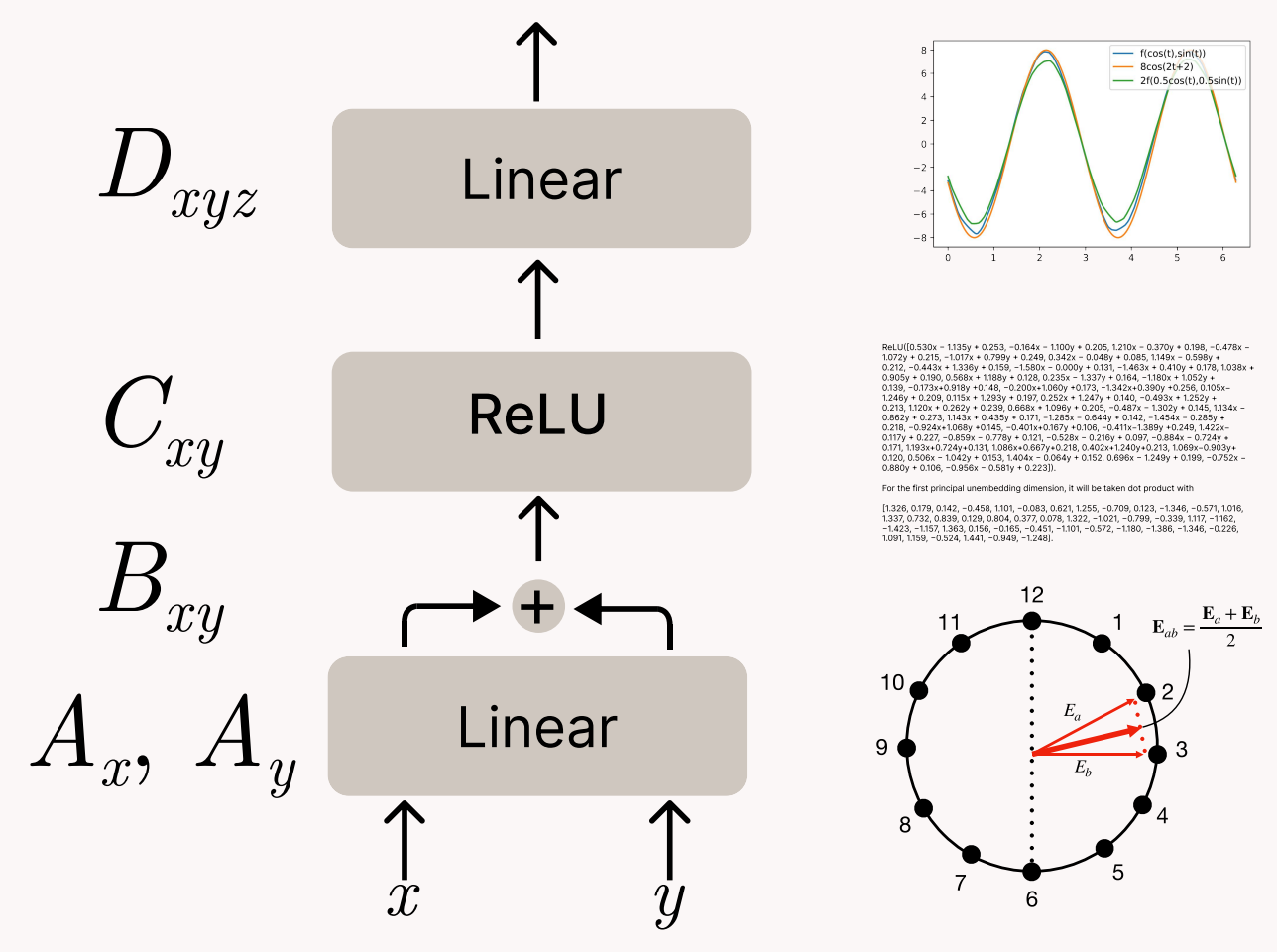
Finite MLPs can be treated as **analytic approximations** of **infinite width MLPs**

ReLU MLPs Can Compute Numerical Integration
Mechanistic Interpretation of a Non-linear Activation

We build upon Nanda 2023 and Zhong 2023 interpretations of the "pizza" modular addition transformer model, which has a **black-box treatment of its MLP**.



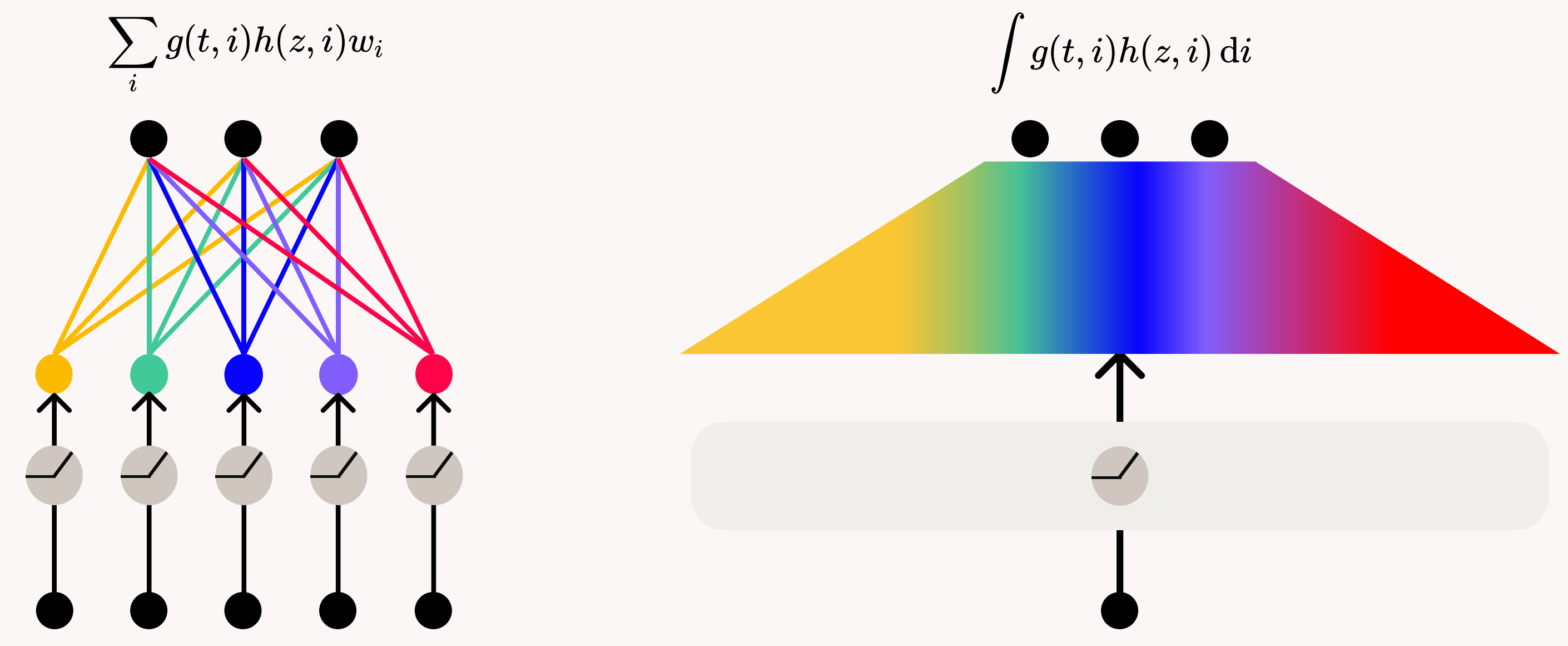
$$\operatorname{argmax}_\phi \cos\phi = 0$$

$$\sin(\alpha)\sin(\beta) + \cos(\alpha)\cos(\beta) = \cos(\alpha - \beta)$$
$$(\alpha = k(x+y); \quad \beta = kz)$$

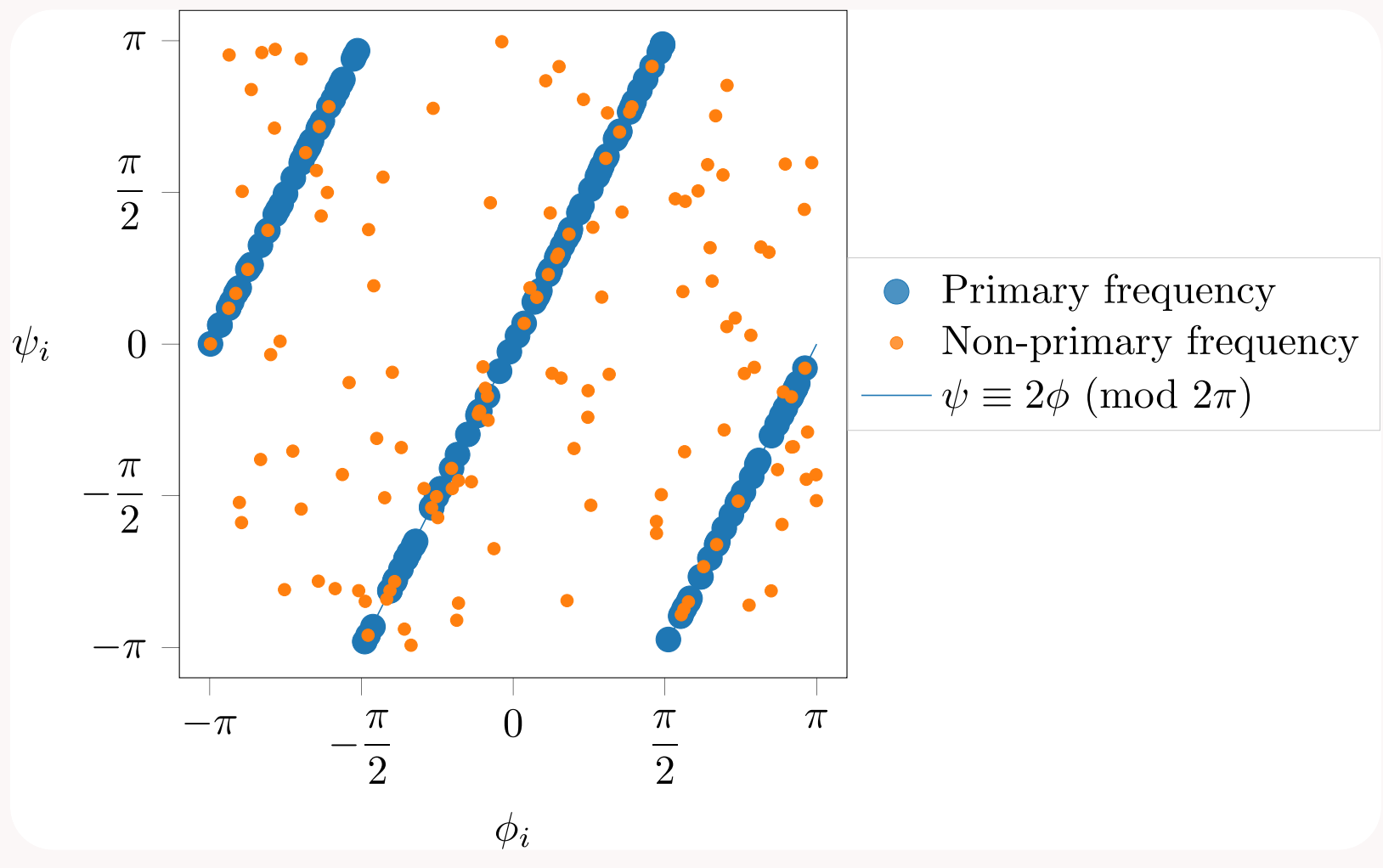
Brute-force

$$\sin(kx) + \sin(ky) = 2\cos(\tfrac{k}{2}(x-y))\sin(\tfrac{k}{2}(x+y))$$
$$\cos(kx) + \cos(ky) = 2\cos(\tfrac{k}{2}(x-y))\cos(\tfrac{k}{2}(x+y))$$

While we cannot compactly describe the behavior of 128 MLP neurons individually, we look for continuous functions capturing the **aggregate input behavior**, treating the finite-width MLP as an approximation of some **infinite-width** counterpart.



We apply **amplitude-phase Fourier transforms** to rewrite each neuron's input and output maps. Neurons are **single frequency** with $k_{\text{in}} = k_{\text{out}}$, output map phases are **2×** the input map phase, and the phases are **uniformly distributed**.

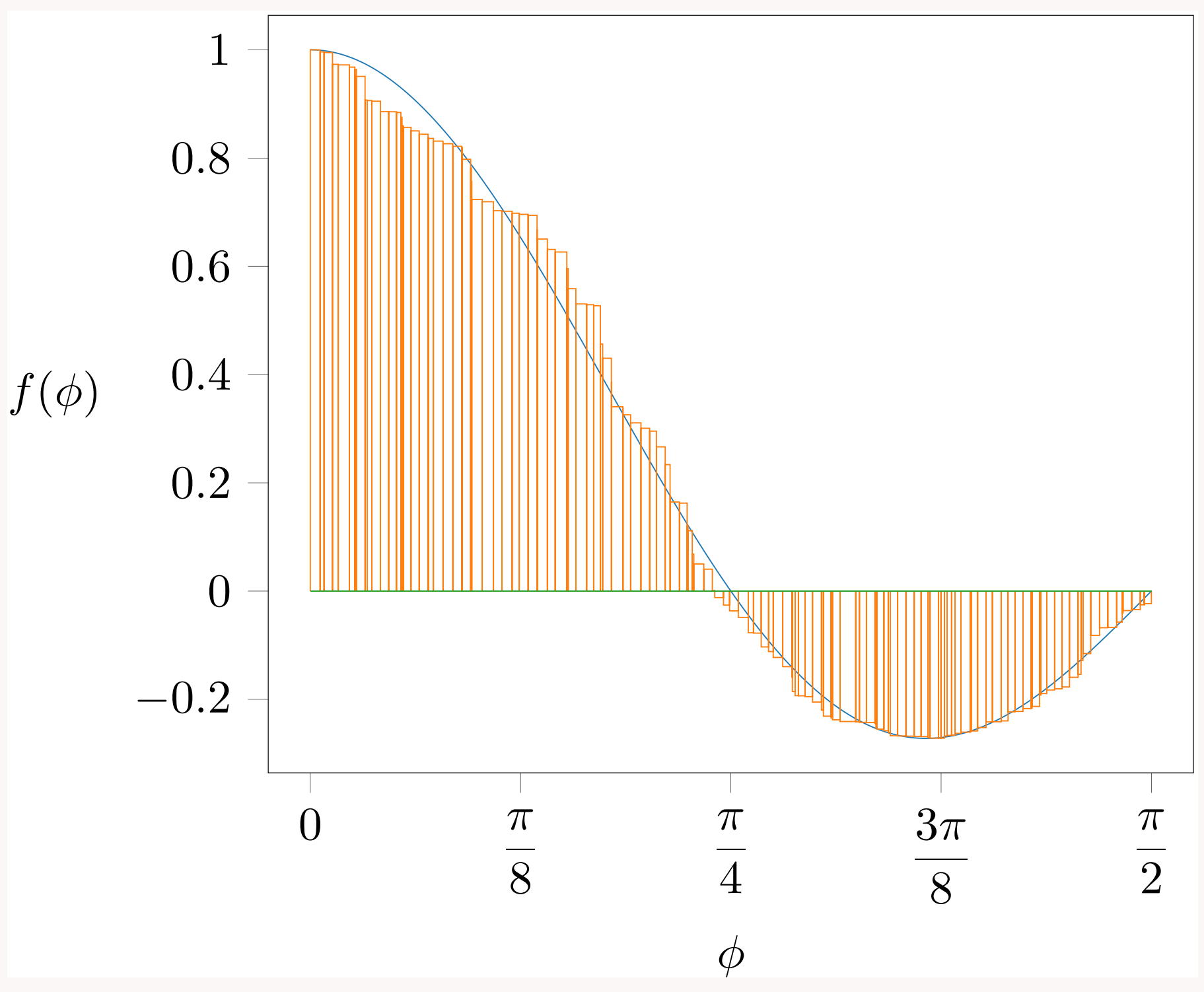


$$(A_x)_i \approx a_i \cos(k_i x + \phi_i)$$
$$(B_{xy})_i = \tfrac{1}{2}(A_x + A_y)$$
$$(C_{xy})_i \approx \mathrm{ReLU}((B_{xy})_i)$$
$$D_{xyz} \approx \textstyle\sum_i (C_{xy})_i \cos(k_i z + \psi_i) d_i$$

$$D_{xyz} \approx \textstyle\sum_i \left|\cos(\tfrac{k_i}{2}(x-y))\right| \underbrace{\mathrm{ReLU}[\cos(\tfrac{k_i}{2}(x+y)+\phi_i)]}_{g(x+y,i)} \underbrace{\cos(z+2\phi_i)}_{h(z,i)} \underbrace{|a_i| d_i}_{w_i}$$

We can sort the neurons by phase and plot one rectangle for each neuron. Given the input $x + y$, the contributions of neurons to the $z = x + y$ logit look remarkably like **numerical integration**.

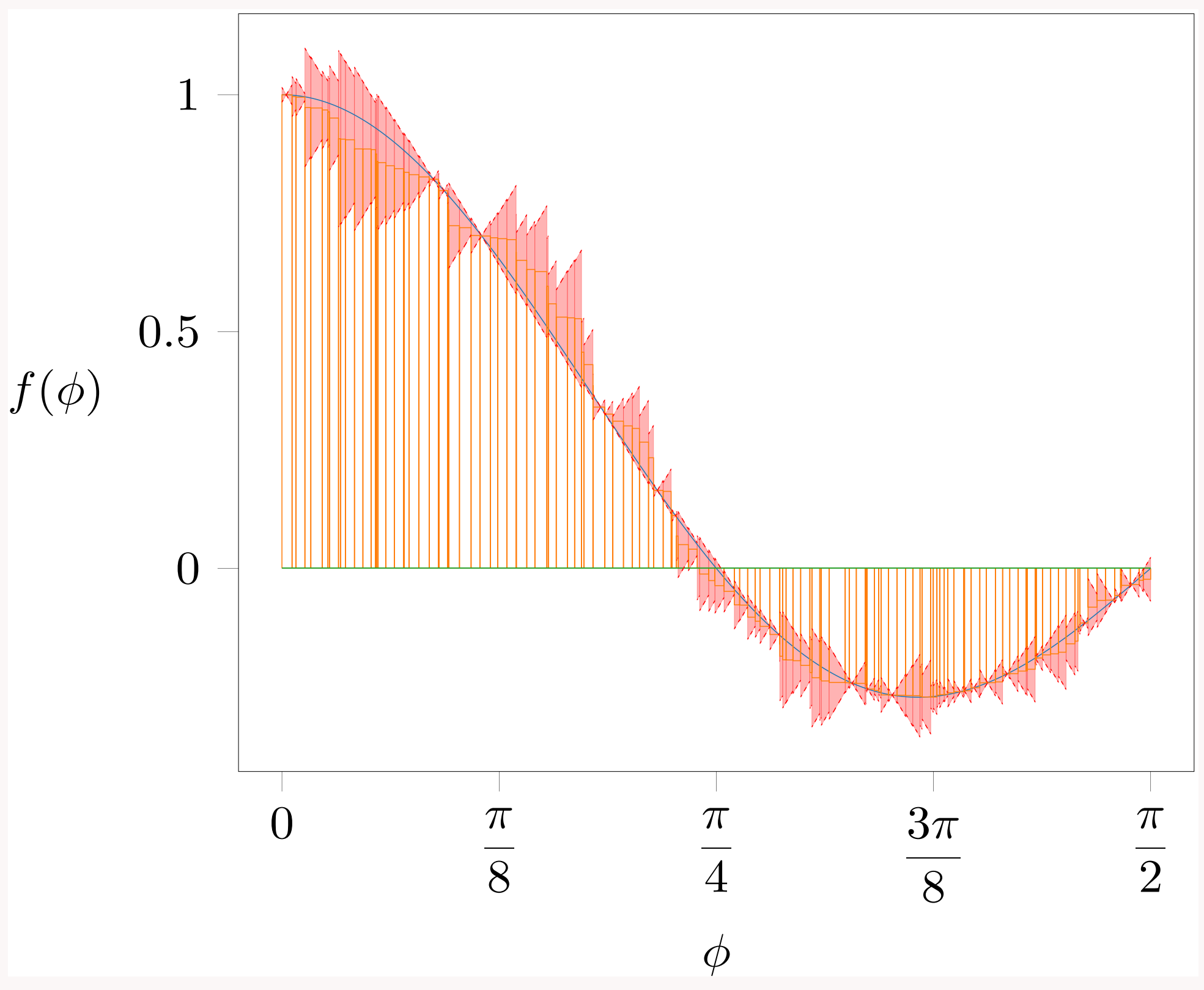
We plot here $x + y = z = 0$.



$$\int_{-\pi}^{\pi} \underbrace{\mathrm{ReLU}(\cos(k(x+y)/2+\phi))\cos(kz+2\phi)}_{f(\phi)}\,\mathrm{d}\phi = \tfrac{2}{3}\cos(k(x+y-z))$$

We confirm this interpretation by using it to **compactly bound error** in the network approximation.

$$\left|\int_{-\pi}^{\pi} f(x) - f(\phi_i)\,\mathrm{d}x\right| \le \int_{-\pi}^{\pi} \underbrace{|f(x) - f(\phi_i)|}_{\le |x-\phi_i| \cdot \underbrace{\sup_x |f'(x)|}_{\le 2}} \mathrm{d}x \le 2\sum_i \left(\int_{a_{i-1}-\phi_i}^{a_i-\phi_i} |x|\,\mathrm{d}x\right)$$



Error Bound Type \ Freq.	12	18	21	22
Normalised abs error	0.04	0.03	0.04	0.03
Normalised id error	0.06	0.05	0.04	0.04
Numerical abs $\int_0^\pi$ bound	0.60	0.41	0.46	0.41
Numerical abs $\int_0^{\pi/2}$ bound	0.45	0.31	0.37	0.30
Naive abs bound	0.74	0.74	0.74	0.74

Chun Hei Yip, Rajashree Agrawal, Jason Gross