Learning to see the hidden world: A perspective on causal representations

Francesco Locatello

Institute of Science and Technology Austria

Causal Learning and AI

What triggers grooming behaviors in ants as a collective hygiene policy?

[Image Credit: Cremer Lab @ ISTA]

Why causal models?

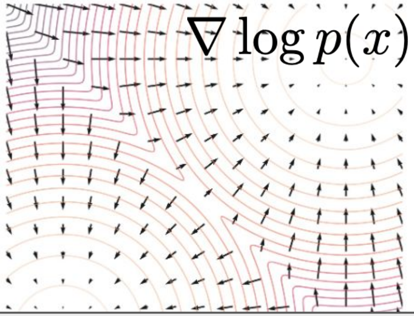
Observational data + causal model predict the result of randomized studies [1]

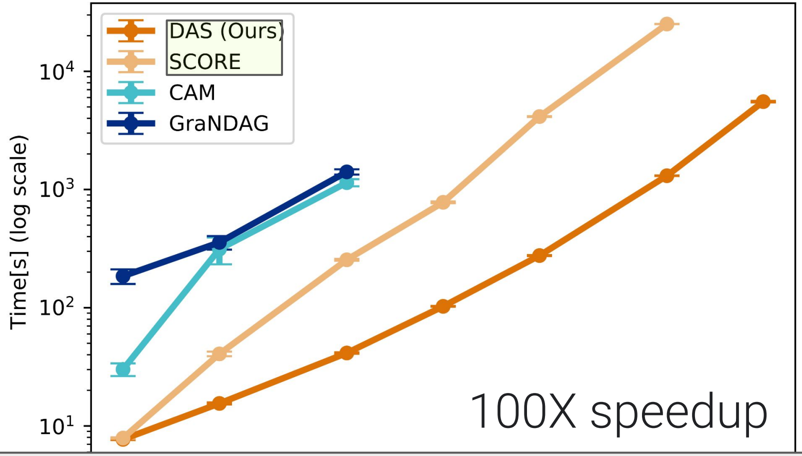
$$P(X_1, \ldots, X_n) = \prod_i P(X_i | \mathbf{PA}_i)$$

$$P(X_1, \ldots, X_n | do(X_j = x)) = \prod_{i \neq j} P(X_i | \mathbf{PA}_i) \delta(X_j = x)$$

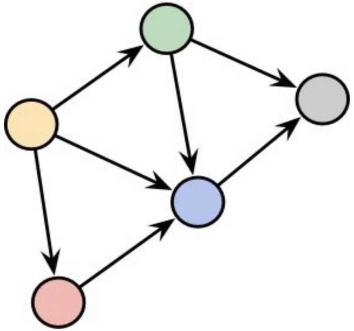
[1] "*Causal Discovery with Additive Noise Models*", Peters et al., JMLR 2014;

Learning the graph





The discovery of the causal order converges linearly as $\exp\left(-\Theta\left(\frac{nC_m^2}{log(m)}\right)\right)$



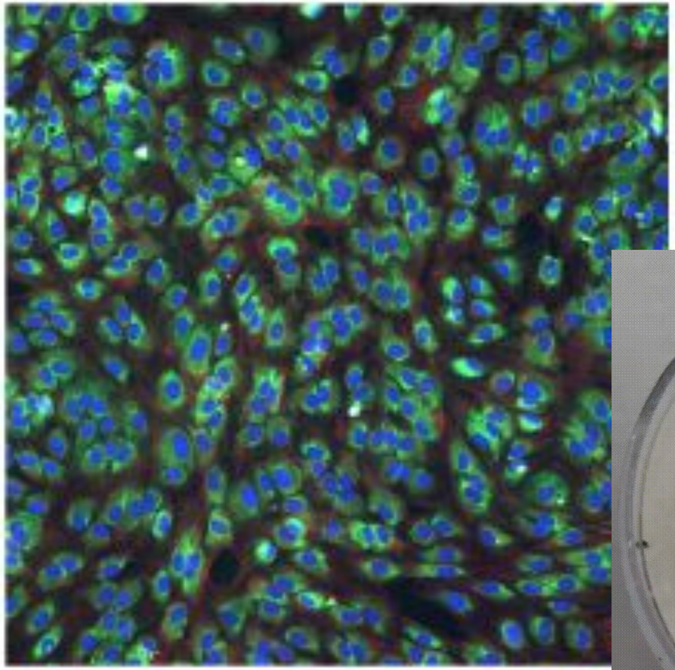
Causal Discovery

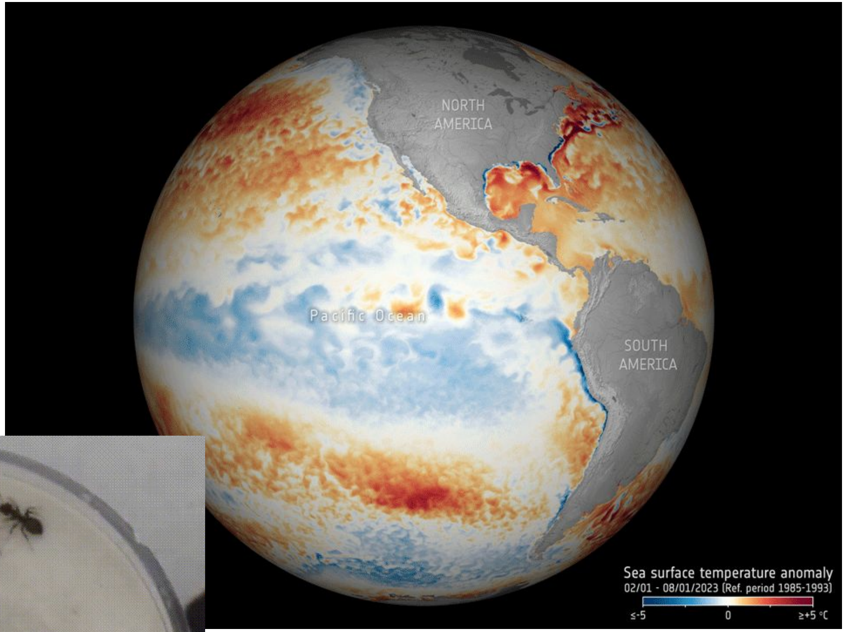
infer

x_1	x_2	x_3	x_4	x_5
0.2	0.1	2.2	1.6	4.3
0.3	0.1	2.1	1.9	5.3
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

“*Score matching enables causal discovery of non-linear additive noise models*”, Rolland, Cevher, Kleindessner, Russel, Schölkopf, Janzing, L; ICML 2022

Data





[Image Credit: rxrx1 Dataset, www.esa.int/, ISTAnt dataset]

- Prediction-powered causal inference
- Exploratory causal inference
- Beyond “standard” causal models

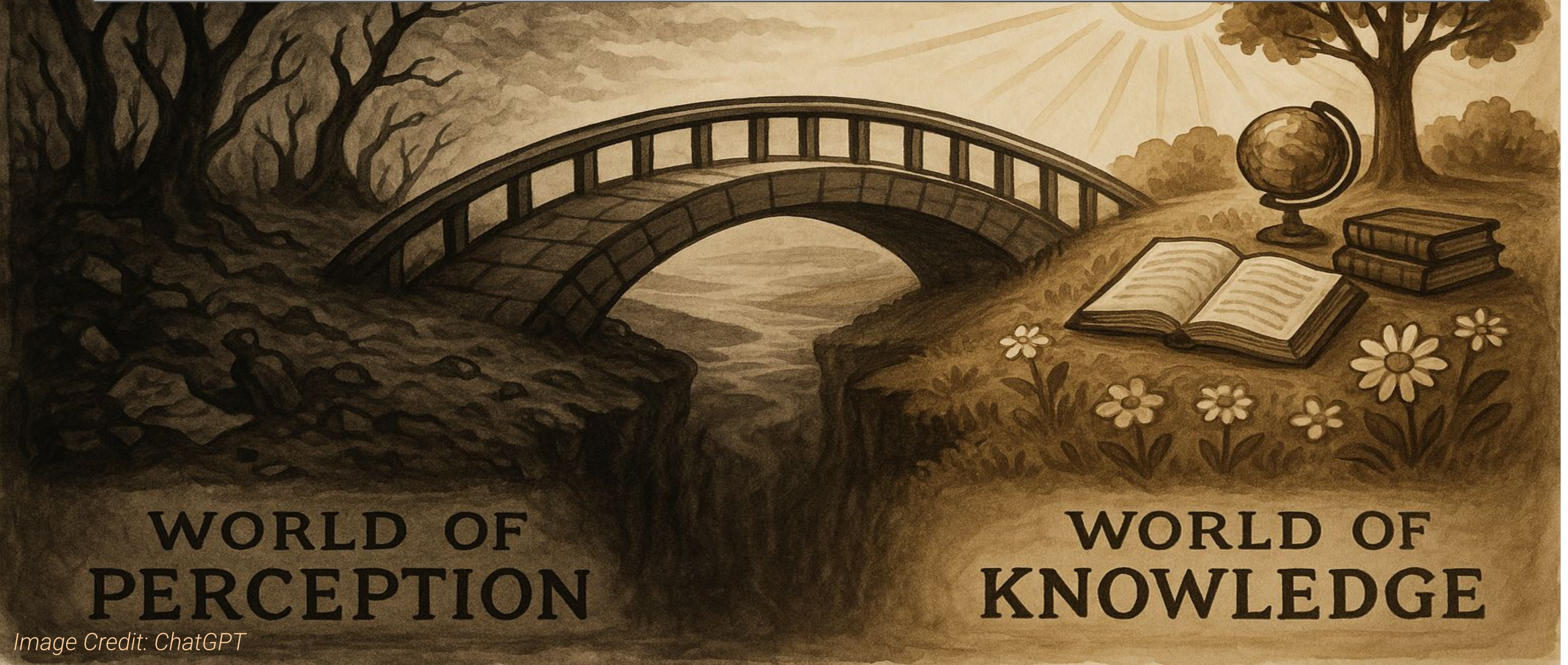


Image Credit: ChatGPT

What do we want?

$$P(X_1, \ldots, X_n) = \prod_i P(X_i | \mathbf{PA}_i)$$

$$\downarrow$$

$$P(X_1, \ldots, X_n | do(X_j = x)) = \prod_{i \neq j} P(X_i | \mathbf{PA}_i) \delta(X_j = x)$$

$$\downarrow$$

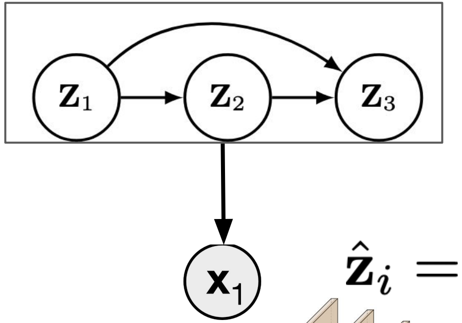
$$= P(X_{V \setminus \{j, \mathbf{PA}_j\}} | X_j = x, X_{\mathbf{PA}_j}) P(X_{\mathbf{PA}_j})$$

Causal estimand that is statistically identified on causal variables is also identifiable from the representation

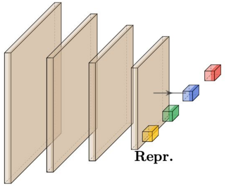
Idea: Representation should make it easier/possible to extract *causal* information with some downstream estimator

"The Third Pillar of Causal Analysis? A Measurement Perspective on Causal Representations", Yao* and Huang*, Cadei, Zhang, L; NeurIPS 2025

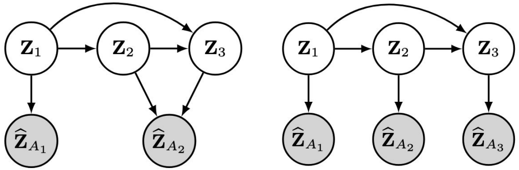
Measurement perspective of CRL and identifiability



$$\hat{\mathbf{z}}_i = h(\mathbf{z}_j)$$



When is a learned representation a valid proxy for a causal variable? "*measurement models*" [1]

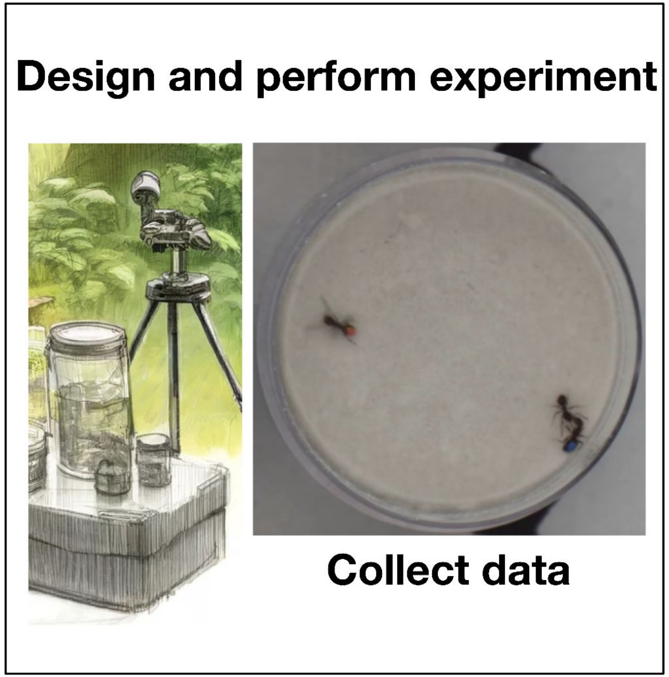


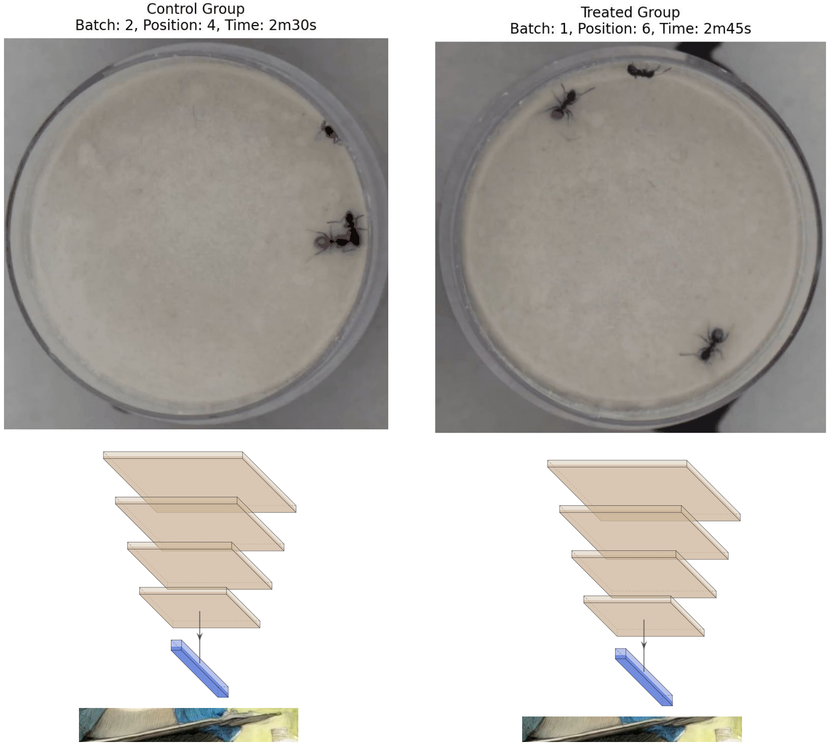
Conditions for causal validity of downstream estimate [2]:

- Know $\widehat{\mathbf{Z}}_{A_j} \perp\!\!\!\perp \mathbf{Z}_i \mid \mathbf{Z}_{[N]\setminus\{i\}}$
- Estimate is invariant to h

[1] "*Learning the structure of linear latent variable models*" R. Silva, R. Scheines, C. Glymour, P. Spirtes, and D. M. Chickering. JMLR, 2006
[2] "*The Third Pillar of Causal Analysis? A Measurement Perspective on Causal Representations*", Yao* and Huang*, Cadei, Zhang, L; NeurIPS 2025
[3] "*Self-supervised Representation Learning Provably Isolates Content From Style*", von Kügelgen*, Sharma*, Gresele*, Brendel, Schölkopf+, Besserve+, L+ NeurIPS 2021

Exemplary pipeline in experimental ecology





ISTAnt dataset

No grooming

Blue to Focal grooming

What? First benchmark to estimate causal effects from real world ecological videos collected in a randomized controlled trial.

"Smoke and Mirrors in Causal Downstream Tasks", R. Cadei, L. Lindorfer, S. Cremer, C. Schmid, L; NeurIPS 2024

Why is it unique?

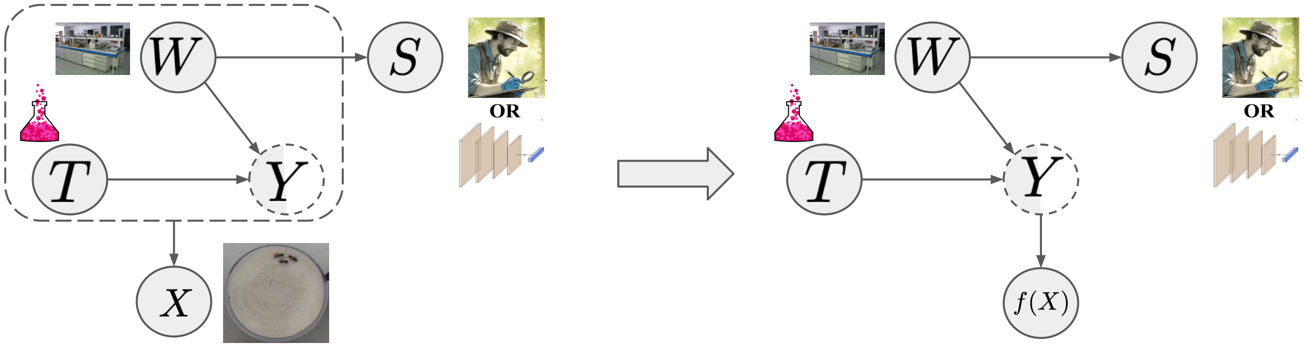
Other benchmarks stop at validating statistical accuracy. We estimate causal effects from real world ecological videos collected in a randomized controlled trial

$$ATE := \mathbb{E}[Y|do(T=1)] - \mathbb{E}[Y|do(T=0)].$$

$$\Big\downarrow \ \textbf{RCT}$$

$$AD := \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0].$$

A measurement perspective on the problem

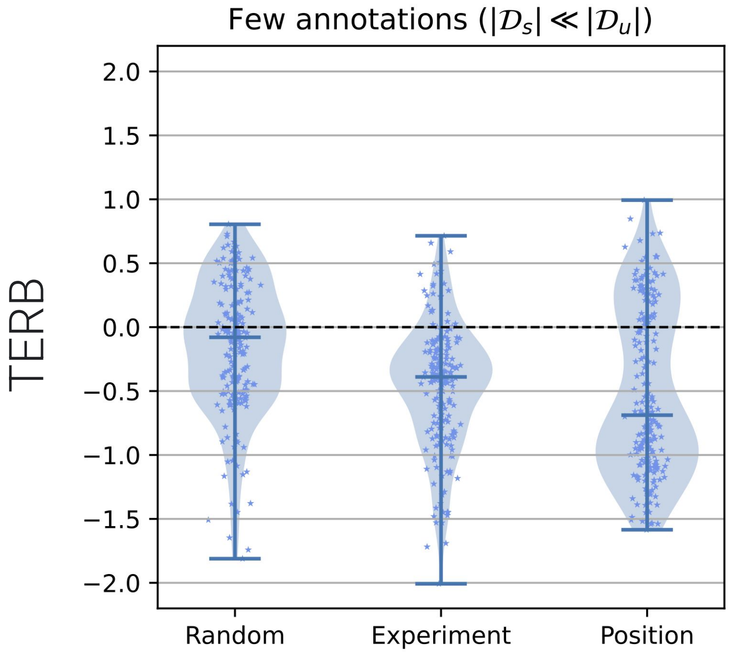


$$TEB := \left(\underbrace{\mathbb{E}_{\boldsymbol{X}|do(T=1)}[f(\boldsymbol{X})] - \mathbb{E}_{Y|do(T=1)}[Y]}_{\textit{Interventional Bias under Treatment}} \right) - \left(\underbrace{\mathbb{E}_{\boldsymbol{X}|do(T=0)}[f(\boldsymbol{X})] - \mathbb{E}_{Y|do(T=0)}[Y]}_{\textit{Interventional Bias under Control}} \right)$$

Errors come from:

- Selection bias: which samples are labelled?
- Pre-training data
- Discretization bias

Problem 1: What data to label?



Implication: Sampling choice matters, but random sampling is not always possible

Optimize for invariance in unified CRL

$$\boxed{\iota(\mathbf{z}_A) = \iota(\tilde{\mathbf{z}}_A)} \Leftrightarrow \boxed{\mathbf{z}_A \sim_\iota \tilde{\mathbf{z}}_A.}$$

Two vectors have the same projection onto the quotient induced by the equivalence relationship

They belong to the same equivalence class

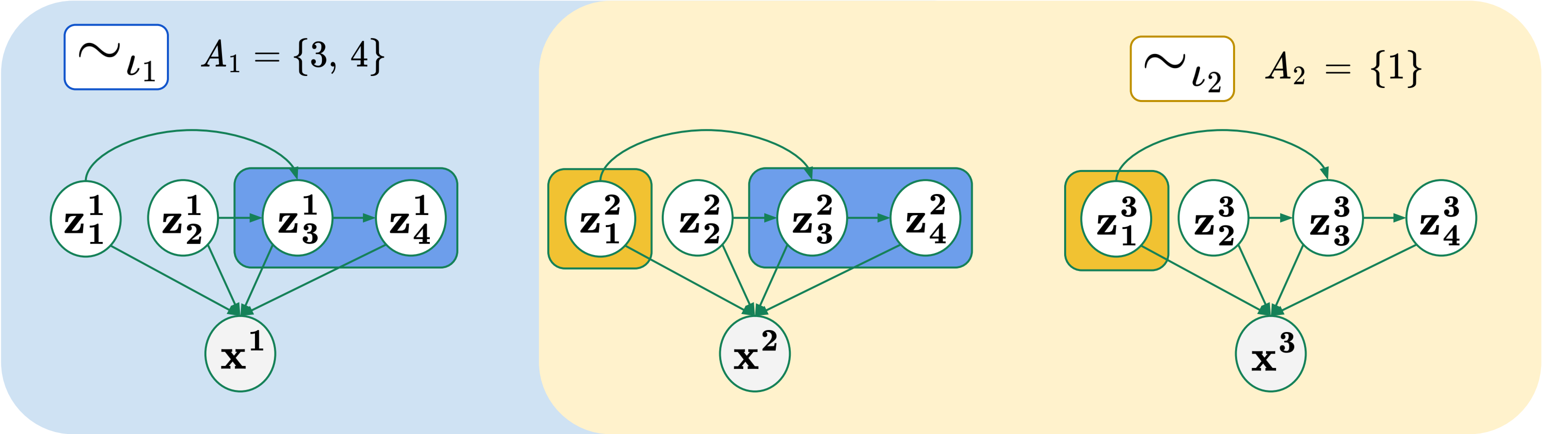
How? Assume access to multiple non-i.i.d. groups of sample. All samples in the same group are *equivalent* in some sense

Why? Because you believe isolating these invariances is relevant to a task

"Unifying Causal Representation Learning with the Invariance Principle" D. Yao, D. Rancati, R. Cadei, M. Fumero, L; ICLR 2025

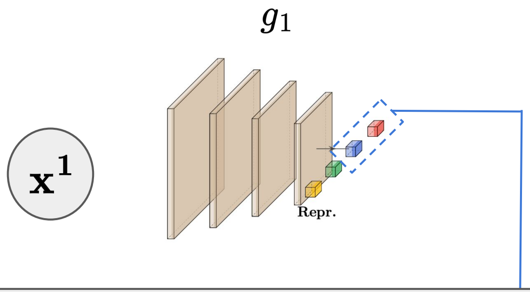
Key idea for unified CRL: invariance principle

$$\iota(\mathbf{z}_A) = \iota(\tilde{\mathbf{z}}_A) \Leftrightarrow \mathbf{z}_A \sim_\iota \tilde{\mathbf{z}}_A.$$



"Unifying Causal Representation Learning with the Invariance Principle" D. Yao, D. Rancati, R. Cadei, M. Fumero, L; ICLR 2025

Key idea for unified CRL: learning

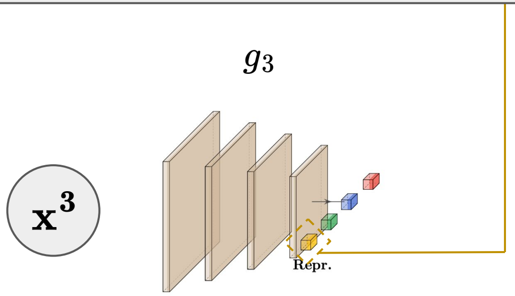


- Invariance Constraint:

$$g_1\left(\mathbf{x}^1\right) \sim_{\iota_1} g_2\left(\mathbf{x}^2\right)$$

$$g_2\left(\mathbf{x}^2\right) \sim_{\iota_2} g_3\left(\mathbf{x}^3\right)$$

*Unify using a single "language" __**31**__ different identification results from 28 paper!*

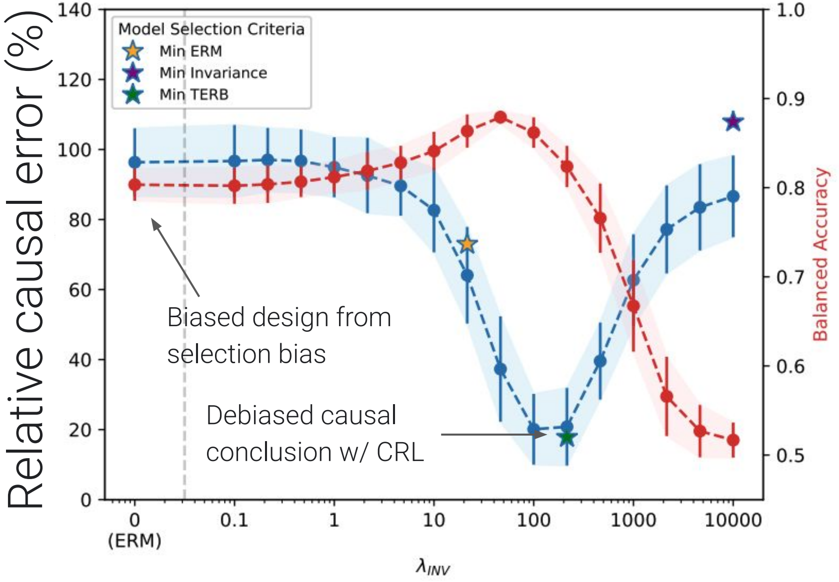


$$I\left(\mathbf{z}^{\mathbf{k}}_{A_2}, g_k\left(\mathbf{x}^{\mathbf{k}}\right)\right) = H\left(\mathbf{z}^{\mathbf{k}}_{\mathbf{A}_2}\right) \quad k = 2,\ 3$$

- **Result**: Smooth encoders satisfying the two constraints block-identify invariant components

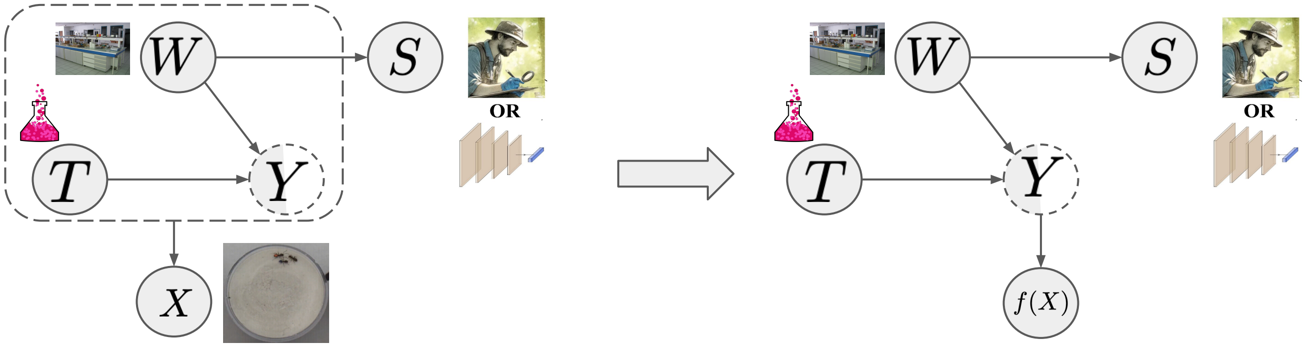
$$\hat{\mathbf{z}}_i = h(\mathbf{z}_j) \quad \boxed{\mathbf{z}^{1,2}_{\mathbf{A}_1}} \quad \boxed{\mathbf{z}^{2,3}_{\mathbf{A}_2}}$$

Debiasing with CRL and the invariance principle



Idea: Assume invariant representation across experiment settings

A measurement perspective on the problem

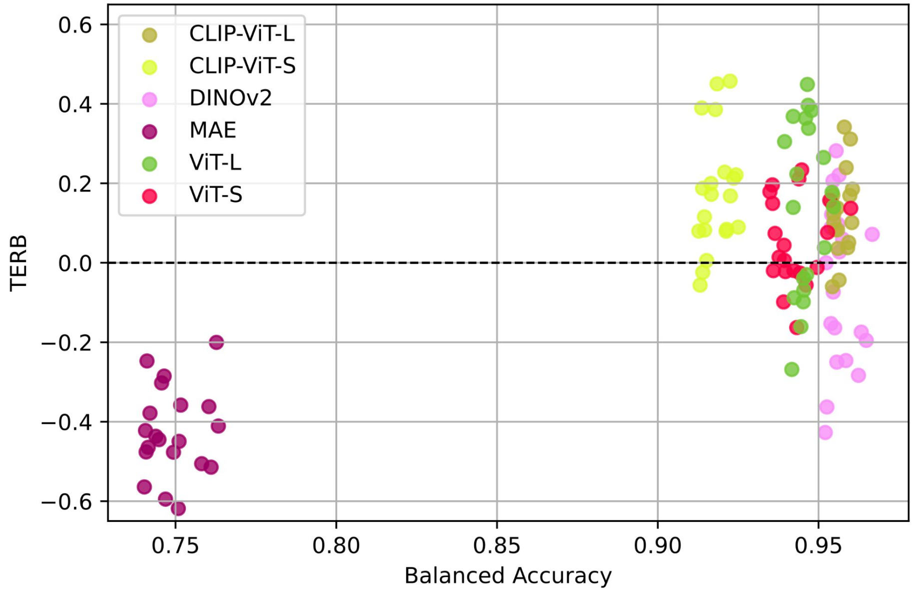


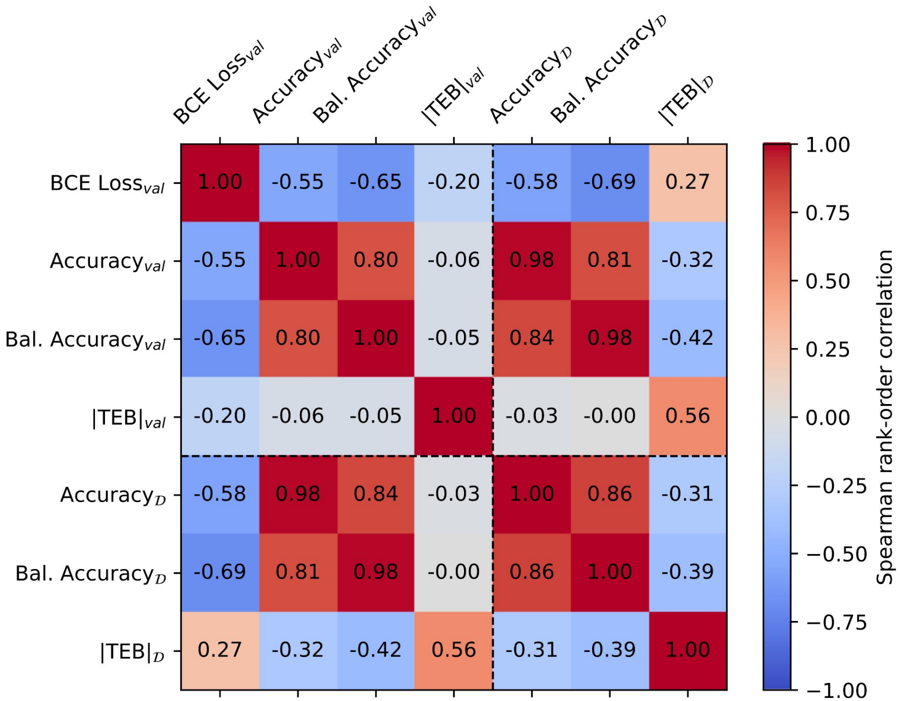
$$TEB := \Bigg(\underbrace{\mathbb{E}_{\boldsymbol{X}|do(T=1)}[f(\boldsymbol{X})] - \mathbb{E}_{Y|do(T=1)}[Y]}_{\textit{Interventional Bias under Treatment}} \Bigg) - \Bigg(\underbrace{\mathbb{E}_{\boldsymbol{X}|do(T=0)}[f(\boldsymbol{X})] - \mathbb{E}_{Y|do(T=0)}[Y]}_{\textit{Interventional Bias under Control}} \Bigg)$$

Errors come from:

- Selection bias: which samples are labelled?
- Pre-training data
- Discretization bias

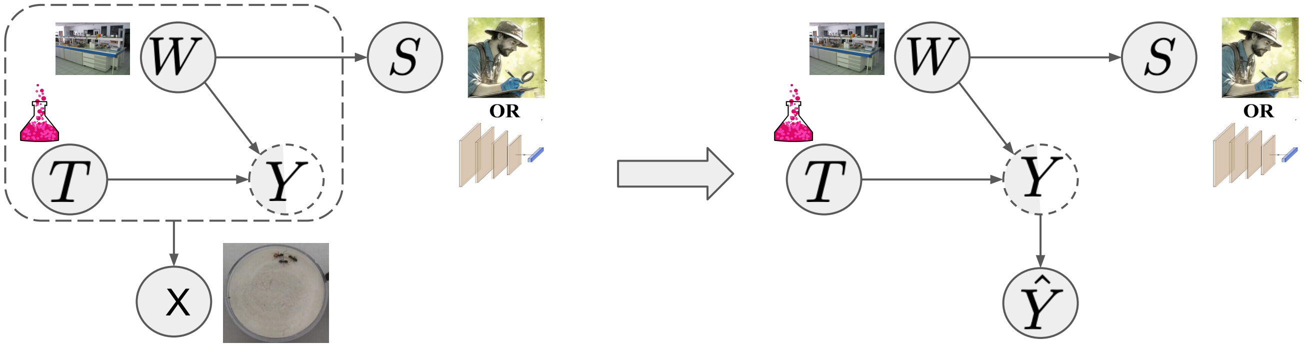
Problem 2: model choice

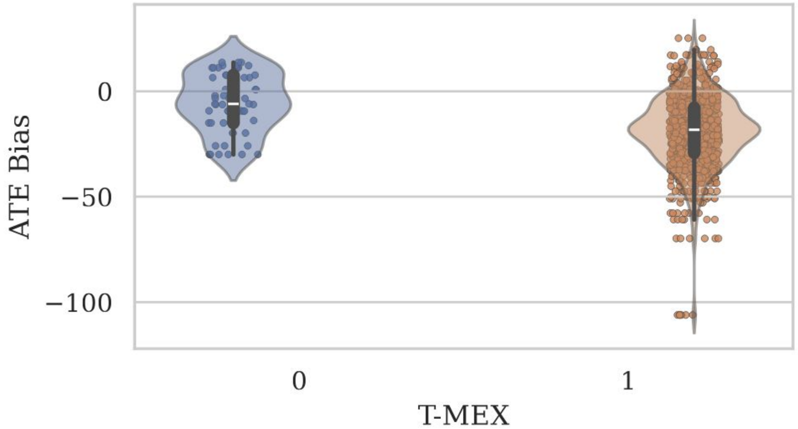




Implication: Models have different TERB and accuracy is not a good indicator of downstream causal performance. TEB on validation data works best

Selecting valid measurements with T-Mex





Conditions for causal validity of downstream estimator:

- $\widehat{\mathbf{Z}}_{A_j} \perp\!\!\!\perp \mathbf{Z}_i \,\big|\, \mathbf{Z}_{[N]\setminus\{i\}}$
- Estimator is invariant to h

Note: T-Mex doesn't need data to follow trial distribution (as long as the conditional indep. still holds)

"The Third Pillar of Causal Analysis? A Measurement Perspective on Causal Representations", Yao* and Huang*, Cadei, Zhang, L; NeurIPS 2025

Problem 3: Postprocessing

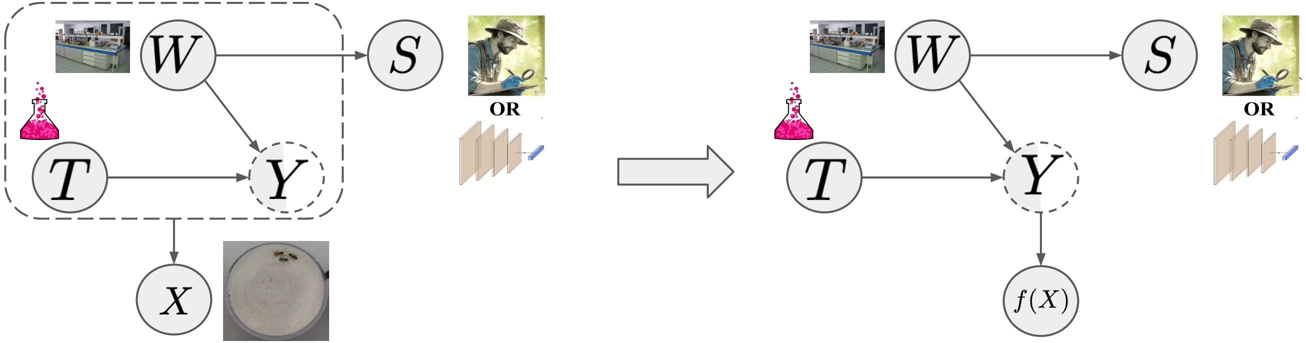
A common choice in ML is to threshold the predictions to make them binary.

Theorem [informal]: This choice can introduce bias.

$$\mathcal{H}_0 : \mathbb{E}[|\mathrm{TEB}(f)|] = \mathbb{E}[|\mathrm{TEB}(\mathbb{1}_{[0.5,1]}(f))|] \quad vs \quad \mathcal{H}_1 : \mathbb{E}[|\mathrm{TEB}(f)|] < \mathbb{E}[|\mathrm{TEB}(\mathbb{1}_{[0.5,1]}(f))|]$$

T-test rejects null hypothesis with p~10^-25

How do we optimize for validity

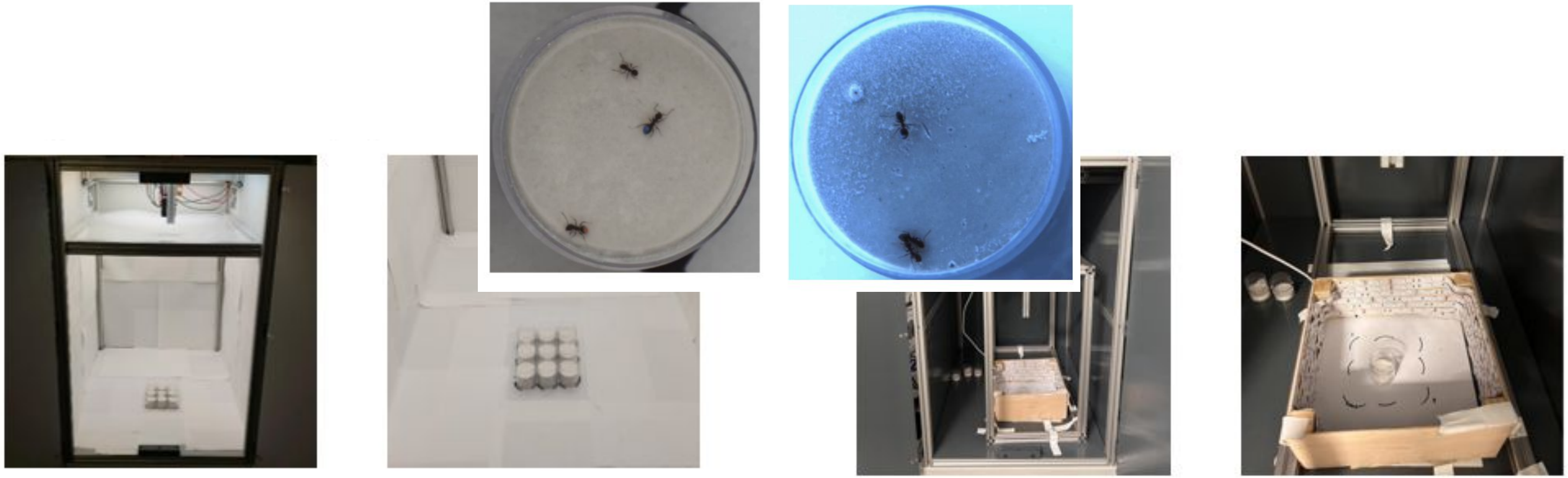


$$TEB := \left(\underbrace{\mathbb{E}_{\boldsymbol{X}|do(T=1)}[f(\boldsymbol{X})] - \mathbb{E}_{Y|do(T=1)}[Y]}_{\textit{Interventional Bias under Treatment}} \right) - \left(\underbrace{\mathbb{E}_{\boldsymbol{X}|do(T=0)}[f(\boldsymbol{X})] - \mathbb{E}_{Y|do(T=0)}[Y]}_{\textit{Interventional Bias under Control}} \right)$$

Thm: Conditional calibration $\mathbb{E}[Y - f(X)|W,T] = 0$ implies valid estimates with *correct confidence intervals* using AIPW. Conditional independence of the measurement model implies conditional calibration.

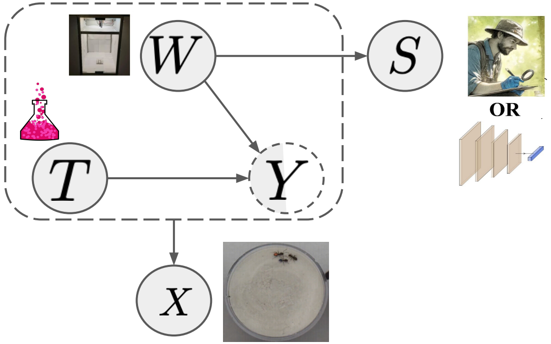
“*Prediction-Powered Causal Inference*”, Cadei, Demirel, De Bartolomeis, Lindorfer, Cremer, Schmid, L; NeurIPS 2025

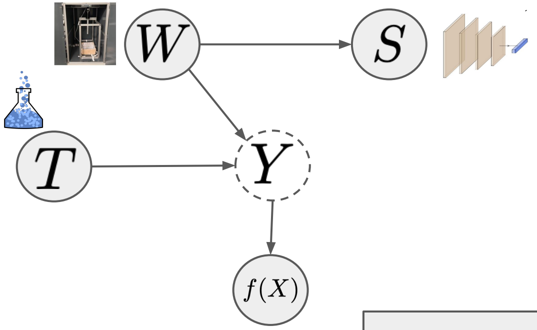
Zero-shot transfer across experiments



“*Prediction-Powered Causal Inference*”, Cadei, Demirel, De Bartolomeis, Lindorfer, Cremer, Schmid, L; NeurIPS 2025

Validity across experiments





$$\min_{h,\phi} \quad \mathbb{E}_{\mathbb{P}^{e_1}}\left[\mathcal{L}\left(Y, \overbrace{h \circ \phi(\boldsymbol{X})}^{f(X)}\right)\right]$$

$$\text{s.t.} \quad \phi(\boldsymbol{X}) \perp\!\!\!\perp \boldsymbol{Z} | Y = y \quad \forall y \in \mathcal{Y}$$

Conditions for causal validity of downstream estimate:

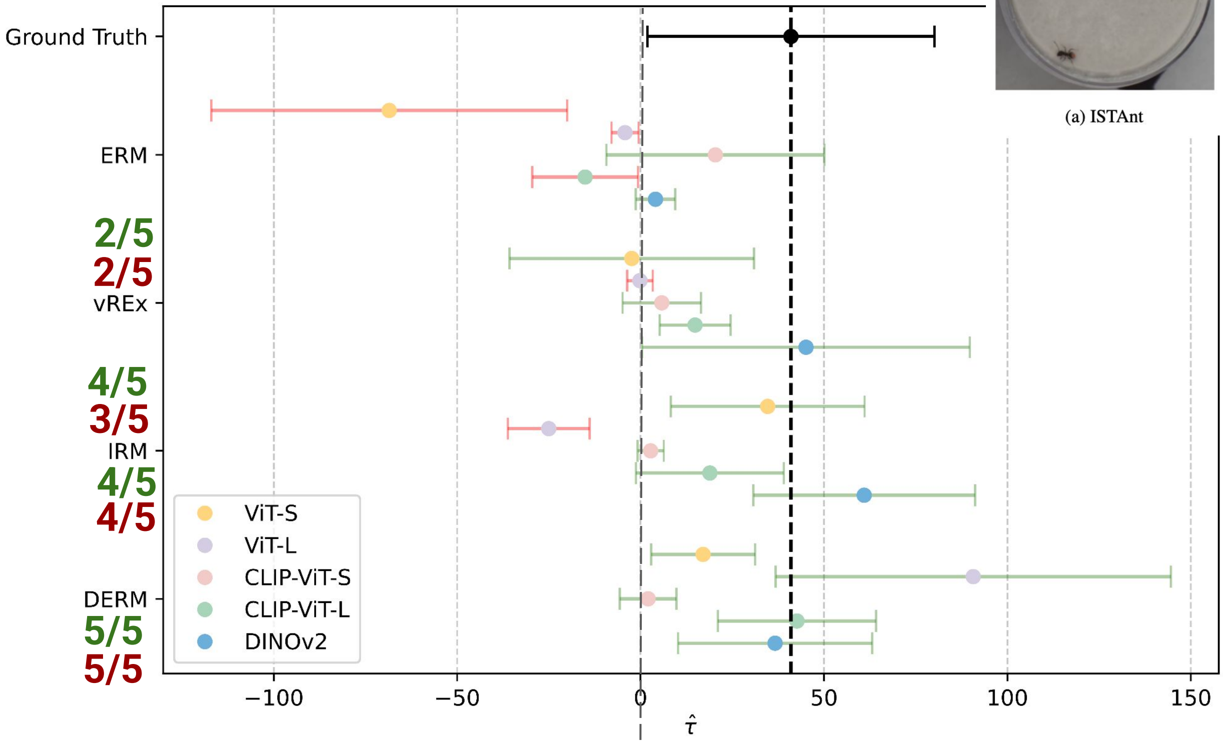
- Know $\widehat{\mathbf{Z}}_{A_j} \perp\!\!\!\perp \mathbf{Z}_i \,|\, \mathbf{Z}_{[N]\setminus\{i\}}$
- Estimate is invariant to h

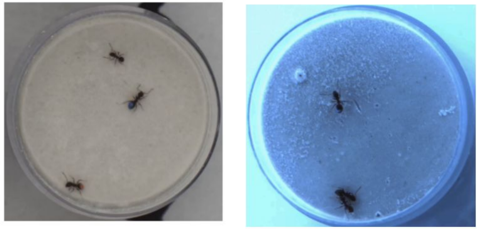
Thm: if a representation is valid on a training experiment and transfers to a target experiment while satisfying $\phi(\boldsymbol{X}) \perp\!\!\!\perp \boldsymbol{Z}|Y = y$, then it remains causally valid.

"*Prediction-Powered Causal Inference*", Cadei, Demirel, De Bartolomeis, Lindorfer, Cremer, Schmid, L; NeurIPS 2025

Results

33% confidence interval overlap





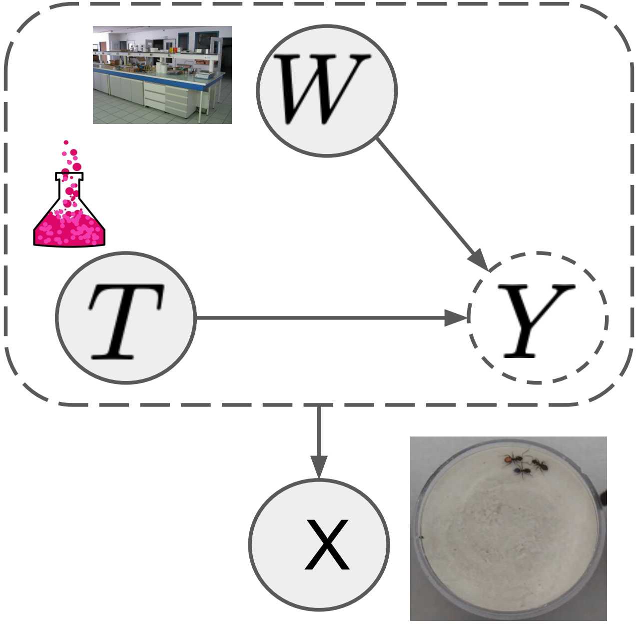
(a) ISTAnt (b) This study

Treatment decreases effect **Treatment increases effect**

"*Prediction-Powered Causal Inference*", Cadei, Demirel, De Bartolomeis, Lindorfer, Cremer, Schmid, L; NeurIPS 2025

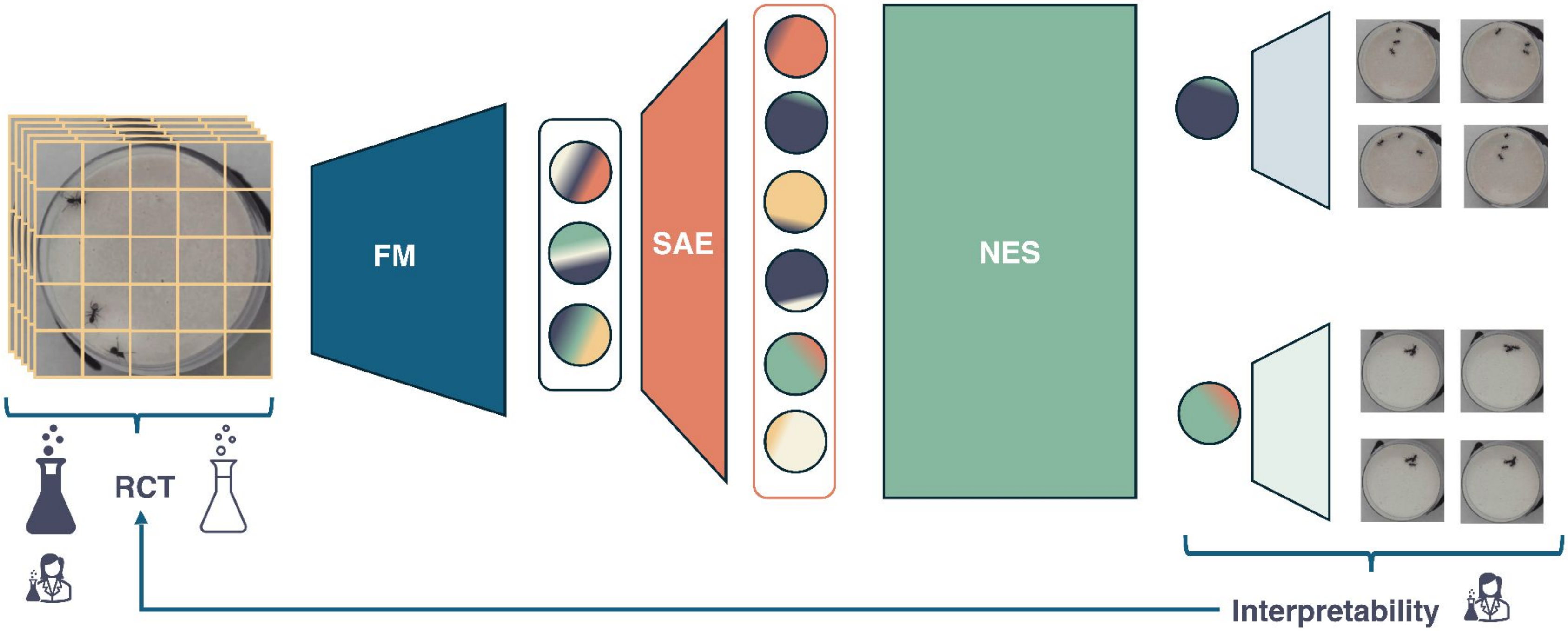
Exploratory causal inference

Exploratory causal inference



“*Exploratory Causal Inference in SAEnce*”, Mencattini*, Cadei*, L. Preprint 2025

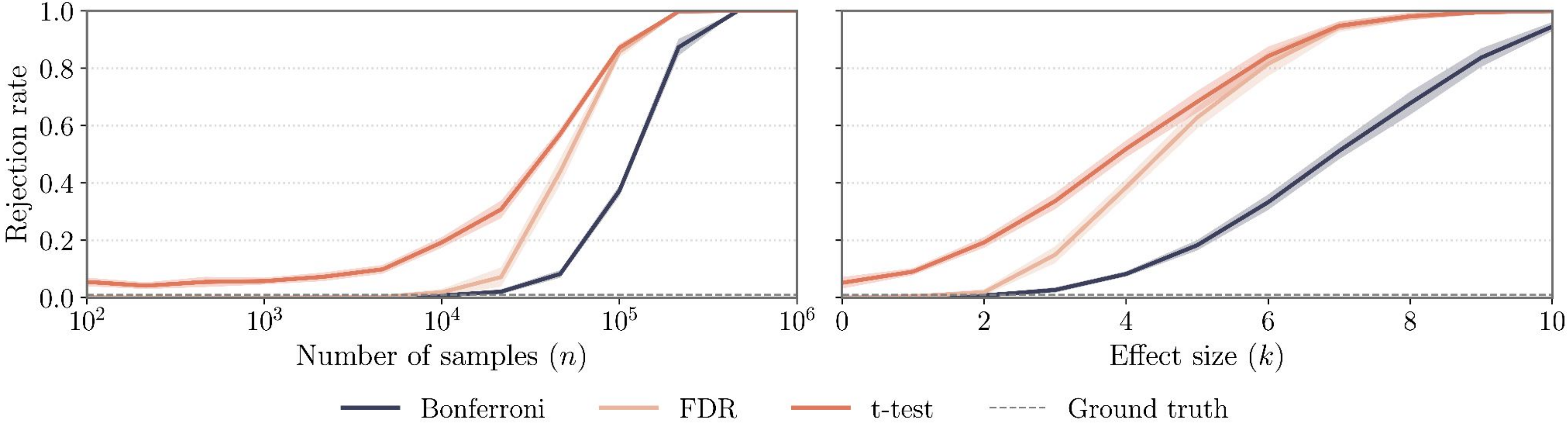
Deep networks as data-driven measurement devices



“*Exploratory Causal Inference in SAEnce*”, Mencattini*, Cadei*, L. Preprint 2025

Hypothesis testing challenges: the paradox of ECI

NES: Consistent recursive testing procedure that corrects the entanglement from previously discovered hypotheses.



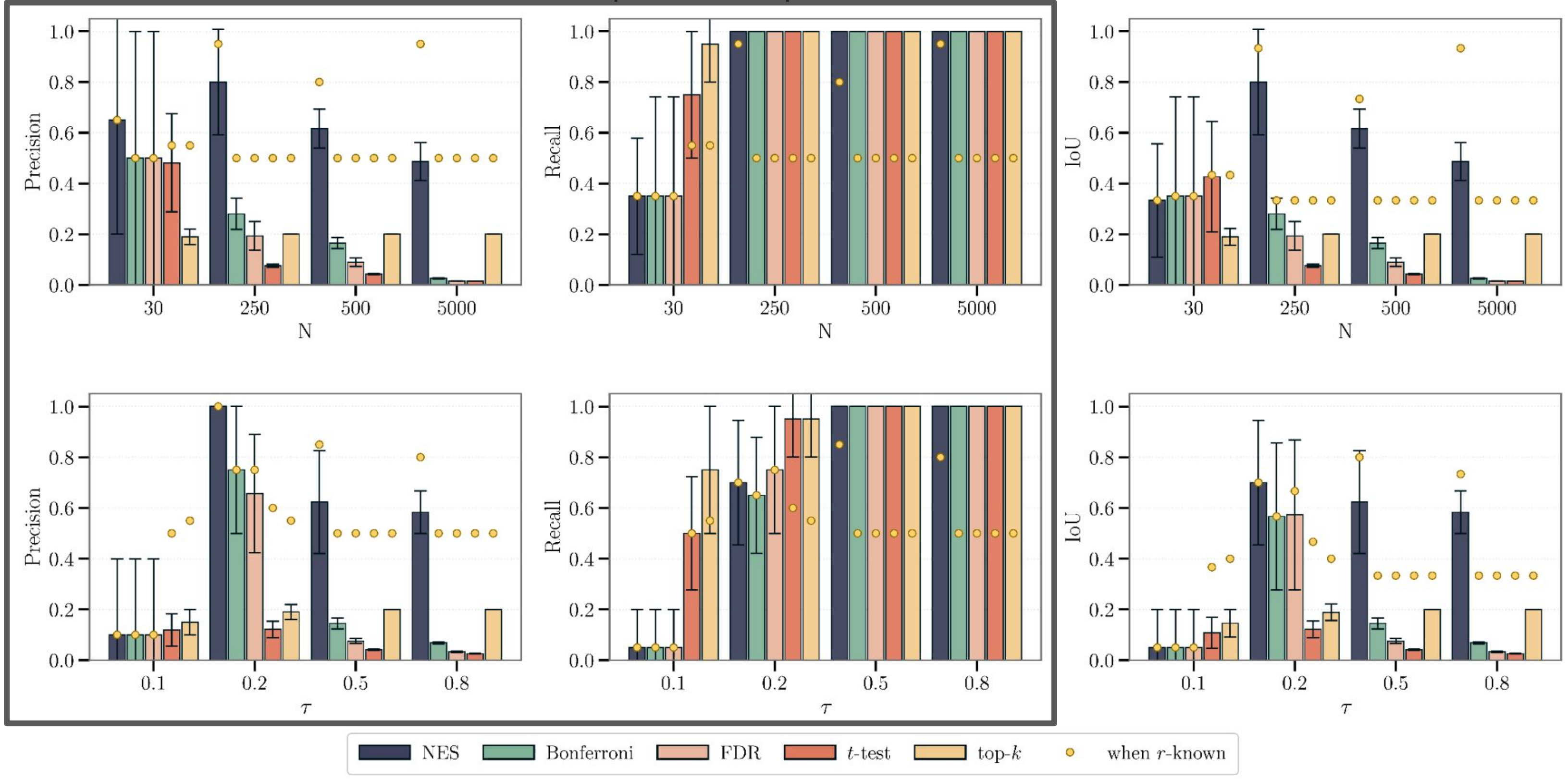
Idea: With powerful tests (strong effects or large sample sizes), all correlations are statistically significant. Small entanglement → all neurons are individual effects.

"*Exploratory Causal Inference in SAEnce*", Mencattini*, Cadei*, L. Preprint 2025

Results on synthetic trials

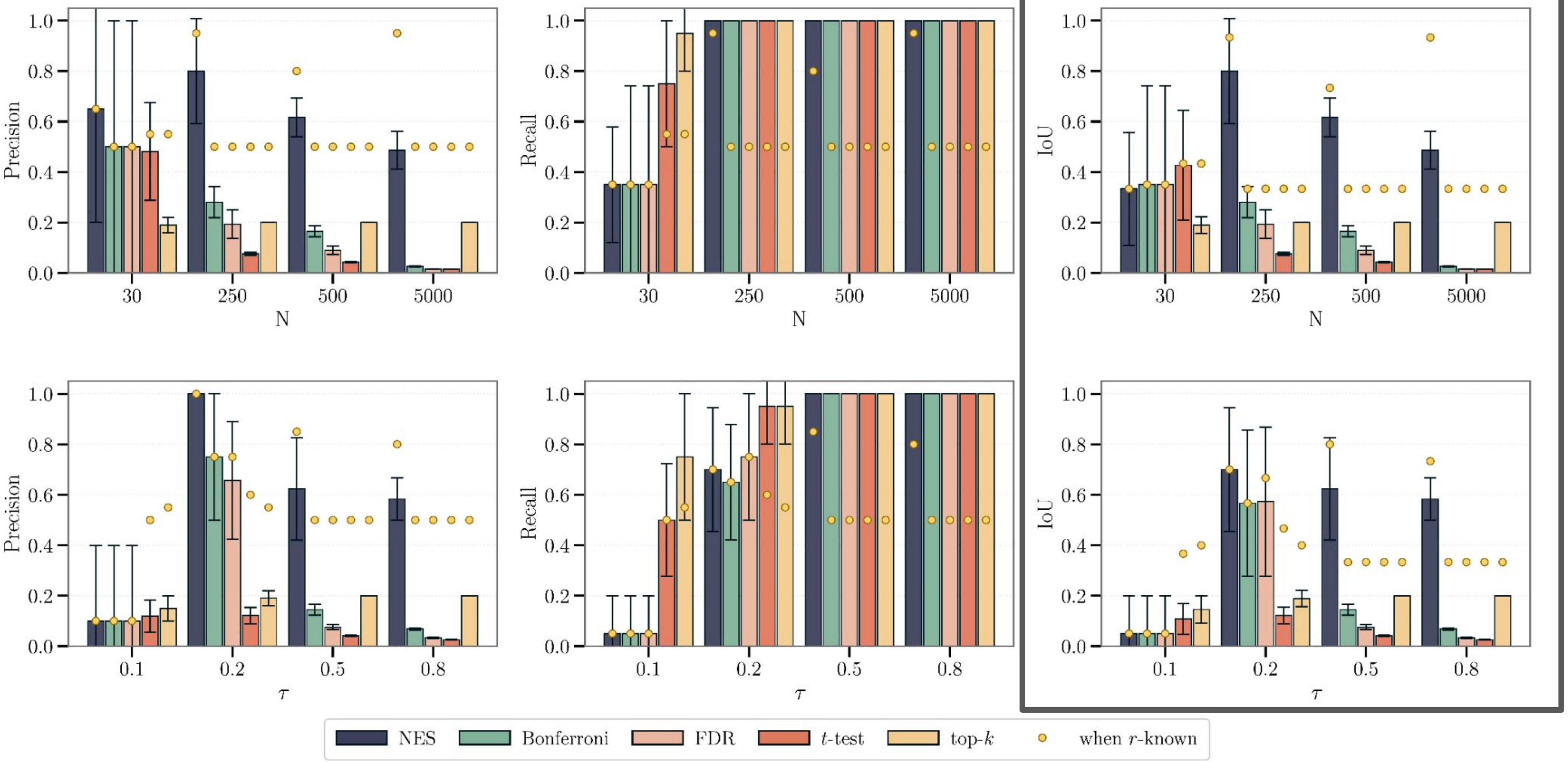
Results on synthetic trials

Paradox of ECI: Precision collapses w/ powerful tests



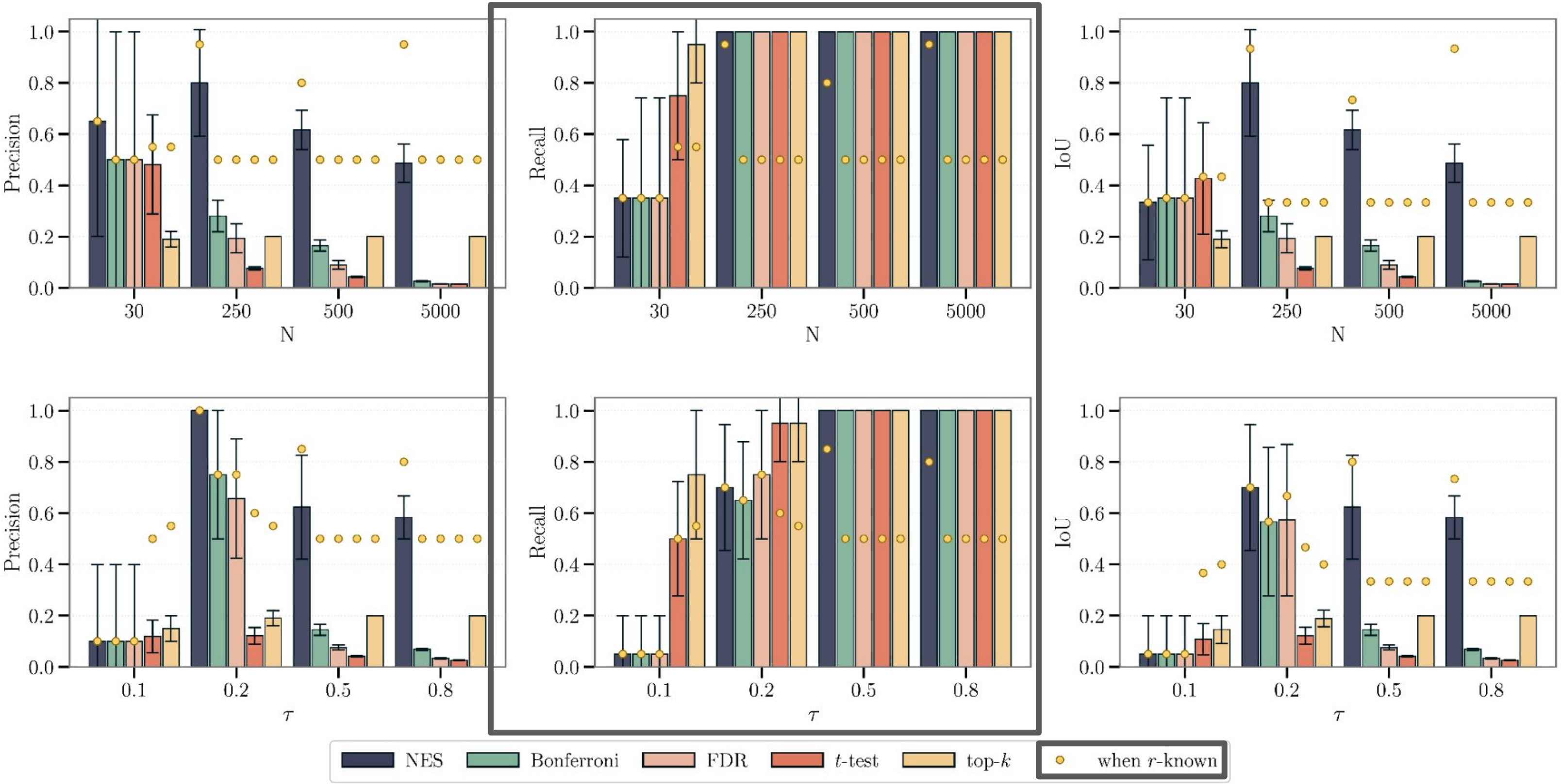
Results on synthetic trials

NES is robust and works well

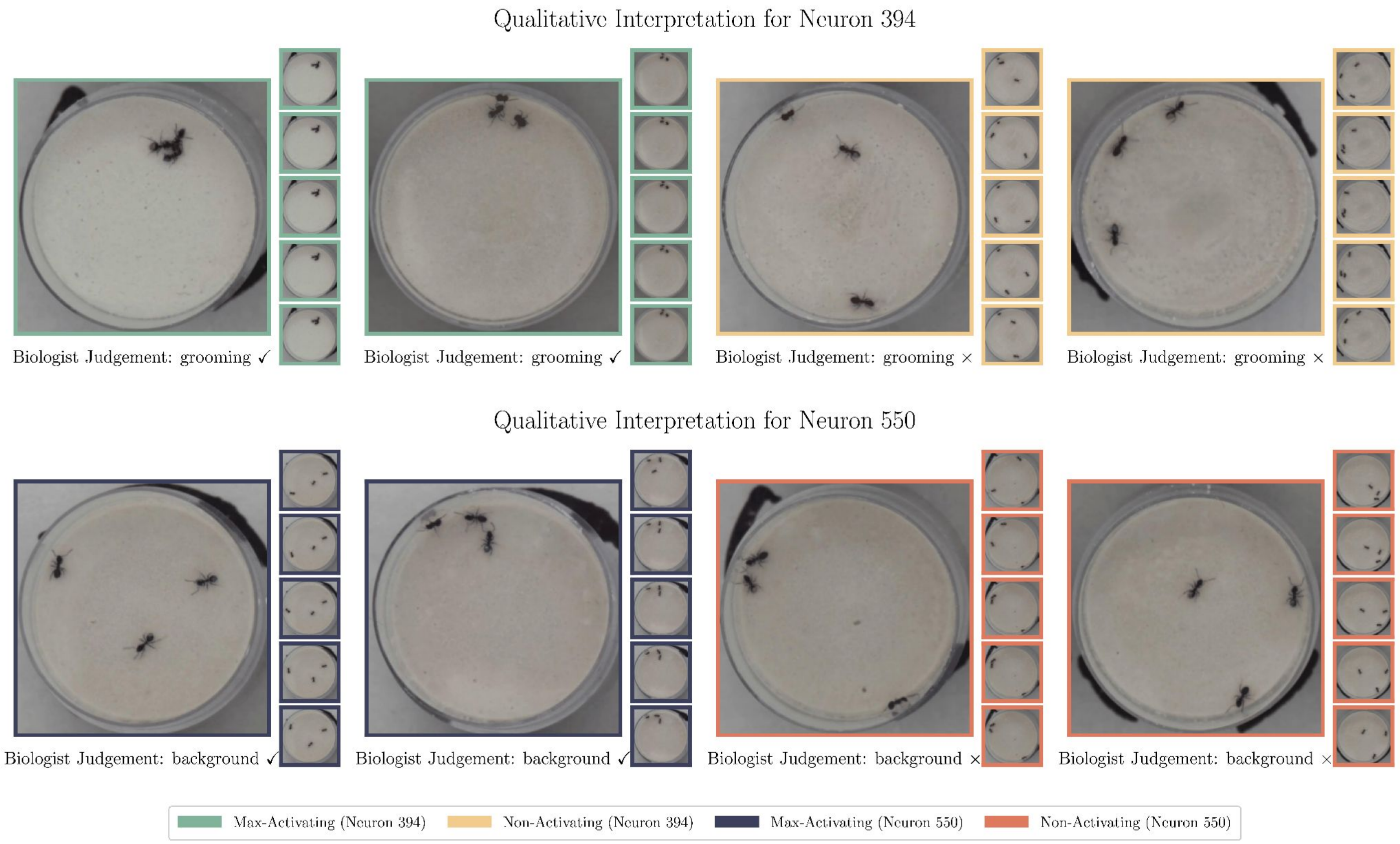


Results on synthetic trials

If r is known, NES does not miss effects

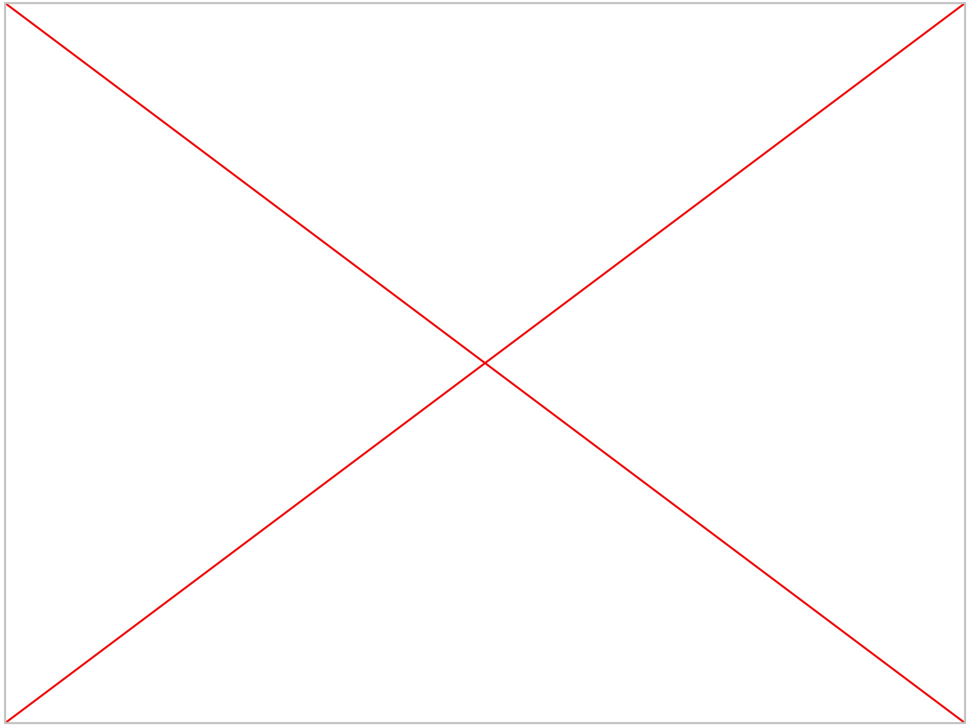


Results on ISTAnt



Beyond “standard” causal models

Temporal dynamics



Dynamical systems as causal models

System of coupled differential equations *modeling physical mechanisms* responsible for time evolution

$$\frac{dx}{dt} = f(x), \quad x \in \mathbb{R}^d$$

Future states are "caused" by immediate past

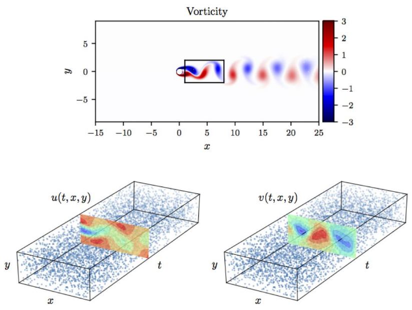
$$x(t + dt) = x(t) + dt \cdot f(x(t))$$

Unclear to which extent these can be learned from data for non-linear systems

For the equation to have causal meaning, f must be a causal mechanism (i.e., interventions are well defined and align with physical experiments)

ML methodologies

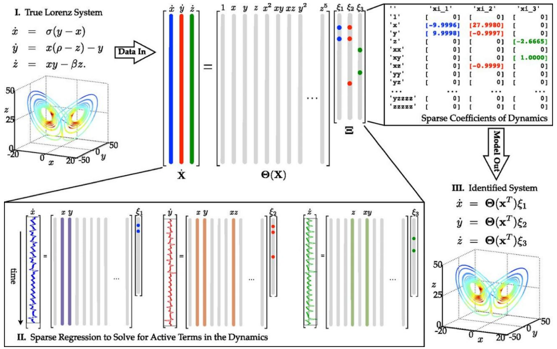
PINNs



Raissi et al., 2019

- Constrain evolution prediction to **follow physics**
- Embeds physics **implicitly via loss**

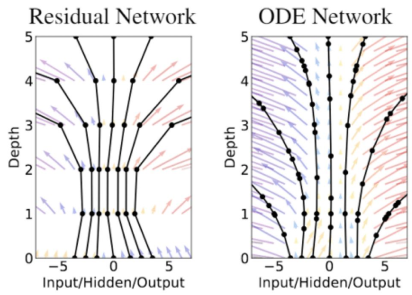
SINDy



Brunton et al., 2016

- **Discover governing equation** regressing time derivatives onto basis functions
- **No learned representations**
- **Reconstruct time derivatives**

NeuralODE

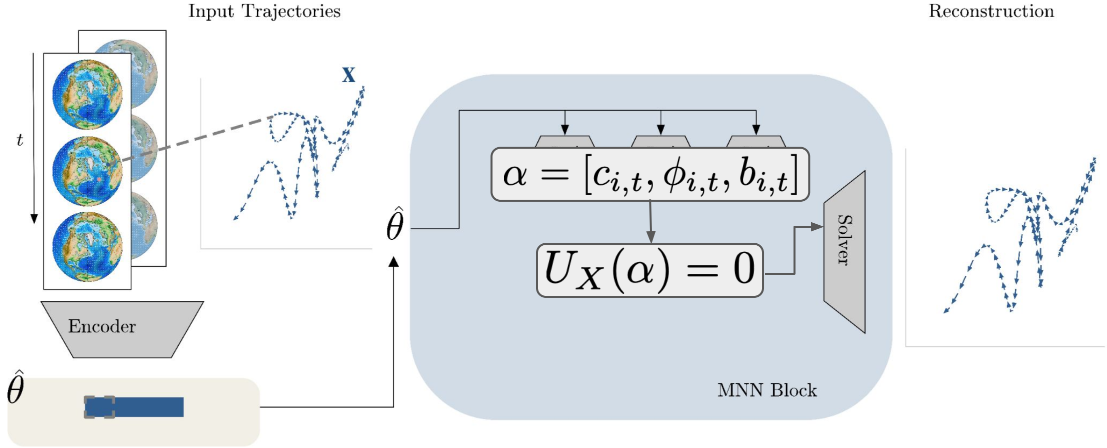


Chen et al., 2018

- Embed **network in solver** to predict next states
- Train by **reconstruction error**

Model Architecture

Parameterize equation as a combination of **learnable** coefficients parameterized by deep neural networks. Trained fully end-to-end.



$$U_X : \sum_i \underbrace{c_i(t, X) u^{(i)}}_{\text{linear terms}} + \sum_k \underbrace{g_k(t, u, u', \ldots; \phi(t, X))}_{\text{nonlinear terms}} = b(t, X)$$

"*Mechanistic Neural Networks for scientific ML*" Pervez, L, Gavves. ICML 2024

Direct least squares solution

	Time Complexity	Space Complexity
Dense Solver	$\boldsymbol{O(T^3)}$	$O(T^2)$
Sparse Solver	$O(T^2)$	$O(T^2)$

- Linear system $\sum_i c_{i,t} u^{(i)} = b_t \longrightarrow Az = b$

$$z = (A^T A)^{-1} A^T b$$

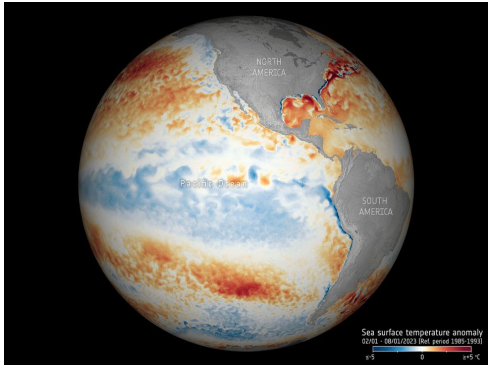
$$z = M^{-1} \beta$$

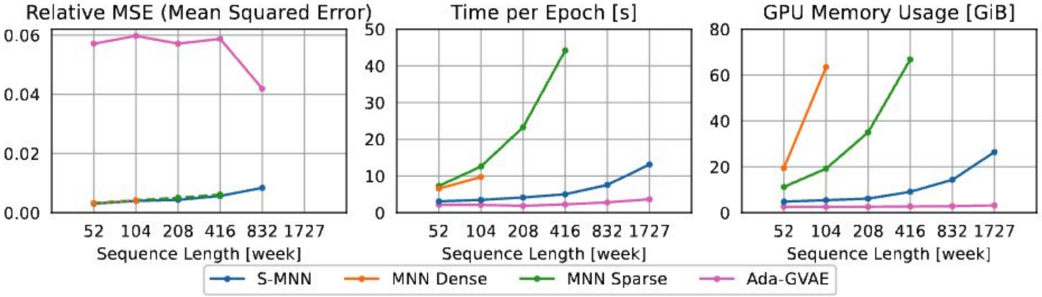
- M is a banded symmetric matrix → solver with linear time and space complexity

- In both cases, we also need to derive the backward gradients for GPU implementation

"*Scalable mechanistic neural networks*", Chen, Yao, Pervez, Alistarh, L. ICLR 2025

Long term sea surface temperature forecasting

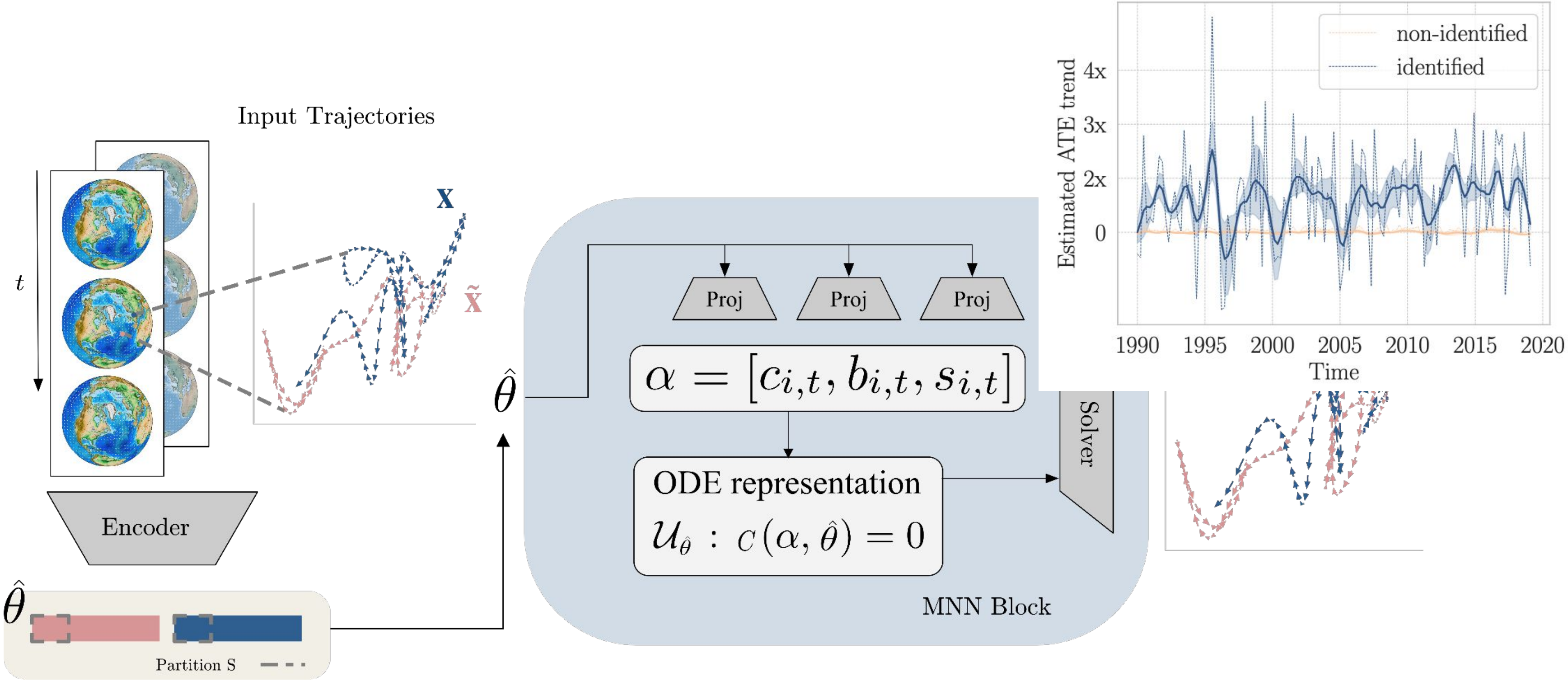




Implication: Better solver immediately improves scalability without affecting accuracy

"*Scalable mechanistic neural networks*", Chen, Yao, Pervez, Alistarh, L. ICLR 2025

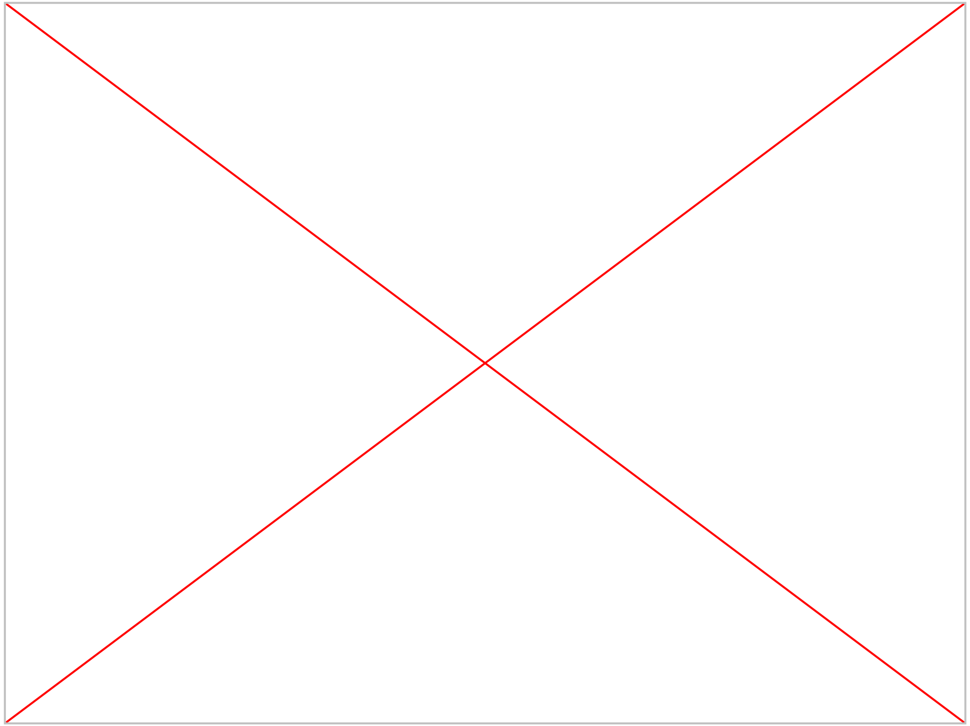
CRL recipes also extend to dynamical systems



"Marrying Causal Representation Learning with Dynamical Systems for Science", Yao, Muller, L; NeurIPS 2024

"The arctic has warmed nearly four times faster than the globe since 1979", Rantanen et al., Nature Communications Earth & Environment 2022.

PDE extension: Ginzburg-Landau Reaction Diffusion



"*Mechanistic PDE Networks for discovery of governing equations*" Pervez, Gavves, L. ICML 2025

Application: Modeling cell differentiation

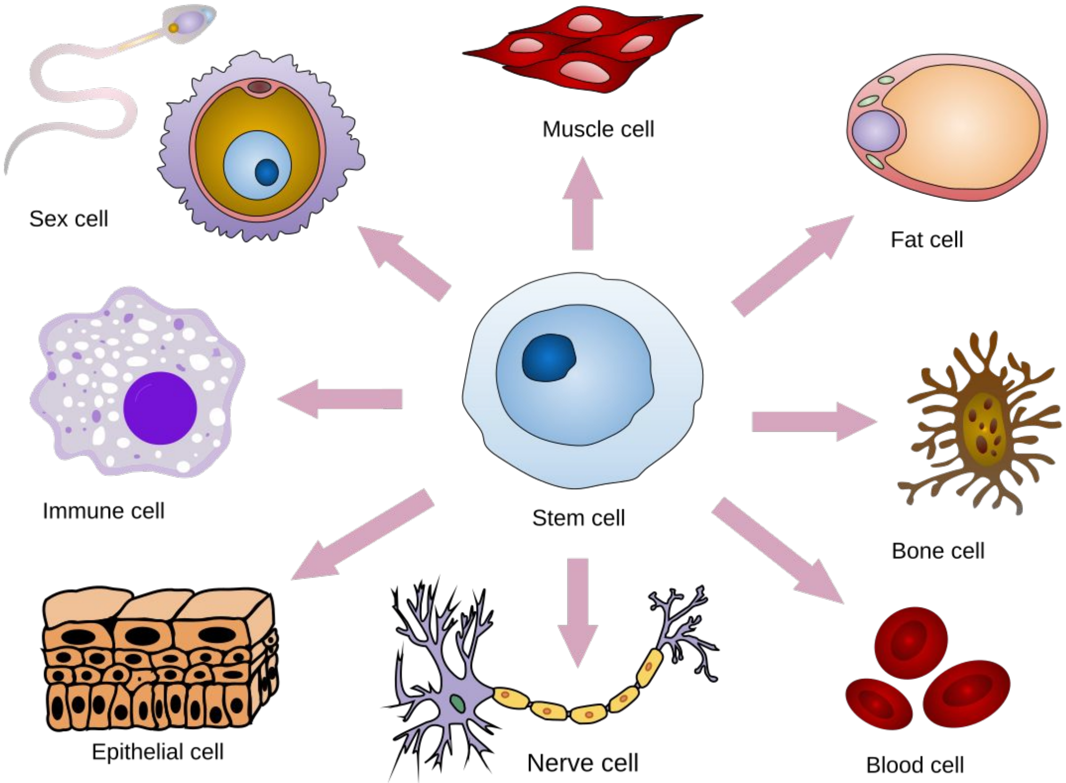
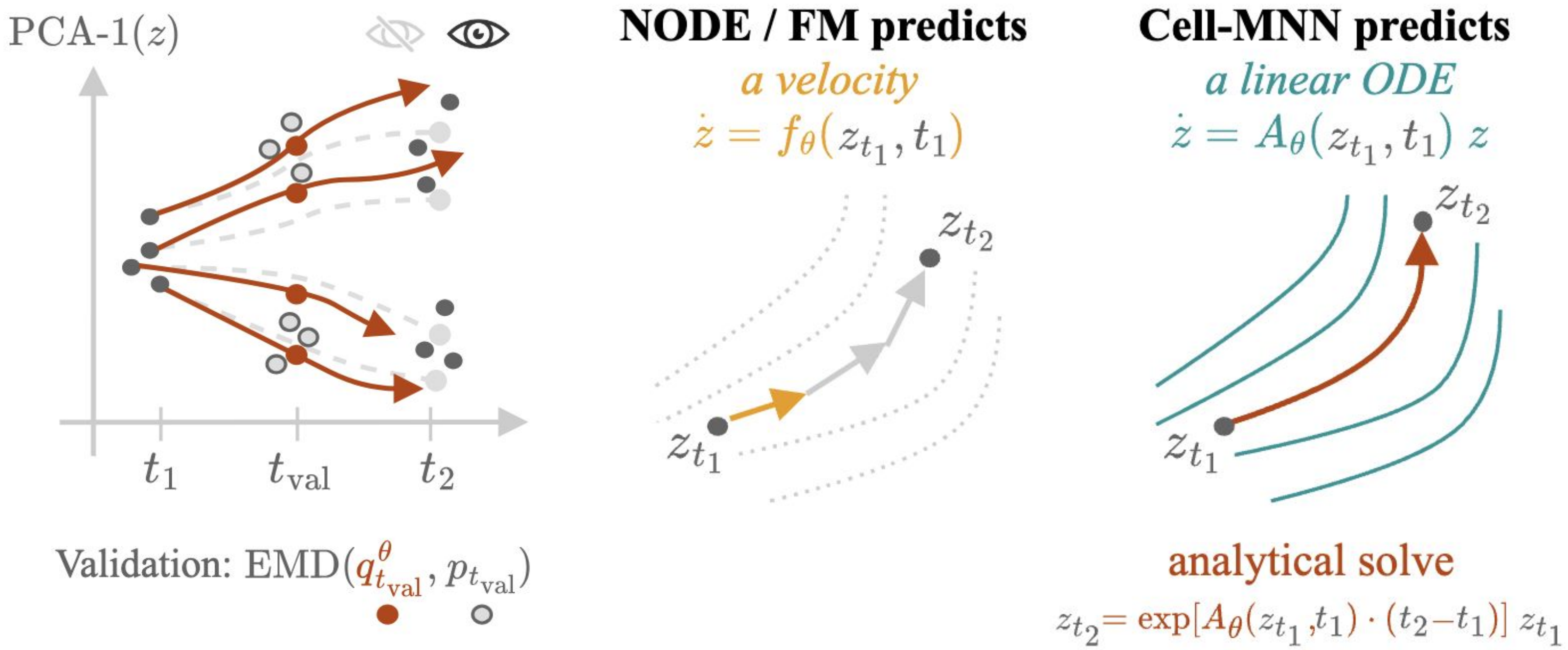


Image credit: Wikipedia (Cellular Differentiation)

Application: Modeling cell differentiation



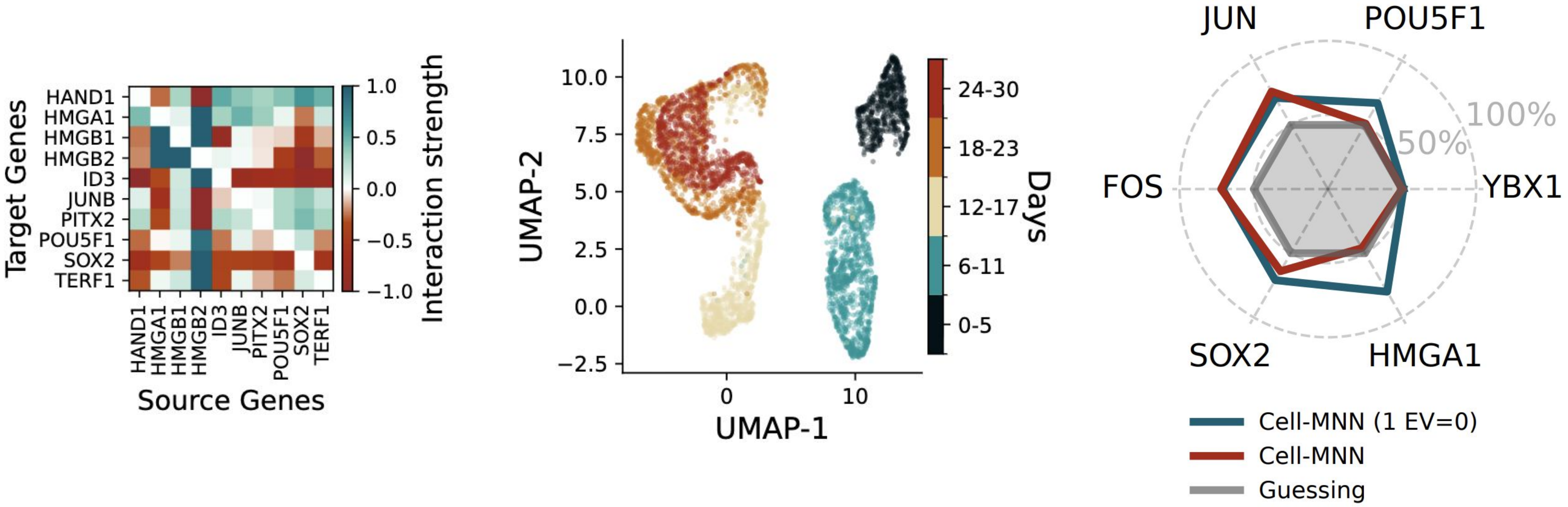
"*Learning Explicit Single-Cell Dynamics Using ODE Representations*", von Bassewitz, Pervez, Fumero, Robinson, Karaletsos, L. Preprint 2025

Performance

Method	Cite	EB	Multi	Average ↓
TrajectoryNet [50]	–	0.848	–	–
WLF-UOT [39]	–	0.800 ± 0.002	–	–
NLSB [27]	–	0.777 ± 0.021	–	–
SB-CFM [52]	1.067 ± 0.107	1.221 ± 0.380	1.129 ± 0.363	1.139 ± 0.077
[SF]^{2}M-Sink [53]	1.054 ± 0.087	1.198 ± 0.342	1.098 ± 0.308	1.117 ± 0.074
[SF]^{2}M-Geo [53]	1.017 ± 0.104	0.879 ± 0.148	1.255 ± 0.179	1.050 ± 0.190
I-CFM [52]	0.965 ± 0.111	0,872 ± 0.087	1.085 ± 0.099	0.974 ± 0.107
DSB [14]	0.965 ± 0.111	0.862 ± 0.023	1.079 ± 0.117	0.969 ± 0.109
I-MFM [24]	0.916 ± 0.124	0.822 ± 0.042	1.053 ± 0.095	0.930 ± 0.116
[SF]^{2}M-Exact [53]	0.920 ± 0.049	0.793 ± 0.066	0.933 ± 0.054	0.882 ± 0.077
OT-CFM [52]	0.882 ± 0.058	0.790 ± 0.068	0.937 ± 0.054	0.870 ± 0.074
DeepRUOT [57]*	0.845 ± 0.167	0.776 ± 0.079	0.919 ± 0.090	0.846 ± 0.071
OT-Interpolate*	0.821 ± 0.004	0.749 ± 0.019	<u>0.830 ± 0.053</u>	0.800 ± 0.044
OT-MFM [24]	**0.724 ± 0.070**	<u>0.713 ± 0.039</u>	0.890 ± 0.123	<u>0.776 ± 0.099</u>
Cell-MNN (ours)*	<u>0.791 ± 0.022</u>	**0.690 ± 0.073**	**0.742 ± 0.100**	**0.741 ± 0.050**

"*Learning Explicit Single-Cell Dynamics Using ODE Representations*", von Bassewitz, Pervez, Fumero, Robinson, Karaletsos, L. Preprint 2025

Discovering GRN



“Learning Explicit Single-Cell Dynamics Using ODE Representations”, von Bassewitz, Pervez, Fumero, Robinson, Karaletsos, L. Preprint 2025

Concluding remarks

Discussion

- ML has a great opportunity in powering causal analysis with key applications in scientific discovery but causal questions are subtle

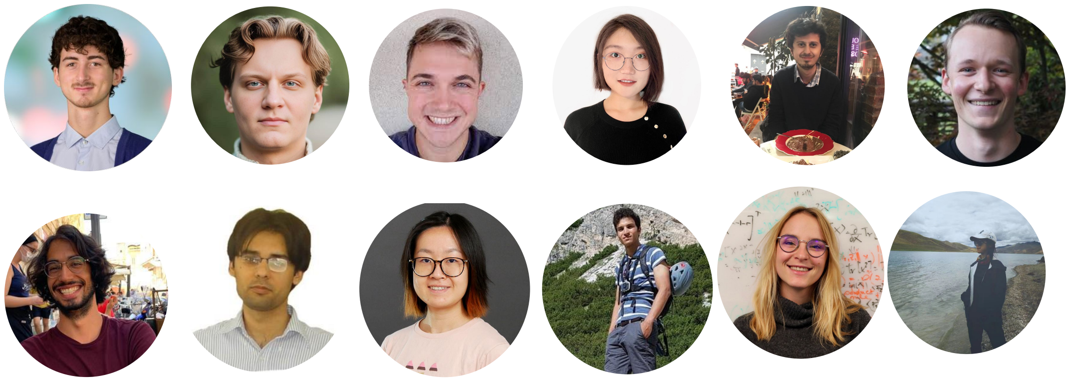
- Chasing predictive accuracy may not lead to more accurate causal conclusions. In AI4Science, scientific questions should be part of the benchmark (especially if they are causal)

- Seeing the hidden world requires assumptions that the statistical causality language can describe

- **CRL:** Representation that makes it easier/ possible to extract causal information with some downstream estimator

Thanks!

FFG Promoting Innovation.
Google
BILAI Bilateral Artificial Intelligence
CHAN ZUCKERBERG INITIATIVE
Google Research Award
ISTA Institute of Science and Technology Austria



You?

I'm hiring PhD students and PostDocs!

Causal Learning and AI