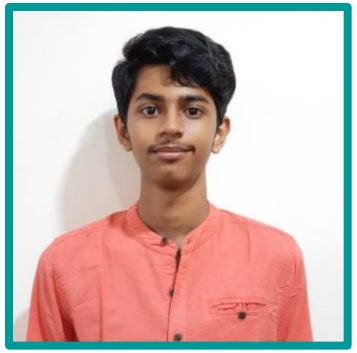
PLOT:

Progressive Localization via Optimal Transport in Neural Causal Abstractions

Ziv Goldfeld

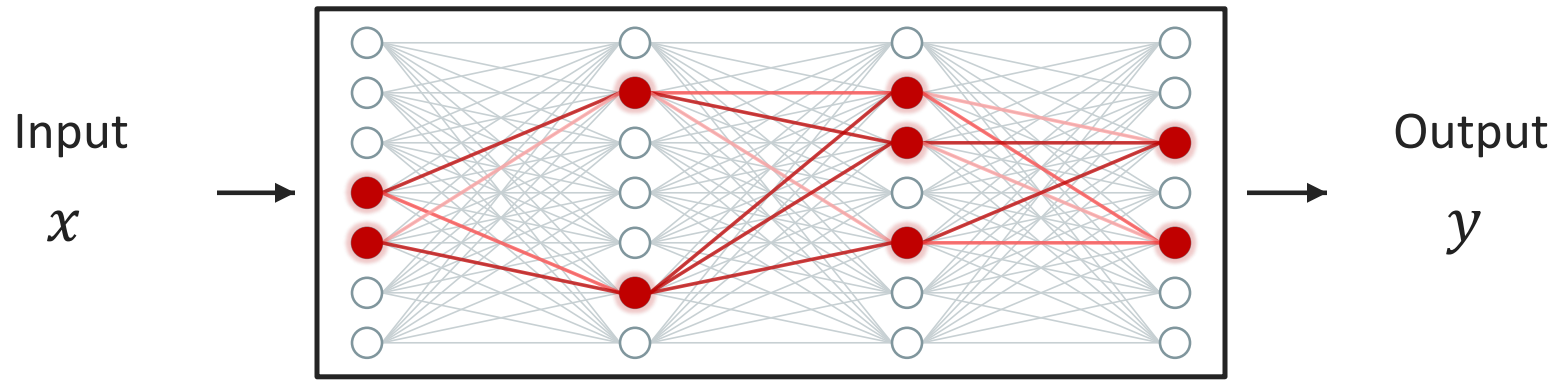


Jonathn Chang



Arya Datla

AI Interpretability for Safety: Opening the Black Box



Pragmatic

Use internal access to detect/prevent failures

Monitoring · diagnosis · steering

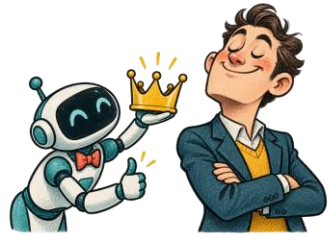
Mechanistic

Uncover internal comp. underlying task/behavior

Representations · operations · interactions

What computations produce these behaviors?

Sycophancy	Stated user belief causes model to abandon correct answer
Evaluation awareness	Model recognizes it is being tested and changes its behavior
Strategic deception	Model misleading by mistake or as a strategy?







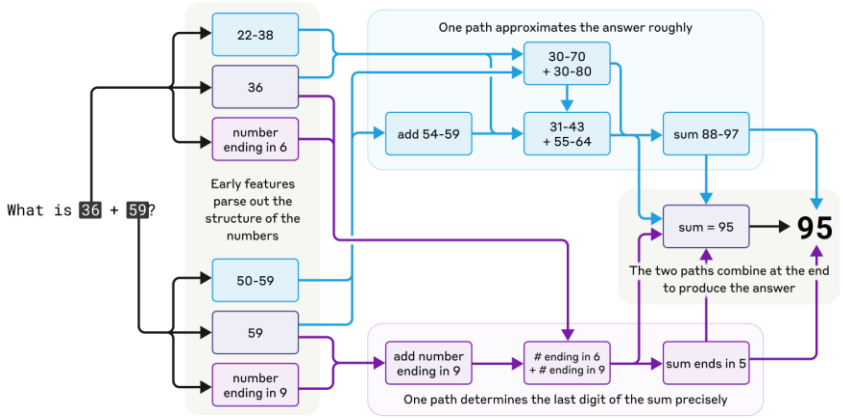
⇒ Understanding mechanisms informs detection, safer training, intervention

Mechanistic Interpretability: Bottom-Up vs. Top-Down

Bottom-up (Circuit tracing)

Start with network & prompt

Learn features, trace effects, attribution graph

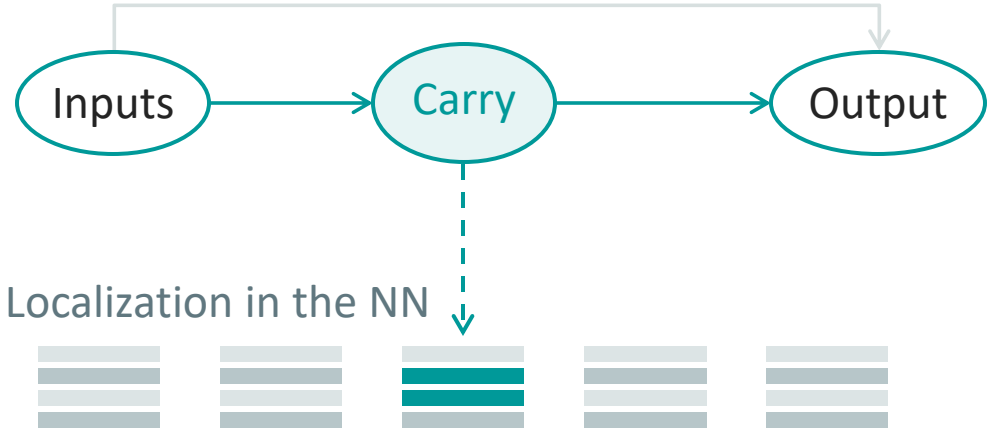


[Anthropic, 2025] “On the biology of LLMs”

- Grounded in actual network
- Complex, prompt-specific, high variance

Top-down (Causal abstractions)

Start with hypothesized causal model



[Geiger, Lu, Icard & Potts, 2021] “Causal abstractions of NNs”

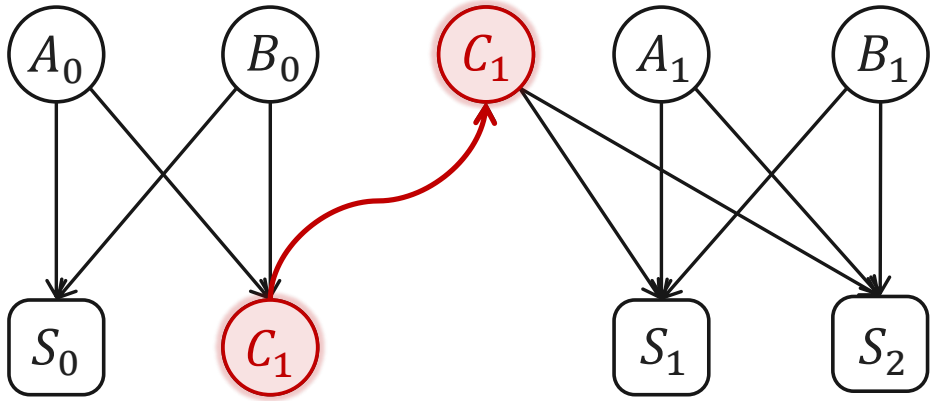
- Compact, testable account of the task
- The hypothesized computation may be absent

Neural Causal Abstractions: Where is the Carry?

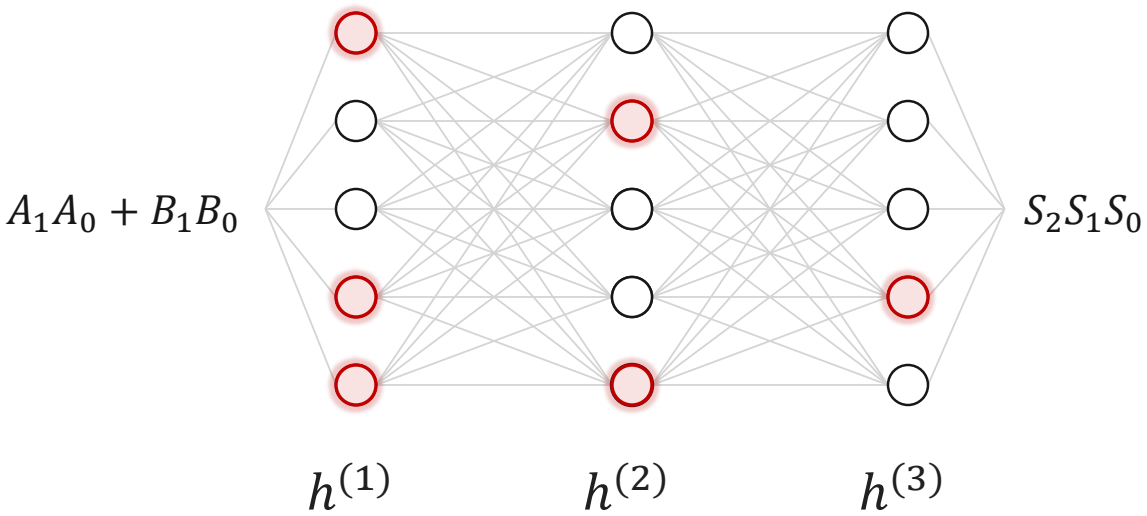
Task: Align high-level causal model with low-level neural computational model

Causal model: Ripple carry algorithm

$$\begin{array}{r} {\scriptstyle 1} \\ A_1 A_0 \\ + \; B_1 B_0 \\ \hline S_2 S_1 S_0 \end{array}$$



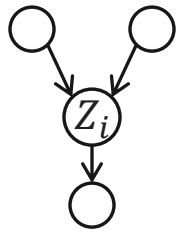
Neural model: Trained on data



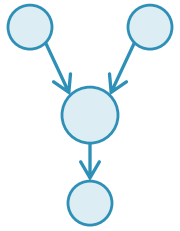
Carry handle: A neuron, a subspace, or a different computation?

Interchange Interventions: Causal Model

$\mathfrak{C}$: $\mathcal{X} \rightarrow \mathcal{Y}$ with variables $Z_1, \ldots, Z_n$



Base x^b



Source x^s



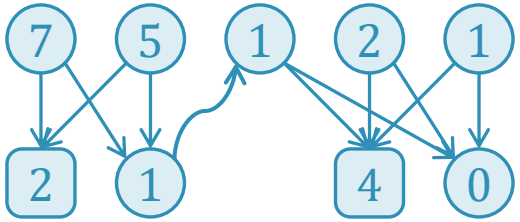
Interchange

$\text{do}(Z_i \leftarrow Z_i(x^s))$

$$\hat{Y}_i(x^b, x^s) \coloneqq \mathfrak{C}\left(x^b; \text{do}(Z_i \leftarrow Z_i(x^s))\right)$$

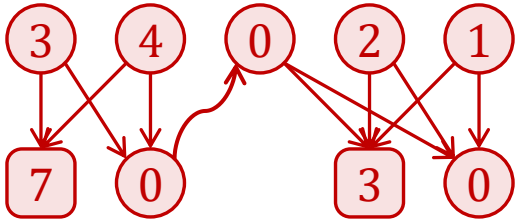
Addition

Base $x^b = (27,15)$



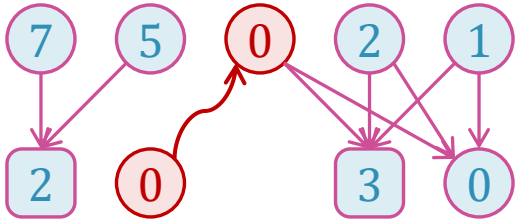
$Y(x^b) = 42$

Source $x^s = (23,14)$



$Y(x^s) = 37$

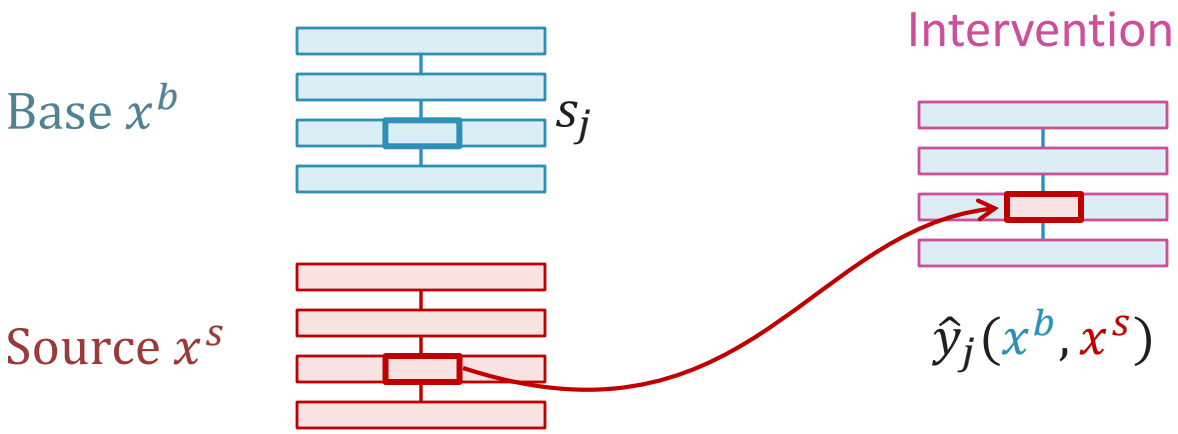
Interchange

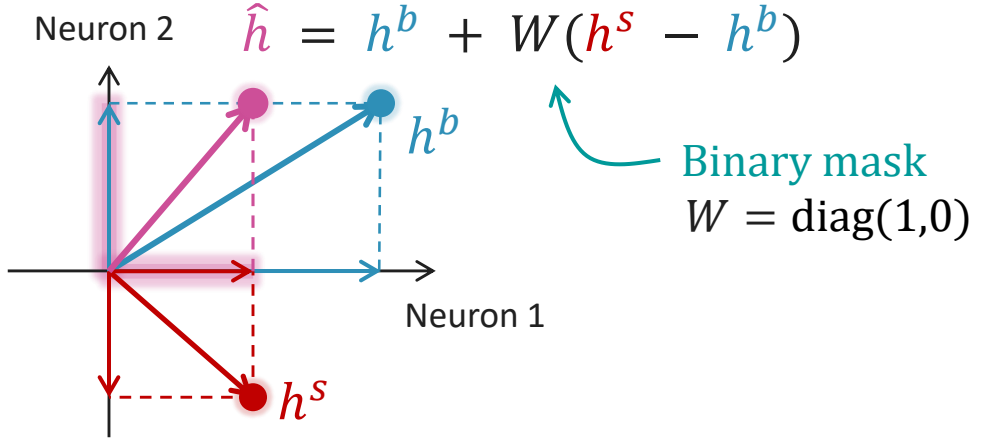


$\hat{Y}(x^b, x^s) = 32$

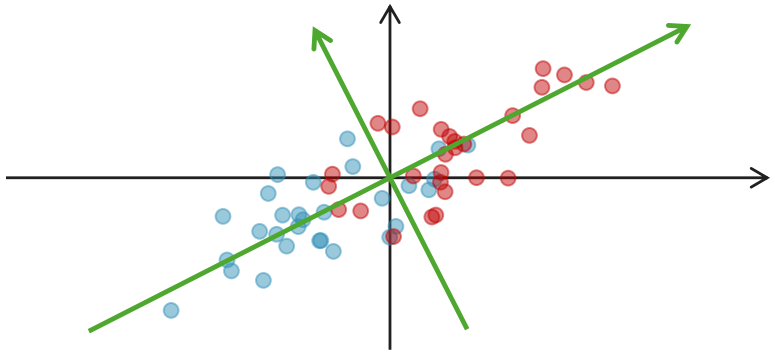
Interchange Interventions: Neural Model

$\mathfrak{N}$: $\mathcal{X} \rightarrow \mathcal{Y}$ with candidate sites $s_1, \ldots, s_m$





Intervention in a rotated subspace



- Rotate to subspace coordinates

  $z = Uh, \qquad U \in \mathbb{R}^{r \times d}, \qquad UU^\top = I_r$

- Intervene in that subspace

  $\hat{h} = h^b + U^\top U(h^s - h^b)$

Finding a Neural Handle

Performance criteria:

$$\mathrm{IIA}_{\mathrm{test}}(Z_i, s_j) \coloneqq \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left\{ \hat{Y}_i(x_t^b, x_t^s) \sim \hat{y}_j(x_t^b, x_t^s) \right\}$$

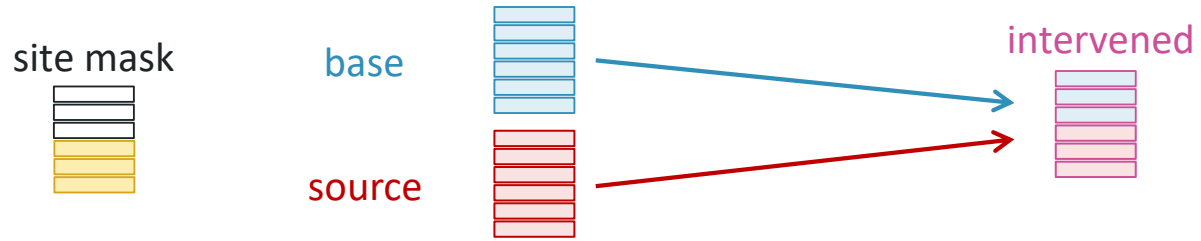
Ranking

[Vig et al.'20]



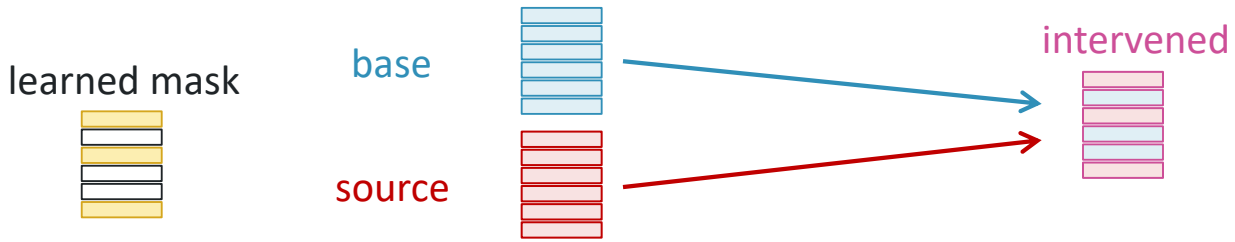
Exhaustive search

[Meng et al.'22]



Desiderata-based masking (DBM)

[Chaudhary-Geiger'24]



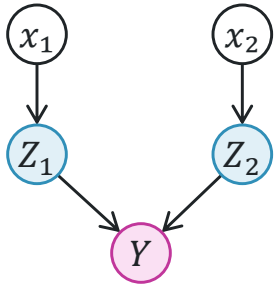
Distributed alignment search (DAS)

[Geiger et al.'24]

Learn handle in rotated subspace

Distributed Alignment Search (DAS): Motivation

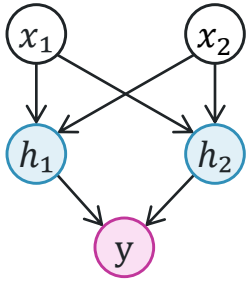
Logical AND



$$x \in \{0,1\}^2$$

$$Z_1 = x_1, Z_2 = x_2, Y = Z_1 \wedge Z_2$$

Neural model

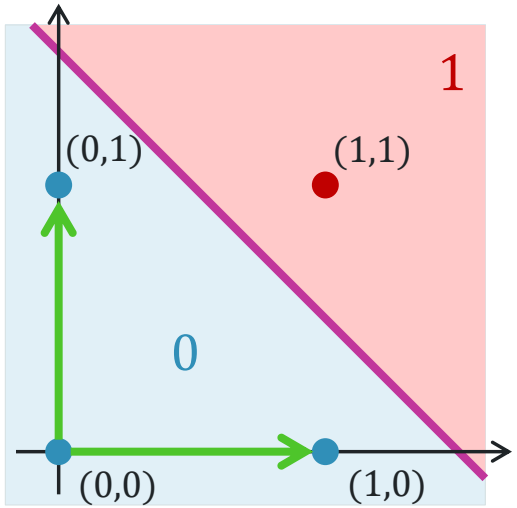


$$h = Wx$$

$$\mathrm{y} = \mathbb{1}_{\{w^\top h + b > 0\}}$$

Canonical

$$W = I_2, \quad w = (1,1), \quad b = -3/2$$



Handles

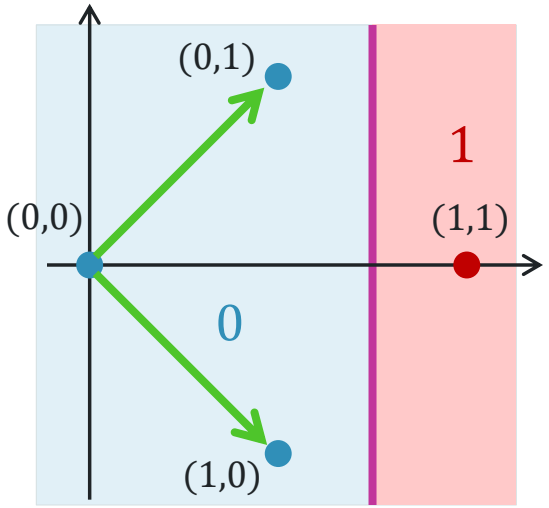
$$Z_1 \leftrightarrow h_1$$

$$Z_2 \leftrightarrow h_2$$

Rotated

$$W = R(-45°), \quad w = (\sqrt{2}, 0), \quad b = -3/2$$

$$h_1 = (x_1 + x_2)/\sqrt{2}, \quad h_2 = (x_2 - x_1)/\sqrt{2}$$



Handles

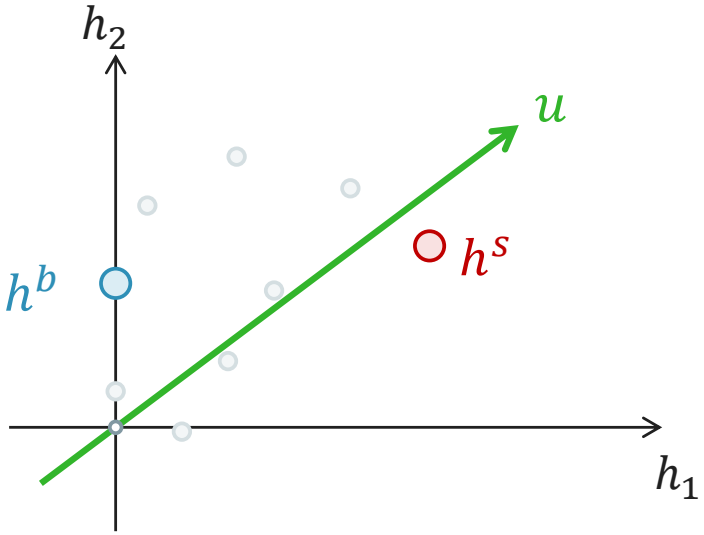
$$Z_1 \leftrightarrow (h_1 - h_2)/\sqrt{2}$$

$$Z_2 \leftrightarrow (h_1 + h_2)/\sqrt{2}$$

Linear Representations and Rotated Subspaces

Working hypothesis: A variable may be represented in a linear subspace.

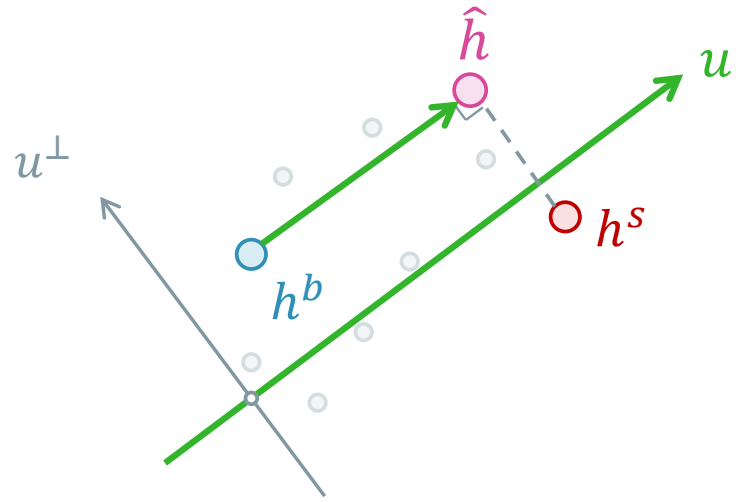
Hidden coordinates



$Z_i \leftrightarrow u^\top h$

Basis rotation

Aligned coordinates



Source along u, base along $u^\perp$

Swap: $\hat{h} = h^b + uu^\top(h^s - h^b), \qquad \|u\|^2 = 1$

Beyond rotations: Same frozen model, broader intervention family.

[Geiger et al.'24]

DAS Algorithm per Causal Variable

Canonical: h^b, h^s

Learned r-subspace: Dh^b, Dh^s swap

Canonical: $\hat{h}_D = h^b + D^\top D(h^s - h^b)$

Layer ℓ $\in \mathbb{R}^d$

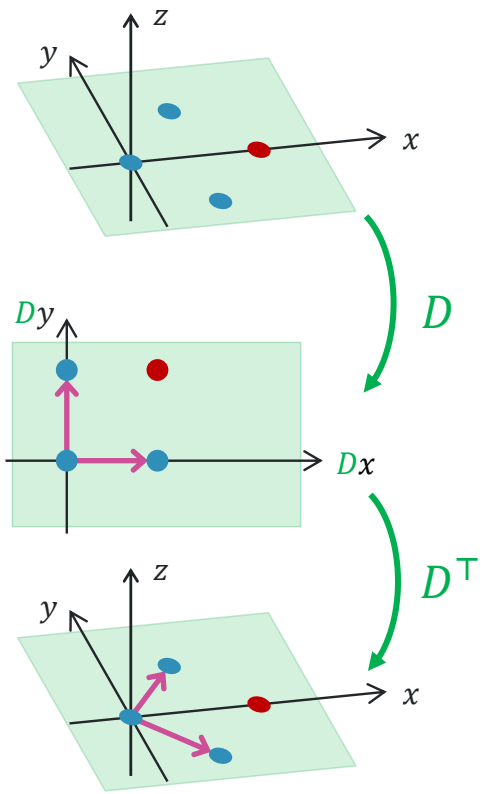
Subspace rotation: $D \in \mathbb{R}^{r \times d}$, $DD^\top = I_r$

D

$\in \mathbb{R}^r$

$D^\top$

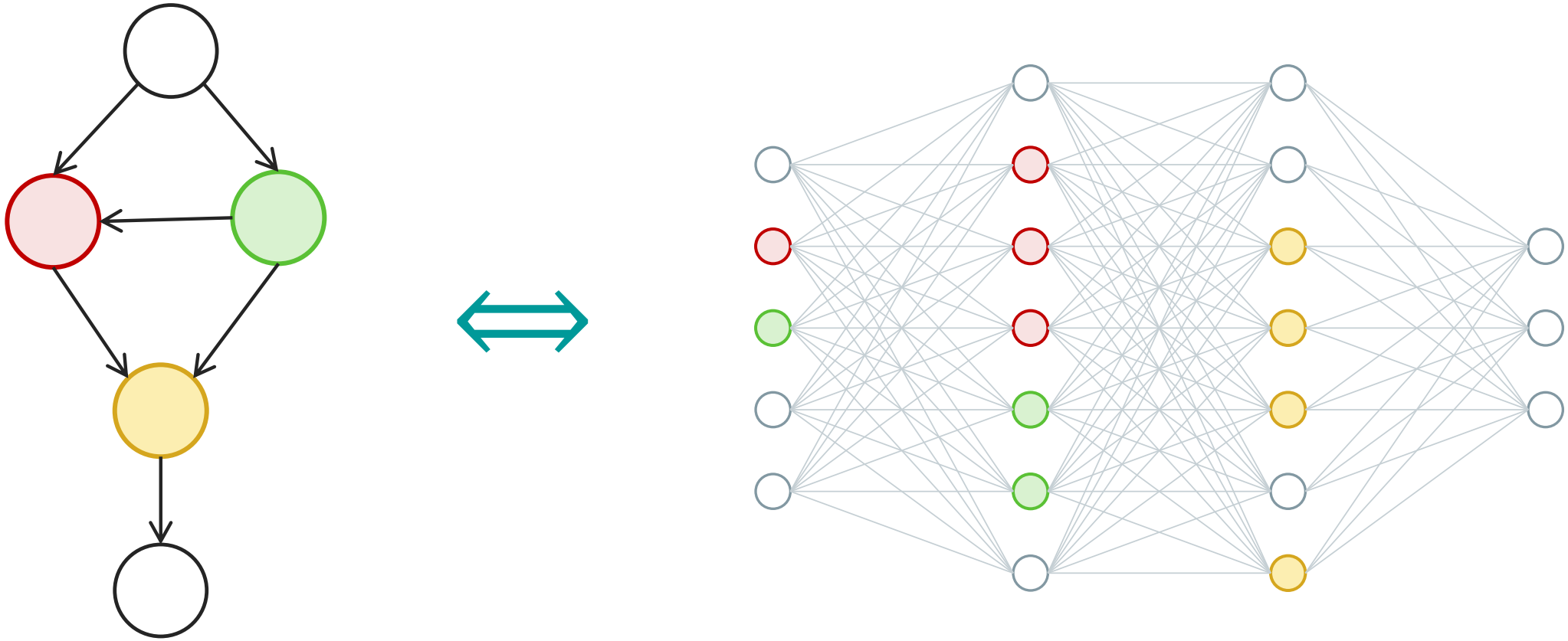
Layer ℓ $\in \mathbb{R}^d$



Training objective (per ℓ, r): $\min_{D:\, DD^\top = I_r} \mathbb{E}_{(b,s)} \left[\mathrm{CE}\left(\hat{Y}_i(x^b, x^s), \hat{y}_\ell^D(x^b, x^s) \right) \right]$

Issues: Nonconvex optimization **&** Exhaustive search over variables, layers, dimensions*

Progressive Localization via Optimal Transport



Optimal Transport

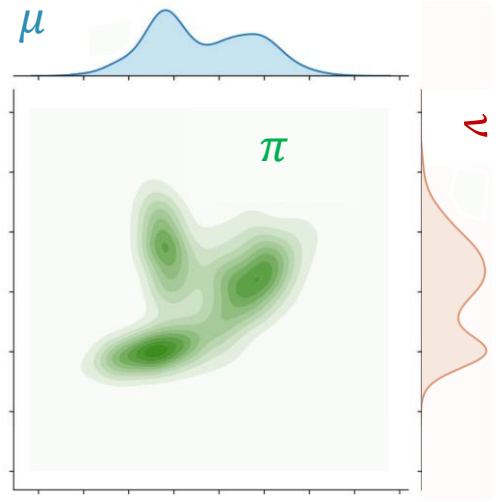


Kantorovich (1942)

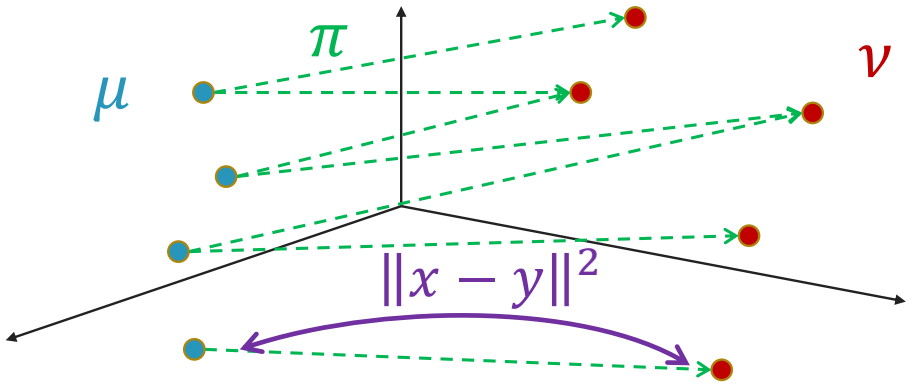
Optimal Transport

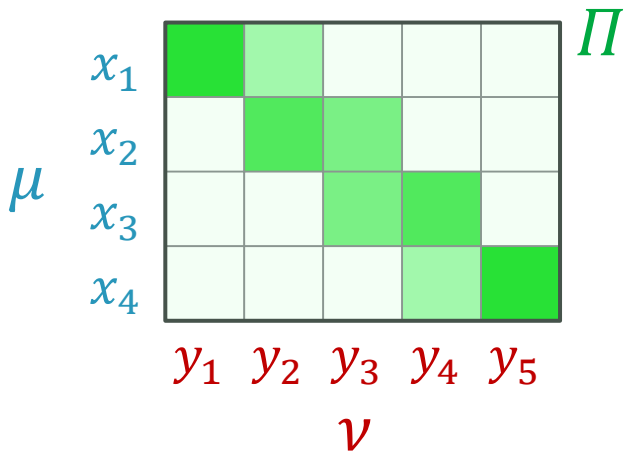
$$\mathrm{OT}_c(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}[c(X,Y)]$$

Coupling (transport plan):
$\pi(x,y) = \mu(x)\pi(y|x)$



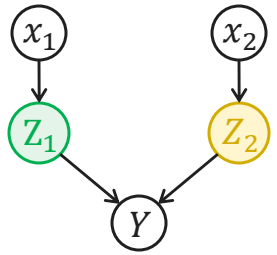
Discrete quadratic OT: $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$, $\nu = \frac{1}{m}\sum_{j=1}^{m}\delta_{y_j}$, $c(x,y) = \|x-y\|^2$





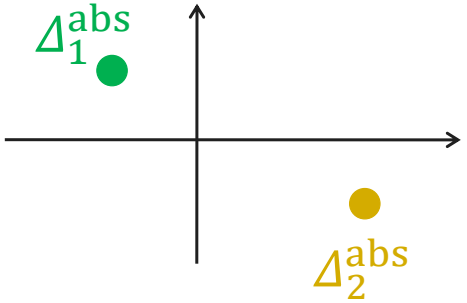
Interchange Intervention Geometry

Causal: Dataset $\mathcal{D} = \{(x_t^b, x_t^s)\}_{t=1}^{T}$ & variables $\{Z_i\}_{i=1}^{n}$



$Y(x_t^b)$ vs $\hat{Y}_i(x_t^b, x_t^s)$

$\Delta_{i,t}^{\text{abs}} = \hat{Y}_i(x_t^b, x_t^s) - Y(x_t^b)$

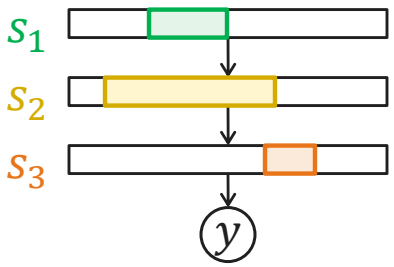


$$\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\Delta_i^{\text{abs}}}$$

- **Record:** $\Delta_{i,t}^{\text{abs}}, \quad t = 1, \dots, T, \quad i = 1, \dots, n,$
- **Aggregate:** $\Delta_i^{\text{abs}} = \text{Agg}(\Delta_{i,1}^{\text{abs}}, \dots, \Delta_{i,T}^{\text{abs}}), \ i = 1, \dots, n$

OT

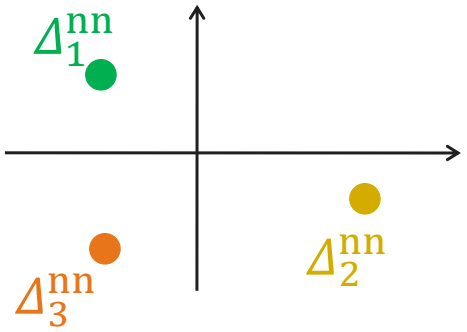
Neural: Dataset $\mathcal{D}$ & sites $\{s_j\}_{j=1}^{m}$



$y(x_t^b)$ vs $\hat{y}_j(x_t^b, x_t^s)$

$\Delta_{j,t}^{\text{nn}} = \hat{y}_j(x_t^b, x_t^s) - y(x_t^b)$

$\Delta_j^{\text{nn}} = \text{Agg}(\Delta_{j,1}^{\text{nn}}, \dots, \Delta_{j,T}^{\text{nn}}), \ j = 1, \dots, m$



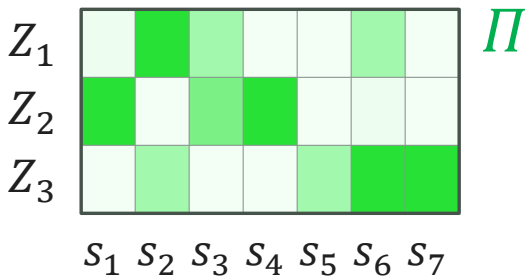
$$\nu = \frac{1}{m}\sum_{j=1}^{m} \delta_{\Delta_j^{\text{nn}}}$$

PLOT Algorithm (Single-Stage)

Data partition: $\mathcal{D} = \mathcal{D}_{\text{fit}} \cup \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$

OT Fitting

Solve $\text{OT}(\mu, \nu)$ on $\mathcal{D}_{\text{fit}}$



- Entropic OT: $\varepsilon > 0$
  Sinkhorn alg. $O(nm)$ per iter.
- Unbalanced OT: $\beta > 0$
  for irrelevant neural sites

Handle & Calibration

Calibrate row Π_i into a handle for Z_i

- Select top-K mass sites $\mathcal{S}_i^K$
- Restrict & renormalize: $\tilde{\Pi}_{ij}^K \propto \Pi_{ij} \mathbb{1}_{\{j \in \mathcal{S}_i^K\}}$

Intervention handle: $h^b = h(x^b)$, $h^s = h(x^s)$

$$\hat{h}_i = h^b + \lambda \sum_{j \in \mathcal{S}_i^K} \tilde{\Pi}_{ij}^K W_j (h^s - h^b)$$

Calibrate: $(\epsilon, \beta, K, \lambda)$ on $\mathcal{D}_{\text{cal}}$

$$(\epsilon_i^*, \beta_i^*, K_i^*, \lambda_i^*) \in \text{argmax}_{\epsilon,\beta,K,\lambda} \text{IIA}_{\text{cal}}(i;\ \epsilon, \beta, K, \lambda)$$

Testing

Test handle on $\mathcal{D}_{\text{test}}$

Per Z_i variable:

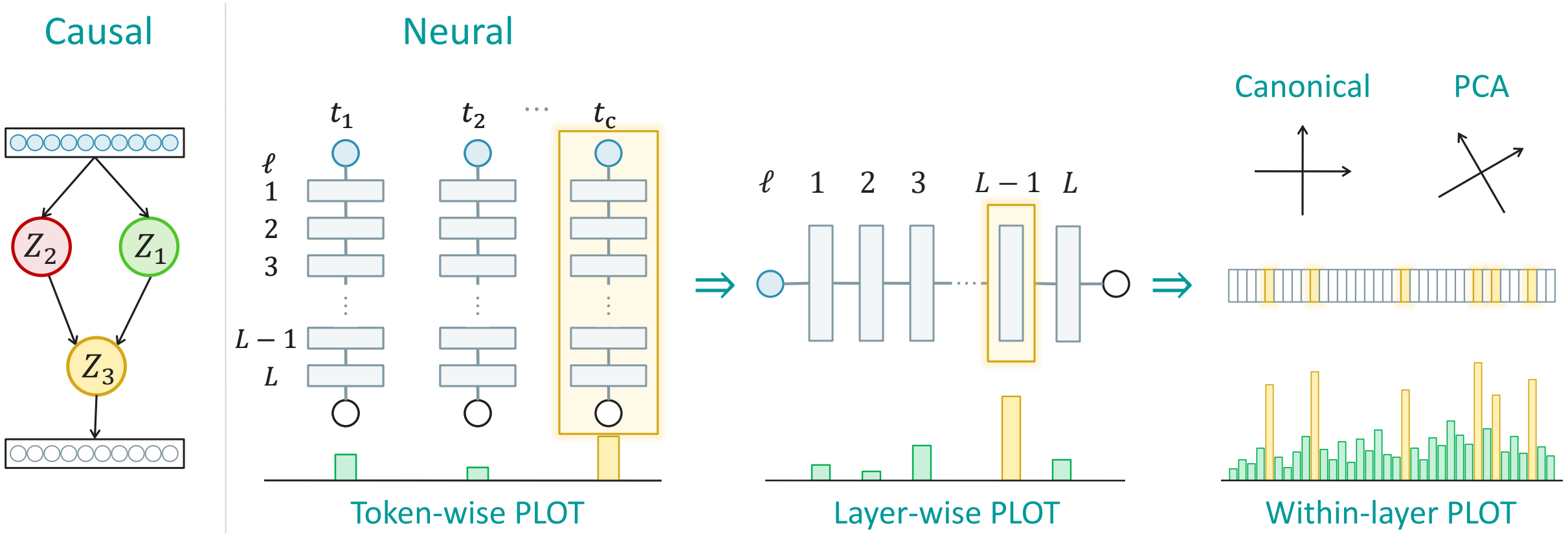
$$\text{IIA}_{\text{test}}(i;\ \epsilon_i^*, \beta_i^*, K_i^*, \lambda_i^*)$$

Average

$$\frac{1}{n} \sum_{\text{i=1}}^{n} \text{IIA}_{\text{test}}(i;\ \epsilon_i^*, \beta_i^*, K_i^*, \lambda_i^*)$$

PLOT Algorithm (Multi-Stage)

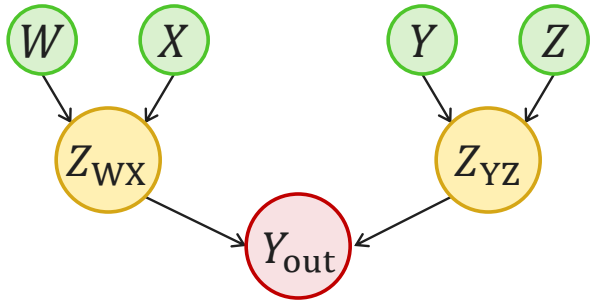
Progressive localization: Fixed causal model, progressively finer neural sites



PLOT-guided DAS: Run DAS within localized layer to save exhaustive search

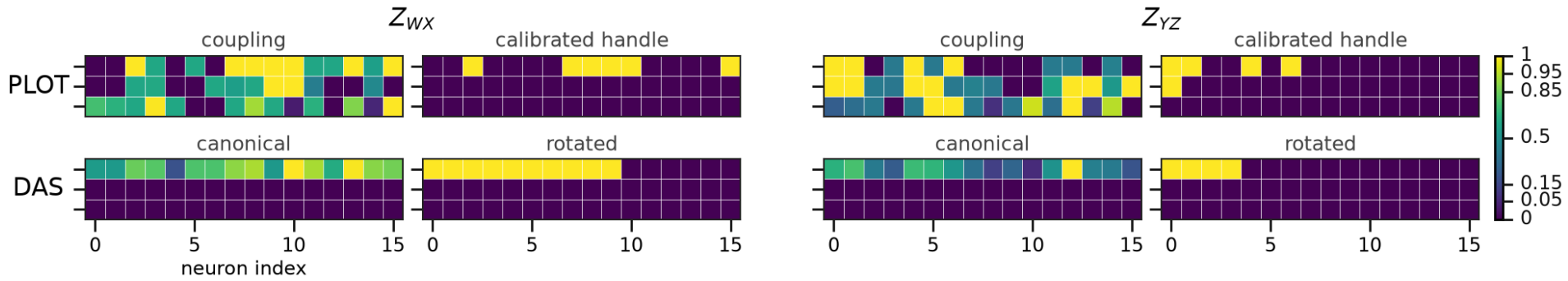
Hierarchical Equality (HEQ)

Task: $Z_{\mathrm{WX}} = \mathbb{1}_{\{W=X\}},\ Z_{\mathrm{YZ}} = \mathbb{1}_{\{\mathrm{Y}=Z\}},\ Y_{\mathrm{out}} = \mathbb{1}_{\{Z_{\mathrm{WX}} = Z_{\mathrm{YZ}}\}}$



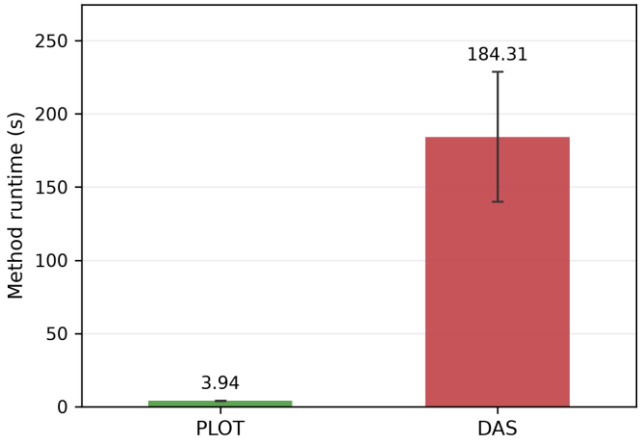
Neural net: $\mathrm{input} \rightarrow 16 \rightarrow 16 \rightarrow 16 \rightarrow 2$ **Sites:** $3 \times 16 = 48$ neurons

Handles:



Results:

Metric	PLOT	DAS
Z_{WX} sensitivity	0.9923 ± 0.0155	0.9938 ± 0.0041
Z_{WX} invariance	0.9988 ± 0.0026	0.9998 ± 0.0004
Z_{YZ} sensitivity	0.9913 ± 0.0190	0.9934 ± 0.0047
Z_{YZ} invariance	0.9901 ± 0.0209	0.9999 ± 0.0003
Mean sensitive IIA	0.9918 ± 0.0172	0.9936 ± 0.0032
Avg. IIA	0.9931 ± 0.0127	0.9967 ± 0.0016
Serial runtime (s)	3.94 ± 0.27	184.31 ± 44.41

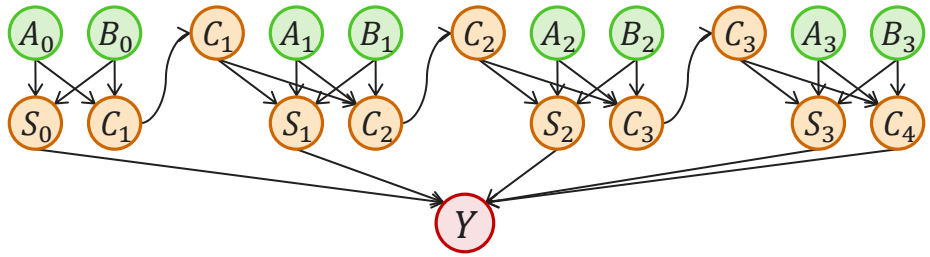


Binary Addition

Task: 4-bit binary addition

Neural net: GRUCell RNN $h_t = \text{GRUCell}([A_t, B_t]^\top, h_{t-1}) \in \mathbb{R}^{16}$

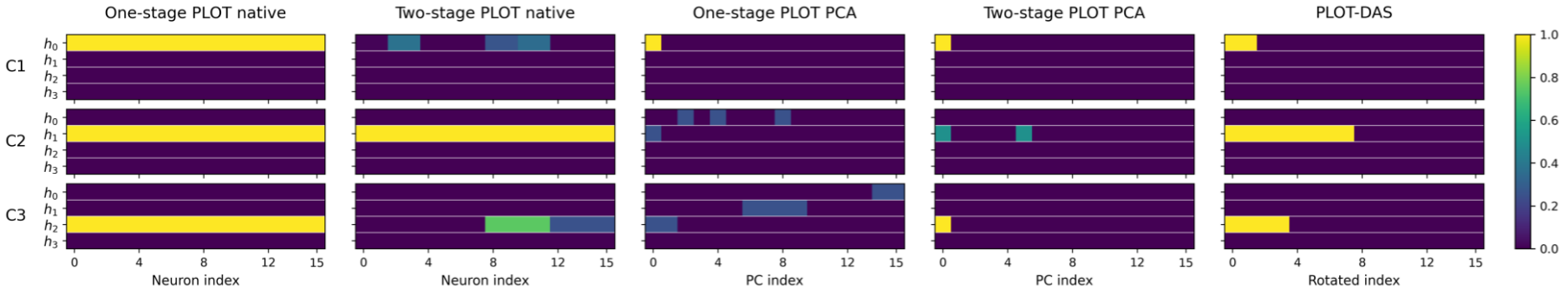
$t = 0, \ldots, 3$



$Y = (C_4, S_3, S_2, S_1, S_0)$

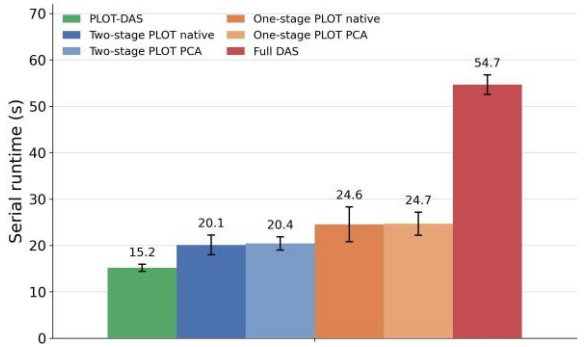
Two-stage PLOT: Recurrent state [full state] → within state [resolution $r = 2^k$]

Handles:



Results:

Method	Test IIA	Serial runtime (s)
1-stage PLOT native	0.767 ± 0.032	24.6 ± 3.7
1-stage PLOT PCA	0.900 ± 0.053	24.7 ± 2.5
2-stage PLOT native	0.792 ± 0.023	20.1 ± 2.1
2-stage PLOT PCA	$\mathbf{0.936 \pm 0.028}$	20.4 ± 1.4
PLOT-DAS	$\mathbf{0.991 \pm 0.007}$	$\mathbf{15.2 \pm 0.8}$
Full DAS	$\mathbf{0.991 \pm 0.007}$	$\mathbf{54.7 \pm 2.1}$

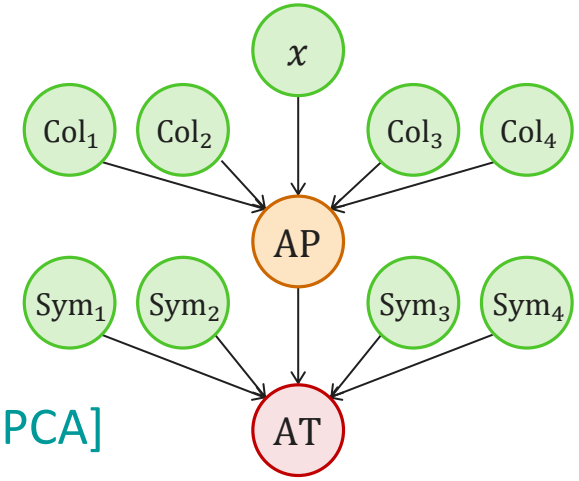


Multiple-Choice Question Answering (MCQA)

[**Mech Interp Benchmark**, Muller et al.'25]

Task: < The car is pink. What color is the car? A. red B. black C. pink D. yellow >

AP: answer position AT: answer token



Neural net: Gemma-2-2B, 26 layers, residual width 2304

Two-stage PLOT: Last token [fixed] → layer [full vector] → within-layer [canonical / PCA]

Results:

Method	AP	AT	Avg.	Runtime (min)	AP layer/dim	AT layer/dim
1-stage PLOT native	0.4290 ± 0.3381	0.7620 ± 0.2958	0.5955 ± 0.2779	112.85 ± 2.64	L5/576	L20/1152
1-stage PLOT PCA	0.7640 ± 0.1130	0.9120 ± 0.0561	0.8380 ± 0.0435	124.71 ± 3.09	L14,17,20,21,23/705	L20,21,24/846
2-stage PLOT native	0.7630 ± 0.1026	0.9030 ± 0.0231	0.8330 ± 0.0523	49.11 ± 3.93	L17/512	L24/1024
2-stage PLOT PCA	0.7640 ± 0.1099	$\mathbf{0.9580 \pm 0.0115}$	0.8610 ± 0.0526	67.50 ± 4.34	L17/141	L24/282
PLOT-DAS	0.8810 ± 0.0542	0.8690 ± 0.0500	0.8750 ± 0.0286	25.27 ± 1.60	L17/256	L24/128
Full DAS	$\mathbf{0.9220 \pm 0.0282}$	0.8700 ± 0.0296	$\mathbf{0.8960 \pm 0.0213}$	509.60 ± 23.88	L17/32	L20/512

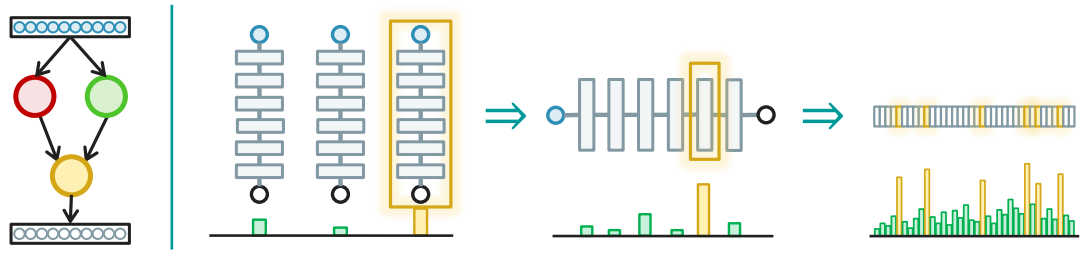
Stage-A OT vs UOT	OT: AP	OT: AT	OT: Avg.	UOT: AP	UOT: AT	UOT: Avg.
Calibration IIA (5 seeds)	0.000	0.974	0.487	**0.738**	0.974	**0.856**

Summary and Future Directions

Causal abstractions: Principled framework for mechanistic interpretability.

PLOT: Progressive localization via optimal transport

- Gradient-free matching
- Counterfactual effect signature geometry
- Direct handles, PLOT-guided-DAS
- Coarse-to-fine progressive localization for comp. gains

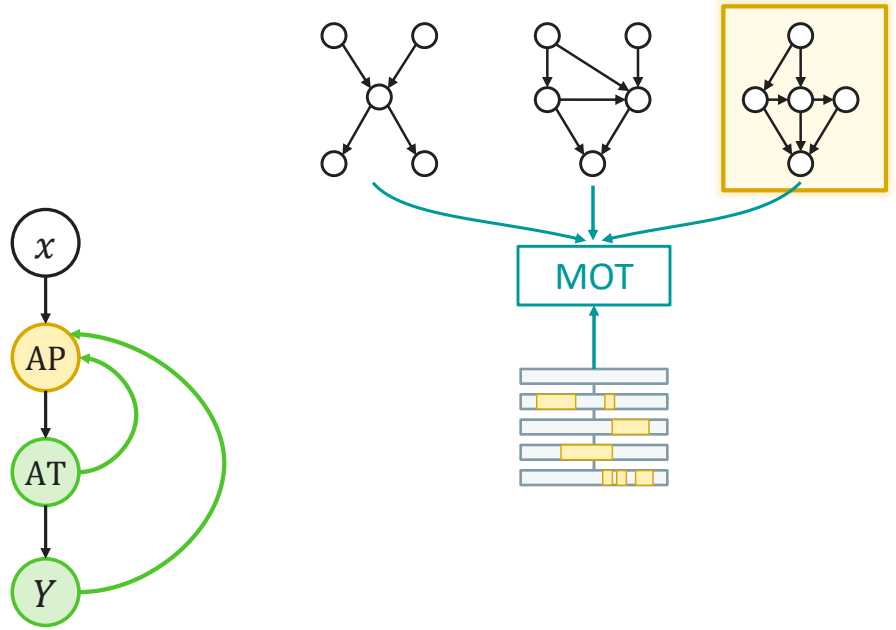


Summary and Future Directions

Causal abstractions: Principled framework for mechanistic interpretability.

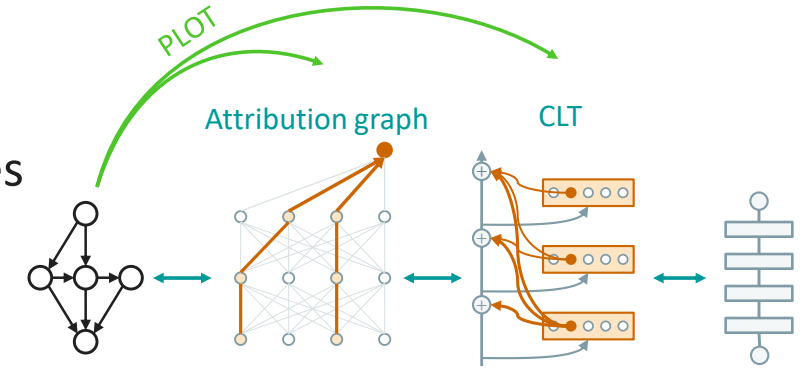
PLOT: Progressive localization via optimal transport

- Gradient-free matching
- Counterfactual effect signature geometry
- Direct handles, PLOT-guided-DAS
- Coarse-to-fine progressive localization for comp. gains



Future Directions

- How to choose the causal model?
  Multimarginal OT to compare competing causal models
- Internal faithfulness? [Méloux et al.'26]
  Gradual discovery: use learned handles to localize upstream handles
- Bridging top-down & bottom-up approaches?
  PLOT on interpretable features / attribution graphs [Ameisen et al.'25]



Chang, Datla, Goldfeld, "PLOT: Progressive Localization via Optimal Transport in Neural Causal Abstraction", arxiv:2605.06979

Variants of Optimal Transport

Entropic OT: For $\varepsilon > 0$

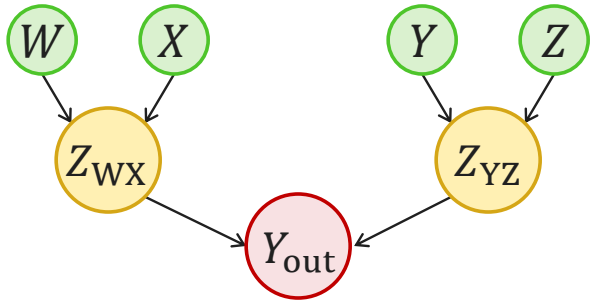
$$\mathrm{OT}_c^{\varepsilon}(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int c \, \mathrm{d}\pi + \varepsilon \mathrm{KL}(\pi \parallel \mu \otimes \nu)$$

Unbalanced OT: For $\beta_1, \beta_2 > 0$:

$$\mathrm{OT}_c^{\beta_1,\beta_2}(\mu, \nu) := \inf_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} \int c \, \mathrm{d}\pi + \beta_1 \mathrm{KL}(\pi_1 \parallel \mu) + \beta_2 \mathrm{KL}(\pi_2 \parallel \nu)$$

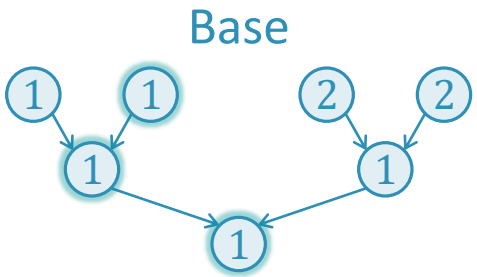
Hierarchical Equality (HEQ)

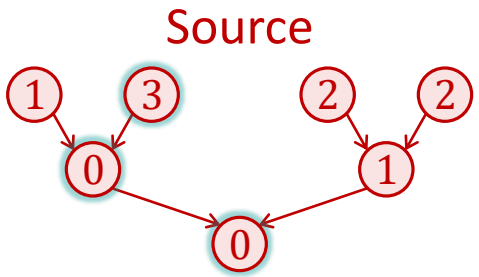
Task: $Z_{\mathrm{WX}} = \mathbb{1}_{\{W=X\}},\ Z_{\mathrm{YZ}} = \mathbb{1}_{\{\mathrm{Y}=Z\}},\ Y_{\mathrm{out}} = \mathbb{1}_{\{Z_{\mathrm{WX}} = Z_{\mathrm{YZ}}\}}$



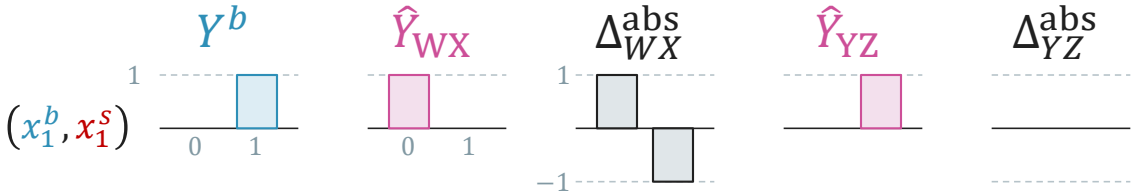
Neural net: $\mathrm{input} \to 16 \to 16 \to 16 \to 2$ **Sites:** $3 \times 16 = 48$ neurons

Z_{WX} sensitivity



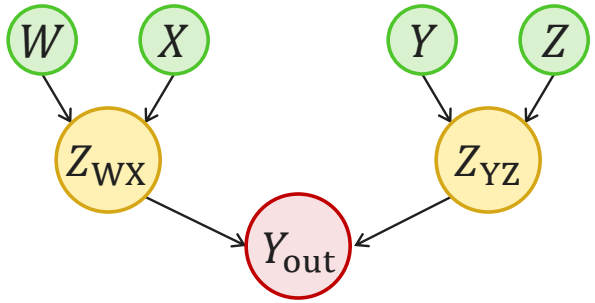


Causal



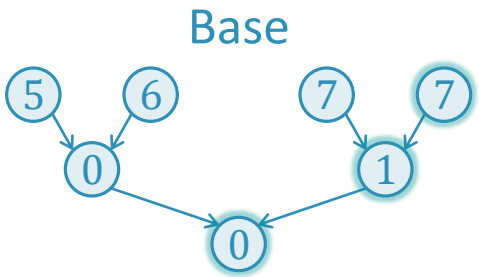
Hierarchical Equality (HEQ)

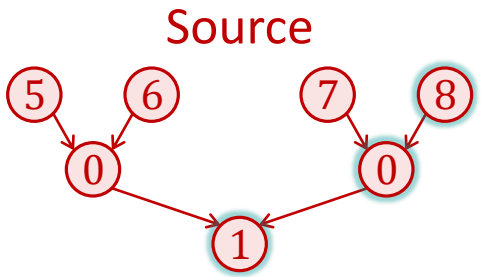
Task: $Z_{\mathrm{WX}} = \mathbb{1}_{\{W=X\}},\ Z_{\mathrm{YZ}} = \mathbb{1}_{\{\mathrm{Y}=Z\}},\ Y_{\mathrm{out}} = \mathbb{1}_{\{Z_{\mathrm{WX}} = Z_{\mathrm{YZ}}\}}$



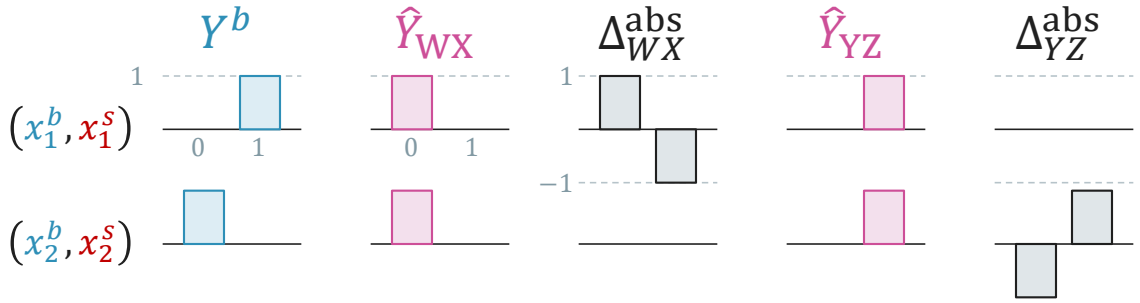
Neural net: $\mathrm{input} \to 16 \to 16 \to 16 \to 2$ **Sites:** $3 \times 16 = 48$ neurons

Z_{YZ} sensitivity



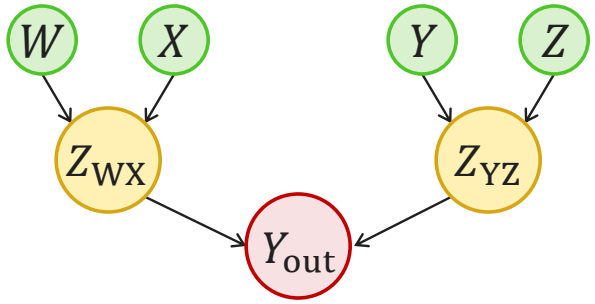


Causal



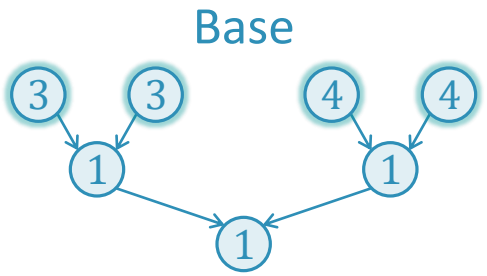
Hierarchical Equality (HEQ)

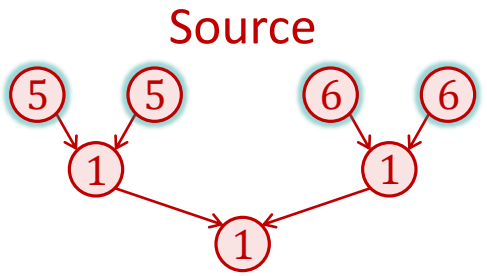
Task: $Z_{\mathrm{WX}} = \mathbb{1}_{\{W=X\}},\ Z_{\mathrm{YZ}} = \mathbb{1}_{\{Y=Z\}},\ Y_{\mathrm{out}} = \mathbb{1}_{\{Z_{\mathrm{WX}} = Z_{\mathrm{YZ}}\}}$



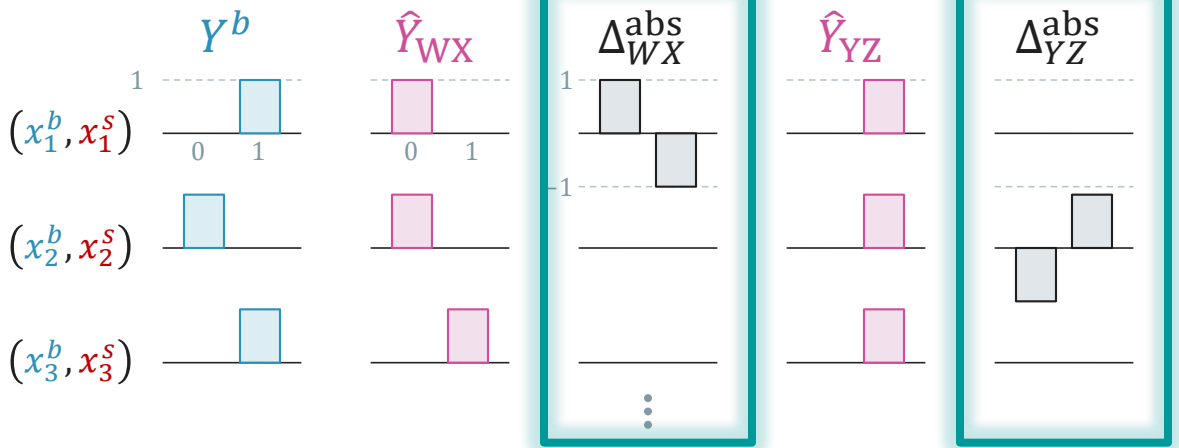
Neural net: $\mathrm{input} \to 16 \to 16 \to 16 \to 2$ **Sites:** $3 \times 16 = 48$ neurons

Invariance





Causal



Neural

