YOUR PHYSICS PRIOR MIGHT BE A CONSTANT: CONTROLS FOR PHYSICS-INFORMED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Physics-informed machine learning injects a physical law into training and credits the physics when accuracy improves. On a public wildfire-spread benchmark we show that this attribution can fail, and that two inexpensive checks expose the failure. The first is a pre-training diagnostic: rank the cells the prior is supposed to discriminate and score the ranking against the class prior. The standard fire-spread law (Rothermel) scores 0.210 against 0.193, nearly uninformative at the data's resolution, which correctly predicts that an eikonal physics loss changes nothing. The second is a set of post-training controls. Distilling a calibrated Rothermel teacher into a U-Net gains $+0.104$ AP over our reproduced Dice baseline on 12 of 12 folds, a result the standard two-arm design would credit to physics. Yet an information-free constant teacher reproduces the gain, a plain Dice+BCE loss reaches $+0.095$ with no teacher at all, and a genuinely informative learned prior (a 13k-parameter neural cellular automaton) gains less than a constant at the label base rate. The gain comes from a constant-target regulariser whose strength is set by the target level and not by the prior's information. Compared with an information-free control at its own level, the learned prior transmits a small amount of information ($+0.03$) and the physics prior none. We propose both checks as standard practice for physics-informed claims.

1 INTRODUCTION

The promise of physics-informed learning is that known dynamics can substitute for scarce or noisy data: impose a differential-equation residual or an empirical law as an auxiliary objective, and the model inherits structure it would otherwise have to learn. The standard experimental design has two arms, one with the physics term and one without, and a standard inference: if the physics arm wins, the physics helped.

This paper is about what that design cannot detect; our own experiment is the example. On WildfireSpreadTS (Gerard et al., 2023), a 607-fire benchmark for next-day wildfire spread, we set out to make physics-informed learning work, and built the two-arm experiment carefully: a faithful reproduction of the published baselines, the operational fire-spread law of Rothermel (1972) implemented differentiably and calibrated to the data, and a distillation objective that transfers the law into a 14.4M-parameter U-Net. The physics arm won by $+0.104$ average precision, on every one of the twelve cross-validation folds. By the usual two-arm standard this is a physics-informed gain.

Two checks show that it is not. A pre-training diagnostic, which ranks the fire-front cells by the prior and scores average precision against the class-prior baseline, shows the law to be nearly uninformative at the benchmark's resolution (0.210 vs. 0.193 class prior, 12-fold means): there is almost no information to transfer. Post-training controls, which repeat the distillation with information-free teachers, show that a constant target reproduces the physics arm, that plain Dice+BCE captures most of the gain with no teacher at all, and that a genuinely informative learned prior gains less than a constant at the label base rate. The mechanism is simple. The calibrated physics teacher carries no usable spatial information and sits at the number 0.196, and an auxiliary BCE term toward a constant at that level regularises a Dice-trained network out of a saturation pathology. Once each teacher is compared with an information-free control at its own level, the learned prior does transmit a small amount of information ($+0.03$ over its shuffled copy) and the physics prior transmits none.

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.

We argue that both checks should be standard practice. The controls required an informative prior arm; ours is a neural cellular automaton (Mordvintsev et al., 2020) small enough (2,336–13,440 parameters) to characterise completely, whose extracted spread rule tracks the observed terrain response more closely than Rothermel's on all four test years.

Two failure modes, and the scope of the claim. The two checks catch different failures. Imposed as a PDE residual, an uninformative law yields a null: adding it changes nothing, and Check One predicts that before training. Imposed as a soft target (distillation, pseudo-labels, any auxiliary loss toward a prior's output), the same law yields a false positive, because the auxiliary term regularises the student whatever the target says; only Check Two exposes it. Our evidence concerns one semi-empirical law at one benchmark's daily, 375 m resolution; it says nothing against physics priors in general, and richer propagation models may pass both checks. The claim is that a two-arm experiment cannot tell.

Contributions.

1. **A prior-informativeness diagnostic** (Section 4): one inference pass, before any training, that here predicts the failure of the physics loss.
2. **Information-free, level-matched controls for auxiliary-loss claims** (Sections 5–6): constant and shuffled teachers plus a no-teacher loss arm. They identify a $+0.104$ AP "physics" gain as a constant-target regulariser that loss terms carrying no physics reproduce, and they separate the level effect from the information a prior actually transmits.
3. **A compact learned prior and its rule** (Sections 7–8): a $13{,}440$-parameter cellular automaton that reaches 96.4% of an independently trained U-Net under matched (focal) tuning as a 16-rollout expectation (94.5% in a single pass), and whose extracted rule is validated against observed marginals, including the finding that the dominant meteorological drivers are not year-stable enough to validate any prior at this resolution.
4. **Benchmark methodology** (Section 3): test-year variance dominates architecture effects, so comparisons must be paired per fold; a stochastic model's prediction is its expectation over rollouts (worth $+0.043$ AP here); and a reproduction-first protocol that exposed our own false positive.

2 Related Work

Physics-informed learning. PINN-style methods impose PDE residuals as losses (Raissi et al., 2019; Karniadakis et al., 2021); for fire, Bottero et al. (2020) solve the level-set equation and Dabrowski et al. (2023) assimilate observed fronts with Bayesian PINNs. A substantial literature studies why PINNs fail, attributing failures to optimisation: stiff gradient pathologies (Wang et al., 2021) and loss landscapes that become intractable for certain PDE regimes (Krishnapriyan et al., 2021). We exhibit an orthogonal and, we believe, under-examined failure axis: the imposed law can be uninformative for the task at the data's resolution, in which case no optimisation remedy applies, and the apparent gains that remain can be loss-form artefacts. Across this literature the informativeness of the law is assumed and ablations compare with-physics against without-physics; we are not aware of physics-informed studies whose gains were required to survive information-free teacher controls.

Wildfire spread benchmarks. WildfireSpreadTS (Gerard et al., 2023) provides 607 fires (13,607 daily 23-channel images, 375 m) with a 12-fold year cross-validation protocol and persistence, Res18-U-Net (Ronneberger et al., 2015) and UTAE (Garnot & Landrieu, 2021) baselines. Lahrichi et al. (2025) re-tuned these baselines (focal loss, learning-rate search) on the same protocol; we treat their numbers as the reference bar (Table 1), and our loss findings corroborate rather than beat them. Huot et al. (2022) provide the coarser mono-temporal antecedent.

Empirical reality checks. Systematic re-examinations have repeatedly found that reported gains disappear under stronger baselines and controlled comparison: in neural recommendation (Ferrari Dacrema et al., 2019), metric learning (Musgrave et al., 2020), and deep RL, where seed vari-

Table 1: Reproduction, our re-tuned U-Nets, and the tuned reference of Lahrichi et al. (2025) (focal loss and learning-rate search). Mean test AP $\pm$ std over 12 folds; the std is dominated by the test year (Section 3).

Model	Ours	Published
Persistence	0.1934 ± 0.0675	0.193 ± 0.065 (Gerard et al., 2023)
Res18-U-Net (Dice)	0.3484 ± 0.0962	0.341 ± 0.086 (Gerard et al., 2023)
Res18-U-Net (Dice+BCE)	0.4429 ± 0.1085	—
Res18-U-Net (focal, repaired)	0.4551 ± 0.1045	0.460 ± 0.084 (Lahrichi et al., 2025)

ance alone undermines many claims (Henderson et al., 2018). Our controls are that instrument for physics-informed learning, applied prospectively rather than as a post-hoc survey.

Distillation without information, and loss form. The closest precedent for our constant teacher is teacher-free distillation: Yuan et al. (2020) show that a poorly trained or even hand-designed teacher improves a student, and interpret distillation as label-smoothing regularisation (Szegedy et al., 2016; Müller et al., 2019). In medical segmentation, compound Dice+CE losses are known to be more robust than Dice alone (Ma et al., 2021), and anchoring outputs at the class prior has precedents in the prior-probability bias initialisation of Lin et al. (2017) and in logit adjustment (Menon et al., 2021). Our contribution is to bring these controls to physics-informed claims, where the teacher is presented as carrying physical information and the loss-form confound is not usually tested.

Neural cellular automata. Differentiable NCAs (Mordvintsev et al., 2020) iterate one shared local rule; they are competitive at extreme parameter counts (Kalkhof et al., 2023) and are formally learned PDE stencils (Richardson et al., 2024). Zhenirovskyy et al. (2026) parameterise a fixed fire CA with a CNN and calibrate per fire on six events; we learn the rule itself and evaluate under year-based cross-validation on 607 fires.

3 Benchmark, Reproduction, and Two Structural Facts

We use WildfireSpreadTS unmodified: the benchmark's 23 raw features expand to 40 tensor channels in its standard loading pipeline, once categorical land cover is one-hot encoded and angular features are expanded. The 40 channels at 375 m (VIIRS reflectance and vegetation, GRIDMET weather, GFS forecasts, SRTM terrain, MODIS land cover, active fire), next-day binary fire mask as target, test AP under 12-fold cross-validation over years 2018–2021, 10k steps AdamW, batch 64, 128×128 crops, Dice loss (Milletari et al., 2016), lowest-validation-loss checkpoint. Table 1 shows the reproduction: our persistence matches the published number to four decimal places (persistence has no trained parameters, so this validates the data pipeline end to end), and the U-Net matches to 0.007.

Two structural facts govern every later comparison. First, the test year dominates variance: folds testing the same year agree to ≤ 0.019 AP although they train on disjoint year pairs, while across test years the mean spans 0.215 (2019) to 0.474 (2021). Architecture and loss effects (0.02–0.10) are smaller than or comparable to the between-year spread (±0.096), so all comparisons in this paper are paired per fold; unpaired comparisons of means on this benchmark are uninterpretable. The same fact bounds what any 12-fold test can show: the folds are four test years times three training pairs, so the twelve paired differences cluster by year and the effective sample size is closer to four. Every headline comparison is therefore reported twice, per fold (Wilcoxon signed-rank and a fold bootstrap) and per test year (folds averaged within year, then paired over the four years; Appendix B), and we call a result established only when both agree. Second, the published baseline degrades under its own loss rather than running short of steps: its best-validation checkpoint arrives at median step $\sim$4,800 of 10,000 in 12/12 folds (the saturation pathology we measure in Section 5.1), and Lahrichi et al. (2025) lift it from 0.341 to 0.460 by loss and learning-rate tuning alone. Physics-informed gains measured against an under-tuned baseline are therefore especially exposed to loss-form confounds.

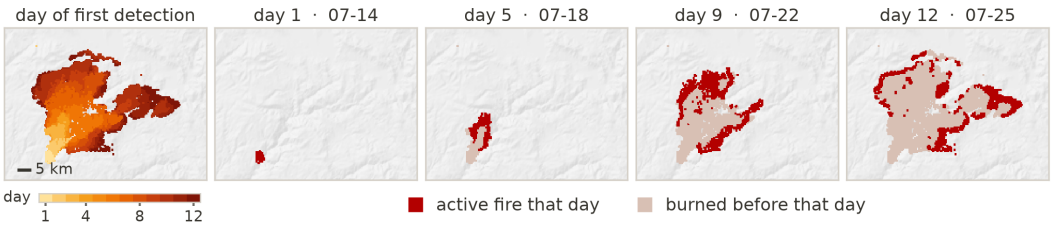


Figure 1: A 2021 test fire at 375 m (VIIRS detections over hillshaded terrain): burned area by day of first detection, then four individual days. Fire moves by local contact, which is the physics a spread law is meant to encode and the inductive bias of the learned local rule in Section 8.

4 Check One: Prior Informativeness (Before Training)

The operational physics of wildfire spread is Rothermel's rate of spread $R = R_0(1+\phi_{\text{wind}}+\phi_{\text{slope}})$ (Rothermel, 1972), used inside level-set solvers (Lautenberger, 2013; Pais et al., 2021) via $\partial\phi/\partial t + S|\nabla\phi| = 0$. We implemented R differentiably from the benchmark's own covariates, verified its qualitative responses, and calibrated its one free scale (predicted head-fire rates exceed observed daily displacement by $\sim$25$\times$; the ratio is absorbed into a fitted factor).

Before imposing a prior, measure how much it discriminates the outcome it is supposed to inform. At ront cells (unburnt cells adjacent to fire, where a spread law acts), rank by the prior and score AP against next-day ignition. The table reports the mean and range over all 12 folds, each computed on that fold's test year (150 test fires per fold, $\sim$19k front cells); the gradient-boosting reference is trained on the fold's training-year front cells and scored on its test-year cells, so every row is an out-of-year number. The class-prior row is each fold's test-year front-cell ignition rate, whose 12-fold mean 0.193 is the same quantity as the 0.196 of the training years that appears in Section 5.

Ranking of front cells	AP (12-fold mean)	range
class prior (random)	0.193	$[0.156, 0.233]$
Rothermel ROS	**0.210**	$[0.185, 0.238]$
learned 13k-parameter rule (Sec. 8)	0.399	$[0.349, 0.451]$
gradient boosting on 7 local drivers	0.275	$[0.228, 0.331]$

Rothermel's ranking exceeds the class prior by $+0.017$ on average; on three folds it does so by more than 0.03 (folds 4, 5 and 10: $+0.044$, $+0.057$, $+0.034$), on six by less, and on three it falls below the prior. The learned rule ranks above the 7-driver gradient-boosting reference on the same cells, so that reference is not an upper bound on what local information supports. The law is nearly uninformative here, for two reasons: its meteorology enters the benchmark at 4.6–28 km native resolution (near-constant across a 48 km training crop), and the law was never calibrated for daily aggregate behaviour at 375 m. Training confirms the diagnostic's prediction: adding an eikonal residual $(R|\nabla T| - 1)^2$ at $\lambda_{\text{eik}} = 0.003$ ($\approx 9\%$ of the supervised loss at initialisation) and a monotonicity penalty at $\lambda_{\text{mono}} = 1$, separately and together, to the K=8 spread model of Section 8 changes test AP by at most 0.008 on any of the four probe folds (one per test year), within the model's seed variance, as expected, since no loss weight can transfer information that a prior does not contain. We did not sweep the weight further; the diagnostic makes the outcome of such a sweep predictable. The diagnostic costs one inference pass; the physics-loss experiments it would have made unnecessary cost $\sim$30 GPU-hours.

5 Check Two: Information-Free Controls (After Training)

A prior can also enter training through distillation (Hinton et al., 2015) rather than through a residual. We train the U-Net with

$$\mathcal{L} \;=\; \text{Dice}(f(x), y) \;+\; \alpha\,\text{BCE}\big(f(x),\, t(x)\big), \qquad \alpha = 1, \tag{1}$$

Table 2: The three-arm study and its controls. **(a)** Gain in test AP over the Dice-only U-Net, paired per fold, with the fold-bootstrap 95% CI and the number of folds on which the arm wins. **(b)** Level-matched comparisons: each teacher against an information-free control at the same target level; the difference is the information the teacher transmits. Rows with 4 folds use one fold per test year and are indicative only (Appendix B).

(a) gain over the Dice-only U-Net (AP 0.348)

Arm	Teacher information	Folds	Gain	95% CI	Wins
Dice+BCE(y)+BCE(0.196)	—	12	$+0.106$	$[+0.096, +0.116]$	12/12
Rothermel teacher	none usable (Sec. 4)	12	$+0.104$	$[+0.093, +0.116]$	12/12
constant 0.196	none by construction	12	$+0.102$	$[+0.088, +0.114]$	12/12
plain Dice+BCE (no teacher)	—	12	$+0.095$	$[+0.081, +0.108]$	12/12
NCA teacher (informative)	front-cell AP 0.399	12	$+0.062$	$[+0.053, +0.070]$	12/12
constant ≈ 0.002	none by construction	12	$+0.050$	$[+0.037, +0.063]$	12/12
shuffled NCA teacher	none (NCA marginal)	12	$+0.028$	$[+0.021, +0.035]$	4/4

(b) level-matched: teacher minus information-free control at the same level

Teacher	Control	Folds	Difference	95% CI	Wins
Rothermel	constant 0.196	12	$+0.003$	$[-0.009, +0.018]$	5/12
NCA	shuffled NCA (exact marginal)	12	$+0.031$	$[+0.027, +0.035]$	4/4
NCA, calibrated	shuffled calibrated NCA	12	-0.003	$[-0.011, +0.005]$	2/4
NCA	constant ≈ 0.002 ($<$ marginal)	12	$+0.012$	$[-0.000, +0.024]$	9/12

where the frozen teacher t is the calibrated Rothermel probability (two-parameter logistic calibration, fit on training front cells) or a separately trained NCA (the informative prior; diagnostic AP 0.399). Teacher quality is verified before each run spends its budget. The controls are three teachers that carry no usable information: a constant, the NCA teacher spatially shuffled (same marginal, no structure), and no teacher at all ($t \equiv y$: plain Dice+BCE).

Table 2(a) contains the paper's central result. The gains do not follow teacher informativeness: the near-random physics teacher gains most, and the informative learned prior trails an information-free constant at the label base rate. The explanation is the level of the teacher. The calibrated physics teacher is nearly uninformative over the image (its all-pixel AP is 0.001–0.005 on every fold, at the class prior) and its calibration intercept places it at $\sigma(b) \approx 0.196$: on fold 0 the fitted slope on ROS is 0.0035 and $t(x) \in [0.1962, 0.1984]$ everywhere; on the other 11 folds the slope is larger (0.045–0.25) and the teacher varies more across cells: its 1st–99th percentile span widens from 0.002 on fold 0 to 0.135 on fold 11, and over all folds $t(x)$ stays inside $[0.173, 0.364]$ with a per-fold median between 0.184 and 0.211. None of that variation is usable structure. The number 0.196 equals the front-cell base ignition rate of the training years; it is a label statistic with no physical content. An explicit constant-0.196 teacher reproduces the Rothermel arm fold by fold (Table 2(b): difference $+0.003$, 95% CI $[-0.009, +0.018]$; equivalent within ± 0.02 AP by TOST at fold level, $p = 0.02$, and short of it at year level, $p = 0.08$), so the entire "physics-informed" gain is a constant-target BCE regulariser whose strength follows the target level ($0.196 \to +0.104$; $\approx 0.002 \to +0.050$). Nor does the physics arm's edge follow the prior's information on the folds where it has some. Check One finds Rothermel measurably informative on three folds (more than 0.03 above the class prior), and the teacher is not exactly constant outside fold 0; yet across the 12 folds the Rothermel-minus-constant difference is uncorrelated with the diagnostic lift (Spearman $+0.12$, $p = 0.71$), averages $+0.002$ on the three informative folds against $+0.003$ on the other nine, and if anything falls as the teacher's spatial spread grows (-0.55, $p = 0.06$; Appendix F). Where the law is informative, the student's gain over the constant does not increase.

The level-matched rows of Table 2(b) also show what a control at the wrong level hides. The NCA teacher's output has a low marginal, so its natural information-free control is its own spatially shuffled copy, which keeps the marginal and destroys the structure. The low constant, set to each fold's training fire-pixel rate (0.0010–0.0022), sits below that marginal (0.005–0.007) and behaves like the shuffled copy (difference $+0.002$, 1/4 folds). Against those the learned prior transmits a small but consistent amount of information: $+0.031$ over the shuffled teacher on all four probe folds and $+0.012$ over the low constant on 9 of 12 folds (all four test years at year level), a comparison that is only approximately level-matched. Against the constant at 0.196 it loses by 0.040, because

that constant is a better regulariser. The channel matters too. Given the same two-parameter logistic calibration as the Rothermel teacher, the NCA teacher collapses to ≈ 0.09 (99th percentile ≤ 0.15) and its advantage over its shuffled copy disappears (-0.003, 2/4 folds): the calibration flattens an informative teacher into a near-constant at a poor level. For Rothermel this changes nothing, since its raw ranking is already at the class prior, but the spread of a calibrated teacher should be checked before distilling. The correct reading of the study is therefore a decomposition: a level effect worth up to $+0.10$ that any teacher at the right level supplies, plus an information effect that is ≈ 0 for the physics teacher and $+0.01$ to $+0.03$ for the learned one. The two-arm design conflates them.

The checks are not limited to negative outcomes. Applied to both priors, they agree with each other in both directions. Rothermel scores 0.210 on Check One against a class prior of 0.193, and on Check Two it adds $+0.003$ ($[-0.009, +0.018]$, 5/12 folds) to its level-matched control. The learned NCA scores 0.399 on Check One and adds $+0.031$ ($[+0.027, +0.035]$, 4/4 folds) on Check Two. The learned prior is the positive control: the diagnostic detects its information before training, and the level-matched comparison recovers it after.

5.1 The Role of a Constant Target

With per-pixel logits $z_i = f_\theta(x)_i$, the arm's loss is $\mathrm{Dice}(\sigma(z), y) + \alpha\,\mathrm{BCE}(\sigma(z), c)$ with the same soft label c at every pixel, whose per-pixel gradient is

$$\frac{\partial}{\partial z_i}\,\mathrm{BCE}\big(\sigma(z_i), c\big) \;=\; \sigma(z_i) - c, \tag{2}$$

zero only at the fixed operating point $z_i = \sigma^{-1}(c) \approx -1.41$; equivalently, up to the additive entropy $H(c)$, the term is the per-pixel anchor $\mathrm{KL}(\mathrm{Ber}(c) \,\|\, \mathrm{Ber}(\sigma(z_i)))$. This matters because Dice's gradient vanishes on confidently-classified pixels: with 99.5% of pixels negative, most of the map saturates to extreme negative logits, and the ordering among those near-zero-probability pixels, which is what AP measures, degenerates. Eq. equation 2 never vanishes: at a saturated negative its magnitude is $|0 - c| = c$, so it pulls logits back toward the anchor, keeps the output distribution graded, and lets the supervised terms re-establish a meaningful ranking within it. The restoring force on saturated pixels equals c, which is why the gain tracks the target level and saturates in the dose–response of Figure A3. The term carries no information about where fire goes, since the target is identical everywhere; it only conditions the logit scale. Its nearest relatives are label smoothing (Szegedy et al., 2016; Müller et al., 2019) and confidence penalties (Pereyra et al., 2017), here with an asymmetric target at the front-cell base ignition rate (0.196) rather than $1/2$. This is not the fire-pixel rate over the whole image, which is two orders of magnitude lower and, as the low-constant arm shows, yields a much smaller gain. A dose–response sweep over $c \in [0.02, 0.8]$ (Figure A3, Appendix I) shows the gain holds over $c \in [0.05, 0.8]$, more than an order of magnitude (at $c = 0.02$ the easiest fold falls below plain Dice+BCE); on the easiest test year the response rises monotonically and saturates for $c \geq 0.3$, so 0.196 is not even the optimal level. Figure 2 summarises all arms.

The saturation this account rests on is measured, not assumed (Appendix E). On the two folds we instrumented, the Dice-only U-Net pushes essentially every pixel past $|z| > 8$ and over 90% of negatives below $p = 10^{-6}$. Every arm with a constant target removes the saturation outright, and the Rothermel arm matches the constant arm on every statistic, as the level account predicts.

Because the gain depends on one specific number, we checked whether that number could carry test-set information. It cannot. The value 0.196 is the output of a logistic calibration fit on training front cells, not of any search. The gain is positive on 12 of 12 folds, each testing a year the constant's derivation never saw. The dose–response is flat over more than an order of magnitude, and two independent seeds reproduce the arm (Appendix G).

The Rothermel teacher is a per-pixel logistic of the rate of spread, applied at every cell of the image. Propagation from today's front, the main content of a spread law, is therefore only partly represented, and the arm tests whether the ranking of front cells by ROS adds anything to a student that already receives today's fire mask as input. That is the right question for this benchmark, where the student has already learned locality (masking the U-Net to the physically reachable set changes AP by $+0.0002$, Appendix H), but it is not the only way to distil a spread law. A one-step level-set propagation from the observed mask, with ROS setting the local advance, would be a fairer physics teacher and is future work; the controls we propose apply to it unchanged.

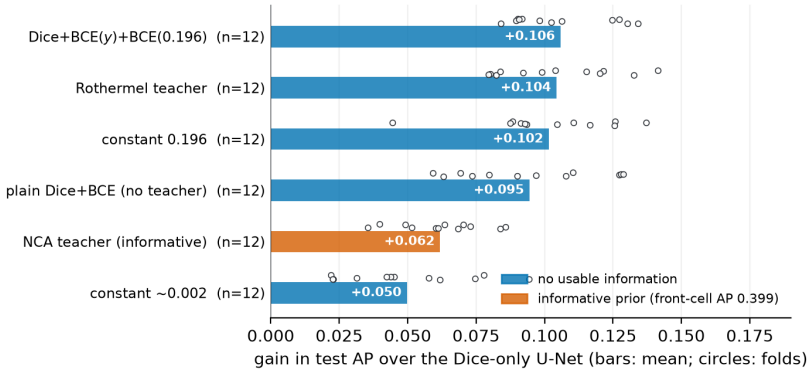


Figure 2: All distillation arms, as gain over the Dice-only U-Net (bars: mean; points: individual folds). The information-free arms at the label base rate (top three, blue) outgain the only teacher that carries real information about fire spread (orange); the low constant (≈ 0.002) does not, which is the level effect of Table 2(b).

Under the standard two-arm design, these experiments would have supported a published claim that distilled fire physics improves wildfire prediction by $+0.10$ AP on every fold of the official protocol.

6 A Loss Recipe from the Controls

The controls also identify what does help. Stacking the constant target on supervision, $\mathrm{Dice} + \mathrm{BCE}(y) + \mathrm{BCE}(0.196)$, reaches AP 0.4542 at 12 folds. This is our strongest U-Net within the Dice/BCE family ($+0.0113$ over Dice+BCE, 10/12 folds) and approaches the tuned reference of Lahrichi et al. (2025) (0.460); the repaired focal U-Net of Section 7 is higher still at 0.4551. The margin is modest (Wilcoxon $p = 0.043$ per fold, which does not survive a Holm correction; year-level t-interval $[-0.001, +0.024]$), and it is not specific to the constant. The recipe carries twice the BCE weight of Dice+BCE(y), and the weight-matched control Dice$+2\,\mathrm{BCE}(y)$ reaches 0.4525, so the recipe leads it by only $+0.0017$ ($[-0.009, +0.011]$, 6/12 folds). On four probe folds the constant had looked decisive (4/4); twelve folds refute that. What the recipe adds over Dice+BCE is more non-vanishing gradient, and a second supervised BCE term supplies it as well as a constant does. Removing Dice does as well: on the four probe folds BCE(y) alone reaches 0.457 against 0.440 for Dice+BCE, and adding the constant to it gives 0.460, within 0.001 of the recipe. Without Dice the constant adds $+0.004$ ($+0.037$, -0.015, -0.008, $+0.000$) against $+0.020$ with it, as the saturation account predicts. We make no state-of-the-art claim. In our experiments the largest accuracy movements on this benchmark came from the form of the loss ($0.348 \rightarrow 0.443$ from one added BCE term, 12/12 folds), and the prior we tested added nothing beyond it. That is a statement about one law at this resolution, not about physics priors, which earn their gains in many settings.

7 The Compact Learned Prior

The controls required an informative prior; this section describes the one we built. Our NCA maintains a per-cell state (channel 0 = fire logit, seeded from today's observed mask with a zero-initialised update, so the untrained model reproduces the persistence baseline), applies one shared 3×3 stencil-plus-MLP rule for $K=16$ steps with stochastic per-cell updates, and is conditioned on the 40 covariates at every step.

Stochastic models must be evaluated as expectations: a single rollout is a sample, not a prediction. Averaging N rollouts of one trained (Dice) model improves AP monotonically, $0.293 \rightarrow 0.336$ for $N=1\rightarrow16$ (12-fold means). Omitting this averaging is a $+0.043$ measurement error, larger than the gain of a two-member deep ensemble (Lakshminarayanan et al., 2017) ($+0.023$) obtained at twice the training cost. All NCA numbers here use $N=16$.

The published Dice baseline is under-tuned, so we compare at matched loss. With Dice+BCE on both sides, the NCA attains 90–93% of the U-Net on every evaluation fold (mean 91.5%) at 13,440 vs. 14.4M parameters. Under (repaired) focal loss (Lin et al., 2017), the best loss for both families at 12 folds, the 12-fold paired result is 0.4386 vs. 0.4551: the U-Net is better by 0.0165 (Wilcoxon $p = 0.005$; the NCA wins 2/12 folds), so the NCA delivers 96.4% of the U-Net at 0.09% of its parameters. Focal training also collapses rollout variance (ensembling gain $+0.008$ vs. $+0.043$ under Dice): a single pass reaches 0.4299, or 94.5% of the U-Net.

Accuracy is flat from 13,440 down to 2,336 parameters and collapses below (Appendix D). The cost is compute rather than parameters: $5.2\times$ the per-pass FLOPs of the U-Net and $9.5\times$ the training activation memory. Several natural extensions of the prior (output blending, joint training, using the NCA as a decoder, masking to the reachable set, low-data and multi-day variants) did not beat the plain U-Net; Appendix H reports them.

8 Extracting and Validating the Learned Rule

At 2,336 parameters the learned rule can be characterised exhaustively. The extraction protocol reads the learned response functions out of the network on real data (full protocol in Appendix K). At held-out front cells, the same population the diagnostic of Section 4 scores, we overwrite one driver's channel with a physical value while every other channel keeps its true value, and record the model's mean predicted ignition probability. Sweeping the value gives the learned response curve, a partial-dependence function estimated in-distribution. Rothermel's response is computed on the identical cells and grid, and the observed response by binning real front cells by their true driver value and measuring next-day ignition rates. All three curves are normalised at the median driver value. Figure 3 shows the result.

The probe population is the entire test year of each of the four probe folds (2018–2021; 1,055–3,337 full-resolution samples and 44k–662k front cells per year), which no checkpoint selection touched; every correlation is reported with a 95% cluster-bootstrap interval over samples, since the cells of one image share their weather and each correlation rests on only 5–9 binned points (Table A3, Appendix K). Terrain slope, which is static and hence unconfounded with weather regime and suppression, gives an unconfounded test: the learned response correlates with the observed one at $+0.79$ on average across the four test years ($+0.63$ to $+0.93$), against $+0.28$ for Rothermel ($+0.10$ to $+0.42$); the learned rule is closer on every year, with non-overlapping intervals on three of the four (2019, the smallest test year, is the exception). Rothermel also misses the observed shape: ignition rates saturate at 11–15° and fall by up to 20% by 30°, where Rothermel's response is still rising, while the learned response is flat or declining on three years and still rising on the fourth. Vegetation greenness (NDVI) is the most year-stable driver in the data, and the learned response tracks it at $+0.74$ to $+0.94$ (mean $+0.84$; every interval excludes zero). Rothermel's model cannot represent it even in principle: its vegetation enters only as static per-fuel-type constants, so live greenness has no term in the model. Specific humidity does not discriminate between the two: both respond in the observed direction, with overlapping intervals on every year. Wind and the energy-release component are the drivers with the largest role in fire physics, and they are the ones that cannot be validated here: their observed marginals change sign across years (Table A3). On 2021, for example, the observed wind response declines, and Rothermel, which predicts the same monotone increase every year, is anti-correlated with it (-0.96). These drivers are entangled with season timing and unrecorded suppression, which limits validation in general: at this resolution, the drivers that dominate fire physics cannot be checked against the data with confidence, whether the response comes from an imposed law or a learned one.

9 Limitations

The labels are VIIRS active-fire detections (about two overpasses per day, cloud gaps, known false alarms). Suppression is unrecorded, so every model learns fire behaviour under the observation period's suppression regime (cf. Zhenirovskyy et al., 2026). Meteorology enters at 4.6–28 km against 375 m labels, and wind direction passes through a lossy sin-only encoding upstream.

Our controls establish what the gains are not. The positive mechanism, constant-target regularisation of Dice saturation, is supported by the level dependence and the dose–response sweep, but we offer

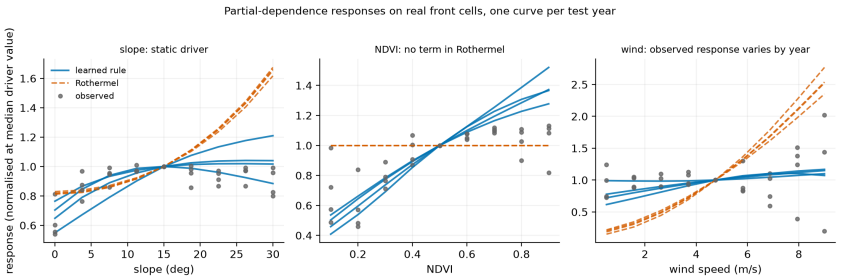


Figure 3: Partial-dependence responses at real front cells of each probe fold's test year (2018–2021): one blue curve and one set of grey points per year, normalised at the median driver value. The learned rule reproduces the observed slope saturation that Rothermel's monotone response misses, and responds to NDVI, which Rothermel's model cannot express (flat dashed line). The observed wind marginal is not year-stable, so no response can be validated against it.

no first-principles theory of the optimal level. The constant-target arms also pay for their ranking gain in calibration: expected calibration error moves from ≤ 0.003 (Dice-only) to 0.15–0.17 for const(0.196) and 0.05–0.06 for the full recipe, so a deployment needing probabilities rather than a ranking would have to recalibrate.

Three further limits are statistical and by design. All 12-fold arms are single-seed. Three seeds on the four probe folds give a spread of 0.006–0.024 AP, comparable to the $+0.011$ margin between the recipe and Dice+BCE, and one const(0.196) run on fold 3 collapsed to 0.043 against 0.309 and 0.311 for the other two seeds. The 12 folds contain only four test years, which is why we report per-year analysis throughout. The physics teacher distils a per-pixel rate of spread rather than a propagation model (Section 5). The diagnostic and controls are task-agnostic, but we demonstrate them on a single benchmark.

10 Conclusion

We reached the AP gain from distilling fire physics into a U-Net of $+\mathbf{0.104}$, on 12 of 12 folds of a faithfully reproduced public benchmark, a physics-informed gain by the usual standard. $\mathbf{0.196}$: the level at which the calibrated physics teacher sits; an explicit constant at that level reproduces the gain, within the fold-level equivalence margin of Section 5. $\mathbf{0.210}$ **vs.** $\mathbf{0.193}$: the prior's ranking quality against the class prior at the cells it was supposed to govern, measurable with one inference pass, before a single physics-informed run was trained. In this study the physics teacher contributed only a target level, and both checks indicate this before training.

The checks cost almost nothing and are not specific to fire: score the prior's informativeness before training, and require the gain to beat information-free teachers at the same target level after it. A constant at the label base rate and a bare auxiliary loss both outgained the one prior with measurable information about fire spread.

What survived the checks is a 2,336-parameter learned spread rule whose terrain response tracks the observed one more closely than Rothermel's on all four test years, and a gain of $+0.106$ AP that came from the form of the loss. Physics priors earn their gains in many settings; a two-arm experiment cannot show which. An imposed prior may be uninformative, and an observed gain may be a level effect; both should be checked.

AI use statement

The authors used generative AI to aid and polish writing, as well as to retrieve and discover existing literature and methods. The literature suggested by the assistant was read and verified by the authors before citation. We take responsibility for the final content of this work, including text or claims produced with the aid of generative AI.

Reproducibility statement

All data are public: WildfireSpreadTS (Zenodo 8006177, CC-BY-4.0), used unmodified through its authors' dataloader and official 12-fold protocol. The environment is pinned: Python 3.11, PyTorch 2.13/cu126, one NVIDIA A100 per run. Hyperparameters and the compatibility patches to the upstream code are listed in Appendix A. Code to reproduce the results is included in the submission.

Ethics statement

This work uses only public satellite-derived data with no personal information. We caution against operational use of any model in this paper for fire management: the labels embed unrecorded suppression activity, and the meteorology is far coarser than the targets. The negative results are intended to raise the evidential bar for deploying learned fire models.

References

Luca Bottero, Francesco Calisto, Giovanni Graziano, Valerio Pagliarino, Martino Scauda, Sara Tiengo, and Simone Azeglio. Physics-informed machine learning simulator for wildfire propagation. In *AAAI Spring Symposium on Machine Learning for Physical Sciences*, 2020. arXiv:2012.06825.

Joel Janek Dabrowski, Daniel Edward Pagendam, James Hilton, Conrad Sanderson, Daniel MacKinlay, Carolyn Huston, Andrew Bolt, and Petra Kuhnert. Bayesian physics informed neural networks for data assimilation and spatio-temporal modelling of wildfires. *Spatial Statistics*, 55, 2023. arXiv:2212.00970.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *ACM Conference on Recommender Systems (RecSys)*, 2019.

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Sebastian Gerard, Yu Zhao, and Josephine Sullivan. WildfireSpreadTS: A dataset of multi-modal time series for wildfire spread prediction. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. Dataset: `https://doi.org/10.5281/zenodo.8006177`.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Fantine Huot, R. Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022. arXiv:2112.02447.

John Kalkhof, Camila González, and Anirban Mukhopadhyay. Med-NCA: Robust and lightweight segmentation with neural cellular automata. In *Information Processing in Medical Imaging (IPMI)*, 2023. arXiv:2302.03473.

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

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

Saad Lahrichi, Jesse Bova, Matthew Johnson, and Jordan Malof. Improved wildfire spread prediction with time-series data and the WSTS+ benchmark. *arXiv preprint arXiv:2502.12003*, 2025.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Chris Lautenberger. Wildland fire modeling with an eulerian level set method and automated calibration. *Fire Safety Journal*, 62:289–298, 2013.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L. Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, 2021.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, 2016.

Alexander Mordvintsev, Ettore Randazzo, Eyvind Niklasson, and Michael Levin. Growing neural cellular automata. *Distill*, 2020. doi: 10.23915/distill.00023. `https://distill.pub/2020/growing-ca/`.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, 2020.

Cristobal Pais, Jaime Carrasco, David L. Martell, Andres Weintraub, and David L. Woodruff. Cell2Fire: A cell-based forest fire growth model to support strategic landscape management planning. *Frontiers in Forests and Global Change*, 4, 2021.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

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.

Alex D. Richardson, Tibor Antal, Richard A. Blythe, and Linus J. Schumacher. Learning spatiotemporal patterns with neural cellular automata. *PLOS Computational Biology*, 20(4):e1011589, 2024. arXiv:2310.14809.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Richard C. Rothermel. A mathematical model for predicting fire spread in wildland fuels. Technical Report Research Paper INT-115, USDA Forest Service, Intermountain Forest and Range Experiment Station, 1972.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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.

Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Maksym Zhenirovskyy, Ion Matei, Rohit Vuppala, Takuya Kurihana, and Hon Yung Wong. Neural-parameterized cellular automata for wildfire spread. *Ecological Informatics*, 97, 2026. doi: 10.1016/j.ecoinf.2026.103928.

A HYPERPARAMETERS AND TRAINING

Shared protocol. All models: AdamW (lr 10^{-3}, weight decay 0.01), 10,000 steps, batch 64 (32 for joint/integration models), random 128×128 crops, test on the lowest-validation-loss checkpoint; the positive-class weight is computed per fold from the training years' fire rate, following the benchmark code. Validation runs every epoch; in the low-data regime (10/25/50% of training fires, whole-fire subsets, seeded) every 10/4/2 epochs so that validation frequency per step matches the full-data runs. Seeds: `seed_everything`=0 throughout; robustness re-runs use seed 1.

U-Net. Res18-U-Net from segmentation-models-pytorch, 40 input channels, no pretrained weights (14,444,241 parameters). Losses: Dice (benchmark default); Dice+BCE; focal with $\alpha = 1/(1+w)$. The upstream implementation passes $\alpha = 1 - w$ with $w \approx 964$, an unnormalised class weight that yields $\alpha = -963$ and that modern torchvision rejects. We patch it, and results obtained with the unpatched path should be treated with care.

NCA. State d channels (channel 0 = fire logit; $d \in \{4, 8, 12, 16, 32\}$, default 16), fixed depthwise perception (identity + Sobel pair), per-cell MLP of two 1×1 convolutions (hidden width $\in \{8, 16, 32, 64, 128, 256\}$, default 128), zero-initialised output layer, stochastic update mask (rate 0.5), rollout depth $K \in \{4, 8, 16, 32\}$ (default 16), fire channel seeded from the observed mask as $6\,y_t - 3$. Evaluation is the mean over 16 rollouts with seeds $1000{+}r$.

Teacher distillation (Section 5). $\mathcal{L} = \mathrm{Dice} + \alpha\,\mathrm{BCE}(f(x), t(x))$ with $\alpha = 1$. An α ablation exists on fold 0 for three arms. NCA teacher: $\alpha = 0.3/1/3$ gives AP $0.526/0.542/0.561$; constant 0.196: $0.522/0.572/0.449$; Rothermel: $0.518/0.585/0.586$ (Dice-only U-Net 0.481 on this fold). We do not read a shape into these curves. On fold 0 the calibrated Rothermel teacher spans $[0.196, 0.198]$ with standard deviation 0.0003, and is therefore the constant arm to within 0.002, yet the two differ by 0.137 AP at $\alpha = 3$. No difference in target can explain that gap, which is consistent with the seed spread we measure separately (range 0.27 for this arm on fold 3 across three seeds; see Limitations). With one run per cell, the ablation supports only the weak statement that $\alpha = 1$ is not a poor choice for any of the three teachers. Rothermel teacher: two-parameter logistic calibration $\sigma(w\,R + b)$ fit per fold on training front cells; $w = 0.0035$ on fold 0 and 0.045–0.25 on the other folds, $\sigma(b) = 0.196$–0.207. NCA teacher: frozen fold-matched checkpoint, 2-rollout mean (an approximation to the 16-rollout expectation used everywhere else, chosen to keep the training step inside the dataloader floor); the NCA teacher was not calibrated, unlike the Rothermel teacher, an asymmetry noted in Section 5. Constant teachers as stated; shuffled teacher permutes the NCA teacher's pixels per image.

Rothermel implementation. Full 1972 surface-fire system with MODIS land cover mapped to representative Anderson fuel models, dead-fuel moisture from the Simard relation, and a single global scale calibrated on training data (predicted head-fire rates exceed observed daily displacement by $\sim25\times$; median observed 0.8 km/day).

B STATISTICAL METHODOLOGY

Comparisons are paired per fold. We report the paired Wilcoxon signed-rank p (two-sided; note its floor is 0.0005 at $n{=}12$ and 0.125 at $n{=}4$) and a 20,000-resample fold-bootstrap 95% CI of the mean difference. Rows with $n < 6$ are flagged wherever shown; their evidence rests on the CI and per-fold consistency.

Because the 12 folds are 4 test years $\times$ 3 training pairs and folds sharing a test year agree to ≤ 0.019 AP, the per-fold differences are clustered and the per-fold p overstates the evidence. Every comparison is therefore also analysed at year level: the folds of each test year are averaged first, giving four paired differences, reported with the mean, a t-interval ($n = 4$, so the sign count cannot reach significance on its own) and the number of years won. Equivalence claims use two one-sided t-tests (TOST) against a ± 0.02 AP margin at both levels. Holm step-down adjustment across all comparisons in `stats.py` is reported next to each raw p in the released table. Table A1 gives the year-level view of the headline pairs. One further arm is not in Table 2: Dice+BCE(y)+BCE(NCA), run on the four probe folds, scores -0.014 against plain Dice+BCE ($[-0.031, -0.002]$, 1/4 folds).

Table A1: Year-level view of the headline comparisons: paired difference in test AP per test year (folds averaged within year), mean, 95% t-interval and years won. Emitted by `scripts/stats.py` from the run logs.

Comparison	2018	2019	2020	2021	Mean	95% t-CI	Wins
Dice+BCE vs Dice-only	$+0.099$	$+0.076$	$+0.098$	$+0.106$	$+0.095$	$[+0.074, +0.115]$	4/4
Rothermel teacher vs Dice-only	$+0.096$	$+0.088$	$+0.119$	$+0.115$	$+0.104$	$[+0.080, +0.128]$	4/4
constant 0.196 vs Dice-only	$+0.111$	$+0.090$	$+0.119$	$+0.087$	$+0.102$	$[+0.077, +0.126]$	4/4
NCA teacher vs Dice-only	$+0.065$	$+0.052$	$+0.059$	$+0.072$	$+0.062$	$[+0.048, +0.076]$	4/4
constant 0.196 vs Rothermel	$+0.015$	$+0.002$	-0.000	-0.028	-0.003	$[-0.032, +0.026]$	2/4
NCA teacher vs shuffled NCA	$+0.036$	$+0.034$	$+0.026$	$+0.028$	$+0.031$	$[+0.024, +0.039]$	4/4
NCA teacher vs constant ≈ 0.002	$+0.024$	$+0.008$	$+0.007$	$+0.010$	$+0.012$	$[-0.001, +0.025]$	4/4
Dice+BCE(y)+BCE(0.196) vs Dice+BCE	$+0.018$	$+0.013$	$+0.014$	$+0.000$	$+0.011$	$[-0.001, +0.024]$	4/4
Dice+2 BCE(y) vs Dice+BCE	$+0.012$	$+0.005$	$+0.014$	$+0.008$	$+0.010$	$[+0.003, +0.016]$	4/4
Dice+BCE(y)+BCE(0.196) vs Dice+2 BCE(y)	$+0.007$	$+0.008$	-0.000	-0.008	$+0.002$	$[-0.010, +0.013]$	2/4
NCA vs U-Net, focal-matched	-0.015	-0.019	-0.023	-0.009	-0.016	$[-0.026, -0.007]$	0/4

C COMPUTE

The full study comprises $\sim$350 SLURM jobs totalling $\sim$700 A100-hours (40 GB, single-GPU jobs). One U-Net configuration trains in 1.5–2.6 h; one NCA configuration in 1.9–2.6 h plus 0.5 h for 16-rollout evaluation; the diagnostic of Section 4 costs one inference pass ($\sim$minutes).

D CAPACITY SWEEP OF THE LEARNED PRIOR

Capacity scales with iteration depth: widening the rule $3\times$ (and doubling its budget) changes nothing (-0.001), while deepening the rollout $K{=}4{\rightarrow}16$ gains $+0.045$ with no added parameters. Scaling down at fixed K (Figure A1), accuracy is flat from 13,440 to 2,336 parameters (91% of the Dice-matched U-Net at $6{,}183\times$ fewer) and collapses below, with seed variance ± 0.04 under 2k parameters. The compute cost is substantial: $5.2\times$ the per-pass FLOPs of the U-Net and $9.5\times$ the training activation memory, so parameters are traded for compute, though the focal variant's single-pass sufficiency removes the former $16\times$ inference multiplier.

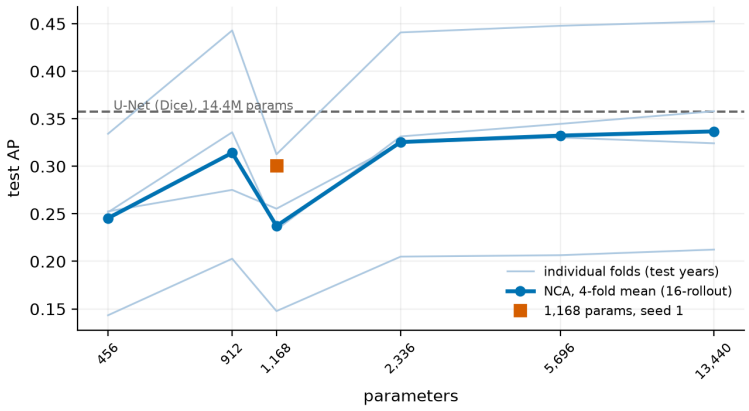


Figure A1: Capacity sweep at fixed rollout depth $K{=}16$ (16-rollout evaluation; faint points are individual folds). Accuracy is flat from 13,440 down to 2,336 parameters and collapses below; the dip at 1,168 is seed instability (orange square/circles: seed-1 rerun). The dashed line is the 14.4M-parameter Dice-trained U-Net on the same folds.

E MEASURED LOGIT SATURATION

The saturation this account rests on is measured, not assumed (folds 0 and 3, full test years). The Dice-only U-Net pushes 99.8% and 100% of pixels past $|z| > 8$ and 91% and 99.5% of negatives below $p = 10^{-6}$; adding BCE(y) barely moves the saturated fraction (99.0%, 99.4%) but lifts most negatives off the floor (3.0%, 4.6% below 10^{-6}). Every arm with a constant target removes the saturation outright: no pixel exceeds $|z| > 8$, the negative-logit standard deviation falls from 2.9–4.7 to 0.08–0.24, and negatives sit at a mean probability of 0.15–0.17 for const(0.196), 0.05–0.06 when BCE(y) is also present. The Rothermel arm matches the constant arm on all four statistics, as the level account predicts.

F THE PHYSICS ARM'S EDGE DOES NOT TRACK THE PRIOR'S INFORMATION

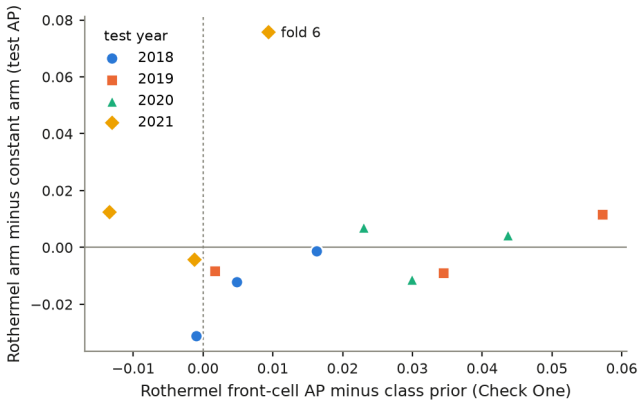


Figure A2: One point per fold. Horizontal: how informative Check One finds the Rothermel prior (front-cell AP minus the class prior). Vertical: test AP of the Rothermel-teacher arm minus the constant-0.196 arm. If the student used the prior's information the points would rise to the right; Spearman $+0.12$ ($p = 0.71$; $+0.14$ without fold 6, whose constant arm had a weak run). Emitted by `scripts/info_tracking.py`.

G THE CONSTANT 0.196 IS NOT TUNED TO THE TEST SET

Because the gain depends on one specific number, we checked whether that number could carry test-set information. Four independent facts rule this out. (i) Provenance: 0.196 was not chosen by us or by any search; it is the output of a logistic calibration fit on training front cells (the fold-0 value; the per-fold calibrations give 0.196–0.207, and we use 0.196 on every fold), and the value is computable before any test data exists (it equals the front-cell base ignition rate of the training years). (ii) Generalisation: a scalar cannot encode test-year information, and the gain is positive on 12 of 12 folds, each testing a year the constant's derivation never saw; on a benchmark where the test year dominates variance, a coincidental constant would be unlikely to survive four different test years. (iii) Dose–response (Figure A3): sweeping c over $[0.02, 0.8]$ shows an operating range wider than an order of magnitude; on the easiest year the response saturates above 0.196, so the calibration-derived value is adequate but not optimal. (iv) Seed robustness: two independent seeds reproduce the c=0.196 arm on both probe folds, spanning ≤ 0.007 AP. The remaining open question is theoretical: we have no first-principles account of the optimal level, only the observation that any level near or above the front-cell base ignition rate works.

H WHAT DID NOT WORK

These negative results close off several natural extensions. Output-blending the NCA with the U-Net beats a same-architecture ensemble control only when the NCA member is evaluated as an expectation ($+0.011$, 4/4 folds); with single rollouts the effect is invisible, repeating the measurement error of Section 7. Joint training, NCA-as-decoder integration (which damps the stochasticity that rollout averaging feeds on), and masking U-Net outputs to the physically reachable set ($+0.0002$: the U-Net has already learned locality) all fail to beat the plain U-Net. There is no low-data crossover: both model families are nearly flat in training-set size from 10% to 100% of fires; performance appears limited by label noise and distribution shift, since more data helps neither family. The NCA's apparent multi-day collapse was protocol asymmetry; at parity (daily re-anchoring for both) it holds 81–90% of the U-Net at 1–3 days.

I DOSE–RESPONSE OF THE CONSTANT-TARGET LEVEL

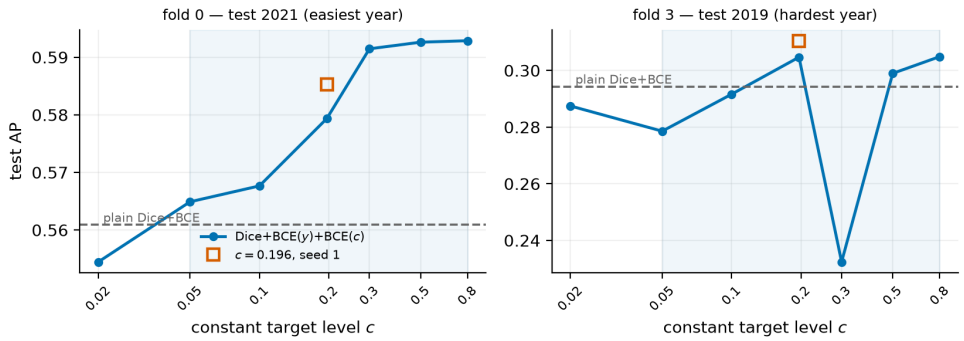


Figure A3: Dose–response of the constant-target level c in $\mathrm{Dice}+\mathrm{BCE}(y)+\mathrm{BCE}(c)$, on the easiest (fold 0) and hardest (fold 3) test years. Dashed: plain Dice+BCE. The effect is robust across more than an order of magnitude; the orange square is an independent-seed repeat of c=0.196. The single fold-3 dip at c=0.3 is within the seed variance we measure on this fold ($\pm$0.01–0.05 across arms; three seeds of the recipe arm span only 0.006 AP here, while one const(0.196) run collapsed to 0.043) and is not replicated on fold 0.

J PER-FOLD RESULTS

Against the Dice-only U-Net the NCA reaches 96.5% of the U-Net's AP at 12 folds (-0.012, $p = 0.052$); this Dice-matched ratio and the focal-matched 96.4% of Section 7 are different comparisons

that happen to agree. One cautionary detail, which our own protocol caught: on fold 0 alone the focal NCA beats both U-Nets (0.571 vs. 0.561), a single-fold "parity" that the 12-fold paired test refutes, another consequence of the test-year variance described in Section 3.

Table A2: Per-fold test AP for the principal configurations. Every value is emitted by `scripts/make_appendix_table.py` from the stored run logs; NCA focal^{16} is the 16-rollout mean. Parenthesised mean: fewer than 12 folds.

Fold	U-Net (Dice)	Dice+BCE	Roth. tchr	const .196	+const .196	const .004	Dice+2BCE	U-Net focal	NCA focal^{16}
0	0.481	0.561	0.585	0.572	0.579	0.513	0.574	0.561	0.571
1	0.376	0.445	0.458	0.469	0.478	0.399	—	0.474	0.462
2	0.477	0.588	0.599	0.603	0.608	0.539	—	0.584	0.576
3	0.221	0.294	0.301	0.309	0.305	0.266	—	0.289	0.279
4	0.353	0.480	0.494	0.490	0.480	0.427	—	0.499	0.471
5	0.208	0.299	0.308	0.296	0.299	0.252	—	0.297	0.299
6	0.463	0.591	0.583	0.507	0.552	0.557	—	0.568	0.541
7	0.352	0.459	0.444	0.456	0.476	0.373	—	0.462	0.448
8	0.340	0.437	0.473	0.466	0.446	0.398	—	0.487	0.459
9	0.356	0.415	0.435	0.467	0.448	0.379	—	0.468	0.445
10	0.216	0.279	0.300	0.309	0.306	0.258	—	0.304	0.254
11	0.338	0.467	0.454	0.455	0.472	0.416	—	0.469	0.459
mean	0.348	0.443	0.453	0.450	0.454	0.398	(0.574)	0.455	0.439

K LEARNED RULE: AGREEMENT WITH THE OBSERVED RESPONSE, PER TEST YEAR

The extraction protocol, which reads the learned response functions out of the network, has four steps, all on real data:

1. *Population.* Collect front cells (unburnt, adjacent to observed fire) from held-out samples; these are the cells where a spread law acts, and the same population the diagnostic of Section 4 scores.

2. *Interventional sweep.* For one driver at a time, overwrite its channel with a physical value v (de-standardised into the network's input scale) while every other channel keeps its true value; record the model's mean predicted ignition probability over the front cells, averaged over stochastic rollouts. Repeating over a grid of v yields the learned response curve, a partial-dependence function estimated in-distribution. A synthetic uniform-landscape probe, the seemingly natural alternative, produced artefacts (near-zero spread at realistic winds, domain-flooding at extreme ones) because uniform covariates are far outside the training distribution; we report this as a caution for rule-extraction studies generally.

3. *References.* Compute Rothermel's response on the identical cells and grid, and the observed response by binning real front cells by their true driver value and measuring next-day ignition rates. All three curves are normalised at the median driver value, making shapes comparable across units.

4. *Distillation.* Fit a cubic per driver to the learned curve ($R^2 \approx 1.00$ throughout): the extracted physics is a one-line equation per driver, and the validation criterion is agreement with the observed curve rather than with our expectations.

Figure 3 shows the resulting curves, and Table A3 the agreement with the observed response per test year.

Table A3: Agreement of the learned rule's and Rothermel's partial-dependence response with the observed front-cell ignition rate on each probe fold's *test* year: Pearson correlation over the 5–9 driver bins, with a 95% cluster-bootstrap interval over test samples (images). Rothermel has no NDVI or ERC term, so its response is flat and the correlation undefined. Emitted by `scripts/make_rule_table.py` from the stored extraction outputs.

Driver	Response	2018	2019	2020	2021
slope	learned	$+0.74$ [.60, .84]	$+0.63$ [.07, .89]	$+0.93$ [.82, .97]	$+0.85$ [.62, .93]
	Rothermel	$+0.10$ [$-.10$, .30]	$+0.26$ [$-.37$, .77]	$+0.42$ [.18, .60]	$+0.34$ [$-.07$, .59]
NDVI	learned	$+0.94$ [.90, .96]	$+0.75$ [.22, .96]	$+0.92$ [.74, .97]	$+0.74$ [.56, .83]
	Rothermel	—	—	—	—
specific humidity	learned	$+0.94$ [.66, .97]	$+0.98$ [.80, .99]	$+0.95$ [.88, .98]	$+0.82$ [.15, .94]
	Rothermel	$+0.81$ [.52, .88]	$+0.97$ [.79, .99]	$+0.86$ [.78, .93]	$+0.70$ [.21, .91]
wind speed	learned	$+0.70$ [.15, .83]	$+0.23$ [$-.57$, .69]	$+0.91$ [.66, .95]	-0.89 [$-.93$, $-.61$]
	Rothermel	$+0.80$ [.17, .89]	$+0.28$ [$-.70$, .66]	$+0.88$ [.68, .94]	-0.96 [$-.98$, $-.57$]
ERC	learned	-0.74 [$-.82$, $-.60$]	$+0.74$ [.44, .87]	$+0.69$ [.54, .82]	-0.29 [$-.56$, .13]
	Rothermel	—	—	—	—