Interpretability at the Network Level: Prior-Guided Drift Diffusion for Neural Circuit Analysis

Tahereh Toosi
Center for Theoretical Neuroscience
Zuckerman Mind Brain Behavior Institute
Columbia University
tahereh.toosi@columbia.edu

Abstract

Interpretability at the neuron level has provided valuable insights into how individual units respond to specific features and patterns. To advance interpretability at the network level, we propose treating networks as generative models to probe their learned statistical priors. We introduce Prior-Guided Drift Diffusion (PGDD), which accesses the implicit statistical structure networks acquire during training. PGDD iteratively refines inputs according to the network's learned priors, essentially probing what patterns emerge from the network's internal statistical knowledge. For adversarially robust networks, this leverages implicit denoising operators shaped by robust training. For standard networks, our extension uses gradient smoothing techniques to stabilize the generative process. Applying this method during early training reveals that networks appear to acquire rich semantic representations well before achieving reliable classification performance. This demonstrates a dissociation between internal representation learning and classification performance, where networks develop structured knowledge before they can reliably use it. Our training-free approach provides direct access to this latent representational structure in the models we tested.

1 Introduction

How can we understand what concepts a network has learned? Interpretability at the neuron level has provided valuable insights into individual unit responses, but understanding network-level knowledge—what populations of neurons collectively know—remains challenging. Current approaches have significant limitations: external generative models impose their own inductive biases [Bau et al., 2019, Xie et al., 2021], while methods requiring additional training face practical challenges including hyperparameter sensitivity, seed-dependent instability, and poor downstream performance [Gao et al., 2024, Authors, 2025, Rajamanoharan et al., 2024]. With reasoning models and inference-time compute becoming central [OpenAI, 2024, Wei et al., 2022], understanding how networks could use their learned knowledge generatively becomes crucial.

We introduce Prior-Guided Drift Diffusion (PGDD), a method that treats networks as implicit generative models. Instead of asking "what activates this neuron?", we ask "what patterns can this network generate from its learned statistical knowledge?" For robust networks, PGDD accesses the implicit denoiser shaped by adversarial training; for standard networks, our extension sPGDD uses gradient smoothing to access learned priors. PGDD works by iteratively refining inputs according to the network's own learned statistical regularities. Applied to early training epochs, PGDD provides evidence of early semantic representation development: networks consistently generate bird-like patterns across different noise initializations by epoch 4, with per-category accuracy analysis confirming preferential learning of avian categories despite poor overall performance at early epochs.

Workshop on Mechanistic Interpretability, NeurIPS 2025.

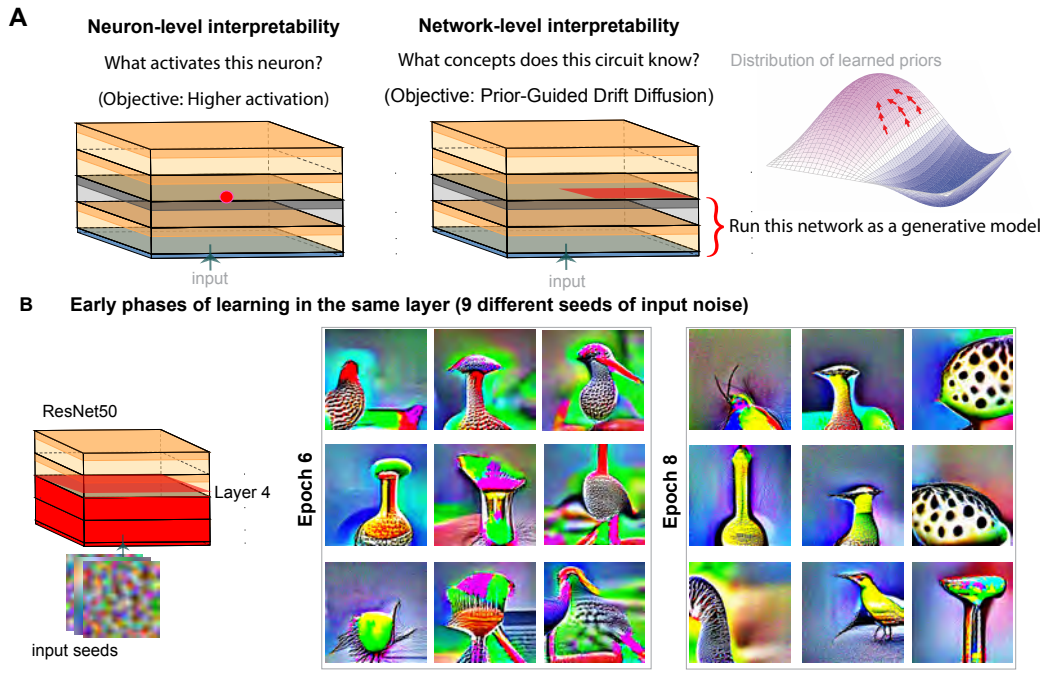


Figure 1: **Network-level interpretability through implicit generative operators.** (A) PGDD shifts from neuron-level analysis to network-level understanding by running networks as generative models to probe learned priors. (B) Across 9 different noise initializations, PGDD applied to ResNet-50 layer 4 at epochs 6-8 consistently produces bird-like patterns with recognizable features (beaks, feathers, wings) at different viewing angles, suggesting an instance of rapid semantic learning in early training. Results from adversarially-trained [Madry et al., 2018] robust network ($\varepsilon = 4$, L_2); PGDD parameters: reference noise $\sigma^2 = 0.2$, diffusion noise=0.01, iterations=500.

Our main contributions are: (1) PGDD, a training-free method that accesses network priors through implicit generative operators, (2) evidence that networks rapidly acquire semantic knowledge within epochs despite poor classification performance, and (3) demonstration of a dissociation between internal knowledge acquisition and external performance metrics.

2 Prior-Guided Drift Diffusion

PGDD iteratively optimizes an input to align its representations with a noisy reference, effectively asking the network to "denoise" according to its learned priors. Given a network f and starting input $\hat{x}$, we minimize:

$$\mathcal{L}_{\text{PGDD}}(\hat{x}) = \|r_\ell(\hat{x}) - \text{sg}(r_\ell(\hat{x} + \varepsilon))\|_2^2 \tag{1}$$

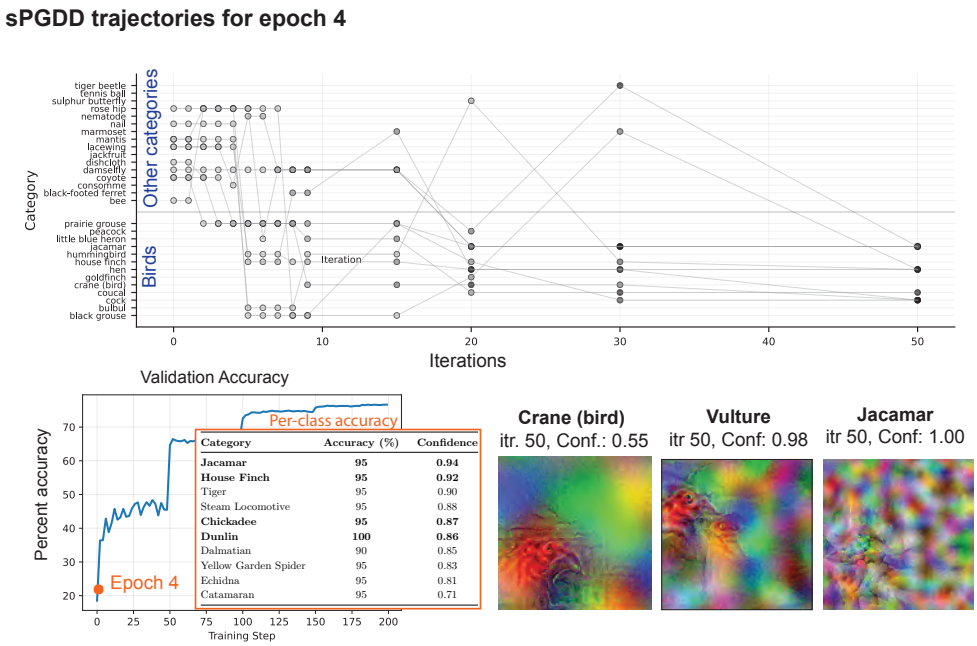
where $r_\ell(\cdot)$ are representations at layer ℓ, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, and $\text{sg}(\cdot)$ stops gradients through the reference.

The update rule follows:

$$\hat{x}_{t+1} = \hat{x}_t + \eta \nabla_{\hat{x}_t} \mathcal{L}_{\text{PGDD}} + \zeta_t \tag{2}$$

where η is the step size and $\zeta_t \sim \mathcal{N}(0, \tau^2 I)$ adds stochastic exploration (diffusion noise). The gradient $\nabla_{\hat{x}} \mathcal{L}_{\text{PGDD}} \approx J(\hat{x})^\top J(\hat{x})\varepsilon$ reveals that PGDD applies the learned denoiser $J^\top J$ to noise ε, where $J(\hat{x}) = \nabla_{\hat{x}} r_\ell(\hat{x})$ is the Jacobian of representations with respect to the input. In robust networks, adversarial training shapes this operator to preserve class-relevant information (see supplementary Section A.2 for detailed theoretical justification).

Extension to Standard Networks (sPGDD) For standard networks that lack the well-structured $J^\top J$ operator shaped by adversarial training, we develop *sPGDD* (smooth PGDD). The core



Category	Accuracy (%)	Confidence
Jacamar	**95**	**0.94**
House Finch	**95**	**0.92**
Tiger	95	0.90
Steam Locomotive	95	0.88
Chickadee	**95**	**0.87**
Dunlin	**100**	**0.86**
Dalmatian	90	0.85
Yellow Garden Spider	95	0.83
Echidna	95	0.81
Catamaran	95	0.71

Figure 2: **A basic category (birds) seems to emerge in early training** (Top) sPGDD applied to standard ResNet-50 at epoch 4 trained on ImageNet consistently produces bird-like patterns across different noise seeds. Y-axis shows detected categories for generated interpretations, with dot intensity indicating classification confidence. (Bottom left) Validation accuracy curve showing epoch 4 position in training progression. (Bottom right) Three example generated patterns at iteration 50 (Crane: 0.55, Vulture: 0.98, Jacamar: 1.00), demonstrating consistent bird-like features. Per-category training accuracy analysis (table inset) shows 4 of top 10 categories are bird classes, suggesting birds as a learned basic level category despite low overall accuracy at epoch 4. (For smoothing, $n_{sample} = 100$ and $\sigma^2 = 0.1$ Control experiments on other epochs in supplementary 5)

idea is to stabilize the update step by smoothing the gradients at each iteration, rather than relying on a single noisy gradient estimate. Specifically, we fix the noisy reference representation $r(x + \epsilon)$ once at the start of the trajectory, and then, at each iteration t, we compute multiple gradients with respect to independently sampled noise perturbations $\{\epsilon_i\}_{i=1}^n$ and average them as $g_t = \frac{1}{n}\sum_{i=1}^n \nabla_{x_t} L_{\text{PGDD}}(x_t; \epsilon_i)$. This gradient smoothing reduces variance, suppresses spurious noise-sensitive directions, and emphasizes the stable prior information embedded in the network. In practice, sPGDD yields smoother and more interpretable trajectories in non-robust networks, though with lower fidelity compared to robust models.

3 Experiments

Understanding how learned representations evolve during training provides crucial insights into neural network development. Previous work like Network Dissection [Bau et al., 2017] has shown that interpretable units emerge gradually across training epochs, with higher layers developing complex patterns only after extensive training. To investigate how network priors develop, we applied PGDD to adversarially robust networks and sPGDD to standard-trained neural networks at different training epochs.

3.1 Learning Trajectories and Semantic Emergence

We applied PGDD to ResNet-50 models trained on ImageNet across epochs 0, 4, 6, 8, 10, 50, 100, 150 (where epoch 0 refers to after the first training epoch, not random initialization). We initialized PGDD with Perlin noise patterns at systematically varying seeds and octaves to ensure robustness across different starting conditions. As shown in Figure 1B, even at early epochs PGDD

arrived at consistent meaningful structure. For example, at epochs 6 and 8, generated patterns show similar bird-like features yet differ in meaningful variations like position, viewing angle, and specific anatomical details, suggesting the network has learned structured representations of avian categories.

To check whether this phenomenon is specific to adversarially trained models, we ran sPGDD on standard-trained networks. Figure 2 shows results for epoch 4 (additional adjacent epochs in supplementary Figure 5). The network consistently arrives at bird-like patterns, though generated images are less sharp than those from adversarially trained networks. This suggests that semantic structure emerges rapidly even in standard training regimes. To validate whether PGDD patterns reflect actual network knowledge, we evaluated per-category accuracy on training images (since priors are built from the training distribution). Despite overall accuracy of only 23% at epoch 4, analysis of the top 10 performing categories across 1000 ImageNet classes revealed that 4 are bird categories (Figure 2B), with consistently high confidence scores. This pattern holds across multiple random seeds and different noise initializations, as shown in supplementary Figure 3. Unlike neuron-level activation maximization methods, which can be unstable and often fail to converge, PGDD operates at the level of network priors and produce stable and strongly convergent trajectories.

This finding is particularly striking because even in Network Dissection [Bau et al., 2017], higher layers at initial epochs did not demonstrate interpretable units or complex semantic concepts beyond those found in lower layers. The rapid emergence of bird categories suggests that basic-level categories [Rosch et al., 1976] may be learned much faster than previously recognized, consistent with theoretical predictions about hierarchical concept acquisition [Saxe et al., 2019]. However, this needs validation across different architectures and training seeds to verify whether this represents a general "bird effect" and whether there are sequential aspects to category emergence (see supplementary Figures 5 and 4 for additional controls). These findings are specific to ResNet-50 trained on ImageNet and require validation across different architectures and datasets to establish broader generality.

4 Discussion

Our main contribution is PGDD, a training-free method for network-level interpretability that treats classifiers as implicit generative models and, through sPGDD, extends to standard networks. Using this tool, we probed training dynamics and found that networks acquire semantic structure (e.g., bird-like features) within a few epochs, well before classification accuracy improves, revealing a dissociation between internal representation learning and external performance. While prior work has shown that classifiers can synthesize images [Santurkar et al., 2019, Grathwohl et al., 2019], our approach is the first to link these generative properties to denoising score matching and extract priors from intermediate layers, extending SmoothGrad into a generative inference setting. This connects to theories of rapid concept acquisition [Saxe et al., 2019, McClelland et al., 2010] and to recent work on emergence in learning [Fort and Jastrzebski, 2019]. PGDD thus highlights a hidden layer of knowledge in networks that is invisible to performance metrics alone. Limitations include reliance on robust models for the clearest results, noisier outputs in sPGDD, and evaluation restricted to ResNet-50 on ImageNet; Control experiments confirm untrained networks show no structured patterns (supplementary Figure 4). future work should extend across architectures, datasets, and domains beyond vision.

5 Conclusion

We introduced PGDD, a training-free method that repurposes classifiers as implicit generative models to access their learned priors. Unlike neuron-level or prediction-focused interpretability methods, PGDD can reveal prior structure from intermediate layers, offering a richer perspective on how networks represent knowledge. Applied across training epochs, PGDD shows that, in the models we probed, semantic categories such as bird emerge well before classification accuracy, suggesting that internal representation development precedes external performance. This dissociation reframes our understanding of learning dynamics and aligns with theories of rapid concept acquisition. Beyond interpretability, PGDD provides a way to track training trajectories, uncover hidden biases in learned representations, and assess how priors shape model behavior. These capabilities may inform both scientific inquiry and the safety of deployed AI systems. Future work should extend PGDD to other

architectures and domains, exploring its role as both an analytical tool and a diagnostic instrument for emerging reasoning models.

Acknowledgments and Disclosure of Funding

I would like to thank Kenneth D. Miller for our ongoing collaboration and discussions on PGDD and related ideas. This work was supported by the NIH National Eye Institute (K99/R00), the Institute for Artificial and Natural Intelligence (ARNI), and the Gatsby Charitable Foundation.

References

Anonymous Authors. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *ICLR*, 2019.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863, 2017.

Harris Drucker and Yann Le Cun. Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, 3(6):991–997, 1992.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. In *International Conference on Machine Learning*, pages 1823–1832, 2019.

Chris Finlay and Adam M Oberman. The robustness of deep networks: a geometrical perspective. *IEEE Signal Processing Magazine*, 36(4):50–62, 2019.

Stanislav Fort and Stanislaw Jastrzebski. Large scale structure of neural network loss landscapes. In *NeurIPS*, 2019.

Leo Gao, Jeffrey Wu, et al. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Will Grathwohl, Kuan-Chieh Wang, Jorn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2019.

Judy Hoffman, Daniel A Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

James L McClelland, Matthew M Botvinick, David C Noelle, David C Plaut, Timothy T Rogers, Mark S Seidenberg, and Linda B Smith. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8):348–356, 2010.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018.

OpenAI. Learning to reason with llms. Technical report, OpenAI, 2024.

Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

Sen Rajamanoharan, Lewis Smith, Arthur Conmy, et al. Negative results for sparse autoencoders on downstream tasks, 2024.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439, 1976.

Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Thirty-second AAAI Conference on Artificial Intelligence*, 2018.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. In *International Conference on Learning Representations*, 2018.

Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *Advances in Neural Information Processing Systems*, pages 1262–1273, 2019.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *PNAS*, 116(23):11537–11546, 2019.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Shaofeng Xie, Xiaoxuan Liang, and Zhenguo Li. The deep generative decoder: using generative models to understand deep representations. *arXiv preprint arXiv:2110.07191*, 2021.

A Supplementary Material

A.1 Prior-Guided Drift Diffusion: Algorithm, Intuition, and Theory

We present the detailed algorithm for Prior-Guided Drift Diffusion (PGDD), together with the underlying intuition and theoretical justification for how PGDD grants access to the learned priors of a network. The method is designed to be both conceptually transparent and practically simple, offering a principled way to leverage the implicit generative structure in networks which were not explicitly trained for pattern generation (notably classifiers).

For reproducibility, we provide a minimal implementation at: `https://anonymous.4open.science/r/PGDD_MechInterp_minimal-21B7/README.md`

A.1.1 PGDD algorithm

Algorithm 1 Prior-Guided Drift Diffusion Objective

Input: Image x_{input}, model f, target layer ℓ, constraint ϵ, step size α, noise ratio σ, iterations T
Output: Refined representations $\{x_t\}_{t=0}^{T}$
// Step 1: Feedforward pass
$x_0 \leftarrow \text{normalize}(x_{\text{input}})$
$f_\ell \leftarrow \text{extract_layers}(f, \ell)$ {Extract model up to layer ℓ}
$x_{\text{noisy}} \leftarrow x_0 + \sigma \cdot \mathcal{N}(0, I)$
$r_{\text{anti-target}} \leftarrow f_\ell(x_{\text{noisy}})$ {Generate noisy reference representation}
for $t = 1$ to T **do**
  // Step 2: Inference objective selection
  anti-target $\leftarrow r_{\text{anti-target}}$ {Use noisy reference as target}
  // Step 3: Feedback error propagation
  $h_t \leftarrow f_\ell(x_{t-1})$ {Forward pass through target layers}
  $\mathcal{L}_t \leftarrow \|h_t - r_{\text{anti-target}}\|^2$ {MSE loss in representation space}
  $g_t \leftarrow \nabla_{x_{t-1}} \mathcal{L}_t$ {Gradient via feedback pathways}
  // Step 4: Constrained activation update
  $\tilde{g}_t \leftarrow \alpha \cdot g_t / (\|g_t\| + \text{1e-10})$ {Normalize gradient}
  $\eta_t \leftarrow \text{diffusion_noise_ratio} \cdot \mathcal{N}(0, I)$ {Add stochastic noise}
  $x'_t \leftarrow x_{t-1} + \tilde{g}_t + \eta_t$ {Move away from representation of noisy input (anti-target)}
  $x_t \leftarrow \text{project}(x'_t, x_0, \epsilon)$ {Enforce $\|x_t - x_0\|_\infty \leq \epsilon$}
  // Step 5: Iteration control
  {Continue to next iteration}
end for

return $\{x_t\}_{t=0}^{T}$

A.2 Intuition behind PGDD and Theory

The Prior-Guided Drift Diffusion (PGDD) objective draws direct inspiration from denoising score matching and Denoising Diffusion Probabilistic Models (DDPMs) **?**. In DDPMs, networks learn to predict added noise by minimizing:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] \tag{3}$$

where $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$. This noise prediction objective implicitly learns the score function of the data distribution, enabling iterative generation through gradient-based sampling.

We show that 1) there is a denoiser hidden in a robust classifier: $J^T J$, it discards class-insensitive perturbations in input 2) We show that the gradient of loss in PGDD between x and $x + \epsilon$ is $J^T J\epsilon$, which means it keeps class-sensitive parts of the perturbation while discarding the rest in each step of PGDD.

PGDD Objective:

$$\mathcal{L}_{\text{PGDD}} = \|r(\hat{x}) - r(\hat{x} + \epsilon)\|^2 \tag{4}$$

where $r(\cdot)$ represents a chosen layer in the network and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

The gradient of the PGDD objective, where $J(\hat{x}) = \nabla_{\hat{x}} r(\hat{x})$ is the Jacobian matrix of representations with respect to input, satisfies:

$$\nabla_{\hat{x}} \mathcal{L}_{\text{PGDD}} \approx 2J^T J\epsilon \tag{5}$$

Linearizing $r(\hat{x} + \epsilon) \approx r(\hat{x}) + J(\hat{x})\epsilon$:

$$\begin{aligned} \nabla_{\hat{x}} \mathcal{L}_{\text{PGDD}} &= 2J(\hat{x})^T (r(\hat{x}) - r(\hat{x} + \epsilon)) && (6) \\ &\approx 2J(\hat{x})^T (r(\hat{x}) - (r(\hat{x}) + J(\hat{x})\epsilon)) && (7) \\ &= 2J(\hat{x})^T J(\hat{x})\epsilon = 2J^T J\epsilon && (8) \end{aligned}$$

This shows PGDD applies a denoising operator $J^T J$ that preserves directions aligned with the learned representations while suppressing orthogonal noise. There is extensive literature showing that robustness suppresses Jacobian norms and local Lipschitz constants Drucker and Le Cun [1992], Ross and Doshi-Velez [2018], Hoffman et al. [2019], Finlay and Oberman [2019]. Adversarial training itself can be understood as a form of operator-norm regularization on J Novak et al. [2018], Cisse et al. [2017]. Empirically, robust networks exhibit markedly reduced input-gradient magnitudes and improved local linearity Tsipras et al. [2019], Etmann et al. [2019]. Recent studies also highlight the low-rank structure of Jacobians in deep networks, linking the spectral decay of $J^\top J$ to both generalization and robustness Novak et al. [2018], Oymak et al. [2019], Sagun et al. [2018]. In particular, a low-rank J implies a low-rank $J^\top J$, which endows $J^\top J$ with filtering properties: in PGDD, at each step, the component of the perturbation that is orthogonal to the target class is suppressed, thereby nudging the trajectory toward a more probable learned prior.

A.3 PGDD roots

We developed Prior-Guided Drift Diffusion (PGDD) not primarily as an interpretability tool, but as a framework to account for human and animal cognitive processes. PGDD extends the broader principle of Generative Inference, which proposes that perception is an active inferential process shaped by the integration of sensory inputs with learned priors.

In biological vision, feedback signals are recruited especially when perception cannot rely on clear feedforward cues alone—for example, in cases of ambiguous, incomplete, or noisy inputs. This feedback-driven integration explains hallmark perceptual effects such as delayed neural responses to illusory contours, laminar-specific activation patterns in early visual cortex, and the flexible interpretation of ambiguous stimuli. PGDD captures these dynamics by iteratively refining internal representations in a way that filters out class-orthogonal perturbations and aligns network activity with more probable priors.

As a result, PGDD provides a computational account of many perceptual and neural phenomena, including figure-ground segregation, Gestalt principles of grouping and closure, illusory contour perception, and imagination-like pattern formation from noise. These effects, long studied in psychophysics and systems neuroscience, emerge naturally when robust classifiers are repurposed to perform inference through feedback. By grounding these phenomena in a single mechanism, PGDD highlights a computational symmetry between learning and inference, offering a unified explanation for how cognitive systems flexibly interpret the world.

A.4 PGDD in early epochs for ResNet50 robust

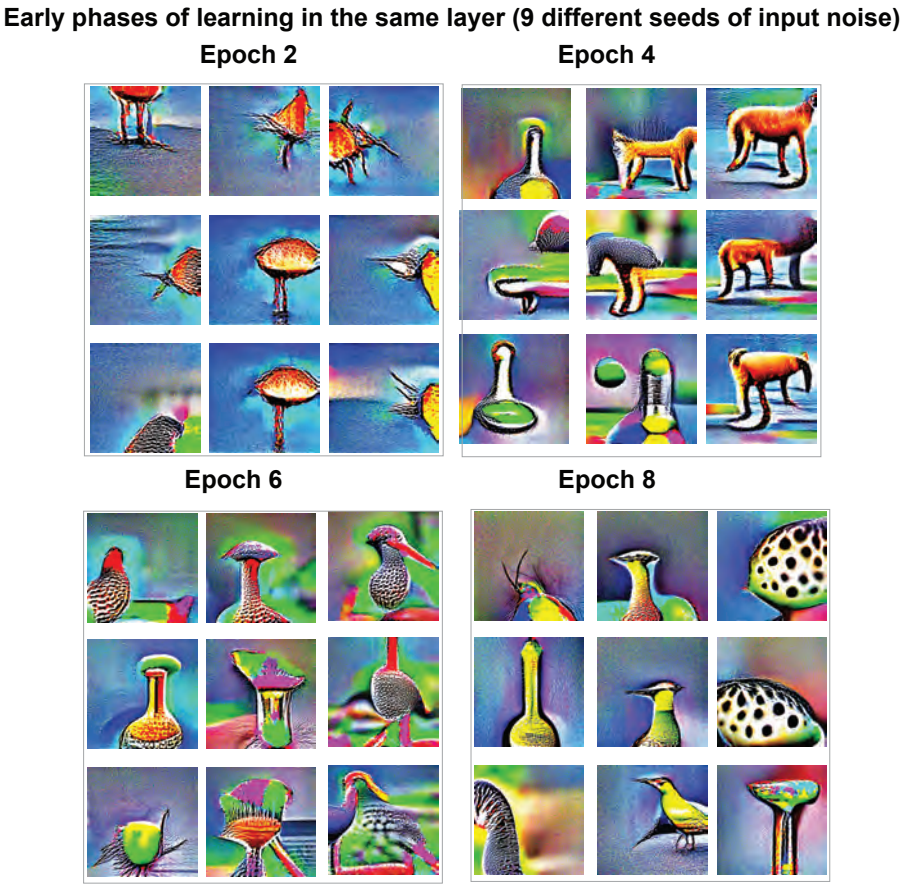


Figure 3: Expanded results corresponding to Figure 1B, showing PGDD applied to ResNet-50 (layer 4) trained on ImageNet across early epochs (2, 4, 6, 8), with 9 different seeds of input noise. PGDD parameters: reference noise variance $\sigma^2 = 0.2$, diffusion noise $= 0.01$, iterations $= 500$, step size $\eta = 0.1$. At epoch 2, generated patterns are largely texture-like and lack coherent structure. By epoch 4, object-like features begin to emerge, including partial animal body from different views.

A.5 sPGDD on untrained network as control

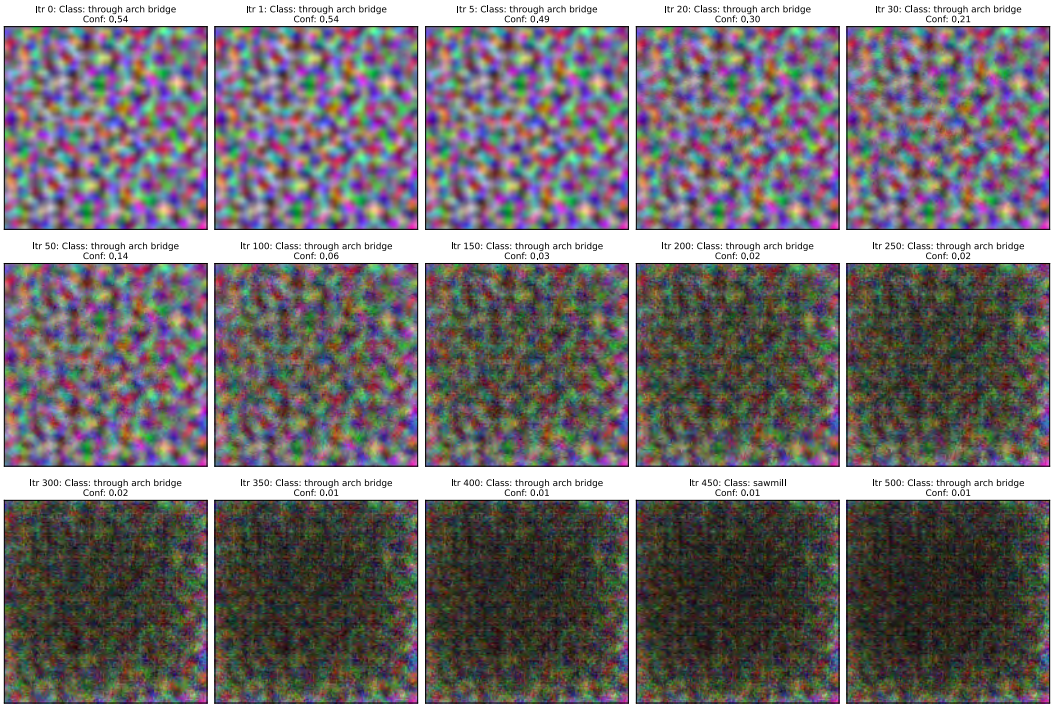


Figure 4: sPGDD on untrained ResNet50 (layer4) with the same input and same sPGDD parameters, does not arrive at structured pattern

A.6 sPGDD trajectory for other epochs

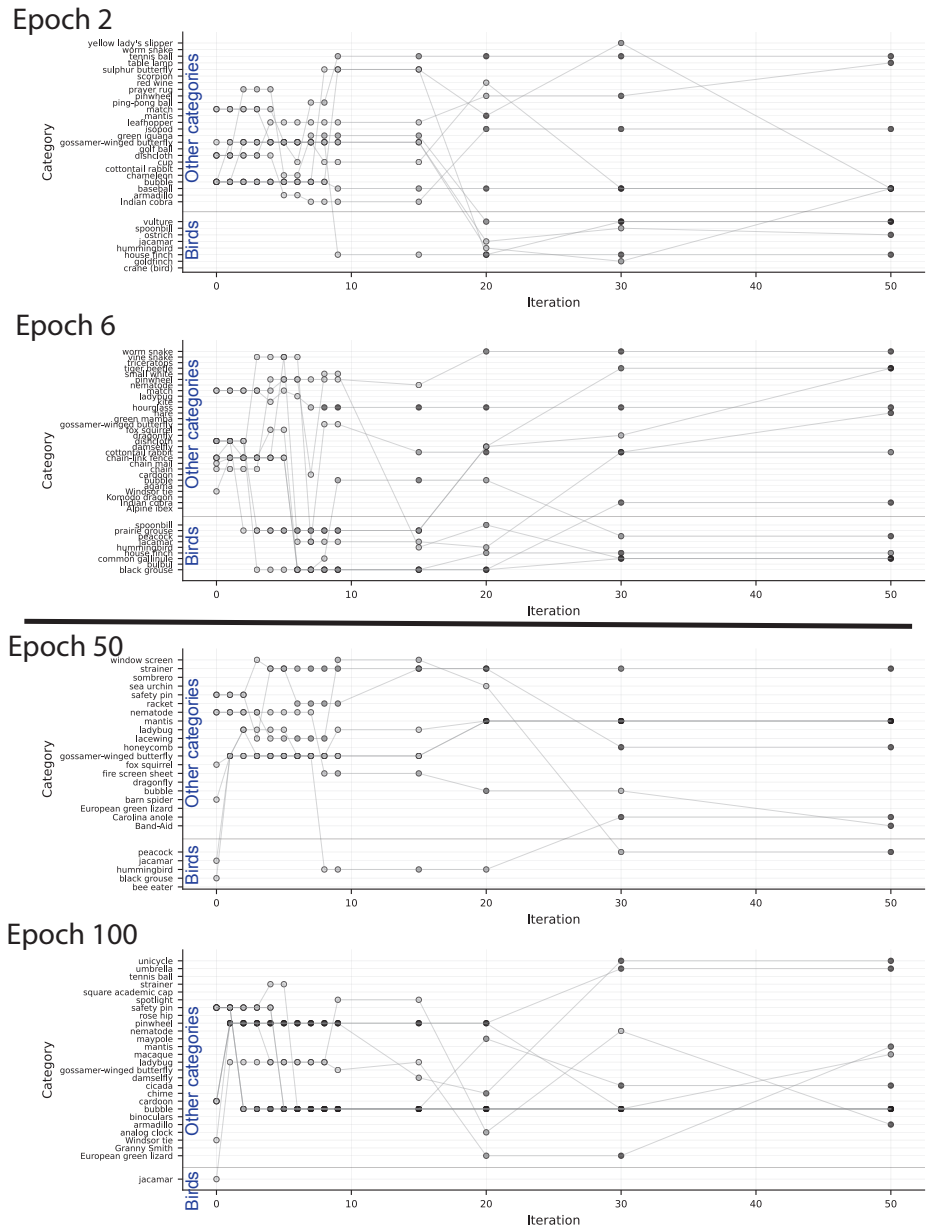


Figure 5: sPGDD trajectories at additional training epochs. Results are shown for epochs 2 and 6, chosen as the two neighboring epochs around the epoch 4 case analyzed in the main text (Figure 2), to demonstrate that the emergence of bird-like patterns is not specific to a single training snapshot but is already visible before and after epoch 4. We also include epochs 50 and 100 to show that while bird-like features are gradually replaced by other structures as discriminative learning unfolds, their early presence was not an artifact of the specific noise input.

A.7 PGDD sweep ϵ robustness in early epochs at layer1

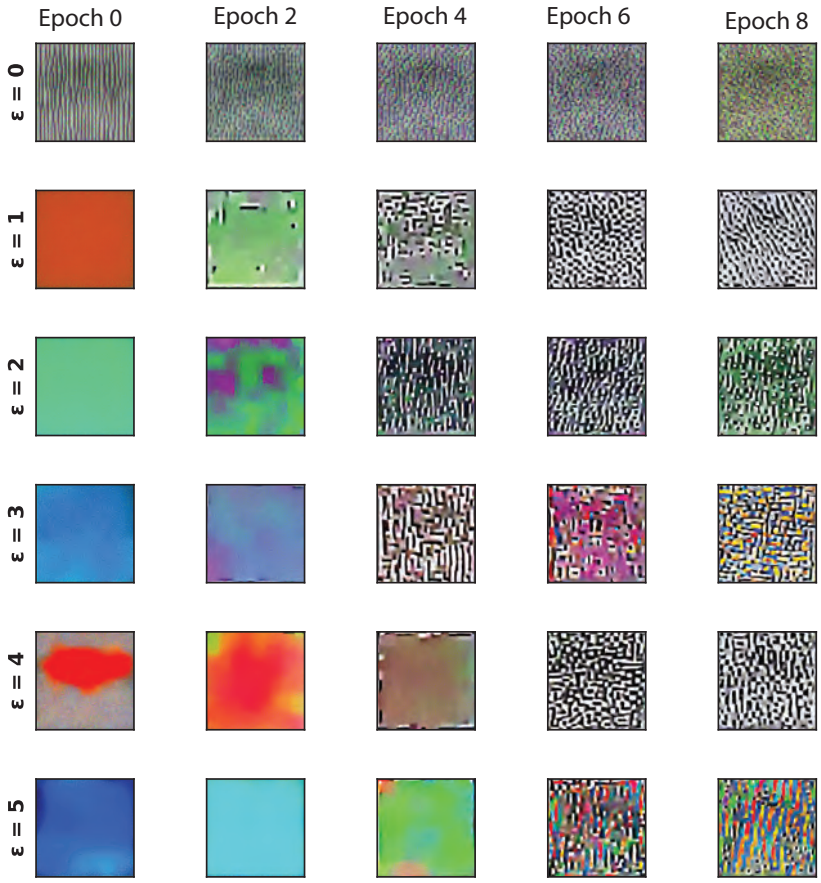


Figure 6: Generated patterns from PGDD applied to ResNet-50 trained on ImageNet, shown across training epochs (0, 2, 4, 6, 8) and robustness levels ($\epsilon = 0$–5). Each row corresponds to a different robustness setting, and each column to a different training epoch. Note that "epoch 0" refers to the network after the *first* epoch of training (not initialization).(ResNet50, Layer 1, at PGDD iteration 1000.). All from the same input seed noise.

A.8 PGDD sweep ϵ robustness across all epochs

Figure 7: Generated patterns from PGDD applied to ResNet-50 trained on ImageNet, shown across training epochs and robustness levels ($\epsilon = 0$–5). Each row corresponds to a different robustness setting, and each column to a different training epoch. Note that "epoch 0" refers to the network after the *first* epoch of training (not initialization).(ResNet50, Layer 3, at PGDD iteration 500.). All from the same input seed noise.