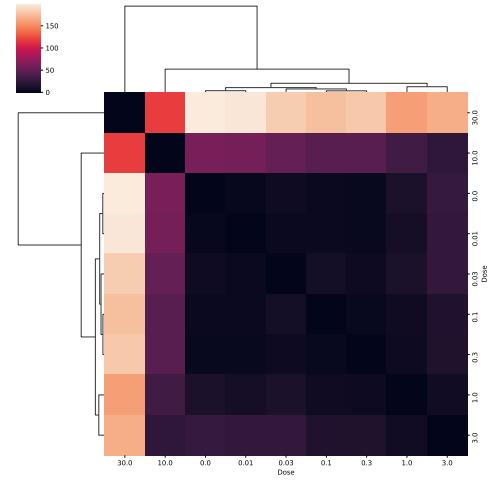


Figure 1.5: MMD

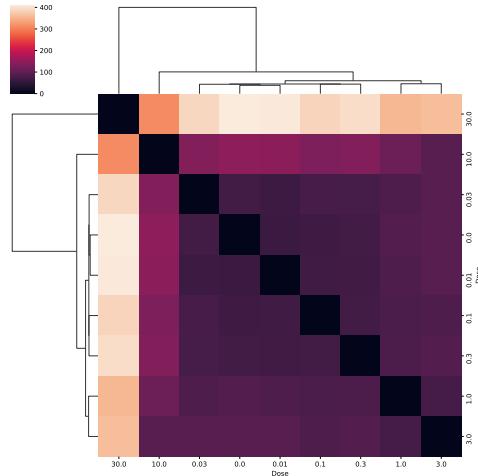


Figure 1.6: Wasserstein

Figure 1.7: Distance metrics across all cell types per dosage

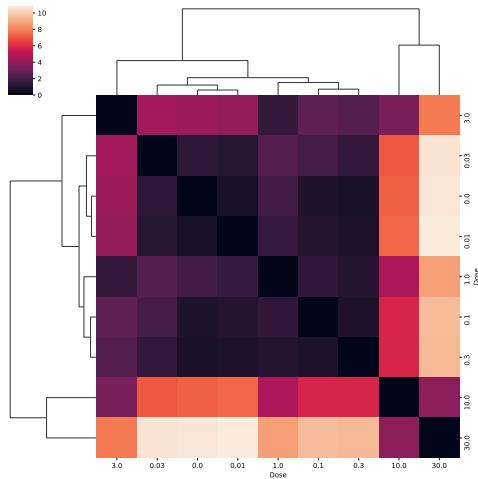


Figure 1.8: E-distance

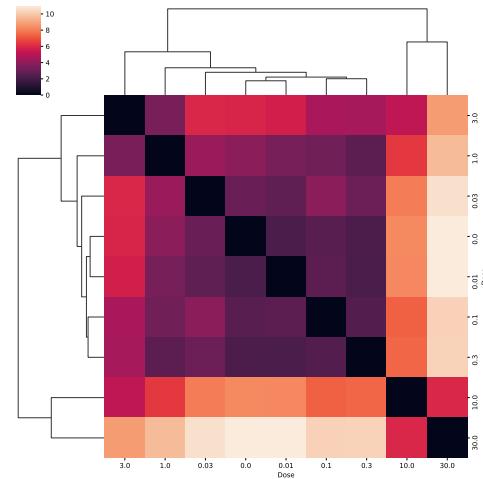


Figure 1.9: Euclidean

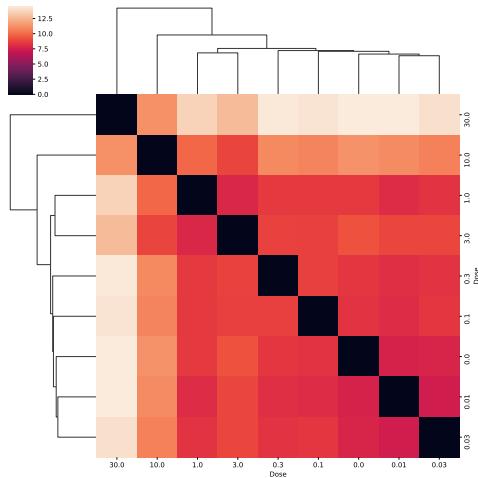


Figure 1.10: Mean pairwise

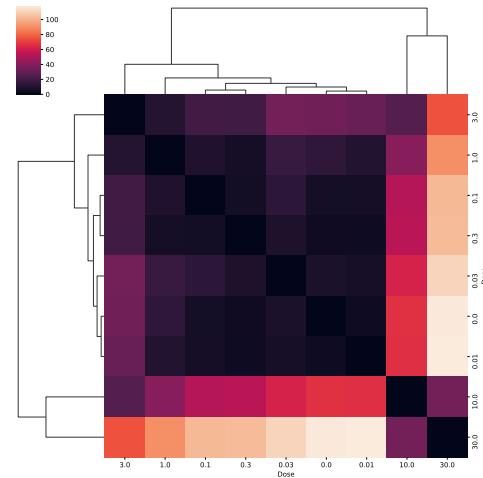


Figure 1.11: MMD

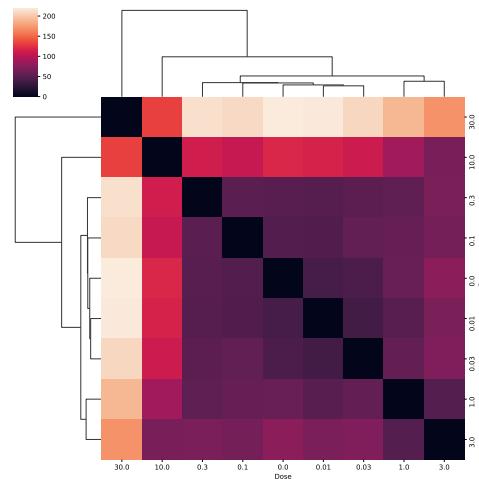


Figure 1.12: Wasserstein

Figure 1.13: Distance metrics for cell type Hepatocytes - portal per dosage

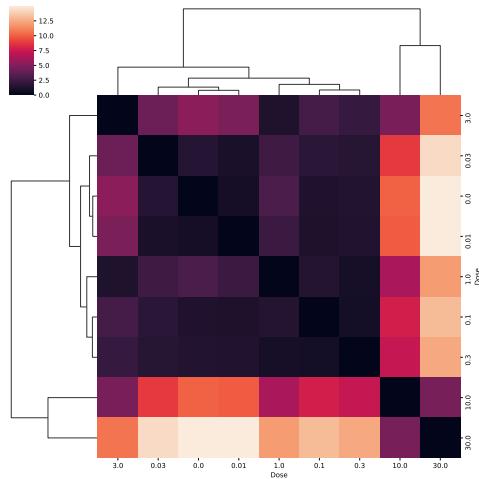


Figure 1.14: E-distance

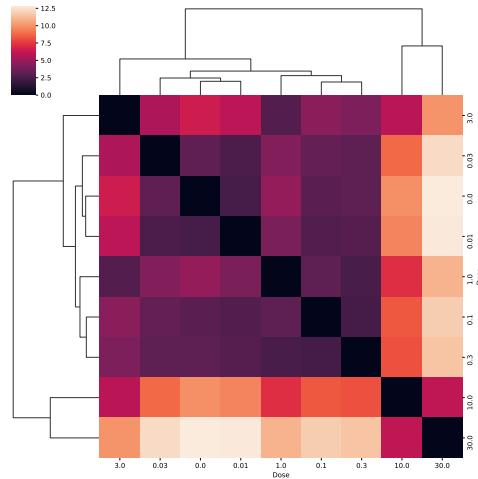


Figure 1.15: Euclidean

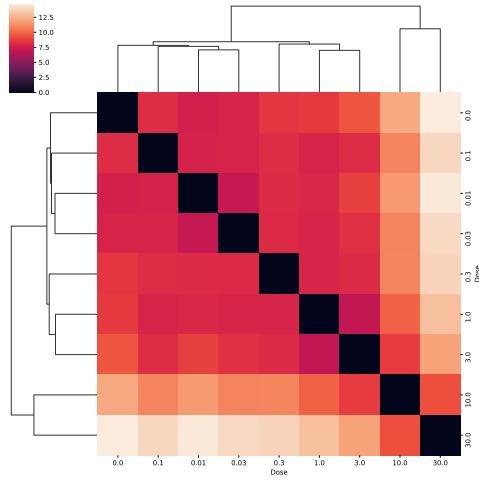


Figure 1.16: Mean pairwise

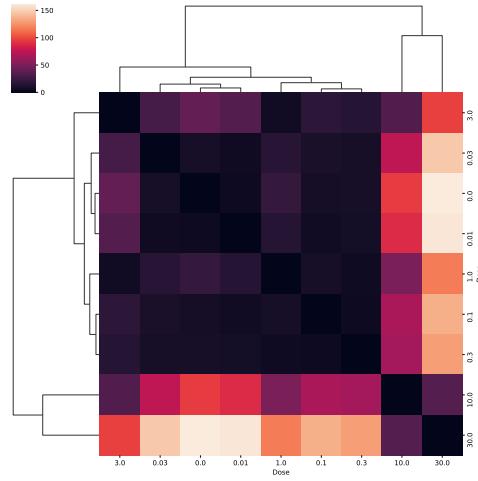


Figure 1.17: MMD

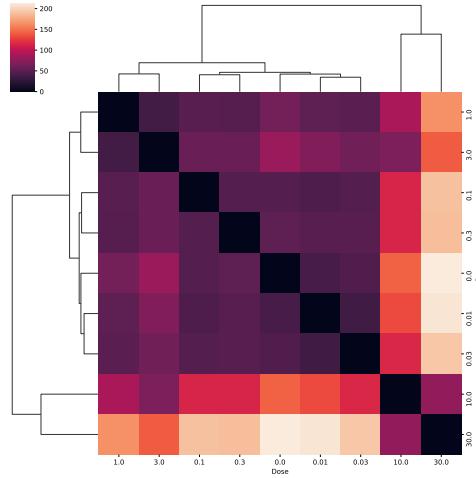


Figure 1.18: Wasserstein

Figure 1.19: Distance metrics for cell type Hepatocytes - central per dosage

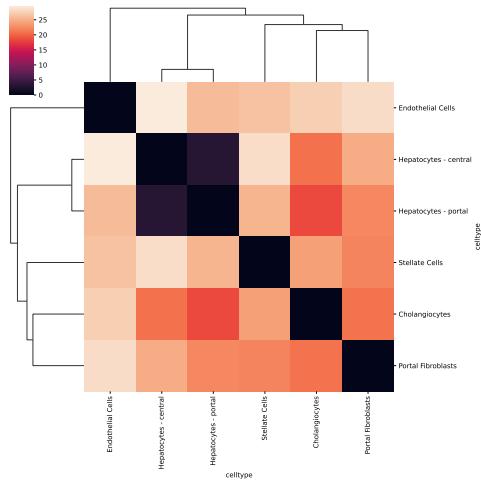


Figure 1.20: E-distance

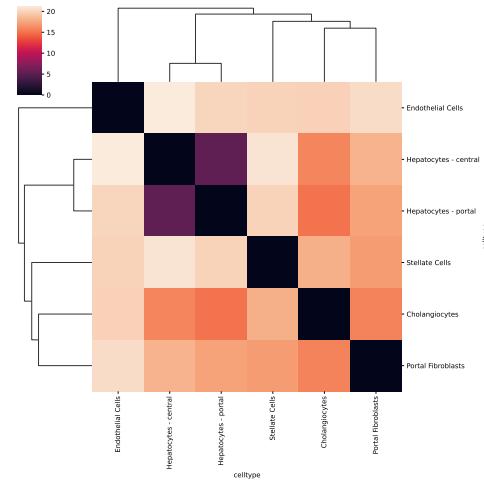


Figure 1.21: Euclidean

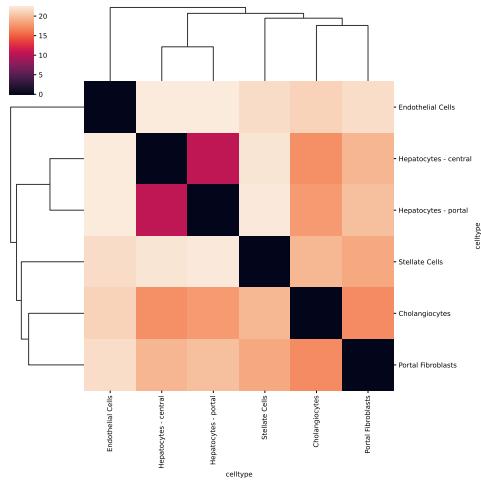


Figure 1.22: Mean pairwise

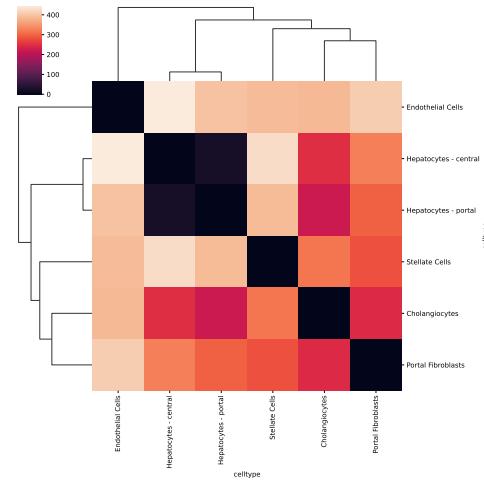


Figure 1.23: MMD

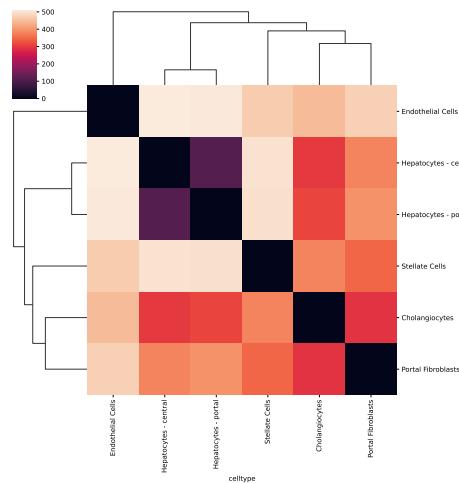


Figure 1.24: Wasserstein

Figure 1.25: Distance metrics for dosage highest $30 \mu\text{g}/\text{kg}$ per cell type

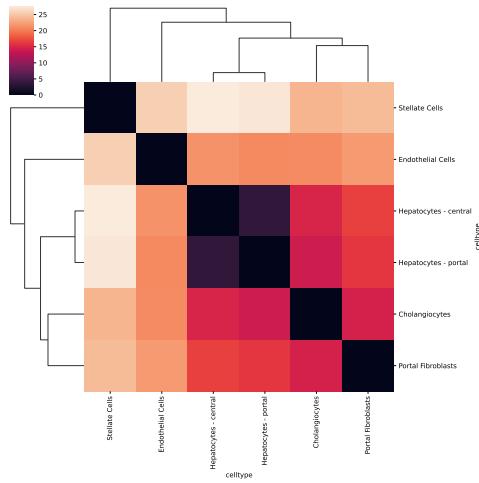


Figure 1.26: E-distance

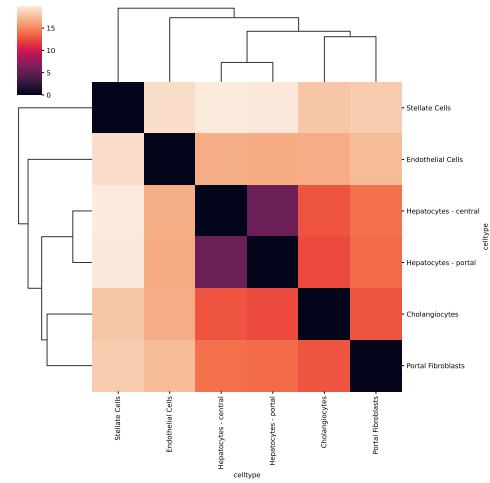


Figure 1.27: Euclidean

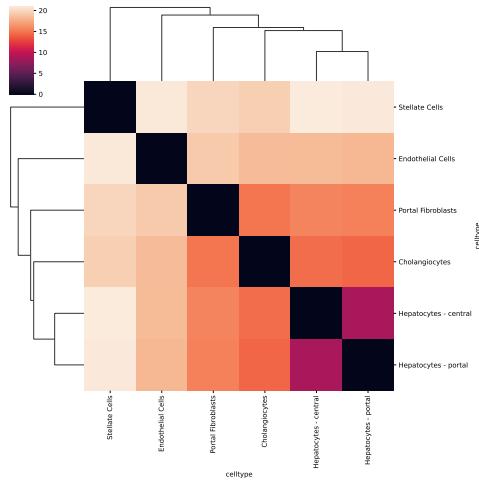


Figure 1.28: Mean pairwise

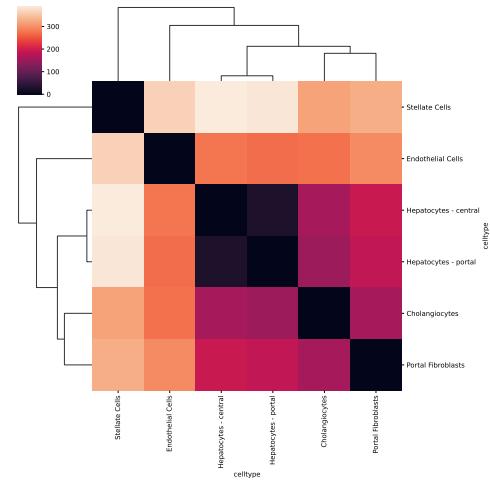


Figure 1.29: MMD

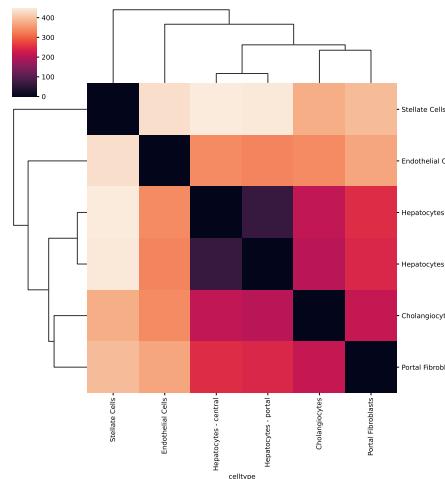


Figure 1.30: Wasserstein

Figure 1.31: Distance metrics for lowest dosage $0.01 \mu\text{g}/\text{kg}$ per cell type

1.1.2 PBMC dataset

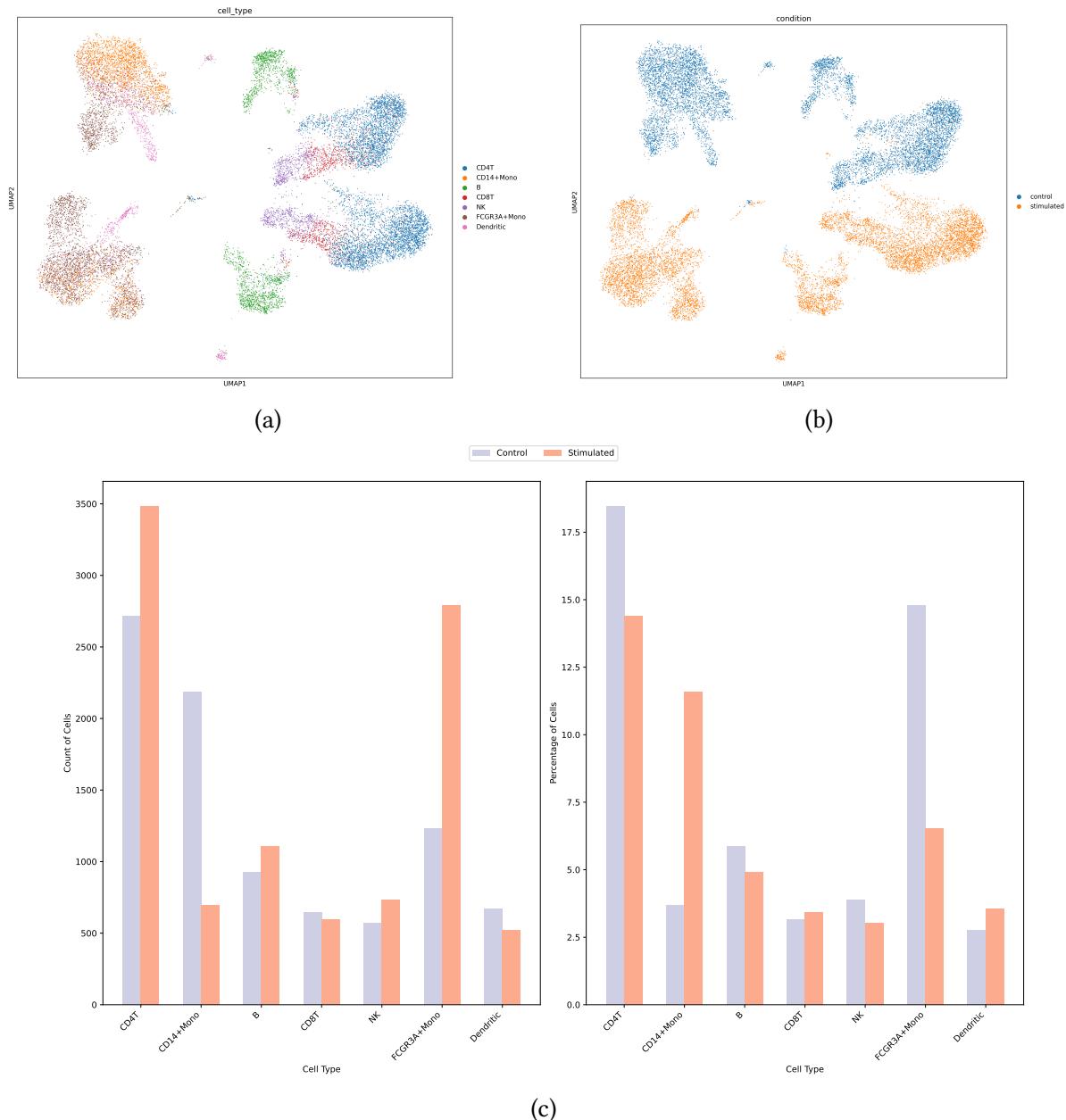


Figure 1.32: PBMC overview

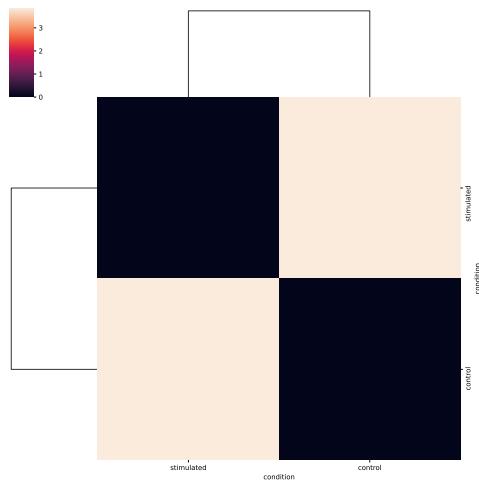


Figure 1.33: E-distance

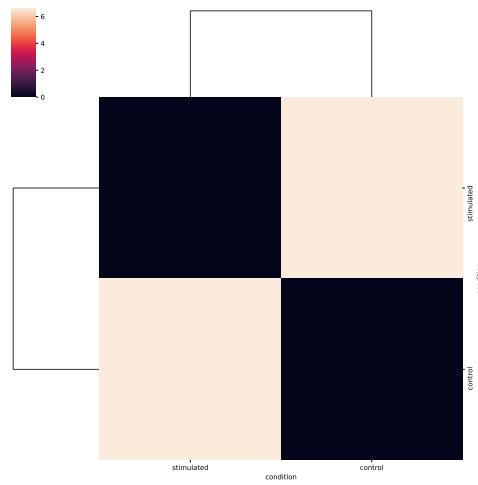


Figure 1.34: Euclidean

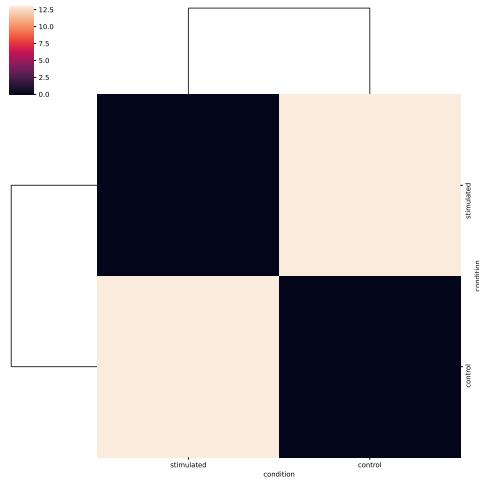


Figure 1.35: Mean pairwise

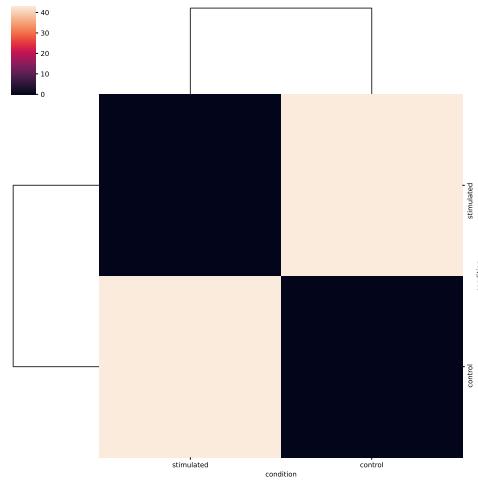


Figure 1.36: MMD

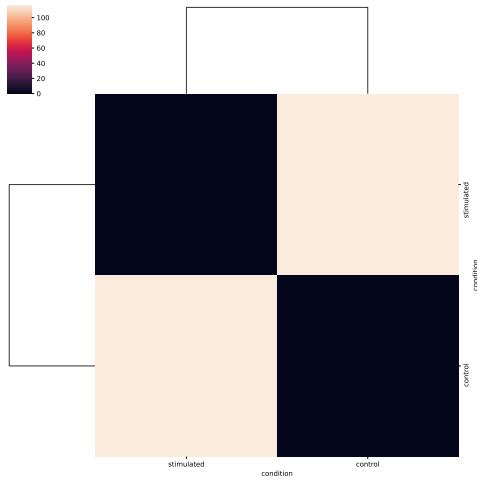


Figure 1.37: Wasserstein

Figure 1.38: Distance metrics per condition

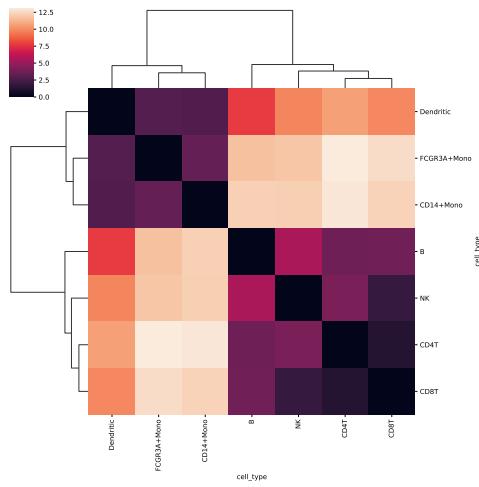


Figure 1.39: E-distance

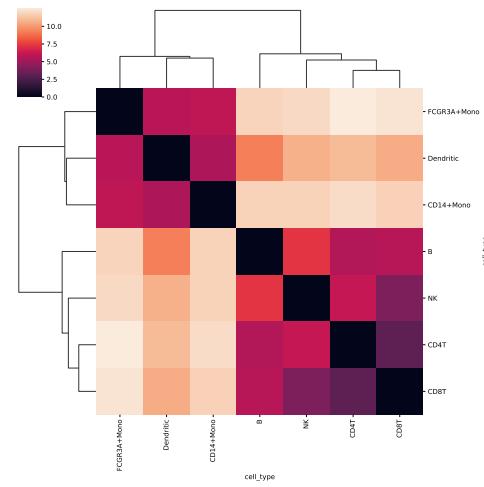


Figure 1.40: Euclidean

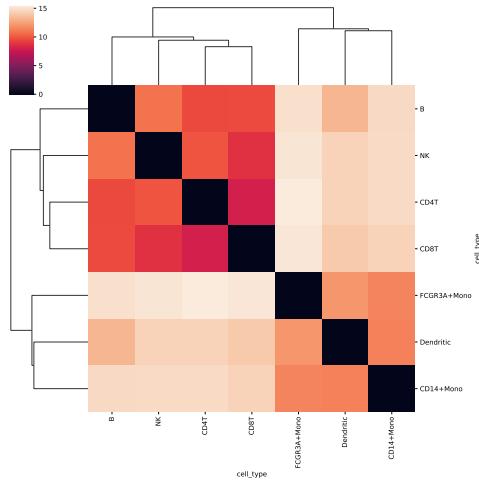


Figure 1.41: Mean pairwise

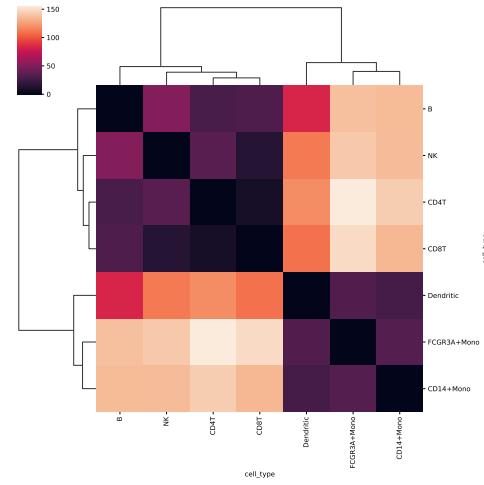


Figure 1.42: MMD

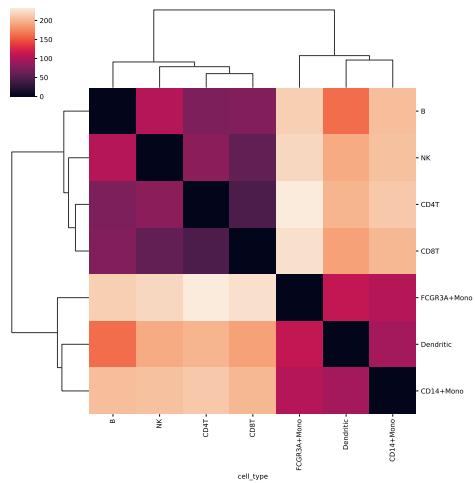
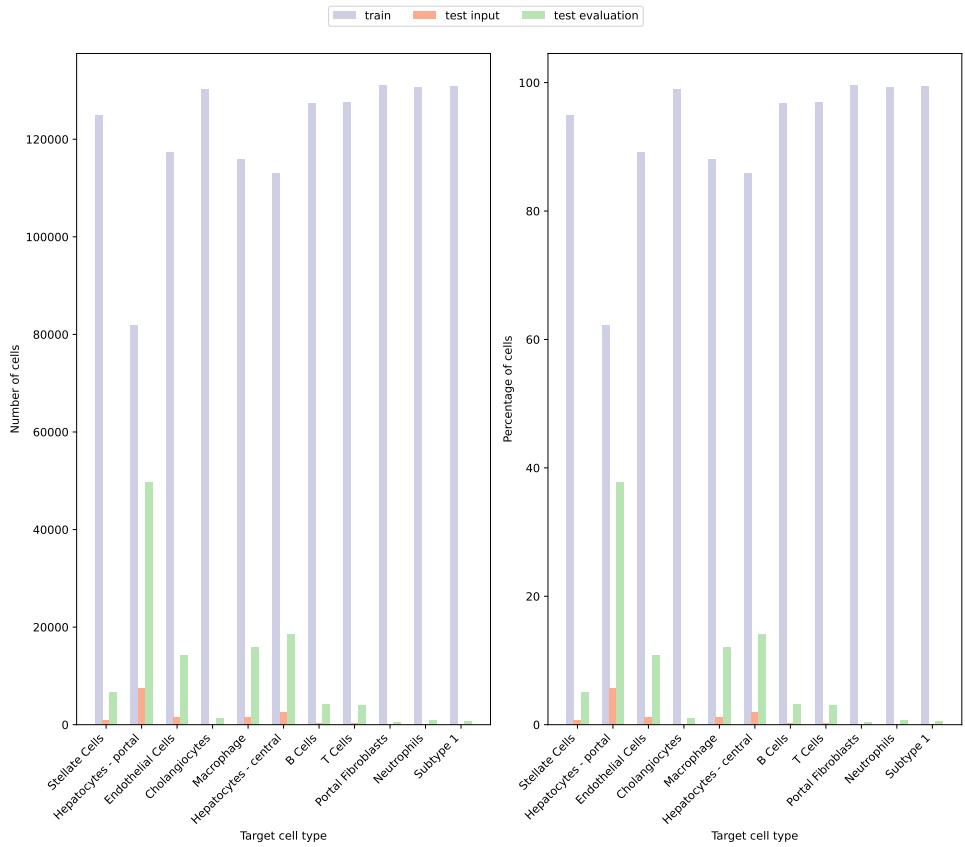
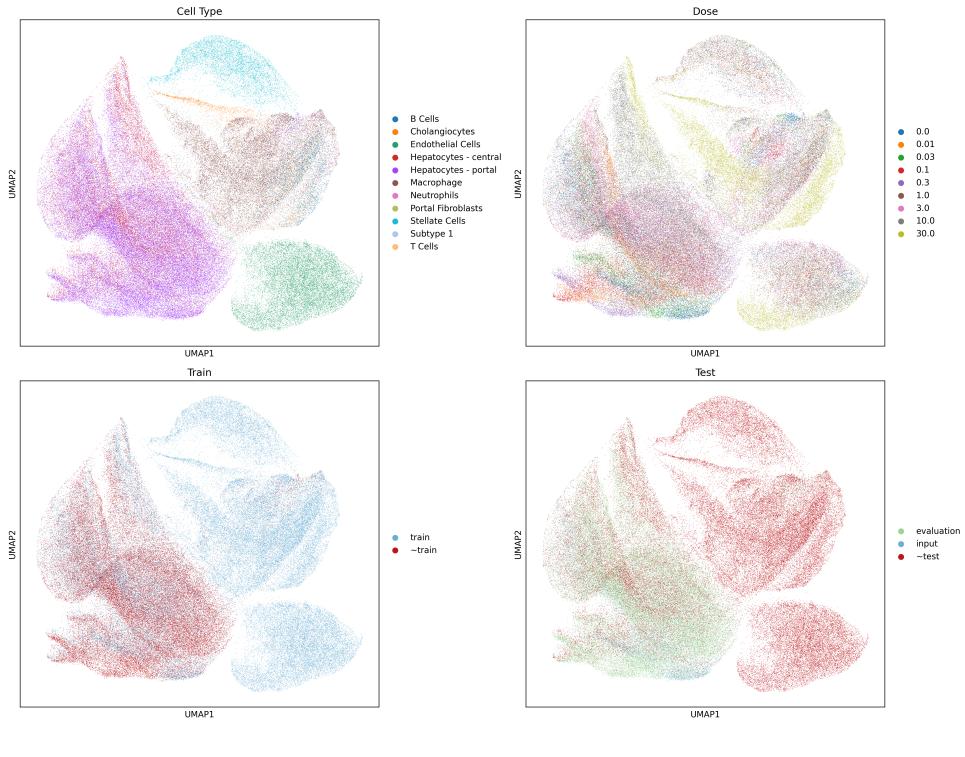


Figure 1.43: Wasserstein

Figure 1.44: Distance metrics per cell type

1.2 Nault all cell types evaluation

1.2.1 Multiple doses



1.2.2 Single dose

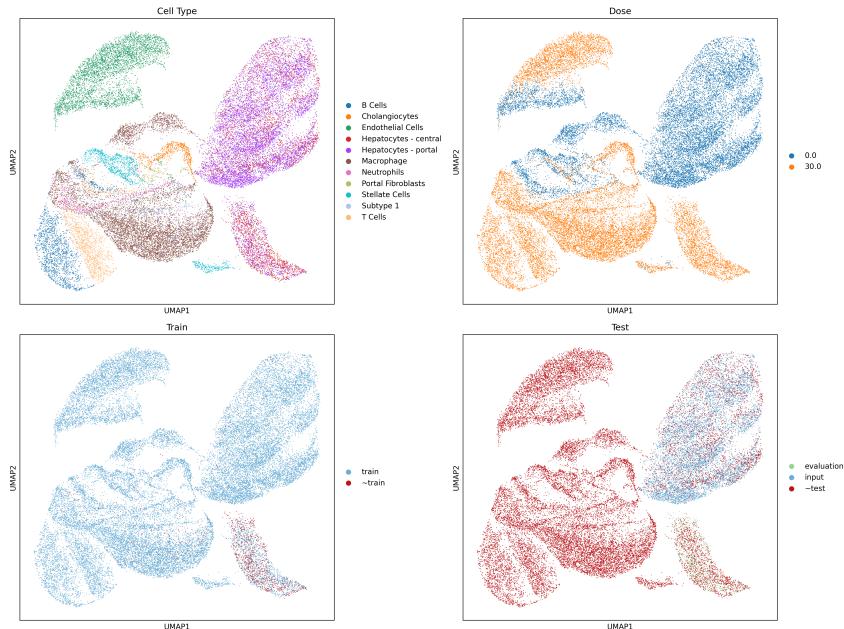


Figure 1.45: Example of $30\mu\text{g}/\text{kg}$

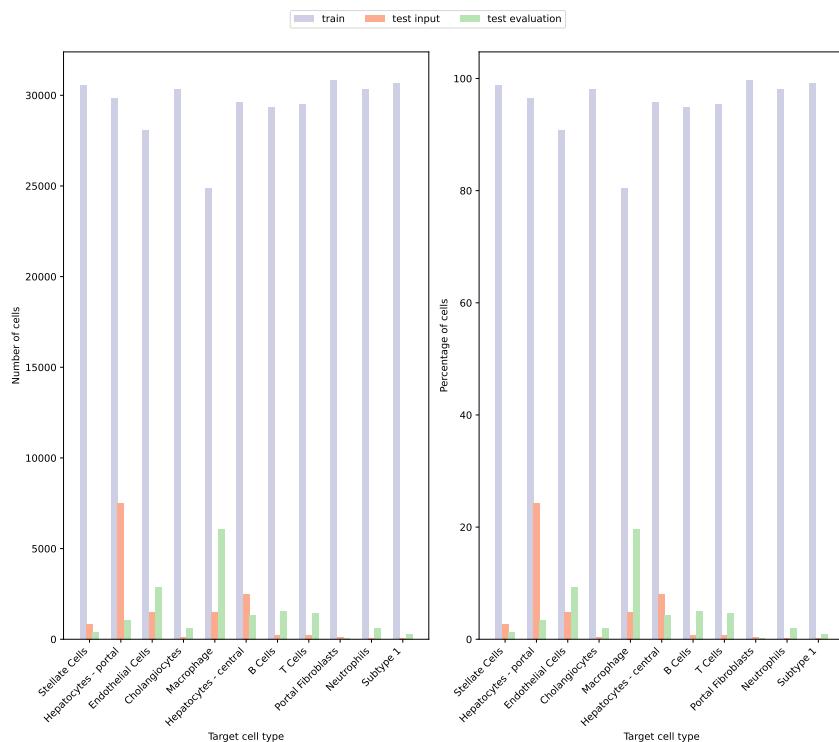


Figure 1.46: Number of cells per cell type for $30\mu\text{g}/\text{kg}$

1.2.3 Comparison

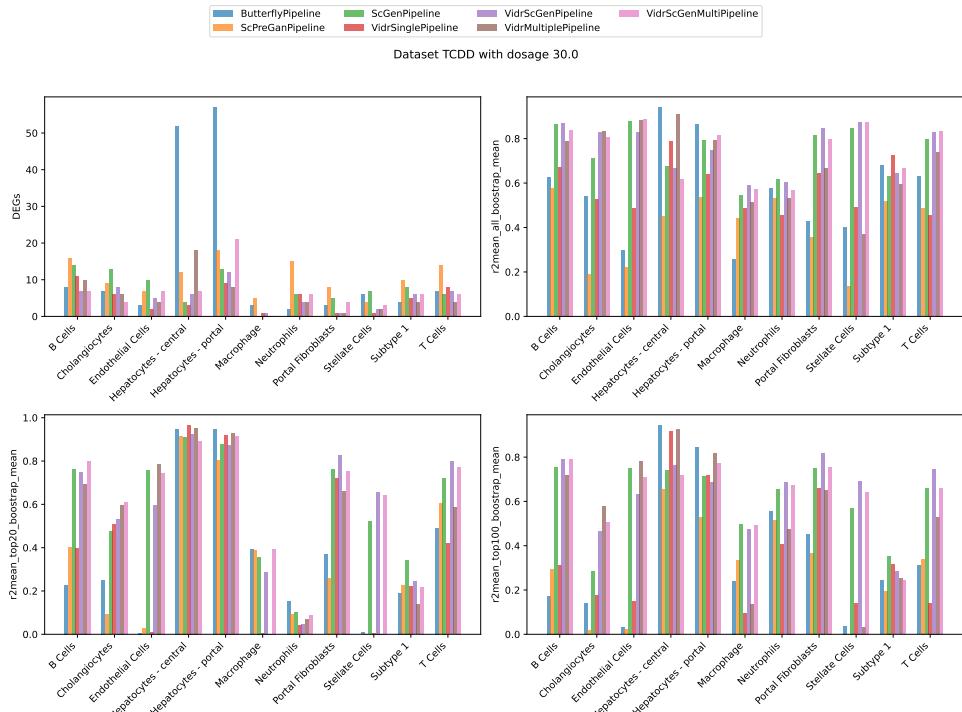


Figure 1.47: Baseline metrics for highest dosage 30 $\mu\text{g}/\text{kg}$ across cell types

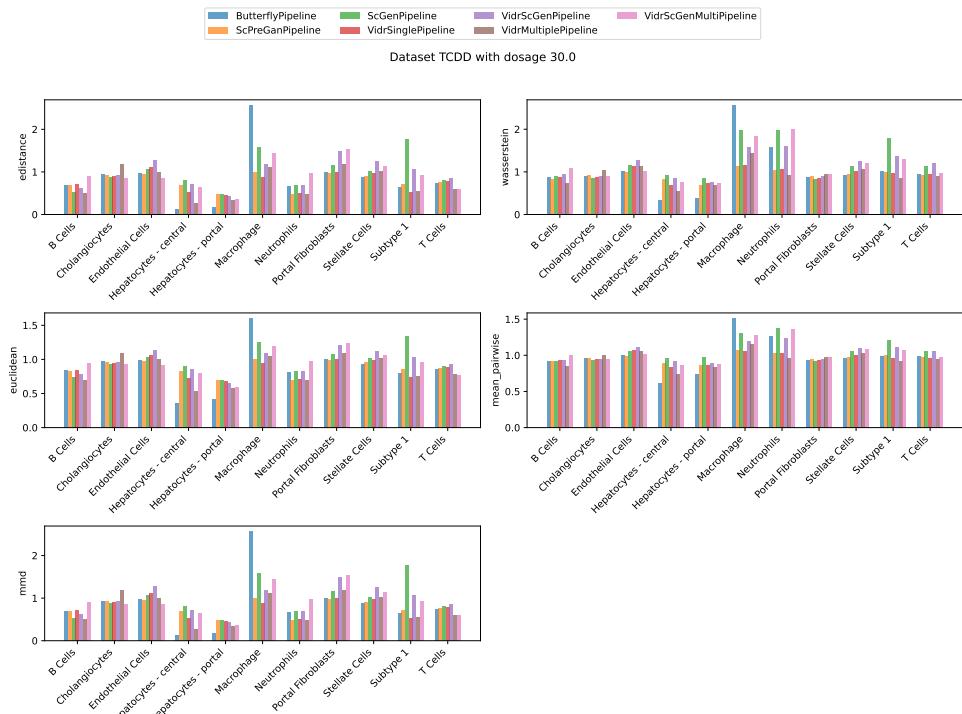


Figure 1.48: Distance metrics for highest dosage 30 $\mu\text{g}/\text{kg}$ across cell types

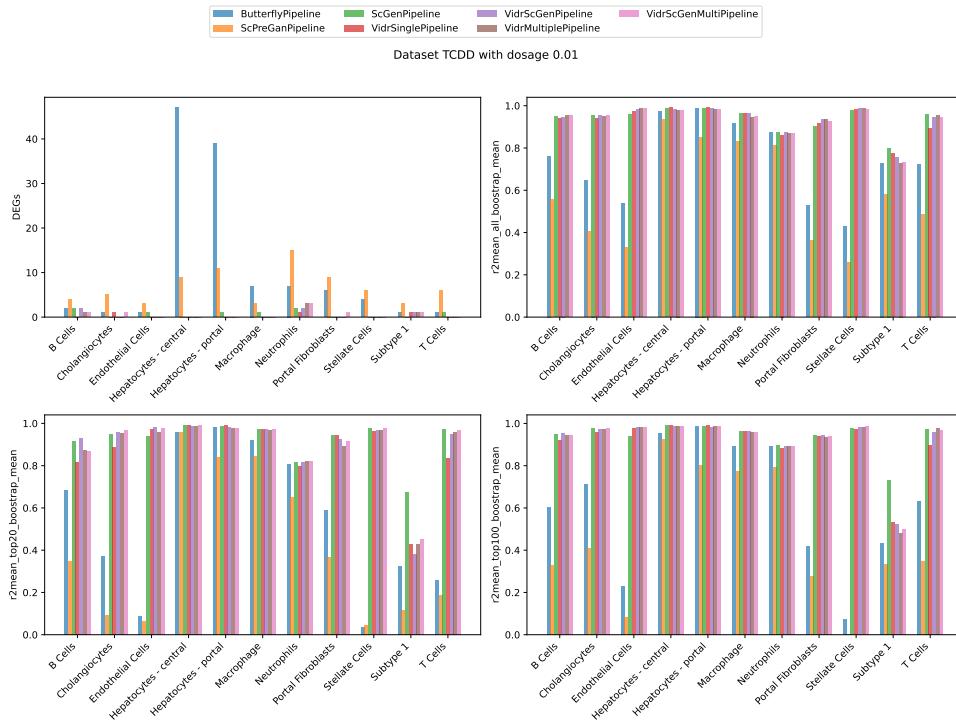


Figure 1.49: Baseline metrics for lowest dosage $0.01 \mu\text{g}/\text{kg}$ across cell types

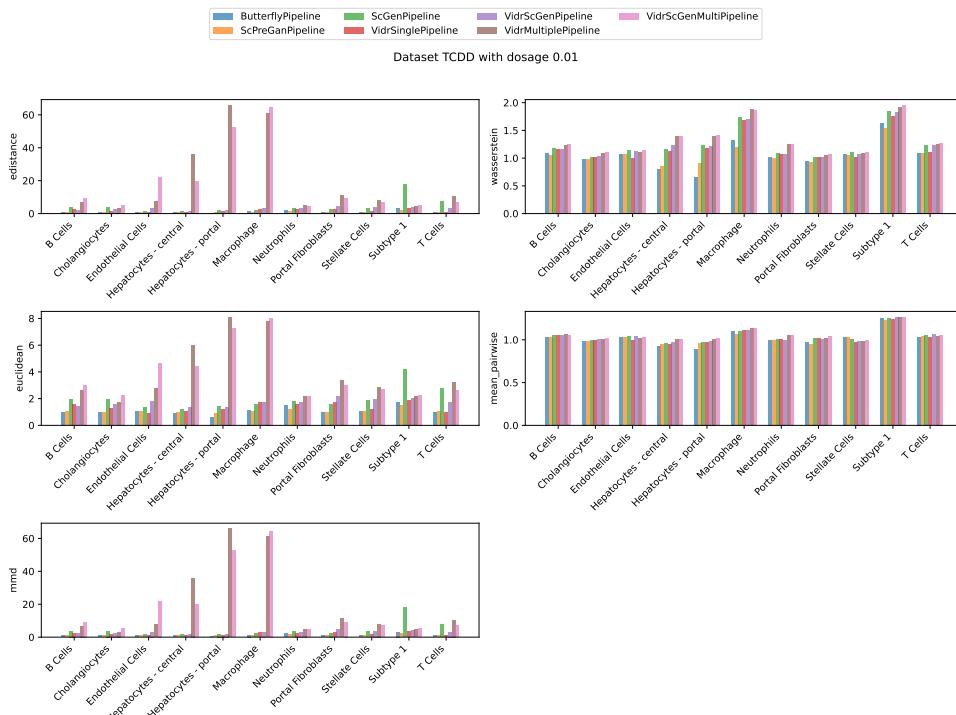


Figure 1.50: Distance metrics for lowest dosage $0.01 \mu\text{g}/\text{kg}$ across cell types

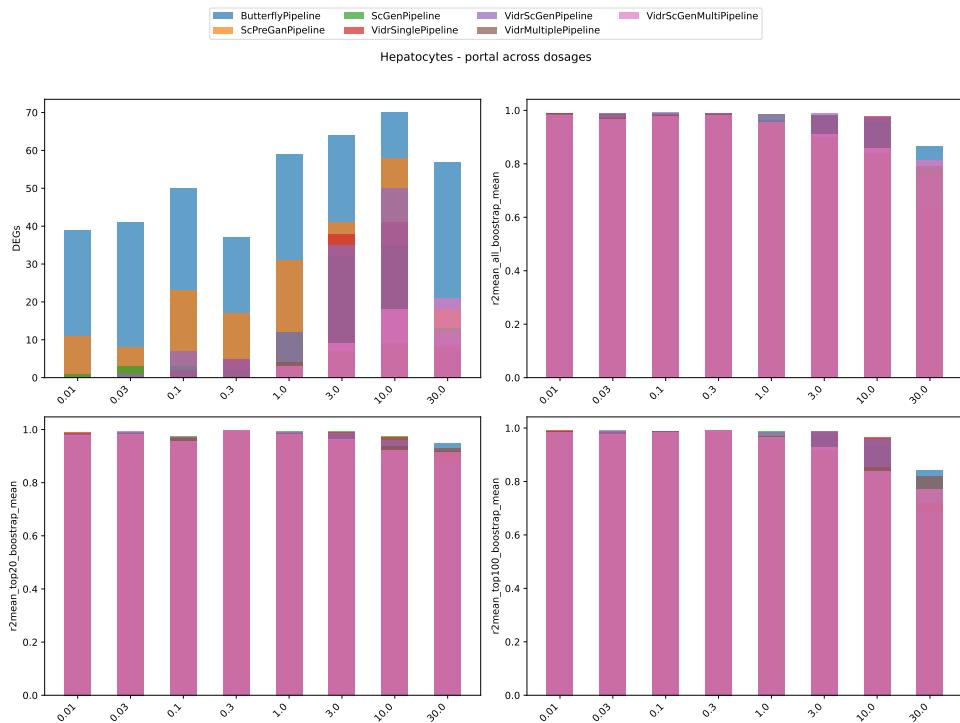


Figure 1.51: Baseline metrics for Hepatocytes - portal across dosages

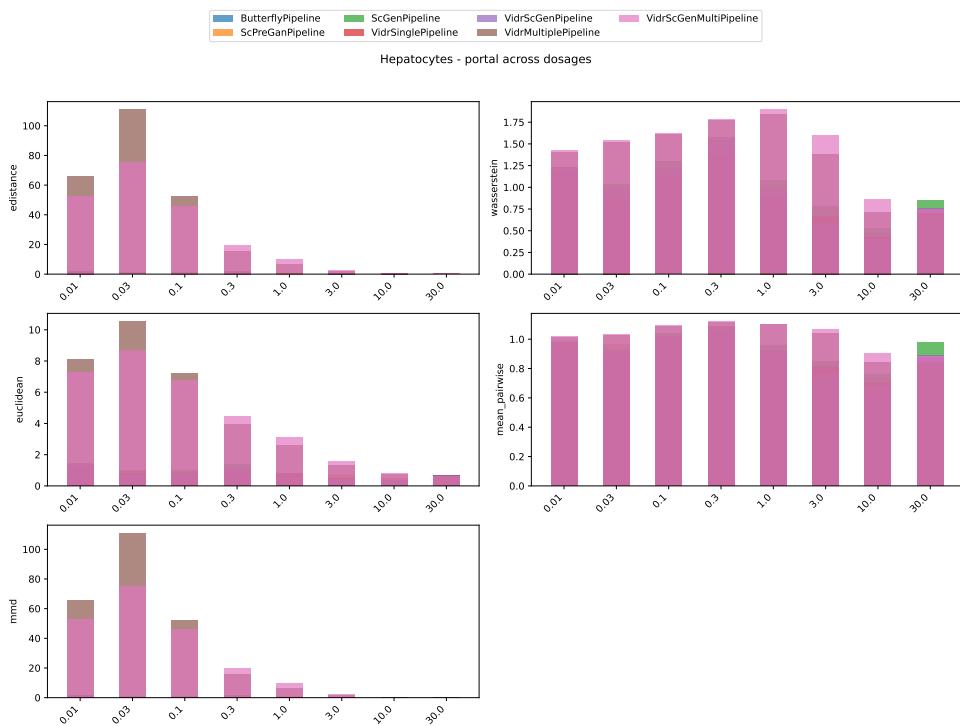


Figure 1.52: Distance metrics for Hepatocytes - portal across dosages

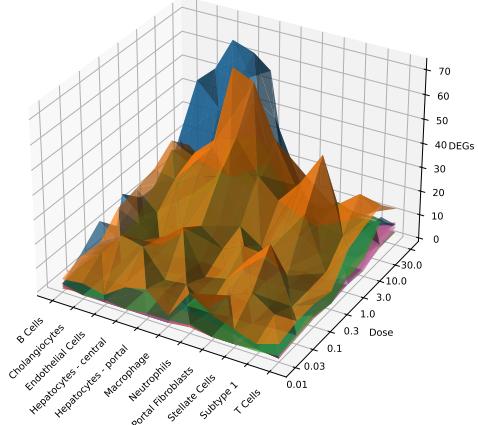


Figure 1.53

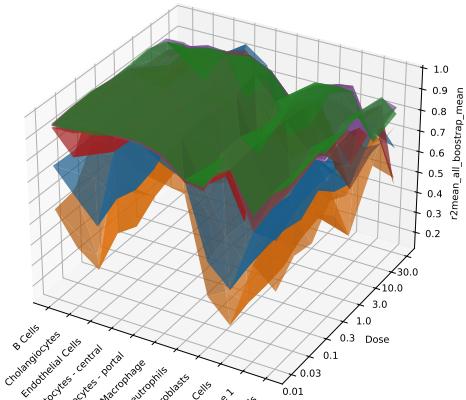


Figure 1.54

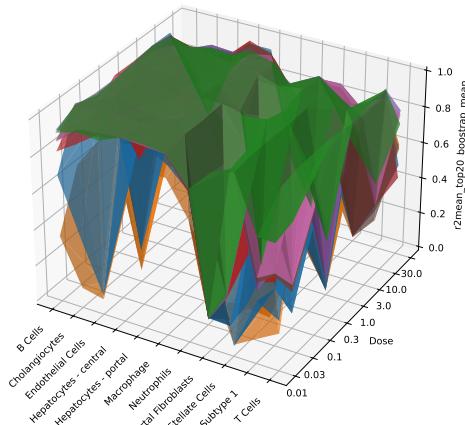


Figure 1.55

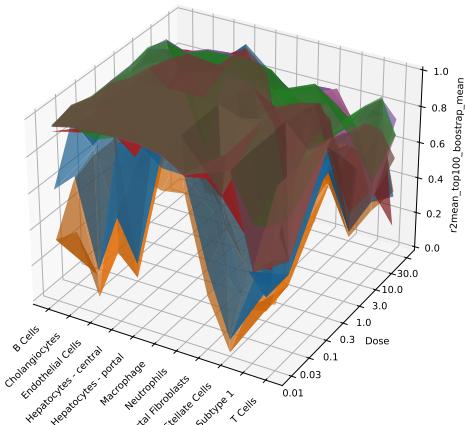


Figure 1.56

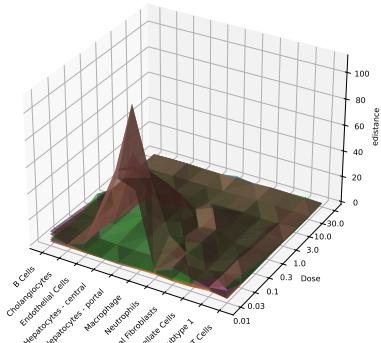


Figure 1.57: E-distance

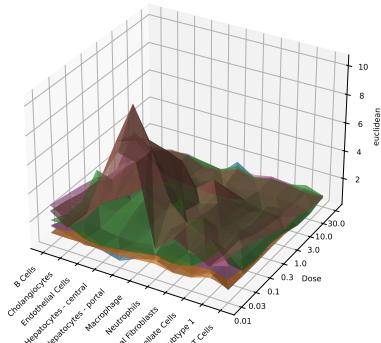


Figure 1.58: Euclidean

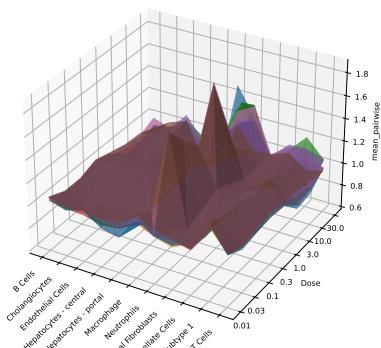


Figure 1.59: Mean pairwise

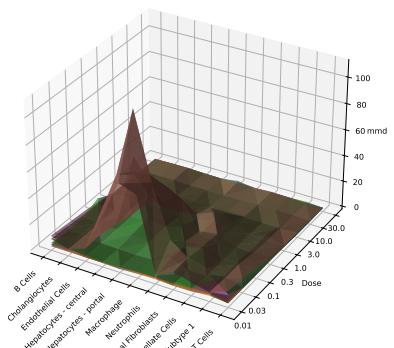


Figure 1.60: MMD

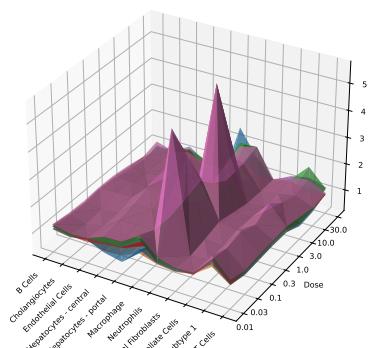
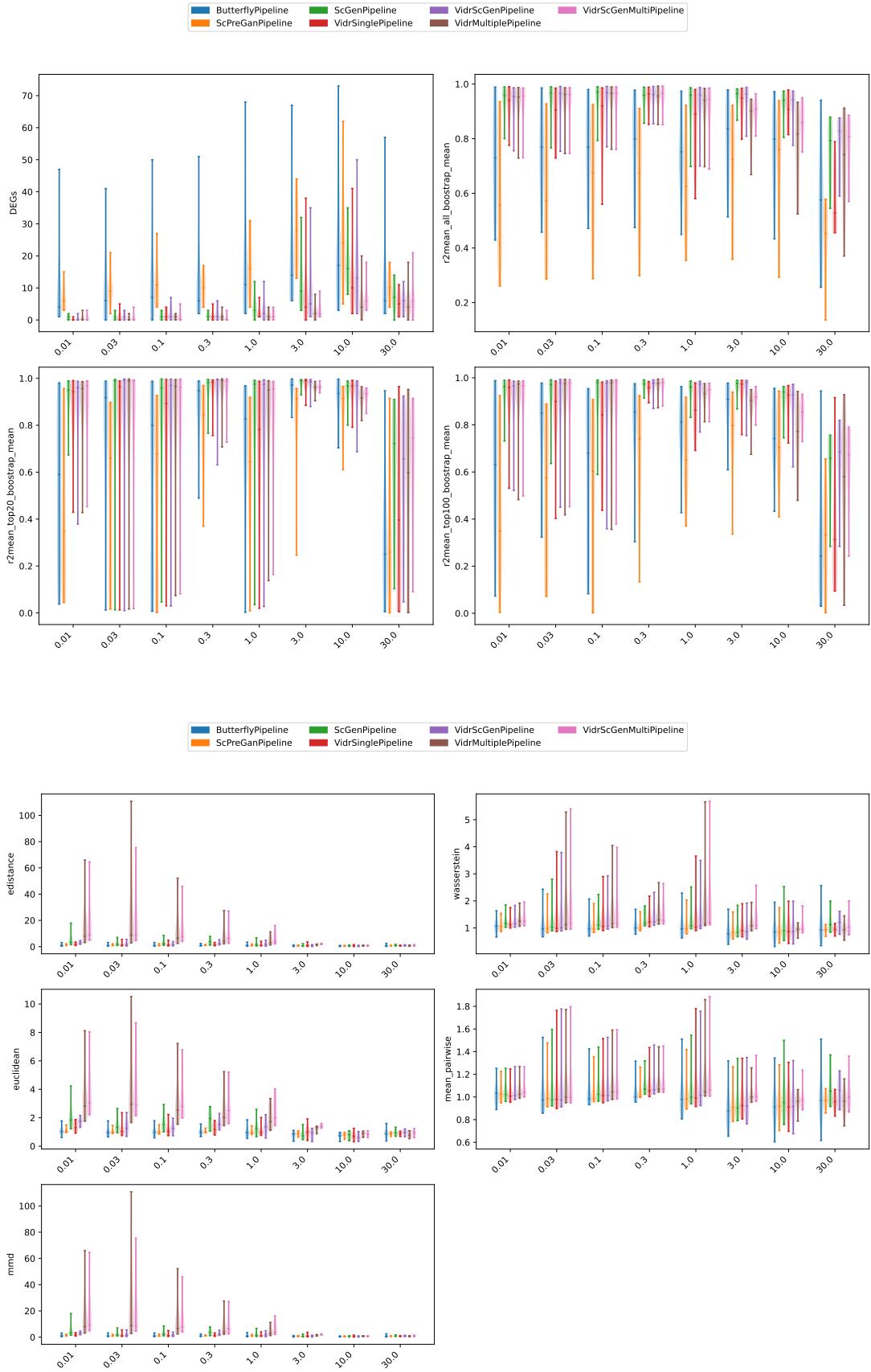
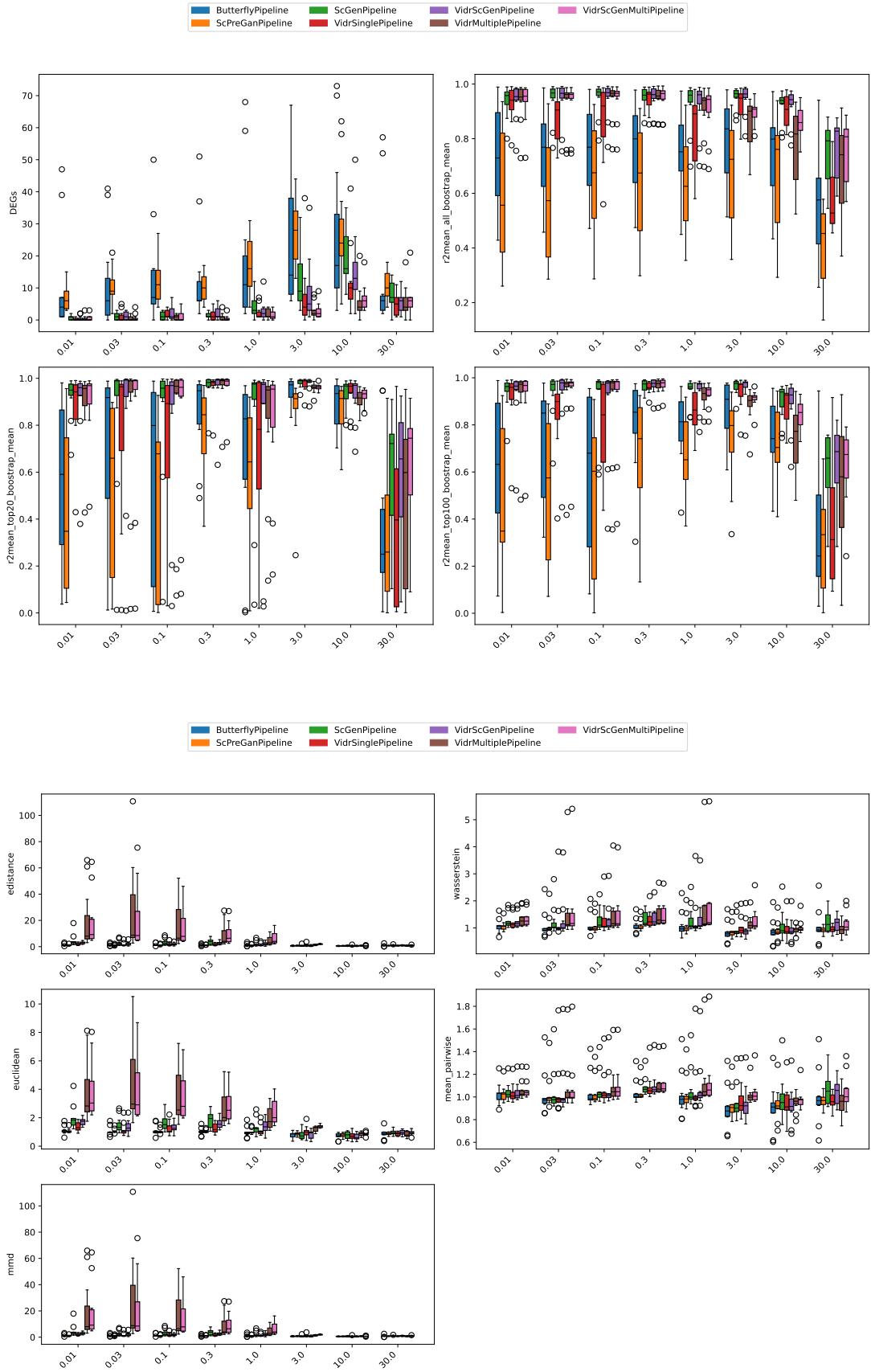
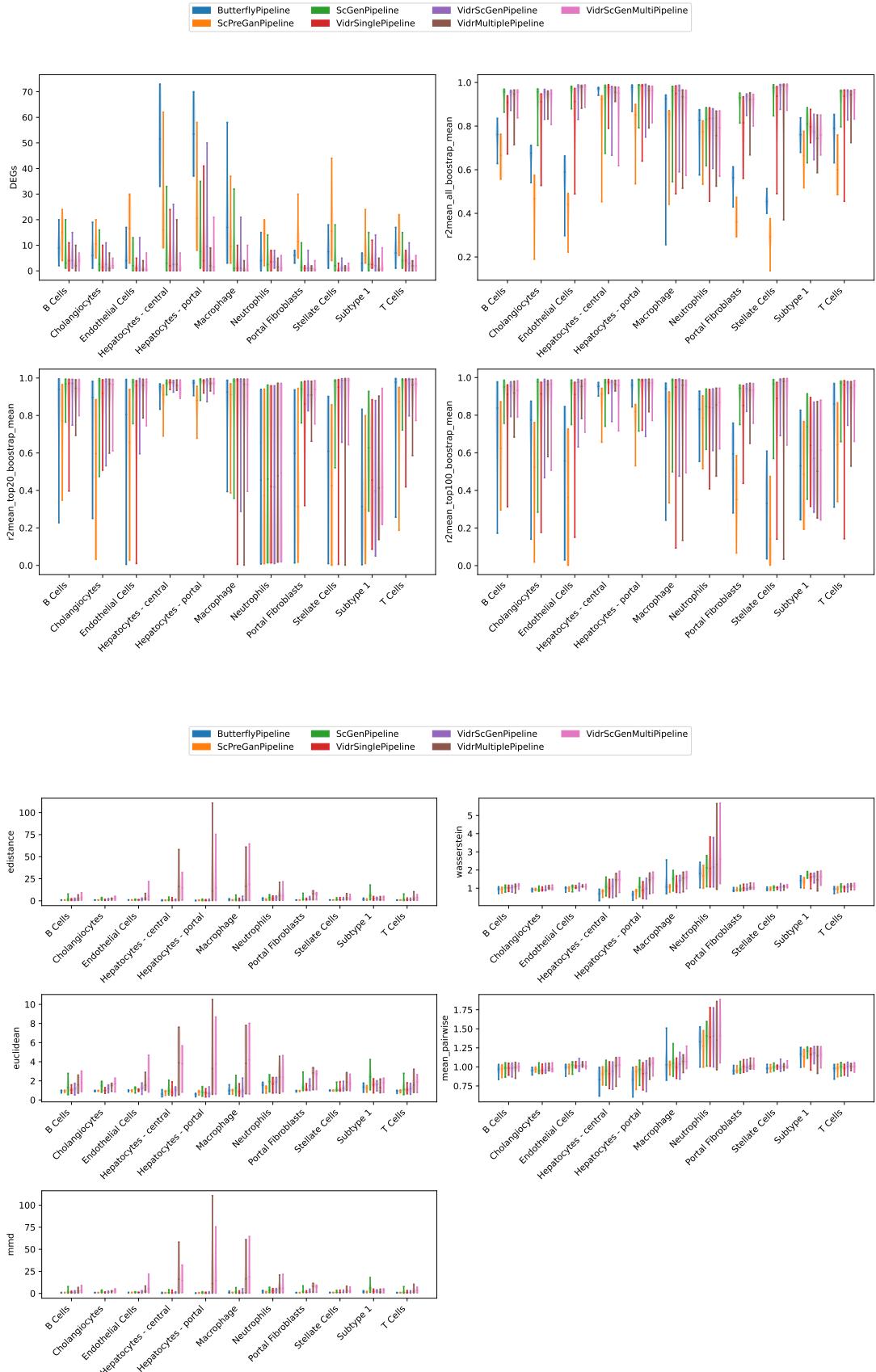


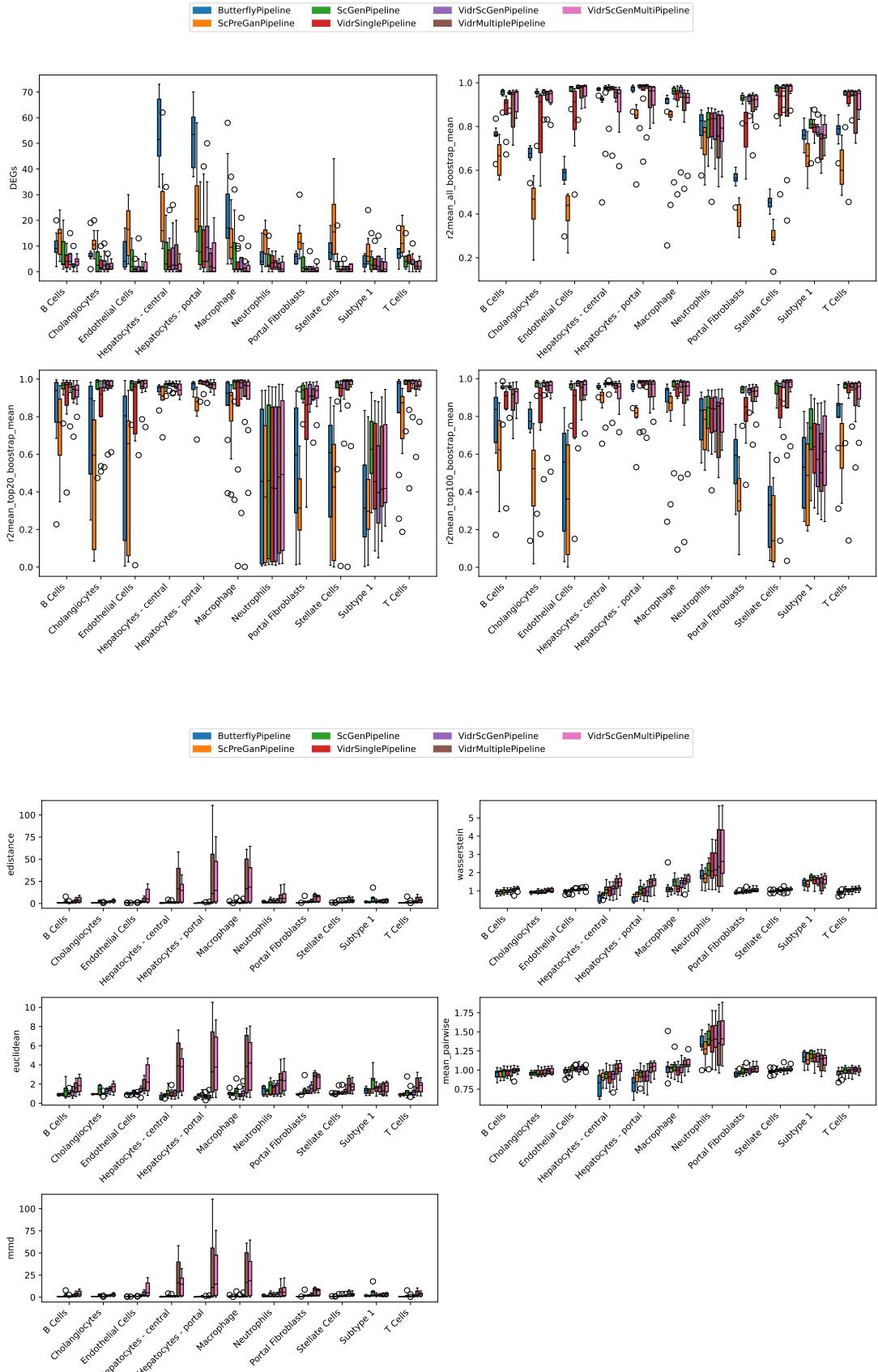
Figure 1.61: Wasserstein

Figure 1.62: Distance metrics per cell type









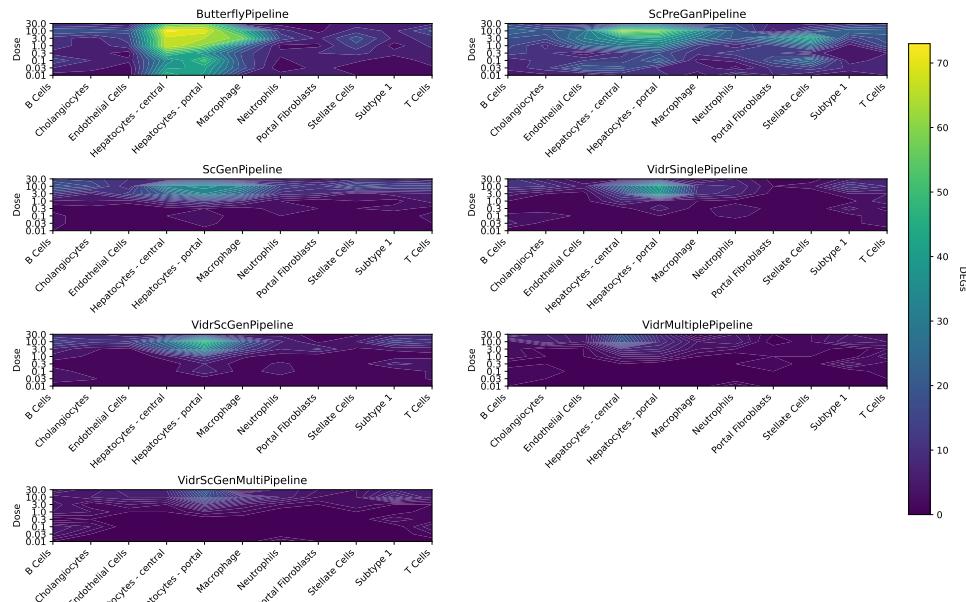


Figure 1.63: DEGs

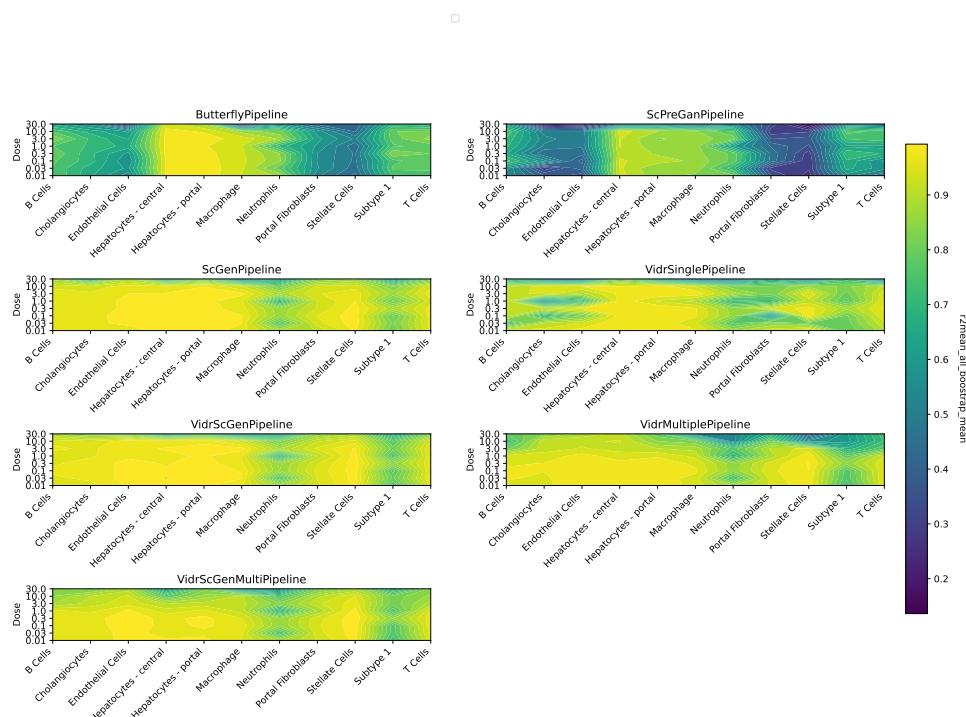


Figure 1.64: r^2 HVGs

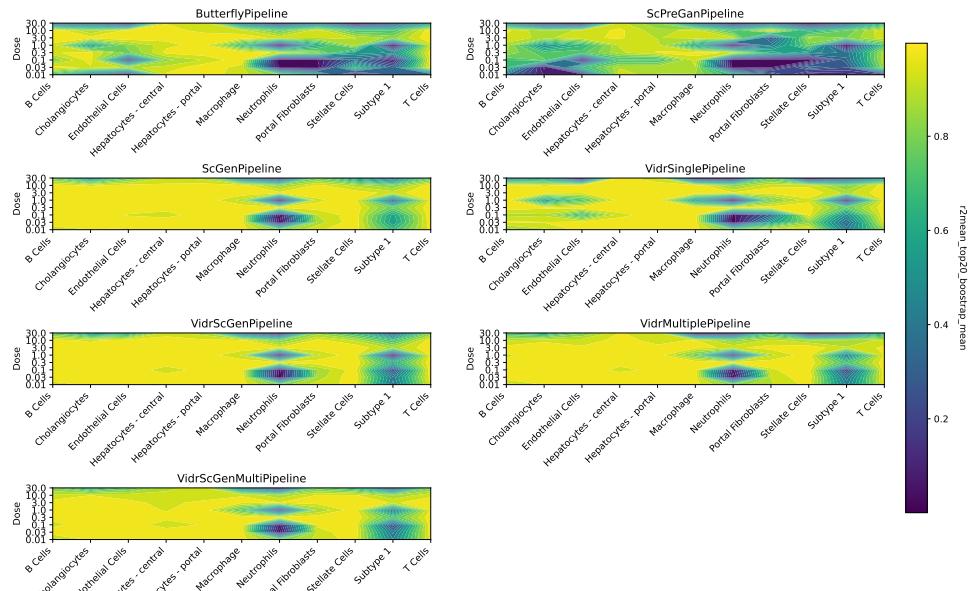


Figure 1.65: r2 top 20

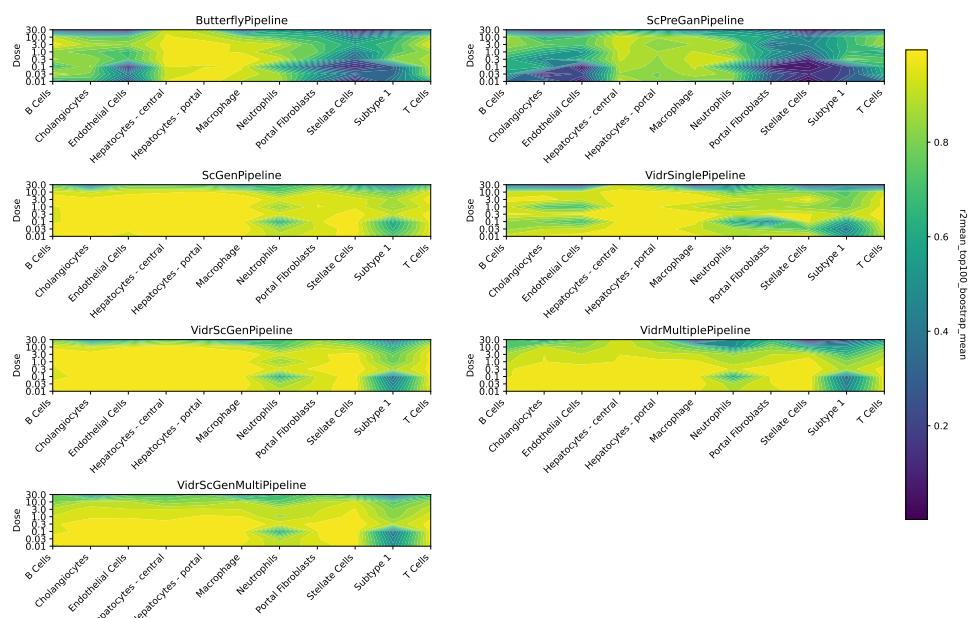


Figure 1.66: r2 top 100

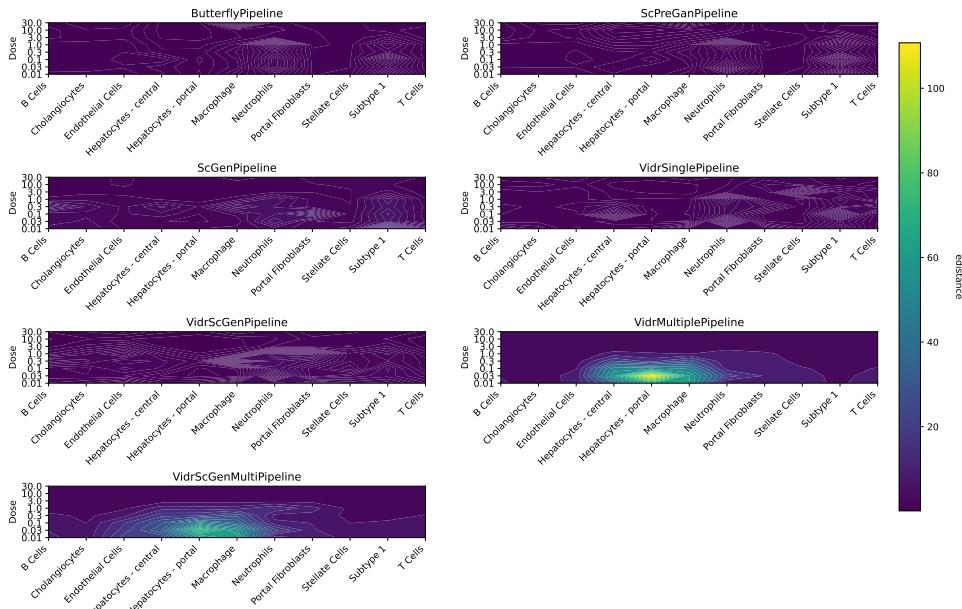


Figure 1.67: E-distance

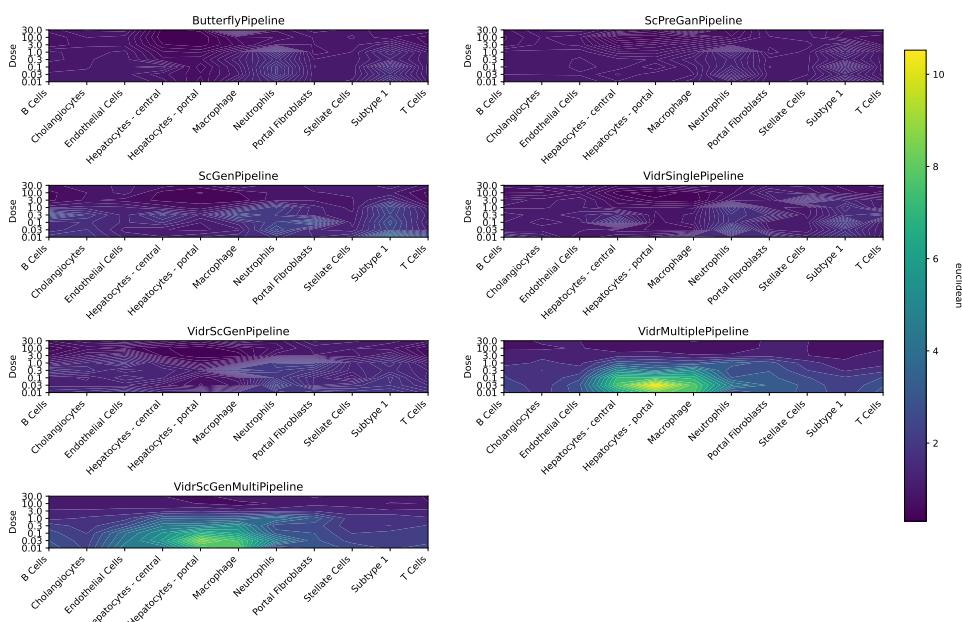


Figure 1.68: Euclidean

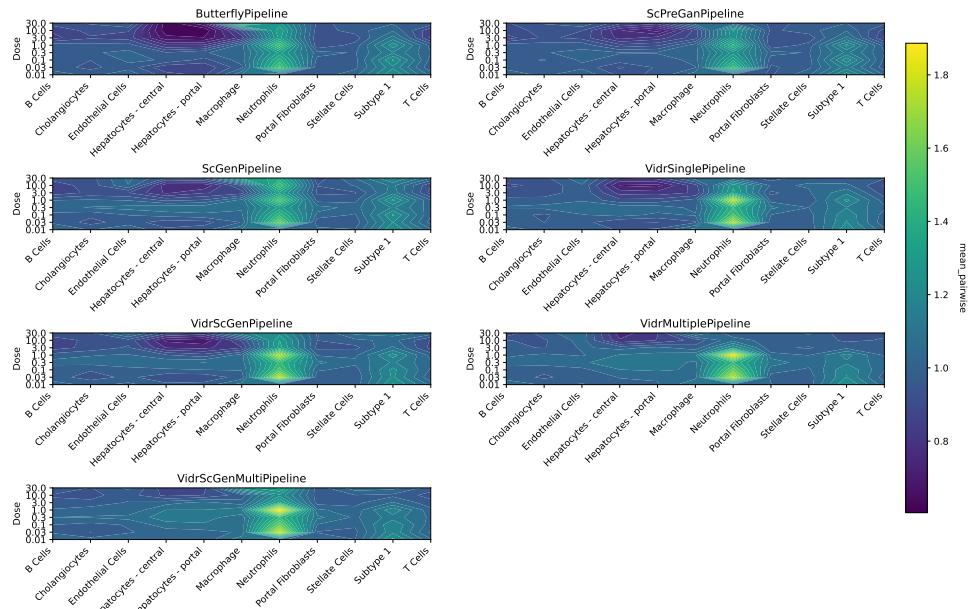


Figure 1.69: Mean pairwise

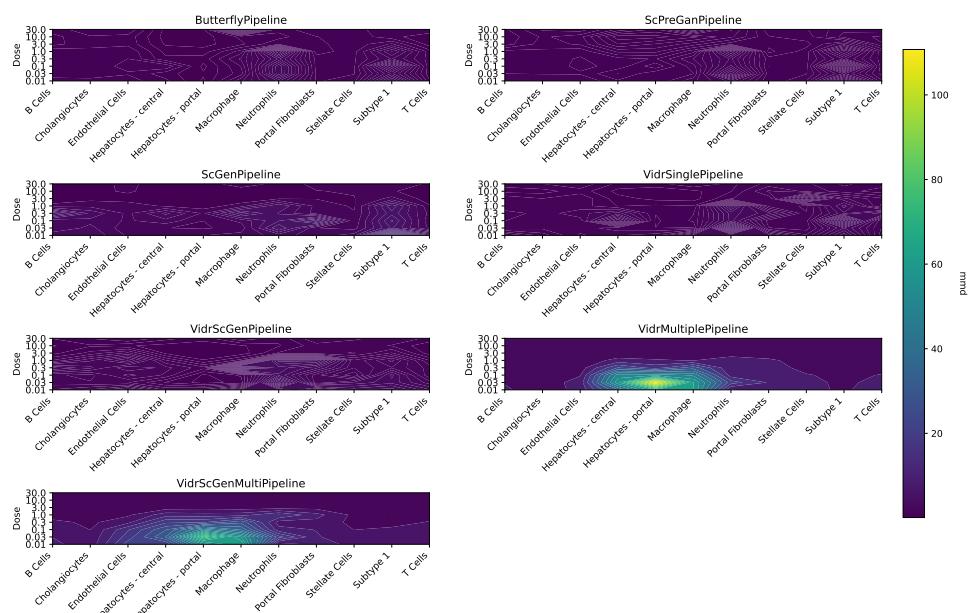


Figure 1.70: MMD

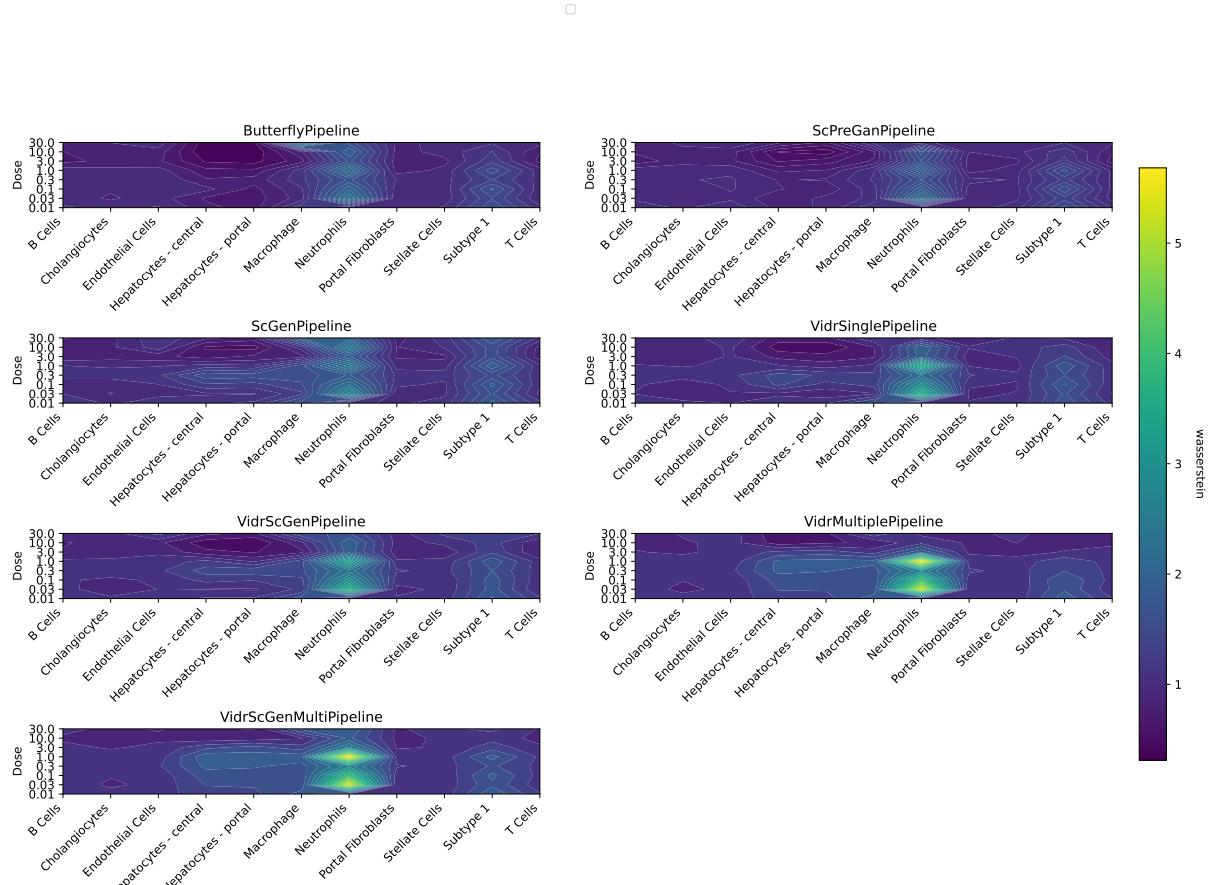
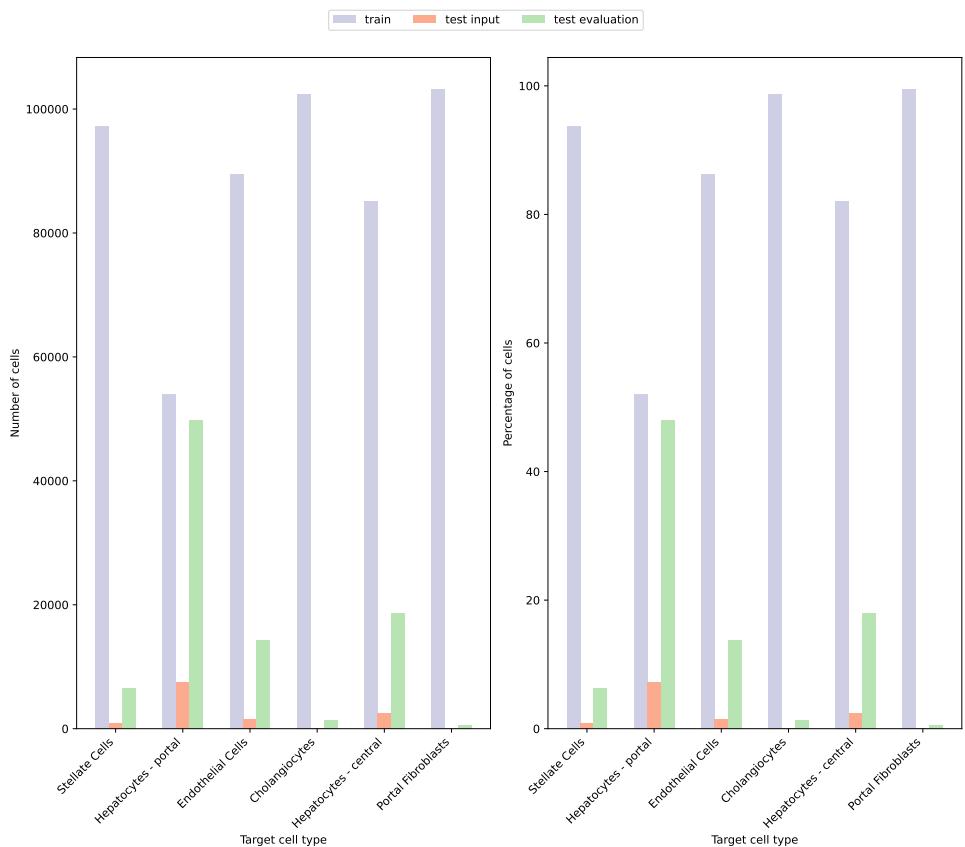
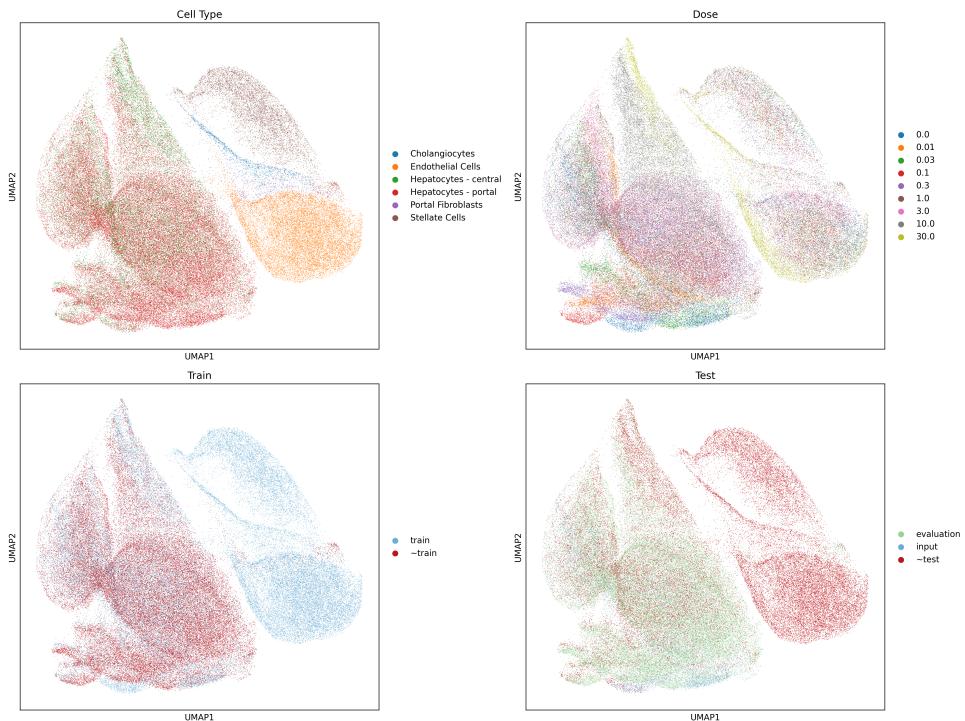


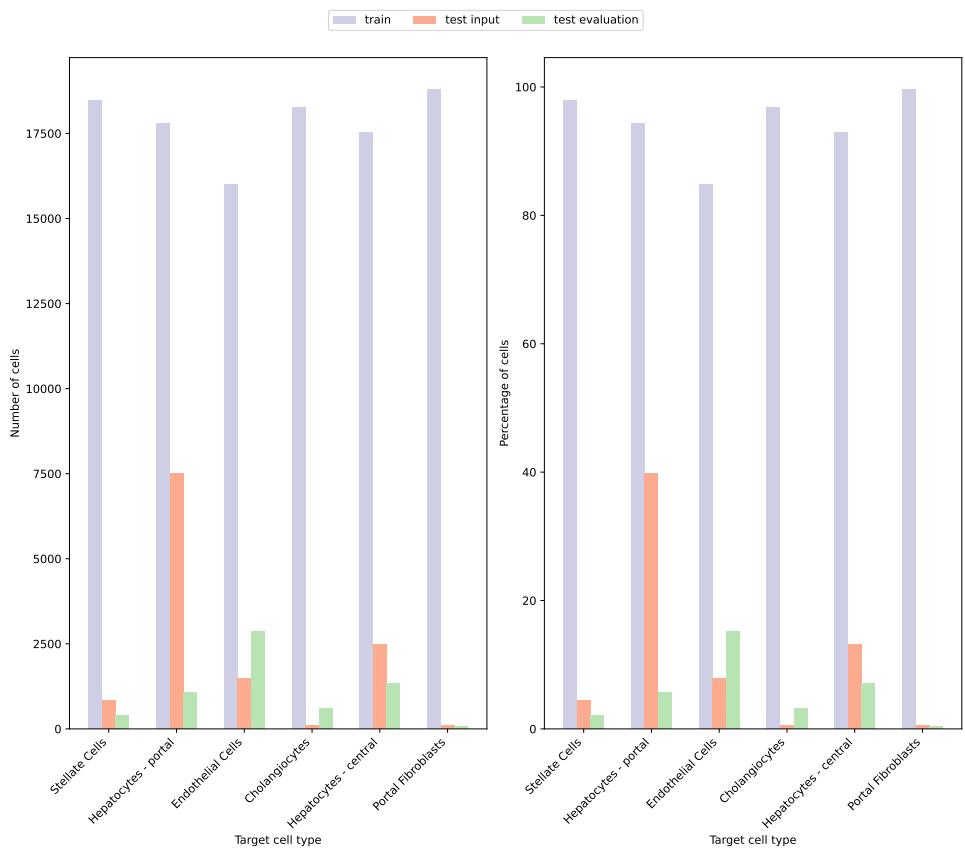
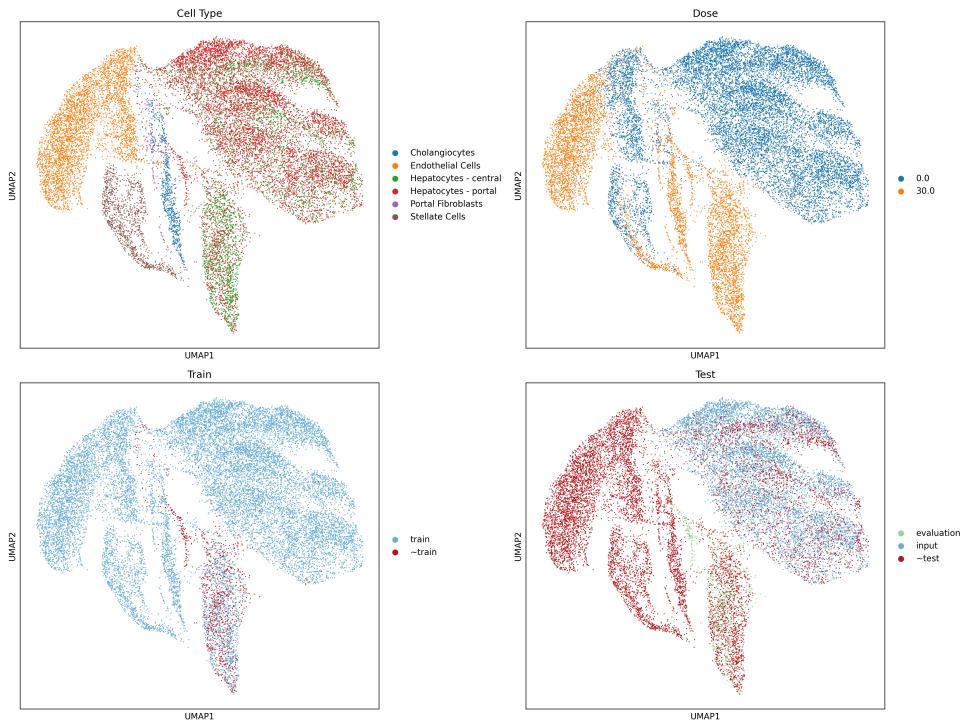
Figure 1.71: Wasserstein

1.3 Nault liver cell types evaluation

1.3.1 Multiple doses



1.3.2 Single dose 30 $\mu\text{g}/\text{kg}$



1.3.3 Comparison

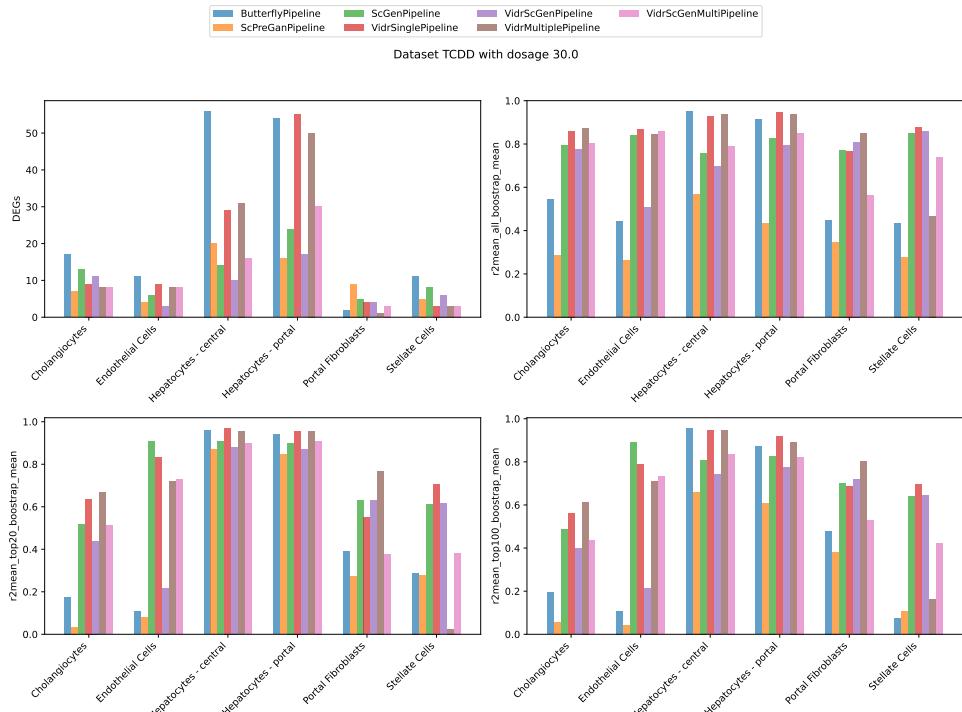


Figure 1.72: Baseline metrics for highest dosage $30 \mu\text{g}/\text{kg}$ across cell types

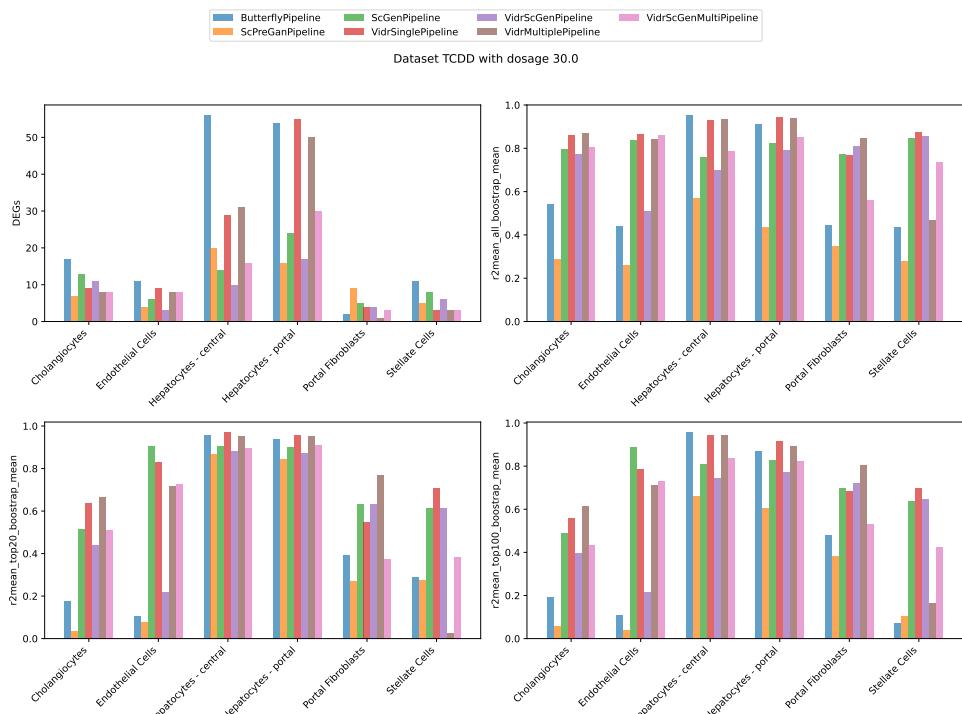


Figure 1.73: Distance metrics for highest dosage $30 \mu\text{g}/\text{kg}$ across cell types

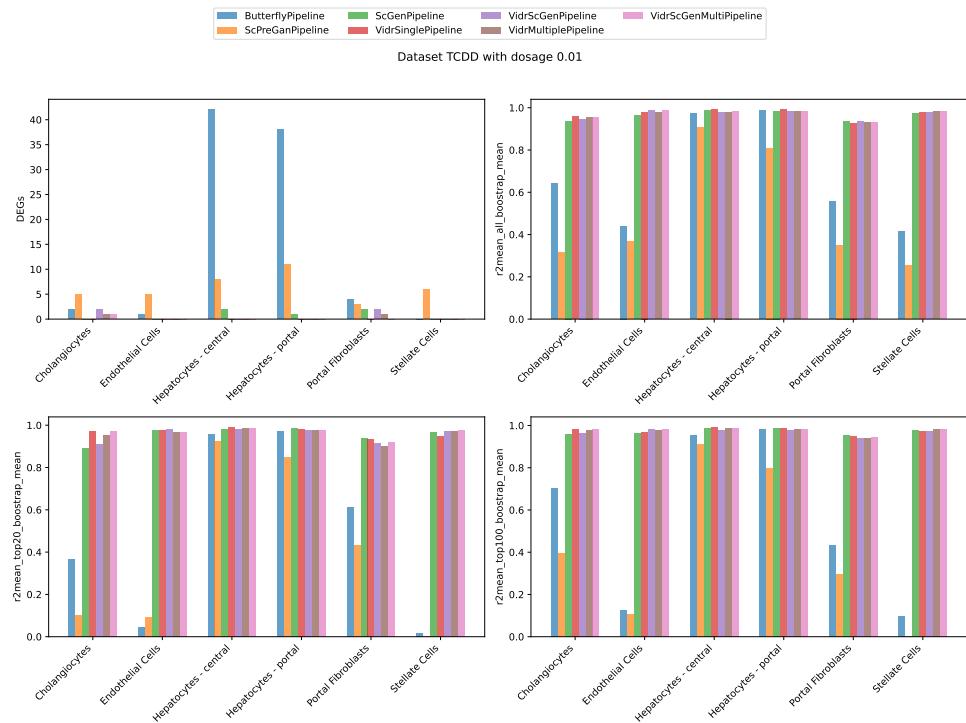


Figure 1.74: Baseline metrics for highest dosage $0.1\mu\text{g}/\text{kg}$ across cell types

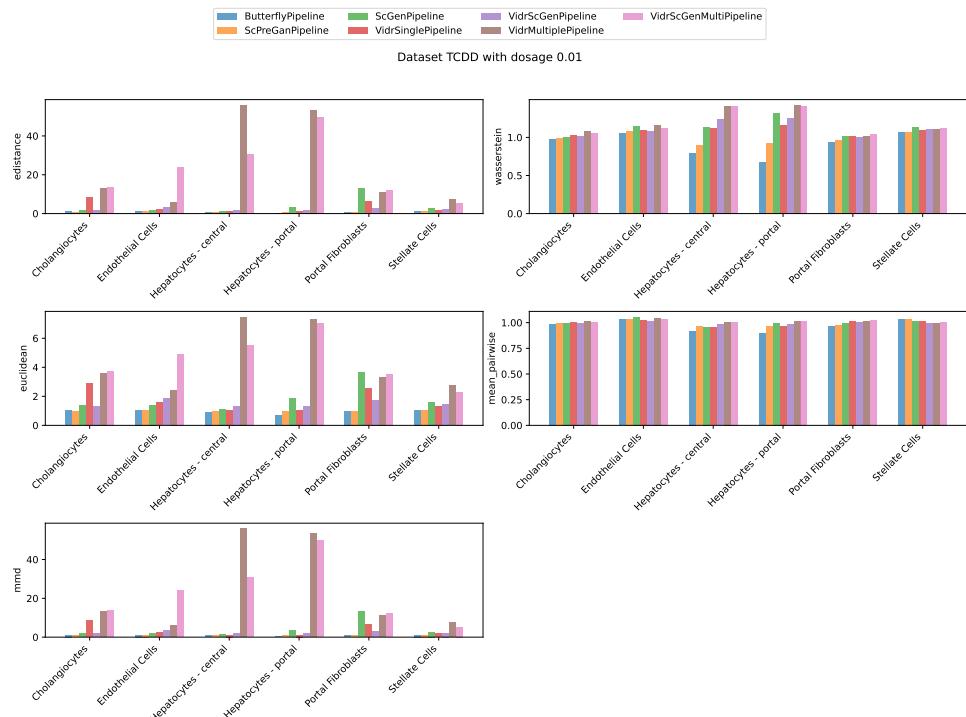


Figure 1.75: Distance metrics for highest dosage $0.1\mu\text{g}/\text{kg}$ across cell types

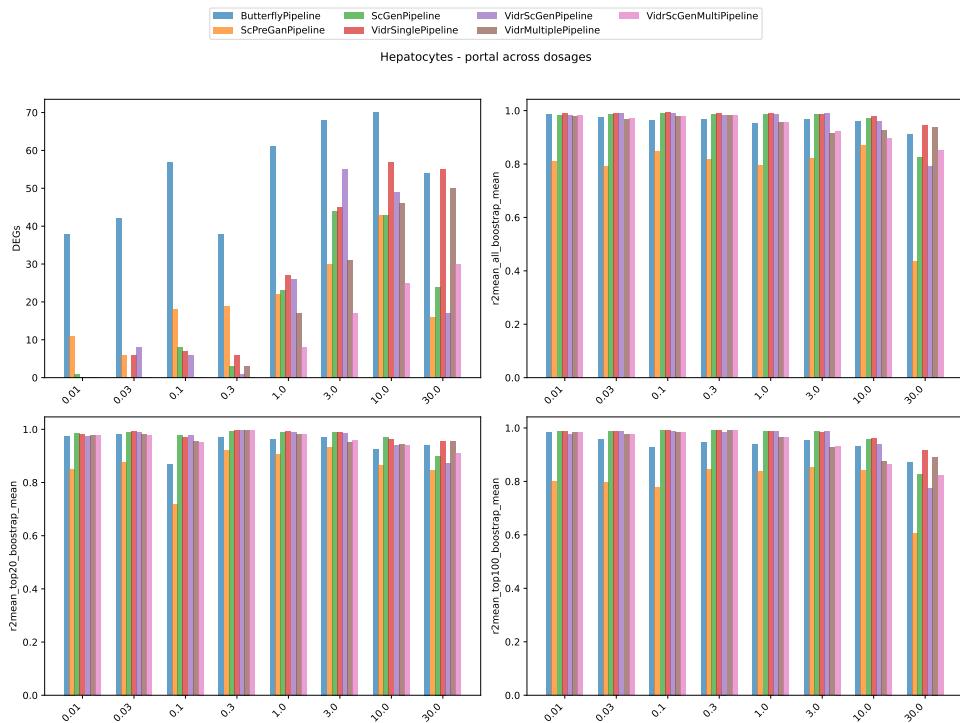


Figure 1.76: Baseline metrics for Hepatocytes - portal across dosages

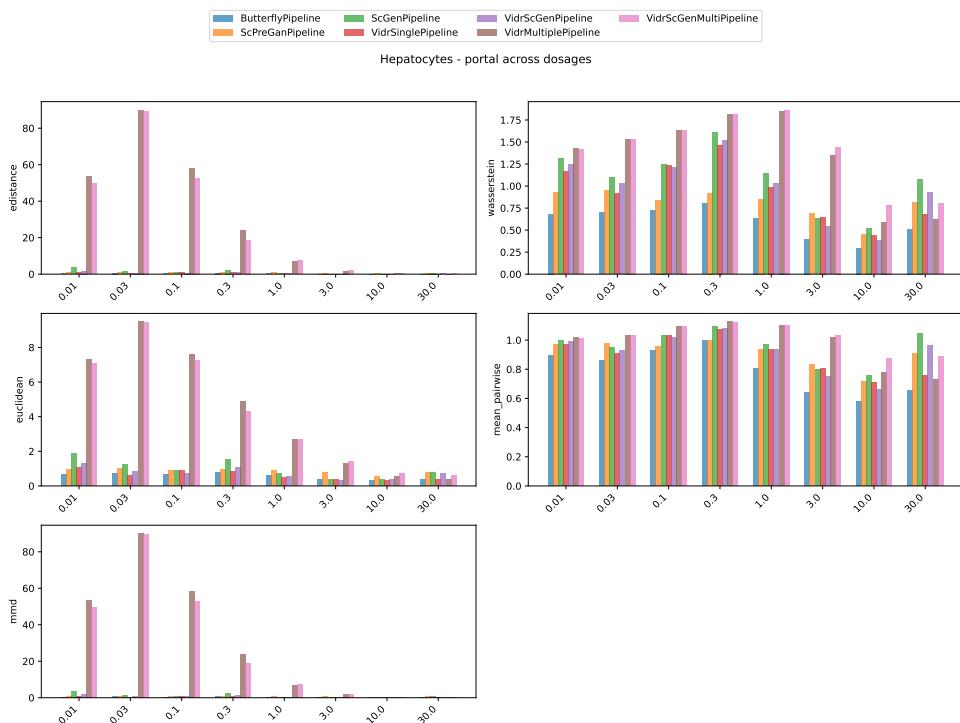


Figure 1.77: Distance metrics for Hepatocytes - portal across dosages

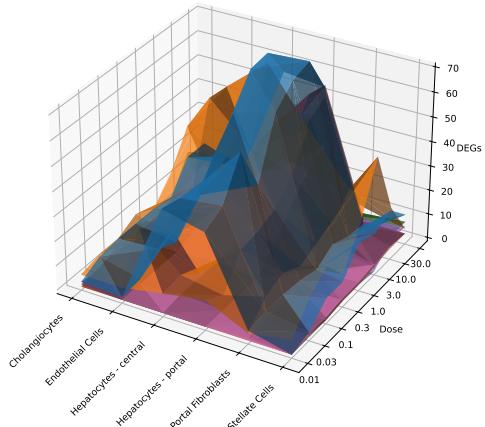
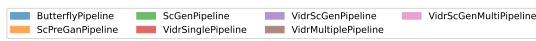


Figure 1.78

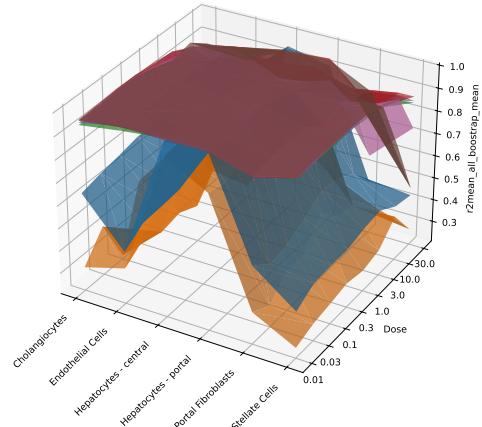


Figure 1.79

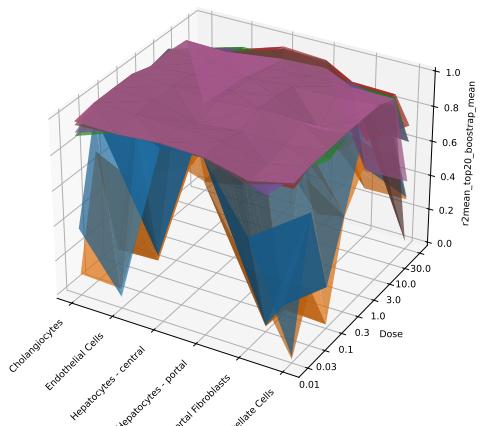


Figure 1.80

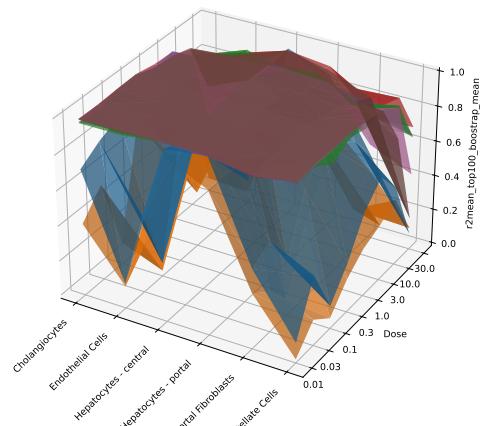


Figure 1.81

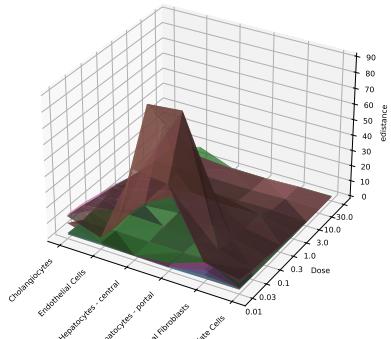


Figure 1.82: E-distance

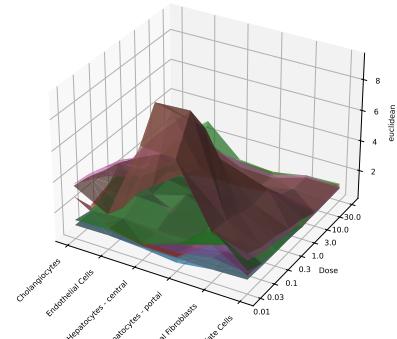


Figure 1.83: Euclidean

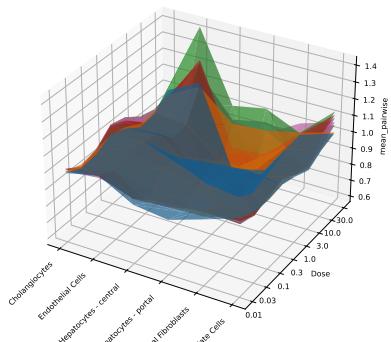


Figure 1.84: Mean pairwise

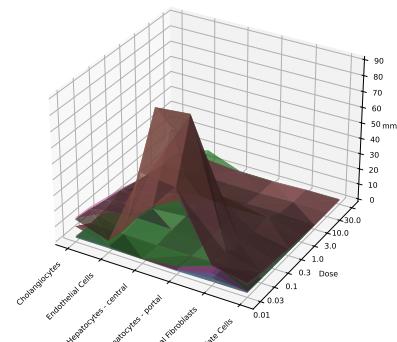


Figure 1.85: MMD

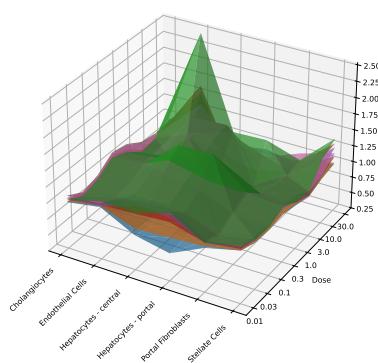
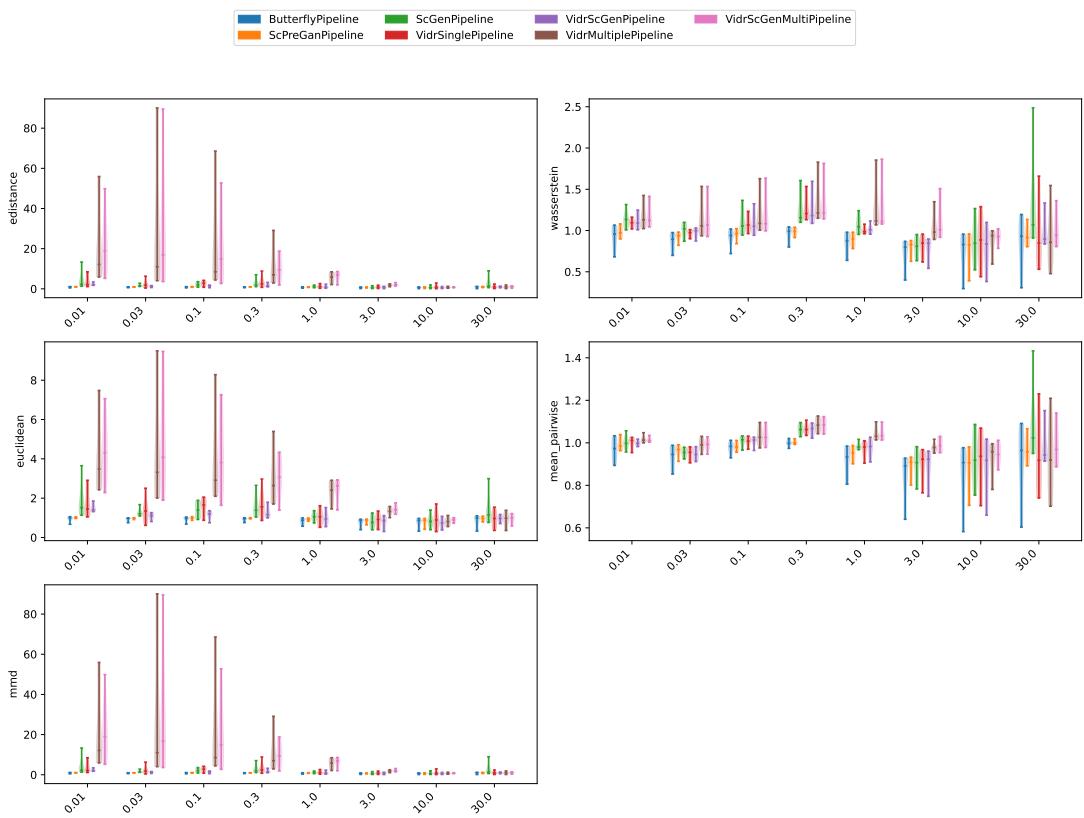
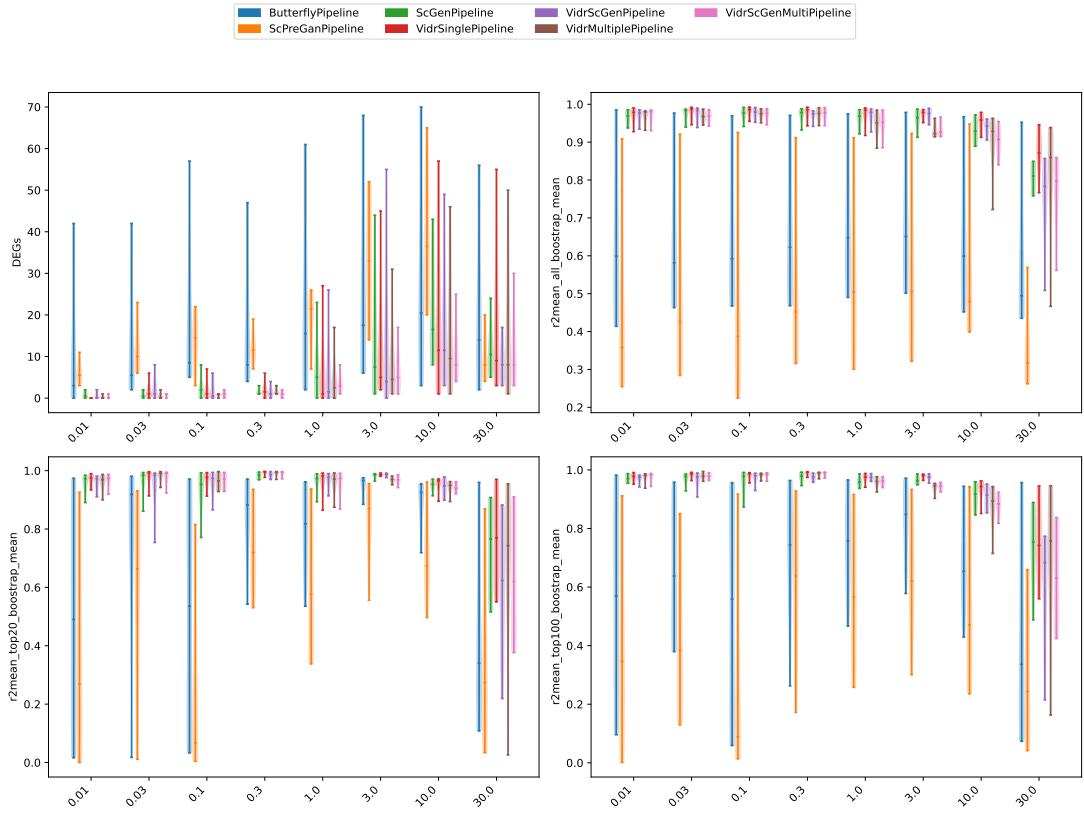
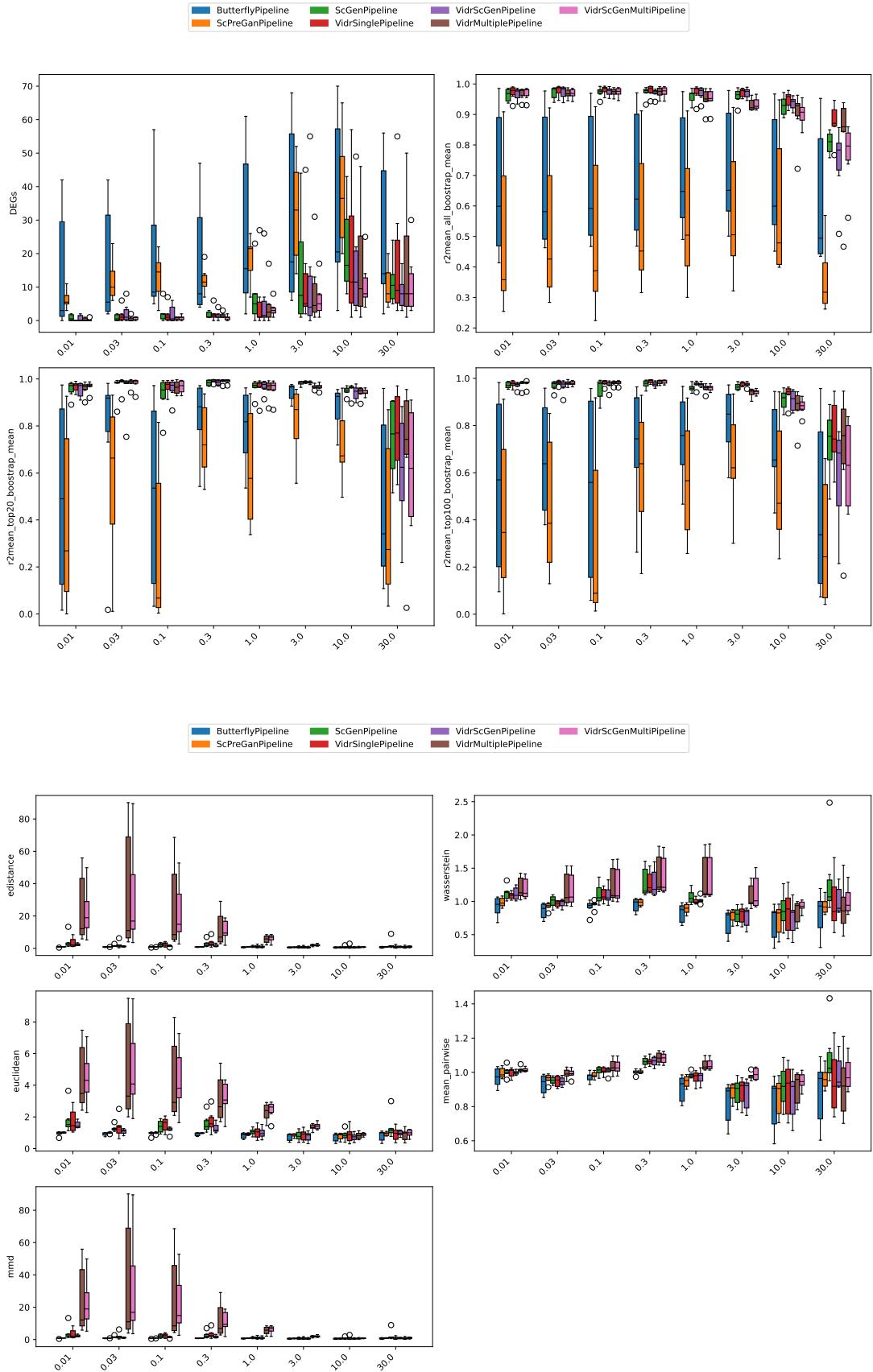
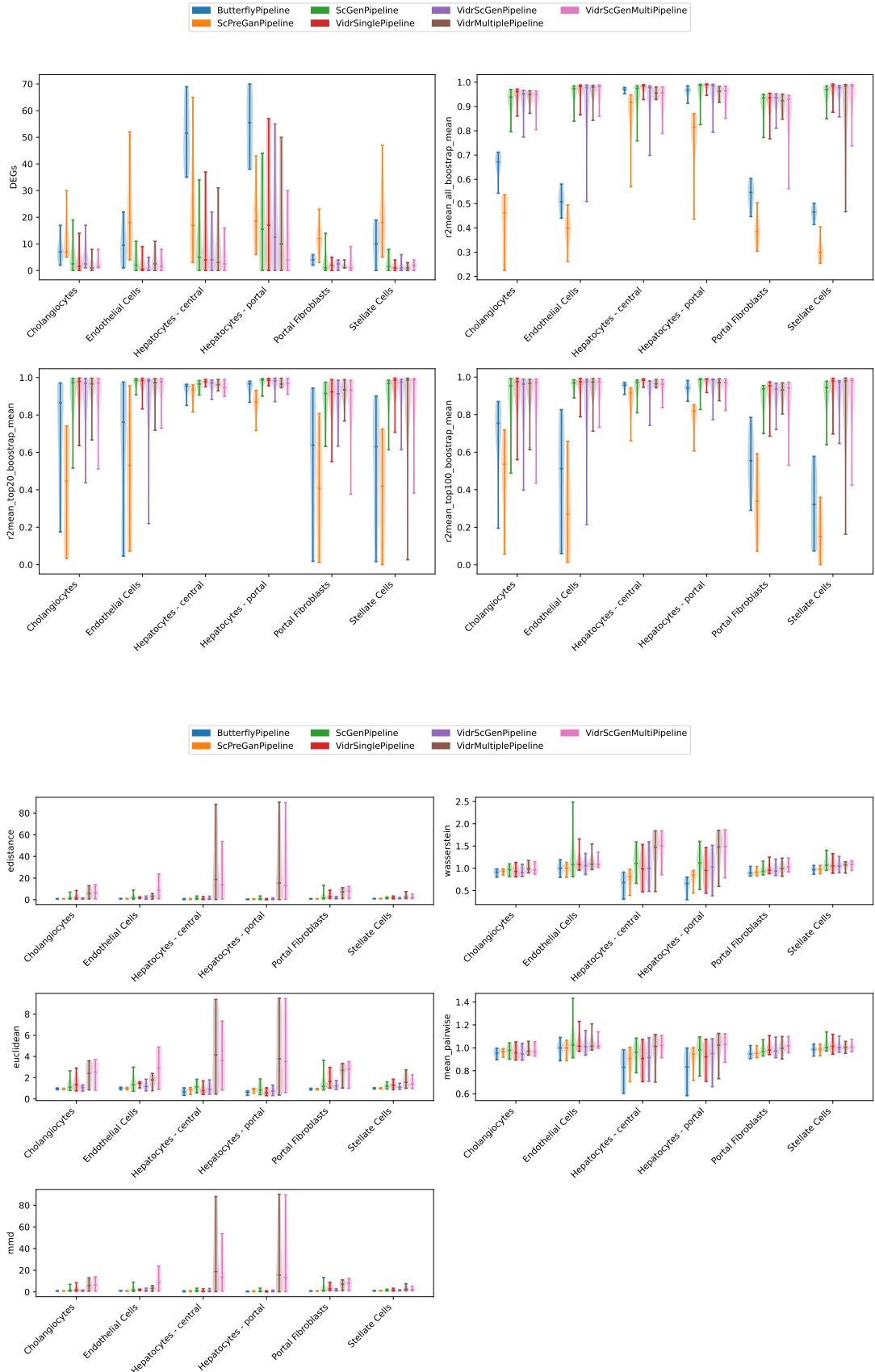


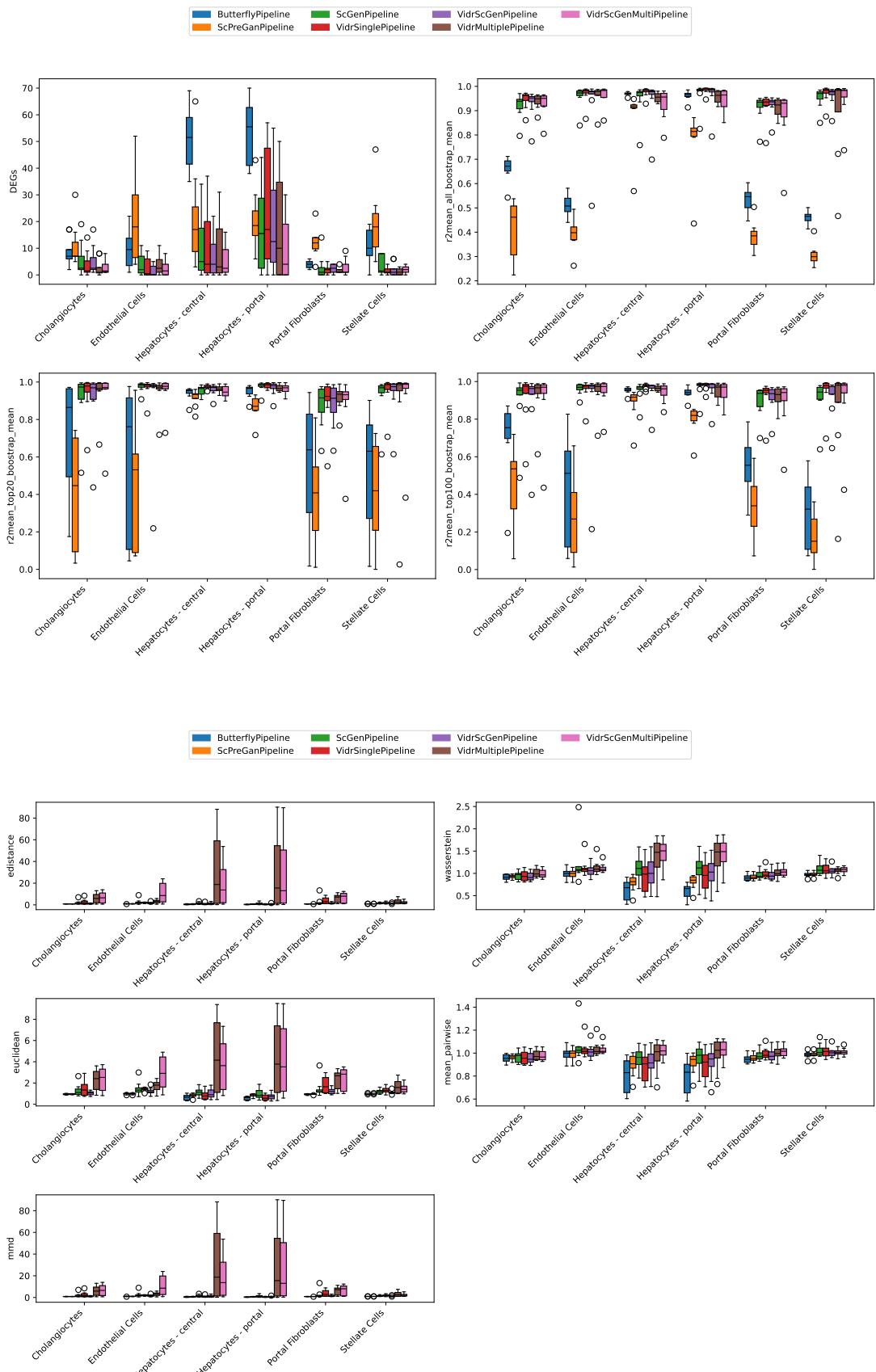
Figure 1.86: Wasserstein

Figure 1.87: Distance metrics per cell type









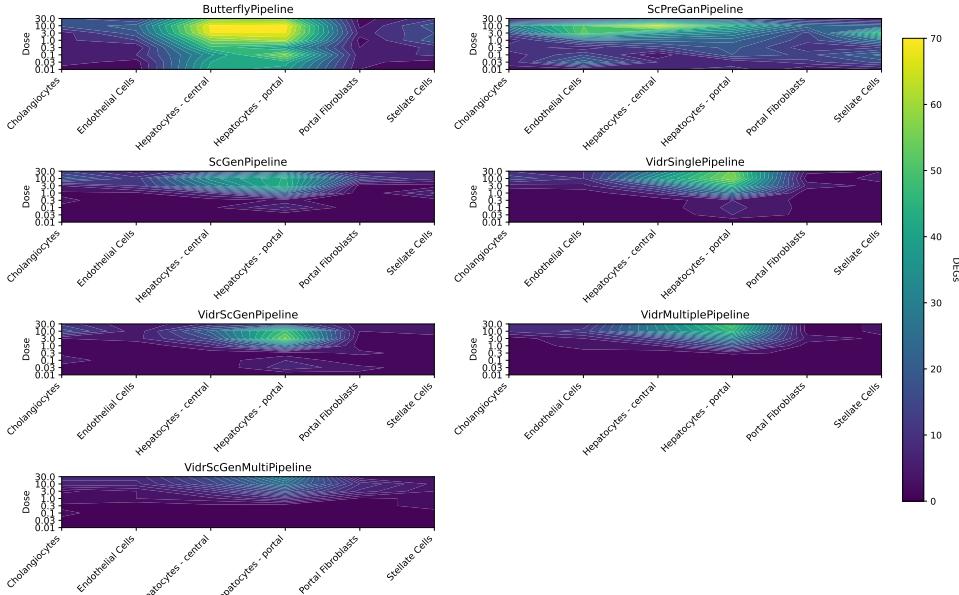


Figure 1.88: DEGs

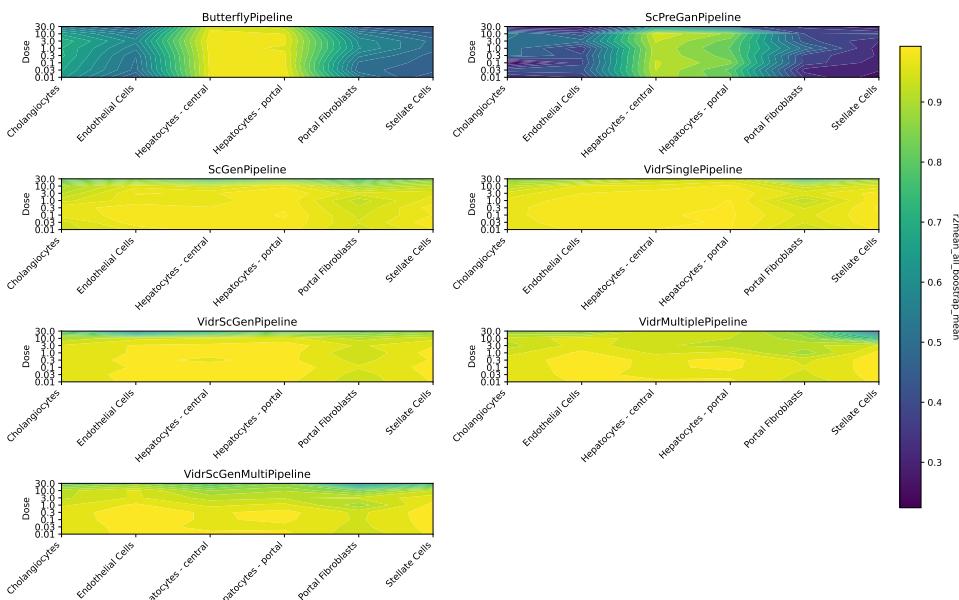


Figure 1.89: r² HVGs

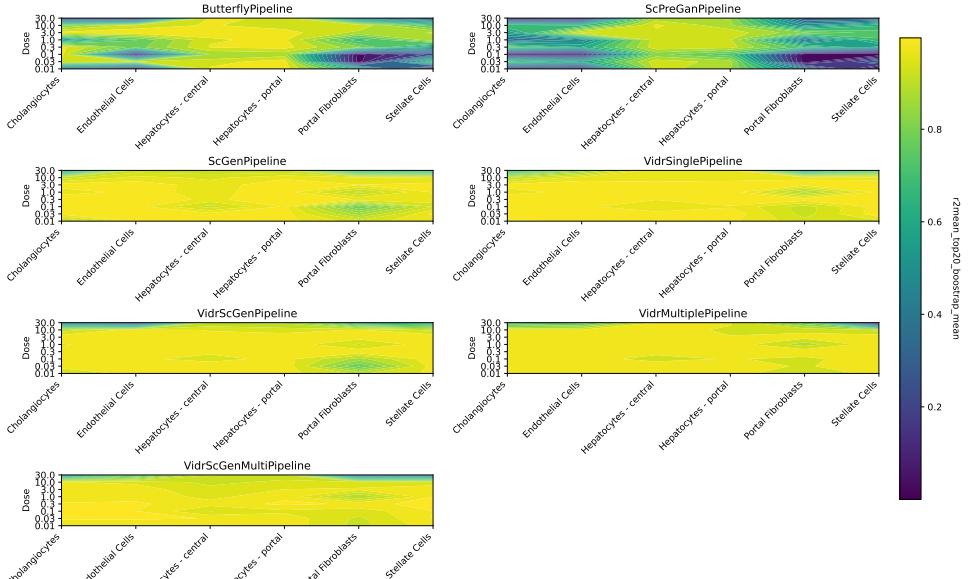


Figure 1.90: r^2 top 20

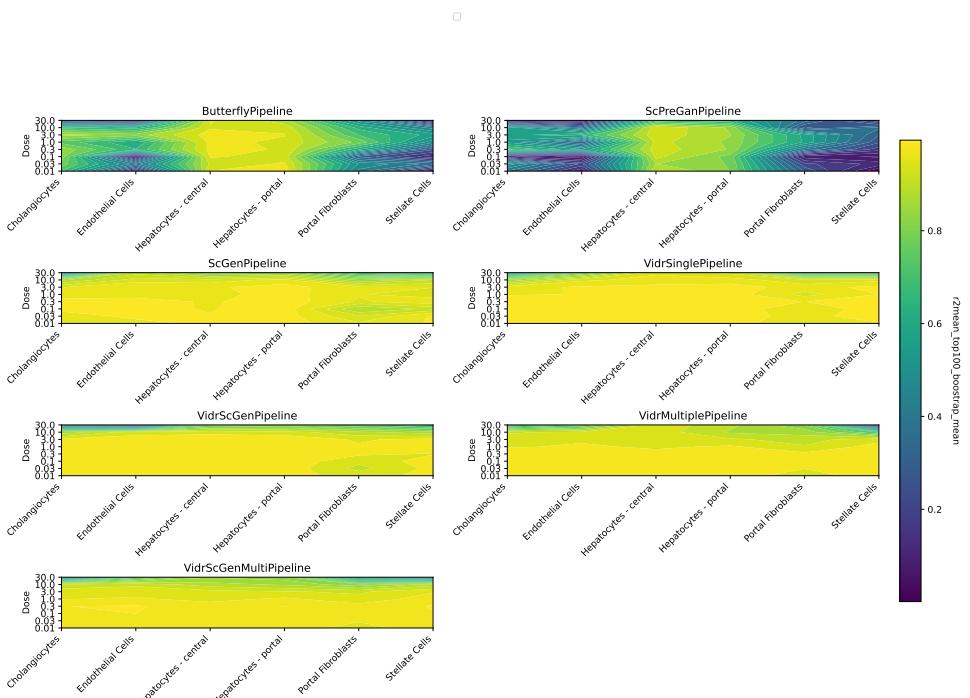


Figure 1.91: r^2 top 100

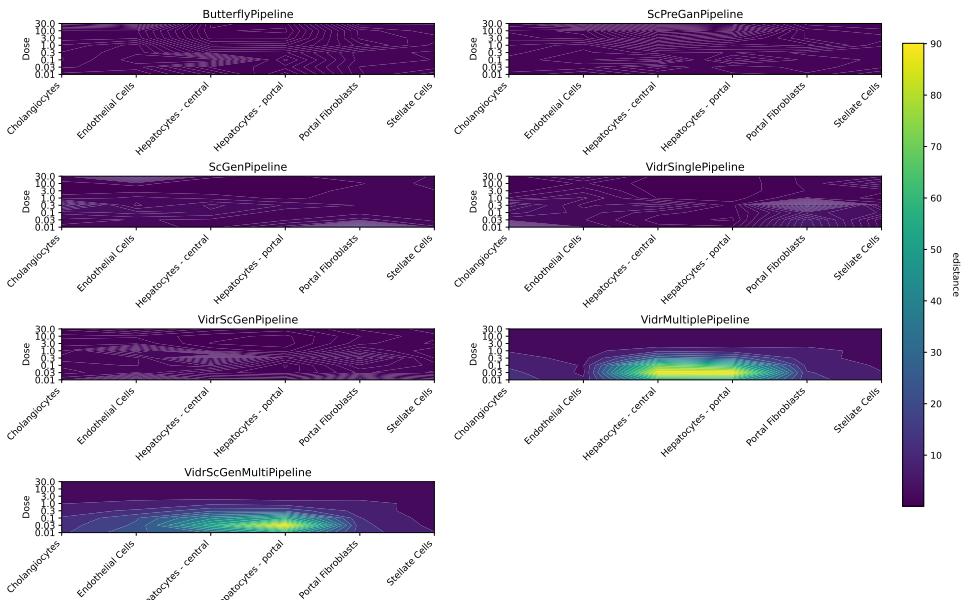
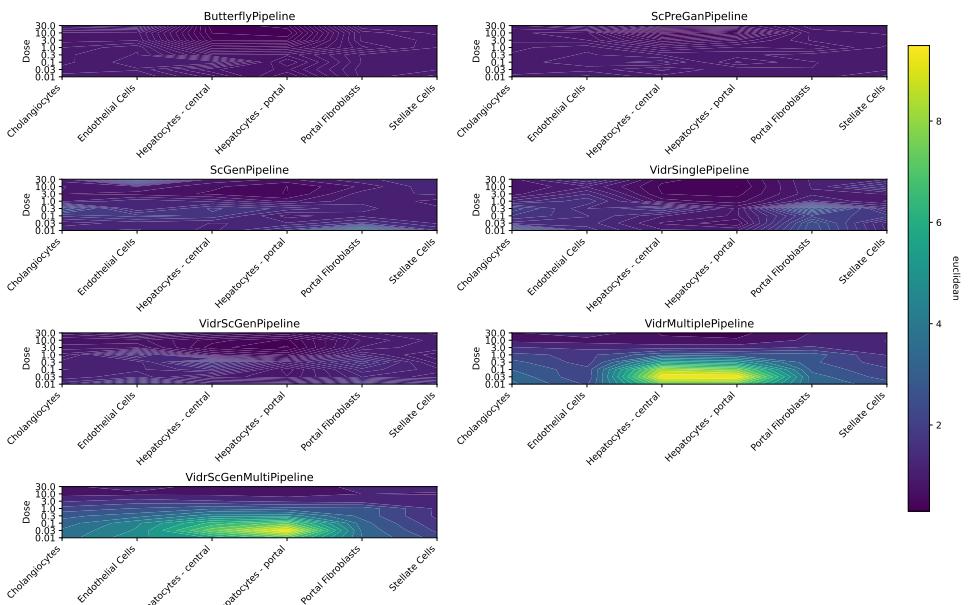


Figure 1.92: E-distance



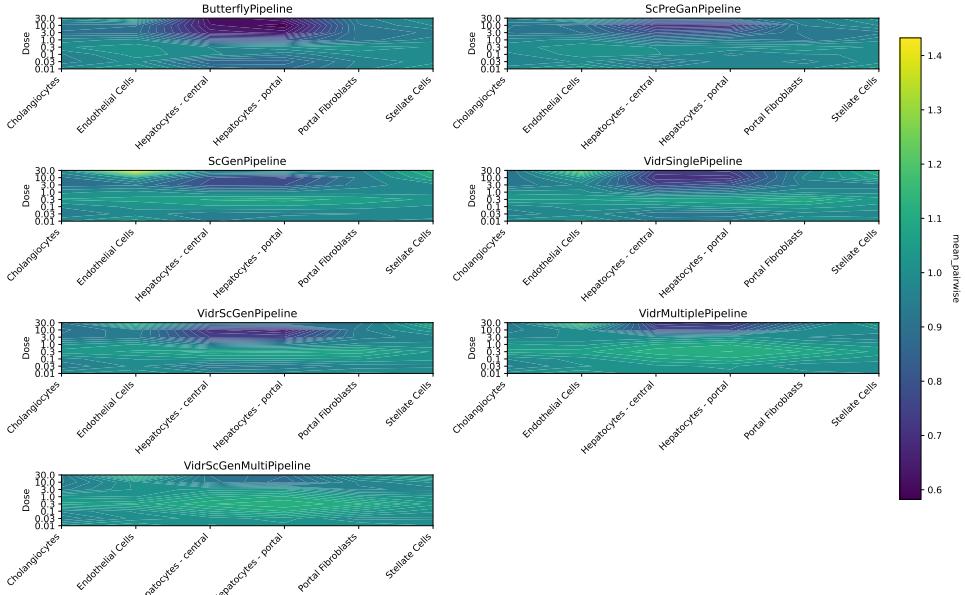


Figure 1.93: Mean pairwise

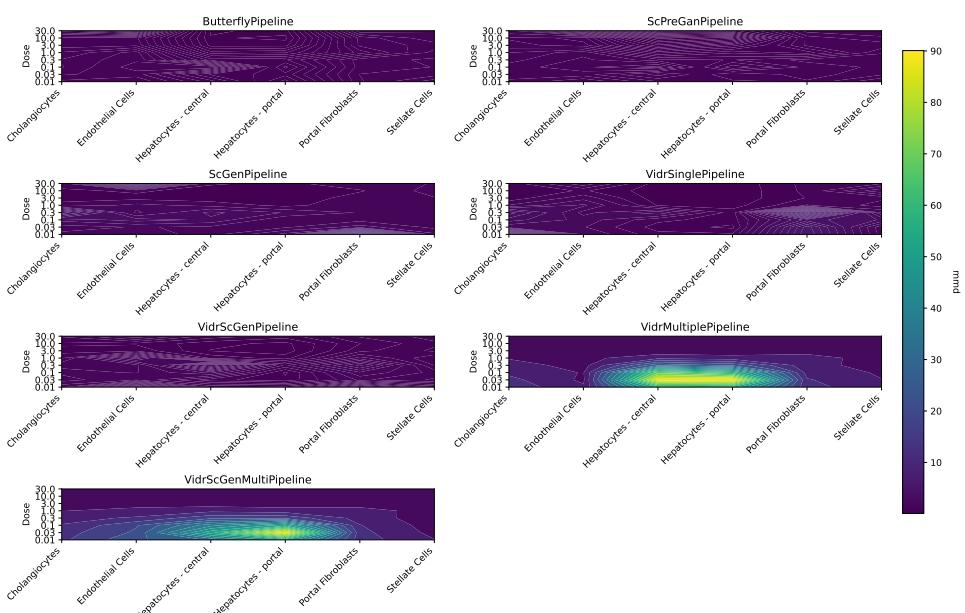


Figure 1.94: MMD

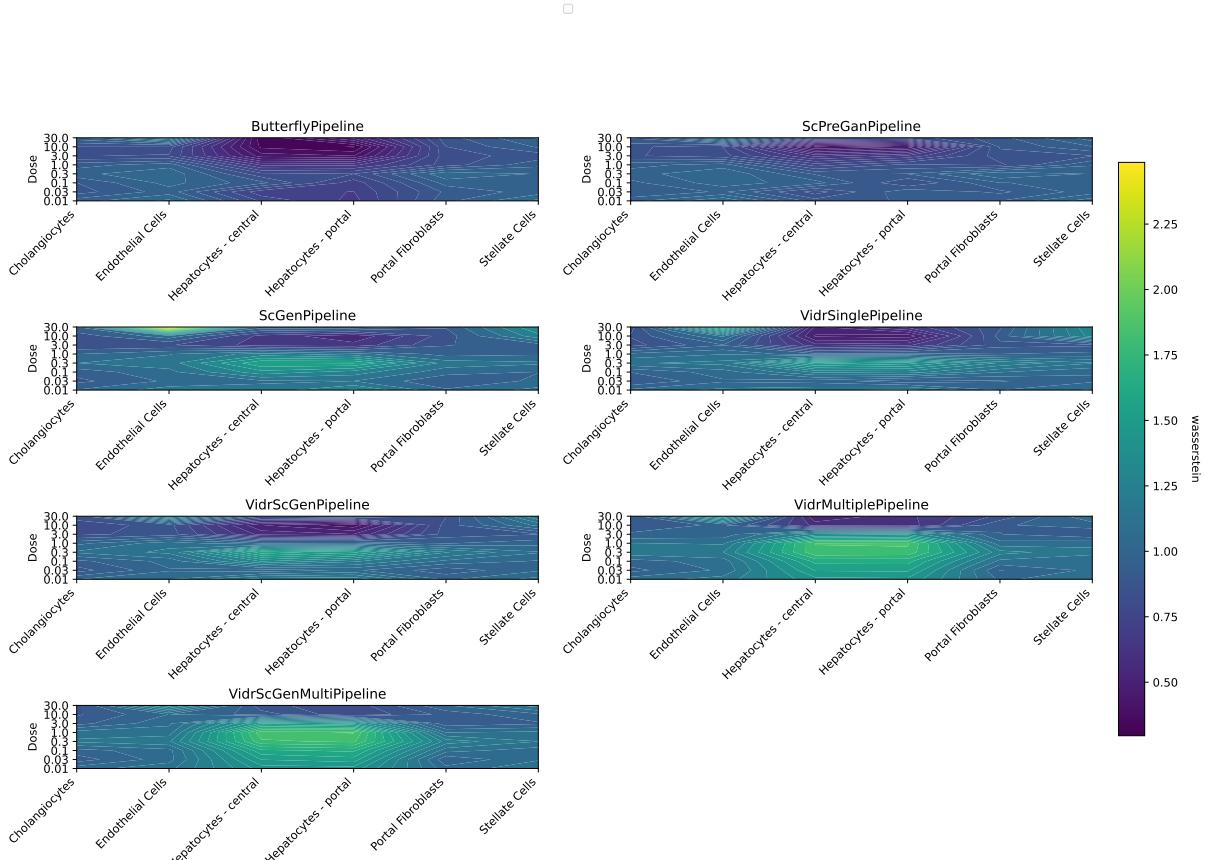
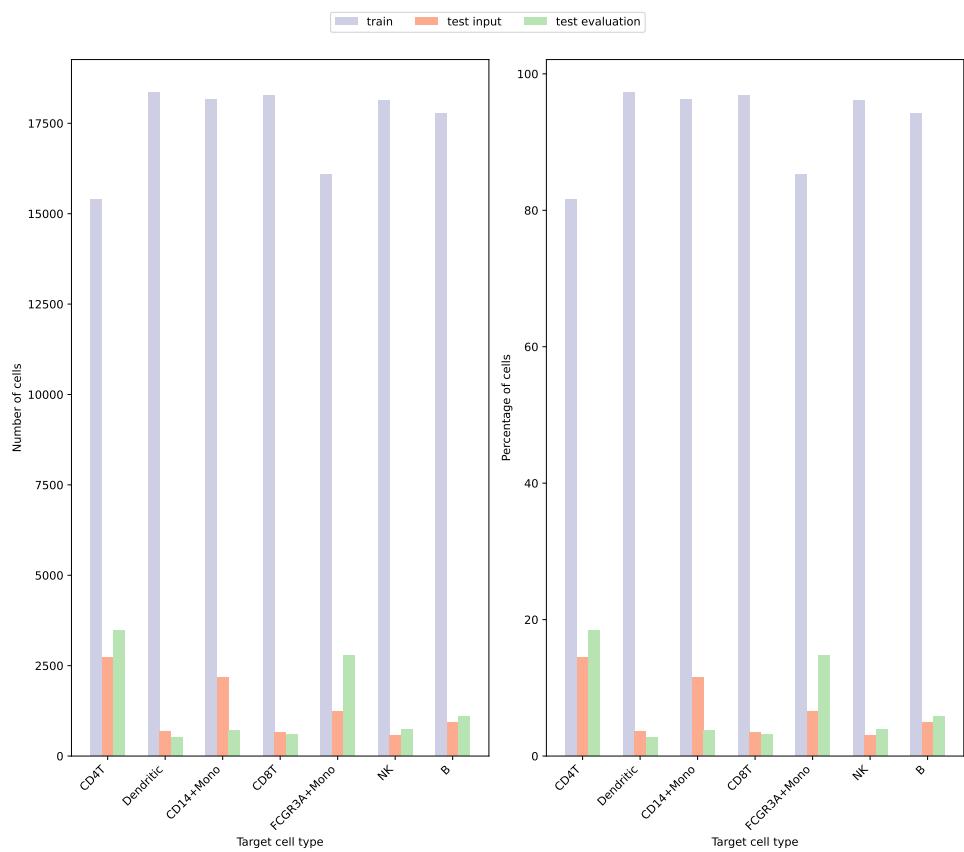
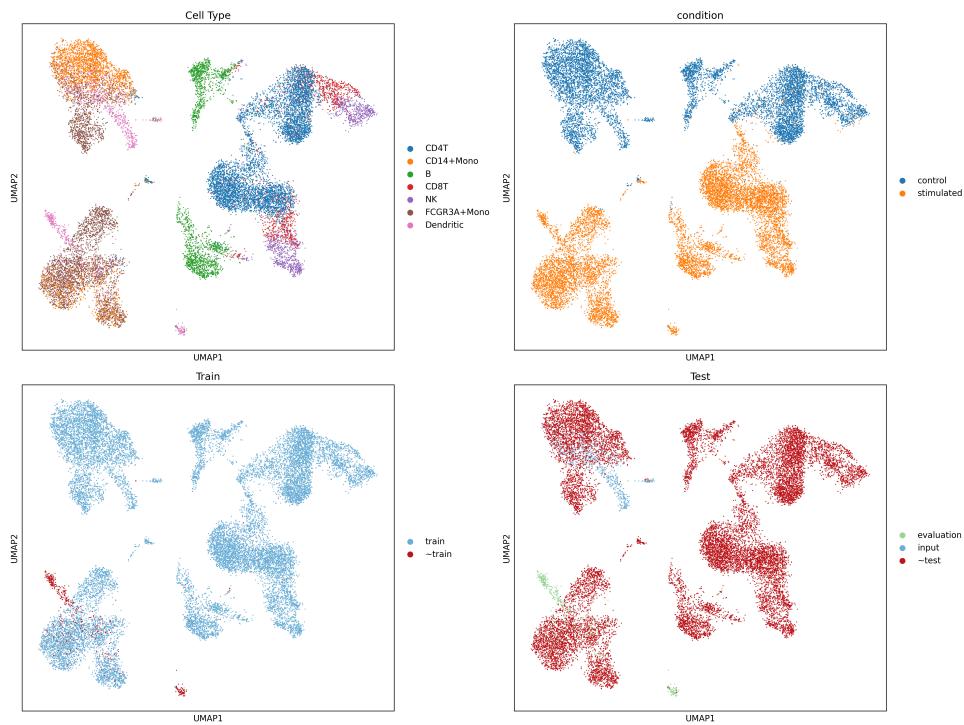


Figure 1.95: Wasserstein

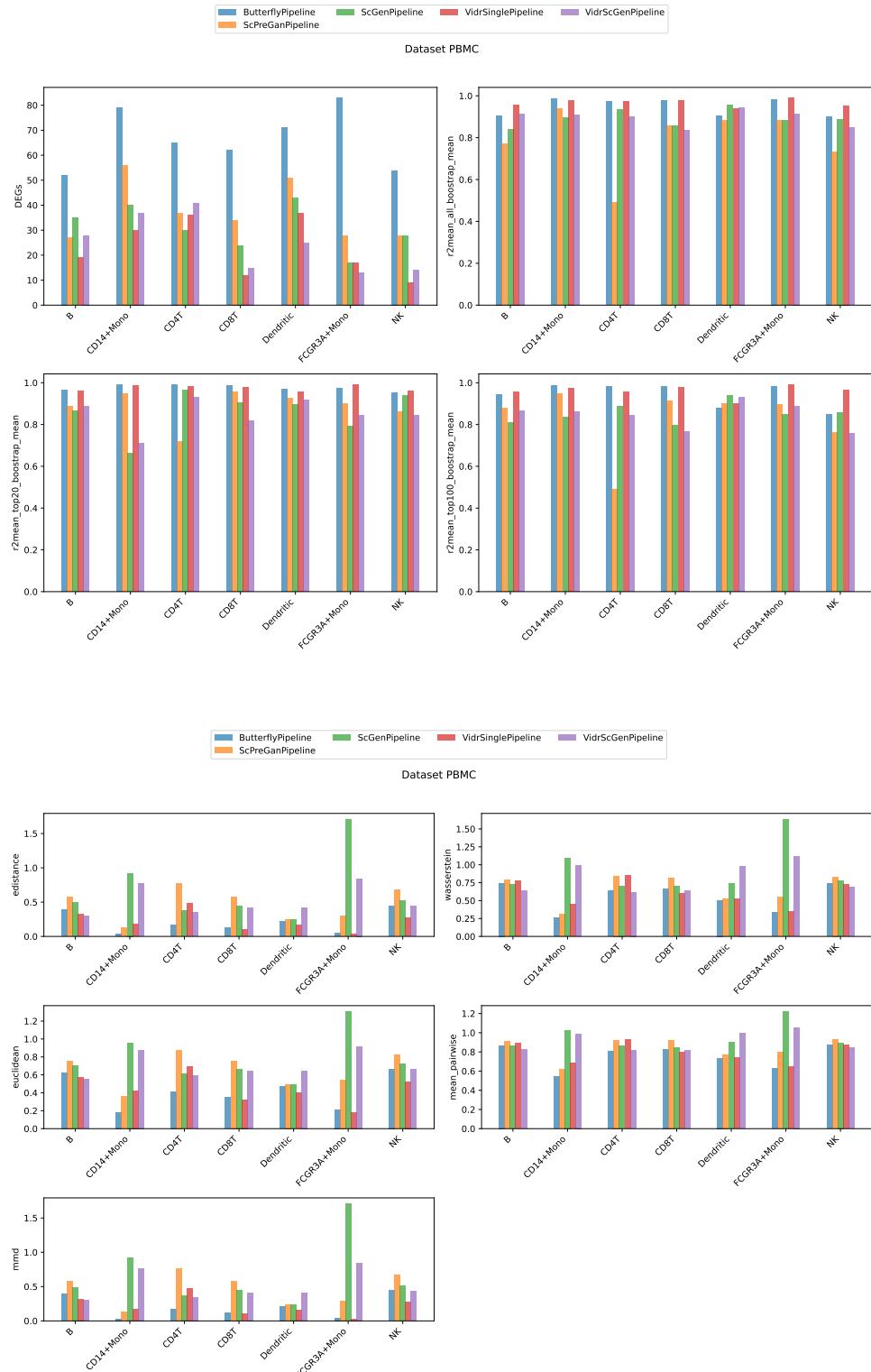
1.3.4 Παρατηρήσεις

- 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).

1.4 PBMC



1.4.1 Comparison



Appendix A

Ακρωνύμια και συντομογραφίες

LAN Local Area Network

Bibliography

- [1] 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.