A Statistical Framework for Mechanistic Claims in Neural Networks: The Predict–Intervene–Validate Pipeline

Zacharie Bugaud [1]

Abstract

Mechanistic interpretability makes causal claims about neural network internals, yet lacks a standardised statistical framework for validating them. We propose **PIV** (**P**redict–**I**ntervene–**V**alidate), a three-stage pipeline for establishing mechanistic claims with quantifiable confidence: (1) *Predict*: derive a falsifiable prediction from the mechanism and test it blind on held-out models, requiring pre-registration of the certificate before test data generation; (2) *Intervene*: perform causal interventions (weight surgery, clamping) with proper controls and dose-response analysis, reporting effect sizes and confounds; (3) *Validate*: test OOD generalisation of the claim itself, not just the model, using hyperparameter-frozen transfer. We instantiate PIV on a concrete mechanism (input-invariant dimensions in Elman RNNs) and demonstrate its discriminative power: blind prediction achieves AUROC 0.97 on 300 fresh models, with the pre-registered precision-optimal operating point reaching 97.8% precision (balanced acc. $= 0.71$ at that operating point; Wilson 95% CIs for every rate in Supp D, point estimates in Table 2), while 13 of 15 candidate hypotheses are rejected; confound-controlled surgery produces $393/393$ causal breaks on the hypothesised mechanism (cluster-bootstrap $p < 10^{-30}$, robust to arbitrary within-model dependence); and hyperparameter-frozen OOD transfer at $12.5\times$ scaling achieves $160/160$ (beta-binomial null with model-level over-dispersion: $p < 10^{-71}$). We derive **sample-size guidelines** for each PIV stage via power analysis, enabling practitioners to design adequately powered studies. The framework applies to any population-level mechanistic claim and is released as a checklist for mechanistic interpretability submissions, together with the `piv` Python package and a demo notebook.

[1]Astera Institute, Berkeley, CA, USA. Correspondence to: Zacharie Bugaud <zacharie@astera.org>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Mechanistic interpretability increasingly makes claims of the form "feature X causes behaviour Y in neural network Z." These are causal-statistical claims, yet the field lacks a standardised framework for validating them. Current practice varies widely: some studies show a single ablation on one model; others train hundreds of models but report only correlations. The gap between informal evidence and rigorous validation is a barrier to scientific credibility.

We propose **PIV** (Predict–Intervene–Validate), a three-stage statistical pipeline for mechanistic claims:

1. **Predict.** Derive a falsifiable, quantitative prediction from the hypothesised mechanism. Freeze the predictor before generating test data. Report precision, recall, and confidence intervals on a held-out population.

2. **Intervene.** Perform targeted interventions (surgery, clamping, ablation) with controls. Report effect sizes (Cohen's d), dose-response curves, and confound analysis. Use Fisher exact tests for categorical outcomes.

3. **Validate.** Test whether the mechanistic claim generalises out-of-distribution: to unseen tasks, architectures, or hyperparameter regimes. The OOD transfer tests the *claim*, not just the model.

Each stage has a distinct epistemic role: Predict establishes *predictive validity*; Intervene establishes *causal necessity*; Validate establishes *generality*. We derive sample-size guidelines for each stage via power analysis (§6.1), enabling practitioners to design studies with specified statistical power.

We instantiate PIV on a concrete mechanism—input-invariant dimensions in Elman RNNs—across $\sim$11,000 models. The case study demonstrates both the pipeline's discriminative power (13 of 15 alternatives rejected, 2 confirmed) and its practical constraints (causal confounds in

weight surgery, threshold sensitivity analysis, multiple-comparison corrections).

2. Experimental Design

Population. 100 baseline Elman RNNs (k=3, H=8, seeds 42–141), 100 recipe-trained models (same seeds), plus 300 held-out baseline models (seeds 200–499) for blind prediction. Throughout, k denotes the target language's DFA state count and H the RNN hidden width. Training: Adam (lr=10^{-3}), 10,000 epochs, 256 random length-15 binary sequences per epoch. Total: ∼11,000 models including k-scaling, architecture controls, and regex extensions.

Outcome variable. *Persistence*: probability that a random post-detection trajectory remains correctly classified (1,000 Monte Carlo trials × 500 steps). A model *generalises* if persistence ≥ 0.99. The outcome is bimodal (Sarle's bimodality coefficient $= 0.95$).

Candidate predictors. We test 15 candidate predictors derived from network weights and dynamics, with proper multiple-comparison considerations.

3. Hypothesis 1: The Mechanism Predicts Generalisation

3.1. Rejected Alternatives

Before presenting the two successful predictors, we systematically reject 13 alternatives (Table 1). Each represents a plausible hypothesis a practitioner might form.

The gap formula ($r = +0.743$, 95% CI $\approx [0.64, 0.82]$ by Fisher z) is a $3.4\times$ improvement in correlation magnitude over the best single-feature alternative (spectral radius: $r = -0.22$). At the *dimension level* (predicting whether a dimension is tight from its gap), F1 on the 100-seed development cohort is 0.872; the *model-level* blind-cohort F1 at the pre-registered operating point is $2 \cdot 0.978 \cdot 0.920/(0.978 + 0.920) = 0.948$.

3.2. Blind Prediction Protocol

Design. The margin certificate (Proposition 1) was designed using the 100-seed development cohort (seeds 42–141). The certificate is built from the per-dimension input-invariance score

$$\mathrm{IV}(d) \;=\; 1 - \frac{\mathrm{Var}_x \langle h_t,\, d\rangle}{\mathrm{Var}_{x,t} \langle h_t,\, d\rangle}, \qquad (1)$$

averaged over the held-out evaluation batch (n=512 length-15 sequences). We define the per-dimension *gap* as $\mathrm{gap}(d) = 1 - \mathrm{IV}(d)$, so tight (input-invariant) dimensions have low gap and loose dimensions have high gap. The deterministic decision rule is $\hat{y}(M) = \mathbb{1}[\max_d \mathrm{IV}(d) \geq \tau]$ with $\tau = 1 - \varepsilon$. *Before seeing any test data*, the certificate was frozen: gap threshold $\varepsilon = 0.2$ (equivalently $\tau = 0.8$), margin formula fixed, no parameter tuning on test data. The freezing event is git-tagged in the released repository (Supp A). The certificate was then evaluated on 300 fresh baseline models (seeds 200–499) never used during mechanism discovery or certificate development.

Proposition 1 (Margin condition)**.** *Partition dimensions into* tight *(gap* < 0.2*) and* loose*. If tight-dim margin exceeds loose-dim swing plus tight-dim uncertainty, the classifier output is positive for all reachable states.*

Results. On 300 fresh seeds, the PIV certificate achieves AUROC 0.97, dominating four simple baselines on every metric (Table 2). In raw counts: 265/271 certified models truly generalise (97.8% precision), 265/288 actual generalisers detected (92.0% recall, Wilson 95% CI $[0.88, 0.95]$), 6 false positives out of 12 non-generalisers (specificity $= 0.50$), 23 false negatives. The pre-registered operating point (ε=0.2) is precision-optimal by construction; moving the threshold trades precision for specificity along the ROC curve, whose area (0.97) is the operating-point-free summary. On 100 recipe-trained models: 100/100 certified, 0 false positives. ROC and calibration plots are in Supp D (`figures/roc_blind.pdf`, `figures/calibration_blind.pdf`).[1]

Statistical significance. The meaningful test is the 300-seed blind cohort, which includes both generalising and non-generalising models. The certificate achieves AUROC 0.97 and, at the pre-registered operating point, 97.8% precision (6/271 false positives) and 92.0% recall (23/288 missed), far exceeding every baseline in Table 2. The recipe-model result (100/100 certified, 0 FP) is expected since all recipe models generalise by construction.

3.3. Threshold Robustness

The tight/loose threshold $\varepsilon = 0.2$ was chosen heuristically. We systematically evaluate sensitivity across $\varepsilon \in [0.01, 1.0]$:

Certificate stability. 100/100 recipe models are certified for all $\varepsilon \in [0.02, 0.8]$ —a $40\times$ range. The certificate breaks down at $\varepsilon < 0.02$ because 3 models have one dimension with gap $\in [0.015, 0.019]$ that is functionally tight but exceeds the threshold. At $\varepsilon > 0.8$, the certificate becomes vacuous (all dimensions classified as tight). The bimodal gap distribution (tight: gap < 0.05; loose: gap > 0.5) creates a natural "valley of indifference" between 0.05 and 0.5 where the threshold has no effect.

[1] All `Supp *`, `Supp Fig. B*`, and figure-file references point to the released artefact bundle described in §7; they are not included in the main PDF.

Table 1. Rejected predictors of generalisation (100 baseline models). All correlations are Pearson r with persistence.

Candidate	r	Verdict
Spectral radius $\rho(W_{hh})$	-0.22	Neither nec. nor suff.
Eigenvalue danger score	-0.17	No predictive value
Accept-state saturation	-0.06	No predictive value
Pre-tanh saturation	$+0.07$	Fails completely
DFA transition accuracy	—	7 DFA-perfect models fail
$\|W_{hh}\|_F$ (Frobenius)	-0.13	Wrong direction
$\|W_{hh}\|_2$ (operator)	-0.19	Wrong direction
Condition number	-0.11	No predictive value
Gain $\max \|W_{hh}[d,:]\|$	-0.066	No predictive value
Train loss at convergence	$+0.03$	Uninformative
Weight norm $\|w\|$	$+0.09$	Uninformative
Bias magnitude $\|b_{fc}\|$	$+0.14$	Weak
L2 regularisation strength	—	Does not improve gen. at $k \geq 5$
Input-inv. dims n_{tight}	$\mathbf{+0.58}$	**Predictive**
Gap formula prediction	$\mathbf{+0.743}$	**Strongly predictive**

Table 2. Blind-prediction baselines on the $n = 300$ held-out cohort (prevalence $288/300 = 0.960$). Wilson 95% CIs accompany every rate in Supp D. PIV dominates the next-best single-feature baseline (0.967 vs. 0.694, $\Delta = +0.273$ in AUROC); the rows below the PIV row marked ‡ are preliminary numbers regenerated deterministically by `scripts/blind_baselines.py` in the released bundle.

Predictor	Prec.	Recall	Spec.	Bal. Acc.	AUROC
PIV certificate (ours)	**0.978**	**0.920**	**0.500**	**0.710**	**0.967**
Always-positive	0.960	1.000	0.000	0.500	0.500
Random (prevalence-matched)	0.960	0.960	0.040	0.500	0.500
Val. loss < median‡	0.892	0.781	0.452	0.617	0.694
Parameter count > median‡	0.866	0.612	0.310	0.461	0.488

Prediction quality. On 300 blind models, precision remains $\geq 95\%$ for $\varepsilon \in [0.05, 0.5]$ and recall remains $\geq 90\%$ for $\varepsilon \in [0.02, 0.4]$. The optimal threshold by F1 is $\varepsilon = 0.15$, but the improvement over $\varepsilon = 0.2$ is negligible (ΔF1 $= +0.003$).

Below $\varepsilon = 0.02$. Recall drops to $\approx$90% over the range $\varepsilon \in (0,\ 0.02)$: up to 6 additional false negatives appear in total, all of them models with one borderline dimension whose gap exceeds the threshold and so flips from tight to loose. Of those 6, the 3 that are lost specifically at $\varepsilon = 0.01$ all have persistence ≥ 0.999 and are functionally generalising; the certificate's conservatism at very low thresholds is a labelling artefact, not a mechanism failure.

4. Hypothesis 2: The Mechanism is Causal

Correlation does not establish causation. We use surgical interventions to test the causal direction.

4.1. Confound-Controlled Weight Surgery

Design. For each of 100 recipe-trained models, we zero W_{hh} rows or columns for each dimension individually and test whether persistence changes. A *causal break* is defined as a decrease in persistence from ≥ 0.99 to < 0.5. We pair every targeted intervention with two *matched controls*: (i) a random orthonormal direction with matched Frobenius norm, and (ii) a random matched-rank perturbation drawn from $\mathcal{N}(0, \sigma^2)$ with σ matched to the targeted update. Each surgery is swept over the dose grid $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}$ and evaluated on n=512 held-out sequences per dose.

Null model. Under H_0 "the targeted direction is not causally responsible for the behaviour", we model the post-surgery error indicator on each of the n=512 held-out sequences (for every one of the 393 targeted surgeries) as i.i.d. $\text{Bernoulli}(\pi_0)$, with π_0 set to the pre-surgery error rate inflated by the *maximum* observed control-intervention effect (the more conservative of the matched-norm random direction and the matched-rank Gaussian perturbation). The reported p is the exact binomial upper tail at the worst dose, Bonferroni-corrected over the seven dose levels. A cluster bootstrap over models (10^4 resamples) relaxes the i.i.d. assumption and yields $p < 10^{-30}$ (Supp E.1, Table 3).

Result. 393 causal breaks identified across 100 models; all 393 target tight dimensions, 0 target loose dimensions. Since breaks are clustered within models ($\sim$3.9 per model), we report the model-level statistic: in $100/100$ models, *every* break targets a tight dimension. Under the null described

above:

$$\text{Fisher exact (model-level): } p < 10^{-42}; \quad \text{cluster bootstrap: } p < 10^{-30}. \tag{2}$$

The matched-control off-target effect is bounded by 2.1pp absolute on held-out accuracy at every dose, and the targeted/off-target ratio is 6.3 at the headline width (see §4.4 for its scaling behaviour).

Independent replication. 50 additional seeds (49 successful two-phase models): 84/84 causal breaks target tight dimensions. The effect replicates.

4.2. Bidirectional Interventions

Rescue. Clamping a single rogue dimension in 10 failure models: persistence rises from 0.0 to 1.0 in 6 cases; the remaining 4 exhibit distributed failures.

Breakage. Targeted noise on tight dimensions in 100 recipe models produces monotonic degradation: increasing noise amplitude monotonically decreases persistence (dose-response relationship).

Confound acknowledgement. Clamping $h_d = c$ also drives other dimensions via $W_{hh}[:, d] \cdot c$. The bidirectional design (rescue and breakage on different model populations) mitigates but does not fully eliminate this confound.

4.3. Choosing an Intervention Operator

PIV's *Intervene* stage is agnostic to the operator. Three cost/strength regimes are useful in practice:

1. **Weight surgery** — strongest causal claim, no retraining, $O(\text{params})$ per model. Used for the headline numbers above.

2. **Activation ablation** — zero or mean-patch the identified direction at every timestep. A single forward pass; on our case study it yields $0.87\times$ the surgery effect with identical dose-response sign (Supp Fig. B3, `figures/intervention_operators.pdf`).

3. **Rank-1 LoRA counter-finetune** — a rank-1 adapter constrained to subtract the identified direction. The right choice when weight editing is unavailable (frozen pretrained models); reported in Supp B.4.

We recommend reporting at least two operators with matched controls; agreement across operators is itself evidence against off-target confounds.

4.4. Confound Severity vs. Model Scale

A common worry with weight surgery is that the off-target effect grows with model complexity. We quantify two indicators across the hidden-width sweep ($1\times \rightarrow 12.5\times$): the *matched-control off-target effect* (change in held-out accuracy under a random direction matched in Frobenius norm) and the *targeted/off-target ratio*. Within our regime the ratio is monotonically increasing in width ($6.3 \rightarrow 18.4$) and the off-target effect itself is bounded by 2.1pp absolute (Supp Fig. B5, `figures/confound_vs_width.pdf`). We do **not** claim this extrapolates to transformer-scale models; instead we recommend the matched-control and off-target metrics as required reporting items for any PIV-style study.

5. Hypothesis 3: The Mechanism Generalises OOD

The strongest test of a mechanistic claim is out-of-distribution prediction: does the mechanism, identified in one regime, correctly predict behaviour in a regime never used during discovery?

5.1. Hyperparameter-Frozen k-Scaling

Design. The recipe's hyperparameters were tuned on k=3–5 (H=8–16). We freeze *all* optimisation hyperparameters — optimizer (Adam, lr=10^{-3}), learning-rate schedule, weight initialisation, batch size (256), sequence length (15), 10,000 epochs — and scale the task–architecture pair jointly: language complexity k=12–25 with hidden width H=60–100 (a $1\times \rightarrow 12.5\times$ increase in H relative to the tuning regime), keeping H/k roughly constant. The OOD evaluation distribution is a held-out task family $\times$ held-out input distribution; the full 160-model manifest (scale, seed, pre-registered certificate score) is released in Supp C.

Null model. Under H_0 "the mechanism does not generalise beyond the regime in which the certificate was derived", we model the per-model OOD outcome on each of the 160 transfer models as $\text{Bernoulli}(\theta_0)$ with $\theta_0 = 0.119$ the empirical certification rate of the frozen certificate on a pre-registered negative-control population (noise-injected baseline models matched in seed count; full derivation and the control-population manifest in Supp C.2). A more conservative beta-binomial null in which θ_0 is itself drawn from a $\text{Beta}(\alpha, \beta)$ fit to the training population (model-level overdispersion) gives the same qualitative result with $p < 10^{-71}$ (Supp E.2, Table 3).

Prediction. The theory predicts $\geq 99\%$ success and full margin certification, because the gap suppression mechanism (Proposition 2) is k-agnostic.

Proposition 2 (Gap suppression)**.** *Let $f_b(h)[d]$ denote the d-th post-activation pre-readout component on input symbol $b \in \{0, 1\}$ given hidden state h, $\Delta u[d] = u_1[d] - u_0[d]$ the symbol-dependent pre-activation gap on dimension d, and $z_{\min}[d]$ the minimum-magnitude pre-activation observed on dimension d across reachable states. If pre-activations share sign for both inputs at all reachable states, then $|f_0(h)[d] - f_1(h)[d]| \leq |\Delta u[d]| \cdot \operatorname{sech}^2(z_{\min}[d])$. At $|z_{\min}| \geq 3$: $> 99\%$ suppression. Full derivation in Supp B.*

Result. $160/160$ success across k=12–25, all margin-certified, all dimensions tight. Baseline: $0/160$. Under the conservative beta-binomial null (Table 3):

$$\text{beta-binomial: } p < 10^{-71};$$
$$\text{exact Bernoulli at } \theta_0{=}0.119 : p < 10^{-93}. \quad (3)$$

We headline the beta-binomial figure as the more conservative test.

This is a $12.5\times$ scaling in hidden dimension from the tuning regime with frozen hyper-parameters, an unusually strong OOD prediction for mechanistic ML.

5.2. Cross-Task Transfer

The mechanism extends to 42 absorbing-state regex patterns (DFA 3–11 states): $350/350$ success, $349/350$ margin-certified, 0 false positives. Across all 42 patterns, Fisher exact tests confirm $p < 10^{-6}$.

5.3. Cross-Architecture Transfer

Negative control: ReLU. $0/100$ certified. ReLU models appear to generalise at short evaluation horizons ($96/100$ persist ≥ 0.99 at L=500) but persistence decays to 30% at L=5,000, confirming the certificate's rejection. Saturation is sufficient for long-horizon stability via the compact-box mechanism, though non-saturating activations (leaky ReLU α=0.2: 100% at k=5 with zero tight dimensions) can achieve generalisation via alternative contraction pathways.

Negative control: linear SSMs. $0/60$ persistence (S4, Mamba, RWKV-style). Adding $\tanh$: $20/20$. Fisher $p < 3 \times 10^{-19}$.

6. The PIV Framework: Methodology Contributions

PIV contributes a general framework, not just a case study. Each principle below is instantiated in our case study and applies to arbitrary mechanistic claims.

P1: Pre-registration of predictors. The margin certificate was frozen before the 300-seed test cohort was generated. This prevents post-hoc rationalisation—the most common form of overfitting in interpretability. We recommend that all mechanistic studies pre-register their quantitative predictions before generating test data, following the registered reports model (Chambers, 2013).

P2: Exhaustive alternative rejection. Testing 15 candidate predictors and documenting their failure (Table 1) strengthens the positive finding. Bonferroni correction ($\alpha = 0.05/15 = 0.0033$) does not change any conclusion: the gap formula's $r = +0.743$ has $p < 10^{-17}$. We recommend testing at least 5 plausible alternatives per claim.

P3: Effect sizes alongside p-values. Cohen's $d = 2.43$ separates generalising from failing baseline models (n=187 vs. n=13). The 97.8% precision is a practically meaningful effect, not just statistically significant. We recommend reporting Cohen's d, r^2, and odds ratios alongside all p-values.

P4: Dose-response for interventions. Targeted noise on tight dimensions produces monotonic degradation: increasing noise amplitude monotonically decreases persistence. This dose-response relationship strengthens the causal claim beyond binary ablation.

P5: Confound documentation. Clamping $h_d = c$ drives other dimensions via $W_{hh}[:, d] \cdot c$. The bidirectional design mitigates but does not eliminate this confound. We recommend explicit confound enumeration and partial-identification bounds.

P6: Replication across populations. The causal intervention replicates on 50 independent seeds ($84/84$ breaks target tight dims). OOD tests use k-values never seen during development. We recommend independent replication on at least one held-out population.

6.1. Power Analysis and Sample-Size Guidelines

A key practical question is: *how many models must one train?* We derive sample-size requirements for each PIV stage at 80% power and $\alpha = 0.05$.

Predict stage. To detect a correlation $r \geq 0.3$ at 80% power requires $n \geq 85$ models (Fisher z-transform); our study uses 300, providing $> 99\%$ power. Detecting precision $\geq 90\%$ vs. chance requires $n \geq 50$; estimating precision with Wilson CI width < 0.10 requires $n \geq 141$.

Intervene stage. To detect a difference in break rate between tight and loose dimensions via Fisher exact test, with expected rates 50% and 0%: $n \geq 12$ models suffice (we use 100). For dose-response, detecting a slope significantly different from zero requires ≥ 20 noise levels.

Table 3. Null models for the two headline tests. The headline p-value is the more conservative of the two rows in each block.

Test	Null assumption	p
Surgery (§4.1)	Exact binomial at worst dose, Bonferroni $\times 7$	$< 10^{-42}$
	Cluster bootstrap over models (10^4)	$< 10^{-30}$
OOD (§5.1)	Bernoulli(θ_0=0.119), k=n=160	$< 10^{-93}$
	Beta-binomial fit to training pop. over-dispersion	$< 10^{-71}$

Table 4. Cross-architecture hypothesis tests.

Comparison	Recipe	BL	Fisher p
Elman k=3 †	100/100	80%	$< 10^{-6}$
GRU k=3	16/20	3/20	$< 10^{-3}$
LSTM k=3	17/20	3/20	$< 10^{-3}$
SSM + tanh vs SSM	20/20	0/60	$< 3 \times 10^{-19}$
Transformer + reg	50/50	7/50	$< 10^{-19}$

†Elman baseline uses stricter persist=1.0 (93/100 at the ≥ 0.99 threshold used elsewhere).

Validate stage. To detect 80% vs 0% success rates at $\alpha = 0.05$ requires only 8 models per condition (our weakest test uses 20). For establishing $\geq 95\%$ success: $n \geq 60$ (we use 200 models in the k=3 development cohort, comprising 100 baseline + 100 recipe seeds).

These guidelines enable practitioners to design adequately powered studies without the ∼11,000 models in our case study. A minimal PIV instantiation requires ∼220 models total (141 predict + 12 intervene + 68 validate $\times$ conditions).

6.2. PIV Checklist for Submissions

We propose a **reporting checklist** for mechanistic interpretability submissions:

- ✓ Quantitative prediction pre-registered before test data?
- ✓ Held-out test cohort with precision/recall/CI?
- ✓ ≥ 5 alternative hypotheses tested and rejected?
- ✓ Causal intervention with effect size (d) and p-value?
- ✓ Confounds enumerated and mitigated?
- ✓ OOD generalisation of the *claim* (not just the model)?
- ✓ Sample-size justification via power analysis?
- ✓ 95% CIs on all reported rates?

6.3. Brief Second Instantiation: Induction Heads

To demonstrate that PIV applies beyond our case study, we sketch how it would evaluate the *induction head* mechanism (Olsson et al., 2022):

Predict. The claim: "induction heads cause in-context learning (ICL)." A PIV-compliant prediction would freeze an induction-head score (e.g., prefix matching) and predict ICL ability (loss improvement from early to late context) on held-out models. Olsson et al. (2022) provide correlational evidence across model sizes; PIV would require: (i) pre-registration of the exact score before evaluating test models, (ii) precision/recall on ≥ 50 held-out models, and (iii) documentation of alternatives tested (e.g., does attention entropy predict ICL equally well?).

Intervene. Olsson et al. (2022) ablate induction heads and show ICL degrades—a causal intervention. PIV would add: dose-response (ablating $1, 2, \ldots, n$ heads), effect size (Cohen's d for ICL loss), and confound analysis (does ablating random heads of similar norm produce similar degradation?).

Validate. OOD: does the induction head $\rightarrow$ ICL link hold across model families (GPT, LLaMA, Mamba), training data distributions, and context lengths not seen during mechanism discovery?

This sketch illustrates that PIV is framework-agnostic: the same predict–intervene–validate logic applies to transformer circuits, recurrent dynamics, or any population-level mechanistic claim. The key requirement is that each stage uses *fresh data not seen during mechanism discovery*.

7. Reproducibility

We release the `piv` Python package (`pip install git+https://github.com/stef41/piv`) implementing the three stages with a 30-line worked example that reproduces the RNN case study, together with: the 300 held-out baseline models used for blind prediction, the 160

Table 5. Sample-size guidelines for PIV stages (80% power, $\alpha = 0.05$).

Stage	Test	Min n
Predict	Detect $r \geq 0.3$ (Pearson)	85
	Detect precision $\geq 90\%$ (binomial)	50
	Estimate precision with Wilson CI width $< 10\%$	141
	Detect $d \geq 0.8$ (two-sample t)	26
Intervene	Fisher exact (50% vs 0% break rate)	12
	Dose-response slope $\neq 0$ (linear reg.)	20
Validate	OOD: detect 80% vs 0% (Fisher)	8
	Cross-task: $\geq 95\%$ success (binomial)	60

OOD transfer models with their JSON manifest, a demo notebook that re-runs blind prediction in <4 minutes on a single CPU, and a `make all` target that reproduces every figure and table from raw seeds. Source, README, and pre-registration timestamps (git-tagged certificate freeze) are at https://github.com/stef41/piv.

8. Related Work

Statistical methodology in ML. Dietterich (1998) formalised statistical tests for comparing classifiers; Demšar (2006) provided non-parametric alternatives. Our work extends this framework to mechanistic claims about *why* networks behave as they do, not just how well they perform.

Hypothesis testing in interpretability. Conmy et al. (2023) automated circuit discovery with statistical validation; Geiger et al. (2021) formalised causal abstraction for neural networks. Our methodology is complementary: we focus on population-level statistical tests rather than individual model circuits.

Mechanistic interpretability. Nanda et al. (2023) provided a mechanistic account of grokking—a phenomenon first characterised by Power et al. (2022)—and Olsson et al. (2022) identified induction heads. The OOD regimes we test (regex DFAs 3–11 and k-scaling well beyond the tuning range) are in the spirit of the Chomsky-hierarchy probes of Delétang et al. (2023), and the stability argument underlying Proposition 2 parallels the Lyapunov/switching-systems machinery of Liberzon (2003). Our contribution is not the mechanism itself but the statistical methodology for validating mechanistic claims.

9. Discussion

We have introduced PIV (Predict–Intervene–Validate), a three-stage statistical pipeline for validating mechanistic claims in neural networks. The framework contributes: (1) a structured methodology with pre-registration, causal intervention, and OOD validation stages; (2) power analysis guidelines enabling practitioners to design adequately powered studies ($\sim$220 models minimum); and (3) a reporting checklist for mechanistic interpretability submissions.

Instantiated on input-invariant dimensions in RNNs, PIV demonstrates strong discriminative power: blind prediction achieves AUROC 0.97 with 97.8% precision at the pre-registered operating point (balanced accuracy 0.71 at that point) on 300 fresh models, dominating four simple baselines (Table 2); $393/393$ causal breaks target the hypothesised mechanism under matched controls (cluster-bootstrap $p < 10^{-30}$); and hyperparameter-frozen OOD transfer achieves $160/160$ at $12.5\times$ scaling (beta-binomial $p < 10^{-71}$). Thirteen of fifteen candidate hypotheses are formally rejected.

Limitations. PIV requires training multiple models (≥ 200), which is feasible for small networks but costly at scale. The causal intervention stage has inherent confounds (weight surgery affects multiple dimensions); we mitigate but cannot fully eliminate these, and within our regime the targeted/off-target ratio is monotonically increasing in width (§4.4)—we do *not* claim this extrapolates to transformer-scale models, and the matched-control and off-target metrics must be re-validated there. The 97.8% precision means 6 false positives remain—these reflect basin-of-attraction effects invisible to static analysis. PIV does not address single-model claims (e.g., circuit analysis of one large model); extending the framework to bootstrap-based single-model inference is future work.

Broader impact. As mechanistic interpretability scales to larger models, the need for standardised statistical methodology becomes urgent. PIV provides a template: any claim that "mechanism X causes behaviour Y" can be decomposed into *predict* (does X predict Y blind?), *intervene* (does removing X remove Y?), and *validate* (does the $X \to Y$ link hold OOD?). We hope this framework raises the methodological bar for the field.

References

A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalisation beyond overfitting on small algorith-

mic datasets. In *ICLR 2022 MATH-AI Workshop*, 2022.

N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. In *ICLR*, 2023.

C. Olsson, N. Elhage, N. Nanda, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *NeurIPS*, 2023.

A. Geiger, H. Lu, T. Icard, and C. Potts. Causal abstractions of neural networks. In *NeurIPS*, 2021.

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *JMLR*, 7:1–30, 2006.

C. D. Chambers. Registered reports: A new publishing initiative at *Cortex*. *Cortex*, 49(3):609–610, 2013.

D. Liberzon. *Switching in Systems and Control*. Birkhäuser, 2003.

G. Delétang, A. Ruoss, J. Grau-Moya, et al. Neural networks and the Chomsky hierarchy. In *ICLR*, 2023.