Prior-informed NeuralODEs to discover sparse regulatory dynamics from temporal gene expression data

Intekhab Hossain

Harvard T.H. Chan School of Public Health

(Department of Biostatistics)

September 13, 2022

Big picture: the problem

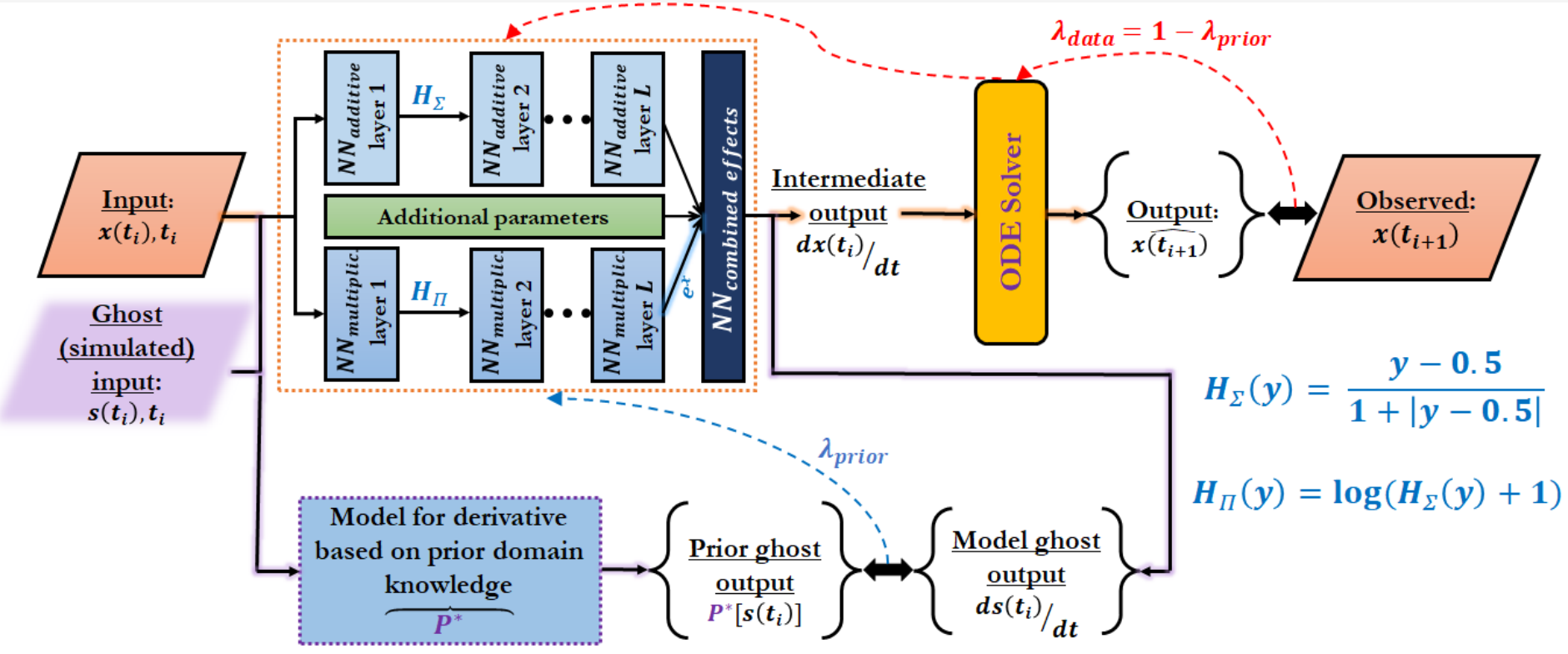
- ➜ We are interested in **gene expression dynamics**
- ☆ Utility:
  1. understanding the nature of biological systems
  2. predicting responses to interventions

Big picture: the problem

- ➜ We are interested in **gene expression dynamics**
- ☆ Utility:
  1. understanding the nature of biological systems
  2. predicting responses to interventions
- ➜ Formal problem:
  - given:
    - $g_1(t = t_0), g_1(t = t_1), \ldots, g_1(t = t_T)$
    - $g_2(t = t_0), g_2(t = t_1), \ldots, g_2(t = t_T)$
    - ……
    - $g_n(t = t_0), g_n(t = t_1), \ldots, g_n(t = t_T)$
  - estimate **dynamics functions** (i.e. ODEs) $f_1, f_2, \ldots, f_n$, where:
    - $dg_1/dt = f_1(g_1, g_2, \ldots, g_n, t)$
    - $dg_2/dt = f_2(g_1, g_2, \ldots, g_n, t)$
    - ……
    - $dg_n/dt = f_n(g_1, g_2, \ldots, g_n, t)$

PHOENIX architecture

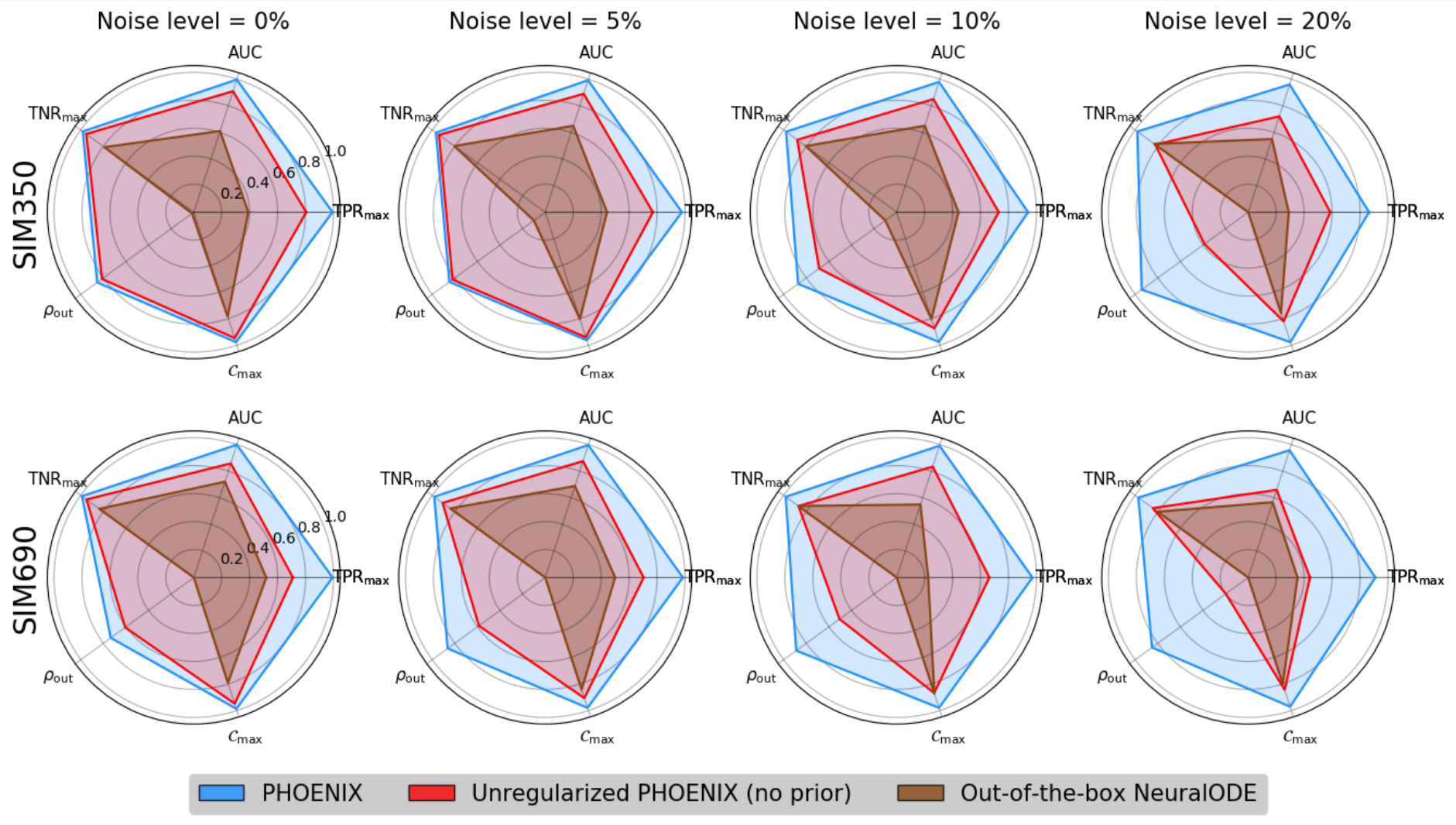
➜ biologically motivated (Hill-like activation functions)



➜ biologically explainable (prior domain knowledge)

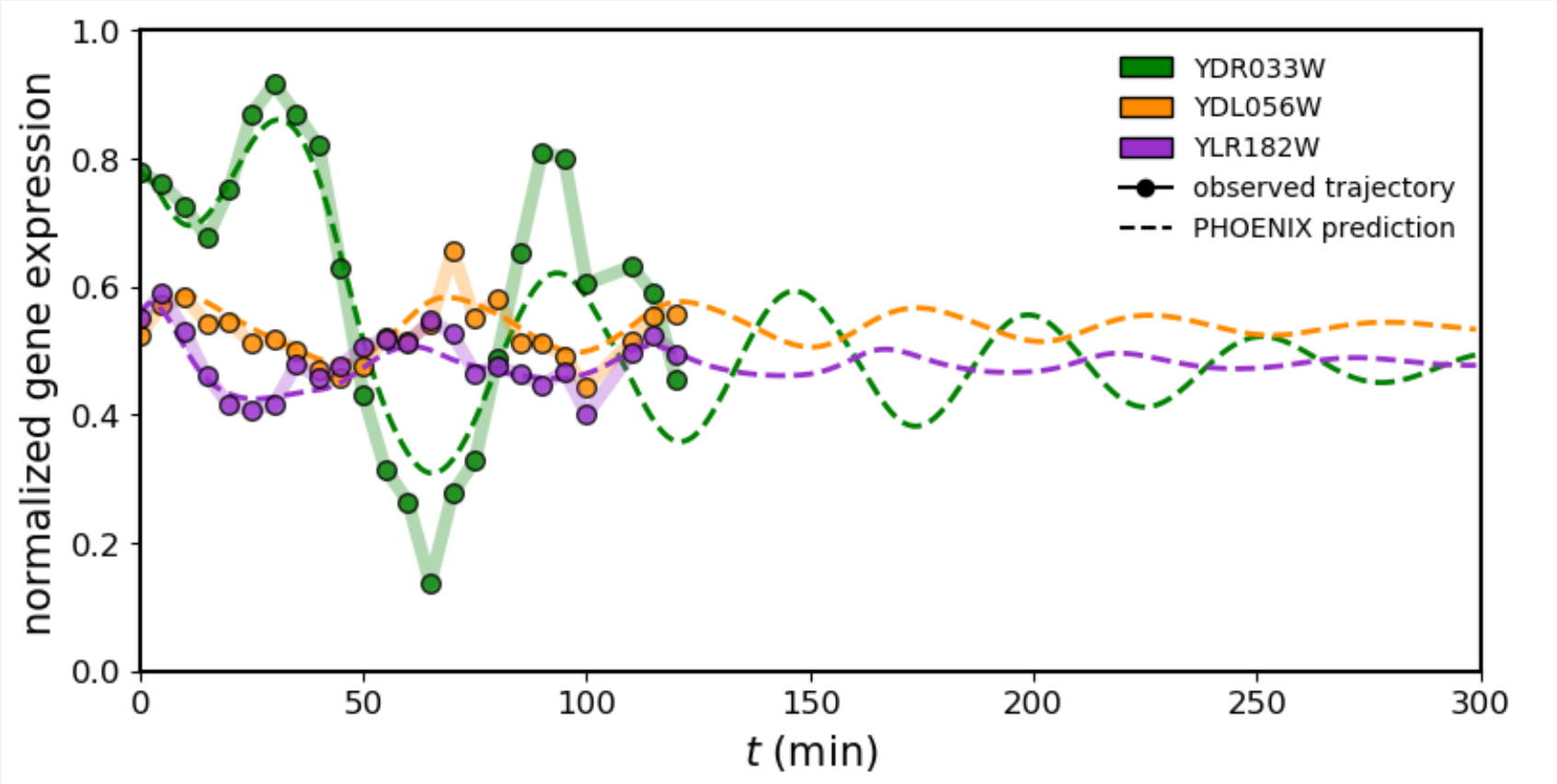
☆ **PHOENIX** = **P**rior-informed **H**ill-like **O**DEs to **E**stimate **N**etwork **I**ntegrals with e**X**plainability

PHOENIX outperforms OOTB models on synthetic data



PHOENIX flexibly estimates yeast cell cycle dynamics

- ➜ Microarray expression for 3551 genes from synchronized yeast cells
- ➜ Prior model based on motif map of promoter targets
- ➜ Explainability validation = ChipSeq data



- ☆ Validation $R^2 = 85\%$; AUC (explainability) $= 0.86$

PHOENIX scales to human breast cancer dynamics

- ➜ Micro-array expression for 11151 genes from breast cancer cells across 186 patients ordered in psuedotime
- ➜ Prior model: used motif map of promoter targets; validation: ChIP-seq data from the MCF7 (breast cancer) cell line in ReMap2018

Number of genes	Val. set R^2	$AUC_{\widehat{GRN}}$	Runtime(AWS \$)
500	99%	0.91	0.06
2000	98%	0.91	0.16
4000	97%	0.86	0.28
11165	97%	0.81	1.63

- ☆ Scaling allows discovery of candidate genes for novel drivers of breast cancer progression

Thank you! (poster #40)

Prior-informed NeuralODEs to Estimate Gene-Regulatory Dynamics

Intekhab Hossain[1], Rebekka Burkholz[2] & John Quackenbush[1,3]

[1] Harvard University dept. Biostatistics (Boston, MA), [2] CISPA - Helmholtz Center for Information Security (Saarbrücken, Germany), [3] Dana Farber Cancer Institute (Boston, MA)

Objectives

➜ Ordinary differential equation (ODE) models that can accurately and causally explain temporal gene expression patterns are crucial for:
1. understanding the nature of biological systems
2. predicting response to interventions

➜ Starting from time-series gene-expression data, we aim to estimate ODEs that describe how expression evolves in time

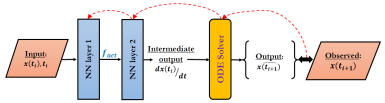
$$\vec{\boldsymbol{x}}(t_0), \boldsymbol{x}(t_1), \vec{\boldsymbol{x}}(t_2), \ldots, \vec{\boldsymbol{x}}(t_T) \xrightarrow{est.} \frac{d\vec{\boldsymbol{x}}}{dt}$$

➜ Current estimation methods suffer from multiple issues:
1. Not **predictive**: simple models that perform poorly out-of-sample [1]
2. Not biologically **explainable**: complex black-box methods that optimize predictions, but not necessarily learn biologically meaningful and sparse representations of the causal dynamics [2, 3]
3. Not **flexible**: parametric models with restrictive functional forms that are not able to flexibly represent arbitrary dynamics across both continuous and discrete time scales [4]
4. Not **scalable**: expensive models that only work on small sets of genes, or a small number of UMAP/PCA projections of genes [5, 6]

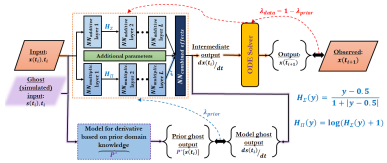
★ We solved these issues concurrently, by employing a modern ML framework regularized by biological domain knowledge, to estimate dynamics that are sparse, explainable, and scalable

Model development

➜ Out-of-the-box (OOTB) NeuralODE [7] architecture:



★ 1. Incorporated Hill-like activations from systems biology [8] to model additive (H_Σ) and multiplicative (H_Π) co-activation/co-repression
2. Explicitly optimized explainability and induced sparsity through user-specified prior model based on expert domain knowledge

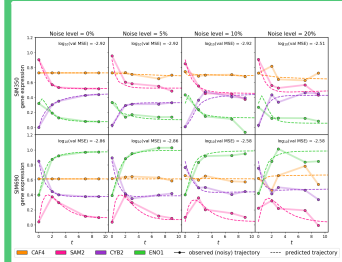


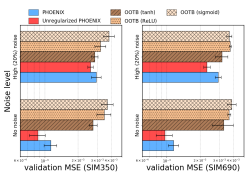
★ **PHOENIX** = **P**rior-informed **H**ill-like **O**DEs to **E**stimate **N**etwork **I**ntegrals with e**X**plainability

Supplemental information

← References & acknowledgements
← Details about figures & results
← Contact information
Tools from our group (`netZoo`) →

#1 Predictive: noisy dynamics from simulated GRN





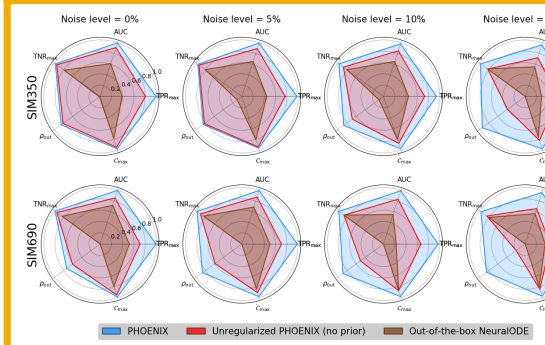
➜ Tested PHOENIX on two simulated [9] systems ($n = 350, 690$ genes) of 150 trajectories each, with varying noise

➜ k-fold cross validation to compare against prior-less PHOENIX and OOTB models

★ PHOENIX is robust and generally outperforms OOTB models, across all noise-settings

★ Prior-less PHOENIX seems to be most predictive, but not as explainable (⟶)

#2 Explainable: recovery of sparse causal biology



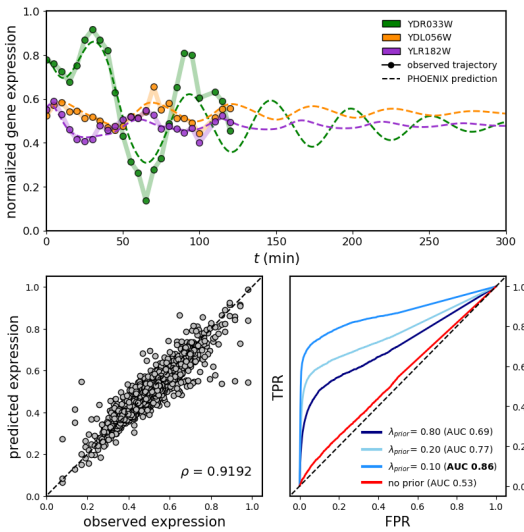
➜ Extracted encoded $\widehat{GRN}$ from trained models, and compared to ground truth GRN in simulation

➜ Explainability metrics: edge-recovery (AUC, TPR, TNR); degree-alignment (ρ_{out}); sparsity ($\mathcal{C}_{\max}$)

★ PHOENIX models heavily outperform OOTB models across all explainability metrics.

★ Prior domain knowledge makes PHOENIX robustly explainable, and induces sparse dynamics.

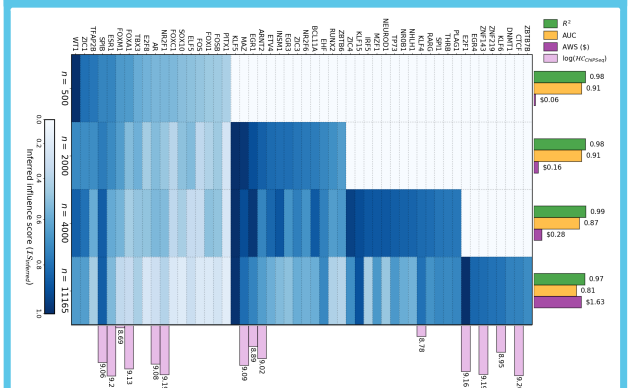
#3 Flexible: oscillating yeast cell cycle dynamics



➜ Applied PHOENIX to microarray trajectories of $n = 3551$ genes from synchronized yeast cells [10] during cell cycle. Used a motif-based prior model and validated explainability using ChipSeq [11]

★ PHOENIX is a universal approximator; it flexibly deviates from Hill-like behavior, to accurately capture oscillating dynamics that dampen over time (as yeast cells fall out of synchronization)

#4 Scalable: large-scale breast cancer dynamics



➜ Applied PHOENIX to microarrays of $n = 500, 2000, 4000, 11165$ genes from 186 breast cancer (BC) samples [12] ordered in psuedotime [13]; motif-based prior model and ChipSeq validation [14]

➜ Performed perturbation analysis on trained models to compute "influence scores' for each gene

★ PHOENIX scales to human-genome-like magnitudes, while remaining predictive and explainable

★ This allows feasible and accurate prediction of the downstream effects of modulating any gene

★ PHOENIX validates known drivers of BC progression and can potentially discover novel drivers