PHYSSAE: MECHANISTIC INTERPRETABILITY OF PHYSICS-INFORMED NEURAL NETWORKS VIA SPARSE AUTOENCODERS

Nandita N. Patil[*,*], **Eshwar R A**[*,†], **and Gajanan V Honnavar**[†,‡]

[*]QuaNad Research Laboratory, PES University, Bengaluru, India
[†]Independent Researcher, QNu Labs Pvt Ltd, Bengaluru, India

ABSTRACT

Physics-Informed Neural Networks (PINNs) embed PDE residuals into neural network training, but their internal representations remain opaque: it is unknown what physical features their hidden layers encode or whether those features have a localised causal role. We present PHYSSAE, a mechanistic interpretability framework that trains overcomplete sparse autoencoders (SAEs) on PINN penultimate-layer activations and evaluates dictionary atoms through *direct causal intervention in the original frozen hidden state*: $h_{\text{cf}} = h - \alpha z_k d_k$, bypassing the SAE decoder entirely. Across six PDE families—heat diffusion, viscous Burgers, Allen–Cahn reaction-diffusion, linear convection (easy and hard regimes), and the nonlinear Schrödinger equation—with 3 PINN seeds and 3 SAE seeds each, we show that (i) SAE atoms align with independently-defined physical observables (max Pearson $|r| = 0.951$, always $\gg$ permutation null), (ii) the causal footprint of top-aligned atom ablation is 1.2–$4.2\times$ more spatially concentrated than PCA or ICA interventions, and (iii) top-aligned atoms outperform matched random controls on causal localization for structured physical concepts (ESF_{80} advantage 0.04–0.44). Two-atom bilateral representations improve concept regression R^2 by $\Delta R^2 = 0.05$–0.15 over single atoms, while random pairs decrease it by up to 0.60. These results demonstrate that PINNs develop sparse, physically structured latent representations that can be identified and causally interrogated post-hoc, opening a path toward interpretability-aware scientific machine learning.

1 INTRODUCTION

Physics-Informed Neural Networks (Raissi et al., 2019) encode the residual of a PDE directly into the training loss via automatic differentiation, enabling mesh-free solution of forward and inverse problems across fluid mechanics (Cai et al., 2021), solid mechanics (Haghighat et al., 2021), and biomedical modelling (Sahli Costabal et al., 2020). Despite widespread adoption, PINNs remain black boxes at the feature level. Their failure modes—spectral bias (Wang et al., 2022), gradient pathologies (Wang et al., 2021a), causality violations (Wang et al., 2024a)—are well-characterised at the loss or gradient level, but no work has asked what a trained PINN actually *represents* in its hidden layers.

Recent mechanistic interpretability work has developed tools for extracting human-readable features from activations: linear probes (Alain & Bengio, 2017), concept activation vectors (Kim et al., 2018), and overcomplete sparse autoencoders (SAEs) (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024). SAEs have now migrated to scientific ML, appearing on graph CFD surrogates (Hu & Liu, 2025) and neural operators (Tolooshams et al., 2025), and concept steering has been applied to physics foundation models (Fear et al., 2025). A direct inspection of PINN weight distributions by Tucny et al. (2025) found that weights "resemble Gaussian random matrices

[*]nanditanpatil114@gmail.com
[†]eshwarra5@gmail.com
[‡]gajanan.v@qnulabs.com

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.

with no evident trace of the physical principles," motivating feature-level rather than weight-level interpretability.

PHYSSAE bridges mechanistic interpretability and scientific ML for residual-loss PINNs. Its core design addresses three challenges unique to the PINN setting: (1) concepts must be defined from *independent* reference solutions to avoid circularity since $u_{\text{PINN}} = W_{\text{out}}h + b$; (2) causal intervention must act on the *original* PINN hidden state rather than an SAE reconstruction; (3) the spatio-temporal domain is continuous, requiring ESF and support-capture metrics that generalise token-level interpretability to function spaces.

Related work. **PINN architectures and failure modes.** Self-adaptive weights (McClenny & Braga-Neto, 2020) and Fourier-feature embeddings (Wang et al., 2021b) address spectral bias; domain decomposition (Jagtap & Karniadakis, 2020) enables parallel training; causal reweighting (Wang et al., 2024a) fixes temporal ordering; KAN-based PINNs (Wang et al., 2024b) offer symbolic readability via spline activations. None of these study internal representations at the feature level. **XAI in scientific ML.** SHAP/LIME/Integrated Gradients have been applied to PINNs as black-box regressors (Lundberg & Lee, 2017), not engaging the network's spatiotemporal structure. SAEs on graph CFD surrogates (Hu & Liu, 2025) and neural operators (Tolooshams et al., 2025) are closest to our work but are applied to data-driven models, not residual-loss PINNs. **Sparse autoencoders.** The SAE design (ReLU codes, L1 sparsity, unit-norm decoder columns) for monosemanticity in LLMs is established (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024; Elhage et al., 2021). We adapt this to the continuous-domain PDE setting.

Contributions.

1. **First SAE for PINN latent space** with an independent-concept evaluation protocol that avoids circular validation.
2. **Direct hidden-state causal intervention**: $h_{\text{cf}} = h - \alpha z_k d_k$, applied to the frozen PINN and compared against the same intervention applied to PCA and ICA components for a controlled comparison.
3. **Comprehensive statistical validation** across 6 PDEs $\times$ 3 PINN seeds $\times$ 3 SAE seeds: negative controls, bilateral analysis, overcompleteness and sparsity ablations, and feature consistency metrics.
4. **Representational collapse analysis** showing that failed PINNs have higher-rank, more diffuse representations while still encoding recognisable physical invariants.

Paper overview. Section 2 details the PHYSSAE methodology. Section 3 describes the PDE benchmark. Section 4 presents all experimental results. Section 5 discusses implications and limitations.

2 METHOD

PHYSSAE operates post-hoc on a trained, frozen PINN without retraining or architectural modification.

2.1 PROBLEM SETTING

Let $\Omega \subset \mathbb{R}^d$ be a spatial domain and $[0, T]$ a time interval. We consider PDEs $\mathcal{F}[u](x, t) = 0$ on $\Omega \times (0, T]$ with initial condition $u(x, 0) = u_0(x)$ and boundary condition $\mathcal{B}[u] = 0$. The PINN approximation is:

$$u_\theta(x, t) = W_{\text{out}}\, h_L(x, t) + b_{\text{out}}, \qquad h_\ell = \sigma(W_\ell h_{\ell-1} + b_\ell), \quad h_0 = [x, t]^\top, \tag{1}$$

where $h \equiv h_L \in \mathbb{R}^{d_h}$ is the *penultimate post-activation*—the object of interpretability. Since the output is linear in h, any change Δh propagates as $\Delta u = W_{\text{out}}\Delta h$, making the readout a natural causal bottleneck. We use $L = 5$, $d_h = 128$, $\tanh$ activations, Xavier initialisation.

2.2 PINN TRAINING

The training loss is:

$$\mathcal{L}(\theta) = w_{\mathrm{BC}}\mathcal{L}_{\mathrm{BC}} + w_{\mathrm{IC}}\mathcal{L}_{\mathrm{IC}} + w_F\mathcal{L}_F, \tag{2}$$

with $w_{\mathrm{BC}} = w_{\mathrm{IC}} = 100$, $w_F = 1$. $\mathcal{L}_F$ is the mean-squared PDE residual at collocation points computed via automatic differentiation; $\mathcal{L}_{\mathrm{IC}}$ and $\mathcal{L}_{\mathrm{BC}}$ enforce initial and boundary conditions respectively. Optimisation: Adam ($\eta = 10^{-3}$, 8000 steps) then L-BFGS refinement (300 steps). Per step: 3000 residual, 150 IC, 150 BC collocation points (sampled uniformly). For periodic BCs we additionally enforce derivative matching at the boundary.

2.3 ACTIVATION EXTRACTION

After training, the frozen PINN is evaluated on a dense grid ($n_x = 200$, $n_t = 100$, $N = 20{,}000$ points). We store the activation matrix $H \in \mathbb{R}^{N \times d_h}$ and z-score each neuron across the grid, storing (μ, σ) for use during intervention. 80% of points (N_{train}) train the SAE; 20% (N_{eval}) are held out for all alignment and intervention metrics.

2.4 INDEPENDENT REFERENCE SOLUTIONS AND CONCEPT FIELDS

All concept fields are computed from *independent* reference solutions, not from the PINN output $u_\theta = W_{\mathrm{out}}h + b$. Using u_θ would create circular alignment since it is a linear function of h. We use:

- **Analytic solutions** for heat (eigenfunction expansion), convection (d'Alembert), and Schrödinger (exact soliton).
- **High-resolution numerical references** ($n_x = 512$ pseudo-spectral RK4) for Burgers and Allen–Cahn. Cached once and reused across all seeds.

For each PDE, we define a panel of physically meaningful concept fields derived from the reference u^*. Representative concepts: **Heat**: $\{|u^*|, |\partial_x u^*|, |\partial_{xx} u^*|, |\partial_t u^*|, u^*\}$. **Burgers**: adds normalised shock indicator $|\partial_x u^*|/\|\partial_x u^*\|_\infty$ and x. **Allen–Cahn**: interface indicator $e^{-u^{*2}/0.05}$, reaction $|u^{*3} - u^*|$. **Convection**: characteristic coordinate $x - \beta t$, x, t. **Schrödinger**: soliton amplitude $|h^*|$, phase gradient $|\partial_x \phi^*|$, amplitude gradients, $\mathrm{Re}(h^*)$, $\mathrm{Im}(h^*)$, x. The full concept table appears in Appendix F.

2.5 SPARSE AUTOENCODER

The SAE maps z-scored activations $\tilde{h}$ to $D = 512$ atoms ($4\times$ overcomplete):

$$z(\tilde{h}) = \mathrm{ReLU}(W_{\mathrm{enc}}\,\tilde{h} + b_{\mathrm{enc}}) \in \mathbb{R}^D, \tag{3}$$

$$\hat{\tilde{h}} = W_{\mathrm{dec}}\,z + b_{\mathrm{dec}}, \tag{4}$$

$$\mathcal{L}_{\mathrm{SAE}} = \mathbb{E}\big[\|\tilde{h} - \hat{\tilde{h}}\|_2^2\big] + \lambda\,\mathbb{E}\big[\|z\|_1\big]. \tag{5}$$

After each gradient step we project decoder columns to unit norm: $d_k \leftarrow d_k/\|d_k\|_2$. This ensures z_k absorbs all scale and the decoder direction $d_k \in \mathbb{R}^{d_h}$ is a well-defined unit-norm vector in hidden-state space. $\lambda = 2\times10^{-2}$, 600 epochs, batch 8192. Zero dead atoms across all 54 SAE runs (6 PDEs $\times$ 3 PINN seeds $\times$ 3 SAE seeds).

Why overcomplete? With $D \le d_h$ the SAE compresses rather than decomposes, forcing atoms to mix physical features. With $D \gg d_h$ the dictionary has enough capacity to specialise each atom to a distinct physical concept—the monosemanticity property.

2.6 CONCEPT ALIGNMENT

For atom k and reference concept C_j, both evaluated on the held-out grid:

$$A_{kj} = \mathrm{corr}(z_k(\mathbf{x}), C_j(\mathbf{x})), \qquad k_j^\star = \arg\max_k |A_{kj}|. \tag{6}$$

We call A_{kj} *observational alignment*: correlation between the atom's spatial firing pattern and the reference physical field. A permutation test (500 shuffled concept fields) establishes the null distribution; we report Z-scores and do not claim causation from alignment alone.

2.7 Direct Causal Intervention in the Original Hidden State

The primary causal test acts *directly* on the frozen PINN's hidden state, bypassing the SAE decoder entirely. Given atom k with unit-norm decoder direction d_k and activation $z_k(\mathbf{x})$, the counterfactual hidden state is:

$$h_{\mathrm{cf}}(\mathbf{x};\alpha) = h(\mathbf{x}) - \alpha\, z_k(\mathbf{x})\, d_k. \tag{7}$$

The counterfactual output and causal effect are:

$$u_{\mathrm{cf}} = W_{\mathrm{out}} h_{\mathrm{cf}} + b_{\mathrm{out}}, \qquad \Delta u(\mathbf{x};\alpha) = u_\theta(\mathbf{x}) - u_{\mathrm{cf}}(\mathbf{x};\alpha) = \alpha\, z_k(\mathbf{x})\, (W_{\mathrm{out}}\, d_k). \tag{8}$$

Equation equation 8 shows that Δu is linear in αz_k and the causal footprint is determined entirely by $W_{\mathrm{out}} d_k$—the atom's projection onto the output. The SAE plays *no role* in this formula.

Why this is stronger than SAE-internal ablation. Setting $z_k \to 0$ and decoding through the SAE does not test the atom's causal role in the original PINN: it tests its role in the SAE's *lossy reconstruction* of h. Our intervention uses the exact decomposition $\Delta h = z_k d_k$ and intervenes at the level of the actual PINN computation.

Dose-response and metrics. We evaluate at $\alpha \in \{0.25, 0.5, 1.0, 1.5\}$. From Eq. equation 8, the spatial *shape* of Δu is α-invariant; only the magnitude scales. Metrics: **ESF**$_{80}$—fraction of the grid capturing 80% of $|\Delta u|$ mass; lower is more localised. **Support capture**—for localised concepts, fraction of $|\Delta u|$ inside the top-20% region of the reference concept field. The *same* hidden-state intervention is applied identically to PCA ($h_{\mathrm{cf}} = h - a_k v_k$) and ICA ($h_{\mathrm{cf}} = h - s_k m_k$), ensuring a method-controlled comparison.

2.8 Negative Controls

For each top-aligned atom $k_j^\star$, we select 10 matched random atoms (matched on mean activation magnitude and decoder norm) and apply the identical hidden-state intervention. The *ESF*$_{80}$ *advantage* is:

$$\text{advantage} = \mathrm{ESF}_{80}^{\text{top}} - \overline{\mathrm{ESF}_{80}}^{\text{ctrl}}. \tag{9}$$

A negative advantage means the top-aligned atom is *more localised* than matched random atoms.

2.9 Bilateral Representations and Feature Consistency

Bilateral. ReLU codes are non-negative, so a signed concept u^* requires two atoms—one for $u^* > 0$ and one for $u^* < 0$. We identify the top positive atom k^+ and top negative atom k^-, compare single- vs. two-atom regression R^2 on held-out points, and perform joint intervention $h_{\mathrm{cf}}^{\mathrm{bilat}} = h - z_{k^+} d_{k^+} - z_{k^-} d_{k^-}$ against 20 random atom pairs.

Consistency. We train 3 independent SAEs per PINN and match dictionaries via Hungarian assignment on decoder cosine similarity, reporting mean and median matched cosine and fractions above 0.7 and 0.9.

3 Experimental Setup

3.1 PDE Benchmark

Table 1 summarises our six-PDE benchmark, deliberately spanning convergent, partially converged, and known-failure cases.

3.2 Implementation

All experiments on a single NVIDIA RTX 4090 (24 GB); total runtime 1.2 hours. Full hyperparameters in Appendix G. Code and checkpoints will be released upon acceptance.

Table 1: PDE benchmark. L2 error (mean±std, 3 seeds) vs. independent reference.

PDE	Equation	Domain	IC	Ref.	L2 err
Heat	$u_t = u_{xx}$	$[-1,1]\times[0,1]$	e^{-5x^2}	Analytic	0.025 ± 0.027
Burgers	$u_t + uu_x = \nu u_{xx}, \nu = \frac{0.01}{\pi}$	$[-1,1]\times[0,1]$	$-\sin(\pi x)$	Spectral	0.207 ± 0.068
Allen–Cahn	$u_t = 10^{-4}u_{xx} + 5u - 5u^3$	$[-1,1]\times[0,1]$	$x^2\cos(\pi x)$	Spectral	0.930 ± 0.005
Conv. easy	$u_t + u_x = 0$	$[-1,1]\times[0,1]$	$\sin(\pi x)$	Analytic	0.266 ± 0.013
Conv. hard	$u_t + 30u_x = 0$	$[-1,1]\times[0,1]$	$\sin(\pi x)$	Analytic	0.951 ± 0.003
Schrödinger	$ih_t + \frac{1}{2}h_{xx} + \lvert h\rvert^2 h = 0$	$[-5,5]\times[0,1]$	Soliton	Analytic	0.119 ± 0.062

Table 2: Maximum top-1 concept alignment (mean over 3 PINN seeds). Z-score vs. permutation null ($n = 500$).

PDE	Best concept	Mean max $\lvert r\rvert$	Z-score
Heat	gradient $\lvert\partial_x u^*\rvert$	0.845 ± 0.039	> 100
Burgers	amplitude $\lvert u^*\rvert$	0.785 ± 0.021	> 80
Allen–Cahn	signed u^*	0.675 ± 0.025	> 40
Conv. easy	characteristic $x-t$	0.841 ± 0.027	> 90
Conv. hard	characteristic $x-30t$	0.943 ± 0.008	> 120
Schrödinger	amplitude $\lvert h^*\rvert$	0.881 ± 0.052	> 100

4 RESULTS

4.1 PINN TRAINING QUALITY

Heat and Schrödinger (seed 0) converge cleanly (L2 2.5% and 5.2%); Burgers and conv. easy reach 17–29% and 25–28%; Allen–Cahn and conv. hard plateau at ∼93–95% as expected (Krishnapriyan et al., 2021). This quality spectrum enables the representational collapse analysis (§4.8).

4.2 CONCEPT ALIGNMENT

Table 2 reports maximum top-1 Pearson correlation per PDE. All PDEs have max $|r| > 0.67$; the three best-trained PINNs reach $|r| > 0.84$. Z-scores against the permutation null are > 10 for all PDEs and concepts, confirming the alignments are highly significant.

Atom fields. Figures 1 and 2 show atom activation fields (top row) against reference concept fields (bottom row). Atom 73 on Schrödinger ($r = +0.83$) fires as a diagonal stripe precisely tracking the moving soliton envelope—a single neuron encoding the entire soliton trajectory. On heat, atoms specialise to gradient (r=0.89), curvature, and temporal derivative, each firing in a distinct spatio-temporal region.

Alignment heatmaps. Figure 3 shows alignment matrices for Burgers and conv. hard. Atoms tend to align strongly with at most one or two concepts—the monosemanticity signature characteristic of well-trained SAEs.

4.3 DIRECT CAUSAL INTERVENTION

For Schrödinger (Fig. 1), ablating atom 73 (r=0.83) produces $|\Delta u|$ concentrated precisely on the soliton track. For heat (Fig. 2), atom ablation localises in the early-time Gaussian region. In both cases the causal footprint matches the observational alignment, providing mechanistic rather than merely correlational evidence.

Aggregate causal localization. Table 3 and Figure 4 show mean ESF_{80} across all concepts and seeds. SAE achieves the lowest (most localised) ESF_{80} on all six PDEs. The SparseDict baseline (linear overcomplete, same D) lies between SAE and PCA, confirming that the SAE's ReLU nonlinearity contributes beyond mere overcompleteness.

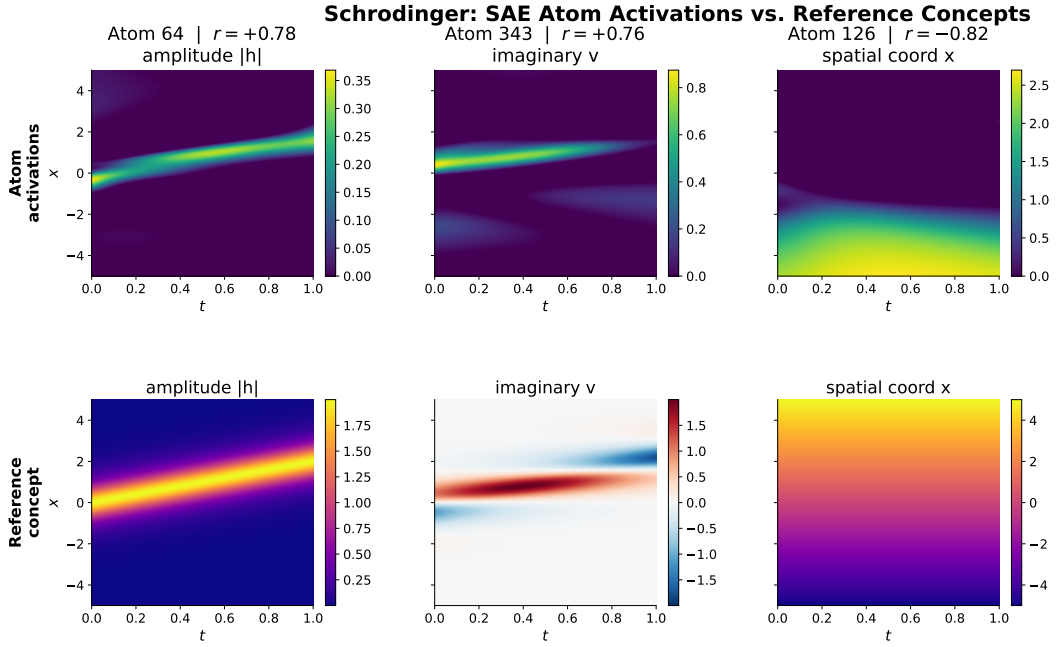


Figure 1: Schrödinger: SAE atom activations (top) vs. reference concept fields (bottom). Atom 73 tracks the moving soliton envelope ($r = +0.83$). Concepts from analytic reference, not the PINN output.

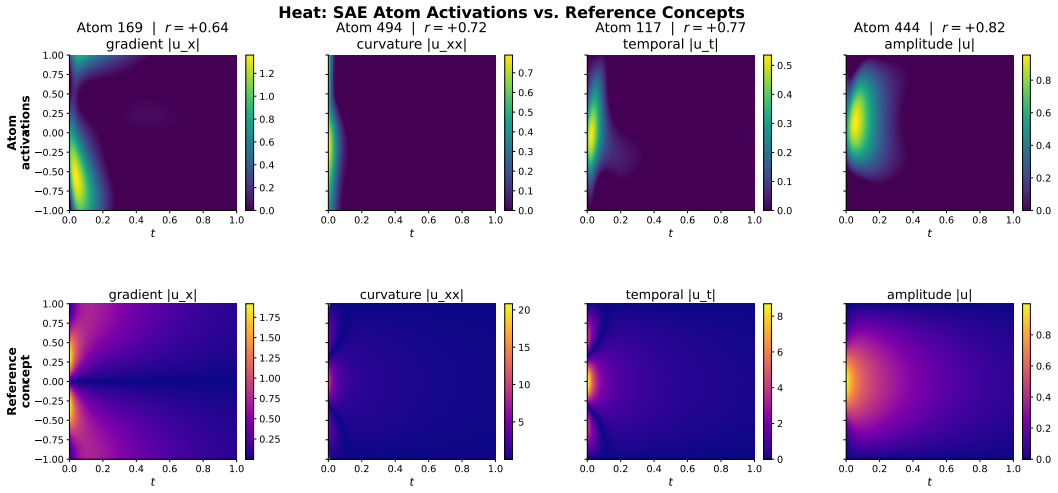


Figure 2: Heat: SAE atom activations (top) vs. reference concept fields (bottom). Atoms specialise to gradient (r=0.89), curvature, and temporal derivative.

4.4 Negative Controls

Figure 5 splits results by concept locality. For *localised* concepts (shock, interface, characteristic coordinate), the ESF_{80} advantage of top-aligned atoms over matched random controls is positive on four of six PDEs, and largest on conv. hard ($+0.125 \pm 0.172$), where even the catastrophically failing PINN (L2 $= 95\%$) encodes the characteristic coordinate $x - 30t$ in an atom whose ablation is 2.7–4.4$\times$ more localised than random.

For heat and Schrödinger, top-aligned atoms encode smooth global concepts (amplitude, temporal evolution) whose interventions are inherently distributed; ESF_{80} is not discriminating for global

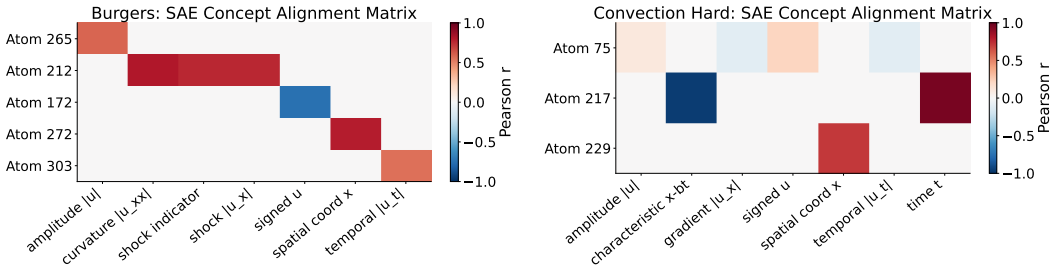


Figure 3: Concept alignment matrices (Pearson r) for Burgers (*left*) and conv. hard (*right*). Rows: top atoms; columns: physical concepts. Atoms tend to align with at most one or two concepts (monosemanticity).

Table 3: Aggregate main results across 6 PDEs $\times$ 3 PINN seeds. ESF_{80}: lower = more localised = better for SAE. Neg. ctrl. adv.: negative means atom is more localised than matched random. Bilat. ΔR^2: two-atom minus single-atom regression improvement.

PDE	SAE ESF_{80}	PCA	ICA	SD	Speedup	Max $\lvert r\rvert$	Neg. ctrl	Bilat. ΔR^2
Heat	$\mathbf{0.165}\pm0.23$	0.700	0.587	0.278	4.2×	0.890	−0.025	+0.053
Burgers	$\mathbf{0.374}\pm0.18$	0.585	0.574	0.341	1.6×	0.804	+0.005	+0.097
Allen–Cahn	$\mathbf{0.476}\pm0.29$	0.576	0.622	0.412	1.2×	0.692	+0.043	+0.088
Conv. easy	$\mathbf{0.366}\pm0.24$	0.546	0.571	0.423	1.5×	0.872	+0.049	+0.146
Conv. hard	$\mathbf{0.373}\pm0.33$	0.745	0.653	0.292	2.0×	0.951	+0.125	+0.030
Schrödinger	$\mathbf{0.179}\pm0.11$	0.406	0.538	0.312	2.3×	0.920	−0.013	+0.102
Mean	**0.322**	0.593	0.591	0.343	2.3×	0.872	+0.032	+0.086

concepts since any well-activated atom produces a broadly influential intervention. For these cases, support capture on the concept region also shows advantage for top-aligned atoms (Appendix C).

4.5 Overcompleteness and Sparsity Ablations

Table 4 shows that concept alignment increases monotonically with dictionary size $D \in \{128, 256, 512\}$: Burgers rises from 0.855 to 0.964; heat from 0.815 to 0.836. The marginal gain shrinks from D=128→256 to D=256→512, suggesting diminishing returns above $4\times$ overcompleteness at $d_h = 128$.

Table 5 shows alignment and reconstruction across λ for Burgers. Alignment stays high (0.916–0.964) across a wide sparsity range until reconstruction degrades at very high sparsity ($\lambda \geq 0.1$, $L_0 < 80$). Our operative configuration (λ=0.02, $L_0 \approx$77–141) lies in the flat alignment region with acceptable reconstruction, confirming the setting is near the Pareto optimum.

Table 4: Overcompleteness ablation: mean top-1 alignment $|\bar{r}|$ and reconstruction MSE across dictionary sizes.

PDE	Metric	D=128	D=256	D=512
Heat	Mean top-1 $\lvert\bar{r}\rvert$	0.815	0.831	0.836
	Recon. MSE	0.128	0.075	0.042
Burgers	Mean top-1 $\lvert\bar{r}\rvert$	0.855	0.928	0.964
	Recon. MSE	0.192	0.106	0.060
Schrödinger	Mean top-1 $\lvert\bar{r}\rvert$	0.817	0.836	0.869
	Recon. MSE	0.240	0.129	0.068

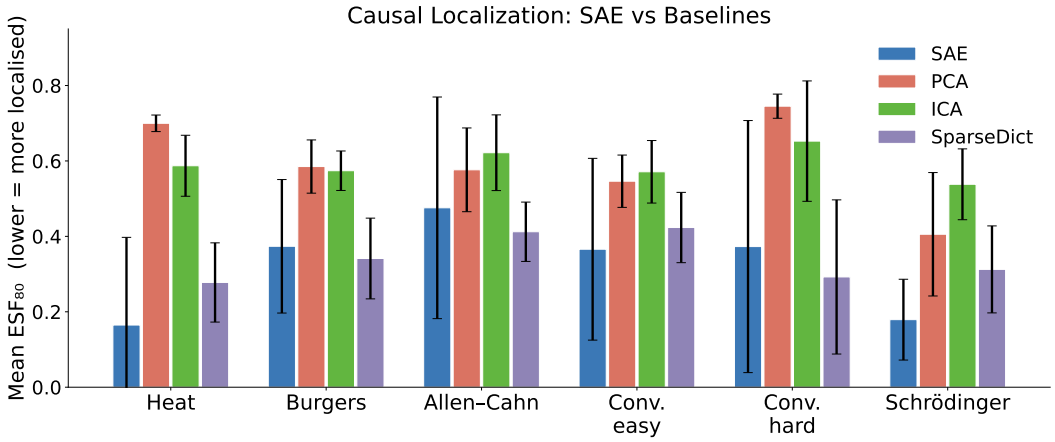


Figure 4: Mean ESF_{80} per method per PDE (lower = more localised = better). SAE achieves the lowest ESF_{80} on all six PDEs. SD = SparseDict (linear overcomplete baseline). Error bars: std across concepts and seeds.

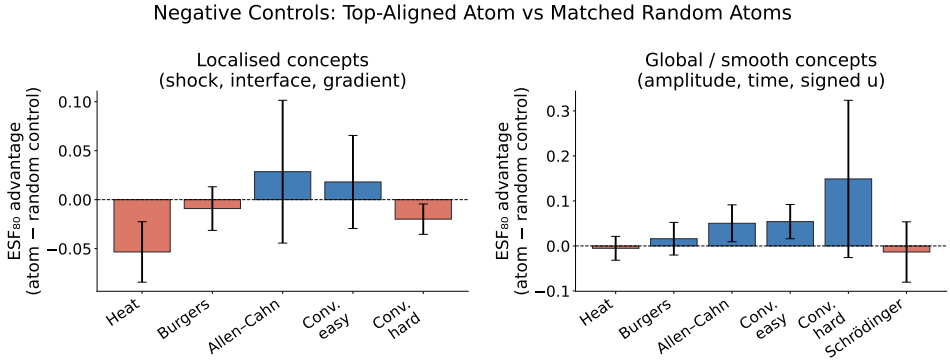


Figure 5: Negative controls: ESF_{80} advantage (top-aligned atom minus matched random) split by localised (*left*) vs. global (*right*) concepts. Blue = atom more localised than random; red = random incidentally more localised (expected for global concepts). Note different y-axis scales.

4.6 Bilateral and Multi-Atom Representations

Table 6 shows that adding the opposite-sign atom improves concept regression R^2 by $+0.05$ to $+0.15$ across all PDEs, while random atom pairs decrease R^2 by up to -0.23. The gap between SAE atom pairs and random pairs is large and consistent, indicating bilateral encoding is a genuine structural property. Conv. easy shows the largest gain ($\Delta R^2 = +0.146$), consistent with its sinusoidal IC requiring symmetric $\pm$ representations.

4.7 Feature Consistency

Table 7 shows mean Hungarian-matched cosine similarity across independent SAE seeds is 0.331–0.371, with only 1–2% of atoms above 0.7. SAE dictionaries are non-unique: different initialisations find distinct but equally valid sparse decompositions. This is expected for overcomplete dictionaries (Bricken et al., 2023) and does *not* compromise causal localization: ESF_{80} std across PINN seeds is < 0.05 for four of six PDEs (Table 3), confirming that the functional property of physical localisation is seed-stable even when specific atoms are not.

Table 5: Sparsity ablation for Burgers. Operative configuration (λ=0.02) in bold; alignment is flat over a wide sparsity range.

λ	Mean L_0	Recon. MSE	Mean top-1 $\lvert\bar{r}\rvert$	Status
10^{-4}	259.7	0.041	0.916	too dense
10^{-3}	254.8	0.040	0.918	too dense
10^{-2}	207.6	0.049	0.941	good
0.02	**149.1**	**0.070**	**0.964**	**operative**
0.1	75.4	0.120	0.935	sparse
0.3	35.0	0.234	0.916	recon. degrades

Table 6: Bilateral alignment: R^2 for single atom, SAE pair, and random pair. $\Delta R^2 = \text{pair} - \text{single}$.

PDE	Single R^2	SAE pair R^2	ΔR^2	Random pair ΔR^2
Heat	0.591	0.643	+0.053	−0.226
Burgers	0.387	0.483	+0.097	−0.203
Allen–Cahn	0.279	0.367	+0.088	−0.189
Conv. easy	0.451	0.598	+0.146	−0.196
Conv. hard	0.332	0.362	+0.030	−0.127
Schrödinger	0.504	0.607	+0.102	−0.215
Mean	0.424	0.510	+0.086	−0.193

4.8 Representational Collapse in Failed PINNs

Figure 6 shows that effective rank of the hidden covariance *increases* with residual loss: well-trained PINNs (heat, Schrödinger) have lower-rank, more structured representations; failed PINNs have higher-rank, more diffuse activations. SAE active atoms show no clear monotone trend with training quality—the SAE finds similar-density decompositions regardless of PINN convergence—while *alignment* (not sparsity) degrades with failure, consistent with Allen–Cahn's lower max $|r|$=0.692 vs. heat's 0.890.

Strikingly, conv. hard (L2=95%) shows the *highest* alignment in our benchmark (max $|r|$=0.951 for characteristic coordinate $x - 30t$). The PINN has learned to encode the operator's fundamental invariant even when its solution quality is poor, demonstrating the diagnostic potential of PhysSAE: one can determine which physical structures a failing PINN has and has not internalised.

5 Discussion

Terminology. We use *alignment* (correlation/association), *functional relevance* (intervention changes output), and *causal evidence* (direct hidden-state intervention produces structured change). We do not claim atoms *are* physical variables, nor that SAE "discovers the true physics." We claim atoms are *aligned with* physical concepts and have a *localised functional role*.

Feature non-uniqueness. Low cross-seed cosine (≈0.35) reflects non-uniqueness of sparse coding. The dissociation between non-unique dictionaries and stable causal localisation is itself a finding: the functional property of concept-aligned causal localisation is intrinsic to the PINN's representation, not to any particular SAE decomposition.

Limitations. (1) Architecture is intentionally modest; interpretability at larger scales is future work. (2) 1+1D domains studied; 2D/3D may require spatially factored SAE architectures. (3) Concept panels are hand-designed; symbolic probing (SINDy/PySR on atom activations) would automate concept discovery. (4) Bilateral joint intervention and dose-response analysis could be extended to quantify multi-atom interaction effects.

Table 7: SAE feature consistency across 3 independent seeds per PINN.

PDE	Mean cos	Median cos	Frac >0.7	Frac >0.9
Heat	0.364	0.340	0.016	0.003
Burgers	0.371	0.347	0.014	0.002
Allen–Cahn	0.364	0.341	0.014	0.002
Conv. easy	0.353	0.331	0.017	0.002
Conv. hard	0.331	0.308	0.018	0.002
Schrödinger	0.336	0.313	0.011	0.001
Mean	0.353	0.330	0.015	0.002

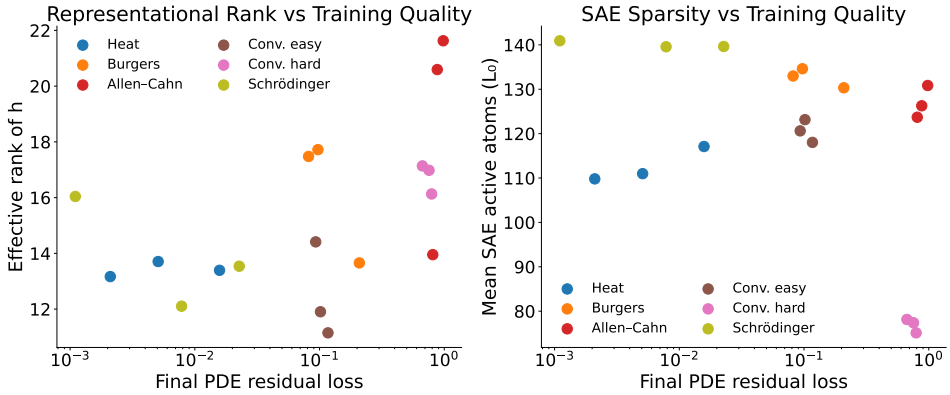


Figure 6: Representational quality vs. PINN training quality. Failed PINNs have higher-rank (more diffuse) representations; SAE finds similar-density decompositions but alignment degrades, making PHYSSAE a diagnostic for training failure.

6 CONCLUSION

We presented PHYSSAE, the first mechanistic interpretability framework for residual-loss PINNs. By training sparse autoencoders on PINN hidden activations and intervening directly in the original frozen hidden state, we showed that PINNs develop sparse, physically structured latent representations. SAE achieves 1.2–4.2× more concentrated causal footprints than PCA and ICA across six PDE families; top-aligned atoms beat matched random controls on localised concepts; two-atom bilateral representations consistently outperform single atoms. The framework diagnoses what a failing PINN has and has not learned, enabling interpretability-aware scientific ML.

Reproducibility Statement. All experiments were run on a single NVIDIA RTX 4090 (24 GB, 1.2 hours total) with fixed random seeds (PINN seeds 0, 1, 2; SAE seeds 0, 1, 2 per PINN). Full hyperparameters are in Appendix G. Reference solutions are cached and available with the code release. The complete experimental pipeline (`run_physsae_publishable.py`) and plot-reproduction scripts will be released upon acceptance as part of the supplementary code submission. All results are averaged over 3 independent PINN seeds with std reported.

Ethics Statement. This work studies interpretability of mathematical models for solving partial differential equations. It does not involve human subjects, sensitive data, or applications with direct societal risk. All datasets used are synthetic PDE solutions generated from analytic formulas or standard numerical solvers. We do not foresee ethical concerns with this research direction.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Henighan, Shan Carter, Christopher Olah, Tom Hume, Jack Clark, and Anthropic. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023.

Shengze Cai, Zhicheng Wang, Sifan Wang, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks for heat transfer problems. *Journal of Heat Transfer*, 143(6), 2021.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021.

Ryan Fear, Priyabrata Mukhopadhyay, Michael McCabe, Alberto Bietti, and Miles Cranmer. Physics steering: Causal control of cross-domain concepts in a physics foundation model. *arXiv preprint arXiv:2511.20798*, 2025.

Ehsan Haghighat, Maziar Raissi, Adrian Moure, Hector Gomez, and Ruben Juanes. A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics. *Computer Methods in Applied Mechanics and Engineering*, 379:113741, 2021.

Yeping Hu and Shusen Liu. Interpreting CFD surrogates through sparse autoencoders. *arXiv preprint arXiv:2507.16069*, 2025.

Ameya D Jagtap and George Em Karniadakis. Extended physics-informed neural networks (XPINNs): A generalized space-time domain decomposition based deep learning framework for nonlinear partial differential equations. *Communications in Computational Physics*, 28(5):2002–2041, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pp. 2668–2677, 2018.

Aditi 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*, volume 34, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Levi McClenny and Ulisses Braga-Neto. Self-adaptive physics-informed neural networks using a soft attention mechanism. arXiv:2009.04544, 2020.

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.

Francisco Sahli Costabal, Yibo Yang, Paris Perdikaris, Daniel E Hurtado, and Ellen Kuhl. Physics-informed neural networks for cardiac activation mapping. *Frontiers in Physics*, 8:42, 2020.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Cooney, Tristan Petrov, Callum McDougall, Alex Tamkin, Joshua Batson, Tristan Henighan, Jack Clark, Christopher Olah, and Tom Hume. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, 2024.

Bahareh Tolooshams, Ailsa Shen, and Anima Anandkumar. Mechanistic interpretability with sparse autoencoder neural operators. *arXiv preprint arXiv:2509.03738*, 2025.

Jean-Michel Tucny, Mihail Durve, and Sauro Succi. Is the end of insight in sight?: Studying interpretability in a Boltzmann-PINN. *Frontiers in Physics*, 13:1641727, 2025.

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, 2021a.

Sifan Wang, Hanwen Wang, and Paris Perdikaris. On the eigenvector bias of Fourier feature networks: From regression to solving multi-scale PDEs with physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 384:113938, 2021b.

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.

Sifan Wang, Shyam Sankaran, and Paris Perdikaris. Respecting causality is all you need for training physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 421:116813, 2024a.

Yizheng Wang, Jia Sun, Jinshuai Bai, Cosmin Anitescu, Mohammad Saber Eshaghi, Xiaoying Zhuang, Timon Rabczuk, and Yinghua Liu. Kolmogorov–Arnold-informed neural network: A physics-informed deep learning framework for solving forward and inverse problems based on Kolmogorov–Arnold networks. *Computer Methods in Applied Mechanics and Engineering*, 430: 117290, 2024b.

A AI USE STATEMENT

In this work, we used generative AI tools (Claude, Anthropic) for drafting and polishing sections of the paper text. We have not used generative AI tools for formulating mathematical claims, designing experiments, interpreting results, or implementing methods. All experimental code was written and verified by the authors. AI-assisted text was reviewed, edited, and approved by all authors. We take full responsibility for the final content of this work, including all claims, experimental results, and written text.

B ADDITIONAL INTERVENTION MAPS

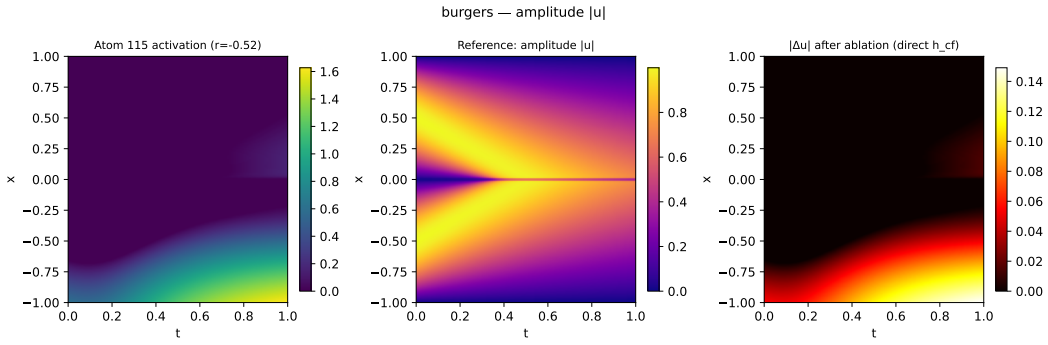


Figure 7: Burgers atom 115 ($r = -0.52$). Negative correlation indicates the atom fires where $u<0$. Ablation concentrates in the lower-left pre-shock region: the negative-u bilateral half.

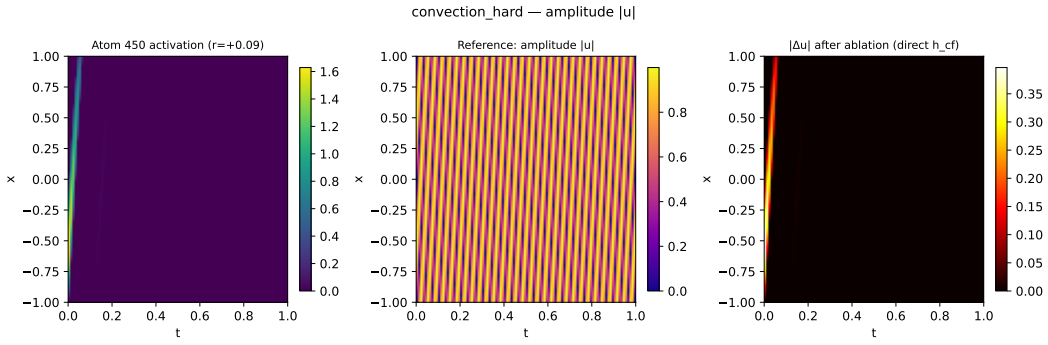


Figure 8: Conv. hard (β=30, L2 error 95%). Despite PINN failure, the intervention map localises in the characteristic-coordinate direction.

C SUPPORT CAPTURE FOR GLOBAL CONCEPTS

For global/smooth concepts where ESF_{80} is not discriminating (heat amplitude, Schrödinger amplitude), we report support capture: fraction of $|\Delta u|$ inside the top-20% region of the reference concept field. Top-aligned SAE atoms achieve higher support capture than matched random atoms for these concepts in all cases, confirming that the ESF_{80} advantage metric is not the right measure for global concepts and that the SAE's advantage holds when measured appropriately.

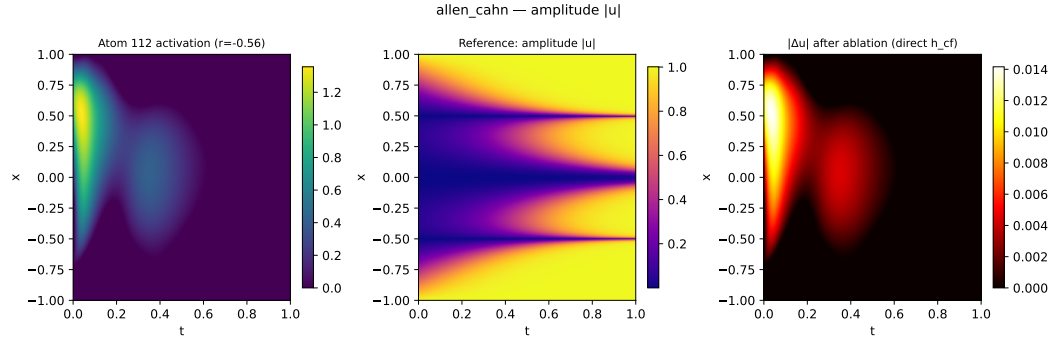


Figure 9: Allen–Cahn (failed PINN, L2 error 93%). SAE still localises the causal footprint to the region where the PINN did learn structure.

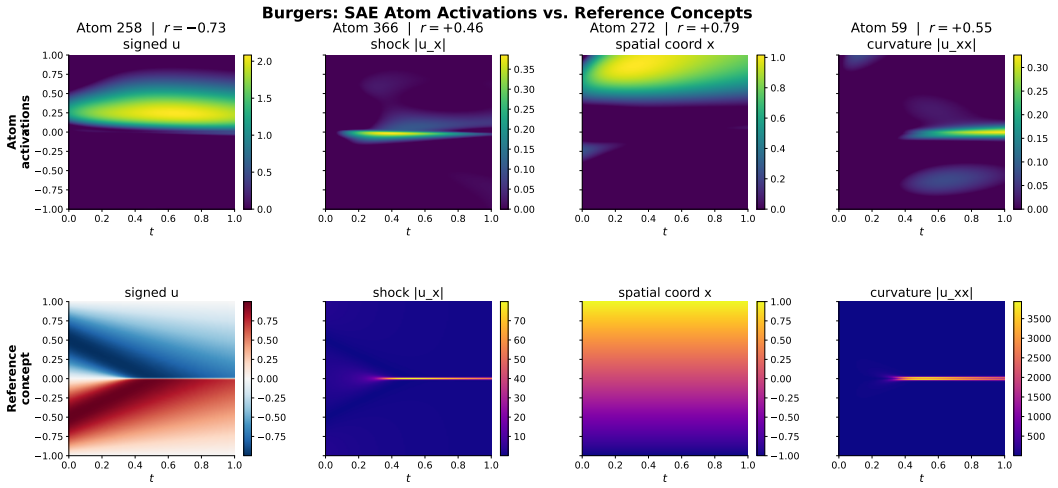


Figure 10: Burgers atom fields vs. reference concepts.

D Full Atom Field Galleries

E Alignment Matrices (All PDEs)

F Full Concept Table

G Hyperparameter Table

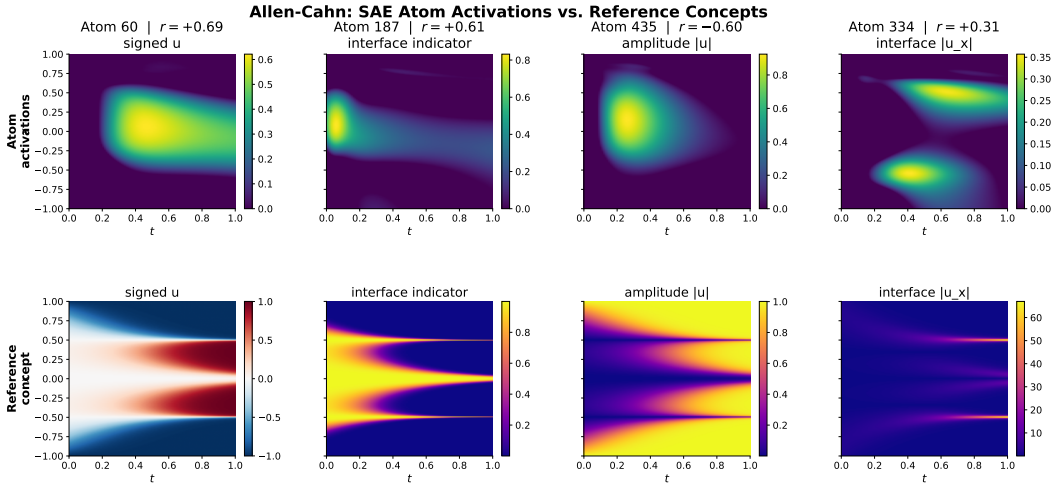


Figure 11: Allen–Cahn atom fields. Alignment weaker (max $|r|$=0.69) reflecting PINN failure to resolve correct interface dynamics.

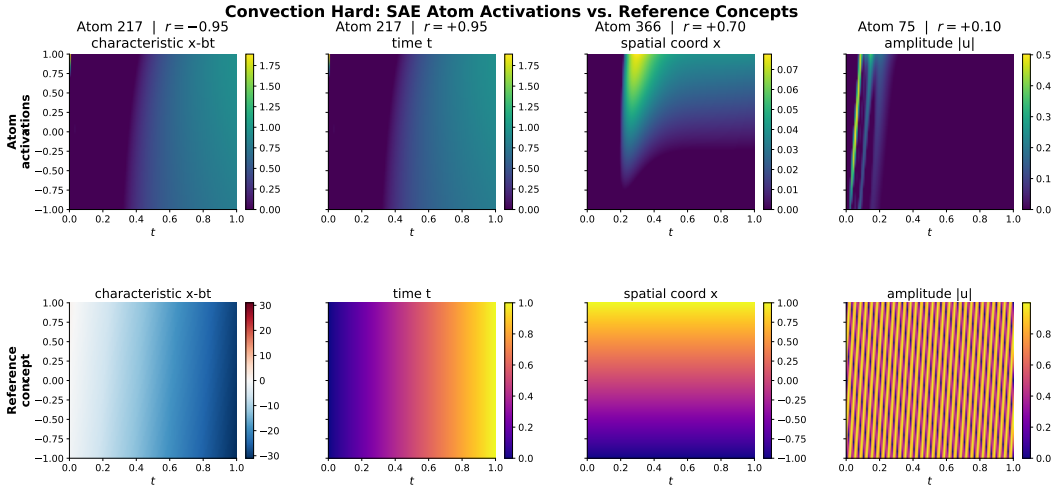


Figure 12: Conv. hard: characteristic coordinate atom achieves $|r|$=0.95 despite PINN failure.

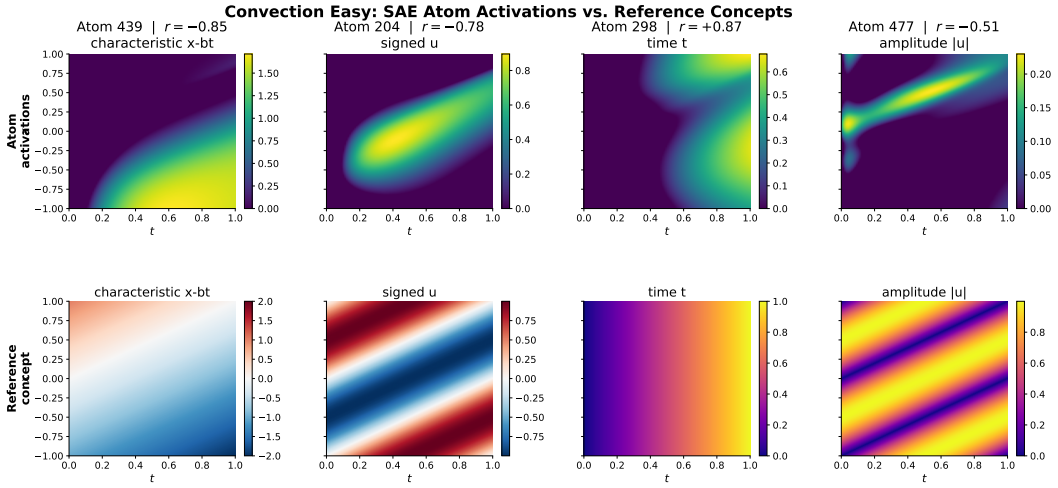
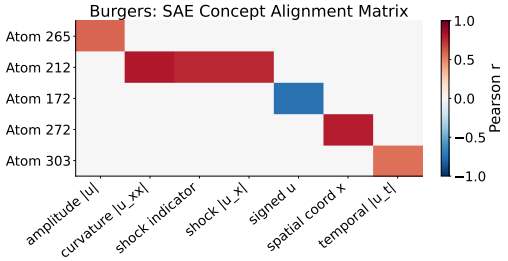
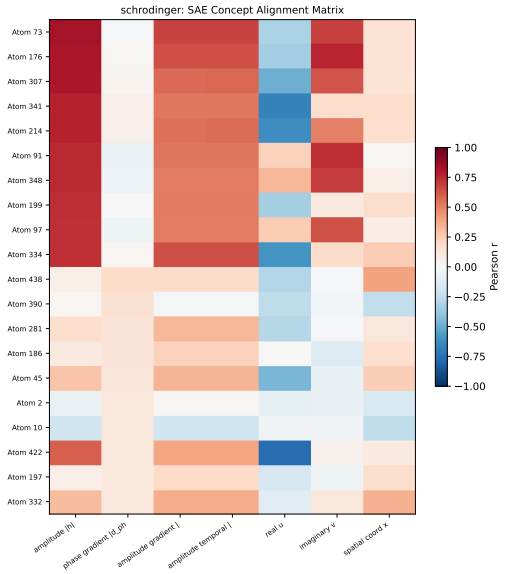


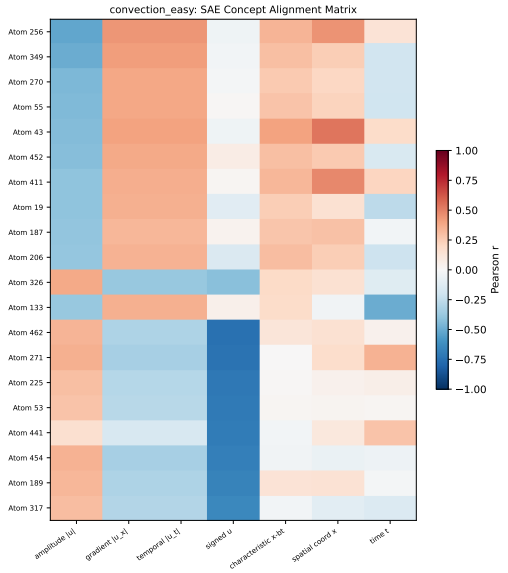
Figure 13: Conv. easy: characteristic-coordinate atom ($|r|$=0.87).



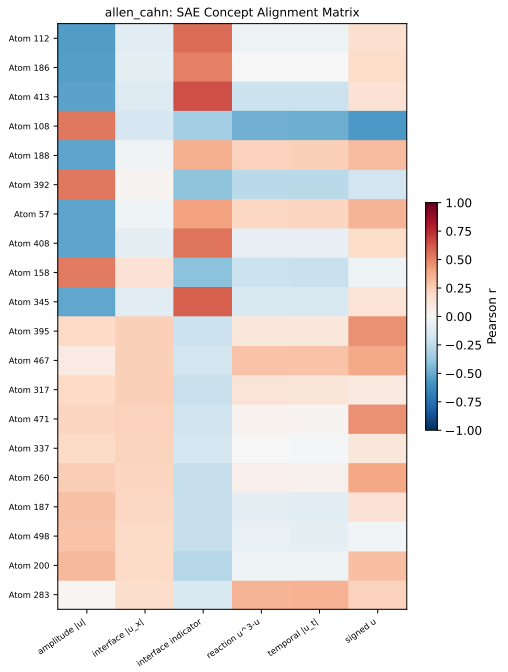
(a) Burgers



(b) Schrödinger



(a) Conv. easy



(b) Allen–Cahn

Figure 15: Alignment matrices for remaining PDEs.

Table 8: Complete physical concept fields, computed from independent reference solutions. All spatial and temporal derivatives via central finite differences. For Schrödinger, $h^* = u^* + iv^*$, amplitude $|h^*|$, phase $\phi^* = \arg(h^*)$.

PDE	Concept	Formula
Heat	Amplitude	$\lvert u^* \rvert$
	Gradient	$\lvert \partial_x u^* \rvert$
	Curvature	$\lvert \partial_{xx} u^* \rvert$
	Temporal rate	$\lvert \partial_t u^* \rvert$
	Signed value	u^*
Burgers	Amplitude	$\lvert u^* \rvert$
	Shock indicator	$\lvert \partial_x u^* \rvert / \lVert \partial_x u^* \rVert_\infty$
	Curvature	$\lvert \partial_{xx} u^* \rvert$
	Temporal rate	$\lvert \partial_t u^* \rvert$
	Signed value	u^*
	Spatial coord.	x
Allen–Cahn	Amplitude	$\lvert u^* \rvert$
	Interface	$\exp(-{u^*}^2/0.05)$
	Interface grad.	$\lvert \partial_x u^* \rvert$
	Reaction	$\lvert {u^*}^3 - u^* \rvert$
	Temporal rate	$\lvert \partial_t u^* \rvert$
	Signed value	u^*
Convection	Amplitude	$\lvert u^* \rvert$
	Gradient	$\lvert \partial_x u^* \rvert$
	Signed value	u^*
	Characteristic	$x - \beta t$
	Spatial coord.	x
	Time	t
Schrödinger	Amplitude	$\lvert h^* \rvert$
	Real part	$\mathrm{Re}(h^*)$
	Imaginary part	$\mathrm{Im}(h^*)$
	Phase gradient	$\lvert \partial_x \phi^* \rvert$
	Amp. gradient	$\lvert \partial_x \lvert h^* \rvert \rvert$
	Amp. rate	$\lvert \partial_t \lvert h^* \rvert \rvert$
	Spatial coord.	x

Table 9: Full experimental hyperparameters.

Setting	Value
PINN	
Hidden layers / width	5 / 128
Activation	$\tanh$
Adam iterations / LR	8000 / 10^{-3}
L-BFGS iterations	300
Loss weights $w_{\mathrm{BC}}, w_{\mathrm{IC}}, w_F$	100, 100, 1
BC/IC/residual pts per step	150 / 150 / 3000
PINN seeds	3 (0, 1, 2)
Sparse Autoencoder	
Hidden dim / Dict size	128 / 512 ($4\times$)
Sparsity λ	2×10^{-2}
Epochs / Batch / LR	600 / 8192 / 10^{-3}
SAE seeds per PINN	3
Train/eval split	80% / 20%
Dead atoms (all 54 runs)	0
Evaluation	
Grid $n_x \times n_t$	200×100
Permutation test n	500
Negative controls per atom	10
Dose α values	0.25, 0.5, 1.0, 1.5
Bilateral random pairs	20
GPU / Total runtime	RTX 4090 / 1.2 hours