WHERE FOURIER COMPONENTS MEET: HOW FEATURES INTERACT IN GROKKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning useful components does not by itself explain how a network combines them into a computation. We develop a theoretical framework for feature interactions, using quadratic networks for modular addition to study how matching Fourier components combine across hidden units. An exact factorization separates interaction power into mode magnitude, block balance, and cross-block overlap, showing how cross-block pairing changes power while preserving each block's distribution of component strengths. We derive an exact state-dependent covariance law for overlap under gradient flow and a finite-step identity for realized optimizer updates. In Adam and Muon networks, within-block concentration accounts for most observed overlap growth, while relative alignment increases consistently. Fixed-marginal interventions that disrupt pairing degrade trained predictions, while phase-only controls preserve interaction power but alter predictions, showing the complementary role of readout compatibility. Within matched training configurations, grokking runs attain higher endpoint overlap. A population mean-squared-error lower bound connects interaction power and readout capacity to reconstruction. Together, these results link where learned components are placed to the interactions they form and the computations they support.

1 INTRODUCTION

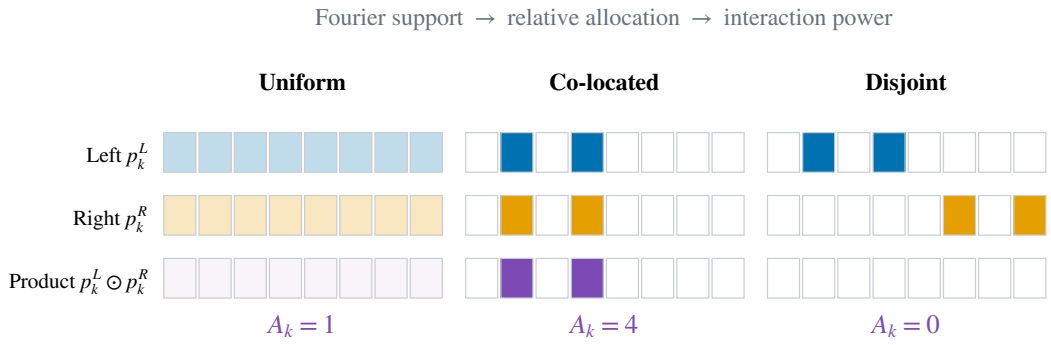
When a task combines inputs processed by separate weight blocks, learning must coordinate where their components are represented. Matching components can occupy different hidden units and contribute little interaction power. Their joint contribution therefore depends on both component strength and placement across neurons. How does training organize separately represented components into functional interactions?

Grokking separates training-set memorization from later generalization (Power et al., 2022), allowing us to study how fitting labels develops into transferable structure (Zhang et al., 2016; Arpit et al., 2017). Fourier analyses identify learned circuits, their formation and cleanup, and competition among frequencies (Nanda et al., 2023; Zhong et al., 2023; Ding et al., 2024). Broad Fourier support is already present at initialization. Identifying these components leaves open how their allocation across independent input blocks controls the interactions available to the network.

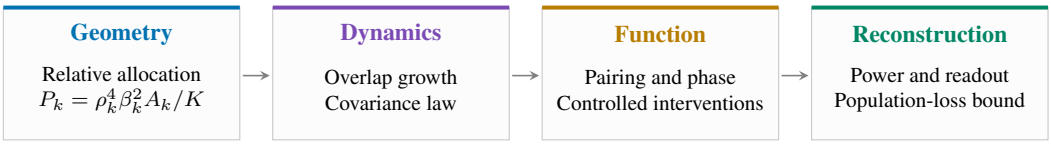
We study *relative allocation*: where matching components meet across independently parameterized input blocks. Identical block energies and within-block concentrations can support different interaction powers when components occupy different neurons (Figure 1(a)). Independent blocks permit changes in pairing at fixed amplitude distributions. This separates the statistics of each block from the geometry of their interaction. Tied blocks share one allocation, so their overlap reduces to the within-block concentration of that shared allocation.

Quadratic modular addition lets us distinguish component strength, relative allocation, interaction power, and readout compatibility through exact analysis and controlled interventions. Readout compatibility describes how the interaction direction projects onto the learned readout, connecting representation geometry to predictions. The network's bilinear cross term supports the same factorization for other frequency pairings, including modular subtraction (Appendix G).

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.



(a) Relative allocation



(b) Research framework

Figure 1: Relative allocation and research framework. **(a)** At fixed ρ_k, β_k, overlap A_k controls interaction power. Blue/orange show allocations p_k^L, p_k^R across eight neurons; purple shows their elementwise product. Uniform, co-located, and disjoint supports give $A_k = 1, 4, 0$. Co-located and disjoint supports share identical amplitude marginals, isolating pairing's effect on power. **(b)** Geometry, dynamics, interventions, and reconstruction link coordination to function (Sections 3–6).

Our analysis asks what determines interaction power, how optimization changes allocation, and how that geometry supports network function (Figure 1(b)). We make three contributions:

- We identify relative pairing across independent input blocks as a geometric degree of freedom at fixed amplitude marginals. An exact factorization separates interaction power into magnitude, block balance, and normalized cross-block overlap, which combines within-block concentration and relative alignment.
- We derive an exact state-dependent covariance law for overlap under gradient flow and a finite-step identity for realized optimizer updates. Finite-sample bounds quantify sampling perturbations to overlap dynamics and give conditional radial growth guarantees.
- We establish the functional role of allocation: disrupting cross-block pairing at fixed amplitude marginals degrades trained predictions, while phase-only controls show the complementary role of readout compatibility. In Adam and Muon networks, within-block concentration accounts for most observed overlap growth, alongside consistent increases in relative alignment. A population-loss bound connects interaction power and readout capacity to reconstruction.

2 Related Work

Accounts of grokking emphasize Fourier circuits, norm dynamics, regularization, and lazy-to-rich transitions (Nanda et al., 2023; Zhong et al., 2023; Liu et al., 2022b; Kumar et al., 2024; Varma et al., 2023; Lyu et al., 2024). Mechanistic analyses recover Fourier solutions and competition among circular representations (Gromov, 2023; Ding et al., 2024); Tian (2025) studies feature emergence and interactions among hidden neurons. Margin maximization characterizes Fourier feature emergence (Morwani et al., 2024), while adaptive-kernel theory connects feature learning to grokking (Rubin et al., 2024). We examine how matching components in independent input blocks jointly produce task-aligned interaction power.

He et al. (2026) characterize phase alignment, frequency diversification, and competition with shared input weights. They study which frequency occupies a neuron; independent blocks additionally ask whether matching components occupy the same neurons at fixed amplitude marginals. Relative pairing changes interaction power at fixed block energies and concentrations; tied blocks eliminate this freedom (Section 3.4). Appendix Table 11 compares these questions.

Optimization and regularization affect the grokking transition (Thilak et al., 2022; Manir & Rupa, 2026). Coupled regularization in Adam (Kingma & Ba, 2014) and decoupled weight decay induce different updates. Our covariance law separates the direct effect of isotropic shrinkage from its effect on the states reached during learning. Appendix H discusses connections to feature learning, Fourier representations, optimization, and subnetworks.

3 Geometry: From Components to Interaction Power

Task-aligned Fourier coordinates make the interaction required by the task explicit and separate its power into mode magnitude, block balance, and relative allocation.

3.1 Task and Network

Let $M \geq 3$ be the modulus and $\mathbb{Z}_M$ the integers from 0 through $M-1$, with arithmetic modulo M. For an input pair $(a, b) \in \mathbb{Z}_M^2$, the target class and one-hot label are

$$c(a, b) = (a + b) \bmod M, \qquad \mathbf{y}(a, b) = \mathbf{e}_{c(a,b)} \in \mathbb{R}^M.$$

The network receives the two concatenated one-hot input vectors. Following quadratic-network studies of grokking (Mohamadi et al., 2024; Tian, 2025), we use a bias-free network with one width-K quadratic hidden layer and a linear readout. Its input weight matrix is $W = [W_L \mid W_R]$, with $W_L, W_R \in \mathbb{R}^{K \times M}$ processing the first and second inputs:

$$h_j(a, b) = \big(W_L[j, a] + W_R[j, b]\big)^2, \qquad \widehat{\mathbf{y}}(a, b) = V\mathbf{h}(a, b),$$

where $j = 1, \ldots, K$ indexes hidden neurons, $\mathbf{h}$ collects their activations, and $V \in \mathbb{R}^{M \times K}$ is the readout matrix. For a training set $\mathcal{S} \subset \mathbb{Z}_M^2$, let $n = |\mathcal{S}|$ and $r = n/M^2$ denote its size and training fraction, respectively. We train with centered mean squared error (MSE), summed over classes and averaged over examples:

$$\mathcal{L}_{\text{train}} = \frac{1}{n} \sum_{(a,b) \in \mathcal{S}} \|C\widehat{\mathbf{y}}(a, b) - C\mathbf{y}(a, b)\|_2^2, \qquad C = I - \mathbf{1}\mathbf{1}^\top/M. \tag{1}$$

Here C centers classes, I is the $M \times M$ identity, and $\mathbf{1}$ is the all-ones vector. Centering removes the constant class mode, so Parseval expresses the loss over nonconstant class modes. Adam uses coupled ℓ_2 regularization; Muon uses decoupled decay on W and Adam on V, with coefficient λ_{reg}. For gradient flow, $\mathcal{L}_{\text{data}}$ denotes the uncentered half-squared loss and λ the isotropic shrinkage coefficient; Appendix B.2 gives the normalization conversion.

3.2 Task-Aligned Fourier Interactions

For a function $u : \mathbb{Z}_M \to \mathbb{C}$, we use the unitary discrete Fourier transform (DFT)

$$\widehat{u}(k) = \frac{1}{\sqrt{M}} \sum_{a=0}^{M-1} u(a)\omega_M^{-ka}, \qquad \omega_M = e^{2\pi i/M},$$

where $k = 0, \ldots, M-1$ is the frequency index and $i^2 = -1$. Matrix notation $[:, k]$ selects column k. The transform acts row-wise on the input weight matrices and has the standard two-dimensional analogue for functions of (a, b). Inner products use $\langle u, v \rangle = \sum_j \overline{u_j} v_j$; the overbar denotes complex conjugation and Re the real part.

For each nonzero frequency k, the interaction vector $q_k \in \mathbb{C}^K$ records neuronwise products of matching input components, and $\widetilde{V}_k \in \mathbb{C}^K$ describes their readout at class frequency k:

$$q_k[j] = \widehat{W}_L[j, k]\widehat{W}_R[j, k], \qquad \widetilde{V}_k[j] = \frac{1}{\sqrt{M}} \sum_{c=0}^{M-1} V[c, j]\omega_M^{-kc}. \tag{2}$$

When $q_k \neq 0$, the normalized vector $q_k/\|q_k\|_2$ defines its interaction direction. Quadratic activation yields the cross-input term $2W_L[j, a]W_R[j, b]$. For $k \neq 0$, the pure-square terms have no support at (k, k), leaving $\widehat{h}_j(k, k) = 2q_k[j]$. The task-aligned interaction power is $P_k := \sum_j |\widehat{h}_j(k, k)|^2 = 4\|q_k\|_2^2$. Class mode k couples to spatial mode $(-k, -k)$, with indices modulo M. Real weights give $P_{-k} = P_k$. Disjoint allocation makes every product vanish despite substantial energy in both blocks.

3.3 Exact Factorization of Interaction Power

At training time t, define the concatenated weight mode and its radius by

$$w_k(t) = \begin{bmatrix} \widehat{W}_L[:,k](t) \\ \widehat{W}_R[:,k](t) \end{bmatrix}, \qquad \rho_k(t) = \|w_k(t)\|_2. \tag{3}$$

Let $a_{jk} = |\widehat{W}_L[j,k]|^2$ and $b_{jk} = |\widehat{W}_R[j,k]|^2$ be the per-neuron block powers, with totals $S_k^L = \sum_j a_{jk}$ and $S_k^R = \sum_j b_{jk}$. Suppressing time, define $p_{jk}^L = a_{jk}/S_k^L$ and $p_{jk}^R = b_{jk}/S_k^R$ whenever block energy is positive. Each entry is a neuron's share of that block's mode energy. At fixed block energies, interaction power depends on how these two normalized allocations overlap across neurons. The following identity separates this dependence from total mode magnitude and left–right energy balance.

Proposition 1 (Exact mode-power factorization)**.** *For every nonzero Fourier mode with $S_k^L, S_k^R > 0$, define the block balance β_k and normalized cross-block overlap A_k by*

$$\beta_k = \frac{2\sqrt{S_k^L S_k^R}}{\rho_k^2}, \qquad A_k = \frac{K \sum_j a_{jk} b_{jk}}{S_k^L S_k^R}.$$

Then its task-aligned interaction power satisfies

$$P_k = \rho_k^4 \beta_k^2 \frac{A_k}{K}, \qquad 0 < \beta_k \le 1, \quad 0 \le A_k \le K. \tag{4}$$

For $P_k > 0$, taking logarithms gives $\log P_k = 4 \log \rho_k + 2 \log \beta_k + \log A_k - \log K$.

The proof is in Appendix B.6. The radius measures total mode magnitude; $\beta_k = 1$ for equal block energies and decreases with imbalance. At fixed amplitude profiles, changes in pairing act through A_k. This product structure extends the decomposition to separable bilinear features, with an overall scale set by the bilinear coefficient. Addition and subtraction select different frequency pairs but share this geometry (Appendix G).

3.4 Cross-Block Overlap and Within-Block Concentration

The normalization gives $A_k = 1$ for uniform allocations and $\mathbb{E}A_k = 1$ for independent, exchangeable allocations. Larger values indicate greater overlap: concentration on one shared neuron gives $A_k = K$, whereas disjoint supports give zero. If either block has zero energy, $P_k = 0$ and normalized overlap is undefined.

Tied blocks share the allocation $p_k = p_k^L = p_k^R$, so $A_k = K\|p_k\|_2^2$. Independent blocks additionally have a relative-alignment degree of freedom:

$$A_k = K\langle p_k^L, p_k^R \rangle = K\|p_k^L\|_2 \|p_k^R\|_2 \cos\theta_k,$$

where θ_k is the angle between allocation vectors. The norms describe each block's concentration; the angle describes their alignment. During training, both can change A_k. At fixed amplitude marginals, changes in A_k isolate relative alignment. For example, let each block be uniform on s neurons with $2s \le K$. Both squared norms equal $1/s$, but coincident supports give $A_k = K/s$ and disjoint supports give zero (Figure 1(a), $K = 8$, $s = 2$). The same freedom motivates the fixed-marginal intervention in Section 5.1. We first derive how it evolves under optimization.

4 Dynamics: Learning Cross-Block Overlap

4.1 Cross-Block Overlap under Gradient Flow

We track two aspects of each mode: its overall magnitude through ρ_k and its cross-block overlap through A_k. Overlap grows when amplification in one block favors neurons carrying the matching component in the other.

Under gradient flow $\dot{W} = -\eta(\nabla_W \mathcal{L}_{\text{data}} + \lambda W)$, $\eta > 0$ is the learning-rate scale, λ the isotropic shrinkage coefficient, and dots denote time derivatives. The radius evolves through data-driven growth and shrinkage:

$$\dot{\rho}_k = \eta\rho_k(\Gamma_k^{\text{data}} - \lambda), \qquad \Gamma_k^{\text{data}} = -\frac{\text{Re}\langle w_k, \nabla_{w_k}\mathcal{L}_{\text{data}}\rangle}{\rho_k^2}. \tag{5}$$

Here $\nabla_{w_k}\mathcal{L}_{\text{data}}$ is the Fourier-transformed weight gradient. The radius grows when its data-driven rate Γ_k^{data} exceeds shrinkage λ. Contraction limits power through $P_k \le \rho_k^4$; balance and overlap determine power at a given radius (Appendix B.2).

Allocation depends on how growth is distributed across neurons. Let $d^L = -\widehat{\nabla_{W_L}\mathcal{L}_{\text{data}}}$ be the left block's negative data gradient in Fourier coordinates, and define its per-neuron growth signal by

$$\gamma_{jk}^L = \frac{\text{Re}(\overline{\widehat{W}_L[j,k]}\, d_{jk}^L)}{a_{jk}}.$$

Define d^R and γ_{jk}^R analogously. Before shrinkage, these signals give $\dot{a}_{jk} = 2\eta a_{jk}\gamma_{jk}^L$ and $\dot{b}_{jk} = 2\eta b_{jk}\gamma_{jk}^R$. To compare growth with existing power, write $\langle\cdot\rangle_a$ and $\langle\cdot\rangle_b$ for neuron averages weighted by a_{jk}/S_k^L and b_{jk}/S_k^R. These averages emphasize neurons carrying more power in the corresponding block; we suppress k inside them.

Proposition 2 (Cross-block overlap dynamics)**.** *On intervals with positive per-unit block powers and $A_k > 0$, under gradient flow with learning rate η and isotropic ℓ_2 regularization λ,*

$$\frac{d\log A_k}{dt} = 2\eta\left(\frac{\text{Cov}_a(b_{\cdot k}, \gamma_{\cdot k}^L)}{\langle b\rangle_a} + \frac{\text{Cov}_b(a_{\cdot k}, \gamma_{\cdot k}^R)}{\langle a\rangle_b}\right). \tag{6}$$

Here $\text{Cov}_a(x,y) = \langle xy\rangle_a - \langle x\rangle_a\langle y\rangle_a$, with Cov_b defined analogously; $b_{\cdot k}$ denotes the vector over neurons at fixed k.

The first covariance is positive when left-block growth favors neurons with greater right-block power; the second compares right-block growth with left-block power. Overlap grows when their normalized sum is positive (Appendix Figure 6(a)). This exact state-dependent rate law uses the current parameters and data gradient. Uniform shrinkage cancels directly while changing the states at which the covariances act. Appendix B.6 proves the identity and gives companion laws for P_k and β_k.

Central finite differences agree with the rate (Appendix E.3). Finite-sample bounds quantify sampling effects and preserve radial growth or decay when task-driven growth and shrinkage differ by more than the residual bound (Appendices B.4–B.5).

4.2 Overlap under Finite-Step Updates

For a finite optimizer step, we use the realized change in power in place of its instantaneous growth rate. Let t index steps and define $\delta_{jk}^L = (a_{jk,t+1} - a_{jk,t})/a_{jk,t}$, with δ_{jk}^R defined analogously. Thus $1 + \delta_{jk}^{L,R}$ is the corresponding power amplification factor. With positive per-unit powers before the step and positive block totals afterward, any optimizer satisfies

$$\frac{A_{k,t+1}}{A_{k,t}} = \frac{\left\langle (1+\delta^L)(1+\delta^R)\right\rangle_P}{\left(1 + \langle\delta^L\rangle_a\right)\left(1 + \langle\delta^R\rangle_b\right)}, \tag{7}$$

Here $\langle z\rangle_P = \sum_j a_{jk}b_{jk}z_j / \sum_j a_{jk}b_{jk}$ weights neurons by interaction power, with all weights evaluated at t. The numerator measures joint amplification where components meet; the denominator is the product of the blockwise-average amplification factors. Overlap increases when the former exceeds the latter. Uniform amplification within each block cancels between numerator and denominator. The overlap change therefore measures redistribution across neurons after accounting for changes in the two blocks' total powers. Momentum, preconditioning, and regularization enter through δ^L, δ^R. Recorded optimizer states reconstruct Adam and Muon updates directly (Appendix E.3); Appendix B.4 gives the radial relation.

4.3 Covariance Decomposition of Mode-Power Growth

We turn from instantaneous covariance across neurons to accumulated changes across modes. Let Δ denote the change from initialization to final weights. The log-power factorization gives

$$\Delta \log P_k = 4\Delta \log \rho_k + 2\Delta \log \beta_k + \Delta \log A_k. \tag{8}$$

Write these three terms as $\mathcal{R}_k = 4\Delta \log \rho_k$, $\mathcal{B}_k = 2\Delta \log \beta_k$, and $\mathcal{C}_k = \Delta \log A_k$, respectively. They record growth through radius, balance, and overlap.

We evaluate every nonconstant mode in 55 Adam and Muon runs, all with positive powers at both endpoints (Appendix E.1). To measure how each factor covaries with power growth across modes, take v to be $\mathcal{R}$, $\mathcal{B}$, or $\mathcal{C}$ and define its additive covariance contribution, when the denominator is positive,

$$s_v = \frac{\mathrm{Cov}_k(\Delta \log P_k, v_k)}{\mathrm{Var}_k(\Delta \log P_k)}.$$

Here Cov_k and Var_k weight all nonconstant modes equally. A positive s_v means that modes with greater power growth also tend to have larger changes in factor v. These descriptive contributions sum to one within each run; individual terms can be negative or exceed one. The overlap term includes changes in both within-block concentration and relative alignment (Section 3.4).

Overlap has the largest median additive covariance contribution: 80.9% (95% cluster-bootstrap interval $[80.4\%, 81.7\%]$). Decomposing $\Delta \log A_k$ gives median contributions of 70.5% for within-block concentration and 9.9% for relative alignment. The relative-alignment covariance term is positive in all 55 runs across both optimizers. Each run's contributions sum exactly; separate across-run medians need not sum to the median overlap contribution (Appendix D.2). We next test allocation's functional role and relation to learning outcomes.

5 Function: Allocation and Generalization

Following prior grokking studies (Power et al., 2022; Mohamadi et al., 2024; Tian, 2025), our core overlap experiments use 65 Adam and Muon trajectories on modular arithmetic data at $M = 127$, $K = 512$, spanning both outcomes near each optimizer's transition region. A configuration specifies optimizer and training fraction; each configuration contains five random initializations (Appendix F.1). Appendix Table 2 maps the additional studies to their measurements. Real weights pair frequencies k and $M - k$, leaving 63 nonredundant nonconstant modes. Their *aggregate cross-block overlap* is the geometric mean

$$G_A(t) = \exp\left(\frac{1}{63}\sum_{k=1}^{63} \log A_k(t)\right). \tag{9}$$

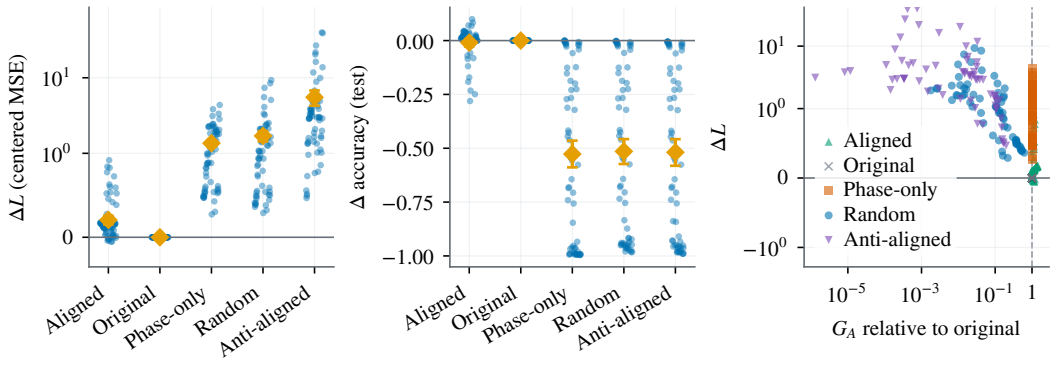
The geometric mean aggregates the multiplicative factor A_k on the logarithmic scale of equation 8. The mode set is fixed across time and outcomes, with equal weight irrespective of power.

5.1 Fixed-Marginal Allocation and Phase Interventions

With the readout fixed, amplitude reallocation changes interaction power and direction at fixed amplitude marginals; phase-only controls probe direction at fixed power. At the final checkpoint, write $\widehat{W}_R[j,k] = r_{jk}e^{i\phi_{jk}}$, with amplitude $r_{jk} \geq 0$ and phase ϕ_{jk}. A neuron permutation π_k reallocates right-block amplitudes; primes denote the transformed model:

$$\widehat{W}'_R[j,k] = r_{\pi_k(j),k}e^{i\phi_{jk}}, \qquad k = 1, \ldots, 63. \tag{10}$$

Each neuron keeps its phase; the left block, readout, and zero-frequency component stay fixed. Conjugate coefficients preserve real weights. The permutation preserves both amplitude multisets, and hence ρ_k, β_k, and within-block concentrations. Equation 4 gives $P'_k/P_k = A'_k/A_k$ for positive reference modes. Anti-alignment pairs the largest left amplitudes with the smallest right amplitudes; alignment pairs them in the same order. Random controls shuffle amplitudes uniformly, while phase-only controls randomize phases.



(a) Held-out loss (b) Held-out accuracy (c) Loss and overlap

Figure 2: Allocation and phase interventions with fixed readout. **(a)–(b)** Held-out loss and accuracy changes across 65 runs. Points show runs; diamonds show equal-weight configuration means with 95% cluster bootstrap intervals. Random-amplitude and phase-only repeats are averaged within runs. **(c)** Loss change versus overlap ratio. Panels (a) and (c) use symmetric-logarithmic vertical scales, linear on $[-1, 1]$; the dashed line in (c), $G_A/G_A^{\text{orig}} = 1$, marks unchanged overlap.

Anti-alignment increases held-out centered MSE by 5.518 (95% interval $[4.264, 6.786]$) and reduces accuracy by 51.9 percentage points (95% interval $[45.8, 58.1]$). Both effects hold across configurations (Figure 2).

Among models with original endpoint accuracy at least 0.95, the mean reduction across configurations is 95.7 points and loss increases by 8.121 (Appendix F.2). Random permutations increase loss by 1.701. Aligned permutations maximize overlap within the permutation family and increase loss by 0.207, with a mean accuracy reduction of 1.0 point. We evaluate each transformation before further training, measuring its immediate effect on the learned computation.

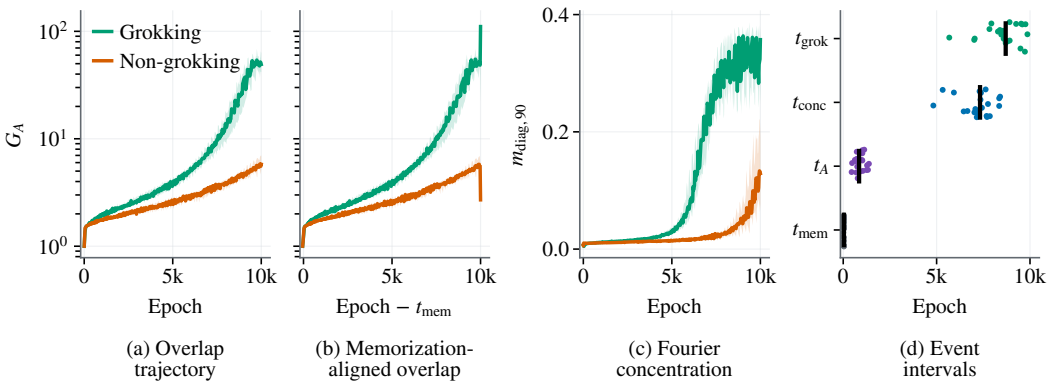
Anti-alignment also reduces the mean absolute readout projection onto interaction directions. Its effect on full-network loss includes changes in diagonal, off-diagonal, and pure-square features (Appendix F.1), establishing functional sensitivity to cross-block pairing at fixed amplitude marginals. Phase-only controls preserve A_k and P_k but increase loss by 1.360, showing the complementary role of readout compatibility. Relabeling both blocks and the readout, or restoring the original pairing, preserves predictions and provides numerical controls.

5.2 Overlap Trajectories and Learning Outcomes

We trace G_A through learning, marking memorization and grokking by the first crossings of 99% training accuracy and 95% held-out accuracy. "Non-grokking" denotes runs below the latter threshold through epoch 10,000. Overlap starts near one and separates by outcome after alignment to memorization (Figure 3(a–b)). The aggregate relative-alignment cosine increases from initialization to endpoint in all 65 runs: its across-run geometric mean rises from 0.505 to 0.874 (Appendix D.2).

We also track *task-aligned Fourier concentration*: each neuron's fraction of nonconstant activation energy at diagonal frequencies (k, k). The statistic $m_{\text{diag},90}$ is its 90th percentile across neurons. This activation statistic measures the expression of task interactions; within-block concentration measures a fixed input component's allocation across neurons.

We compare accuracy events with overlap doubling, $G_A(t) \geq 2G_A(0)$, and concentration reaching $m_{\text{diag},90} \geq 0.30$. Events lie between the last observation below threshold and the first at or above it; separated intervals establish their order (Appendix D.3). Among trajectories reaching both thresholds, overlap doubling follows memorization and precedes concentration; concentration precedes generalization in all 22 grokking runs in this sample. Overlap crossing also precedes grokking at thresholds of 1.5 and 4 times its initial value. Doubling also occurs in most non-grokking runs, motivating comparison of final overlap levels.



(a) Overlap trajectory (b) Memorization-aligned overlap (c) Fourier concentration (d) Event intervals

Figure 3: Overlap and event ordering. **(a)–(b)** G_A at absolute and memorization-aligned epochs. **(c)** Task-aligned concentration $m_{\text{diag},90}$. Lines show outcome-group medians; bands show pointwise 95% cluster bootstrap intervals. Panels (a,c) use 22 grokking and 43 non-grokking runs; (b) uses 4–22 and 4–43 per offset, respectively, taking the nearest observation within 12.5 epochs without interpolation. **(d)** Memorization (t_{mem}), overlap doubling (t_A), concentration (t_{conc}), and grokking (t_{grok}) intervals for 22 grokking runs; black marks show median upper endpoints. Appendix Figure 5 shows grokking-aligned overlap and held-out accuracy.

Within six configurations containing both outcomes, grokking runs have higher endpoint overlap: the configuration-adjusted geometric-mean ratio is 4.976 (95% interval $[3.738, 6.976]$; Figure 4). The same comparison gives ratios of 4.562 for within-block concentration and 1.091 for relative alignment. Concentration supplies most of the contrast; alignment adds a smaller, consistent difference, positive in all six configuration means, with a pooled 95% interval $[1.058, 1.123]$ (Appendix D.2). Accounting for initialization differences also preserves the positive alignment association, linking learning outcomes to changes acquired during training. These comparisons hold the optimizer and training fraction fixed. Together with the trajectories, they show that overlap growth accompanies both outcomes, while grokking runs reach higher endpoint levels. The fixed-marginal interventions complement this association by directly changing cross-block pairing in trained networks.

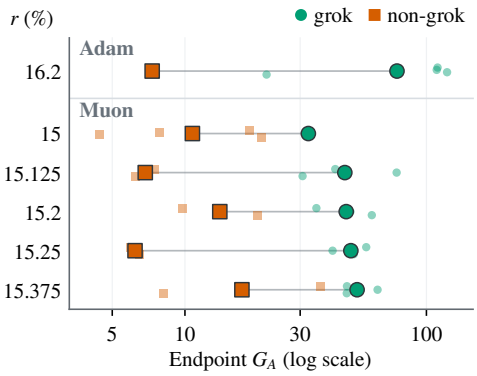


Figure 4: Endpoint overlap (30 runs, six mixed-outcome configurations). Small markers show runs; large outlined markers show geometric means. Thin horizontal lines connect the two outcome-group geometric means within each configuration. r is the training fraction.

5.3 Task-Aligned Concentration and Generalization

We next examine how neuronwise interactions appear in the hidden-activation spectrum. In a complementary study of 250 Adam and Muon runs, grokking models have higher final task-aligned concentration within mixed-outcome configurations, with a weighted difference of 0.141 in this energy-fraction statistic (Appendix D.6). At a larger width and smaller modulus, a comparison with matched initializations links coupled regularization to increased concentration and grokking (Appendix Figure 8). Overlap measures how input components share hidden units; concentration summarizes the per-neuron fraction of nonconstant activation energy devoted to task-aligned interactions. The nonconstant spectrum also includes off-diagonal and pure-square contributions. Greater overlap increases diagonal power at fixed radius and balance, while concentration also depends on the remaining activation energy. We next connect this interaction power to reconstruction through the readout.

6 Reconstruction: Interaction Power and Readout Capacity

The geometry and interventions identify two requirements for reconstruction: task-aligned interaction power and readout compatibility. A mode-wise lower bound connects the power requirement to population loss at a given readout norm. Let $Y_k(a,b)$ and $T_k(a,b)$ denote the class-axis DFT coefficients of the network output $\widehat{\mathbf{y}}(a,b)$ and target $\mathbf{y}(a,b)$, respectively. The population loss L_{pop} evaluates the centered MSE in equation 1 over all M^2 input pairs:

$$L_{\text{pop}} = \frac{1}{M^2} \sum_{a,b} \sum_{k \neq 0} \left| Y_k(a,b) - T_k(a,b) \right|^2. \tag{11}$$

For each nonconstant class mode k, the spatial DFT of T_k has target amplitude $\sqrt{M}$ at $(-k,-k)$. At this task-aligned spatial frequency, the network amplitude is controlled jointly by the interaction power P_k and the class-frequency readout norm $\|\widetilde{V}_k\|_2$.

Corollary 3 (Mode-wise lower bound on population loss)**.** *On the full grid,*

$$L_{\text{pop}} \;\geq\; \underline{L}_{\text{pop}} := \frac{1}{M^2} \sum_{k \neq 0} \left(\sqrt{M} - \|\widetilde{V}_k\|_2 \sqrt{P_k} \right)_+^2, \qquad \sqrt{P_k} = \rho_k^2 \beta_k \sqrt{A_k / K}. \tag{12}$$

The factorized expression applies when $S_k^L, S_k^R > 0$; if either block has zero energy, $P_k = 0$ and the bound uses $\sqrt{P_k} = 0$ directly.

Here $(x)_+ = \max(x, 0)$ and $\underline{L}_{\text{pop}}$ denotes the bound; Appendix B.7 gives its proof. Each summand is the squared shortfall between the target amplitude and the maximum supported by interaction power and readout norm. At fixed radius, balance, and readout norm, larger overlap raises this capacity and weakly lowers the bound. Phase-only transformations leave the bound unchanged while increasing measured loss: direction determines how much capacity the readout realizes.

Figure 1(a) illustrates the extremes: disjoint supports give $P_k = 0$, leaving the target coefficient unreconstructed. Greater overlap enlarges the attainable amplitude. Once it reaches the target, that mode contributes zero to the bound; reconstruction depends on the readout projection and unwanted spatial frequencies. Amplitude and phase interventions in Figure 2 act on these complementary requirements.

The bound concerns the centered MSE objective. Argmax classification admits sparser exact solutions, including a two-mode cosine classifier (Appendix B.7); its spectral requirements differ from reconstructing target outputs over all input pairs. At trained checkpoints from sixteen runs, the bound-to-loss ratio has median 23.24% (Appendix Figure 9). This ratio quantifies the tightness of the capacity bound relative to measured population loss.

7 Discussion

This work identifies relative allocation as a geometric link between represented components and functional interactions. In the quadratic setting, matching Fourier components contribute jointly through their placement across input blocks, and the learned readout determines how the resulting interaction power supports computation. The analysis connects this geometry to optimization and network function in an analytically tractable setting.

Natural training and fixed-marginal interventions address complementary aspects of coordination. Training strengthens overlap mainly through within-block concentration, alongside systematic growth in relative alignment. Holding the marginals fixed exposes the functional role of pairing, while the readout determines how the resulting interactions support predictions. This distinction connects the development of representation geometry to its use in computation.

Future work can test whether analogous coordination mechanisms arise across depth and through nonlinear interactions, including those in attention circuits (Olsson et al., 2022). Extending this approach calls for interaction variables suited to each architecture and interventions that preserve relevant component statistics. Such studies could clarify how the organization of learned components shapes the computations they support.

AI Use Statement

Generative AI tools assisted with manuscript polishing, checking and refining proof steps, suggesting revisions to existing analysis and plotting code, and checking the implementation and presentation of methods already specified by the authors. The authors independently determined the core research direction, experimental design, statistical methods, and interpretation of scientific results. All mathematical claims, proofs, experimental results, and AI-assisted content were independently verified by the authors. Code changes suggested during debugging were manually reviewed and tested. The authors take full responsibility for the final content of the paper.

Ethics Statement

This work studies learning dynamics using synthetically generated modular arithmetic tasks. It involves no human participants, personal data, or sensitive real-world records.

Reproducibility Statement

Proofs and their assumptions appear in Appendix B; training settings, numerical details, software versions, and hardware appear in Appendix I.

References

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019. 37

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pp. 233–242. PMLR, 2017. 1

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. PMLR, 2020. 37

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. 38

Etienne Boursier, Scott Pesme, and Radu-Alexandru Dragomir. A theoretical framework for grokking: Interpolation followed by riemannian norm minimisation. *Advances in Neural Information Processing Systems*, 38:151231–151267, 2026. 31

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019. 37

Xander Davies, Lauro Langosco, and David Krueger. Unifying grokking and double descent. *arXiv preprint arXiv:2303.06173*, 2023. 38

Xiaoman Delores Ding, Zifan Carl Guo, Eric J Michaud, Ziming Liu, and Max Tegmark. Survival of the fittest representation: A case study with modular addition. *arXiv preprint arXiv:2405.17420*, 2024. 1, 2, 37, 38

Darshil Doshi, Aritra Das, Tianyu He, and Andrey Gromov. To grok or not to grok: Disentangling generalization and memorization on corrupted algorithmic datasets. In *International Conference on Learning Representations*, volume 2024, pp. 3782–3810, 2024a. 37

Darshil Doshi, Tianyu He, Aritra Das, and Andrey Gromov. Grokking modular polynomials. *arXiv preprint arXiv:2406.03495*, 2024b. 37

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. 38

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019. 38

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018. 38

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International conference on machine learning*, pp. 3259–3269. PMLR, 2020. 38

Hiroki Furuta, Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. Interpreting grokked transformers in complex modular arithmetic. *arXiv preprint arXiv:2402.16726*, 2024. 37

Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023. 2, 37

XY Han, Vardan Papyan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021. 23

Jianliang He, Leda Wang, Siyu Chen, and Zhuoran Yang. On the mechanism and dynamics of modular addition: Fourier features, lottery ticket, and grokking. *arXiv preprint arXiv:2602.16849*, 2026. 2, 37, 38

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018. 37

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019. 38

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

Tanishq Kumar, Blake Bordelon, Samuel Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *International Conference on Learning Representations*, volume 2024, pp. 23010–23035, 2024. 2, 37, 38

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663, 2022a. 37

Ziming Liu, Eric J Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. *arXiv preprint arXiv:2210.01117*, 2022b. 2, 37

Ziming Liu, Ziqian Zhong, and Max Tegmark. Grokking as compression: A nonlinear complexity perspective. *arXiv preprint arXiv:2310.05918*, 2023. 37

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Du, Jason Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *International Conference on Learning Representations*, volume 2024, pp. 33897–33936, 2024. 2, 37

Neil Mallinar, Daniel Beaglehole, Libin Zhu, Adityanarayanan Radhakrishnan, Parthe Pandit, and Mikhail Belkin. Emergence in non-neural models: grokking modular arithmetic via average gradient outer product. *arXiv preprint arXiv:2407.20199*, 2024. 37, 38

Shalima Binta Manir and Anamika Paul Rupa. A systematic empirical study of grokking: Depth, architecture, activation, and regularization. *arXiv preprint arXiv:2603.25009*, 2026. 3, 38

Jack Miller, Charles O'Neill, and Thang Bui. Grokking beyond neural networks: An empirical exploration with model complexity. *arXiv preprint arXiv:2310.17247*, 2023. 38

Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. Bridging lottery ticket and grokking: Understanding grokking from inner structure of networks. 2025. 37, 38

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J Sutherland. Why do you grok? a theoretical analysis of grokking modular addition. *arXiv preprint arXiv:2407.12332*, 2024. 3, 6, 37, 38

Depen Morwani, Benjamin Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: case studies in algebraic tasks. In *International Conference on Learning Representations*, volume 2024, pp. 29077–29114, 2024. 2, 37

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021. 38

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023. 1, 2, 37, 38

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018. 38

Pascal Jr Tikeng Notsawo, Guillaume Dumas, and Guillaume Rabusseau. Grokking beyond the euclidean norm of model parameters. *arXiv preprint arXiv:2506.05718*, 2025. 37

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022. 9

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021. 37

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022. 1, 6

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. 37

Noa Rubin, Inbar Seroussi, and Zohar Ringel. Grokking as a first order phase transition in two layer networks. In *International Conference on Learning Representations*, volume 2024, pp. 23881–23904, 2024. 2, 37

Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pp. 39–48, 1974. 15

Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001, 2019. 38

Anand Swaroop. Latent algorithmic structure precedes grokking: A mechanistic study of relu mlps on modular arithmetic. *arXiv preprint arXiv:2603.23784*, 2026. 37

Hu Tan, Kuo Gai, and Shihua Zhang. Beyond neural collapse: Task-intrinsic geometry governs neural representations in modular arithmetic. *arXiv preprint arXiv:2606.08985*, 2026. 23

Zhiquan Tan and Weiran Huang. Understanding grokking through a robustness viewpoint. *arXiv preprint arXiv:2311.06597*, 2023. 37

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. *arXiv preprint arXiv:2206.04817*, 2022. 3, 38

Yuandong Tian. Provable scaling laws of feature emergence from learning dynamics of grokking. *arXiv preprint arXiv:2509.21519*, 2025. 2, 3, 6, 37, 38

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023. 2, 37

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020. 37

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016. 1

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. *Advances in neural information processing systems*, 36:27223–27250, 2023. 1, 2, 37

A NOTATION

Table 1 summarizes the notation used in the main argument and its proofs. Symbols specific to individual bounds are defined where those bounds are stated.

Table 1: Principal notation, grouped by its role in the analysis. Normalized allocations, balance, and overlap require positive energy in both input blocks.

Symbol	Meaning
Network and Fourier coordinates (Section 3)	
$M, K, \mathcal{S}, n, r$	Modulus, hidden width, training set, sample count $n = \lvert\mathcal{S}\rvert$, and training fraction $r = n/M^2$.
W_L, W_R, V	Left and right input-weight blocks and the linear readout.
$q_k, \widetilde{V}_k$	Neuronwise products of matching Fourier components and the class-frequency readout.
w_k, ρ_k	Concatenated input-weight mode and its norm $\rho_k = \lVert w_k \rVert_2$.
Interaction geometry (Sections 3.3–3.4)	
$a_{jk}, b_{jk};\ S_k^L, S_k^R$	Per-neuron block powers $\lvert\widehat{W}_L[j,k]\rvert^2$, $\lvert\widehat{W}_R[j,k]\rvert^2$ and their sums; $\rho_k^2 = S_k^L + S_k^R$.
p_{jk}^L, p_{jk}^R	Shares of mode energy across neurons: a_{jk}/S_k^L, b_{jk}/S_k^R.
β_k	Block balance $2\sqrt{S_k^L S_k^R}/\rho_k^2 \in (0, 1]$.
A_k, P_k	Normalized cross-block overlap $K\langle p_k^L, p_k^R\rangle \in [0, K]$ and interaction power $P_k = \rho_k^4 \beta_k^2 A_k / K$.
G_A	Geometric mean of A_k over the fixed set of nonconstant conjugate-pair representatives, equation 9.
Learning dynamics (Section 4.1; Appendix B.4)	
$\eta, \lambda, \lambda_{\text{reg}}$	Gradient-flow time scale, isotropic shrinkage coefficient, and the implemented regularization/decay coefficient.
Γ_k^{data}	Normalized radial contribution of the data gradient, equation 5.
$\gamma_{jk}^L, \gamma_{jk}^R$	Per-neuron relative power-growth signals from the two blocks' data gradients.
$\delta_{jk}^L, \delta_{jk}^R$	Relative block-power changes in one optimizer step, equation 7.
d_k^v, C_k^v	Negative-gradient component and its contribution to $d \log A_k / dt$.
$\mathcal{R}_k, \mathcal{B}_k, \mathcal{C}_k$	Logarithmic changes in ρ_k^4, β_k^2, and A_k. Their normalized covariance contributions $s_{\mathcal{R}}, s_{\mathcal{B}}, s_{\mathcal{C}}$ sum to one.
$\kappa_k, B_{n,\zeta}$	Maximum weighted coefficient of variation controlling overlap-rate sensitivity, and the uniform finite-sample gradient bound (Appendices B.4–B.5).
Representation and reconstruction (Sections 5.2–6)	
$m_{\text{diag},90}$	90th percentile across neurons of the task-aligned fraction of nonconstant activation energy.
$t_{\text{mem}}, t_A, t_{\text{conc}}, t_{\text{grok}}$	First crossings of the memorization, overlap, concentration, and grokking thresholds.
S_{diag}	Total diagonal interaction power per hidden unit, equation 37.
D_k, Q_k	Stabilized readout projection onto the interaction direction and squared readout norm (Appendix C.2).
Y_k, T_k	Class-axis Fourier coefficients of the network output and target.
$L_{\text{pop}}, \underline{L}_{\text{pop}}$	Centered, class-summed population MSE and its mode-wise lower bound, equation 11–equation 12.

B PROOFS

We first establish the Fourier decomposition and radial dynamics, then quantify finite-sample and optimizer effects (Appendices B.1–B.5). Appendix B.6 derives the power factorization, overlap dynamics, and initialization scale. Appendix B.7 connects interaction power to population reconstruction.

B.1 EMPIRICAL CORRELATIONS BETWEEN FOURIER CHARACTERS

For a finite training set, let $\chi_k(a,b) = \omega_M^{-k(a+b)}$ denote the kth modular-addition character. The empirical Gram matrix of these task characters has entries

$$G_{k\ell}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{(a,b)\in\mathcal{S}} \chi_k(a,b)\overline{\chi_\ell(a,b)}. \tag{13}$$

Distinct task frequencies are orthogonal on the full input grid. Finite sampling introduces correlations; the next bound controls the largest one across all frequency pairs. Model-dependent gradient terms enter the remainder in Appendix B.2.

Proposition 4 (Concentration of the empirical Fourier Gram matrix)**.** *Let $M \geq 3$, $1 \leq n \leq M^2$, and $0 < \zeta < 1$. If $\mathcal{S}$ is sampled uniformly from the size-n subsets of $\mathbb{Z}_M^2$, then an absolute constant $C > 0$ satisfies, with probability at least $1-\zeta$,*

$$\max_{\substack{1\leq k,\ell\leq M-1\\ k\neq\ell}} |G_{k\ell}(\mathcal{S})| \leq C\sqrt{\frac{\log(2(M-1)^2/\zeta)}{|\mathcal{S}|}}. \tag{14}$$

For fixed $k \neq \ell$, the summands in equation 13 have modulus one and their full-grid mean is zero:

$$\frac{1}{M^2}\sum_{a,b=0}^{M-1} \omega_M^{-(k-\ell)(a+b)} = 0.$$

Serfling's inequality for sampling without replacement, applied to the real and imaginary parts (Serfling, 1974), therefore gives

$$\mathbb{P}(|G_{k\ell}(\mathcal{S})| \geq t) \leq 4\exp\bigl(-c|\mathcal{S}|t^2\bigr)$$

for an absolute constant $c > 0$. There are fewer than $(M-1)^2$ ordered off-diagonal pairs. A union bound and $t = C\sqrt{\log(2(M-1)^2/\zeta)/|\mathcal{S}|}$ prove equation 14.

B.2 RADIAL DYNAMICS AND SUFFICIENT DECAY CONDITIONS

Loss convention. For the analytic radial formulas we use

$$\mathcal{L}_{\text{data}}(W,V) = \frac{1}{2|\mathcal{S}|}\sum_{(a,b)\in\mathcal{S}} \|\widehat{\mathbf{y}}(a,b) - \mathbf{y}(a,b)\|^2, \qquad \mathcal{J} = \mathcal{L}_{\text{data}} + \frac{\lambda}{2}\bigl(\|W\|_F^2 + \|V\|_F^2\bigr),$$

where λ is the isotropic gradient-flow regularization coefficient. The actual centered training objective is equation 1. Converting the nonconstant label-interaction term to that convention multiplies it by two; empirical gradient reconstruction uses the centered training loss throughout. The mode-power factorization, covariance form, and realized-update identities hold under either convention with the corresponding data gradient.

Let g_k^{data} be the same unitary Fourier transform applied to the real-space gradient $\nabla_W \mathcal{L}_{\text{data}}$, restricted to mode k at a checkpoint on the trajectory trained with $\mathcal{S}$. Define g_k^{task} from the W-gradient of the *full-grid* nonconstant label interaction $\mathcal{L}_{\text{label}}^{\neq 0}$ in equation 23, evaluated at the same parameters, and set $g_k^{\text{rem}} = g_k^{\text{data}} - g_k^{\text{task}}$. Thus the remainder contains the prediction-norm gradient, the empirical–population label discrepancy, and every other term absent from the population interaction. For $\rho_k > 0$, define the normalized radial component for each gradient term by

$$\Gamma_k^u(t,\mathcal{S}) = -\frac{\mathrm{Re}\langle w_k, g_k^u\rangle}{\rho_k^2}.$$

Here u denotes the data, task, or remainder component. This definition gives the exact decomposition $\Gamma_k^{\text{data}} = \Gamma_k^{\text{task}} + \Gamma_k^{\text{rem}}$. The remainder is controlled over a bounded parameter class below.

Corollary 5 (Task-aligned radial component)**.** *For the quadratic model and every nonzero mode with $\rho_k > 0$, the full-grid task component is*

$$\Gamma_k^{\mathrm{task}} = \frac{4}{M^{3/2}} \frac{\mathrm{Re}\langle \widetilde{V}_k, q_k \rangle}{\rho_k^2}. \tag{15}$$

It also obeys the readout-norm bound

$$|\Gamma_k^{\mathrm{task}}| \le \frac{2}{M^{3/2}} \|\widetilde{V}_k\|_2. \tag{16}$$

Thus feature–readout alignment determines the sign of the task-aligned radial component. Together, this component, the residual gradient, and regularization determine net growth.

Proposition 6 (Exact gradient-flow radial dynamics)**.** *Under gradient flow*

$$\dot{W} = -\eta \left(\nabla_W \mathcal{L}_{\mathrm{data}} + \lambda W \right), \qquad \eta > 0, \quad \lambda \ge 0, \tag{17}$$

where the dot denotes the continuous-time derivative and η sets its time scale, every mode with $\rho_k(t) > 0$ satisfies

$$\dot{\rho}_k(t) = \eta \rho_k(t) \left(\Gamma_k^{\mathrm{data}}(t, \mathcal{S}) - \lambda \right) = \eta \rho_k(t) \left(\Gamma_k^{\mathrm{task}}(t, \mathcal{S}) + \Gamma_k^{\mathrm{rem}}(t, \mathcal{S}) - \lambda \right). \tag{18}$$

Consequently, mode k grows exactly when $\Gamma_k^{\mathrm{data}} > \lambda$ and shrinks exactly when $\Gamma_k^{\mathrm{data}} < \lambda$. The task-only sign rule $\Gamma_k^{\mathrm{task}} \gtrless \lambda$ is valid whenever $|\Gamma_k^{\mathrm{rem}}| < |\Gamma_k^{\mathrm{task}} - \lambda|$.

The residual can change the task-predicted growth direction only when it is at least as large as the gap between task-driven growth and shrinkage. The proof is in Appendix B.3.

Corollary 7 (Sustained mode decay)**.** *Suppose there exist $t_0 \ge 0$ and $\gamma > 0$ such that $\rho_k(t) > 0$ for $t \ge t_0$ and*

$$\Gamma_k^{\mathrm{data}}(t, \mathcal{S}) \le \lambda - \gamma \quad \textit{for every } t \ge t_0. \tag{19}$$

The mode radius therefore satisfies

$$\rho_k(t) \le \rho_k(t_0) \exp[-\eta\gamma(t - t_0)] \quad \textit{for every } t \ge t_0. \tag{20}$$

The same conclusion follows from the stronger, directly decomposed condition $\Gamma_k^{\mathrm{task}} + |\Gamma_k^{\mathrm{rem}}| \le \lambda - \gamma$ for every $t \ge t_0$.

On any interval where $\rho_k > 0$, integrating equation 18 gives

$$\log \frac{\rho_k(t_1)}{\rho_k(t_0)} = \eta \int_{t_0}^{t_1} \left(\Gamma_k^{\mathrm{data}}(t, \mathcal{S}) - \lambda \right) dt.$$

A persistent shortfall below shrinkage therefore produces exponential decay; the integral also handles rates that change sign. The quadratic identity gives

$$P_k(t) = 4\|q_k(t)\|_2^2 \le \rho_k(t)^4. \tag{21}$$

Hence Corollary 7 implies $P_k(t) \le \rho_k(t_0)^4 e^{-4\eta\gamma(t-t_0)}$. Imbalance between the left and right weight blocks provides another route to reduced interaction power: it can reduce q_k while preserving ρ_k.

B.3 Proofs of the Population Decomposition and Radial Balance

Task-aligned interaction. The model–label interaction in the squared loss is

$$\mathcal{L}_{\mathrm{label}} = -\frac{1}{|\mathcal{S}|} \sum_{(a,b) \in \mathcal{S}} \langle \widehat{\mathbf{y}}(a,b), \mathbf{y}(a,b) \rangle. \tag{22}$$

Lemma 8 (Population mode decomposition)**.** *Under the full-grid loss $\mathcal{S} = \mathbb{Z}_M^2$, the nonconstant part of the model–label interaction is*

$$\mathcal{L}_{\mathrm{label}}^{\neq 0} = -\frac{2}{M^{3/2}} \sum_{k=1}^{M-1} \mathrm{Re} \left\langle \widetilde{V}_k, q_k \right\rangle. \tag{23}$$

Hence its mode-k derivative depends only on $(\widetilde{V}_k, \widehat{W}_L[:,k], \widehat{W}_R[:,k])$, subject to the usual conjugate pairing for real parameters.

The label interaction rewards matching input frequencies through their readout projection. Each nonzero frequency contributes separately; the zero-frequency term contains the class mean.

Proof of Lemma 8. Write $h_j(a,b) = W_L[j,a]^2 + W_R[j,b]^2 + 2W_L[j,a]W_R[j,b]$. After summing over the full grid, the two one-input terms depend only on the class-mean coefficient $\widetilde{V}_0$: for example,

$$\sum_{a,b} V[(a+b) \bmod M, j] W_L[j,a]^2 = \left(\sum_c V[c,j]\right)\left(\sum_a W_L[j,a]^2\right).$$

They therefore contribute exclusively to the constant component of the task interaction.

For the cross term, Fourier inversion gives

$$\begin{aligned}&\sum_{a,b} V[(a+b) \bmod M, j] W_L[j,a] W_R[j,b] \\ &\qquad = \sqrt{M} \sum_{k=0}^{M-1} \overline{\widetilde{V}_k[j]}\, \widehat{W}_L[j,k] \widehat{W}_R[j,k].\end{aligned}$$

Multiplying by the factor $-2/M^2$ from equation 22, summing over neurons, and removing $k=0$ yields equation 23. Real vectors obey the usual conjugate symmetry $\widehat{u}(M-k) = \overline{\widehat{u}(k)}$. □

Proof of Corollary 5. Fix a non-self-conjugate pair of modes k and $M-k$ and scale both its left and right input-weight coefficients by a real scalar s, together with their conjugate partners. The pair contribution to the task interaction is $-4M^{-3/2}\,\mathrm{Re}\langle \widetilde{V}_k, q_k\rangle$, and q_k scales as s^2. Its derivative at $s=1$ is therefore

$$-\frac{8}{M^{3/2}}\,\mathrm{Re}\langle \widetilde{V}_k, q_k\rangle.$$

Unitarity and conjugate symmetry express the same directional derivative as $2\,\mathrm{Re}\langle w_k, g_k^{\mathrm{task}}\rangle$. Dividing by $-2\rho_k^2$ gives equation 15. If M is even, the self-conjugate mode $k = M/2$ contributes once to the Fourier sum and once to the directional derivative; both factors of two disappear and the same formula follows. Finally, $|\,\mathrm{Re}\langle \widetilde{V}_k, q_k\rangle| \le \|\widetilde{V}_k\|_2 \|q_k\|_2$ and

$$\|q_k\|_2 \le \|\widehat{W}_L[:,k]\|_2 \|\widehat{W}_R[:,k]\|_2 \le \frac{\rho_k^2}{2}.$$

Substitution into equation 15 proves equation 16. □

Proof of Proposition 6. Because the DFT is unitary, applying it to the real-space gradient flow equation 17 yields

$$\dot{w}_k = -\eta\left(g_k^{\mathrm{data}} + \lambda w_k\right).$$

For $\rho_k = \|w_k\|_2 > 0$,

$$\begin{aligned}\dot{\rho}_k &= \frac{\mathrm{Re}\langle w_k, \dot{w}_k\rangle}{\rho_k} \\ &= -\eta \frac{\mathrm{Re}\langle w_k, g_k^{\mathrm{data}}\rangle}{\rho_k} - \eta\lambda\rho_k \\ &= \eta\rho_k\left(\Gamma_k^{\mathrm{data}} - \lambda\right).\end{aligned}$$

Substituting $g_k^{\mathrm{data}} = g_k^{\mathrm{task}} + g_k^{\mathrm{rem}}$ gives the second equality in equation 18. If the remainder magnitude is smaller than the task-only margin, that margin retains its sign. □

Proof of Corollary 7. The identities in equation 18 and equation 19 imply $\dot{\rho}_k \le -\eta\gamma\rho_k$ for $t \ge t_0$. Gronwall's inequality gives equation 20. Finally, $\Gamma_k^{\mathrm{task}} + |\Gamma_k^{\mathrm{rem}}| \le \lambda - \gamma$ implies $\Gamma_k^{\mathrm{data}} \le \lambda - \gamma$. □

B.4 Finite Samples and Discrete Updates

Finite-sample bounds control gradient error across modes and a compact parameter set. A growth or decay prediction remains valid when its margin exceeds the residual bound (Proposition 10). We first extend the radial and overlap identities to finite optimizer steps.

For any optimizer, write its realized mode update as

$$w_{k,t+1} = w_{k,t} - \alpha_t u_{k,t}, \qquad \alpha_t > 0.$$

The direction $u_{k,t}$ includes all transformations and regularization or decay terms applied by the optimizer. For $\rho_{k,t} > 0$, let

$$\Gamma^{\mathrm{upd}}_{k,t} = -\frac{\mathrm{Re}\langle w_{k,t}, u_{k,t}\rangle}{\rho^2_{k,t}}, \qquad \Xi_{k,t} = \frac{\|u_{k,t}\|_2^2}{\rho^2_{k,t}}.$$

Expanding the squared norm gives

$$\frac{\rho^2_{k,t+1}}{\rho^2_{k,t}} = 1 + 2\alpha_t \Gamma^{\mathrm{upd}}_{k,t} + \alpha_t^2 \Xi_{k,t}, \tag{24}$$

The linear term measures the update's radial projection; the quadratic term accounts for its squared length. Thus mode k contracts exactly when $\Gamma^{\mathrm{upd}}_{k,t} + \alpha_t \Xi_{k,t}/2 < 0$. If the realized direction admits $u_{k,t} = u^{\mathrm{task}}_{k,t} + u^{\mathrm{rem}}_{k,t} + d_t w_{k,t}$, where $d_t \geq 0$ is the scalar shrinkage coefficient, define

$$\Gamma^{v,\mathrm{upd}}_{k,t} = -\frac{\mathrm{Re}\langle w_{k,t}, u^v_{k,t}\rangle}{\rho^2_{k,t}}.$$

Here v denotes the task or remainder component. Then $\Gamma^{\mathrm{upd}}_{k,t} = \Gamma^{\mathrm{task,upd}}_{k,t} + \Gamma^{\mathrm{rem,upd}}_{k,t} - d_t$. Consequently, the residual bound $|\Gamma^{\mathrm{rem,upd}}_{k,t}| \leq B^{\mathrm{upd}}_{k,t}$ gives the sufficient contraction condition

$$\Gamma^{\mathrm{task,upd}}_{k,t} + B^{\mathrm{upd}}_{k,t} - d_t + \frac{\alpha_t}{2}\Xi_{k,t} \leq -\gamma_{k,t} < 0. \tag{25}$$

Under equation 25, if $2\alpha_t\gamma_{k,t} < 1$, then $\rho_{k,t+1} \leq e^{-\alpha_t\gamma_{k,t}}\rho_{k,t}$.

The identity in equation 24 applies directly to realized Adam and Muon updates, including momentum, preconditioning, orthogonalization, regularization, and decoupled decay. Cross-block overlap admits an exact realized law of the same kind. For positive initial per-unit powers, write $a'_j = a_j(1+\delta^L_j)$ and $b'_j = b_j(1+\delta^R_j)$, suppressing the mode index. Then

$$\frac{\sum_j a'_j b'_j}{\sum_j a_j b_j} = \langle (1+\delta^L)(1+\delta^R)\rangle_P, \qquad \frac{S^{L\prime}}{S^L} = 1 + \langle \delta^L\rangle_a,$$

and similarly for the right block. Substitution into $A = K\sum_j a_j b_j/(S^L S^R)$ proves equation 7 whenever the new block totals are positive. For a gradient-flow step of duration Δt, $\delta^{L,R} = 2\eta(\gamma^{L,R} - \lambda)\Delta t + o(\Delta t)$; the first-order log-ratio is precisely equation 6. For adaptive optimizers the realized changes include the gradient transformation.

Effect of finite-sample gradients on cross-block overlap. Write $g^{\mathrm{samp}}_k = g^{\mathrm{data}}_{k,\mathcal{S}} - g^{\mathrm{data}}_{k,\Omega}$ for the empirical–population gradient difference, where $\Omega = \mathbb{Z}^2_M$. Its negative-gradient contribution is $d^{\mathrm{samp}}_k = -g^{\mathrm{samp}}_k$. For any negative-gradient component $d^v_k = [d^{L,v}_k; d^{R,v}_k]$, define $\gamma^{L,v}_{jk} = \mathrm{Re}(\overline{\widehat{W}_L[j,k]} d^{L,v}_{jk})/a_{jk}$ and similarly on the right. Its contribution to $d\log A_k/dt$ is

$$C^v_k = 2\eta\left(\frac{\mathrm{Cov}_a(b, \gamma^{L,v})}{\langle b\rangle_a} + \frac{\mathrm{Cov}_b(a, \gamma^{R,v})}{\langle a\rangle_b}\right).$$

Let $B_{n,\zeta} = B_{n,\zeta}(\varepsilon_{\mathrm{net}})$ be the uniform gradient bound in Proposition 10. Define the weighted coefficients of variation for the two blocks,

$$\kappa^L_k = \frac{\mathrm{std}_a(b_{\cdot k})}{\langle b\rangle_a}, \qquad \kappa^R_k = \frac{\mathrm{std}_b(a_{\cdot k})}{\langle a\rangle_b}, \qquad \kappa_k = \max\left(\kappa^L_k, \kappa^R_k\right).$$

Here std_a and std_b denote standard deviations under the same normalized weights as $\langle\cdot\rangle_a$ and $\langle\cdot\rangle_b$.

Proposition 9 (Sampling contribution to overlap dynamics)**.** *Assume positive per-unit block powers, as in Proposition 2. At fixed parameters and for a uniformly drawn training set,*

$$\mathbb{E}_{\mathcal{S}}\big[C_k^{\mathrm{samp}}\big] = 0, \tag{26}$$

and for any negative-gradient component v,

$$\big|C_k^v\big| \le 2\eta \left(\kappa_k^L \frac{\|d_k^{L,v}\|_2}{\sqrt{S_k^L}} + \kappa_k^R \frac{\|d_k^{R,v}\|_2}{\sqrt{S_k^R}} \right) \le \frac{4\sqrt{2}\,\eta\,\kappa_k\,\|d_k^v\|_2}{\rho_k \beta_k}. \tag{27}$$

On the uniform-bound event of Proposition 10, if the trajectory remains in Θ with positive per-unit powers over $[t_0, T]$, then

$$\left| \int_{t_0}^{T} C_k^{\mathrm{samp}}(t)\,dt \right| \le 4\sqrt{2}\,\eta\,B_{n,\zeta} \int_{t_0}^{T} \frac{\kappa_k(t)}{\rho_k(t)\beta_k(t)}\,dt.$$

Equation 27 separates gradient size from geometric sensitivity. The weighted variation κ_k measures how unevenly the opposite block distributes power across neurons; $\rho_k\beta_k$ sets the scale in the denominator. Using the measured $\|g_k^{\mathrm{samp}}\|_2$ evaluates sampling effects along the trajectory. Substituting $B_{n,\zeta}$ gives a bound uniform over modes and parameters, including the dimension $D = 4K$ and parameter-set covering number.

The bound controls the instantaneous overlap rate. At $\kappa_k = 0$ the first-order overlap rate vanishes, but the gradient can create dispersion that changes A_k later. For example, both coefficient vectors following $(1 + t, 1 - t)$ give $A(t) = 1 + 4t^2/(1 + t^2)^2$, with $A'(0) = 0$ and $A(t) > 1$ for small nonzero t. Comparisons across trained modes combine $1/(\rho_k\beta_k)$ with their gradient scale and dispersion, which evolve with the same parameters. The sample-size bound controls a fixed compact parameter class; its constants and the actual trajectory depend on width and optimizer.

At a parameter chosen independently of $\mathcal{S}$, the sampling contribution has zero mean. During training, the parameters $\theta_t = (W_t, V_t) = \theta_t(\mathcal{S})$ depend on the sample, allowing both accumulated sampling deviations and adaptive feedback. The inequality bounds the change in overlap that the sampling contribution can produce along the realized trajectory.

B.5 Finite-Sample Gradient Bound

Let $\theta = (W, V)$, $\Omega = \mathbb{Z}_M^2$, and $N = M^2$. For $x \in \Omega$, define

$$\ell_x(\theta) = \frac{1}{2}\|\widehat{\mathbf{y}}(x;\theta) - \mathbf{y}(x)\|_2^2, \qquad \psi_k(x;\theta) = \mathcal{F}_k[\nabla_W \ell_x(\theta)] \in \mathbb{C}^{2K},$$

where $\mathcal{F}_k$ restricts the unitary Fourier transform to the two mode-k input blocks. Set

$$g_{k,\mathcal{S}}^{\mathrm{data}}(\theta) = \frac{1}{n}\sum_{x\in\mathcal{S}} \psi_k(x;\theta), \qquad g_{k,\Omega}^{\mathrm{data}}(\theta) = \frac{1}{N}\sum_{x\in\Omega} \psi_k(x;\theta),$$

and $g_{k,A}^{\mathrm{rem}} = g_{k,A}^{\mathrm{data}} - g_k^{\mathrm{task}}$ for either the training set $A = \mathcal{S}$ or the full grid $A = \Omega$.

Proposition 10 (Uniform finite-sample gradient bound)**.** *Let Θ be a compact parameter set, fix a failure probability $0 < \zeta < 1$, and let $\mathcal{N}_\Theta(\varepsilon_{\mathrm{net}})$ be the size of an $\varepsilon_{\mathrm{net}}$-net. Split $\psi_k \in \mathbb{C}^{2K}$ into $D = 4K$ real coordinates. Suppose each coordinate has range at most R over inputs and is L-Lipschitz in $\theta \in \Theta$ under the Euclidean parameter norm. If $\mathcal{S}$ is a uniform size-n subset of the $N = M^2$ inputs, then with probability at least $1 - \zeta$, simultaneously for all nonzero k and $\theta \in \Theta$,*

$$\big\|g_{k,\mathcal{S}}^{\mathrm{data}}(\theta) - g_{k,\Omega}^{\mathrm{data}}(\theta)\big\|_2 \le B_{n,\zeta}(\varepsilon_{\mathrm{net}}), \tag{28}$$

$$B_{n,\zeta}(\varepsilon_{\mathrm{net}}) := \sqrt{D}\left[R\sqrt{\frac{N-n+1}{2nN}\log\frac{2D(M-1)\mathcal{N}_\Theta(\varepsilon_{\mathrm{net}})}{\zeta}} + 2L\varepsilon_{\mathrm{net}} \right].$$

Consequently, for $\rho_k(\theta) > 0$,

$$|\Gamma_k^{\mathrm{rem}}(\theta, \mathcal{S})| \le \mathcal{B}_{k,n,\zeta}(\theta;\varepsilon_{\mathrm{net}}) := \frac{\|g_{k,\Omega}^{\mathrm{rem}}(\theta)\|_2 + B_{n,\zeta}(\varepsilon_{\mathrm{net}})}{\rho_k(\theta)}. \tag{29}$$

The square-root term controls sampling error; the $2L\varepsilon_{\text{net}}$ term extends the bound from a finite parameter net to all of Θ. The radial remainder combines this error with the full-grid residual, divided by mode radius.

Corollary 11 (Sign preservation)**.** *On the event above, for any $\gamma > 0$,*

$$\Gamma_k^{\text{task}} \le \lambda - \mathcal{B}_{k,n,\zeta} - \gamma \quad \Longrightarrow \quad \Gamma_k^{\text{data}} \le \lambda - \gamma, \tag{30}$$

$$\Gamma_k^{\text{task}} \ge \lambda + \mathcal{B}_{k,n,\zeta} + \gamma \quad \Longrightarrow \quad \Gamma_k^{\text{data}} \ge \lambda + \gamma. \tag{31}$$

For a trajectory in Θ with $\rho_k > 0$, if the first condition holds throughout $[t_0, t]$, then $\rho_k(t) \le \rho_k(t_0)e^{-\eta\gamma(t-t_0)}$ and $P_k(t) \le \rho_k(t_0)^4 e^{-4\eta\gamma(t-t_0)}$. If the second condition holds throughout that interval, then $\rho_k(t) \ge \rho_k(t_0)e^{\eta\gamma(t-t_0)}$; power additionally depends on β_k and A_k through Proposition 1.

The residual bound sets a margin around shrinkage. Outside that margin, the task-driven term determines growth or decay; sustaining the margin gives the exponential bounds above.

Proof of Proposition 10. Fix a mode, a real coordinate of ψ_k, and a point in an ε_{net}-net of Θ. Serfling's inequality for sampling without replacement gives

$$\Pr\left(\left|g_{k,\mathcal{S},j}^{\text{data}} - g_{k,\Omega,j}^{\text{data}}\right| > t\right) \le 2\exp\left(-\frac{2nt^2}{R^2(N-n+1)/N}\right).$$

A union bound over D real coordinates, $M-1$ nonzero modes, and $\mathcal{N}_\Theta(\varepsilon_{\text{net}})$ net points, followed by the stated choice of t, controls every net point with probability at least $1-\zeta$.

For arbitrary $\theta \in \Theta$, choose a net point θ_0 within ε_{net}. Both the empirical and population coordinate means are L-Lipschitz, so their difference changes by at most $2L\varepsilon_{\text{net}}$ between θ and θ_0. Converting the simultaneous coordinate bound to an ℓ_2 bound costs the factor $\sqrt{D}$ and proves equation 28. The radial remainder satisfies

$$\begin{aligned} \|g_{k,\mathcal{S}}^{\text{rem}}\|_2 &\le \|g_{k,\Omega}^{\text{rem}}\|_2 + \|g_{k,\mathcal{S}}^{\text{data}} - g_{k,\Omega}^{\text{data}}\|_2, \\ |\Gamma_k^{\text{rem}}(\theta,\mathcal{S})| &\le \|g_{k,\mathcal{S}}^{\text{rem}}\|_2/\rho_k, \end{aligned}$$

where the second line is Cauchy–Schwarz. This proves equation 29. □

Proof of Corollary 11. The decomposition $\Gamma_k^{\text{data}}(\theta,\mathcal{S}) = \Gamma_k^{\text{task}}(\theta) + \Gamma_k^{\text{rem}}(\theta,\mathcal{S})$ and equation 29 preserve the task-margin sign under either condition in equations 30 and 31. Substitution into Proposition 6 and integration give the upper and lower bounds for ρ_k. Under the contraction condition, equation 21 gives the upper bound on interaction power. □

Affine realized updates and frequency mixing. Let g_t^{all} concatenate all Fourier gradient modes, including the zero-frequency (DC) mode. Conditional on a fixed Adam denominator, the update has the form

$$u_{k,t} = \mathcal{T}_{k,t} g_t^{\text{all}} + h_{k,t},$$

where $\mathcal{T}_{k,t}$ is a block row of the transformed linear operator. Spatially varying coordinatewise preconditioning generally mixes frequencies, so the block row acts on the complete gradient. The term $h_{k,t}$ contains fixed history and the transformed coupled-regularization contribution. Here regularization is included in $h_{k,t}$, so the separate scalar d_t in Appendix B.4 is zero. Taking R, L uniformly over all M modes extends Proposition 10 to include DC (replace $M-1$ by M in its union bound), giving a per-mode error $B_{n,\zeta}^{\text{all}}$. For $u_{k,t}^{\text{task}} = \mathcal{T}_{k,t} g_t^{\text{task,all}}$, Cauchy–Schwarz gives $|\Gamma_{k,t}^{\text{rem,upd}}| \le B_{k,t}^{\text{upd}}$ with

$$B_{k,t}^{\text{upd}} := \frac{\|\mathcal{T}_{k,t}\|_{\text{op}} \left(\|g_{\Omega,t}^{\text{rem,all}}\|_2 + \sqrt{M}\, B_{n,\zeta}^{\text{all}}\right) + \|h_{k,t}\|_2}{\rho_{k,t}}.$$

This uniform bound includes frequency mixing. The realized-update analysis in Appendix E.3 evaluates Adam and Muon directly from the complete gradient and optimizer state.

B.6 Factorization and Cross-Block Overlap Dynamics

Proof of Proposition 1. The quadratic interaction gives $P_k = 4\sum_j a_{jk}b_{jk}$, while $\rho_k^2 = S_k^L + S_k^R$. Substitution yields the stated factorization. The balance bound follows from $2\sqrt{S_k^L S_k^R} \le S_k^L + S_k^R$, and the overlap bound follows from $0 \le \sum_j p_{jk}^L p_{jk}^R \le 1$ for the normalized allocations. □

Continuous-time factor dynamics. Under the positivity conditions of Proposition 2, the block-power and product-weighted averages in Sections 4.1–4.2 give three companion rate identities:

$$\frac{d\log \rho_k}{dt} = \eta\big(\Gamma_k^{\mathrm{data}} - \lambda\big), \qquad \Gamma_k^{\mathrm{data}} = \frac{\sum_j a_{jk}\gamma_{jk}^L + b_{jk}\gamma_{jk}^R}{\rho_k^2}, \tag{32}$$

$$\frac{d\log P_k}{dt} = 2\eta\big(\langle \gamma^L + \gamma^R\rangle_P - 2\lambda\big), \tag{33}$$

$$\frac{d\log \beta_k}{dt} = \eta\big(\langle\gamma^L\rangle_a + \langle\gamma^R\rangle_b - 2\Gamma_k^{\mathrm{data}}\big), \tag{34}$$

Together with equation 6, these identities give the exact continuous-time decomposition of mode-power dynamics. The first is the radial balance of Proposition 6.

Proof of Proposition 2. Under gradient flow $\dot{\widehat{W}}_L[j,k] = \eta(d_{jk}^L - \lambda\widehat{W}_L[j,k])$, so $\dot a_{jk} = 2\eta a_{jk}(\gamma_{jk}^L - \lambda)$ and likewise for b_{jk}. Summing gives $\frac{d}{dt}\rho_k^2 = 2\eta\sum_j[a_{jk}(\gamma_{jk}^L - \lambda) + b_{jk}(\gamma_{jk}^R - \lambda)]$, which is equation 32 after dividing by $2\rho_k^2$.

For the power, $|q_k[j]|^2 = a_{jk}b_{jk}$, so $\frac{d}{dt}|q_k[j]|^2 = b_{jk}\dot a_{jk} + a_{jk}\dot b_{jk} = 2\eta a_{jk}b_{jk}(\gamma_{jk}^L + \gamma_{jk}^R - 2\lambda)$. Each block's per-unit growth is weighted by the other block's power on the same unit. With $P_k = 4\sum_j a_{jk}b_{jk}$, dividing by P_k gives equation 33.

For the balance, $\log\beta_k = \log 2 + \frac{1}{2}(\log S_k^L + \log S_k^R) - \log\rho_k^2$ and $\dot S_k^L/S_k^L = 2\eta(\langle\gamma^L\rangle_a - \lambda)$, which yields equation 34; the shrinkage terms cancel.

The overlap law then follows from the factorization identity $\log A_k = \log P_k - 4\log\rho_k - 2\log\beta_k + \log K$. Substituting the three rates leaves $d\log A_k/dt = 2\eta[(\langle\gamma^L\rangle_P - \langle\gamma^L\rangle_a) + (\langle\gamma^R\rangle_P - \langle\gamma^R\rangle_b)]$, and under the $a_{\cdot k}$-weighted measure $\langle\gamma^L\rangle_P = \langle b\gamma^L\rangle_a/\langle b\rangle_a$, so the bracket is $\mathrm{Cov}_a(b_{\cdot k}, \gamma_{\cdot k}^L)/\langle b\rangle_a$ and symmetrically for the right block. All shrinkage terms have cancelled, which proves cancellation of the direct isotropic shrinkage term. The rate identity is immediate from the state identity. □

Sampling perturbations.

Proof of Proposition 9. For a uniformly drawn training set the empirical gradient is unbiased for the population gradient at fixed parameters, so $\mathbb{E}_{\mathcal S}[d_k^{\mathrm{samp}}] = 0$ and hence $\mathbb{E}_{\mathcal S}[\gamma_{jk}^{L,\mathrm{samp}}] = 0$ for every j. The weights a_{jk} and b_{jk} are functions of the parameters alone, and C_k^v is linear in the negative gradient, so expectation passes through the covariance and equation 26 follows.

For equation 27, Cauchy–Schwarz under the $a_{\cdot k}$-weighted measure gives $|\mathrm{Cov}_a(b_{\cdot k}, \gamma_{\cdot k}^{L,v})| \le \mathrm{std}_a(b_{\cdot k})\,\mathrm{std}_a(\gamma_{\cdot k}^{L,v})$, so the first term of C_k^v is at most $2\eta\,\kappa_k^L\,\mathrm{std}_a(\gamma_{\cdot k}^{L,v})$. Since $|\gamma_{jk}^{L,v}| = |\mathrm{Re}(\overline{\widehat W_L[j,k]}\,d_{jk}^{L,v})|/a_{jk} \le |d_{jk}^{L,v}|/\sqrt{a_{jk}}$,

$$\mathrm{std}_a\big(\gamma_{\cdot k}^{L,v}\big)^2 \le \big\langle(\gamma^{L,v})^2\big\rangle_a = \frac{1}{S_k^L}\sum_j a_{jk}\big(\gamma_{jk}^{L,v}\big)^2 \le \frac{\|d_{\cdot k}^{L,v}\|_2^2}{S_k^L},$$

and symmetrically for the right block, which is the first inequality.

For the second, $S_k^L S_k^R = (\beta_k\rho_k^2/2)^2$ and $S_k^L + S_k^R = \rho_k^2$ give $\min(S_k^L, S_k^R) \ge \rho_k^2\beta_k^2/4$, so both denominators are at least $\rho_k\beta_k/2$. Bounding κ_k^L, κ_k^R by κ_k and using $\|d_k^{L,v}\|_2 + \|d_k^{R,v}\|_2 \le \sqrt{2}\,\|d_k^v\|_2$ completes the proof. □

Dependence on width at initialization. Independent Gaussian blocks provide a reference for overlap before learning. For every nonzero, non-self-conjugate mode, the orthonormal DFT makes a_{jk} and b_{jk} independent exponentials of common mean μ. This covers all nonconstant modes at odd moduli. At even moduli, the self-conjugate Nyquist mode $k = M/2$ has real Gaussian coefficients and requires a separate calculation.

Proposition 12 (Overlap at initialization)**.** *For iid $\mathcal{N}(0, \sigma^2)$ initialization with $\sigma > 0$, fixed M, and every nonzero mode that is not self-conjugate, so $k \neq 0$ and $2k \not\equiv 0 \pmod M$, $A_k \to 1$ in probability as $K \to \infty$, with*

$$\sqrt{K}\,\log A_k \xrightarrow{d} \mathcal{N}(0, 1), \tag{35}$$

so $A_k = 1 + O_p(K^{-1/2})$.

Thus independent blocks start near unit overlap, with fluctuations that shrink as width grows. This supplies a reference scale for measuring later coordination.

Proof of Proposition 12. For iid $\mathcal{N}(0, \sigma^2)$ weights and a non-self-conjugate nonzero mode, the orthonormal DFT gives $\widehat{W}_L[j, k]$ circularly symmetric complex Gaussian with $\mathbb{E}|\widehat{W}_L[j, k]|^2 = \sigma^2$, so a_{jk} and b_{jk} are independent exponentials with common mean $\mu = \sigma^2$, independent across j and across blocks. Write $\langle\cdot\rangle$ for the average over the K units. Then $\mathbb{E}\langle ab\rangle = \mu^2$ and $\mathbb{E}\langle a\rangle = \mathbb{E}\langle b\rangle = \mu$, so $A_k = \langle ab\rangle/(\langle a\rangle\langle b\rangle) \to 1$.

$\sqrt{K}$ times the centered vector of the three sample means is asymptotically normal by the multivariate central limit theorem. For its covariance, $\mathrm{Var}(a) = \mu^2$ and, using $\mathbb{E}[a^2] = 2\mu^2$, $\mathrm{Var}(ab) = 4\mu^4 - \mu^4 = 3\mu^4$ and $\mathrm{Cov}(ab, a) = \mathbb{E}[a^2]\mu - \mu^3 = \mu^3$. Dividing by K for the sample means and applying the delta method to $\log A_k = \log\langle ab\rangle - \log\langle a\rangle - \log\langle b\rangle$,

$$\mathrm{Var}_{\mathrm{asymp}}(\sqrt{K}\log A_k) = \left(\frac{3\mu^4}{\mu^4} + \frac{\mu^2}{\mu^2} + \frac{\mu^2}{\mu^2} - \frac{2\mu^3}{\mu^3} - \frac{2\mu^3}{\mu^3}\right) = 1,$$

the two blocks being independent so their means do not covary. The delta method therefore proves the distributional limit in equation 35. □

Simulation gives $\mathrm{Var}(\log A_k) \cdot K$ between 0.94 and 1.05 for tested widths from $K = 32$ to 512, with mean A_k within 0.8% of one. The Gaussian reference gives a $K^{-1/2}$ scale for initial overlap fluctuations. The training experiments use uniform `nn.Linear` initialization and measure how the initially broad spectrum evolves at each width. Together these analyses separate initialization geometry from the subsequent amplification process.

B.7 Population-Loss Bound and Reconstruction

The bound in Corollary 3 relates interaction power and readout norm to the amplitude required for target reconstruction.

Uniform readout-norm relaxation. If in addition $\|\widetilde{V}_k\|_2 \le B$ for every nonzero mode, then

$$L_{\mathrm{pop}} \;\ge\; \frac{1}{M^2}\sum_{k\neq 0}\big(\sqrt{M} - B\sqrt{P_k}\big)_+^2.$$

Population-risk conventions. The uncentered risk per output coordinate is

$$L_{\mathrm{pop}}^{\mathrm{mean}} = \frac{1}{M^3}\sum_{a,b,c}\big(\widehat{y}_c(a, b) - y_c(a, b)\big)^2.$$

Its lower bound has the same form with an M^{-3} prefactor. The centered and uncentered risks satisfy $L_{\mathrm{pop}} = M\,L_{\mathrm{pop}}^{\mathrm{mean}} - M\,\mathrm{mean}_{a,b}\,\bar{e}(a, b)^2$ with $\bar{e}(a, b) = \frac{1}{M}\sum_c\big(\widehat{y}_c - y_c\big)$ the per-example class-mean error, so their nonconstant-mode lower bounds differ by the factor M. The centered risk corresponds to the training objective.

Proof of Corollary 3. Take the orthonormal DFT along the class axis and then the unitary 2D DFT along (a, b). The target $y_c(a, b)$ has class-mode-k component $M^{-1/2}\omega_M^{-k(a+b)}$, whose only nonzero spatial coefficient is $\sqrt{M}$ at $(u, v) = (-k, -k)$. In $h_j = (W_L[j, a] + W_R[j, b])^2$ the squared terms depend on one coordinate each, so their spatial transforms are supported on $u = 0$ and $v = 0$; for $k \neq 0$ the diagonal frequency $(-k, -k)$ has both coordinates nonzero, and only the cross term $2W_L[j, a]W_R[j, b]$ contributes there, with coefficient $2\widehat{W}_L[j, -k]\widehat{W}_R[j, -k] = 2q_{-k}[j]$. Hence the network's coefficient is $2\sum_j \widetilde{V}_k[j]q_{-k}[j]$.

Cauchy–Schwarz gives $\left|2\sum_j \widetilde{V}_k[j]q_{-k}[j]\right| \leq 2\|\widetilde{V}_k\|_2\|q_{-k}\|_2 = \|\widetilde{V}_k\|_2\sqrt{P_{-k}}$, and $P_{-k} = P_k$ for real weights. Because equation 11 is centered, its class-mode-zero term vanishes, and by Parseval along the class axis and then along (a, b) it equals M^{-2} times the sum of squared coefficient differences over all nonconstant class modes and all spatial frequencies; retaining only the diagonal frequency $(-k, -k)$ for each $k \neq 0$ discards non-negative terms and leaves $M^{-2}\sum_{k\neq 0}(\sqrt{M} - \|\widetilde{V}_k\|_2\sqrt{P_k})_+^2$. The uncentered mean risk has the same diagonal bound with M^{-3} in place of M^{-2}, since its nonconstant-mode contribution is L_{pop}/M. For positive block energies, $\sqrt{P_k} = \rho_k^2\beta_k\sqrt{A_k/K}$ is equation 4. If either block energy vanishes, $q_k = 0$ and $P_k = 0$ directly, and the same loss inequality holds. □

Reconstruction and classification. The bound relates insufficient interaction power and readout norm to the trained squared-loss objective. Its target is one-hot reconstruction. Classification has a different spectral requirement: eight quadratic neurons can represent $\widehat{y}_c(a, b) = \cos(2\pi(a+b-c)/M)$ through $(u + v)^2 - (u - v)^2 = 4uv$. This predictor has the unique correct argmax on every input, finite readout norms, and only the pair $k = 1, M - 1$ in its diagonal feature spectrum, giving effective rank $N_{\text{eff}} = 2$ (Appendix D.5). The construction can be padded to width 512. This example makes the objective distinction explicit: the population-loss bound quantifies squared reconstruction error, while the grokking definition measures classification accuracy.

The geometry depends on the representation being measured. Neural-collapse analyses under mean squared error (MSE) study within-class concentration and class-mean/classifier geometry (Han et al., 2021); modular-arithmetic studies identify task-dependent cyclic alternatives to simplex geometry (Tan et al., 2026). Our effective-rank statistic describes the distribution of power across task-aligned frequencies.

C Fourier Measurements and Algorithms

The Fourier measurements below are computed at each recorded checkpoint. Aggregate overlap uses the full, fixed mode set in equation 9.

C.1 Relation between Interaction Power and Weight Norm

Let $h_j^{\times}(a, b) = 2W_L[j, a]W_R[j, b]$ denote the cross term in the quadratic hidden feature. Its DFT factorizes as

$$\widehat{h}_j^{\times}(k, \ell) = 2\widehat{W}_L[j, k]\widehat{W}_R[j, \ell].$$

On the diagonal, this identity yields $P_k = 4\|q_k\|_2^2$ and the upper bound equation 21 in terms of the weight norm in equation 3.

C.2 Computation and Update Decomposition

The stabilized readout projection onto the interaction direction and the squared readout norm are

$$D_k(t) = \frac{\text{Re}\langle\widetilde{V}_k(t), q_k(t)\rangle}{\|q_k(t)\|_2 + \epsilon}, \qquad Q_k(t) = \|\widetilde{V}_k(t)\|_2^2,$$

where $\epsilon = 10^{-8}$ keeps the denominator positive. Algorithm 1 uses $\odot$ for elementwise products.

Algorithm 1: Computation of Fourier mode statistics

Input: Weights $W_L(t), W_R(t), V(t)$ at each recorded time $t \in \mathcal{T}$.
Output: Continuous Fourier measurements, feature–readout statistics, and the factors ρ_k, β_k, A_k that determine P_k through equation 4.

for $t \in \mathcal{T}$ **do**
  Compute row-wise DFTs of $W_L(t), W_R(t)$ and the class-axis DFT of $V(t)$
  for $k = 1, \dots, M-1$ **do**
    $q_k(t) \leftarrow \widehat{W}_L[:,k](t) \odot \widehat{W}_R[:,k](t)$ and $P_k(t) \leftarrow 4\|q_k(t)\|_2^2$
    $a_k \leftarrow |\widehat{W}_L[:,k]|^2$, $b_k \leftarrow |\widehat{W}_R[:,k]|^2$, $S_k^L \leftarrow \mathbf{1}^\top a_k$, $S_k^R \leftarrow \mathbf{1}^\top b_k$
    $\rho_k^2 \leftarrow S_k^L + S_k^R$, $\beta_k \leftarrow 2\sqrt{S_k^L S_k^R}/\rho_k^2$, $A_k \leftarrow K\, a_k^\top b_k/(S_k^L S_k^R)$
    Compute $Q_k(t)$ and $D_k(t)$ from $q_k(t)$ and $\widetilde{V}_k(t)$
  Normalize the mode powers $P_k(t)$ to obtain $N_{\text{eff}}(t)$, and evaluate $m_{\text{diag},90}(t)$
Compute G_A on the fixed set of conjugate-pair representatives and locate event intervals using Appendix D.3

Algorithm 2: Mode-wise decomposition of discrete optimizer updates

Input: One recorded optimizer step $(W_t, g_t, \text{optimizer state}_t, W_{t+1})$ and step size $\alpha_t > 0$, where g_t is the input-weight gradient.
Output: Changes in mode norm and overlap, logarithmic contributions to power changes, and an optimizer-update reconstruction residual.

Suppress the step index t on α_t, $\rho_{k,t}$, $\Gamma_{k,t}^{\text{upd}}$, and $\Xi_{k,t}$ below
$\Delta W \leftarrow W_{t+1} - W_t$
Transform W_t, ΔW and g_t into the orthonormal Fourier basis
for $k = 1, \dots, M-1$ **do**
  $u_{k,t} \leftarrow -\Delta w_k/\alpha_t$; compute $\Gamma_{k,t}^{\text{upd}}$ and $\Xi_{k,t}$ as in Appendix B.4
  $\Delta\rho_k^2 \leftarrow 2\alpha\Gamma_k^{\text{upd}}\rho_k^2 + \alpha^2\Xi_k\rho_k^2$ by equation 24, and compare with $\rho_k^2(t{+}1) - \rho_k^2(t)$
  $\delta_{\cdot k}^L, \delta_{\cdot k}^R \leftarrow$ realized per-unit relative power changes; evaluate $A_k(t{+}1)/A_k(t)$ by equation 7 and compare with the direct ratio
  $\mathcal{R}_k, \mathcal{B}_k, \mathcal{C}_k \leftarrow$ logarithmic changes of ρ_k^4, β_k^2, A_k; check $\mathcal{R}_k + \mathcal{B}_k + \mathcal{C}_k = \log\big(P_k(t{+}1)/P_k(t)\big)$
Independently reconstruct the optimizer update from g_t and state_t to obtain ΔW_{rec}; report $\|\Delta W_{\text{rec}} - \Delta W\|_F/\|\Delta W\|_F$

Normalized allocations require positive block energies; relative per-unit changes require positive initial powers, and logarithmic changes require positive factors at both endpoints. Algorithms 1–2 evaluate these quantities on their stated domains. Zero interaction power is retained as zero; undefined normalized quantities are left undefined.

Algorithm 2 computes the mode-wise power decomposition from individual optimizer updates. Computing the Fourier statistics costs $O(KM \log M)$ per update; update reconstruction adds the optimizer's cost, including Muon orthogonalization. Appendices E.1–E.3 describe the recorded steps and reconstruction samples used for these checks.

The per-mode calculations in Algorithm 1 operate directly on weight DFTs in $O(KM \log M)$ time and $O(KM)$ memory per checkpoint. We evaluate task-aligned concentration separately over all input pairs.

D Cross-Block Overlap and Representation Results

Table 2 maps the core overlap experiments, complementary representation studies, and structural checks to their evidence. Each row specifies the setting and measurement supporting its conclusion.

Table 2: Experimental scope and measurements. All studies use quadratic networks. The 55-run covariance study draws on a subset of the 65-run core study.

Study and setting	Measurement and evidence
Core overlap: 65 Adam/Muon runs, $M = 127$, $K = 512$	Fixed-marginal and phase interventions, overlap trajectories, and endpoint associations; concentration/alignment decomposition (Section 5; Appendix D.2).
Mode growth: 55 of the core runs	Additive covariance contributions to cross-mode log-power growth (Appendices E.1, D.4).
Transition study: 250 Adam/Muon runs, $M = 127$, $K = 512$	Concentration–outcome associations and event timing; optimizer-specific transition regions (Appendices D.6, F.3).
Width and modulus comparisons	Spectrum and power at $K = 512, 1024$; endpoint concentration across five moduli (Appendices D.5, I.2).
Regularization: 15 pairs, $M = 71$, $K = 2048$	Changes in concentration and generalization under a matched regularization change (Appendix Figure 8).
Modular subtraction: $M = 23, 41$	Numerical factorization checks for the subtraction frequency pairing (Appendix G).
Population-loss bound: 16 runs, 48 checkpoints	Measured loss and bound tightness at initialization and trained states (Appendix Figure 9).
Gradient and optimizer-update checks	Agreement with continuous-time identities and reconstruction of realized Adam/Muon updates (Appendix E).

D.1 Aggregate Overlap Trajectories and Within-Configuration Outcomes

The aggregate G_A in equation 9 uses modes $1, \ldots, 63$, one representative of every nonconstant conjugate pair at $M = 127$. The mode set is fixed across time and models, independently of power. For positive overlaps, $\Delta \log G_A$ is the equally weighted mean of $\Delta \log A_k$, the overlap terms in equation 8. Thus a given multiplicative change contributes equally at every frequency. If any mode has zero block energy, the aggregate is undefined; if all modes are defined but any overlap is exactly zero, then $G_A = 0$ and $\log G_A = -\infty$. All 38,545 states in this sample have finite positive overlap for all 63 modes.

Continuous trajectories. Each run has 593 recorded states: every epoch from 0 through 199, every 25 epochs from 200 through 9975, and the final state at epoch 10,000. We report G_A, task-aligned Fourier concentration, effective rank, and train/test accuracy and loss. Absolute-time curves use recorded epochs. For alignment to memorization or grokking, each run contributes its nearest recorded state within 12.5 epochs of a 25-epoch offset grid point. Alignment uses actual observations. The grokking-aligned panel contains the 22 runs reaching that event. Pointwise 95% intervals use 2,000 stratified cluster-bootstrap resamples, with the same draws across offsets. Figures 3 and 5 show points supported by at least three runs and report effective sample counts. Curves pool outcome groups across configurations; the endpoint analysis below compares outcomes within configurations.

Endpoint association within configurations. Let $z_i = \log G_A(10{,}000)$ and let y_i indicate whether run i reaches held-out accuracy 0.95 by the horizon. We estimate the coefficient of y_i in a regression with a separate intercept for each configuration, defined by optimizer and training fraction. Equivalently,

$$\widehat{\Delta}_G = \frac{\sum_c w_c(\bar{z}_{c,1} - \bar{z}_{c,0})}{\sum_c w_c}, \qquad w_c = \frac{n_{c,1} n_{c,0}}{n_{c,1} + n_{c,0}}.$$

Six configurations contain both outcomes (30 runs); seven have a single outcome and receive zero weight in this contrast. All six log-overlap differences are positive, with aggregate $\widehat{\Delta}_G = 1.605$ and $\exp(\widehat{\Delta}_G) = 4.976$. The 95% cluster-bootstrap interval for the geometric-mean ratio is $[3.738, 6.976]$. Counts and weights are recalculated in each bootstrap resample. All 32 within-configuration cross-

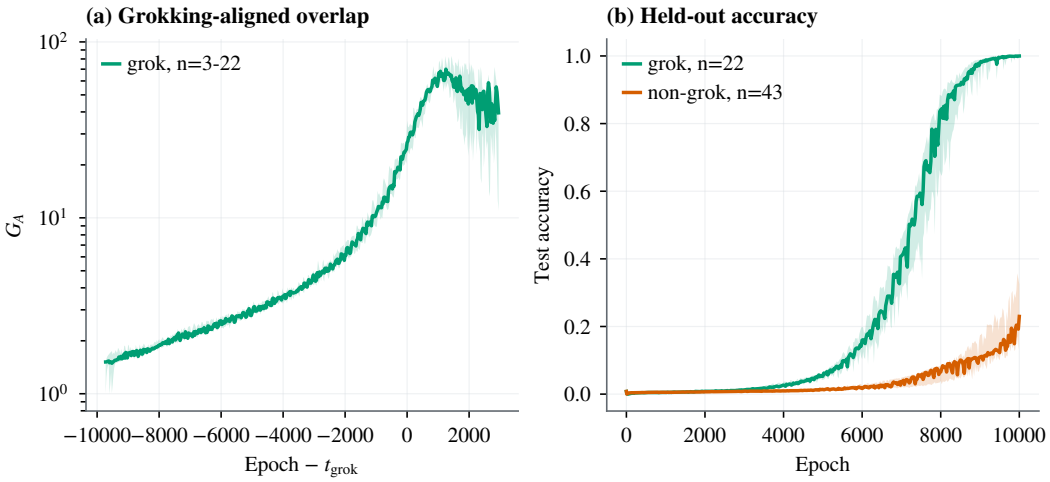


Figure 5: Complementary views of the overlap trajectories. **(a)** Aggregate overlap aligned to grokking, conditional on the 22 runs reaching the held-out accuracy threshold. **(b)** Held-out accuracy in absolute time for both outcome groups. Lines are medians; bands are pointwise 95% cluster bootstrap intervals. Legends give effective run counts. Grokking alignment uses recorded states within 12.5 epochs of each offset.

outcome pairs have higher overlap in the grokking run; uncertainty estimates account for shared runs through cluster resampling.

Table 3: Endpoint overlap across training configurations. Each configuration contains five runs; the sample contains 22 grokking and 43 non-grokking runs. The overlap ratio compares geometric means of G_A between the two outcomes within a configuration. A dash denotes a configuration with only one outcome. Common model and optimizer settings are given in Appendix F.1.

Optimizer	Training fraction (%)	Grokking	Non-grokking	Overlap ratio
Adam	15.000	0	5	–
	15.500	0	5	–
	15.600	0	5	–
	15.800	0	5	–
	16.000	0	5	–
	16.200	4	1	10.32
Muon	14.800	0	5	–
	15.000	1	4	3.02
	15.125	3	2	6.71
	15.200	3	2	3.34
	15.250	3	2	7.83
	15.375	3	2	3.00
	15.600	5	0	–

Mode-level robustness. We group modes by whether their maximum power in the final tenth of training reaches at least 70% of their trajectory peak, omitting modes whose peak power is at most 10^{-12}. In 23 Muon runs containing both groups, we compare endpoint overlap separately for modes remaining near peak power and for the remaining modes. Applying the configuration-weighted contrast above to each run's log median endpoint A_k gives geometric-mean ratios of 4.52 and 8.71, respectively, across five mixed-outcome configurations (19 runs). The corresponding within-configuration pairwise orderings are 15 of 15 and 14 of 15 in the grokking direction. Configuration-level mean contrasts for the latter group are positive in four of five configurations. The pooled endpoint association therefore extends beyond modes that remain near their trajectory peak.

D.2 Within-Block Concentration and Relative Alignment in Training

Section 3.4 factors overlap into within-block concentration $K\|p_k^L\|_2\|p_k^R\|_2$ and relative alignment $\cos\theta_k$. For positive factors,

$$\Delta \log A_k = \Delta \log\big(K\|p_k^L\|_2\|p_k^R\|_2\big) + \Delta \log \cos\theta_k. \tag{36}$$

The first term measures changes in the concentration of each frequency's energy across neurons within the two input blocks. The second measures relative alignment of their nonnegative energy-allocation vectors. Complex phases and amplitudes jointly set interaction direction. Task-aligned Fourier concentration $m_{\mathrm{diag},90}$ summarizes the per-neuron task-aligned share of nonconstant activation energy (Appendix D.3). Fixed-marginal amplitude permutations preserve the first factor and change the second; natural training changes both.

Measurements and uncertainty. We extract both factors from all 65 runs at epochs 0, 250, 5000, 9975, and 10,000, giving 325 weight checkpoints. The covariance analysis uses 55 runs and all 126 nonconstant modes, as in Section 4.3. Aggregate factors use the geometric mean over the same 63 nonredundant modes as G_A, with equal log weights at every time and outcome. All measured factors are finite and positive. The decomposition reproduces the overlap covariance contributions and endpoint contrast; the maximum absolute discrepancy in covariance contributions is 6.6×10^{-15}. Intervals use 10,000 stratified cluster-bootstrap draws, retaining all runs with the same initialization identifier. Strata group clusters covering the same set of configurations (Appendix F.1).

Covariation with mode-power growth. For each term on the right of equation 36, we compute its covariance with $\Delta \log P_k$, divided by $\mathrm{Var}_k(\Delta \log P_k)$, exactly as in Section 4.3. Linearity makes the two contributions sum to the overlap contribution within each run. Table 4 reports their separate medians and additive equal-run means. These statistics quantify descriptive covariation across modes. Within-block concentration has the larger contribution under both summaries, and the relative-alignment covariance term is positive in every run. The alignment medians are 12.63% for 25 Adam runs (95% interval $[10.46\%, 16.18\%]$) and 8.08% for 30 Muon runs ($[7.25\%, 11.23\%]$). Radius and balance retain median contributions of 19.3% and 0.015% to total log-power growth. We compute each median separately across runs, so the medians need not sum to the median overlap contribution. Equal-run means preserve the additive identity before rounding.

Table 4: Concentration and alignment components of overlap. Panels (a) and (b) summarize covariance contributions to cross-mode log-power growth over the 55-run covariance sample, using separate medians and additive means, respectively. Panel (c) compares endpoint geometric means between grokking and non-grokking runs within six mixed-outcome configurations (30 runs). All intervals use stratified cluster resampling.

Factor	Estimate	95% interval
(a) Median covariance contribution (%)		
Overlap	80.87	80.40–81.70
Within-block concentration	70.52	68.87–71.70
Relative alignment	9.92	8.51–11.54
(b) Mean covariance contribution (%)		
Overlap	76.81	71.46–80.27
Within-block concentration	61.45	54.14–68.33
Relative alignment	15.35	11.35–20.52
(c) Configuration-adjusted endpoint ratio		
Overlap	4.976	3.738–6.976
Within-block concentration	4.562	3.504–6.261
Relative alignment	1.091	1.058–1.123

Endpoint association. We apply the within-configuration contrast in Appendix D.1 separately to the log aggregate concentration and cosine factors. The same configuration weights give an additive log contrast and hence a multiplicative ratio decomposition: $4.976 \simeq 4.562 \times 1.091$. Concentration

supplies the larger contrast; relative alignment adds a positive difference in all six configuration means and in 30 of 32 cross-outcome pairs. Using initialization-to-final changes in log aggregate cosine gives a ratio of 1.088 (95% interval $[1.050, 1.124]$), again with positive mean differences in all six configurations. These estimates link relative alignment and its growth to outcomes at fixed training settings. All 65 runs enter resampling; each draw determines its mixed-outcome configurations. Of 10,000 draws, 9,998 yield an estimable contrast, and these define the reported intervals.

Growth during training. The aggregate cosine factor increases from initialization to epoch 10,000 in all 65 runs, with its across-run geometric mean rising from 0.505 to 0.874. The mean log-growth decomposition gives 2.4185 for overlap (95% interval $[2.2695, 2.5685]$), 1.8697 for within-block concentration ($[1.7262, 2.0126]$), and 0.5488 for relative alignment ($[0.5392, 0.5585]$). These equal-run means quantify average temporal growth; the covariance contributions quantify cross-mode variation, and the configuration contrasts quantify differences between outcomes. The five weight checkpoints support coarse growth summaries. The event analysis below uses the denser recorded overlap and activation-concentration trajectories to resolve threshold order.

D.3 Overlap, Concentration, and Accuracy Event Intervals

For each neuron with positive nonconstant activation energy, define its diagonal energy fraction and the checkpoint concentration statistic by

$$d_j = \frac{\sum_{k=1}^{M-1} |\widehat{h}_j(k,k)|^2}{\sum_{(u,v)\neq(0,0)} |\widehat{h}_j(u,v)|^2}, \qquad m_{\mathrm{diag},90} = \mathrm{percentile}_{90,j}(d_j).$$

Excluding $(0,0)$ is equivalent to subtracting each neuron's mean activation over the full input grid before computing its Fourier energy. Thus d_j measures the share of input-dependent variation carried by task-aligned interactions. The memorization, grokking, concentration, and overlap thresholds are train accuracy 0.99, test accuracy 0.95, $m_{\mathrm{diag},90} = 0.30$, and $G_A(t) = 2G_A(0)$, respectively. Accuracy events use the complete per-epoch evaluation history and final checkpoint. Overlap and concentration events use the 593 recorded states described above.

For each event, $(\ell, u]$ brackets the first threshold crossing between the last below-threshold sample and the first sample at or above threshold. The first recorded accuracy crossing is used even if accuracy later falls below the threshold. We infer an order only when the intervals are separated: event a precedes b if $u_a \leq \ell_b$. Overlapping intervals are unresolved; events not observed by the horizon are right-censored, and two right-censored events have no inferred order.

Table 5: Event ordering and observation counts across 65 runs. Whenever both events are observed, their intervals are separated in the stated order; no pair is reversed or unresolved. The other columns count runs in which only the first event, or neither event, is observed. The default overlap event t_A uses a twofold increase; the last two rows vary this threshold.

Event order	Both observed	First event only	Neither event
$t_{\mathrm{mem}} \to t_A$	61	4	0
$t_A \to t_{\mathrm{conc}}$	35	26	4
$t_{\mathrm{conc}} \to t_{\mathrm{grok}}$	22	13	30
$t_A \to t_{\mathrm{grok}}$	22	39	4
$t_{A,1.5\times} \to t_{\mathrm{grok}}$	22	43	0
$t_{A,4\times} \to t_{\mathrm{grok}}$	22	28	15

Memorization is observed in all 65 runs, the $2\times$ overlap event in 61, concentration in 35, and grokking in 22. For pairs with both events observed, every interval comparison has the same direction: memorization precedes overlap in 61 runs, overlap precedes concentration in 35, and concentration precedes grokking in 22. In particular, every grokking run has the ordered sequence $t_{\mathrm{mem}} < t_A < t_{\mathrm{conc}} < t_{\mathrm{grok}}$ at the observed resolution. Overlap crossings are also common among non-grokking runs: 39 of the 43 such runs cross $2G_A(0)$. The event therefore locates a representation change shared by trajectories with different generalization outcomes.

The fixed sensitivity thresholds $1.5G_A(0)$ and $4G_A(0)$ are reached in 65 and 50 runs, respectively. At both thresholds, all 22 grokking runs have strictly separated overlap-before-grokking intervals. At $4\times$, 28 non-grokking runs also cross, while 15 have neither event observed. These censored cases remain distinct from overlapping observed intervals and from observed events occurring in the reverse order.

D.4 Covariance Contributions to Mode-Power Growth

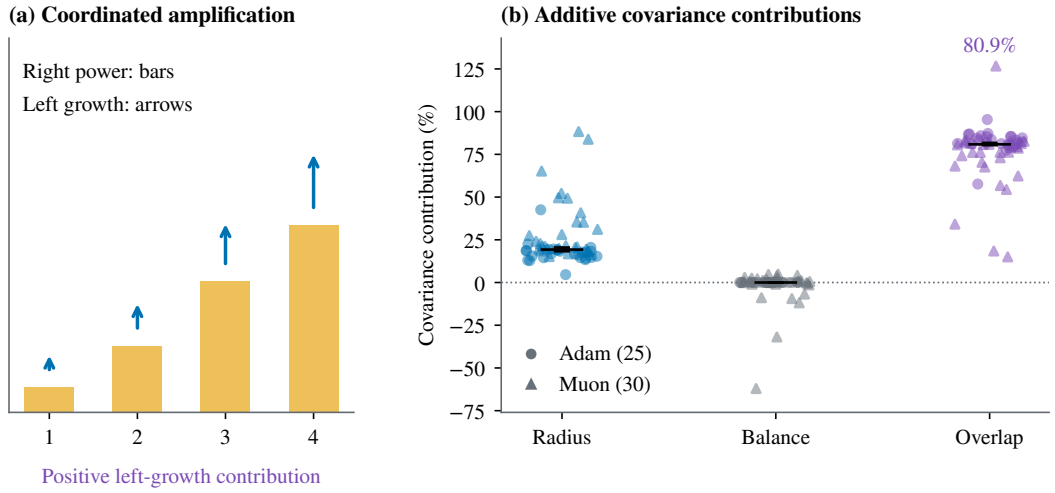


Figure 6: Cross-block overlap and differential mode growth. **(a)** Positive weighted covariance arises when left-block amplification favors neurons with larger right-block power. **(b)** Additive covariance contributions across all 55 runs; markers distinguish optimizers, and black bars show medians with 95% cluster bootstrap intervals. Contributions can be negative; each run's three contributions sum to one. The overlap term includes both within-block concentration and relative alignment; Table 4 separates these components.

Figure 6 connects the weighted-covariance mechanism to the observed contributions of radius, balance, and overlap to differential mode-power growth. The decomposition includes every nonconstant mode in all 55 runs; Table 6 gives the pooled and optimizer-specific estimates.

Table 6: Overlap's additive covariance contribution to cross-mode log-power growth. Median contributions and 95% cluster bootstrap intervals are expressed as percentages.

Group	Runs	Median contribution	95% interval
All	55	80.9	80.4–81.7
Adam	25	83.2	81.3–84.5
Muon	30	77.4	75.2–80.5

D.5 Spectral Diversity and Interaction Power

Fourier-mode effective rank. Within the diagonal subspace, define mode power and, when total power is positive, its normalized distribution

$$P_k(t) = \sum_{j=1}^{K} |\widehat{h}_j(k,k)|^2, \qquad m_k(t) = \frac{P_k(t)}{\sum_{q=1}^{M-1} P_q(t)}.$$

For absolute amplification we separately use the width-normalized power

$$S_{\mathrm{diag}}(t) = \frac{1}{K} \sum_{k=1}^{M-1} P_k(t), \tag{37}$$

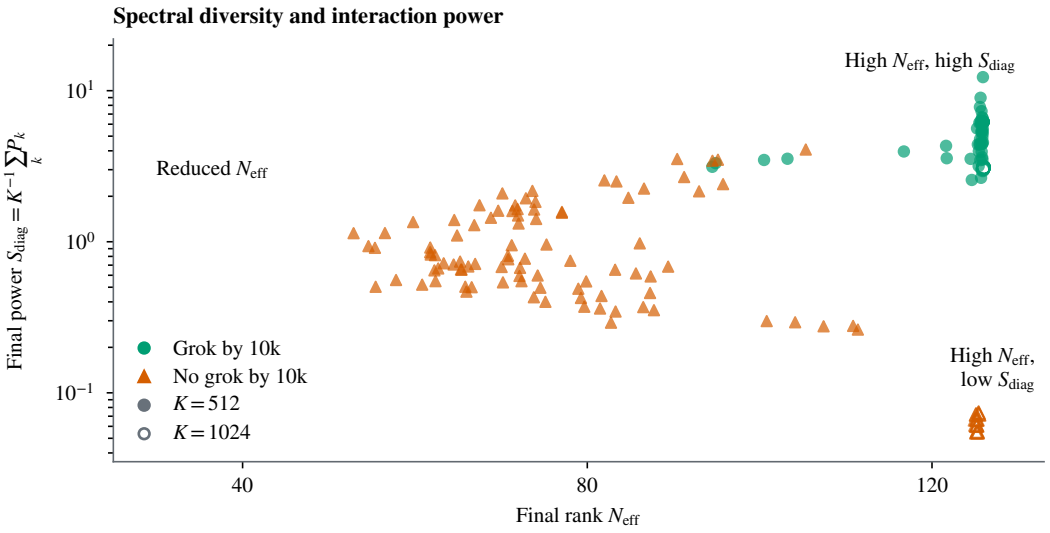


Figure 7: Broad Fourier spectra can carry little task-aligned interaction power. Each point shows final effective rank and total diagonal power per neuron for one of 150 width-512 or 40 width-1024 runs. Shape denotes the outcome by epoch 10,000; fill denotes width. At width 1024, both outcomes retain broad spectra but differ substantially in absolute interaction power.

which retains feature amplitude. The entropy-based effective rank is

$$N_{\text{eff}}(t) = \exp\big(-\sum_k m_k(t) \log m_k(t)\big),$$

where $H(m) = -\sum_k m_k \log m_k$ is Shannon entropy, with $0 \log 0 = 0$. Effective rank ranges from 1 to $M - 1$ and is invariant to total power. For this network $P_k = 4\|q_k\|_2^2$; Appendix C gives the computation. We use effective rank as an auxiliary description of spectral diversity alongside absolute interaction power.

At width 512, mean final effective rank is 123.6 in the 59 grokking runs and 75.5 in the 91 non-grokking runs. Effective rank summarizes spectral diversity; total diagonal power measures how strongly these components are jointly expressed.

Spectral diversity and interaction power. At $K = 1024$, both outcome groups retain broad spectra: mean final N_{eff} is 125.96 in the 19 grokking runs and 125.22 in the 21 non-grokking runs. Median S_{diag} distinguishes the groups: 3.09 and 0.066 (width-512 medians: 5.63 and 0.75). Together, the comparisons reveal two ways task-aligned interactions can remain weak: the spectrum can lose diversity, or remain broad while carrying little absolute task-aligned interaction power (Figure 7).

The link $A_k \longrightarrow P_k \longrightarrow$ aggregate representation connects neuronwise allocation to the distribution (N_{eff}) and scale (S_{diag}) of task-aligned power. Concentration $m_{\text{diag},90}$ additionally compares per-neuron task-aligned power with nonconstant feature energy, including off-diagonal and pure-square components. Section 5 relates task-aligned concentration to generalization.

D.6 Fourier Concentration and Regularization

Figure 8(a) first compares continuous values, without a concentration cutoff: median final $m_{\text{diag},90}$ is 0.379 for the 128 grokking runs and 0.026 for the other 122 runs. Within all seven configurations containing both outcomes (100 runs), grokking runs also have higher mean and median concentration; the weighted within-configuration mean difference is 0.141 (permutation $p < 5 \times 10^{-5}$; Table 7). The distributions partially overlap. Grokking means reaching 0.95 test accuracy within training, even if accuracy later declines.

We use $m_{\text{diag},90} = 0.30$ as a fixed event threshold. Crossing precedes grokking in 127 of the 128 grokking runs at the recorded resolution; one saved crossing occurs 264 epochs later. Among

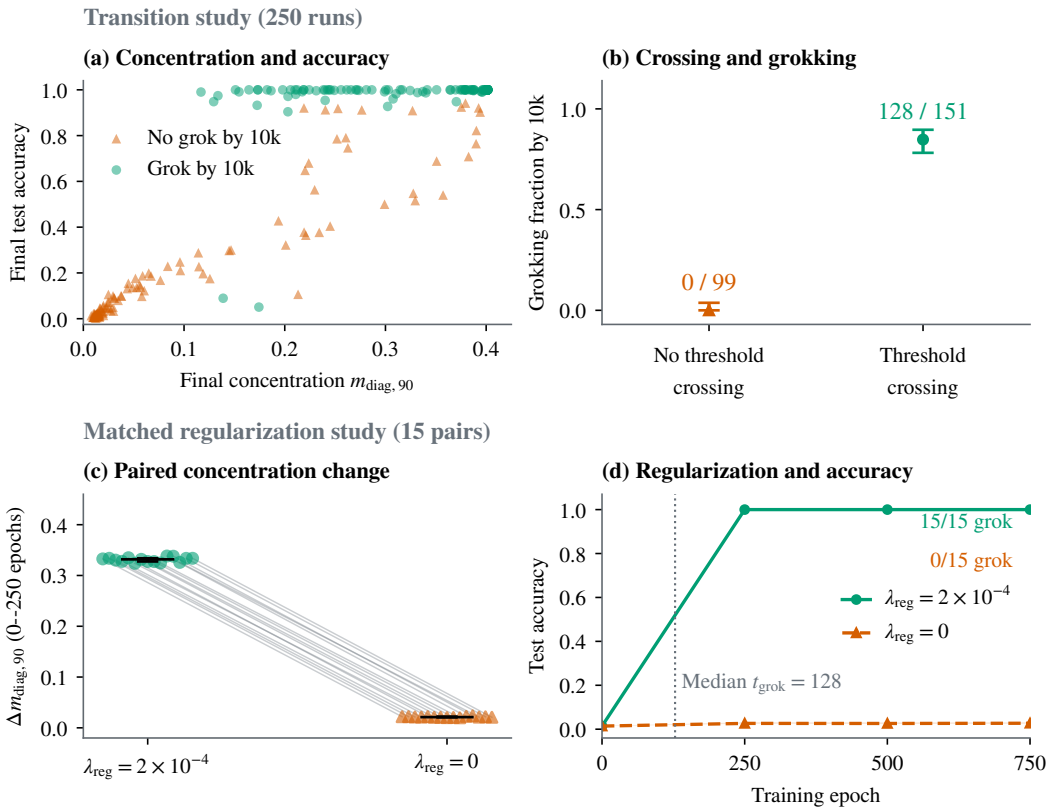


Figure 8: Task-aligned Fourier concentration typically precedes grokking and responds to coupled ℓ_2 regularization. **(a)** Continuous final concentration versus test accuracy for all 250 transition-study runs. **(b)** Grokking frequencies conditional on ever crossing 0.30, with Wilson 95% confidence intervals (CIs). **(c)–(d)** Separate matched Adam regularization experiment: concentration changes over epochs 0–250 and test accuracy over epochs 0–750, using 15 initializations per condition. Lines in (c) join matched initializations; bars/curves show medians with 95% run-bootstrap CIs. Accuracy markers show saved checkpoints, joined by lines; the dotted line marks the median recorded grokking time of 128. The concentration measurement at 250 follows that median event.

all 151 crossing runs, 128 grok within the 10,000-epoch horizon, while 23 remain below the grokking criterion. None of the 99 non-crossing runs groks (Figure 8(b)). Across thresholds 0.20–0.35, crossing precedes grokking in 121–128 of the 128 grokking runs. Appendix D.7 reports the corresponding event counts.

Matched regularization comparison. To study a single regularization change across matched initializations, we use $M = 71$, $K = 2048$, $n = 2016$, and Adam learning rate 10^{-3}. Across 15 paired initializations, coupled ℓ_2 regularization $\lambda_{\text{reg}} = 2 \times 10^{-4}$ yields grokking in all 15 runs; with $\lambda_{\text{reg}} = 0$, all unregularized runs remain below the grokking threshold through epoch 10,000. Median concentration gains by epoch 250 are 0.332 and 0.021, respectively (Figure 8(c–d)). Median memorization and grokking times in the regularized condition are 15 and 128 epochs. The epoch-250 measurement therefore captures the joint shift in representation and generalization induced by regularization. The overlap dynamics provide a complementary mechanistic view: under gradient flow, isotropic shrinkage cancels directly from the A_k equation while changing the radial trajectory and hence the states at which the cross-block covariances are evaluated. This complements fast/slow norm-dynamics accounts (Boursier et al., 2026).

Concentration within training configurations. We compare final concentration within the 21 configurations defined by optimizer and training fraction in the 250-run study. Seven contain both outcomes (100 runs); the other 14 receive zero weight in the comparison. Table 7 gives

the configuration-specific contrasts: mean differences range from 0.007 to 0.325, and the Muon distributions overlap within the transition region.

Let x_i be final $m_{\mathrm{diag},90}$, and let $n_{c,1}$ and $n_{c,0}$ count grokking and non-grokking runs in configuration c. We summarize the mixed-outcome configurations by

$$\widehat{\Delta}_{\mathrm{within}} = \frac{\sum_c w_c(\bar{x}_{c,1} - \bar{x}_{c,0})}{\sum_c w_c}, \qquad w_c = \frac{n_{c,1}n_{c,0}}{n_{c,1} + n_{c,0}}.$$

This is the outcome coefficient in least squares with a separate intercept for each configuration. Scaling each difference by the sample standard deviation across all runs in that configuration gives a weighted difference of 1.365 standard deviations. Among 274 opposite-outcome pairs within configurations, grokking runs have higher concentration with concordance 0.839 (ties receive half weight). Uncertainty estimates account for shared runs across pairs through resampling of initialization clusters.

For a conditional association test, we permute whole outcome vectors for each initialization within strata sharing the same configuration set, preserving each configuration's grokking count and dependence across configurations sharing an initialization. None of 20,000 permutations has an absolute weighted difference at least as large as observed; the two-sided Monte Carlo value, with the plus-one correction, is $1/20{,}001 < 5 \times 10^{-5}$. This reference assumes joint exchangeability of these outcome vectors within these strata under the null of no concentration–outcome association.

Table 7: Endpoint concentration in the seven configurations containing both outcomes (100 runs). Differences subtract non-grokking from grokking concentration. Concordance is the fraction of opposite-outcome pairs in which the grokking run has higher concentration, assigning half weight to ties.

	Runs		Concentration difference		
Training fraction (%)	Grokking	Non-grokking	Mean	Median	Concordance
Adam					
15.875	1	9	+0.325	+0.349	1.000
16.000	6	14	+0.222	+0.296	1.000
16.125	7	3	+0.247	+0.345	1.000
16.250	18	2	+0.204	+0.205	1.000
16.375	9	1	+0.007	+0.007	1.000
Muon					
15.125	4	6	+0.059	+0.087	0.792
15.250	13	7	+0.025	+0.069	0.571

D.7 Sensitivity of Concentration Event Times

We compare first crossings at 0.20, 0.25, 0.30, and 0.35 in the 250-run concentration study, using checkpoints recorded every 250 epochs from initialization to epoch 10,000. Table 8 compares each first saved crossing with the finer recorded grokking time, retaining all 128 grokking and 122 non-grokking runs. All 128 grokking runs cross every tested cutoff; 121–128 cross before grokking. Higher cutoffs occur later and produce more late recorded crossings (seven at 0.35, one at 0.30). Thus concentration usually precedes generalization throughout this range, while exact event counts and waiting times depend on the chosen cutoff and checkpoint resolution.

Table 8: Sensitivity of concentration timing to the threshold. All 128 grokking runs cross every tested threshold, with no crossing recorded at the same epoch as grokking. The first two count columns locate those crossings relative to grokking; the last counts crossings among the 122 non-grokking runs.

Threshold	Before grokking	After grokking	Non-grokking crossings
0.20	128	0	31
0.25	128	0	25
0.30	127	1	23
0.35	121	7	14

D.8 Population-Loss Bound at Trained Checkpoints

Across sixteen runs, the bound has median tightness 23.24% at the 32 trained checkpoints (epochs 5,000 and 10,000), compared with 99.32% at the sixteen initialization checkpoints. Tightness is the ratio of the lower bound to measured population loss; checkpoints repeat within runs. At trained states, the remaining gap includes error from interaction direction and unwanted spatial frequencies, as described in Section 6.

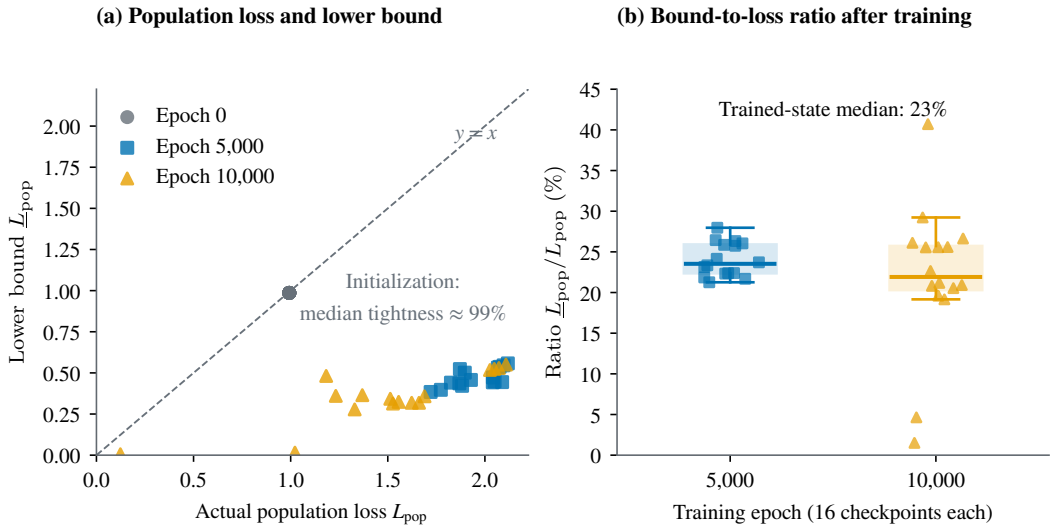


Figure 9: Interaction capacity and population-loss lower bounds at trained checkpoints. **(a)** Population loss and its lower bound at 48 checkpoints from sixteen runs; the dashed line is equality. **(b)** Trained-checkpoint tightness at epochs 5,000 and 10,000, shown on a 0–45% axis. Boxes show the interquartile range (IQR) and median; every checkpoint is plotted. The pooled trained-state median is 23% (32 points). Panel (a) includes initialization checkpoints and annotates their median tightness near 99%. Checkpoints repeat within runs. The bound quantifies the reconstruction cost imposed by mode-wise capacity shortfalls.

E Optimizer Dynamics and Finite-Step Evaluation

E.1 Mode-Growth and Optimizer-Update Experiments

The 55-run covariance analysis uses 25 Adam and 30 Muon trajectories, all at $M = 127$, $K = 512$, regularization/decay coefficient 10^{-4}, and a 10,000-epoch horizon. Each configuration contains five initializations. Adam uses learning rate 0.01 and training fractions $0.150, 0.155, 0.158, 0.160, 0.162$; Muon uses learning rate 0.03 and fractions $0.148, 0.150, 0.15125, 0.152, 0.1525, 0.15375$. The analysis takes the first and final saved weight checkpoints, corresponding to initialization and epoch 10,000. The decomposition requires $P_k > 10^{-12}$ at both endpoints and positive balance and overlap. All 126 nonconstant modes satisfy these conditions in every run, so the decomposition uses the

Table 9: Agreement between observed radial changes and their small-step predictions on nine full-batch gradient-descent trajectories. At every interval length, Spearman correlation exceeds 0.99999998 and growth/decay signs agree in all comparisons. A calibration slope of one denotes exact agreement.

Interval length (steps)	Mode–interval comparisons	Calibration slope
1	56,700	1.000000
2	55,566	0.999982
5	52,164	0.999926

full nonconstant spectrum. Each mode is included once per frequency index; conjugate duplication leaves covariance contributions unchanged. Confidence intervals use 10,000 cluster bootstrap resamples, stratified by configuration set, retaining shared initializations across configurations. The 80.9% estimate is a descriptive additive covariance contribution over the full initialization-to-final interval. Overlap combines within-block concentration and relative alignment, both of which change along these trajectories; Appendix D.2 separates their contributions. Table 6 reports the pooled and optimizer-specific estimates.

Optimizer-update trajectories. We record gradients, optimizer states, and adjacent weights in 45 quadratic networks (15 Adam and 30 Muon) to evaluate the realized-update identities and reconstruct optimizer steps. These comprise the nine configurations above other than Adam at $r = 0.150$ and 0.155, with five initializations each; Appendix I gives the recording schedule and implementation settings.

E.2 Gradient-Descent Evaluation of the Radial Identity

On nine deterministic full-batch gradient-descent trajectories, we compute Γ_k^{data}, Γ_k^{task}, and Γ_k^{rem} directly, then compare $\Delta \log \rho_k$ with the small-step prediction

$$\widehat{\Delta \log \rho_k} = \Delta s\, \eta \big(\Gamma_k^{\mathrm{data}} - \lambda\big),$$

where Δs is the checkpoint interval. Direct gradients and directional finite differences agree, and Table 9 shows near-exact agreement over one-, two-, and five-step windows before task-aligned concentration crosses its threshold.

The gradient-descent trajectories confirm the small-step prediction of the radial law. The realized-update analysis next extends the measurement to the Adam and Muon configurations used in the main experiments.

E.3 Discrete Update Reconstruction

Central finite differences reproduce the continuous-time radius, balance, overlap, and power laws, including cancellation of direct isotropic shrinkage in the overlap rate. Across 1,800 optimizer steps from the 45 networks (40 per run), the squared-radius identity reproduces the measured change with relative error below 7.2×10^{-13}. Independent input-layer update reconstruction uses all 15 Adam runs and 36 recorded updates from six of the 30 Muon runs. Reconstructed updates agree with stored updates to relative error below 1.6×10^{-5}. The update rules and numerical conventions are given in Appendix I.

F Allocation Interventions and Optimizer Settings

We evaluate the functional effects of Fourier interventions and describe the optimizer configurations used to sample near the generalization transition.

F.1 Fixed-Marginal Allocation Experiments

Models and evaluation. The 65-run modular-addition analysis spans 13 configurations with five initializations each (Table 3); 55 runs also contribute to the mode-growth analysis in Section 4.3.

We evaluate the endpoint input-block weights and readouts on the original train/test splits. All have modulus $M = 127$, hidden width $K = 512$, square activation, regularization/decay coefficient 10^{-4}, and a 10,000-epoch training horizon. Learning rates are 0.01 for Adam and 0.03 for Muon; the readout uses Adam in the Muon configuration. Initialization and optimizer settings follow Appendix I.

We apply each intervention to the epoch-10,000 weights and immediately evaluate the transformed network with the trained readout fixed. Train and held-out examples use each model's original random split. Loss is the centered, class-summed, sample-mean MSE in equation 1; accuracy uses the argmax of the original uncentered logits. Each effect subtracts the same model's unmodified endpoint metric, so positive ΔL denotes increased error.

Preserving marginals while changing their relative allocation. For mode k, write the right coefficient as $r_{jk}e^{i\phi_{jk}}$ and apply equation 10. The permutation acts on the amplitudes alone, preserving each neuron's original phase. For initially zero coefficients, the phase is defined as zero. The DC coefficient is unchanged, and the coefficient at $M - k$ is set to the conjugate of the new coefficient at k before the inverse DFT. The permutation preserves both input-block amplitude multisets, their energy sums, ρ_k, β_k, and within-block concentrations. The overlap depends on the cross-block pairing of these amplitudes.

For anti-alignment, sort left amplitudes in increasing order and assign right amplitudes in decreasing order; for alignment, assign them in increasing order. The rearrangement inequality minimizes and maximizes the product sum $\sum_j a_{jk}b_{jk}$ within this permutation family. Random controls use independent uniform permutations for each mode. Changes in interaction power then satisfy $P'_k/P_k = A'_k/A_k$ for the positive reference modes, directly linking the intervention to Proposition 1.

Readout compatibility. With each neuron's original phase fixed, amplitude reallocation changes both the norm and direction of q_k. We measure the stabilized readout projection D_k and squared norm Q_k defined in Appendix C.2 for every treatment. D_k is a stabilized signed real projection onto the interaction direction, with magnitude bounded by $\sqrt{Q_k}$. The readout norm remains fixed. Averaging $|D_k|$ over all 63 modes and then over runs gives 1.726 for the original weights, 0.673 after anti-alignment, 1.155 for random amplitude permutations, and 1.684 after alignment. The simultaneous changes in power and readout projection describe how the allocation transform changes the network's task-aligned output. Full-network loss and accuracy also include the induced changes in off-diagonal interactions and pure-square features. Phase controls probe direction at fixed power, while consistent relabeling and inverse transformations verify prediction preservation.

Uncertainty estimates. Confidence intervals use 10,000 stratified cluster-bootstrap resamples. Runs sharing the same initialization identifier form one cluster across configurations; strata group clusters covering the same configuration set. Repeated random transformations are averaged within each run before statistical aggregation. The reported effect is the equally weighted mean of the configuration-level paired changes.

F.2 Allocation Controls and Functional Responses

We evaluate each model under 30 treatments: the original weights, one anti-aligned and one aligned allocation, 16 random amplitude permutations, eight phase-only controls, one global hidden-neuron permutation, and two inverse-restoration controls. Phase-only controls retain all amplitudes while independently randomizing each neuron's phases and restoring conjugate symmetry. They preserve every A_k and P_k while changing the interaction direction. A global neuron permutation applies the same relabeling to both input blocks and the corresponding readout columns. Inverse controls undo the anti-aligned or aligned amplitude permutation using each neuron's original phases.

Table 10 summarizes the functional responses. Anti-alignment increases loss and reduces accuracy in all 65 models and all 13 configurations. Random amplitude permutations produce smaller but substantial changes. Aligned permutations increase overlap, with mean changes of 0.207 in MSE and -0.010 in accuracy. The aligned permutation maximizes power at fixed amplitude marginals while changing interaction direction. The fixed readout uses both, so the resulting loss depends on their joint change. The phase-only response likewise shows the functional contribution of interaction direction at unchanged power. These observations complement the norm-based reconstruction capacity in Corollary 3.

Table 10: Paired changes in held-out loss and accuracy after allocation and phase transformations. Means weight configurations equally; intervals are 95% cluster bootstrap intervals. Repeated random transformations are averaged within each run. Global neuron relabeling and both inverse transformations preserve logits and predictions to numerical precision.

	Loss change ΔL		Accuracy change $\Delta\mathrm{Acc}$	
Transformation	Mean	95% interval	Mean	95% interval
Anti-aligned	$+5.518$	$(4.264,\ 6.786)$	-0.519	$(-0.581,\ -0.458)$
Aligned	$+0.207$	$(0.152,\ 0.256)$	-0.010	$(-0.027,\ 0.007)$
Random amplitudes	$+1.701$	$(1.436,\ 1.969)$	-0.515	$(-0.573,\ -0.458)$
Phase only	$+1.360$	$(1.222,\ 1.489)$	-0.527	$(-0.589,\ -0.465)$

High-accuracy endpoints. Among the 22 models with original endpoint accuracy at least 0.95, anti-alignment gives descriptive configuration-averaged changes of $\Delta L = 8.121$ and $\Delta\mathrm{Acc} = -0.957$. These models span seven configurations and coincide with the runs reaching the grokking threshold in this sample.

The transformations preserve the stated Fourier invariants to numerical precision. Global relabeling and inverse transformations preserve logits and predictions to numerical precision.

F.3 Adam and Muon Generalization Transition Regions

The 250-run transition study uses $M = 127$, $K = 512$, with a common loss, regularization/decay coefficient 10^{-4}, and 10,000-epoch horizon. Adam uses learning rate 0.01 and ten training fractions from 0.150 to 0.165; Muon uses 0.03 and eleven fractions from 0.1425 to 0.160. Each configuration has ten independent initializations, with twenty at Adam $r = 0.160, 0.1625$ and Muon $r = 0.1525, 0.155$. We record whether each run reaches test accuracy 0.95 by the horizon.

The optimizer-specific transition regions motivate using different training fractions for Adam and Muon in the allocation and trajectory experiments. Adam's grokking rate rises from 0.30 at $r = 0.160$ to 0.90 at $r = 0.1625$ (20 runs per configuration), whereas Muon rises from 0.40 at $r = 0.15125$ to 1.00 at $r = 0.15375$ (10 runs per configuration). Figure 10 shows Wilson 95% intervals.

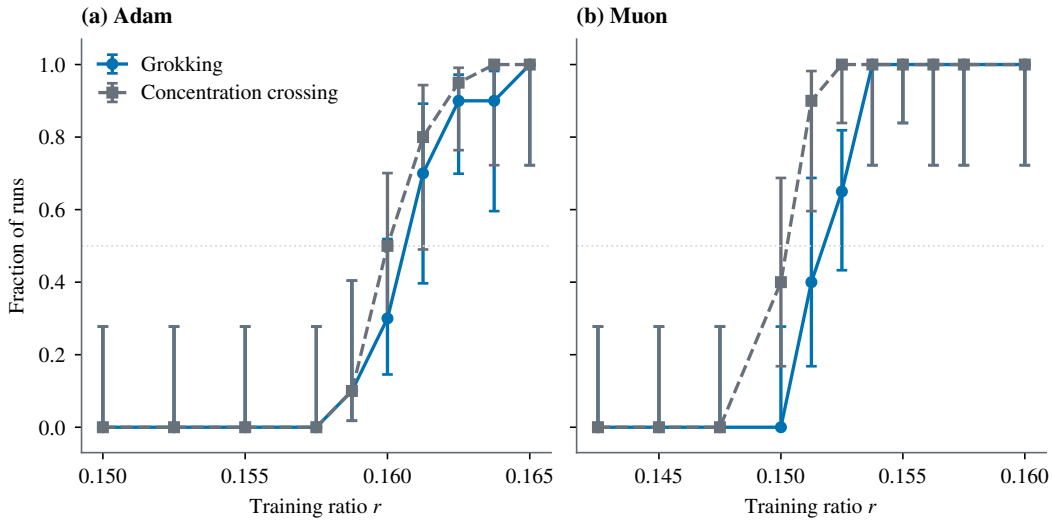


Figure 10: Training configuration shifts the generalization boundary. **(a)** Adam. **(b)** Muon. Circles show grokking rates and squares show concentration-threshold crossing rates; error bars are Wilson 95% intervals. The 250-run transition study is summarized by training fraction. Each point aggregates independent initializations at one training fraction.

G Bilinear Structure and Modular Subtraction

Separable bilinear interactions. The same geometry applies to $h_j^{\times}(a,b) = c\,u_j(a)v_j(b)$: the unitary two-dimensional Fourier transform gives $\widehat{h}_j^{\times}(k,\ell) = c\,\widehat{u}_j(k)\widehat{v}_j(\ell)$. Defining $\rho_{k\ell}$, $\beta_{k\ell}$, and $A_{k\ell}$ from these two coefficient vectors as above yields

$$P_{k\ell}^{\times} = \frac{|c|^2}{4}\rho_{k\ell}^4\beta_{k\ell}^2\frac{A_{k\ell}}{K}.$$

For the quadratic interaction $c = 2$, modular addition selects (k,k) and subtraction selects $(-k,k)$. The task chooses the frequency pairing; separability supplies the factorization.

Modular subtraction. For subtraction, pairing $\widehat{W}_L[j,-k]$ with $\widehat{W}_R[j,k]$ preserves the mode-power factorization. The same factorization matches measured subtraction-mode power to numerical precision at moduli $M = 23$ and $M = 41$. The relevant change from addition is the task-aligned frequency pairing; the definitions of mode norm, block balance, and normalized cross-block overlap otherwise remain the same. Relabeling the second input by $b \mapsto -b \bmod M$ transforms subtraction into addition. These numerical checks validate the factorization under a change of frequency pairing within the shared separable bilinear structure.

H Additional Related Work

H.1 Context for Feature Learning and Optimization

Grokking and feature learning. Grokking has been linked to Fourier circuits, norm dynamics, regularization, implicit bias, and the transition from lazy to feature-learning regimes (Nanda et al., 2023; Zhong et al., 2023; Liu et al., 2022b; Kumar et al., 2024; Varma et al., 2023; Lyu et al., 2024; Notsawo et al., 2025). Effective representation theories and adaptive-kernel analyses describe structured representations and phase transitions (Liu et al., 2022a; Rubin et al., 2024); compression and robustness offer complementary accounts (Liu et al., 2023; Tan & Huang, 2023). Corrupted-label experiments further separate memorizing and generalizing components (Doshi et al., 2024a). Other work studies subnetworks and latent algorithmic structure before the accuracy transition (Minegishi et al., 2025; Swaroop, 2026). We study task-aligned interaction formation through the relative allocation of matching Fourier components across hidden units in the two input blocks.

Fourier representations and dynamics. Mechanistic studies recover Fourier algorithms and phase-structured solutions (Gromov, 2023; Nanda et al., 2023; Furuta et al., 2024). Tian (2025) describes three stages from random-feature memorization to interactive feature learning; He et al. (2026) analyze phase symmetry, frequency diversification, and within-neuron competition; and Ding et al. (2024) study which circular Fourier representations survive. Margin-maximization analyses characterize Fourier features in algebraic tasks (Morwani et al., 2024). Studies of learned solutions and aggregate Fourier mass further characterize representation changes (Mohamadi et al., 2024; Doshi et al., 2024b; Mallinar et al., 2024). We analyze how the two input-side components of a *fixed frequency* jointly produce task-aligned interaction power. This adds the geometry and dynamics of cross-block allocation to the analysis of Fourier representations.

The shared-input formulation of He et al. (2026) applies the same weights to the sum of the two embeddings and studies which frequency occupies a neuron. Independent input blocks introduce the relative alignment of two energy allocations; tying the blocks reduces overlap to the concentration of a single allocation (Section 3.4). We study this additional geometry through fixed-marginal pairing interventions and track both concentration and relative alignment during training. Table 11 compares the questions and representation quantities.

Finite width and optimization. Random-feature and neural-tangent-kernel models provide reference descriptions of learning with fixed or nearly fixed features (Rahimi & Recht, 2007; Jacot et al., 2018; Arora et al., 2019). Initialization scale controls lazy versus rich behavior (Chizat et al., 2019; Woodworth et al., 2020), while spectral bias, gradient starvation, and increasing function complexity characterize which directions are learned during training (Basri et al., 2020; Pezeshki et al., 2021;

Kalimeris et al., 2019). Adaptive-optimizer slingshots reveal additional late-training dynamics associated with grokking (Thilak et al., 2022); learning dynamics also depend on regularization (Kumar et al., 2024; Manir & Rupa, 2026). Our outcome comparisons condition on the training configuration, including optimizer, learning rate, regularization, and training fraction.

Generalization and model capacity. Grokking also occurs in non-neural models (Miller et al., 2023). Its relation to double descent has been studied through differences in pattern learning speeds (Davies et al., 2023), building on observations of non-monotone generalization with model size and training time (Belkin et al., 2019; Nakkiran et al., 2021). Capacity bounds for overparametrized networks (Neyshabur et al., 2018) and jamming transitions near interpolation (Spigler et al., 2019) provide further context for width-dependent behavior. The allocation geometry separates spectral content from interaction power as model width and training configuration change.

Feature allocation and subnetworks. The lottery-ticket hypothesis concerns sparse trainable subnetworks (Frankle & Carbin, 2018), with subsequent work relating their trainability to linear mode connectivity (Frankle et al., 2020). These questions complement the subnetwork view of grokking (Minegishi et al., 2025). Superposition studies how multiple features share representational dimensions (Elhage et al., 2022); our overlap measures whether the two input-side components of a specified Fourier mode occupy the same neurons. Diversity across independently trained predictors (Fort et al., 2019) is also distinct from the within-network mode diversity measured by N_{eff}.

Table 11: Comparison with closely related analyses of modular-addition grokking.

Work	Main focus	Representation quantity
Nanda et al. (2023)	Circuit formation and cleanup before the accuracy transition	Learned Fourier circuits and their computational roles
Tian (2025)	Staged feature emergence and sample-size scaling	Frequency features transitioning from random-feature memorization to interactive learning
He et al. (2026)	Phase symmetry, frequency competition, and diversification	Shared-input Fourier allocation; which frequency occupies a neuron
Ding et al. (2024)	Competition and survival of circular representations	Selection among frequencies during learning
Mohamadi et al. (2024); Mallinar et al. (2024)	Kernel-to-rich transition and emergence of learned structure	Feature emergence in kernel, random-feature, and neural models
This work	How matching components in independent input blocks form functional interactions	Relative cross-block allocation and task-aligned interaction power

I Reproducibility Details

We provide experimental code and analysis scripts as supplementary material. The training configurations appear in Appendix I.2.

I.1 Implementation and Numerical Conventions

Allocation interventions and aggregate-overlap analyses use Python 3.11 and PyTorch 2.3.1 on CPU. Functional effects are computed in float64 and agree with float32 evaluations. Other numerical analyses use Python 3.12 and PyTorch 2.3.1; the additional optimizer-update analyses use Python 3.12.13 and PyTorch 2.11.0+cu126.

The Adam/Muon networks use bias-free linear maps $W \in \mathbb{R}^{K\times 2M}$ and $V \in \mathbb{R}^{M\times K}$ with default uniform initialization and readout multiplier 1.0. The input blocks are the two halves of W. We

write λ_{reg} for the implementation's `weight_decay` coefficient: it gives coupled ℓ_2 regularization in Adam and decoupled input-weight decay in Muon. The gradient-flow derivation uses λ. Adam uses moment-decay coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$, $\varepsilon = 10^{-8}$, and coupled ℓ_2 regularization. Our Muon implementation uses $\beta_1 = 0.9$, $\beta_2 = 0.99$, float32, and five Newton–Schulz iterations with iteration-specific coefficients, with Nesterov and RMS/spectral compensation disabled. Muon multiplies W by $1 - \eta\lambda_{\text{reg}}$ before adding the orthogonalized update; V uses Adam with coupled ℓ_2 at the same learning rate and coefficient. Model checkpoints occur every 250 epochs; optimizer states and updates are recorded every epoch below 200 and every 25 epochs thereafter.

The main experiments used approximately 200 GPU-hours on NVIDIA L40 hardware; additional runs used NVIDIA A40 GPUs. Radial controls use float64, fixed random-number states, disabled TF32, and deterministic algorithms. Paired runs use the same training split and are executed in interleaved order. Allocation, aggregate-overlap, and covariance-contribution intervals use stratified cluster bootstrap resampling, with 10,000 endpoint/summary draws and 2,000 trajectory draws (Appendix F.1). Other median summaries use 10,000 percentile bootstrap resamples of runs with pointwise 95% intervals for trajectories. Curves display observed checkpoint summaries directly. Wilson intervals use independent runs within each configuration; concentration permutations preserve initialization clusters across configurations.

I.2 Training Configurations

Listed training fractions are nominal values rounded to integer sample counts. Here η denotes the constant optimizer step size α_t.

Radial-law evaluation. The radial evaluation uses full-batch gradient descent at $M = 127$, $K = 512$, training fractions 0.155, 0.160, and 0.165 with three initializations each, learning rate 10^{-3}, coupled ℓ_2 coefficient 10^{-4} on W alone, and 50 updates. These controls use the uncentered half-squared loss and independent Gaussian initialization, with variances $1/(2M)$ for W and $1/K$ for V. Every step is saved; gradients and Fourier projections are recomputed in float64. We use one-step windows and evaluate robustness with two- and five-step windows, with intervals measured in optimizer steps.

Representation comparisons across scales. The continuous spectrum–power comparison uses $M = 127$ and widths $K = 512$ and $K = 1024$ (Figure 7). Additional Adam measurements at moduli $M = 23, 41, 71, 97, 127$ show higher endpoint Fourier concentration in grokking runs at each modulus.

Coupled ℓ_2 regularization comparison. The 10,000-epoch comparison in Appendix D.6 uses 15 pairs of Adam runs at $M = 71$ and $K = 2048$. Each pair shares the training split and initialization; the regularization coefficient is the only configuration change.

Optimizer transition configurations. Appendix F.3 gives the settings and replication counts for the 250-run transition study. These runs supply the concentration and outcome comparisons in Figure 8(a–b) and the optimizer-specific transition rates in Figure 10.

Modular subtraction. The subtraction experiments use the same architecture, Adam optimizer, MSE objective, and 10,000-epoch horizon as addition. At $M = 23$, $K = 512$, training fractions are $r = 0.35, 0.375, 0.40$ (3 initializations each). At $M = 41$, we evaluate $r = 0.20, 0.22, 0.24, 0.26, 0.28, 0.30$ and additional initializations at $r = 0.22, 0.24, 0.26$. Appendix G gives the task-aligned frequency pairing and numerical identity evaluation.

J Broader Impact

This work develops tools for understanding how learned components combine into functional interactions. Experiments use synthetic arithmetic data and small neural networks. The analysis may inform methods for interpreting representations whose features depend on interactions across parameter blocks.