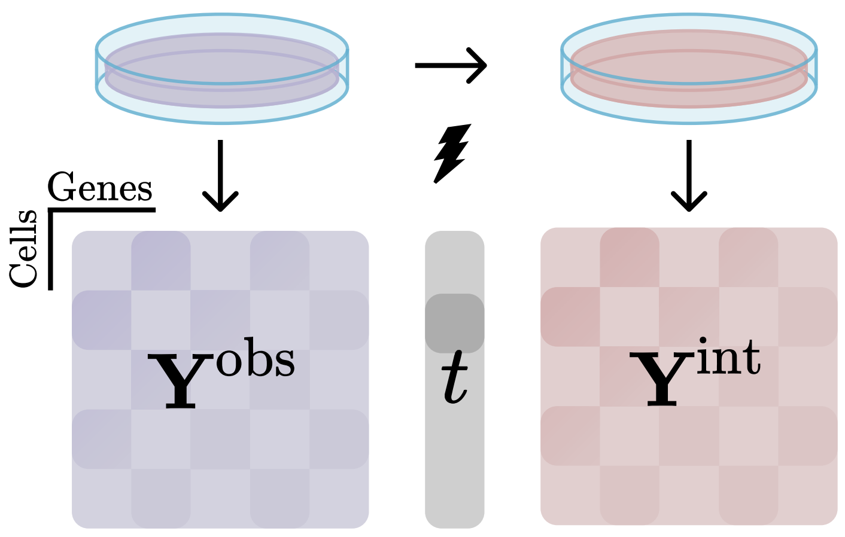
How to build virtual cell foundation models that adapt to unseen biological contexts?

MapPFN: Learning Causal Perturbation Maps in Context

Marvin Sextro[1,2,3], Weronika Kłos[1,2], Gabriel Dernbach[1,4,2,3]

[1]Machine Learning Group, TU Berlin [2]BIFOLD [3]Aignostics [4]Charité

The Data



- CRISPR knockout screens with **single-cell RNA-seq** readout
- Cells destroyed by measurement and **unpaired** across treatments
- Measured across distinct **biological contexts** (cell lines)

The Problem

- Perturbation effects are **context-dependent**
- Combinatorial space **infeasible to cover experimentally**
- Existing methods **retrain and tune per dataset**
- No method conditions on **interventional data**

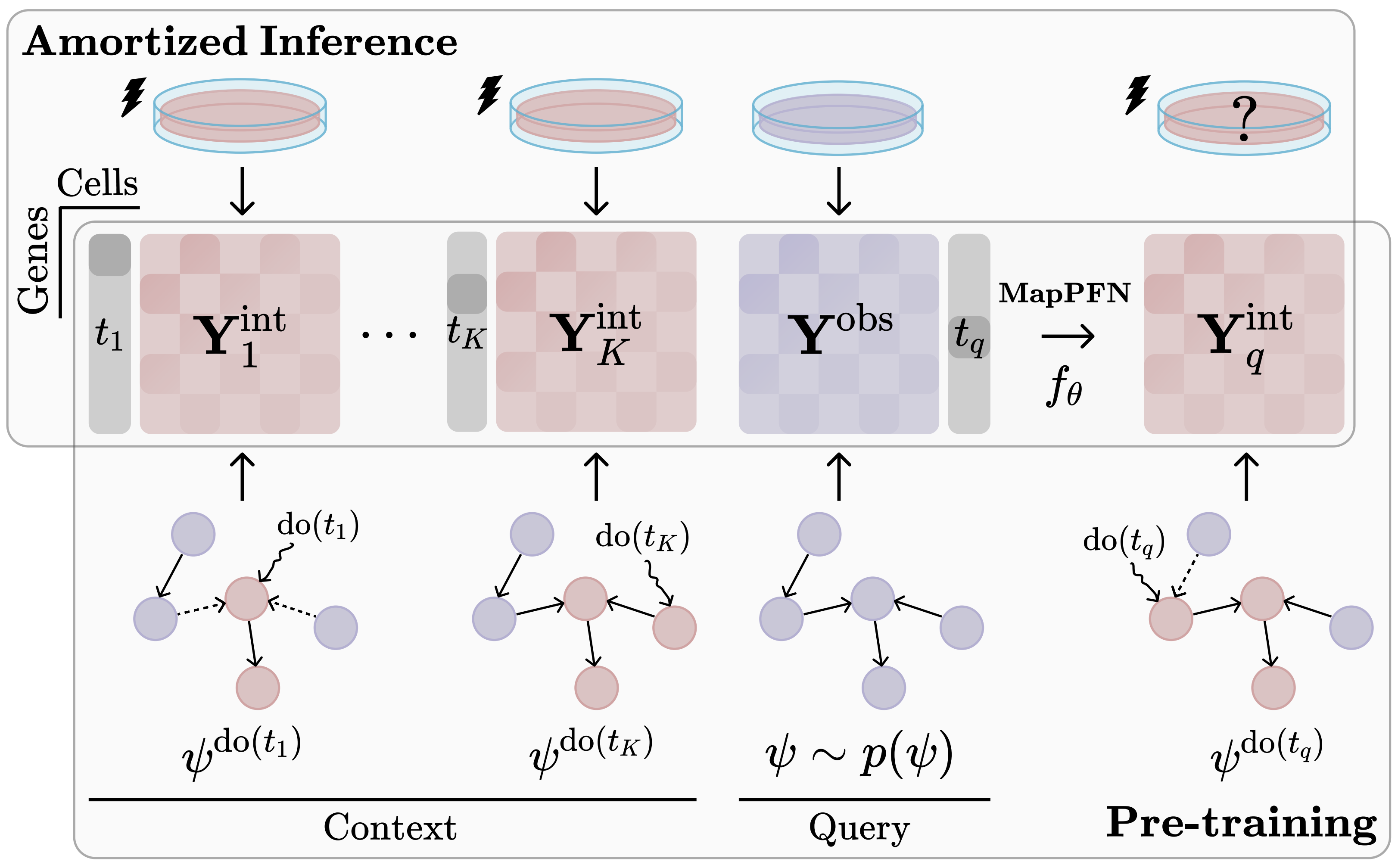
Our Solution

- **Pre-training on synthetic data:** a prior-data fitted network (PFN) that meta-learns causal perturbation prediction
- **Interventional context conditioning:** leverage observed interventions to improve identifiability of causal structure
- **In-Context Learning (ICL):** adapt to unseen biological contexts and gene sets at inference time

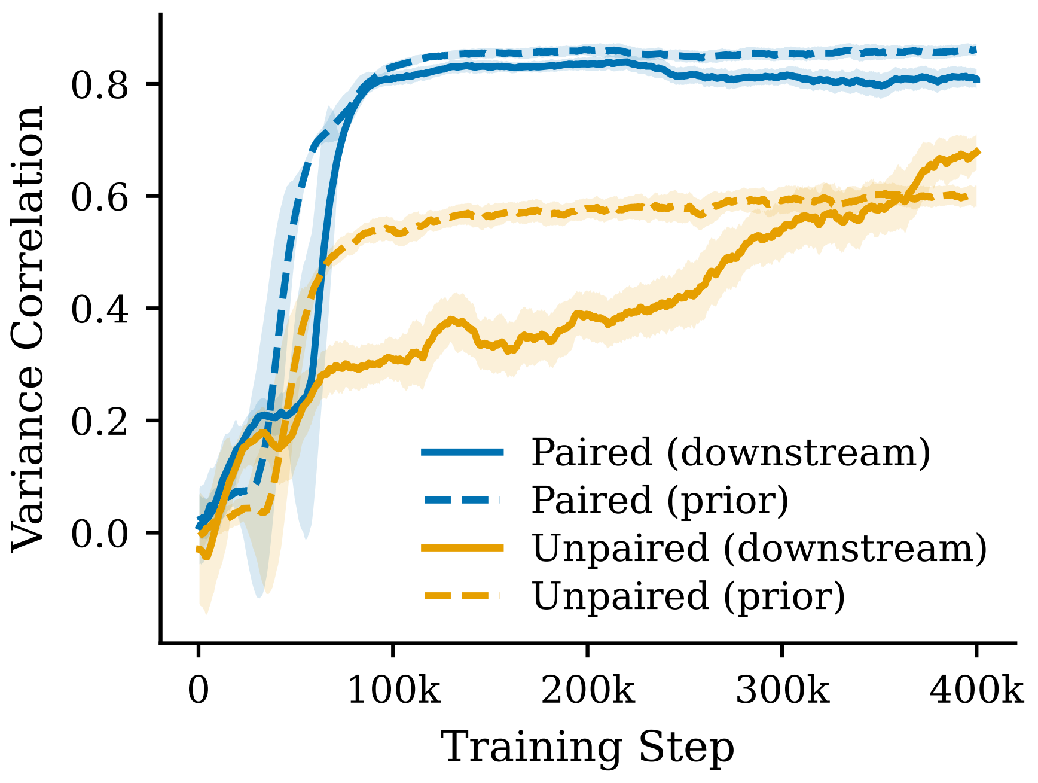
Method

Draw causal model ψ from prior $p(\psi)$, generate $\mathbf{Y}^{\text{obs}}$ and build context $\mathcal{C} = \{(t_k, \mathbf{Y}^{\text{int}}_k)\}_{k=1}^{K}$ by intervening with treatments $t_k \in \mathcal{T}$. MapPFN uses a Multimodal Diffusion Transformer (MMDiT) to approximate the **posterior predictive distribution**:

$$p(\mathbf{y}_q^{\text{int}} \mid \text{do}(t_q), \mathbf{Y}^{\text{obs}}, \mathcal{C}) = \int p(\mathbf{y}_q^{\text{int}} \mid \text{do}(t_q), \mathbf{Y}^{\text{obs}}, \psi)\, p(\psi \mid \mathbf{Y}^{\text{obs}}, \mathcal{C})\, d\psi$$



Synthetic Biological Prior



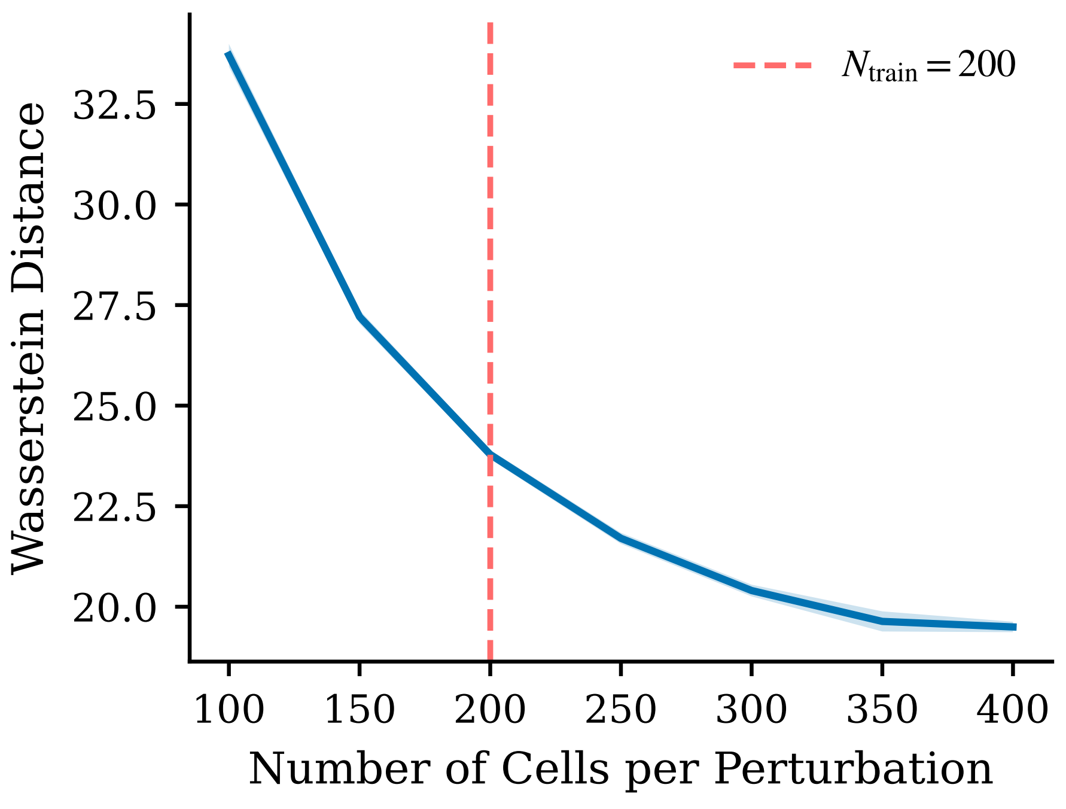
- **Preferential attachment** gene regulatory network (GRN) generator with sparse, directed, modular graphs
- Simulation of **gene expression dynamics** and ***in silico* knockouts** via stochastic differential equations (SDEs)
- **Counterfactual (paired)** prior improves identifiability and downstream performance

Benchmark Results

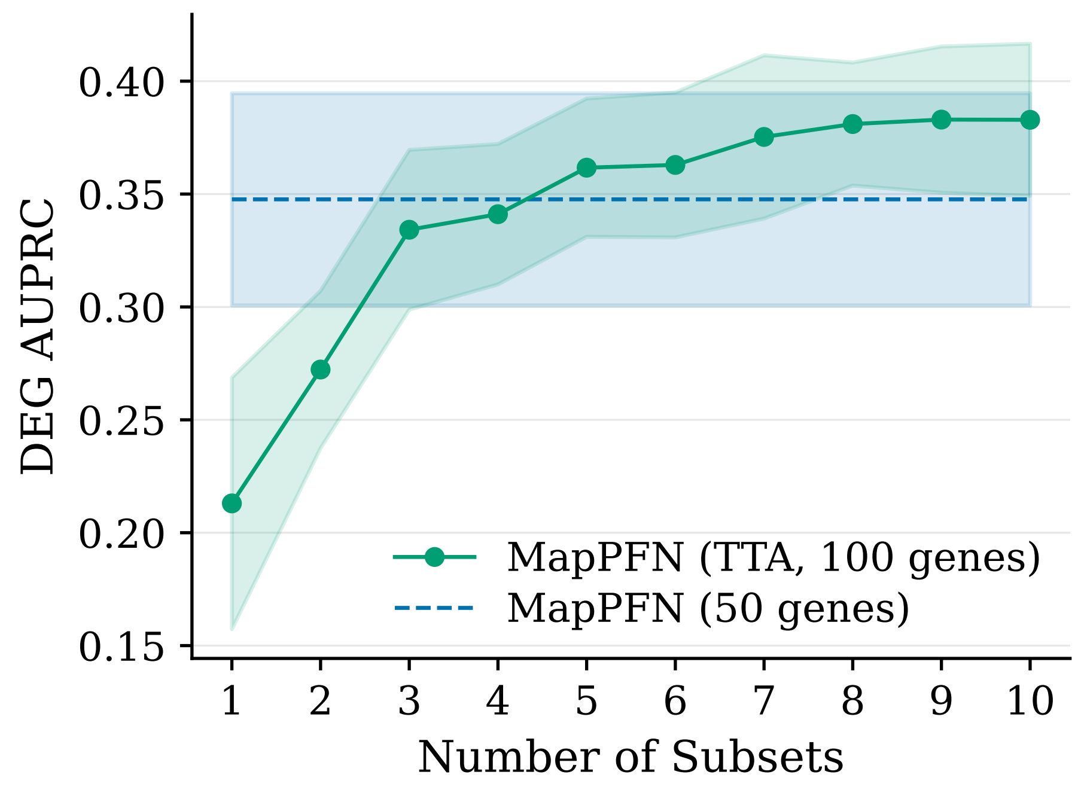
A single pre-trained MapPFN adapts across datasets. Zero-shot, it recovers differentially expressed genes (AUPRC) on par with baselines trained on real data. MMD ×10^{-3}. Mean ± std over 10 seeds.

Cell Line	Method	MMD ↓	RMSE ↓	PDS ↓	AUPRC ↑
Melanoma	CPA	140.09 ±0.35	0.13 ±0.00	0.49 ±0.01	0.04 ±0.00
	CondOT	**7.11** ±0.12	0.10 ±0.00	0.06 ±0.01	0.34 ±0.05
	MFM	7.28 ±0.13	0.10 ±0.00	0.09 ±0.02	0.28 ±0.04
	CellFlow	**7.16** ±0.17	0.10 ±0.00	0.41 ±0.01	0.10 ±0.02
	STATE	7.82 ±0.09	**0.08** ±0.00	0.07 ±0.02	0.33 ±0.04
	MapPFN (pre-trained)	10.07 ±0.19	0.13 ±0.00	0.17 ±0.01	0.34 ±0.02
	MapPFN (fine-tuned)	7.84 ±0.14	0.10 ±0.00	**0.03** ±0.01	**0.38** ±0.03
Leukemia	CPA	78.74 ±1.27	0.17 ±0.00	0.50 ±0.02	0.15 ±0.01
	CondOT	26.51 ±0.68	0.27 ±0.01	0.54 ±0.04	0.14 ±0.01
	MFM	105.64 ±0.73	0.71 ±0.00	0.51 ±0.02	0.16 ±0.01
	CellFlow	14.55 ±0.46	0.17 ±0.00	0.50 ±0.01	0.16 ±0.01
	STATE	15.28 ±0.44	0.17 ±0.00	0.47 ±0.03	**0.17** ±0.01
	MapPFN (pre-trained)	191.88 ±1.46	0.78 ±0.00	0.49 ±0.01	0.16 ±0.01
	MapPFN (fine-tuned)	**12.24** ±0.58	**0.15** ±0.00	**0.42** ±0.03	**0.18** ±0.01

Test-Time Adaptation



Predictions **improve beyond the number of cells** seen during pre-training.



Scales to larger gene sets at inference time through test-time augmentation.

Ablations

Both **interventional context** and **paired prior** are critical. Melanoma. MMD ×10^{-3}.

Configuration	MMD ↓	RMSE ↓	PDS ↓	AUPRC ↑
MapPFN	**10.07**	**0.13**	**0.17**	**0.34**
− paired prior	21.84	0.23	0.20	0.21
− interventional context	15.88	0.20	0.20	0.13

Meta-learning on synthetic biological priors enables context-adaptive virtual cell models through in-context learning.

TECHNISCHE UNIVERSITÄT BERLIN BIFOLD aignostics CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN

Project Page Paper Code