US HUPO 2023
Chicago, IL | March 4-8

Single-cell mass spectrometry-based proteomics enables causal inference in observational studies

Devon Kohler[1], Karen Sachs[2;3;4], Charles Tapley Hoyt[5], Benjamin M. Gyori[5], Jeremy Zucker[6], Olga Vitek[1]

1. Khoury College of Computer Sciences, Northeastern University, Boston, MA;
2. Answer ALS, Washington, DC; 3. Next Generation Analytics, Palo Alto, CA; 4. Modulo Bio, Inc. San Diego CA;
5. Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA; 6. Pacific Northwest National Laboratory, Richland, WA;

Poster P8.07

Abstract & Motivation

Single cell proteomics greatly increases the number of replicates and cellular resolution in MS-based proteomic experiments. However, these experiments are expensive (especially when targeting perturbations). Causal inference methods, which are typically challenging to apply to traditional bulk-MS due to the lack of replicates, allow us to estimate the impact of perturbations from purely observational data. We propose a method and workflow for predicting the effect of interventions in observational single cell MS experiments and apply the workflow to a recent observational single cell experiment[1].

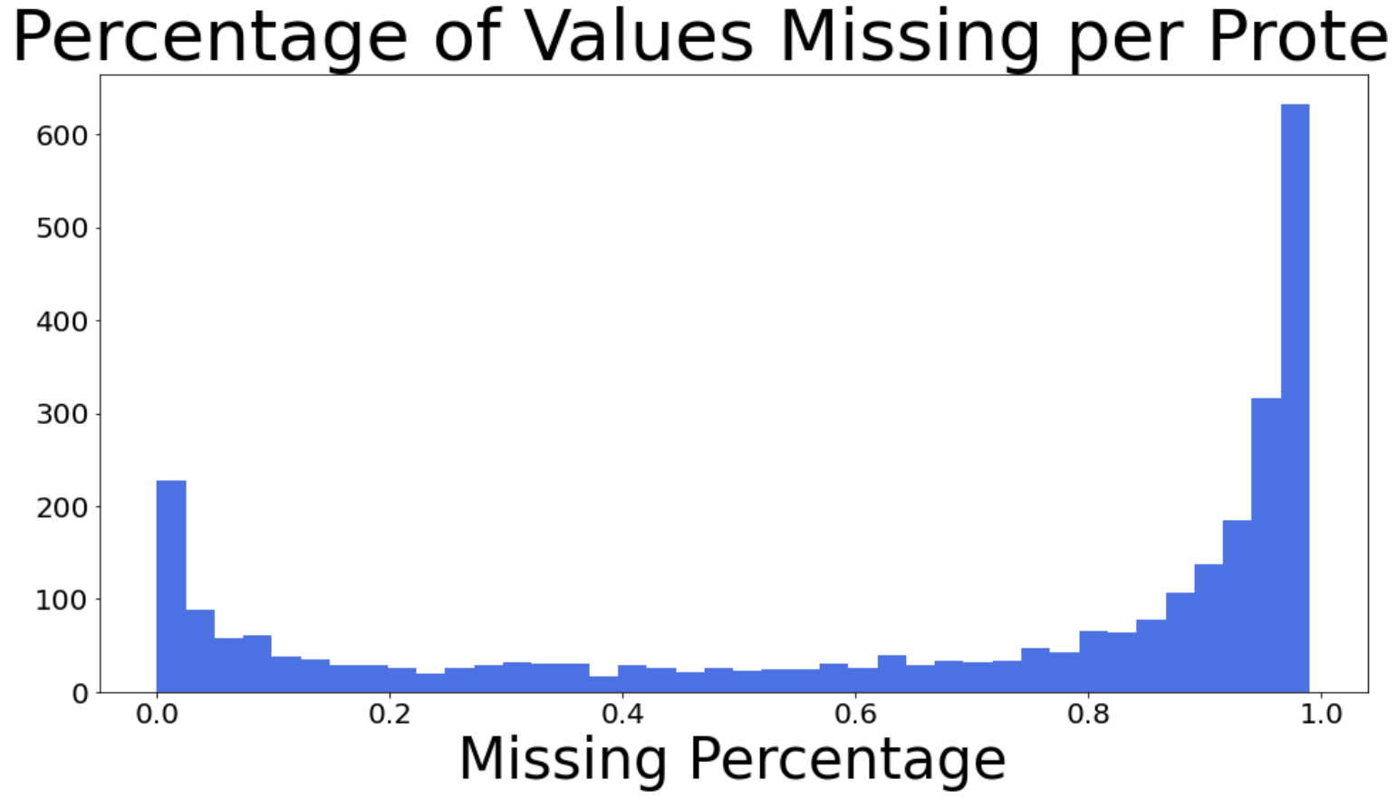
Upstream processing of quantified data impacts downstream results[2]

1. Batch effect correction

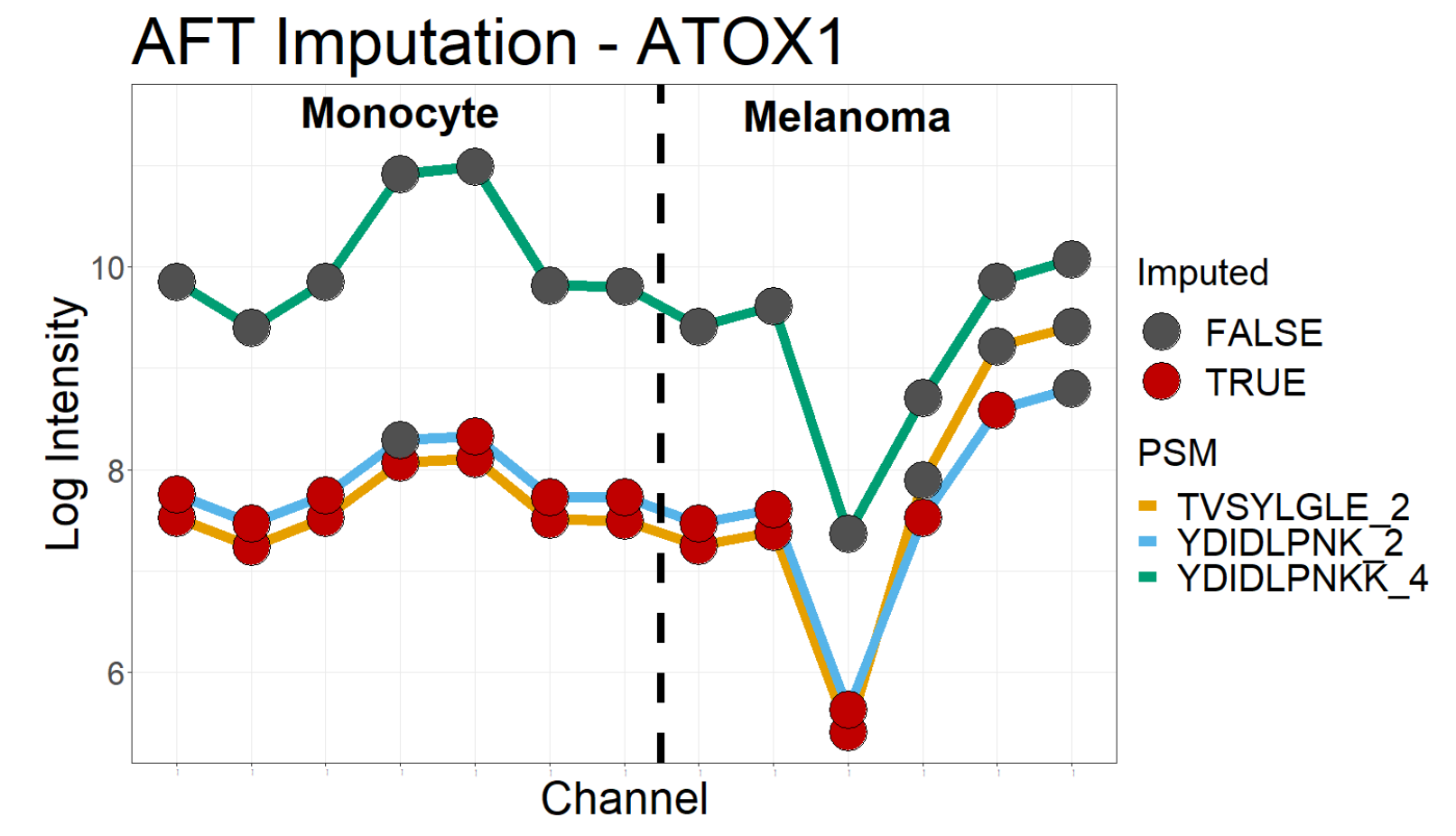
Many runs are required to measure large numbers of cells. Measuring over multiple runs creates technical artifacts that need to be corrected for. In tandem mass tag (TMT) experiments, this can be done by leveraging a reference normalization channel.

2. Missing value imputation

Percentage of Values Missing per Protein



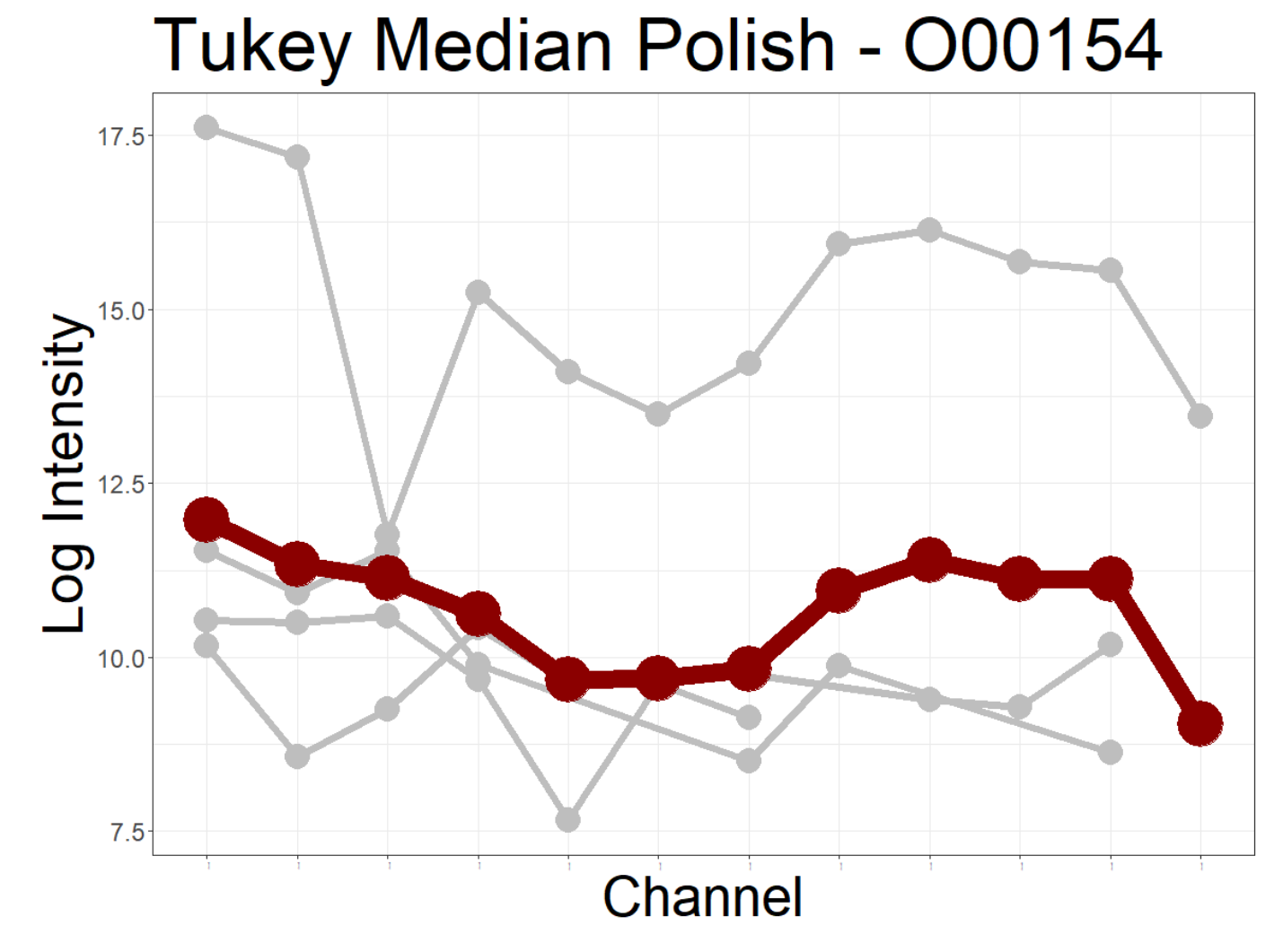
AFT Imputation - ATOX1



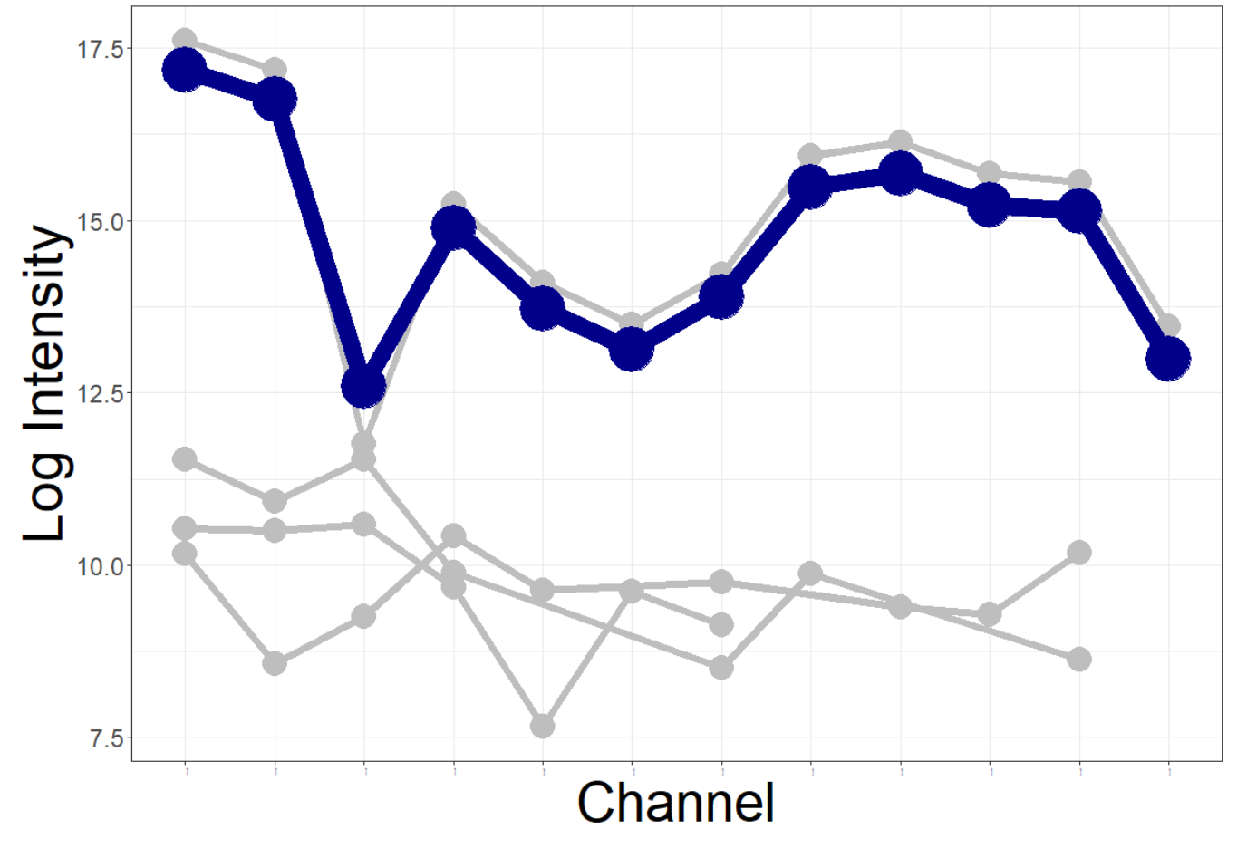
Missing values are a major challenge in single-cell experiments. Naïve imputation methods can create data that is homogenous, masking between cell variation. Here, we use an Accelerated Time Failure (AFT) model which only imputes when there is enough evidence to do so.

3. Feature level summarization

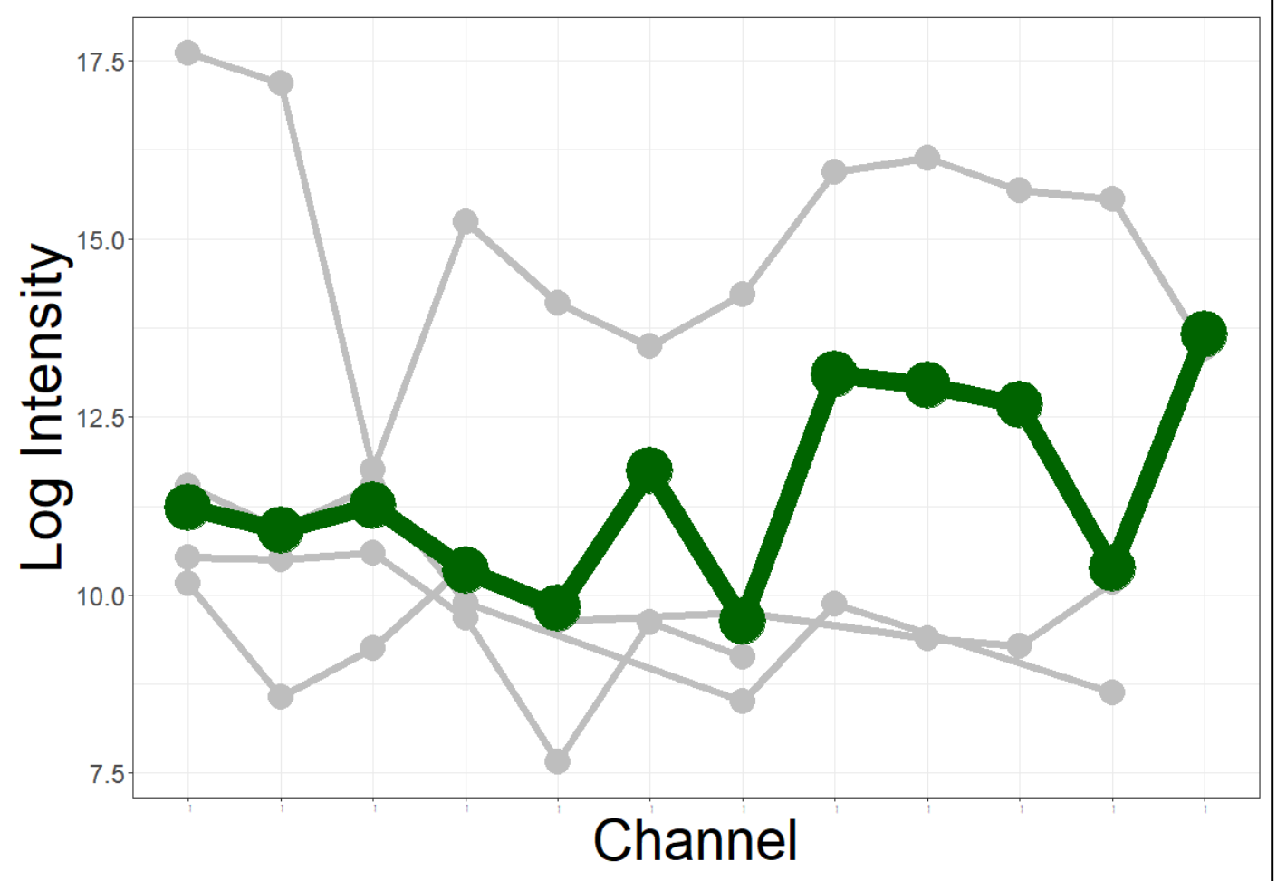
Tukey Median Polish - O00154



LogSum - O00154



Linear - O00154



Feature-level summarization rolls all features into a single value per protein per cell. Summarization has a major impact on the final value. Tukey's Median Polish is a robust method, which is less effected by outliers and noisy features, compared to other methods.

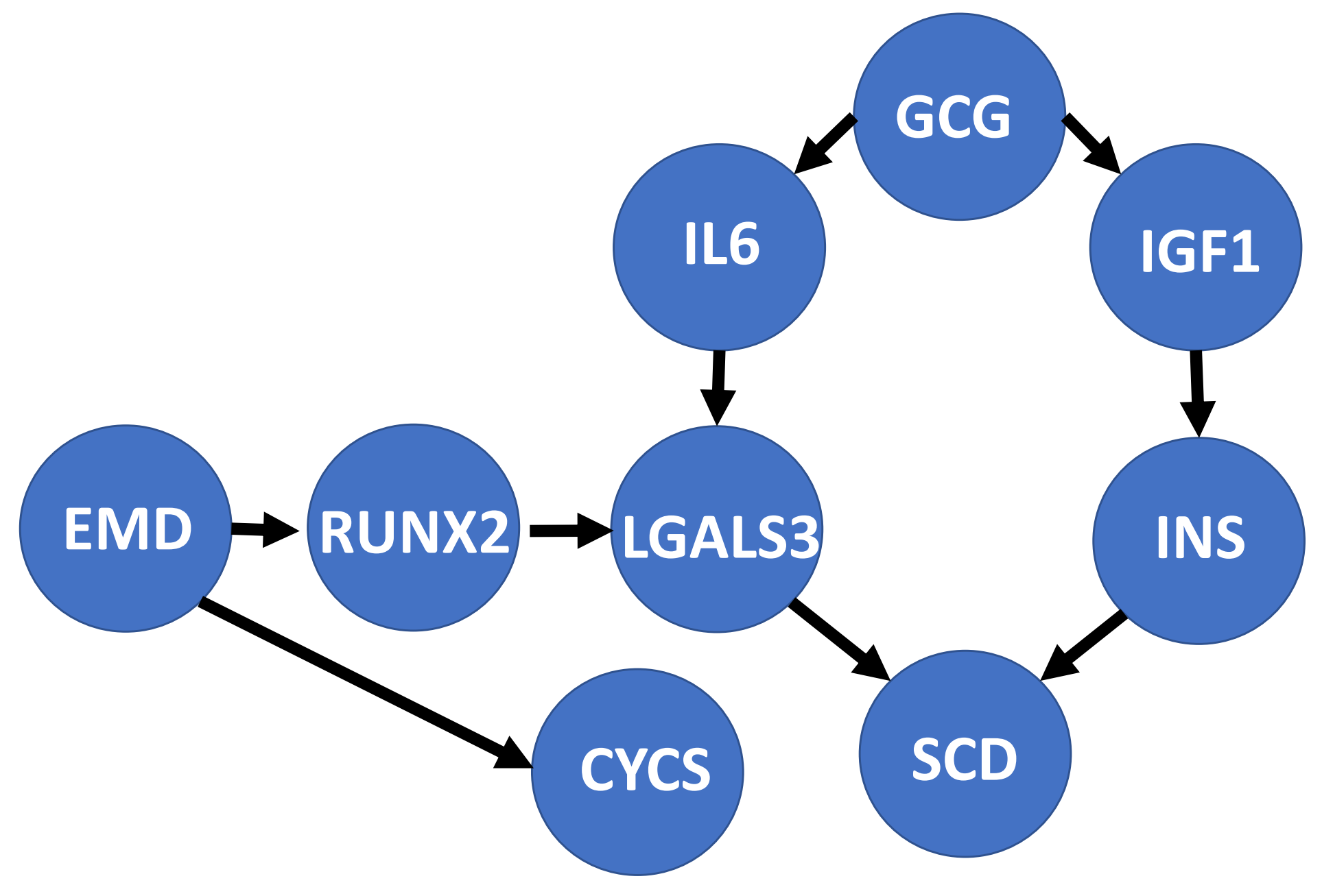
Leverage INDRA to build a prior knowledge network (PKN)[3]

INDRA
Integrated Network and Dynamical Reasoning Assembler

INDRA Relationship Table

Out Gene	In Gene
EMD	RUNX2
EMD	CYCS
RUNX2	LGALS3
IL6	LGALS3
GCG	IL6
GCG	IGF1
IGF1	INS
INS	SCD
LGALS3	SCD

INDRA Graphical Relationships

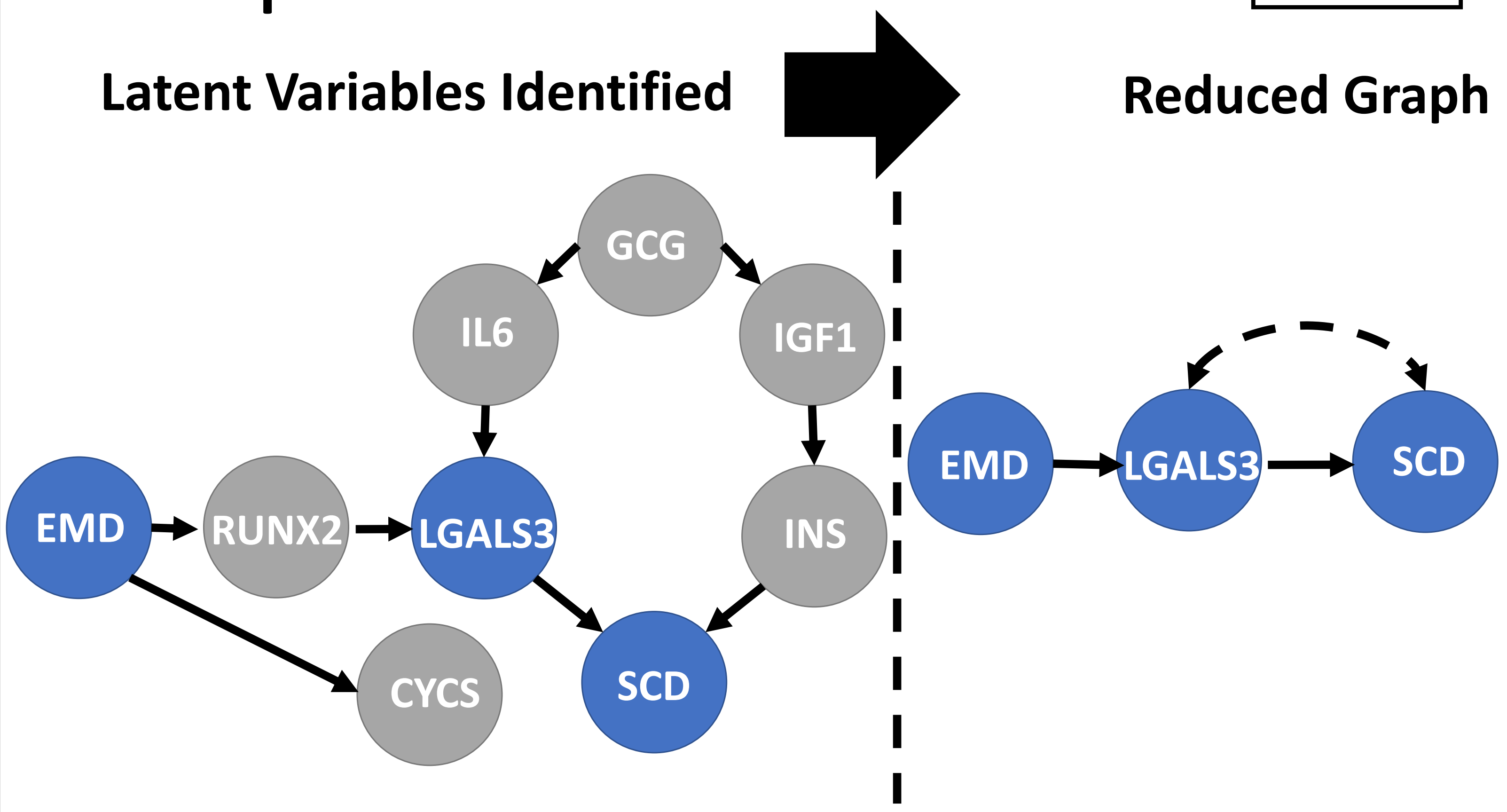


Use INDRA to extract relationships between proteins measured in single cell experiment. Since we are only measuring abundance, we only use upregulation or down regulation events (as opposed to activation or phosphorylation).

Integrate PKN with experimental data[4,5,6]

Y_0

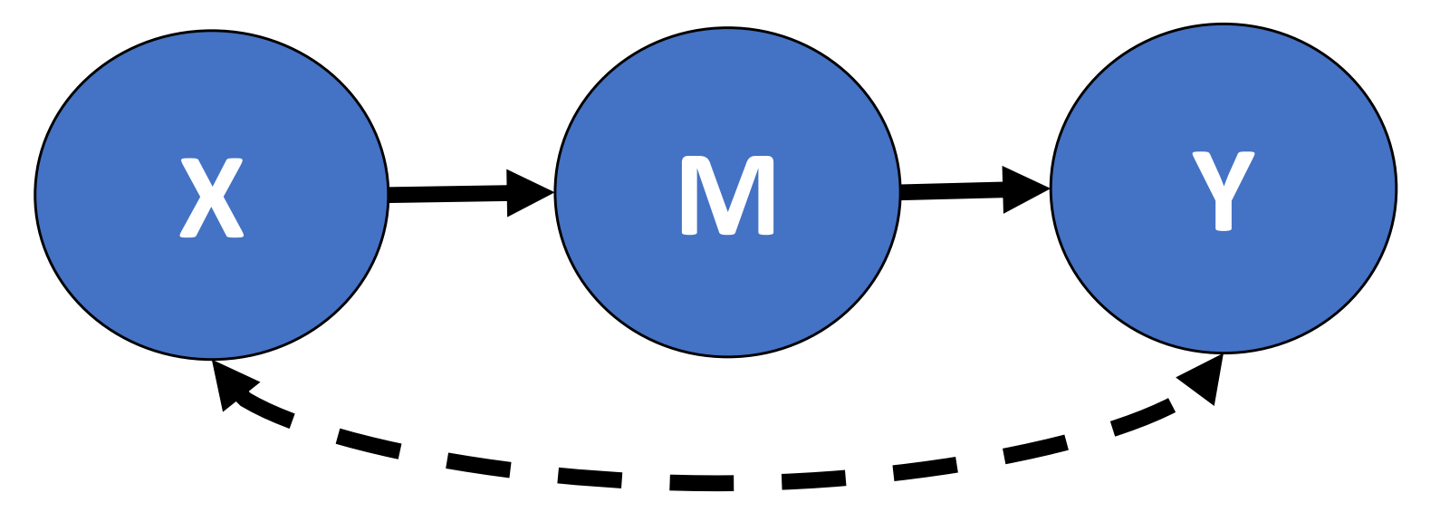
Latent Variables Identified **Reduced Graph**



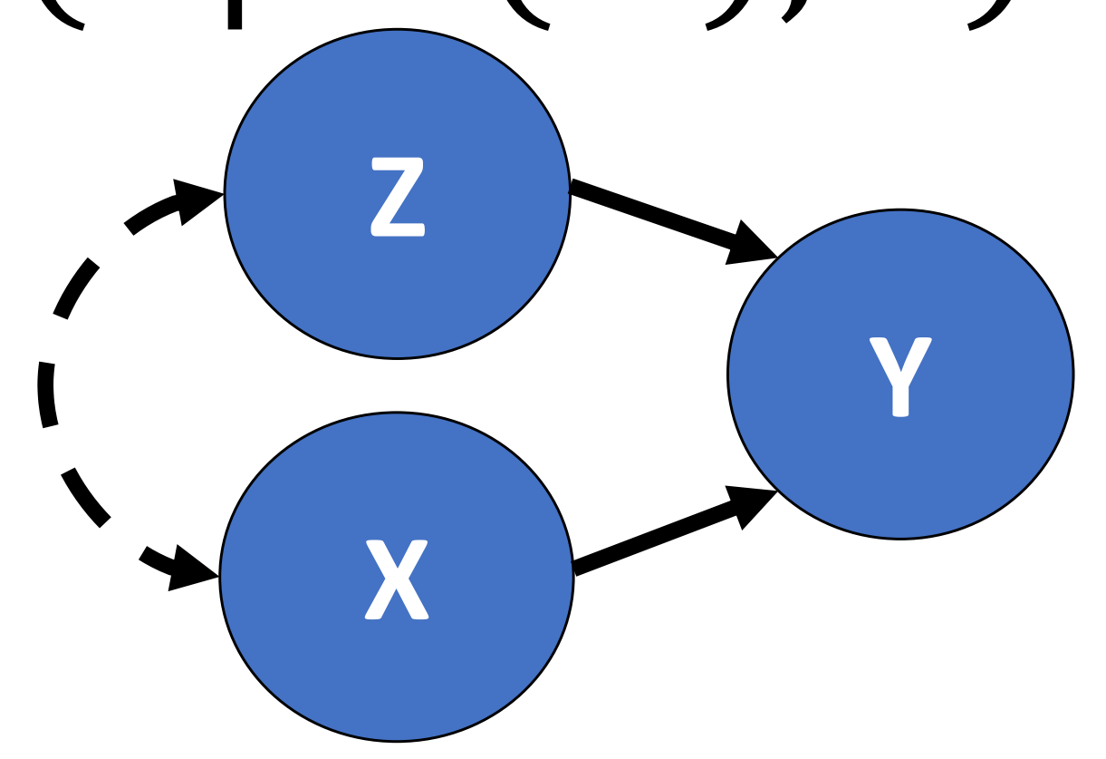
The observed (blue) and latent (grey) proteins are encoded into the graph. With this representation we create a latent variable directed acyclic graph. The Y0 package is used to reduce the graph using graph reduction algorithms. [4] This final reduced graph can then be used to determine identifiability.

Do-calculus allows us to determine what interventions are estimable given the graph topology and observed nodes[7]

$P(Y|do(x'))$ $P(Y|do(x'),Z)$



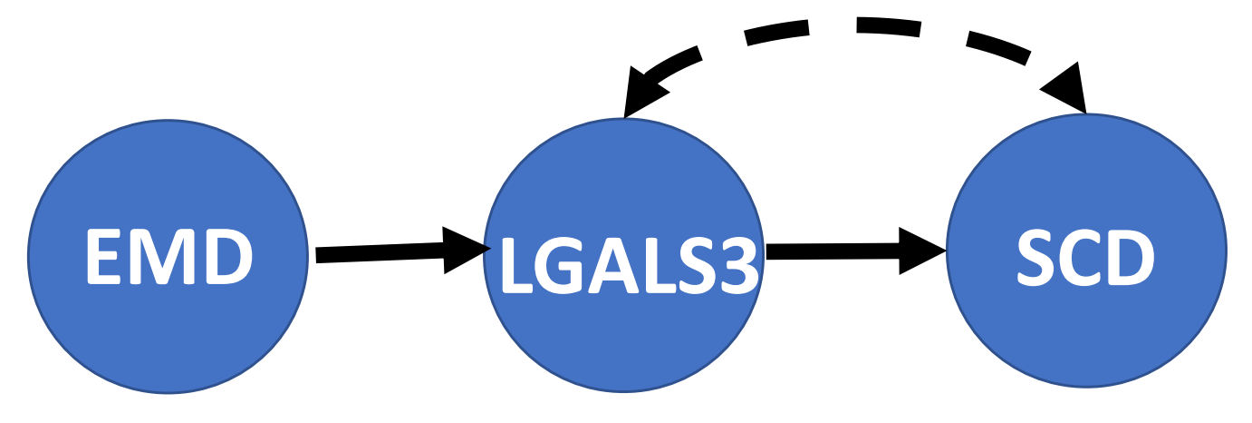
Mediator Adjustment



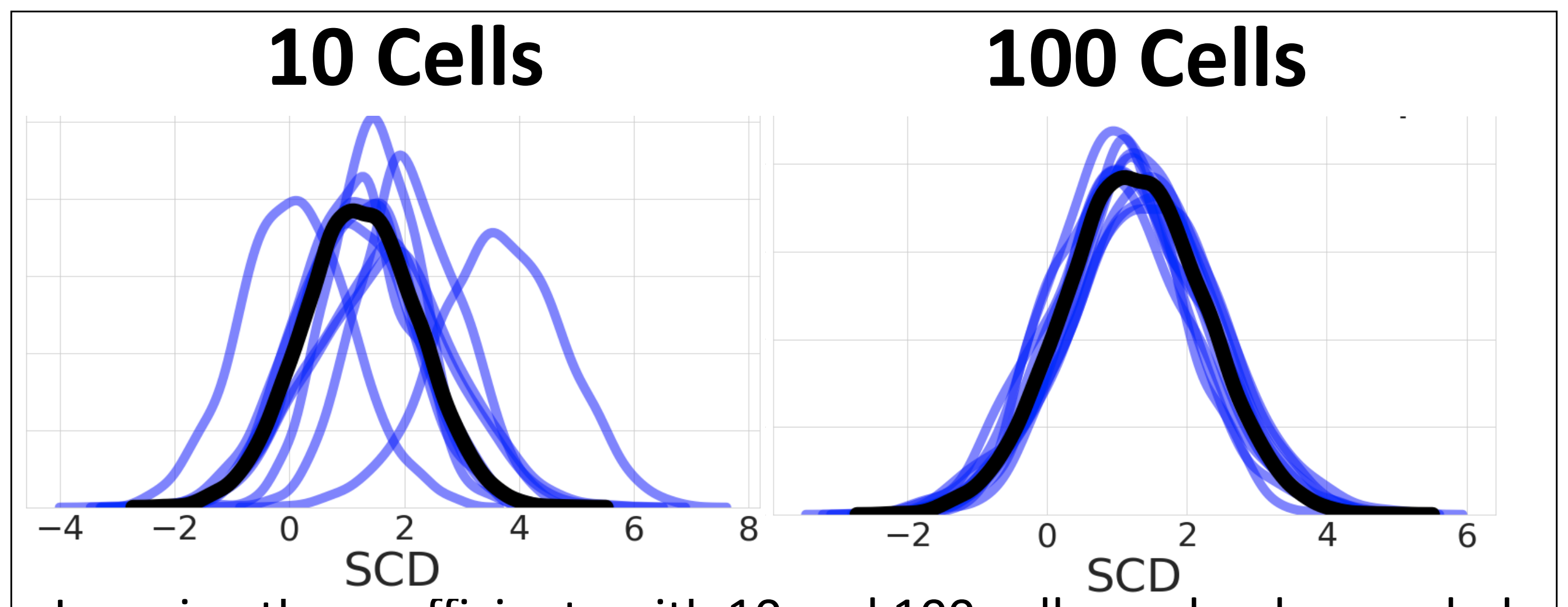
Backdoor Adjustment

Build latent variable model (LVM) over latent DAG[8]

Express graph as linear combinations of Gaussian distributions

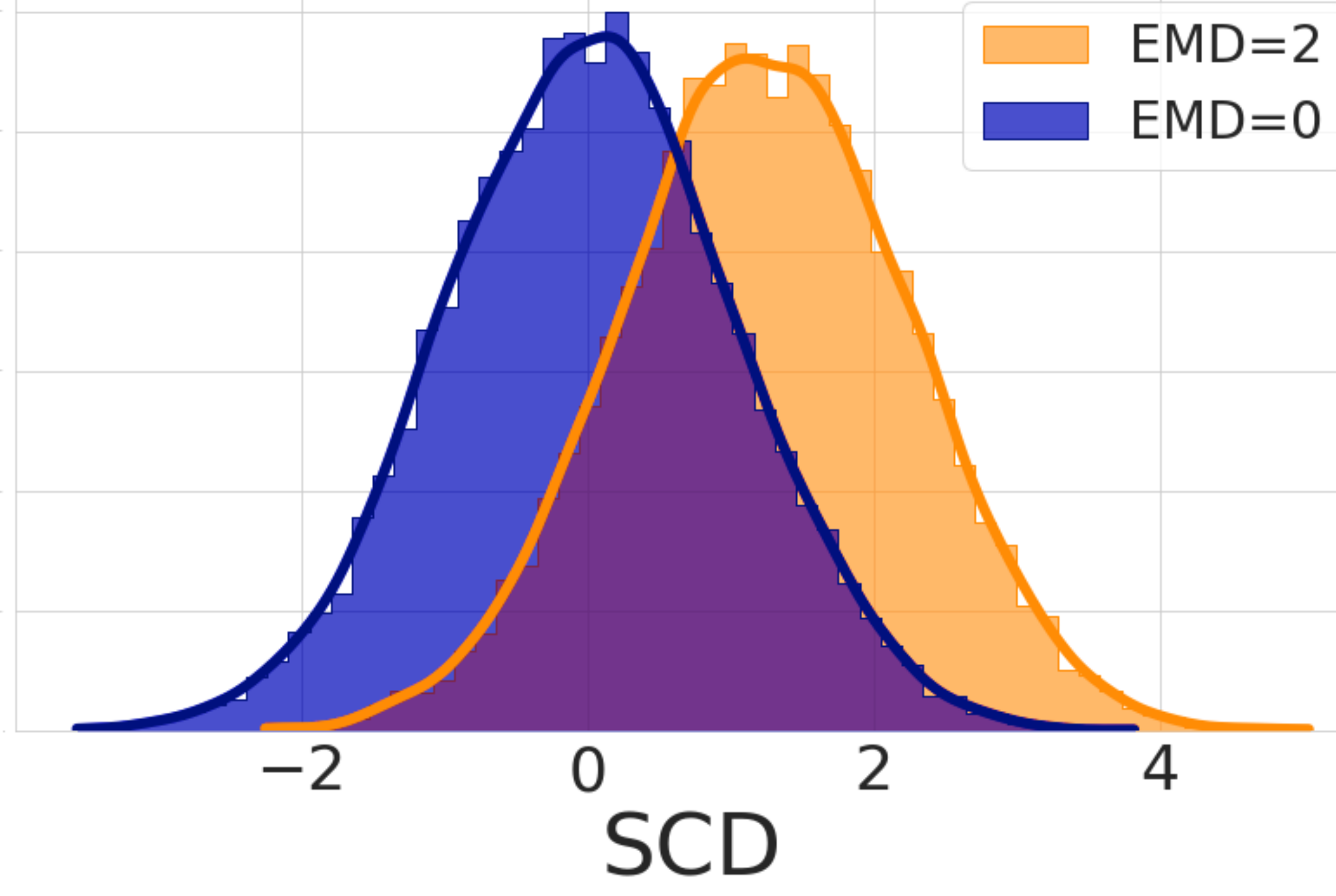


$$Latent = \mathrm{N}(\mu_1, \sigma_1^2)$$
$$EMD = \mathrm{N}(\mu_2, \sigma_2^2)$$
$$LGALS3 = \mathrm{N}(\mu_3 + \beta_1 EMD + \beta_2 Latent\ , \sigma_3^2)$$
$$SCD = \mathrm{N}(\mu_4 + \beta_3 LGALS3 + \beta_4 Latent, \sigma_4^2)$$



Learning the coefficients with 10 and 100 cells randomly sampled from the original dataset. With less samples the interventional distribution is inconsistent.

$$P(SCD|do(EMD' = 2))$$



Average Causal Effect (ACE) = 1.24

Interventional distribution of EMD=2 vs EMD=0. The model was trained and intervened on by forcing EMD to be a specific value. The intervened model was then sampled to generate an interventional distribution. We can compare the average effect of the interventions taking the difference of their expected values.

References & Acknowledgments

1. Leduc, A. et al. Genome Biology, 23, 261 (2022).
2. Choi, M., et al. Bioinformatics, 30(17), 2524-6 (2014).
3. Gyori, B.M. et al. Molecular Systems Biology, 13, 954 (2017).
4. Evans, R.J. Scand. J. Statist., 43, 625–648 (2016).
5. Mohammad-Taheri, S. et al. Bioinformatics, 38, Supplement_1 (2022).
6. Charles Tapley Hoyt, et al. (2021). y0-causal-inference/y0: (v0.1.0). Zenodo.
7. Pearl, J. (2009) Causality. Cambridge University Press, Cambridge, England.
8. Bingham, E. et al. Journal of Machine Learning Research (2018).

We would like to thank Sara Mohammad-Taheri. CTH was funded under the Defense Advanced Research Projects Agency (DARPA) Young Faculty Award [W911NF-20-1-0255]. Jeremy Zucker's work was supported by the Data-Model Convergence Initiative, a component of the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830. KS was funded in part by MDA Award 574137