A HIDDEN-LAYER PHASE SCALE PREDICTS WHEN SINUSOIDAL NEURAL FIELDS COLLAPSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Sinusoidal activations remove the low-frequency bias of ReLU networks and are now standard in neural fields. Recent work replaces the single sinusoid of SIREN with a sum of sinusoids whose frequencies are drawn at random when the network is built. Such a network either fits its target or outputs the mean colour of that target, and we found nothing in between. One quantity predicts which of the two, from a single forward pass before training. It is the largest phase scale over the hidden sinusoidal layers. The phase scale of a layer is the slowest frequency of its sinusoids multiplied by the standard deviation of that layer's input. Computing it does not need the target. Over 300 runs on three images, 181 of 181 runs train below a phase scale of 1.81 and none of 51 trains above 4.0. The failure does not sit in the first layer, the one that sees the coordinates. Rescaling that layer by a factor of 20 leaves every run at the mean colour, while rescaling the hidden layers changes the outcome. The correlation map of a random sinusoidal layer gives the phase scale its form. It also accounts for the boundary moving down as the network deepens. A failed network reaches the mean colour by shrinking the weights of its output head. Between those cutoffs the bank of frequencies does not decide the outcome by itself. On identical initial weights, moving only the starting value of the output bias turns runs that train into runs that collapse, and runs that collapse into runs that train. Across 1859 runs the phase scale also orders the risk of collapse on Kodak-24, and it carries to audio, to shapes and to a PDE.

1 INTRODUCTION

Neural fields map a coordinate to a signal value with a multilayer perceptron (Xie et al., 2022). They now carry shape (Park et al., 2019; Mescheder et al., 2019), scenes (Mildenhall et al., 2020; Sitzmann et al., 2019), images (Chen et al., 2021) and compressed signals (Dupont et al., 2021; Strümpler et al., 2022). ReLU networks fit low frequencies first and miss fine detail (Rahaman et al., 2019; Basri et al., 2020). Sinusoidal activations remove that bias and are now standard (Sitzmann et al., 2020; Liu et al., 2024; Saragadam et al., 2023). STAF (Morsali et al., 2026) generalizes SIREN's single sinusoid to a sum of τ sinusoids with learnable frequencies, phases and amplitudes. We call the set of frequencies a bank, and STAF draws it at initialization from $\omega_i \sim U[0, 30)$.

A network built on a random bank either fits its target or outputs the mean colour of that target. We drew 100 banks and trained each of them on three CelebA images. Across those 300 runs nothing landed between the two outcomes, and the failures sit at the mean colour to within the precision we measure. The same failure appears in STAF on Kodak-24. Its activation collapses in 56 of 120 runs inside a five-layer network, and in none of 120 inside STAF's own three-layer network. Two networks that differ only in the frequencies they drew can end on opposite sides, and the draw gives no warning.

We answer with a quantity that any implementation can print before training starts. At initialization, sinusoidal layer l receives a preactivation z_l with standard deviation $\mathrm{std}(z_l)$, and the slowest term of the bank sweeps $\kappa_l = \omega_{\min} \mathrm{std}(z_l)$ radians across one standard deviation of that input. We call κ_l the phase scale of layer l and summarize a network by κ, its largest value over the hidden layers.

The first layer sees the coordinates, which makes it the natural suspect. Two experiments rule it out. A first-layer account predicts that widening the coordinate range by a factor a moves the boundary as $1/a$, and at $a = 0.5$ and $a = 2$ the boundary did not move. Rescaling the first layer alone leaves

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

45 of 45 runs at the mean colour. Rescaling the hidden layers instead, down to a phase scale of 1.5, trains 5 of the same 9 networks. We had written down the opposite expectation for both experiments.

Contributions.

- **A measured boundary.** Below a phase scale of 1.81 all 181 runs train, and above 4.0 none of 51 does. Between the two values the outcomes mix, and the best single threshold still misclassifies 1 to 3 runs in 100. Section 4.4 derives the form of the quantity from the correlation map of a random sinusoidal layer.
- **What sets the position of the boundary.** The phase scale counts the slowest sinusoid of the bank, and the rest of the bank still matters. We held the slowest at 5 and raised only the faster ones. The phase scale stayed between 4.4 and 5.0, while the outcome went from training to not training. The boundary also falls as the network deepens.
- **What a failed network does.** Its output head shrinks to a fraction of its initial size before the output reaches the mean colour. The hidden layers keep their size and change which features they extract. Inside the mixed region, on identical initial weights, moving only the starting value of the output bias reverses the outcome in 18 of 18 pairs.

The phase scale also orders the risk of collapse over 120 STAF networks on Kodak-24, and in an exploratory experiment it carries to audio, to two shapes and to a PDE.

2 Related work

Sinusoidal neural fields. SIREN (Sitzmann et al., 2020) uses a single sinusoid with an initialization rule that holds the preactivation distribution fixed across layers. Later work widens the activation family. FINER (Liu et al., 2024) varies the frequency with the bias, WIRE (Saragadam et al., 2023) uses a Gabor wavelet, MFN (Fathony et al., 2021) multiplies filters, BACON (Lindell et al., 2022) bounds the band, multiplicative filter networks constrain the compositional structure itself (Shekarforoush et al., 2022), and other families leave sinusoids altogether (Ramasinghe & Lucey, 2022). None of these works measures where a bank network stops leaving the mean colour, or which layer carries that boundary.

Frequency generation by layer composition. A sinusoidal network expresses integer harmonic combinations of its initial frequencies, with a support that grows with depth (Yüce et al., 2022; Benbarka et al., 2022), and TUNER (Novello et al., 2025) holds that spectrum under control during training. This line explains how composition expands the frequency content of a network, and not whether the network fits its target at all.

Weight scaling. Yeom et al. (2025) multiply the initial weights of every layer but the last by a constant α, which trains SIREN about ten times faster at some cost in generalization. In our terms that scaling raises the phase scale of every layer. They report a trade-off between speed and generalization and not a failure to train, and their object is a single sinusoid while ours is a bank.

Spectral bias and input encodings. One line of work asks which frequencies a network fits first (Rahaman et al., 2019; Basri et al., 2020; Tancik et al., 2020; Jacot et al., 2018; Xu et al., 2019; Basri et al., 2019), and progressive encodings anneal those frequency bands during optimization (Lin et al., 2021; Park et al., 2021; Hertz et al., 2021). Both concern frequencies measured over the input domain, and Section 4.2 shows that the boundary we measure stays put when that domain is rescaled.

Signal propagation. Mean field theory describes trainability through the correlation map of a random layer (Poole et al., 2016; Schoenholz et al., 2017; Pennington et al., 2017; Xiao et al., 2018), and its order parameter is the slope of that map at a correlation of one. Section 4.4 uses the same map at the other end, where the slope depends on the slowest term of the bank, and Appendix E shows that the mean-field quantity separates our runs less well.

3 SETUP

The network and the phase scale. The network has L sinusoidal layers followed by a linear head, and layers 2 to L are the hidden layers. A sinusoidal layer sums τ sinusoids, $\sigma(z) = \sum_{i=1}^{\tau} \alpha_i \sin(\omega_i z + \phi_i)$, with frequencies ω_i, phases ϕ_i and amplitudes α_i that are drawn at initialization and trained afterwards. The set of frequencies is the bank. We write $[5] \times 8$ for a bank whose eight terms all sit at 5, and $[5] + [10] \times 7$ for one slow term at 5 with seven faster terms at 10. Unless stated otherwise $L = 5$, the width is 256, $\tau = 8$ with $\alpha_i = 1/\tau$ and zero phases, and the weights come from He initialization (He et al., 2015). Every layer starts from the same bank and then trains its own frequencies, phases and amplitudes, which means the layers share a starting point and not a parameter. Two fixed banks appear throughout. The low bank is $[2, 4, 8, 16]$ repeated twice, and the high bank is $[30] \times 8$. Random banks draw $\omega_i \sim U[0, 30)$. The phase scale is measured at initialization on the training coordinates, and $\kappa = \max_{l \geq 2} \kappa_l$.

The three settings, collapse and departure. Three training settings run through the paper. The CelebA setting fits one CelebA image (Liu et al., 2015) at 178×178 with five sinusoidal layers of width 256, He initialization, Adam (Kingma & Ba, 2015) at 10^{-3}, 2000 epochs, batch 4096 and a sigmoid head. The Kodak setting fits one Kodak image at 256×256 the same way, with a linear head. The STAF setting follows Morsali et al. (2026) on Kodak-24 at 256×256, with Glorot initialization (Glorot & Bengio, 2010), 16,000 optimizer steps at a learning rate of 10^{-3} decayed to 0.05 of its value, and the best of the checkpoints taken every 100 passes. A run has collapsed when its final PSNR sits within 0.5 dB of the PSNR of the mean colour of its target. A run departs at the first recorded epoch after 150 at which its PSNR passes the mean colour by more than 1 dB. The loss is recorded every 10 epochs, the runs traced in Section 6 are checked every 25 epochs, and Appendix B compares the two records.

Predictions written before the results. For each of 30 experiments we wrote down, before seeing any of its results, what we expected its runs to show, and Appendix A lists every prediction with its outcome. The predictions failed in whole or in part in 23 of the 30 experiments, and each failure is reported where it changed the paper.

4 A THRESHOLD IN THE HIDDEN LAYERS

A Monte Carlo over random banks shows that the phase scale separates the two outcomes. Two experiments then rule the first layer out. The correlation map of a random sinusoidal layer explains why the quantity has this form.

4.1 MONTE CARLO OVER RANDOM BANKS

We drew 100 random banks, 50 with $\tau = 5$ and 50 with $\tau = 8$, and trained each of them on three CelebA images under the CelebA setting, one seed per image. The outcome is bimodal on every image, and Figure 1(a) shows the 300 runs. A run either fits the face or outputs the mean colour of the target, and the failures sit at the PSNR of that mean colour, 12.59, 12.95 and 11.31 dB on the three images.

We find that the phase scale separates the two sides. Below $\kappa = 1.81$ every run trains, 60 of 60, 60 of 60 and 61 of 61 across the three images. We fixed that lower cutoff on the first two images and predicted that it would hold on the third, which we had not yet run. It held, 61 of 61.

The upper cutoff took two attempts. On the first two images no bank above 2.36 trains, and we wrote 2.36 down. The third image refuted it, with seven banks above 2.36 that train there. Over all three images the upper cutoff sits at $\kappa = 4.0$, and none of the 51 banks above it trains, 0 of 17 per image. Section 6.3 accounts for the difference between the images.

Between the cutoffs the two outcomes mix. Fitting one threshold to each image gives 2.31, 2.33 and 2.95, and each of them misclassifies 1 to 3 of the 100 banks. The lower cutoff marks where every run trains, and these thresholds mark where the error is smallest.

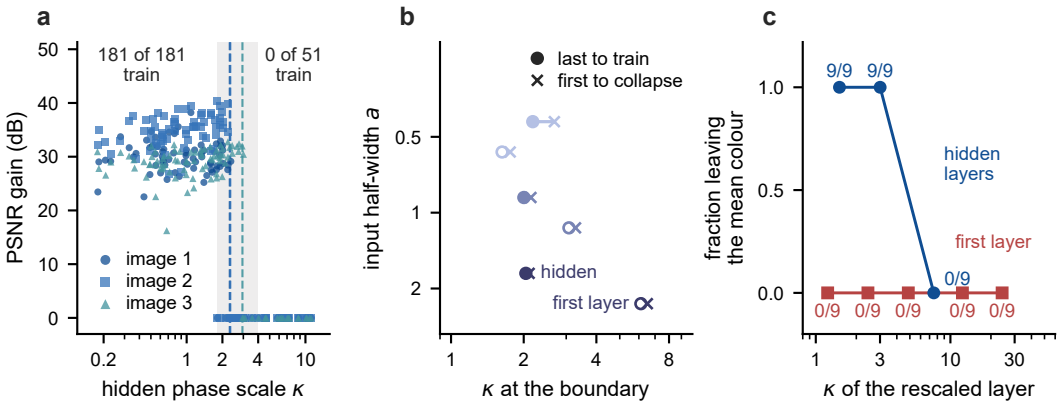

Figure 1: **The threshold sits in the hidden layers.** (a) PSNR above the mean colour against the hidden phase scale, 300 runs on three CelebA images. Dashed lines are the best threshold per image. (b) The input-range test. Open markers are the first-layer phase scale at the boundary, filled markers the hidden-layer value. (c) Per-layer rescaling. Fraction of runs that depart.

4.2 The input-range test falsifies the first-layer account

The natural account puts the failure in the first layer, which sees the coordinates directly. A sinusoid of frequency $\omega_{\min}$ over a coordinate range $[-a, a]$ completes $\omega_{\min} a/\pi$ periods, and once it completes many the integrals that carry the gradient average out, the cancellation the Riemann-Lebesgue lemma describes in the limit. This account predicts that the boundary in nominal $\omega_{\min}$ moves as $1/a$. We wrote that prediction down first, then ran the coordinate range at $a = 0.5$ and $a = 2$.

The boundary stayed where it was. In nominal $\omega_{\min}$, the last bank that trains and the first that fails sit at 3.98 and 4.36 for $a = 0.5$, and at 3.72 and 3.98 for both $a = 1$ and $a = 2$. The two phase scales at that same boundary behave differently. The first-layer value follows the coordinate range, from 1.63 and 1.76 at $a = 0.5$, to 3.08 and 3.28 at $a = 1$, to 6.12 and 6.50 at $a = 2$. The hidden-layer value stays in a narrow band, from 2.18 and 2.68 at $a = 0.5$ to 2.04 and 2.11 at $a = 2$.

The reason is that a hidden layer never sees the coordinates. Its input is the output of the previous sinusoid, which is bounded. The median standard deviation of that input stays between 0.52 and 0.54 across all three ranges, while the first layer has a median from 0.41 to 1.64. One bank of the 50 changed its outcome with the coordinate range, and it sits inside the mixed region, a count made after the outcomes were known.

4.3 Per-layer rescaling

The coordinate-range test changes the first-layer phase scale through the coordinates. Rescaling the first-layer weights changes it directly, and it gives the same answer. We took the high bank, which collapses under both Kodak settings. Under the Kodak setting, on three images and three seeds, we multiplied its first-layer weights by factors from 1 down to 0.05. The first-layer phase scale falls from 24 to 1.2, and 45 of 45 runs still end at the mean colour.

Rescaling the hidden layers instead changes the outcome. With the first layer left at its He value, a phase scale of 7.5 in the hidden layers gives 0 of 9 runs that train, 3.0 gives 6 of 9, and 1.5 gives 5 of 9.

Most failures of this grid below the threshold are late collapses, which Section 6.1 describes.

4.4 Why the variable has this form

The form of κ_l follows from the correlation map of a wide random sinusoidal layer, which Appendix F derives. At initialization the preactivations of two coordinates in hidden layer l are jointly Gaussian with mean zero, variance $q_l = \mathrm{std}(z_l)^2$ and a correlation c_l that depends on the pair. With

He kernels the next layer has

$$c_{l+1} = F_l(c_l) = \frac{K_l(c_l)}{K_l(1)}, \qquad K_l(c) = \sum_{i,j} \alpha_i \alpha_j \, e^{-q_l(\omega_i^2+\omega_j^2)/2} \, \sinh(q_l \, \omega_i \omega_j \, c), \tag{1}$$

which keeps $c = 1$ and pulls every correlation between zero and one toward zero. For a small c_l the map is linear, $c_{l+1} \approx \rho_l \, c_l$, with

$$\rho_l = \frac{q_l \, \mathbb{E}[\sigma'(z_l)]^2}{\mathbb{E}[\sigma(z_l)^2]}, \qquad \mathbb{E}[\sigma'(z_l)] = \sum_i \alpha_i \, \omega_i \, e^{-\omega_i^2 q_l/2}. \tag{2}$$

Each term of the mean derivative carries $e^{-\kappa^2/2}$ in the phase scale of its own frequency, and the diagonal terms of the power in the denominator have no such factor. The faster terms can contribute little to that sum, a condition that Appendix F.2 states and that the banks of Section 5.2 meet. The slowest term is then the only one left. With $\alpha_{\min}$ its total amplitude,

$$\rho_l \approx \frac{\alpha_{\min}^2}{\mathbb{E}[\sigma(z_l)^2]} \, \kappa_l^2 \, e^{-\kappa_l^2}, \qquad (F_L \circ \cdots \circ F_2)'(0) = \prod_{l=2}^{L} \rho_l. \tag{3}$$

To first order, a small correlation that enters the hidden stack leaves it multiplied by the product. The approximation has four consequences. First, hold the hidden layers at initialization and fit the head by least squares. That best head gains under 0.03 dB over the mean colour for the six banks above the boundary that Appendix F.4 checks, against a reference bound of 0.017 dB. Second, the factor $\kappa^2 e^{-\kappa^2}$ falls above a phase scale of 1, which is why we summarize a network by its largest hidden phase scale. Third, in this approximation the first layer enters only through q_2 and c_2, and the power of a sine approaches saturation once its phase scale passes 1. This fits the sweep of Section 4.3, where the hidden phase scale stayed at 30. Fourth, a deeper network has more factors. The map gives the form of the variable and not the threshold, which we measure in the next section.

5 Where the threshold sits

Where the threshold sits depends on the family of the bank, on the frequency of its fast terms and on the depth of the network.

5.1 The boundary by bank family

Figure 2(a) groups the runs by the kind of bank they use. For each kind we read off the largest phase scale at which a run still trained and the smallest at which one collapsed. Together these two values bracket the boundary for that kind of bank.

The bracket moves a long way. Random banks give 3.01 against 1.81. A bank whose eight terms all sit at one frequency gives 8.02 against 7.05, about three times as much. A bank with one slow term among seven at 30 gives 2.66 against 2.65. Two kinds never collapsed in the runs we have. One is $[5] + [10] \times 7$, up to 4.47. The other is a set of banks with several fixed frequencies, up to 2.31. Appendix C lists them by experiment. Where the two values cross, the phase scale does not separate the outcomes inside that kind of bank.

Inside the mixed region the fraction of random banks that train falls gradually. It goes from 29 of 36 to 6 of 22 to 1 of 10 as κ rises through 2.36, 3.00 and 4.00, pooled over the three CelebA images.

5.2 The frequency of the fast terms

The phase scale counts the slowest term of the bank, and the rest of the bank still matters. We fixed the slowest term at 5, set the other seven terms to a common frequency f, and swept f over 5, 8, 10, 15, 20 and 30 under the Kodak setting on three images. Across the whole sweep the phase scale stays between 4.4 and 5.0, because the slowest term and the preactivation scale barely move. The outcome changes all the same, and Figure 2(b) shows the sweep.

At $f = 5$, which is the single-frequency bank $[5] \times 8$, all 15 runs depart, at a median epoch of 510. At $f = 8$ all 3 depart at a median of 320, and at $f = 10$ all 9 depart at a median of 260. At $f = 15$ one run of 3 departs, at epoch 1210. At $f = 20$ none of 3 departs, and at $f = 30$ none of 6 does.

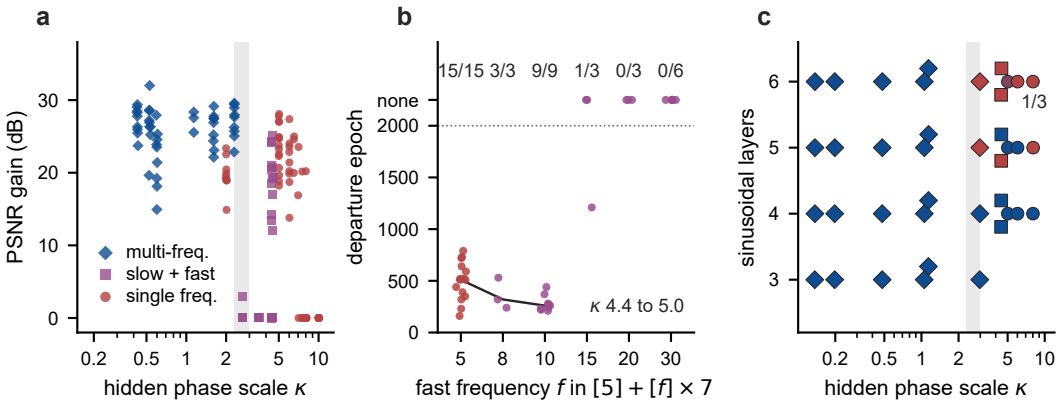

Figure 2: **Where the threshold sits.** (a) Outcome against the phase scale, by bank family. (b) Departure epoch as the frequency of the fast terms changes, with the slowest term held at 5. The top row holds the runs that do not depart within 2000 epochs. (c) Outcome against the phase scale and the depth, with mixed cells counted.

The trend is not monotone. Moderate fast terms make the network leave the mean colour earlier than no fast terms at all, and only past $f = 15$ do they delay the departure and then prevent it. All amplitudes start at $1/\tau$ and stay trainable, which makes the sweep an axis in initial frequency and not in initial power.

5.3 Depth

The same bank changes outcome when the network gets deeper. We trained banks at depths of 3, 4, 5 and 6 sinusoidal layers under the Kodak setting. Banks with a phase scale of 1.14 and below train at every depth. A bank at $\kappa \approx 2.98$ trains 3 of 3 at depths 3 and 4 and collapses in all three runs at depths 5 and 6. Among the five banks between $\kappa = 4.4$ and $\kappa = 8.1$, all five train 3 of 3 at depth 4, three of them still train at depth 5, and at depth 6 only $[5] \times 8$ survives, at 1 of 3.

The threshold is therefore a surface over the phase scale and the depth, which Figure 2(c) draws. The direction matches the product of Section 4.4 and the depth trend of Yeom et al. (2025), whose best weight-scaling factor falls as the network deepens.

6 What a collapsed run is doing

A run at the mean colour can still leave it late. While it sits there the weights of its output head have shrunk to a fraction of their initial norm. Whether the run ever leaves depends in part on where its output bias started.

6.1 Late departures and late collapses

The single-frequency family shows how late a departure can come. $[5] \times 8$ leaves the mean colour at a median epoch of 510, $[6.5] \times 8$ at 570 and $[7.5] \times 8$ at 770, while at $\omega = 10$ none of 11 runs leaves at all. Five times the budget, 10,000 epochs, left all 6 runs of $[8] \times 8$, $[10] \times 8$ and $[5] + [30] \times 7$ where they were.

A departure is not a crossing of the boundary. Traced over 2000 epochs, $[5] \times 8$ reaches 43.4 dB with a phase scale that never leaves the range 4.69 to 5.03. The runs that fail raise theirs only after they are already stuck.

A network can also fit the image first and fall back to the mean colour later, and this is a different event. It appears in the rescaled grid of Section 4.3. There the high bank was brought below the boundary by shrinking its hidden weights, not by changing its frequencies, and training undoes that. The second-layer weight norm grows from 1.66 back to between 25.5 and 36.7, the phase scale follows it from about 2.4 up past 6, and the output falls back to the mean colour. Of the 14 collapses

in that grid at a phase scale of 3 or below, 13 follow this path. Appendix D gives the traces. A bank that contains a slow frequency cannot drift this way, because its slowest frequency is fixed and the input of a hidden layer is bounded.

6.2 THE OUTPUT HEAD COLLAPSES FIRST, AND THE HIDDEN LAYERS CHANGE DIRECTION

At initialization the output of these networks is worse than a constant image. On one Kodak image it starts between 6.9 and 8.5 dB against 21.1 dB for the mean colour. The fastest way to cut the loss from there is for the head to give up its weights and let the bias carry a constant. That is what the runs do. In $[5] \times 8$ the weights of the head are down to 0.266 of their initial norm by epoch 25, and the output reaches the mean colour by epoch 50.

How far the head falls separates the two outcomes. Among the nine traced runs of Experiment 5c that later depart, the head bottoms out between 0.20 and 0.28 of its initial norm, early, between epochs 50 and 175. The six runs that never depart go further down, to between 0.15 and 0.19, and they get there between epochs 925 and 1925. The low bank only falls to between 0.49 and 0.54. At departure the head has recovered to between 1.19 and 2.31 times its own low point.

The hidden layers keep the size of their weights and change what they extract. Over the 13 traced runs of Experiment 5h that depart, the mean hidden weight norm at departure sits between 1.003 and 1.042 of its initial value. The phase scale itself rises by at most 16%. The largest single kernel has meanwhile moved between 0.10 and 0.44 of its own norm. The runs that stay do not hold their phase scale, and five of them end between 2.0 and 4.9 times theirs.

The features this motion produces do not settle. While a run sits at the mean colour, the last hidden layer's features have a linear CKA of 0.11 with the same features 25 epochs earlier. That is the value expected for two independent random feature sets of that size. The low bank holds between 0.64 and 1.00 from epoch 50, and the runs that depart settle to between 0.88 and 1.00 as they go.

6.3 ABOVE THE THRESHOLD, THE STARTING VALUE OF THE OUTPUT BIAS CHANGES THE OUTCOME

Above the boundary the bank does not decide the outcome by itself. Seven banks with a phase scale between 2.42 and 3.01 train on the third CelebA image and none of them trains on the second. A 2×2 over two banks and two seeds puts the difference in the image and not in the weight draw. Two accounts of that difference failed, each written down before its runs. Scaling the contrast of either image over three levels left both banks where they were. Over nine further images two banks at $\kappa = 3.00$ and $\kappa = 2.93$ swapped outcomes twice, which rules out any account that looks at the target alone.

What matters about the image is its mean colour, and what matters about the network is where its output bias starts. The output is a weighted sum of the last hidden layer's features plus a bias, and the bias alone can produce a constant. For the mean colour to form, that bias has to travel from its starting value to the value that produces the mean. Under a sigmoid head the distance it travels is $|\mathrm{logit}(m) - b_0|$, for a target mean m and an initial bias b_0 averaged over the three channels. Under a linear head it is $|m - b_0|$.

Ordered by that distance, the 84 runs of Figure 3(c) separate. None of 42 trains at a distance of at most 0.5. Two of 10 train between 0.5 and 0.95. All 32 train at 0.95 or more.

The intervention settles the direction. We build one set of initial weights per pair and change only the starting value of the output bias. All 18 pairs reverse, across three settings of six pairs each. Moving the start onto $\mathrm{logit}(m)$ turns 6 runs that train into 6 that collapse. Moving it one unit away turns 6 collapses into 6 runs that train.

The bias distance is also our account of why the upper cutoff of Section 4.1 moved when the third image was added and stays put under the two Kodak settings.

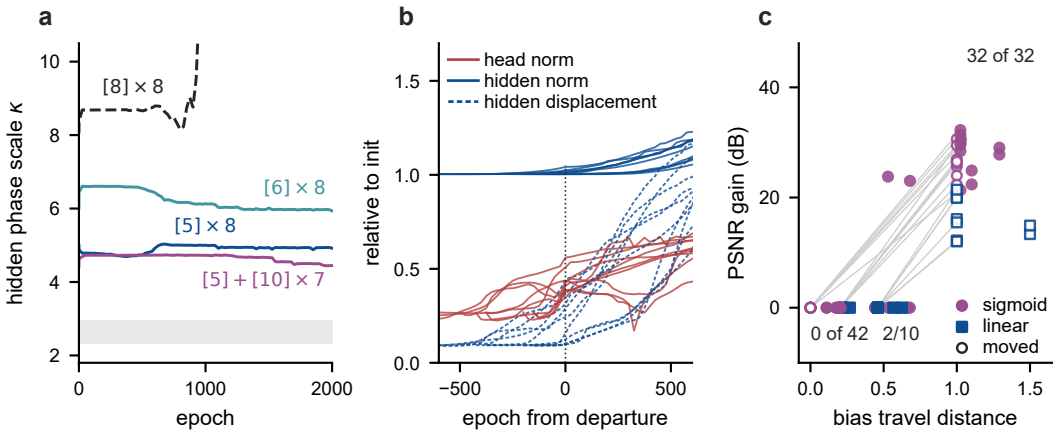

Figure 3: **What a run does while it sits at the mean colour.** (a) The hidden phase scale over 2000 epochs for four banks. The shaded band is the random-bank threshold range, and no run enters it. (b) Ten runs above the threshold that train, aligned at departure. Red is the head norm and solid blue the mean hidden norm, each relative to its initial value. Dashed blue is the largest displacement of a hidden kernel relative to its initial norm. (c) Outcome against the distance the output bias must travel, 84 runs. Paired runs share their initial weights.

7 Consequences

7.1 STAF's seed variance

STAF draws a fresh bank in every layer, which makes its phase scale a property of the draw. Under the STAF setting on Kodak-24, 120 runs of the activation in a five-layer network without a residual path collapse 56 times and average 24.78 ± 9.99 dB. STAF's default three-layer network with its residual path collapses in none of 120 and averages 32.34 ± 2.85 dB. Section 7.2 separates the depth from the residual path.

These runs did not record their phase scale, and we rebuild each one. STAF draws its frequencies, phases and amplitudes from the run's seed. That lets us recreate the activation exactly and measure its phase scale over three fresh Glorot weight draws. Ordered by that estimate, the risk of collapse rises. None of the 10 runs below 2.31 collapses, 6 of 20 collapse between 2.31 and 2.95, 8 of 28 between 2.95 and 4.0, and 42 of 62 above 4.0. The separation is softer than the one in Section 4.1. The best single threshold misclassifies 31 of 120 runs at 3.71, and 20 runs above 4.0 train anyway.

The amplitude draw belongs to that risk. One seed gives $[30] \times 8$ amplitudes that nearly cancel and a phase scale of 1.23, and with that draw it trains on all 24 images where uniform amplitudes collapse on all 24.

7.2 What removes the collapse

A residual path adds the input of a layer to its output, and it removes the collapse at a fixed depth. With five sinusoidal layers on three images and three seeds, $[30] \times 8$ collapses in 9 of 9 runs without the path and in 0 of 9 with it, reaching 24.42 ± 1.53 dB. The path is not free. The low bank, which never collapses either way, loses 7.4 dB when the path is added. With $h_l = \sigma(z_l) + z_l$, the residual form of He et al. (2016), the derivative with respect to the previous layer is $(D_{\sigma'} + I)W$, which keeps a path through W when the mean of σ' approaches zero.

Two other configurations avoid the failure without a residual path. Under the STAF setting the low bank has a phase scale of 0.86 and collapses in 0 of 24 runs, at 38.74 ± 1.98 dB. SIREN's initialization holds the phase scale near 1 by construction, with a measured value between 0.97 and 1.00 on four signals. At a learning rate of 10^{-4} it reaches 48.23 ± 1.11 dB, above every bank under the STAF setting. We compare on trainability and not on fit quality, and Appendix I shows that a lower learning rate explains neither bank result.

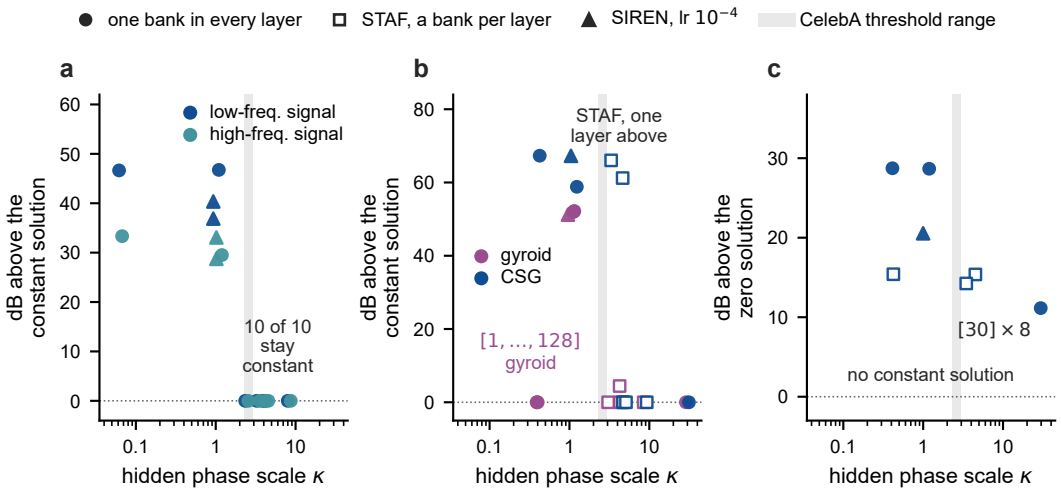

Figure 4: **Three other signals on the axis of Figure 1.** (a) audio, (b) two shapes, (c) the Burgers equation. The vertical axis is the gain over the constant output. Filled circles start every layer from the same bank, open squares are STAF networks and triangles are SIREN at 10^{-4}. The grey band is the CelebA range.

7.3 Beyond images

The same measurement carries to three other signals, with each setting fixed before the runs, and Figure 4 puts them on the axis of Section 4.1. We saw these runs before any prediction existed for them, and we mark the experiment as exploratory. Their phase scale comes from rebuilding each initialization, which Appendix H describes.

On two audio signals the switch falls between 1.14 and 2.44. Four networks at or below 1.14 all train, between 44.8 and 55.3 dB of SNR, and ten at or above 2.44 all stay at the constant output. The bracket agrees with the image boundary.

The two shapes are given as signed fields, a gyroid level set and a solid built from boolean operations. There the bracket is far wider, between 1.20 and 30. The low bank reaches an IoU of 0.993 to 0.994 in four runs and $[30] \times 8$ stays at the constant output on both shapes. One configuration below the boundary fails on one shape, and SIREN at 10^{-3} fails on audio and on both shapes despite a phase scale near 1. A learning rate of 10^{-4} removes the SIREN failure, and Appendix G gives the other.

The Burgers equation marks the boundary of the account. In a physics-informed network (Raissi et al., 2019; Wang et al., 2022; 2021) the initial condition pins the output at one edge of the domain. A constant output cannot satisfy that condition, and there is nothing to collapse into. The phase scale shows up in the size of the error instead. The low bank and $[1, \ldots, 128]$ reach a relative L^2 error of 0.037, while $[30] \times 8$ at a phase scale of 29.7 reaches 0.277. The error of the three STAF networks does not follow their phase scale.

8 Conclusion

A sinusoidal neural field built on a bank has two possible outcomes. It fits its target, or it ends at the mean colour of that target. The largest phase scale over the hidden layers predicts which of the two, from one forward pass before training and without the target. The failure belongs to those layers. Rescaling the first layer does not remove it, and the correlation map of a random sinusoidal layer gives the phase scale its form.

Where the boundary falls is measured, and it depends on the kind of bank and on the depth of the network. A failed network reaches the mean colour by shrinking its output head, and inside the mixed region the output bias can send the same network either way. SIREN's initialization and any bank with one slow frequency keep the phase scale low by construction.

REPRODUCIBILITY STATEMENT

Appendix B gives the three settings in full, Appendix A lists every experiment and every prediction with its outcome, and Appendix H lists every cross-domain run. We will release the code, the results files behind every number in the paper and the scripts that produce every figure when the paper is published.

THE USE OF LARGE LANGUAGE MODELS

Large language models were used throughout this project under the direction of the authors, in four ways. They took part in research ideation, proposing hypotheses and writing down the predictions of Appendix A before seeing the results they concern. They took part in research execution, writing the training, analysis and plotting code. They proved the mathematical claims of Appendix F and Appendix E, including the correlation map and its two propositions. They drafted and edited every section of the text and checked the quoted numbers against the stored results, and a second language model reviewed the manuscript. The authors directed the work, decided which experiments to run, and take full responsibility for the content.

REFERENCES

Ronen Basri, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density. In *International Conference on Machine Learning*, pp. 685–694, 2020.

Nuri Benbarka, Timon Höfer, Hamd ul Moqeet Riaz, and Andreas Zell. Seeing implicit neural representations as Fourier series. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2283–2292, 2022.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8628–8638, 2021.

Emilien Dupont, Adam Goliński, Milad Alizadeh, Yee Whye Teh, and Arnaud Doucet. Coin: Compression with implicit neural representations. *arXiv preprint arXiv:2103.03123*, 2021.

Rizal Fathony, Anit Kumar Sahu, Devin Willmott, and J Zico Kolter. Multiplicative filter networks. In *International Conference on Learning Representations*, 2021.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-Hornung, and Daniel Cohen-Or. SAPE: Spatially-adaptive progressive encoding for neural optimization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8820–8832, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5721–5731, 2021.

David B Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. Bacon: Band-limited coordinate networks for multiscale scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16231–16241, 2022.

Zhen Liu, Hao Zhu, Qi Zhang, Jingde Fu, Weibing Deng, Zhan Ma, Yanwen Guo, and Xun Cao. Finer: Flexible spectral-bias tuning in implicit neural representation by variable-periodic activation functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2713–2722, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4455–4465, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020.

Alireza Morsali, MohammadJavad Vaez, Mohammad Hossein Soltani, Amirhossein Kazerouni, Babak Taati, and Morteza Mohammad-Noori. A unified theory of sinusoidal activation families for implicit neural representations. *Transactions on Machine Learning Research*, 2026. First circulated as STAF: Sinusoidal trainable activation functions for implicit neural representation, arXiv:2502.00869, 2025.

Tiago Novello, Diana Aldana, Andre Araujo, and Luiz Velho. Tuning the frequencies: Robust training for sinusoidal neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3071–3080, 2025.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 165–174, 2019.

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5845–5854, 2021.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in Neural Information Processing Systems*, volume 30, 2017. arXiv:1711.04735.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, volume 29, 2016. arXiv:1606.05340.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310, 2019.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Sameera Ramasinghe and Simon Lucey. Beyond periodicity: Towards a unifying framework for activations in coordinate-mlps. In *European Conference on Computer Vision*, pp. 142–158, 2022.

Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G Baraniuk. Wire: Wavelet implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18507–18516, 2023.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017. arXiv:1611.01232.

Shayan Shekarforoush, David B. Lindell, David J. Fleet, and Marcus A. Brubaker. Residual multiplicative filter networks for multiscale reconstruction. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8550–8563, 2022.

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7462–7473, 2020.

Yannick Strümpler, Janis Postels, Ren Yang, Luc Van Gool, and Federico Tombari. Implicit neural representations for image compression. In *European Conference on Computer Vision*, pp. 74–91, 2022.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020.

Sifan Wang, Hanwen Wang, and Paris Perdikaris. On the eigenvector bias of Fourier feature networks: From regression to solving multi-scale PDEs with physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 384:113938, 2021.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why pinns fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel S. Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, pp. 5393–5402, 2018. arXiv:1806.05393.

Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 41(2):641–676, 2022.

Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *International Conference on Neural Information Processing*, pp. 264–274, 2019.

Taesun Yeom, Sangyoon Lee, and Jaeho Lee. Fast training of sinusoidal neural fields via scaling initialization. In *International Conference on Learning Representations*, 2025. arXiv:2410.04779.

Gizem Yüce, Guillermo Ortiz-Jiménez, Beril Besbinar, and Pascal Frossard. A structured dictionary perspective on implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19206–19216, 2022.

A THE EXPERIMENTS AND THEIR PREDICTIONS

Before each of the 30 experiments that carried predictions we wrote down which runs the experiment contained, what we expected each of them to show, and what would count against that expectation. Nothing was edited afterwards, and we will release that file with the code when the paper is published. Every set of predictions but two was written before its runs were launched. The sets for

Experiments 0 and 3 were written after their first runs had started and before any result had been seen. Table 2 in Section A.1 lists every prediction with its outcome.

A prediction can state several outcomes. It held when all of them occurred, held in part when some did, and was refuted when none did. One counterexample refutes an outcome stated for every run of a group, and a prediction is untested when no run could decide it. Of the 105 predictions, 45 held, 36 were refuted, 20 held in part and 4 could not be tested. Table 1 gives the counts for each experiment. Seven experiments had no failure, seven had a prediction that held in part and none that was refuted, and the remaining sixteen contain at least one refuted prediction. The main text reports each failure where it changed a section, and the failures of Experiments 0 and 1 carry the argument of Section 4.

Table 1 also gives what each experiment varies and how many runs it holds. The 1859 training runs of the paper are the rows of the first five groups apart from the two 12-epoch runs of Experiment 4f, which record a loss trace and no score. The last group holds the two measurements that rebuild networks without training them.

Table 1: Every experiment, its run count, and how many of its predictions held, held in part, were refuted or were untested.

Experiment	What varies	Runs	Held	Partly	Refuted	Untested
0	High bank with the first layer at 1 to 0.05, one cell with both at 0.1, low bank, He and Glorot, three Kodak images, three seeds	81	2	0	4	0
0b	High bank with the hidden layers rescaled, five cells	45	0	1	5	0
0c	Four runs of the cell at hidden scale 0.05, checked every 25 epochs	4				
1	100 random banks on CelebA image 1, the $\tau = 5$ banks also at $a = 0.5$ and $a = 2$, one reference run	201	1	0	2	0
1r	The same banks on image 2 with seed 7	101	2	0	2	0
1s	The same banks on image 3 with seed 13	101	1	0	3	0
1t	Two banks, two images, two seeds	8				
1u	Two banks over a contrast sweep of two images	12	0	0	3	0
1v	Two banks on nine further images	18	1	0	4	1
1w	Linear head on five images and affine targets under the sigmoid head	14	3	0	0	0
1x	Four settings of the initial head bias	8	4	0	0	0
1y	18 pairs on identical initial weights	36	4	0	0	0
2	STAF setting on Kodak-24, six bank arms, SIREN, STAF with and without the residual path over five seeds	408	1	1	1	1
2b	SIREN at three depth and learning-rate settings	72				
2c	The two fixed banks at lower learning rates	72				
3a	Low and random banks at equal τ, eight images, three seeds	96	0	1	0	0
3a2	Ten random banks on eight images	80	2	0	1	1
3b	Random banks with and without two slow terms	75	0	1	0	0
3c	Both fixed banks with and without the residual path	36	0	1	0	0
3d	Frozen frequencies or frozen weights	27	0	1	0	0
3e	Eight deterministic banks	72	0	1	0	0
3f	Six banks at depths 3 to 6	96	1	0	0	0
3g	Six banks with one slow term or one frequency	18	2	2	0	0
4a, 4b, 4c	Audio, Burgers and signed distance functions	41				
4d	SIREN at its own learning rate and further gyroid seeds	16	2	1	1	0

Table 1 continued.

Experiment	What varies	Runs	Held	Partly	Refuted	Untested
4e	First-layer scale swapped between two banks	2	0	0	1	0
4f	Two 12-epoch runs, no score	2				
5a	Five banks over their first 100 epochs	15	1	2	2	0
5b	The same banks at depths 4 and 6	30	3	0	0	0
5c	The same banks at depth 5, checked every 25 epochs	15				
5d	Four banks for 10,000 epochs	8	2	0	2	0
5e	Single-frequency banks at 6.5, 7 and 7.5	9	2	0	0	1
5f	Fast terms at 8, 15 and 20	9	4	0	0	0
5g	Two boundary banks and the low bank, traced	6				
5h	Four banks and the low bank with displacement and CKA records	15	0	3	3	0
5i	Three random banks on each side of the threshold	12	2	1	1	0
2g	Phase scale at initialization of every STAF run of Experiment 2 and of its six fixed-bank arms	246	3	1	1	0
4g	Phase scale at initialization of every run of Experiment 4	59	2	3	0	0

Experiments 4a, 4b and 4c ran before any prediction existed for them, and we had seen their results when the predictions for Experiment 3 were written. The main text marks the cross-domain experiment as exploratory for that reason. Experiments 0c, 1t, 2b, 2c, 4f and 5g were diagnostic runs added after a result, with a stated purpose and no numbered prediction. Experiment 5c carried no new prediction and tested the escape hypothesis written for Experiment 5a, which it refuted.

Some analyses were run on stored results after their outcomes were known. They are the departure counts over every Kodak run in Appendix D, the count of Experiment 1 banks that change outcome across input ranges in Section 4.2, the comparison of the phase scale with the Gaussian mean and the mean-field quantity in Appendix E, the map-quantity comparison in the same appendix, the least-squares head and the other numerical checks of Appendix F, the per-layer reading of the STAF networks in Appendix G, and the settling times of Appendix K.

A.1 Every prediction

Table 2 lists the 105 predictions in the order they were written, with the outcome of each. A primed label marks a prediction that replaced an earlier line before the runs it concerned had given a result, and the earlier line stays in the table with its own outcome. Two groups, for Experiments 4g and 2g, are measurements made on runs whose outcomes were already known, and their predictions concern the measured values.

Table 2: Every prediction with its outcome.

	Prediction	Outcome
Experiment 0		
P0.1	The high bank under He with its first layer at the He scale fails below 20 dB on all nine runs	Refuted as stated. All 9 sit at the mean colour, which is 21.10 dB on kodim02, and 6 of 9 lie below 20 dB
P0.2	The first layer at 0.1 with hidden layers at He rescues most runs	Refuted, 9 of 9 collapse
P0.3	Glorot initialization rescues the high bank	Refuted, 0 of 9 with uniform amplitudes. Replaced by P0.3′ before any Glorot result
P0.4	The first layer and the hidden layers both at 0.1 fail or stay far below the low bank	Refuted, 6 of 9 train at 31 to 45 dB
P0.5	The low bank trains under He and under Glorot	Held, 18 of 18
Experiment 1		

Table 2 continued.

	Prediction	Outcome
P1.1	At $a = 1$ the outcome is bimodal with a boundary in $\omega_{\min}$ between 3 and 5	Held, boundary between 3.72 and 3.98
P1.2	The boundary in $\omega_{\min}$ doubles at $a = 0.5$ and halves at $a = 2$	Refuted, 3.98 to 4.36 at $a = 0.5$ and 3.72 to 3.98 at both other ranges
Experiment 2		
P2.1	SIREN reaches about 28 to 32 dB under the STAF protocol	Refuted, 15.81 ± 3.49 dB under that protocol. At a learning rate of 10^{-4} it reaches 48.23 dB
P2.2	STAF without the residual path has a spread near 9 dB, and with it a higher mean and a smaller but still large spread	Partly. 9.99 dB and 32.34 dB held. The spread with the residual path fell to 2.85 dB, and the entry gave no number for still large
P2.3	The high bank under Glorot reproduces the earlier Kodak column, and under He it fails on every image	Untested as written. Its two amplitude draws became P2.3′ before Experiment 2 ran
Revised after two runs of Experiment 0		
P0.3′	Glorot with uniform amplitudes fails under the base protocol	Held, 9 of 9
P2.3′	The seed-42 Laplace amplitudes reproduce the earlier Kodak column above 35 dB and uniform amplitudes fail on every image	Held, 39.60 dB with 0 of 24 against 24 of 24
Added after Experiment 0, before Experiments 0b, 1 and 2 ran		
P0b.1	The first layer at 0.1 with hidden layers at 0.04 trains in at least 8 of 9 runs	Refuted, 6 of 9
P0b.2	The first layer at 0.1 with hidden layers at 0.02 trains, at a lower PSNR than 0.04	Refuted, 8 of 9 train and the mean PSNR is higher
P0b.3	The first layer at He with hidden layers at 0.05, 0.10 or 0.25 fails in every run, because the first layer alone violates the criterion	Refuted, 0.05 and 0.10 train 5 of 9 and 6 of 9. Only 0.25 fails, 0 of 9
P1.3	The first-layer phase scale at the boundary sits near π for every input range	Refuted, 1.63, 3.08 and 6.12
Revised after Experiment 1, before Experiment 0b ran		
P0b.1′	The first layer at 0.1 with hidden layers at 0.04 trains	Refuted, 6 of 9. Two fit and then fell back, one never started
P0b.2′	The first layer at 0.1 with hidden layers at 0.02 trains	Refuted, 8 of 9. One fit and then fell back
P0b.3′	With the first layer at He, hidden 0.05 trains, 0.10 is mixed and 0.25 fails	Partly. 5 of 9, 6 of 9 and 0 of 9
Experiment 3, written after Experiment 3a had started and before any result		
P3a	Random banks collapse in 35 to 55 percent of runs, the low bank never, and trained random banks sit within 3 dB of it	Partly. The rate could not be tested with three banks per arm, the low bank collapsed in 0 of 24, and one of the five trained banks sits 3.2 dB below it
P3b	Two added slow terms remove every collapse, pure random banks collapse in about 30 percent, both within 3 dB of the low bank	Partly. 0 of 36 and 0.4 dB held, pure random banks collapsed 3 of 36 from one bank
P3c	The residual path lets the high bank train at a lower PSNR and changes the low bank by under 2 dB	Partly. 0 of 9 at 24.42 dB held, the low bank lost 7.4 dB
P3d	Frozen frequencies within 1 dB of full learning, frozen weights below 25 dB, frequency shifts under 5 percent	Partly. 0.4 dB and 19.15 dB held, shifts of 2.6 to 6.5 percent
P3e	Banks with $\omega_{\min} \leq 4$ train every run, banks with $\omega_{\min} \geq 8$ collapse every run, and the range between is mixed	Partly. Six banks and $[10] \times 8$ held, $[5] \times 8$ trains 9 of 9
P3f	Random-bank collapse rises with depth and the low bank collapses at no depth	Held, 0, 0, 3 and 3 of 15 across depths, and 0 of 36
Experiment 4d		
P4d.1	SIREN at 10^{-4} passes 40 dB on both audio signals	Held, 45.5 and 44.1 dB
P4d.2	SIREN at 10^{-4} passes an IoU of 0.9 on both shapes, and three layers at 10^{-3} as well	Partly. 0.994 and 1.000 held, three layers at 10^{-3} give 0.000 and 0.002
P4d.3	SIREN at 10^{-4} reaches a relative error below 0.1 on the Burgers equation	Held, 0.094

Table 2 continued.

	Prediction	Outcome
P4d.4	$[1, \ldots, 128]$ trains on the gyroid in two of three seeds at each learning rate	Refuted at 10^{-3}, 0 of 3. At 10^{-4} one run reaches 0.973, which cannot decide two of three
Experiment 4e		
4e	One of two readings holds. Either the first-layer scale is the cause and swapping it flips both banks, or the hidden variance is the cause and swapping changes nothing	Refuted, both swapped configurations train at 0.993
Experiment 3a2		
P3a2.1	Every bank below 2.0 trains on 8 of 8 images	Held, three banks
P3a2.2	Every bank above 2.7 collapses on 8 of 8	Held, seven banks
P3a2.3	Only a bank between 2.0 and 2.7 splits across images	Untested, no bank fell inside
P3a2.4	Trained banks sit within 3 dB of the low bank	Refuted, two within 2.3 dB and one 6.4 dB below
Experiment 3g		
P3g.1	$[6] \times 8$ trains 3 of 3 and $[8] \times 8$ at least 1 of 3	Partly. 3 of 3 held, $[8] \times 8$ 0 of 3
P3g.2	$[5] + [30] \times 7$ collapses 3 of 3	Held
P3g.3	$[3] + [30] \times 7$ trains 3 of 3 and $[4] + [30] \times 7$ is mixed or collapsed	Partly. The first collapses 2 of 3 with one partial fit, the second collapses 3 of 3
P3g.4	$[5] + [10] \times 7$ trains 3 of 3	Held
Experiments 5a and 5b		
P5a.1	$[5] \times 8$ and $[6] \times 8$ bring their phase scale below 3 within 10 epochs	Refuted, it rises to 5.45 and 6.62 by epoch 99
P5a.2	$[8] \times 8$ stays above 3 and at the mean colour	Held, and every bank sits at the mean colour at epoch 100
P5a.3	$[5] + [10] \times 7$ drops below 3 and $[5] + [30] \times 7$ does not	Partly. The first rises from 4.43 to 4.55, the second from 4.42 to 4.61
P5a.4	In every run whose phase scale falls the hidden norms fall with it, and in every run whose phase scale does not fall they are flat or rising	Partly. Two runs end 0.3 and 0.7 percent below their starting phase scale with hidden norms up by 0.3 to 0.5 percent, and the other 13 rise with rising norms
Escape	A successful run lowers its phase scale into the band before it departs	Refuted by Experiment 5c, no run enters 2.0 to 2.7
P5b.1	The six-layer boundary lies below the five-layer one, with $[6] \times 8$ the likelier to fail first	Held, $[6] \times 8$ 0 of 3 and $[5] \times 8$ 1 of 3
P5b.2	$[5] + [10] \times 7$ collapses at six layers	Held, 0 of 3
P5b.3	$[8] \times 8$ or $[5] + [30] \times 7$ trains at four layers	Held, both 3 of 3
Experiment 5d		
P5d.1	$[8] \times 8$ departs before epoch 10000 on at least one image	Refuted
P5d.2	$[10] \times 8$ does not depart	Held
P5d.3	$[5] + [30] \times 7$ does not depart	Held
P5d.4	The random bank at the Monte Carlo boundary does not depart	Refuted, departs at 160 on both images, 55.50 and 46.98 dB
Experiment 5e		
P5e.1	The median departure rises across 6.5, 7 and 7.5	Held, 570, 630 and 770
P5e.2	$[7] \times 8$ departs on at least one image	Held, 2 of 3
P5e.3	A run departing after epoch 1500 ends below 30 dB	Untested
Experiment 5f		
P5f.1	The departure epoch rises across f of 8, 15 and 20	Held, 320, 1210 and none
P5f.2	$f = 15$ departs on at least one image	Held, at 1210
P5f.3	$f = 20$ departs after epoch 800 or not at all	Held, none of 3
P5f.4	$f = 8$ departs earlier than $[5] \times 8$	Held, 320 against 515
Experiment 1r		
P1r.1	Every bank below 2.0 trains, at the rate of the first image	Refuted, 64 of 66 against 118 of 119
P1r.2	Every bank above 2.7 collapses	Held, 0 of 25
P1r.3	Any bank that flips lies between 2.0 and 2.7	Refuted, a bank at 1.89 flips
P1r.4	28 successes at $\tau = 5$, within 4	Held, 27
Experiment 1s		
P1s.1	Every bank below 1.81 trains	Held, 61 of 61
P1s.2	No bank above 2.36 trains	Refuted, 7 train
P1s.3	Any flip lies between 1.81 and 2.36	Refuted

Table 2 continued.

	Prediction	Outcome
P1s.4	At most 4 banks disagree with the first image	Refuted
Experiment 1u		
P1u.1	The third image scaled down to the contrast of the first collapses both banks	Refuted, both train
P1u.2	The first image scaled up trains both banks	Refuted, both collapse
P1u.3	Both banks flip at the same side of the sweep	Refuted, neither bank flips anywhere in the sweep
Experiment 1v		
P1v.1	The high-frequency fraction orders the outcome over twelve images	Refuted
P1v.2	The four images richest in high frequencies train	Refuted
P1v.3	The three images poorest in them collapse	Held
P1v.4	Image 000016 collapses	Refuted, both banks train
P1v.5	Image 000008 sits at the cutoff and is not called	Untested
P1v.6	The two banks agree on every image	Refuted, opposite outcomes on two images
Experiment 1w		
P1w.1	With a linear head all ten runs collapse	Held
P1w.2	The first image mapped to a mean of 0.25 trains under the sigmoid head	Held, 41.00 and 43.52 dB
P1w.3	The third image mapped to a mean of 0.5 collapses	Held
Experiment 1x		
P1x.1	The sigmoid bias started at the logit of the mean collapses both banks on the third image	Held
P1x.2	The sigmoid bias started 1.0 below that logit trains both on the first image	Held, 43.21 and 38.23 dB
P1x.3	The linear bias started 1.0 below the mean trains both on the first image	Held, 33.45 and 28.69 dB
P1x.4	The linear bias started 1.5 below trains both	Held, 25.95 and 27.50 dB
Experiment 4g, a measurement on known outcomes		
P4g.0	The rebuild reproduces the seven stored phase scales	Held, to every decimal
P4g.1	The audio phase scales fall in the named ranges	Held
P4g.2	The SDF phase scales fall in the named ranges, and the two STAF networks that recover the CSG shape have the two smallest values, both below 2.95	Partly. The ranges held and the two are the smallest, at 3.19 and 4.44
P4g.3	The Burgers phase scales fall in the named ranges, the STAF networks sit above the low bank, and the error rises with the phase scale	Partly. One STAF network sits at 0.42, and the error rises over the bank arms only
P4g.4	No network above 2.95 leaves the constant level on audio or on the shapes, and every network below 1.81 trains apart from two known cases	Partly. Two STAF networks above 2.95 reconstruct the CSG shape, the second part held
Experiment 2g, a measurement on known outcomes		
P2g.1	The median phase scale of the STAF runs that collapse exceeds that of the runs that train	Held, 5.78 against 3.16
P2g.2	One threshold misclassifies at most 12 of 120	Refuted, 31
P2g.3	Runs with no hidden layer above 2.95 train, runs with two or more collapse, runs with exactly one are mixed, and the count separates at least as well as one threshold	Partly. 6 of 30 with none collapse and 9 of 45 with two or more train. The exactly-one class is mixed, 14 of 45 collapse, and the count errs on 29 runs against 31
P2g.4	At least 20 of STAF's default networks start with a layer above 2.95 and none collapses	Held, 76 and 0 of 120
P2g.5	The Laplace-amplitude high bank sits below 2 and the uniform one above 7	Held, 1.23 and 21.25
Experiments 1y and 5h		
P1y.1	Sigmoid head on the third image, the bias moved onto the logit of the mean. In all six networks the default run trains and the moved run collapses	Held, 6 of 6
P1y.2	Sigmoid head on the first image, the bias moved 1.0 below that logit. In all six the default run collapses and the moved run trains	Held, 6 of 6

Table 2 continued.

	Prediction	Outcome
P1y.3	Linear head on the first image, the bias moved 1.0 below the mean. In all six the default run collapses and the moved run trains	Held, 6 of 6
P1y.4	Paired runs share their hidden phase scale to 10^{-4}	Held
P5h.1	At departure every hidden kernel has moved by under 10 percent and the head by over 50	Partly. The hidden kernels moved by 0.10 to 0.44 of their norm, the head by 0.74 to 1.03
P5h.2	At departure every hidden layer keeps a CKA of 0.9 with its start	Refuted, 0.05 to 0.12 in layers 3 to 5
P5h.3	The hidden gradient norms fall below their initial value on the plateau, as a median over the plateau probes, and in at least 7 of 9 training runs their maximum within 200 epochs after departure exceeds the plateau median	Partly. Fell in 13 of 13, rose in 4 of the 9 and in 4 of all 13
P5h.4	Each of the three $[8] \times 8$ runs ends within 10 percent of its start with a CKA of at least 0.9 in every hidden layer	Refuted. One run trains, and the two that collapse end at 1.29 to 1.67 with a CKA of 0.11 to 0.14
P5h.5	The Experiment 5c outcomes replicate on new weight draws	Refuted in one run, $[8] \times 8$ trains on one image
P5h.6	After one epoch the low bank keeps a CKA of 0.9 in every hidden layer, and every bank above 4.4 has a hidden layer below 0.7	Partly. The low bank falls to 0.31, 0.12 and 0.11 in layers 3 to 5, the second part held
Experiment 5i		
P5i.0	The three banks below train and the three above collapse on both images	Held, 12 of 12
P5i.1	Above, in every run the median consecutive-probe CKA of the last layer over the plateau is below 0.2	Held, 6 of 6 at 0.111
P5i.2	Below, in every run that median over epochs 50 to 500 is above 0.5	Refuted, 4 of 6
P5i.3	Above, in every run the median over epochs 50 to 400 of the largest hidden kernel step between probes is below 1 percent of the initial norm, while the median last-layer CKA over the same probes is below 0.2	Partly. The step is 2.3 to 7.6 percent, the CKA stays at 0.111

B SETTINGS

B.1 THE NETWORK

Every bank network has L layers of width 256 with a sinusoidal activation, followed by a linear head. The activation of a unit is a sum of τ terms $\alpha_i \sin(\omega_i z + \phi_i)$, and ω_i, ϕ_i and α_i are trainable in every layer. Every layer starts from the same bank, with $\alpha_i = 1/\tau$ and $\phi_i = 0$ unless an experiment says otherwise. The dense kernels use He normal initialization in the Kodak and CelebA settings, Glorot uniform in the STAF setting, and a uniform draw of half-width $\sqrt{6/n}$ in the SDF and Burgers experiments, where n is the fan-in. That last draw is scaled by $1/\max_i \omega_i$ in the first layer of those two experiments. Biases start at zero.

The STAF networks follow STAF's own code. Each layer draws its own bank of $\tau = 5$ frequencies as ω_0 times a uniform draw on $[0, 1)$, with $\omega_0 = 60$ in the first layer and 30 in the hidden layers on images, as in STAF's image setting, and 30 throughout on the other signals. Phases are uniform on $[-\pi, \pi]$. Amplitudes are the signed square root of Laplace samples of scale $1/(2\tau)$. STAF's default image network has three layers and adds the preactivation to the activation of every hidden layer, which is the residual path of Section 7.2. The five-layer arm removes that path.

SIREN follows Sitzmann et al. (2020), with $\omega_0 = 30$ inside every sine and hidden kernels drawn uniformly with half-width $\sqrt{6/n}/\omega_0$. Its phase scale is ω_0 times the standard deviation of the kernel output.

B.2 THE THREE IMAGE SETTINGS

The CelebA setting fits one CelebA image (Liu et al., 2015) resized to 178×178, with coordinates on a grid in $[-a, a]^2$ and $a = 1$ unless the experiment varies it. The network has five sinusoidal layers, a sigmoid head, Adam at 10^{-3}, 2000 epochs and batches of 4096 pixels. Each epoch is one pass over a fresh permutation of the pixels. The Kodak setting repeats the CelebA setting on one Kodak image resized to 256×256, with a linear head. That gives 16 optimizer steps per epoch and 32,000 steps per run. The STAF setting follows Morsali et al. (2026) on all 24 Kodak images at 256×256, with Glorot initialization, 1000 passes over the pixels in batches of 4096, a learning rate of 10^{-3} multiplied by $0.05^{k/K}$ at step k of K, and the best full-image PSNR over checkpoints taken every 100 passes.

The seed of a run fixes the data order and every draw made through TensorFlow, which covers the uniform draws of the SDF, Burgers and SIREN networks. The He and Glorot kernels of Keras are seeded separately, and the experiments before Experiment 1y drew them fresh in every run. Experiments 1y, 5h and 5i fix every draw, and Experiment 1y reloads one set of initial weights for both runs of a pair.

The phase scale stored with each run was measured on the network that trained and is exact for that run. Where a later measurement rebuilds such a network, it takes the median over fresh kernel draws and marks the value as an estimate. The median runs over ten draws for the networks of Section 7.3 and three for the 240 STAF networks of Section 7.1. Ten draws of one recipe stay within 5.4 percent of their median.

B.3 THE THREE OTHER SIGNALS

The audio experiment fits two synthetic signals of 8192 samples on $[-1, 1]$. The low-frequency signal sums three fundamentals at 2, 5 and 8 cycles with three harmonics each under a slow envelope. The high-frequency signal sums fundamentals at 20, 35 and 50 cycles with two harmonics each under a decaying envelope. Both are scaled to unit peak. The network has five sinusoidal layers of width 256 under He initialization, Adam at 10^{-3}, 3000 epochs and batches of 1024. The score is the SNR of the fit, and the constant output is the SNR of the signal's mean. The random banks of this experiment draw five frequencies as 30 times a uniform draw, and one bank serves every layer.

The shape experiment fits two implicit fields on $[-1, 1]^3$. The gyroid target is $|g(x)| - 0.1$ for the gyroid function g with a period of $1/2$, which is a level-set field and not a Euclidean distance. The CSG target is built from the distance primitives of a torus and a box with three spherical holes cut out. The 50,000 training points are half uniform in the cube and half drawn near the surface. The near-surface half takes uniform candidates whose field value is below 0.3 in absolute value, and fills up with uniform points when the candidates run short. That supplies 12,078 points on the gyroid and 14,903 on the CSG shape. The network has five sinusoidal layers of width 256, Adam at 10^{-3} with cosine decay, 5000 epochs and batches of 4096. The score is the intersection over union of the predicted and true signs on the training points.

The Burgers experiment fits $u_t + uu_x = \nu u_{xx}$ with $\nu = 0.01/\pi$ on $x \in [-1, 1]$ and $t \in [0, 1]$, with $u(x, 0) = -\sin(\pi x)$ and $u = 0$ at both ends. The network has three sinusoidal layers of width 128. The loss is the mean squared residual of the equation on 1000 of 5000 collocation points per step, plus ten times the squared error on 512 initial and 512 boundary points. Adam runs at 10^{-3} with cosine decay for 10,000 steps, with the gradient clipped to a global norm of 1. The score is the relative L^2 error against a finite-difference reference on a 100×100 grid.

B.4 COLLAPSE AND DEPARTURE

A run has collapsed when its final PSNR lies within 0.5 dB of the PSNR of the mean colour of its target. The mean-colour PSNR of each Kodak image is stored once and reused by every experiment. On audio the constant output is the SNR of the signal's mean, and on the shapes it is an IoU of zero, because the constant these runs settle on is positive and puts every point outside the surface.

Two measures of departure appear in the paper. Sections 5.2 and 6.1, Figure 2(b) and Figure 6 use the loss history, which records the mean training loss of every tenth epoch. The PSNR of that loss is compared with the mean colour, and the departure is the first recorded epoch after 150 at which it

passes the mean colour by more than 1 dB. Section 6.2 and Table 5 evaluate the whole image every 25 epochs, and the departure is the first such check after epoch 150 that passes the mean colour by 1 dB. The two agree within 25 epochs in 17 of the 22 departing runs of Experiments 5c and 5h. Table 3 lists the five that differ. In the first of them the loss history catches a short early crossing while the network is still settling toward the mean colour, which the whole-image check does not see.

Table 3: The five traced runs whose two departure measures differ by more than 25 epochs.

Experiment	Run	Loss history	Whole image
5c	$[5] \times 8$, kodim02	160	700
5c	$[5] \times 8$, kodim13	520	225
5c	$[5] + [10] \times 7$, kodim02	240	400
5h	$[5] \times 8$, kodim22	440	400
5h	$[5] + [10] \times 7$, kodim02	230	400

The 1 dB margin was fixed before the departure experiments ran. Over the 264 runs above a phase scale of 2.1 in Figure 6, the set of 80 runs that ever departs does not change when the margin is set to 0.5, 1, 2 or 3 dB. The median departure epochs of the single-frequency family do change, and under a 3 dB margin they become 540, 580, 630, 920 and 1340 for ω of 5, 6, 6.5, 7 and 7.5, against 510, 220, 570, 630 and 770 under the 1 dB margin. The 3 dB medians are monotone and the 1 dB medians are not, because a small early crossing makes $[6] \times 8$ look faster than $[5] \times 8$.

B.5 The gradient norm at initialization

The gradient norm at initialization does not point the way the outcome does. Among the $\tau = 5$ banks of the first Monte Carlo image, its median is 300 for the banks that collapse and 46 for the banks that train. Under the base Kodak setting the high bank starts with a gradient norm of 6.6×10^5 and the low bank with 58, and it is the high bank that never leaves the mean colour. A large derivative at initialization is what a fast sinusoid produces, and it is the runs with the larger gradient that stay at the mean colour.

B.6 Controls that do not change the outcome

Three experiments check properties of the bank that could have looked decisive. Figure 5 shows the first two.

The number of terms in the bank does not decide the outcome. What it changes is the chance of drawing a slow term. At equal τ the low bank reaches 45.18 ± 4.51 dB with five terms and 44.49 ± 3.61 dB with eight, over eight images and three seeds, and neither collapses in 24 runs. Random banks at $\tau = 5$ reach 43.04 ± 4.71 dB with no collapse in 24 runs. At $\tau = 8$ they collapse in 8 of 24, all from one bank with a phase scale of 3.14.

A slow term from anywhere removes the collapse. Twelve random banks with two terms drawn from $U[0, 4]$ and six from $U[0, 30)$ train in 36 of 36 runs on three images, at 42.43 ± 5.26 dB and within 0.4 dB of the low bank. Twelve banks with all eight terms from $U[0, 30)$ collapse in 3 of 36, all from the one bank with a phase scale of 7.41.

Learning the frequencies does not decide it either. With the low bank on three images and three seeds, freezing the frequencies gives 42.74 ± 5.52 dB against 43.14 ± 6.64 dB with everything learned. Freezing the dense weights and learning only the bank gives 19.15 ± 2.47 dB. With everything learned the frequencies move by 2.6 to 6.5 percent on average per layer.

C The boundary by bank family

Table 4 gives the two values behind Section 5.1. For each family the first column is the largest phase scale at which a run still trained. The second is the smallest at which one collapsed. Both are read off the runs of the experiments in the last column. They are measured extrema and not confidence

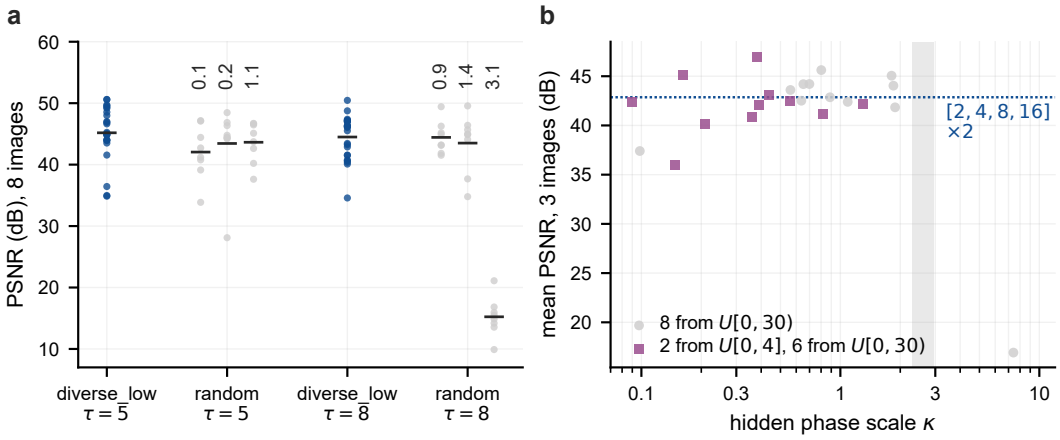

Figure 5: **Two controls on the bank.** (a) The low bank and random banks at equal τ on eight images, with the phase scale of each random bank written above it. (b) Random banks with two slow terms against pure random banks, mean PSNR over three images.

bounds, and where they overlap the phase scale does not separate the outcomes inside that family. An entry of none means that no run of the family collapsed.

Table 4: The boundary by bank family, as measured extrema.

Family	Bank	Largest κ, no collapse	Smallest κ, with collapse	Experiments
Random	$U[0, 30)$, $\tau = 5$ or 8	3.01	1.81	1, 1r, 1s
Slow + seven at 30	$[s] + [30] \times 7$, $s \in \{3, 4, 5\}$	2.66	2.65	3g, 5c, 5d
Slow + seven at 10	$[5] + [10] \times 7$	4.47	none	3g, 5c, 5h
Single frequency	$[\omega] \times 8$	8.02	7.05	3e, 3g, 5c, 5d, 5e, 5h
Deterministic	five multi-frequency banks	2.31	none	3e

The random-bank row pools the three CelebA images of the CelebA setting. Its largest non-collapsing value of 3.01 comes from the third image, where the output bias starts far from the target mean, and Section 6.3 gives the account. On the first two images no bank above 2.36 trains.

The single-frequency row overlaps because the same bank on the same image gives opposite outcomes under different weight draws. $[8] \times 8$ trains in 1 of 11 five-layer runs, on kodim02 with a departure at epoch 750, and $[7] \times 8$ trains in 2 of 3 and $[7.5] \times 8$ in 1 of 3. The measured extrema are 8.02 for a run that trains and 7.05 for a run that collapses.

The deterministic row holds the five multi-frequency banks of Experiment 3e, which are $[1, 1, 2, 2, 4, 4, 8, 16]$, $[1, 2, 4, 8, 16, 32, 1, 2]$, $[1, 3, 5, 10, 15, 20, 25, 30]$, $[3, 3, 6, 6, 12, 12, 24, 24]$ and $[5, 10, 15, 20, 25, 30, 5, 10]$. All five train in 9 of 9 runs, up to a phase scale of 2.31. The three single-frequency banks of that experiment, $[2] \times 8$, $[5] \times 8$ and $[10] \times 8$, belong to the single-frequency row, where $[10] \times 8$ at 10.09 collapses in 9 of 9.

D THE PHASE SCALE DOES NOT CROSS THE THRESHOLD

D.1 THE TRACES BEHIND SECTION 6.1

A natural account of the delayed departures in Section 6.1 is that the network lowers its phase scale while it waits and crosses the boundary from above. We had predicted exactly that, and the traces go the other way. Over the first 100 epochs of five banks on three images the phase scale rises or holds. Averaged over the images it goes from 5.03 to 5.45 for $[5] \times 8$, from 6.02 to 6.62 for $[6] \times 8$, from 8.05 to 8.94 for $[8] \times 8$, from 4.43 to 4.55 for $[5] + [10] \times 7$ and from 4.42 to 4.61 for $[5] + [30] \times 7$,

and two of the fifteen runs end within one percent below their start. The hidden weight norms stay within half a percent of the He value of 22.7 throughout that window.

The full traces say the same over 2000 epochs, and Table 5 lists them. The phase scale enters the range 2.0 to 2.7 in none of the fifteen runs. Nine successes end within 0.9 of where they started. The six failures end at 3.6 to 10 times their starting value, and that growth starts hundreds of epochs after the run has settled at the mean colour. $[8] \times 8$ on kodim02 reaches the mean colour at epoch 75 and its phase scale first exceeds 1.5 times its starting value at epoch 975. The later growth cannot explain the arrival at the mean colour.

Table 5: The fifteen traced runs of Experiment 5c. Each row is one bank on three Kodak images, listed in the same order in every column.

Bank	κ at start	κ at the end	Outcome	Departure
$[5] \times 8$	5.00 to 5.03	4.92, 5.68, 5.88	trains 3 of 3, 34.9 to 43.4 dB	700, 225, 225
$[6] \times 8$	6.01 to 6.02	5.94, 6.69, 6.48	trains 3 of 3, 32.9 to 41.0 dB	750, 525, 225
$[5] + [10] \times 7$	4.42 to 4.46	4.45, 4.40, 4.44	trains 3 of 3, 26.2 to 45.4 dB	400, 375, 225
$[5] + [30] \times 7$	4.43 to 4.44	24.3, 20.2, 22.9	collapses 3 of 3	none
$[8] \times 8$	8.04 to 8.07	28.9, 80.7, 35.3	collapses 3 of 3	none

The two banks that define the random-bank boundary on the first CelebA image behave the same way on Kodak. Traced on two images, the bank at 2.00 and the bank at 2.14 both train, depart at epoch 160 or 180, and hold their phase scale between 1.91 and 2.29 throughout. The bank at 2.14 is the one that collapses on that CelebA image, and on Kodak it trains in 4 of 4 runs, including two runs of 10,000 epochs that reach 55.50 and 46.98 dB.

D.2 Departure across every Kodak run

Figure 6 places all 631 Kodak runs with a loss history on one axis. A run is early when it sits above the mean colour at epoch 200 and at the end, late when it sits at the mean colour at epoch 200 and ends more than 3 dB above it, fallen when it sits above at 200 and at the mean colour at the end, and never when it sits at the mean colour at both. This analysis came after the outcomes were known.

The two ends behave as the phase scale says. Below 2.0, 339 of 361 runs are early, 14 are late, 6 fell and 2 never leave. Above 7.0, 114 of 118 never leave. Departure is late in the middle, where between 4.5 and 7.0 it comes after epoch 200 in 24 of the 28 runs that train. Late departures are not confined to one family, and 22 multi-frequency runs depart after epoch 200, the latest of them at epoch 1720.

The 1 dB margin was fixed before these runs. Changing it does not change which runs depart. The same 80 runs depart at a margin of 0.5, 1, 2 or 3 dB, over the 264 runs above a phase scale of 2.1. It does change the median epoch. Under 3 dB the medians for ω of 5, 6, 6.5, 7 and 7.5 are 540, 580, 630, 920 and 1340, which rise with the phase scale. Under 1 dB they are 510, 220, 570, 630 and 770, and the 220 of $[6] \times 8$ comes from one small early crossing that the wider margin ignores.

D.3 The late collapses of the rescaled grid

Section 6.1 reports that 13 of the 14 collapses of the rescaled high bank at a hidden phase scale of 3 or below first fit the image and then fall back. The low phase scale of this grid comes from shrinking the hidden weights of a bank whose frequencies all sit at 30, and training returns those weights toward their initial scale. Figure 7 shows the whole grid, and Figure 8 traces four runs of the cell at a hidden scale of 0.05.

In the three runs that fall, the largest hidden phase scale sits between 2.3 and 3.3 during the fit and leaves that range upward, to 8.6, 6.3 and 11.8, within 25 epochs. Two of the three falls happen inside that same interval, at epochs 1250 and 1025. In the third run the phase scale leaves at epoch 1775 and the output reaches the mean colour at epoch 1975. The run that never falls keeps its phase scale between 2.4 and 3.4 for 2000 epochs. The second-layer weight norm grows in all four runs, and the norm alone does not mark which one falls.

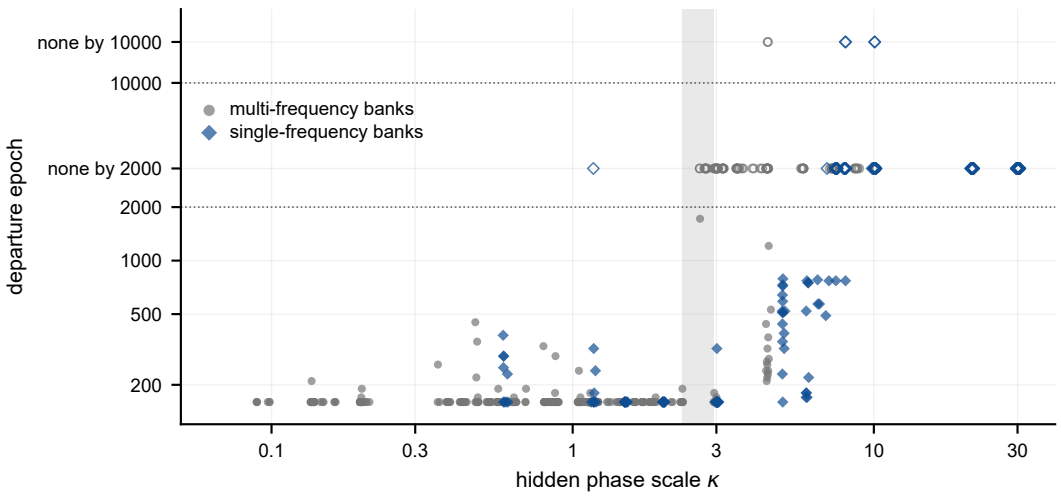

Figure 6: **Departure epoch against the phase scale at initialization.** Every Kodak run with a loss history. Open markers never departed within their budget, and the grey band is the random-bank threshold range. Single-frequency banks include the rescaled high bank of Experiments 0 and 0b.

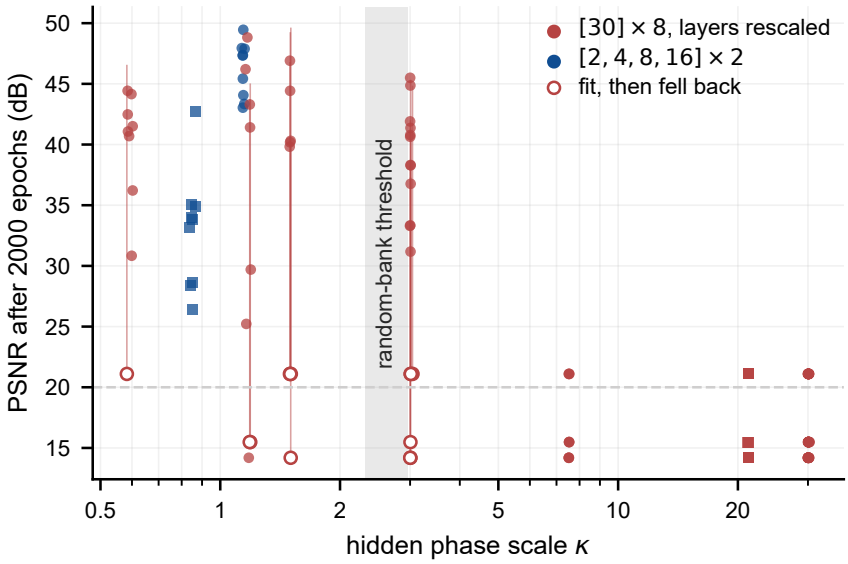

Figure 7: **The rescaled high bank and the low bank against the hidden phase scale.** Every point is one run of Experiments 0 and 0b. Open markers fit the image and then fell back to the mean colour.

E THE GAUSSIAN MEAN OF THE DERIVATIVE

E.1 THE CLOSED FORM

Let the preactivation z of a sinusoidal layer be Gaussian with mean zero and standard deviation $s > 0$. The derivative of the activation is $\sigma'(z) = \sum_i \alpha_i \omega_i \cos(\omega_i z + \phi_i)$, and its mean over z is

$$\mathbb{E}[\sigma'(z)] = \sum_i \alpha_i \, \omega_i \, e^{-\omega_i^2 s^2/2} \cos \phi_i. \tag{4}$$

The proof is one line. For any real ω and ϕ, the mean of $\cos(\omega z + \phi)$ is the real part of $e^{i\phi}$ times the characteristic function of z at ω, which is $e^{-\omega^2 s^2/2}$. Linearity over the terms gives Equation 4. With

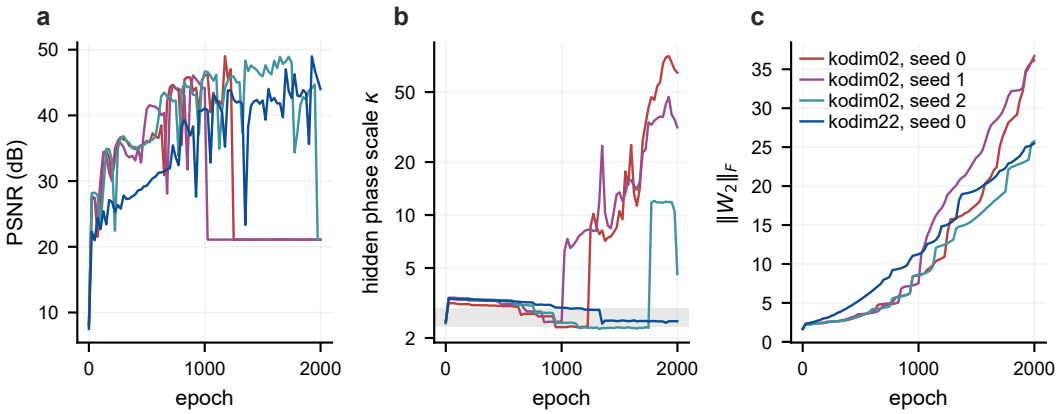

Figure 8: **Four traced runs at a hidden scale of** 0.05**.** (a) PSNR. (b) The largest hidden phase scale, with the random-bank threshold range. (c) The norm of the second-layer kernel.

the phases at zero and the amplitudes at $1/\tau$, each term is at most $s^{-1}e^{-1/2}/\tau$, which it reaches at $\omega s = 1$, and the whole mean is bounded by $s^{-1}e^{-1/2}$.

Equation 4 gives the direction of the effects of the phase scale in Section 5. A fast term is suppressed in its own frequency without ever reaching zero. The equation does not give the non-monotone effect of the fast terms in Section 5.2.

E.2 What it does not give

The equation does not give the value of any boundary, and the single-frequency family is the direct counterexample. A pure bank has a preactivation scale close to 1.0 at every hidden layer, and its phase scale is close to its frequency. At $[6] \times 8$ Equation 4 returns 9.1×10^{-8}, and those networks train in 9 of 9 runs under the Kodak setting. $[5] \times 8$ returns $5e^{-12.5}$ and trains in 15 of 15. The boundary is therefore measured and not read off this equation.

One could build the discriminating quantity out of this mean instead of the phase scale. Two banks rule that out. $[5]+[10]\times 7$ and $[5]+[30]\times 7$ start at the same phase scale of 4.4, and the first trains in 3 of 3 while the second collapses in 3 of 3. A correlation-map quantity built from the mean puts both near 10^{-9} and cannot tell them apart. It also orders other pairs the wrong way, giving 2.2×10^{-9} to $[5] + [30] \times 7$, which collapses, and 1.4×10^{-14} to $[6] \times 8$, which trains.

E.3 A post hoc comparison with the mean-field quantity

Signal propagation theory locates the order-to-chaos transition of a deep network at $\chi_1 = 1$, with $\chi_1 = \sigma_w^2\, \mathbb{E}[\sigma'(z)^2]$ (Poole et al., 2016; Schoenholz et al., 2017). We compared three scores against the outcome of the 300 Monte Carlo runs. They are the phase scale, the largest χ_1 over the hidden layers, and the smallest mean of Equation 4 over the hidden layers, each taken at the measured preactivation scale of its layer. This analysis came after the outcomes were known.

The phase scale and the mean order the outcome equally well, and χ_1 orders it far worse. The areas under the ROC curve are 0.989, 0.989 and 0.794. The best single threshold misclassifies 1 to 3 runs of 100 per image on the first two scores and 13 to 22 on the third. The mean-field value exceeds one in every run, from 21 to 290, which places all 300 runs on the chaotic side and separates none of them. This is a comparison of per-layer scores and not a test of the fixed point and the depth scales that the theory is about.

A broader comparison gives the same ranking by a small margin. We pooled the banks of seven experiments into 281 cells, one bank at one depth and one coordinate range, of which 280 are all-or-nothing over their runs. The best single threshold misclassifies 9 cells on the phase scale, 10 on the correlation-map quantity and 10 on the mean, and eight of the errors are shared by all three.

Appendix F derives the map quantities. They explain why the variable has this form, and they do not predict its value, which changes with the family of the bank.

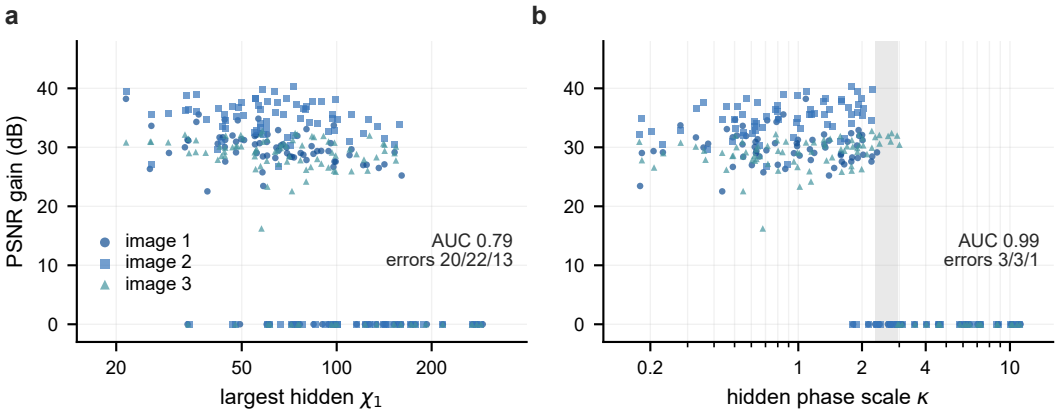

Figure 9: **The mean-field quantity against the phase scale.** The 300 Monte Carlo runs, PSNR above the mean colour against (a) the largest hidden χ_1 and (b) the phase scale.

F WHY THE VARIABLE HAS THIS FORM

The phase scale was found by measurement, and Section 5.1 gives its value family by family. This appendix derives its form from the correlation map of a random sinusoidal layer. The derivation gives the form of the variable and not its value, and Appendix E measures where it fails. Throughout, the phases are zero, the amplitudes are positive, and the hidden preactivations are treated as Gaussian with a common variance across coordinates, which makes every $K_l(1)$ positive. The numerical checks of this appendix were made after the training outcomes were known.

F.1 ONE LAYER AS A MAP ON CORRELATIONS

Take a hidden layer l with preactivation standard deviation s_l and write $q_l = s_l^2$. For two coordinates x and x', the preactivations z and z' of one unit are sums of 256 independent terms over the weight draw. In the wide limit they are jointly Gaussian with mean zero, common variance q_l and a correlation c_l that depends on the pair. The activation is $\sigma(z) = \sum_i \alpha_i \sin(\omega_i z)$ with the phases at zero. One computation carries the whole map.

Lemma 1. *Let z and z' be jointly Gaussian with mean zero, variance q and correlation c. For any ω and ω',*

$$\mathbb{E}\big[\sin(\omega z)\sin(\omega' z')\big] = e^{-q(\omega^2+\omega'^2)/2}\,\sinh(q\,\omega\omega' c). \tag{5}$$

Proof. Write the product as $\frac{1}{2}[\cos(\omega z - \omega' z') - \cos(\omega z + \omega' z')]$. Each argument is Gaussian with mean zero and variance $q(\omega^2 + \omega'^2 \mp 2\omega\omega' c)$, and the mean of the cosine of such a variable is $e^{-V/2}$ with V its variance. The difference of the two exponentials is $e^{-q(\omega^2+\omega'^2)/2}$ times $2\sinh(q\omega\omega' c)$. □

Proposition 1. *Let $K_l(c) = \sum_{i,j} \alpha_i\alpha_j\, e^{-q_l(\omega_i^2+\omega_j^2)/2}\, \sinh(q_l\,\omega_i\omega_j\,c)$. With He kernels of variance $2/n$ and zero biases, the next layer has*

$$q_{l+1} = 2\,K_l(1), \qquad c_{l+1} = \frac{K_l(c_l)}{K_l(1)} = F_l(c_l). \tag{6}$$

Proof. $\mathbb{E}[\sigma(z)\sigma(z')]$ is a sum of τ^2 terms, each given by Lemma 1, and the sum is $K_l(c_l)$. The next preactivation is $z_{l+1} = \sum_k W_k \sigma(z_k)$ over n units with independent weights of mean zero and variance $2/n$. Its variance is n times $2/n$ times $K_l(1)$, the covariance of the pair is $2K_l(c_l)$, and their ratio is the correlation. □

F_l runs from $F_l(0) = 0$ to $F_l(1) = 1$. It is increasing and convex because $\sinh$ is and every coefficient is positive, which gives $F_l(c) \leq c$ on $[0, 1]$ and a slope at zero of at most one. A layer keeps a correlation of one, sends a correlation of zero to zero, and pulls every value in between downward.

Table 6 checks one step of the variance recursion at width 256. It takes the measured standard deviation of the second layer of four networks at initialization, applies Equation 6 once, and compares the result with the measured third layer. The four agree to within 0.005. For a single frequency the recursion is $q_{l+1} = 1 - e^{-2\omega^2 q_l}$, whose fixed point sits just below 1, with a deficit close to $e^{-2\omega^2}$ that is 3.4×10^{-4} at $\omega = 2$ and below 10^{-21} from $\omega = 5$ up, and every pure bank in the paper has a hidden standard deviation close to one for that reason.

Table 6: The map applied once to the measured second layer.

Bank	s_2 measured	s_3 from the map	s_3 measured
$[2, 4, 8, 16] \times 2$	0.536	0.571	0.570
$[5] \times 8$	0.990	1.000	0.995
$[5] + [10] \times 7$	0.885	0.884	0.881
$[5] + [30] \times 7$	0.887	0.884	0.884

F.2 The slope at zero

For small c, $\sinh(x) \approx x$ and the double sum in $K_l(c)$ factorizes,

$$c_{l+1} \approx \rho_l \, c_l, \qquad \rho_l = F_l'(0) = \frac{q_l \, m_l^2}{K_l(1)}, \qquad m_l = \sum_i \alpha_i \omega_i \, e^{-q_l \omega_i^2/2}. \tag{7}$$

m_l is the Gaussian mean of the derivative from Appendix E. ρ_l is the fraction of a small correlation that survives layer l. Its denominator is the power of the activation, $\mathbb{E}[\sigma(z)^2]$, whose diagonal terms approach $\alpha_i^2/2$ once their phase scales are large and whose cross terms depend on the spacing of the frequencies. Its numerator holds only the terms whose factor $e^{-q_l \omega_i^2/2}$ has not vanished. The ratio is the power the slow terms pass on, over the power of the whole bank. For the low bank at its measured second layer $\rho = 0.26$, and for $[5] \times 8$ it is 1.1×10^{-9}.

Each term of m_l carries $e^{-\omega_i^2 s_l^2/2}$, which is $e^{-\kappa^2/2}$ in the phase scale of that term, and the term with the smallest frequency decays slowest. When the other terms contribute little to the sum,

$$\sum_{\omega_i > \omega_{\min}} \alpha_i \, \omega_i \, e^{-q_l(\omega_i^2 - \omega_{\min}^2)/2} \ll \alpha_{\min} \, \omega_{\min}, \tag{8}$$

the slowest term is the only one left, $m_l \approx \alpha_{\min} \omega_{\min} \, e^{-\kappa_l^2/2}$ with $\alpha_{\min}$ the total amplitude at the slowest frequency and $\kappa_l = \omega_{\min} s_l$, and ρ_l falls as $e^{-\kappa_l^2}$. The condition is on the spacing of the frequencies and not on the phase scales alone. Two frequencies close together share the slow term, and the low bank is one case where the term at 4 still contributes. For a pure bank the sum has one term, and for the mixed banks of Section 5.2, with the slow term at 5 and the fast terms at 8 or more, the other terms contribute less than 10^{-5} of the slowest.

The amount of long-range correlation a hidden layer passes on is therefore set by the phase scale of its slowest term, through a Gaussian tail. From $\kappa = 1$ to $\kappa = 3$ the factor $e^{-\kappa^2}$ falls by a factor of 3000, which is why the dependence on the phase scale is steep.

F.3 Depth, and the other end of the map

The map is applied once per hidden layer, and the chain rule gives the slope of the composed map at zero,

$$(F_L \circ \cdots \circ F_2)'(0) = \prod_{l=2}^{L} \rho_l, \tag{9}$$

which is the factor a small correlation keeps between the second layer and the head. Since $F_l(c) \leq c$ on $[0, 1]$, every correlation below one also shrinks at every hidden layer before it reaches the linear

region, and there the full map acts and not its slope. For $[5] \times 8$ at $q = 1$ the map sends 0.9 to 0.08 while ρ times 0.9 is 6×10^{-10}, which is why Equation 9 gives the rate near zero and not the path there. The product describes the decay near zero. Proposition 2 bounds what a head can do when the features of different coordinates are unrelated, and Table 7 checks the features the networks actually have. Both describe the network before the first optimizer step.

The slope at the other end, $\chi_l = F_l'(1) = q_l\, \mathbb{E}[\sigma'(z)^2]/K_l(1)$, is the rate at which nearby coordinates separate. It equals the mean-field χ_1 of Appendix E at a fixed point of q, and it is the quantity that places the edge of chaos (Poole et al., 2016; Schoenholz et al., 2017). $\mathbb{E}[\sigma'(z)^2]$ is dominated by the fastest terms of the bank, and χ is 19.6 for the low bank, 24.5 for $[5] \times 8$ and 694 for $[5] + [30] \times 7$ at their measured second layers. These three networks sit far on the chaotic side by that measure, the two that train and the one that collapses alike, and over the 300 Monte Carlo runs of Appendix E the mean-field value runs from 21 to 290. The separation is at the other end of the map.

F.4 What the head can do at initialization

The constant output is the mean of the target. Write Φ for the $N \times n$ matrix of last-layer features at the N training coordinates, with $n = 256$, and fit the linear head with its bias by least squares while the hidden layers stay at their initial values. The bias alone reproduces the mean of the target. What the features can add is bounded in the following idealized case, and the bound is the reference for Table 7.

Proposition 2. *Let the units of the last hidden layer be independent, with preactivations at the N training coordinates that are jointly Gaussian with mean zero, a common variance and zero correlation between distinct coordinates, and let $\tilde{y}$ be one fixed channel of the target with its mean removed. The least-squares fit removes, in expectation over the weight draw, at most a fraction $n/(N-1)$ of $\|\tilde{y}\|^2$.*

Proof. The fit projects $\tilde{y}$ onto $V' = V \cap \mathbf{1}^\perp$, where V is the span of $\mathbf{1}$ and the n columns of Φ, which gives $\dim V' \le n$. Under the assumption the feature vectors of distinct coordinates are independent and identically distributed, which makes the law of V' invariant under every permutation of the coordinates. Then $\mathbb{E}[P_{V'}]$ commutes with every permutation matrix, and

$$\mathbb{E}[P_{V'}] = aI + b\,\mathbf{1}\mathbf{1}^\top, \qquad \operatorname{tr}\mathbb{E}[P_{V'}] = \mathbb{E}[\dim V'] \le n, \tag{10}$$

$$P_{V'}\mathbf{1} = 0 \quad \Rightarrow \quad b = -a/N, \qquad a = \frac{\mathbb{E}[\dim V']}{N-1} \le \frac{n}{N-1}, \tag{11}$$

$$\mathbb{E}\|P_{V'}\tilde{y}\|^2 = \tilde{y}^\top \mathbb{E}[P_{V'}]\,\tilde{y} = a\,\|\tilde{y}\|^2. \qquad \square$$

On a 256×256 image the bound is 0.39 percent of the energy, and a fit that removed that fraction would gain 0.017 dB over the mean colour. Table 7 does the fit on the initial features of eight networks, with one draw of the weights per bank and three Kodak images. The actual features are not exchangeable, since with zero biases and phases the whole stack before the head is odd in the coordinate, and every bank above the boundary still sits within a factor of 1.5 of the reference. Those banks give up 0.3 to 0.6 percent of the energy, under 0.03 dB, with a mean cosine between neighbouring pixels below 0.001. The low bank and the rescaled high bank of Section 4.3 keep a neighbour correlation of 0.05 and 0.58, and their heads do a little better, up to 6.1 percent and 0.27 dB.

The table says that the initial features give a head almost nothing beyond the mean, for the banks that train and the banks that collapse alike. What separates the outcomes is what training does to the features afterwards, which the map does not describe and Section 6 measures. This analysis came after the outcomes were known.

F.5 What the map explains

The map bears on five findings of the main text, within the conditions stated above.

The form of the variable. ρ_l depends on the bank through m_l, and m_l through the phase scale of its slowest term. The paper takes the largest value over the hidden layers because $\kappa^2 e^{-\kappa^2}$ falls once κ passes 1. This is a trend and not an identity, since ρ_l also carries q_l and $K_l(1)$. Over the 60 banks of

Table 7: The least-squares head on the initial features, one weight draw and three Kodak images per bank.

Bank	κ	Energy removed percent	Gain over the mean dB	Neighbour correlation
Bound of Proposition 2		0.39	0.017	
$[2, 4, 8, 16] \times 2$	1.14	1.5 to 6.1	0.07 to 0.27	0.05
$[30] \times 8$, hidden layers at 0.05	1.50	2.8 to 5.3	0.12 to 0.24	0.58
$[5] + [10] \times 7$	4.43	0.36 to 0.53	0.016 to 0.023	0.00
$[5] + [30] \times 7$	4.43	0.33 to 0.48	0.014 to 0.021	0.00
$[5] \times 8$	5.02	0.33 to 0.46	0.015 to 0.020	0.00
$[6] \times 8$	6.02	0.37 to 0.55	0.016 to 0.024	0.00
$[8] \times 8$	8.02	0.38 to 0.46	0.016 to 0.020	0.00
$[30] \times 8$	30.1	0.35 to 0.45	0.015 to 0.020	0.00

Experiment 1 whose hidden phase scales all exceed 1, the layer with the largest phase scale supplies the smallest factor in 59.

The first layer. Its preactivation has a variance that changes with the coordinate, and Proposition 1 does not apply to it as a step. Within the approximation it enters the hidden stack through q_2 and c_2 only, and the power of a sine approaches saturation once its phase scale passes 1, at 86 percent of the limit at a phase scale of 1. The sweep of Section 4.3 fits this. Its first-layer phase scale fell from 24 to 1.2, the largest hidden phase scale of all 45 runs stayed between 30.0 and 30.3, and all 45 collapsed. The map does not say what a first layer below that range would do.

Depth. Equation 9 has one factor per hidden layer. A bank that leaves enough correlation after three hidden layers can leave too little after four, which is what the bank at $\kappa = 2.98$ does between depths 4 and 5 in Section 5.3.

The coordinate range and the target. Neither appears in Equation 6. The hidden variance $q_{l+1} = 2K_l(1)$ is bounded because the sine is, and for the low bank it moves only from 0.41 to 0.27 as q_l goes from 0.05 to 1. That is why the hidden standard deviation stays near 0.52 to 0.54 at every coordinate range of Section 4.2 while the first-layer value moves fourfold. Where the target does change the outcome, in Section 6.3, it acts through the output bias, which the map does not describe.

The mean colour. Proposition 2 bounds what a head can add to the mean of the target when the features are exchangeable, and the failures of Section 4.1 sit at that mean to within the precision we measure.

F.6 What it does not give

The value. Equation 9 says how much long-range correlation survives and not how much is needed. That amount has to be measured, and Section 5.1 measures it family by family. The single-frequency banks are the sharpest case. $[5] \times 8$ and $[6] \times 8$ have ρ of 10^{-9} and 10^{-14} and both train. $[5] + [10] \times 7$ trains while $[5] + [30] \times 7$ collapses at the same ρ, because fast terms at 10 and at 30 enter the denominator of Equation 7 in the same way.

The dynamics. The map describes the network before the first step. Everything in Section 6 lies outside it. Appendix D shows that no run lowers its phase scale into the band before it departs, which rules out the simplest dynamical reading of the map, that a run trains by walking into the region where Equation 9 allows it.

Two assumptions. The lemma uses phases at zero, which holds for every bank network of the paper and not for STAF, whose phases are uniform on $[-\pi, \pi]$. The map is also a wide-network statement, of which Table 6 checks the variance recursion at width 256 to within 0.005 in the standard deviation.

G The two shapes in detail

The two configurations whose layers start from the same bank bracket the switch and no more. The low bank covers both shapes and three seeds on the gyroid, at phase scales of 1.17, 1.18, 1.18 and

1.20, and reaches an IoU of 0.994, 0.993, 0.994 and 0.993 in that order. $[30] \times 8$, at a phase scale of 30, stays at the constant output on both shapes.

One configuration below the boundary fails, and we report it as a case the phase scale does not cover. The bank $[1, \ldots, 128]$, at a phase scale of 0.41, reconstructs the CSG shape at an IoU of 0.998 and stays at the constant output on the gyroid in three seeds. Changing any one of the bank, the first-layer scale and the learning rate recovers the shape, at an IoU between 0.973 and 0.993. A 12-epoch run shows the hidden phase scale between 0.4 and 0.7 while the loss stalls, which is a second route into the constant output that the phase scale does not see.

STAF networks on these shapes give five data points and no rule, and this reading is post hoc. Two of the five reconstruct the CSG shape, at reconstructed phase scales of 3.19 and 4.44, while the three that fail sit at 4.48, 4.89 and 8.89. On the gyroid none of the five reconstructs.

H EVERY CROSS-DOMAIN RUN

Table 8 lists every run of the cross-domain experiment. The first four groups hold the 41 scored runs of Experiments 4a to 4c, the 16 runs of Experiment 4d and the 2 runs of Experiment 4e. The last two rows of the gyroid group are the 12-epoch runs of Experiment 4f, which record a loss trace and no IoU, against a constant output of 0.197. The phase scale is the exact rebuild of the run's own initialization where the recipe carries a seed, and the median of ten fresh draws, marked with a dagger, where it does not.

Table 8: Every cross-domain run. A dagger marks a phase scale taken as the median of ten fresh draws.

Task	Network	lr	κ	Result
audio low-freq	SIREN 5L	10^{-3}	0.97	8.5 dB SNR
audio low-freq	diverse_low	10^{-3}	$1.14^{\dagger}$	55.3 dB SNR
audio low-freq	[8,16,32,64]x2	10^{-3}	$4.04^{\dagger}$	8.6 dB SNR
audio low-freq	random bank seed 0	10^{-3}	$8.34^{\dagger}$	8.6 dB SNR
audio low-freq	random bank seed 1	10^{-3}	$0.06^{\dagger}$	55.2 dB SNR
audio low-freq	random bank seed 2	10^{-3}	$3.40^{\dagger}$	8.6 dB SNR
audio low-freq	random bank seed 3	10^{-3}	$4.36^{\dagger}$	8.6 dB SNR
audio low-freq	random bank seed 4	10^{-3}	$2.44^{\dagger}$	8.6 dB SNR
audio low-freq	SIREN 5L	10^{-4}	0.97	45.5 dB SNR
audio low-freq	SIREN 3L	10^{-3}	0.97	8.6 dB SNR
audio low-freq	SIREN 3L	10^{-4}	0.97	48.9 dB SNR
audio high-freq	SIREN 5L	10^{-3}	0.97	15.3 dB SNR
audio high-freq	diverse_low	10^{-3}	$1.14^{\dagger}$	44.8 dB SNR
audio high-freq	[8,16,32,64]x2	10^{-3}	$4.04^{\dagger}$	15.3 dB SNR
audio high-freq	random bank seed 0	10^{-3}	$8.34^{\dagger}$	15.3 dB SNR
audio high-freq	random bank seed 1	10^{-3}	$0.06^{\dagger}$	48.7 dB SNR
audio high-freq	random bank seed 2	10^{-3}	$3.40^{\dagger}$	15.3 dB SNR
audio high-freq	random bank seed 3	10^{-3}	$4.36^{\dagger}$	15.3 dB SNR
audio high-freq	random bank seed 4	10^{-3}	$2.44^{\dagger}$	15.3 dB SNR
audio high-freq	SIREN 5L	10^{-4}	0.97	44.1 dB SNR
audio high-freq	SIREN 3L	10^{-3}	0.97	15.3 dB SNR
audio high-freq	SIREN 3L	10^{-4}	0.97	48.4 dB SNR
Burgers PINN	diverse_low	10^{-3}	1.19	0.037 rel. L^2
Burgers PINN	diverse_wide	10^{-3}	0.41	0.037 rel. L^2
Burgers PINN	collapsed_30	10^{-3}	29.66	0.277 rel. L^2
Burgers PINN	SIREN 3L	10^{-3}	1.00	0.226 rel. L^2
Burgers PINN	STAF-style seed 0	10^{-3}	$0.42^{\dagger}$	0.170 rel. L^2
Burgers PINN	STAF-style seed 1	10^{-3}	$4.50^{\dagger}$	0.170 rel. L^2
Burgers PINN	STAF-style seed 2	10^{-3}	$3.46^{\dagger}$	0.194 rel. L^2
Burgers PINN	SIREN 3L	10^{-4}	1.00	0.094 rel. L^2
SDF gyroid	diverse_low	10^{-3}	1.17	IoU 0.994

Table 8 continued from the previous page.

Task	Network	lr	κ	Result
SDF gyroid	diverse_wide	10^{-3}	0.41	IoU 0.000
SDF gyroid	collapsed_30	10^{-3}	30.03	IoU 0.000
SDF gyroid	SIREN 5L	10^{-3}	1.00	IoU 0.000
SDF gyroid	STAF-style seed 0	10^{-3}	$8.90^{\dagger}$	IoU 0.000
SDF gyroid	STAF-style seed 1	10^{-3}	$4.44^{\dagger}$	IoU 0.108
SDF gyroid	STAF-style seed 2	10^{-3}	$4.49^{\dagger}$	IoU 0.000
SDF gyroid	STAF-style seed 3	10^{-3}	$4.89^{\dagger}$	IoU 0.000
SDF gyroid	STAF-style seed 4	10^{-3}	$3.19^{\dagger}$	IoU 0.000
SDF gyroid	SIREN 5L	10^{-4}	1.00	IoU 0.994
SDF gyroid	SIREN 3L	10^{-3}	1.00	IoU 0.000
SDF gyroid	diverse_wide seed 1	10^{-3}	0.41	IoU 0.000
SDF gyroid	diverse_wide seed 2	10^{-3}	0.42	IoU 0.000
SDF gyroid	diverse_wide	10^{-4}	0.41	IoU 0.973
SDF gyroid	diverse_low seed 1	10^{-3}	1.18	IoU 0.994
SDF gyroid	diverse_low seed 2	10^{-3}	1.20	IoU 0.993
SDF gyroid	diverse_wide, first layer 1/16	10^{-3}	0.41	IoU 0.993
SDF gyroid	diverse_low, first layer 1/128	10^{-3}	1.15	IoU 0.993
SDF gyroid	diverse_wide, 12-epoch probe	10^{-3}	0.42	loss $0.53 \to 0.21$
SDF gyroid	diverse_wide, 12-epoch probe	10^{-4}	0.41	loss $0.63 \to 0.46$
SDF CSG	diverse_low	10^{-3}	1.18	IoU 0.993
SDF CSG	diverse_wide	10^{-3}	0.41	IoU 0.998
SDF CSG	collapsed_30	10^{-3}	30.00	IoU 0.000
SDF CSG	SIREN 5L	10^{-3}	1.00	IoU 0.000
SDF CSG	STAF-style seed 0	10^{-3}	$8.89^{\dagger}$	IoU 0.000
SDF CSG	STAF-style seed 1	10^{-3}	$4.44^{\dagger}$	IoU 0.998
SDF CSG	STAF-style seed 2	10^{-3}	$4.48^{\dagger}$	IoU 0.000
SDF CSG	STAF-style seed 3	10^{-3}	$4.89^{\dagger}$	IoU 0.000
SDF CSG	STAF-style seed 4	10^{-3}	$3.19^{\dagger}$	IoU 0.998
SDF CSG	SIREN 5L	10^{-4}	1.00	IoU 1.000
SDF CSG	SIREN 3L	10^{-3}	1.00	IoU 0.002

I KODAK-24 IN DETAIL

Table 9 and Figure 10 give the per-image results behind Sections 7.1 and 7.2. The low bank with the seed-42 Laplace amplitudes reaches 41.64 ± 1.58 dB, above its 38.74 ± 1.98 dB with uniform amplitudes, and $[30] \times 8$ with the Laplace amplitudes reaches 39.60 ± 2.19 dB while collapsing on every image with uniform amplitudes. The Laplace draw for $[30] \times 8$ sums to -0.059 over its eight terms against a sum of absolute values of 2.02, which is the near cancellation of Section 7.1.

A lower learning rate explains neither bank result. The low bank still trains at 3×10^{-4} and 10^{-4}, reaching 36.15 and 34.79 dB, and at 10^{-4} all 24 runs of $[30] \times 8$ end at or below the mean colour.

Ordered by the reconstructed phase scale, the 120 five-layer STAF networks collapse in 0 of 7 below 1.81, 0 of 3 between 1.81 and 2.31, 6 of 20 between 2.31 and 2.95, 8 of 28 between 2.95 and 4.0 and 42 of 62 above 4.0. The runs that train also fit worse as the phase scale rises. Their median PSNR above the mean colour is 21.6 dB below 2.31, 18.7 dB between 2.31 and 4.0 and 16.2 dB above. Among STAF's default three-layer networks with the residual path, 76 of 120 start with a hidden layer above 2.95 and 55 above 4.0, and none of the 120 collapses.

J THE BIAS INTERVENTION IN DETAIL

Section 6.3 rests on a chain of six experiments, and this appendix gives each one. The two banks are the $\tau = 5$ trials 1 and 29 of the Monte Carlo draw, at phase scales of 3.00 and 2.93, which collapse on the first CelebA image and train on the third.

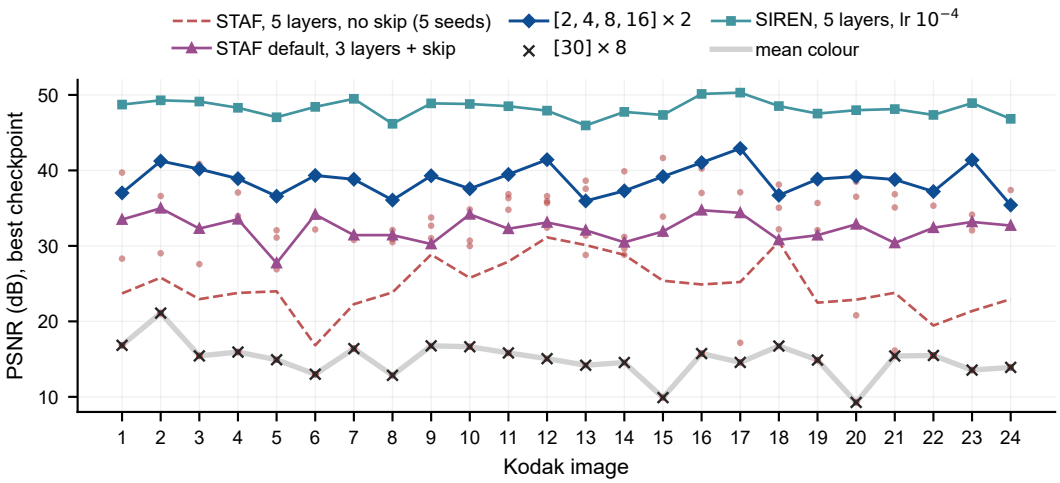

Figure 10: **Kodak-24 under the STAF setting.** Best checkpoint per image. Small dots are the five seeds of the five-layer STAF network.

Table 9: Kodak-24 per image under the STAF protocol, best checkpoint, in dB. Mean colour is the PSNR of the constant image at the target's mean. The amplitude columns are the uniform start at $1/8$ and the seed-42 Laplace draw. STAF rows are means over five seeds, with the number of collapsed seeds for the five-layer network without the residual path.

		$[2, 4, 8, 16] \times 2$		$[30] \times 8$		STAF 5L, no res.		STAF 3L	SIREN
Image	Mean colour	$1/8$	Laplace	$1/8$	Laplace	mean	collapsed	default	5L, 10^{-4}
01	16.83	37.02	40.92	16.83	40.86	23.70	3/5	33.49	48.72
02	21.10	41.24	45.20	21.10	40.00	25.79	3/5	34.98	49.29
03	15.43	40.18	43.26	15.43	38.34	22.94	3/5	32.29	49.13
04	15.94	38.91	40.72	15.94	39.45	23.77	3/5	33.53	48.30
05	14.92	36.58	38.85	14.92	38.47	23.98	2/5	27.72	47.04
06	13.00	39.33	41.71	13.00	40.08	16.83	4/5	34.16	48.42
07	16.39	38.82	41.82	16.39	42.08	22.27	3/5	31.43	49.49
08	12.85	36.07	38.55	12.85	36.16	23.84	2/5	31.43	46.18
09	16.76	39.29	42.24	16.76	42.16	28.86	1/5	30.25	48.88
10	16.64	37.57	42.03	16.64	41.90	25.76	2/5	34.18	48.80
11	15.82	39.47	41.65	15.82	39.23	27.92	2/5	32.27	48.50
12	15.06	41.42	42.74	15.06	38.26	31.13	1/5	33.11	47.92
13	14.19	35.96	39.33	14.19	35.70	30.11	1/5	32.07	45.96
14	14.55	37.29	39.89	14.55	37.80	28.82	1/5	30.48	47.75
15	9.90	39.18	42.57	9.90	38.82	25.39	2/5	31.91	47.35
16	15.73	41.04	44.50	15.73	39.88	24.89	3/5	34.73	50.13
17	14.57	42.91	43.03	14.57	44.09	25.22	2/5	34.37	50.31
18	16.73	36.70	41.47	16.73	38.99	30.55	1/5	30.79	48.53
19	14.88	38.84	41.47	14.88	36.98	22.48	3/5	31.42	47.53
20	9.27	39.20	41.51	9.27	40.83	22.87	2/5	32.86	47.98
21	15.42	38.79	41.61	15.42	41.77	23.78	2/5	30.39	48.12
22	15.48	37.20	41.73	15.48	38.12	19.45	4/5	32.40	47.35
23	13.55	41.37	42.36	13.55	43.27	21.37	3/5	33.17	48.92
24	13.91	35.41	40.29	13.91	37.25	22.94	3/5	32.69	46.83
mean	14.95	38.74	41.64	14.95	39.60	24.78	56/120	32.34	48.23

J.1 The image and not the seed

Under seeds 42 and 13 both banks collapse on the first image and train on the third, at 41.1 to 42.0 dB. Neither seed changed the outcome of either bank.

J.2 Contrast and spectrum

Scaling the pixels of the first image about their mean to a standard deviation of 0.274 and 0.306, near and above the 0.279 of the third image, leaves both banks at the mean colour. Scaling the third image down to 0.240 and 0.206 leaves both banks training, and one reaches 46.36 dB at the lowest contrast.

Over nine further images the fraction of spectral energy outside a radius of 16 does not order the outcome. The image richest in high frequencies collapses one bank and trains the other, and the two banks give opposite outcomes on two images, which rules out any account that looks at the target alone. The two images on which both banks train, the third image and one of the nine, have channel means between 0.15 and 0.35, and every image on which both collapse has a mean between 0.34 and 0.62.

J.3 The head and the target mean

With a linear head the two banks collapse on all five images tested, 10 of 10, including the two images where they train under the sigmoid head. Under the sigmoid head, the first image mapped affinely to a mean of 0.25 trains both banks at 41.00 and 43.52 dB, and the third image mapped to a mean of 0.50 collapses both.

Starting the sigmoid bias at the logit of the target mean, which leaves it nothing to travel, collapses both banks on the third image. Starting it 1.0 below that logit trains both on the first image, at 43.21 and 38.23 dB. A linear head started 1.0 or 1.5 below the target mean trains both on the first image as well, at 33.45 and 28.69 dB and at 25.95 and 27.50 dB.

J.4 Identical weights

Experiment 1x rebuilt the network for every arm, and the He kernels of Keras were fresh each time, which means no two runs shared their weights. Experiment 1y builds each network once, reloads its initial weights before every run, and resets the data order. The runs of one network differ only in the head and in the start of the head bias. Kernel fingerprints agree within each network, and the hidden phase scale agrees to four decimals. Figure 11 shows the eighteen pairs. In every pair the run whose bias starts farther from the mean colour trains and the run whose bias starts nearer collapses. The nearer start is the logit of the mean in the six pairs on the third image and zero in the twelve pairs on the first.

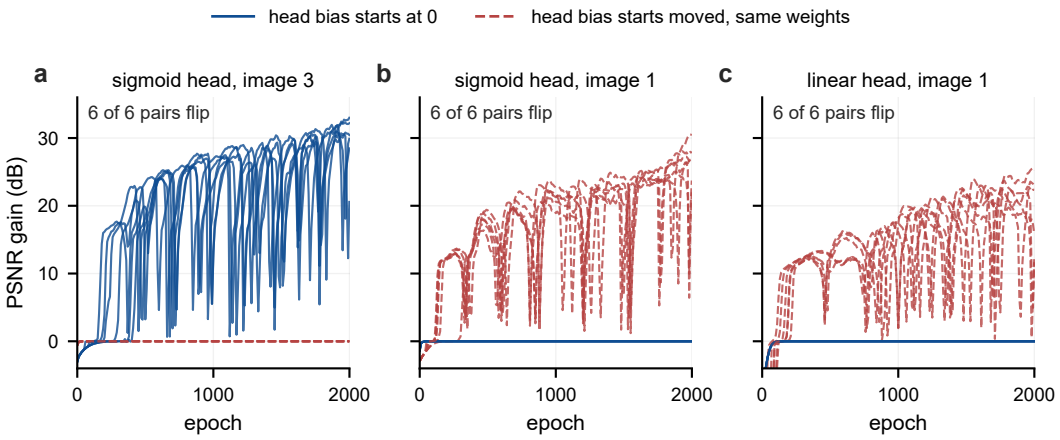

Figure 11: **The paired bias intervention.** Six networks per panel, each run twice from the same weights and data order. (a) Sigmoid head on the third image, bias moved onto the logit of the mean. (b) Sigmoid head on the first image, bias moved 1.0 below it. (c) Linear head on the first image, bias moved 1.0 below the mean.

The 84 runs of Figure 3(c) are the runs of Experiments 1t, 1v, 1w, 1x and 1y on these two banks. Ordered by the distance the bias must travel, none of 42 trains at a distance of at most 0.5, 2 of 10 train between 0.5 and 0.95, and 32 of 32 train at 0.95 or more.

K WHAT THE HIDDEN LAYERS DO BEFORE DEPARTURE

Experiment 5h reruns the four banks of Experiment 5c and the low bank on three Kodak images. Every 25 epochs it records how far each kernel has moved from its start, how much each layer's output still resembles its output at the start on 2048 fixed coordinates, and the gradient norm of each kernel on a fixed batch. Figure 12 shows three of these. Section 6.2 reports the displacements and the norms.

The step size does not follow the gradient. The first epoch moves every hidden kernel by 5.3 to 5.8 percent of its norm, while the norms of their gradients run from 0.4 to 1100. The gradient norm then falls to between 0.1 and 1.0 while a run sits at the mean colour, in 13 of 13 runs, and it does not rise again after departure in 9 of 13. The two $[8] \times 8$ runs that collapse drift instead, from near epoch 600, and end with a hidden displacement of 1.29 to 1.67 and a mean hidden norm at 1.67 and 1.83 times its start.

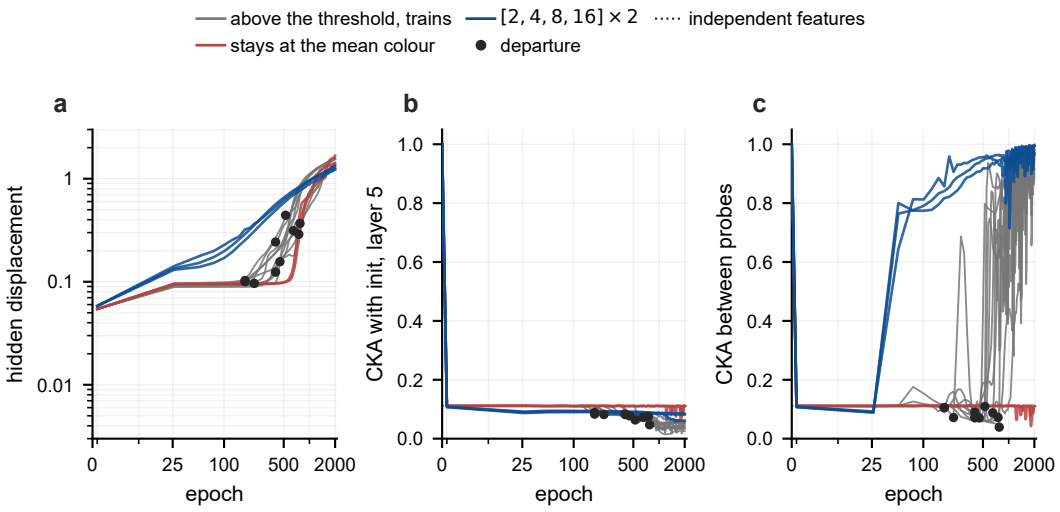

Figure 12: **Hidden layers before and after departure.** Fifteen runs of Experiment 5h. (a) Largest relative displacement of a hidden kernel. (b) CKA of the last hidden layer with its start. (c) CKA of the last hidden layer between consecutive records.

Panel (c) shows the churn that Section 6.2 reports, and Experiment 5h was not built to test it. Experiment 5i did, on six random $\tau = 5$ banks chosen by a rule written before any of them ran, three with a phase scale between 0.5 and 1.5 and three between 3.0 and 5.0, on two images. The three above collapse and the three below train, 12 of 12. The head norm falls to 0.01 to 0.02 of its start in the runs above and to 0.12 to 0.29 in the runs below.

Two of the four predictions of Experiment 5i failed. The runs below the boundary were expected to keep their deep features from epoch 50 to 500, and 4 of 6 did. The two that failed sat at the mean colour until epochs 325 and 375 and churned until then. The largest hidden kernel step between records was expected to stay under 1 percent of the initial norm, and it failed in 6 of 6, at 2.3 to 7.6 percent.

By the rule written down with the experiment, the churn does not enter the main text as evidence. It occurs on both sides of the boundary in these twelve runs and subsides near departure. The phase scale predicts whether a run leaves. It does not predict whether the features churn while it waits.