Escaping Local Optima in the Waddington Landscape: A Multi-Stage TRPO-PPO Approach for Single-Cell Gene Perturbation Analysis

Boabang Francis

Content

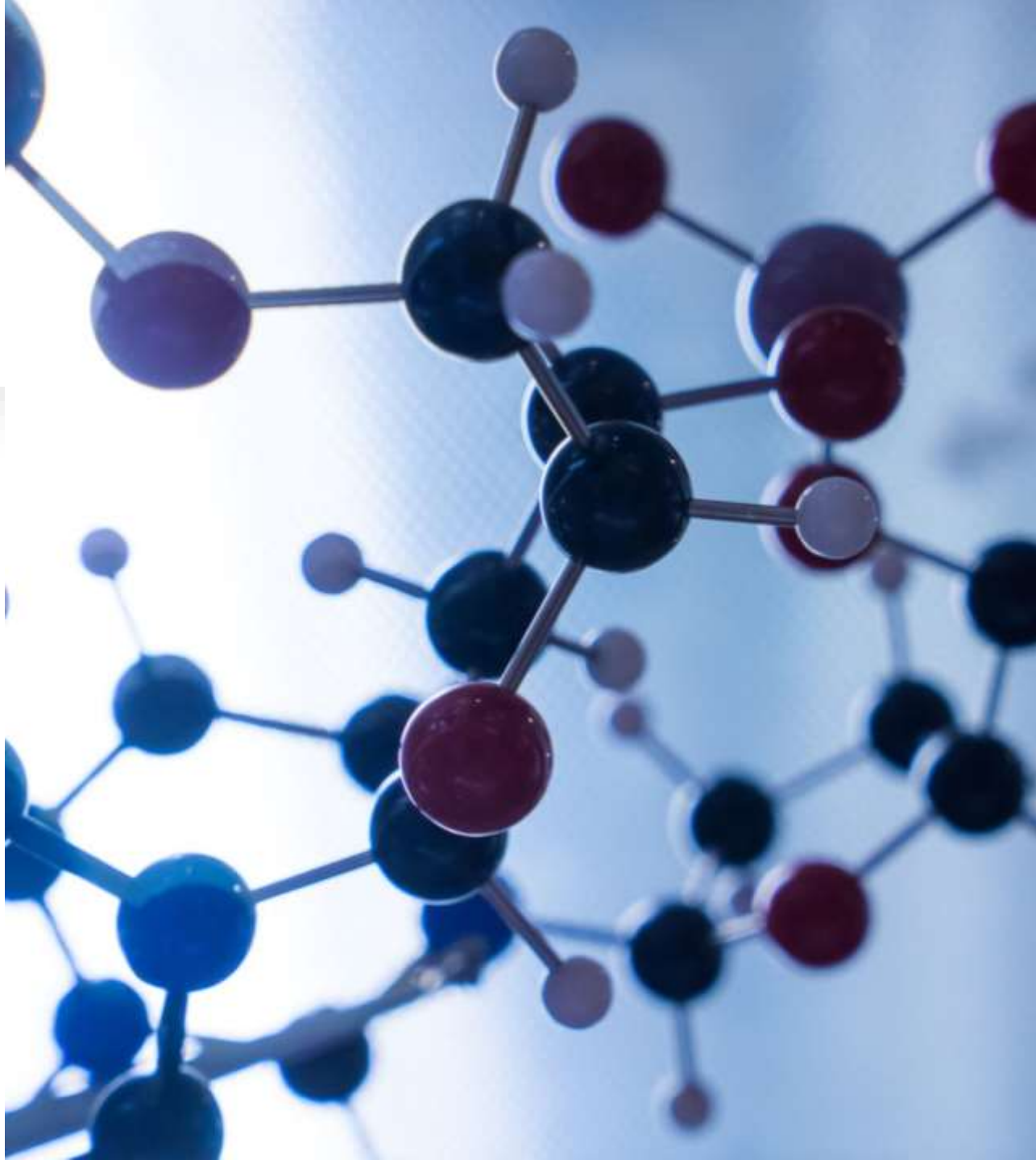
- Background
- Gaps in the Research
- Proposed Solution
- Datasets
- Model Design
- System Design (Hardware and Software)
- Crisper Knockout Evaluation Setting
- Drug Development Evaluation Setting
- Expected Outcome
- Future Work

Background

- Predicting the outcome of genetic perturbations (e.g., CRISPR knockouts) is key to understanding cell differentiation and disease response.

- Perturbation datasets (Perturb-seq, Mixscape)

enable modeling the causal mapping between perturbations and transcriptional outcomes.



Gaps in the Research

Known:

- Linear models (LASSO, Elastic Net) identify sparse gene–gene dependencies. Linear models can fail to capture nonlinear gene–gene or protein to protein dependencies or combinatorial perturbations.
- Gaussian Process Regression captures limited nonlinearity.
- Reinforcement learning (RL) shows promise for modeling dynamic perturbation sequences

- **Unknown:**
- The nonlinear extensions such as Gaussian Process regression, graph neural networks, or reinforcement learning frameworks that better represent biological complexity and adaptive perturbation effects however they often get stuck in local optima in the Waddington landscape.
- Most existing models rely exclusively on either *in silico* perturbation data or experimental perturbation data but rarely integrate both limiting their ability to generalize and validate predictions across simulated and real biological contexts. **Hence, g**eneralization to *out-of-sample* experimental perturbations.
- Interpretability of perturbation effects across multi-omics layers (RNA, ATAC, ADT) with experimental Validation.

References:

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). ***Inferring regulatory networks from expression data using tree-based methods***. PLoS ONE, 5(9): e12776.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., ... & Kellis, M. (2012). ***Wisdom of crowds for robust gene network inference***. Nature Methods, 9(8): 796–804.

Xing, H., & Yau, C. (2025). ***GPerturb: Gaussian Process Modelling of Single-Cell Perturbation Data***. Nature Communications, 16(1): 5423.

Gavriilidis, G. I., et al. (2024). ***A Mini-review on Perturbation Modelling Across Single-Cell Omic Modalities***. Computational and Structural Biotechnology Journal, 23: 1886–1895.

Proposed Solution: Using Initialization Strategies to Address Nonlinearity in Developmental Biology

- We leverage machine-learning initialization strategies to overcome the inherent nonlinearity in developmental biology, improving the modeling of cell-differentiation trajectories.
- Concept: Convex or structured initializations (e.g., trust-region) provide stable starting points for highly nonlinear optimization landscapes in biological systems [1] .
- Goal: To ensure that nonlinear models such as deep or reinforcement-learning frameworks converge toward biologically meaningful minima rather than arbitrary local optima.
- Impact: Enhances the fidelity of cell-fate prediction and differentiation modeling, leading to more accurate simulations of lineage commitment and reprogramming dynamics.
- Outcome: A unified approach where machine-domain initialization bridges computational optimization and biological interpretability, improving predictive performance in cell differentiation and regenerative modeling.

[1] Boabang, Francis, and Samuel Asante Gyamerah. "Escaping Local Optima in the Waddington Landscape: A Multi-Stage TRPO-PPO Approach for Single-Cell Perturbation Analysis." *arXiv preprint arXiv:2510.13018* (2025).

Datasets

Perturb-seq / Mixscape – CRISPR perturbations with transcriptomic readouts for **validation.**

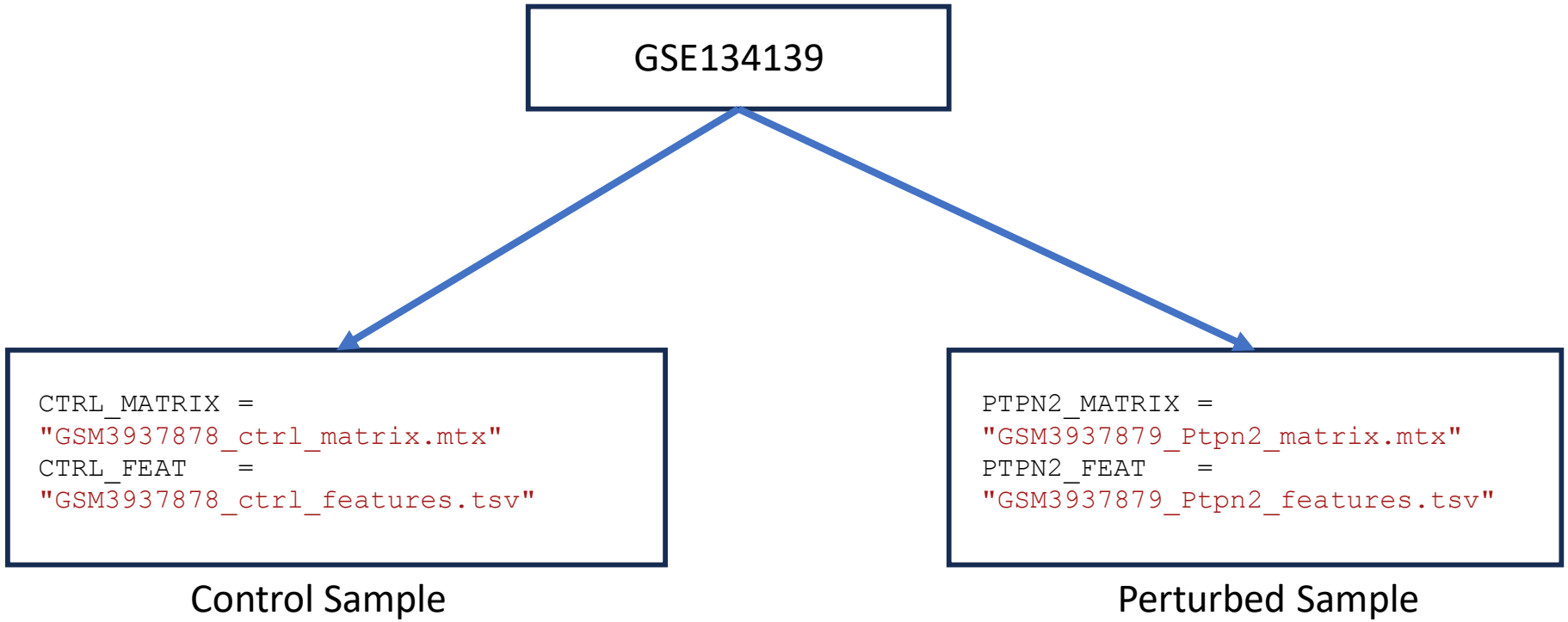
scRNA-seq – Multi-omic single-cell data.

Simulated GRNs – In-silico perturbation environments for **model training.**

http://www.perturbase.cn/

https://www.ncbi.nlm.nih.gov/bioproject/?term=cite

Sample Data Example

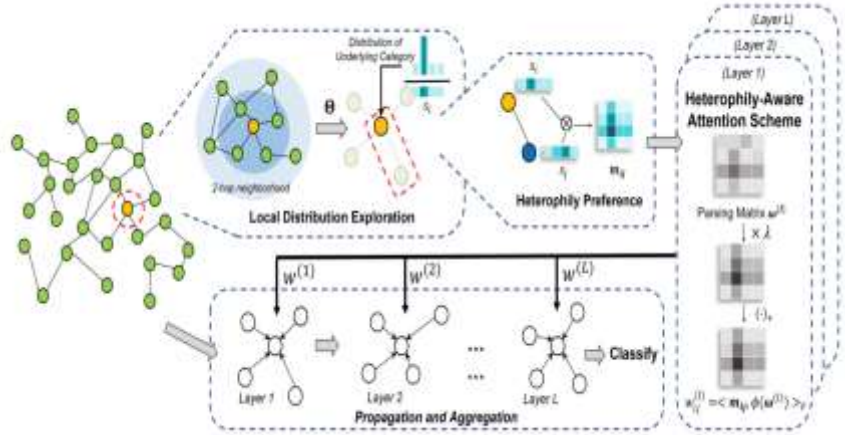


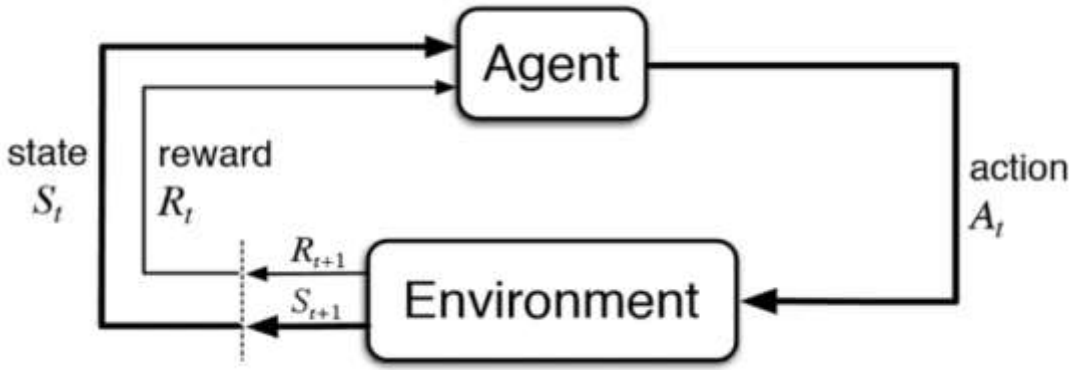
https://drive.google.com/drive/folders/1AAI4KU9G-UHoKo72Ne6iGAaCgFPADIW3?usp=drive_link

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134139

Model Design

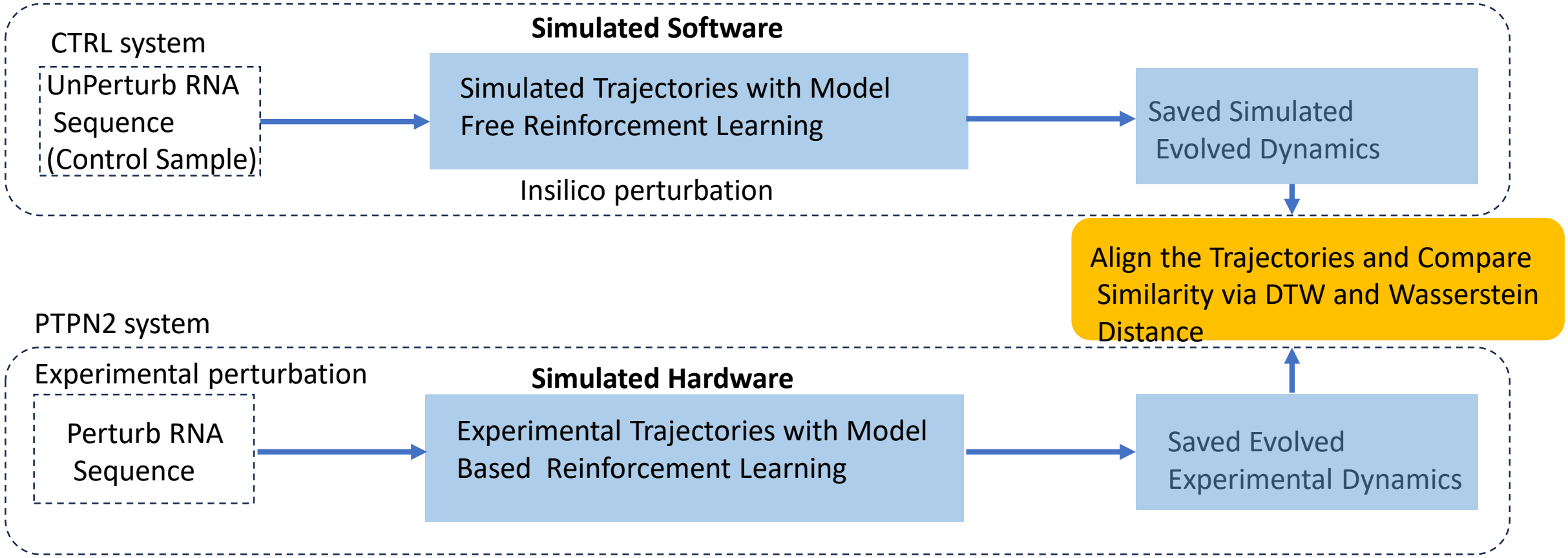
Graph Attention Network for the Gene Embedding





Reinforcement Learning for Trajectory Prediction

Simulated Hardware and Software for Gene Perturbation Analysis



Crisper Knockout

Crispr Knockout (Single Perturbation)

Unperturb data (Control Sample) and Perturb data with CRISPRKO

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134139

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156478

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126310

RNA Dataset Description

Organism: Mus musculus

RNA sequence

Perturbation: Crisper Knockout: Crisprko

Unperturb data (Control Sample) –CTRL_data

Perturb data with CRISPRKO-PTPN2 data

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134139

https://drive.google.com/drive/folders/1AAI4KU9G-UHoKo72Ne6iGAaCgFPADIW3?usp=drive_link

LaFleur MW, Nguyen TH, Coxe MA, Miller BC et al. PTPN2 regulates the generation of exhausted CD8(+) T cell subpopulations and restrains tumor immunity. *Nat Immunol* 2019 Oct;20(10):1335-1347.

Code Link

- Colab:

https://colab.research.google.com/drive/1pJJpmq2cz5iY0ctqEYmCx7U8p4kUNRa-?usp=sharing

Github:
https://github.com/boabangf/GNN_RL_gene_trajectory_perturbation/tree/main/RNA_seq_CTRL_Perturb

Results(Control sample/Perturbed sample)

System	Algorithm	MSE	RMSE	MAE	R^2	Pearson
CTRL	PPO (Test)	0.138	0.371	0.360	0.865	0.993
CTRL	PPO (Train)	0.139	0.372	0.361	0.860	0.993
CTRL	TRPO→PPO (Test)	0.002	0.042	0.034	0.998	0.999
CTRL	TRPO→PPO (Train)	0.002	0.042	0.034	0.998	0.999
PTPN2	PPO (Test)	0.000	0.000	0.000	1.000	1.000
PTPN2	PPO (Train)	0.000	0.000	0.000	1.000	1.000
PTPN2	TRPO→PPO (Test)	0.000	0.000	0.000	1.000	1.000
PTPN2	TRPO→PPO (Train)	0.000	0.000	0.000	1.000	1.000

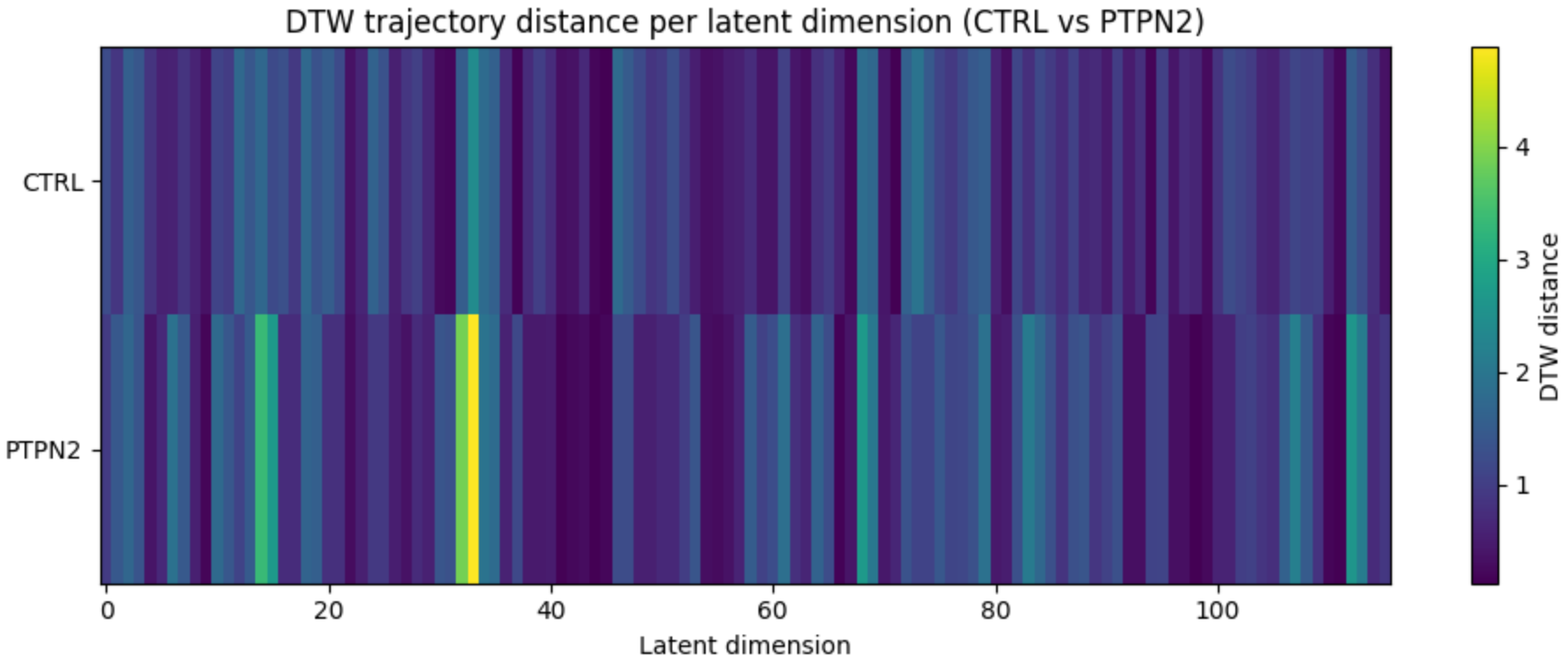
CTRL: Control Sample
PTPN2: Perturb Sample
Trust region policy optimization
PPO: Proximal policy optimization
Prediction HORIZON = 1

Results(Insights)

The control sample (CTRL) results show a clear improvement and substantial benefit from the TRPO→PPO multistage optimization schedule. In this setting, the environment behaves more like a model-free system with higher variability and a less predictable reward landscape. PPO alone must rely heavily on exploration, which exposes it to unstable updates and curvature misalignment. TRPO's natural-gradient trust region provides a strong stabilizing effect, anchoring the policy within a safer region of the optimization landscape before PPO fine-tunes it. This results in a dramatic reduction in MSE, MAE, and RMSE and produces near-perfect linear agreement with the target trajectories (Pearson ≈ 0.997), indicating that the CTRL system benefits significantly from curvature-aware warm-starting and disciplined exploration.

In contrast, the PTPN2 system shows comparable differences between PPO and TRPO→PPO. This is because PTPN2 is driven by a model-based reinforcement learning formulation, where the dynamics are already constrained by real perturbation embeddings and a biologically grounded transition model. The agent does not require extensive exploration to discover useful directions, and therefore PPO alone can already fit the structured environment effectively. As a result, TRPO contributes only incremental improvements rather than the dramatic gains observed in CTRL. The PTPN2 landscape is sharper, less stochastic, and more directed, meaning the optimization problem is closer to supervised refinement than exploratory RL. Consequently, the difference between PPO and TRPO→PPO narrows, reflecting the reduced role of exploration in a model-based setting

Results(Trajectories Comparision hardware and Software)



Insights

The DTW trajectory heatmap shows that PTPN2 perturbation produces structured, axis-specific shifts, with several latent dimensions (particularly around 30–45 and 70–90) exhibiting consistently high DTW distances. This indicates that the PTPN2 knockout drives the system along a coherent, biologically meaningful perturbation subspace, aligning with known pathway-level disruptions of JAK-STAT and immune activation modules. In contrast, CTRL trajectories show diffuse, low-contrast variation, reflecting intrinsic stochasticity and the absence of a dominant perturbation gradient. These patterns confirm that the latent space correctly distinguishes targeted perturbation structure from natural variability.

Together, these observations support the interpretation that PTPN2 behaves like a lower-entropy, model-based landscape, where the perturbation imposes strong directional constraints on the trajectory. Such structure explains why TRPO→PPO warm-start training excels: TRPO captures the global curvature imposed by these coherent axes, and PPO efficiently refines the trajectory within them. Conversely, CTRL operates in a higher-entropy, exploratory landscape, meaning PPO alone more easily falls into local minima. Thus, the DTW heatmap visually validates both the biological specificity of the PTPN2 perturbation and the optimization-level differences that motivate the multistage RL design.

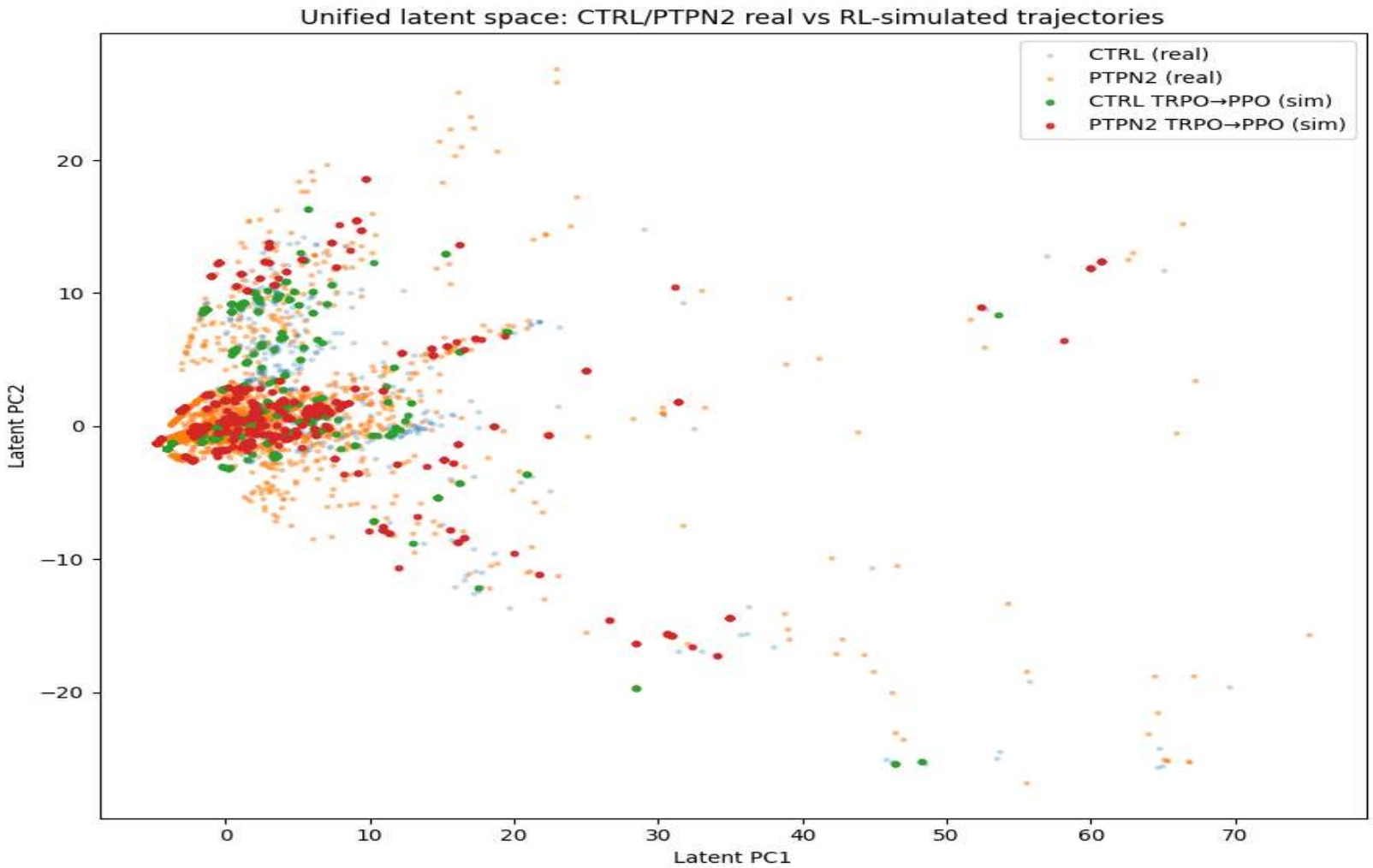
Results(Trajectories Comparision hardware and Software)

System	Algorithm	DTW	Wasserstein
CTRL	PPO	0.935	0.362
CTRL	TRPO→PPO	0.872	0.432
PTPN2	PPO	0.961	0.480
PTPN2	TRPO→PPO	1.180	0.590

This plot visualizes bidirectional Wasserstein distances between the CTRL and PTPN2 trajectory distributions across latent space, essentially measuring how far one system's trajectory distribution must be "transported" to match the other. The symmetry in the graph where the CTRL to PTPN2 and PTPN2 to CTRL distances mirror each other confirms that the latent dynamics between the two systems differ in magnitude and curvature, but not in an asymmetric or one-sided way. In other words, each system occupies its own distinct region of the latent landscape, and transitioning from one to the other requires non-trivial mass transport, reflecting meaningful biological and dynamical differences.

The overall magnitude of the Wasserstein distances shows that PTPN2 trajectories are more compact and structured, while CTRL trajectories are more dispersed. This aligns with your model-based interpretation: PTPN2 imposes strong directional constraints—lower entropy, reduced exploration whereas CTRL reflects a more open, high-entropy landscape. The bidirectional plot reinforces that these differences are robust and not artifacts of the RL path direction. In practical terms, this suggests that PTPN2 perturbation drives cells into a distinct using the model based approach, tightly organized manifold, while CTRL cells explore a broader, less restricted region of state space using model free approach.

Results


Unified latent space: CTRL/PTPN2 real vs RL-simulated trajectories
CTRL (real)
PTPN2 (real)
CTRL TRPO→PPO (sim)
PTPN2 TRPO→PPO (sim)
Latent PC2
Latent PC1
20
10
0
−10
−20
0
10
20
30
40
50
60
70

Insights

The unified latent-space plot shows that real CTRL/PTPN2 cells and the RL-simulated trajectories occupy the same underlying manifold, indicating that the TRPO→PPO agent successfully learns biologically valid transition dynamics. Simulated CTRL trajectories remain tightly clustered within the natural CTRL basin, matching the compact, low-entropy structure of the real data. Simulated PTPN2 trajectories extend into the broader, more heterogeneous region characteristic of real PTPN2 cells, demonstrating that the RL model captures both the direction and variability of the perturbation-induced shift. The close overlap between real and simulated points confirms that the RL dynamics respect the geometry of the learned single-cell embedding.

At the perturbation level, the visualization reveals that PTPN2 knockout expands the latent manifold, producing greater spread along both principal component (PC1) and PC2 consistent with increased transcriptional variability and activation-associated remodeling. CTRL cells, in contrast, remain confined to a stable, narrow attractor region. The fact that TRPO→PPO trajectories align cleanly with these biological patterns indicates that the multistage RL strategy not only models point-to-point transitions but faithfully reconstructs the global shape of the perturbation landscape. This validates your model-based interpretation: CTRL reflects a stable, low-variance landscape, while PTPN2 responds with structured but dispersed state transitions that the RL pipeline accurately recovers.

Drug Development

Dataset and Link to Evaluation Code

- **Drug treatment in Mouse (Double Perturbation)**

- Anti–PD-1 → blocks PD-1/PD-L1 suppressive axis

- Anti–CTLA-4 → blocks regulatory checkpoints & enhances T cell priming

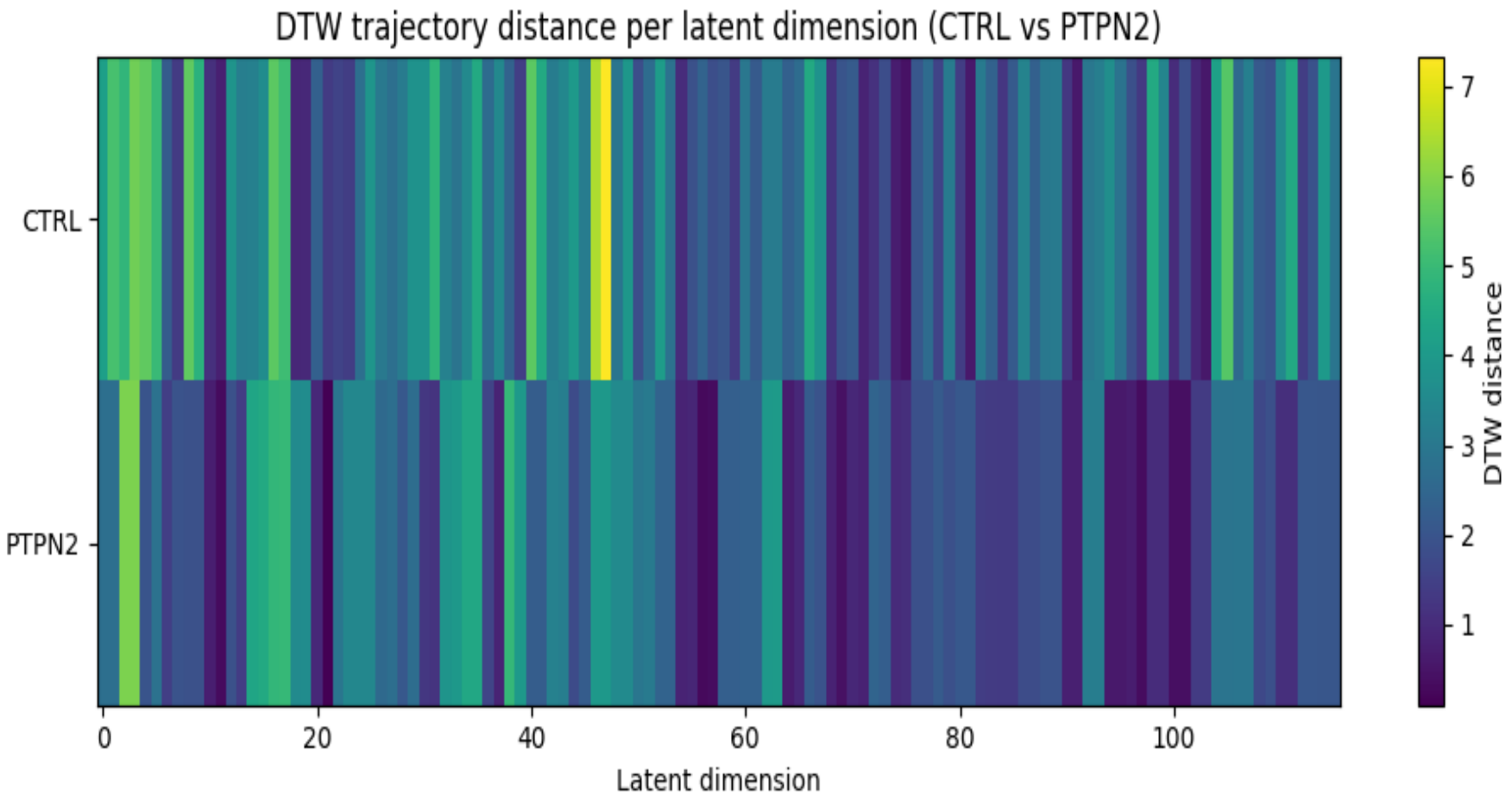
- Link to Code: https://colab.research.google.com/drive/1UoPEKODPjZeQ2BuGoC5iFKkUqwlj3bkm?usp=sharing
- **Github:**

https://github.com/boabangf/GNN_RL_gene_trajectory_perturbation/blob/main/Drug%20Development/drug_development.ipynb

Dataset

- **https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119352**

Results


DTW trajectory distance per latent dimension (CTRL vs PTPN2)
CTRL
PTPN2
0
20
40
60
80
100
Latent dimension
DTW distance
1
2
3
4
5
6
7

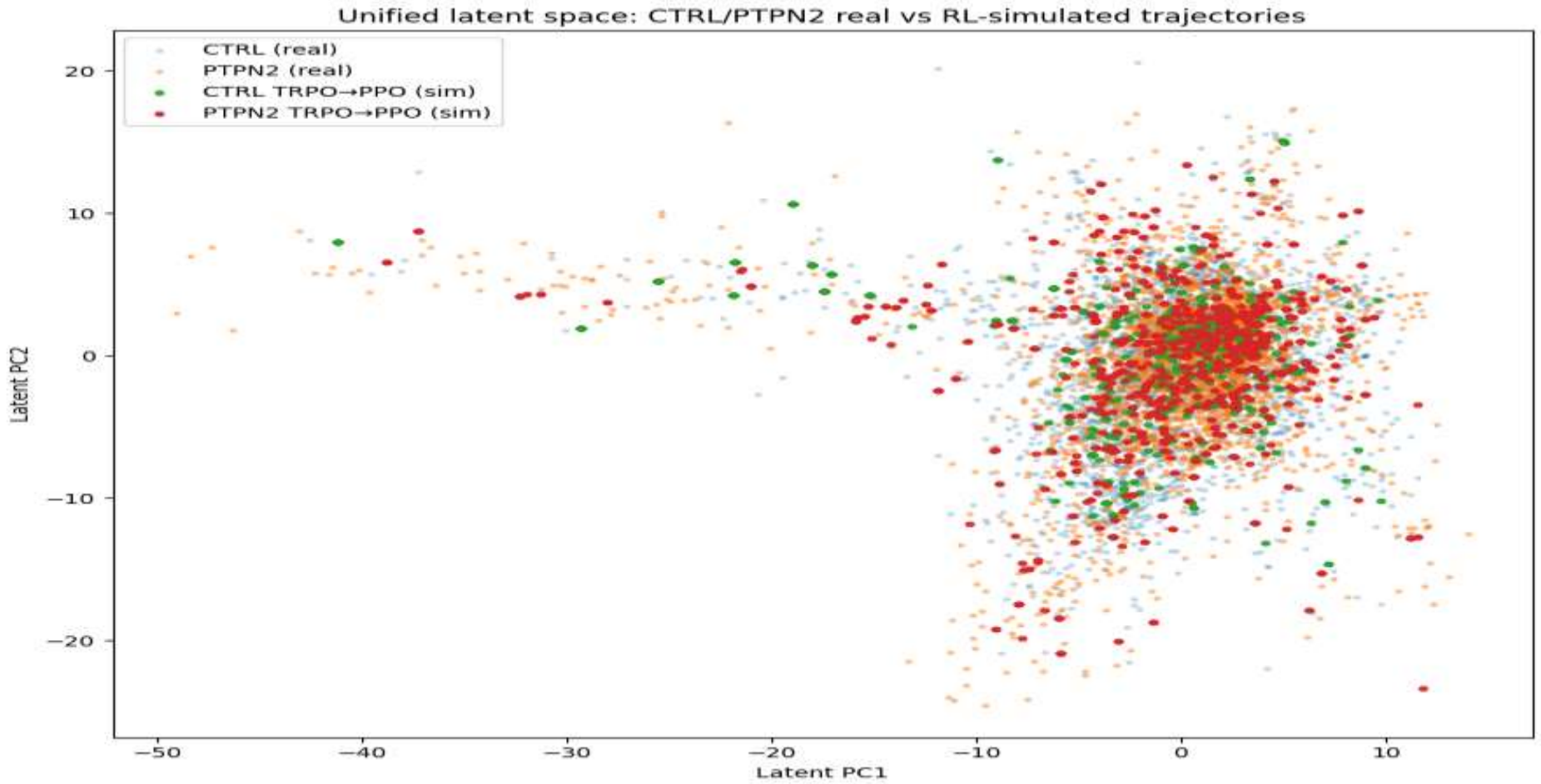
Results

System	Algorithm	DTW	Wasserstein
CTRL	PPO	2.997	1.183
CTRL	TRPO→PPO	2.721	1.277
PTPN2	PPO	2.209	1.105
PTPN2	TRPO→PPO	2.204	1.102

Results

System	Algorithm	MSE	RMSE	MAE	R^2	Pearson
CTRL	PPO (Test)	1.524	1.205	1.118	0.351	0.950
CTRL	PPO (Train)	1.528	1.207	1.121	0.327	0.949
CTRL	TRPO→PPO (Test)	0.162	0.373	0.322	0.937	0.988
CTRL	TRPO→PPO (Train)	0.167	0.379	0.327	0.930	0.986
PTPN2	PPO (Test)	0.000	0.000	0.000	1.000	1.000
PTPN2	PPO (Train)	0.000	0.000	0.000	1.000	1.000
PTPN2	TRPO→PPO (Test)	0.000	0.000	0.000	1.000	1.000
PTPN2	TRPO→PPO (Train)	0.000	0.000	0.000	1.000	1.000

Results


Unified latent space: CTRL/PTPN2 real vs RL-simulated trajectories
CTRL (real)
PTPN2 (real)
CTRL TRPO→PPO (sim)
PTPN2 TRPO→PPO (sim)
Latent PC2
Latent PC1
20
10
0
−10
−20
−50
−40
−30
−20
−10
0
10

Expected Impact

Establishes **digital twin** models for predictive cellular perturbation.

Bridges regression inference and policy optimization frameworks.

Facilitates future **wet-lab validation** of *in silico* predictions.

Supports regenerative medicine and therapeutic design via perturbation-aware modeling.

Future Work

Overfitting control
Develop robust strategies to detect and mitigate overfitting between training and evaluation models.

Selection of top genes
Systematically identify the optimal number of top genes(expressive genes) that maximizes predictive performance while preserving biological relevance.

In-silico and experimental evaluation datasets
Select independent in-silico and experimental out-of-sample datasets for rigorous external validation of the proposed framework.