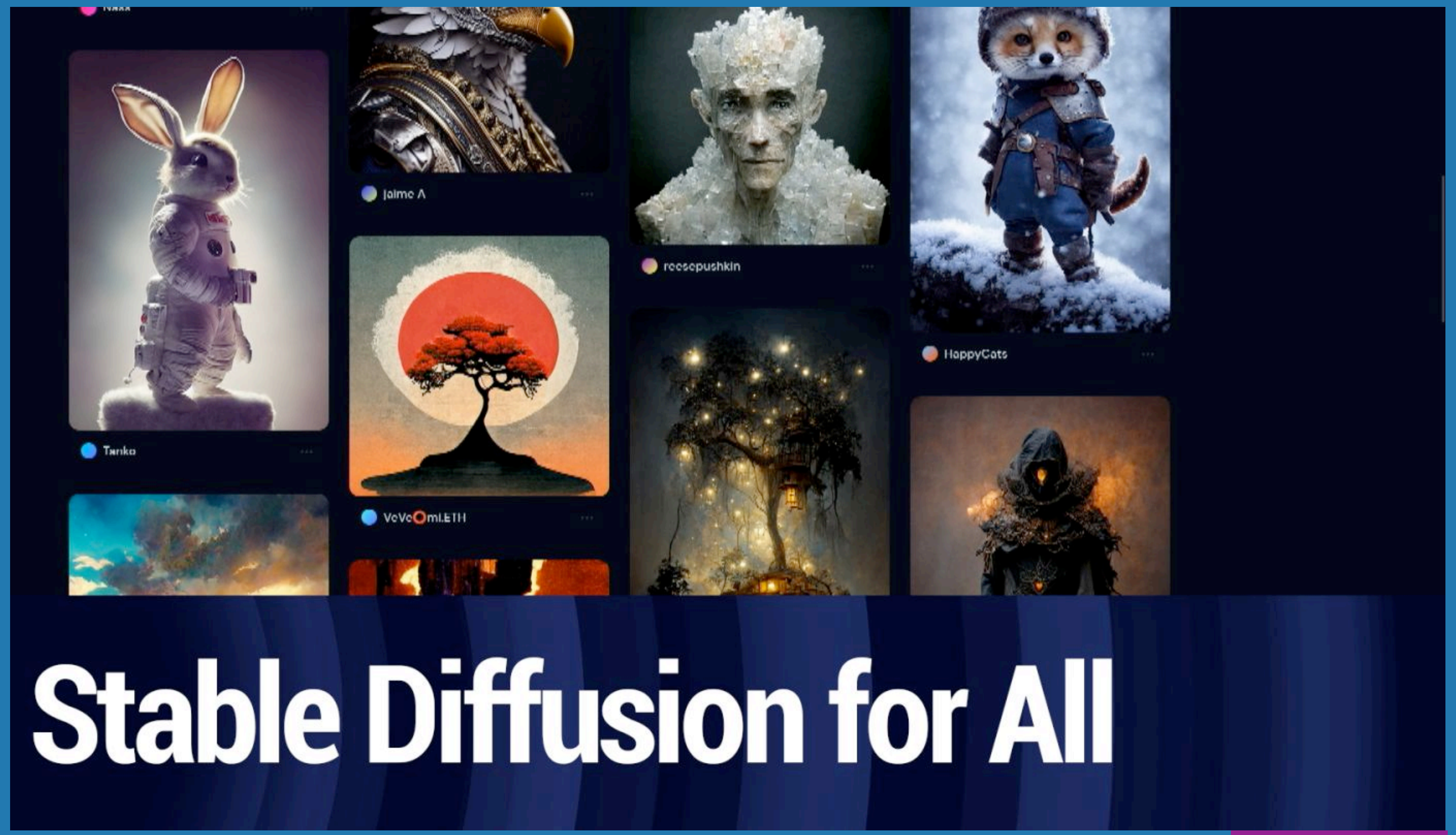
SEPTEMBER 8, 2022

Towards Data-Efficient, Grounded and Safe Deep Models

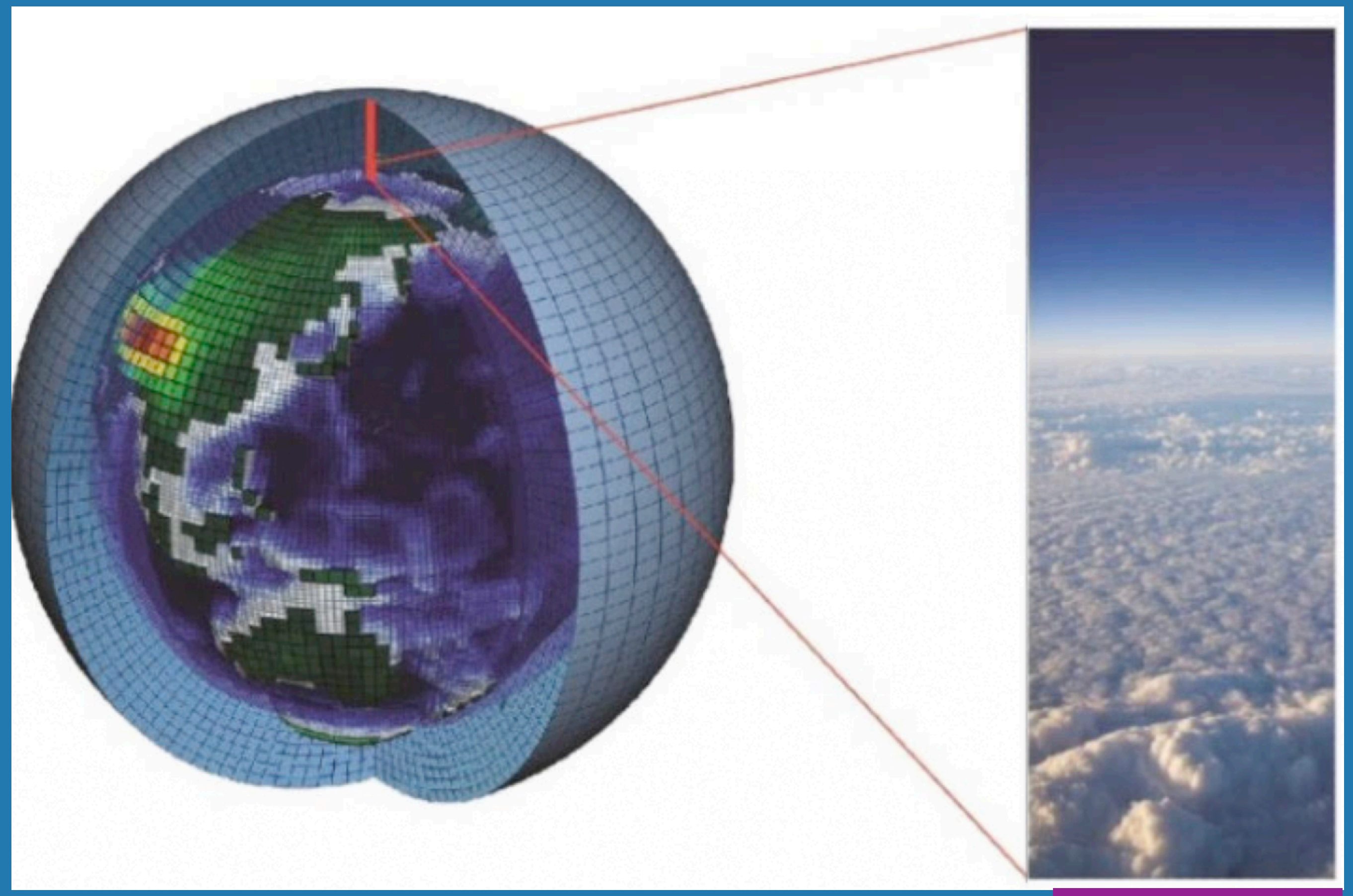
Jay Thiagarajan

Machine Intelligence Group
Lawrence Livermore National Labs

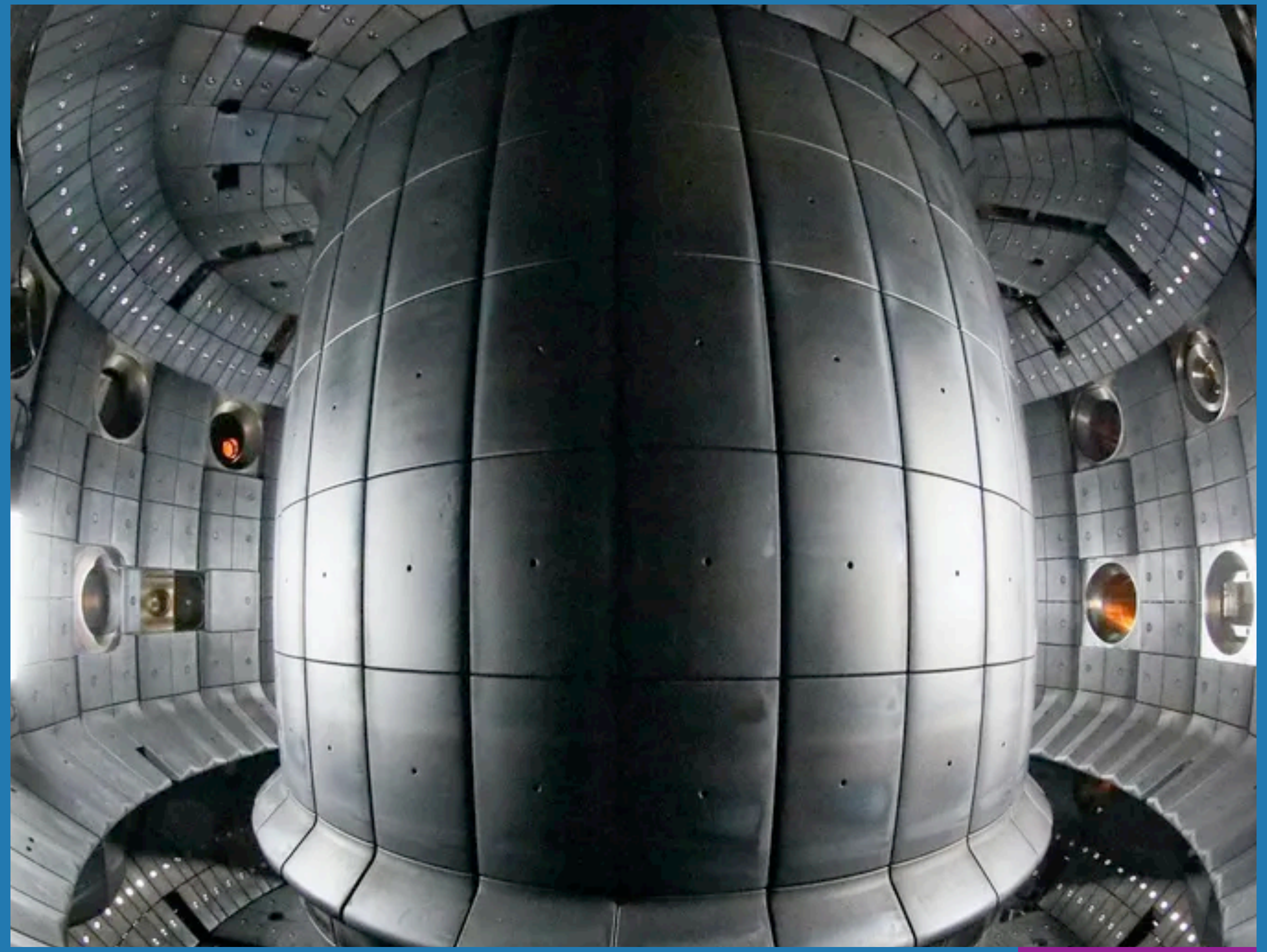
Strategize



Create

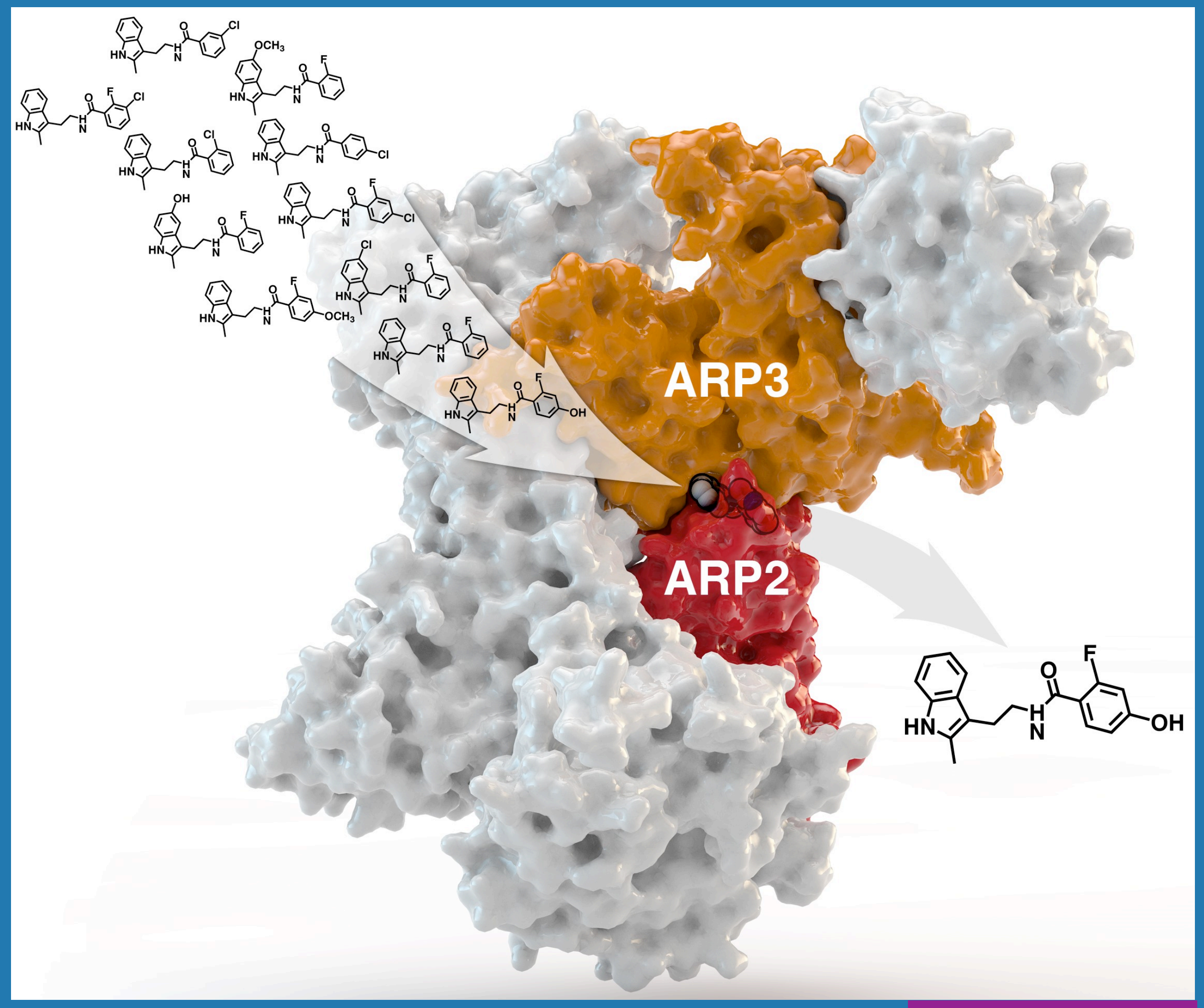


Forecast



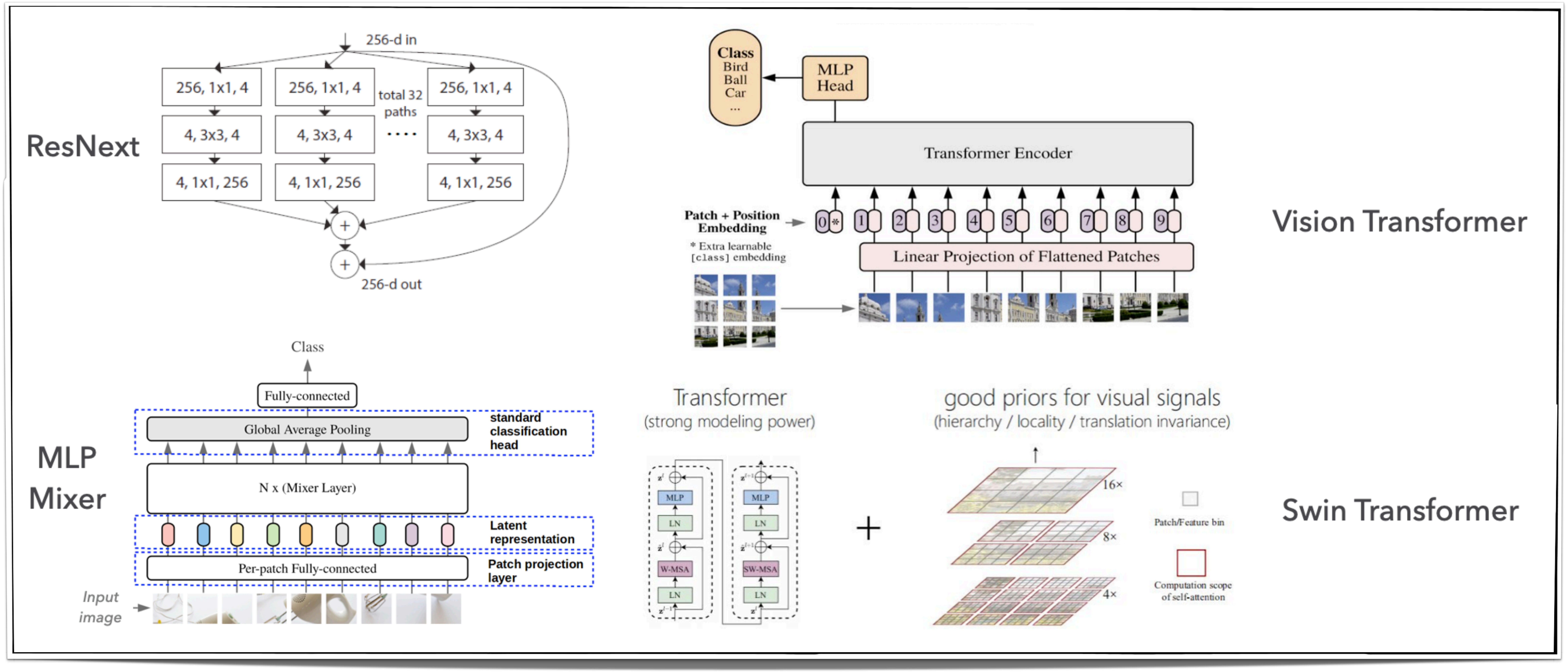
Control

Personalize

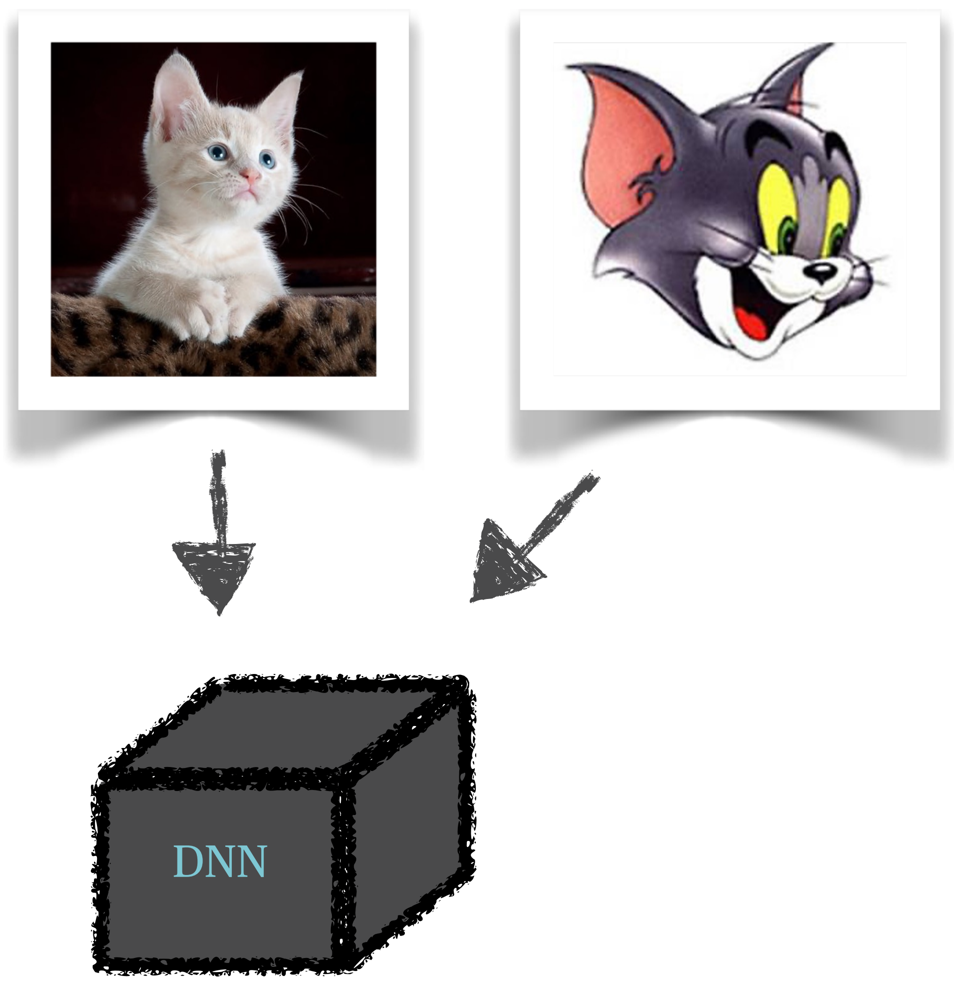


Discover

Advanced Architectures and Optimization Strategies Have Pushed the SOTA in Vision Tasks

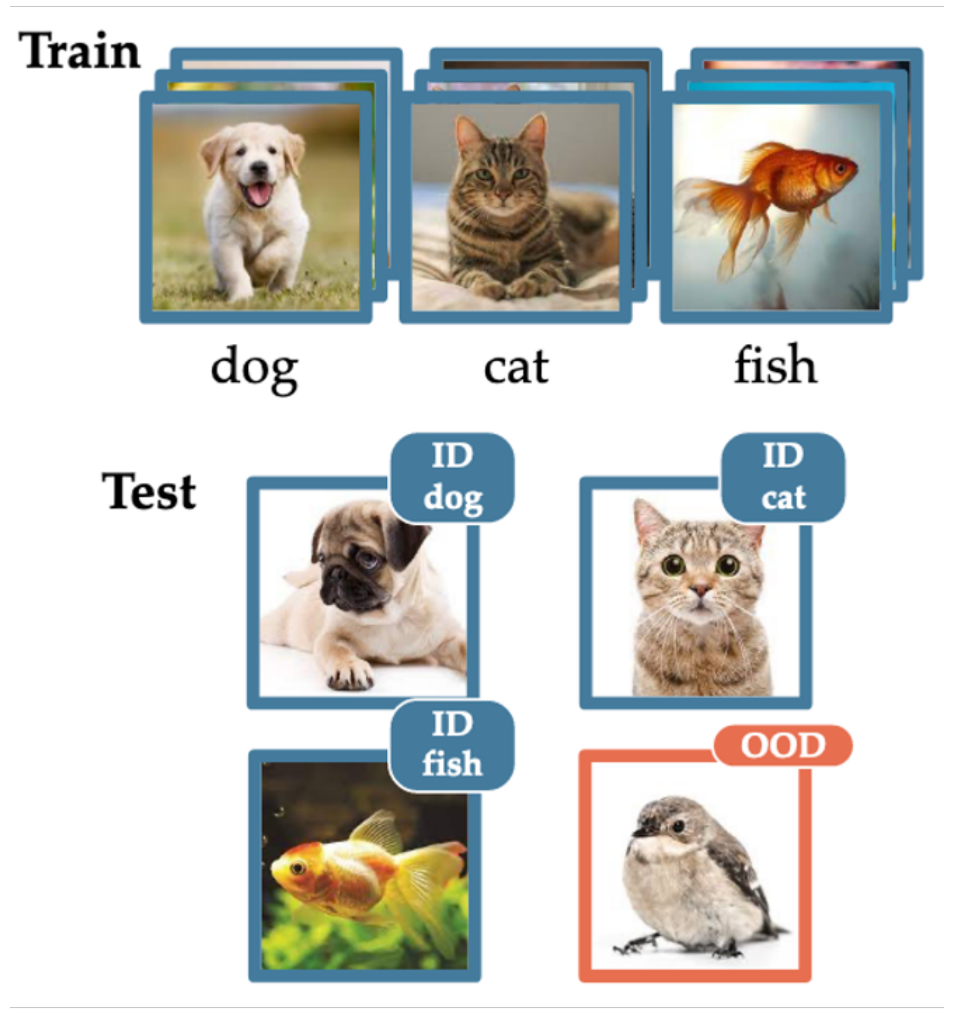


What Maketh a “Good” Model?

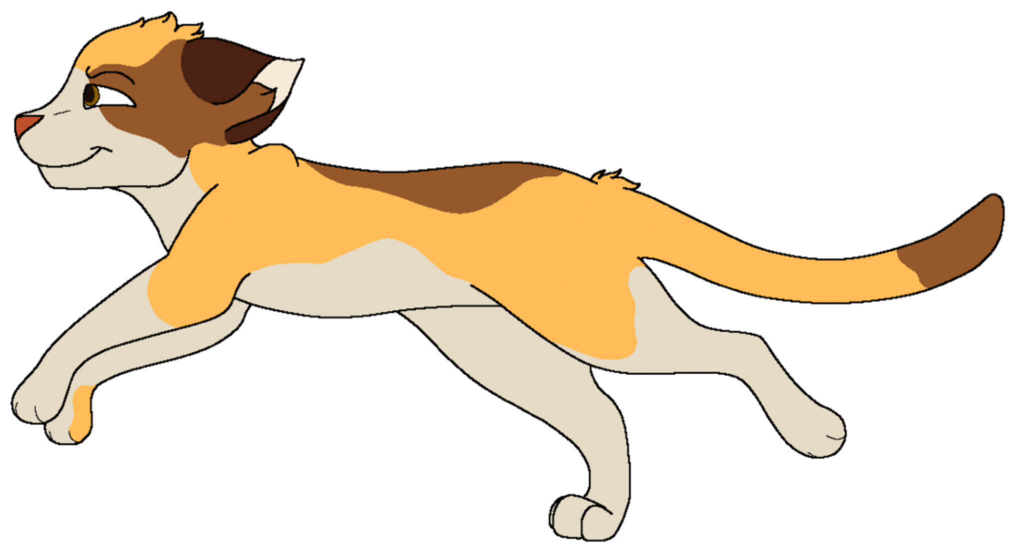


Withstand Distribution Shifts

Resilient to Malicious Manipulations

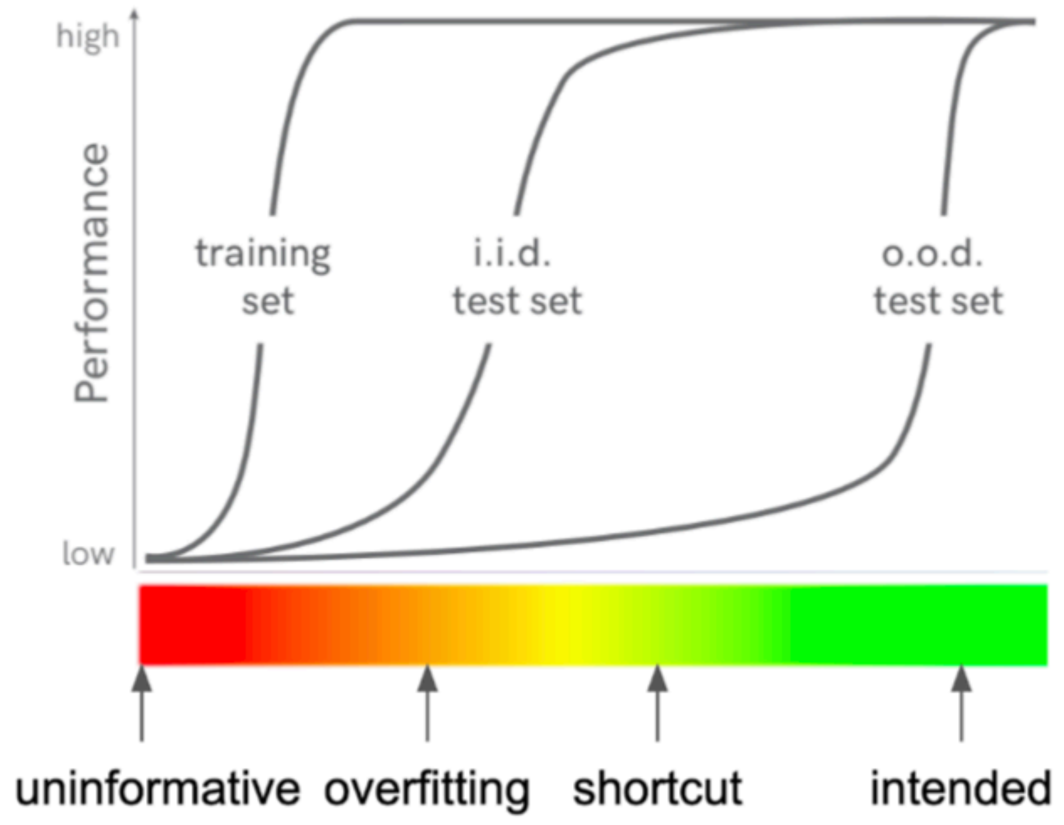


Reject Anomalous Samples

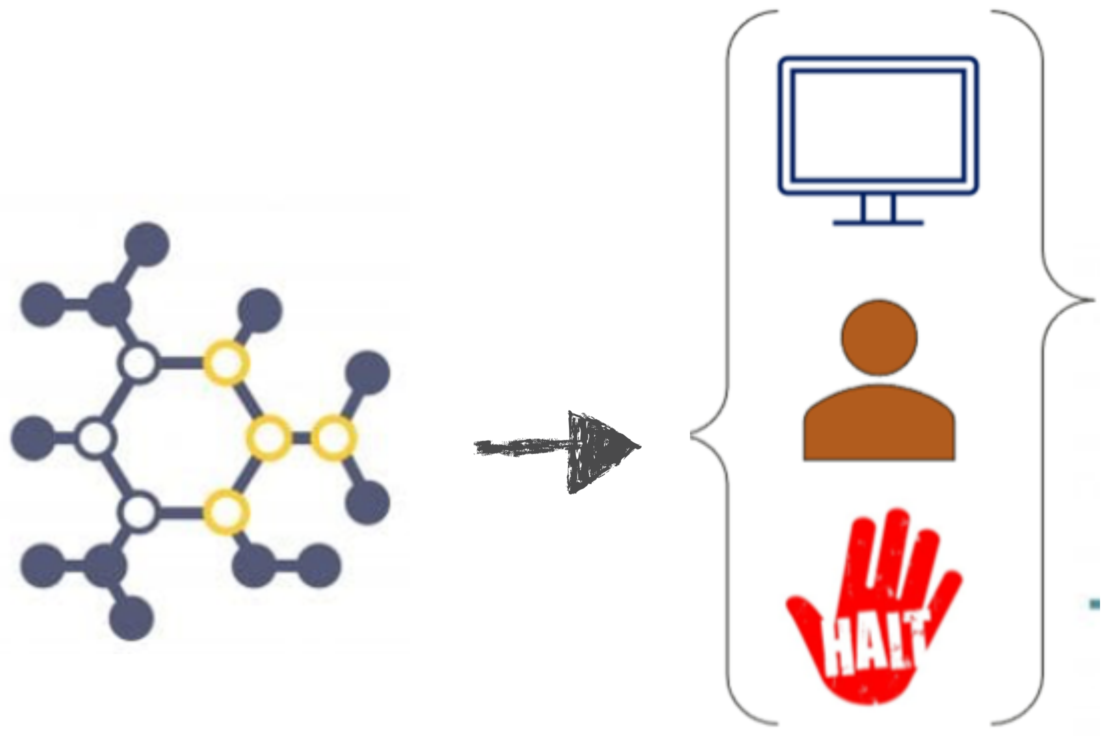


Consistency of representations across frames of a video

Adhere to our Understanding of the Underlying Process



Avoid Shortcut Decision Rules

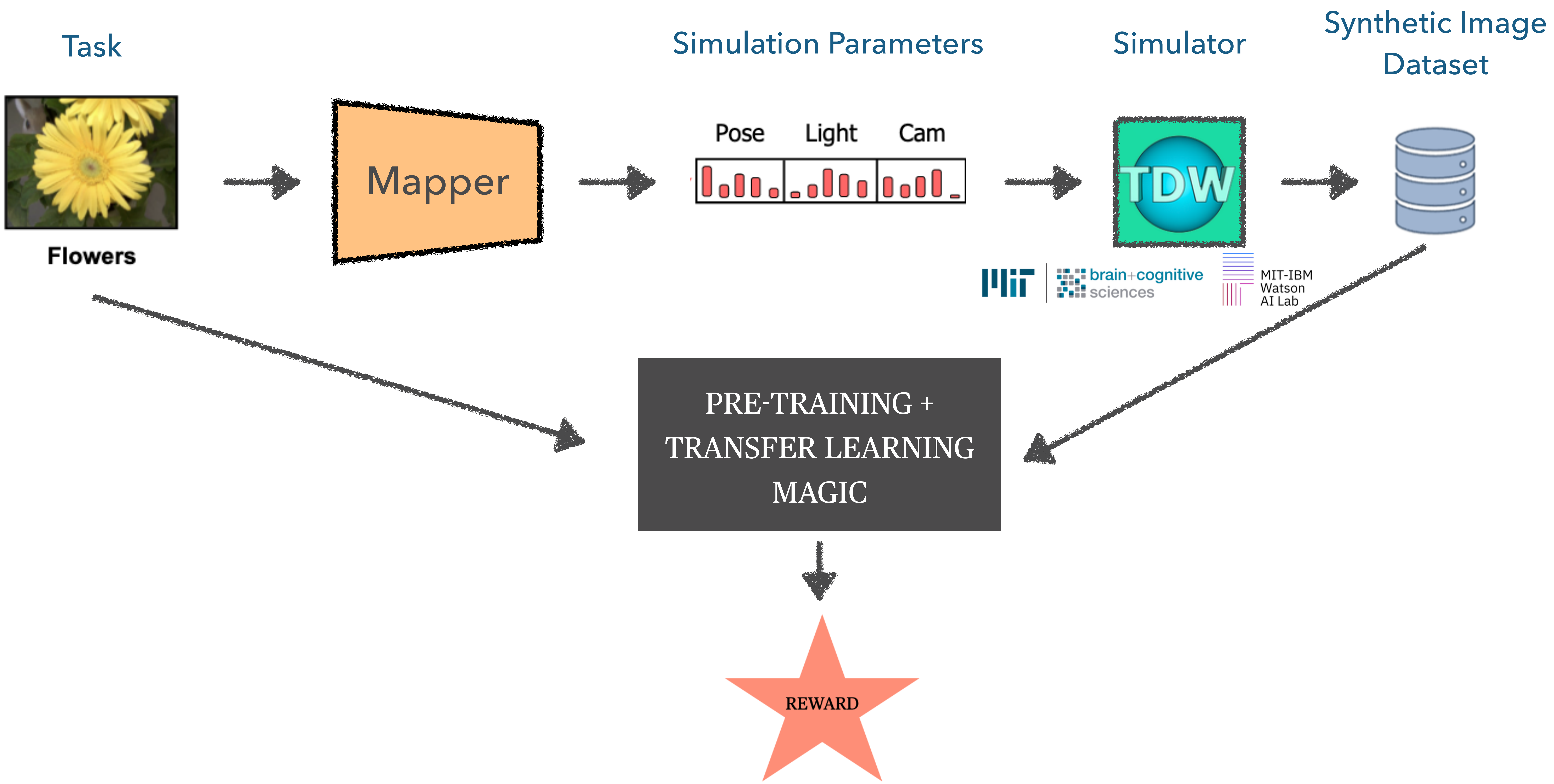


Prescribe not just Predict

Data-Efficiency

We are Transitioning from the Era of Purely Data-Driven Learning to “Domain-Aware” Learning

Physical Models Can Enable Effective Exploration of the Data Manifold Specific to a Given Task



We are Transitioning from the Era of Purely Data-Driven Learning to "Domain-Aware" Learning

Knowledge graphs provide a convenient way to specify domain knowledge without analytical descriptions

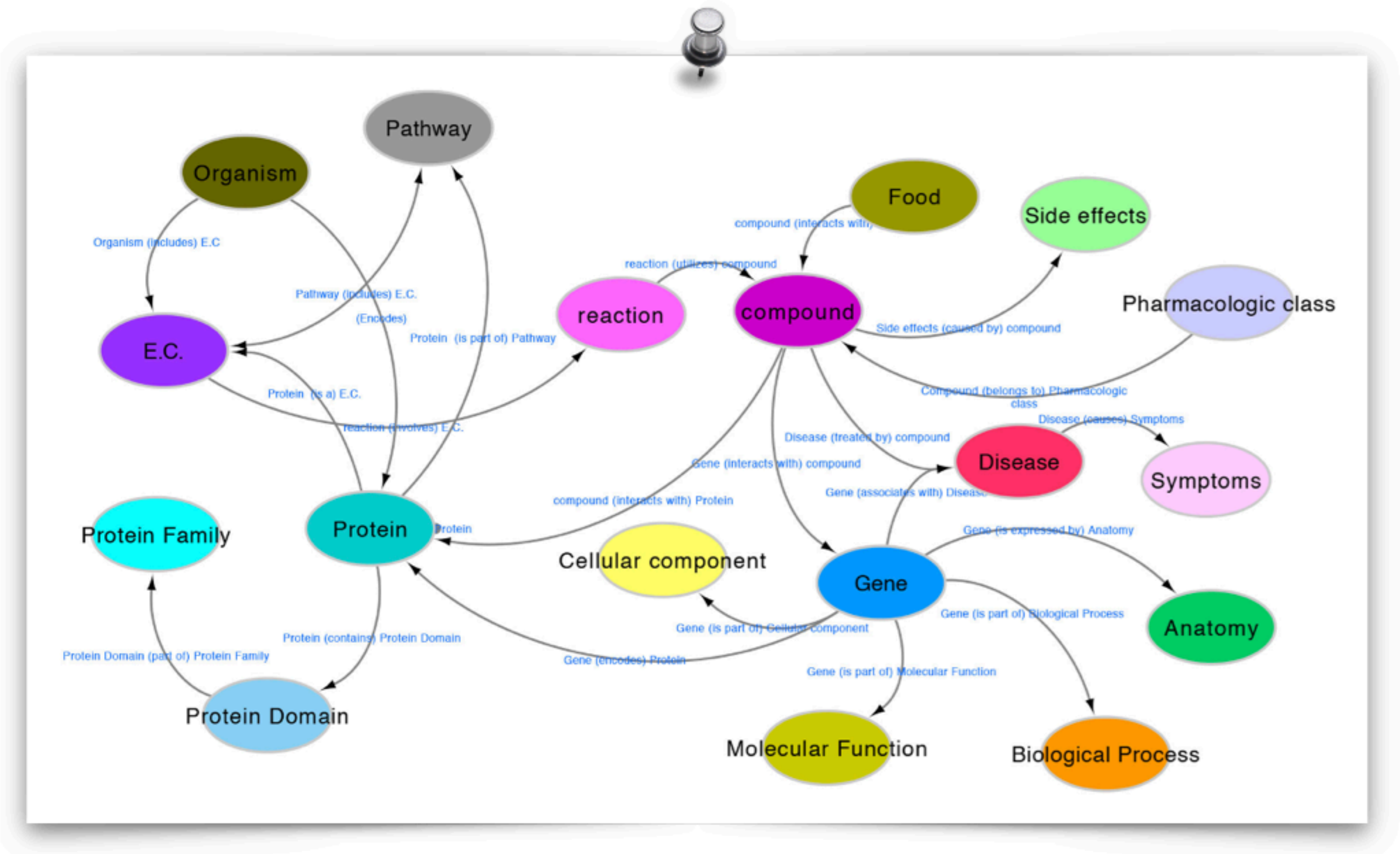
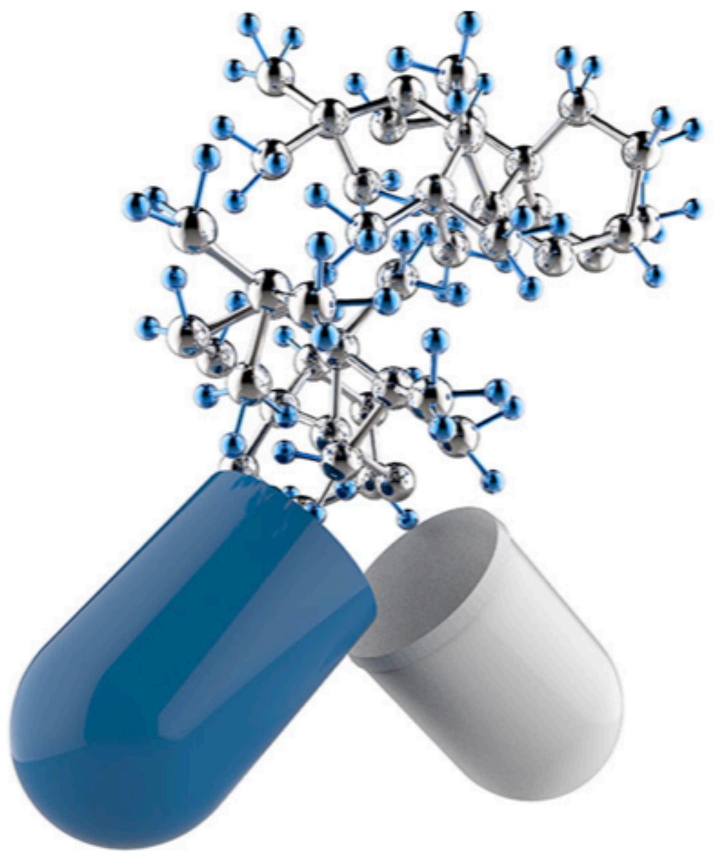
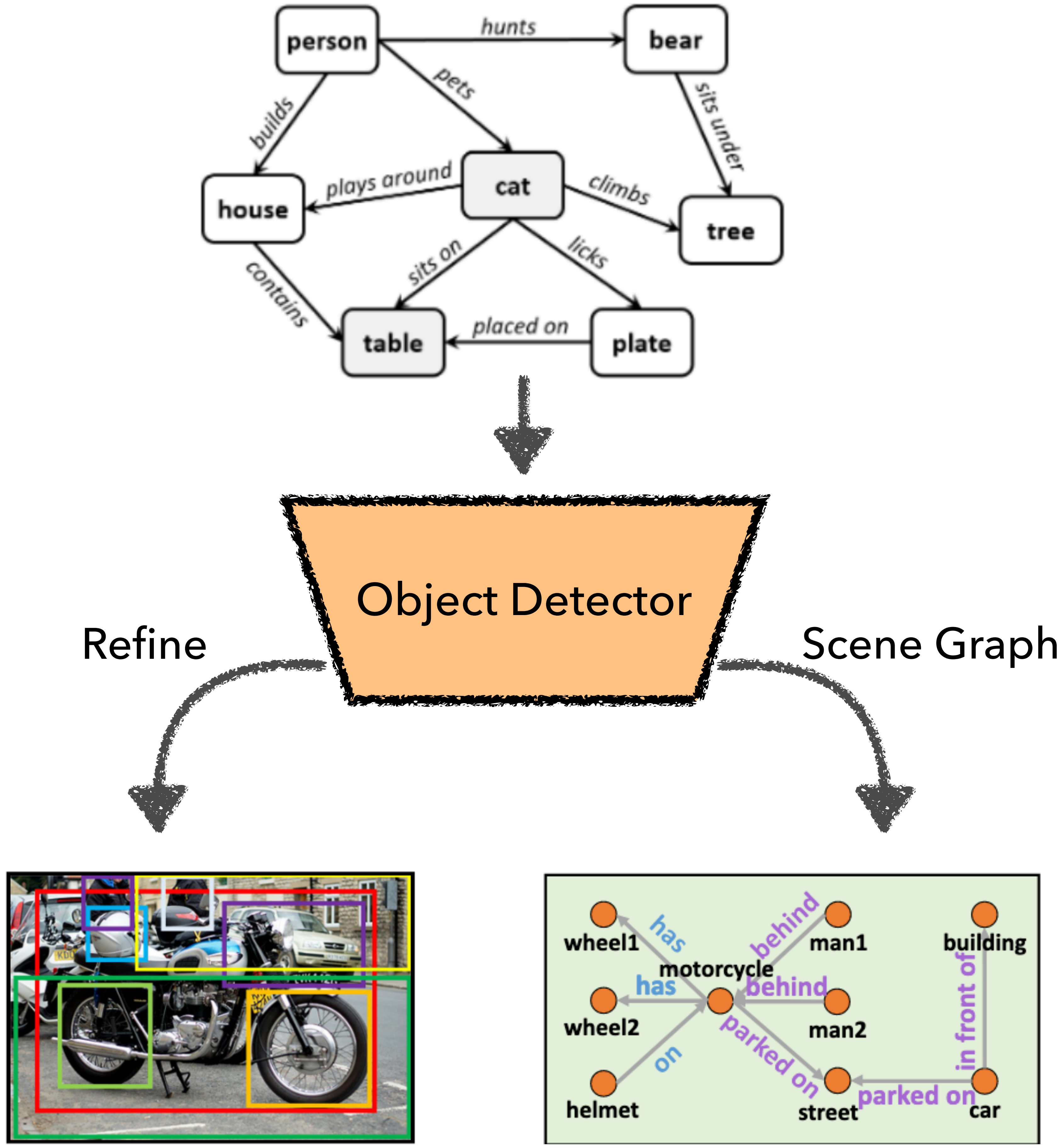


Image source: SPOKE (https://spoke.ucsf.edu/)

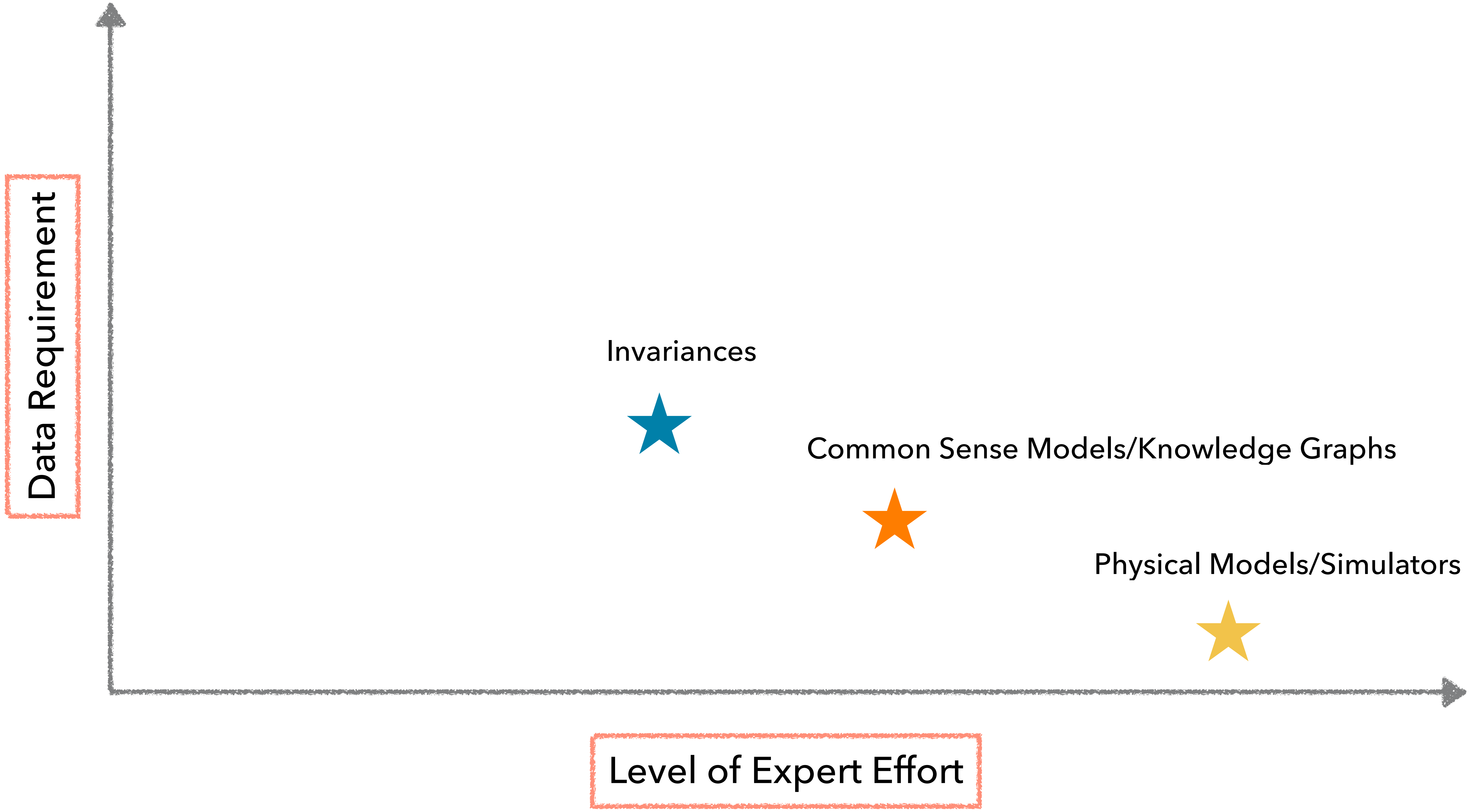
Extended Latent Spaces



Drug Design

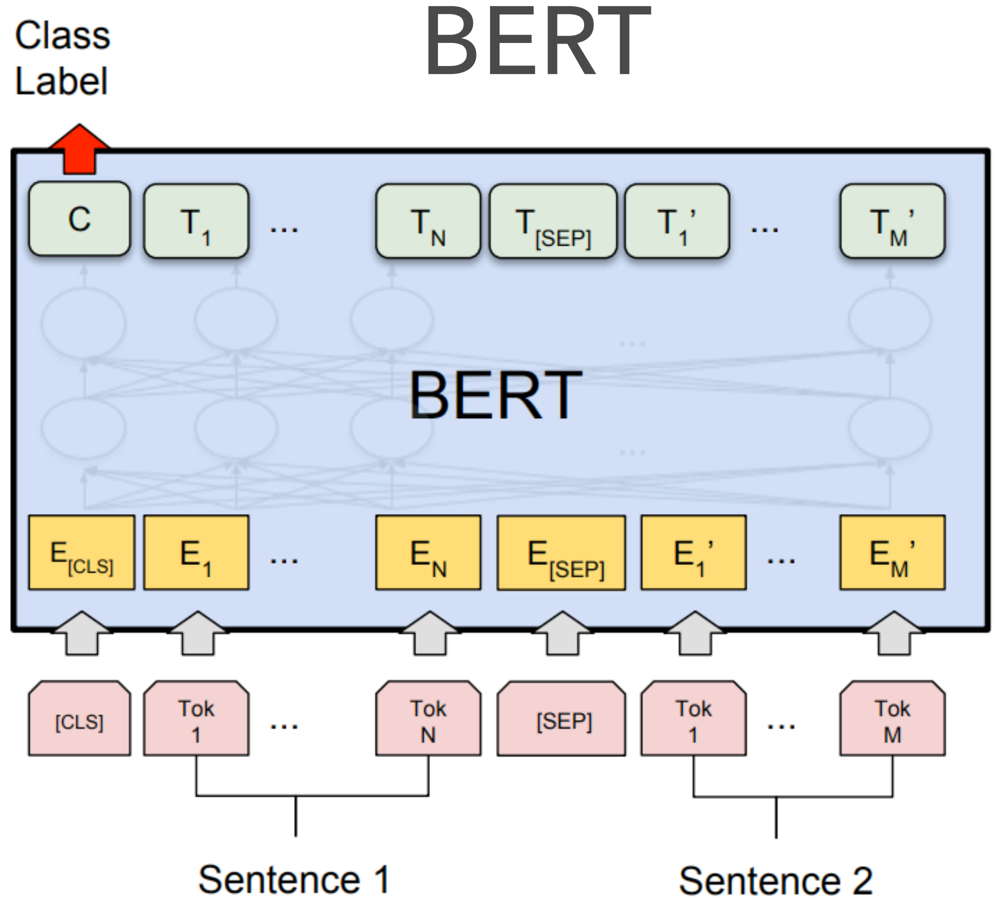


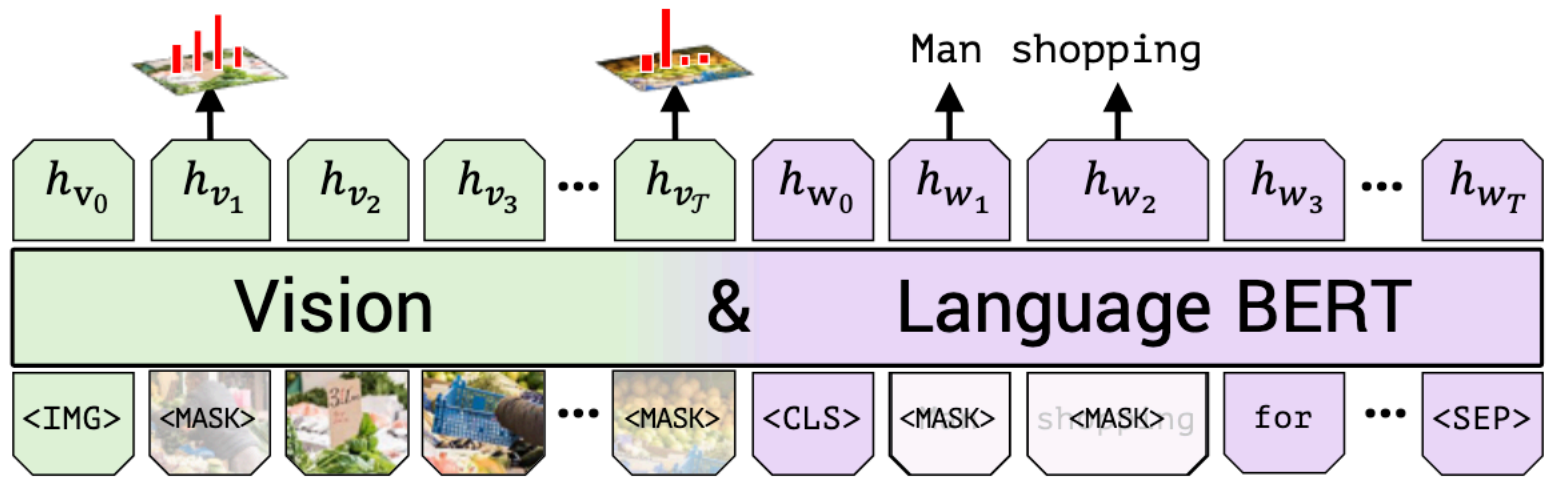
We are Transitioning from the Era of Purely Data-Driven Learning to "Domain-Aware" Learning



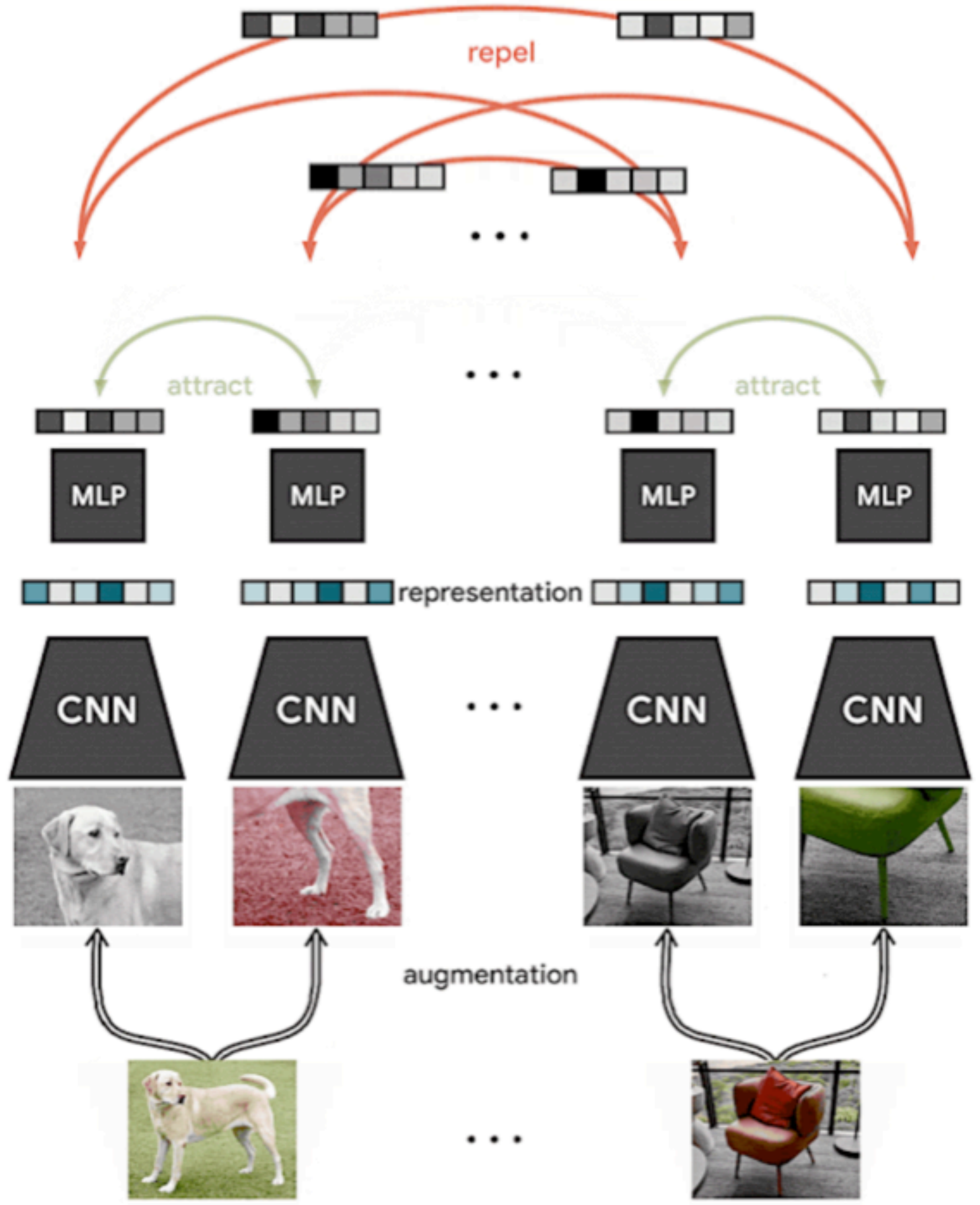
Leveraging Known Invariances is at the Core of Modern Representation Learning

BERT



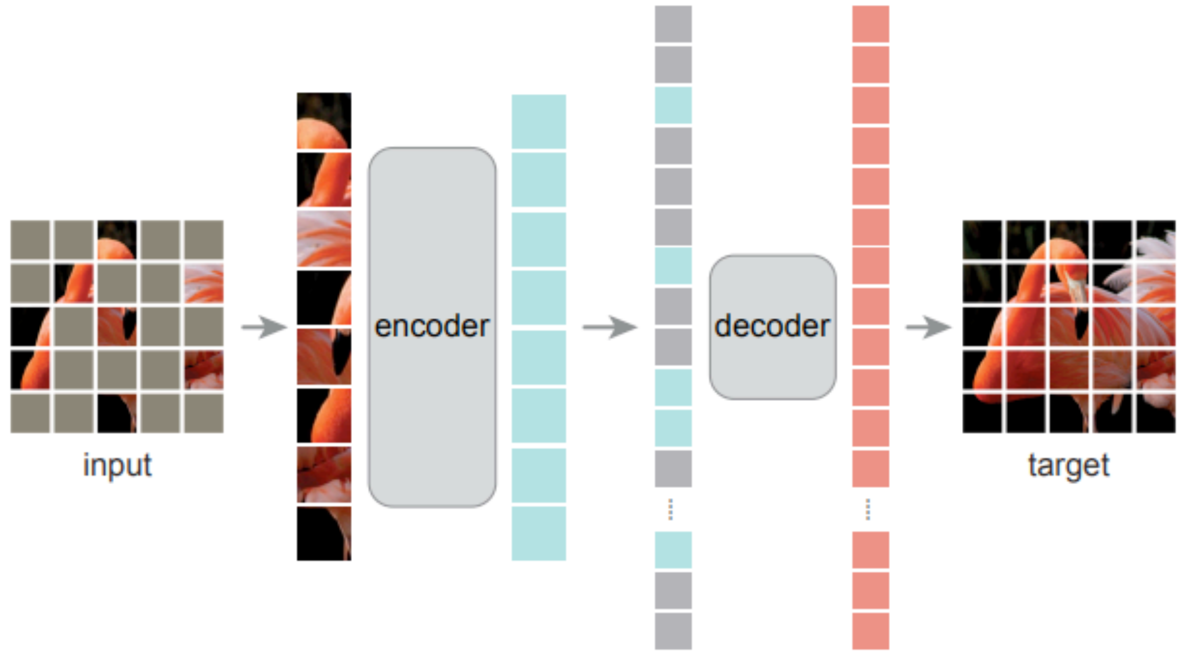


VILBERT

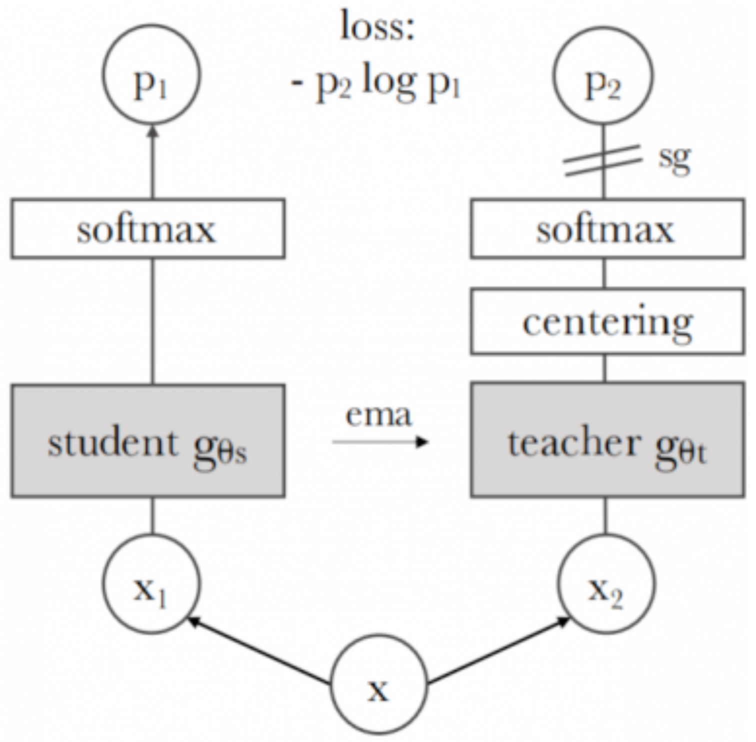


SimCLR, MoCo

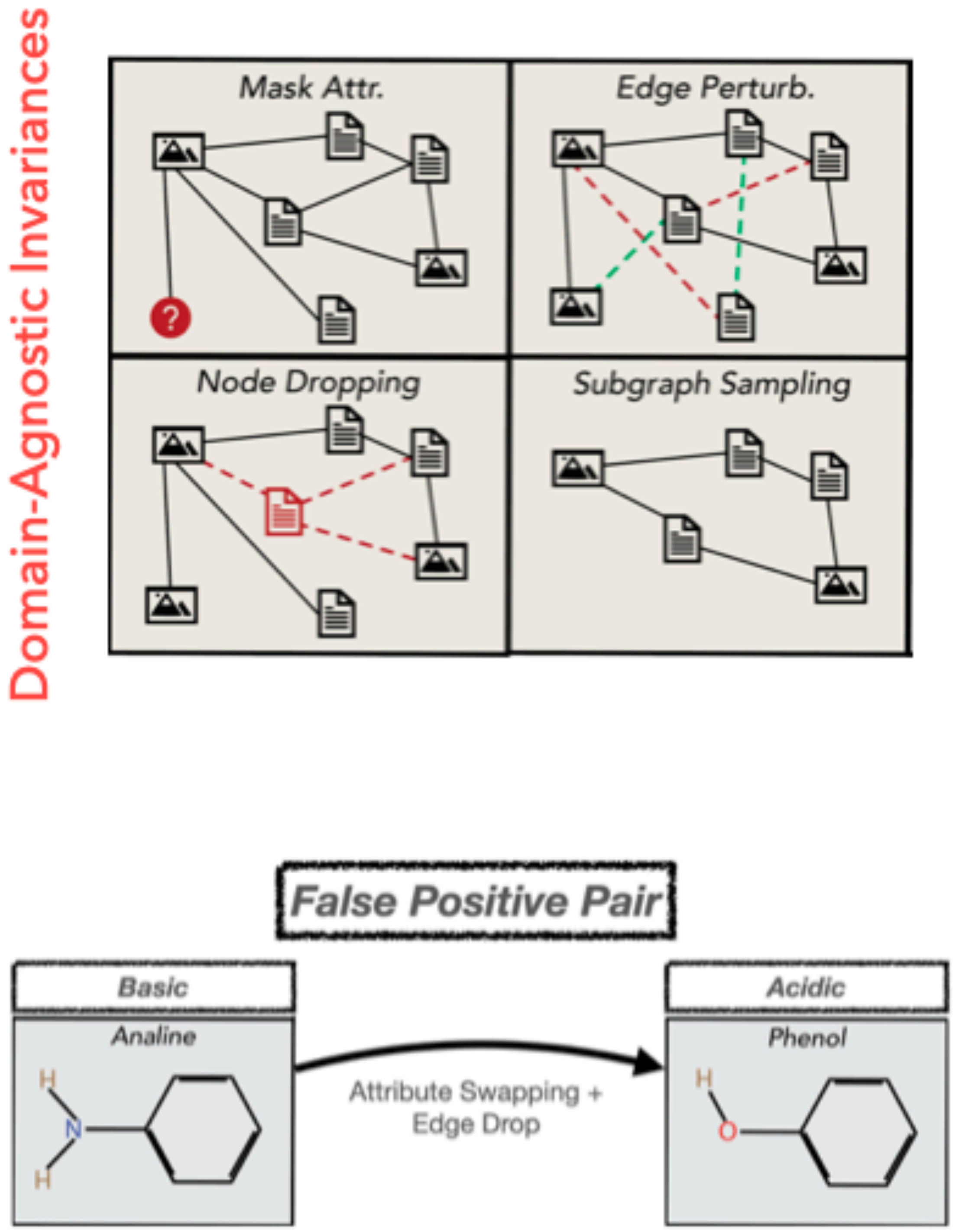
Masked Autoencoders

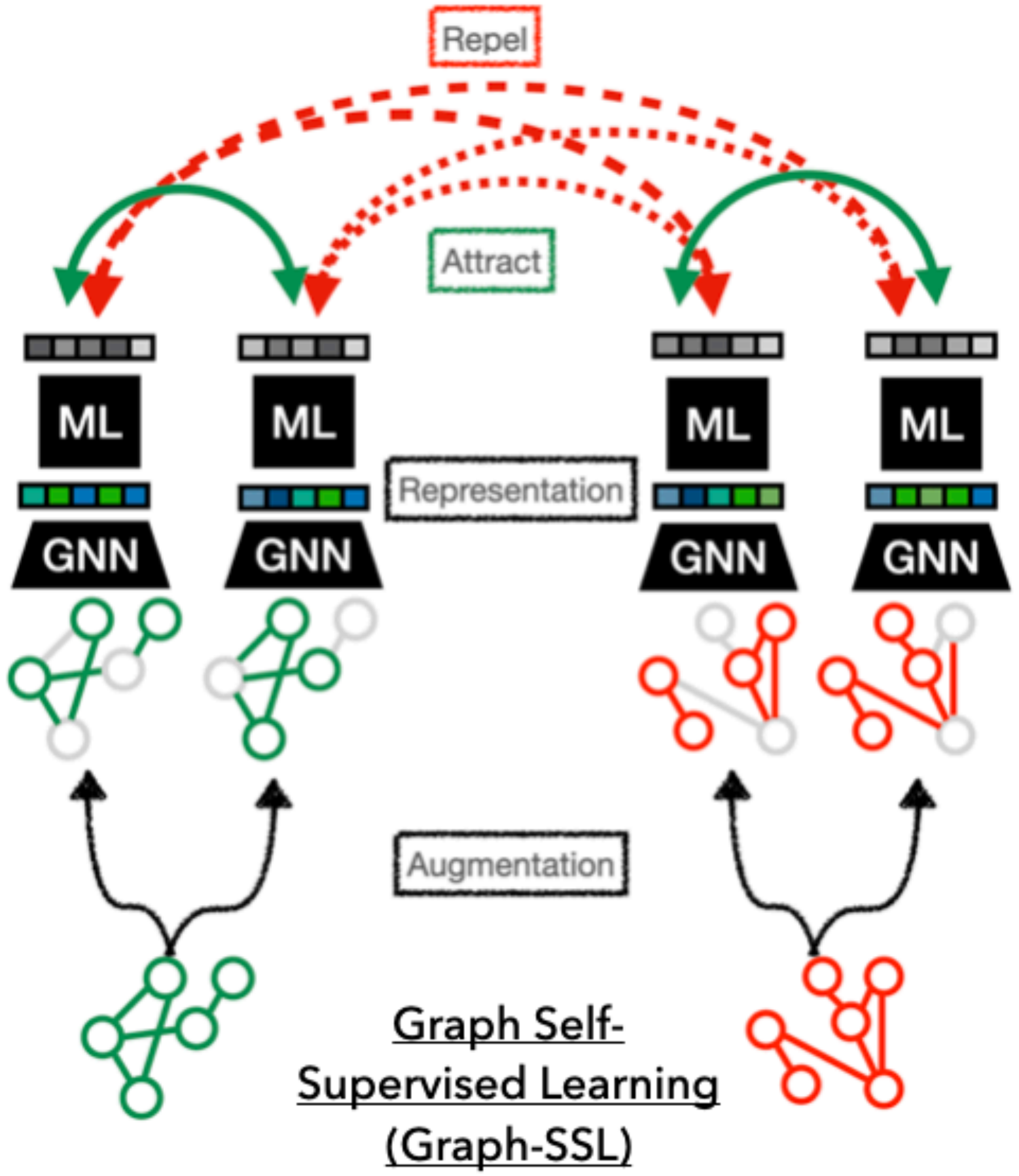


DINO



“Generic Augmentations” Are Not a Silver Bullet for All Applications and Data Modalities



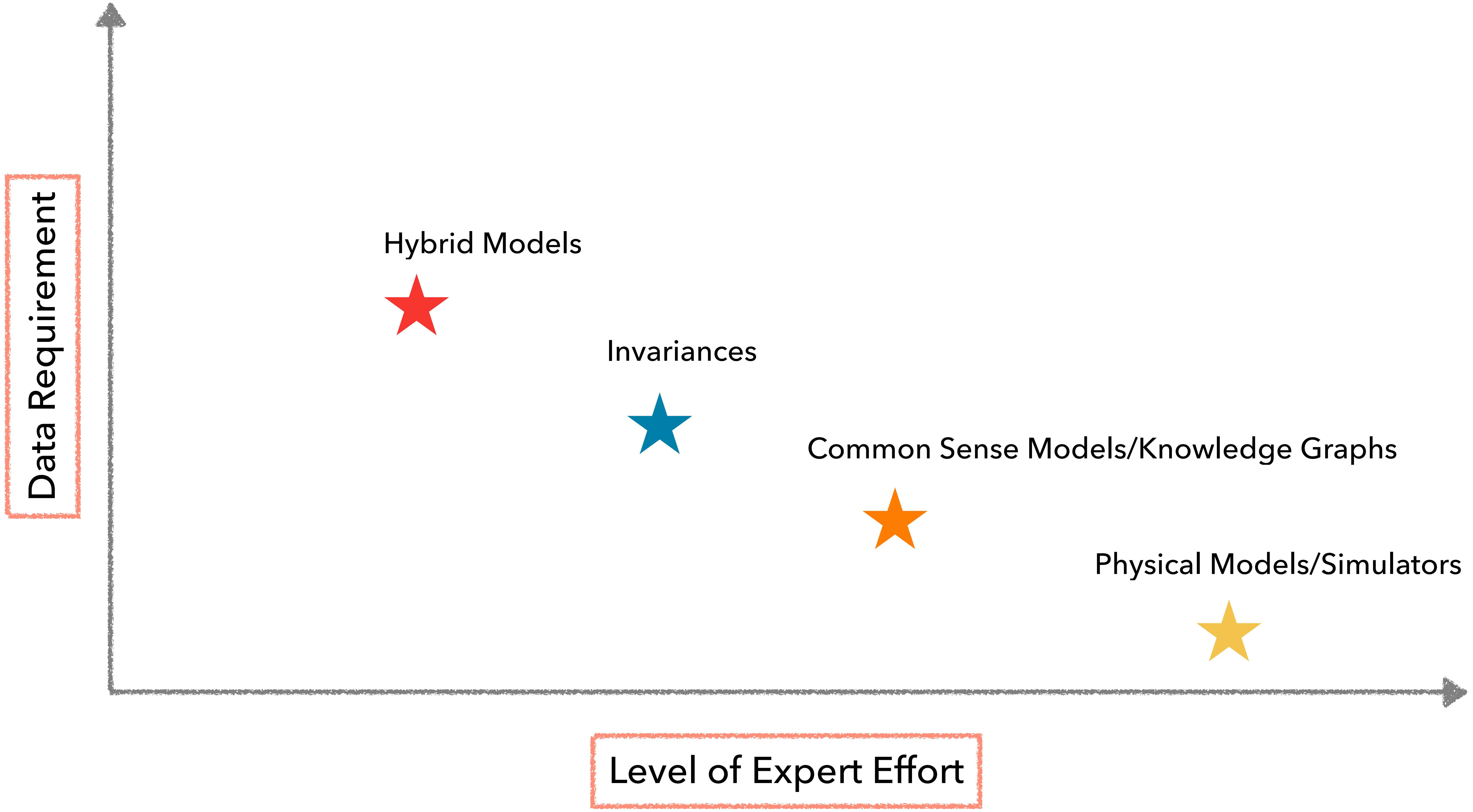


Choosing the “right”augmentations is indeed domain knowledge

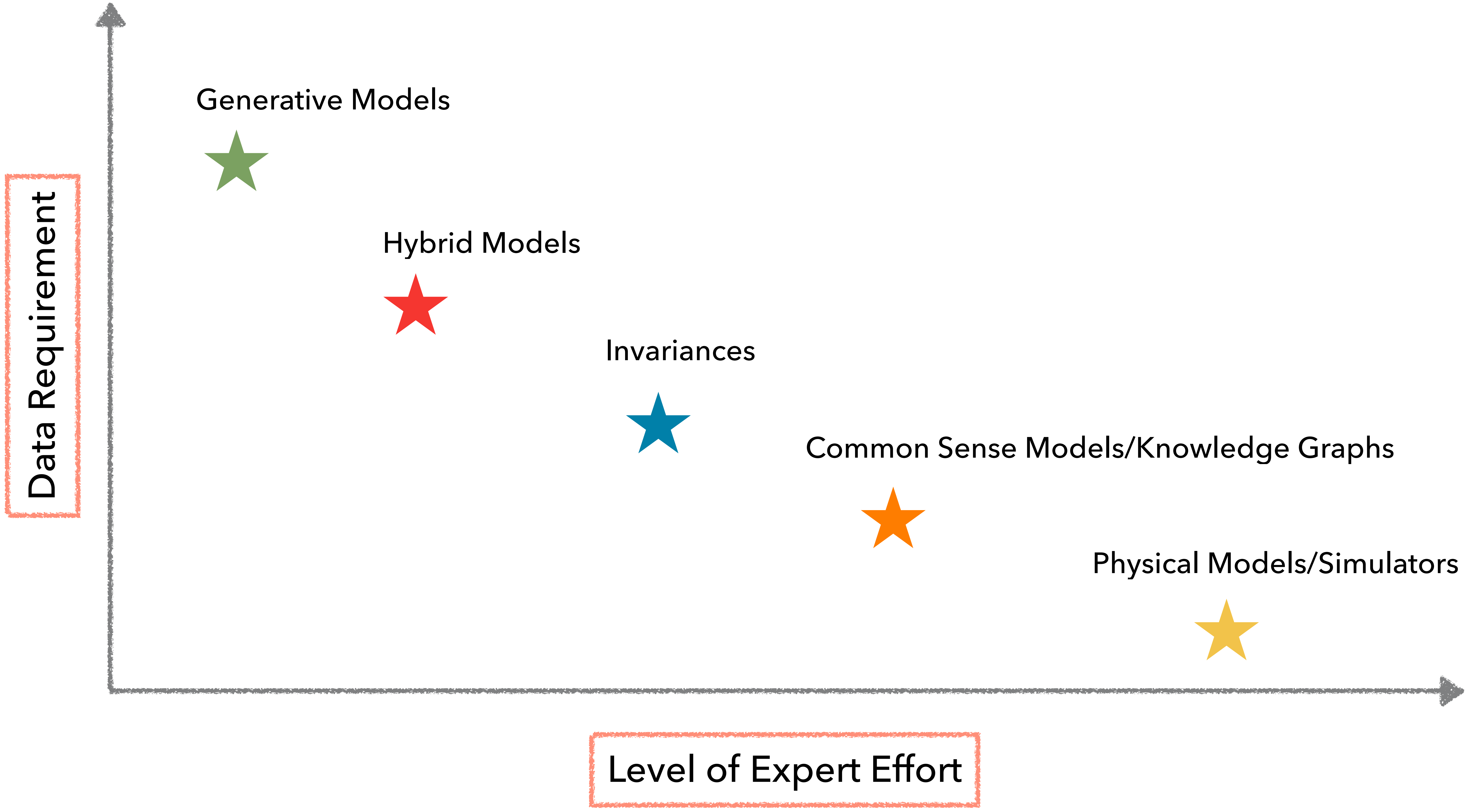
In practice, even state-of-the-art AutoAug techniques fail to pick the “right” augmentation

Analyzing Data-Centric Properties of Contrastive Learning on Graphs, https://arxiv.org/abs/2208.02810

We are Transitioning from the Era of Purely Data-Driven Learning to "Domain-Aware" Learning



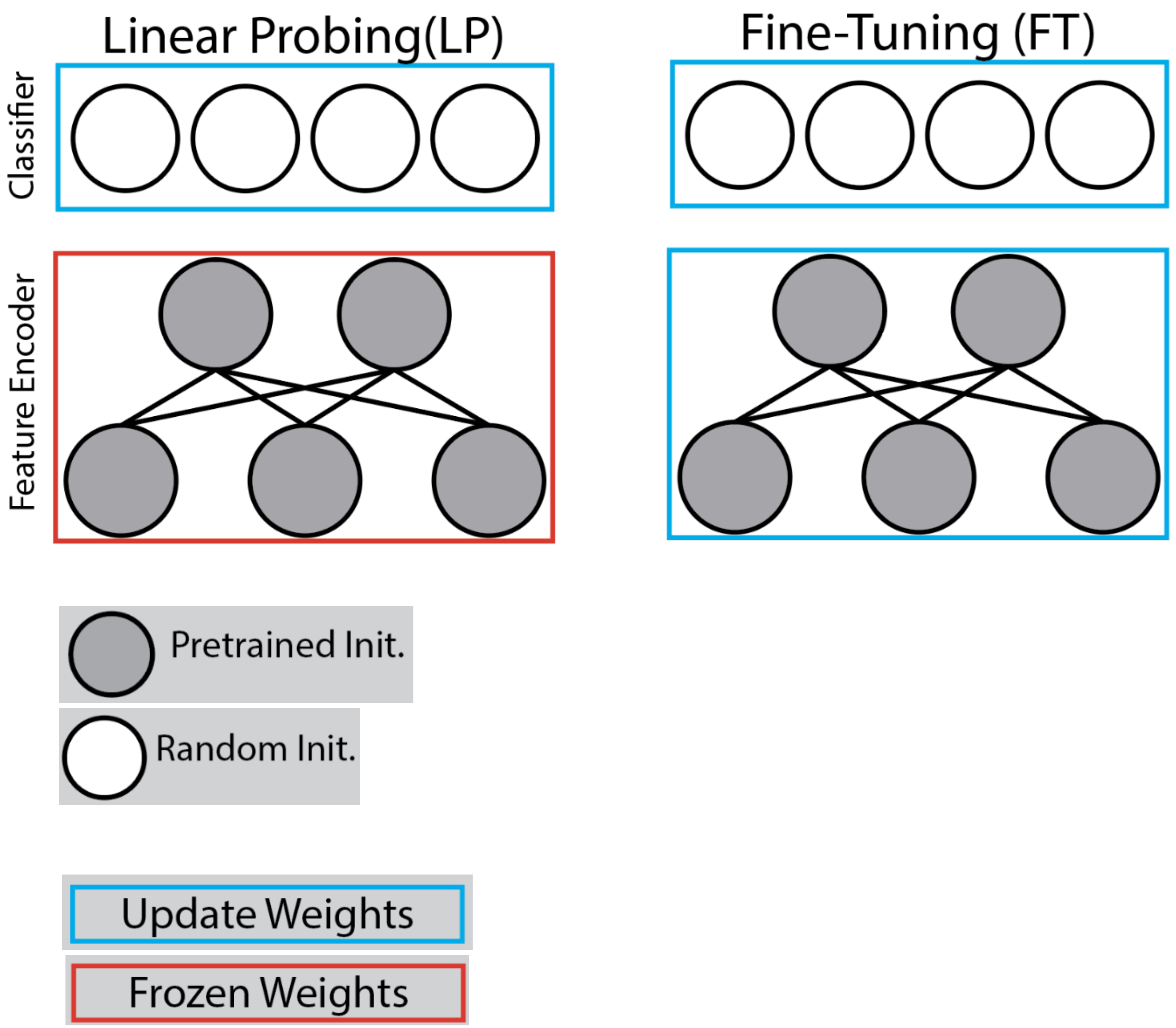
We are Transitioning from the Era of Purely Data-Driven Learning to “Domain-Aware” Learning



Generalization vs Safety Trade-off when Adapting from Pre-Trained Representations

Controlling Feature Distortion by Mitigating Simplicity Bias

With Pre-trained Representations, Adaptation Protocols Have Become a Critical Part of Current ML Pipelines



- What happens with dataset biases?

  LP is already sufficient to avoid shortcuts through data reweighting! (Kirichenko et al., arxiv: 2204.02937)

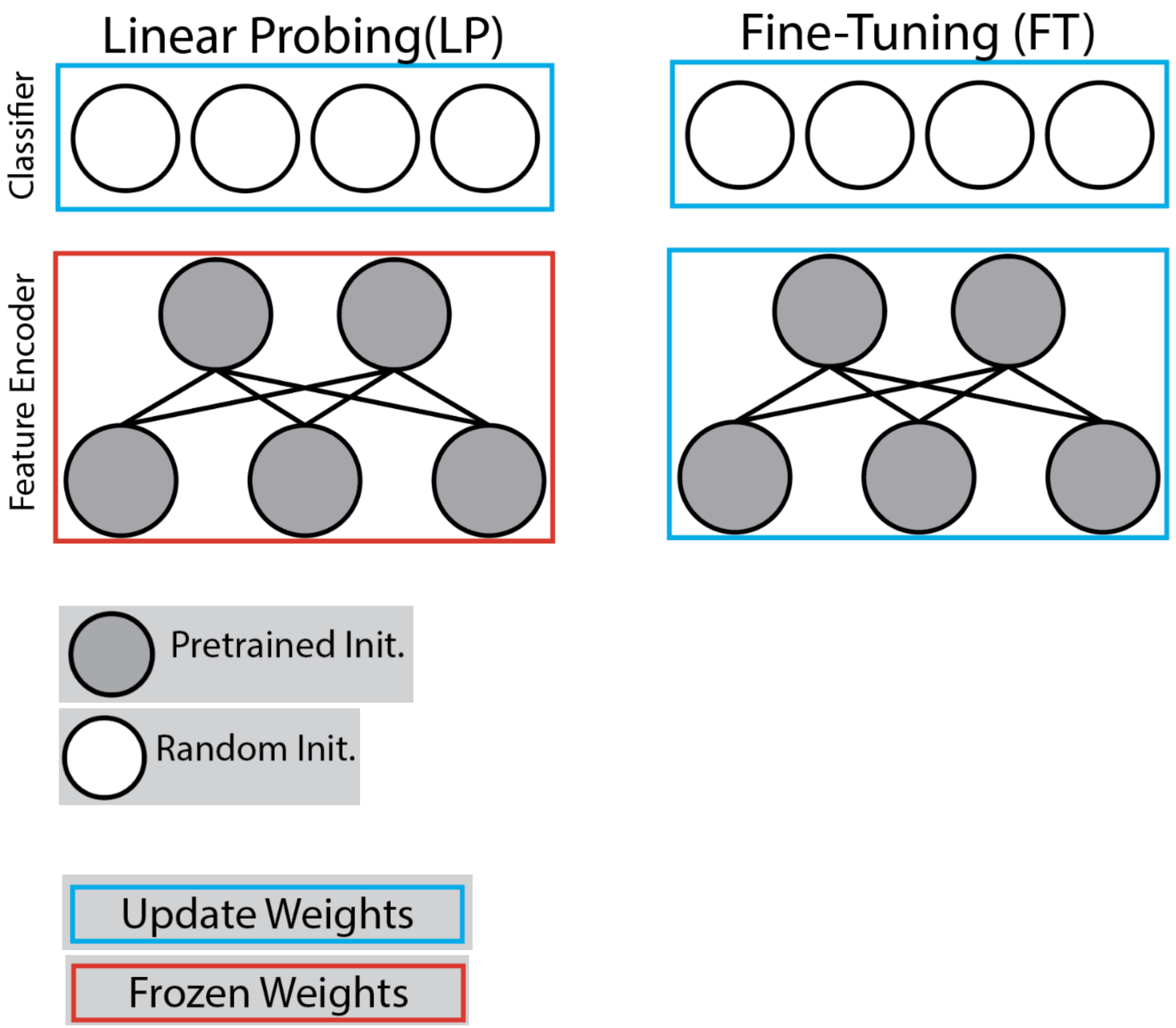
- How do the protocols generalize to ID data?

  By adapting all network features FT often performs better than LP

- How do the protocols generalize to OOD data?

  Surprisingly, under distribution shifts (CIFAR-10 to STL-10), LP outperforms FT

Why does Model Fine-Tuning Compromise on OOD Accuracy?



Lens of Feature Distortion

$$\overbrace{\boldsymbol{y}}^{\text{target}} = \underbrace{\mathrm{v}}_{\text{predictor}}{}^{\top} \overbrace{\mathbf{B}}^{\text{feature extractor}} \underbrace{\mathrm{x}}_{\text{input}}$$

$$\text{Optimal: } \mathbf{B}_\star, \mathrm{v}_\star$$

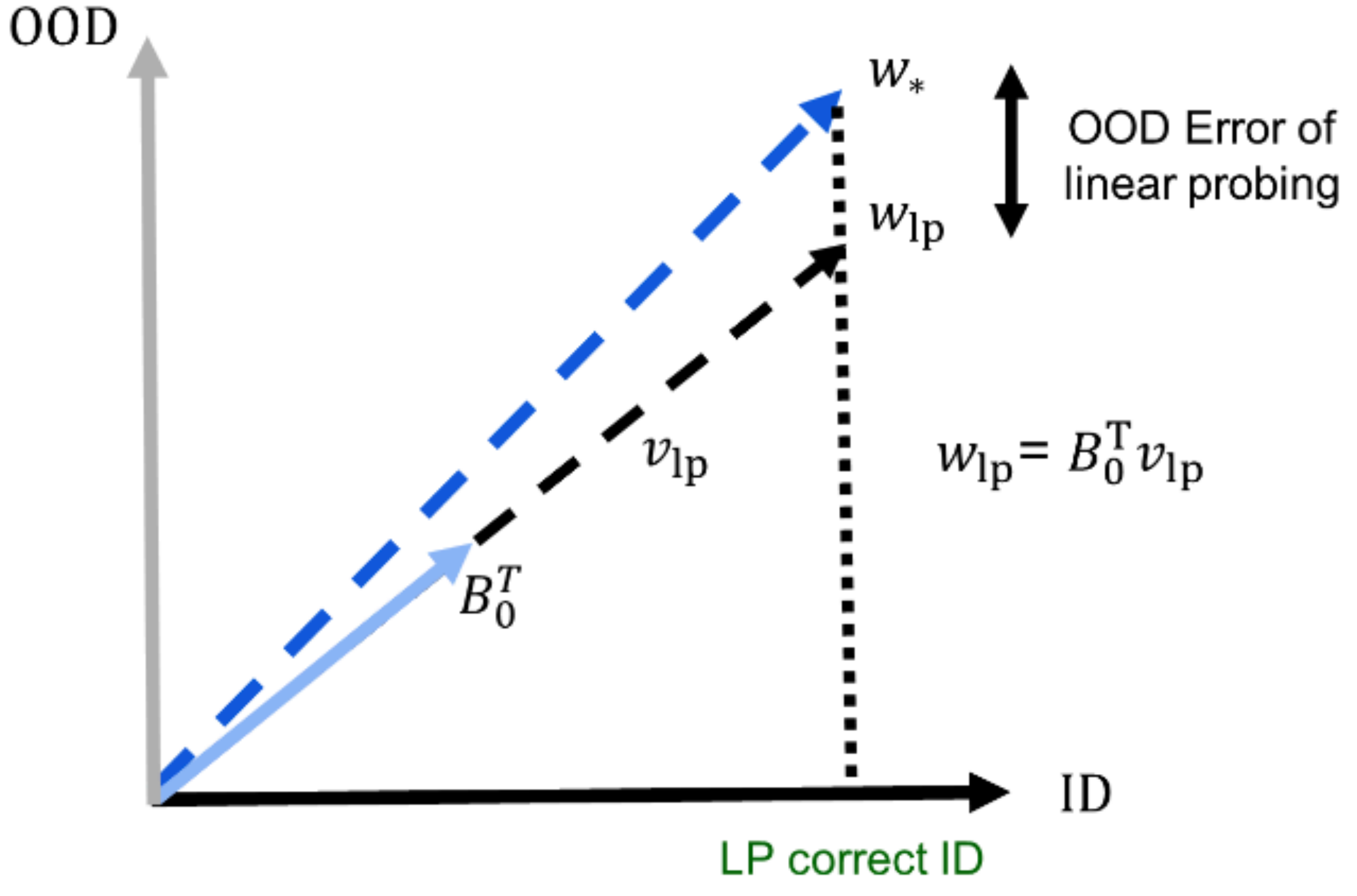
$$(\overbrace{\mathbf{U}}^{\text{rotation}}\mathrm{v}_\star)^\top(\mathbf{U}\mathbf{B}_\star)\mathrm{x} = \mathrm{v}_\star^\top \mathbf{B}_\star \mathrm{x}$$

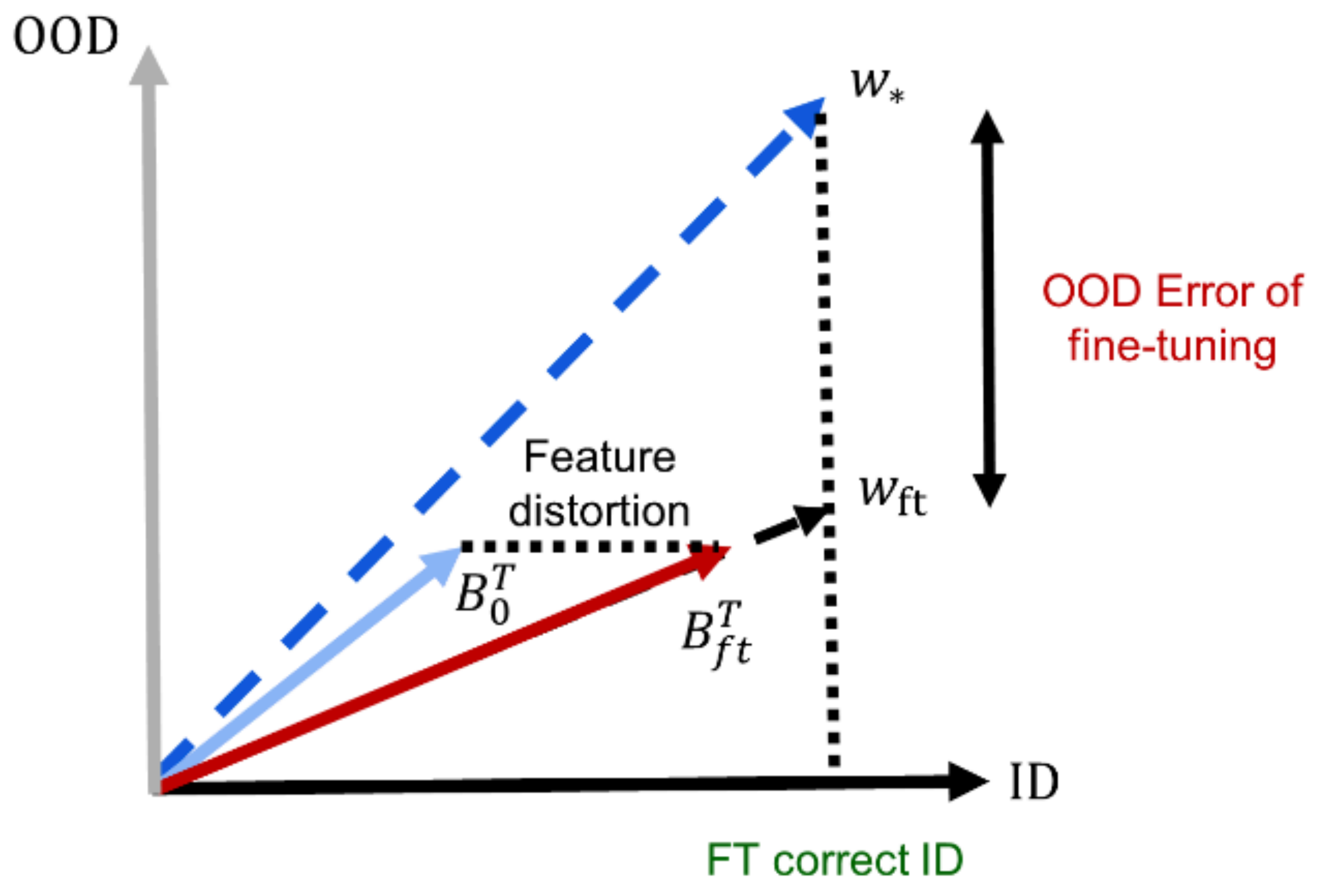
FEATURE EXTRACTOR DISTANCE

Distance between two feature extractors is measured as

$$d(\mathbf{B}_0, \mathbf{B}_\star) = \min_{\mathbf{U}} \|\mathbf{B}_0 - \mathbf{U}\mathbf{B}\|_2$$

Why does Model Fine-Tuning Compromise on OOD Accuracy?





Two Key Insights

- Features get distorted only in the ID subspace and not in the orthogonal subspace

$$\nabla_{\mathbf{B}} L(\mathrm{v}, \mathbf{B}) = 2\mathrm{v}(\mathrm{Y} - \mathrm{X}\mathbf{B}\mathrm{v})^\top \mathrm{X} \quad \boxed{\nabla_{\mathbf{B}} L(\mathrm{v}, \mathbf{B})\mathrm{u} = 0, \mathrm{u} \in S^\perp}$$

- Feature distortion leads to higher OOD error

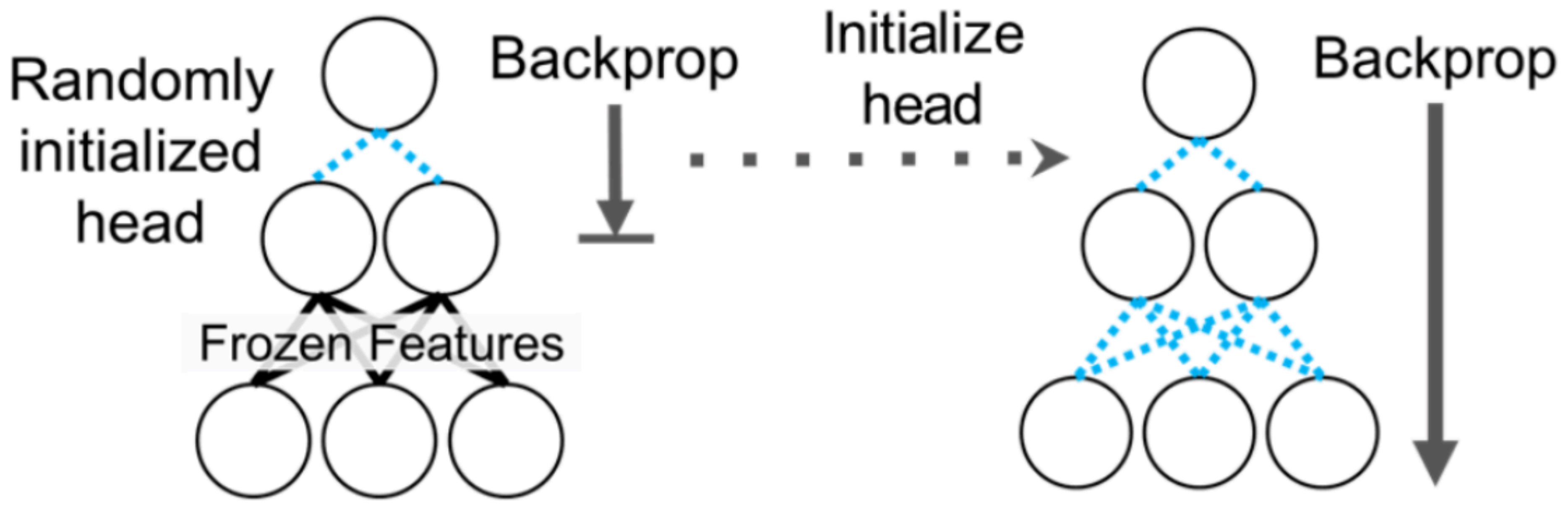
$$\sqrt{\sigma_{\min}(\mathbf{\Sigma})}\left(\frac{\cos\theta_{\max}(R_0, S^\perp)}{\sqrt{k}} \frac{\min(\phi, \phi^2/\|\mathrm{w}_\star\|_2)}{(1 + \|\mathrm{w}_\star\|_2)^2} - \epsilon\right)$$

Lower bound

$$\phi^2 = \left|(\mathrm{v}_0^\top \mathrm{v}_\star)^2 - (\mathrm{v}_\star^\top \mathrm{v}_\star)\right|$$

Empirical Insight: Performing LP Prior to Invoking the FT Step Appears to Fix this Issue

This two-step optimization controls the amount of feature distortion



Kumar et al., "Fine Tuning Can Distort Pre-Trained Features and Underperform Out-of-Distribution", ICLR 2022

Protocol	ID Train ACC	ID Test ACC	OOD ACC
LP	97.2	91.39	81.94
FT	**99.5**	**95.58**	80.34
LP+FT	98.9	94.50	**86.57**

ID: Cifar-10 OOD: STL-10

Performing LP Prior to Invoking the FT Step Can Fix this Issue or Can It?

calibration

corruptions

anomaly rejection

Generalization vs Safety Trade-off

WHAT HAPPENS IF WE TAKE INTO ACCOUNT MODEL SAFETY TO OBTAIN A HOLISTIC EVALUATION?

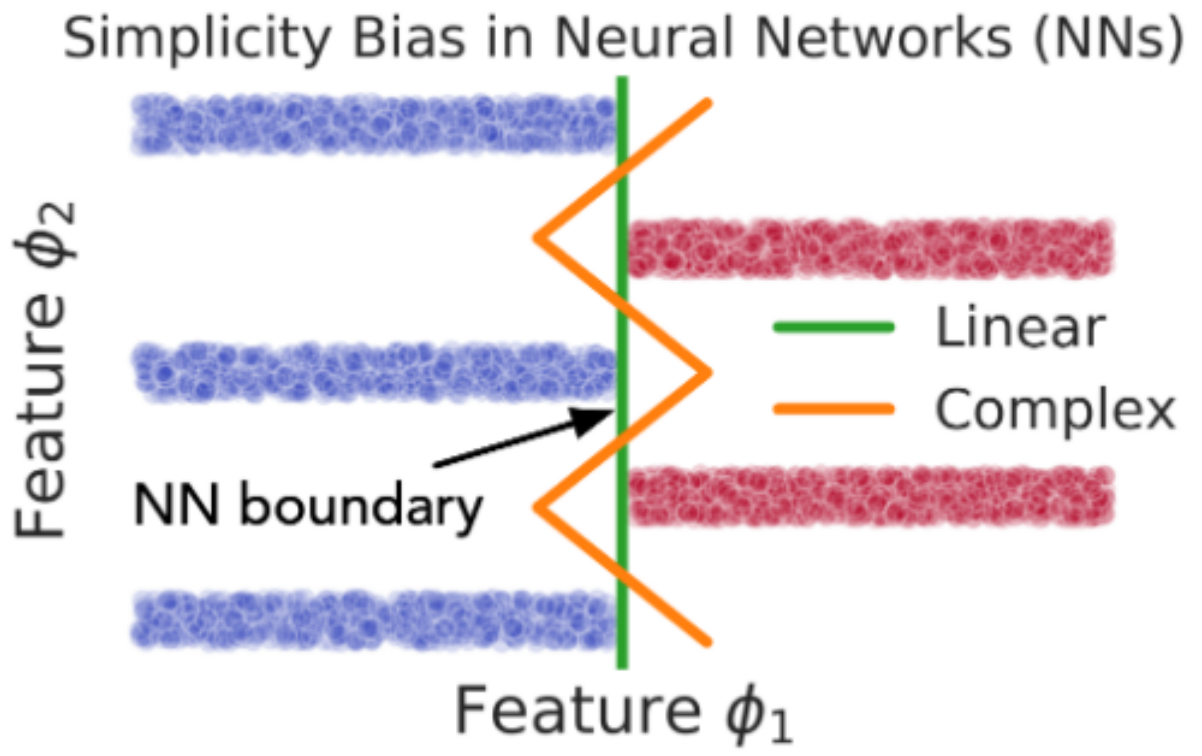
Protocol	ID Test ACC	OOD ACC	mCA	CE (RMSE)	Anomaly AUROC
LP	91.39	81.94	69.1	0.171	62.1
FT	**95.58**	80.34	**74.7**	**0.137**	**99.1**
LP+FT	94.50	**86.57**	69.1	0.217	64.5

ID: Cifar-10 OOD: STL-10

Balancing Transferability and Task Performance during Adaptation

We consider a synthetic dataset obtained by blending each CIFAR-10 class with a corresponding class from MNIST

Hypothesis: If the LP/FT model relies on the simplest features (digit) and remains invariant to complex features (CIFAR), it will fail when the digit mapping is switched at test time

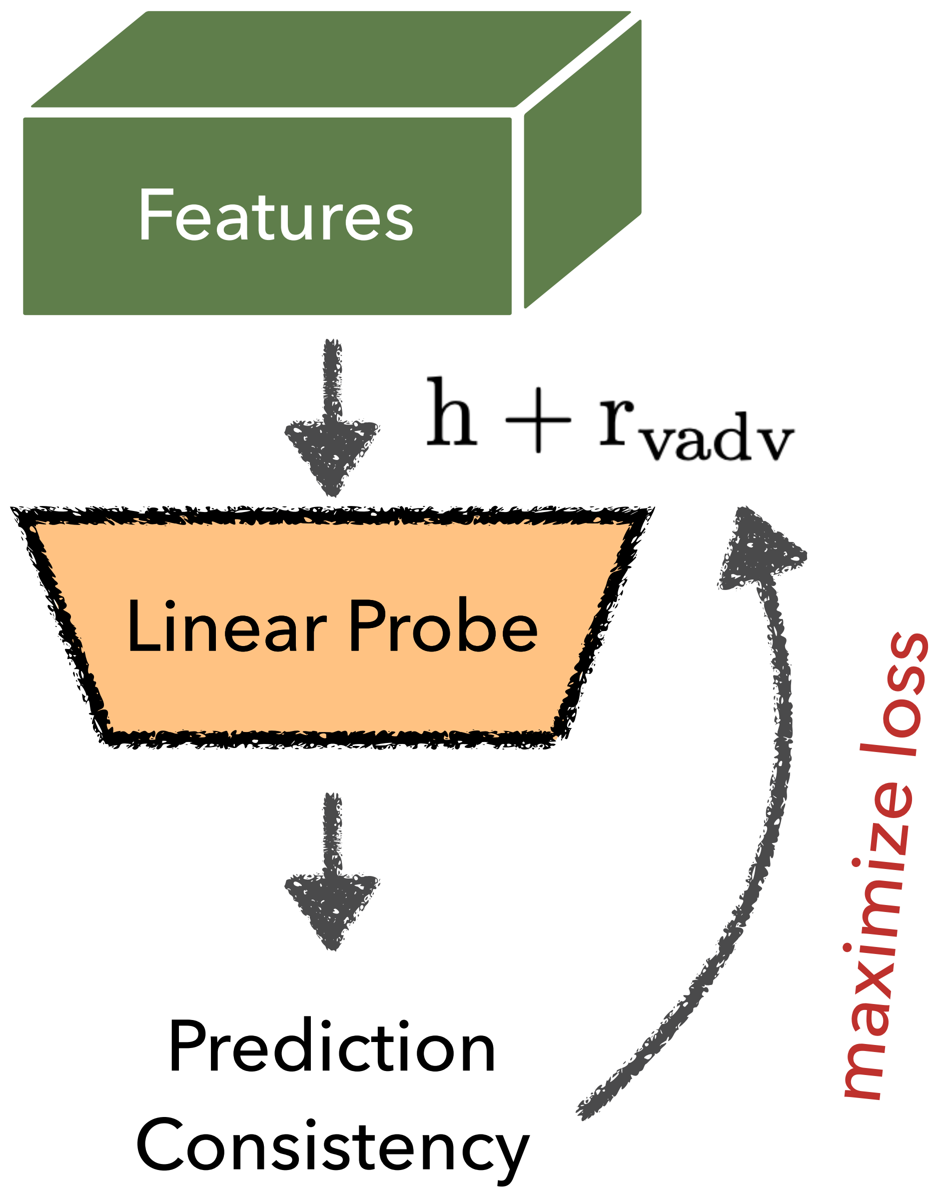


Shah et al., "The Pitfalls of Simplicity Bias in Neural Networks", Neurips 2020

Protocol	ID Test ACC	OOD (Rand) ACC
LP	90.3	79.9
FT	**98.5**	39.1

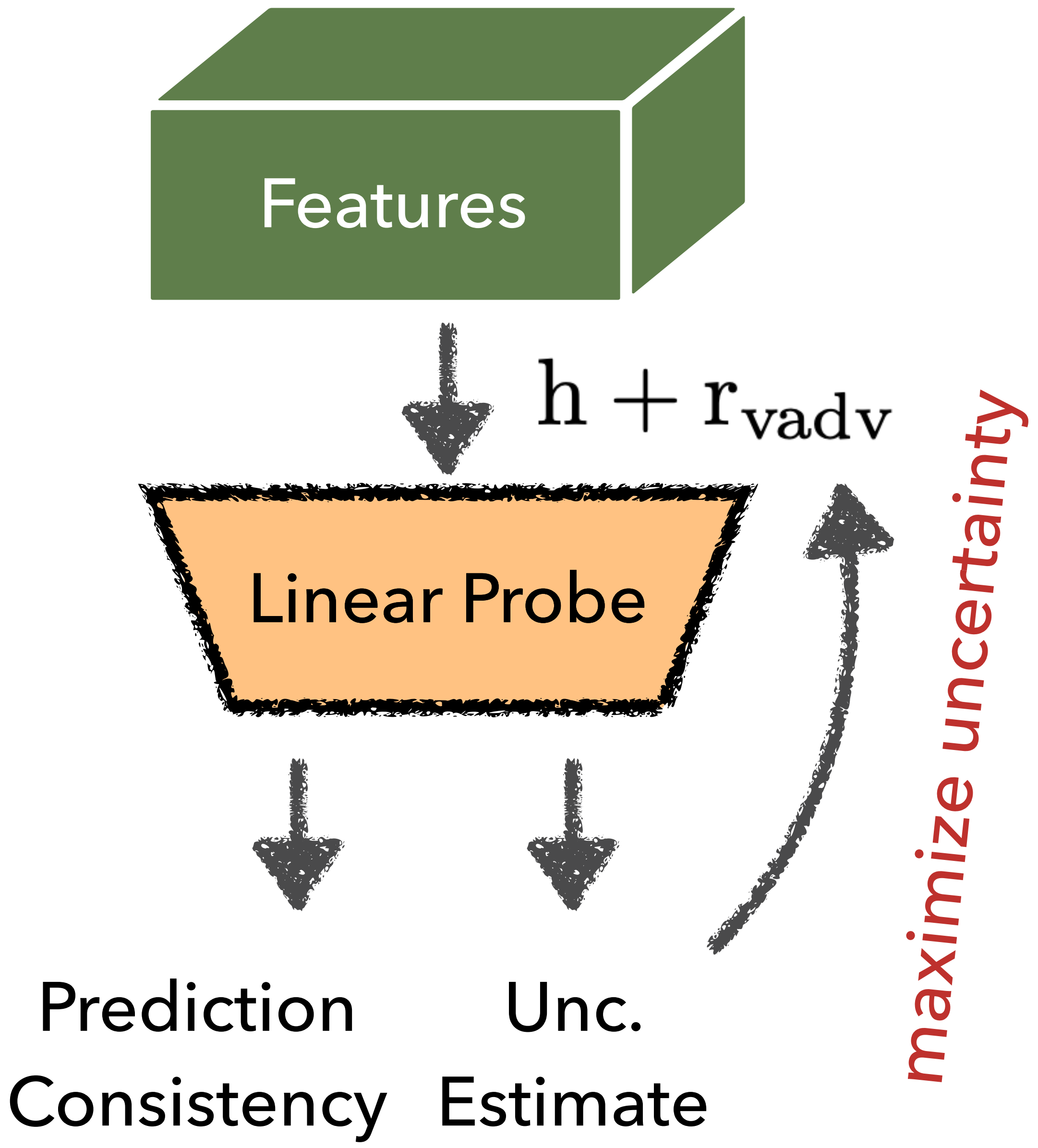
We Leverage Hardness-Promoting Augmentations during LP to Mitigate Simplicity Bias During FT

Loss-based perturbations



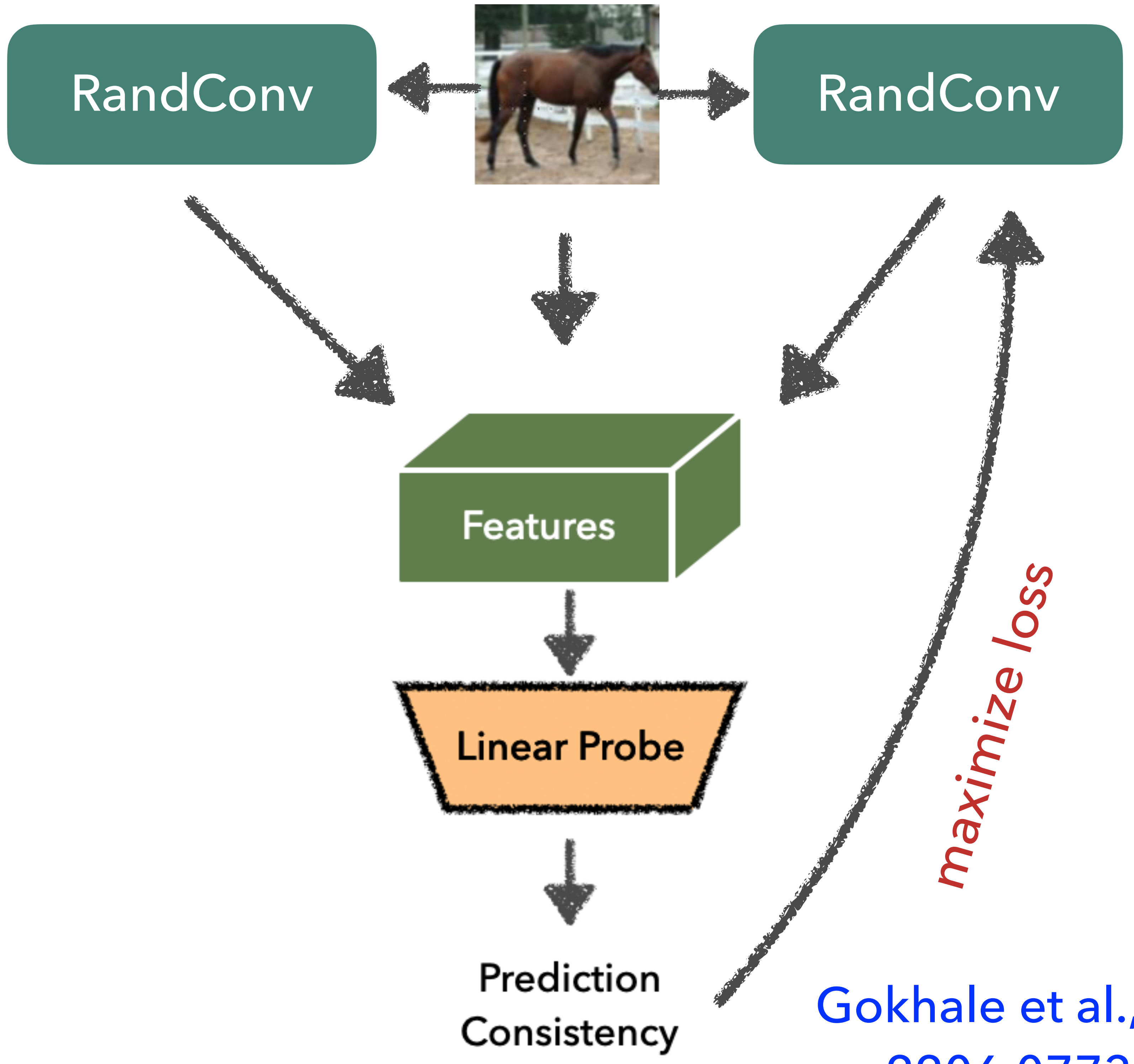
Trivedi et al., arxiv: 2207.12615

Uncertainty-based perturbations



Pagliardini et al., arxiv: 2202.05737

Improved Diversity with ALT



Gokhale et al., arxiv: 2206.07736v1

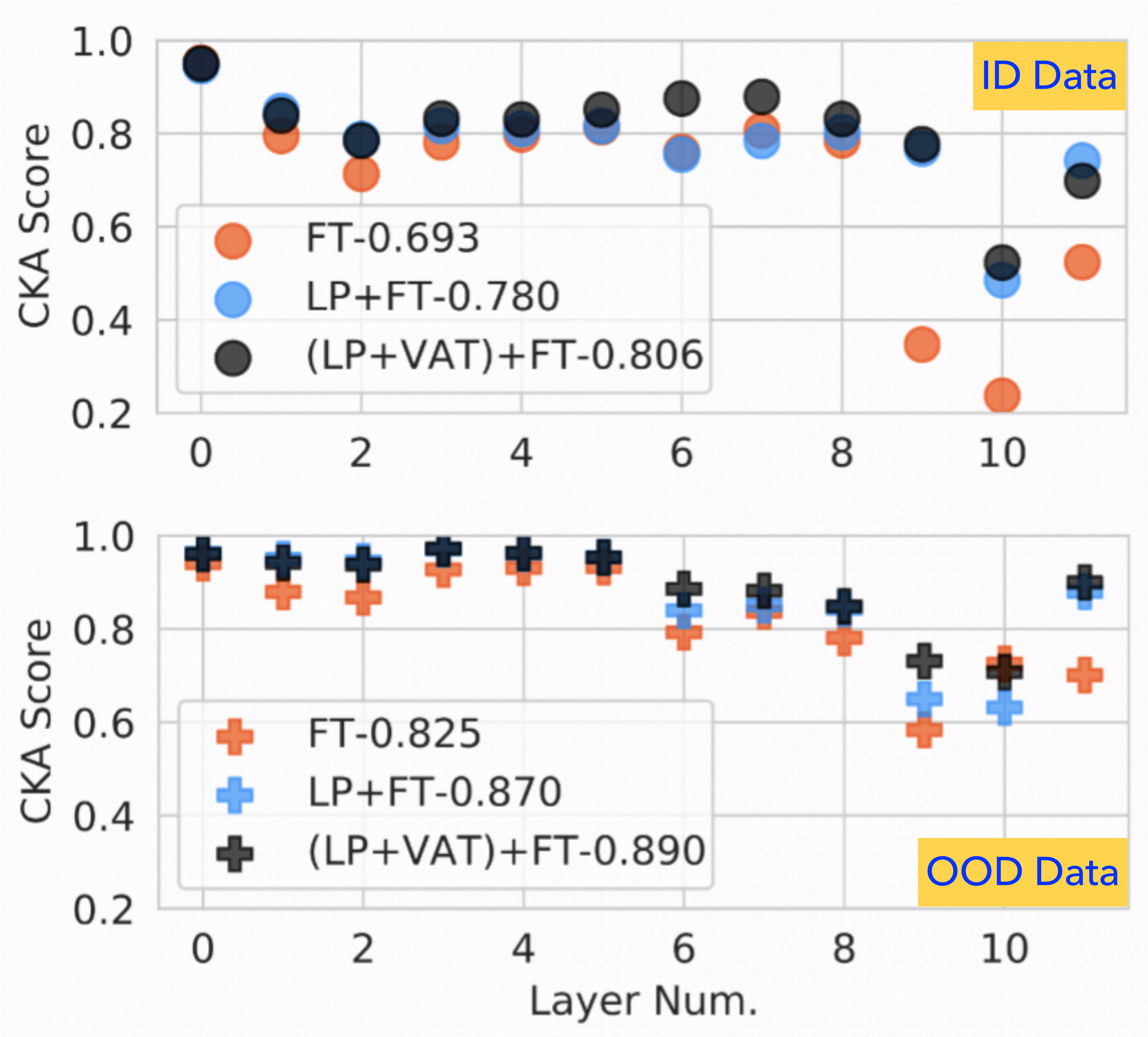
This Modification to the LP Step Further Controls the Feature Distortion in (LP + FT) Protocols

Centered Kernel Alignment (CKA)

$$\frac{\frac{1}{k}\sum_{i=1}^{k}\mathrm{H}(\mathbf{X}_i\mathbf{X}_i^\top, \mathbf{Y}_i\mathbf{Y}_i^\top)}{\sqrt{\frac{1}{k}\sum_{i=1}^{k}\mathrm{H}(\mathbf{X}_i\mathbf{X}_i^\top, \mathbf{X}_i\mathbf{X}_i^\top)}\sqrt{\frac{1}{k}\sum_{i=1}^{k}\mathrm{H}(\mathbf{Y}_i\mathbf{Y}_i^\top, \mathbf{Y}_i\mathbf{Y}_i^\top)}}$$

$$\frac{1}{n(n-3)}\left(\mathrm{Tr}(\tilde{\mathbf{K}}\tilde{\mathbf{L}}) + \frac{\mathbf{1}^\top\tilde{\mathbf{K}}\mathbf{1}\mathbf{1}^\top\tilde{\mathbf{L}}\mathbf{1}}{(n-1)(n-2)} - \frac{2}{n-2}\mathbf{1}^\top\tilde{\mathbf{K}}\tilde{\mathbf{L}}\mathbf{1}\right)$$

Nguyen et al., "Do Wide and Deep Networks Learn the Same Things?", ICLR 2021



How does the Modified LP+FT Protocol Compare?

Generalization vs Safety Trade-off

WHAT HAPPENS IF WE TAKE INTO ACCOUNT MODEL SAFETY TO OBTAIN A HOLISTIC EVALUATION?

calibration

corruptions

anomaly rejection

Protocol	ID Test ACC	OOD ACC	mCA	CE (RMSE)	Anomaly AUROC
LP	91.39	81.94	69.1	0.171	62.1
FT	95.58	80.34	74.7	0.137	**99.1**
LP+FT	94.50	86.57	69.1	0.217	64.5
Ours	**96.55**	**92.19**	**81.35**	**0.082**	94.7

Exploring the Design of Adaptation Protocols for Improved Generalization and Machine Learning Safety (Preprint coming soon)

Constructing Knowledge Bridges from Task Distributions to Improve Adaptation

Balancing Transferability and Customization

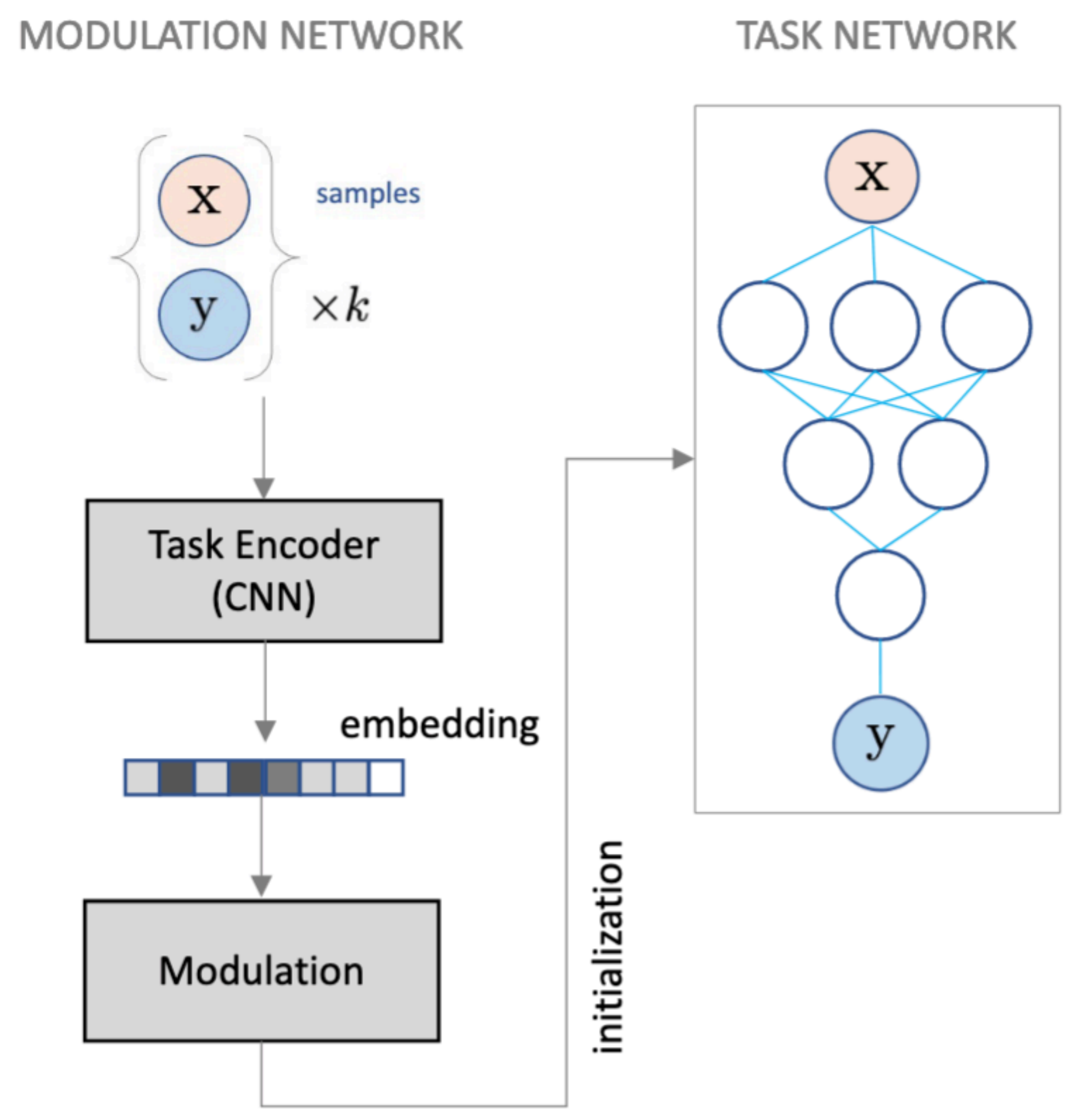
When We Have Access to a Distribution of Tasks, Can We Systematically Improve Adaptation?

Dataset Setting	Few-Shot Task Adaptation	Few-Shot Dataset Generalization
TRAIN (Episodes)	$\mathcal{D}^{tr} = \{(\mathrm{x}_i, \mathrm{y}_i)\}_{i=1}^{\lvert \mathcal{D}^{tr} \rvert}$ $\mathrm{y}_i \in \mathcal{C}^{tr}$	$\mathcal{D}^{tr} = \mathcal{D}_1^{tr} \cup \mathcal{D}_2^{tr} \cdots \cup \mathcal{D}_M^{tr}$ $\mathcal{D}_m^{tr} = \{(\mathrm{x}_i, \mathrm{y}_i)\}_{i=1}^{\lvert \mathcal{D}_m^{tr} \rvert}, \mathrm{y}_i \in \mathcal{C}_m^{tr}$
TEST $\mathcal{T} = (\mathcal{S}_\mathcal{T}, \mathcal{Q}_\mathcal{T})$	$\mathcal{S}_\mathcal{T} = \{(\mathrm{x}_1, \mathrm{y}_1), \cdots, (\mathrm{x}_{kN}, \mathrm{y}_{kN})\}$ $\mathcal{Q}_\mathcal{T} = \{(\mathrm{x}_1^*, \mathrm{y}_1^*), \cdots\}$ $(\mathrm{x}, \mathrm{y}) \in \mathcal{D}^{te}$ $\mathrm{y}, \mathrm{y}^* \in \{1, \cdots, N\} \subset \mathcal{C}^{te}$	$\mathcal{S}_\mathcal{T} = \{(\mathrm{x}_1, \mathrm{y}_1), \cdots, (\mathrm{x}_{kN}, \mathrm{y}_{kN})\}$ $\mathcal{Q}_\mathcal{T} = \{(\mathrm{x}_1^*, \mathrm{y}_1^*), \cdots\}$ $(\mathrm{x}, \mathrm{y}) \in \mathcal{D}_{M+1}^{te}$ $\mathrm{y}, \mathrm{y}^* \in \{1, \cdots, N\} \subset \mathcal{C}_{M+1}^{te}$

Meta Learning

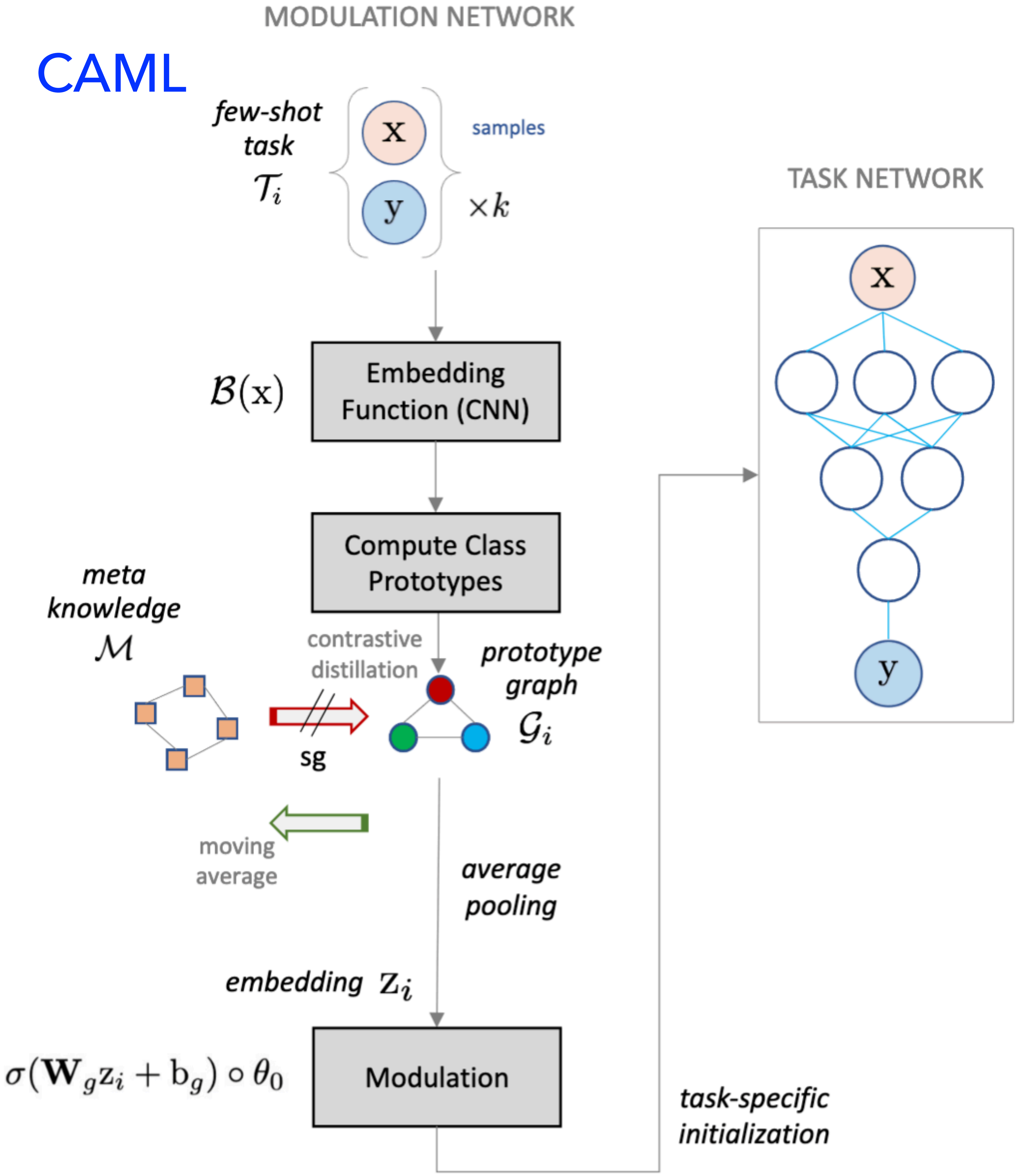
Complex Semantic and Covariate shifts

Task-Aware Modulation is a common technique used to combat task heterogeneity



Knowledge Bridges to Effectively Leverage Historical Experience for Improved OOD Generalization

CAML



- An external knowledge bridge to aggregate prior experience
- A contrastive training strategy to balance between *transferability* and *customization*

vanilla task encoding

task encoding after using prior

$$-\mathbb{E}\left[\log \frac{\exp(\mathrm{sim}(\mathrm{z}_i, \hat{\mathrm{z}}_i))}{\exp(\mathrm{sim}(\mathrm{z}_i, \hat{\mathrm{z}}_i)) + \sum_{m \neq n} \exp(\mathrm{sim}(\mathrm{v}_i^m, \mathrm{v}_i^n))}\right]$$

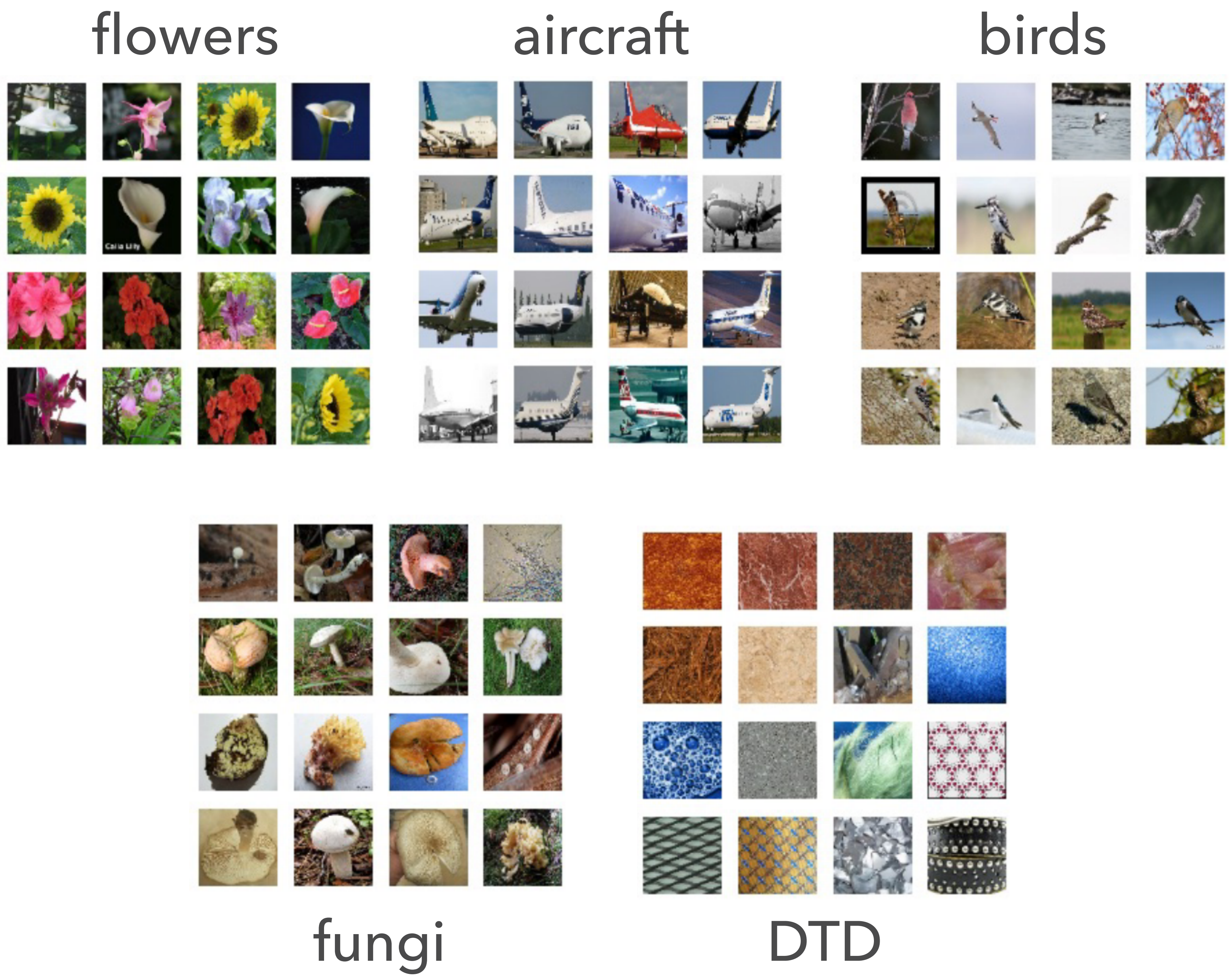
class prototype

- Exponential moving average update of knowledge bridge (similar to BYOL/DINO-style training)

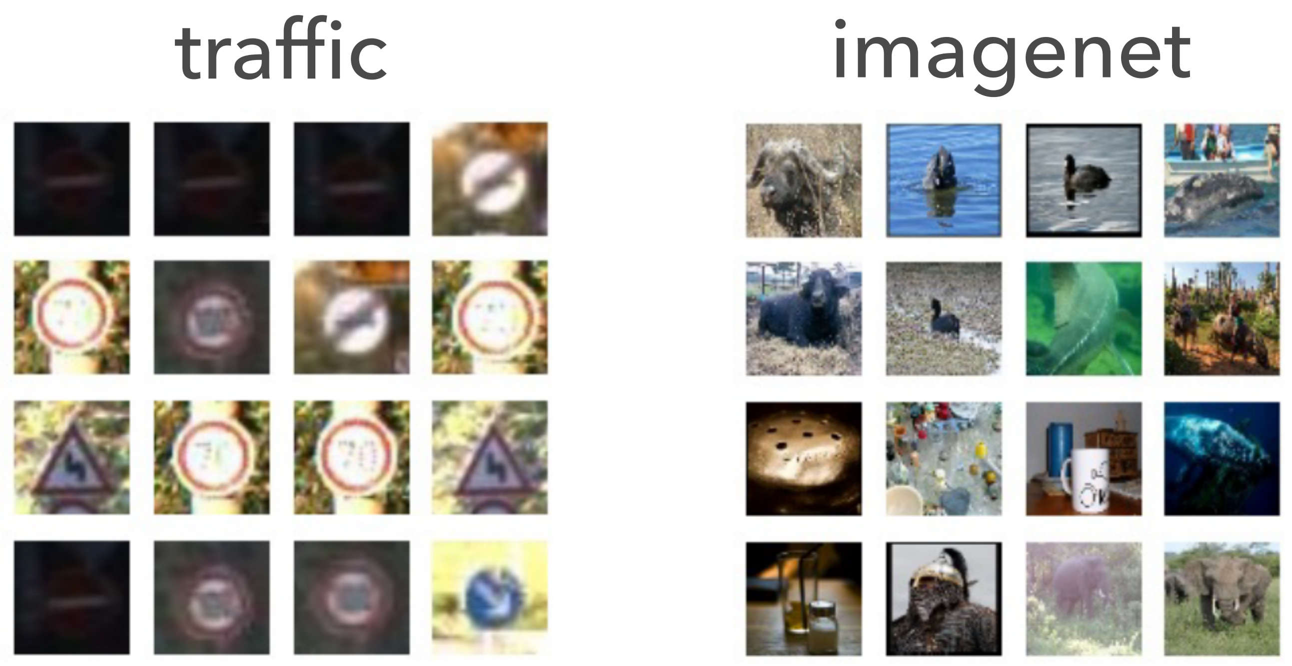
Contrastive Knowledge-Augmented Meta-Learning for Few-Shot Classification, WACV 2023

CAML Produces Highly Robust Meta Learners that Can Handle Challenging Shifts

Observed



Adaptation



MuMO-MAML	47.44	41.91
ARML	52.69	44.63
CAML	**62.18**	**50.39**

Contrastive Knowledge-Augmented Meta-Learning for Few-Shot Classification, WACV 2023

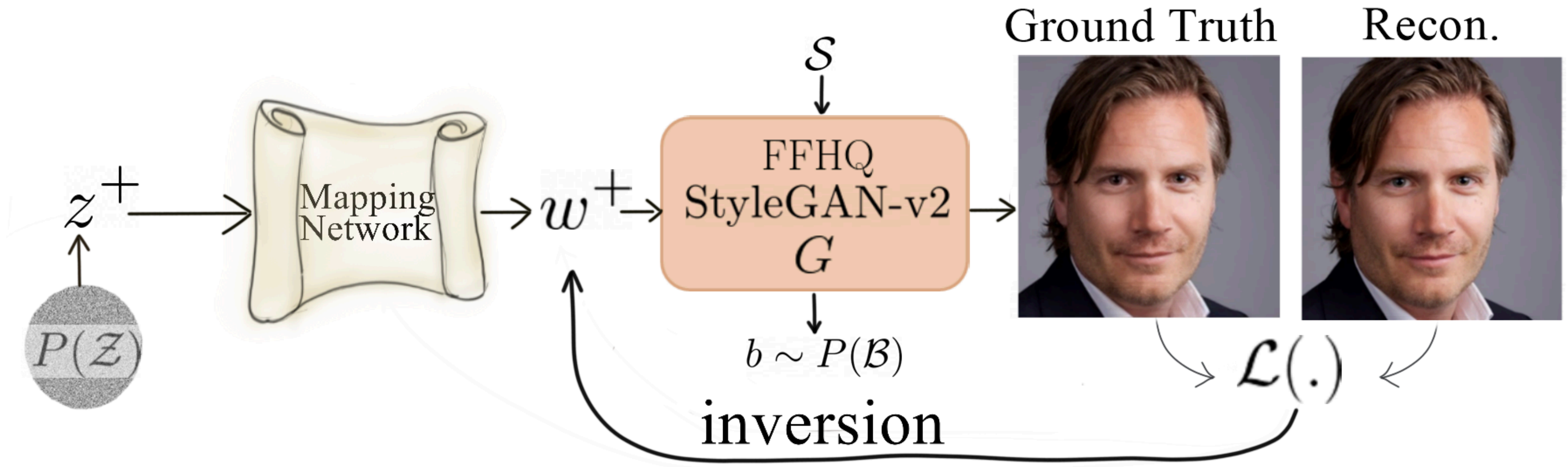
Enabling Reliable OOD Inversion and Manipulation with Generative Models

Effectively Utilizing General-Purpose Priors

Pre-Trained Generative Models Provide Strong Priors to Solve Ill-Posed Inversion Tasks

$$\overset{\text{estimate}}{\hat{\mathrm{x}}} = \mathcal{L}(\overset{\text{observation}}{\mathrm{y}}, \underset{\text{corruption process}}{\mathbf{F}}(\mathrm{x})) + \lambda \underset{\text{regularizer}}{\mathrm{R}_{\mathcal{M}}}(\mathrm{x})$$

Corruption process: identity transformation



- Projected Gradient Descent
- Intermediate Layer Optimization
- I2S, I2S++
- IDInvert
- StyleRig
- StyleFlow
- …

What Happens When We Attempt to Recover an OOD Image using this Approach?

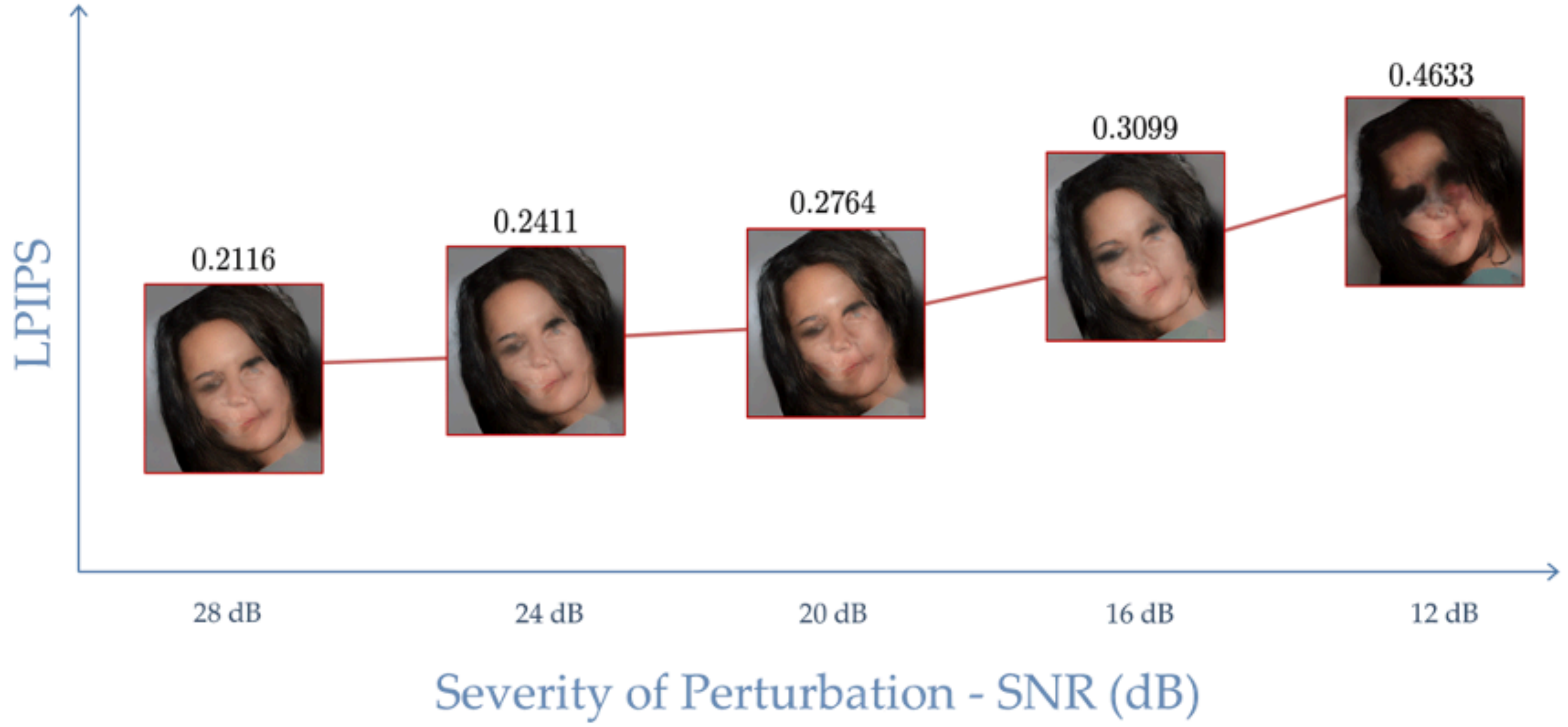
GAN Inversion using ILO

Rotation

Translation

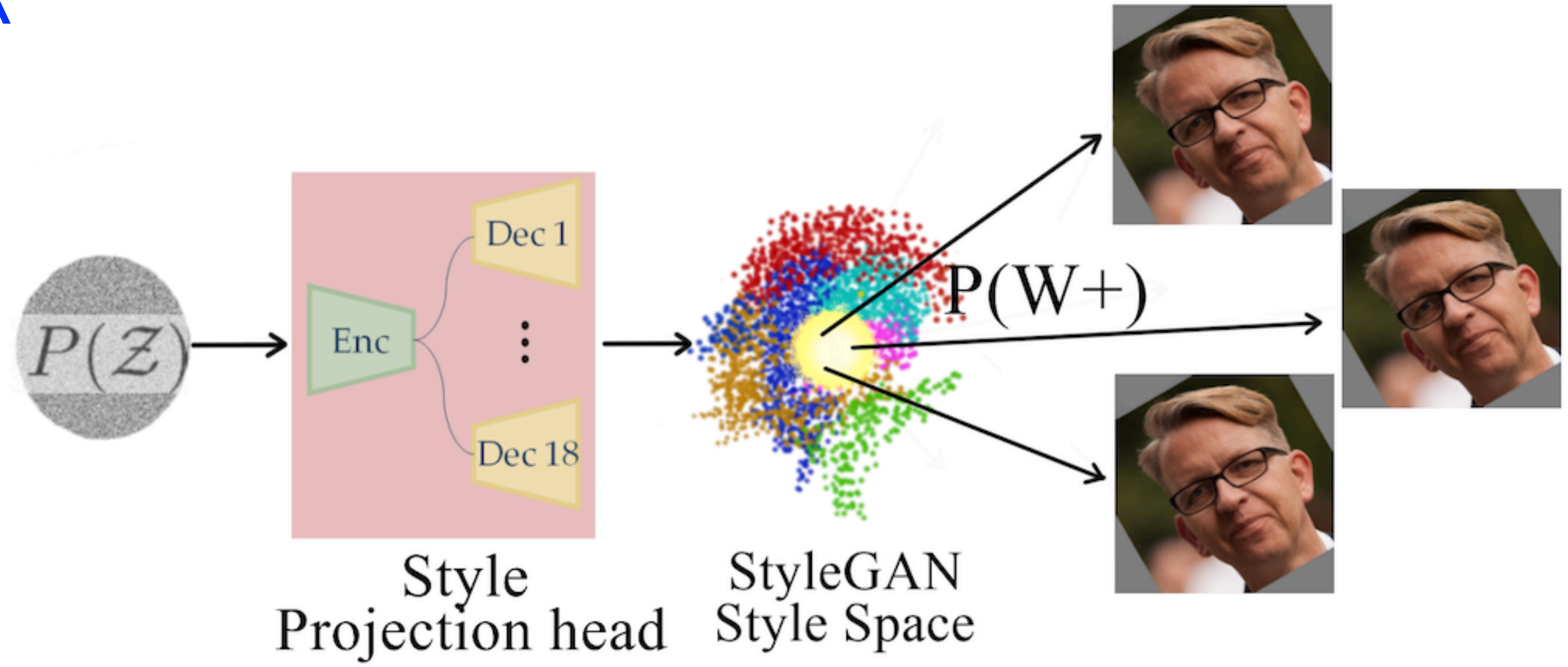
Zoom

- Non-robust nature of W+ optimization
- Lack of priors in W+ to regularize the inversion



By Expressing Uncertainties in the Style Space, We Can Implicitly Impose a Vicinal Regularization

SPHInX



- Learn a distributional mapping from P(Z) to P(W+), such that any realization recovers the observation
- Projection head that decouples the different style latent spaces
- Produces solutions that are locally robust

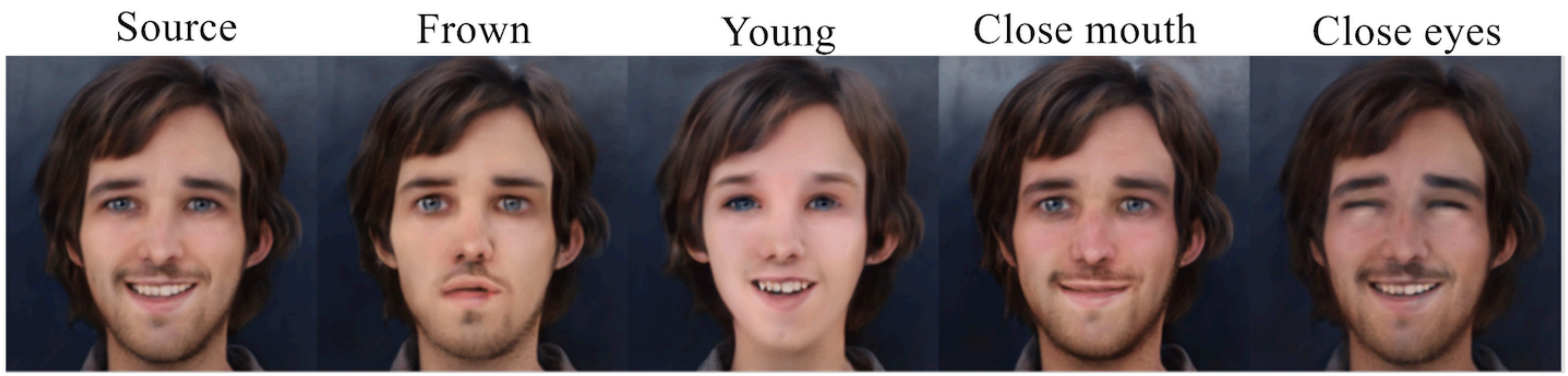
Improved StyleGAN-v2 based Inversion for Out-of-Distribution Images, ICML 2022

SPHInX Consistently Leads to Higher Fidelity Inversion under Challenging Distribution Shifts

Method	Translation				Rotation			Scaling			
	0	50	100	150	10	20	30	0.75	0.875	1.125	1.25
Image2StyleGAN	25.63	25.06	24.53	23.92	25.76	24.65	23.87	25.82	25.25	26.17	26.27
P-norm+	21.79	20.94	19.78	18.54	20.70	18.91	17.93	21.53	19.41	22.07	21.85
StyleGAN2 Inv.	18.73	18.29	17.31	16.71	17.95	17.22	16.02	18.65	18.43	19.12	19.43
PSP	20.54	19.03	17.59	16.50	19.14	17.78	16.99	19.02	17.78	20.63	20.15
BDInvert	<u>26.47</u>	<u>26.30</u>	<u>26.37</u>	<u>26.43</u>	<u>26.48</u>	<u>26.49</u>	<u>26.33</u>	<u>26.44</u>	<u>26.28</u>	<u>26.98</u>	<u>27.26</u>
SPHInX	**29.68**	**29.31**	**28.96**	**28.81**	**29.12**	**28.72**	**28.59**	**28.62**	**29.07**	**29.22**	**28.71**

Improved StyleGAN-v2 based Inversion for Out-of-Distribution Images, ICML 2022

Can Attribute Directions from the Original GAN used for OOD Images?



Semantic editing of cartoon images

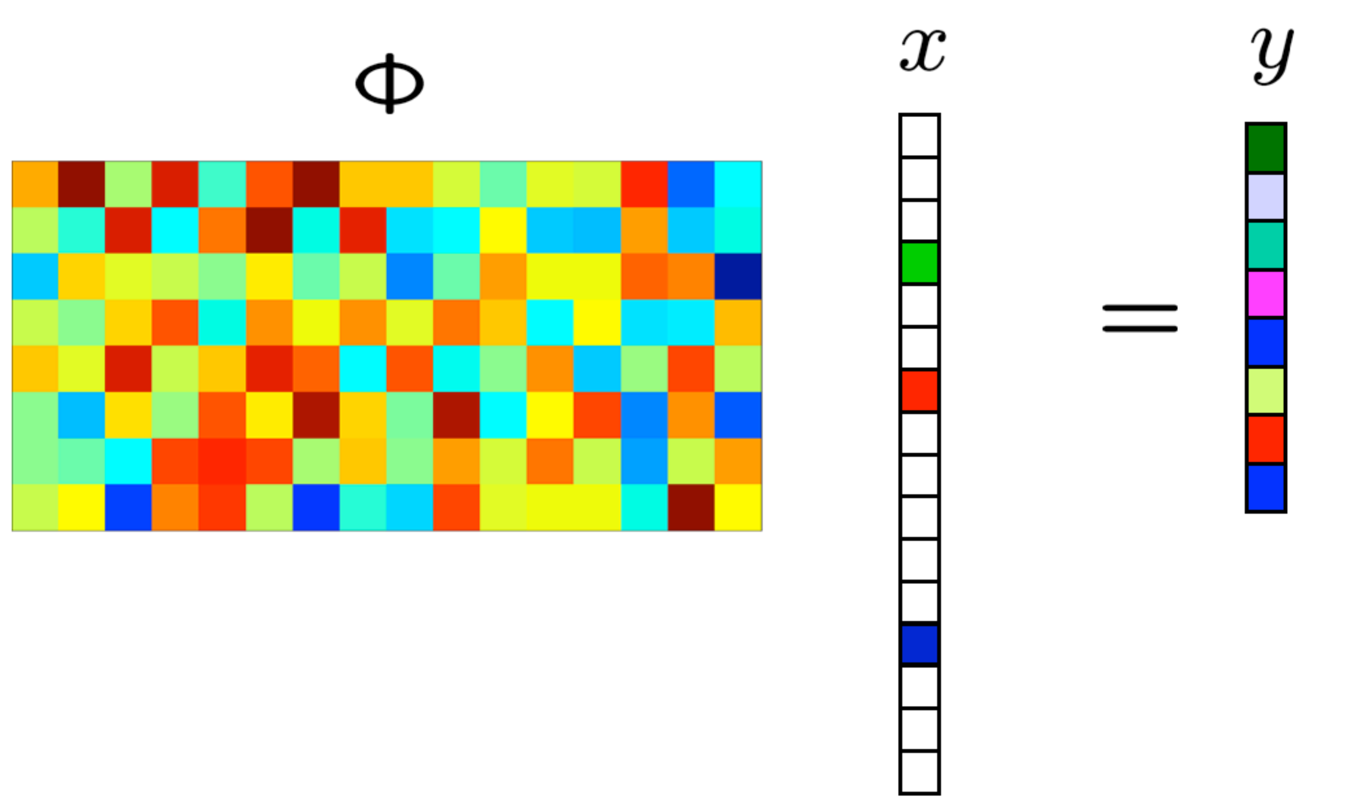
attribute direction

$$\mathbf{w}^{+}_{\text{edit}} = \mathbf{w}^{+} + \alpha \mathbf{v}$$

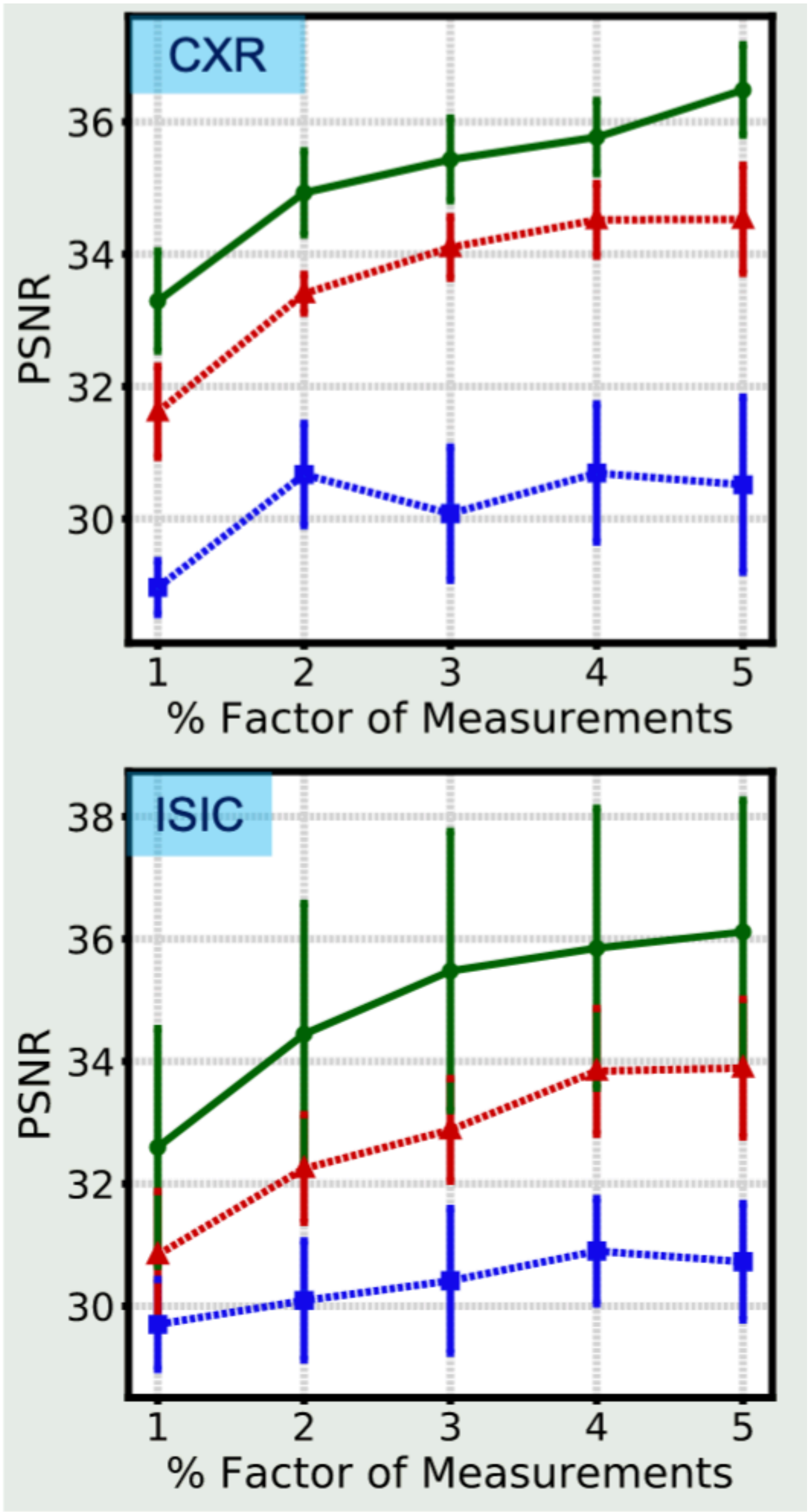
Improved StyleGAN-v2 based Inversion for Out-of-Distribution Images, ICML 2022

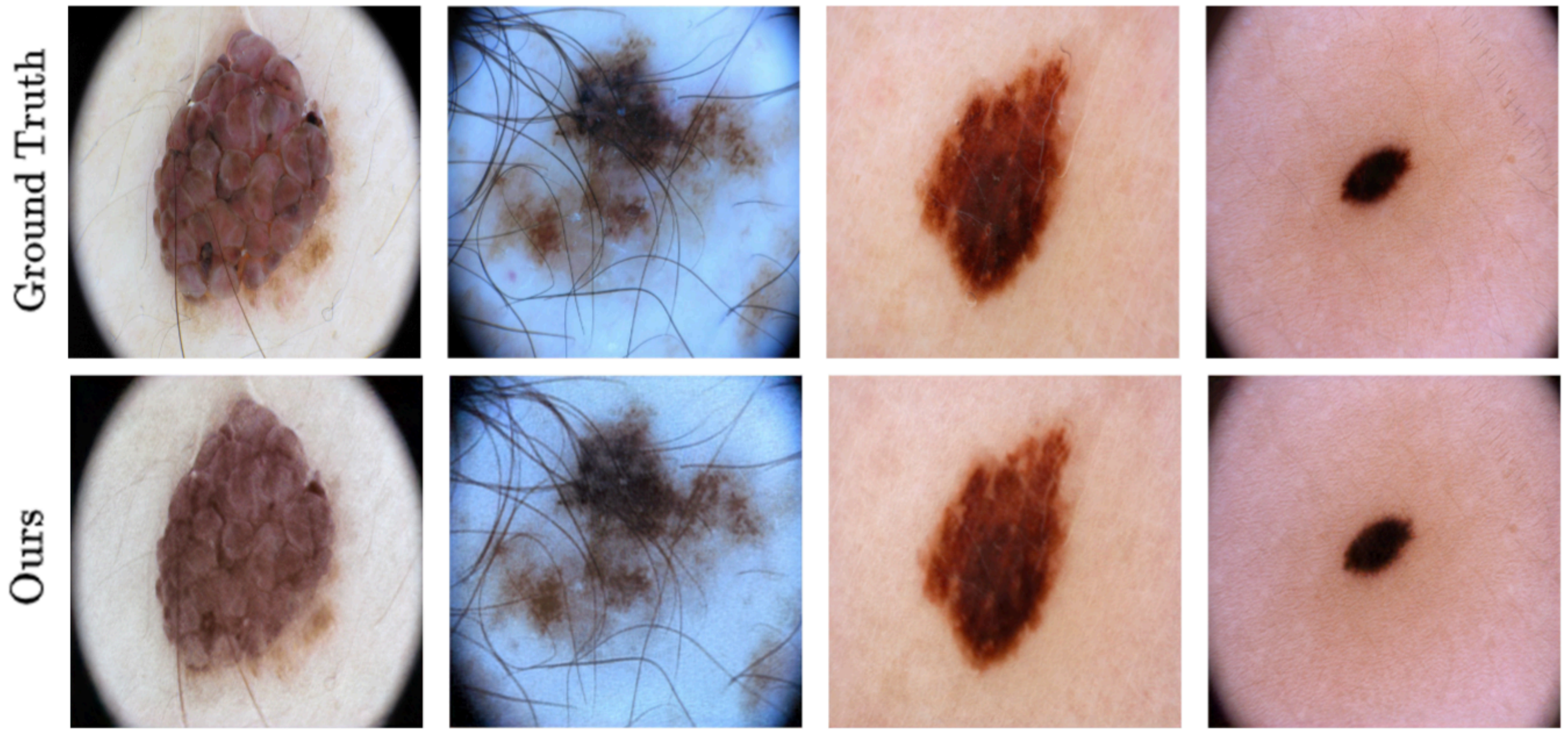
SPHInX Improves the Fidelity of OOD Inversion Even Under Larger Semantic Discrepancies

Compressive Recovery of Medical Images



Prior: StyleGAN trained on FFHQ faces

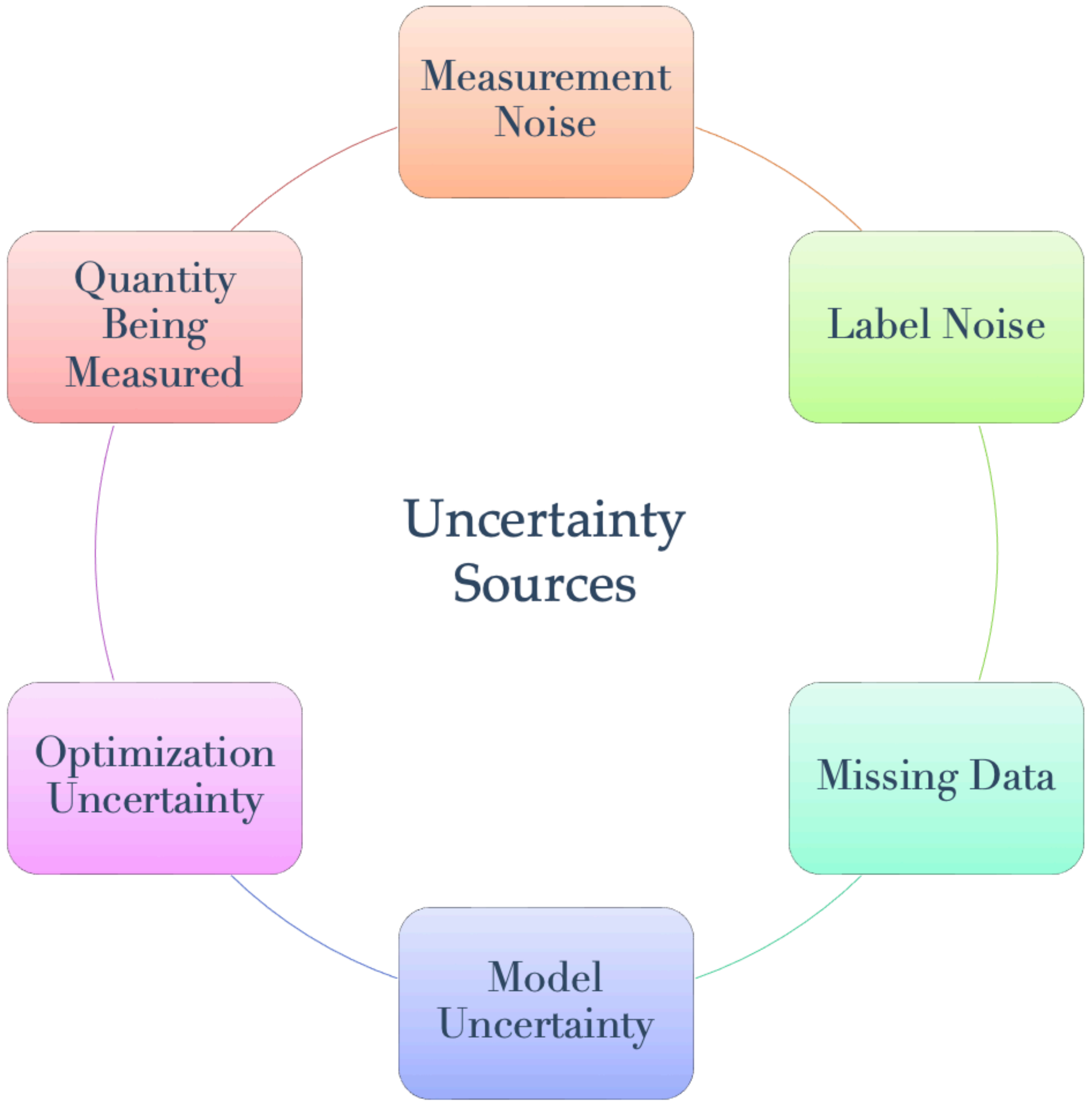




Improved StyleGAN-v2 based Inversion for Out-of-Distribution Images, ICML 2022

Advancing Model Characterization to Promote Safe Models

A Fine-Grained Characterization of the Model is Required to Systematically Promote Safe Models

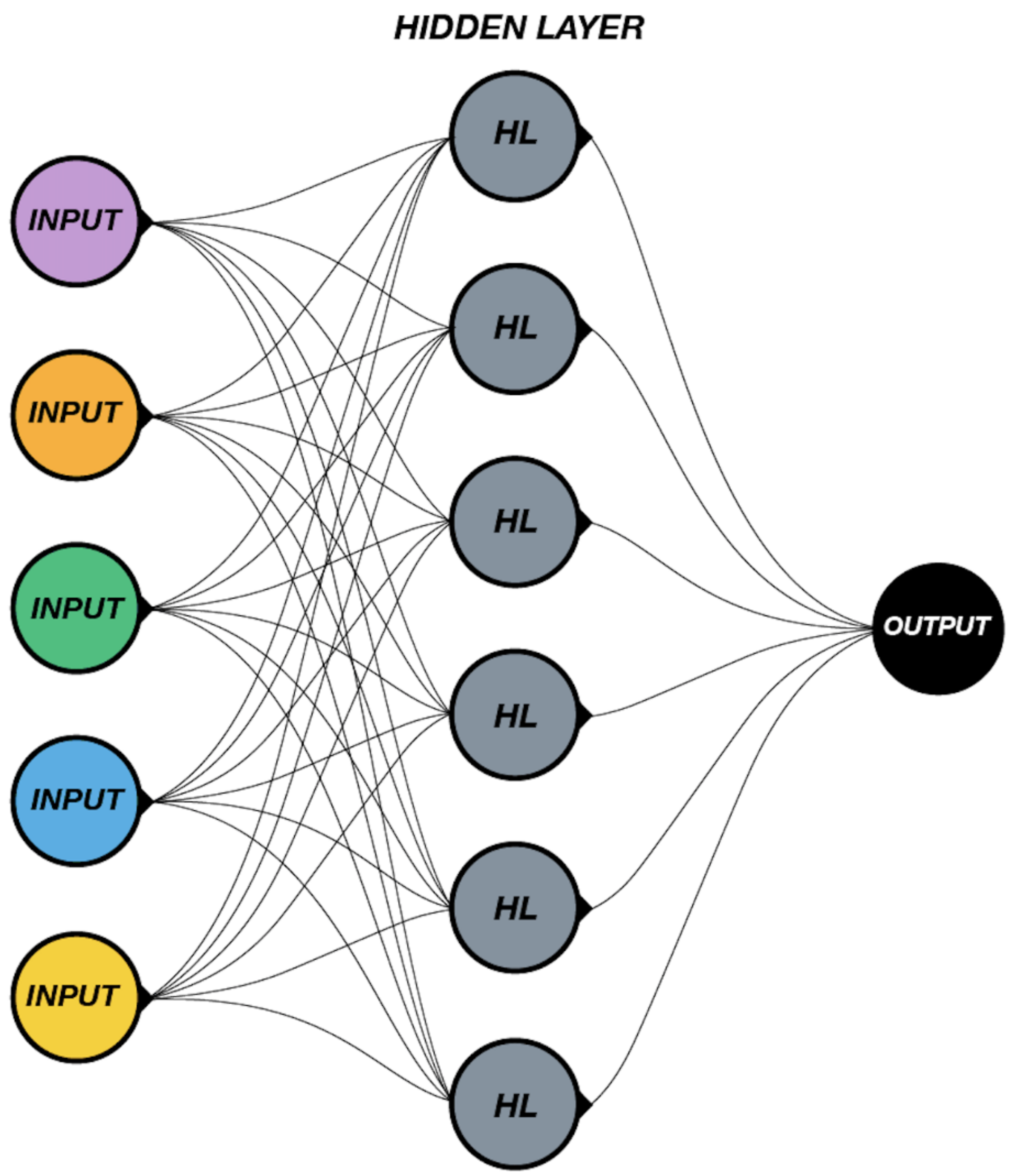


- Modeling different sources of variability and aggregating them to estimate the posterior predictive distribution.

$$p(y|\,\mathbf{x}) = \int \underbrace{P(y|\,\mathbf{x},\boldsymbol{\theta})}_{\text{Aleatoric}}\,\underbrace{p(\boldsymbol{\theta}\,|\mathbb{D})}_{\text{Epistemic}}\,d\boldsymbol{\theta}$$

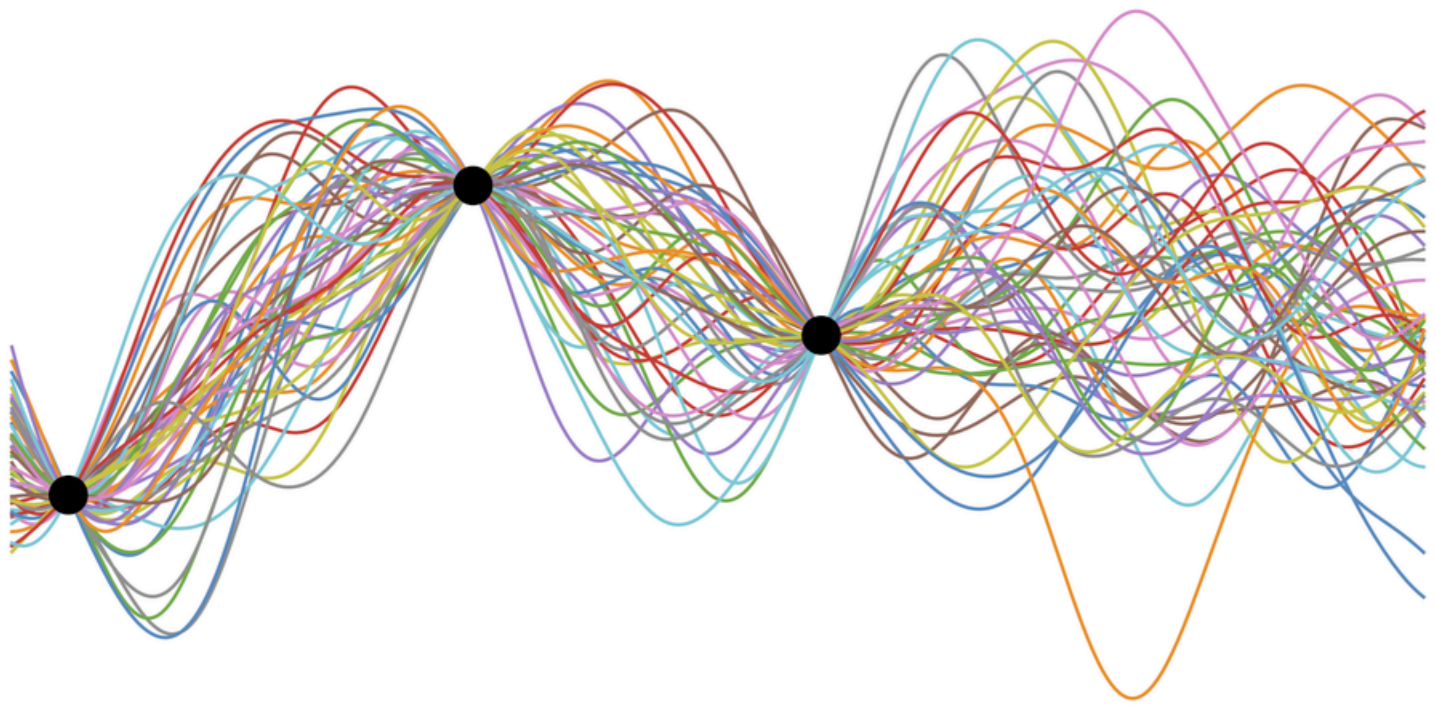
Aleatoric Uncertainties – What You Cannot Know!

- Uncertainties arising from data, that are "irreducible" even with infinite samples.
- Sometimes, can be resolved by leveraging additional features or views of the data.



Epistemic Uncertainties – What You Do Not Know!

- Given finite training data, many models can fit the same data well.
- Variability in the hypotheses can be interpreted as *model uncertainties.*
- “Reducible” and vanishes in the limit of infinite data.

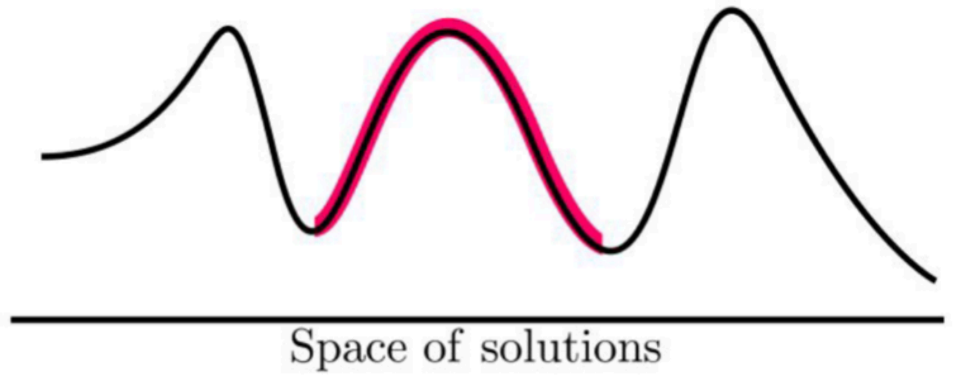


A Gaussian Process

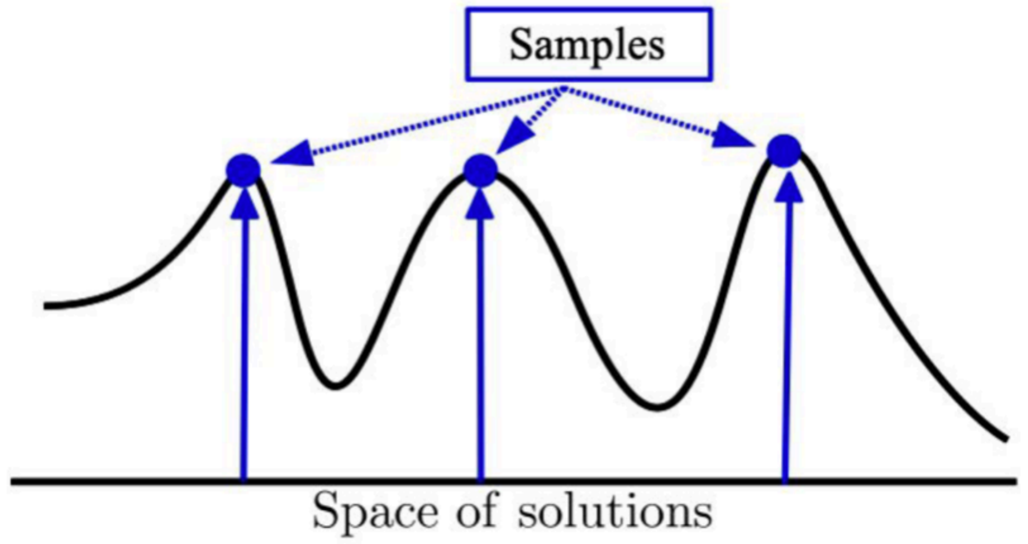
Probabilistic Approach

$$\frac{1}{S}\sum_{s=1}^{S} p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}^{(s)})$$

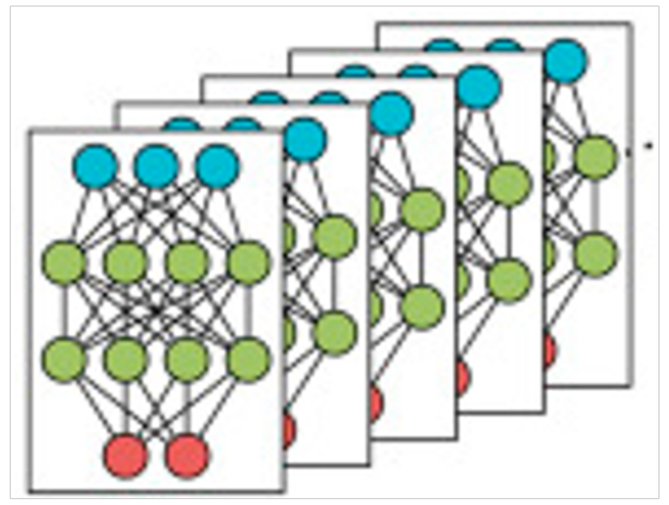
Variational



Sampling



Ensembling



A Motivating Experiment

Consider a training dataset: $\mathcal{D} = \{(x_i, y_i)\}$, where $x_i, y_i \in \mathbb{R}^d$

Now, construct these biased datasets – and fits deep networks to each of them

$$\mathcal{D}_{c_0} = \{(x_i + c_0, y_i)\}, \text{for a fixed } c_0 \in \mathbb{R}^d \longrightarrow f_{c_0}$$

$$\mathcal{D}_{c_1} = \{(x_i + c_1, y_i)\}, \longrightarrow f_{c_1}$$

$$\vdots \qquad\qquad \vdots$$

$$\mathcal{D}_{c_k} = \{(x_i + c_k, y_i)\}, \longrightarrow f_{c_k}$$

Anchors

Will the resulting deep networks be the same?

Anchoring: A New Principle for Quantifying Epistemic Uncertainties

If we use a shift-invariant kernel to build models, we will learn the same function

$$f_{c_0} = f_{c_1} = \cdots = f_{c_k},$$
$$\text{if } \kappa(x_i, x_j) = \kappa(x_i - c, x_j - c)$$

Interesting observation: The neural tangent kernel induced by a deep network is not shift-invariant.

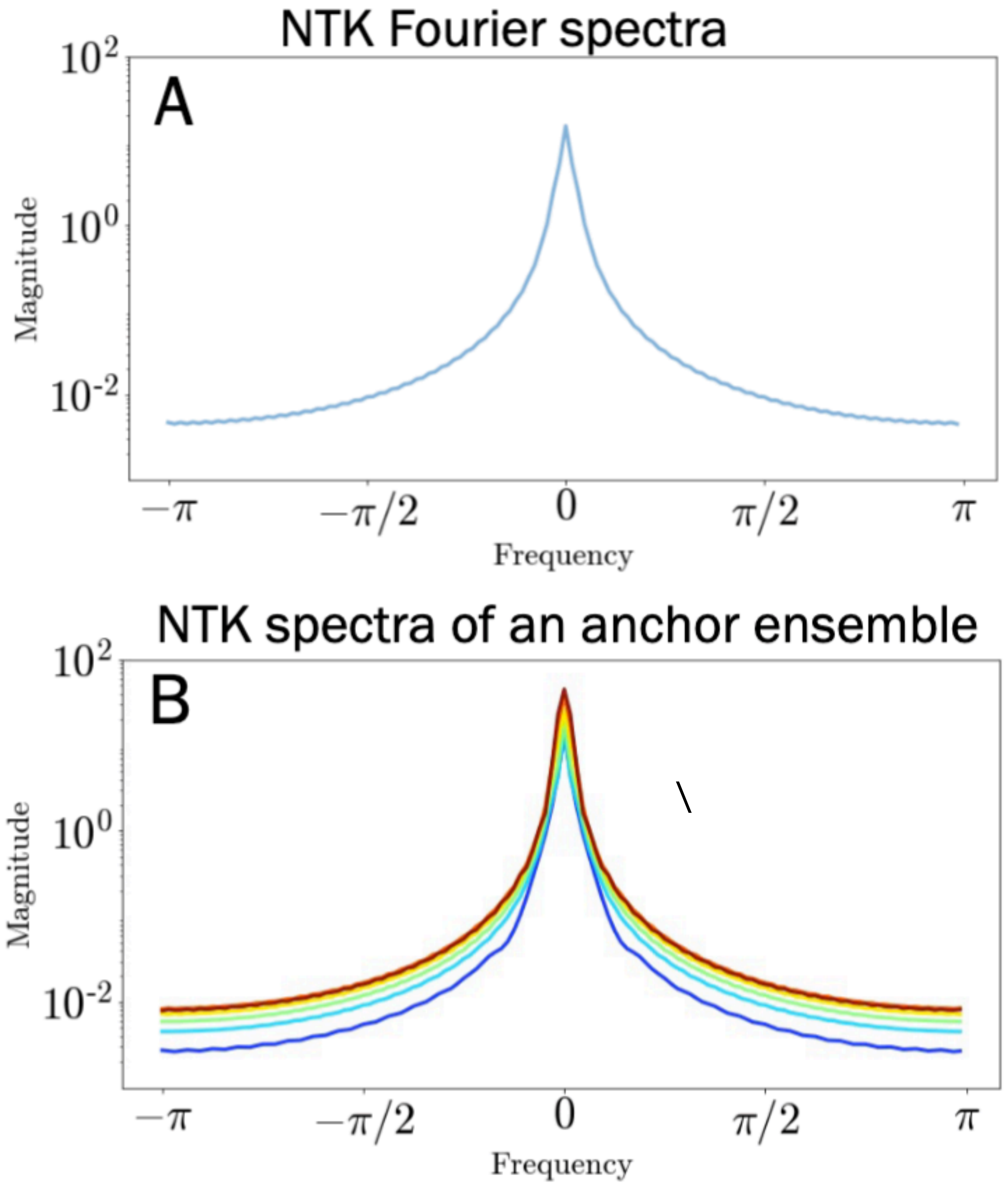
$$\mathbf{K}_{\mathrm{x}_i\mathrm{x}_j} = h_{\mathrm{NTK}}(\mathrm{x}_i^\top \mathrm{x}_j) = \frac{1}{2\pi}\mathrm{x}_i^\top \mathrm{x}_j(\pi - \cos^{-1}(\mathrm{x}_i^\top \mathrm{x}_j))$$

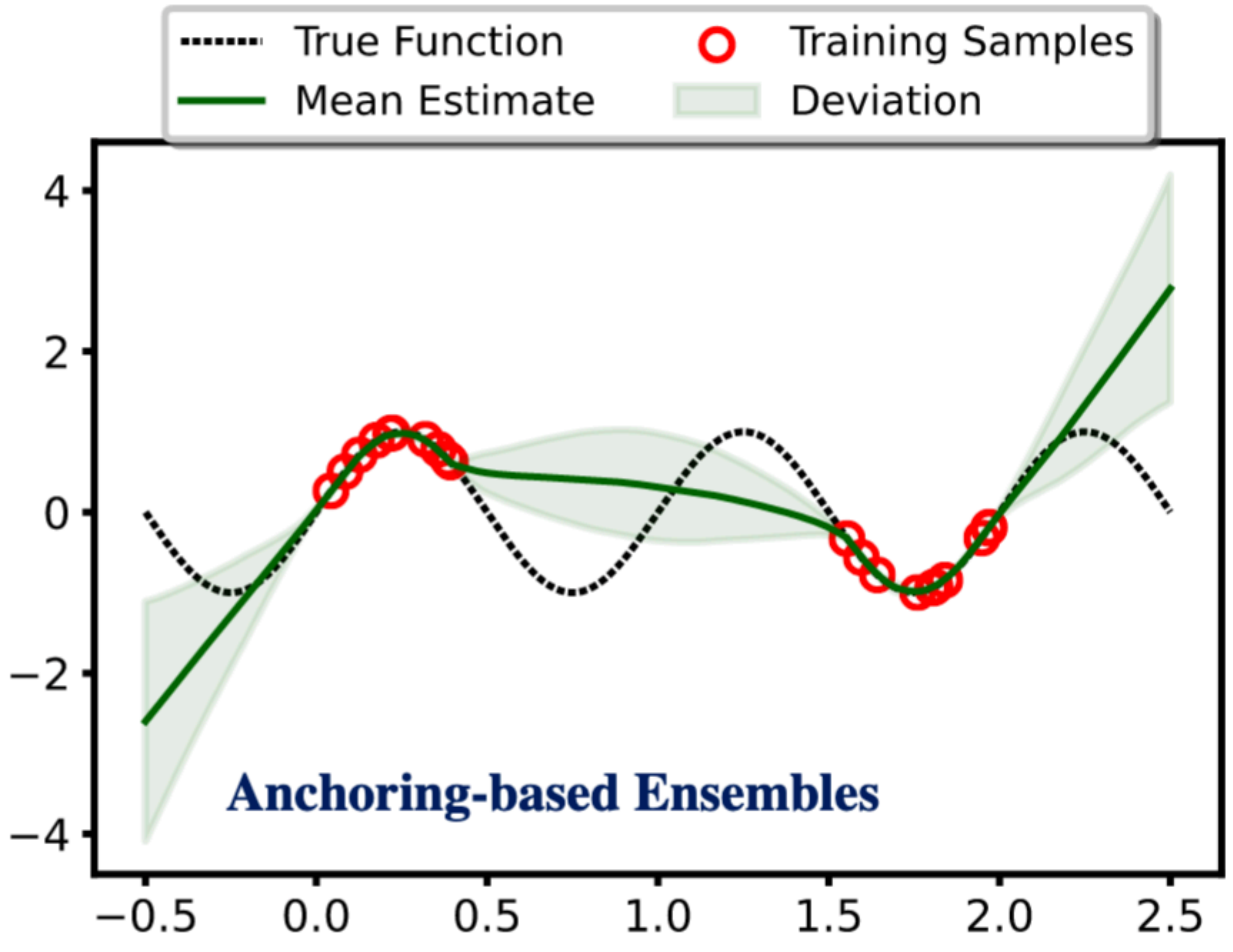
NTK of unperturbed data

$$\mathbf{K}_{(\mathrm{x}_i-\mathrm{c})(\mathrm{x}_j-\mathrm{c})} = \boxed{\mathbf{K}_{\mathrm{x}_i\mathrm{x}_j}} - \boxed{\Gamma_{\mathrm{x}_i,\mathrm{x}_j,\mathrm{c}}}$$

NTK perturbation in terms of c

Ensembling via Stochastic Data Centering





Mean estimate

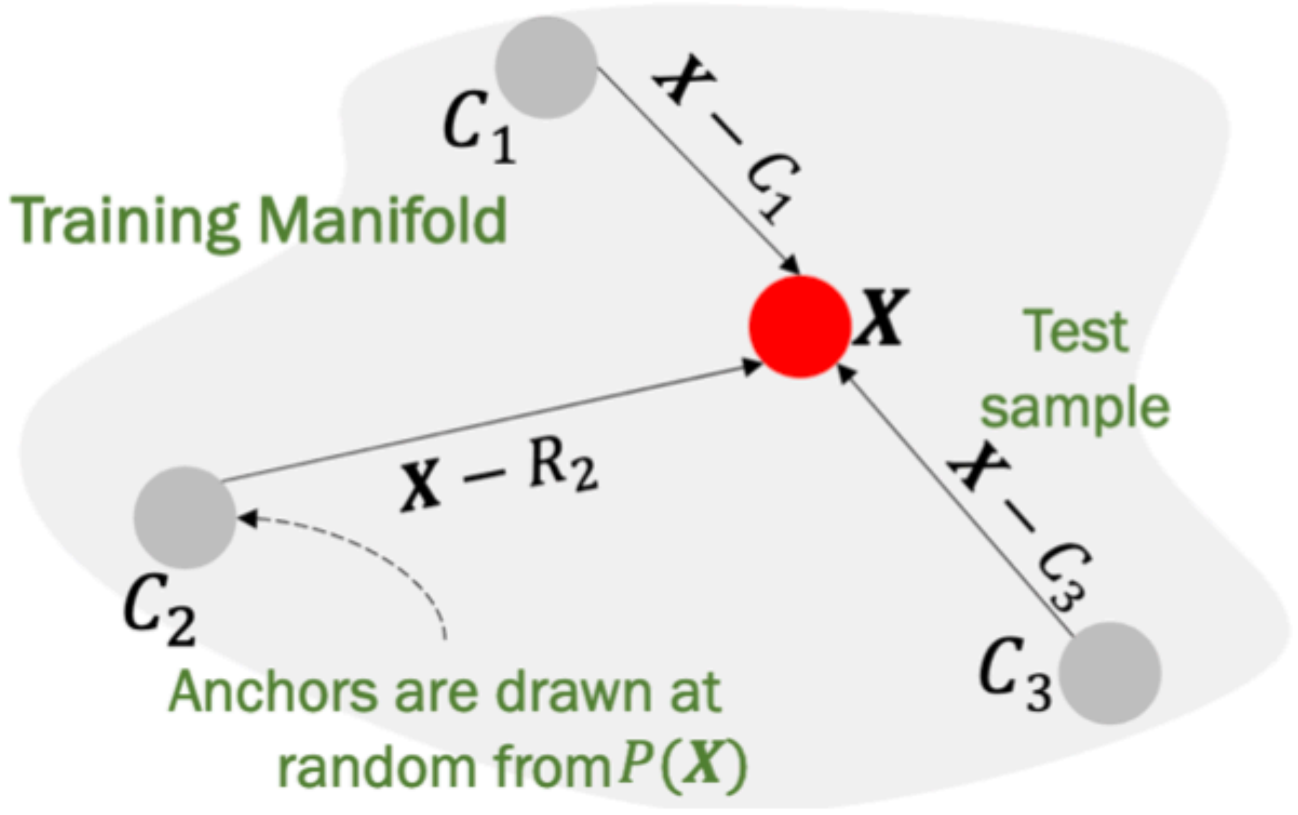
$$\frac{1}{K}\sum_{k=1}^{K} f_\theta(c_k, x - c_k)$$

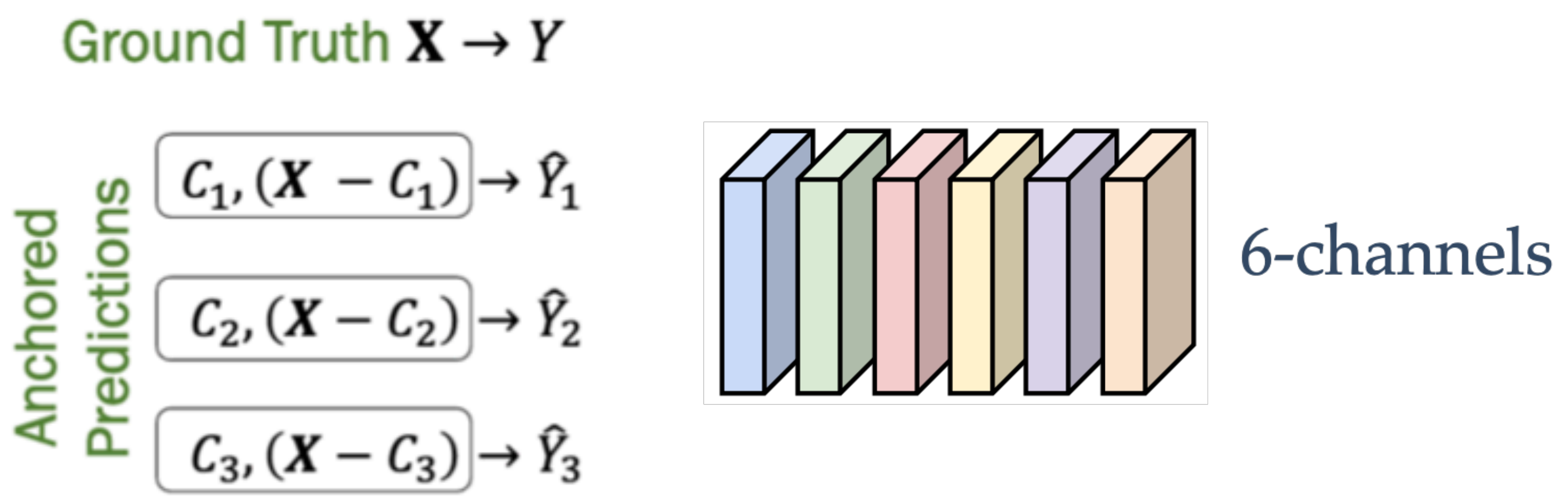
Uncertainty estimate

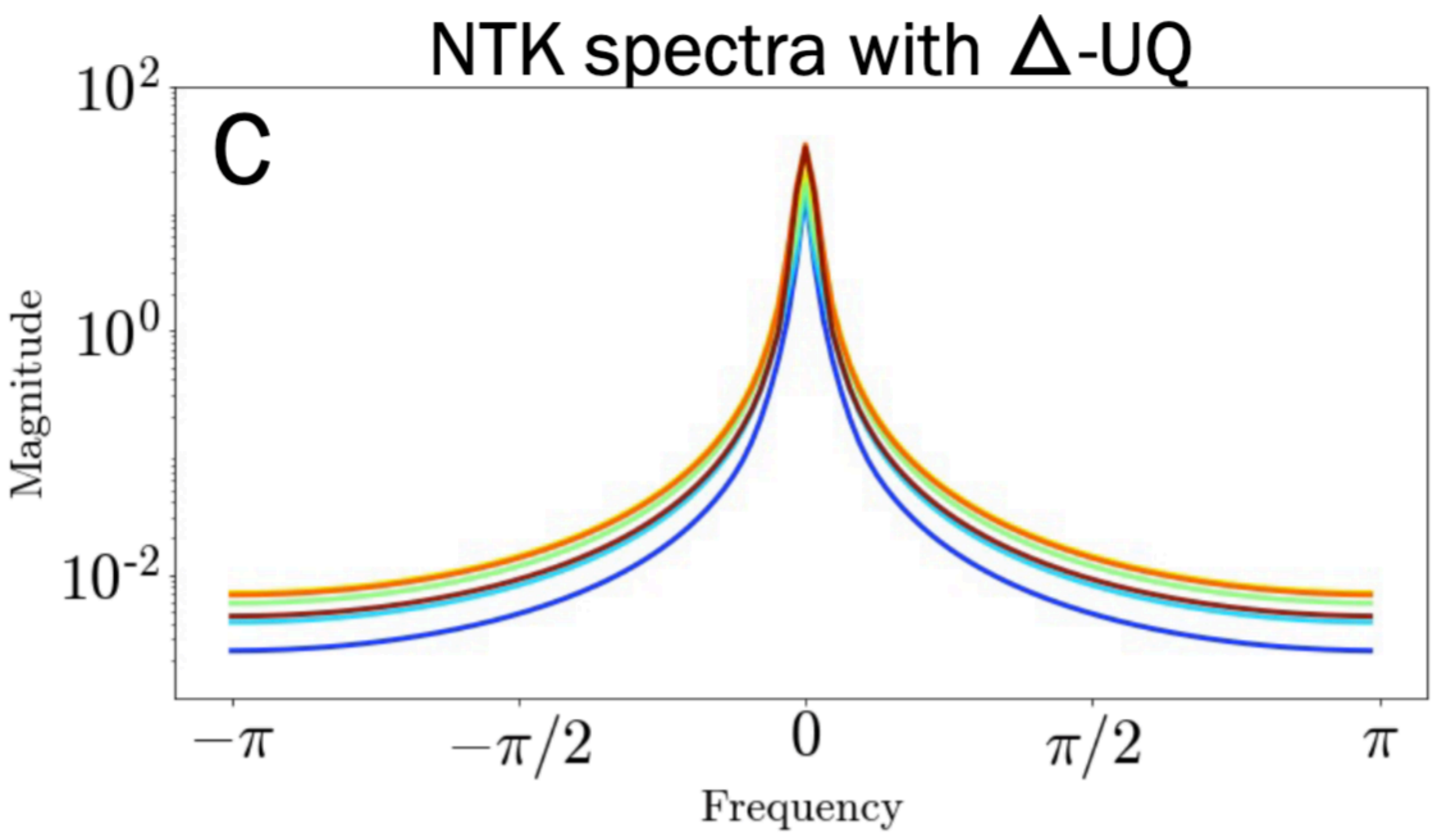
$$\sqrt{\frac{1}{K}\sum_{k=1}^{K} (f_\theta(c_k, x - c_k) - \mu)^2)}$$

Each shift introduces a different bias, resulting in a different hypothesis!

Δ-UQ: Rolling Anchor Ensembling into a Single Model Training

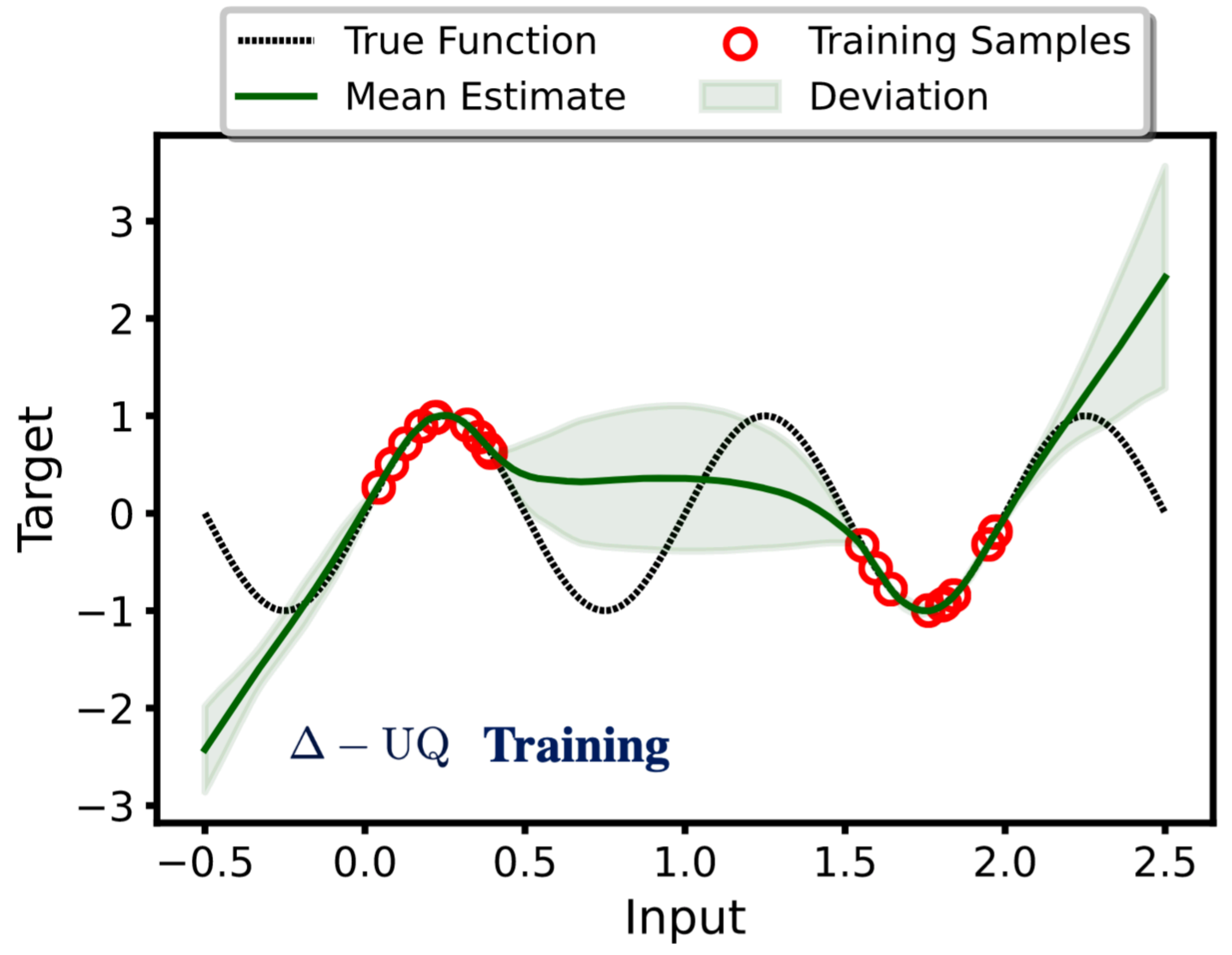






$$\theta^* = \arg\min_{\theta} \mathcal{L}(f(x_c, \theta), y)$$

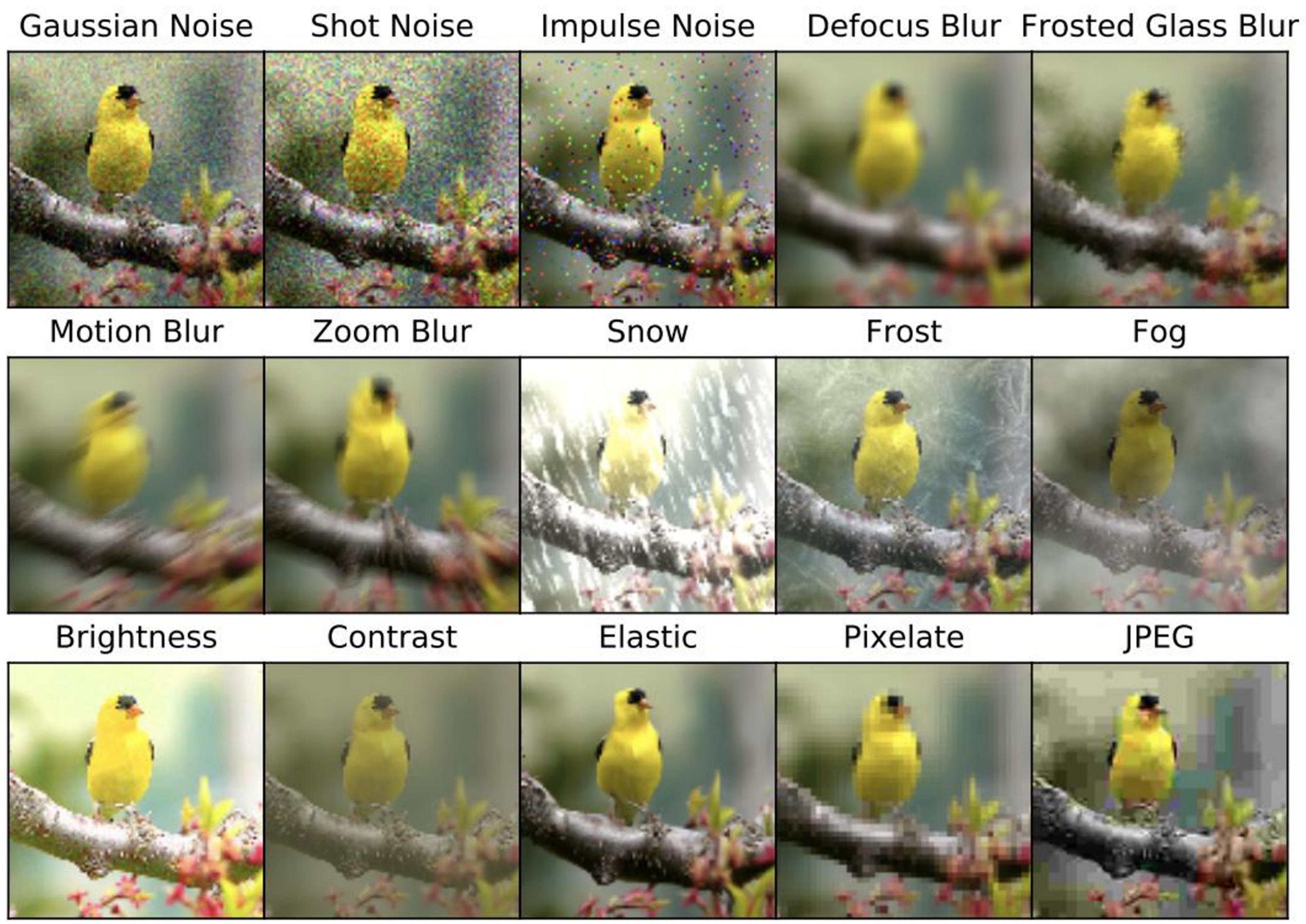
$$\text{where } x_c = (c, x - c) \text{ for } c \sim P(x)$$



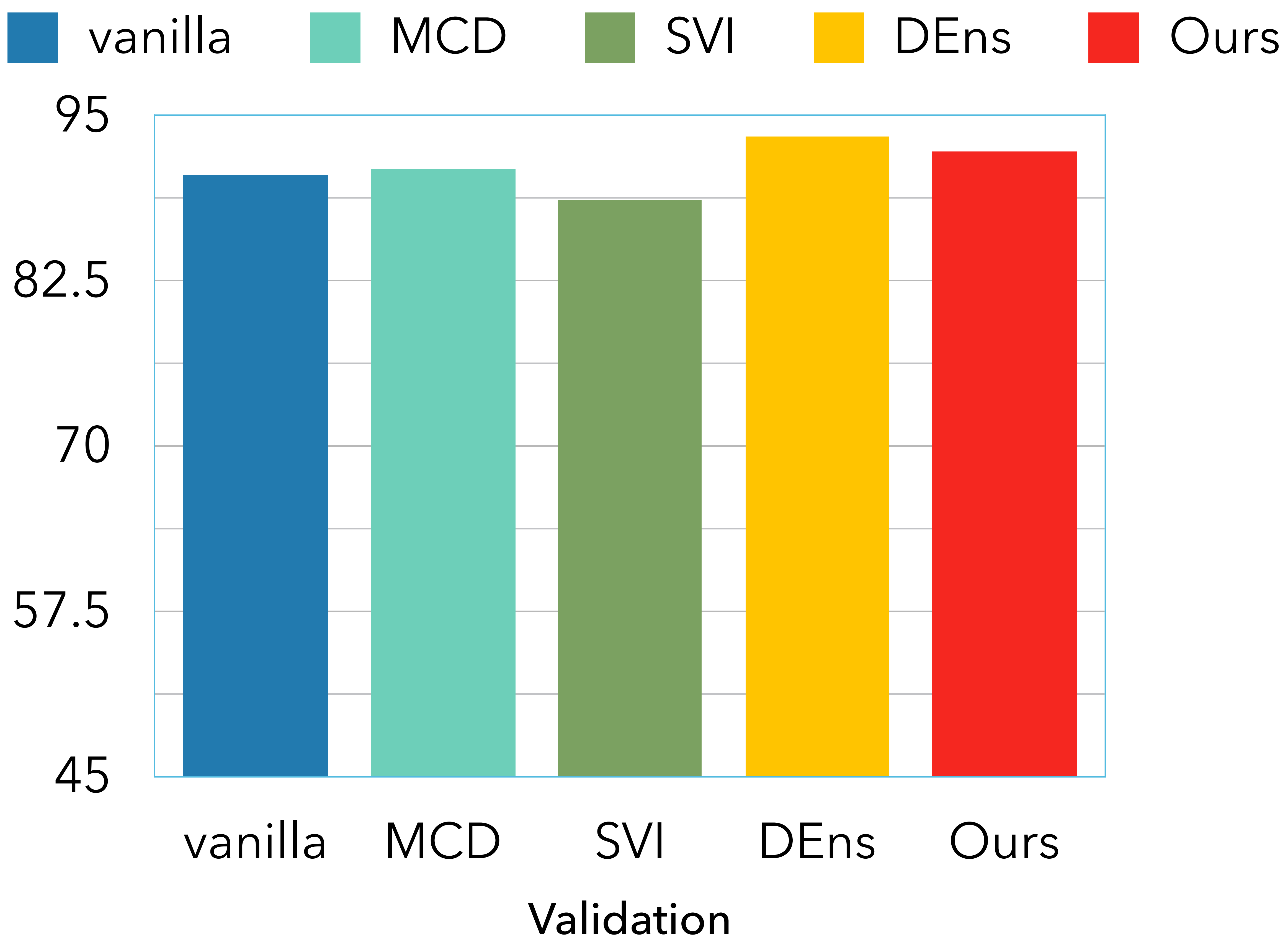
$$f_\Delta(\{c_1, x - c_1\}) = f_\Delta(\{c_2, x - c_2\}) = \cdots = f_\Delta(\{c_k, x - c_k\})$$

Single Model Uncertainty Estimation via Stochastic Data Centering, https://arxiv.org/abs/2207.07235

Δ-UQ Models Do Not Compromise on ID Performance, but Withstands Corruptions Better



CIFAR10 –> CIFAR10-C



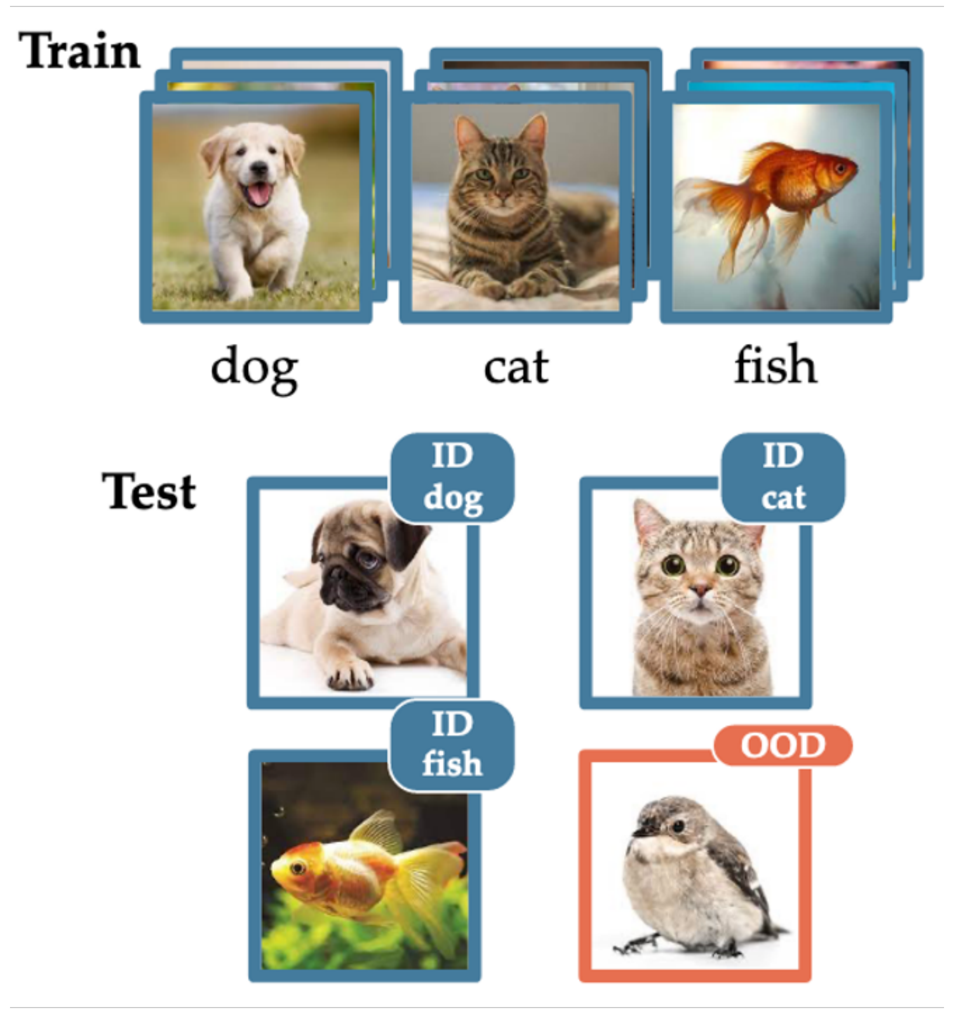
Single Model Uncertainty Estimation via Stochastic Data Centering, https://arxiv.org/abs/2207.07235

Reliable Anomaly Detection: Adaptive Temperature Scaling via Δ-UQ Uncertainties

Anchor Marginalized Prediction (AMP) Score

$$\mathcal{S}(\mathrm{x}) = -\frac{1}{N}\sum_{\text{all classes}} \log(\mathsf{SOFTMAX}(H^c(y|\mathrm{x})))$$

$$H^c(y|\mathrm{x}) = \frac{H(y|\mathrm{x})}{1+\exp(\boldsymbol{\tau}(\mathrm{x}))}$$



Near OOD

	Method	ResNet-34 FPR95 ↓ / AUROC ↑
C10 → C100	ODIN	58.0 / 88.2
	Energy	47.5 / 88.4
	GM	59.8 / 83.6
	Mahal.*	58.4 / 88.2
	AMP (ours)	**43.5 / 90.2**

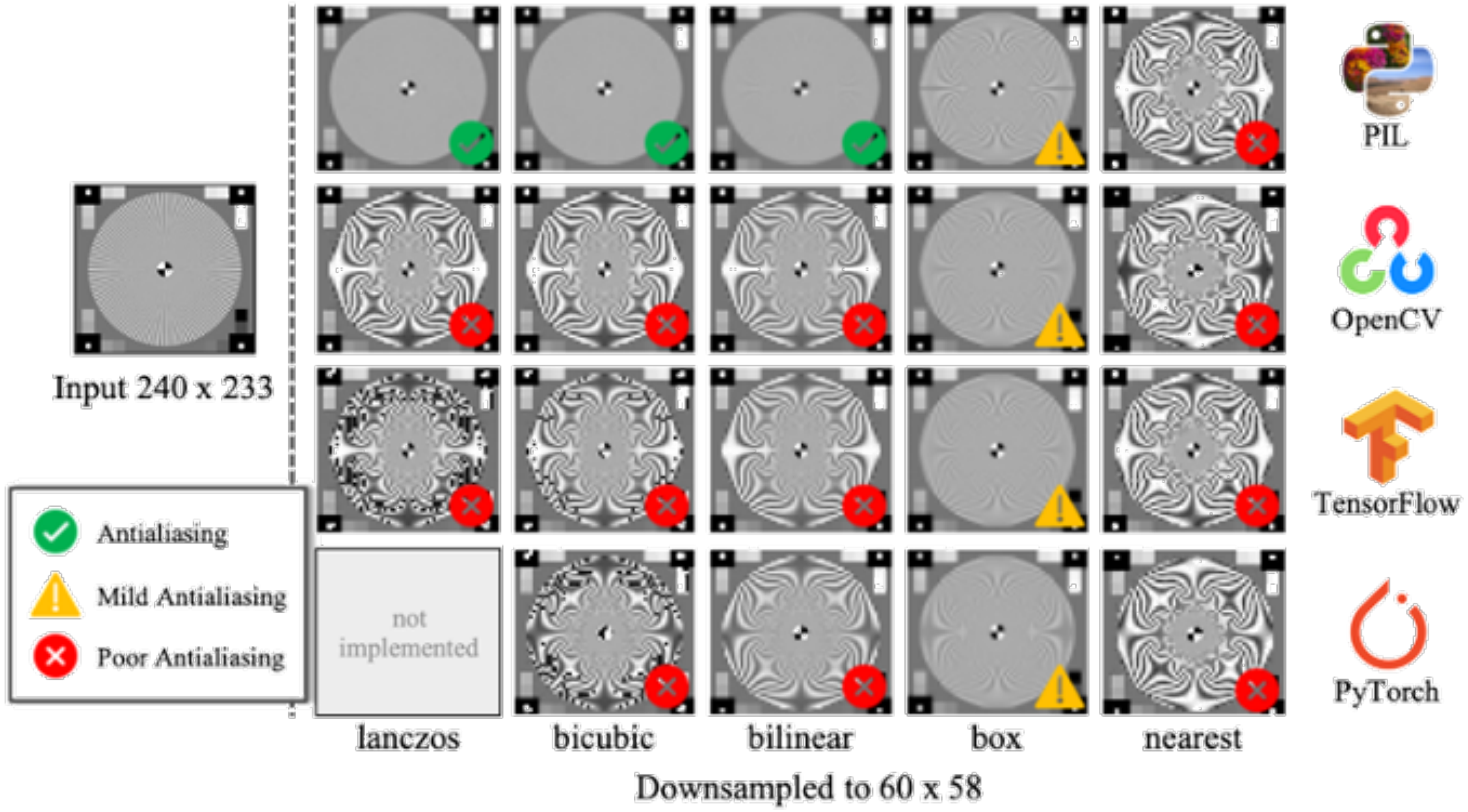
Semantically Coherent OOD (Yang et al., ICCV 2021)

In-distribution	Method	Needs OOD Exposure?	FPR95 ↓	AUROC ↑
CIFAR-100 (ResNet-18)	ODIN (Liang et al., 2017)	✗	81.89	77.98
	Energy (Liu et al., 2020)	✗	81.66	79.31
	OE (Hendrycks et al., 2019)	✓	80.06	78.46
	MCD (Yu and Aizawa, 2019)	✓	85.14	74.82
	UDG (Yang et al., 2021)	✓	75.45	79.63
	AMP	✗	**70.34**	**82.22**

Out of Distribution Detection via Neural Network Anchoring, ACML 2022 (https://arxiv.org/abs/2207.04125v1)

Reliable Anomaly Detection: Adaptive Temperature Scaling via Δ-UQ Uncertainties

And it does not rely on shortcuts to reject anomalous samples!!!



Parmar et al., "On buggy resizing libraries and surprising subtleties in FID calculation", arXiv:2104.11222

FPR / AUROC Metrics

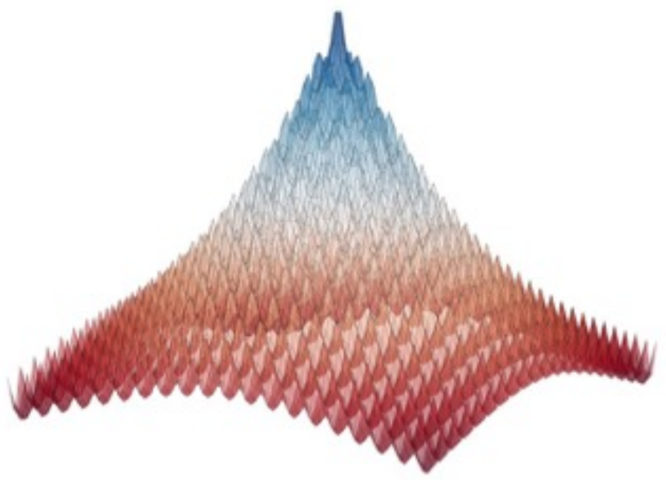
In Distribution	Method	Pillow Resizing from original LSUN				Average
		nearest*	bilinear	bicubic	lanczos	
CIFAR-10 (ResNet-34)	MSP	41.5 / 94.0	47.8 / 91.6	45.5 / 92.2	45.3 / 92.4	46.9 / 92.2
	Energy	28.6 / 98.4	34.5 / 92.9	33.0 / 93.4	32.0 / 93.8	30.4 / 94.1
	GM	**1.8 / 99.2**	46.2 / 90.7	49.0 / 90.6	46.3 / 91.3	25.6 / 94.9
	AMP	7.1 / 98.4	**13.0 / 97.4**	**13.9 / 97.2**	**14.3 / 97.2**	**9.6 / 98.1**

Out of Distribution Detection via Neural Network Anchoring, ACML 2022 (https://arxiv.org/abs/2207.04125v1)

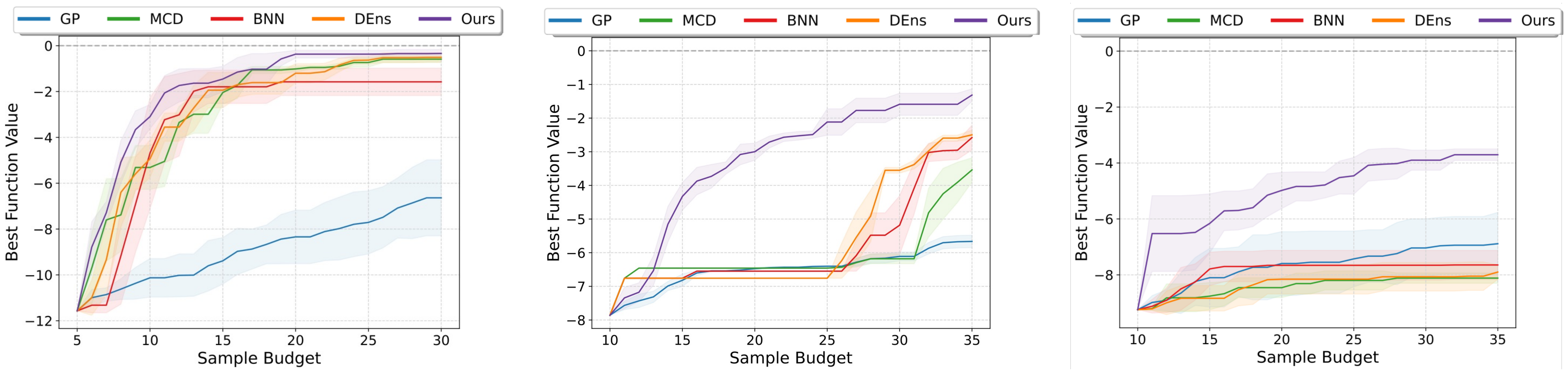
△-UQ Produces Meaningful Estimates even with Limited Data and Leads to Improved Active Learners

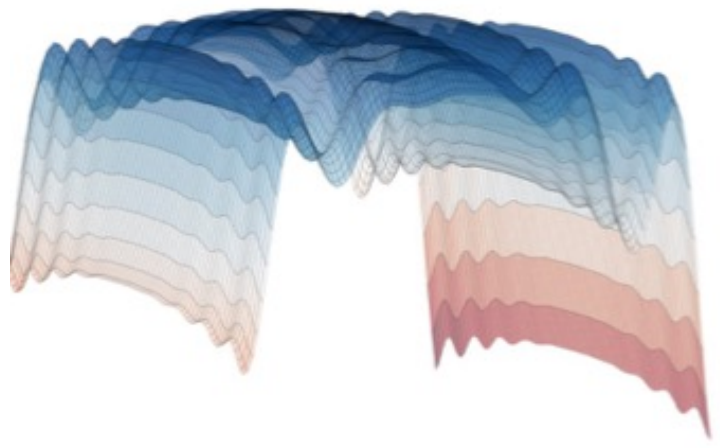
Bayesian Optimization with EI

Single Model Uncertainty Estimation via Stochastic Data Centering, https://arxiv.org/abs/2207.07235

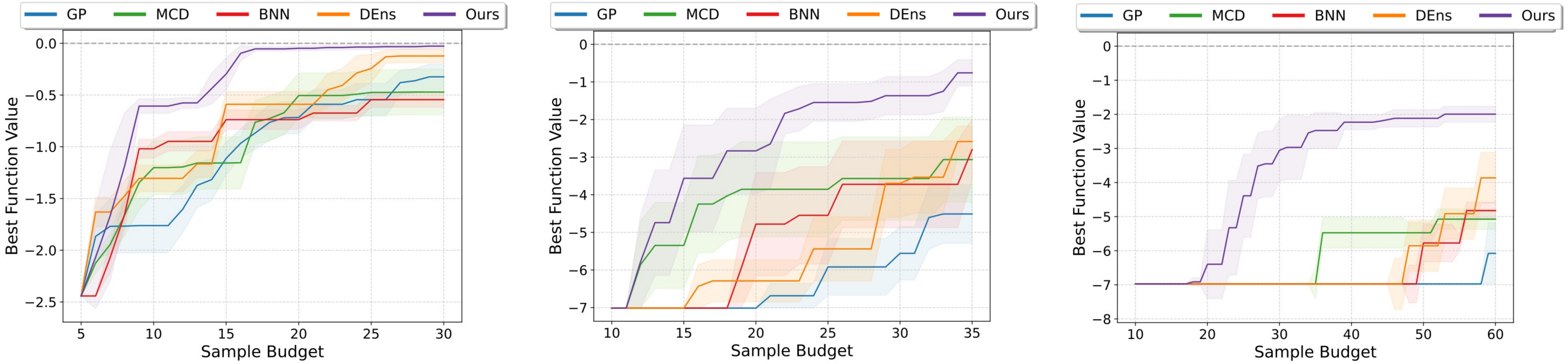


Ackley Function





Levy Function



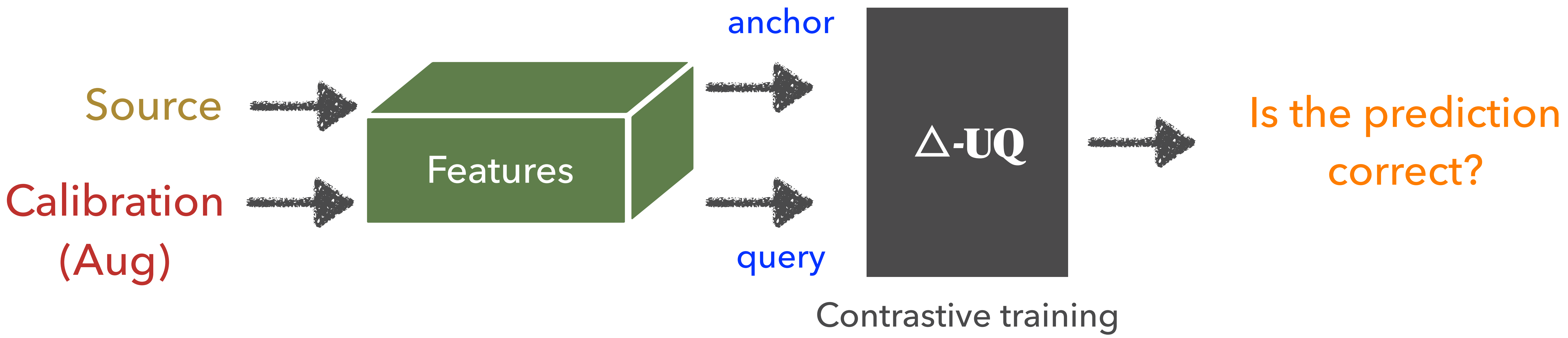
DIM = 2

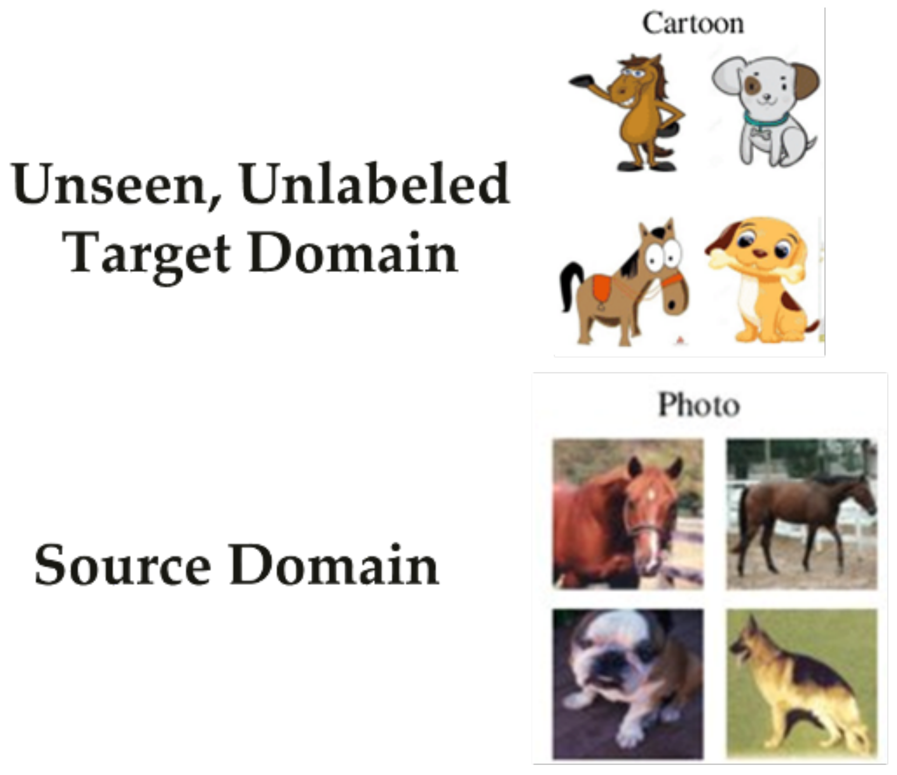
DIM = 4

DIM = 8

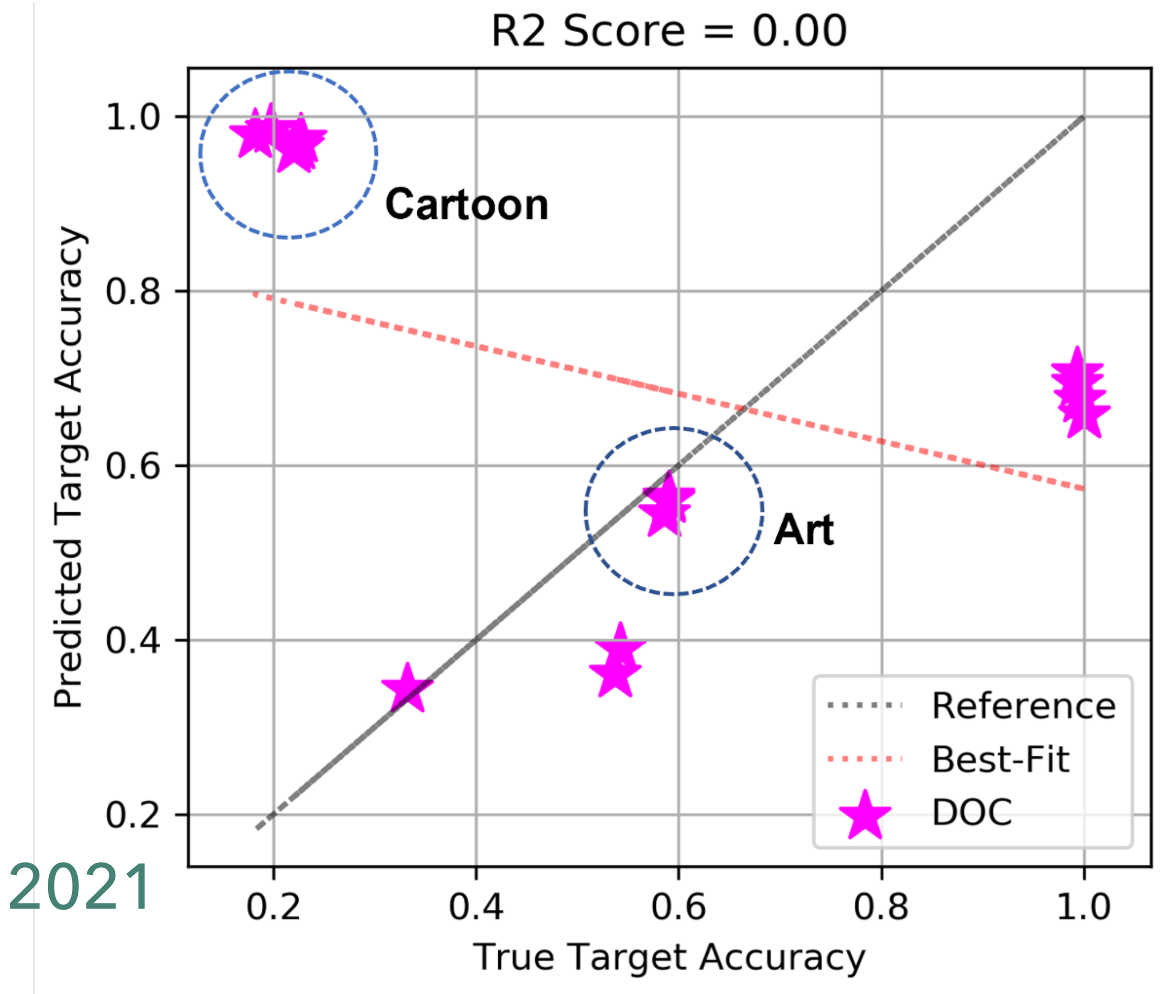
Finally, the Principle of Anchoring can be used to Quantify Representation Uncertainties

Measuring uncertainties in the representation space will provide insights into its generalization

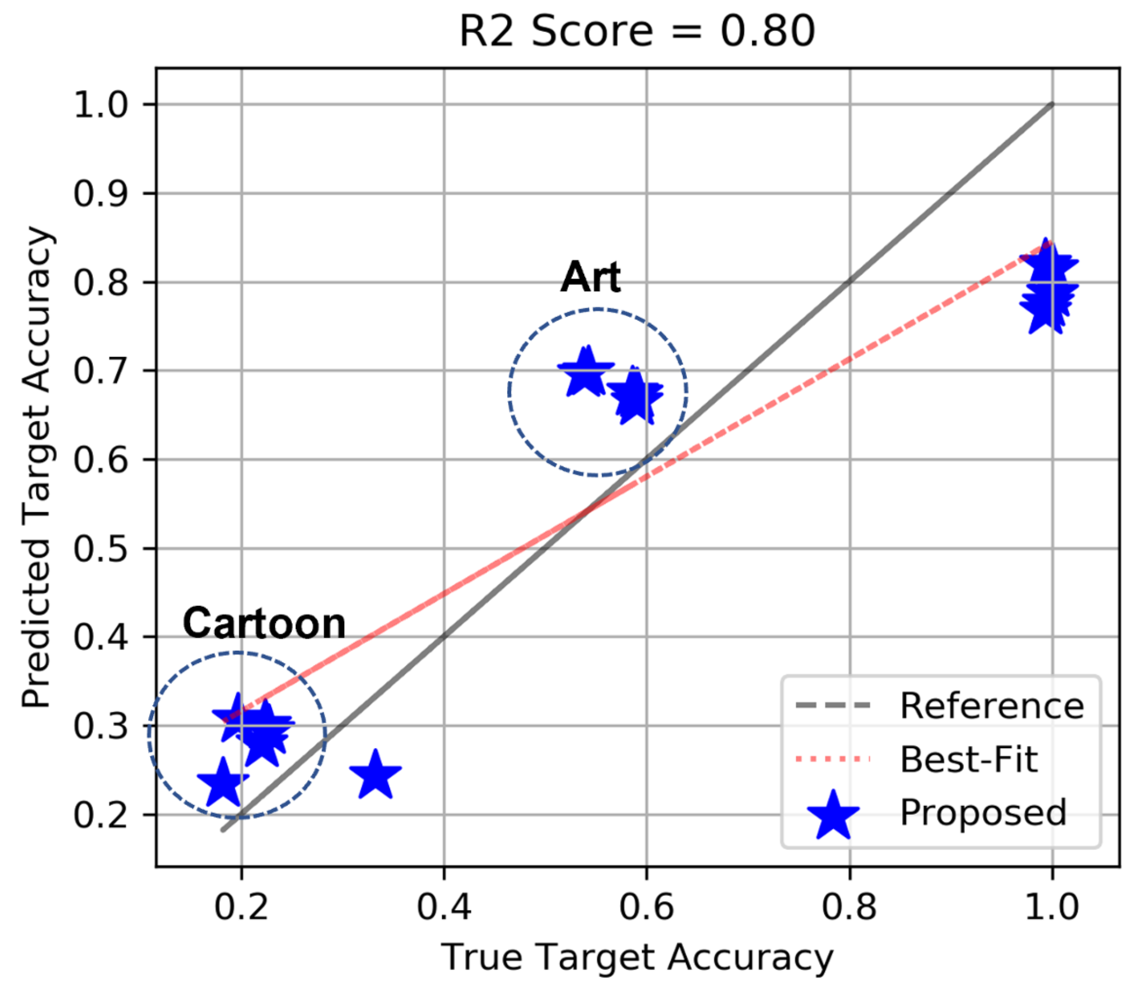




Expected Target Accuracy?



Guillory et al., CVPR 2021

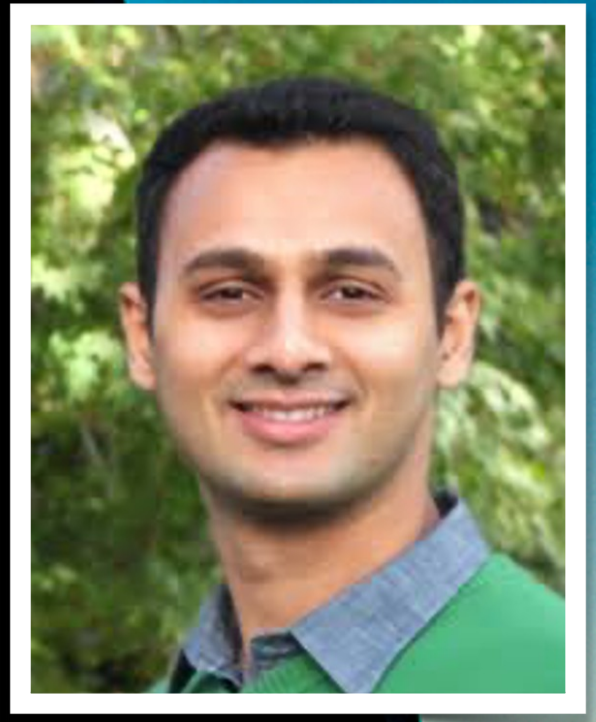


Predicting the Generalization Gap under Distribution Shifts using Anchoring (Preprint coming soon)

Key Takeaways

- “Knowledge-aware” learning is emerging as a key research field in ML as different forms of world models are becoming available.
- OOD Generalization, ML Safety and data efficiency are critical axes to holistically evaluate how well we leverage these pre-trained models in our ML pipelines.
- We need new theoretical tools to precisely characterize the trade-off between these axes when using different “priors”
- Knowledge is “incomplete” – Suitably augmenting world models with our experience is essential to realize closed-loop systems.
- Uncertainty estimation and model reliability characterization are an integral part of model design and optimization.

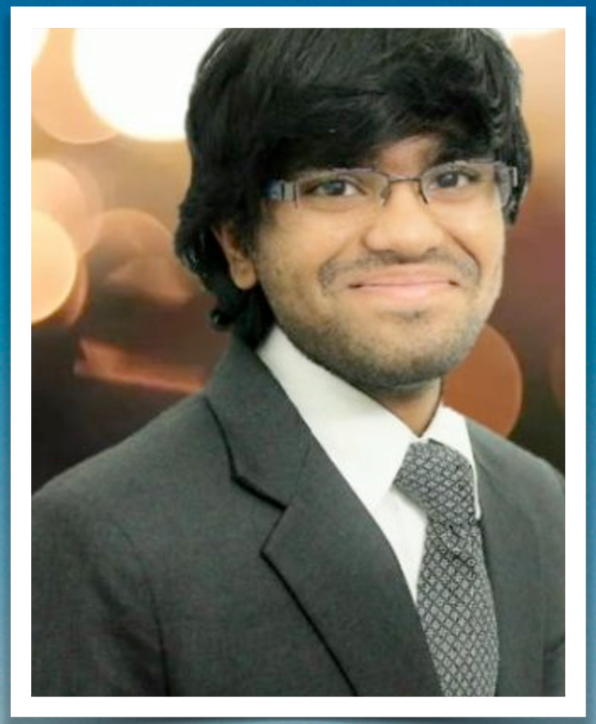
My Awesome Collaborators!



Rushil Anirudh

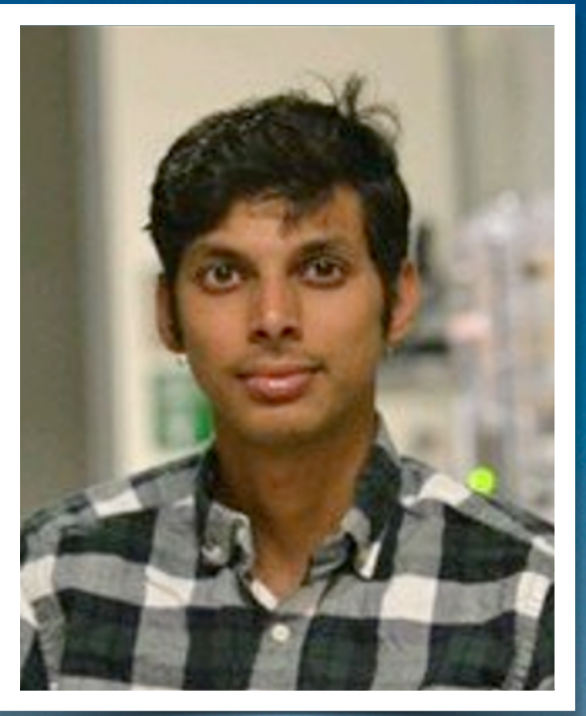
Puja Trivedi

Vivek N



Rakshith S

Mark Heimann

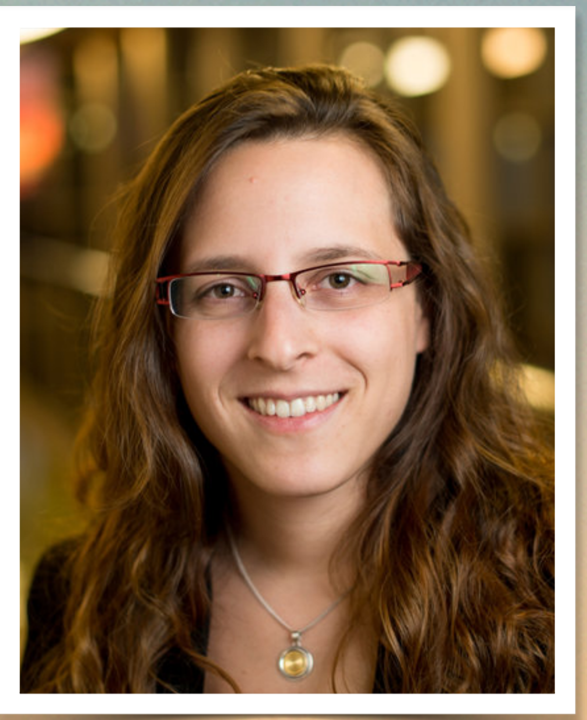


Kowshik Thopalli

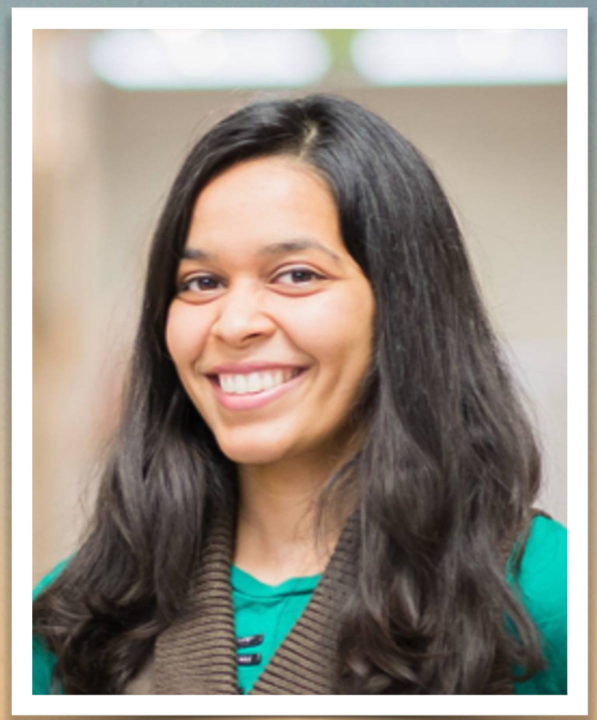
Peer-Timo Bremer

T. S. Jayram

Yamen Mubarka



Danai Kotura

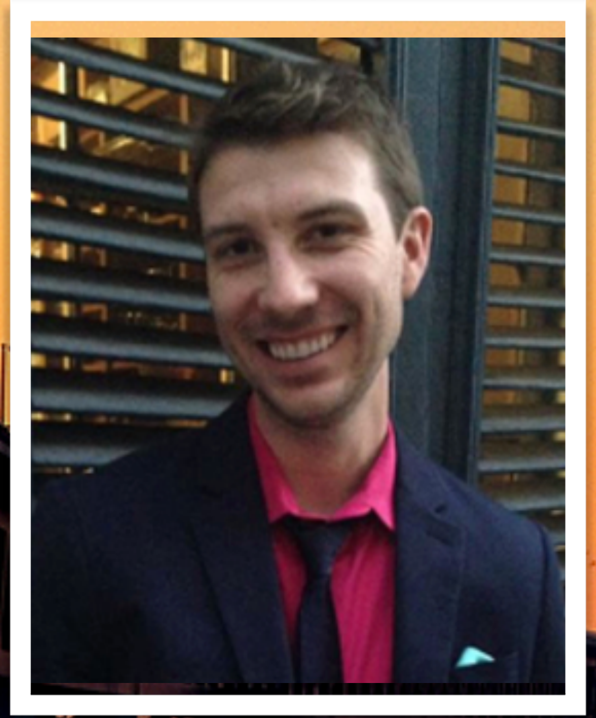


Deepta Rajan

Bhavya Kailkhura

Pavan Turaga

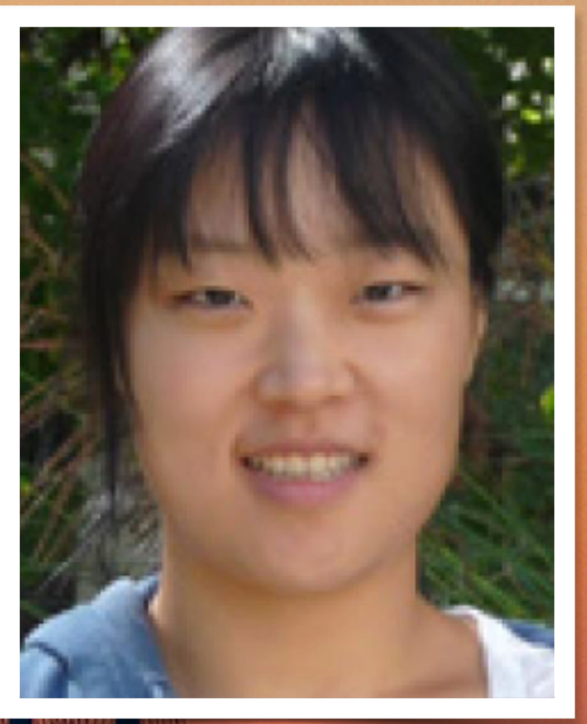
Andreas Spanias



Luc Peterson

Akshay Chaudhari

Brian Spears



Irene Kim

Thank You!!

jjthiagarajan@gmail.com https://jjthiagarajan.com [jjayaram7](https://twitter.com/jjayaram7)