On the Relationship Between Explanation & Prediction: A Causal View

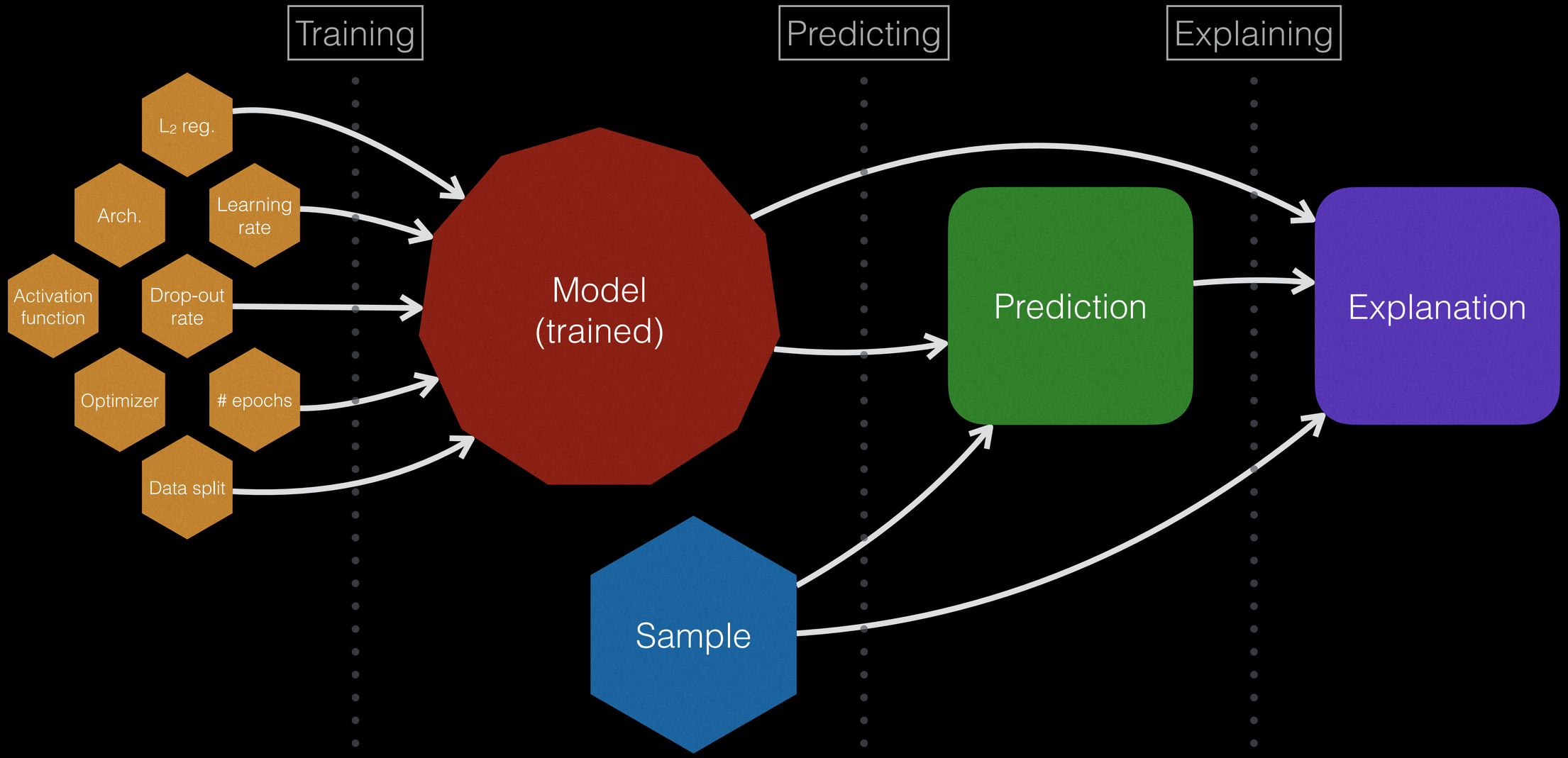
Amir-Hossein **Karimi**

in collaboration w/
Krikamol **Muandet**, Simon **Kornblith**, Bernhard **Schölkopf**, Been **Kim**

July, 2023

ETH zürich

Motivation: Explanation Generation Process



Treatment Effects

Treatment Effect

the **effect** that a **treatment** (i.e., the indep. var.) has upon the **response** variable (i.e., the dep. var.) in a study.

$$Y_{T=1}(X) - Y_{T=0}(X)$$

Individual Treatment Effect (ITE)

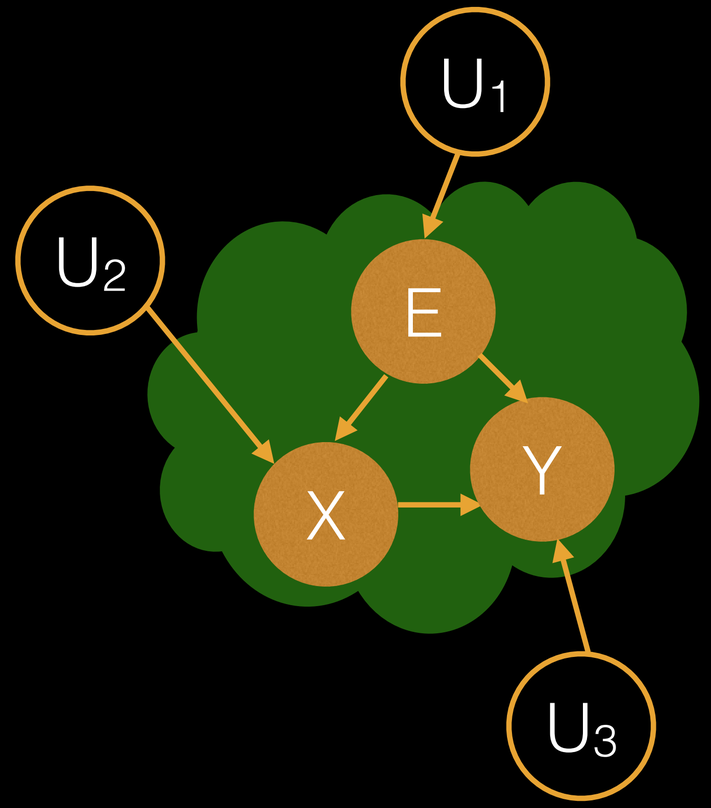
The treatment effect on one individual; **impractical.**

$$Y_{T=1}(X=x) - Y_{T=0}(X=x)$$

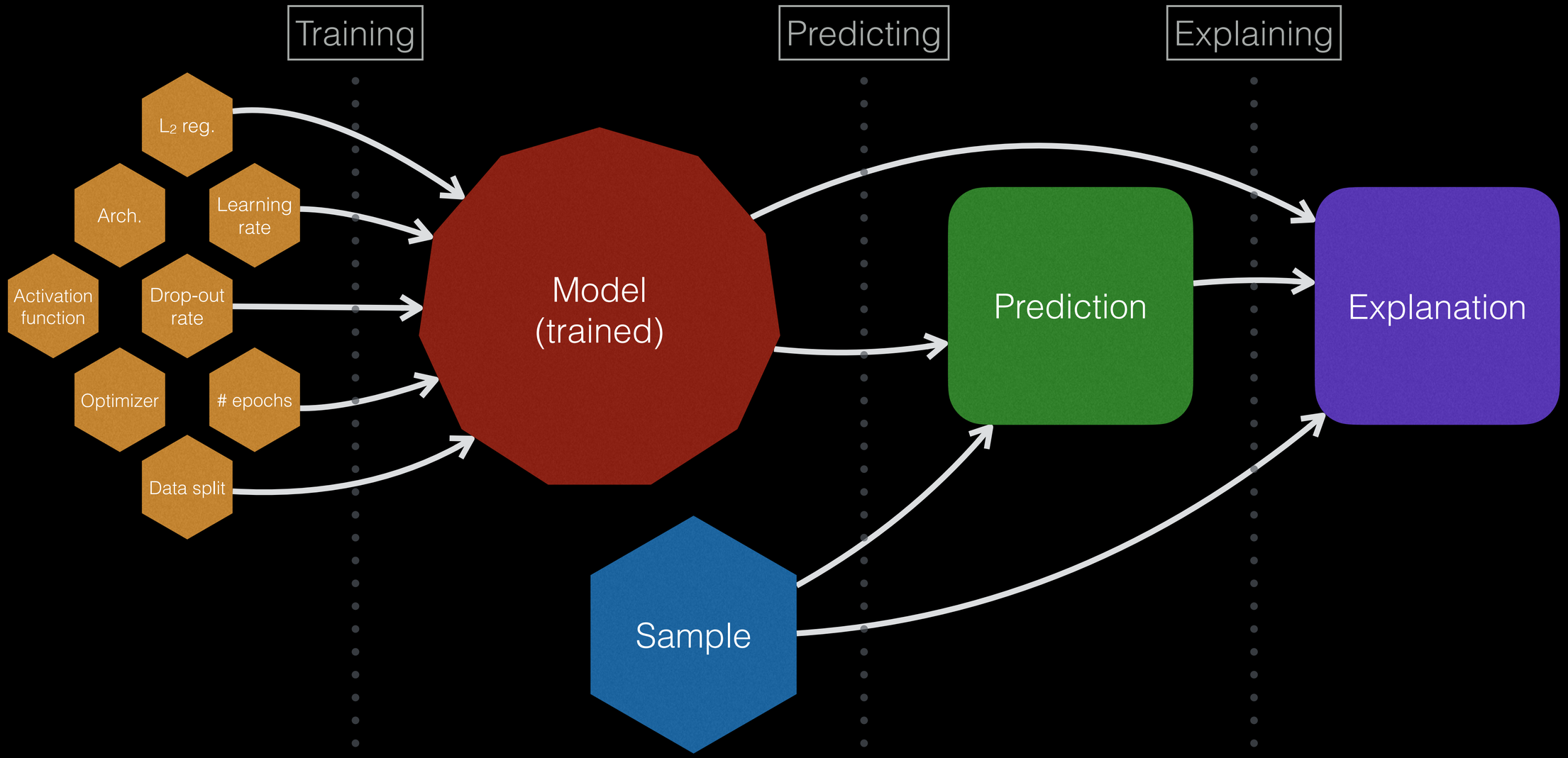
Average Treatment Effect (ATE)

The treatment effect on individuals within a population.

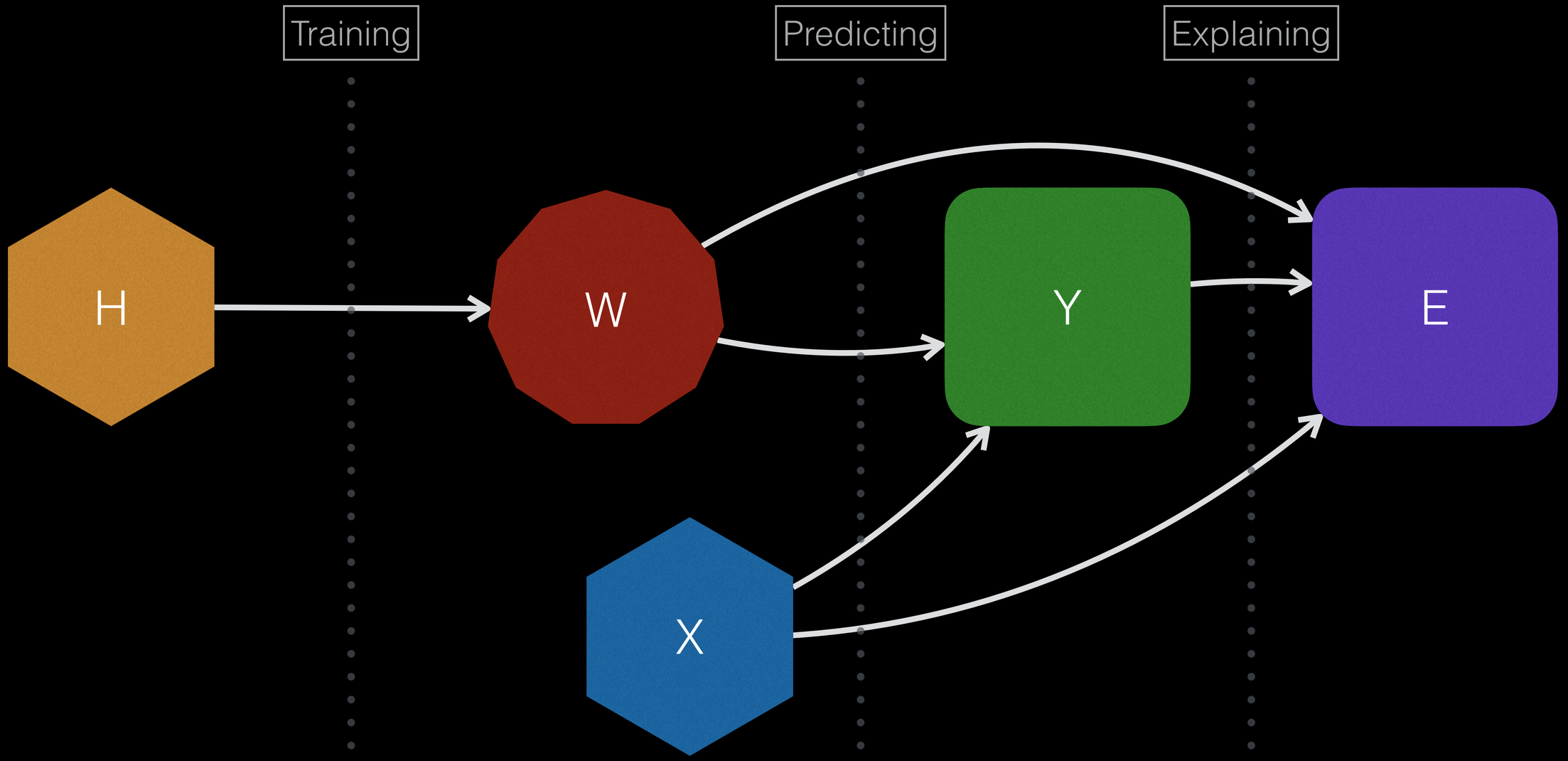
$$\mathbb{E}_{x' \sim P(X)} [\, Y_{T=1}(X=x') - Y_{T=0}(X=x') \,]$$



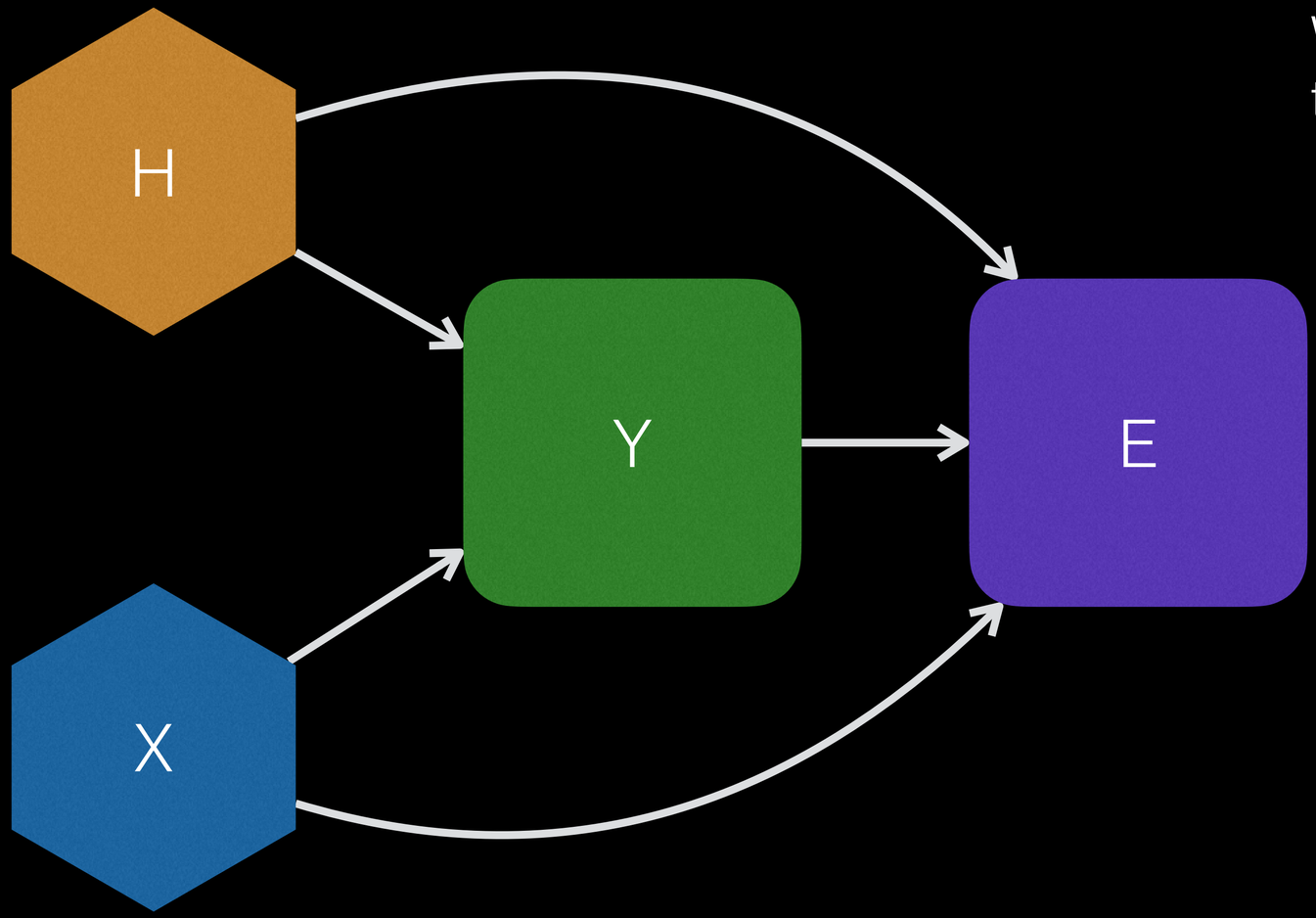
Explanation Generation Process



Explanation Generation Process

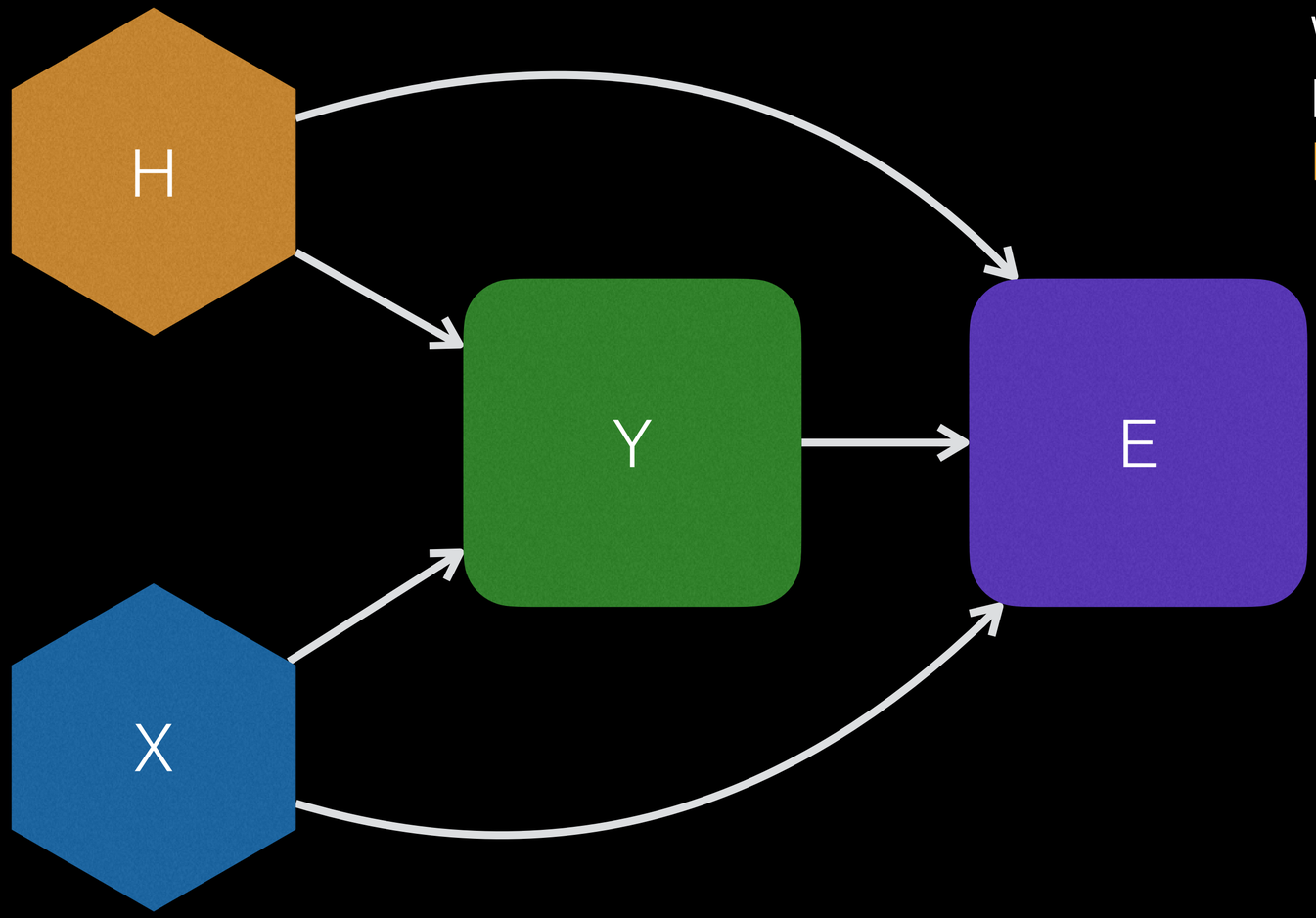


Hyperparameters as Treatments



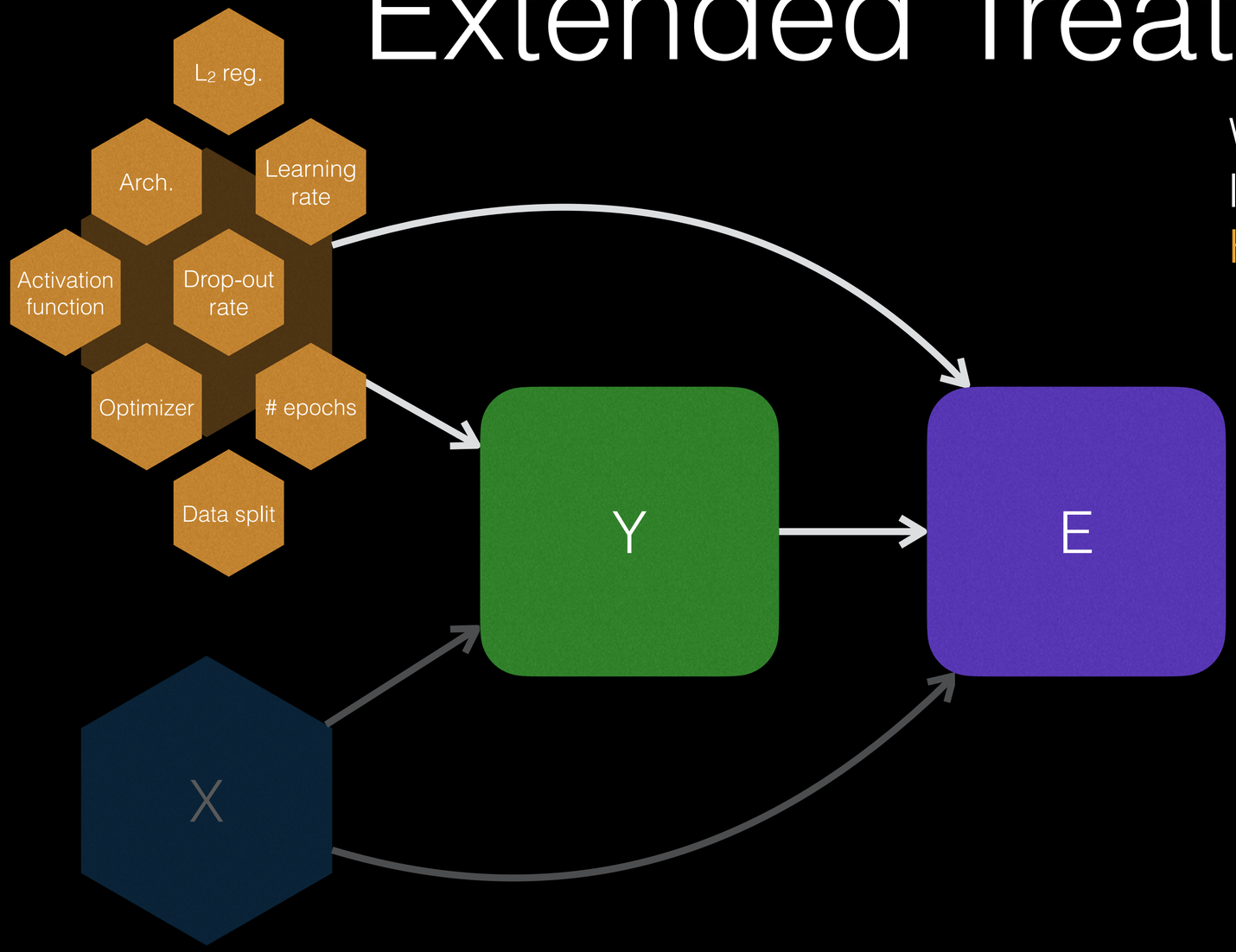
What is the effect of the hyperparameters on the resulting prediction/explanation?

Hyperparameters as Treatments



What does the prediction/explanation for X = x look like, if the hyperparameters take on value H = h rather than H = h', *all else being equal*?

Extended Treatment Effects



What does the prediction/explanation for X = x look like, if the hyperparameters take on value H = h rather than H = h', *all else being equal*?

$Y_{h=1} - Y_{h=0}$
single binary treatment

$\mathbb{E}_{m \neq n} [Y_{h=n} - Y_{h=m}]$
single non-binary treatment

$\mathbb{E}_{h \setminus i} [\mathbb{E}_{m \neq n} [Y_{hi=n,\, h \setminus i} - Y_{hi=m,\, h \setminus i}]]$
multiple non-binary treatment

$\mathbb{E}_{h \setminus i} [\mathbb{E}_{m \neq n} [\| \varphi(Y_{hi=n,\, h \setminus i}) - \varphi(Y_{hi=m,\, h \setminus i}) \|_G]]$
multiple non-binary treatments
& a non-binary target

Natural vs. simulation-based potential outcomes

i	$Y_{h=0}$	$Y_{h=1}$	$Y_{h=2}$
1	a	-	-
2	-	f	-
3	-	-	k
4	-	h	-
...	...	...	...

i	$Y_{h=0}$	$Y_{h=1}$	$Y_{h=2}$
1	a	e	-
2	b	f	-
3	c	g	-
4	d	h	-
...	...	...	...

Model Zoo & Explanations

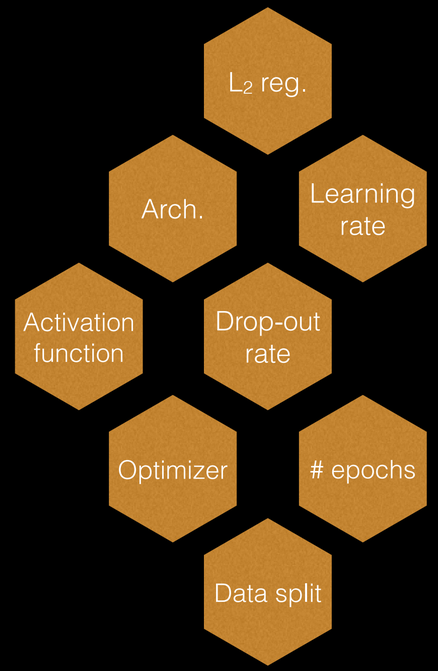
30,000 pre-tained models:
3 layer CNNs (4,970 parameters);
trained to convergence (max 86 epochs)

4 datasets:
MNIST, FASHION, SVHN, CIFAR10

8 hyparparameters:
drawn “independently at random” from pre-specified ranges

Fixed architecture. Fixed random seed.

4+1 saliency-based explanations:
Gradient, SmoothGrad, Integrated Gradients, Grad-CAM
Reference explanation: “identity”, i.e., E = Y —> ITE_E = ITE_Y



[Unterthiner et al. 2020]

Most types of H influence Y (and E) in a similar way

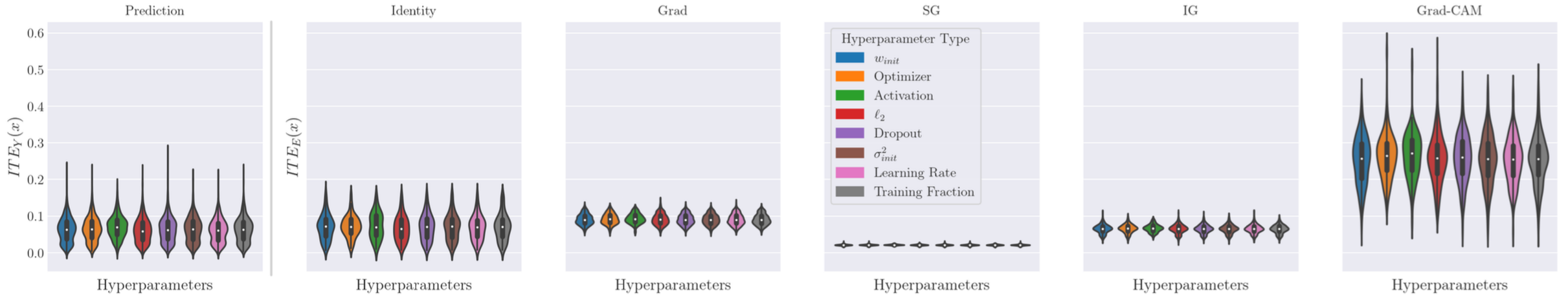


Figure 3: Comparison of ITE_Y and ITE_E for CIFAR10 shows that different types of H influence E and Y in a similar way.

H influences Y (and E) differently across performance buckets

Performance buckets:

- 0 - 20 pctl.
- 20 - 40 pctl.
- 40 - 60 pctl.
- 60 - 80 pctl.
- 80 - 90 pctl.
- 90 - 95 pctl.
- 95 - 99 pctl.
- 99 - 100 pctl.

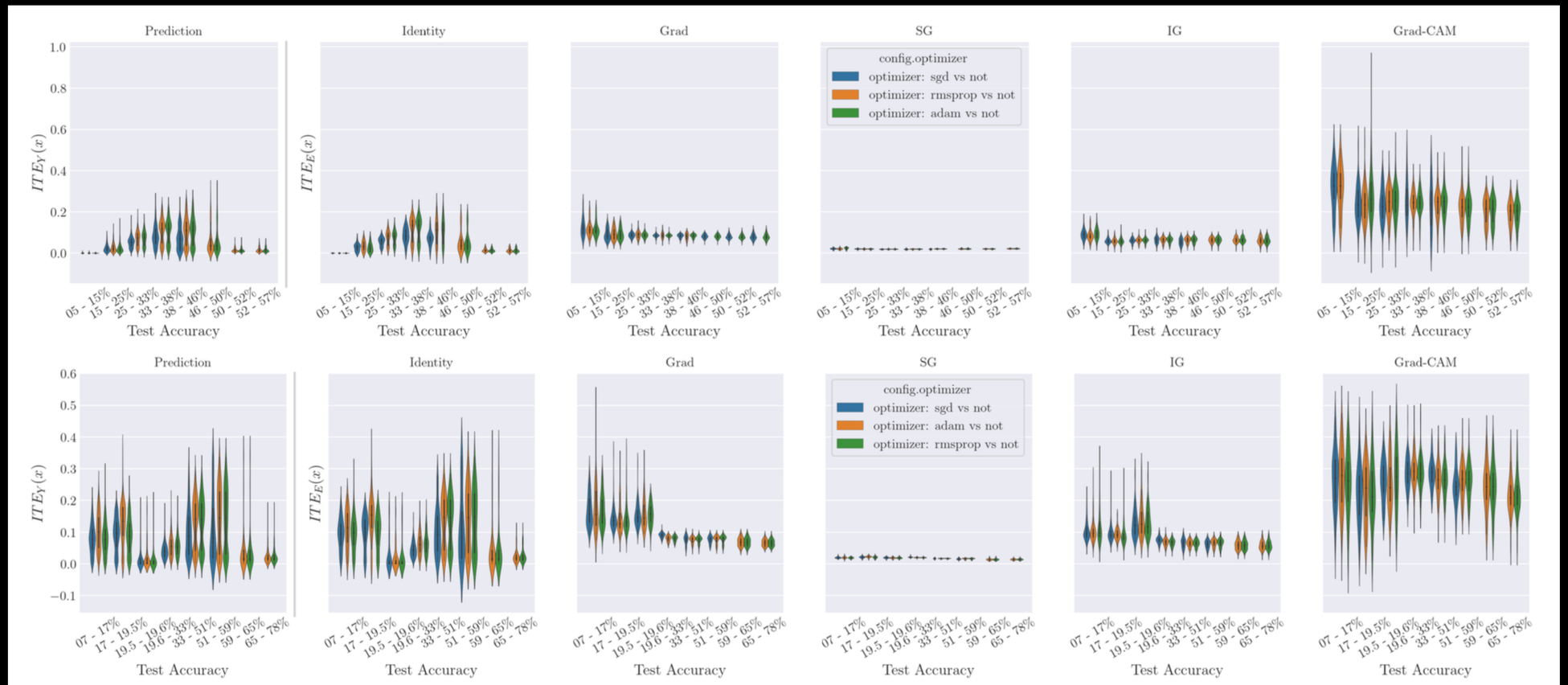


Figure 4: Comparison of ITE values of $h_{\text{optimizer}}$ on Y (left) and E (right) for models across different performance buckets, showing the discrepancy in the effect of H on Y vs. that on E (top: CIFAR10; bottom: SVHN). Interestingly, there is a difference of ITE_E across accuracy buckets, and more importantly, none of the explainability methods resemble ITE_Y.

Explanations may still be explaining something other than the prediction

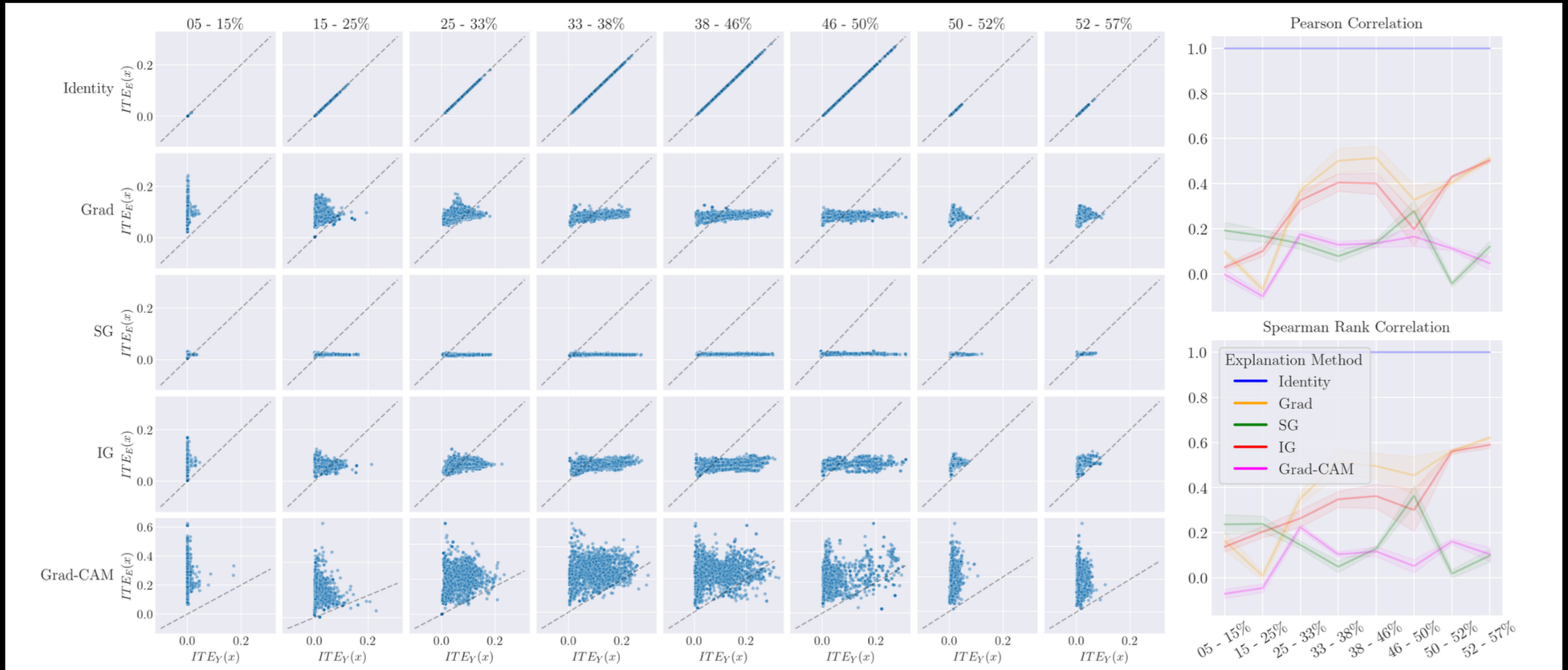
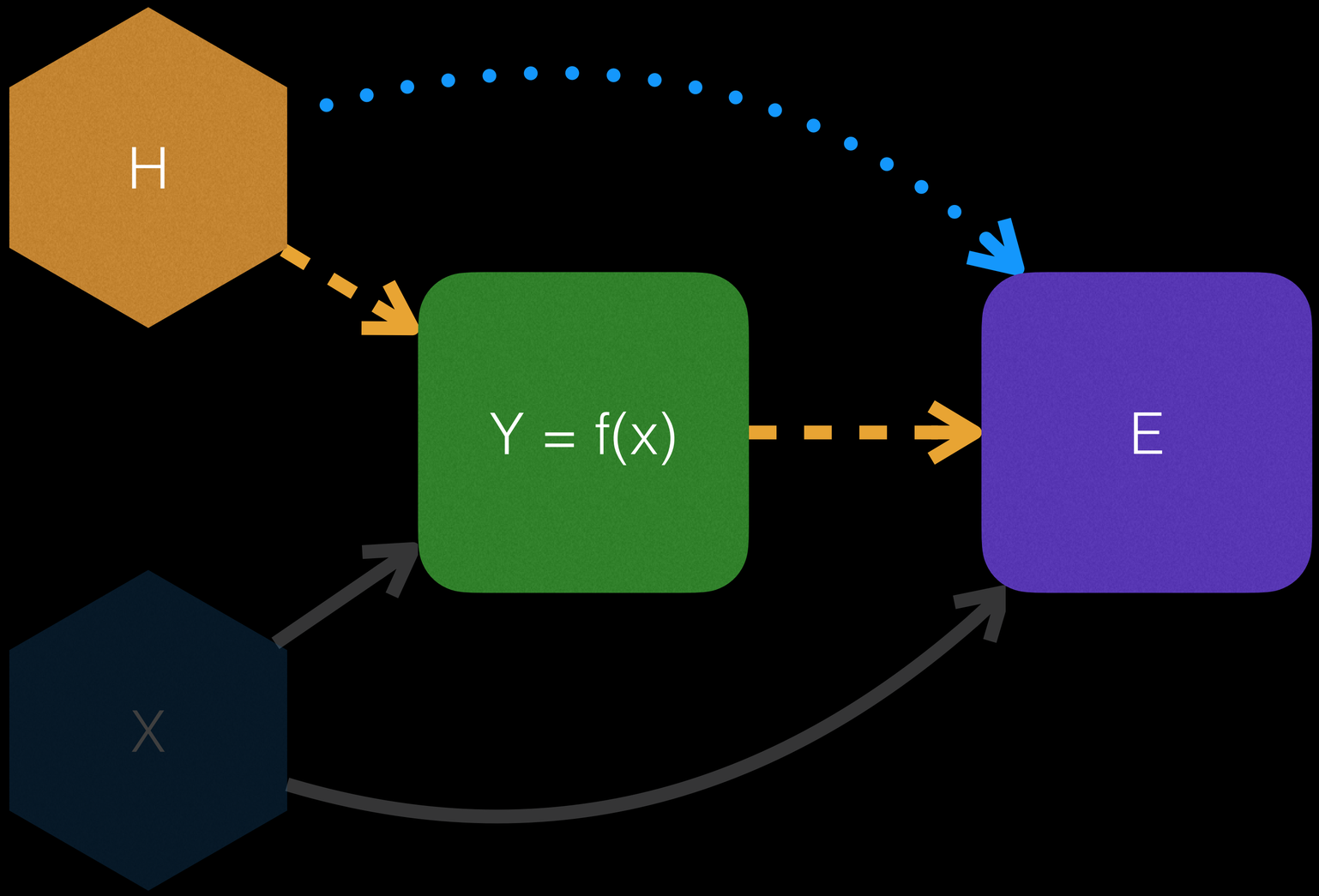


Figure 5: (left) Each column is a subset of models at each accuracy bucket, each row is a different explanation method. Whereas low-performing CIFAR10 models (first column) show little change in predictions as their explanations differ, top-performing models show the reverse of this trend. (right) Correlation measures of the scatter plots on the left show a decreased correlation in the top 1% models.

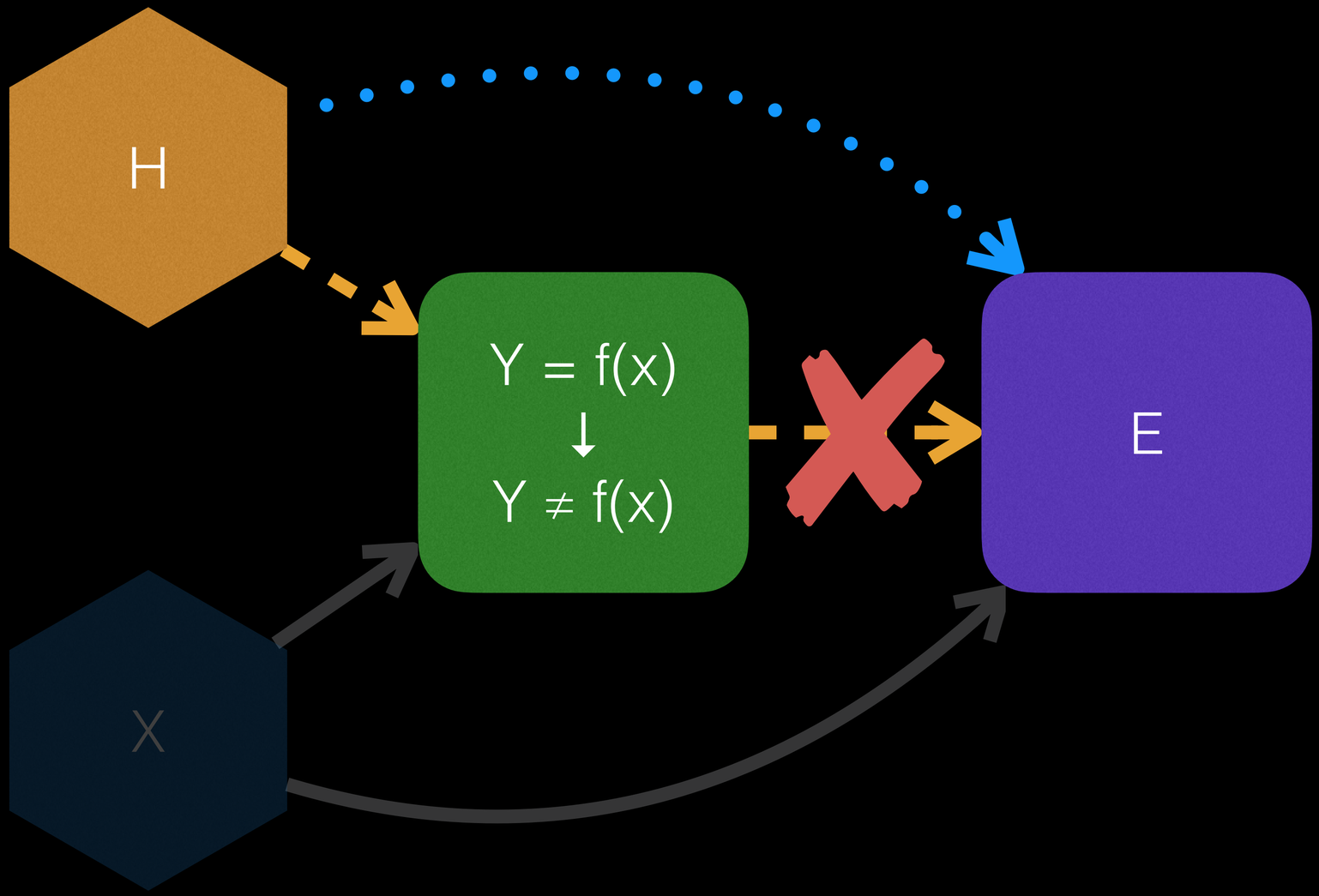
Direct vs indirect effects



$\mathbf{ITE}_{\mathbf{E}}$ measures the **total** effect:
* **direct effect**
* **indirect effect**

How to tease them apart?

Direct vs indirect effects



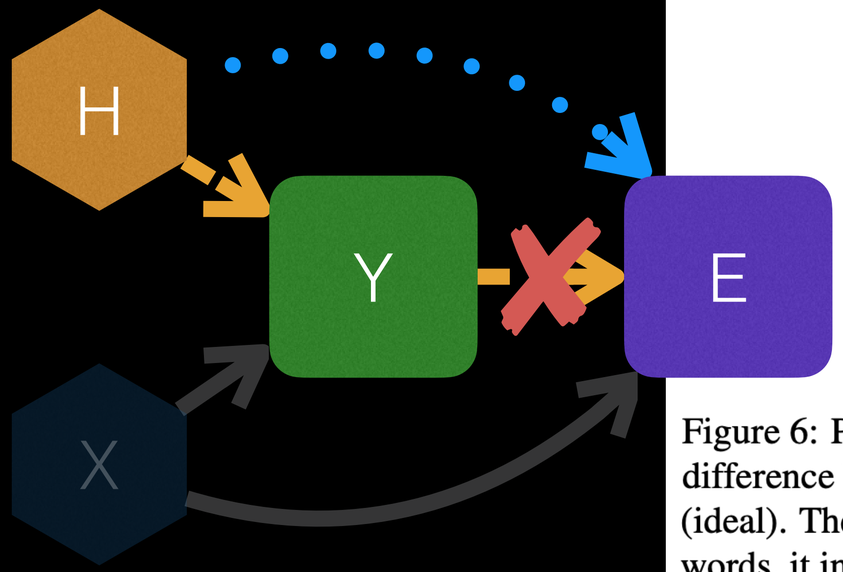
ITE_E measures the **total** effect:
* **direct effect**
* **indirect effect**

How to tease them apart?

We can sever the flow of dependence from H to E by randomizing Y

* **total effect**: ITE_E, $y=f(x)$
* **direct effect**: ITE_E, $y \neq f(x)$
* **indirect effect**: Δ above

Explanations from the highest performing models may be comparatively less reliable



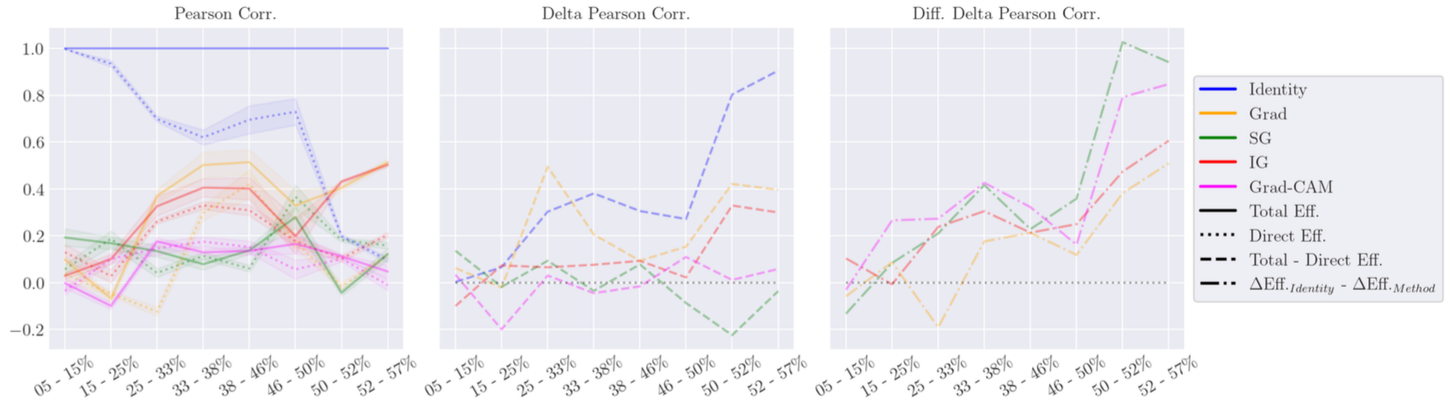
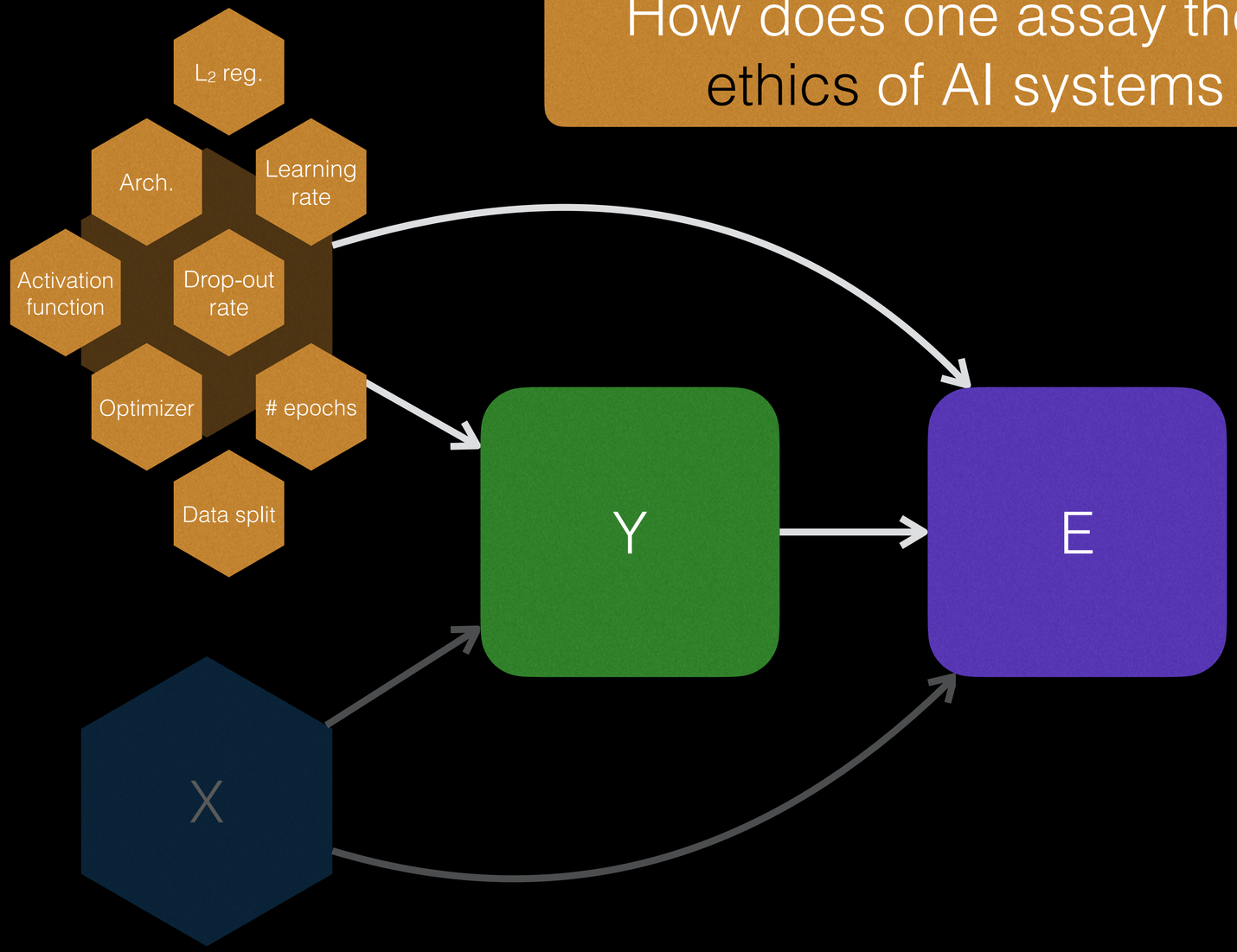


Figure 6: Pearson correlation between ITE_Y and ITE_E in total and direct effect (first column). The second column is the difference between total and direct effect, where higher values mean that the influence of H on E flows more through Y (ideal). The third column plots the difference of delta correlations between ideal case (Identity) and each method. In other words, it indicates how far each method moves away from ideal case, as a model performs better.

How does one assay the safety, factuality, and ethics of AI systems to foster trust in AI?



Common answer: use explanations

Preliminary work:

- Cautionary tale: explanations may still be explaining something other than the prediction
- We propose a causally-grounded quantitative metric to study the relationship between predication and explanation

Future work:

- Extension beyond saliency map
  e.g., SHAP, LIME, recourse, etc.
- Creating a OSS tool to measure causal effect of Y on E for any given black-box model

Thank you!