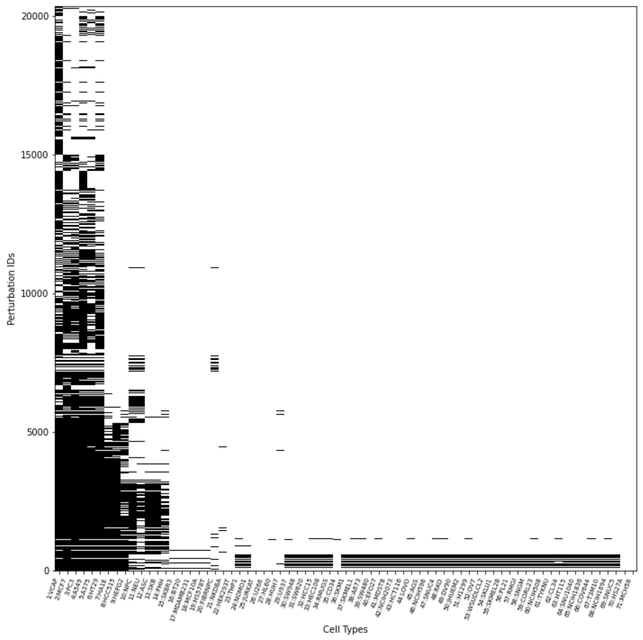
Causal Imputation via Synthetic Interventions

Chandler Squires[123*], Dennis Shen[1*], Anish Agarwal[1], Devavrat Shah[1], Caroline Uhler[12]

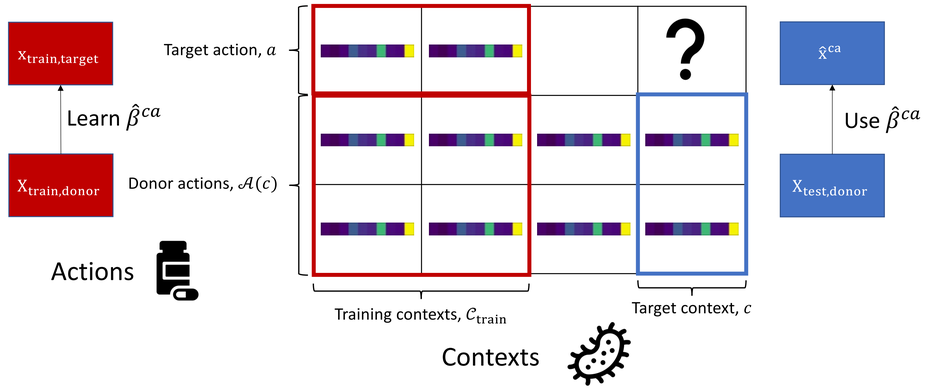
[1]LIDS, MIT [2]IDSS, MIT [3]CSAIL, MIT *Equal contribution

GOAL: TRANSPORT THE EFFECT OF A DRUG FROM ONE CELL TYPE TO OTHER CELL TYPES



Availability of (cell type, perturbation) outcomes in the Connectivity Map (CMap) dataset [1]. Each outcome $\boldsymbol{x^{ca}}$ in cell type c under drug a is a $p = 978$-dimensional vector of gene expression levels.

METHOD: SYNTHETIC INTERVENTIONS ON ACTIONS (SI-A)



Target context: the context (cell type) for which we want a prediction.

Target action: the action (drug) for which we want a prediction.

Donor actions: actions whose outcomes have been measured in the target context.

Training contexts: contexts for which the outcomes of both donor actions and the target action are measured.

ASSUMPTIONS

Linear Factor Model

The outcome $\boldsymbol{x^{ca}} \in \mathbb{R}^p$ can be written as $\boldsymbol{x^{ca}} = U^c \boldsymbol{v^a}$ for $U^c \in \mathbb{R}^{p \times r}$ and $\boldsymbol{v^a} \in \mathbb{R}^r$.

Sufficient Donor Actions

There exists $\boldsymbol{\beta_{a,c}} \in \mathbb{R}^{|\mathcal{A}(c)|}$ such that $\boldsymbol{v^a} = \boldsymbol{\beta_{a,c}^\top}\, \boldsymbol{v^{\mathcal{A}(c)}}$.

Sufficient Training Contexts

$rowspan(X_{test,donor})$ is a subset of $rowspan(X_{train,donor})$.

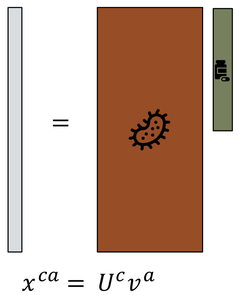
THEORETICAL RESULTS

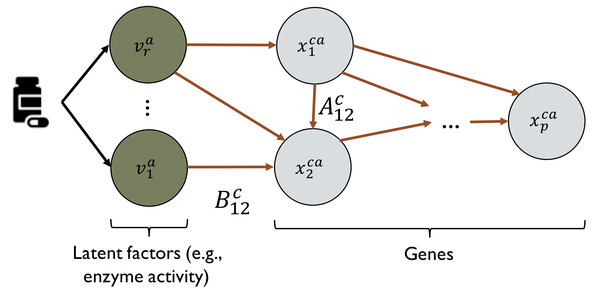
IDENTIFIABILITY (THEOREM 1):

Under the above assumptions, the SI-A method identifies the outcome $\boldsymbol{x^{ca}}$.

STRUCTURAL EQUATION MODELS INDUCE LINEAR FACTOR MODELS (PROPOSITION 1):

If $\boldsymbol{x^{ca}} = A^c \boldsymbol{x^{ca}} + B^c \boldsymbol{v^a}$ with A^c acyclic, then it satisfies a linear factor model.





DESIDERATA

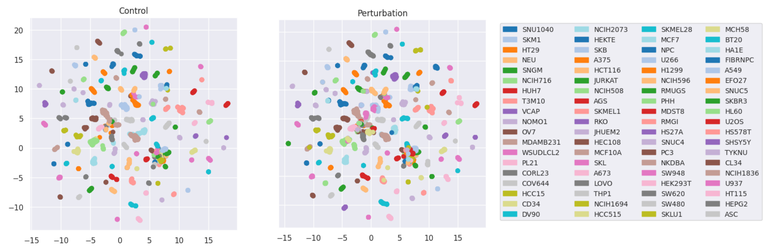
- **Fast**
- Flexible/effective model of **heterogenous effects**
- Theoretical consistency **guarantees**
- Detailed causal model (i.e., **selection diagram**) not necessary

	Fast	Heterogeneous Effects	Guarantees	Needs selection diagram
Fixed Effects (FE)	✓	✗	✓	✗
Autoencoding + FE	✓	✓	✗	✗
Causal Transportability	✓	✓	✓	✓
MICE/MissForest	✗	✓	✗	✗
Synthetic Interventions	✓	✓	✓	✗

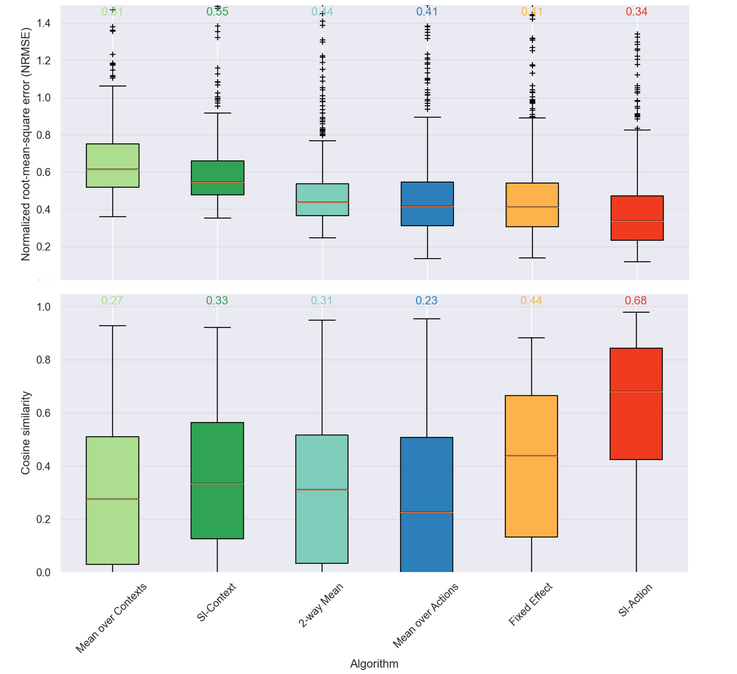
Existing methods do not satisfy our desiderata:

- Fixed effects [2] isn't expressive enough.
- Fixed effects + autoencoding [3] and traditional imputation methods [4,5] don't have guarantees.
- Transportability methods [6] need selection diagrams.

EMPIRICAL RESULTS



Most of the variation in the CMAP dataset is attributed to cell type, not drug. The plots show a UMAP embedding of gene expression levels, colored by cell type.



SI-A outperforms baselines on a randomly selected subset of the CMAP dataset.

REFERENCES:

[1] Subramanian, Aravind, et al. "A next generation connectivity map: L1000 platform and the first 1,000,000 profiles." *Cell*171.6 (2017): 1437-1452.

[2] Dixit, Atray, et al. "Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens." *cell* 167.7 (2016): 1853-1866.

[3] Lotfollahi, Mohammad, F. Alexander Wolf, and Fabian J. Theis. "scGen predicts single-cell perturbation responses." *Nature methods* 16.8 (2019): 715-721.

[4] Stekhoven, Daniel J., and Peter Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data." *Bioinformatics* 28.1 (2012): 112-118.

[5] Van Buuren, Stef, and Karin Oudshoorn. *Flexible multivariate imputation by MICE*. Leiden: TNO, 1999.

[6] Lee, Sanghack, Juan Correa, and Elias Bareinboim. "General transportability–synthesizing observations and experiments from heterogeneous domains." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 06. 2020.