Causal Discovery and Prediction with Interventional Data: Application to Cell Perturbation Experiments

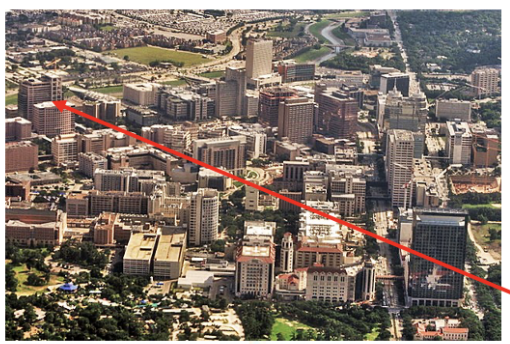
James P. Long

University of Texas MD Anderson Cancer Center

Shiga Seminar – June 7, 2023

Contact: jplong@mdanderson.org

My Background



TMC | TEXAS MEDICAL CENTER

I work there.

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

My Background

- Research interests:

- Visiting Shiga University for June.
- Collaborating with Professor Shimizu.
- Please come visit me in 317 (jplong@mdanderson.org).

Today's Talk: 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

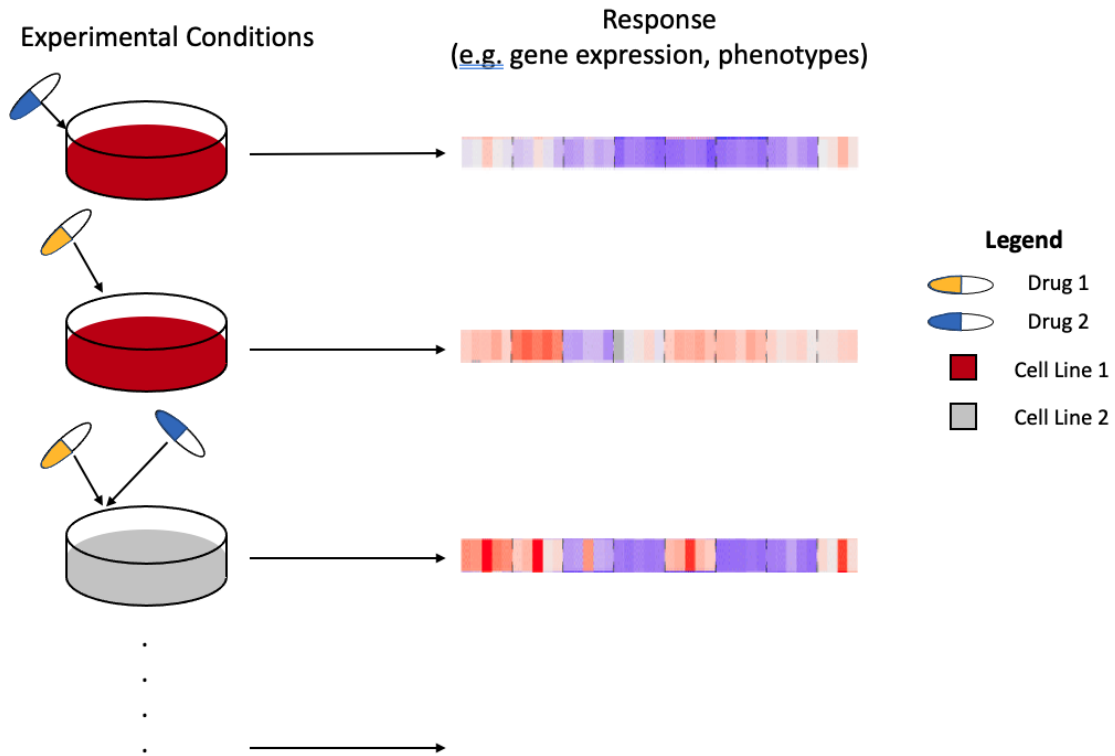
Research partially supported by NIH grants: P30CA016672, CCTS UL1TR003167, P50CA140388, and P50CA127001

Outline

Outline

Cell 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]

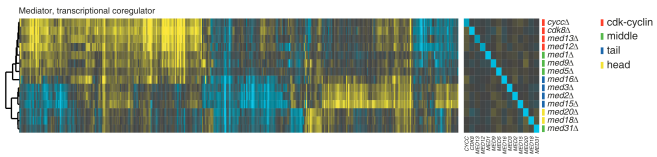
In Silico Perturbation Modeling

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

Perturbation Data Sets

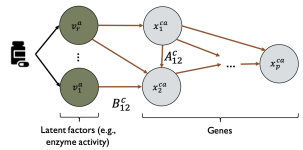
NIH LINCS PROGRAM

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

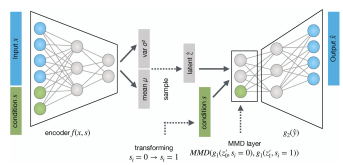


Deleteome: $\sim$ 1500 single gene KO with 6000 mRNA responses [Kemmeren et al., 2014]

Modeling Approaches



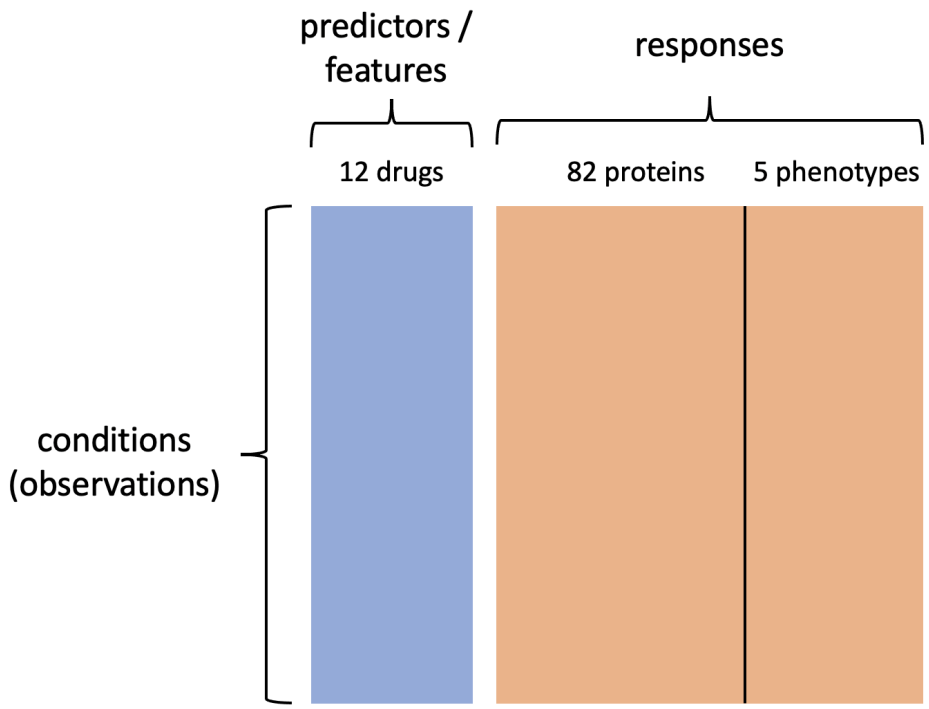
Causal DAGS: Squires et al. [2022], Meinshausen et al. [2016], Peters et al. [2016]



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

Melanoma (SK-Mel-133) Perturbation Experiments

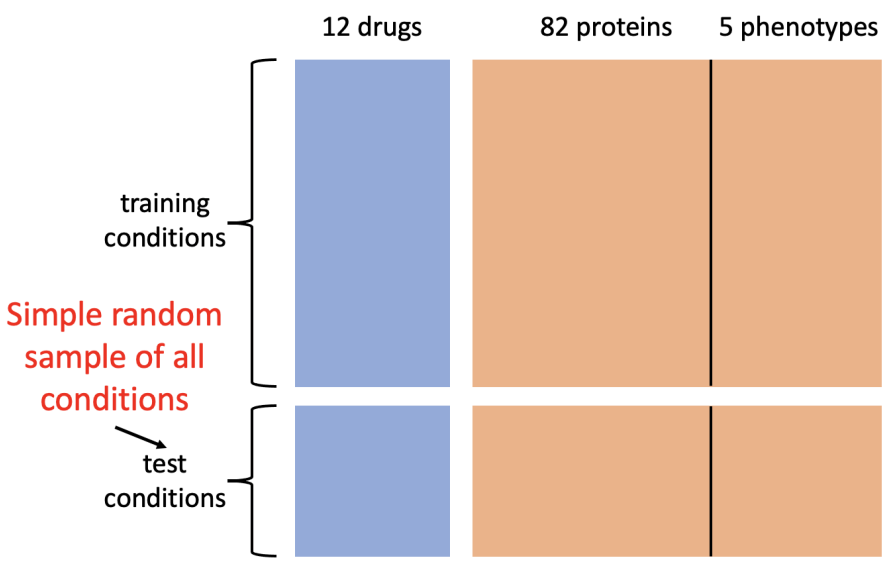
- Data collected in Korkut et al. [2015]
- Single cancer cell line SK-Mel-133
- 12 drugs applied to cell line at various doses



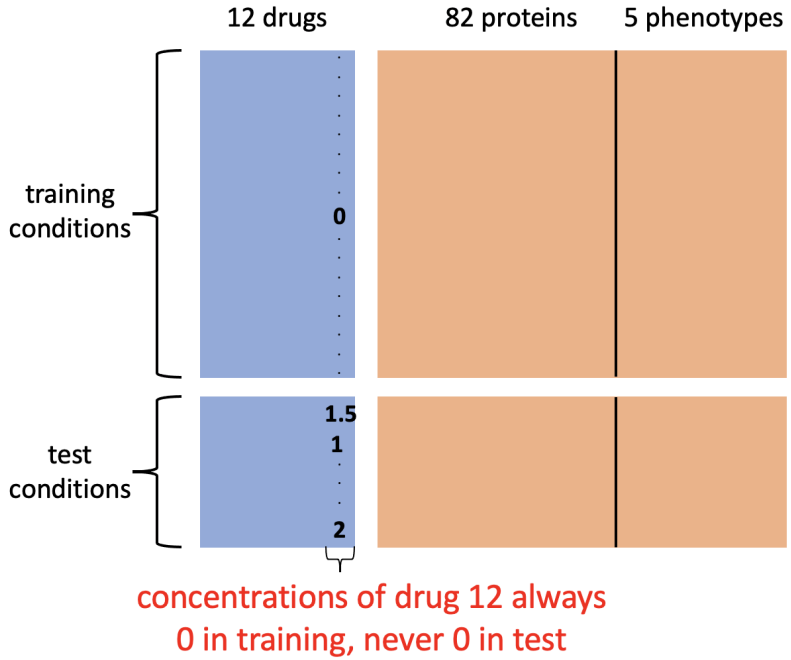
Goal 1: Predict responses to combinations of these 12 drugs.
Goal 2 (more ambitious): Predict responses to new drugs.

Two Forms of Model Validation

Random Fold (RF)



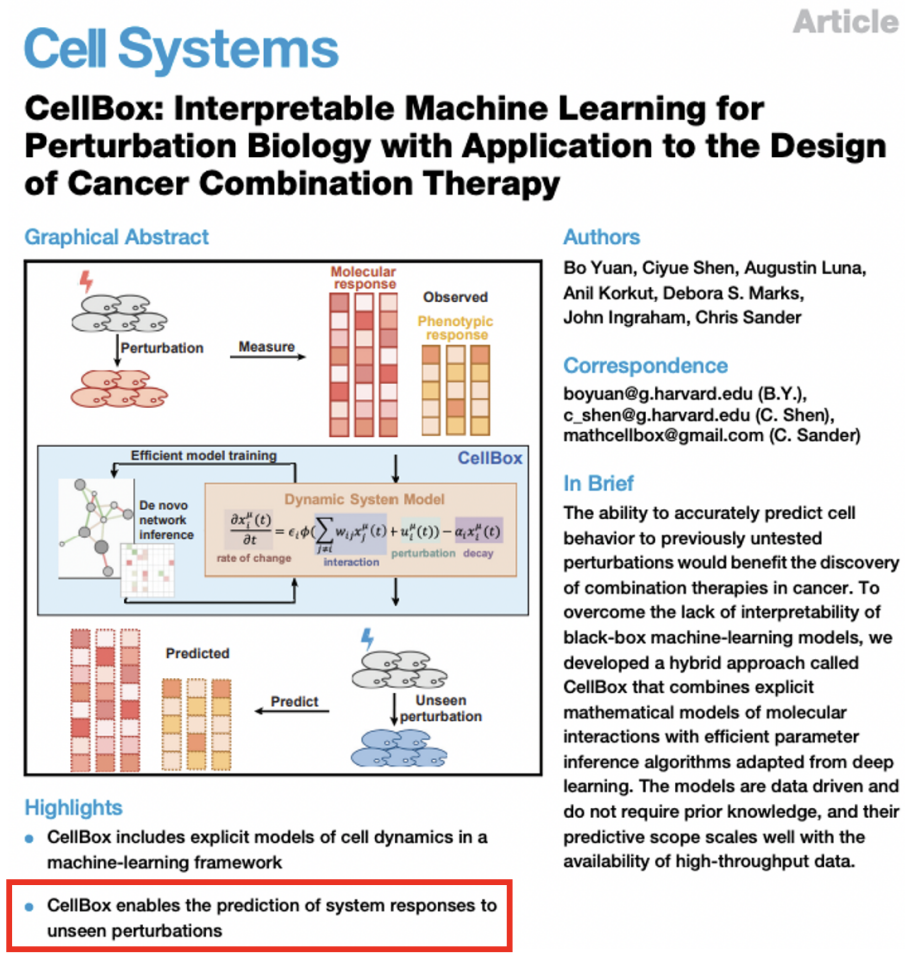
Leave One Drug Out (LODO)



- RF used for assessing model performance on Goal 1.
- LODO used for assessing model performance on Goal 2.

Cellbox Model

- Cellbox: Computational model to predict cell responses to drugs
- Designed to predict responses in LODO (unseen perturbations)



Article

Cell Systems

CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy

Graphical Abstract

Highlights

- CellBox includes explicit models of cell dynamics in a machine-learning framework
- CellBox enables the prediction of system responses to unseen perturbations

Authors

Bo Yuan, Ciyue Shen, Augustin Luna, Anil Korkut, Debora S. Marks, John Ingraham, Chris Sander

Correspondence

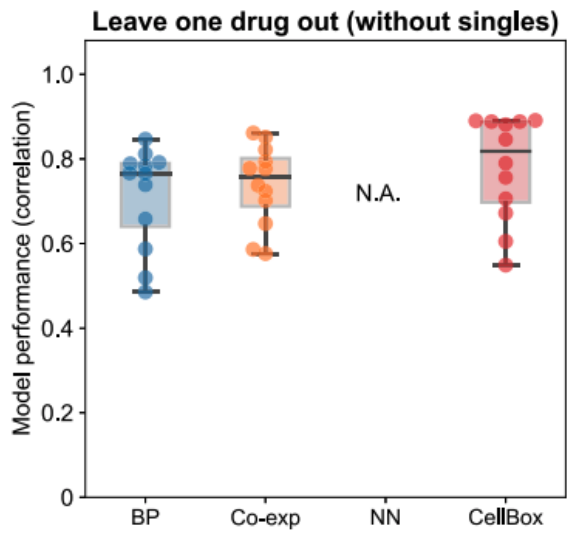
boyuan@g.harvard.edu (B.Y.), c_shen@g.harvard.edu (C. Shen), mathcellbox@gmail.com (C. Sander)

In Brief

The ability to accurately predict cell behavior to previously untested perturbations would benefit the discovery of combination therapies in cancer. To overcome the lack of interpretability of black-box machine-learning models, we developed a hybrid approach called CellBox that combines explicit mathematical models of molecular interactions with efficient parameter inference algorithms adapted from deep learning. The models are data driven and do not require prior knowledge, and their predictive scope scales well with the availability of high-throughput data.

Cellbox Results on LODO

- For drug $i = \{1, \ldots, 12\}$
  - Training is all conditions where drug i not used.
  - Fit model (cellbox or other) on training
  - Compute correlation between predictions and experimental response on test (drug i is always used)



Cellbox had best performance of methods tested.

Overview of Remainder of Talk

- Two Prediction Strategies
  - Regression
  - Causal Discovery / Structure Learning
- Cellbox
  - Introduce Cellbox model
  - Theory connecting Cellbox and Causal Structure Learning
- Comparison of Modeling Strategies
  - Simulations
  - Application to Melanoma data

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:

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

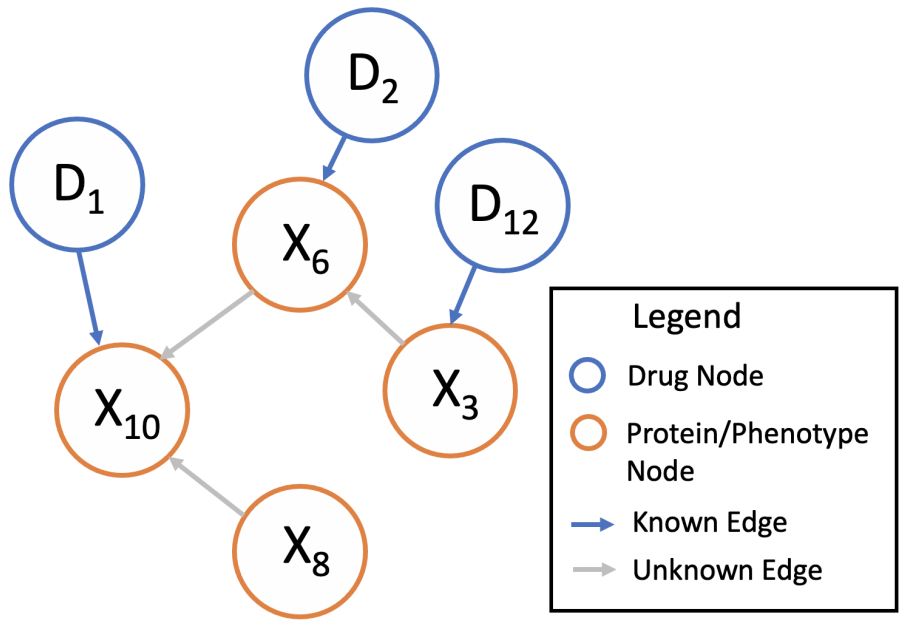
- RF Validation:
  - Invertibility will typically hold when $n > q$ (e.g. holds in Melanoma data set). Could regularize (use $\lambda > 0$) if many drugs q relative to number of experiments n.
- LODO Validation:
  - Invertibility never holds (If drug i is left out of training $(\mathbf{D}^T\mathbf{D})_{ii}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

Drugs and Regulatory Networks

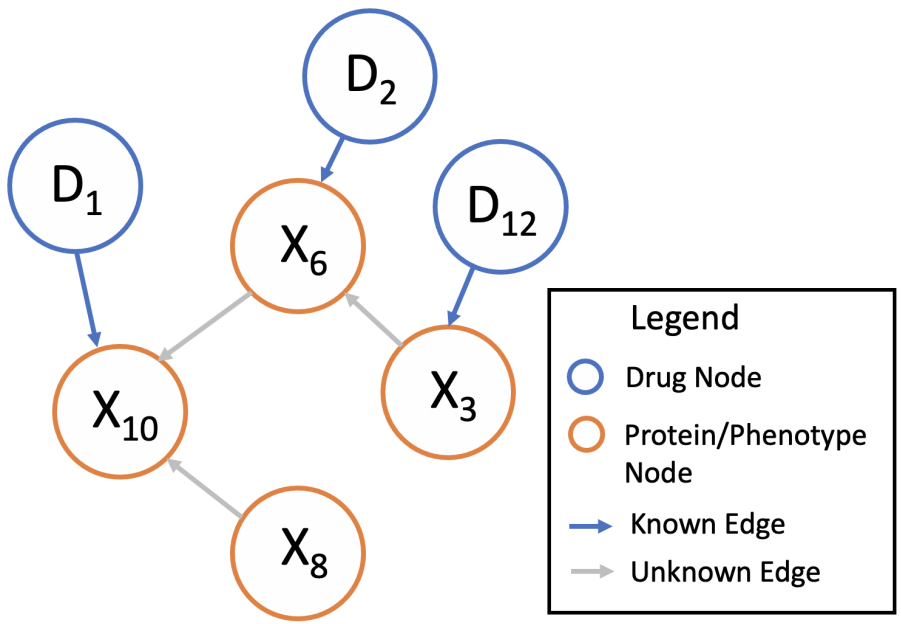
- Many drugs directly target a particular protein
  - AKT inhibitor drug (D_{12}) reduces the expression of AKT protein (X_3)
- Proteins regulate expression of other proteins / phenotypes according to some causal structure.
  - AKT (X_3) effects expression of "downstream" proteins (X_6 and X_{10})



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 arrows)
3. Propagate effect of new drug through the inferred causal structure.



Causal Structure Learning + Prediction

Simple implementation of idea:

- Linear intervention + linear causal structure model:

$$\mathbf{X} = \mathbf{X}A + \mathbf{D}B^T + \epsilon$$

  - $\mathbf{X} \in \mathbb{R}^{n \times p}$ protein responses
  - $\mathbf{D} \in \mathbb{R}^{n \times q}$ drug concentrations
  - $A_{ij} =$ causal effect of X_i on X_j
  - $B_{il} =$ effect of 1 unit of drug l on response X_i

- Causal Structure Learning (CSL):

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

- Predict response

$$\widehat{x} = d^T B^T (I - \widehat{A})^{-1}.$$

Regression versus CSL

With $\lambda = 0$ (no regularization):

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

$$\textbf{CSL: } \widehat{A} = \underset{A}{\operatorname{argmin}} \, ||\mathbf{X} - \mathbf{D}B^T(I - A)^{-1}||_F^2$$

- Regression model estimates total effect of drug on response.
- Causal model estimates direct effect of response variables on each other
  - Requires knowledge of B, direct targets of drugs
  - $\widehat{A}$ may 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.
- **Main Point:** Causal model can predict for left out drug (not used in training) by using other drugs which have same direct protein targets as left out drug.

Outline

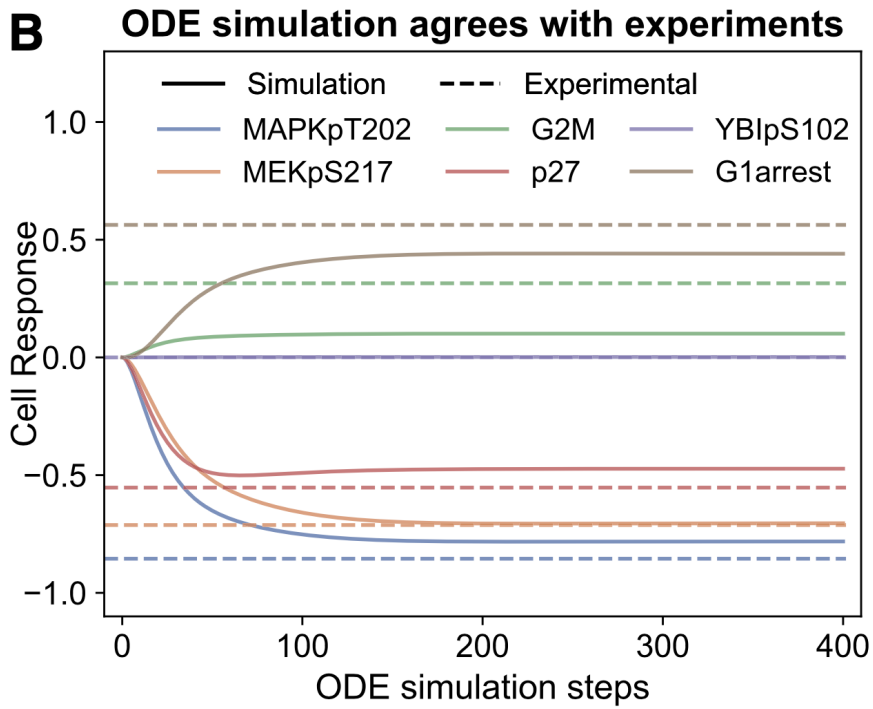
Background on Cellbox

- Cellbox model proposed in Yuan et al. [2021]
- Obtained best prediction performance on RF and LODO
- Presented system of ODEs. No discussion of causality in paper.

Cellbox (Linear Version):

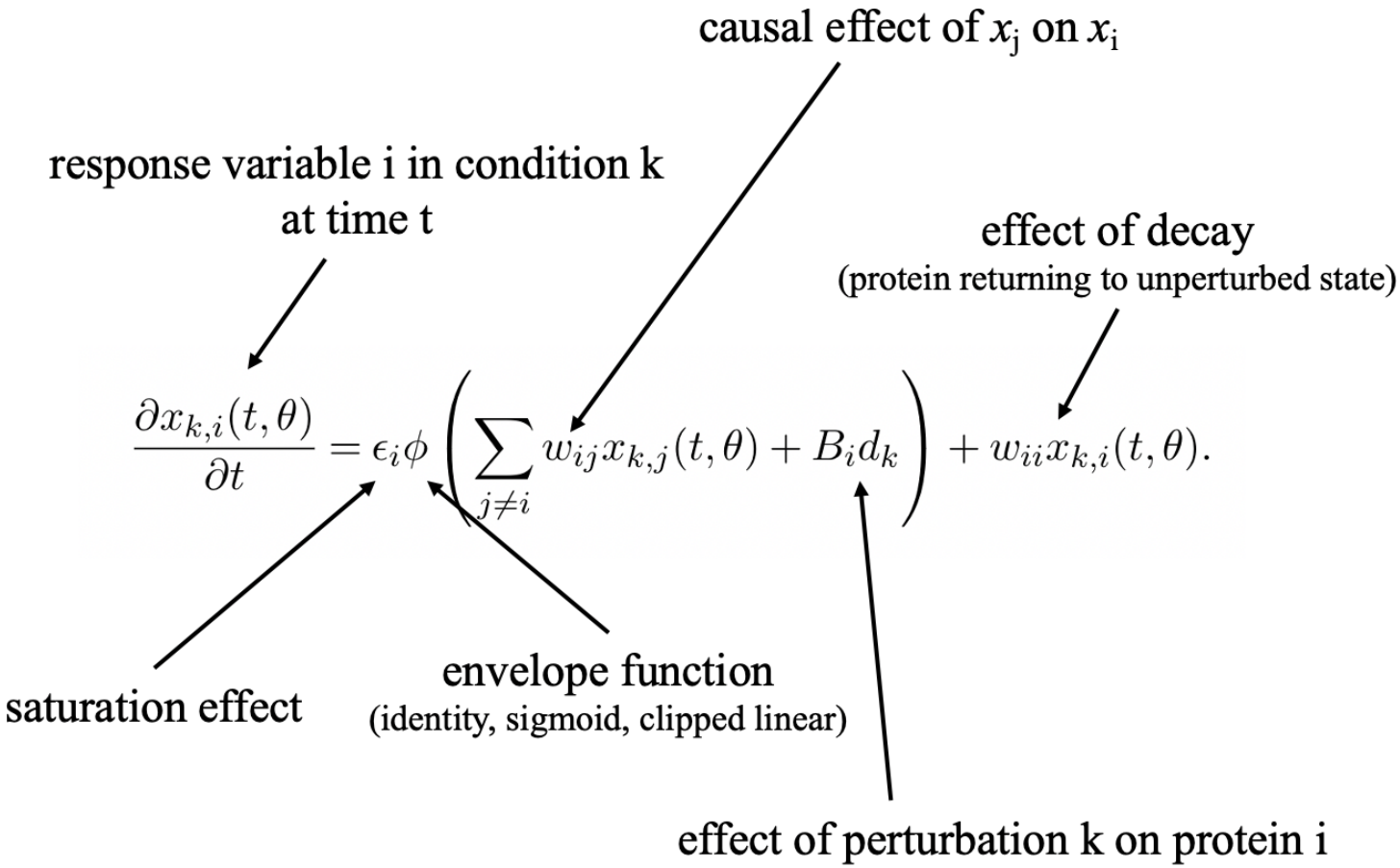
- $W \in \mathbb{R}^{p \times p}$
- $x_k(t, W) =$ predicted response to condition k at time t

$$\frac{\partial x_k(t, W)}{\partial t} = W x_k(t, W) + B d_k$$



Yuan et al. [2021]

Full Cellbox ODE Model



$$\theta = (W, \epsilon)$$

Parameter Estimation

Steady State:

$$x_k(\theta) \equiv \lim_{t \to \infty} x_k(t, \theta).$$

Loss Function:

$$L(\theta) = \sum_k ||x_k - x_k(\theta)||_2^2 + \lambda ||W - diag(W)||_1$$

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.

Cellbox as Causal Structure Learning Model

Theorem

For Linear Cellbox with $W \prec 0$, *steady state is*

$$x_k(\theta) = -W^{-1}Bd_k$$

and

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

Proof Idea.

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

□

Comments on Result

- Without regularization the models are:

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

$$\textbf{CSL: } \widehat{A} = \underset{A}{\operatorname{argmin}} \, ||\mathbf{X} - \mathbf{D}B^T(I - A)^{-1}||_F^2$$

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

- Linear Cellbox is equivalent to CSL model with $W = -(I - A)^T$.
- Similar results connecting dynamical systems and causal graphs with cycles [Lacerda et al., 2012, Dash, 2005]

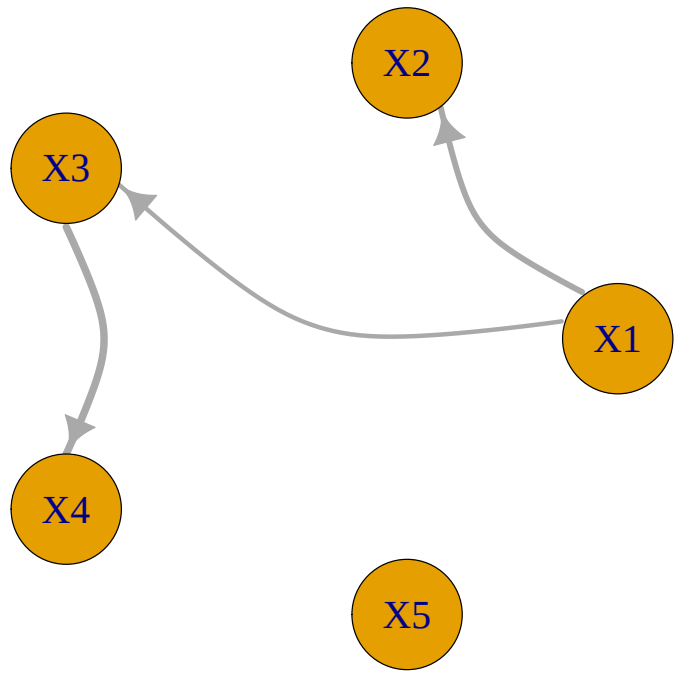
Outline

Outline

Simulation 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
- $\epsilon \in \mathbb{R}^{105 \times 5}$, all elements independent distributed $N(0, 0.1^2)$

$$\mathbf{X} = \mathbf{X}A + \mathbf{D}B^T + \epsilon$$

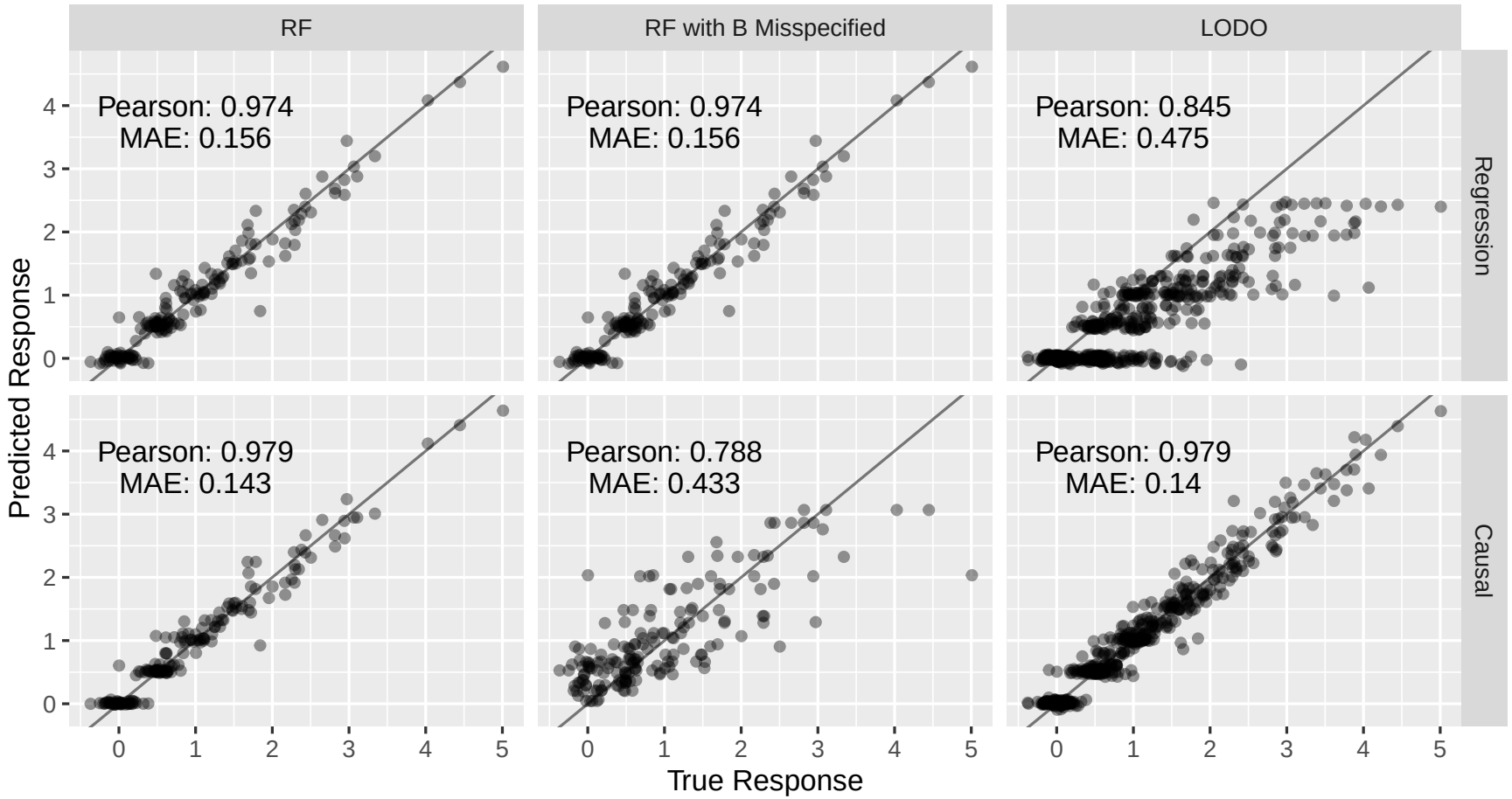


True Causal DAG A

Compare Regression with CSL in 3 Settings

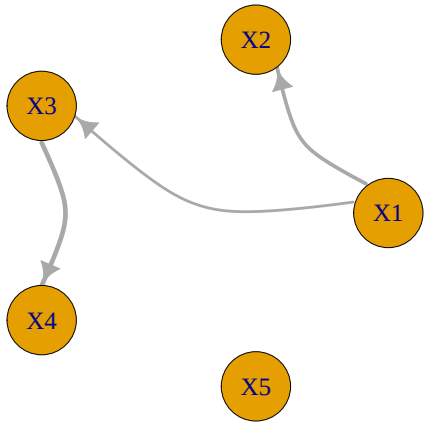
1. **Random Fold (RF):**
   - Data is divided randomly into 2/3 training and 1/3 test
2. **RF with B Misspecified:**
   - Training–test set split is identical to RF.
   - B matrix (direct effect of drugs) is misspecified.
   - 10 drugs with 2 targets are assumed to influence their targets with a strength of 1
3. **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

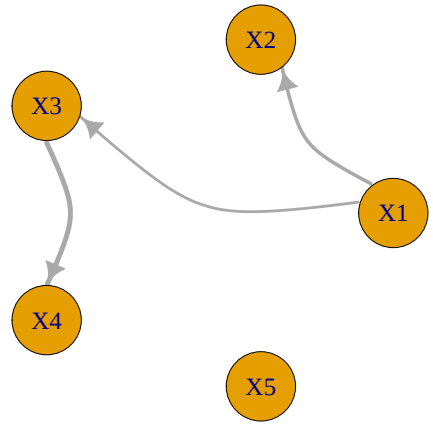


- **RF:** Regression and Causal model both obtain good performance
- **RF with B Misspecified:** Causal model has poor performance
  - Regression is completely robust to misspecification
- **LODO:** Regression has poor performance

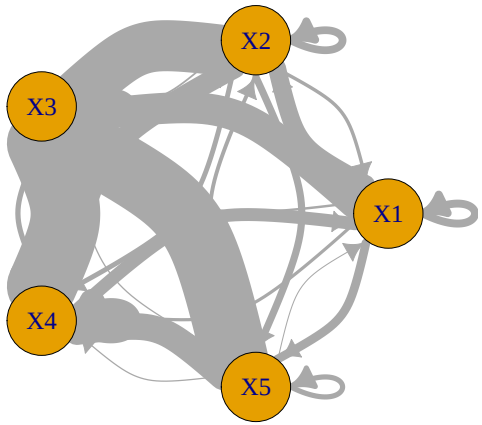
Estimated Causal Graphs using CSL Model



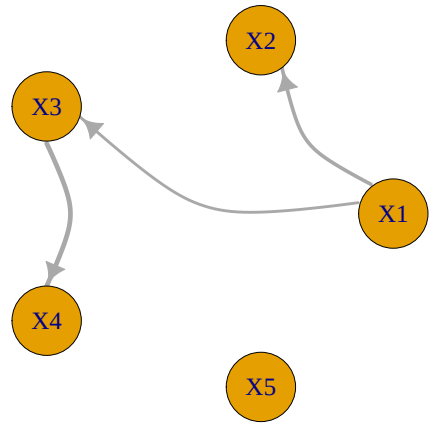
a) True Graph



b) Random Fold



c) Random Fold, Misspecified B

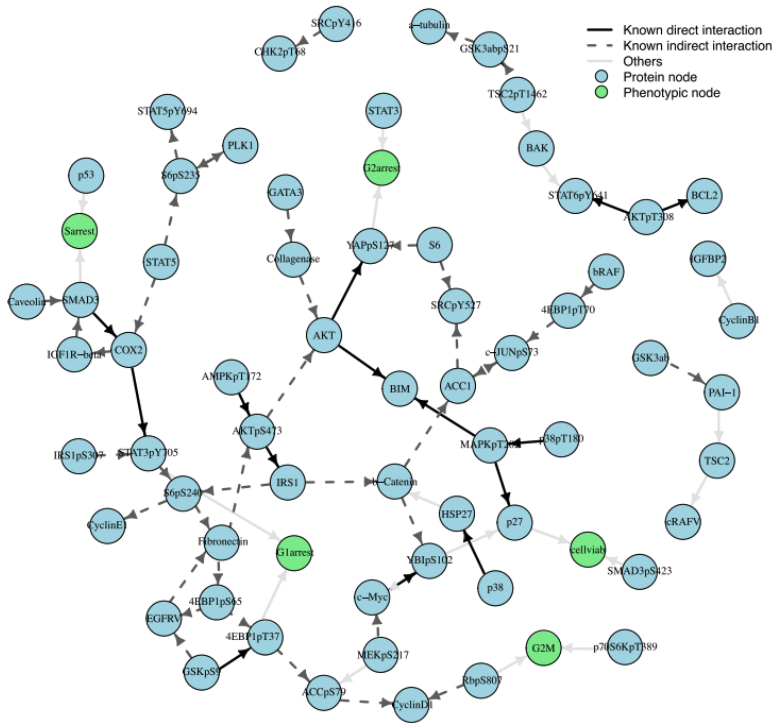


d) LODO

Outline

Background

- Data collected in Korkut et al. [2015]
- 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

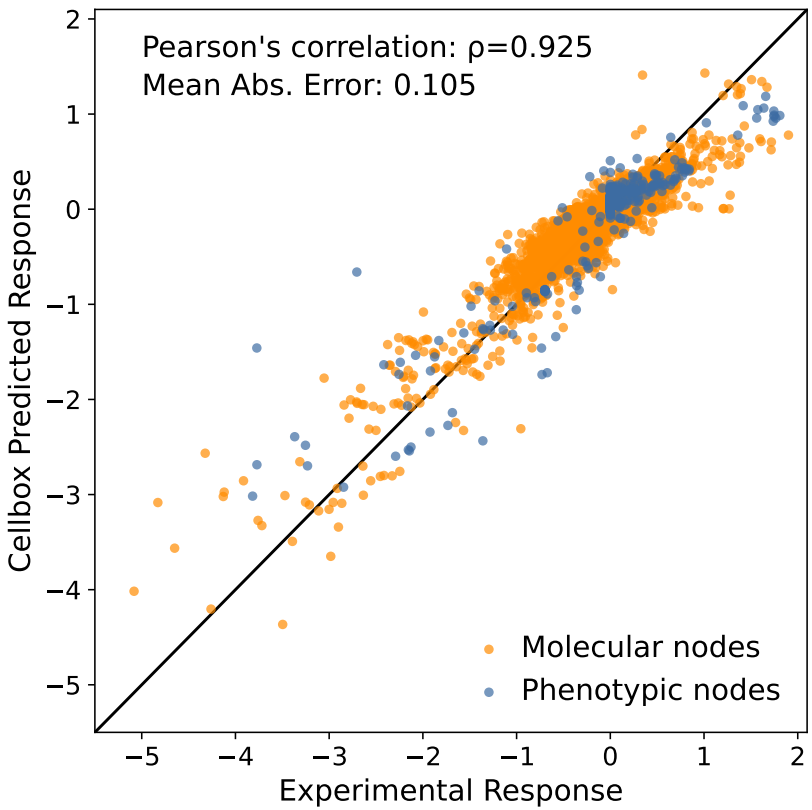


Graph learned by Cellbox

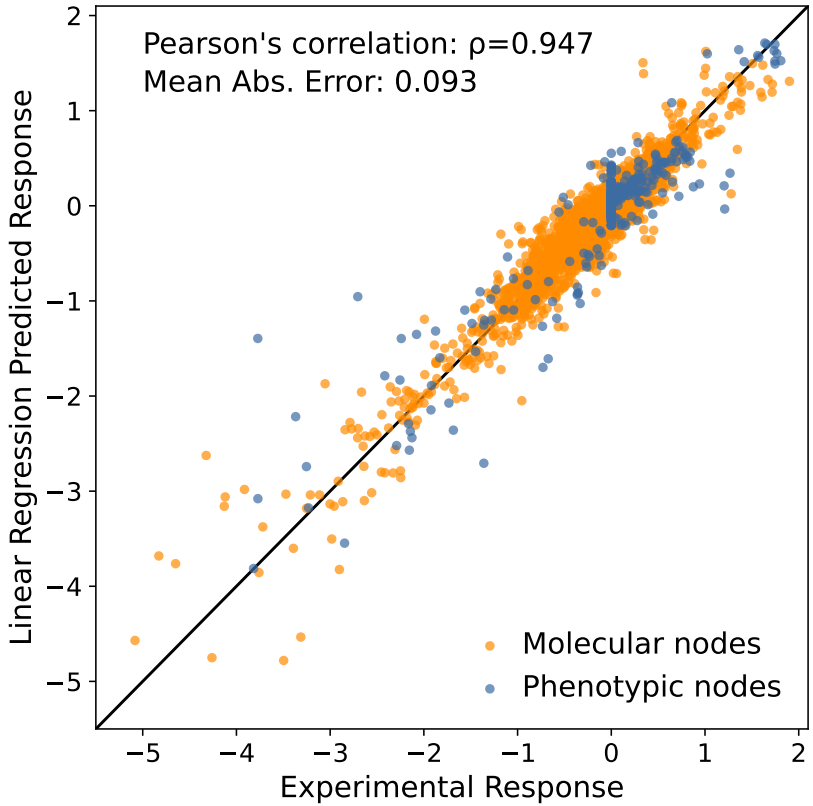
Random Fold Validation Setup

- 70% training / 30% testing data split
- Repeated 1000 times
- ≈ 300 (0.3×1000) predictions for each condition
- Average predicted responses for each condition.
- Compare predictions and experimentally observed responses
  - $n \times 87$ predictions

Random Fold Validation Results



Cellbox



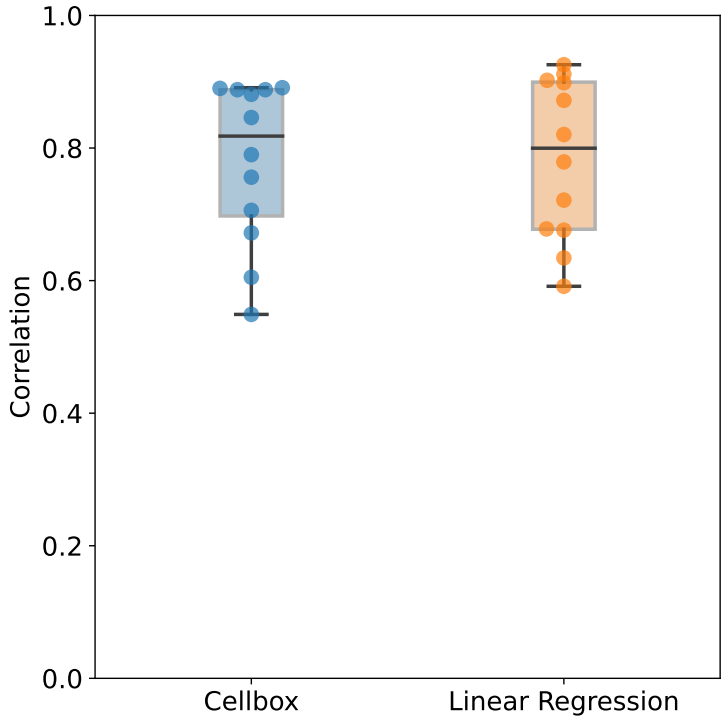
Linear Regression

Linear regression outperforms Cellbox in RF validation.

LODO Setup

- For drug $i \in \{1, \ldots, 12\}$
  - Training set is all conditions where drug i not used
  - Test set is all other conditions
  - For linear regression, set effect of drug i 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

- Evaluated causal model (Cellbox) using prediction performance
  - Traditionally causal inference focuses on parameter estimation. Conclusions sensitive to difficult to verify confounding assumptions.
  - Prediction performance is more objective / unbiased.
- Linear regression achieved best prediction performance on Melanoma data set.
  - Message: Start with Linear Regression when modeling.
  - Reasons for Poor Performance of Causal Model:
    - Only 12 drugs and ~ 90 response variables. Very challenging to estimate causal structure.
    - Direct effect of drugs (B matrix) misspecified.

Summary

- Causal structure learning requires more assumptions, but can predict in LODO validation where regression fails.
  - Focused on linear modeling case, but conclusions apply to non–linear models.
- Future / ongoing work: Replace A learning model with causal discovery algorithm (e.g. LiNGAM) or hybrid (LiNGAM+Intervention).

$$\mathbf{X} = \mathbf{X}A + \mathbf{D}B^T + \epsilon$$

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

D. Dash. Restructuring dynamic causal systems in equilibrium. In *International Workshop on Artificial Intelligence and Statistics*, pages 81–88. PMLR, 2005.

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.

G. Lacerda, P. L. Spirtes, J. Ramsey, and P. O. Hoyer. Discovering cyclic causal models by independent components analysis. *arXiv preprint arXiv:1206.3273*, 2012.

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.