Causal Models, Prediction, and Extrapolation in Cell Line Perturbation Experiments

James Long

University of Texas MD Anderson Cancer Center

TAMU 5th Annual Bioinformatics Symposium – October 14, 2022

Contact: jplong@mdanderson.org

Collaboration

James Long
MDACC

Summer Yang
UT Health

Kim-Anh Do
MDACC

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

UTHealth
The University of Texas
Health Science Center at Houston

Research partially supported by NIH grants: P30CA016672, CCTS UL1TR003167, P50CA140388, and P50CA127001

Outline

Cell Line Perturbation Experiments

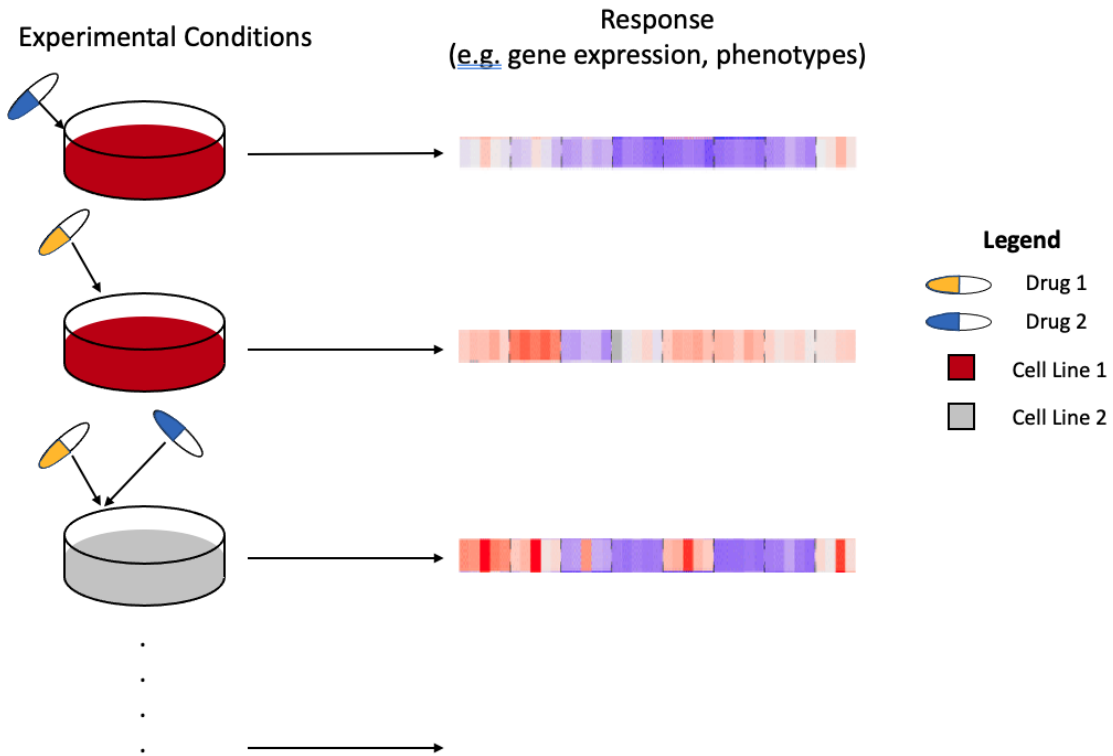
Modeling Strategies
- Regression
- Causal Model (Cellbox)

Comparison
- Analytic
- Simulation
- Melanoma Cell Line Data

Outline

Cell Line Perturbation Experiments

- Groups of cells are perturbed (e.g. drug applied)
- Responses measured (e.g. cell survival, gene expression)



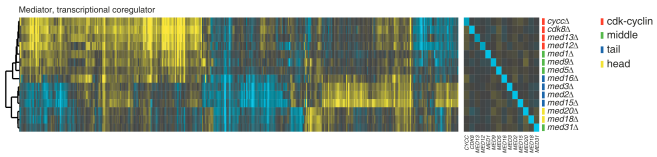
- Many scientific uses for data including identification of synergistic therapies in cancer [Zhao et al., 2020]

In Silico Perturbation Modeling

- **Challenge:** Experimental resources are limited (time, money)
- **Solution:** In silico (computational) models are used to predict the responses to untested conditions.

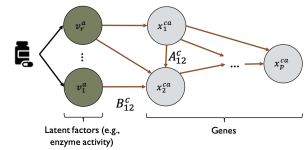
Perturbation Data Sets

LINCS L1000: $\sim$ 1 million experiments, expression of 1000 response genes measured [Subramanian et al., 2017]

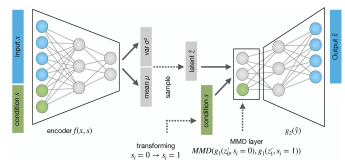


Deleteome: $\sim$ 1500 single gene KO with 6000 mRNA responses [Kemmeren et al., 2014]

Modeling Approaches



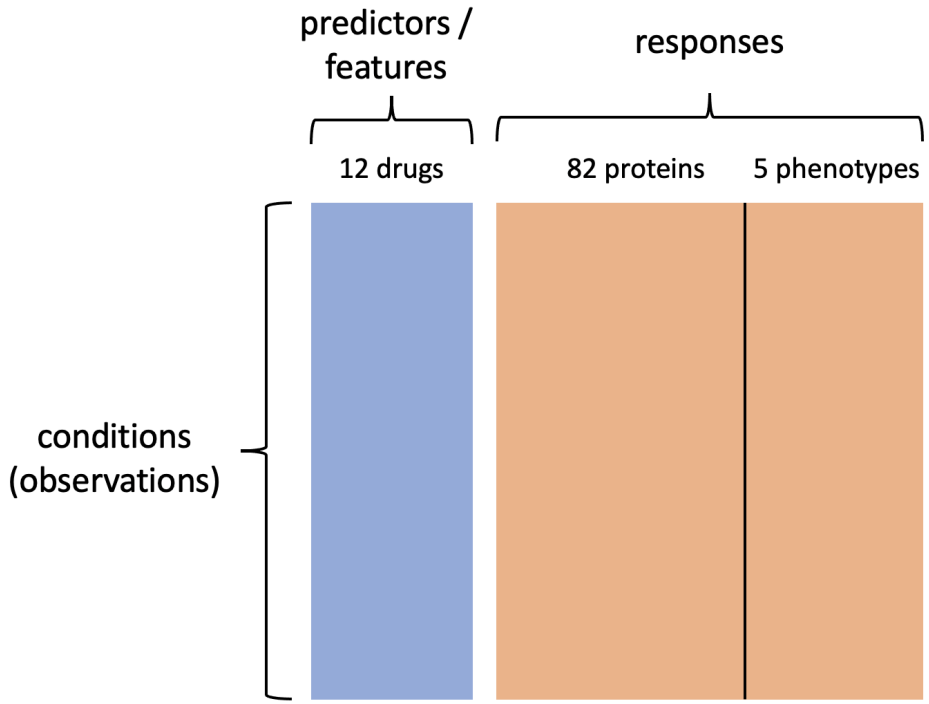
Causal DAGS: Squires et al. [2022], Meinshausen et al. [2016], Peters et al. [2016]



Transfer Learning / VAE: Lotfollahi et al. [2019, 2020, 2021]

Melanoma (SK-Mel-133) Perturbation Experiments

- Data collected in Korkut et al. [2015]
- Single cancer cell line SK-Mel-133
- 12 drugs applied to cell line at various doses

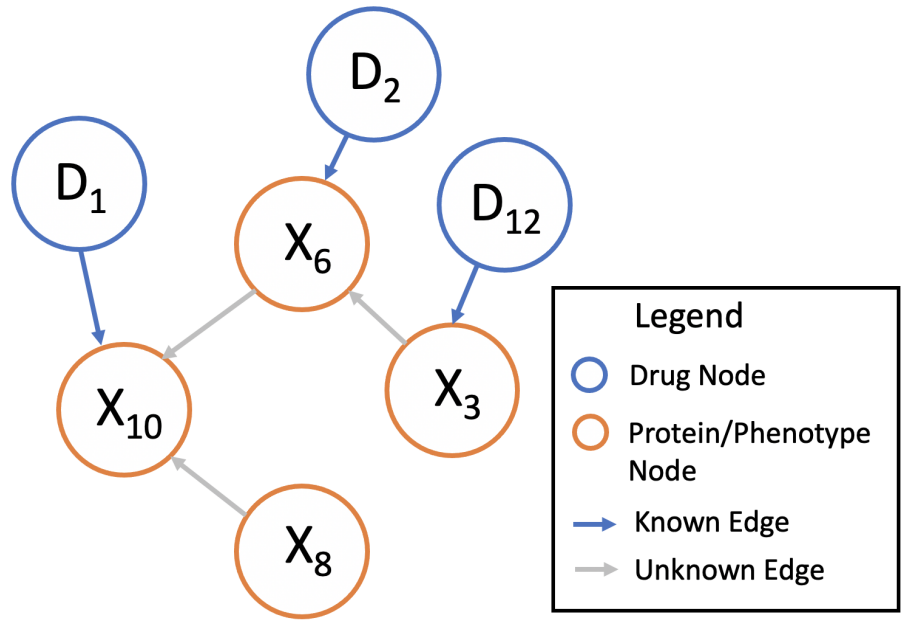


Goal 1: Construct model which can predict cellular responses to these 12 drugs.

Goal 2 (more ambitious): Construct model which can predict cellular responses to untested drugs.

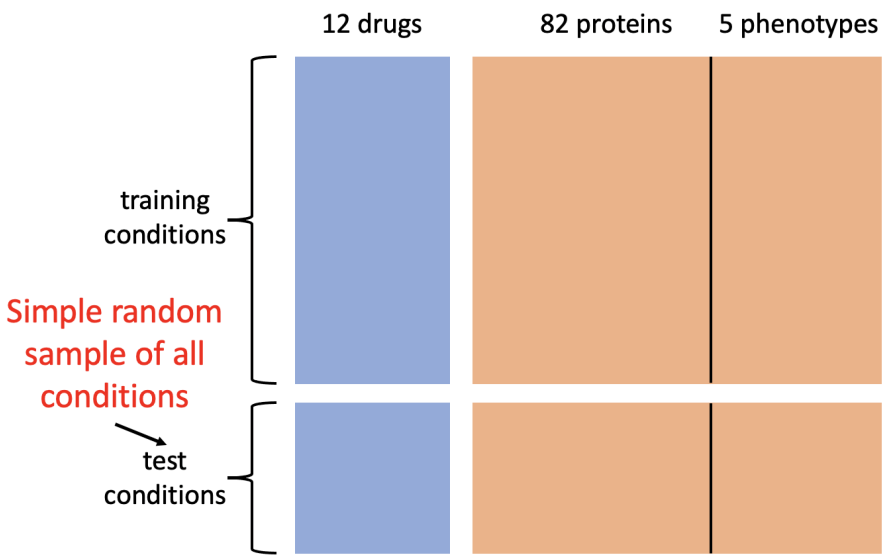
Drugs and Regulatory Networks

- Many drugs directly target a particular protein
  - An AKT inhibitor drug reduces the expression of AKT protein
  - Drug D_{12} is an inhibitor of protein X_3.
- Proteins regulate expression of other proteins / phenotypes according to some causal structure.
  - An AKT inhibitor will effect expression of proteins which are "downstream" from AKT

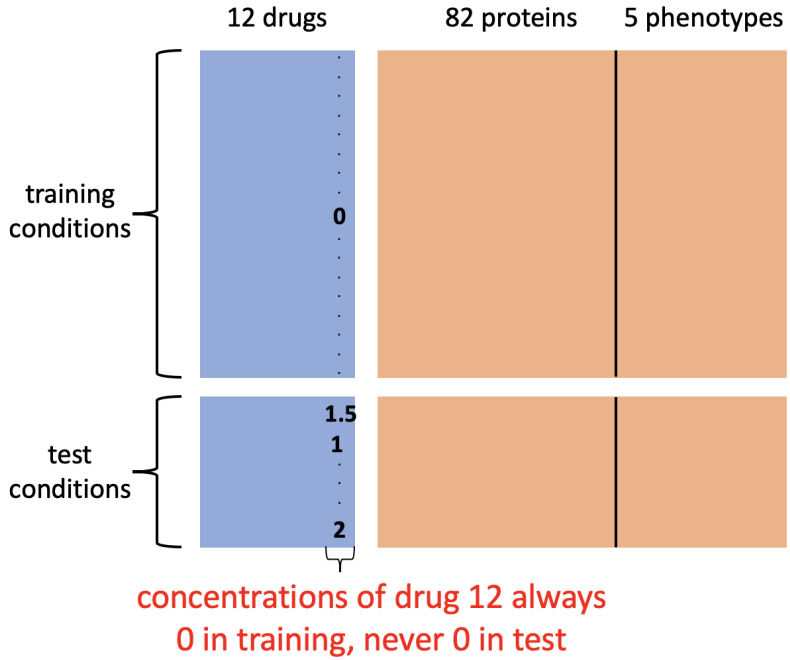


Two Model Validation Strategies

Random Fold (RF)



Leave One Drug Out (LODO)



LODO validation address how well a model can predict effects of yet to be tested drugs (Goal 2).

Overview of Remainder of Talk

- Regression Modeling
  - Struggles with LODO validation.
- Cellbox Causal Model
  - Makes more assumptions than regression models, but can (in principal) achieve Goal 2.
- Comparison of Causal versus Regression Modeling

Outline

Outline

Regression Model and Predictions

- $\mathbf{D} \in \mathbb{R}^{n \times q}$ are training drug concentrations (features)
- $\mathbf{X} \in \mathbb{R}^{n \times p}$ are training protein/phenotype (responses)
- $d \in \mathbb{R}^{q}$ test drug concentrations (features)
- $x \in \mathbb{R}^{p}$ test protein/phenotype (responses)
- Least squares regression fit:

$$\widehat{R} = \underset{R}{\operatorname{argmin}} \; ||\mathbf{X} - \mathbf{D}R||_F^2 + \lambda ||R||_1.$$

- Predict response:

$$\widehat{x} = d^T \widehat{R}.$$

- Compare prediction $\widehat{x}$ with ground truth x

Regression with RF and LODO Validation

- With $\lambda = 0$, $\widehat{R}$ is unique iff $\mathbf{D}^T\mathbf{D}$ is invertible:

$$\widehat{R} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{X}$$

- RF Validation:
  - Invertibility will typically hold when $n > q$ (e.g. holds in Melanoma data set). Could regularize (use $\lambda > 0$) if many drugs q relative to number of experiments n.
- LODO Validation:
  - Invertibility never holds (If drug i is held out of training $(\mathbf{D}^T\mathbf{D})_{ii} = 0$)
  - If $\lambda > 0$ and drug i is held out, then $\widehat{R}_{i\cdot} = \vec{0}$

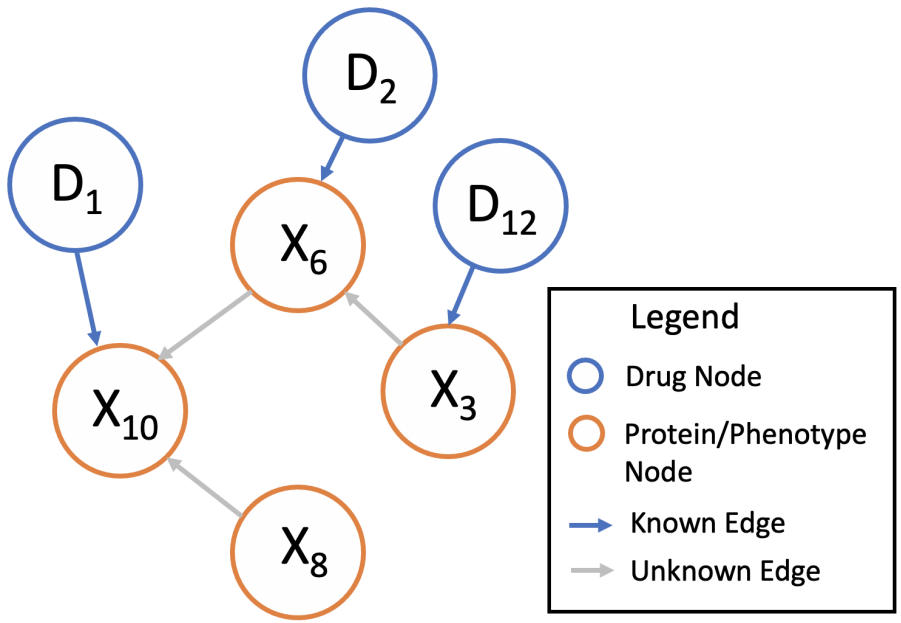
Qualitative Point: LODO validation requires extreme form of extrapolation: predicting the effect of drug that has never been used in training.

Outline

Causal Modeling and LODO Prediction

How can causal models predict effect of untested drugs?

1. Use training data to learn causal structure (grey arrows)
2. For new drug (not used in training data), assume direct target is known (blue arrow)
3. Propagate effect of new drug through the inferred causal structure.

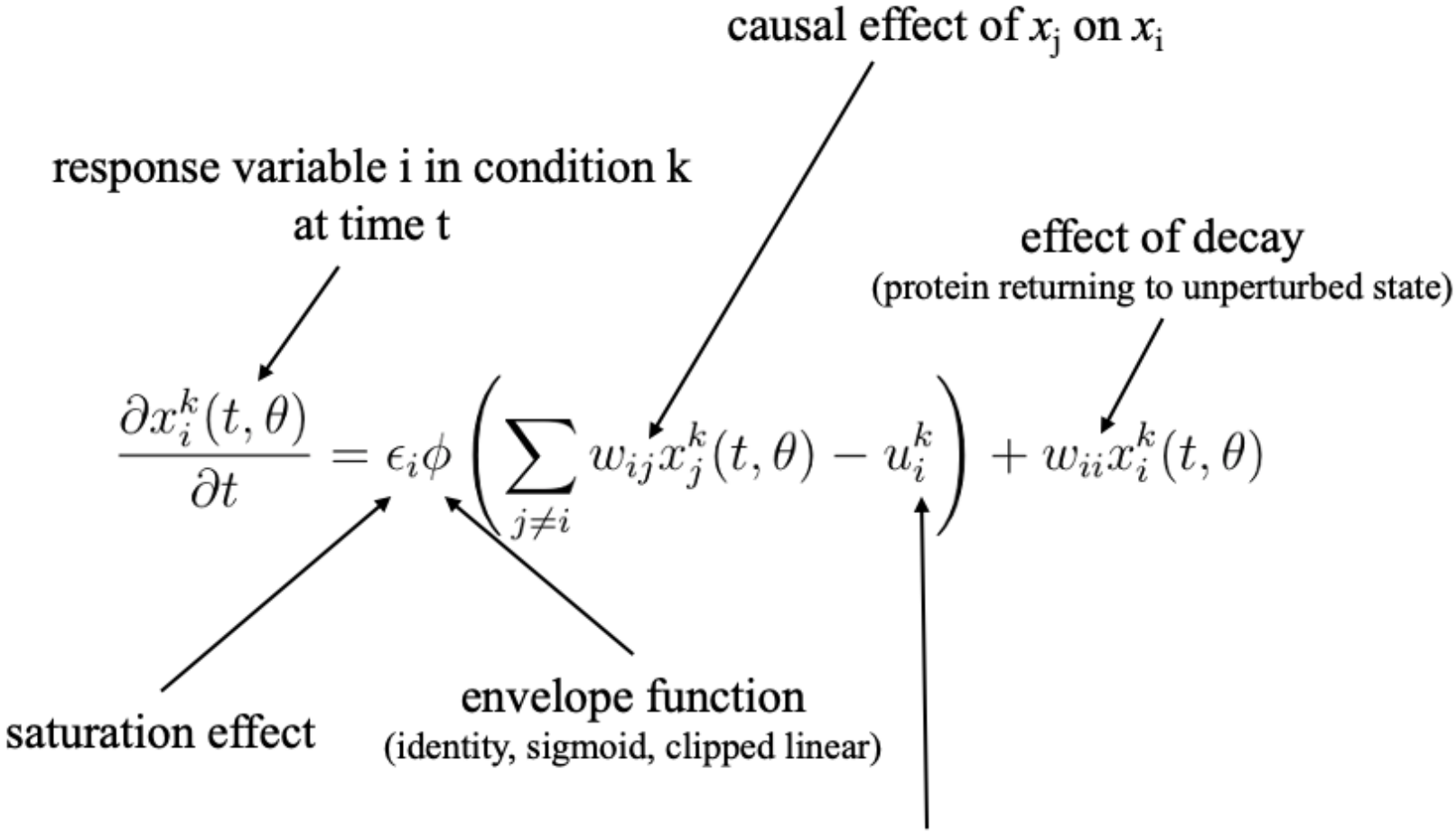


Cellbox implements this idea with ODEs.

Background on Cellbox

- Proposed in Yuan et al. [2021] in Cell Systems
- Ordinary Differential Equations (ODE) model
- Yuan et al. [2021] proposed LODO validation as a more rigorous form of model testing
- Cellbox outperformed competitor methods in RF and LODO validation. Competitors:
  - Neural networks
  - Belief propagation
  - Co–expression models

Cellbox ODE Model



$\theta = (W, \epsilon)$

- $B \in R^{p \times q}$
- B_{il} is effect of 1 unit of drug l on protein i
- $u_i^k = \sum_{l=1}^{q} B_{il}\, d_l^k$

Parameter Estimation

Steady State:

$$x_i^k(\theta) \equiv \lim_{t \to \infty} x_i^k(t, \theta).$$

Loss Function:

$$L(\theta) = \sum_k \sum_i |x_i^k - x_i^k(\theta)|^2 + \lambda ||W - diag(W)||_1$$

Minimize Loss Over θ:

$$\widehat{W}, \widehat{\epsilon} = \underset{\theta=(W,\epsilon)}{\operatorname{argmin}} \; L(\theta). \tag{1}$$

Notes:

- Only steady state data collected on Melanoma, so only steady state value implied by model influences loss.
- Heun's ODE solver + Adam optimizer used to fit parameters.

Outline

Outline

Closed Form Steady State for Linear Cellbox

Theorem

Suppose ϕ is identity envelope function, $\epsilon = 1$, and W is invertible. Then

$$x^k(\theta) = (x_1^k(\theta), \ldots, x_p^k(\theta))^T = -W^{-1}Bd^k$$

and

$$\widehat{W} = \underset{W}{\operatorname{argmin}} \, ||\mathbf{X} - \mathbf{D}B^T(-W^{-1^T})||_F^2 + \lambda||W - diag(W)||_1.$$

Proof Sketch.

- Assumptions imply linear systems of ODEs.
- Linear systems of ODEs have closed form solutions.
- Take time limit ($t \to \infty$) of solution.

□

Causal versus Regression Comparison

Ignoring regularization terms:

$$\widehat{W} = \underset{W}{\operatorname{argmin}} \, ||\mathbf{X} - \mathbf{D}B^T(-W^{-1^T})||_F^2$$

- Estimates W, direct effect of response variables on each other
- Requires knowledge of B, direct targets of drugs
- $\widehat{W}$ can be uniquely defined even when drug is never used, i.e. column of $\mathbf{D}$ is 0. Only need $\mathbf{D}B^T$ to be full column rank.
- **Qualitative Idea:** Model can predict for held out drug (not used in training) by using other drugs which have same direct protein targets as held out drug.

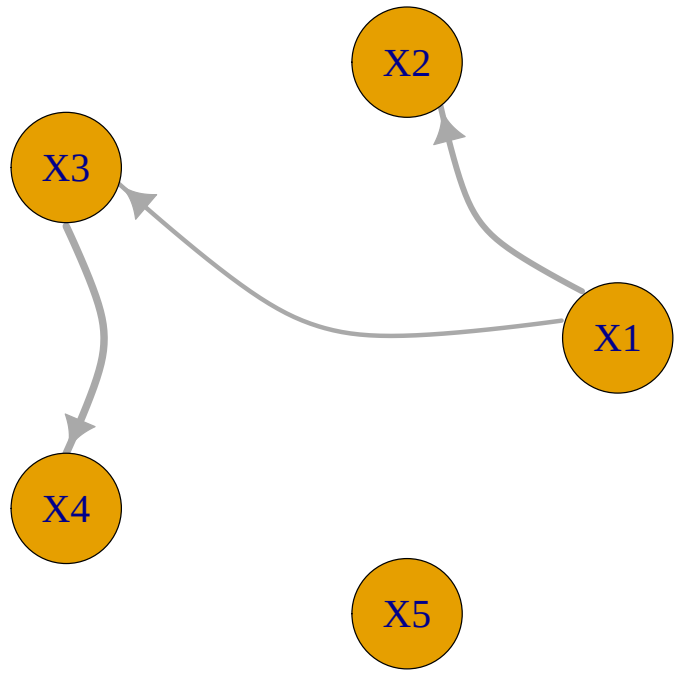
Regression: $\widehat{R} = \underset{R}{\operatorname{argmin}} \, ||\mathbf{X} - \mathbf{D}R||_F^2$

Outline

Parameters

- $p = 5$ response variables
- $q = 15$ drugs
  - 5 drugs target 1 response
  - 10 drugs target 2 response
- All combinations of 2 drugs tested so $n = \binom{15}{2} = 105$
- $\mathbf{D} \in \{0, 1\}^{n \times 15}$
- B matrix
  - Drugs with 1 target have effect 1 on target
  - Drugs with 2 targets have effect $1/2$ on each target
- $\delta_X \in \mathbb{R}^{105 \times 5}$, all elements independent distributed $N(0, 0.2^2)$

True Causal DAG A



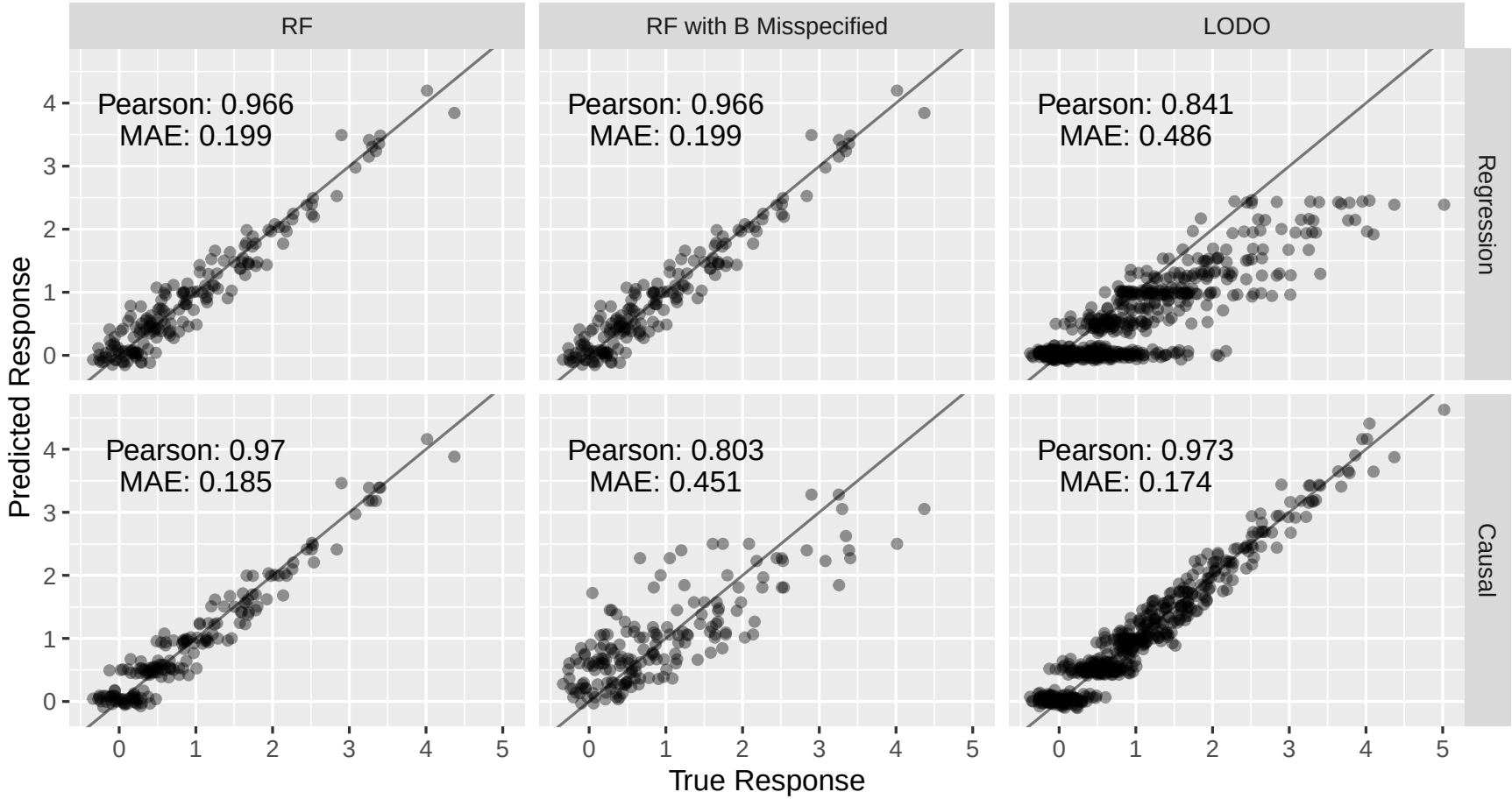
$$\mathbf{X} = \mathbf{D}B^T (I - A)^{-1} + \delta_X$$

Testing Conditions

Compare Regression with Causal Estimator in 3 settings:

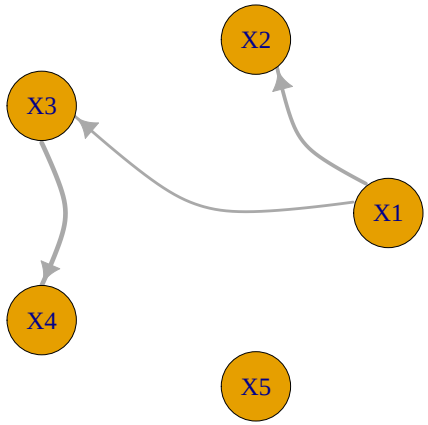
- **Random Fold (RF):**
  - Data is divided randomly into 2/3 training and 1/3 test
- **RF with B Misspecified:**
  - Training–test set split is identical to RF.
  - B matrix (direct effect of drugs) is misspecified.
  - 10 drugs with 2 targets are assumed to influence their targets with a strength of 1
- **Leave-one-drug-out (LODO):**
  - One drug is left out of the training set.
  - For the regression estimator, the coefficient on the left out drug is set to 0.

Simulation Prediction Performance

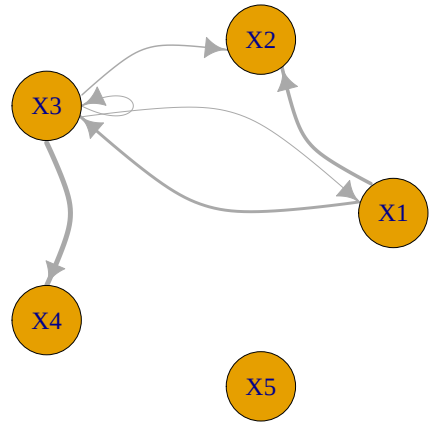


- **RF:** Regression and Causal model both obtain good performance
- **RF with B Misspecified:** Causal model has poor performance
  - Regression is completely robust to misspecification
- **LODO:** Regression has poor performance

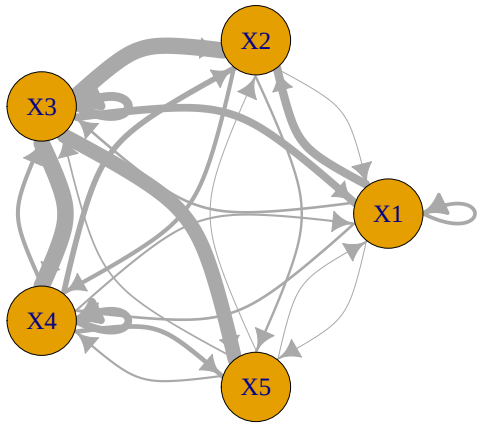
Simulation Imputed DAGs with Causal Model



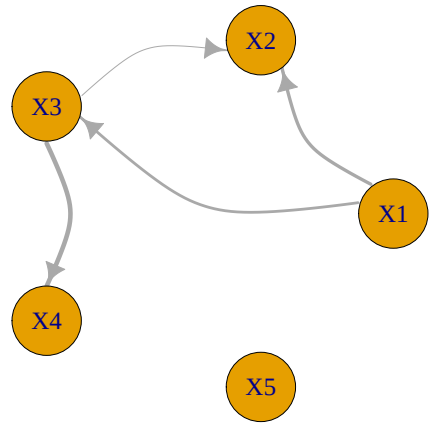
a) True DAG



b) Random Fold



c) Random Fold, Misspecified B



d) LODO

Outline

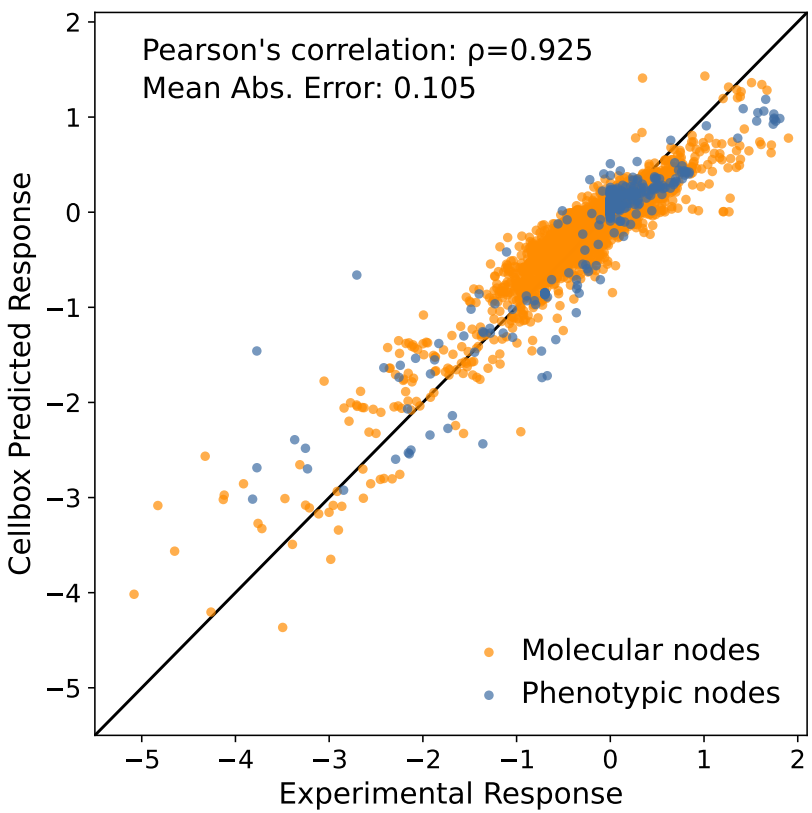
Background

- Data collected in Korkut et al. [2015]
  - Korkut modeled with Belief Propagation algorithms
- Yuan et al. [2021] developed / tested Cellbox on data
  - Proposed LODO validation
  - Cellbox outperformed all competitors
  - Did not compare Cellbox to Linear Regression on RF or LODO
- We follow Yuan et al. [2021] for model validation setup
- Use Cellbox results from paper
  - Sigmoid function ϕ

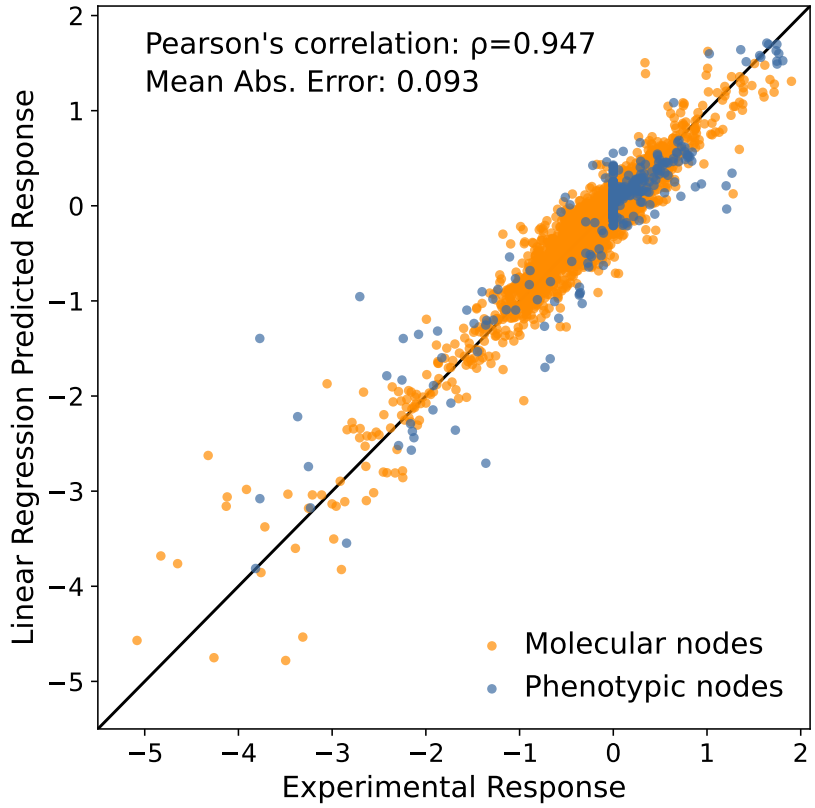
Random Fold Validation Setup

- 70% training / 30% testing data split
- Repeated 1000 times
- Obtain roughly $300 = 0.3 \times 1000$ predictions for each condition
- Average predicted responses.
- Compute correlation between predictions and experimentally observed responses
    - $n \times 87$ predictions

Random Fold Validation Results



Cellbox



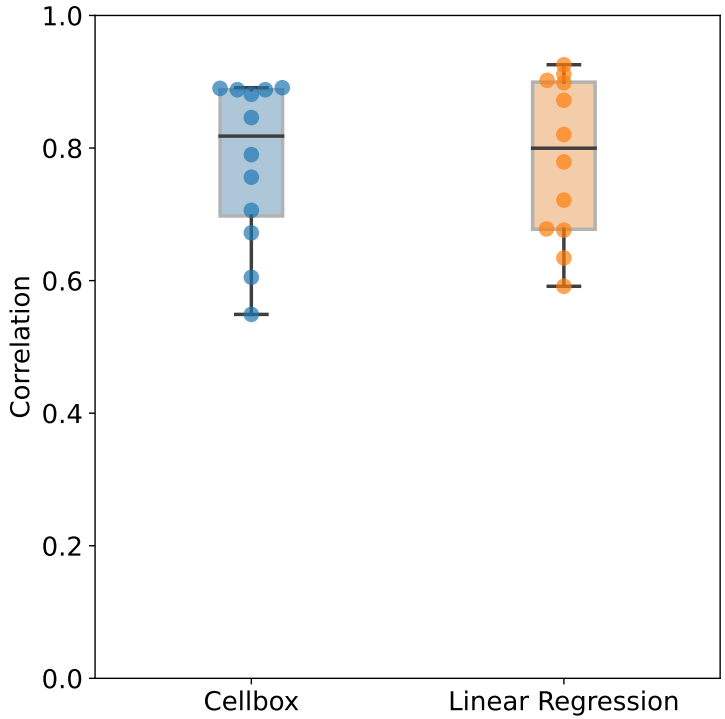
Linear Regression

Linear regression outperforms Cellbox in RF validation.

LODO Setup

- For drug $A \in \{1, \ldots, 12\}$
  - Training set is all conditions where drug A not used
  - Test set is all other conditions
  - For linear regression, set effect of drug A on responses to 0
  - Compute correlation between observed and predicted responses for Cellbox and Linear Regression
- Results in 12 correlations (1 / drug) for each model

LODO Results



Average correlation coefficient:

- 0.780 for Cellbox
- 0.784 for Linear regression

Conclusion: Linear regression and Cellbox obtain very similar performance in LODO.

Summary

- We derived some of the first analytic results comparing causal discovery models for prediction with regression models.
- Causal discovery models make more assumptions than the regression approach, but can extrapolate to predict effect of untested drugs.
    - Focused on linear modeling case, but qualitative concepts apply to non–linear models.
- Achieved state–of–the art prediction performance on the Melanoma cell line using linear regression. This highlights the importance of benchmarking in bioinformatics.

References

- Full details:
  - Long, James P., Yumeng (Summer) Yang, and Kim-Anh Do. "Causal Models, Prediction, and Extrapolation in Cell Line Perturbation Experiments." arXiv preprint arXiv:2207.09991 (2022).
  - `https://github.com/longjp/causal-pred-drug-code`
- Cellbox Paper:
  - Yuan, B., Shen, C., Luna, A., Korkut, A., Marks, D. S., Ingraham, J., & Sander, C. (2021). "CellBox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy." Cell systems, 12(2), 128-140.
  - `https://github.com/sanderlab/CellBox`

Thank you. Questions?

Bibliography I

P. Kemmeren, K. Sameith, L. A. Van De Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. van Wageningen, C. W. Ko, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014.

A. Korkut, W. Wang, E. Demir, B. A. Aksoy, X. Jing, E. J. Molinelli, Ö. Babur, D. L. Bemis, S. O. Sumer, D. B. Solit, et al. Perturbation biology nominates upstream–downstream drug combinations in raf inhibitor resistant melanoma cells. *Elife*, 4, 2015.

M. Lotfollahi, F. A. Wolf, and F. J. Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

M. Lotfollahi, M. Naghipourfar, F. J. Theis, and F. A. Wolf. Conditional out-of-distribution generation for unpaired data using transfer vae. *Bioinformatics*, 36(Supplement_2):i610–i617, 2020.

M. Lotfollahi, A. K. Susmelj, C. De Donno, Y. Ji, I. L. Ibarra, F. A. Wolf, N. Yakubova, F. J. Theis, and D. Lopez-Paz. Compositional perturbation autoencoder for single-cell response modeling. *BioRxiv*, 2021.

N. Meinshausen, A. Hauser, J. M. Mooij, J. Peters, P. Versteeg, and P. Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

C. Squires, D. Shen, A. Agarwal, D. Shah, and C. Uhler. Causal imputation via synthetic interventions. In *Conference on Causal Learning and Reasoning*, pages 688–711. PMLR, 2022.

A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

B. Yuan, C. Shen, A. Luna, A. Korkut, D. S. Marks, J. Ingraham, and C. Sander. Cellbox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. *Cell systems*, 12(2):128–140, 2021.

W. Zhao, J. Li, M.-J. M. Chen, Y. Luo, Z. Ju, N. K. Nesser, K. Johnson-Camacho, C. T. Boniface, Y. Lawrence, N. T. Pande, et al. Large-scale characterization of drug responses of clinically relevant proteins in cancer cell lines. *Cancer Cell*, 38(6): 829–843, 2020.