REGINA

aka **R**egularized **E**ncoder with Latent Cycle-**G**AN for **I**n-vitro **N**eural Cell Perturbation **A**pproximation

presented by Regina Fiam

2026 GPU Day 28th May

Bence Dudas[1], Regina Nora Fiam[2], Csaba Kiss[2,3], Gábor Papp[1], István Csabai[2]
[1]Department of Theoretical Physics, Eötvös Loránd University, Budapest, Hungary
[2]Department of Physics of Complex Systems, Eötvös Loránd University, Budapest, Hungary
[3]HUN-REN Research Centre for Natural Sciences, Budapest, Hungary
dudas.bence@ttk.elte.hu, fregin@student.elte.hu, kmcs1@student.elte.hu
pg@elte.hu, istvan.csabai@ttk.elte.hu

Problem: Unpaired nature of the datas

Fundamental problem with scRNA-seq: requires lysing the cell (measurements not repeatable)

- never observe the exact same cell in both its control state and its perturbed state
- Unpaired distributions: a pool of control cells and a pool of perturbed cells and the goal is to learn a mapping function

$$\mathbb{P}_{\text{ctrl}} \xrightarrow{T_p} \mathbb{P}_{\text{pert}}$$

Previous methods

Additive latent shifts (scGen, CPA)

-encode cells into the latent space

-assumption: perturbation is an additive vector

-problem: fails on non linear interactions (epistasis)

Knowledge graph priors (GEARS)

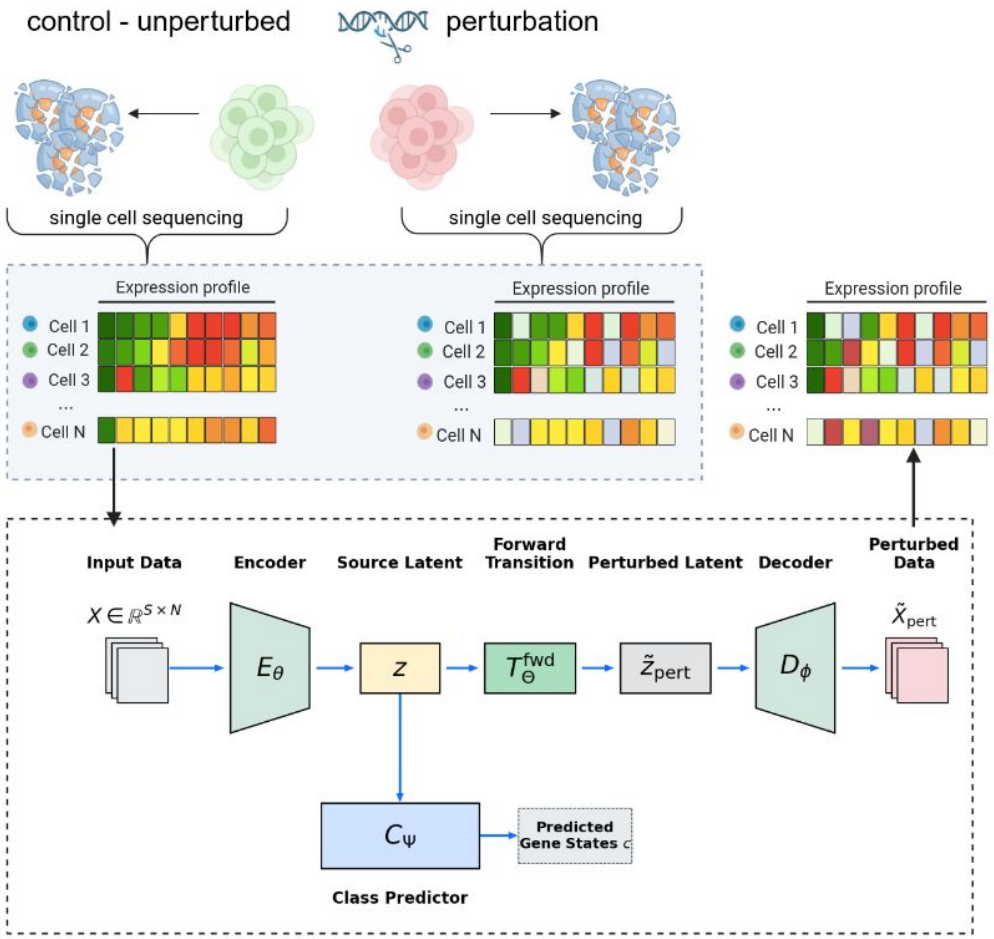
-inject external biological databases (e.g. GEO)

-assumption: pruning data only to highly variable genes is enough

-problem: completeness of the database

Our method

- fully data-driven
- joint unpaired distribution matching problem
- no additive vectors →can capture non linearities
- no prior external knowledge graph
- retains all genes (not only HVGs)



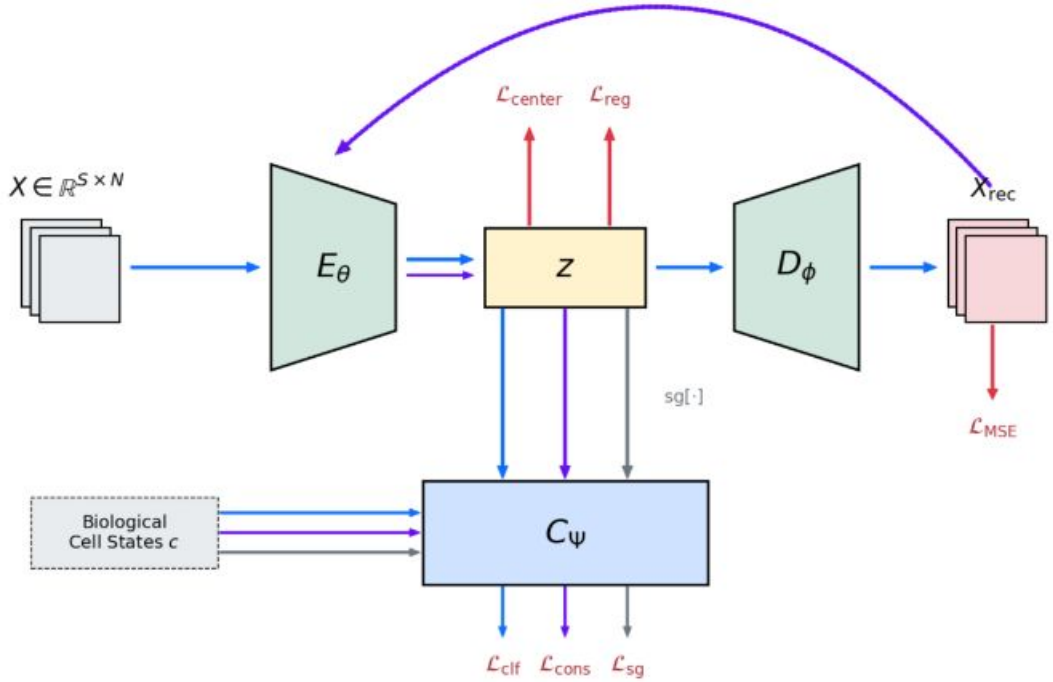
Architecture: Phase I. (latent regularization)

Input: raw gene expressions

Then: *tokenized*, feed into *encoder* (capture global dependencies)

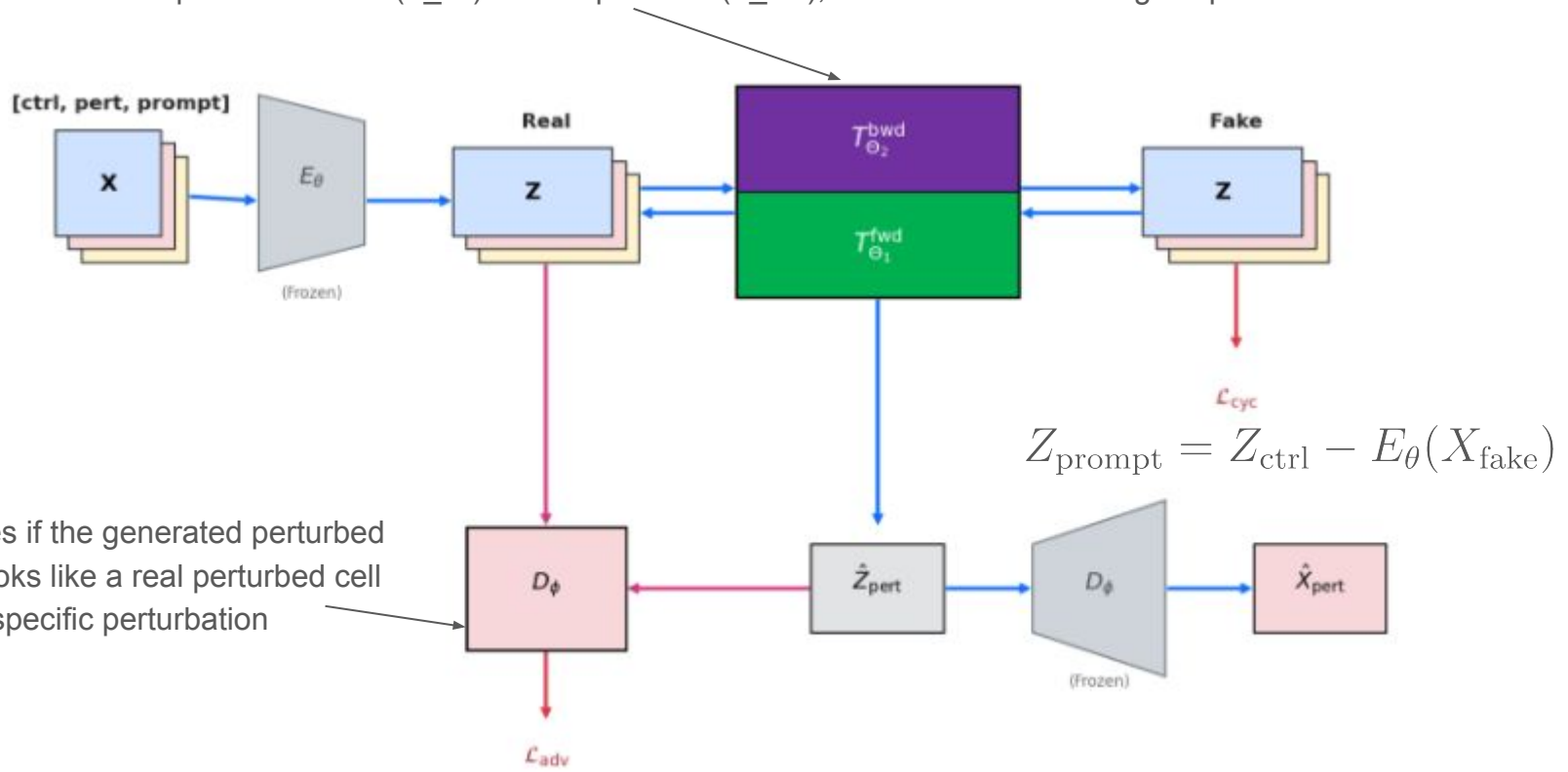
CLF: regularize the compressed data in latent space

Center Loss: minimize the distance in the latent space between the same classes



Architecture: Phase II. (latent cycle-GAN)

map a control cell to a perturbed state (T_fw) and map it back (T_bw), it must return to its original position.



Running a Cycle-GAN directly on 20,000+ dimensions of raw gene expression is highly unstable. Doing it in a regularized latent space makes training robust and scalable.

Generalization to unseen perturbations

How do you predict the impact of a gene knock-out that was never seen during training?

1. synthetic prompting signal (masking out target genes in control)
2. Feeding to the encoder and compute:

$$Z_{\text{prompt}} = Z_{\text{ctrl}} - E_\theta(X_{\text{fake}})$$

Understanding the evaluations

Local precision

Pearson (local-point wise) → predict the specific individual gene levels

Global distributional fidelity

MMD/Wasserstein distance → how well the entire predicted population distribution matches the real population

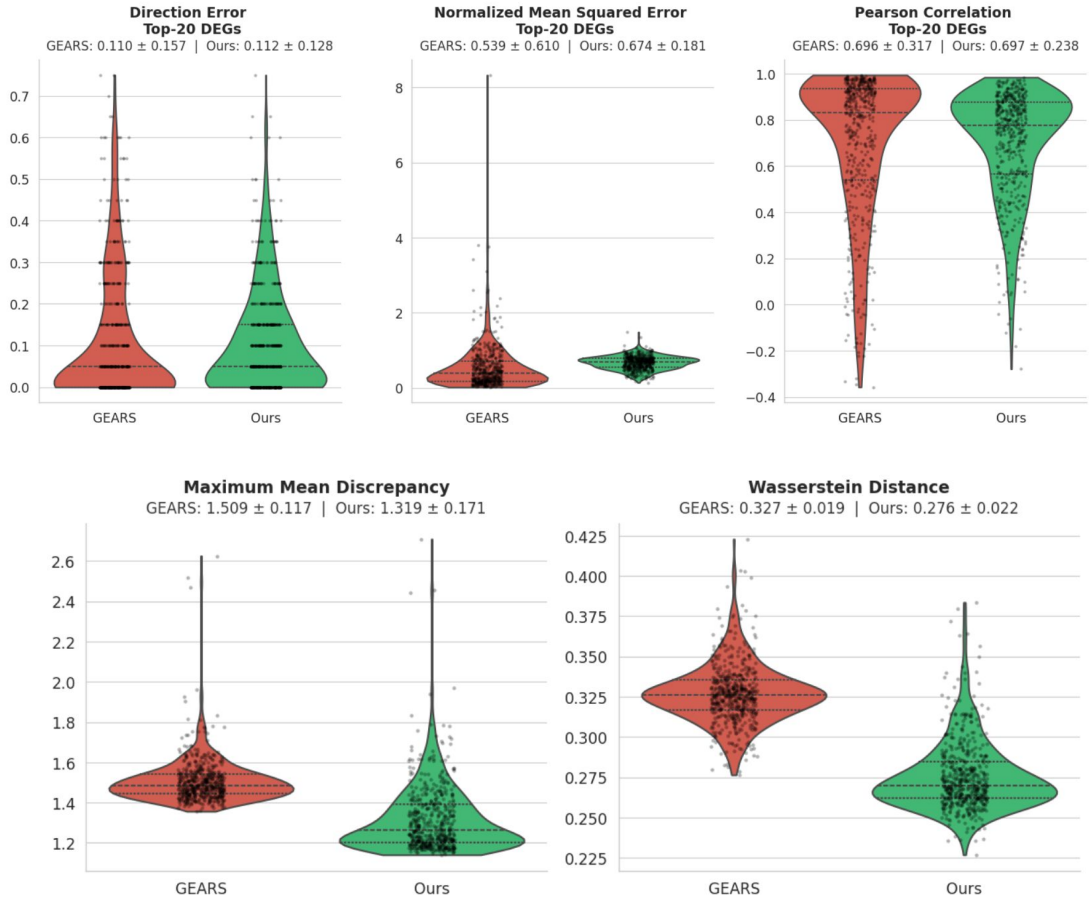
Dataset	Model	Pearson	Wasserstein
Adamson (All Genes)	GEARS	$\underline{0.5854 \pm 0.3548}$	0.2612 ± 0.0187
	PerturbNet	0.2166 ± 0.2742	$\mathbf{0.2101 \pm 0.0258}$
	REGINA (Ours)	$\mathbf{0.6429 \pm 0.3948}$	$\underline{0.2547 \pm 0.0171}$
Adamson (Top 20 DEGs)	GEARS	$\underline{0.6419 \pm 0.5984}$	
	PerturbNet	0.4059 ± 0.3911	
	REGINA (Ours)	$\mathbf{0.6510 \pm 0.5554}$	

Dataset landscape

Replogle: genome-wide (8749 genes, 236k cells) → massive diversity (best for REGINA)

Norman: combinatorial (gene-pair) interaction focus

Dixit: small dataset (24 perturbation)



Ablation study

Latent Classifier → without it, severe manifold collapse (Pearson down)

Center Loss →vital for local precision (Pearson down)

Latent prompting → replacing it with fix embeddings complete failure on unseen genes

Model Configuration	Pearson	Dir. Error	MMD	Wasserstein
REGINA (Full Model)	$\mathbf{0.6844 \pm 0.2944}$	$\mathbf{0.1276 \pm 0.1057}$	$\mathbf{1.4776 \pm 0.1167}$	$\mathbf{0.1583 \pm 0.0257}$
w/o Center Loss	0.5599 ± 0.4169	0.2117 ± 0.1937	1.8696 ± 0.2696	0.1697 ± 0.0292
w/o Latent Classifier	0.0820 ± 0.3043	0.4322 ± 0.1619	4.1214 ± 0.6969	0.3157 ± 0.0555
w/o Adversarial Training	0.6617 ± 0.2292	0.1776 ± 0.1259	1.5742 ± 0.1247	0.1603 ± 0.0246
w/o Latent Prompting	-0.0547 ± 0.2016	0.5136 ± 0.1179	5.6422 ± 0.3384	7.8000 ± 3.4901

Conclusion, future

REGINA successfully replaces rigid biological graph assumptions with an flexible, unpaired optimal transport framework in latent space, but….

REGINA also requires ***high training diversity;*** its performance is fundamentally bounded by the size and variety of the training data, so…

on very small datasets, graph-based methods like GEARS perform better because they can cheat using external knowledge, so…

it balances *local* point-wise accuracy with superior *global* distributional fidelity.

Thank you for your attention!

fregin@student.elte.hu

(Special thanks to my colleagues!)

We acknowledge the Digital Government Development and Project Management Ltd. for awarding us access to the Komondor HPC facility based in Hungary.