Interpreting deep neural networks trained on elementary p groups reveals algorithmic structure

Gavin McCracken[*] McGill University, Mila
Arthur Ayestas Hilgert[*] Mechanica Labs
Sihui Wei McGill University, Mila
Gabriela Moisescu-Pareja McGill University, Mila
Zhaoyue Wang McGill University, Mila
Jonathan Love Leiden University

Abstract

We interpret deep neural networks (DNNs) trained on elementary p group multiplication, examining how our results reveal some of the nature within major deep learning hypotheses. Using tools from computational algebra and geometry, we perform analyses at multiple levels of abstraction, and fully characterize and describe: 1) *the global algorithm* DNNs learn on this task—the multidimensional Chinese remainder theorem; 2) *the neural representations*, which are 2-torus $\mathbb{T}^2$ embedded in $\mathbb{R}^4$ encoding coset structure; 3) *the individual neuron activation patterns*, which activate solely group coset structure. Furthermore, we find neurons organize their activation strengths via the Lee metric. Overall, our work is an exposition toward understanding how DNNs learn group multiplications.

1 Introduction

As deep neural networks (DNNs) scale and are deployed in increasingly high-stakes settings, it's imperative to develop a concrete understanding of how these models make decisions. Black-box models are especially dangerous in domains like healthcare [Basu et al., 2025] or self-driving where Clever Hans-type failures may have catastrophic consequences [Dreyer et al., 2025]. Many argue that ensuring the safety and reliability of such systems will require insights from a broad range of scientific disciplines [Longo et al., 2024, Krueger et al., 2025]. In particular, Eberle et al. [2025] argue that empirical evidence of DNNs learning algorithms warrants an algorithmic understanding of what DNNs learn.

Inspired by how exactly solved models influenced the progression of physics [Baxter, 2016], we draw on the field of group theory to study how DNNs learn elementary p group multiplication. By uncovering *what* the solutions learned are, we gain insights into the nature of three empirical hypotheses we believe fundamental to hopes to create generalizable interpretability tools. i) Universality: DNNs trained with different settings on similar training distributions will learn solutions that utilize similar principles [Li et al., 2015, Olah et al., 2020] ii) The Manifold hypothesis: all natural datasets are distributed on the surface of a low-dimensional manifold; DNNs recover and make use of this manifold [Goodfellow et al., 2016]. iii) Platonic Representation hypothesis: different architecture DNNs trained on different datasets will converge to a *platonic* (shared) understanding of the dataset, reflected by internal distances between similar data aligning across sufficiently overparameterized architectures [Huh et al., 2024].

Contribution 1. We discover that DNNs learn an algorithm that we can fully describe and characterize at all levels of abstraction. We call this global computation the multidimensional Chinese Remainder

[*]Equal contribution. Correspondence: gavin.mccracken@mail.mcgill.ca, aahilgert@gmail.com

Proceedings of the 1st Conference on Topology, Algebra, and Geometry in Data Science(TAG-DS 2025).

Theorem (mCRT). This name comes from our identification that all the neural representations learned correspond to the coset structures in the group.

Contribution 2. Our work, while limited to the study of elementary p and homocyclic groups, offers insight into the aforementioned hypotheses by deepening our collective understanding of how DNNs learn group multiplications. **Universality:** our discovery that neural representations activate on coset structure aligns with prior work on both Abelian (commutative) group multiplications, being the cyclic group [McCracken et al., 2025a], non-Abelian (non-commutative) group multiplications, being the permutation [Stander et al., 2024] and dihedral group McCracken et al. [2025b]. **Manifolds:** we discover neural representations are the surface of the 2-torus $\mathbb{T}^2$, embedded in $\mathbb{R}^4$. This is verified via persistent homology, principal component analyses (PCA) and diffusion maps. **Platonic Representations:** we identify that neurons learn the Lee metric ($\mathbb{T}^2$ geodesic) to order cosets into level sets of approximately constant activation strength.

2 Related work

Analogous to how exact models in physics [Baxter, 2016] progressed practical and theoretical understanding of physical phenomena by allowing for more grounded and holistically complete studies [McCracken, 2021], recent research has looked for datasets providing full understanding of solutions learned by DNNs. Group multiplication tasks have become standard benchmarks for both the mechanistic interpretability [Nanda et al., 2023, Chughtai et al., 2023a,b, He et al., 2024, Tao et al., 2025, Doshi et al., 2023, Stander et al., 2024, McCracken et al., 2025b] and theoretical deep learning [Gromov, 2023, Morwani et al., 2024, Mohamadi et al., 2023, McCracken et al., 2025a] communities. In fact, they've given both empiricists and theoreticians a common ground for proving scientific hypotheses. Notably, group multiplication plays a prominent role in validating the *Universality Hypothesis* [Li et al., 2015, Olah et al., 2020, Chughtai et al., 2023b, Huh et al., 2024], which posits that DNNs learning related tasks will converge to similar internal circuits.

On the empirical side, it is the case that the viral phenomenon of grokking was first identified while training networks on modular addition, which is a group multiplication [Power et al., 2022]. This led to Nanda et al. [2023]'s work, which provided surprisingly clean algorithmic interpretations of transformer architectures to explain grokking. Subsequently, an empirical investigation into the Universality Hypothesis was conducted, generalizing the algorithm from Nanda et al. [2023] by studying both cyclic and permutation group multiplications [Chughtai et al., 2023b]. They claimed that networks universally learned matrix representations of the group and multiplied them to compute answers. Later, it was revealed that this was not the case, Zhong et al. [2023] claimed that two entirely different circuits were being learned by different transformer architectures on modular addition. Thereafter, Stander et al. [2024] showed Chughtai et al. [2023b]'s results weren't reproducible, finding coset circuits, not Chughtai et al. [2023b]'s claimed algorithm were utilized on DNNs learning the permutation group. Thus, claims of universality fell apart modular addition had two competing interpretations and neither aligned with the cosets interpretation on permutation groups.

Recently, both these disparities were resolved. McCracken et al. [2025a] proved that DNNs trained on modular addition were universally utilizing approximate coset structure, conjecturing that DNNs would use such structure to solve all group multiplications. Subsequently, Moisescu-Pareja et al. [2025] failed to reproduce Zhong et al. [2023]'s claims of two disparate circuits, instead finding their architectures learned things the same way, thus reopening the Universality Hypothesis.

Meanwhile, the theoretical community made breakthroughs using cyclic group multiplication as well. Gromov [2023] provided an analytical solution for minimizing cross-entropy loss in networks with quadratic activations. Lyu et al. [2023] argued that smoothness was an inductive bias that could provably induce grokking. This was followed by Morwani et al. [2024], who rigorously proved $\mathcal{O}(n)$ features were required in 1-layer networks. Furthermore, Morwani et al. [2024] argued the reason sinusoidal frequencies emerged during training was due to the theory that DNNs seek to maximize the margin between the correct and second largest output logits, utilizing smoothness norms in their arguments, which was simultaneously, proposed by Mohamadi et al. [2023].

3 Background

Elementary p groups occupy a privileged place in the landscape of abelian structures: they extend the cyclic group—the one-dimensional archetype of abelianness—into higher dimensions while retaining complete uniformity of order. In that sense, they form the maximally symmetric, or isotropic, abelian groups. Let $\mathbb{Z}$ be the set of all integers. We define $\mathbb{Z}_p$ as the additive group resulting from the quotienting of $\mathbb{Z}$ by $p\mathbb{Z}$ where $p \in \mathbb{Z}$ is prime.

In the elementary p group $(\mathbb{Z}_p)^n$, each element is an n-tuple of coordinates taken modulo p. Addition is performed component-wise, each element being reset to 0 when "achieving" the value of p. For example, in $(\mathbb{Z}_3)^2$, adding $(2,1)+(2,2)=(4,3)$ yields $(1,0)$ after we apply mod 3 component-wise. In the group $(\mathbb{Z}_p)^n$, a *coset* is obtained by taking a subgroup $H \leq (\mathbb{Z}_p)^n$ and translating it by some fixed element $g \in (\mathbb{Z}_p)^n$ as such: $g + H = \{g + h \mid h \in H\}$. The set of cosets generated from a subgroup of a group G form a partition of G.

Homocyclic groups. We use "elementary p groups" for $(\mathbb{Z}_p)^n$ with p prime. We call $(\mathbb{Z}_m)^n$ homocyclic when m is arbitrary (i.e., not necessarily prime).

Cayley graph structure. We can visualize the group $(\mathbb{Z}_m)^n$ as an n-dimensional grid with side length m, where opposite faces are identified, as seen in Fig. 1 left panel. In the $n=2$ case, this grid can be rolled into a tube by gluing together one pair of identified opposite faces, then this tube can be turned into a donut by gluing together the remaining pair of identified faces. This results in a 2-torus, as seen in Fig. 1 right panel. In general, all groups of the form $(\mathbb{Z}_m)^n$ can be visualized as a discrete n-torus. Geometrically, cosets appear as parallel affine slices of the n-dimensional discrete torus, as seen in Fig. 1 right panel. More generally, cosets of $(\mathbb{Z}_m)^n$ can be visualized as evenly spaced hyperplanes wrapping around the torus, tiling it into parallel families of points.

Additive functionals. An *additive functional* on $G \cong (\mathbb{Z}_m)^n$ is a group homomorphism:

$$f_\xi : G \to \mathbb{Z}_m, \qquad f_\xi(x) := \langle \xi, x \rangle = \sum_{i=1}^{n} \xi_i x_i \pmod{m}$$

where $\xi = \xi_1, \ldots, \xi_n, x = (x_1, \ldots, x_n) \in G$. Every group homomorphism $f : (\mathbb{Z}_m)^n \to \mathbb{Z}_m$ arises as f_ξ for some ξ, and the correspondence $\xi \mapsto f_\xi$ identifies $\mathrm{Hom}((\mathbb{Z}_m)^n, \mathbb{Z}_m)$ with $(\mathbb{Z}_m)^n$. We can therefore fully describe these linear functionals with the form f_ξ with $\xi \in G$.

For a function $f : X \to Y$, the *kernel* $\ker f_\xi$ is the set of inputs that map to the identity of Y and the *image* $\mathrm{Im}(f_\xi)$ is the set of all outputs attained by f. Now, consider an arbitrary $\xi = (\xi_1, \ldots, \xi_n) \in G$. As f_ξ can be expressed as a dot product, each output of f_ξ is divisible by $d = \gcd(\xi_1, \ldots, \xi_n, m)$. Therefore, $\mathrm{Im}(f_\xi) \subseteq d\mathbb{Z}_m$. Moreover, Bézout's identity Bézout [1779] (see Appendix B for details) guarantees integers $u_1, \ldots, u_n, w$ such that $\sum_{i=1}^{n} u_i \xi_i + wm = d$. Hence $f_\xi(u_1, \ldots, u_n) \equiv d$

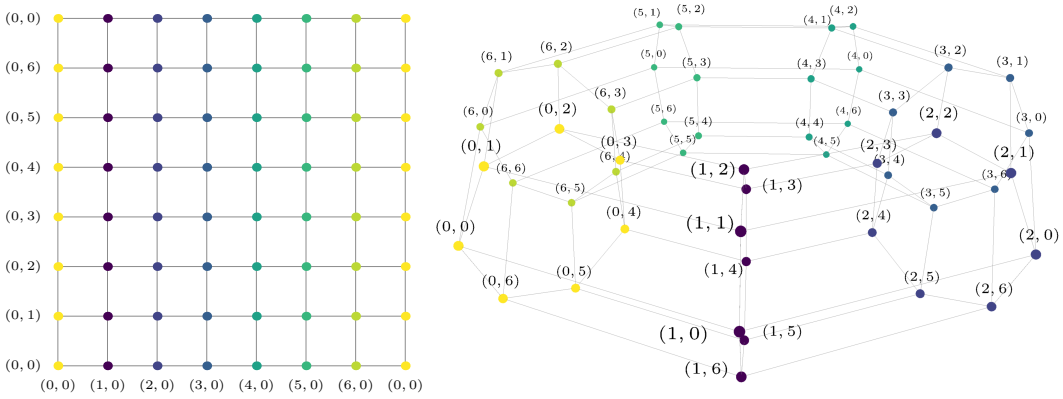

Figure 1: Two equivalent geometric interpretations of the Cayley graph of $(\mathbb{Z}_7)^2$. The left panel shows the Cayley graph of $(\mathbb{Z}_7)^2$ generated by the unit vectors and embedded in $\mathbb{R}^2$. Vertex colorings correspond to cosets induced by $(1,0) \in (\mathbb{Z}_7)^2$. The right panel shows Cayley graph of $(\mathbb{Z}_7)^2$ generated by the unit vectors and embedded in $\mathbb{R}^3$. The vertex colorings correspond to cosets induced by $(1,0) \in (\mathbb{Z}_7)^2$.

$(\mathrm{mod}\ m)$, and by scaling we obtain $f_\xi(tu_1, \ldots, tu_n) \equiv td \pmod m$ for every $t \in \mathbb{Z}$. This shows that every multiple of d in $\mathbb{Z}_m$ is attained, so $\mathrm{Im}(f_\xi) = d\mathbb{Z}_m$. Consequently, $\ker f_\xi$ forms a subgroup of $(\mathbb{Z}_m)^n$ and the cosets induced by this subgroup are precisely the fibers of f_ξ.

Irreducible representations. A *representation* of a group is a way of assigning a matrix to each group element such that multiplying elements in the group corresponds exactly to multiplying their matrices within a vector space. We call a representation *irreducible* if there is no proper, smaller subspace of the vector space that all those matrices leave invariant (i.e., all proper subspaces are altered by at least one matrix).

Characters. For a finite abelian group, such as $(\mathbb{Z}_m)^n$, every irreducible representation is one-dimensional. Let $G = (\mathbb{Z}_m)^n$. A one-dimensional representation is a homomorphism from G into the multiplicative group of nonzero complex numbers $\mathbb{C}^\times$. These one-dimensional irreducible representations are called *characters*.

Given an additive functional f_ξ of $(\mathbb{Z}_m)^n$, we obtain a character χ_ξ by exponentiation, as follows: $\chi_\xi(x) := e^{\frac{2\pi i}{m} f_\xi(x)}$.

As $f_\xi(x + y) = f_\xi(x) + f_\xi(y)$, the character is therefore multiplicative: $\chi_\xi(x + y) = \chi_\xi(x)\chi_\xi(y)$.

Therefore, each χ_ξ is an irreducible representation of $(\mathbb{Z}_m)^n$. Moreover, each irreducible representation of $(\mathbb{Z}_m)^n$ can be derived in such a fashion, by which vectors $\xi \in (\mathbb{Z}_m)^n$ have a one-to-one correspondence with the characters χ_ξ.

In this way, the "period" of a character is equivalent to the index of its kernel in $(\mathbb{Z}_m)^n$, and the oscillatory pattern of the characters is the cyclic repetition of its values across these cosets. For instance, in $\mathbb{Z}_6 = (\mathbb{Z}_6)^1$, the character χ_3 has image $\{-1, 1\}$ of size 2, therefore its kernel has index 2, and χ_3 partitions $\mathbb{Z}_6$ into two cosets $\{0, 2, 4\}$ and $\{1, 3, 5\}$. χ_3 oscillates between -1 and 1 as it traverses $\mathbb{Z}_6$ in the path prescribed by its *directional frequency vector* $\xi = 3$ used to construct the additive functional and thereby its associated character as seen in Fig. 2.

Slopes on the Product Groups. Consider the product group $G \times G = (\mathbb{Z}_m)^n \times (\mathbb{Z}_m)^n$, with elements written as $(a, b) \in G \times G$. Additive functionals on $G \times G$ can be constructed from a pair of directional frequency vectors $\xi_1, \xi_2 \in (\mathbb{Z}_m)^n$ and a pair of coefficients $(\alpha, \beta) \in (\mathbb{Z}_m)^2 \setminus \{(0, 0)\}$. The functional is defined as $\ell_{\xi_1,\xi_2;\alpha,\beta}(a, b) := \alpha f_{\xi_1}(a) + \beta f_{\xi_2}(b) \pmod m$, where f_{ξ_1} and f_{ξ_2} are the additive functionals.

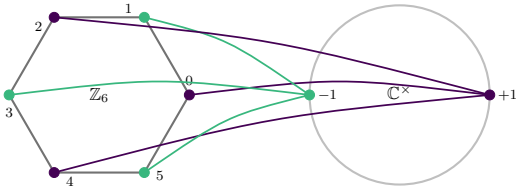

Figure 2: The mapping from $\mathbb{Z}_6$ to $\mathbb{C}^\times$ defined by the character χ_3 of the element $3 \in \mathbb{Z}_6$

The pair (α, β) specifies the *slope* of the functional. Each functional $\ell_{\xi_1,\xi_2;\alpha,\beta}$ partitions $G \times G$ into cosets of its kernel. Geometrically, these cosets are elements of the Cartesian products of two slices of the n-torus defined by the relation $\alpha f_{\xi_1}(a) + \beta f_{\xi_2}(b) \equiv c \pmod m$. Vertical and horizontal slopes correspond to parallel slices aligned with the coordinate axes, while diagonal slopes correspond to tilted slices that cut across both coordinates simultaneously.

The slope structure indicates whether a neuron is acting independently on one argument, or detecting correlations between a and b.

Metrics and Geometry. The group $(\mathbb{Z}_m)^n$ carries a natural geometric structure that depends on the modulus m. This geometry is expressed in terms of distances, graphs, and higher-dimensional complexes built directly from the group elements.

Hamming metric. When $m = 2$, each group element is a binary vector $x \in \{0, 1\}^n$. The natural metric is the *Hamming distance*, defined as the number of coordinates in which two vectors differ. The Cayley graph of $(\mathbb{Z}_2)^n$ under this metric is the n-dimensional hypercube with a vertex for each binary vector and edges that connect vectors differing in exactly one coordinate.

Lee metric. When $m > 2$, each coordinate of the vector $\xi \in (\mathbb{Z}_m)^n$ is an element of $\mathbb{Z}_m$. The natural generalization of the Hamming metric is the *Lee metric* that measures cyclic distance. For $a, b \in \mathbb{Z}_m$, the Lee distance is: $d_{\text{Lee}}(a, b) = \min(|a - b|, m - |a - b|)$. The total Lee distance between $\xi_1, \xi_2 \in (\mathbb{Z}_m)^n$ is the sum of Lee distances in each coordinate.

The directional discrete Fourier transform (dDFT). The dDFT probes a neuron's preactivation $N(a, b)$ on $G \times G = (\mathbb{Z}_m)^n \times (\mathbb{Z}_m)^n$ by projecting onto one-dimensional fibers cut out by additive functionals $f_\xi(x) = \langle \xi, x \rangle \pmod{m_\xi}$ (with $m_\xi = m/\gcd(\xi_1, \ldots, \xi_n, m)$) and their characters $\chi_\xi(x) = e^{2\pi i f_\xi(x)/m_\xi}$. For a slope (α, β), we form the fiber coordinate $t = \alpha\langle \xi_a, a \rangle + \beta\langle \xi_b, b \rangle \pmod{L}$ (with $L = \mathrm{lcm}(m_{\xi_a}, m_{\xi_b})$) and take a 1-D DFT in the basis $\{e^{2\pi i r t/L}\}$ to read off the dominant irreducible representation. Directions are taken up to units and, when m is composite, slopes factor per prime via the Chinese Remainder Theorem (CRT) with energies combined across components; see Appendix C for details.

Betti numbers from algebraic topology are used to distinguish the structure of neural representations in layers. The k-th Betti number B_k counts k dimensional holes: B_0 counts connected components, B_1 counts loops, B_2 counts voids enclosed by surfaces. For reference, a disc has Betti numbers $(B_0, B_1, B_2) = (1, 0, 0)$, a circle has $(1, 1, 0)$, and a 2-torus has $(1, 2, 1)$.

4 Results

Architectures studied. The dataset is all input pairs of two tokens $(a \in (\mathbb{Z}_m)^n, b \in (\mathbb{Z}_m)^n)$, with answer c to $a + b = c$. We use 1- and 2-multilayer perceptrons (MLPs) with rectified linear (ReLU) activations. We use one trainable embedding matrix of size $|(\mathbb{Z}_m)^n| \times 128$ features. The embeddings for inputs a, b to the DNN are rows E_a and E_b, which are selected from this matrix and presented to the first layer by concatenating E_a and E_b. All layers have 1024 neurons. The last layer of the network is a standard linear layer with $|(\mathbb{Z}_m)^n|$ output logits, and networks are trained to minimize cross-entropy loss, *i.e.,* place most of their output mass on the correct logit c. We train with Adam optimizer [Kingma and Ba, 2014], weight decay (i.e., standard ℓ_2-regularization), and a 90/10 train/test split. We use the novel methodology of Moisescu-Pareja et al. [2025] for studying group multiplications, in order to both directly identify the neuron's participating in each neural representation, and visualize and quantify the manifolds corresponding to the neural representations.

Our primary result is that we can give mathematical descriptions for what neurons learn (cosets), for neural representations ($\mathbb{T}^2$) and even the global computational algorithm learned.

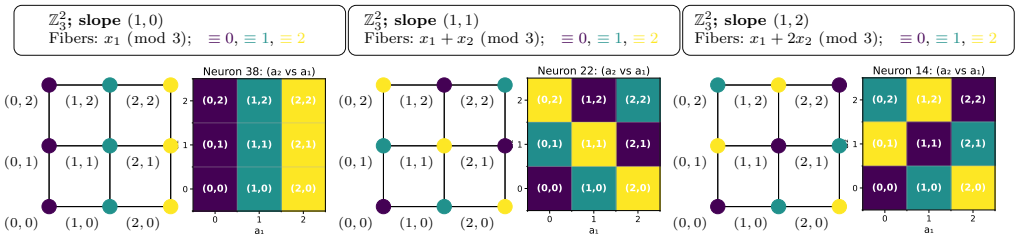

Figure 3: Discrete torus (Cayley graph of $(\mathbb{Z}_3)^2$). The left subfigure shows the schematic fibers, while the right subfigure shows the corresponding 3×3 heatmap obtained by plotting a neuron's $a = (a_1, a_2)$ contribution preactivation matrix. Slopes are $(1, 0)$, $(1, 1)$, and $(1, 2)$ respectively.

Neurons activate on cosets. We can decompose the preactivation value of a neuron into two terms, being the contribution from a, and the contribution from b. We find that neurons in the first layer activate on level sets, being the cosets induced by an additive functional $f_{\xi_a}(a)$ and another $f_{\xi_b}(b)$. Remarkably, all neurons in all models trained learn $\xi_a = \xi_b$, meaning one neuron uses the same projection vector ξ for both a and b. Thus, it has learned cosets of the same "class" of coset.

We visualize the cosets of $(\mathbb{Z}_3)^2$ to show they match the structure learned in the neural preactivations in a trained network in Fig. 3 and visualize $(\mathbb{Z}_3)^3$ in Fig. 4. Note the contribution to the preactivation coming from a can be on a different coset than the preactivation term coming from b (Fig. 5). A theoretical model fitting the neural preactivations follows. We identify that layer-1 neurons utilize the functional Lee metric, and neurons thereafter use the cross-functional Lee metric.

Functional Lee metric. The Lee metric measures distances along the coordinate axes of $(\mathbb{Z}_m)^n$. However, in many situations the geometry of interest is not determined by the standard coordinates, but rather by an additive functional. Recall that an additive functional is a homomorphism where $\xi \in (\mathbb{Z}_m)^n$ specifies the direction.

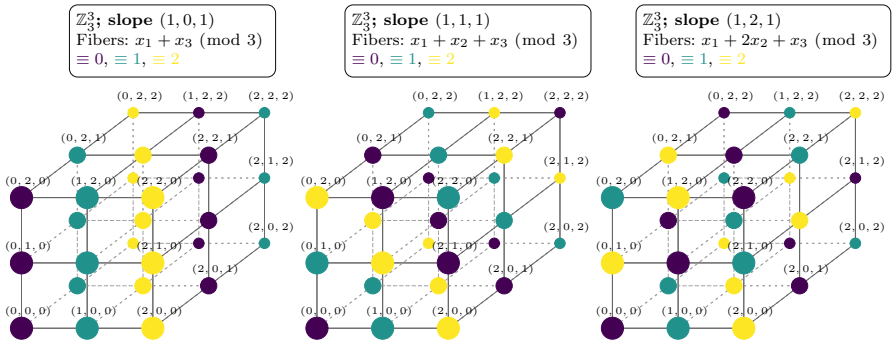

Figure 4: Cayley graph of $(\mathbb{Z}_3)^3$. Three schematic panels corresponding to neuron preactivations (left to right) with slopes $(1,0,1)$, $(1,1,1)$, and $(1,2,1)$. Colors indicate fiber level sets of $x_1 + x_3 \pmod 3$, $x_1 + x_2 + x_3 \pmod 3$, and $x_1 + 2x_2 + x_3 \pmod 3$, respectively.

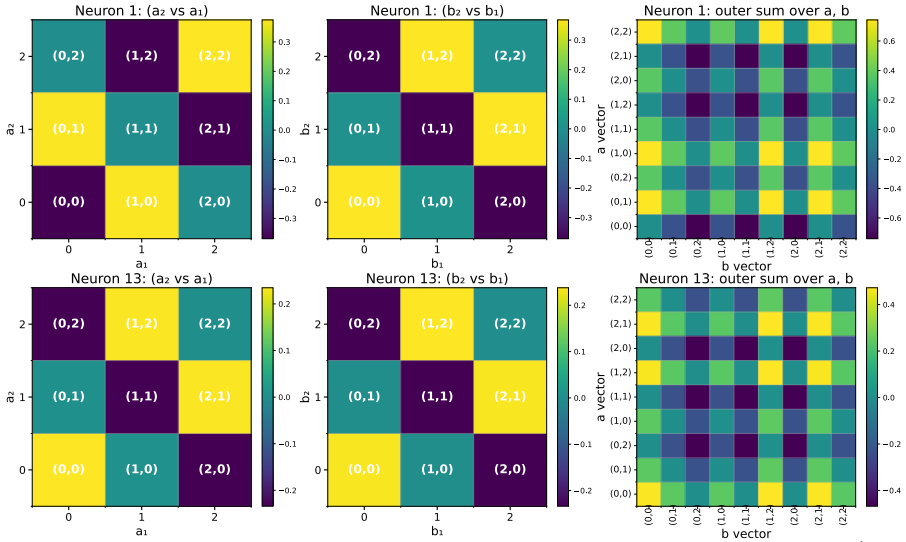

Figure 5: **Layer 1 neurons:** (Top row) preactivations of neuron 1 from inputs $a = (a_1, a_2)$ and $b = (b_1, b_2)$, concentrating on different cosets, $a_1 + a_2 \equiv 1 \pmod 3$ versus $b_1 + b_2 \equiv 0 \pmod 3$. (Bottom row) preactivations of neuron 13 from a and b, activate on the same coset, $a_1 + a_2 \equiv 0 \pmod 3$ and $b_1 + b_2 \equiv 0 \pmod 3$. Neuron 1 and 13 are members of the same neural representation. Both learn to project a and b onto $\xi = (1, 1)$, meaning $1(a_1)+1(a_2) \bmod 3$ and $1(b_1)+1(b_2) \bmod 3$ give the cosets.

This functional partitions the group into cosets of $\ker f_\xi$. The image $\mathrm{Im} f_\xi \cong d\mathbb{Z}_m$, where $d = \gcd(\xi_1, \ldots, \xi_n, m)$, of f_ξ is a subgroup of $\mathbb{Z}_m$, and hence has a natural Lee metric. Pulling this metric back along f_ξ defines a distance on $(\mathbb{Z}_m)^n$ that reflects separation along the direction ξ.

For $a, b \in (\mathbb{Z}_m)^n$ and $\xi \in (\mathbb{Z}_m)^n$, the *functional Lee distance* is: $d_\xi(a, b) := d_{\text{Lee}}\big(f_\xi(a), f_\xi(b)\big)$.

Geometrically, $d_\xi(a, b)$ measures how many steps one must move along the direction ξ in order to pass from the coset of a to the coset of b, with wrap-around on the m-cycle. All points in the same coset of $\ker f_\xi$ are identified at distance zero, while different cosets are spaced evenly according to their residues modulo the image of f_ξ. This coset identification is shown by vertex colors in Fig. 1.

Cross-functional Lee metric. In the product group $G \times G = (\mathbb{Z}_m)^n \times (\mathbb{Z}_m)^n$, the geometry may depend on directions frequency vectors in each factor separately. This is captured by the two-parameter functional $\ell_{\xi_1,\xi_2;\alpha,\beta}(a, b)$.

As before, this functional partitions the product group into cosets of its kernel, which are affine slices of the discrete $2n$-torus. Its image is a subgroup of $\mathbb{Z}_m$ given by

$$\mathrm{Im}(\ell_{\xi_1,\xi_2;\alpha,\beta}) \cong \gcd\big(m, \alpha d_1, \beta d_2\big)\,\mathbb{Z}_m, \qquad d_i = \gcd(\xi_{i1}, \ldots, \xi_{in}, m) \quad i \in \{1, 2\}.$$

This subgroup carries a natural Lee metric, which induces a metric on $G \times G$ via pullback through $\ell_{\xi_1,\xi_2;\alpha,\beta}$.

For $(a,b),(a',b') \in G \times G$, the cross-functional Lee distance is $d_{\xi_1,\xi_2;\alpha,\beta}\big((a,b),(a',b')\big) := d_{\text{Lee}}\big(\ell_{\xi_1,\xi_2;\alpha,\beta}(a,b),\ \ell_{\xi_1,\xi_2;\alpha,\beta}(a',b')\big)$.

Geometrically, $d_{\xi_1,\xi_2;\alpha,\beta}$ measures cyclic distance after projecting (a,b) onto the one-dimensional subgroup generated jointly by ξ_1 and ξ_2 in the slope ratio (α,β). Cosets of $\ker \ell_{\xi_1,\xi_2;\alpha,\beta}$ collapse to distance zero, while distinct cosets are spaced evenly along the cycle of size $m/\gcd(m,\alpha d_1,\beta d_2)$.

In layer 1. The slope (α,β) is (0,1) and (1,0) in this case as the neuron preactivation is independent. Let $\xi \in (\mathbb{Z}_m)^n$ and $\alpha,\beta \in \mathbb{Z}_m$. Let $M := m/\gcd(m,\alpha d,\beta d)$, where $d = \gcd(\xi_1,\ldots,\xi_n,m)$. Let $\ell_{\xi;\alpha,\beta} := \ell_{\xi,\xi;\alpha,\beta}$ be an additive functional on the product group $G \times G = (\mathbb{Z}_m)^n \times (\mathbb{Z}_m)^n$ For $(a,b) \in (\mathbb{Z}_m)^n \times (\mathbb{Z}_m)^n$, a *simple neuron* is a neuron that has the preactivation $N(a,b)$:

$$N(a,b) = \cos\Big(\tfrac{2\pi}{M}\,\ell_{\xi;\alpha=1,\beta=0}(a,b) + \phi_a\Big) + \cos\Big(\tfrac{2\pi}{M}\,\ell_{\xi;\alpha=0,\beta=1}(a,b) + \phi_b\Big) \tag{1}$$

where $\phi_a,\phi_b \in [0,2\pi)$. Note: $(\alpha=1,\beta=0)$ in the a functional and $(\alpha=0,\beta=1)$ in the b functional.

In layers after 1. Let $\xi_1,\xi_2 \in (\mathbb{Z}_m)^n$ and $\alpha,\beta \in \mathbb{Z}_m$. Let $M := m/\gcd(m,\alpha d_1,\beta d_2)$, where $d_i = \gcd(\xi_{i1},\ldots,\xi_{in},m)$ for $i \in \{1,2\}$. For $(a,b) \in (\mathbb{Z}_m)^n \times (\mathbb{Z}_m)^n$, a *simple neuron* is a neuron that has the preactivation: $N(a,b) = \cos\Big(\tfrac{2\pi}{M}\,\ell_{\xi_1,\xi_2;\alpha,\beta}(a,b) + \phi\Big)$, where $\phi \in [0,2\pi)$.

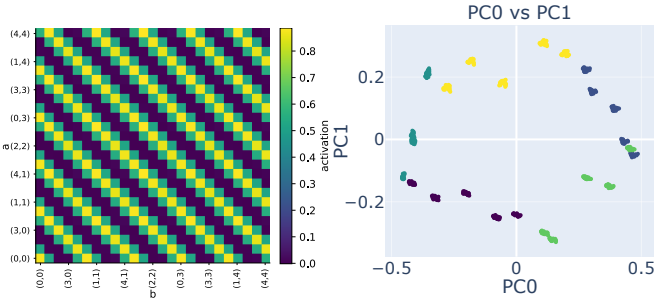

Figure 6: **Layer 2.** (left) neurons in layer 2 have activation patterns with correlations between a and b; they understand cosets in the cross space $(\mathbb{Z}_m)^n \times (\mathbb{Z}_m)^n$ unlike in the preactivation grid (rightmost) in Fig. 5. A PCA reveals a circle was learned, confirmed by persistent homology (Betti numbers (1, 1, 0)). Note, the PCA is colored by the cosets of the answer c—*i.e.,* second layer neurons encode c's coset information.

We test the Simple neuron model quantitatively by training 100 random seeds after hyperparameter tuning via grid search for a good learning rate and weight decay, and inspecting the R^2 of fitting simple neurons in layers one and two. Training 2 layer multilayer perceptrons (MLPs) on $(\mathbb{Z}_5)^2$ achieves a layer 1 avg R^2: 0.99769 and layer 2 avg R^2: 0.89509, on $(\mathbb{Z}_7)^2$ Layer 1 avg R^2: 0.99703 Layer 2 avg R^2: 0.92657, on $(\mathbb{Z}_3)^3$ an R^2 of 0.999 in both layers and on $(\mathbb{Z}_5)^3$ an average R^2 of 0.989 in layer 1 and 0.971 in layer 2. The reason for the lower R^2 values in layer 2 is that we only fit one simple neuron and due to the fully connected nature of MLPs it's easy for the neurons in layer 2 to be composed of linear combinations of any of the irreducible representations found in layer 1.

Neural representations. Using the dDFT, we take all neurons in layer 1 that concentrate on the same direction pair (ξ_a,ξ_b) and get the preactivation vector of length equal to the dataset $|(\mathbb{Z}_m)^n \times (\mathbb{Z}_m)^n|$ for each neuron and stack these vectors into a $\mathcal{M} = |\#\text{ neurons}| \times |(\mathbb{Z}_m)^n \times (\mathbb{Z}_m)^n|$ matrix. We perform principal component analyses on these matrices to reduce dimensionality, finding the first four components explain $> 99.5\%$ of the variance. In layer 2, we take all neurons that concentrate on the same slope and direction pair (u_a,u_b,α,β), and do PCA.

When training on $(\mathbb{Z}_m)^n$, with $m \in \{59,62,97\}$ for $n=1$; $m \in \{5,6,7,11\}$ for $n=2$; and $m \in \{3,4,5,6\}$ with $n=3$, if neurons in a neural representation have learned a coset with at least 3 points in each fiber (coset equivalence class) (*e.g.* Fig. 1 has 6) then: **layer 1** persistent homology will identify the neural representation has Betti numbers (1, 2, 1) implying it's homologically equivalent to $\mathbb{T}^2$, **layers** >1 have (1, 1, 0), implying it's homologically equivalent to a circle. This can be qualitatively seen in Figs. 6, 7, showing a circle and torus respectively.

4.1 Global algorithmic strategy: Chinese Remainder Theorem on Homocyclic Groups

The *Chinese Remainder Theorem* describes how congruences with respect to different moduli interact. In the finite abelian case, it provides a canonical decomposition of groups into factors. For homocyclic

groups, the CRT admits two complementary formulations: an algebraic formulation in terms of direct sum decompositions of subgroups, and a geometric formulation in terms of intersecting cosets and hyperplanes in the cube complex (i.e., the grid-geometric view of the homocyclic group).

Classical CRT. Suppose m has prime factorization $m = \prod_i p_i^{k_i}$. The Chinese Remainder Theorem says that arithmetic modulo m can be broken into simpler arithmetic, one piece for each $p_i^{k_i}$. Consider $\mathbb{Z}_m$ as a circle with m equally spaced points along its circumference. Each divisor $p_i^{k_i}$ marks out a coarser cycle: it partitions the m points into cosets of size $m/p_i^{k_i}$, evenly spaced around the circle. The CRT says that if you know which coset x belongs to, for each of these prime-power partitions of the underlying cyclic group, then you can know exactly which of the m points x is. Equivalently, if a system of congruences of the form $x \equiv a_i \pmod{p_i^{k_i}}$ describes which coset x belongs to on each prime-power cycle, then CRT guarantees that the intersection of these cosets is a single point in $\mathbb{Z}_m$.

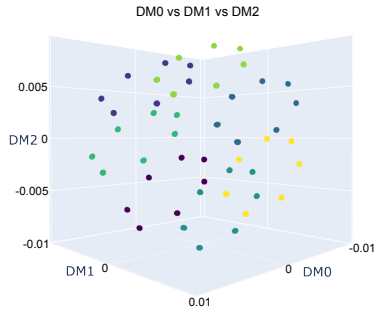

Figure 7: Diffusion map of a Layer 1 neural representation in $(\mathbb{Z}_7)^2$, composed of neurons learning $\xi_a = \xi_b = (1, 1)$, showing a torus $\mathbb{T}^2$.

Multidimensional CRT. Now consider the higher-dimensional group $(\mathbb{Z}_m)^n$, which, as shown in Fig. 1, can be represented as a n-torus. Just as in the cyclic case, divisors of m create coarser partitions of this space. Each additive function f_ξ of $(\mathbb{Z}_m)^n$ partitions the elements of $(\mathbb{Z}_m)^n$ into level sets, and $\ker f_\xi$ defines a subgroup of $(\mathbb{Z}_m)^n$. A subgroup of $(\mathbb{Z}_m)^n$ carves the torus into cosets, which, geometrically, are *parallel affine slices* of the torus, as they consist of all points satisfying the same linear congruence modulo m. The resulting cosets form a family of slices that are parallel when projected onto the Cayley graph of $(\mathbb{Z}_m)^n$, since they are all defined by the same linear relation with only the constant term varying. These slices partition the torus into exactly m/d distinct layers, where $d = \gcd(\xi_1, \ldots, \xi_n, m)$. Together, the slices cover all of $(\mathbb{Z}_m)^n$.

The *multidimensional CRT* asserts that if we select several such functionals, and the subgroups defined by their kernels are independent (i.e., their sum is $(\mathbb{Z}_m)^n$), then the intersection of cosets, each chosen from a distinct functional induced level set partition of $(\mathbb{Z}_m)^n$, consists of exactly one point. In other words, specifying which parallel affine slice x belongs to in each partition uniquely determines x modulo m. This is the abstract algorithm utilized by DNNs learning homocyclic groups (an abstract algorithm is a template, allowing for variations of implementation across random seeds). DNNs perform exactly this: neurons in the last layer specialize to outputting on the cosets of c, which they compute from being fed the coset membership of a and b on $\mathbb{T}^2$ from the preceding layer. The positive and negative interference from different c cosets contributing to the outputs results in argmax selecting the correct logit c. Analogously, argmax can be thought of as taking a set intersection since the correct logit c, is a member of many different cosets (non-parallel affine slices).

5 Discussion

By using a classical algebraic group as a training task, we've provided a toy model where we fully characterize everything DNNs with non-linear ReLU activations learn. We describe everything from neuron activations, to identifying all neurons in each neural representation, to the manifolds learned by DNNs, to the abstract algorithm instantiated across random seeds (mCRT). We claim our qualitative and quantitative results—neurons learn coset structures with high R^2, and layer-1 neural representations form tori before becoming circles encoding the cosets of the answer c—provide sufficient evidence to empirically claim DNNs learn implementations of the mCRT algorithm.

Our results tie into literature: 1) finding cosets critical in DNNs learning homocyclic multiplication aligns with work conjecturing DNNs will universally use coset structure to learn group multiplications [McCracken et al., 2025a]; 2) finding that the functional and cross-functional Lee metrics are learned by neurons to encode activation strengths (cosets near each other w.r.t. the metric have similar activation strengths) ties into the manifold and Platonic Representation hypotheses by matching predictions. This is because DNNs recover the natural geodesic metric on the correct manifolds that describe the task. 3) Our multi-scale analysis allows for the network to be understood at the

neuron, neural representation and global computation levels of abstraction, aligning with calls for theoreticians to progress an algorithmic understanding of how DNNs learn tasks [Eberle et al., 2025].

Limitations, cautions and future work. Due to studying one task we can't make broad claims of generality, but we don't see this as a limitation as much as a need for future work on other group multiplications. Stander et al. [2024] found errors in highly influential work studying the permutation group, and suggested that deep, careful interpretability studies are needed instead of broad ones to ensure such errors aren't repeated in future literature. Additionally, the finding by Moisescu-Pareja et al. [2025]—that multiple prior works incorrectly interpreted DNNs trained on cyclic group multiplications—motivate rigorous and accurate analyses further. In light of these findings, we claim that the utilization of methods from computational algebra can aid future work trying to uncover whether there's universality in how DNNs learn group multiplications.

References

Sumana Basu, Adriana Romero-Soriano, and Doina Precup. Reward the reward designer: Making reinforcement learning useful for clinical decision making. In *Women in Machine Learning Workshop @ NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=2vEeeXzUTh.

Maximilian Dreyer, Jim Berend, Tobias Labarta, Johanna Vielhaben, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Mechanistic understanding and validation of large ai models with semanticlens. *Nature Machine Intelligence*, pages 1–14, 2025.

Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.

Frank Krueger, René Riedl, Jennifer A Bartz, Karen S Cook, David Gefen, Peter A Hancock, Sirkka L Jarvenpaa, Lydia Krabbendam, Mary R Lee, Roger C Mayer, et al. A call for transdisciplinary trust research in the artificial intelligence era. *Humanities and Social Sciences Communications*, 12(1):1–10, 2025.

Oliver Eberle, Thomas McGee, Hamza Giaffar, Taylor Webb, and Ida Momennejad. Position: We need an algorithmic understanding of generative ai. *arXiv preprint arXiv:2507.07544*, 2025.

Rodney J Baxter. *Exactly solved models in statistical mechanics*. Elsevier, 2016.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL https://arxiv.org/abs/2405.07987.

Gavin McCracken, Gabriela Moisescu-Pareja, Vincent Letourneau, Doina Precup, and Jonathan Love. Uncovering a universal abstract algorithm for modular addition in neural networks, 2025a. URL https://arxiv.org/abs/2505.18266.

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets. In *Forty-first International Conference on Machine Learning*, 2024.

Gavin McCracken, Sihui Wei, Gabriela Moisescu-Pareja, Harley Wiltzer, and Jonathan Love. The representations of deep neural networks trained on dihedral group multiplication. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025b. URL https://openreview.net/forum?id=weKecpFYNf.

Gavin McCracken. *Using Exact Models to Analyze Policy Gradient Algorithms*. McGill University (Canada), 2021.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning*, pages 6243–6267. PMLR, 2023a.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. Neural networks learn representation theory: Reverse engineering how networks perform group operations. In *ICLR 2023 Workshop on Physics for Machine Learning*, 2023b.

Tianyu He, Darshil Doshi, Aritra Das, and Andrey Gromov. Learning to grok: Emergence of in-context learning and skill composition in modular arithmetic tasks. *arXiv preprint arXiv:2406.02550*, 2024.

Tao Tao, Darshil Doshi, Dayal Singh Kalra, Tianyu He, and Maissam Barkeshli. (how) can transformers predict pseudo-random numbers? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=asDx9sPAUN.

Darshil Doshi, Aritra Das, Tianyu He, and Andrey Gromov. To grok or not to grok: Disentangling generalization and memorization on corrupted algorithmic datasets. *arXiv preprint arXiv:2310.13061*, 2023.

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

Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham M. Kakade. Feature emergence via margin maximization: case studies in algebraic tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=i9wDX850jR.

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica Sutherland. Grokking modular arithmetic can be explained by margin maximization. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. URL https://openreview.net/forum?id=QPMfCLnIqf.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S5wmbQc1We.

Gabriela Moisescu-Pareja, Gavin McCracken, Harley Wiltzer, Colin Daniels, Vincent Létourneau, and Jonathan Love. On the geometry and topology of neural circuits for modular addition. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=9VCl4UcTZm.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon S Du, Jason D Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. *arXiv preprint arXiv:2311.18817*, 2023.

Etienne Bézout. *Théorie générale des équations algébriques*. Ph.-D. Pierres, 1779.

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

A Glossary of Terms and Notation

$\mathbb{Z}$ The set of all integers.

$\mathbb{Z}_m$ The additive group of integers modulo m, i.e. the quotient $\mathbb{Z}/m\mathbb{Z}$. Elements are residue classes $\{0, 1, \ldots, m-1\}$ with addition taken modulo m.

$(\mathbb{Z}_m)^n$ The direct product of n copies of $\mathbb{Z}_m$. Elements are n-tuples $(x_1, \ldots, x_n)$ with componentwise addition modulo m. This is called a *homocyclic group*.

$\mathbb{R}$ The set of real numbers.

$\mathbb{C}$ The set of complex numbers.

$\mathbb{C}^\times$ The multiplicative group of nonzero complex numbers.

gcd The greatest common divisor of a set of integers, i.e. the largest positive integer dividing all of them.

lcm The least common multiple of a set of integers, i.e. the smallest positive integer divisible by all of them.

Group A set G with a binary operation $\cdot$ such that (i) the operation is associative, (ii) there is an identity element $e \in G$ with $e \cdot g = g \cdot e = g$ for all $g \in G$, and (iii) every $g \in G$ has an inverse g^{-1} with $g \cdot g^{-1} = g^{-1} \cdot g = e$.

Abelian group A group in which the operation is commutative: $g \cdot h = h \cdot g$ for all $g, h \in G$.

Subgroup A subset $H \subseteq G$ that is itself a group under the same operation. Written $H \leq G$.

Normal subgroup A subgroup $N \leq G$ such that $gNg^{-1} = N$ for all $g \in G$. In abelian groups every subgroup is normal.

Quotient group G/H The set of cosets $\{gH : g \in G\}$, with multiplication defined by $(gH)(hH) = (gh)H$.

Coset For $H \leq G$ and $g \in G$, the set $gH = \{gh : h \in H\}$. Cosets partition G into disjoint subsets.

Index $[G : H]$ The number of cosets of H in G, equivalently the size of the quotient group G/H.

Homomorphism A function $f : G \to H$ between groups such that $f(g_1 g_2) = f(g_1) f(g_2)$ for all $g_1, g_2 \in G$.

Kernel $\ker f$ The set of elements in G mapped to the identity of H under a homomorphism $f : G \to H$.

Image $\mathrm{Im}(f)$ The set of elements in H that are outputs of a homomorphism $f : G \to H$.

Additive functional f_ξ A group homomorphism

$$f_\xi(x) := \langle \xi, x \rangle = \sum_{i=1}^{n} \xi_i x_i \pmod{m},$$

with $\xi = (\xi_1, \ldots, \xi_n) \in (\mathbb{Z}_m)^n$. Every homomorphism $(\mathbb{Z}_m)^n \to \mathbb{Z}_m$ is of this form.

Fiber For a function $f : X \to Y$ and $y \in Y$, the fiber $f^{-1}(y)$ is the set of inputs mapping to y. For additive functionals, fibers are cosets of $\ker f_\xi$.

Character χ_ξ A one-dimensional irreducible representation of $(\mathbb{Z}_m)^n$ associated to f_ξ:

$$\chi_\xi(x) = \exp\Big(\tfrac{2\pi i}{m} f_\xi(x)\Big).$$

Irreducible representation (irrep) A group representation that admits no nontrivial invariant subspace. For finite abelian groups, all irreps are one-dimensional and correspond to characters.

Dual group $\widehat{G}$ The set of all characters of a finite abelian group G, with pointwise multiplication. For $G = (\mathbb{Z}_m)^n$, the dual group is isomorphic to G itself.

Period of a character The number of cosets of $\ker \chi$ in G; equivalently, the index $[G : \ker \chi]$.

Torus $\mathbb{T}^n$ The topological space $\mathbb{R}^n/\mathbb{Z}^n$, equivalently the product of n circles. Discretely, $(\mathbb{Z}_m)^n$ can be viewed as an n-dimensional grid with side length m and opposite faces identified, i.e. a discrete n-torus.

Hypercube The Cayley graph of $(\mathbb{Z}_2)^n$, with vertices the binary vectors of length n and edges between vectors differing in exactly one coordinate.

Cube complex A cell complex formed by gluing together cubes (in various dimensions) along their faces. The Cayley graph of $(\mathbb{Z}_m)^n$ naturally forms a cube complex, with m-cycles in each coordinate direction.

Affine slice A coset of a subgroup defined by a linear congruence. For example, $f_\xi(x) \equiv c \pmod m$ defines a slice perpendicular to ξ.

Hamming metric For $m = 2$, the distance between two vectors in $(\mathbb{Z}_2)^n$ given by the number of coordinates where they differ.

Lee metric For $m > 2$, the cyclic distance on $\mathbb{Z}_m$: $d_{\text{Lee}}(a, b) = \min(|a - b|, m - |a - b|)$, extended coordinate-wise to $(\mathbb{Z}_m)^n$.

Functional Lee metric The Lee metric on the image of f_ξ, pulled back to $(\mathbb{Z}_m)^n$ by defining $d_\xi(x, y) = d_{\text{Lee}}(f_\xi(x), f_\xi(y))$.

Slope (α, β) Coefficients specifying how two group elements $(a, b) \in G \times G$ combine in the functional $\ell_{\xi;\alpha,\beta}(a, b) = \alpha f_\xi(a) + \beta f_\xi(b)$. Vertical/horizontal slopes act on one argument, diagonal slopes couple both.

Chinese Remainder Theorem (CRT) The statement that if $m = \prod_i p_i^{k_i}$, then $\mathbb{Z}_m \cong \prod_i \mathbb{Z}_{p_i^{k_i}}$, so congruences modulo m decompose into simultaneous congruences modulo prime powers.

Multidimensional CRT (mCRT) Generalization to $(\mathbb{Z}_m)^n$: specifying cosets with respect to a family of independent functionals determines a unique element in G.

Bézout's identity If $a_1, \ldots, a_n, m \in \mathbb{Z}$ and $d = \gcd(a_1, \ldots, a_n, m)$, then there exist integers $u_1, \ldots, u_n, w$ such that $\sum_i u_i a_i + wm = d$.

B Bézout's Identity and Surjectivity of Additive Functionals

We recall a standard number-theoretic fact that guarantees the surjectivity of additive functionals f_ξ onto their image subgroups.

Let $a_1, \ldots, a_n, m \in \mathbb{Z}$ and let $d = \gcd(a_1, \ldots, a_n, m)$. Then there exist integers $u_1, \ldots, u_n, w \in \mathbb{Z}$ such that

$$u_1 a_1 + \cdots + u_n a_n + wm \;=\; d.$$

Applying this to the coefficients $(\xi_1, \ldots, \xi_n)$ defining an additive functional $f_\xi(x) = \sum_i \xi_i x_i \pmod m$ on $G = (\mathbb{Z}_m)^n$, we obtain integers $u_1, \ldots, u_n, w$ such that $\sum_i u_i \xi_i + wm = d$. Hence

$$f_\xi(u_1, \ldots, u_n) \equiv d \pmod m.$$

By scaling, $f_\xi(tu_1, \ldots, tu_n) \equiv td \pmod m$ for every $t \in \mathbb{Z}$. Thus every multiple of d in $\mathbb{Z}_m$ is attained, so

$$\operatorname{Im}(f_\xi) \;=\; d\mathbb{Z}_m.$$

This shows that the image of an additive functional is exactly the subgroup of $\mathbb{Z}_m$ consisting of multiples of $d = \gcd(\xi_1, \ldots, \xi_n, m)$, justifying the classification of additive functionals in Section 3.1.

C Directional Discrete Fourier Transform

The directional DFT analyzes a neuron's preactivation $N(a, b)$ over $G \times G = (\mathbb{Z}_m)^n \times (\mathbb{Z}_m)^n$ by restricting to one-dimensional fibers cut out by the additive functionals already defined. Fix a pair of elements $\xi_1, \xi_2 \in (\mathbb{Z}_m)^n$ and write its canonical representative ξ with effective modulus $m_\xi = m/\gcd(\xi_1, \ldots, \xi_n, m)$; the associated functional is $f_\xi(x) = \langle \xi, x \rangle \pmod{m_\xi}$ and its character is $\chi_\xi(x) = \exp\big(2\pi i\, f_\xi(x)/m_\xi\big)$. For a *slope* $(\alpha, \beta) \in \mathbb{Z}_m^2$ (axis cases $(1, 0)$, $(0, 1)$; "cross" cases otherwise), we collapse the 2D grid along the coset fibers of

$$t = \ell_{\xi_1,\xi_2;\alpha,\beta}(a, b) \;:=\; \alpha \langle \xi_1, a \rangle + \beta \langle \xi_2, b \rangle \pmod L,$$

where $L = \mathrm{lcm}(m_a, m_b)$ with m_a, m_b the effective moduli of the chosen directions for a and b. Along this one-dimensional coordinate t, we expand $N(a, b)$ in the character basis $\{e^{2\pi i\, rt/L}\}_{r\in\mathbb{Z}_L}$; the resulting spectrum reveals which *irrep frequency* dominates that fiber. In the axis marginals (slopes $(1, 0)$ and $(0, 1)$) this reduces to a 1-D DFT of the a-only or b-only contribution using $\chi_\xi(a)$ or $\chi_\xi(b)$; in the cross detector we use both copies and test correlations by varying (α, β). When m is composite, directions are taken *up to units* and slopes factor *per prime* (Chinese Remainder Theorem), with energies combined across components; when m is prime, slopes can be represented projectively as $(1, t)$ plus $(0, 1)$.

Operationally (see the code), phase tables cache $\exp\big(-2\pi i\, c\,\langle \xi, x\rangle/m_u\big)$ to project onto these characters efficiently, pick peak non-DC frequencies, and bucket neurons by whether the winning a and b directions.

D Experimental settings

If not specified, the learning rate = weight decay and the batch size is p^n, and the networks are trained much longer than needed to ensure convergence. Most experiments except those specified below used an 80%, 20% train/test split.

$\mathbb{Z}_7^2$: p=7, n=2, bs=49, nn=512, wd=0.001, epochs=15008, training set size=4704.

$\mathbb{Z}_5^2$: p=5, n=2, bs=25, nn=512, wd=0.005, epochs=25008, training set size=1200.

$\mathbb{Z}_3^2$: p=3, n=3, bs=9, nn=512, wd=0.001, epochs=25008, training set size=2160.

$\mathbb{Z}_3^3$: p=3, n=3, bs=9, nn=512, wd=0.0001, epochs=25008, training set size=2160.