Mathematical Modeling of a Bi-factor Stem Cell Differentiation System

Yuanchuan (Robert) Shao
Fangrui (Lori) Liu
Zhixiang (Carl) Yao
Youyuan Hu

Dec. 2023

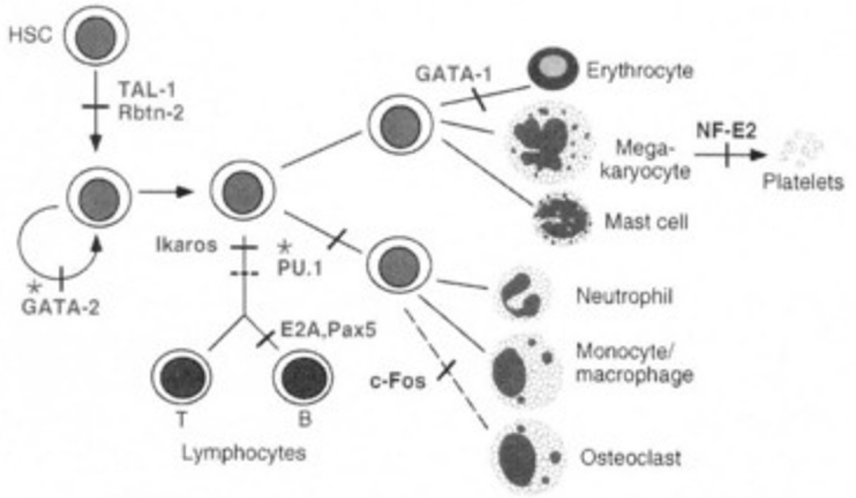
Duke

The main purpose

- **Model the hematopoietic stem cell differentiation process using PU.1 and GATA-1.**
- Elucidate the regulatory mechanisms governing differentiation into erythrocyte/megakaryocyte and granulocyte/macrophage lineages.
- Seek insights into the interplay between transcription factors PU.1 and GATA-1.
- Provide opportunities for controlling stem cell development, with significant therapeutic implications.
- Integrate machine learning and our mechanism-based biological model.

GATA-1 and PU.1 play important roles in HSC differentiation

- GATA-1: Drives differentiation into erythroid and megakaryocytic cells
- PU.1: Drives differentiation into myeloid and lymphoid
- Mutual antagonism: GATA-1 and PU.1 form a heterodimer that inhibits both genes' expression



(Duff et al., 2011)

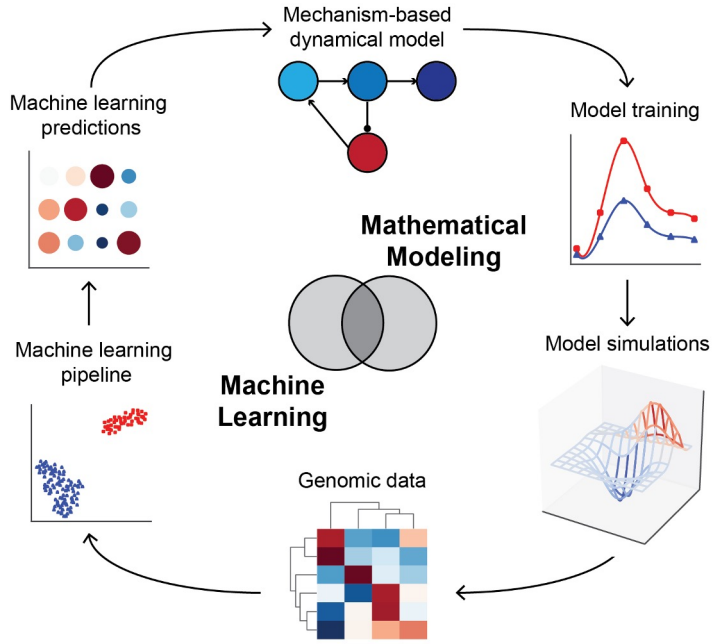
Shivdasani et al., 1996

Computational Challenges in Mechanism-Based Biological Models

- Mechanism-based models often require exploration of massive parametric spaces, demanding significant computational resources
- Generating time courses for dynamic models consisting of differential equations is a time-consuming process, particularly for models with numerous parameters
- Accurate modeling, especially in predicting complex biological behaviors like oscillations or spatial patterns, demands precision, further escalating computational needs

Biological models can be enhanced with Machine learning

- Efficiently handles large datasets and complex parameters
- Improves model prediction accuracy and efficiency
- Enables effective exploration of vast parameter spaces

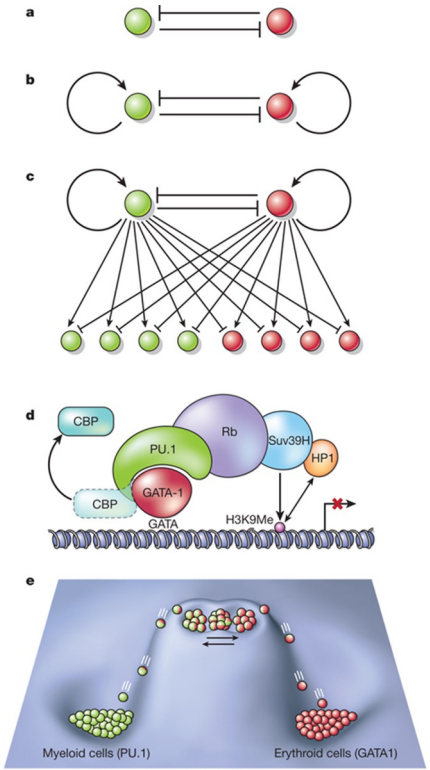


Kirouac et al., 2023

(Wang et al, 2019)

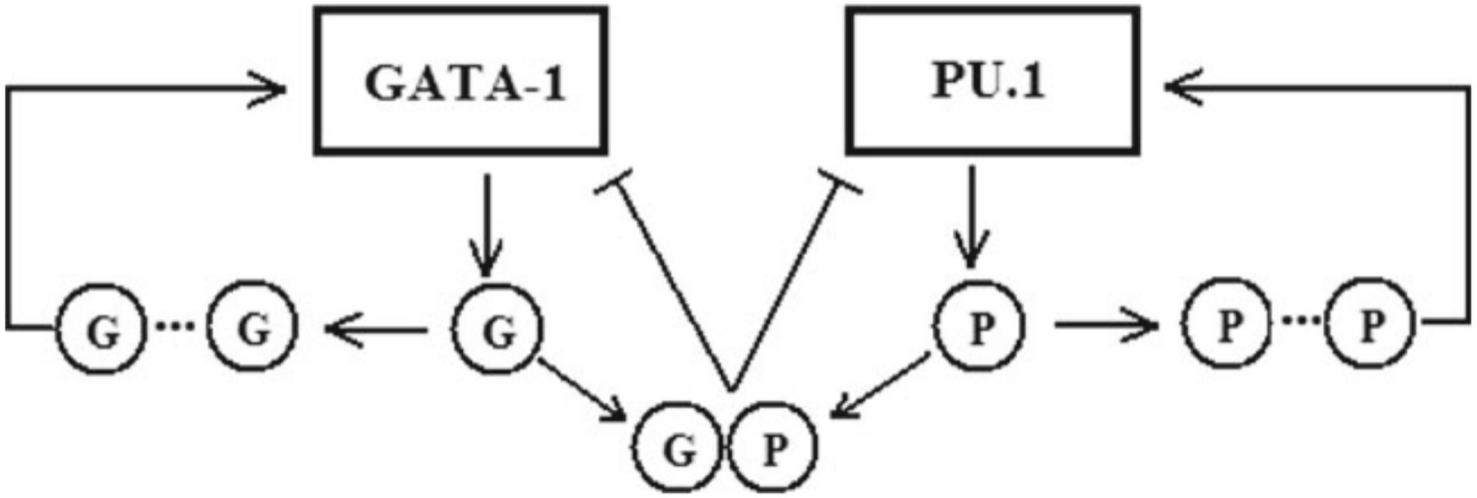
In General

In our study, we focus on hematopoietic stem cell differentiation, emphasizing the roles of GATA-1 and PU.1 in guiding blood lineage. We aim to improve mathematical models using deep learning, simplifying complexity and increasing predictive accuracy.



Graf et al., 2009

A mathematical model that describes the differentiation process



* 0th order degradation

Duff et al., 2011

A mathematical model that describes the differentiation process

$$\frac{d[G]}{dt} = a_1 \frac{[G]^n}{\theta_{a1}^n + [G]^n} + b_1 \frac{\theta_{b1}^m}{\theta_{b1}^m + [G]^m[P]^m} - k_1[G]$$

$$\frac{d[P]}{dt} = a_2 \frac{[P]^n}{\theta_{a2}^n + [P]^n} + b_2 \frac{\theta_{b2}^m}{\theta_{b2}^m + [G]^m[P]^m} - k_2[P],$$

Autoregulation | Cross inhibition | Degradation

Duff et al., 2011

Non-dimensionalization simplification reduces parameter size

- $a = a_1 = a_2$
- $b = b_1 = b_2$
- $k = k_1 = k_2$
- $\theta_a = \theta_{a_1} = \theta_{a_2}$
- $\theta_b = \theta_{b_1} = \theta_{b_2}$

Under these assumptions, the ODEs are simplified to:

1. $\frac{d[G]}{dt} = a\frac{[G]^n}{\theta_a^n + [G]^n} + b\frac{\theta_b^m}{\theta_b^m + [G]^m[P]^m} - k[G]$
2. $\frac{d[P]}{dt} = a\frac{[P]^n}{\theta_a^n + [P]^n} + b\frac{\theta_b^m}{\theta_b^m + [G]^m[P]^m} - k[P]$

Non-dimensionalization simplification reduces parameter size

Units of Parameters

The units for the parameters in the ODEs are described as follows:

$[G], [P], \theta_a$: Units for concentrations

θ_b: Squared units for concentrations

a, b: Concentration divided by time

k: Inverse of time

m, n: Dimensionless (Hill coefficients)

Non-dimensionalization simplification reduces parameter size

- $X = \frac{[G]}{\theta_a}$
- $Y = \frac{[P]}{\theta_a}$
- $\theta = \frac{\theta_a^2}{\theta_b}$
- $\tau = kt$

$$\alpha = \frac{a}{k\theta_a}$$

$$\beta = \frac{b}{k\theta_a}$$

$$\frac{dX}{d\tau} = \alpha \frac{X^n}{1 + X^n} + \beta \frac{1}{1 + (\theta XY)^m} - X$$

$$\frac{dY}{d\tau} = \alpha \frac{Y^n}{1 + Y^n} + \beta \frac{1}{1 + (\theta XY)^m} - Y$$

Non-dimensionalization simplification reduces parameter size

- 12 parameters → 7 parameters → 5 parameters
- Assuming symmetry as a primary observation of the differentiation system
- Non-dimensionalization frees up computational burden

$$\frac{dX}{d\tau} = \alpha \frac{X^n}{1 + X^n} + \beta \frac{1}{1 + (\theta XY)^m} - X$$

$$\frac{dY}{d\tau} = \alpha \frac{Y^n}{1 + Y^n} + \beta \frac{1}{1 + (\theta XY)^m} - Y$$

Data Generation

1. Use a loop to go through all combinations of parameters (a, b, m, n, k, θ_a, θ_b).
2. Generate a grid of (G, P) pairs (5 by 5) as initial conditions for each parameter combination.
3. Add random perturbations to the integral every several time steps to avoid unstable steady states (saddle points).
4. Remove repeating steady states and unreasonable data points.
5. Non-dimensionalize all the data points.
6. Perform data augmentation (generate more data points for cases with multiple steady states).

Data Summary

Number of Steady State	1	2	3	4	Total
Data Size	79,390	16,221	22,839	12,507	130,957

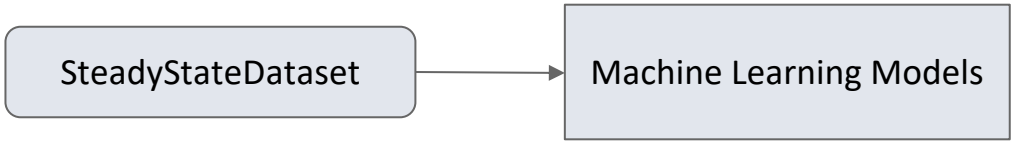
Data Normalization

- Center the data around zero:
  - Subtract the mean value of the training dataset.
- Bring the data to a comparable scale:
  - Scale the data by dividing it through the standard deviation of the training data.

Datasets

- **SteadyStateDataset:**
  - Independent Variables: 5 parameters
  - Dependent Variables: the associated count of steady states
- **DistributionDatasets**:
  - Independent Variables: 5 parameters
  - Dependent Variables: the non-dimensionalized distributions of G and P.
  - DistributionDataset_1, DistributionDataset_2, DistributionDataset_3, DistributionDataset_4

Machine Learning Classifier Models



- To predict the number of steady state based on the 5 parameters
- Decision Tree
- Random Forest

Results – ML Classifier Models

Accuracy: 0.888515577275504

	precision	recall	f1-score	support
1	0.96	0.96	0.96	7922
2	0.72	0.73	0.72	1662
3	0.81	0.81	0.81	2276
4	0.82	0.81	0.81	1236
accuracy			0.89	13096
macro avg	0.83	0.83	0.83	13096
weighted avg	0.89	0.89	0.89	13096

Decision Tree

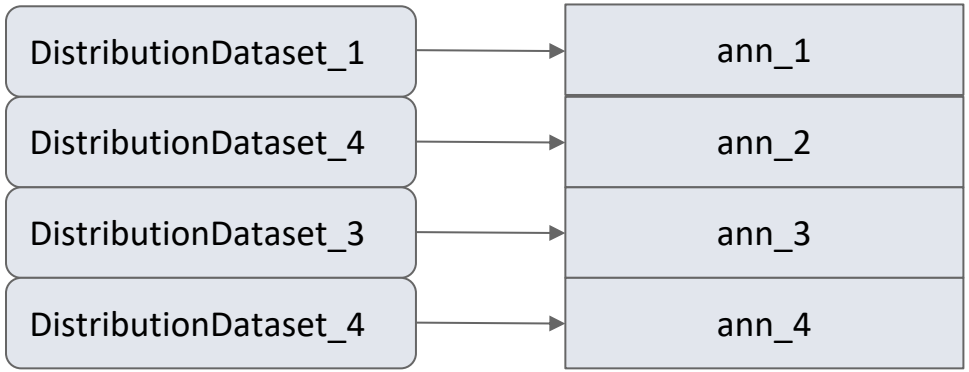
Accuracy: 0.9085980452046426

	precision	recall	f1-score	support
1	0.97	0.96	0.97	7922
2	0.77	0.78	0.78	1662
3	0.84	0.85	0.84	2276
4	0.86	0.84	0.85	1236
accuracy			0.91	13096
macro avg	0.86	0.86	0.86	13096
weighted avg	0.91	0.91	0.91	13096

Random Forest

- Both models perform well
- The Random Forest seems to outperform the Decision Tree model

Neural Network



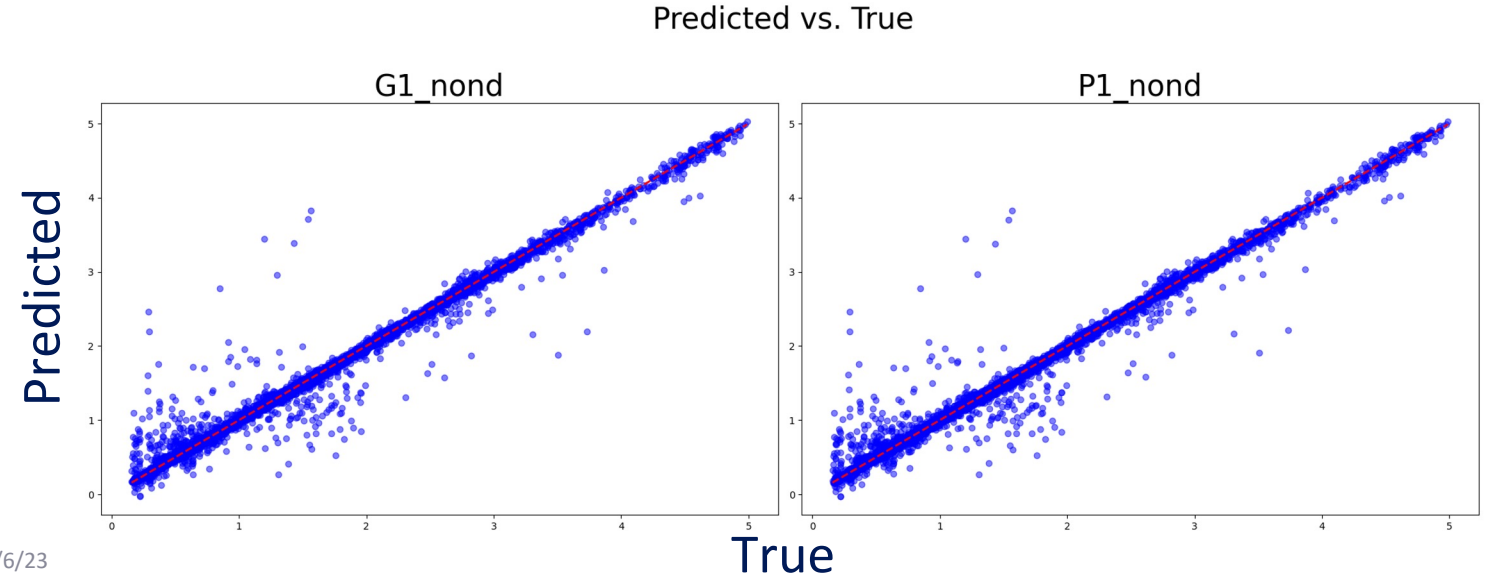
- **Input Layer**
- **Hidden Layers**
  - ***Hidden Layer 1***: 64 neurons with ReLU activation function
  - ***Hidden Layer 2***: 32 neurons with ReLU activation function
  - ***Hidden Layer 3***: 16 neurons with ReLU activation function
- **Output Layer**

- Feed DistributionDatasets from different steady state status into the respective neural network model for training

Results – Neural Network Models

- DistributionDataset_1: 1 steady state

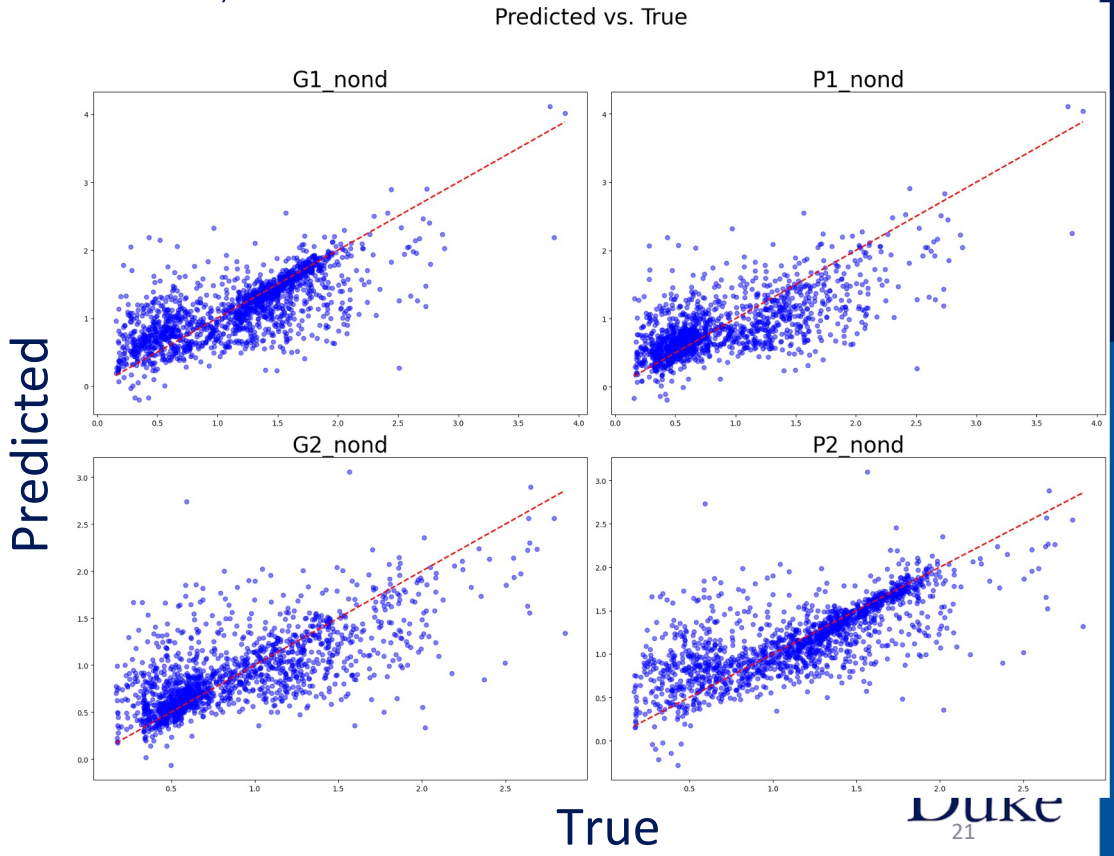
	MSE	MAE	R2 Score
Training Set	0.0167	0.0492	0.983
Testing Set	0.0231	0.0525	0.977



Results – Neural Network Models

- DistributionDataset_2: 2 steady states

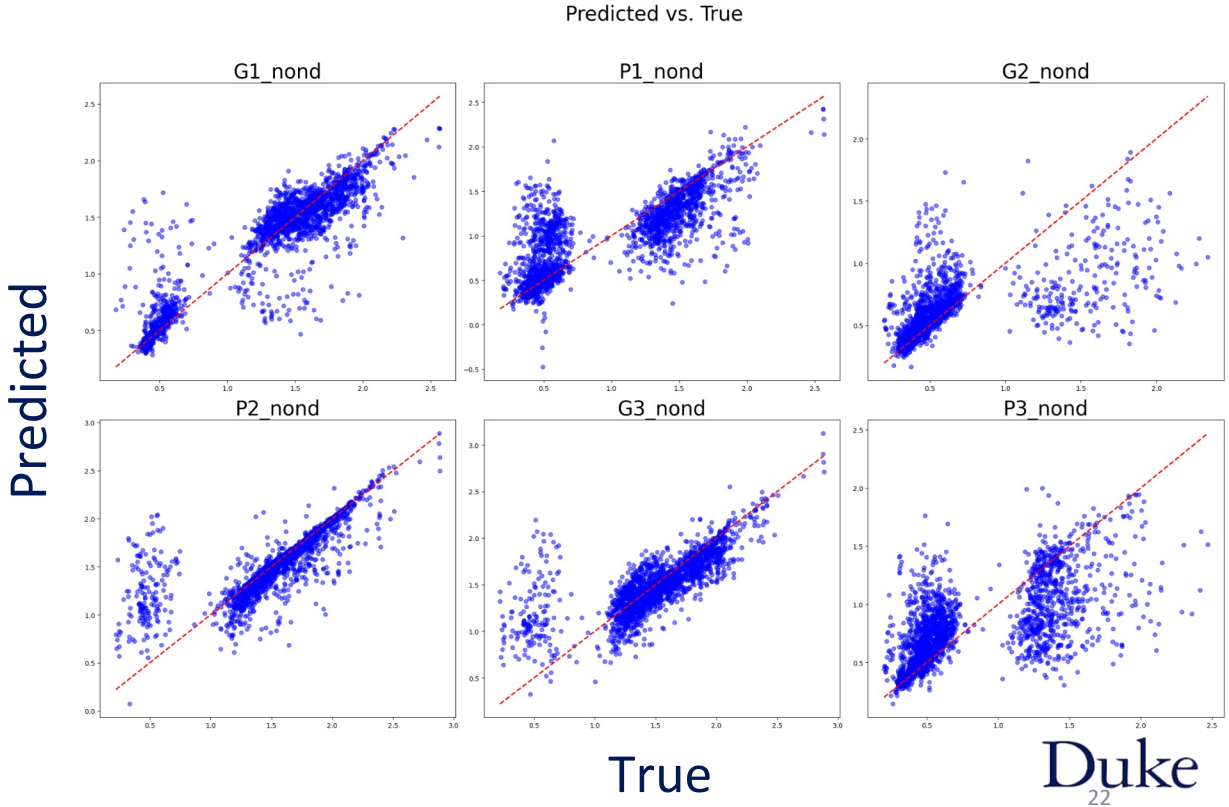
	Training Set	Testing Set
MSE	0.0890	0.143
MAE	0.0202	0.248
R2 Score	0.653	0.466



Results – Neural Network Models

- DistributionDataset_3: 3 steady states

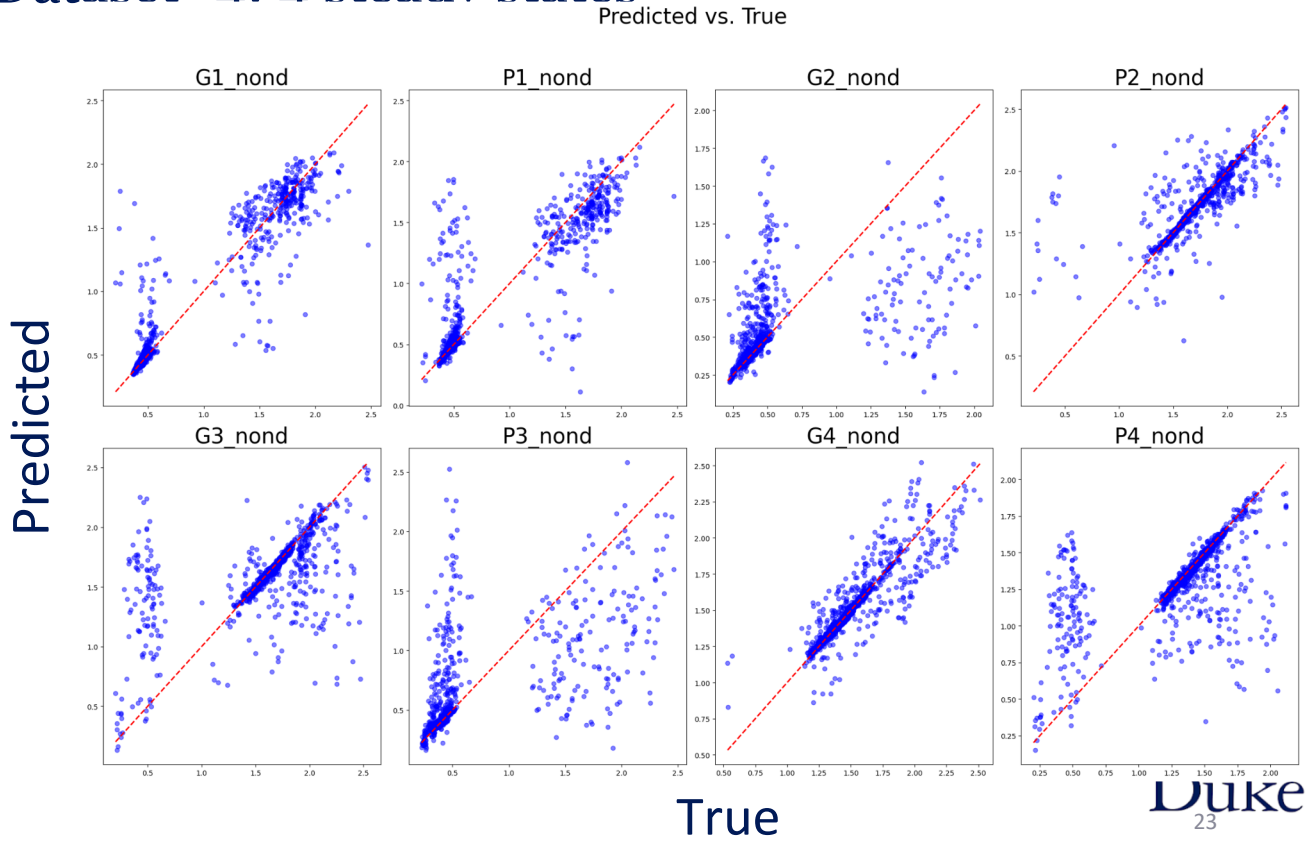
	Training Set	Testing Set
MSE	0.0744	0.083
MAE	0.166	0.174
R2 Score	0.606	0.566



Results – Neural Network Models

- DistributionDataset 4: 4 steady states

	Training Set	Testing Set
MSE	0.0487	0.0641
MAE	0.0944	0.109
R2 Score	0.669	0.579



Future work

- Reorder the steady states to maintain the continuity of all the entries in the output vector.
- Perform a classification on distribution datasets based on the symmetry of the location of these steady states.
- Combine machine learning classifier model with simplified ann model to prevent overfitting on training datasets.

Questions?

Reference

1. Shivdasani, R., & Orkin, S. (1996). The transcriptional control of hematopoiesis [see comments]. Blood, 87(10), 4025–4039. https://doi.org/10.1182/blood.v87.10.4025.bloodjournal87104025
2. Duff, C., Smith-Miles, K., Lopes, L., & Tian, T. (2011). Mathematical modelling of stem cell differentiation: The pu.1–GATA-1 interaction. Journal of Mathematical Biology, 64(3), 449–468. https://doi.org/10.1007/s00285-011-0419-3
3. Wang, S., Fan, K., Luo, N., Cao, Y., Wu, F., Zhang, C., Heller, K. A., & You, L. (2019). Massive computational acceleration by using neural networks to emulate mechanism-based biological models. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-12342-y
4. Kirouac, D. C., Zmurchok, C., Deyati, A., Sicherman, J., Bond, C., & Zandstra, P. W. (2023). Deconvolution of clinical variance in car-T cell pharmacology and response. Nature Biotechnology, 41(11), 1606–1617. https://doi.org/10.1038/s41587-023-01687-x
5. Graf, T., & Enver, T. (2009). Forcing cells to change lineages. Nature, 462(7273), 587–594. https://doi.org/10.1038/nature08533