Diagnosing Generalization Failures from Representational Geometry Markers

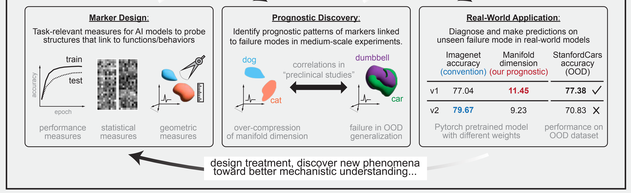
Chi-Ning Chou[1,2], Artem Kirsanov[3], Yao-Yuan Yang[4], SueYeon Chung[1,2,5]

1: Center for Computational Neuroscience, Flatiron Institute.
2: Department of Physics, Harvard University.
3: Program in Neuroscience, Harvard University.
4: Google DeepMind.
5: Applied Mthematics, Harvard University.

Diagnostic, System-Level Analysis for Generalization Failures

Motivation: Current research often takes bottom-up, mechanistic approaches such as mechanistic interpretability methods that aim to reverse-engineer DNNs' internal computations by identifying interpretable features [1,2,3], functional circuits [4,5], or causal structures [6,7]. While powerful, these methods typically yield insights at a microscopic level (e.g., detailed features or circuits), yet they struggle to provide a macroscopic view of network behavior that can guide the prediction and prevention of system-level failures.

Our method: We propose a complementary perspective inspired by the history of medicine: a diagnostic, system-level paradigm for understanding neural networks. Rather than attempting to reconstruct their internal mechanisms, we focus on a top-down approach: develop task-relevant measurements—markers for AI models—that serve as reliable indicators of potential failure modes.



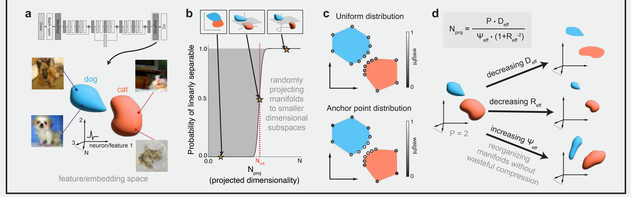
Marker Design to Probe Task-Relevant Information

Overview: A central step in our diagnostic framework is selecting and designing markers—**scalar quantities computed entirely from ID data**—that capture aspects of a pretrained model relevant for downstream generalization. In image classification, we focus on penultimate-layer feature vectors, as the final classification decision obtained by a linear readout from this layer.

Performance measures: Train/test accuracy, logit-based measures.

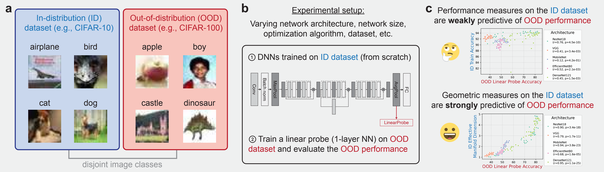
Statistical measures: Feature sparsity, covariance etc.

Geometrical measures: Feature distance/angle, participation ratio, dimensionality measures (e.g., numerical rank [11,12]), neural collapse measures [13,14,15], task-relevant geometric measures (e.g., GLUE [16]).



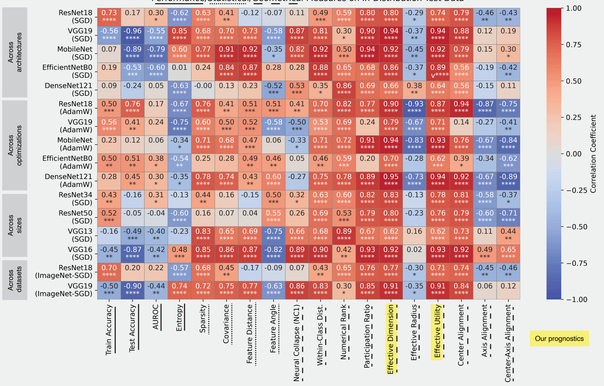
Prognostic Discovery for Failures in OOD Generalization

Overview: We study medium-scale models as a testbed for identifying prognostic indicators of failure modes. Our goal is to detect **ID signals** that reliably predict how a model will behave under distribution shift—**without any access to OOD data**.



Setting: We trained multiple DNN architectures (e.g., ResNet, VGG) from scratch on CIFAR-10. For each architecture, we swept over four initial learning rates, four weight decay values, and three random seeds, using both SGD and AdamW optimizers. In all cases, we ensured that the training accuracy was above 99% and the test accuracy ranged from 88% to 95%.

To assess the OOD generalization of learned representations, we adopt a linear probing framework [17,18]. After ID training, the network's feature extractor was frozen. A new linear classifier was then trained on top of these features using the OOD dataset (CIFAR-100). The test accuracy of this linear probe served as our measure of OOD performance.



Result: We found that **larger ID effective dimension and effective utility (from GLUE [16] analysis) indicate failure in OOD performance**. Other ID dimensionality and collapse measures also exhibit positive correlations.

Applications to Predicting Performance of Transfer Learning

Overview: A common scenario in applied machine learning involves selecting a pretrained model from a public repository like PyTorch Hub or Hugging Face. For a given architecture, multiple sets of weights are often available, each trainedwith different optimization recipes, regularization schemes, or data preprocessing pipelines. The standard heuristic is to choose the model with the highest reported ID accuracy. Here, we apply the prognostic indicators discovered in our exploratory experiments to this practical challenge.

Setting: We analyzed 20 popular architectures from PyTorch's official repository, each released with two sets of weights (v1 and v2). By design, the v2 weights achieve higher accuracy on the ID ImageNet benchmark. This heterogeneity makes for a challenging and realistic testbed for our diagnostic framework. For each v1/v2 pair, we first measured the effective dimensionality and utility of their ImageNet object manifolds. We then evaluated their OOD transfer performance on 9 image classification datasets, e.g., Flowers102, Stanford Cars, Places365, etc.

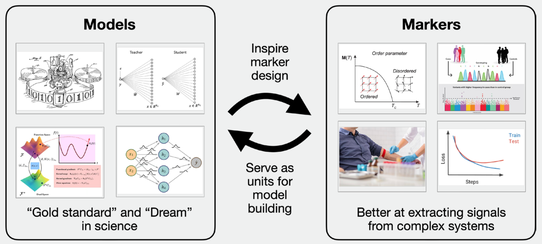
Model		Marker		ID Acc	OOD Acc (measured by a linear probe)									OOD pred. acc.
Arch.	Weight	D_{eff}	Ψ_{eff}	ImageNet	Pet	Flower	Food	Nat.	Place	Car	Texture	Aircraft	Skin	
MobileNet (V3_Large)	v1	**13.07**	**0.956**	74.04	89.16	**89.71**	**72.79**	36.56	**44.39**	**53.28**	**67.64**	**41.60**	**80.29**	7/9
	v2	10.06	0.924	**75.27**	**91.05**	85.82	71.94	**37.66**	43.70	49.97	66.28	34.86	80.12	
RegNet_X (16GF)	v1	**10.50**	**0.929**	80.06	92.72	**88.74**	**75.86**	**42.09**	45.72	**56.13**	70.96	**41.77**	**82.65**	6/9
	v2	8.78	0.899	**82.72**	**93.13**	85.88	74.95	38.84	**46.02**	49.43	**72.30**	28.01	80.49	
RegNet_X (1_6GF)	v1	**11.35**	**0.935**	77.04	91.89	**88.28**	**72.36**	**39.33**	**44.27**	**53.96**	**68.60**	**40.37**	**79.45**	8/9
	v2	9.19	0.908	**79.67**	**92.43**	79.22	68.03	32.68	43.48	40.46	67.06	22.99	78.26	
RegNet_X (32GF)	v1	**10.15**	**0.930**	80.62	93.00	**88.06**	**76.00**	**40.93**	45.72	**54.86**	**71.67**	**38.54**	**82.68**	7/9
	v2	8.81	0.905	**83.01**	**93.38**	85.94	75.73	38.59	**46.35**	47.22	70.41	26.42	81.19	
RegNet_Y (16GF)	v1	**9.87**	**0.923**	80.42	**92.81**	**88.19**	**76.66**	**41.08**	45.89	**54.62**	71.10	**41.60**	**83.42**	7/9
	v2	9.29	0.906	**82.89**	92.30	85.07	75.99	37.58	**46.72**	48.63	**71.19**	27.26	80.02	
RegNet_Y (1_6GF)	v1	**11.06**	**0.933**	77.95	92.30	**87.46**	**73.64**	**40.26**	**44.72**	**53.89**	**69.17**	**42.75**	**80.49**	8/9
	v2	8.60	0.909	**80.88**	**93.13**	79.46	69.61	31.87	43.67	40.58	65.11	25.86	77.56	
RegNet_Y (400MF)	v1	**10.98**	**0.933**	74.05	90.55	**84.88**	**68.07**	**34.74**	**42.68**	**43.00**	**65.60**	**38.32**	**76.56**	8/9
	v2	9.16	0.914	**75.80**	**91.06**	76.81	63.61	27.75	40.77	33.52	63.71	22.36	75.92	
RegNet_Y (800MF)	v1	**11.27**	**0.937**	76.42	91.79	**86.25**	**72.28**	**38.97**	**44.34**	**52.91**	**69.34**	**42.29**	**79.09**	8/9
	v2	9.42	0.913	**78.83**	**92.44**	79.48	68.00	32.28	42.92	43.37	65.80	30.10	77.39	
RegNet_Y (8GF)	v1	**10.32**	**0.926**	79.34	92.40	**88.23**	**76.39**	**42.38**	45.86	**55.70**	70.09	**44.32**	**83.65**	6/9
	v2	8.61	0.903	**81.68**	**93.15**	85.10	75.53	38.19	**46.24**	50.16	**70.60**	28.30	80.99	
ResNeXt (101_32X8D)	v1	**10.14**	**0.925**	77.62	92.10	**83.77**	69.04	**34.76**	43.27	**45.46**	67.71	**32.90**	**79.39**	5/9
	v2	9.61	0.922	**81.20**	**92.59**	80.11	**69.51**	31.76	**44.93**	40.37	**68.17**	24.76	78.55	
ResNet (152)	v1	**9.48**	**0.920**	78.31	**92.86**	**84.73**	**71.88**	**36.26**	44.81	**47.96**	67.73	**35.47**	**81.22**	7/9
	v2	9.14	0.910	**82.28**	92.66	81.17	71.70	32.23	**46.02**	40.62	**69.82**	23.16	79.19	
ViT (B_16)	v1	**9.56**	**0.956**	81.87	**95.24**	**99.56**	**94.04**	**61.09**	**50.51**	**79.98**	**78.51**	43.05	81.42	7/9
	v2	8.86	0.928	**85.30**	95.18	97.71	89.98	60.03	50.14	77.93	76.12	**50.10**	**83.58**	
ViT (H_14)	v1	**9.89**	**0.946**	85.71	96.33	99.64	**95.95**	59.60	52.02	85.48	77.66	46.91	83.98	1/9
	v2	8.32	0.907	**88.55**	**96.73**	99.64	95.60	**69.76**	**53.29**	**88.88**	**79.73**	**55.17**	**84.75**	
ViT (L_16)	v1	**9.56**	**0.947**	85.15	95.99	**99.58**	**95.74**	58.39	51.37	84.16	**79.15**	45.48	**81.78**	4/9
	v2	6.49	0.888	**88.06**	**96.51**	99.43	95.02	**66.18**	**51.95**	**87.20**	78.35	**55.84**	81.45	
WideResNet (50_2)	v1	9.19	0.914	78.47	92.02	81.14	67.46	32.46	43.26	42.34	66.81	27.64	**79.92**	8/9
	v2	**10.59**	**0.925**	**81.60**	**93.20**	**82.29**	**72.67**	**35.41**	**46.61**	**47.50**	**70.90**	**28.03**	79.89	

Results: The prediction accuracy from our prognostics (73.02%) is significantly higher than the prediction accuracy of using ID test accuracy (37.22%).

Discussion

Connection to shortcut learning: Shortcut learning refers to a model relies on simple, highly predictive but non-robust features that correlate with the label in the training data, instead of learning the underlying task-relevant structure. Our prognostics of OOD generalization failure via an **overcompression of effective dimension and utility** can be understood as measures that quantify the effective amount of relevant features learned by a model. It is an interesting future direction to establish a rigorous connection between shortcut learning and the overcompression of object manifolds in feature space.

Model-first vs. Marker-first approach: Scientific inquiry often oscillates between two complementary paradigms. Model-first approaches, rooted in physics and increasingly prominent in ML theory, construct mechanistic frameworks to generate predictions and causal understanding. In contrast, marker-first approaches prioritize empirically observable summaries—such as order parameters, biomarkers, or training metrics—that track system behavior without requiring a full model. Marker-first strategies are often effective at uncovering unexpected regularities, which can subsequently guide the development of deeper mechanistic theories. The two approaches form a productive feedback loop: markers reveal phenomena worth explaining, and models consolidate and generalize those insights. Neural networks, as emergent high-dimensional systems, may benefit from a diagnostic science with marker-first approach that anticipates vulnerabilities and unseen failure modes, and later guides future mechanistic insight.



References & Affiliations & Fundings

[1] Olah et al., 2017
[6] Mueller et al., 2024
[11] Masarczyk et al., 2023
[16] Chou et al., 2025
[2] Yun et al., 2023.
[7] Geiger et al., 2025
[12] Harun et al., 2024
[17] Alain & Bengio, 2016
[3] Cunningham et al., 2023
[8] Hendrycks & Dietterich, 2018
[13] Papyan et al., 2020
[18] Zhu et al., 2023
[4] Olah et al., 2020
[9] Guillory et al., 2021
[14] Ammar et al., 2023
[5] Dunefsky et al., 2024
[10] Liu et al., 2020
[15] Harun et al., 2025

This work was supported by the Center for Computational Neuroscience at the Flatiron Institute, Simons Foundation, and by NIH grant R01DA059220. S.C. was partially supported by a Sloan Research Fellowship, a Klingenstein-Simons Award, and the Samsung Advanced Institute of Technology project, "Next Generation Deep Learning: From Pattern Recognition to AI." All experiments wereperformed using the Flatiron Institute's high-performance computing cluster.

Contact: cchou@flatironinstitute.org

paper link