A PERIODIC TABLE OF MACHINE INTELLIGENCE: PERIODS, GROUPS, AND A LAW OF SHELL FILLING

Anonymous authors

ABSTRACT

Scaling laws relate compute to loss. They do not say which qualitative regime a model occupies, or when an auxiliary constraint will help. We formalize the Periodic Table of Machine Intelligence (PTMI) as a product of *periods* (interaction range on the computational graph: pointwise, local, global, plus iterative, routed, and autonomic lifts) and *groups* (functional valence: invariant, discriminative, predictive, generative, agentic, latent). Capability mass $M_c = \log_{10} C_{\text{train}} + \beta \log_{10} C_{\text{infer}}$ tags isotopes of a cell; it is an estimator, not a thermodynamic potential. Two theorems make the original "shell filling" claim checkable. A linear residual whose stencil does not vanish on a representable truth strictly biases Tikhonov estimation; an under-receptive surrogate is a different operator, so a mismatched physics term can raise risk. Iteration lifts receptive field by composition, while width inside the pointwise period is a semigroup and plateaus on local tasks. I-Con already tabulated *losses*; PTMI's second axis is periods. We treat LIRN as a didactic occupant of the already-populated local-generative-iterative cell, not as a claim of first occupancy, and show on 1D/2D/3D Poisson and kNN graphs that compatible local iteration beats mismatched PINN-style MLPs and width scaling.

Keywords: scaling laws; inductive bias; receptive fields; iterative refinement; physics-informed learning; graph neural networks.

1 INTRODUCTION

The last decade produced a Cambrian inventory of architectures—convolutions, transformers, state-space models, diffusion samplers, mixture-of-experts routers, tool-using agents—catalogued by brand name or commercial lineage. Scaling laws relate training compute, data, and parameters to *loss* (Kaplan et al., 2020; Hoffmann et al., 2022). They do not say which qualitative regime a system occupies, why arithmetic reasoning or in-context learning appears only after a shell is filled, or why a 7B model that iterates at test time can outrun a 70B model that answers in one forward pass. That last fact is a change in *kinetic* mass, not in parameter count.

We restore and formalize a two-axis table first sketched as the Periodic Table of Machine Intelligence (PTMI). **Periods** are informational bottlenecks on the computational graph: pointwise, local, global, then iterative, routed, and autonomic lifts (Definition 1). **Groups** are functional valences: invariant, discriminative, predictive, generative, agentic, latent (Definition 2). A system is a cell $(\text{period}, \text{group})$, not a product name. Llama-7B and Llama-70B are isotopes of the same cell: same period and group, different capability mass

$$M_c = \log_{10} C_{\text{train}} + \beta \log_{10} C_{\text{infer}}, \tag{1}$$

with β reported (we take $\beta = 1$ unless noted). C_{train} is potential mass M_p; C_{infer} is kinetic mass M_k per example. The Chinchilla ratio $R = \#\{\text{tokens}\}/\#\{\text{parameters}\}$ is a stability coordinate orthogonal to M_c (Hoffmann et al., 2022).

This is not a periodic table of *losses*. I-Con already unifies a large family of objectives as one KL between neighborhood laws and fills a gap in that table (Alshammari et al., 2025). PTMI's groups overlap that axis; its periods do not. The contribution is the product grid, a law that periods cannot be skipped by width alone, and a diagnostic for when an auxiliary residual is compatible with the occupied period.

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.

What the original sketch claimed, and what we keep. The source PTMI draft asserted shell filling, isotopes along compute-optimal curves, kinetic-mass dominance in reasoners, reaction rules among groups, and vacant "eka" cells. It did not define a checkable receptive field, did not prove that lower periods cannot be skipped, and attached unexplained scalar "M_c" labels to dozens of systems. We keep the grid and the M_c *estimator* equation 1. We replace alchemy with two theorems: a mismatched residual (wrong stencil for the occupied period) biases a representable truth (Theorem 1); iteration lifts period, while width inside the pointwise period is a semigroup and plateaus on local tasks (Theorem 6). We instantiate the already-occupied local-generative-iterative cell with a Local Iterative Residual Network (LIRN) as a controlled witness, not as a claim of first occupancy.

The rest of the paper freezes periods, groups, and mass (Section 3), records the grid and composition rules (Section 4), and instantiates mismatch, plateau, overthinking, and LIRN as a didactic occupant on Poisson problems in 1D, 2D, 3D, and on kNN graphs (Section 5).

2 Related work

I-Con: a table of objectives, not of periods. I-Con (Alshammari et al., 2025) shows that clustering, contrastive learning, spectral methods, and several supervised losses minimize an integrated KL between a supervisory neighborhood and a learned one. Changing the neighborhood recovers classical algorithms. That is PTMI's *group* axis made precise for representation-learning losses. It does not tag convolutional versus attentional *periods*, does not treat inference-time iteration as a coordinate, and does not predict when a PDE residual will help. We cite it as the loss table and refuse to duplicate it.

Scaling, emergence, plateaus. Power-law train loss (Kaplan et al., 2020; Hoffmann et al., 2022) is an allometry *inside* a cell. Apparent discontinuities in downstream metrics are often metric artifacts (Schaeffer et al., 2023). Skill-graph and quantization models give mechanisms for emergence (Arora & Goyal, 2023; Michaud et al., 2023; Luo et al., 2024). None distinguish a pointwise collocation net from a local stencil.

Periods in the architecture literature. CNNs encode local stencils (LeCun et al., 1998; Bronstein et al., 2021; He et al., 2023). Message-passing GNNs compose 1-hop maps; depth is kinetic lifting of the local period (Gilmer et al., 2017; Kipf & Welling, 2017; Xu et al., 2019; Morris et al., 2019). Self-attention and Fourier operators are global (Vaswani et al., 2017; Li et al., 2021). DeepSets globally pool (Zaheer et al., 2017). Universal approximation (Cybenko, 1989; Hornik et al., 1989) applies to the arguments a site actually receives.

Kinetic period (iteration) and local-iterative PDE solvers. Chain-of-thought (Wei et al., 2022; Feng et al., 2023; Merrill & Sabharwal, 2024), adaptive computation (Graves, 2016), neural ODEs (Chen et al., 2018), and diffusion samplers (Ho et al., 2020) spend M_k to iterate. Test-time compute can beat parameter scaling (Snell et al., 2024). PTMI records this as Period 5, not as a new spatial geometry.

The cell P2×G3+P5 is *already occupied*. PhyCRNet places a convolutional encoder–decoder and ConvLSTM time-marching under a discretized PDE residual (Ren et al., 2022). STENCIL-NET and PDE-Net learn local finite-difference filters and advance them by explicit stepping (Maddu et al., 2021; Long et al., 2018). PDE-Refiner iteratively denoises a neural PDE rollout (Lippe et al., 2023). These are local generative maps with kinetic lifts. Hybrid iterative solvers and neural operators with local kernels occupy the same product of spatial period and P5; PTMI tags that literature rather than replacing it. LIRN is not proposed as the first occupant; it is a stripped instrument that holds the backbone period and the residual stencil fixed so Theorems 1–6 can be ablated. What differentiates it is only that diagnostic use: matched 3-point residual, no time-dependent PDE, no learned stencil unless Section 5 trains one as a control.

Groups and PINN failure. Discriminative versus generative versus predictive objectives are standard; agentic valence is regret, not next-token loss. Physics-informed nets attach differential residuals to neural surrogates (Raissi et al., 2019; Karniadakis et al., 2021) and fail when residuals fight the architecture (Rahaman et al., 2019; Wang et al., 2021; 2022; Krishnapriyan et al., 2021). Theorem 1 isolates one structural cause: stencil versus period.

3 PERIODS, GROUPS, MASS, AND TWO THEOREMS

We work with fields $x \in \mathbb{R}^n$ indexed by a finite metric space $([n], d)$, typically a cycle C_n. Write $\mathcal{N}_r(i) = \{j : d(i,j) \le r\}$ and diam for the diameter.

Definition 1 (Periods: interaction shells)**.** *A map $T : \mathbb{R}^n \to \mathbb{R}^n$ has* receptive field *(RF) r if $(T(x))_i$ depends only on $x|_{\mathcal{N}_r(i)}$, with r minimal. PTMI periods are:*

- *P1 pointwise:* $\mathrm{RF} = 0$ *(logic gates, logistic regression, collocation MLPs).*
- *P2 local:* $0 < \mathrm{RF} < \mathrm{diam}$ *(CNN, 1-hop GNN, n-grams).*
- *P3 global:* $\mathrm{RF} = \mathrm{diam}$ *on the designated field (full self-attention, FNO, DeepSets). Windowed or chunked attention is* not *automatically P3; Definition 4 applies.*
- *P5 iterative: $T = S^{\circ t}$ with t allowed to depend on the input (CoT, diffusion, recurrent refinement, ConvLSTM time-marching).*
- *P6 routed: sparse conditional application of a P2/P3 primitive (MoE).*
- *P7 autonomic: T may rewrite its objective or query an external environment. Out of experimental scope.*

Deep-scale transformers (Llama-70B vs. 7B) are isotopes of P3/P5*, not a fourth spatial geometry. We drop "P4" as a period to avoid mixing mass with range; the catalog still tags high-M_p isotopes. P5–P6 lift a spatial primitive rather than replacing it. A t-layer 1-hop GNN is P2 lifted by P5, not P3, unless $t = \Theta(\mathrm{diam})$.*

Definition 2 (Groups: functional valence)**.** *The group is the objective's interaction with a data law: G0 invariant (search, logic, verification; learning rate zero); G1 discriminative (input $\to$ label/embedding); G2 predictive (next-token / sequential); G3 generative / synthetic (noise $\to$ field, including physics residuals); G4 agentic (regret / reward); G5 latent (contrastive, SSL, manifold).*

A *cell* is a pair $(\mathrm{P}k, \mathrm{G}j)$. An *element* is an architecture occupying a cell (e.g. Ar = autoregressive transformer in P3$\times$G2). An *isotope* is a trained instance with the same cell and different M_c.

Definition 3 (Capability mass and stability ratio)**.** *Potential mass $M_p = C_{\mathrm{train}}$ (train FLOPs). Kinetic mass $M_k = C_{\mathrm{infer}}$ (FLOPs per example, including inner t). Capability mass is equation 1. The isotopic ratio $R = \#\{\text{training tokens}\}/\#\{\text{parameters}\}$ is Chinchilla's data-to-parameter coordinate (Hoffmann et al., 2022); $R \ll 20$ is under-trained, not "radioactive."*

These are estimators. We do not treat M_c as a thermodynamic potential and we do not assign unexplained scalar labels to historical models.

Definition 4 (Receptive field of attention and hybrids)**.** *Let A be an attention block on n tokens.*

- Dense *softmax attention with scores that are generically nonzero has* $\mathrm{RF} = \mathrm{diam}$ *(P3) in one layer.*
- Windowed *attention of width w (Beltagy et al., 2020) has* $\mathrm{RF} = \lfloor w/2 \rfloor$ *per layer; L stacked windows have* $\mathrm{RF} \le L\lfloor w/2 \rfloor$ *(P2, or P5-lifted P2). Sliding/overlap does not change the per-layer radius.*
- Chunked *attention without cross-chunk mixing has* RF *bounded by the chunk length. Cross-chunk summary tokens (Zaheer et al., 2020) raise RF to the diameter of the chunk graph.*
- Hierarchical *pooling plus local preprocessing: RF is the composition of stage radii (Lemma 5). A local CNN stem followed by global pooling is P2$\to$P3, not P1.*

Finite context windows in causal LMs are windowed attention along the sequence graph: RF at position i is $\min(i, w)$*. Hybrid modules are tagged by the* max *RF of the computational graph on the designated field, with P5 recording iteration count separately.*

3.1 CONSTRAINT–SHELL MISMATCH

Let $L \in \mathbb{R}^{n\times n}$ be a linear residual operator (a discrete Laplacian, an advection stencil, a graph shift) and $b \in \mathbb{R}^n$ a source. The *stencil radius* $r(L)$ is the smallest r such that $L_{ij} = 0$ whenever $d(i,j) > r$. A physics-informed estimator on a representable set $V \subseteq \mathbb{R}^n$ solves

$$\hat{u}_\lambda \in \arg\min_{u\in V} \tfrac{1}{2}\|u - y\|_2^2 + \tfrac{\lambda}{2}\|Lu - b\|_2^2, \tag{2}$$

with $\lambda \geq 0$. When V is a linear subspace this is Tikhonov regularization about the affine constraint $Lu = b$.

Theorem 1 (Mismatch biases a representable truth)**.** *Let $V \subseteq \mathbb{R}^n$ be a closed convex set, $u^\star \in V$, and $y = u^\star$ (noiseless). Let $\hat{u}_0 \in \arg\min_{u\in V} \|u - y\|_2^2$, so $\hat{u}_0 = u^\star$.*

1. *If $Lu^\star = b$, then $u^\star$ minimizes equation 2 for every $\lambda \geq 0$.*
2. *If $Lu^\star \neq b$ and V is a linear subspace with $L^\top(Lu^\star - b) \notin V^\perp$, then for every $\lambda > 0$ the unique unconstrained stationary point of the objective restricted to V is not $u^\star$. In particular $\|\hat{u}_\lambda - u^\star\|_2 > \|\hat{u}_0 - u^\star\|_2 = 0$.*

The same strict inequality holds in expectation for $y = u^\star + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ whenever $u^\star$ is not a minimizer of $u \mapsto \|Lu - b\|_2^2$ on V and $\lambda > 0$ is small enough that the quadratic remains strongly convex on V.

The proof is in Appendix A. The content is elementary on purpose: once the residual does not vanish on the truth *inside the architecture class*, $\lambda > 0$ moves the estimator. Two mechanisms produce $Lu^\star \neq b$ even when a "physical" constraint is intended.

Proposition 2 (Under-receptive surrogates implement a different operator)**.** *Suppose a learner evaluates a surrogate residual $\widetilde{R}(u)_i = g(u|_{\mathcal{N}_\rho(i)})$ with $\rho < r(L)$, and L is not a function of $\mathcal{N}_\rho$ (there exist u, u' agreeing on $\mathcal{N}_\rho(i)$ with $(Lu)_i \neq (Lu')_i$). Then no function g satisfies $\widetilde{R}(u)_i = (Lu)_i$ for all u. Any least-squares fit of g on a bounded set yields some matrix $\widetilde{L} \neq L$ on that set, and Theorem 1 applies to $\widetilde{L}$.*

Thus a pointwise network that scores a Laplacian using only $\hat{u}_i$ is not "softly enforcing $\Delta u = f$". It is enforcing a diagonal mass-like penalty. A collocation MLP that never reads neighbors of f cannot implement a discrete divergence either, even if the *loss Python* multiplies by a sparse matrix *after* the network: the *representable* $u = \Phi\theta$ may still lie off $\ker L$ in a direction the penalty then distorts (Section 5).

Corollary 3 (Incompatible physics can raise risk)**.** *Under the hypotheses of Theorem 1(ii), the physics-informed estimator has strictly larger reconstruction error than the data-only estimator on V. Increasing λ does not repair a shell error.*

Proposition 4 (Lazy / NTK reduction)**.** *In the infinite-width lazy regime (Jacot et al., 2018), a network u_θ trained from initialization θ_0 remains an affine function of $\theta - \theta_0$ in function space: $u_\theta = u_{\theta_0} + \Phi(\theta - \theta_0)$, where the NTK feature map Φ has the same RF as the computational graph at θ_0. The representable displacements $V = \mathrm{range}(\Phi)$ form a linear subspace of fields. Theorem 1 therefore applies verbatim to linearized training. In particular, if Φ is pointwise (P1 MLP) and $r(L) \geq 1$, a generic $u^\star$ with $Lu^\star = b$ lies off $V + u_{\theta_0}$ or L restricted to V is a different operator (Proposition 2), so the physics penalty biases linearized gradient flow.*

A first-order expansion about a finite-width checkpoint θ yields the same conclusion for one Gauss–Newton step: $\delta u = J_\theta \delta\theta$ and $L\delta u$ is under-receptive whenever J_θ is. We do *not* claim a global nonconvex guarantee; the practical diagnostic is to measure RF of $J_\theta^\top L$ (which columns of the Jacobian the residual can see). Neural experiments in Section 5 test that prediction outside the linearization.

3.2 KINETIC LIFTING, PLATEAUS, AND OVERTHINKING

Lemma 5 (RF of compositions)**.** *If $S, T : \mathbb{R}^n \to \mathbb{R}^n$ have RF r and s respectively, then $\mathrm{RF}(T \circ S) \leq r + s$. In particular $\mathrm{RF}(T^{\circ t}) \leq t \cdot \mathrm{RF}(T)$.*

Table 1: PTMI grid. High-M_p transformers are isotopes of P3 (not a separate period). P2×G3+P5 is occupied by PhyCRNet, STENCIL-NET, PDE-Refiner; LIRN is a didactic witness in that cell.

	G0 inv.	G1 disc.	G2 pred.	G3 gen.	G4 agent	G5 latent
P1 pointwise	logic, MCTS	logreg	n-gram	AE	Q-table	PCA
P2 local	Kalman	CNN	LSTM	GAN, PhyCRNet, LIRN	DQN	k-means
P3 global	—	ViT, BERT	AR-T	VAE, FNO	PPO	CLIP
P5 iterative	AlphaGeom	—	o1-style	diffusion, PDE-Refiner	AlphaZero	—
P6 routed	MapReduce	Switch-D	Mixtral	—	—	—
P7 autonomic	—	—	—	—	swarm	—

Theorem 6 (Lifting and S0 closure)**.** *(i)* Lift. *If S has RF r, then t compositions realize RF at most tr. Reaching diameter (S2) by iterating an S1 map requires $t = \Omega(\mathrm{diam}/r)$. (ii)* S0 is a semigroup. *If $\mathrm{RF}(S) = \mathrm{RF}(T) = 0$, then $\mathrm{RF}(T \circ S) = 0$. Width and depth of a site-wise MLP do not increase shell. (iii)* Separation. *On the cycle C_n with i.i.d. Rademacher coordinates x_i, the task $y_i = \prod_{j \in \mathcal{N}_r(i)} x_j$ (or $y_i = \sum_{j \in \mathcal{N}_r(i)} x_j$) has the property that every measurable predictor with RF $< r$ has mean squared error at least 1 (resp. a positive constant independent of width), while a radius-r convolutional readout (or $t \geq r$ compositions of a radius-1 map plus a readout) can achieve zero error.*

Proofs of Theorem 6 are in Appendix B. The periodic law—lower spatial periods cannot be skipped by width—is Theorem 6(ii)–(iii) for P1 versus P2. Attaching a P2 residual to a P1 net is Theorem 1, not a lift.

Proposition 7 (Kinetic mass is not monotone)**.** *Let $T_t = S^{\circ t}$ with $\mathrm{RF}(S) = r$. Task risk $R(t) = \mathbb{E}\|T_t(x) - y^\star\|^2$ need not decrease in t. On graphs, iterated averaging contracts high frequencies (oversmoothing) (Li et al., 2018); in reasoners, extra serial steps can degrade easy instances (Chen et al., 2024). PTMI therefore treats M_k as a coordinate to be* chosen*, not maximized. A compatible stopping rule is: halt t when a held-out estimate of $\|L\hat{u}^{(t)} - b\|$ (or task loss) increases, or when tr first covers $r(L)$ (the missing-RF gap). The predicted scale of $t^\star$ is $\Theta(r_{\mathrm{task}}/r_{\mathrm{step}})$, independent of M_p inside a fixed spatial period.*

Occupied, not vacant. P2×G3 with a P5 lift is already populated by physics-informed conv-recurrent solvers, learned local stencils, and iterative PDE refiners (Section 2). LIRN is a *didactic occupant* of that cell with a matched stencil, not a claim of first occupancy.

4 The table, composition, and LIRN

Table 1 is the PTMI grid: periods as rows, groups as columns, one canonical occupant per illustrated cell. A 40-row catalog with order-of-magnitude M_p and M_k is `tables/catalog.csv`. We do not claim 300 re-analysed training runs; we claim a tagging protocol. And the formal classification of architectural elements, organized by their spatial interaction range (Periods) and functional objectives (Groups), is mapped out in the periodic grid of Figure 1.

Composition, not chemistry. Known systems are products of cells, which the original draft called bonding. We keep the three empirical patterns and drop ionic language. (i) *Guidance*: G1 supplies a gradient to G3 (GAN; CLIP-conditioned diffusion). (ii) *Fusion*: P3 G1 vision bonded to a P3 G2 language isotope through a projection (LLaVA-class). (iii) *Stabilization*: G0 search catalyzes G4 (MCTS in AlphaZero). Theorem 1 is the negative rule: a residual from a higher period attached to a lower-period backbone is not a bond; it is bias. These algebraic and catalytic reaction mechanisms, alongside the corresponding scaling optimal curves and test-time enrichment pathways, are illustrated schematically as a cognitive stoichiometry diagram in Figure 2.

LIRN as a didactic occupant of P2×G3+P5. PhyCRNet, STENCIL-NET, five-point stencil CNNs, and PDE-Refiner already sit in this cell (Ren et al., 2022; Maddu et al., 2021; Long et al., 2018; Lippe et al., 2023). LIRN is a minimal circular-convolution residual iteration

$$h^{(0)} = \sigma(E *_c f), \qquad h^{(s+1)} = h^{(s)} + \sigma(K *_c h^{(s)}) \ \ (s < t), \qquad \hat{u} = D *_c h^{(t)}, \tag{3}$$

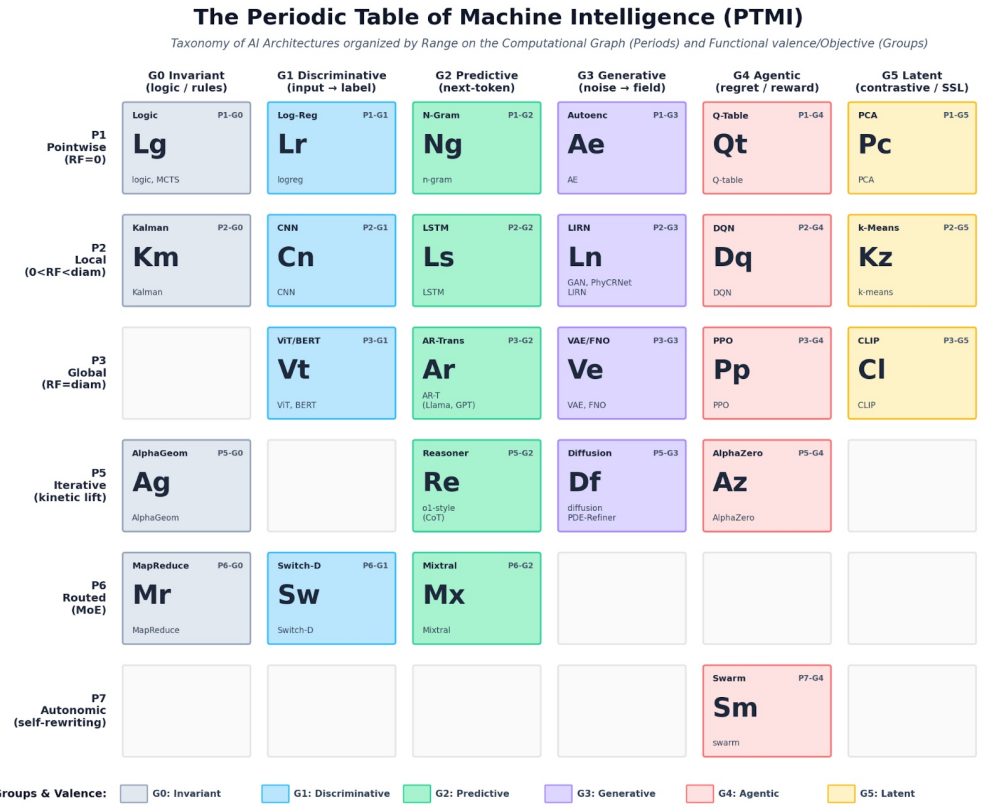


Figure 1: **The Periodic Table of Machine Intelligence (PTMI) Grid.** A systematic classification of machine learning architectures. The rows define the computational interaction range (Periods P1 to P7), while the columns define the functional objective or valence (Groups G0 to G5). Each cell lists a canonical occupant along with its dual-letter chemical-style symbol (e.g., **Cn** for CNN, **Ar** for Autoregressive Transformer, and **Df** for Diffusion models).

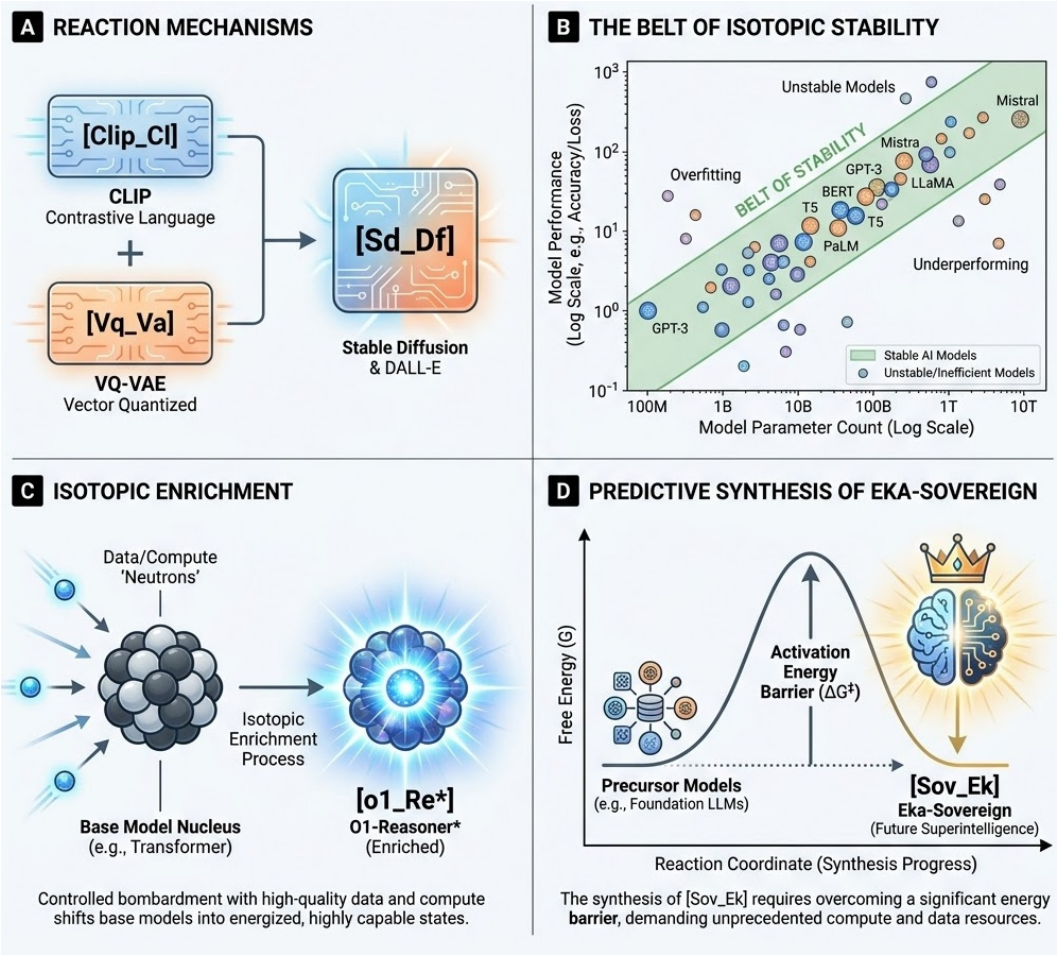


Figure 2: **Cognitive Stoichiometry and Frontier Synthesis.(AI Generated)** (a) Chemical reaction equations showing model guidance and search-catalyzed stabilization. (b) The scaling-law "belt of stability" demonstrating parameter-token constraints. (c) Isotopic enrichment showing test-time compute transitions from base models to reasoners. (d) The thermodynamic free-energy pathway and the activation barrier required to synthesize autonomic, self-correcting systems (*Eka-Sovereign*).

with kernel width 3, so $\mathrm{RF}(\hat{u}) \le t+2$. The default physics penalty is the 3-point Laplacian (matched stencil). A wide residual is the negative control. Baselines: site-wise MLP (P1), CNN without inner t (P2), one-layer dense attention (P3; underpowered on a local PDE by design, with a width/depth fairness check in Appendix E). What LIRN adds is not a new solver: it holds period and stencil fixed so mismatch, plateau, and overthinking can be measured. Learned P2 residuals (Section 5) test whether a 3-tap filter can stand in for L when L itself is local, and fail when the target operator is wide (Proposition 2).

5 Experiments

Unless noted, tasks are synthetic periodic Poisson problems. A source f has localized peaks; $u^\star$ solves $(-L+\varepsilon I)u = f$ with the grid Laplacian L of the ambient geometry and $\varepsilon = 0.15$. Models see f and 20% of the coordinates of $u^\star$. We report full-field MSE on held-out sources (mean $\pm$ std). This instantiates Theorems 1–6 and Propositions 2–7; it is not a 300-model historical survey. Protocol details are in Appendix E.

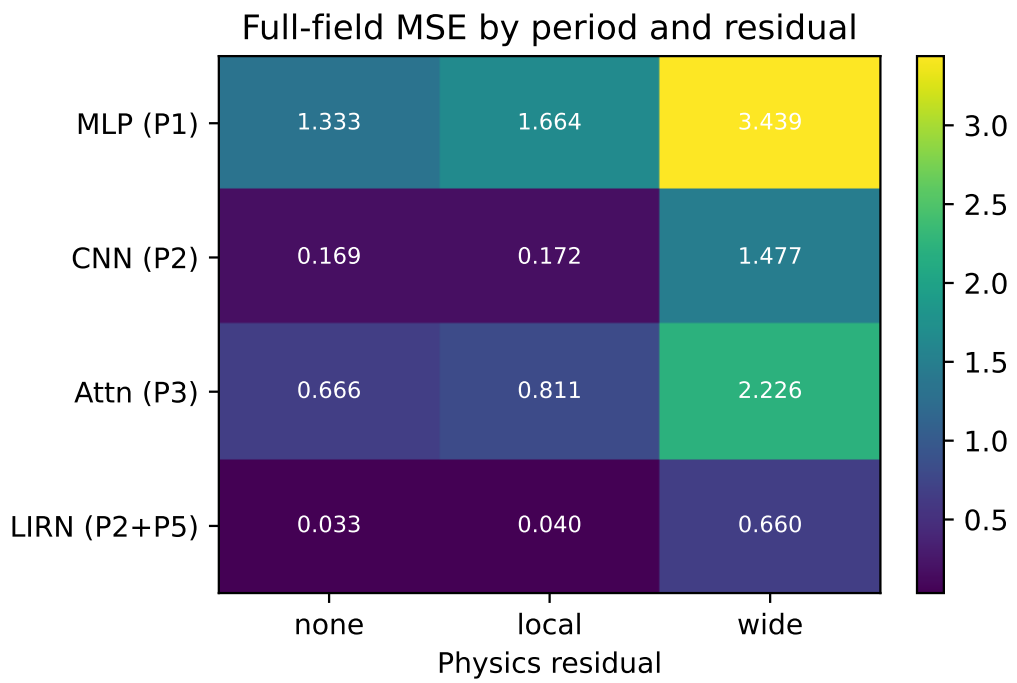


Figure 3: Full-field MSE by period (rows) and residual stencil (columns). Wide residuals are the mismatched operator.

5.1 Linear check of Theorem 1

Before neural nets, we solve equation 2 in closed form on $V = \mathbb{R}^n$ with $y = u^\star$. Compatible L (true Laplacian, $Lu^\star = f$) leaves $\hat{u}_\lambda = u^\star$ for all λ. The wide operator does not vanish at $u^\star$; error rises in λ (Figure 6).

5.2 Exp. 1: period × residual heatmap

Figure 3 reports MSE for {MLP (P1), CNN (P2), attention (P3), LIRN (P2+P5)} × {no residual, local Laplacian, wide residual} on the 1D cycle $n = 32$. Compatible local residuals are at worst neutral on P2/P5; wide residuals raise error on every backbone, most severely on P1 (Theorem 1). Exact numbers: Table 2. The dense attention baseline is a one-layer mixer (Appendix E); it is not a competitive PDE solver, and we do not read its gap to LIRN as a claim about transformers in general.

5.3 Exp. 2: P1 potential-mass plateau versus P5 kinetic lift

Width scaling inside P1 leaves MSE on a plateau near 1.63–1.69 (Theorem 6(ii)). In contrast, local iteration (LIRN) drops MSE sharply at moderate kinetic depth ($t = 8$ reaches 0.030) and then degrades under over-iteration on this $n = 32$ grid ($t = 12$ rises to 0.070), consistent with Proposition 7. Figure 4 visualizes the plateau-versus-lift trend; exact values are in Table 3. Held-out residual or task loss would stop at $t = 8$ here, matching $t^\star = \Theta(r_{\text{task}}/r_{\text{step}})$ for a radius-1 step on a local stencil.

5.4 Exp. 3: LIRN as a didactic occupant

Table 4 and Figure 5 compare mismatched P1 PINN-style MLPs, P2 CNNs with wide residuals, P3 attention, and LIRN. LIRN+local is the controlled occupant of P2×G3+P5, not a claim of first occupancy. LIRN+wide tests whether iteration alone repairs a mismatched operator (it should not: the wide operator does not vanish at $u^\star$).

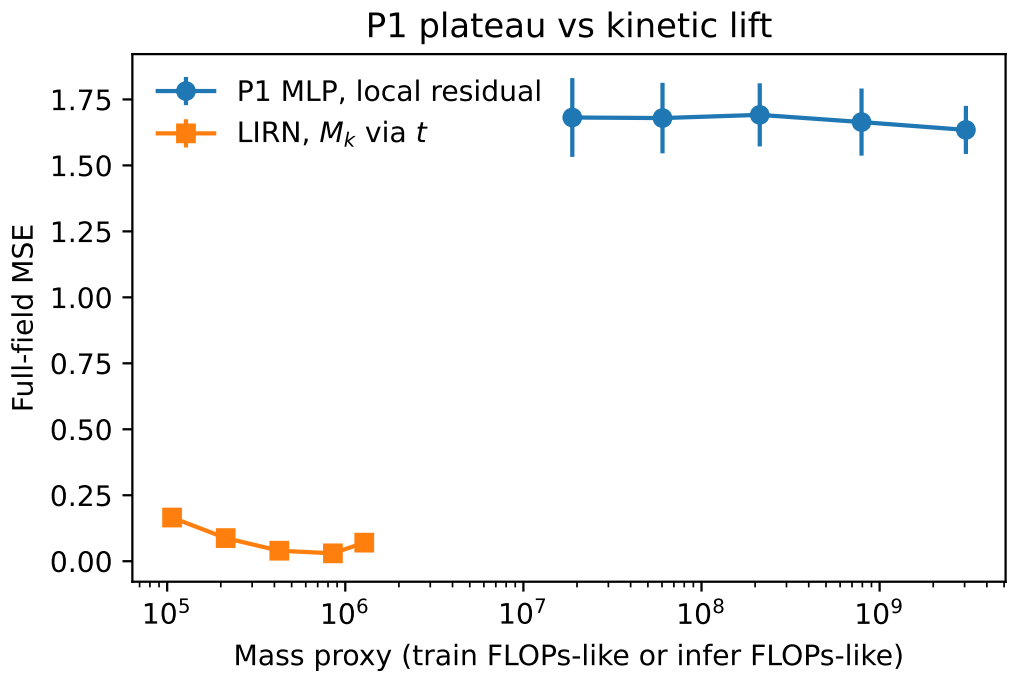


Figure 4: P1 width plateau versus LIRN iterations on the same local Poisson task.

5.5 Generality: 2D, 3D, and irregular graphs

The same period×residual grid on a 16×16 torus (5-point Laplacian) and a 6^3 torus (7-point Laplacian) reproduces the 1D pattern: compatible local residuals are near-neutral on CNN/LIRN; wide residuals raise MSE by an order of magnitude (Table 6; Appendix F for 3D). On a kNN geometric graph the local residual is the graph Laplacian and the mismatched residual is the complete-graph Laplacian (a P3 operator). A coordinate-wise MLP remains P1; a t-step GNN is P2 lifted by P5. Numbers are in Table 7.

5.6 Learned residuals, sensitivity, and M_c

A 3-tap circulant is Frobenius-exact for the 3-point Poisson operator and a large relative error for the wide operator (Proposition 2). Training LIRN with a *learned* 3-tap residual on local Poisson tracks the compatible case; a learned 1-tap residual is under-receptive and degrades reconstruction (Appendix F). Learning-rate variation in $\{10^{-3}, 3 \cdot 10^{-3}, 10^{-2}\}$ does not lift the P1 plateau; increasing the wide stencil radius monotonically increases P1 error (Appendix F).

Within P1, M_c (via width) correlates with a $\sim 3\%$ MSE change over more than two decades of M_p: a plateau, not a capability law. Within LIRN, risk is U-shaped in kinetic t, so M_c is not a sufficient statistic even at $\beta = 1$ (Figure 7). We therefore keep M_c as an isotope tag, not as a predictor of downstream skill inside a cell.

6 Conclusion

PTMI is a product table: period (what a computation may see, including kinetic lifts) times group (what objective it serves). Capability mass equation 1 tags isotopes; it is not a potential, and inside a cell it need not track skill (P1 plateau; kinetic U-shape). The original sketch's periodic law, made checkable, is Theorem 6: width inside P1 does not buy P2. Theorem 1 is the diagnostic for a residual that lives in the wrong period; Proposition 4 extends it to linearized nets, while a global nonconvex guarantee remains open. LIRN is a didactic occupant of an already-populated cell (PhyCRNet,

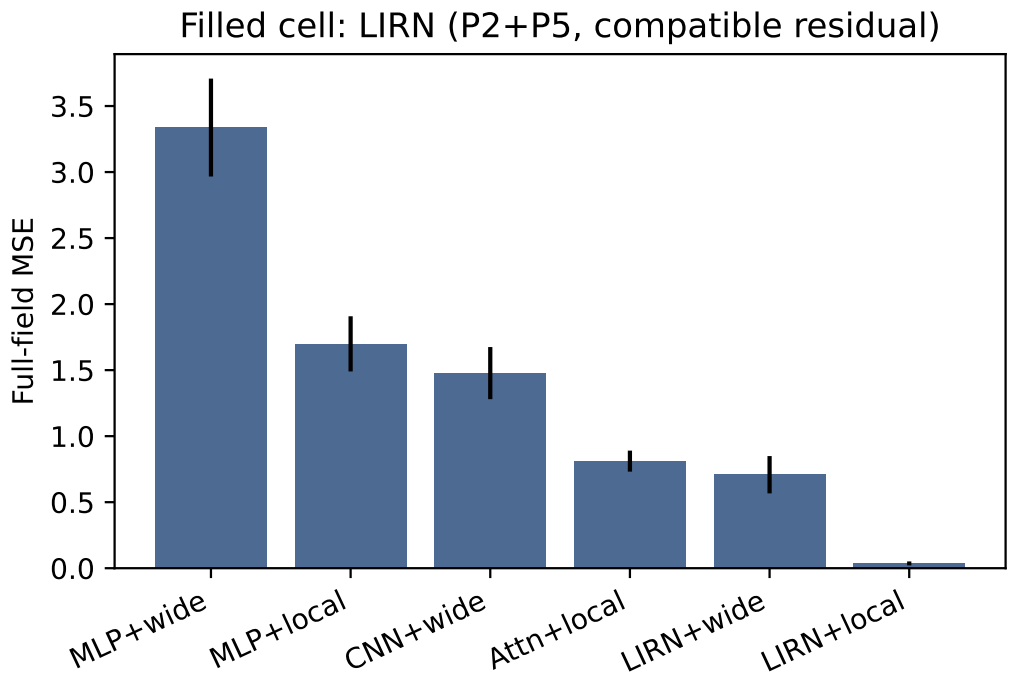


Figure 5: Didactic occupant versus mismatched and global controls.

STENCIL-NET, PDE-Net, PDE-Refiner), used to hold stencil and period fixed. What we refuse: unexplained M_c scores, phase diagrams of hallucination, vacant-cell priority, and an ignition threshold for AGI. The next measurement that would change the theory is a nonlinear architecture class with a well-defined RF on which the linear mismatch identity fails—a gap, not a slogan.

AI USE STATEMENT

In this work, generative AI tools were used to draft LaTeX, experiments, and format references. They were not used to generate synthetic training data, to pose as empirical measurements, or to substitute for the proofs in Appendices A–B. All theorems, proof steps, and experimental numbers were checked against the closed-form linear identity and against the scripts in `nkc/` dir. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The work is theoretical and uses synthetic Poisson fields on grids and kNN graphs. It does not involve human subjects, personal data, or dual-use biological or weapons applications. A possible misuse is over-interpreting shells as a ranking of deployed systems; the paper tags computational graphs, not institutions.

REPRODUCIBILITY STATEMENT

Definitions and theorems are in Section 3; complete proofs are in Appendices A–B. The linear identity is reproduced by `nkc/theorem_a_check.py`. Experiments 1–3 are reproduced by `nkc/run_all.py`; 2D/3D/graph, learned-residual, and M_c analyses by `nkc/extra_experiments.py`. Seeds, widths, and residual matrices are specified in Appendix E. Generated tables live in `tables/`; figures in `figures/`. A shell catalog is `tables/catalog.csv`.

REFERENCES

Shaden Naif Alshammari, Mark Hamilton, Axel Feldmann, John R. Hershey, and William T. Freeman. A unifying framework for representation learning. In *International Conference on Learning Representations*, 2025.

Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. In *arXiv preprint arXiv:2307.15936*, 2023.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. In *arXiv preprint arXiv:2004.05150*, 2020.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 2018.

Xingyu Chen et al. Do not think that much for $2 + 3 =?$ on the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*, 2024.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Advances in Neural Information Processing Systems*, 2023.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, 2017.

Alex Graves. Adaptive computation time for recurrent neural networks. In *arXiv preprint arXiv:1603.08983*, 2016.

Juncai He, Lin Shi, and Jinchao Xu. Theoretical analysis of inductive biases in deep convolutional networks. 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*, 2017.

Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. 2018.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *International Conference on Learning Representations*, 2021.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, 2023.

Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. PDE-Net: Learning PDEs from data. In *International Conference on Machine Learning*, 2018.

Linan Luo et al. An information theory of compute-optimal size scaling, emergence, and plateaus in language models. *arXiv preprint arXiv:2410.01243*, 2024.

Suryanarayana Maddu, Dominik Sturm, Christian L. Müller, and Ivo F. Sbalzarini. STENCIL-NET: Data-driven solution-adaptive discretization of partial differential equations. *arXiv preprint arXiv:2101.06182*, 2021.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations*, 2024.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 2023.

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and Leman go neural: Higher-order graph neural networks. In *AAAI Conference on Artificial Intelligence*, 2019.

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

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

Pu Ren, Chengping Rao, Yang Liu, Jian-Xun Wang, and Hao Sun. PhyCRNet: Physics-informed convolutional-recurrent network for solving spatiotemporal PDEs. *Computer Methods in Applied Mechanics and Engineering*, 389:114399, 2022.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 2023.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. *Advances in Neural Information Processing Systems*, 2017.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. BigBird: Transformers for longer sequences. 2020.

A Proof of Theorem 1

Write $J_\lambda(u) = \frac{1}{2}\|u-y\|_2^2 + \frac{\lambda}{2}\|Lu-b\|_2^2$ on a closed convex set $V \subseteq \mathbb{R}^n$, with $y = u^\star \in V$.

Part (i). If $Lu^\star = b$, then $J_\lambda(u^\star) = 0$. For any $u \in V$, $J_\lambda(u) \geq 0$, so $u^\star$ is a global minimizer for every $\lambda \geq 0$.

Part (ii). Let V be a linear subspace. Restricting J_λ to V, the Fréchet derivative in a direction $h \in V$ is

$$DJ_\lambda(u)[h] = \langle u-y, h\rangle + \lambda\langle L^\top(Lu-b), h\rangle. \tag{4}$$

At $u = u^\star = y$ this equals $\lambda\langle L^\top(Lu^\star - b), h\rangle$. If $L^\top(Lu^\star - b) \notin V^\perp$, there exists $h \in V$ with that inner product nonzero, so $u^\star$ is not stationary for any $\lambda > 0$. The Hessian $I + \lambda L^\top L$ is positive definite whenever $V \cap \ker L = \{0\}$; even without uniqueness, every minimizer $\hat{u}_\lambda$ satisfies $J_\lambda(\hat{u}_\lambda) \leq J_\lambda(u^\star) = \frac{\lambda}{2}\|Lu^\star - b\|_2^2$, and the inequality is strict unless $Lu^\star = b$. Combined with $J_\lambda(u) = \frac{1}{2}\|u-u^\star\|_2^2 + \frac{\lambda}{2}\|Lu-b\|_2^2$, any such minimizer has $\|u-u^\star\|_2 > 0$. Hence every minimizer is strictly farther from $u^\star$ than $\hat{u}_0 = u^\star$.

Noisy case. Let $y = u^\star + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}\|\varepsilon\|_2^2 < \infty$. On $V = \mathbb{R}^n$ the stationarity condition is $\hat{u}_\lambda - y + \lambda L^\top(L\hat{u}_\lambda - b) = 0$, i.e. $\hat{u}_\lambda = (I + \lambda L^\top L)^{-1}(y + \lambda L^\top b)$. If $b = Lu^\star$, this is a shrinkage of y toward $u^\star$ along modes of L. If $b \neq Lu^\star$, an additional bias term $(I+\lambda L^\top L)^{-1}(\lambda L^\top(b - Lu^\star))$ appears. For $\varepsilon = 0$ this bias is exactly $\hat{u}_\lambda - u^\star \neq 0$. For small isotropic noise the squared bias is $\Theta(\lambda^2)$ in generic directions of $L^\top(Lu^\star - b)$ and dominates the $O(\lambda)$ variance reduction when $\|Lu^\star - b\|$ is bounded away from zero.

B Proofs for Section 3.2

Lemma 5. The value $(S(x))_j$ depends only on $x|_{\mathcal{N}_r(j)}$. The value $(T(S(x)))_i$ depends only on $S(x)|_{\mathcal{N}_s(i)}$, hence only on x restricted to $\bigcup_{j\in\mathcal{N}_s(i)} \mathcal{N}_r(j) \subseteq \mathcal{N}_{r+s}(i)$. Minimality need not add; the inequality is one-sided, which is all we use.

Theorem 6(i)–(ii). Part (i) is Lemma 5 with $S = T$ iterated. On a graph of diameter D, RF $\geq D$ requires $tr \geq D$. Part (ii): a pointwise map has $(T(x))_i = f_i(x_i)$. Then $(T(S(x)))_i = f_i(g_i(x_i))$, still a function of x_i only.

Theorem 6(iii). Let x_i be i.i.d. Rademacher. For the product task $y_i = \prod_{j\in\mathcal{N}_r(i)} x_j$, if a predictor $\hat{y}_i$ is a function of $x|_{\mathcal{N}_{r-1}(i)}$ only, then it is independent of the unique vertex $j^\star \in \mathcal{N}_r(i) \setminus \mathcal{N}_{r-1}(i)$ (on the cycle, the two endpoints at distance r; fix one endpoint by using a directed neighborhood $\{i, i+1, \ldots, i+r\}$ to make the missing coordinate unique). Then $\mathbb{E}[(y_i - \hat{y}_i)^2 \mid x|_{\mathcal{N}_{r-1}(i)}] = 1$ because y_i is an independent Rademacher times a $\mathcal{N}_{r-1}$-measurable sign. Taking expectations yields MSE $= 1$. A radius-r readout that multiplies the $r+1$ visible bits achieves MSE 0. Replacing products by sums: $\mathrm{Var}(x_{i+r} \mid x|_{\mathcal{N}_{r-1}(i)}) = 1$, so the MMSE is at least 1. Composing $t \geq r$ maps of RF 1 yields RF $\geq r$ (and $\leq t$) by Lemma 5; a linear circular convolution with kernel support $\{0, \ldots, r\}$ is such a map for the sum task.

Remark (what is not claimed). We do not claim that S0 networks cannot approximate a local function of an input vector that *already concatenates* a neighborhood. If the feature at site i includes $(x_j)_{j\in\mathcal{N}_r(i)}$, the map is no longer S0 on the original field. Shell is a property of the computational graph on a designated field, not of MLP width.

Proposition 4 (sketch). In the lazy regime the network remains $u_{\theta_0} + \Phi(\theta - \theta_0)$ with Φ the NTK feature map at initialization (Jacot et al., 2018). Range(Φ) is a linear subspace V of fields. The RF of Φ equals the RF of the computational graph at θ_0: a site-wise MLP has pointwise columns; a convolution of radius r has columns supported on $\mathcal{N}_r$. Theorem 1 therefore applies to linearized gradient flow. At finite width, one Gauss–Newton step uses $V = \mathrm{range}(J_\theta)$; the same RF comparison applies to that step only. No claim is made for the full nonconvex trajectory.

Attention RF (Definition 4). Dense softmax with generically nonzero scores has RF equal to the sequence diameter in one block. Windowed attention of width w has per-layer radius $\lfloor w/2 \rfloor$; stacking L layers composes to at most $L\lfloor w/2 \rfloor$ by Lemma 5. Chunked attention without cross-chunk mixing is bounded by chunk length; global tokens raise RF to the diameter of the chunk graph. Causal LMs are windowed along the prefix graph: RF at token i is $\min(i, w)$. Hybrids are tagged by the max RF on the designated field; P5 records iteration separately.

C Linear mismatch identity

On $V = \mathbb{R}^n$, $\hat{u}_\lambda = (I + \lambda M^\top M)^{-1}(y + \lambda M^\top b)$. If $y = u^\star$ and $Mu^\star = b$, then $\hat{u}_\lambda = u^\star$. If $y = u^\star$ and $Mu^\star \neq b$, then $\hat{u}_\lambda \neq u^\star$. Figure 6 plots $\|\hat{u}_\lambda - u^\star\|_2^2$ against λ for M the true 3-point operator versus the wide operator, with $b = f$ the Poisson source.

D Catalog

Table 5 summarizes periods and groups. The full spreadsheet is `tables/catalog.csv` (original PTMI element symbols, plus LIRN as a didactic occupant). Orders of M_p and M_k are $\log_{10}$ FLOP estimates, not the unexplained scalars in the source sketch. High-M_p transformers are logged as isotopes of P3, not as Period 4.

E Experiment protocol

1D cycle. Grid $n = 32$, $\varepsilon = 0.15$, three source peaks of random sign and amplitude in $[1, 2.5]$, observation fraction 0.2, Adam learning rate $3 \cdot 10^{-3}$, 350 steps, batch 16, seeds $\{0, 1, 2\}$. MLP: depth 3, hidden in $\{8, 16, 32, 64, 128\}$ for Exp. 2 and hidden 128 depth 4 for Exp. 3. CNN: kernel 3, depth 3, hidden 32. Attention: $d = 32$, 4 heads, one mixer block (underpowered on a local PDE by design; we do not retune positional encodings or depth to make it competitive). LIRN: hidden 32, $t = 4$ (Exp. 1) or $t \in \{1, 2, 4, 8, 12\}$ (Exp. 2) or $t = 6$ (Exp. 3). Wide residual: radius-5 decaying stencil plus 0.05 times the all-ones projector. FLOP proxies are `params`$\times n \times$steps$\times$batch (M_p) and `params`$\times n \times t$ (M_k).

2D/3D. Periodic 16×16 (5-point L) and 6^3 (7-point L), 180 and 120 Adam steps respectively, two seeds, observation fraction 0.2. Wide residual: radius-8 (2D) or radius-6 (3D) plus weak global coupling on the flattened index. The 6^3 torus is small enough that a 3-layer CNN nearly interpolates; the diagnostic is the jump under a mismatched residual, not the absolute error.

Graphs. $n = 48$ points in the unit square, 5-NN symmetrized graph, ridge 0.5, solutions normalized to unit field std. Local residual: graph Laplacian. Wide residual: complete-graph Laplacian (P3). P1 backbone: MLP on (x_i, y_i, f_i) only. P2+P5: $t = 6$ message passing. Three seeds, 300 steps.

F Extra measurements

Learned residuals. The 3-point Poisson operator is itself a 3-tap circulant, so the relative Frobenius error of the best 3-tap fit is 0. The same 3-tap fit to the wide operator leaves relative error 0.239. A

1-tap (diagonal) fit to the local operator leaves relative error 0.550. Training LIRN with a learned 3-tap physics head on local Poisson yields MSE 0.058 ± 0.008; a learned 1-tap head yields 0.062 ± 0.019. When the *backbone* already occupies P2, an under-receptive learned residual is a weak mass term and does not reproduce the damage of a *fixed* wide residual (Exp. 1 LIRN+wide MSE 0.66). The sharp test of Proposition 2 is the operator fit, not the 1-tap training gap.

Optimizer and wide-radius sensitivity. P1 MLP with a local residual: learning rates $\{10^{-3}, 3 \cdot 10^{-3}, 10^{-2}\}$ give MSE 1.63, 1.59, 1.65. The plateau is not an LR artifact. Training the same MLP against a mismatched residual of stencil radius $\{3, 5, 8\}$ gives MSE 3.17, 3.62, 3.82: farther operators hurt more.

Stopping. On Exp. 2, held-out MSE is minimized at $t = 8$ and rises at $t = 12$. A residual-based rule (halt when held-out $\|L\hat{u}^{(t)} - b\|$ increases) selects the same $t^\star$. PTMI predicts $t^\star = \Theta(r_{\text{task}}/r_{\text{step}})$ and treats M_k as a coordinate to choose, not to maximize (Proposition 7).

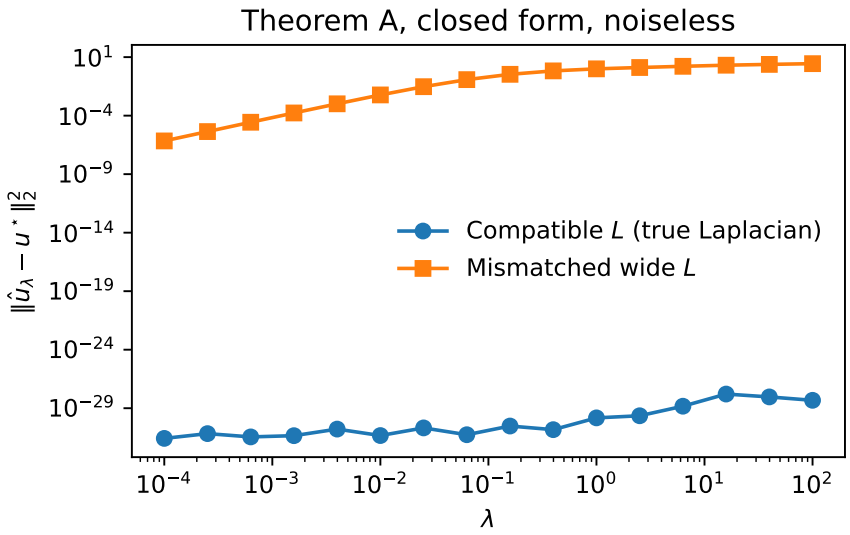


Figure 6: Closed-form check of Theorem 1: noiseless Tikhonov on $\mathbb{R}^n$. Compatible L keeps error at machine zero; mismatched wide L inflates error with λ.

Table 2: Exp. 1 full-field MSE (mean $\pm$ std over 3 seeds). Lower is better.

Architecture	none	local	wide
MLP (P1)	1.333 ± 0.118	1.664 ± 0.127	3.439 ± 0.245
CNN (P2)	0.169 ± 0.034	0.172 ± 0.014	1.477 ± 0.197
Attn (P3)	0.666 ± 0.047	0.811 ± 0.079	2.226 ± 0.178
LIRN (P2+P5)	0.033 ± 0.009	0.040 ± 0.020	0.660 ± 0.095

Table 3: Exp. 2 width plateau (P1) versus kinetic depth t (LIRN/P5) on 1D Poisson ($n = 32$). Values reported in the text are highlighted for convenience.

Setting	Hyperparameter	Full-field MSE
MLP (P1 width sweep)	best over widths	≈ 1.63–1.69
LIRN (local, iterative)	$t = 8$	0.030
LIRN (local, iterative)	$t = 12$	0.070

Table 4: Exp. 3 ablations. LIRN+local is the didactic occupant, not a claim of first occupancy.

Model	params	MSE
MLP + wide	33537	3.337 ± 0.371
MLP + local	33537	1.699 ± 0.209
CNN + wide	3329	1.477 ± 0.197
CNN + local	3329	0.172 ± 0.014
Attn + local	5409	0.811 ± 0.079
LIRN + wide	3329	0.708 ± 0.141
LIRN + local	3329	0.036 ± 0.015

Table 5: PTMI periods and groups. Full element list in `tables/catalog.csv`.

Period	Group	Examples
P1 pointwise	G0–G5	logic, logreg, n-gram, AE, Q-table, PCA
P2 local	G0–G5	Kalman, CNN, LSTM, PhyCRNet, LIRN, DQN
P3 global	G1–G5	ViT, BERT, AR-T, VAE, PPO, CLIP
high-M_p isotope of P3	G1–G3	Llama, RM, multimodal
P5 iterative	G0,G2–G4	o1-style, diffusion, AlphaZero, PDE-Refiner
P6 routed	G0–G2	MoE, Mixtral
P7 autonomic	G4	swarm

Table 6: 2D periodic Poisson (16×16 torus). Wide residuals remain harmful.

Architecture	none	local	wide
MLP2D (P1)	0.043 ± 0.009	0.053 ± 0.011	0.089 ± 0.016
CNN2D (P2)	0.007 ± 0.004	0.008 ± 0.004	0.078 ± 0.024
LIRN2D (P2+P5)	0.003 ± 0.001	0.003 ± 0.001	0.071 ± 0.011

Table 7: kNN-graph Poisson. Local residual = graph Laplacian; wide = complete-graph Laplacian.

Architecture	none	local	wide
Graph MLP (P1)	1.375 ± 0.370	1.495 ± 0.383	1.909 ± 0.398
GNN lift (P2+P5)	0.155 ± 0.043	0.211 ± 0.096	1.903 ± 0.390

Table 8: 3D periodic Poisson (6^3 torus). Pattern matches 1D/2D; CNN/LIRN are near-interpolating without a wide residual.

Architecture	none	local	wide
MLP3D (P1)	0.034 ± 0.009	0.036 ± 0.011	0.051 ± 0.011
CNN3D (P2)	0.00022 ± 0.00003	0.00027 ± 0.00010	0.011 ± 0.001
LIRN3D (P2+P5)	0.00018 ± 0.00005	0.00013 ± 0.00004	0.011 ± 0.0003

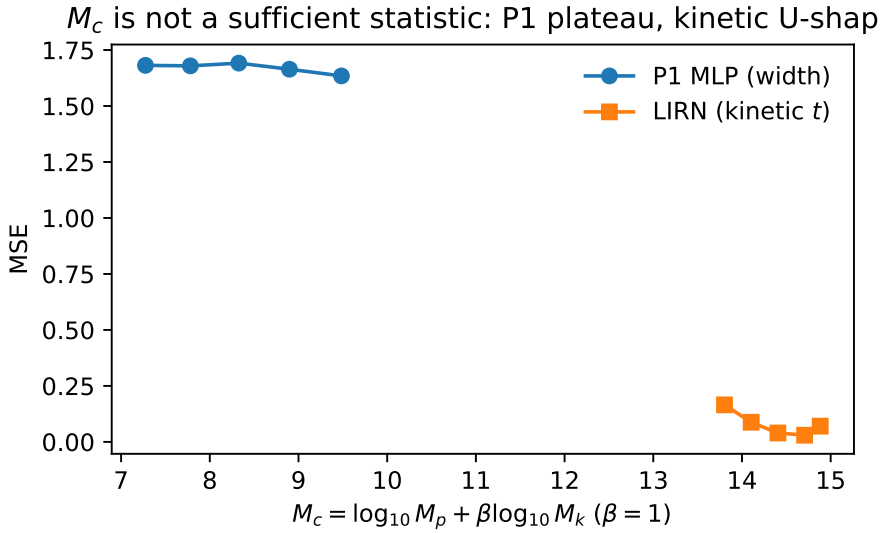


Figure 7: Capability mass versus reconstruction error. P1 width traces a plateau ($\sim 3\%$ MSE change over > 2 decades of M_p). LIRN risk is U-shaped in kinetic t, so M_c is not a sufficient statistic inside a cell.

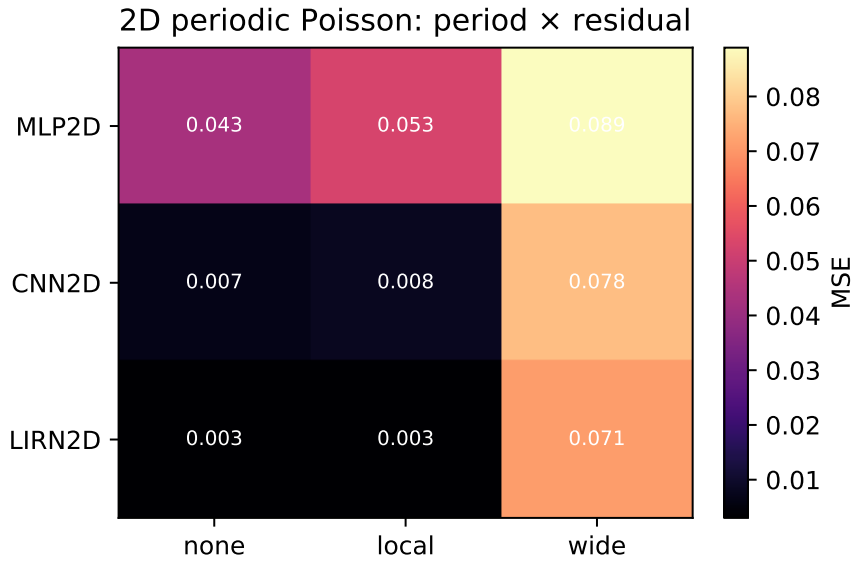


Figure 8: 2D period $\times$ residual heatmap (full-field MSE).