Northeastern University
Khoury College of Computer and Information Sciences

Causal inference enables the estimation of outcomes of interventions from observational mass spectrometry (MS)-based proteomics experiments

Devon Kohler[1,2], Karen Sachs[3,4,5], Jeremy Zucker[6], Benjamin M. Gyori[1,2], Lindsay Pino[7], Olga Vitek[1,2]

Presentation outline

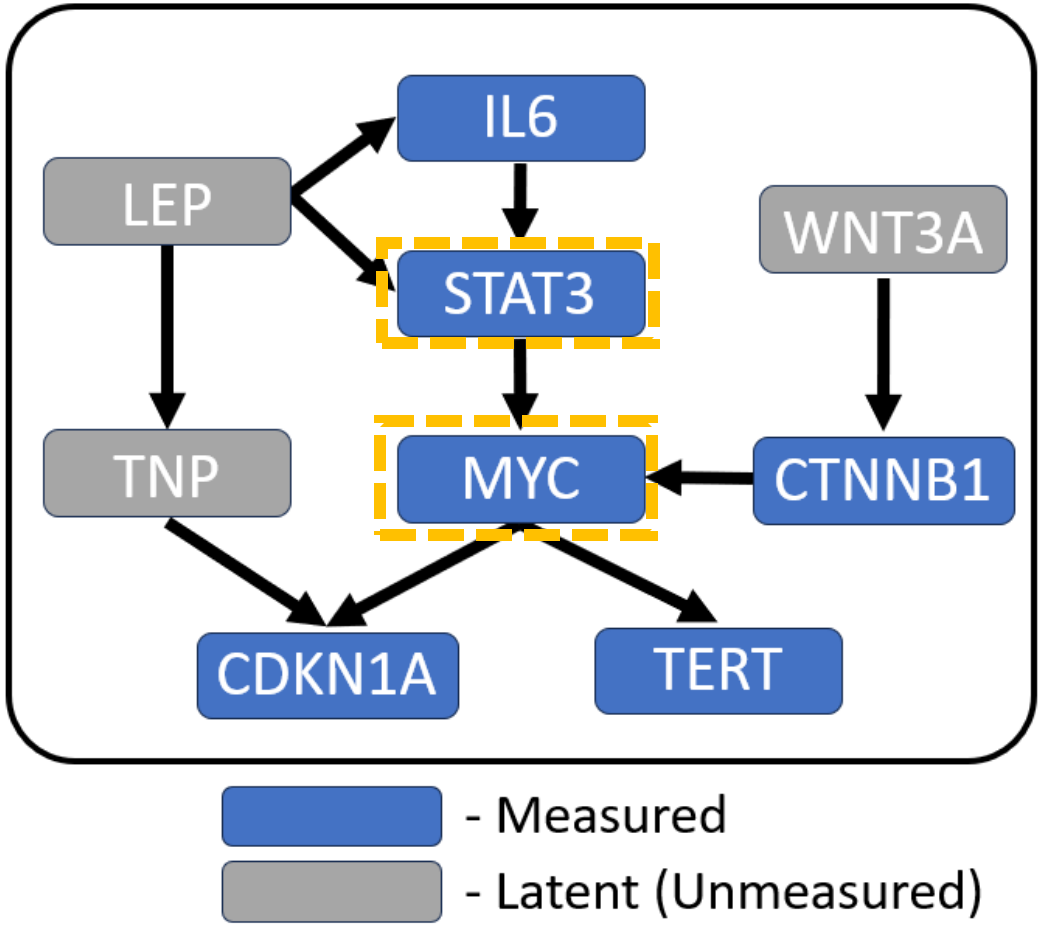
- Motivation
- Methods and implementation
- Results
  - Simulation: Classical ML
  - Simulation: Causal ML
  - Biological experiment: Chromatin-binding activity of transcription factors

Understanding the proteome response to perturbations is important in understanding protein function

- This can be done experimentally using external mechanisms such as drugs but can be challenging and expensive
- Using machine learning methods, we can try to predict the effect of the perturbation without physically performing it

Causal inference allows us to estimate the proteome response to perturbations from purely observational data

- Causal network (in the form of a directed acyclic graph (DAG))
- Graph is expert and literature-derived
- Observational experimental data

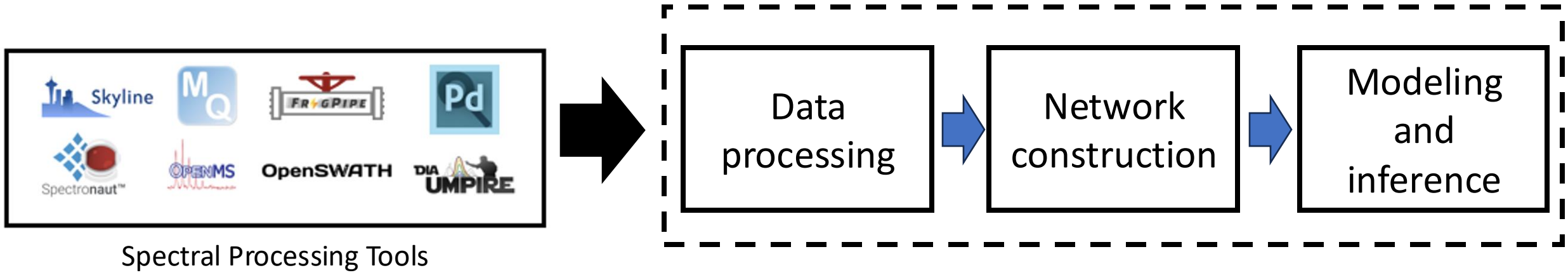


Presentation outline

- Motivation
- Methods and implementation
- Results
  - Simulation: Classical ML
  - Simulation: Causal ML
  - Biological experiment: Chromatin-binding activity of transcription factors

Method broken down into three main steps

- Takes as input the output of spectral processing tools used for identification and quantification
- Three main modeling steps



Spectral Processing Tools

Data processing required to input reasonable data into model

- Batch effect correction (normalization)
- Missing value imputation (on the fragment/precursor-level)
- Summarization to protein-level

Relative quantification of proteins and of post-translational modifications with shared peptides: a weight-based approach

Mateusz Staniak

P-III-0835

Network Construction

INDRA
Integrated Network and Dynamical Reasoning Assembler

- Manually curated or automated
- Automation can be done with biological knowledge databases
  - INDRA (Integrated Network and Dynamical Reasoning Assembler)

Integrating MSstats with INDRA for enhanced interpretation of proteomic differential analysis results

Anthony Wu

P-III-0847

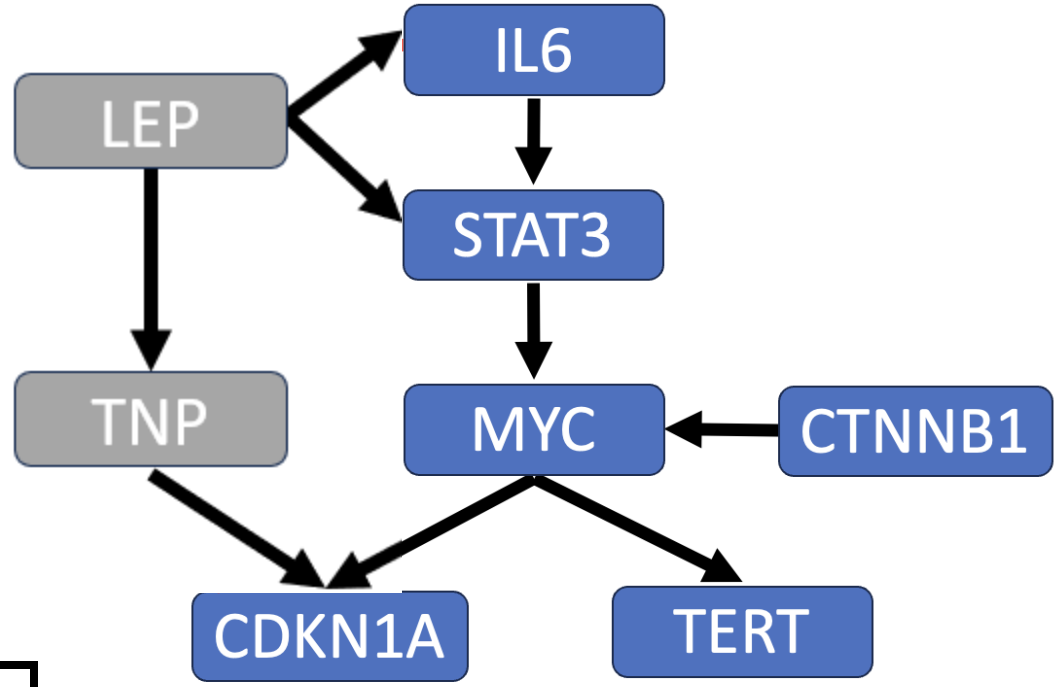
Results

I found statements that are not only from medscan, have an agent where HGNC=7553, and have type Activation

MYC affects **apoptotic process** 2	1 | 4 7 2002
MYC affects **cell population proliferation**	25 3 1921
MYC affects **transcription, DNA-templated**	6 216 688
MYC affects **cell cycle**	2 1 | 9 2 427
MYC affects **cell growth**	1 | 7 336
MYC affects **cell differentiation**	1 1 337
MYC affects **glycolytic process**	1 | 15 4 292
MYC affects **TP53**	1 28 223
MYC activates **TP53**.	1 28 223

Implement latent variable model (LVM)

- Leverage Pyro, probabilistic programming language, which can encode causal relationship between proteins
- Include latent variables
- Bayesian model
  - Informed priors learned from literature



Example System of Linear Equations

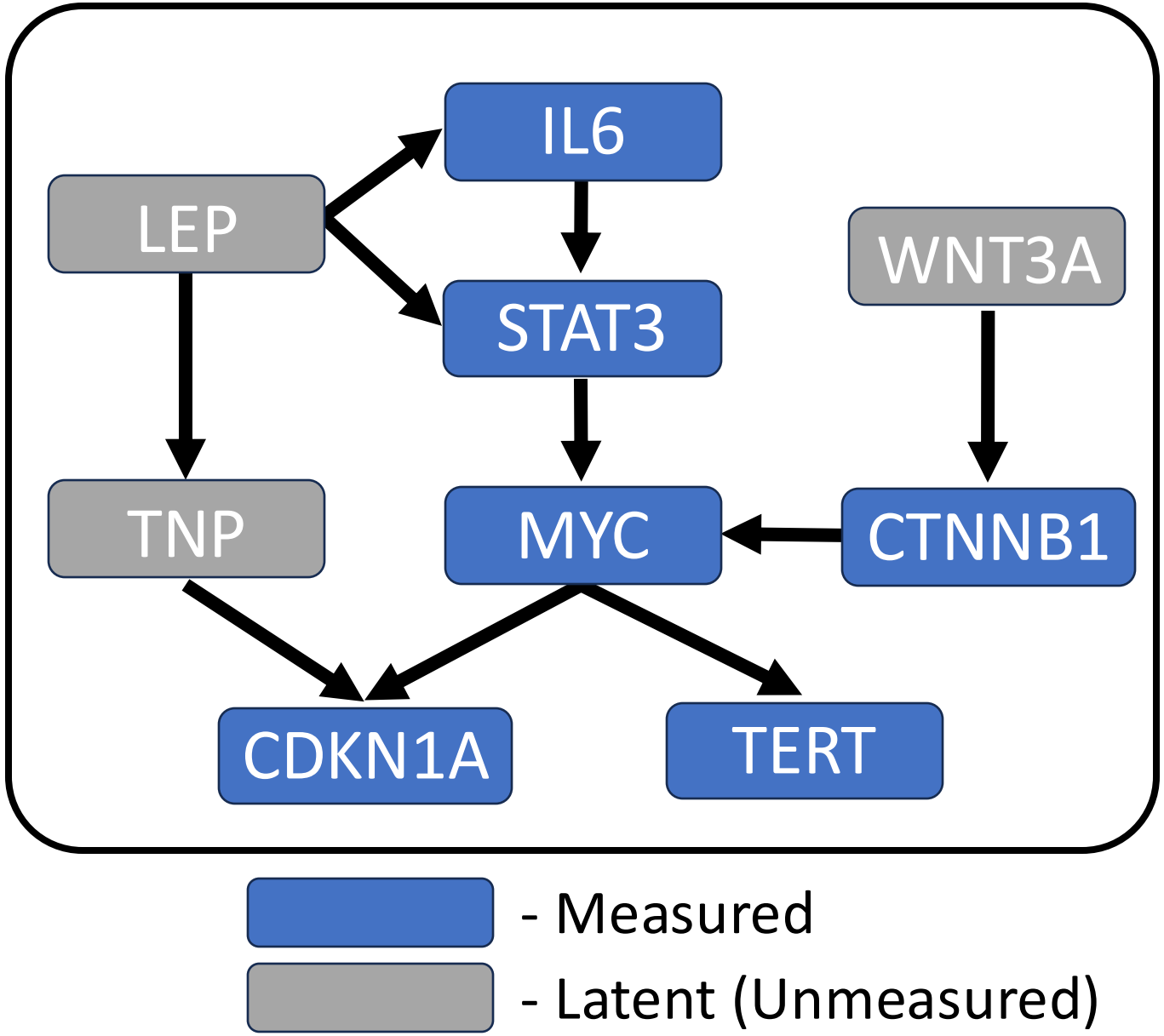
$$P(STAT3) = N(\beta_{STAT_0} + \beta_{STAT3_1} * IL6, \sigma^2_{STAT3})$$
$$P(MYC) = N(\beta_{MYC_0} + \beta_{MYC_1} * STAT3 + \beta_{MYC_2} * CTNNB1, \sigma^2_{MYC})$$
$$P(TERT) = N(\beta_{TERT_0} + \beta_{TERT_1} * MYC, \sigma^2_{TERT})$$

Presentation outline

- Motivation
- Methods and implementation
- Results
  - Simulation: Classical ML
  - Simulation: Causal ML
  - Biological experiment: Chromatin-binding activity of transcription factors

Biological Network – MYC transcription factor pathway

- Simulate observational data with causal structure and experimental properties
- Simulate perturbational data for validation



Classic machine learning approach

- Fit model with MYC as the outcome variable
- Remaining measured proteins as predictors
- Find proteins that have a high association with MYC

Linear Regression

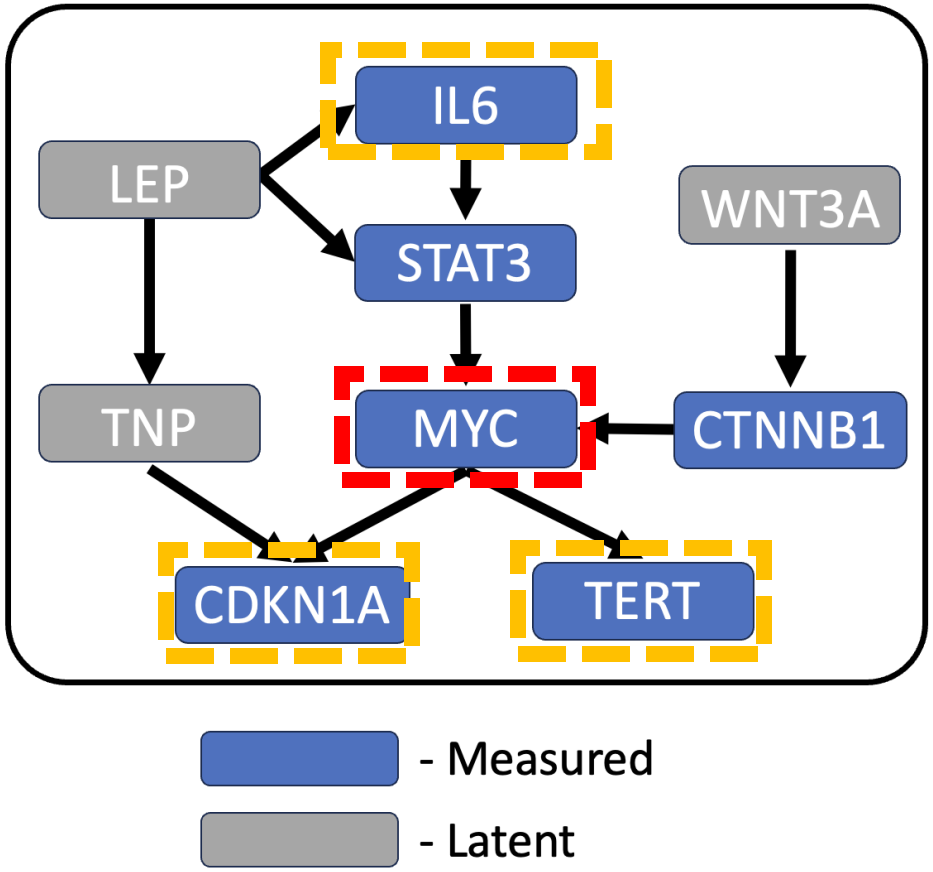
$$MYC_i = \mu_i + IL6_i + STAT3_i + CTNNB1_i + CDKN1A_i + TERT_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Classic ML methods cannot capture network relationships

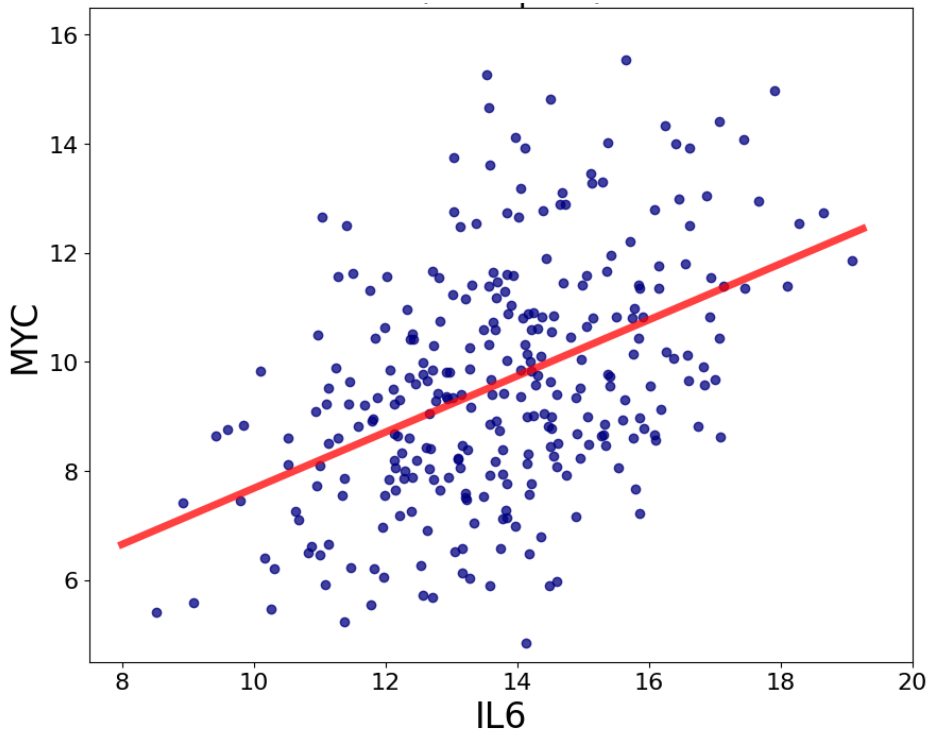
Method	MYC-Inhibition Candidate	MYC-Inhibition Candidate
Linear Regression	TERT	CDKN1A
Random Forest	TERT	IL6
XGBoost	CDKN1A	IL6

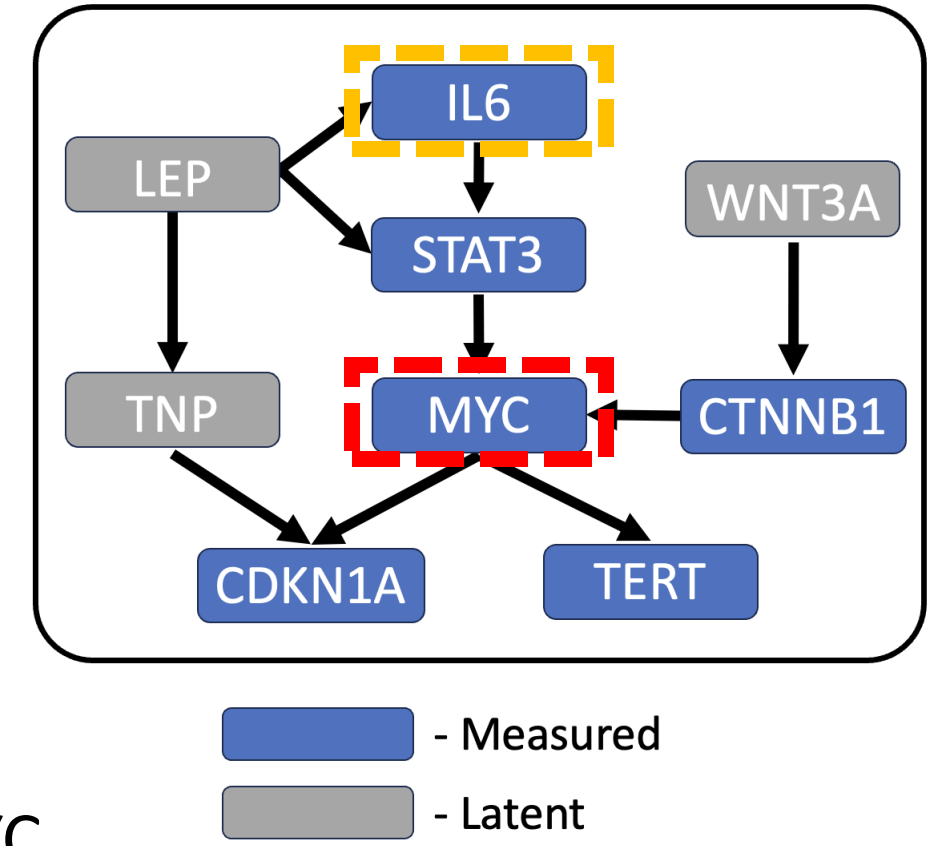
- TERT and CDKN1A are downstream of MYC
- IL6 is upstream of MYC but is confounded by LEP, biasing the results



Classic ML methods cannot identify bias caused by latent confounding

MYC and IL6 appear to be positively correlated

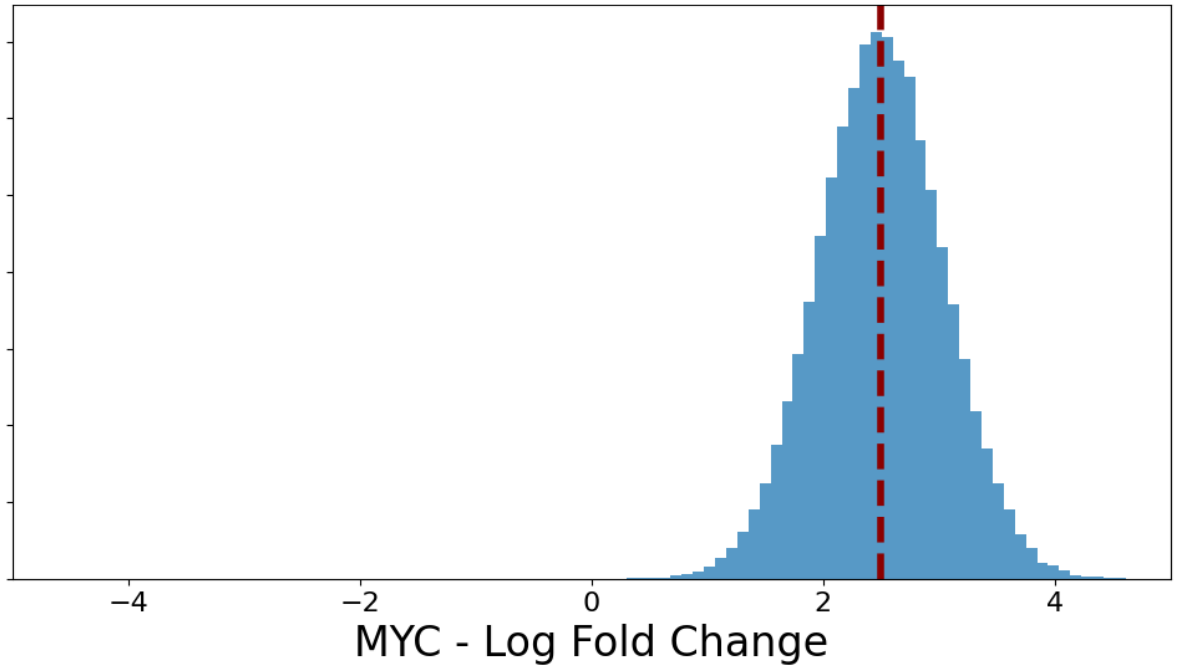


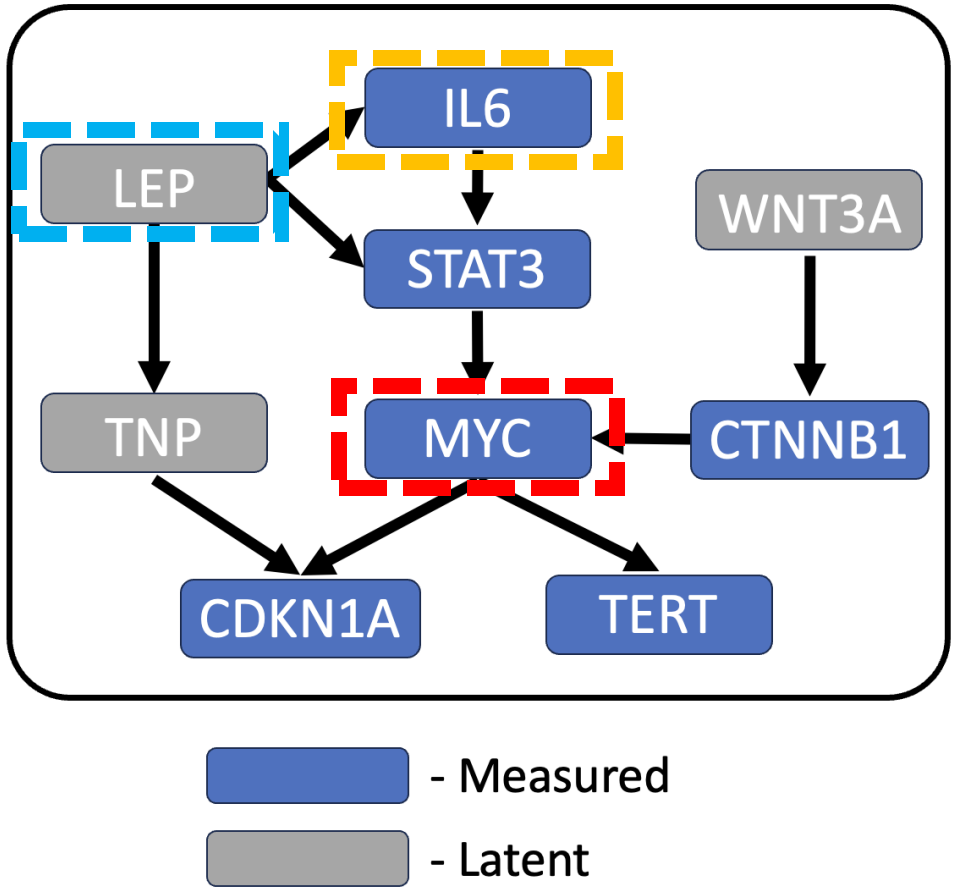


LEP is a latent confounder between IL6 and MYC

Classic ML methods cannot identify bias caused by latent confounding

MYC interventional distribution after inhibiting IL6 by 2 log FC





Inhibiting IL6 causes MYC to increase

Presentation outline

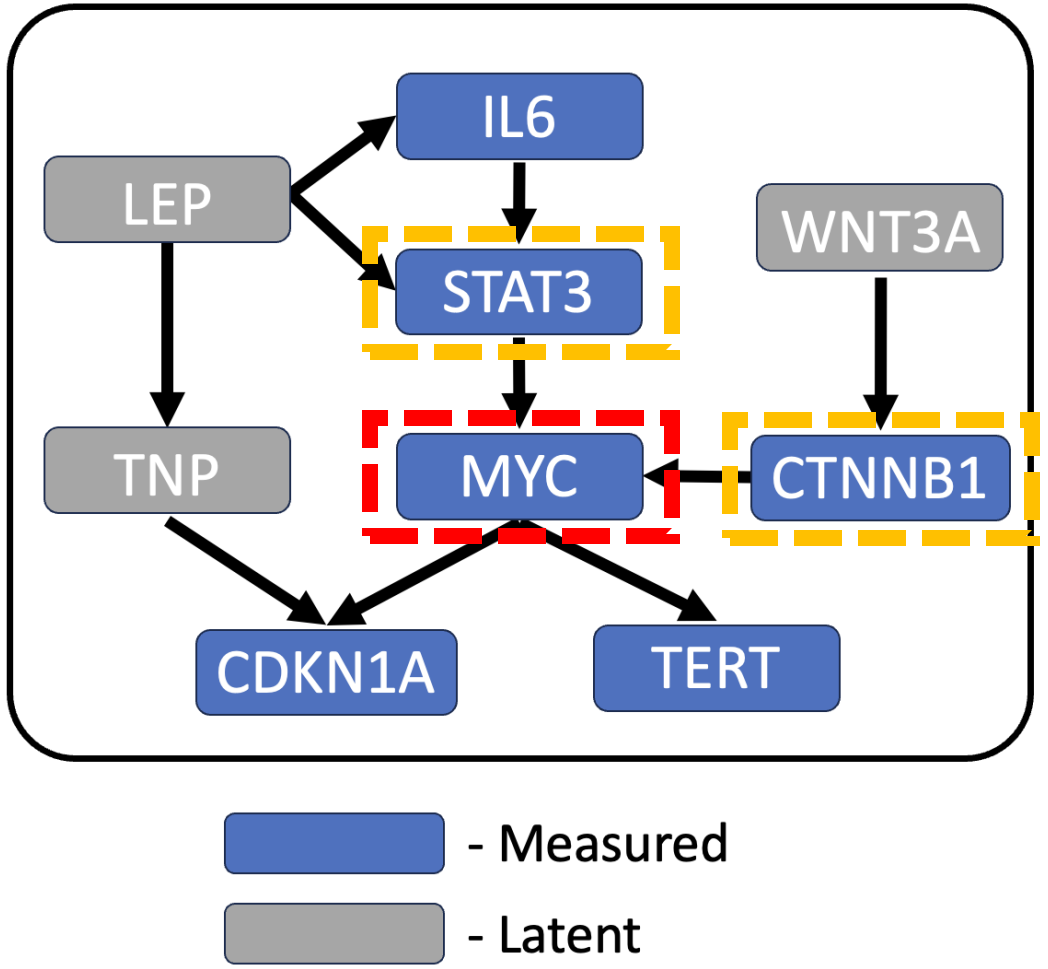
- Motivation
- Methods and implementation
- Results
  - Simulation: Classical ML
  - Simulation: Causal ML
  - Biological experiment: Chromatin-binding activity of transcription factors

Causal inference can overcome these problems

- Specifically encode causal relationship between proteins into the model
- Use a latent variable model (LVM) to specify unmeasured proteins
- Confirm whether questions can be answered (i.e., if they are identifiable) in a non-biased way

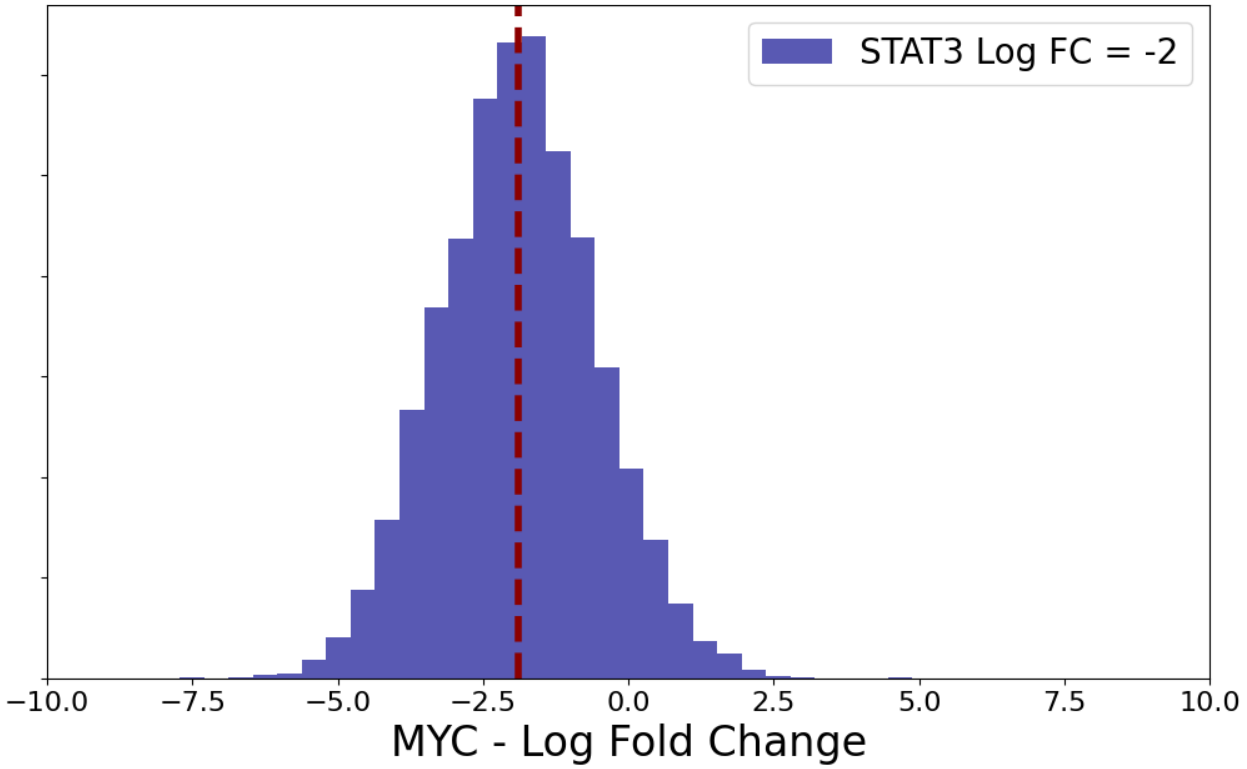
Causal inference suggested drug targets

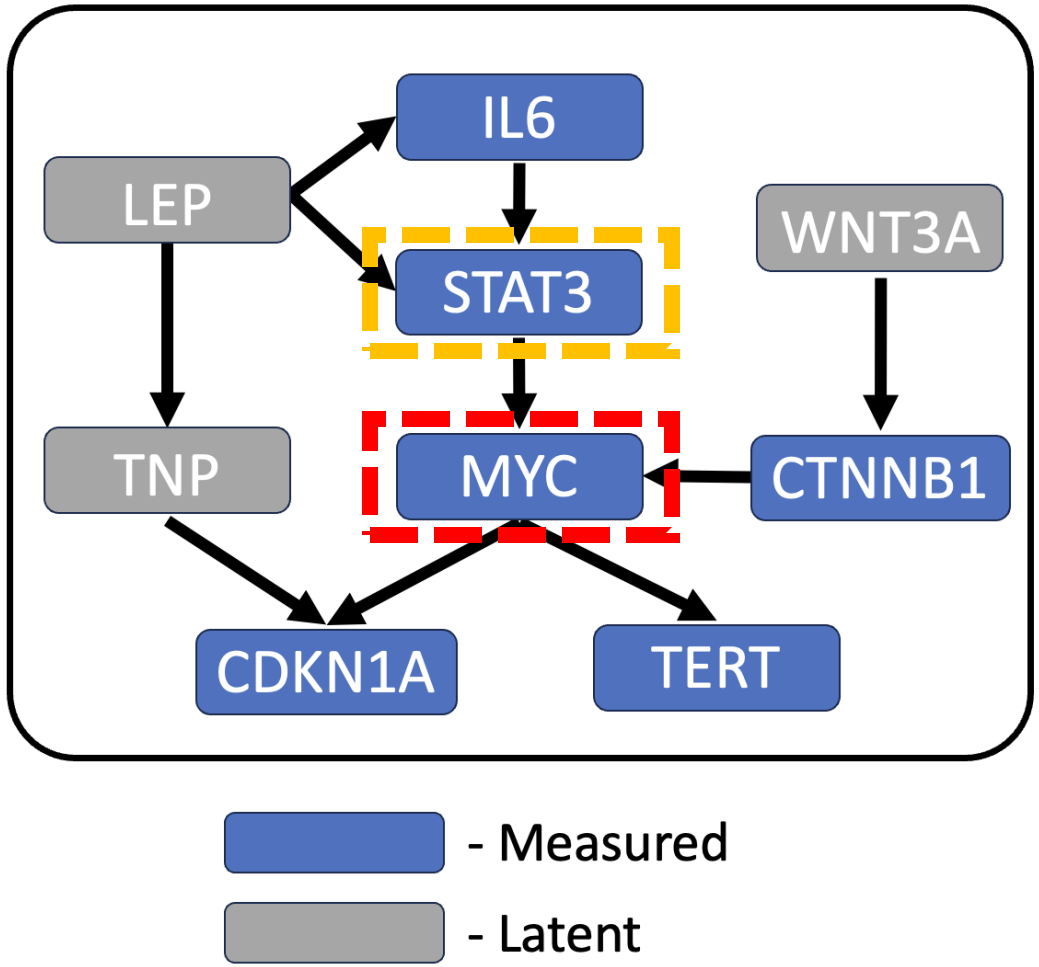
- STAT3 and CTNNB1 are identifiable and upstream of MYC
- Potential MYC-Inhibition candidates



STAT3 shows strong affect on MYC

MYC interventional distribution after inhibiting STAT3 by 2 log FC





STAT3 would be a good follow up candidate

Presentation outline

- Motivation
- Methods and implementation
- Results
  - Simulation: Classical ML
  - Simulation: Causal ML
  - Biological experiment: Chromatin-binding activity of transcription factors

Biological experiment: Chromatin-binding activity of transcription factors – Talus Bioscience

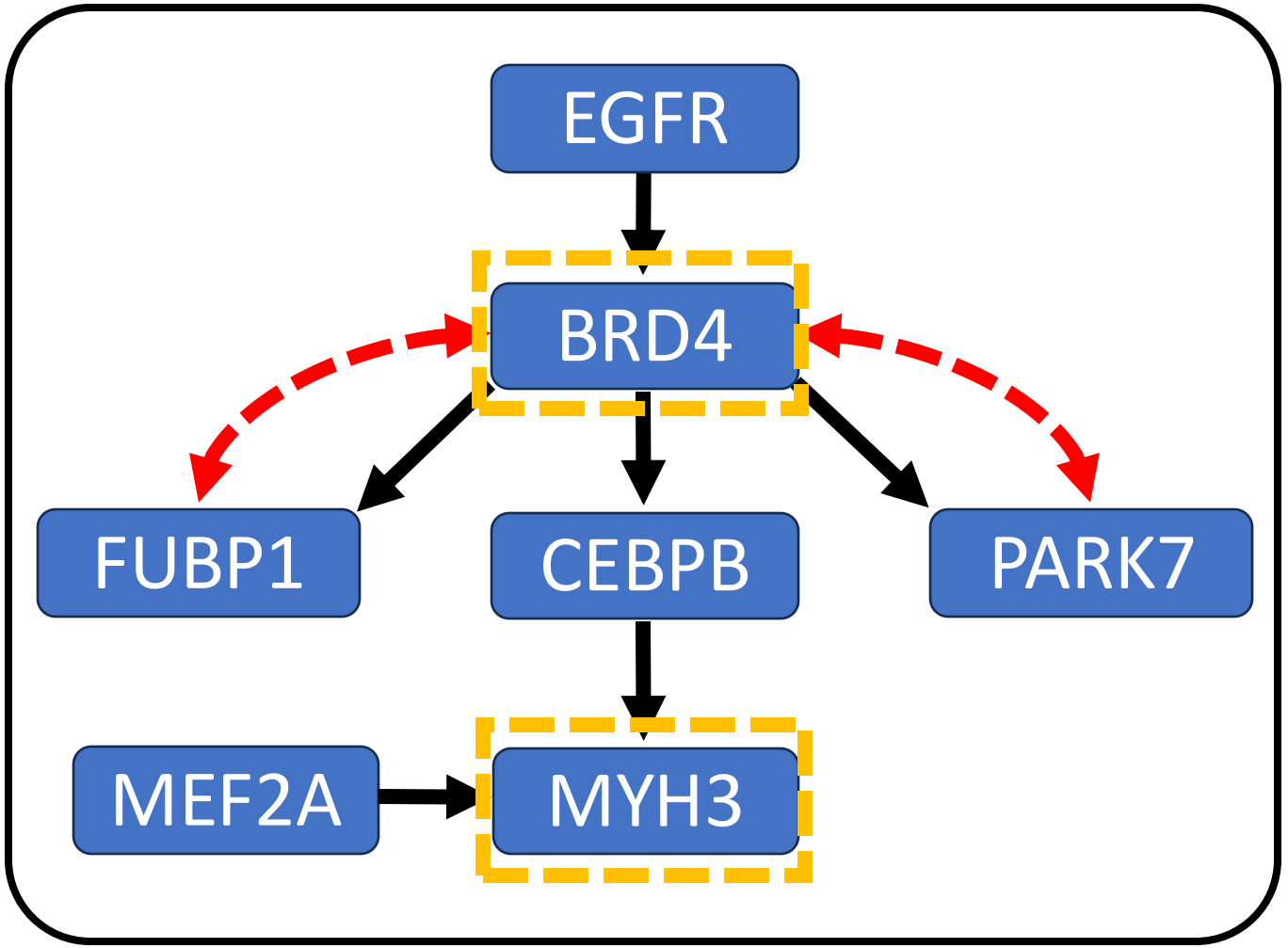
- Effect of eight drug compounds on chromatin-binding activity of transcription factors

Treatment	Sample Type	Replicates
DMSO	Observational	132
dBET6	Interventional	17

- dBET6 known to inhibit BRD proteins
- Can we predict the effect of inhibiting BRD proteins, using only DMSO replicates to train the model?

Objective: estimate the interventional effect of BRD4 on MYH3

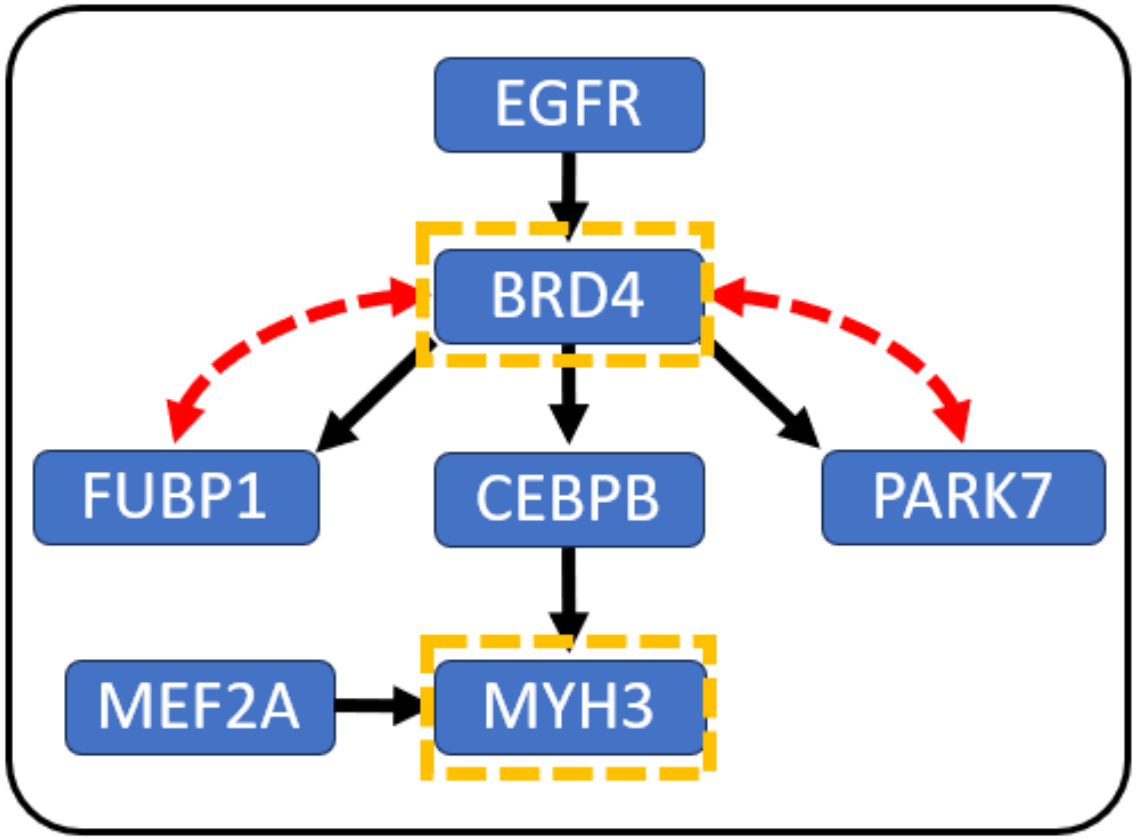
- Network extracted from INDRA
- Model trained on DMSO to learn relationships between proteins



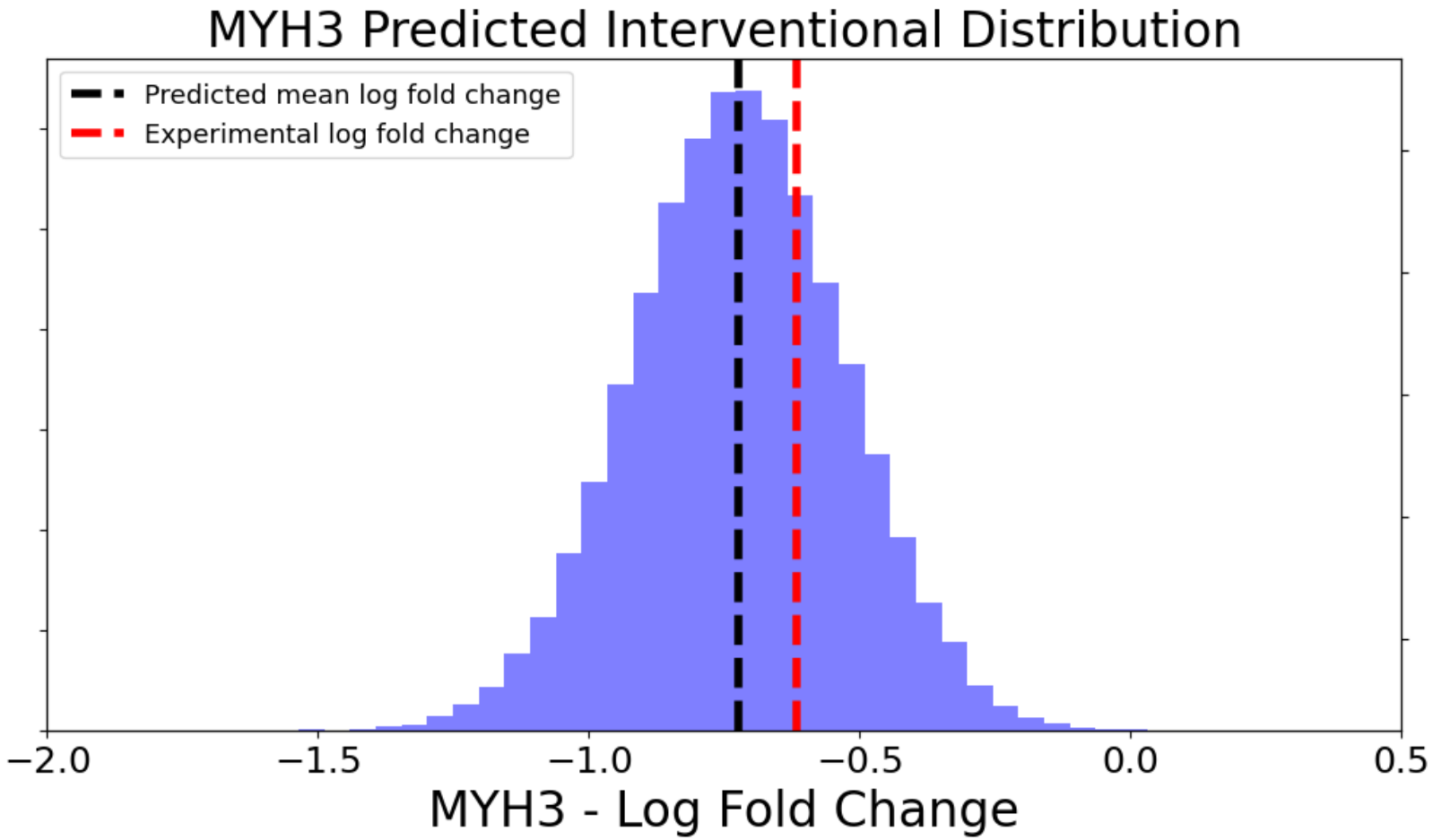
Objective: validate the model estimate with dBET6 molecule

Protein	Experimental log_2 fold change (DMSO - dBET6)
BRD4	2.669
MYH3	-.617

- Experimentally, when BRD4 is inhibited, MYH3 increases
- Validate if model estimates the same response



Experimentally measured MYH3 fold change falls within the estimated distribution



Conclusions

- Causal inference is better suited for the estimation of perturbational effects compared to traditional ML algorithms
- We were able to correctly predict the causal effect both in simulations and a real-world study
- Challenges remain, such as building an accurate graphical network

Acknowledgements

Northeastern University

OLGA VITEK LAB

Statistical Methods For Studies Of Biomolecular Systems

Northeastern

Olga Vitek
Benjamin Gyori
Sara Mohammad Taheri
Charlie Hoyt
Anthony Wu

Talus Bioscience

Daniele Canzani
Julia E Robbins
Andrea I Gutierrez
William E Fondrie
Alexander J Federation
Lindsay K Pino

NextGen Analytics

Karen Sachs

PNNL

Jeremy Zucker