Epigenetically Informed Control of Gene Regulatory Networks

From Network Inference to Control Design

Aryaman Bahl Autrio Das
Team Twin - 9

IIIT Hyderabad
Dynamical Processes & Complex Networks

Problem Statement

Research Question

How can we construct a **control-ready** gene regulatory network (GRN) that integrates **epigenetic information** for therapeutic intervention design in breast cancer?

Biological Context:

1. Genes regulate each other through transcription factors (TFs)
2. Cancer = dysregulated gene expression
3. Network rewiring through mutations & epigenetics
4. Control objective: restore healthy regulation

Control Theory Prespective:

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$$

- $x \in \mathbb{R}^n$: gene expression states
- $A \in \mathbb{R}^{n \times n}$: regulatory network (to infer)
- $B \in \mathbb{R}^{n \times m}$: control inputs (to design)
- $u \in \mathbb{R}^m$: therapeutic interventions

Dataset

We use data from the METABRIC breast cancer cohort, focusing on tumor samples for which both DNA methylation and gene expression measurements are available.

Data Type	Dimension	Coverage
Gene Expression (RNA-seq)	1,417 samples $\times$ 11,171 genes	100%
DNA Methylation (450K array)	1,417 samples $\times$ 11,171 genes	100%
After Preprocessing	**8,378 genes**	**High-quality**

Preprocessing Steps:

1. Median imputation for missing values
2. Removal of zero-variance genes
3. Filtering of low-variance genes (bottom 25%)
4. Result: 8,378 high-quality genes for analysis

1 Methodology

Correlation

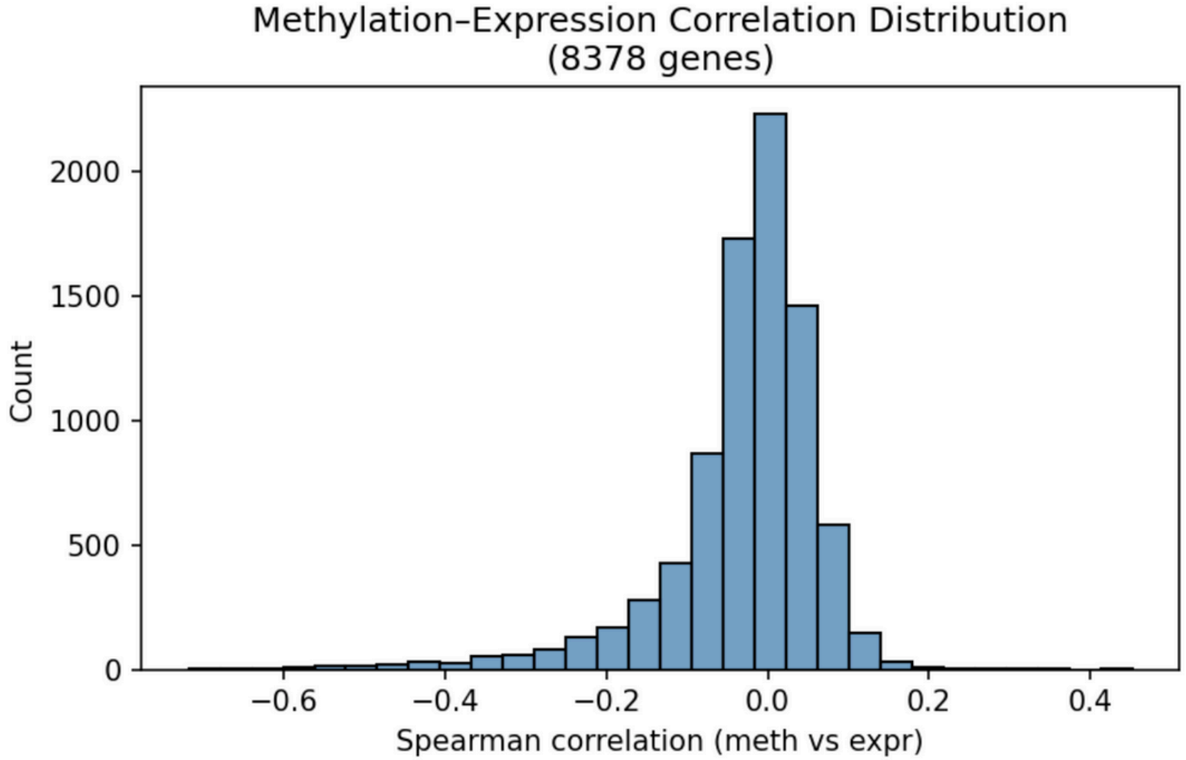
Objective: Quantify epigenetic regulation

Method:

- Compute Spearman correlation between methylation and expression
- For each gene: $\rho_i = \text{corr}(\text{meth}_i, \text{expr}_i)$
- $\rho_i < 0$: methylation silences gene
- $\rho_i > 0$: methylation activates gene

Result:

- 100% coverage (8,378/8,378 genes)
- Mean correlation: -0.029
- Distribution: broad, bimodal



GRN Inference with ElasticNet

✓ Signed weights (activation/repression)

ElasticNet combines L1 (Lasso) sparsity with L2 (Ridge) stability

Algorithm:

1: Select 500 most variable genes as TFs
2: **for** each target gene i **do**
3: Fit ElasticNet:
$\text{expr}_i \sim \sum_{j \in \text{TFs}} w_{ij} \cdot \text{expr}_j$
4: Extract non-zero coefficients w_{ij}
5: Keep top-3 regulators by $|w_{ij}|$
6: **end for**
7: Return edge list: $(TF_j \rightarrow \text{gene}_i, w_{ij})$

Parameters:

- $\alpha = 0.01$ (regularization)
- l_1 ratio $= 0.9$ (90% Lasso, 10% Ridge)
- TOP_K $= 3$ per target

Output:

Metric	Value
Total edges	25,134
Positive (activation)	18,879 (75%)
Negative (repression)	6,255 (25%)
Runtime	∼15-20 min

Adjacency Matrix Construction

Construction:

$$\mathbf{A}_{expr}[i,j] = \begin{cases} w_{ij} & \text{if } TF_j \rightarrow \text{gene}_i \\ 0 & \text{otherwise} \end{cases}$$

Property	Value
Shape	8,378 × 8,378
Non-zeros	25,134
Density	0.036%
Sparsity	99.964%
Mean degree	6.0
Max out-degree	353

Matrix Statistics

Properties:

- Dimension: 8,378 × 8,378
- Non-zeros: 25,134 (0.036% dense)
- Format: CSR sparse matrix
- Weight range: $[-0.56, 0.99]$
- Signed: positive = activation, negative = repression

Validation:

- No duplicate edges
- No self-loops (diagonal = 0)
- All genes represented

Methylation Integration

Objective

Modulate regulatory edges based on epigenetic state

Mathematical Formulation:

$$\text{scale}_i = \begin{cases} 1 - |\rho_i| & \text{if } \rho_i < 0 \quad \text{(methylation silences)} \\ 1 + |\rho_i| & \text{if } \rho_i \geq 0 \quad \text{(methylation activates)} \end{cases}$$

$$\mathbf{A}_{final} = \text{diag}(\text{scale}) \cdot \mathbf{A}_{expr}$$

Interpretation:

- scale < 1: weaken incoming edges
- scale $= 1$: no epigenetic effect
- scale > 1: strengthen incoming edges

Example:

- TP53: $\rho = -0.5 \Rightarrow$ scale $= 0.5$ (suppressed)
- MYC: $\rho = +0.8 \Rightarrow$ scale $= 1.8$ (enhanced)

Results:

Category	Count (%)
Suppressed	4,842 (57.8%)
Enhanced	3,536 (42.2%)
Mean scale	0.971

5 Gershgorin Stabilization

Challenge

Unstabilized matrix may have positive eigenvalues
$\Rightarrow$ unstable dynamics

Before vs After:

Metric	Before	After
Spectral abscissa	+0.72	-1.26
Status	Unstable	Stable
All $\lambda < 0$	No	Yes

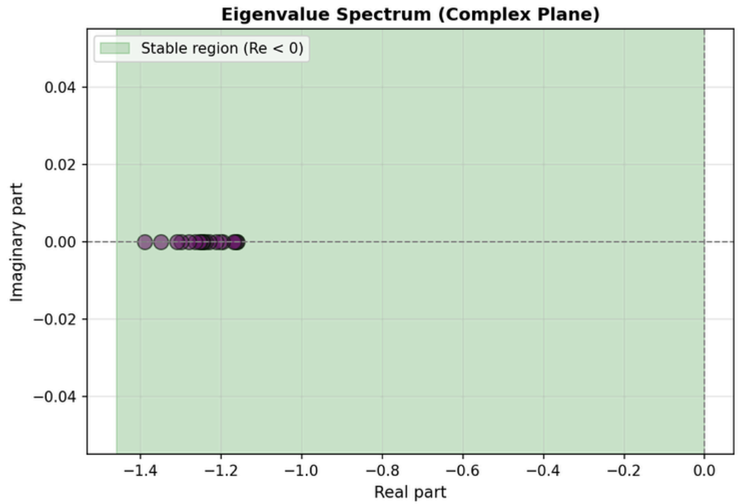
Gershgorin Circle Theorem: Every eigenvalue λ lies in at least one disc:

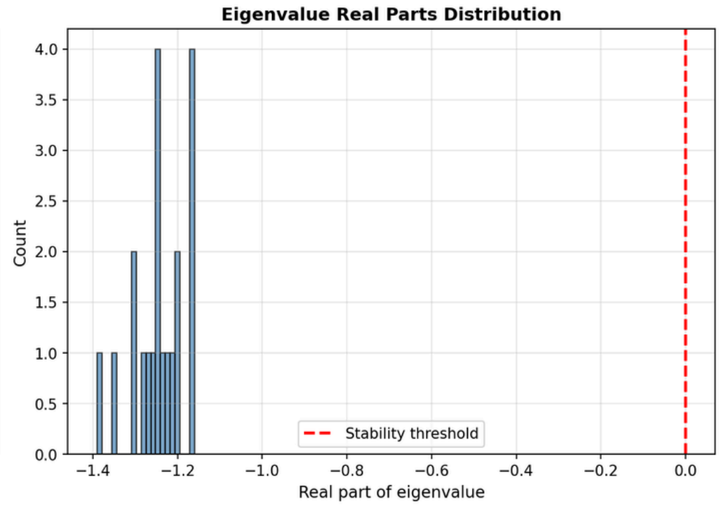
$$|\lambda - \mathbf{A}[i,i]| \leq r_i$$

where $r_i = \sum_{j \neq i} |\mathbf{A}[i,j]|$ (off-diagonal sum)

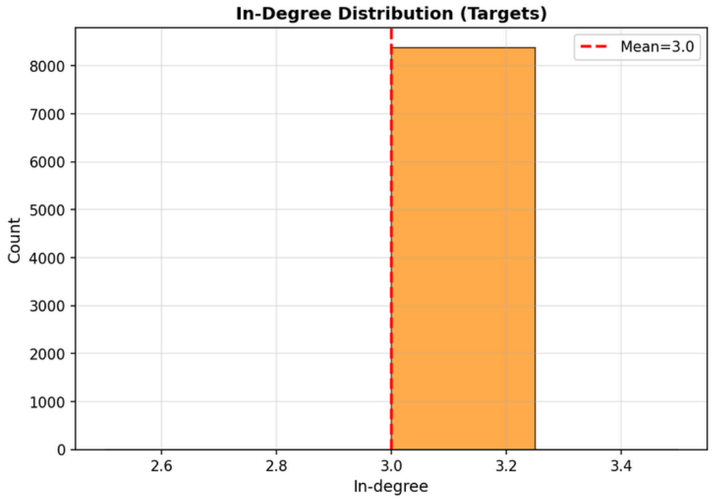
Stabilization Algorithm:

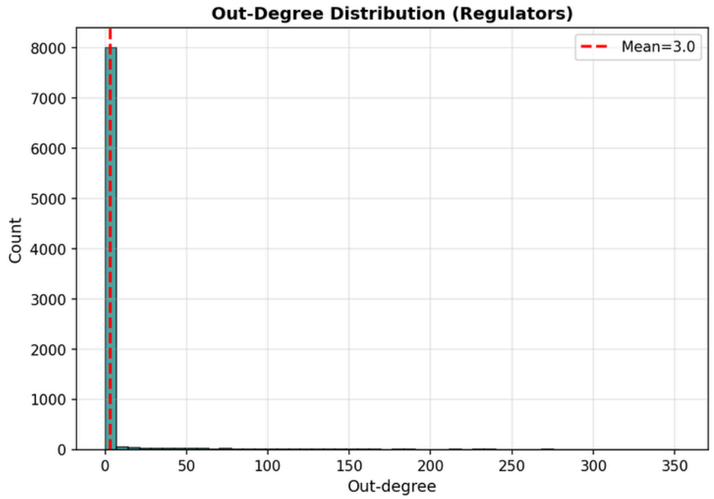
1. For each gene i: compute r_i
2. Set diagonal: $\mathbf{A}[i,i] = -(r_i + \text{margin})$
3. Guarantees: $\text{Re}(\lambda) < -\text{margin}$ for all λ

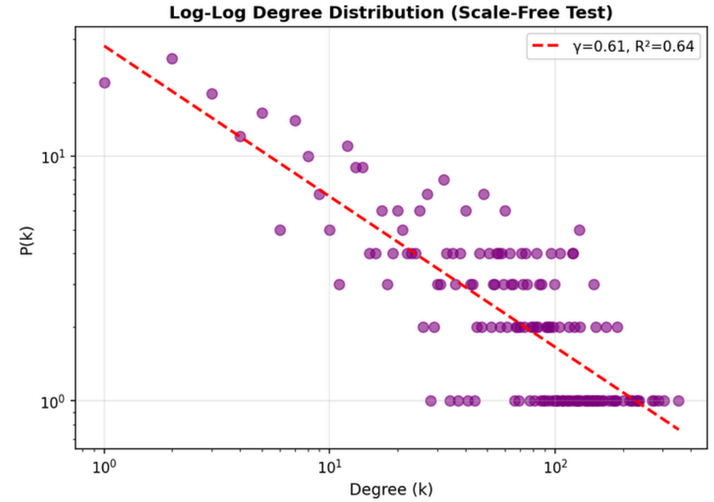


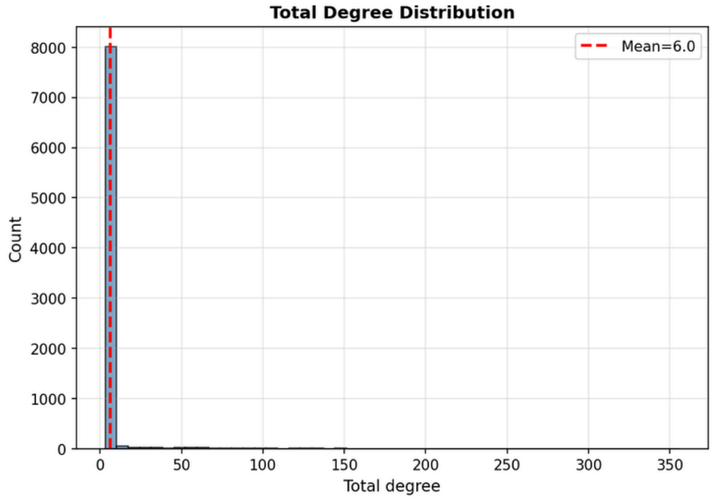


Network Topology: Degree Distributions





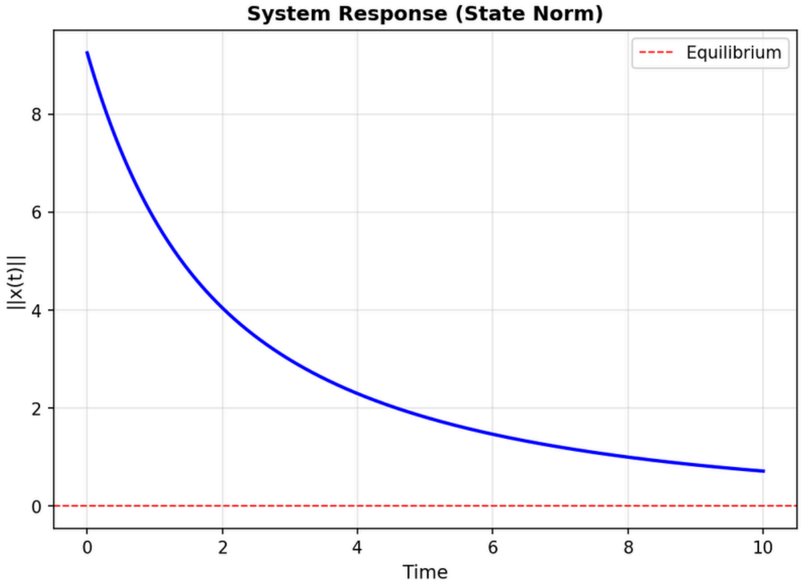


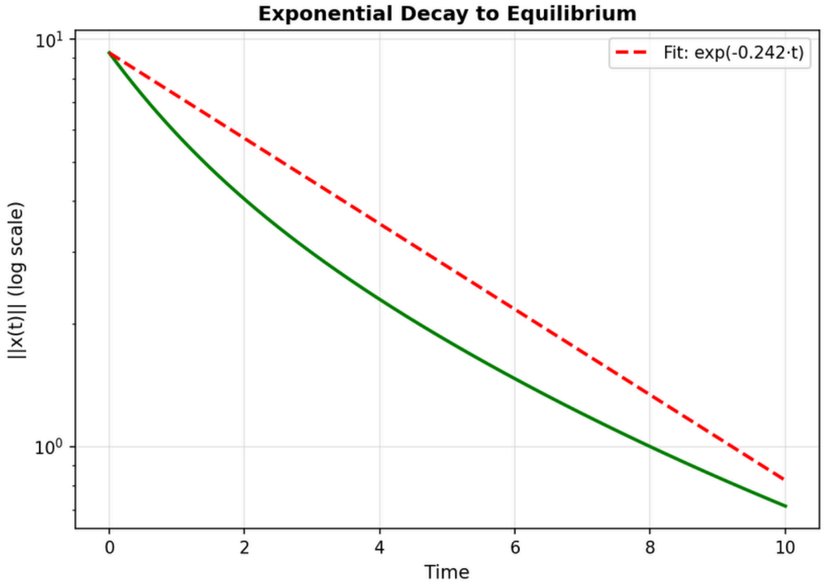


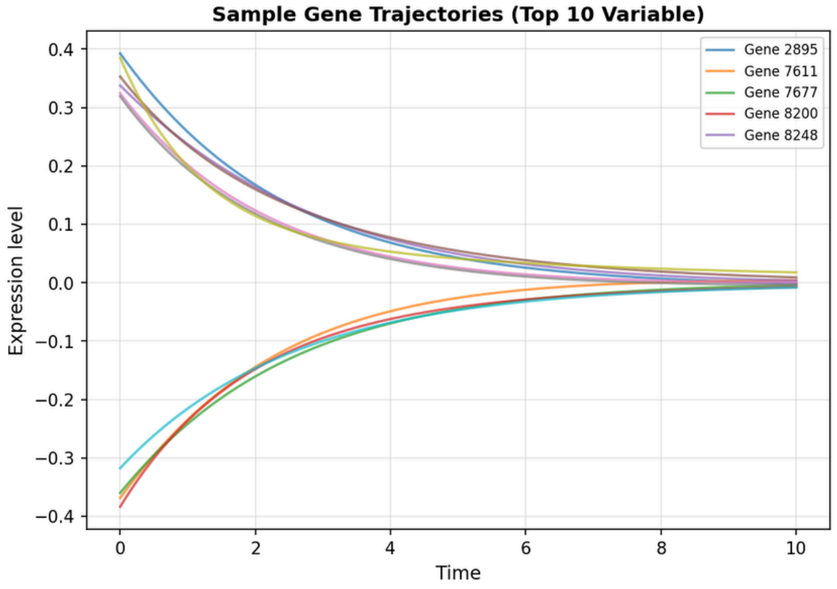
Network Statistics Summary

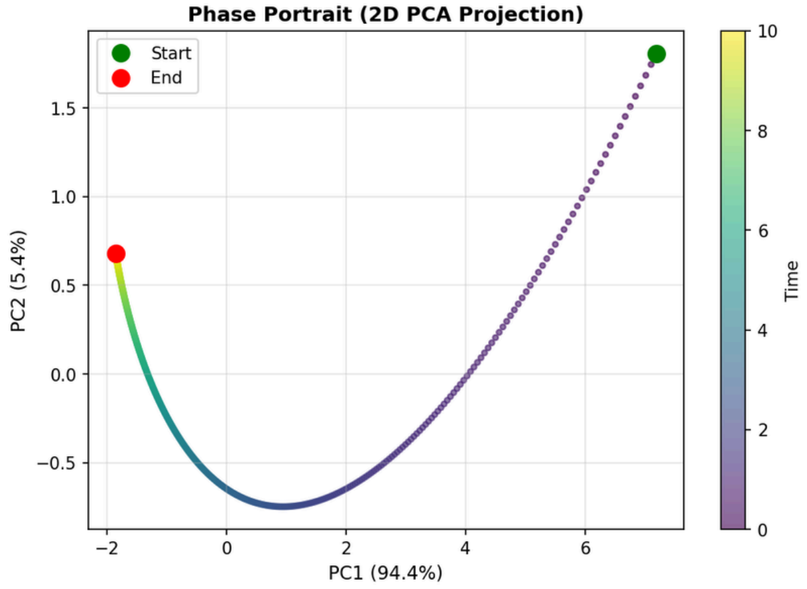
Property	Value	Interpretation
Nodes	8,378	Genes in network
Edges	25,134	Regulatory interactions
Density	0.036%	Highly sparse
Avg degree	6.0	3 in + 3 out
Max out-degree	353	Master regulator
Connectivity		
WCC	1 (100%)	Fully connected
SCC	8,073	Mostly hierarchical
Isolates	0	All genes participate
Small-World		
Clustering coeff.	0.133	370× higher than random
Avg path length	3.77	Efficient information flow
Small-world?	Yes	High clustering + short paths
Motifs		
Feed-forward loops	633	Signal filtering
Feedback loops	81	Homeostasis
Reciprocity	2.5%	Mostly unidirectional

Dynamics Simulation

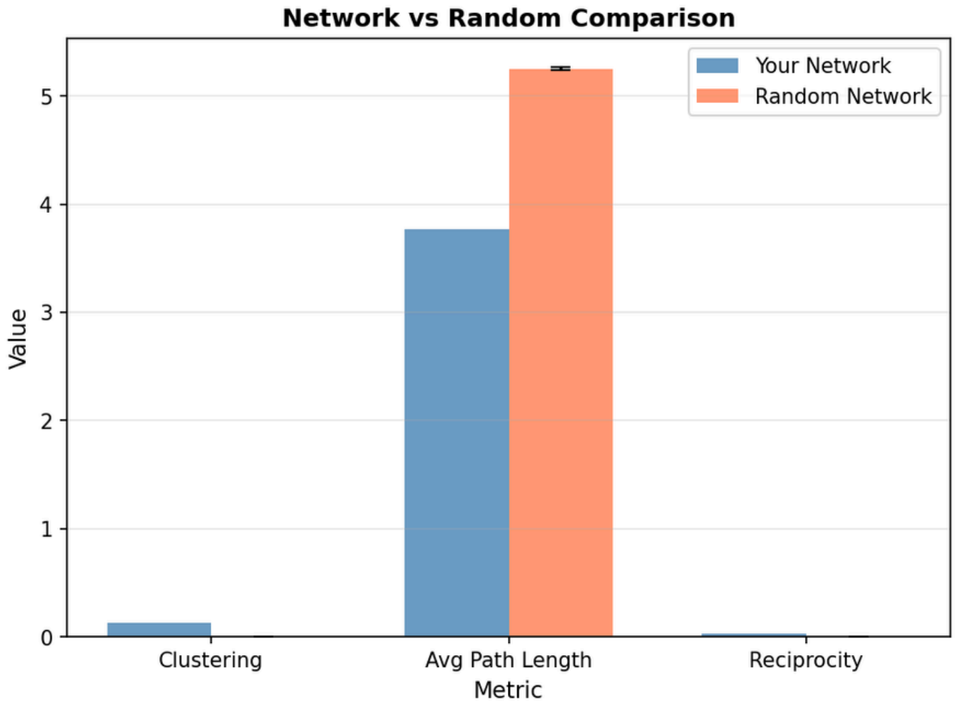


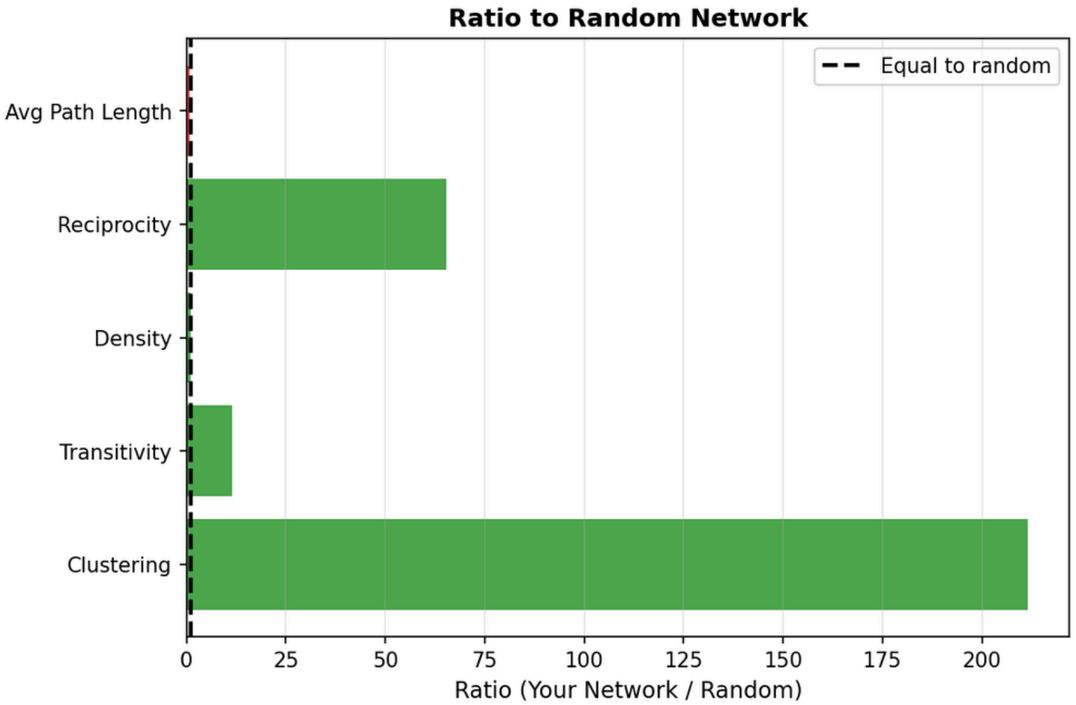






Comparison to Random Networks





Interpretation:

- Higher clustering indicates strong functional modules
- Shorter paths imply efficient signaling across the network
- Non-random structure reflects underlying biological organization

System Formulation for Control

Linear Time-Invariant (LTI) System

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$$

Control Objectives:

1. Set-point regulation (healthy state)
2. Trajectory tracking
3. Minimize control effort

System Matrix (A)

- Size: 8378 × 8378
- Stable: max eigenvalue = −1.26
- Highly sparse: 25,134 non-zeros (99.96% sparse)
- Signed edges: activation / repression

Control Design

- Control input matrix B (n × m)
- Control signals u(t) (dimension m)
- Model reduction: 8378 → 100
- LQR controller
- MPC controller

Top Hub Genes (by Degree):

Gene	Degree
IGFBP7	66.27
TUBA1C	54.58
UBE2G2	52.87
ITGB1	48.07
SPCS1	44.95
RPS27A	44.54
HNRNPK	40.96
GSG1L	39.18
GABARAPL2	38.43
MXD3	37.71

Master regulators with high out-degree

Biologically Motivated Selection:

- **TP53**: Tumor suppressor (tested)
- **BRCA1**: DNA repair (tested)
- **EGFR**: Growth factor receptor (tested)

Final B Matrix Design:

- **Dimension:** 8,378 $\times$ 27 control inputs
- **Strategy:** Top-27 hub genes by degree
- **Actuation:** Direct control via unit vectors

$$B_{i,j} = \begin{cases} 1.0 & \text{if gene } i \text{ is control gene } j \\ 0 & \text{otherwise} \end{cases}$$

Controllability Analysis

Structural Controllability:

Metric	Value
Total genes	8,378
Driver nodes	7,899
Controllability	**94.3%**
Control inputs (m)	27

Interpretation:

- ✓ 94% of genes are potential drivers
- ✓ Sparse actuation possible
- ✓ Network highly controllable
- ✓ Multiple control strategies feasible

Controllability Rank Check:

For selected $\mathbf{B}$ matrix ($m = 27$ inputs):

$$\mathcal{C} = [\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{m-1}\mathbf{B}]$$

Challenge:

- Full rank check: $O(n^3)$ with $n = 8{,}378$ (infeasible)
- Solution: Model reduction

Model Reduction Strategy:

- **Original:** $n = 8{,}378$ states
- **Reduced:** $r = 100$ effective dimensions
- **Method:** POD/SVD-based projection
- **Control inputs:** $m = 27$ maintained

Control Law: LQR with Model Reduction

Discrete-Time LQR Design

Cost function:

$$J = \sum_{k=0}^{\infty} [\mathbf{x}_k' Q \mathbf{x}_k + \mathbf{u}_k' R \mathbf{u}_k]$$

Control law:

$$\mathbf{u}_k = -K\mathbf{x}_k, \quad K = (R + \mathbf{B}'P\mathbf{B})^{-1}\mathbf{B}'P\mathbf{A}$$

where P solves the discrete-time algebraic Riccati equation (DARE).

System model

$$x_{k+1} = Ax_k + Bu_k.$$

Finite-horizon quadratic cost

$$J = \sum_{k=0}^{N-1} (x_k^\top Q x_k + u_k^\top R u_k) \; + \; x_N^\top Q_f x_N.$$

Quadratic value function (DP)

$$V_t(x) = x^\top P_t x, \qquad P_N = Q_f.$$

Stage minimization

$$J_t(x,u) = x^\top (Q + A^\top P_{t+1} A)x + 2u^\top (B^\top P_{t+1} A)x + u^\top (R + B^\top P_{t+1} B)u.$$

Define:

$$S_t = R + B^\top P_{t+1} B, \qquad G_t = B^\top P_{t+1} A.$$

Optimal feedback

$$u_t^\star = -K_t x_t, \qquad K_t = S_t^{-1} G_t = (R + B^\top P_{t+1} B)^{-1} B^\top P_{t+1} A.$$

Riccati recursion

$$P_t = Q + A^\top P_{t+1} A - A^\top P_{t+1} B K_t.$$

Interpretation

- Produces time-varying gains $\{K_0, \ldots, K_{N-1}\}$.
- Equivalent to unconstrained finite-horizon LQR.

MPC Implementation

Receding-horizon MPC

- At each time t: solve finite-horizon LQR.
- Apply only first control: $u_t = -K_0 x_t$.
- Shift horizon and repeat.

Reference tracking

$$\tilde{x}_t = x_t - x_{\text{ref}}, \qquad u_t = -K_0 \tilde{x}_t + u_{\text{ref}}.$$

Steady-state feedforward satisfies:

$$x_{\text{ref}} = A x_{\text{ref}} + B u_{\text{ref}}, \qquad (I - A) x_{\text{ref}} = B u_{\text{ref}}$$

Solve via least-squares or pseudoinverse for u_{ref}.

Key equations (compact)

$$K_t = (R + B^\top P_{t+1} B)^{-1} B^\top P_{t+1} A,$$

$$P_t = Q + A^\top P_{t+1} A - A^\top P_{t+1} B K_t,$$

$$u_t = -K_0 (x_t - x_{\text{ref}}) + u_{\text{ref}}.$$

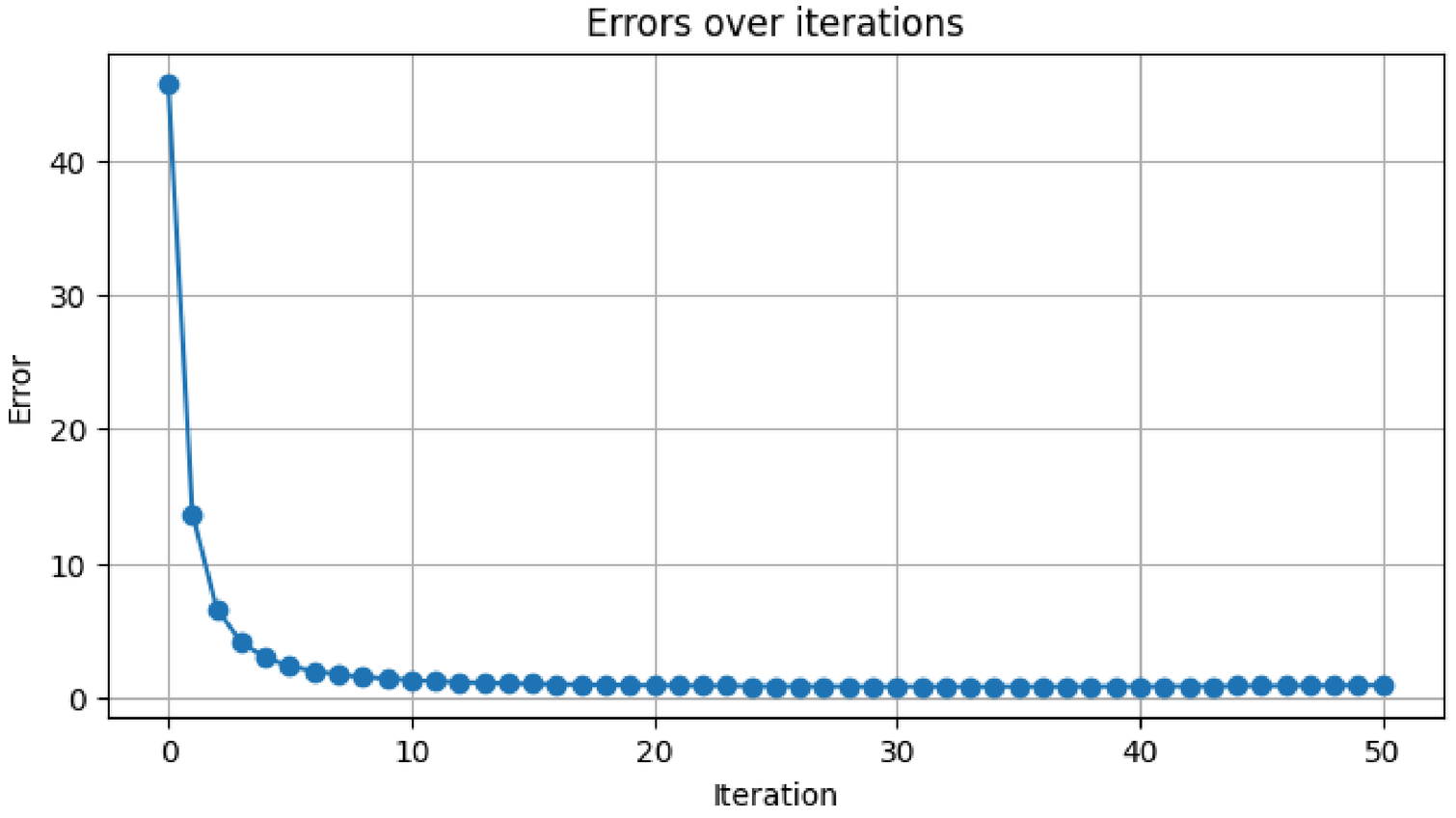
Numerical considerations

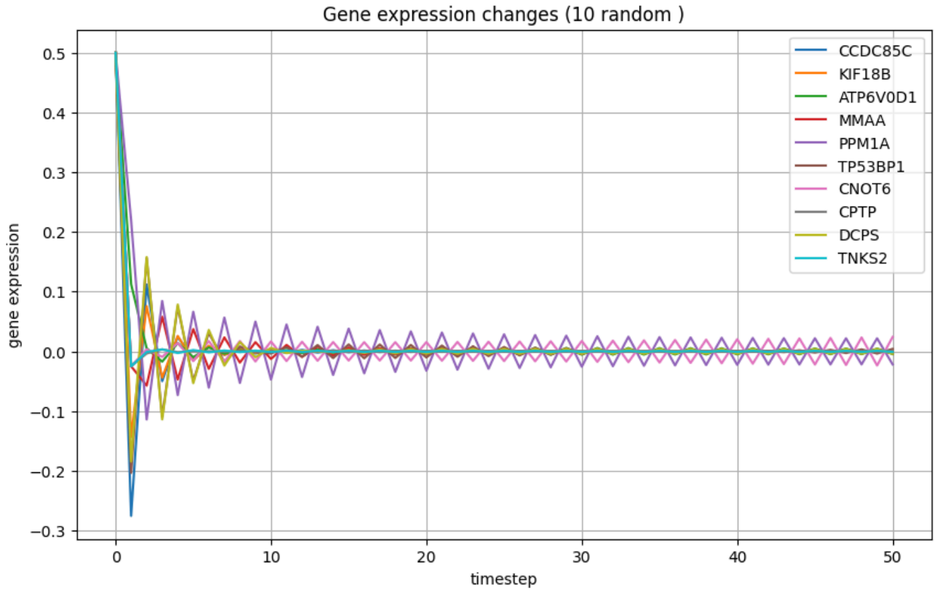
- Regularize $S_t = R + B^\top P_{t+1} B$ (e.g. $S_t + \epsilon I$) if ill-conditioned.
- Symmetrize P_t to avoid drift: $P_t \leftarrow (P_t + P_t^\top)/2$.
- Scaling of Q, R strongly affects conditioning and performance.

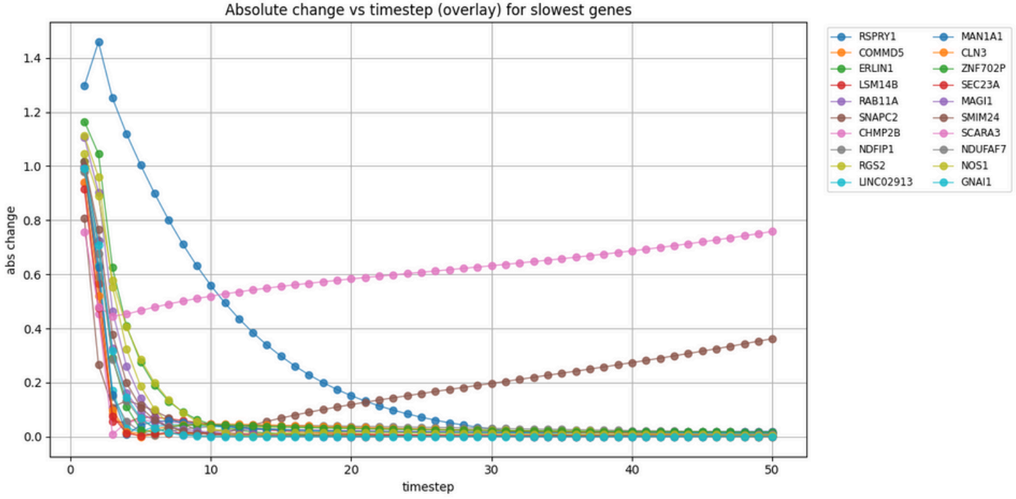
Simulation Results

Configuration	Number of Drivers
Set 1	837
Set 2	500
Set 3	300
Best Selection	27

Top Degree Selection N drivers : 837



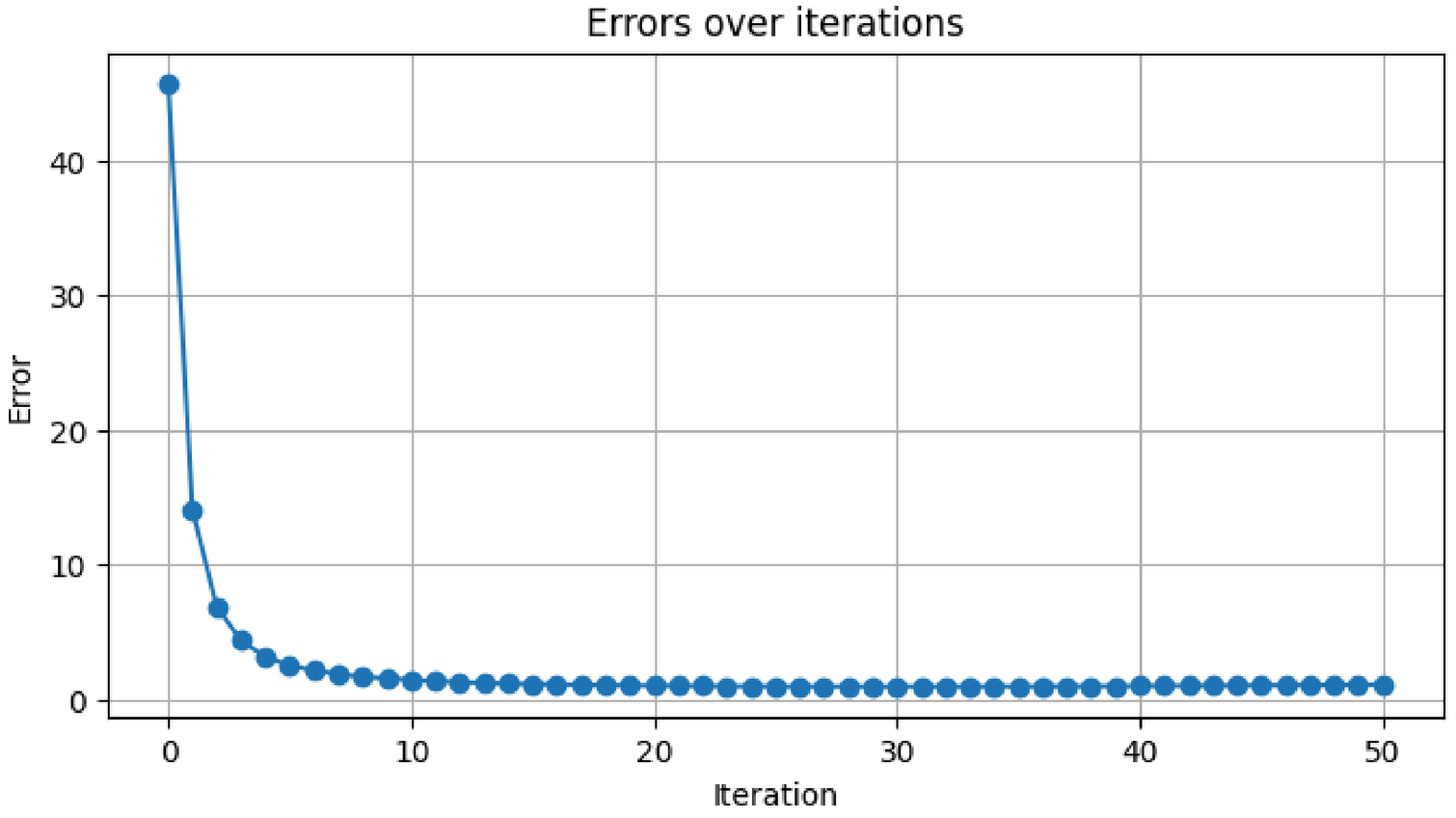


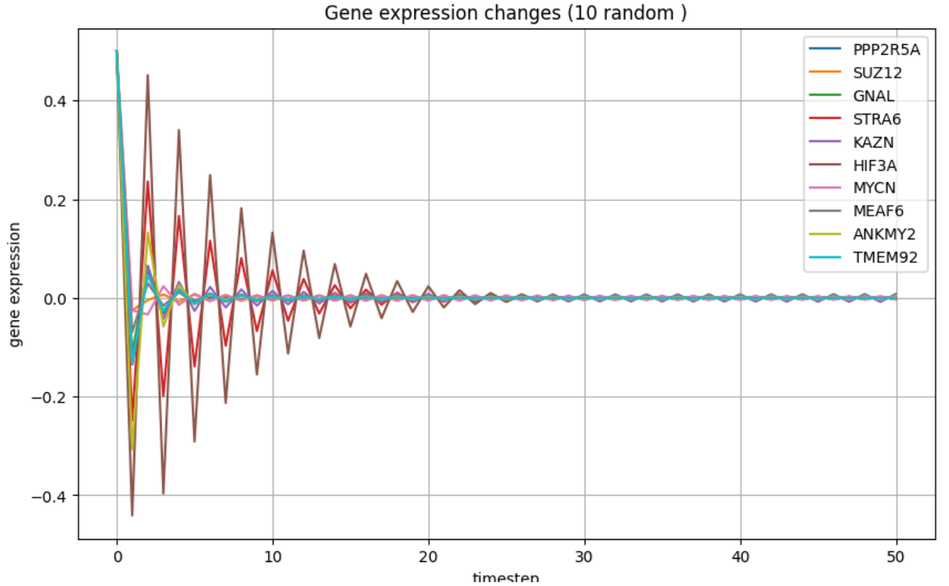


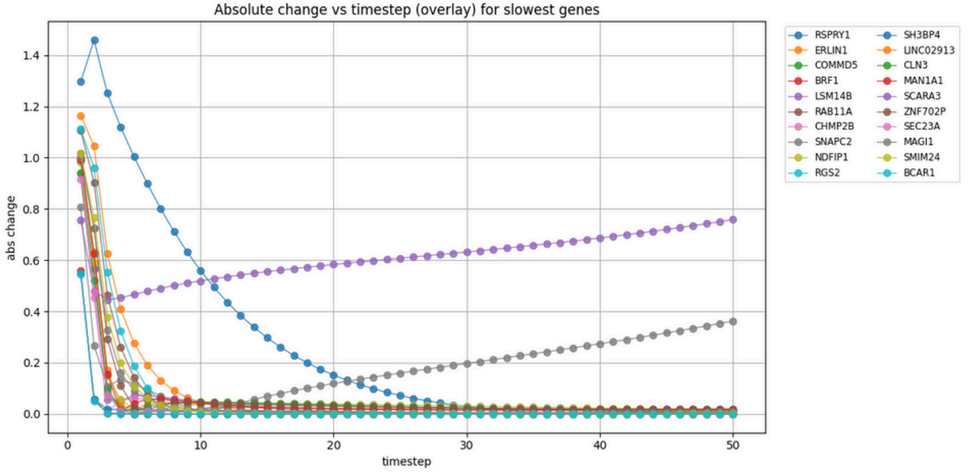
Run — 837 driver nodes (top-degree selection)

- Error curve: rapid decline and low final error → good convergence.
- Trajectories: most gene signals settle quickly toward steady values.
- Variability: overall amplitude and step changes are reduced across genes.
- Tradeoff: very effective stabilization but uses many driver nodes (low sparsity).

Top Degree Selection N drivers : 500



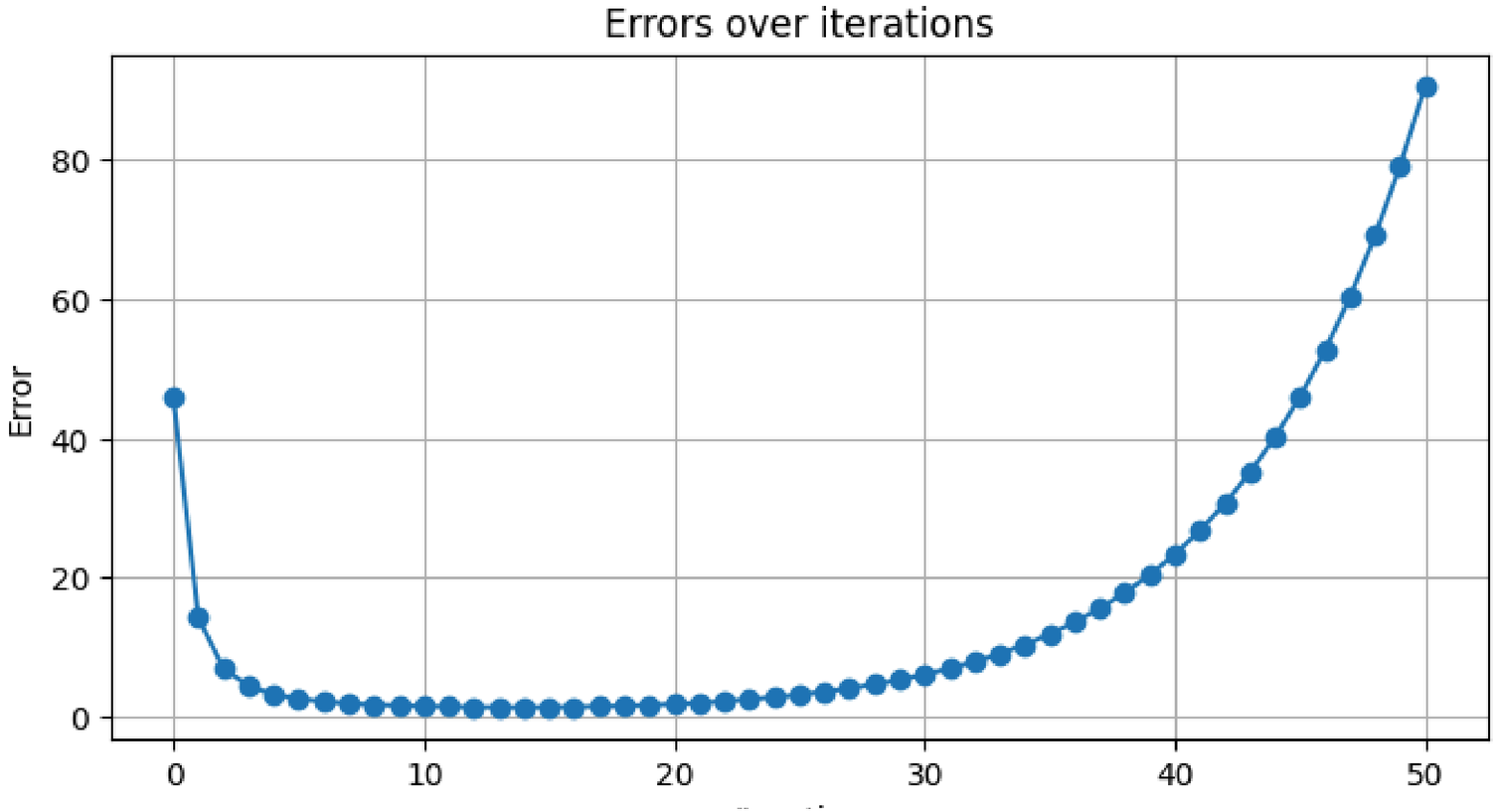


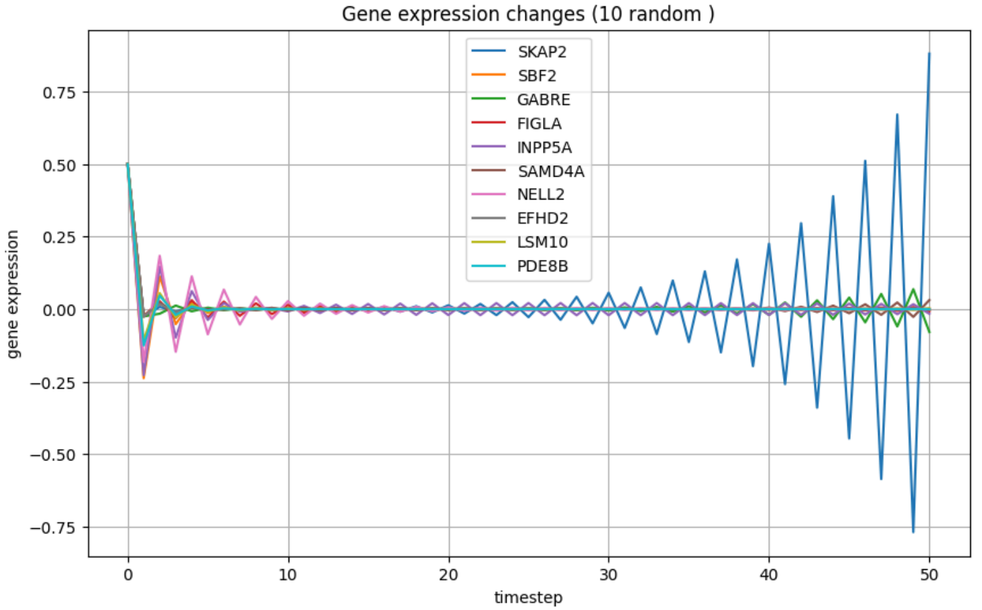


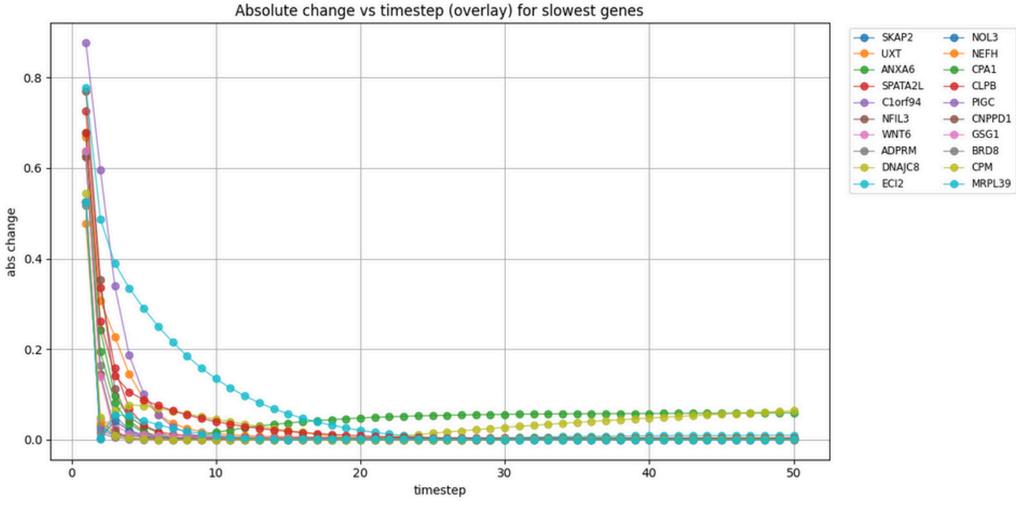
Run — 500 driver nodes (top-degree selection)

- Error curve: clear decline but slightly slower / higher final error than 837.
- Trajectories: many genes stabilize, though more residual fluctuations remain.
- Variability: moderate reduction in per-gene variance compared with uncontrolled case.
- Tradeoff: good balance — fewer drivers with modest loss in control performance.

Top Degree Selection N drivers : 300



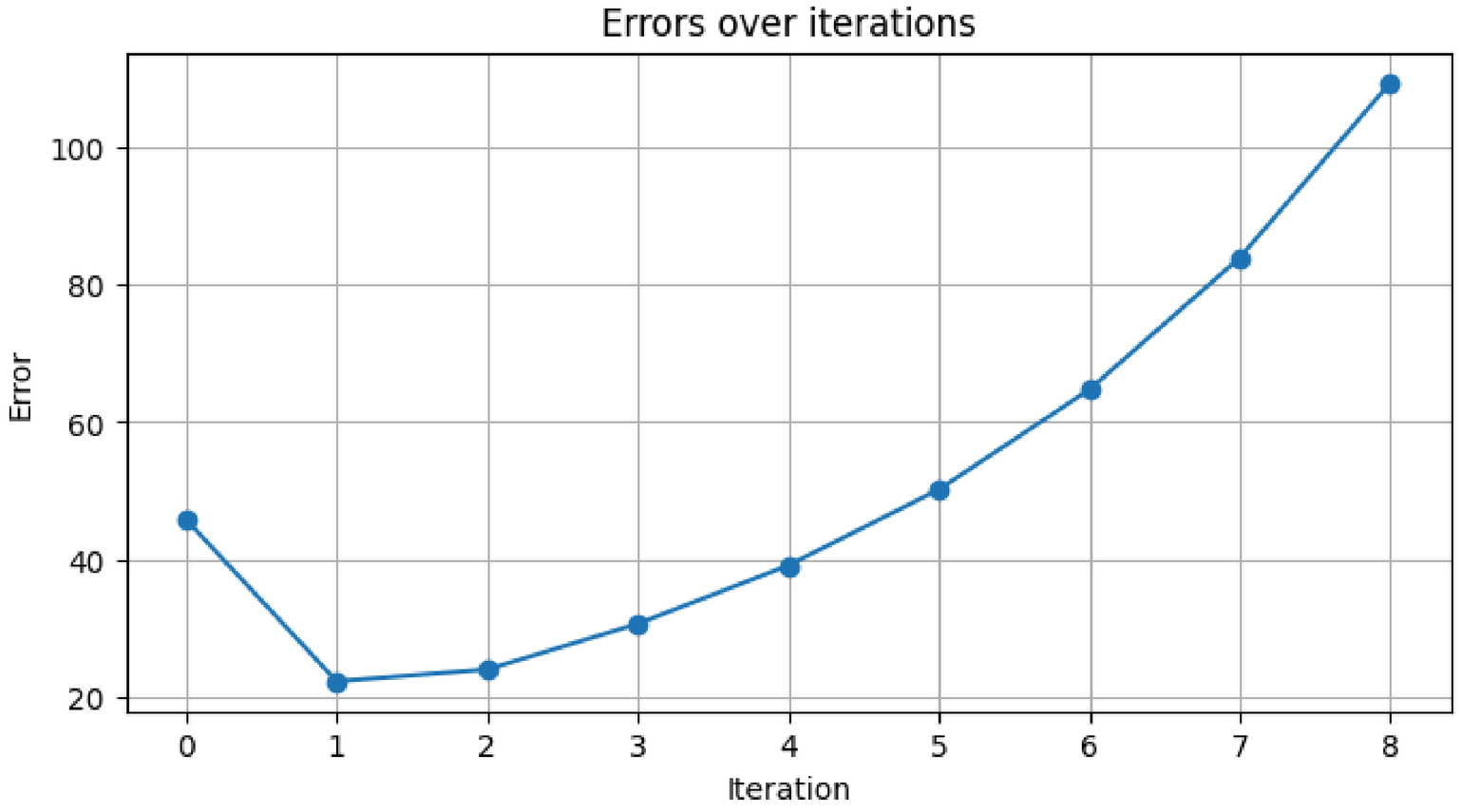


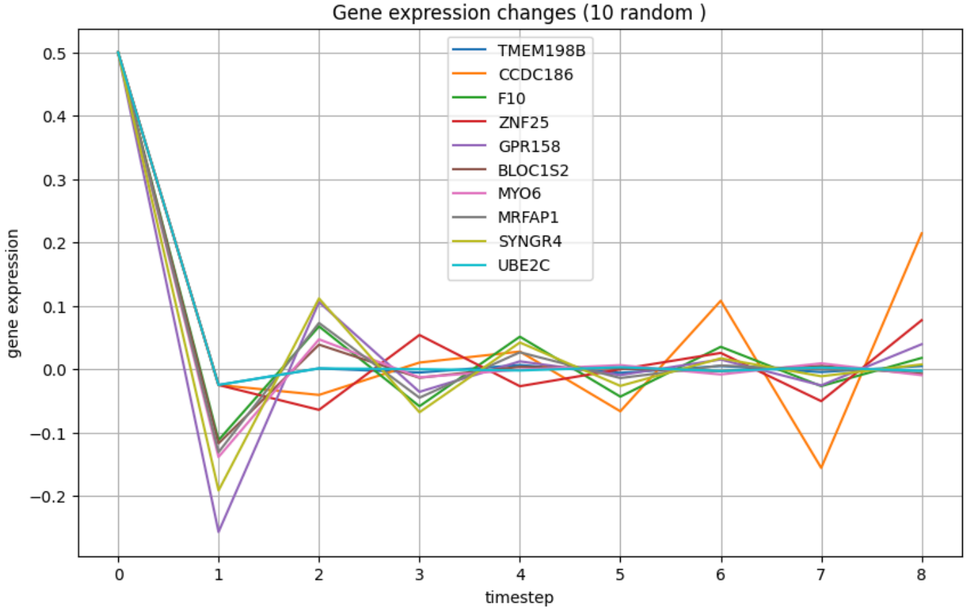


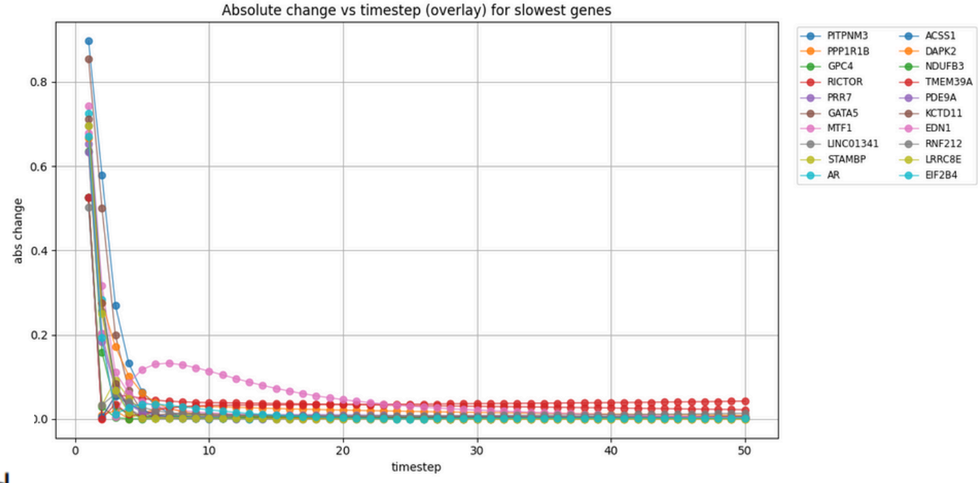
Run — 300 driver nodes (top-degree selection)

- Error curve: slower convergence and noticeably higher final error.
- Trajectories: several genes show persistent deviations or larger oscillations.
- Variability: increased spread and larger step-changes for peripheral genes.
- Tradeoff: control effectiveness drops substantially when driver count is reduced this far.

Gramian Analysis selection N drivers : 27







Run — 27 driver nodes (Gramian controllability selection)

- Error curve: initial drop for controllable modes but higher final error overall.
- Trajectories: key modes / hub genes are well-regulated; many others remain uncontrolled.
- Variability: non-target genes show large changes and higher variance.
- Tradeoff: very sparse, mode-specific control — efficient for targeted regulation but insufficient for full-network stabilization.

Limitations and Future Work

GRN Inference Limitations:

1. **Time-invariance:**
   - **A** is constant (no cell cycle dynamics)
   - Valid only for short horizons
   - Missing circadian regulation
2. **Linearity assumption:**
   - No saturation or Hill kinetics
   - Approximation near equilibrium
   - Neglects gene expression bounds
3. **Top-K constraint:**
   - Forces uniform in-degree $= 3$
   - Sacrifices scale-free property
   - Trade-off: sparsity vs biological realism

Control Implementation Limitations:

1. **Model reduction instability:**
   - Critical issue: $r = 100$ too aggressive
   - Lost critical system dynamics
   - Numerical artifacts in discrete-time
   - **Lesson:** Need $r \geq 500$ or continuous-time
2. **No constraint enforcement:**
   - Gene expression bounds ignored
   - Control saturation not handled
   - MPC can address this (future work)
3. **Reference state ambiguity:**
   - No validated "healthy" profile
   - $\mathbf{x}_{ref} = 0$ is arbitrary
   - Need clinical data for targets

References

- The Cancer Genome Atlas Network (2012). *Comprehensive molecular portraits of human breast tumours.* Nature, 490(7418), 61-70.
- Zou, H., & Hastie, T. (2005). *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B, 67(2), 301-320.
- Gershgorin, S. (1931). *Über die Abgrenzung der Eigenwerte einer Matrix.* Bulletin de l'Académie des Sciences de l'URSS, 6, 749-754.
- Liu, Y. Y., Slotine, J. J., & Barabási, A. L. (2011). *Controllability of complex networks.* Nature, 473(7346), 167-173.
- Pedregosa, F., et al. (2011). *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 12, 2825-2830.
- Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX.* Los Alamos National Lab (LANL).

Thank You

Questions?

Aryaman Bahl
aryaman.bahl@students.iiit.ac.in

Autrio Das
autrio.das@research.iiit.ac.in